AUGMENTED LAGRANGIAN WITHOUT ALTERNATING DIRECTIONS: PRACTICAL ALGORITHMS FOR INVERSE PROBLEMS IN IMAGING

Rahul Mourya [*], *Loïc Denis, Jean-Marie Becker*

Université Jean Monnet,
CNRS, UMR 5516, Laboratoire Hubert Curien,
F-42000, Saint-Étienne, France

Eric Thiébaut

Université de Lyon 1,
CNRS, UMR 5574, Observatoire de Lyon,
F-69561, Saint Genis Laval Cedex, France

ABSTRACT

Several problems in signal processing and machine learning can be casted as optimization problems. In many cases, they are of large-scale, nonlinear, have constraints, and may be nonsmooth in the unknown parameters. There exists plethora of fast algorithms for smooth convex optimization, but these algorithms are not readily applicable to nonsmooth problems, which has led to a considerable amount of research in this direction. In this paper, we propose a general algorithm for nonsmooth bound-constrained convex optimization problems. Our algorithm is instance of the so-called augmented Lagrangian, for which theoretical convergence is well established for convex problems. The proposed algorithm is a blend of superlinearly convergent limited memory quasi-Newton method, and proximal projection operator. The initial promising numerical results for total-variation based image deblurring show that they are as fast as the best existing algorithms in the same class, but with fewer and less sensitive tuning parameters, which makes a huge difference in practice.

***Index Terms*—** Constrained convex optimization, nonsmooth optimization, hierarchical optimization, ADMM, proximal operator, deblurring, total variation.

1. INTRODUCTION

Many problems in signal processing and machine learning can be stated follows:

$$\mathbf{x}^* := \arg\min_{\mathbf{x}} F(\mathbf{x}) \quad \text{subject to} \quad \mathbf{x} \in \Omega, \tag{1}$$

where $F(\mathbf{x}) = \{f(\mathbf{x}) + r(\mathbf{x})\}$; $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, and convex; $r : \mathbb{R}^n \to \mathbb{R}$ is continuous and convex but not necessarily differentiable everywhere, and Ω is a bound constraint set. Often, f corresponds to a loss, and r to a regularizer. This class of problems, in general, do not have close-form solutions, and rely on iterative method. There exists numerous well established algorithms [1] (LBFGS-B), [2] (VMLM-B), [3] (BLMVM), [4] (ASA-CG), [5] (minConf-TMP), based on limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) method [6] with superlinear convergence rate for the problem (1) when $\forall \mathbf{x}, r(\mathbf{x}) = 0$, but these techniques do not work off-the-shelf when $\forall \mathbf{x}, r(\mathbf{x}) \neq 0$. Some efforts have been made in an ad-hoc manner to apply LBFGS methods directly to nonsmooth convex problems that are differentiable almost everywhere, and convergence to the optimum has been noted [7] in the cases where no nonsmooth point is encountered, otherwise [8, 9, 10] report catastrophic failures (convergence to a non-optimum) of the direct methods. The traditional algorithms for nonsmooth optimization are based on a stabilization steepest descent by exploiting gradient or subgradient information evaluated at multiple points, which is the essential idea behind subgradient methods [11, 12], the bundle methods [9, 13], and the gradient sampling algorithm [14, 15]. These algorithms are successful for nonsmooth problems with sublinear rate of convergence.

Another class of algorithms for solving the problem (1) use the general idea of variable splitting: introduction of auxiliary variables in order to decouple minimizations of f and of r. The proximal forward-backward iterative scheme introduced in [16] and [17] is a notable representative algorithm of this class; see the recent survey [18] and the references therein for very general convergence results of proximal forward-backward algorithms under various conditions and settings. relevant to problem (1). Unlike the algorithms using second-order information (quasi-Newton methods), the proximal forward-backward algorithms use only first order information, and are slower having only sublinear convergence rate, but enjoy global convergence guarantees for nonsmooth convex problems. The convergence rates of these algorithms have been further improved in [19, 20] by using the idea of Nesterov [21] developed in 1983, which suggest to use the information from the previous iterations in a smart way at each new iteration.

The two instances of the general proximal forward-backward algorithm are Alternating Minimization Algorithm (AMA) [22], and closely related Alternating Direction Method of Multipliers (ADMM) [23] whose developments

[*]The author is supported by a PhD grant funded by the ARC6, Région Rhône-Alpes.

backs to 1970s. These algorithms use augmented Lagrangian (AL) terms to handle the constraints, and are closely related to algorithms such as dual decomposition, the method of multipliers, Douglas-Rachford splittings, Dykstra's alternating projections, Bregman iterative algorithm, and others. ADMM is a blend of decomposability of dual ascent and superior convergence properties of the method of multipliers. Because of its flexibility to in handling different types of objective functions/constraints, and its simplicity in implementation for distributed optimization, ADMM has gained popularity in both signal processing and machine learning communities since the last two decades. However, only sublinear rate of convergence have been achievable by the ADMM, and convergence speed is highly dependent upon the chosen penalty parameters. Optimal tuning of those parameters remains largely an open question. In this paper we propose an algorithm for solving nonsmooth bound-constrained convex optimization problem, which uses variable splitting and AL to handle the nonsmooth part, and the robust quasi-Netwon method to handle the bound-constraint and smooth part. The proposed algorithm is as fast as the other algorithms in the same class with fewer number of tuning parameters, and is immune to large range variation in the tuning parameter.

2. PROPOSED ALGORITHM

The problem (1) without the bound constraint can be written equivalently after variable splitting as:

$$\arg\min_{\mathbf{x},\,\mathbf{z}} \{\ f(\mathbf{x}) + r(\mathbf{z})\ \} \quad \text{s.t.} \quad \mathbf{x} - \mathbf{z} = 0. \tag{2}$$

The Augmented Lagrangian of problem (2) can be written:

$$\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}, \mathbf{u}) = f(\mathbf{x}) + r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z} + \mathbf{u}\|_2^2, \tag{3}$$

where $\mathbf{u}$ is *dual scaled variable*, and $\rho > 0$ is a augmented penalty parameter.

The method of multipliers solves the problem (2) with bound constraint, ($\mathbf{x} \in \Omega$), by the following iterations (k being the iteration counter):

$$\{\mathbf{x}^{(k+1)}, \mathbf{z}^{(k+1)}\} := \arg\min_{\mathbf{z},\mathbf{x}\in\Omega} \{\ f(\mathbf{x}) + r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z} + \mathbf{u}^{(k)}\|_2^2\ \} \tag{4}$$

$$\mathbf{u}^{(k+1)} := \mathbf{u}^{(k)} + \mathbf{x}^{(k+1)} - \mathbf{z}^{(k+1)} \tag{5}$$

Rather than jointly minimizing with respect to $\mathbf{x}$ and $\mathbf{z}$ variables, ADMM solves the same problem by following iterations:

$$\mathbf{x}^{(k+1)} := \arg\min_{\mathbf{x}\in\Omega} \{\ f(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z}^{(k)} + \mathbf{u}^{(k)}\|_2^2\ \} \tag{6}$$

$$\mathbf{z}^{(k+1)} := \arg\min_{\mathbf{z}} \{\ r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{x}^{(k+1)} - \mathbf{z} + \mathbf{u}^{(k)}\|_2^2\ \} \tag{7}$$

$$\mathbf{u}^{(k+1)} := \mathbf{u}^{(k)} + \mathbf{x}^{(k+1)} - \mathbf{z}^{(k+1)} \tag{8}$$

The above two methods for solving the problem (2) are very similar. ADMM is viewed as a version of the method of multipliers with a single *Gauss-Seidel* iteration over $\mathbf{x}$ and $\mathbf{z}$. The real advantage of these methods is their convergence to the solution under very general conditions [23]: (i) the extended-real-valued functions f and r in problem (2) are closed, proper, and convex; and (ii) the unaugmented Lagrangian $\mathcal{L}_0$ has a saddle point. Moreover, the two methods converge even when $\mathbf{x}$-, $\mathbf{z}$-updates are not carried out exactly, provided that the errors between the approximate and exact solution at each iteration are summable [24].

2.1. Motivation

In general the global convergence rate of ADMM is limited to sublinear rate, however, recently numerous algorithms [25, 26, 27] based on ADMM for solving the instances of the general problem (1) in applications related to image reconstruction and other linear inverse problems have proven to be more efficient, and converges faster in comparison to many state-of-the-art algorithms [28, 19, 20]. The convergence speed of ADMM is highly dependent upon the penalty parameters, and only optimally tunned parameters result in high convergence speed. The optimal tunning of the penalty parameters is still an open problem. It has also been observed that the convergence speed is higher when separate penalty parameters are used for each augmented term introduced in the ADMM. The experimental results in [27] clearly demonstrate that the ADMM with i). sufficient number of variable splittings (so that each variable updates can be carried out in closed-form), and ii). optimally tunned penalty parameters, converges faster than the other variants of ADMM. Following this suggestion, a nonsmooth convex problem such as image deblurring problem with total variation(TV) regularization, the one considered in section (3), requires multiple variable splittings, which requires to tune multiple parameters hindering the convergence speed. In this paper, we consider going back to the method of multipliers and avoiding as many splitting as possible so as to reduce the number of penalty parameters. We show that by using a hierarchical optimization approach, an efficient algorithm is obtained.

2.2. Derivation of the algorithm

Going back to the method of multipliers (4)-(5), the joint minimization problem can be formulated in a hierarchical way:

$$\mathbf{x}^* := \arg\min_{\mathbf{x}\in\Omega} \quad f(\mathbf{x}) + r(\mathbf{z}^*(\mathbf{x})) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z}^*(\mathbf{x}) + \mathbf{u}\|_2^2$$

$$\text{with } \mathbf{z}^*(\mathbf{x}) = \arg\min_{\mathbf{z}} \quad r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z} + \mathbf{u}\|_2^2 \tag{9}$$

The inner optimization problem is often known in closed form. We use a quasi-Newton method to compute the solution of the outer minimization problem.

Proposition 1. *The gradient of the partially optimized augmented Lagrangian with respect to* $\mathbf{x}$ *is given by:*

$$\nabla_{\mathbf{x}}\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}^*(\mathbf{x}), \mathbf{u}) = \nabla f(\mathbf{x}) + \rho(\mathbf{x} - \mathbf{z}^*(\mathbf{x}) + \mathbf{u})\ . \tag{10}$$

Proof. When r is a smooth function, this follows directly from the chain rule formula since the gradient of augmented

Lagrangian with respect to $\mathbf{z}$ is zero at the optimal value $\mathbf{z}^*$. In the non-smooth case, the augmented Lagrangian is still differentiable by virtue of the augmented term that smooths r [29]. In order to prove our proposition in the non-smooth case, we use a classical characterization of the gradient of a convex function, namely that the function lies above its gradient.

As $\mathbf{z}^*(\mathbf{x})$ is the solution of the inner minimization:

$$\forall \mathbf{z},\ \forall \mathbf{x},\ r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{x}-\mathbf{z}+\mathbf{u}\|_2^2 \geq r(\mathbf{z}^*(\mathbf{x})) + \frac{\rho}{2}\|\mathbf{x}-\mathbf{z}^*(\mathbf{x})+\mathbf{u}\|_2^2 . \quad (11)$$

For a fixed $\mathbf{z}$, the gradient of the augmented Lagrangian $\mathcal{L}_\rho(\mathbf{x},\mathbf{z},\mathbf{u})$ with respect to $\mathbf{x}$ is:

$$\nabla_{\mathbf{x}}\mathcal{L}_\rho(\mathbf{x},\mathbf{z},\mathbf{u}) = \nabla f(\mathbf{x}) + \rho(\mathbf{x}-\mathbf{z}+\mathbf{u}) . \quad (12)$$

By convexity of f, we have:

$$\forall \mathbf{x},\ \forall \mathbf{z},\ \mathcal{L}_\rho(\mathbf{x},\mathbf{z},\mathbf{u}) \geq \mathcal{L}_\rho(\hat{\mathbf{x}},\mathbf{z},\mathbf{u}) + [\nabla f(\hat{\mathbf{x}}) + \rho(\hat{\mathbf{x}}-\mathbf{z}+\mathbf{u})]^t(\mathbf{x}-\hat{\mathbf{x}}) . \quad (13)$$

Equation (11) applied to $\mathbf{z}^*(\mathbf{x})$ and $\hat{\mathbf{x}}$ and equation (13) applied to $\mathbf{z}^*(\mathbf{x})$ lead to:

$$\forall \mathbf{x},\ \mathcal{L}_\rho(\mathbf{x},\mathbf{z}^*(\mathbf{x}),\mathbf{u}) \geq \mathcal{L}_\rho(\hat{\mathbf{x}},\mathbf{z}^*(\hat{\mathbf{x}}),\mathbf{u}) + [\nabla f(\hat{\mathbf{x}}) + \rho(\hat{\mathbf{x}}-\mathbf{z}^*(\mathbf{x})+\mathbf{u})]^t(\mathbf{x}-\hat{\mathbf{x}}) . \quad (14)$$

This inequality is close to the gradient inequality sought, except that $\mathbf{z}^*(\mathbf{x})$ appears instead of $\mathbf{z}^*(\hat{\mathbf{x}})$ in the second line. We thus need to prove that $(\mathbf{z}^*(\hat{\mathbf{x}})-\mathbf{z}^*(\mathbf{x}))^t(\mathbf{x}-\hat{\mathbf{x}}) \geq 0$ for all $\mathbf{x}$. By expanding the squared norm in equation (11), we can show that

$$\forall \mathbf{z},\ \forall \mathbf{x},\ r(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{z}\|_2^2 - \rho\mathbf{z}^t(\mathbf{x}+\mathbf{u}) \geq r(\mathbf{z}^*(\mathbf{x})) + \frac{\rho}{2}\|\mathbf{z}^*(\mathbf{x})\|_2^2 - \rho\mathbf{z}^*(\mathbf{x})^t(\mathbf{x}+\mathbf{u}) . \quad (15)$$

Combining equation (15) applied to $(\mathbf{z} := \mathbf{z}^*(\mathbf{x}),\ \mathbf{x} := \hat{\mathbf{x}})$ and to $(\mathbf{z} := \mathbf{z}^*(\hat{\mathbf{x}}),\ \mathbf{x})$ gives the desired result. In conclusion, the following inequality holds:

$$\forall \mathbf{x},\ \mathcal{L}_\rho(\mathbf{x},\mathbf{z}^*(\mathbf{x}),\mathbf{u}) \geq \mathcal{L}_\rho(\hat{\mathbf{x}},\mathbf{z}^*(\hat{\mathbf{x}}),\mathbf{u}) + [\nabla f(\hat{\mathbf{x}}) + \rho(\hat{\mathbf{x}}-\mathbf{z}^*(\hat{\mathbf{x}})+\mathbf{u})]^t(\mathbf{x}-\hat{\mathbf{x}}) , \quad (16)$$

which proves the proposition in Eq.(10). □

Proposition 2. *The gradient difference is independent of the value of dual variable* $\mathbf{u}$*:*

$$\nabla_{\mathbf{x}}\mathcal{L}_\rho\big(\mathbf{x}^{(2)},\mathbf{z}^*(\mathbf{x}^{(2)}),\mathbf{u}\big) - \nabla_{\mathbf{x}}\mathcal{L}_\rho\big(\mathbf{x}^{(1)},\mathbf{z}^*(\mathbf{x}^{(1)}),\mathbf{u}\big) = \nabla f(\mathbf{x}^{(2)}) - \nabla f(\mathbf{x}^{(1)}) + \rho\big[\mathbf{x}^{(2)}-\mathbf{x}^{(1)}+\mathbf{z}^*(\mathbf{x}^{(1)})-\mathbf{z}^*(\mathbf{x}^{(2)})\big]$$

Using proposition 1, any smooth optimization method based solely on cost function and gradient evaluations can be used to perform joint minimization over $\mathbf{x}$ and $\mathbf{z}$. If the method uses gradient differences to collect second order information (e.g., quasi-Newton methods), the memorized previous steps can be used even after the dual variables have been updated, since gradient differences are not affected by dual updates, see proposition 2. The applicability of smooth optimization methods is thus extended to non-smooth problems typically encountered in image reconstruction.

Table 1: Relative difference between the cost function after 1500 FFT evaluations and at the solution (in %)

method	$[F(x^{(1500)}) - F(x^*)]/F(x^*)$		
ADMM 3x		.35	
ADMM 1x	2.6 (30 it)	2.7 (60 it	2.2 (100 it)
proposed(rst)	2.2 (30 it)	.55 (60 it)	.20 (100 it)
proposed	.06 (30 it)	.04 (60 it)	.038 (100 it)

3. ILLUSTRATION ON IMAGE DEBLURRING

We evaluate the efficiency of our algorithm on an image deblurring problem. We consider maximum a posteriori deblurring with TV regularization and positivity constraints:

$$\mathbf{x}^* := \arg\min_{\mathbf{x}\geq 0}\ \frac{1}{2}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_{\mathbf{W}}^2 + \lambda\,\mathrm{TV}(\mathbf{x}) , \quad (17)$$

where $\mathbf{y} \in \mathbb{R}^n$ is the observed (blurry and noisy) image, $\mathbf{H} \in \mathbb{R}^{n\times n}$ is the blurring operator (discrete convolution matrix), $\mathbf{W} \in \mathbb{R}^{n\times n}$ is the inverse of noise covariance matrix (diagonal in our case), $\lambda > 0$ is tuned to reach good compromise between smoothness of the solution and data fitting. $\mathrm{TV}(\mathbf{x})$ is the isotropic-TV on the unknown image, $\mathbf{x}$, defined as: $\mathrm{TV}(\mathbf{x}) = \sum_{i=1}^n \|(\nabla\mathbf{x})_i\|_2$, where $\nabla = [\nabla_x^T\ \nabla_y^T]^T$ is first-order finite difference operator in two dimensions (see e.g., [30]). The $\|\boldsymbol{v}\|_{\mathbf{W}}^2$ is defined as $\boldsymbol{v}^t\mathbf{W}\boldsymbol{v}, \forall \boldsymbol{v}$. As done in [27], to correctly handle borders, we reconstruct $\mathbf{x}$ that is larger than the available blurred image. Pixel values outside the field-of-view are given zero weight by matrix $\mathbf{W}$ (i.e., if pixel k is not seen, then $W_{k,k} = 0$). We apply the deblurring model (17) on a portion of blurry and noisy Lena image shown in Figure 3. Note the zero-padding on the border of the observed image to handle it correctly.

Previous research [27] has shown the superiority of ADMM with multiple splittings over other state-of-the-art methods. We compare the convergence speed of (i) ADMM (hereafter method **ADMM 3x**) with three splittings: $\mathbf{y}-\mathbf{H}\mathbf{x} = \xi;\ \nabla\mathbf{x} = \omega;\ \mathbf{z} = \mathbf{x}$, and the resulting AL:

$$\frac{1}{2}\|\xi\|_{\mathbf{W}}^2 + \frac{\rho}{2}\|\mathbf{y}-\mathbf{H}\mathbf{x}-\xi+\mathbf{u}_1\|_2^2 + g(\mathbf{z}) + \frac{\nu}{2}\|\mathbf{x}-\mathbf{z}+\mathbf{u}_2\|_2^2 + \lambda\sum_{i=1}^n\|\omega_i\|_2 + \frac{\gamma}{2}\sum_{i=1}^n\|(\nabla\mathbf{x}-\omega+\mathbf{u}_3)_i\|_2^2 \quad (18)$$

that makes possible to solve each sub-problems in closed-form. g is an indicator function for the constraint $\mathbf{z} \geq 0$. The three penalty parameters are tuned after many experimental trails; (ii) ADMM (hereafter **ADMM 1x**) with a single splitting: $\nabla\mathbf{x} = \omega$ with similar AL as in (18). The $\mathbf{x}$-update is solved approximately by a quasi-Newton method (VMLM-B) and ω-update in closed form by the proximal operator (soft-thresholding); (iii) our proposed algorithm (hereafter **proposed**) with or without restarting the quasi-Newton method (VMLM-B) with the same variable splitting and AL as in **ADMM 1x**. All these methods are implemented

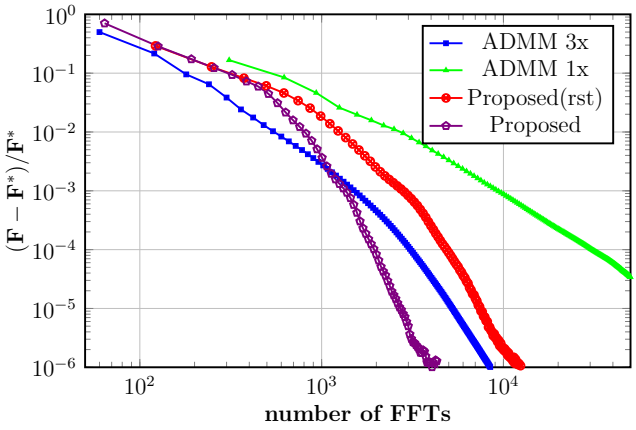

Fig. 1: Convergence speed comparison of the optimization methods with possible best penalty parameters (ρ, γ) for each.

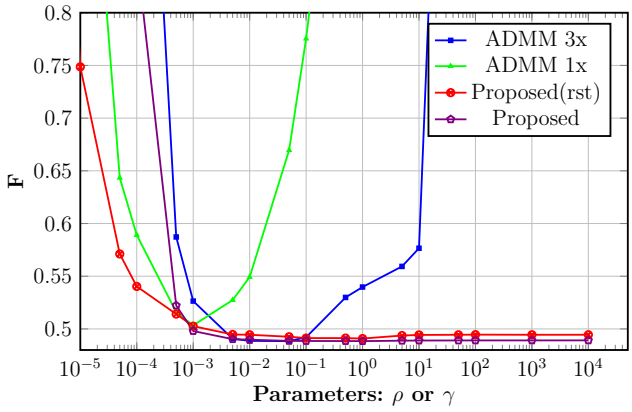

Fig. 2: Influence of the penalty parameters on the convergence speed: function cost at 1500 FFT evaluations vs. penalty parameters, ρ for **ADMM 3x** (ν and γ at possible best values), and γ for remaining two methods.

in MATLAB, except the core of the VMLM-B (LBFGS update and line-search) is implemented in C language. In all of methods the most expensive operation is the circular convolution, which is done using Fast Fourier Transform (FFT). Other operations have linear cost and are moreover same in all the methods, thus we neglect them in the convergence speed comparison. The Figure 1 compares convergence speed of the four methods against the number of FFTs consumed, and clearly the **proposed** method without restarting VMLM-B has speed comparable to the fastest **ADMM 3x**. In **ADMM 1x**, and **proposed** methods, the number of maximum successful iterations in VMLM-B was fixed to 150 and 60 respectively.

Table 1 reports the relative distance between the cost function after 1500 FFTs evaluations and the optimal value (estimated by the value reached after 10000 FFTs by the best performing algorithm). When the sub-minimizations are performed using an iterative method (VMLM-B), the number of

Fig. 3: On left, the observed image (241x241, BSNR=37.25dB); on top right corner, the blur kernel (15x15); on right, the estimated image (255x255, ISNR=6.12dB) after 1000 FFT evaluations by the **proposed** method. λ is tunned by hand to have possible best image quality.

iterations influence the convergence speed. We therefore report different values corresponding to 30, 60 or 100 inner iterations. As observed in [27], doing several splittings and using well tuned penalty parameters is preferable than performing an approximate minimization of the more complex sub-problem posed with a single splitting. Our proposed method reaches a better solution after the considered number of iterations if previous gradient steps are re-used after *dual variable* update (last row of the table). Loosing this second-order information noticeably impacts the performance.

Figure 2 illustrates the influence of the penalty parameters on the convergence speed of the four optimization methods, and the **proposed** methods are immune to large range of variation in the penalty parameter. **ADMM 3x** can converge very quickly for well tunned parameters, the convergence degrades strongly for a bad tuning. Given the number of parameters to be jointly tuned, this is a clear practical drawback. Reducing the number of splitting by resorting to iterative minimization of the sub-problems, as done with **ADMM 1x**, simplifies parameter tuning at the cost of a degraded convergence. Our **proposed** method offers a remarkable robustness to changes of the single penalty parameter while showing a convergence speed comparable to **ADMM 3x** with optimal tuning.

4. CONCLUSION

We proposed a general approach to handle non-smooth large-scale optimization problems, which follows the augmented Lagrangian approach. We have shown that hierarchical optimization is preferable to performing an alternate minimization, as usually done with ADMM. Quasi-Newton methods can then be applied efficiently thanks to a simple expression of the gradient, and the possibility to accumulate gradient difference knowledge independently of *dual variable* updates. Our results on image deblurring problem are promising: competitive convergence speed is reached with a method that is much simpler to tune.

5. REFERENCES

[1] Ciyou Zhu, Richard Byrd, Jorge Nocedal, and Jose Luis Morales, "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM Journal on Scientific and Statistical Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[2] Éric Thiébaut, "Optimization issues in blind deconvolution algorithms," in *Proc. SPIE 4847, Astronomical Data Analysis II*, Waikoloa, Hawaii, 2002, SPIE.

[3] Steven J Benson and Jorge J Moré, "A Limited Memory Variable Metric Method in Subspaces and Bound Constrained Optimization Problems," Tech. Rep. ANL/MCS-P909-0901, Math. and Computer Science Division, Argonne National Laboratory, 2001.

[4] William W. Hager and Hongchao Zhang, "A New Active Set Algorithm for Box Constrained Optimization," *SIAM Journal on Optimization*, vol. 17, no. 2, pp. 526–557, 2006.

[5] Mark W Schmidt, Ewout van den Berg, Michael P. Friedlander, and Kevin Murphy, "Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm," in *International Conference on Artificial Intelligence and Statistics*, Clearwater Beach, Florida, April 2009.

[6] Dong C. Liu and Jorge Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, pp. 503–528, 1989.

[7] Claude Lemaréchal, "Numerical Experiments in Nonsmooth Optimization," in *IIASA Workshop on Progress in Nondifferentiable Optimization*, Laxenburg, Austria, 1982, pp. 61–84.

[8] L Luksan and J Vlcek, "Globally Convergent Variable Metric Method for Convex Nonsmooth Unconstrained Minimization," *Journal of Optimization Theory and Applications*, vol. 102, no. 3, pp. 593–613, 1999.

[9] M. Haarala, K. Miettinen, and M.M. Mäkelä, "New limited memory bundle method for large-scale nonsmooth optimization," *Optimization Methods and Software*, vol. 19, no. 6, pp. 673–692, 2004.

[10] Adrian S. Lewis and Michael L. Overton, "Nonsmooth optimization via quasi-Newton methods," *Mathematical Programming*, vol. 141, no. 1, pp. 135–163, 2012.

[11] Angelia Nedic and Dimitri P Bertsekas, "Convergence Rate of Incremental Subgradient Algorithms," *Stochastic Optimization: Algorithms and Applications*, vol. 54, pp. 223–264, 2001.

[12] Jin Yu, S V N Vishwanathan, and Nicol N Schraudolph, "A Quasi-Newton Approach to Nonsmooth Convex Optimization Problems in Machine Learning," *Journal of Machine Learning Research*, vol. 11, pp. 1145–1200, 2010.

[13] Choon Hui Teo, S V N Vishwanathan, Alex Smola, and Quoc V Le, "Bundle Methods for Regularized Risk Minimization," *Journal of Machine Learning Research*, vol. 11, pp. 311–365, 2010.

[14] James V. Burke, Adrian S. Lewis, and Michael L. Overton, "A Robust Gradient Sampling Algorithm for Nonsmooth, Nonconvex Optimization," *SIAM Journal on Optimization*, vol. 15, no. 3, pp. 751–779, 2005.

[15] Krzysztof C. Kiwiel, "Convergence of the Gradient Sampling Algorithm for Nonsmooth Nonconvex Optimization," *SIAM Journal on Optimization*, vol. 18, no. 2, pp. 379–388, 2007.

[16] Gregory B. Passty, "Ergodic convergence to a zero of the sum of monotone operators in Hilbert space," *Journal of Mathematical Analysis and Applications*, vol. 72, no. 2, pp. 383–390, 1979.

[17] Ronald E. Bruck, "On the Weak Convergence of an Ergodic Iteration for the Solution of Variational Inequalities for Monontone Operators in Hilbert Space," *Journal of Mathematical Analysis and Applications*, vol. 61, pp. 159–164, 1977.

[18] Patrick L. Combettes and Valérie R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[19] José M. Bioucas-Dias and Mário A. T. Figueiredo, "A New TwIST: Two-Step Iterative Shrinkage/Thresholding Algorithms for Image Restoration," *IEEE Transaction on Image Processing*, vol. 16, no. 12, pp. 2992–3004, 2007.

[20] Amir Beck and Marc Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM Journal of Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[21] Yurii Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence O (1/k2)," *Doklady Akademii Nauk SSSR*, vol. 269, no. 3, pp. 543–547, 1983.

[22] Paul Tseng, "Applications of a Splitting Algorithm to Decomposition in Convex Programming and Variational Inequalities," *SIAM Journal on Control and Optimization*, vol. 29, no. 1, pp. 119–138, 1991.

[23] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Journal of Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 001–122, 2011.

[24] Jonathan Eckstein and Dimitri P Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, pp. 293–318, 1992.

[25] Manya V. Afonos, José M. Bioucas-Dias, and Mario A. T. Figueiredo, "Fast Image Recovery Using Variable Splitting and Constrained Optimization," *IEEE Transaction on Image Processing*, , no. 3, pp. 1–11, 2009.

[26] Manya V Afonso, José M Bioucas-dias, and Mário A T 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, 2011.

[27] Antonios Matakos, Sathish Ramani, and Jeffrey A. Fessler, "Accelerated Edge-Preserving Image Restoration Without Boundary Artifacts," *IEEE transactions on image processing*, vol. 22, no. 5, pp. 2019–2029, 2013.

[28] Stephen J. Wright, Robert D. Nowak, and Mário A.T. Figueiredo, "Sparse Reconstruction by Separable Approximation," *IEEE Transcation on Signal Processing*, vol. 57, no. 7, pp. 2479–2493, 2009.

[29] Neal Parikh and Stephen Boyd, "Proximal Algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2013.

[30] Yilun Wang, Junfeng Yang, Wotao Yin, and Yin Zhang, "A new alternating minimization algorithm for total variation image reconstruction," *SIAM Journal of Imaging Sciences*, vol. 1, no. 3, pp. 248–272, 2008.