Linearized Alternating Direction Method with Penalization for Nonconvex and Nonsmooth Optimization

Yiyang Wang,[1] Risheng Liu,[2] Xiaoliang Song,[1] and Zhixun Su[1,3]
[1] School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China
[2] School of Software Technology, Dalian University of Technology, Dalian 116024, China
[3] National Engineering Research Center of Digital Life, Guangzhou 510006, China
{ywerica, ericsong507}@gmail.com, {rsliu, zxsu}@dlut.edu.cn

Abstract

Being one of the most effective methods, Alternating Direction Method (ADM) has been extensively studied in numerical analysis for solving linearly constrained convex program. However, there are few studies focusing on the convergence property of ADM under nonconvex framework though it has already achieved well-performance on applying to various nonconvex tasks. In this paper, a linearized algorithm with penalization is proposed on the basis of ADM for solving nonconvex and nonsmooth optimization. We start from analyzing the convergence property for the classical constrained problem with two variables and then establish a similar result for multi-block case. To demonstrate the effectiveness of our proposed algorithm, experiments with synthetic and real-world data have been conducted on specific applications in signal and image processing.

1 Introduction

Though plenty of problems in machine learning and image processing can be modeled as convex optimization (Candès and Wakin 2008; Wright et al. 2010; Yang et al. 2009; Liu et al. 2014), many recent applications like distributed clustering, tensor factorization, dictionary learning and gradient based minimization have shown the great success (Liavas and Sidiropoulos 2014; Bao et al. 2014; Xu et al. 2012; Wang et al. 2014) and led to growing interest in nonconvex and nonsmooth (NCNS) optimization.

Many of these problems, either convex or nonconvex can be formulated/reformulated as a linearly constrained separable program with n blocks of variables:

$$\min_{\mathbf{x}_i,\dots,\mathbf{x}_n} \Psi(\mathbf{X}) = \sum_{i=1}^{n} f_i(\mathbf{x}_i), \ \text{ s.t. } \sum_{i=1}^{n} \mathcal{A}_i(\mathbf{x}_i) = \mathbf{c}, \tag{1}$$

where in this paper, variables $\{\mathbf{x}_i\}_{i=1}^n$, constant $\mathbf{b}$ can be either vectors or matrices and we denote $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ for simplicity of discussion; $\{\mathcal{A}_i\}_{i=1}^n$ are linear mappings without additional restrictions; the objective function Ψ can be either convex or nonconvex, smooth or nonsmooth which satisfies the following two conditions:

1. The f_i's are proper, lower semi-continuous functions with $\inf f_i > -\infty$ for $\forall i$ and $\inf \Psi > -\infty$.
2. $\Psi(\mathbf{X})$ is a coercive, Kurdyka-Łojasiewicz (KL) function[1], namely, $\lim_{\min\{\|\mathbf{x}_i\|:i=1,\dots,n\}\uparrow\infty} \Psi(\mathbf{X}) = +\infty$[2].

Based on different choices of $\Psi(\mathbf{X})$, problem (1) covers a variety of problems. Signal representation based on ℓ_0 sparse model is used as a technique for decomposing a signal into an optimal superposition of bases (Chen and Donoho 1994), which can be seen as a one-block nonconvex problem of (1). Being as a basic but vital task, image denoising can be modeled as sparse coding with ℓ_0-regularization (Bao et al. 2014; Gregor and LeCun 2010), which is a special two-block case of (1) that the objective function is a combination of a convex function and a nonconvex one.

There have been extensive literatures in optimization and numerical analysis on solving the problem (1) with the case that all the f_i's are convex functions. For solving the convex constrained problem, efficient algorithms like Alternating Direction Method (ADM) (Lin, Liu, and Li 2013), Alternating Minimization Algorithm (Tseng 1991) and Accelerated Proximal Gradient (Zuo and Lin 2011) have been widely applied. A good summary of these methods can be found in (Goldstein et al. 2014), but we in this paper only focus on ADM since it has been extensively studied from smooth and strongly convex function to nonsmooth convex function (Boyd et al. 2011); from naive algorithm to its linearized version (Lin, Liu, and Su 2011); from two-block case to multi-block case (Chen et al. 2014) and have been proved to be quite efficient by many application-driven tasks.

Despite of the numerous methods for solving convex constrained problem, few studies are conducted on the nonconvex frameworks. Greedy methods, including Matching Pursuit (MP) (Mallat and Zhang 1993), Orthogonal Matching Pursuit (OMP) (Tropp and Gilbert 2007), Weak Matching Pursuit (WMP) (Temlyakov 2011) are designed for the ℓ_0 sparse approximation task due to the hopelessness of a straightforward approach. Another common strategy for solving nonconvex linearly constrained problem is to reformulate it to an unconstrained optimization, e.g. penalizing

[1] We ignore the definition of KL inequality and KL function (Bolte, Sabach, and Teboulle 2014) due to space limit.

[2] $\|\cdot\|$ denotes the ℓ_2 norm of vector and the Frobenius norm of matrix in here and after.

the linear constraint of ℓ_0 sparse coding problem with a certain parameter and then it can be solved by Iterative Hard Thresholding Algorithm (IHTA) (Bach et al. 2012). Moreover, (Bolte, Sabach, and Teboulle 2014) proposes a proximal alternating linearized minimization (PALM) for unconstrained optimization with objective function that satisfies KL property. Follow that, (Xu and Yin 2014) extends the PALM to Block Coordinate Descent and gives analysis on its accelerated version. Very recently, (Wang, Xu, and Xu 2014) propose a Bregman modification of ADM for nonconvex constrained optimization with special assumptions on objective functions and linear constraints.

In this paper, we aim to propose an algorithm based on linearized ADM with penalization (LADMP) for solving general nonconvex, nonsmooth problems. Our development and analysis begin on the classical linearly constrained problem with two variables and then straightforward to establish a similar result for multi-block problems. Specifically, by introducing an auxiliary variable, we penalize its bringing additional constraint on the objective function with an increasing parameter. For the linearly constraint optimization with fixed penalization, we propose a method based on linearized ADM for the purpose of avoiding the difficulties of solving subproblems. We prove that our LADMP converges to a KKT point of the primal optimization. Furthermore, though it seems like that the auxiliary variable brings double computational cost to LADMP, we show detailed skills to reduce the complexity of LADMP less than linearized ADM (LADM). We test LADMP on ℓ_0 sparse approximation task with synthetic clean data and propose a speed-up strategy for this special problem. In addition, an experiment on ℓ_0 sparse coding is conducted on real-world data for the application of image denoising. The experimental results verify the convergence property of LADMP and also indicate the effectiveness of our proposed algorithm.

2 Preliminaries

We start the analysis of the classical linearly constrained optimization with two variables, denoting as $\mathbf{x}$ and $\mathbf{y}$ to simplify the subsequent derivation (the same to f, g, $\mathcal{A}$ and $\mathcal{B}$):

$$\min_{\mathbf{x},\mathbf{y}} f(\mathbf{x}) + g(\mathbf{y}), \quad \text{s.t.} \mathcal{A}(\mathbf{x}) + \mathcal{B}(\mathbf{y}) = \mathbf{c}. \tag{2}$$

Firstly, we briefly review the ADM and its linearized version for convex optimization and then give some notations and assumptions that will be used throughout the paper.

2.1 ADM and LADM for Convex Optimization

For the case that f and g in problem (2) are convex functions, the optimal point of problem (2) can be obtained through iteratively minimizing the augmented Lagrange function $L_\beta(\mathbf{x},\mathbf{y},\mathbf{p}) = f(\mathbf{x}) + g(\mathbf{y}) + \langle \mathbf{p}, \mathcal{A}(\mathbf{x}) + \mathcal{B}(\mathbf{y}) - \mathbf{c}\rangle + \frac{\beta}{2}\|\mathcal{A}(x) + \mathcal{B}(\mathbf{y}) - \mathbf{c}\|^2$ with the Lagrange multiplier $\mathbf{p}$:

$$\begin{aligned}
&\mathbf{x}_{k+1} \in \arg\min_{\mathbf{x}} L_\beta(\mathbf{x}, \mathbf{y}_k, \mathbf{p}_k),\\
&\mathbf{y}_{k+1} \in \arg\min_{\mathbf{y}} L_\beta(\mathbf{x}_{k+1}, \mathbf{y}, \mathbf{p}_k),\\
&\mathbf{p}_{k+1} = \mathbf{p}_k + \beta(\mathcal{A}(x_{k+1}) + \mathcal{B}(\mathbf{y}_{k+1}) - \mathbf{c}).
\end{aligned}$$

The ADM is appealing when $\mathcal{A}$ and $\mathcal{B}$ are identities (Yin 2010), however, for the case that $\mathcal{A}$ and $\mathcal{B}$ are not identities, a common strategy is to introduce auxiliary variables to substitute $\mathbf{x}$ and $\mathbf{y}$ in the objective function. Though this strategy ensures the closed-form solution of each subproblem, it brings additional problem like high storage and computational cost and weak theoretical results (Lin, Liu, and Su 2011; Chen et al. 2014). Another simple but efficient strategy to tackle this problem is to linearize the quadratic term of $L_\beta(\mathbf{x},\mathbf{y},\mathbf{p})$. This linearized version of ADM is proved to be effective for convex problem by applications (Lin, Liu, and Su 2011), however, both of ADM and LADM can not be used to nonconvex optimization directly due to the failure of the Féjer monotonicity of iterates (Wang, Xu, and Xu 2014).

2.2 Notations and Assumptions

1. We in this paper denote the variable $\mathbf{w} = (\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{p})$, $\widehat{\mathbf{w}} = (\mathbf{w},\mathbf{r})$ with variable $\mathbf{r}$ satisfies $\mathbf{r}_k^l = \mathbf{z}_k^{l-1}$; then denote $\mathbf{dx}_k^{l,l+1} = \|\mathbf{x}_k^{l+1} - \mathbf{x}_k^l\|$, $\mathbf{dy}_k^{l,l+1} = \|\mathbf{y}_k^{l+1} - \mathbf{y}_k^l\|$, $\mathbf{dz}_k^{l,l+1} = \|\mathbf{z}_k^{l+1} - \mathbf{z}_k^l\|$ and $\mathbf{dp}_k^{l,l+1} = \|\mathbf{p}_k^{l+1} - \mathbf{p}_k^l\|$. With $\alpha_k^0 = 2(\eta_k^3)^2/\beta_k$, $\widehat{L}_{\beta_k}^k(\widehat{\mathbf{w}}) = L_{\beta_k}^k(\mathbf{w}) + \alpha_k^0\|\mathbf{z} - \mathbf{r}\|^2$.
2. The sequence $\{\mathbf{w}_k^l\}_{l\in\mathbb{N}}$ generated by LADMP is bounded and the primal problem (2) has a stable point.
3. Parameter β_k satisfies $\beta_k\eta_k^3 > 4(\eta_k^3 + 2\mu_k)^2 + 4(\eta_k^3)^2$; η_k^1, η_k^2 satisfy $\eta_k^1 > L_k^{\mathcal{A}}$, $\eta_k^2 > L_k^{\mathcal{B}}$ where $L_k^{\mathcal{A}}$, $L_k^{\mathcal{B}}$ are the Lipschitz constants of the partial gradients of function $\frac{\beta_k}{2}\|\mathcal{A}(\mathbf{x}) + \mathcal{B}(\mathbf{y}) - \mathbf{c}\|^2$ with respect to $\mathbf{x}$ and $\mathbf{y}$.

3 LADMP for NCNS Problem

Instead of solving the primal problem (2), we deal with an equivalent problem by introducing an auxiliary variable $\mathbf{z}$ as

$$\begin{aligned}
\min_{\mathbf{x},\mathbf{y},\mathbf{z}} \quad & f(\mathbf{x}) + g(\mathbf{y}),\\
\text{s.t.} \quad & \widetilde{\mathcal{A}}(\mathbf{x}) + \widetilde{\mathcal{B}}(\mathbf{y}) + \mathcal{T}(\mathbf{z}) = \widetilde{\mathbf{c}},\\
& \mathcal{R}(\mathbf{z}) = \mathbf{0},
\end{aligned} \tag{3}$$

where the notations are $\widetilde{\mathcal{A}} = [\mathcal{A}; \mathcal{O}]$, $\widetilde{\mathcal{B}} = [\mathcal{O}; \mathcal{B}]$, $\widetilde{\mathbf{c}} = [\mathbf{c}; \mathbf{0}]$, $\mathcal{T} = [\mathcal{I}, \mathcal{O}; \mathcal{O}, -\mathcal{I}]$ and $\mathcal{R} = [\mathcal{I}, -\mathcal{I}]$, where $\mathbf{0}$ denotes a constant with all-zero elements, $\mathcal{O}$ denotes zero mapping and $\mathcal{I}$ denotes identity mapping. It is easy to check that the reformulated problem (3) is equivalent to the primal problem (2), then our proposed algorithm is proposed directly on solving the problem (3).

3.1 The Proposed Algorithm

We penalize the square of the last constraint violation to the objective function with penalty parameter $\mu > 0$ as

$$\begin{aligned}
\min_{\mathbf{x},\mathbf{y},\mathbf{z}} \quad & f(\mathbf{x}) + g(\mathbf{y}) + \frac{\mu}{2}\|\mathcal{R}(\mathbf{z})\|^2,\\
\text{s.t.} \quad & \widetilde{\mathcal{A}}(\mathbf{x}) + \widetilde{\mathcal{B}}(\mathbf{y}) + \mathcal{T}(\mathbf{z}) = \widetilde{\mathbf{c}}.
\end{aligned} \tag{4}$$

By driving μ to ∞, we consider a sequence of $\{\mu_k\}_{k\in\mathbb{N}}$ with $\mu_k \uparrow \infty$ as $k \uparrow \infty$, and to seek the approximate minimizer $(\mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k)$ of (4) for each fixed μ_k. Then, a proximal method based on LADM is proposed for each constrained problem (4) with fixed μ_k.

To be specific, by writing down the augmented Lagrange function $L^k_{\beta_k}(\mathbf{w}) = f(\mathbf{x})+g(\mathbf{y})+\frac{\mu_k}{2}\|\mathcal{R}(\mathbf{z})\|^2+\frac{\beta_k}{2}\|\widetilde{\mathcal{A}}(\mathbf{x})+\widetilde{\mathcal{B}}(\mathbf{y}) + \mathcal{T}(\mathbf{z}) - \widetilde{\mathbf{c}} + \frac{1}{\beta_k}\mathbf{p}\|^2 - \frac{1}{2\beta_k}\|\mathbf{p}\|^2$ of problem (4) with fixed μ_k, then $\mathbf{w}_k$ is obtained by solving the following subproblems iteratively with initial point $\mathbf{w}^0_k = \mathbf{w}_{k-1}$:

$$\begin{aligned}
\mathbf{x}^{l+1}_k &\in \arg\min_{\mathbf{x}} f(\mathbf{x}) + \frac{\eta^1_k}{2}\|\mathbf{x} - \mathbf{u}^l_k\|^2,\\
\mathbf{y}^{l+1}_k &\in \arg\min_{\mathbf{y}} g(\mathbf{y}) + \frac{\eta^2_k}{2}\|\mathbf{y} - \mathbf{v}^l_k\|^2,\\
\mathbf{z}^{l+1}_k &= \arg\min_{\mathbf{z}} \frac{\beta_k}{2}\|h(\mathbf{x}^{l+1}_k, \mathbf{y}^{l+1}_k, \mathbf{z}, \mathbf{p}^l_k)\|^2\\
&\qquad + \frac{\mu_k}{2}\|\mathcal{R}(\mathbf{z})\|^2 + \frac{\eta^3_k}{2}\|\mathbf{z} - \mathbf{z}^l_k\|^2,\\
\mathbf{p}^{l+1}_k &= \beta_k h(\mathbf{x}^{l+1}_k, \mathbf{y}^{l+1}_k, \mathbf{z}^{l+1}_k, \mathbf{p}^l_k),
\end{aligned} \tag{5}$$

where $h(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{p}) = \widetilde{\mathcal{A}}(\mathbf{x}) + \widetilde{\mathcal{B}}(\mathbf{y}) + \mathcal{T}(\mathbf{z}) + \frac{1}{\beta_k}\mathbf{p} - \widetilde{\mathbf{c}}$ is denoted for simplicity; $\mathbf{u}^l_k = \mathbf{x}^l_k - \frac{\beta_k}{\eta^1_k}\widetilde{\mathcal{A}}^* h(\mathbf{x}^l_k, \mathbf{y}^l_k, \mathbf{z}^l_k, \mathbf{p}^l_k)$ and $\mathbf{v}^l_k = \mathbf{y}^l_k - \frac{\beta_k}{\eta^2_k}\widetilde{\mathcal{B}}^* h(\mathbf{x}^{l+1}_k, \mathbf{y}^l_k, \mathbf{z}^l_k, \mathbf{p}^l_k)$. The iteration stops and the $\mathbf{w}_{k+1}$ is obtained when reaching

$$\|\nabla_{\mathbf{z}} L^k_{\beta_k}(\mathbf{w}^{l+1}_k)\| \leq \tau_k, \tag{6}$$

where $\tau_k \to 0$ is a nonnegative tolerance to ensure that the minimization is carried out more accurately as the iterations progress. We summarize the proposed LADMP in Alg. 1 and give the convergence analysis in the following section.

Algorithm 1 Solving problem (3) by LADMP

```
Given $\mu_0 > 0$, $\nu > 1$, sequence $\{\tau_k\} > 0$ with $\tau_k \downarrow 0$.
Initialize variables $\mathbf{w}_0$.
repeat
    Given appropriate parameters $\eta^1_k$, $\eta^2_k$, $\eta^3_k$ and $\beta_k$.
    Set starting point $\mathbf{w}^0_k$ as $\mathbf{w}_{k-1}$.
    while $\|\nabla_{\mathbf{z}} L^k_{\beta_k}(\mathbf{w}^{l+1}_k)\| > \tau_k$ do
        Compute $\mathbf{w}^{l+1}_k$ by Eq. (5).
    end while
    Set $\mathbf{w}_{k+1}$ as $\mathbf{w}^{l+1}_k$.
    Update $\mu_{k+1} = \nu\mu_k$.
until converges.
```

3.2 Convergence Analysis of LADMP

The methodology that describes the main steps to achieve convergence analysis of LADMP can be split into two parts. Firstly, it is necessary to prove the accessibility of the stopping criterion (6). Secondly, we prove that a limit point $\mathbf{w}^*$ of sequence $\{\mathbf{w}_k\}_{k\in\mathbb{N}}$ is a KKT point of problem (3).

Accessibility of the Stopping Criterion With the decrease of τ_k, the minimization is carried out more accurately so that $\mathbf{w}_k$ is nearly the optimal point of problem (4) with fixed μ_k. Therefore, instead of proving the accessibility of the stopping criterion, we prove the convergence property of our algorithm for solving problem (4) with fixed μ_k.

Before proving the main theorem, we first provide a key lemma of the proposed algorithm, i.e. the sufficient descent property and lower boundedness of the subgradient of the iterates gap which are two quite standard requirements shared by essentially most descent algorithms (Attouch et al. 2010).

Lemma 1 *Suppose that the assumptions in the Preliminaries hold, $\{\widehat{\mathbf{w}}^l_k\}_{l\in\mathbb{N}}$ is a sequence generated by LADMP for the problem (4) with fixed μ_k , then there exist positive integers γ^1_k and γ^2_k such that the following two assertions hold:*

1. (Sufficient Descent Property)

$$\begin{aligned}
&\widehat{L}^k_{\beta_k}(\widehat{\mathbf{w}}^l_k) - \widehat{L}^k_{\beta_k}(\widehat{\mathbf{w}}^{l+1}_k)\\
\geq &\gamma^1_k((\mathbf{dx}^{l,l+1}_k)^2 + (\mathbf{dy}^{l,l+1}_k)^2 + (\mathbf{dz}^{l,l+1}_k)^2).
\end{aligned}$$

2. (Lower Boundedness of Subgradient)[3]

$$\begin{aligned}
&\mathit{dist}(0, \partial\widehat{L}^k_{\beta_k}(\widehat{\mathbf{w}}^l_k))\\
\leq &\gamma^2_k(\mathbf{dx}^{l-1,l}_k + \mathbf{dy}^{l-1,l}_k + \mathbf{dz}^{l-1,l}_k + \mathbf{dz}^{l-2,l-1}_k).
\end{aligned}$$

Lemma 1 can be straightforward proved in a similar way as the previous works (Bolte, Sabach, and Teboulle 2014; Bao et al. 2014; Xu and Yin 2014; Wang, Xu, and Xu 2014), however, we ignore the detailed proof due to the space limit. Then with this vital lemma, we can prove the global convergence property of LADMP for solving problem (4) with fixed μ_k, as stated in the following theorem.

Theorem 2 *Suppose that $\{\widehat{\mathbf{w}}^l_k\}_{l\in\mathbb{N}}$ be a sequence generated by LADMP with the assumptions in the Preliminaries hold. With the help of the KL property, we can get the following inequality that there exists a positive integer M such that*

$$\begin{aligned}
&\mathbf{dx}^{l-1,l}_k + \mathbf{dy}^{l-1,l}_k + \mathbf{dz}^{l-1,l}_k + \mathbf{dz}^{l-2,l-1}_k + M \vartriangle^{l,l+1}_k\\
\geq &3(\mathbf{dx}^{l,l+1}_k + \mathbf{dy}^{l,l+1}_k + \mathbf{dz}^{l,l+1}_k),
\end{aligned} \tag{7}$$

where for any $l \geq N_1$, $\sum_{l=N_1}^{N_2} \vartriangle^{l,l+1}_k < \infty$. In addition, $\{\mathbf{w}^l_k\}_{l\in\mathbb{N}}$ is a Cauchy sequence that converges to a KKT point of the problem (4) with fixed μ_k.

Proof Let $\omega(\mathbf{w}^0_k)$ be the set of the limit points of $\{\mathbf{w}^l_k\}_{l\in\mathbb{N}}$ generated by LADMP from initial point $\mathbf{w}^0_k$. Then $\widehat{L}^k_{\beta_k}(\widehat{\mathbf{w}})$ has the uniformized KL property from the analysis of the previous work (Bolte, Sabach, and Teboulle 2014; Xu and Yin 2014; Wang, Xu, and Xu 2014) on $\omega(\mathbf{w}^0_k)$, that is, there exists function ϕ corresponding to KL property and $N_1 \in \mathbb{N}$ such that for any $l \geq N_1$:

$$\mathrm{dist}(0, \partial\widehat{L}^k_{\beta_k}(\widehat{\mathbf{w}}^l_k)) \vartriangle^{l,l+1}_k \geq \widehat{L}^k_{\beta_k}(\widehat{\mathbf{w}}^l_k) - \widehat{L}^k_{\beta_k}(\widehat{\mathbf{w}}^{l+1}_k).$$

Then by bringing in the conclusions in Lemma 1, we can easily get the inequality (7) with $M = 27\gamma^2_k/4\gamma^1_k$.

Then we summarize the inequality (7) from N_1 to any $N_2 > N_1$ which yields that

$$\begin{aligned}
&\sum\nolimits_{l=N_1}^{N_2} 2\mathbf{dx}^{l,l+1}_k + 2\mathbf{dy}^{l,l+1}_k + \mathbf{dz}^{l,l+1}_k\\
\leq &C + M\phi(\widehat{L}^k_{\beta_k}(\widehat{\mathbf{w}}^{N_1}_k) - \widehat{L}^k_{\beta_k}(\widehat{\mathbf{w}}^*_k)),
\end{aligned}$$

[3] For any subset $\mathcal{S} \subset \mathbb{R}^d$ and any point $\mathbf{u} \in \mathbb{R}^d$, $\mathrm{dist}(\mathbf{u}, \mathcal{S}) := \inf\{\|\mathbf{v} - \mathbf{u}\| : \mathbf{v} \in \mathcal{S}\}$, when $\mathcal{S} = \emptyset$, we have $\mathrm{dist}(\mathbf{u}, \mathcal{S}) = \infty$.

where C is a constant less than infinite and is not affected by the value of l. Since N_2 is chosen arbitrarily, this easily shows that $\{\mathbf{x}_k^l, \mathbf{y}_k^l, \mathbf{z}_k^l\}_{l\in\mathbb{N}}$ has finite length:

$$\sum_{l=0}^{\infty} \mathbf{dx}_k^{l,l+1} + \mathbf{dy}_k^{l,l+1} + \mathbf{dz}_k^{l,l+1} < +\infty. \tag{8}$$

Due to the penalization strategy of the LADMP, we can indicates that $\sum_{l=0}^{\infty} \mathbf{dp}_k^{l,l+1} < +\infty$ (an intermediate proof of the Sufficient Descent Property in Lemma 1), which together with Eq. (8) implies that $\{\mathbf{w}_k^l\}_{l\in\mathbb{N}}$ is a Cauchy sequence and in addition converges to a KKT point though writing down the first order optimality condition of each subproblem.

Remark 3 *The Theorem 2 above ensures that the sequence $\{\mathbf{w}_k^l\}_{l\in\mathbb{N}}$ generated by LADMP converges to a KKT point of the problem (4) with fixed μ_k, which indicates the accessibility of the stopping criterion (6) with the decreasing of τ_k. In addition, it is natural to get the assertion that for the linearly constrained optimization in a similar form of (4), solving it through (5) ensures the convergence of the algorithm itself.*

Main Convergence Theorem of LADMP After getting the sequence $\{\mathbf{w}_k\}_{k\in\mathbb{N}}$ where each $\mathbf{w}_k$ is an approximate solution of problem (4) with fixed μ_k. Then we have that any limit point of $\{\mathbf{w}_k\}_{k\in\mathbb{N}}$ is a KKT point for the problem (3) from the following theorem.

Theorem 4 *Suppose $\{\mathbf{w}_k\}_{k\in\mathbb{N}}$ be a sequence generated by LADMP with the assumptions in the Preliminaries hold, then we can conclude that a limit point $\mathbf{w}^*$ of $\{\mathbf{w}_k\}_{k\in\mathbb{N}}$ is a KKT point of the problem (3).*

Proof With the assumptions in the Preliminaries, $\{\mathbf{w}_k\}_{k\in\mathbb{N}}$ has a convergent subsequence $\{\mathbf{w}_{k_j}\}_{j\in\mathbb{N}}$ with $\mathbf{w}_{k_j} \to \mathbf{w}^*$ as $j \uparrow \infty$. On the other hand, we can obtain from the stopping criterion (6) that

$$\|\mu_k \mathcal{R}^*\mathcal{R}(\mathbf{z}_{k+1}) + \mathcal{T}^*(\mathbf{p}_{k+1}) + \beta_k \mathcal{T}^*(\widetilde{\mathbf{d}}_{k+1})\| \leq \tau_k, \tag{9}$$

where $\widetilde{\mathbf{d}}_{k+1} = \widetilde{\mathcal{A}}(\mathbf{x}_{k+1}) + \widetilde{\mathcal{B}}(\mathbf{y}_{k+1}) + \mathcal{T}(\mathbf{z}_{k+1}) - \widetilde{\mathbf{c}}$. Then with the triangle inequality we reformulate Eq. (9) as

$$\|\mathcal{R}^*\mathcal{R}(\mathbf{z}_{k+1})\| \leq \frac{1}{\mu_k}(\tau_k + \|\mathcal{T}^*(\mathbf{p}_{k+1})\| + \|\beta_k \mathcal{T}^*(\widetilde{\mathbf{d}}_{k+1})\|).$$

Since $\{\mathbf{w}_k\}_{k\in\mathbb{N}}$ is bounded, then there exists an integer M_0 such that $\|\mathcal{T}^*(\mathbf{p}_{k+1})\| + \|\beta_k \mathcal{T}^*(\widetilde{\mathbf{d}}_{k+1})\| \leq M_0$. When we take limits as $k \uparrow \infty$, the bracketed term on the right-hand-side approaches M_0 since $\tau_k \downarrow 0$. In addition, the right-hand-side term approaches to zero since $\mu_k \uparrow \infty$ as $k \uparrow \infty$. From the corresponding limit on the left-hand-side, we obtain $\mathcal{R}^*\mathcal{R}(\mathbf{z}^*) = 0$. Therefore, we have $\mathcal{R}(\mathbf{z}^*) = 0$. since the mapping $\mathcal{R}$ is linearly independent. On the other hand, it is an explicit expression of KKT point in Theorem 2 that

$$\begin{aligned} &-\partial f(\mathbf{x}^*) \in \widetilde{\mathcal{A}}^*(\mathbf{p}^*), \quad \mu_k \mathcal{R}^*\mathcal{R}(\mathbf{z}^*) + \mathcal{T}^*(\mathbf{p}^*) = 0, \\ &-\partial g(\mathbf{y}^*) \in \widetilde{\mathcal{B}}^*(\mathbf{p}^*), \quad \widetilde{\mathcal{A}}(\mathbf{x}^*) + \widetilde{\mathcal{B}}(\mathbf{y}^*) + \mathcal{T}(\mathbf{z}^*) - \widetilde{\mathbf{c}} = 0, \end{aligned}$$

which together with $\mathcal{R}(\mathbf{z}^*) = 0$. ensures the desired assertion that $\mathbf{w}^*$ is a KKT point of the problem (3). ■

Remark 5 *For nonconvex nonsmooth problems, converging to KKT point is so far the best result (Bolte, Sabach, and Teboulle 2014; Wang, Xu, and Xu 2014; Yuan and Ghanem 2013) identified by researchers as far as we know. It should be emphasized that a KKT point of nonconvex problem could be a local minimizer for the problem under some conditions, e.g. the second-order sufficient conditions (Nocedal and Wright 2006). Moreover, converging to a KKT point for nonconvex problem is significant as a reference for the optimal point and seems to be acceptable in various applications.*

3.3 Extension to Multi-block Problem

Different from the conclusion that direct extension of the ADM for multi-block convex optimization is not necessarily convergent (Chen et al. 2014). The convergence analysis of multi-block, nonconvex and nonsmooth optimization is a straightforward extension of the problem with two variables.

Consider the multi-block, nonconvex and nonsmooth optimization with linearly constraint described as problem (1), we solve the following equivalent problem by introducing an auxiliary variable $\mathbf{z}$ and penalize the bring constraint on the objective function with μ:

$$\begin{aligned} \min_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}} \quad & \sum_{i=1}^{n} f_i(\mathbf{x}_i) + \frac{\mu}{2}\|\mathcal{R}(\mathbf{z})\|^2, \\ \text{s.t.} \quad & \sum_{i=1}^{n} \widetilde{\mathcal{A}}_i(\mathbf{x}_i) + \mathcal{T}(\mathbf{z}) = \widetilde{\mathbf{c}}, \end{aligned} \tag{10}$$

where for $i = 1, \ldots, n-1$, $\widetilde{\mathcal{A}}_i = [\mathcal{A}_i; \mathcal{O}]$; $\widetilde{\mathcal{A}}_N = [\mathcal{O}; \mathcal{A}_N]$. $\mathcal{R}$, $\mathcal{T}$ and $\widetilde{\mathbf{c}}$ are defined in a similar way of two-block case.

Solve the problem (10) by our proposed LADMP, it is the same with two-block case that the penalize parameter $\mu_k \uparrow \infty$ and the stopping criterion $\tau_k \downarrow 0$ as $k \uparrow \infty$. Obviously, the assumption of the parameters should be modified for multi-block case, but the methodology of proving the convergence is a straightforward extension of Theorem 2 and Theorem 4. Hence we in this paper ignore the detailed proof for multi-block problem due to space limit.

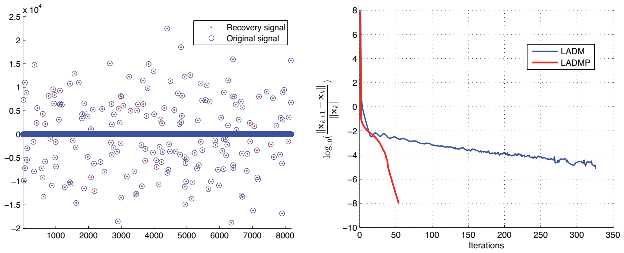
3.4 Discussion of Implementation Details

We in this section show implementation details when applying LADMP for nonconvex and nonsmooth optimization with linearly constraint. With these well designed tips, our LADMP becomes easier to apply and its computational cost turns smaller than the conventional LADM.

Alternative Condition Our convergence analysis is conducted under the stopping criterion (6), which is indispensable to the main theorem of LADMP. However, this stopping criterion may not be easily confirmed and can be replaced by another practical termination condition as follows:

$$\max\{\frac{\|\mathbf{x}_k^{l+1} - \mathbf{x}_k^l\|}{\max\{1, \|\mathbf{x}_k^l\|\}}, \frac{\|\mathbf{y}_k^{l+1} - \mathbf{y}_k^l\|}{\max\{1, \|\mathbf{y}_k^l\|\}}\} < \tau_k. \tag{11}$$

This alternative condition is equivalent to the stopping criterion (6) while at the same time much easier to be checked.



(a) Accuracy of LADMP (b) Convergence property

Figure 1: Convergence property on (a) the accuracy of the recovery signal generated by our proposed LADMP and (b) the recovery error on iterations compared with LADM.

Hence, the new condition does not affect the convergence analysis so that we suggest using the (11) rather than the stopping criterion (6) in practice.

Computational Cost For the convenience of discussion, we consider the two-block problem with $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$, linear mappings $\mathcal{A}$ and $\mathcal{B}$ are matrices $\mathbf{A} \in \mathbb{R}^{t\times m}$, $\mathbf{B} \in \mathbb{R}^{t\times n}$. Though it seems like that our LADMP brings double complexity due to the introduction of auxiliary variable $\mathbf{z}$, the special properties that $\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}} = [\mathbf{A}^\top\mathbf{A}; \mathbf{0}]$, $\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{B}} = \mathbf{0}$, $\widetilde{\mathbf{B}}^\top\widetilde{\mathbf{B}} = [\mathbf{0}; \mathbf{B}^\top\mathbf{B}]$ and $\widetilde{\mathbf{B}}^\top\widetilde{\mathbf{A}} = \mathbf{0}$ help simplify the computation. On the other hand, the subproblem of $\mathbf{z}$ seems like that there is a necessary to compute the inverse matrix of $\mu_k\mathbf{R}^\top\mathbf{R} + (\beta_k + \eta_k^3)\mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. However, its inverse matrix can be explicitly represented as $[c_k^1\mathbf{I}, c_k^2\mathbf{I}; c_k^2\mathbf{I}, c_k^1\mathbf{I}]$, where $c_k^1 = (\mu_k+\beta_k+\eta_k^3)/((\mu_k+\beta_k+\eta_k^3)^2-\mu_k^2)$ and $c_k^2 = \mu_k/((\mu_k+\beta_k+\eta_k^3)^2-\mu_k^2)$. With these simplifications, the complexity of LADMP is $\mathcal{O}(3mt+3nt)$ which is less than $\mathcal{O}(3mt+4nt)$ of LADM at each iteration.

4 Experimental Results

Though there are many tasks in machine learning and image processing that can be formulated/reformulated to problem (1), we consider to apply LADMP to the applications of signal representation and image denoising. For signal representation, the input signals are represented using a sparse linear combination of basis vectors which is popular for extracting semantic features. On the other hand, as a special kind of signal, images are widespread in daily life and image denoising is a basic but vital task in image processing. All the algorithms, including comparative methods are implemented by Matlab R2013b and are tested on a PC with 8 GB of RAM and Intel Core i5-4200M CPU.

4.1 Signal Representation

Signal representation aims to construct succinct representations of the input signal, i.e. a linear combination of only a few atoms of the dictionary bases. It can be typically formulated as the following optimization:

$$\min_{\mathbf{x}} \ \|\mathbf{x}\|_0, \text{ s.t. } \mathbf{Ax} = \mathbf{c}, \tag{12}$$

s/λ	Methods	Time(s)	Error
1024/ 2×10^6	MP	0.101	1.068×10^{-4}
	WMP	0.114	1.068×10^{-4}
	OMP	**0.094**	4.657×10^{-30}
	LADMP	0.392	1.361×10^{-7}
	F-LADMP	0.305	$\mathbf{1.492\times10^{-30}}$
2048/ 1.5×10^7	MP	0.808	1.051×10^{-4}
	WMP	**0.535**	1.050×10^{-4}
	OMP	0.816	6.206×10^{-30}
	LADMP	1.01	1.227×10^{-7}
	F-LADMP	0.846	$\mathbf{3.735\times10^{-30}}$
4096/ 1×10^8	MP	4.498	1.157×10^{-4}
	WMP	4.511	1.157×10^{-4}
	OMP	4.786	7.16×10^{-5}
	LADMP	4.030	1.978×10^{-6}
	F-LADMP	**3.182**	$\mathbf{8.045\times10^{-7}}$
8192/ 1×10^9	MP	35.686	1.220×10^{-4}
	WMP	35.702	1.220×10^{-4}
	OMP	44.841	6.664×10^{-5}
	LADMP	15.917	1.415×10^{-6}
	F-LADMP	**11.117**	$\mathbf{9.837\times10^{-30}}$
16384/ 7×10^9	MP	212.701	1.133×10^{-4}
	WMP	217.707	1.133×10^{-4}
	OMP	415.069	6.77×10^{-5}
	LADMP	62.089	1.107×10^{-6}
	F-LADMP	**47.161**	$\mathbf{3.056\times10^{-8}}$

Table 1: Comparison of running time (in seconds) and recovery error of different methods for solving signal representation problem.

where $\mathbf{A} \in \mathbb{R}^{t\times s}$ is the bases combined dictionary and $\mathbf{c}$ is the input signal. The problem (12) is a one-block case of (1) so that it can be solved by the LADMP. In addition, we choose greedy methods, including MP (Mallat and Zhang 1993), OMP (Tropp and Gilbert 2007), WMP (Temlyakov 2011) for comparison and verify the convergence property of our LADMP with the comparison of the traditional LADM.

In this paper, the synthetic signals used for comparison are designed as: $\mathbf{A}$ is a Gaussian random matrix with the sizes $s = 1024, 2048, 4096, 8192, 16384$ and $t = s/2$. The ideal signal $\mathbf{x}$ is randomly generated using the same strategy of (Li and Osher 2009) and its sparsity is set as the nearest integer of $t/20$. In order to see the relationship between data and parameters, we add a parameter λ to the objective function, which is set differently with different data sizes (see Table 1). Other parameters of LADMP are empirically set as: $\mu_0 = 0.17$, $\nu = 0.1$, $\eta_k^1 = 10^{-5}$, $\eta_k^2 = \mu_k \times 10^{-5}$ and all the algorithms are stopped when $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|/\|\mathbf{x}_k\| < 10^{-7}$.

Given $\mathbf{A}$ and $\mathbf{c}$, we can represent the signal $\widehat{\mathbf{x}}$ by different algorithms. The relative recovery error $\|\widehat{\mathbf{x}} - \mathbf{x}\|/\|\mathbf{x}\|$ is used to measure the performance of representation. The reported error listed in Table 1 is a mean error of 20 trials. In addition, we also list the running time of each algorithm to show the time consuming problem. Then, we can easily see from the Table 1 that when the data size is small ($s = 1024, 2048$), LADMP performs comparative with the best-performed algorithm. However, LADMP outperforms the other methods on both accuracy and efficiency when meeting big data size.

We choose a sample with $s = 8192$ to show the convergence property of our proposed algorithm. Fig. 1(a) gives a visual example on the accuracy of the algorithm. On the other hand, we draw the differences of $\mathbf{x}$ between iterations, i.e. $\log_{10}(\|\mathbf{x}_{k+1} - \mathbf{x}_k\|/\|\mathbf{x}_k\|)$ in Fig. 1(b). To better describe the effectiveness of the proposed algorithm, we compare the convergence curve of LADMP with the one generated by LADM. It can be seen from the figure that applying LADM to problem (12) directly sometimes does not converge.

Furthermore, a speed-up strategy is tailored specifically to solve this signal representation problem (12). Inspired by the strategy of OMP that after getting the support set of the non-zero values, the value of the recovery signal can be easily computed by plain Least Squares. On the other hand, we found that in the early several steps, LADMP can always detect the correct support of the solution; we use the same strategy of OMP to speed up the algorithm. Empirically speaking, LADMP finds the correct support when $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|/\|\mathbf{x}_k\| < 10^{-4}$, which is the stopping criteria of the fast version of LADMP (F-LADMP). Thanks to the correctness of finding the support set of the solution, the recovery errors are extremely small when using F-LADMP. From Table 1, it shows that F-LADMP performs better than LADMP on both speed and recovery performance.

4.2 Image Denoising

Being as an important method for image denoising, sparse coding based models require solving a class of challenging nonconvex and nonsmooth optimization problems and we in this paper use the one with fundamental form:

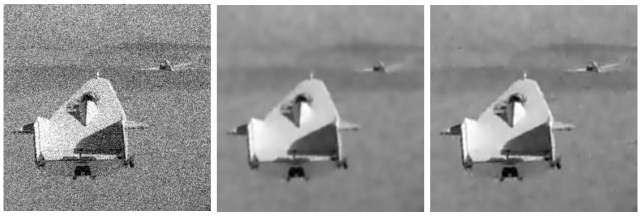
$$\min_{\mathbf{x},\mathbf{y}} \|\mathbf{x}\|_0 + \frac{\gamma}{2}\|\mathbf{y}\|^2, \text{ s.t. } \mathbf{A}\mathbf{x} + \mathbf{y} = \mathbf{c}, \qquad (13)$$

where $\mathbf{A}$ denotes dictionary, $\mathbf{y}$ denotes the corrupted noise in the observed image $\mathbf{c}$. We apply LADMP to problem (13) and compare the performance with IHTA which reformulates the primal problem (13) into unconstrained optimization. The convergence property of IHTA is proved by adding proximal term under proper parameters (Bach et al. 2012).

Similar to the experiment of signal representation, a parameter λ is multiplied on the regular term of the objective function to help coding. All the parameters of LADMP are set as: $\lambda = 1.0$, $\gamma = 0.4$, $\mu_0 = 0.2$, $\eta_k^1 = 10^{-5}$ and $\eta_k^2 = \mu_k$ for dealing with all the images. We list the running time and PSNR values of LADMP and IHTA in Table 2. The added noises in Table 2 are Gaussian randomly noises with level $\sigma_s = 20$. Compared with IHTA, LADMP performs better on removing noises from the images which may result from the approximate solutions obtained through penalizing the constraint with certain parameter. In addition, an example is given in Fig. 2 to show the image denoising performance.

Image	Method	Time(s)	PSNR
Barbara	IHTA	77.009	28.053
	LADMP	**62.636**	**28.814**
Liftingbody	IHTA	61.336	31.653
	LADMP	**60.764**	**32.199**
Pepper	IHTA	73.153	27.658
	LADMP	**53.342**	**28.728**
Child	IHTA	69.832	29.347
	LADMP	**61.062**	**29.977**

Table 2: Comparison of the running time (in seconds) and PSNRs for image denoising.



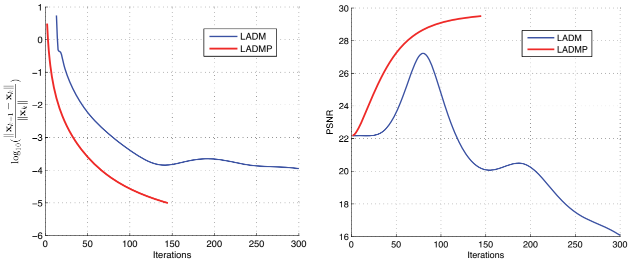
(a) Noisy image (b) IHTA (c) LADMP

Figure 2: An example on the recovery image. (a) noisy image with noisy level $\sigma_s = 30$, recovery image by (b) IHTA with PSNR: 30.201 and (c) LADMP with PSNR: 31.443.

As explained in the previous sections, there is no evidence of the convergence property for directly applying LADM to nonconvex and nonsmooth problem. Hence we conduct another experiment on the comparison of the algorithm itself between directly applying LADM and our proposed LADMP to solve the problem (13). In Fig. 3(a), the convergence curve indicates that our LADMP algorithm converges to an optimal point ($\|\mathbf{x}_{k+1} - \mathbf{x}_k\|/\|\mathbf{x}_k\| < 10^{-5}$) while at the same time the LADM does not converge within the maximum iteration steps (300). In addition, Fig. 3(b) shows the change of PSNR values during iterations which demonstrate the stable and effective performance of our proposed LADMP. point ($\|\mathbf{x}_{k+1} - \mathbf{x}_k\|/\|\mathbf{x}_k\| < 10^{-5}$) while at the same time the LADM does not converge within the maximum iteration steps (300).

5 Conclusion

We first propose LADMP based on ADM for a general nonconvex and nonsmooth optimization. By introducing an auxiliary variable and penalize its bringing constraint to the objective function, we prove that any limit point of our proposed algorithm is a KKT point of the primal problem. In addition, our algorithm is a linearized method that avoids the difficulties of solving subproblems. We start the convergence analysis of LADMP from two-block case and then establish a similar convergence result for multi-block case. Experiments on signal representation and image denoising have shown the effectiveness of our proposed algorithm.



(a) Convergence property

(b) PSNR values

Figure 3: Comparisons between LADM and the proposed LADMP algorithm on both (a) convergence performance and (b) the change of PSNR values during iterations.

6 Acknowledgments

Risheng Liu is supported by the National Natural Science Foundation of China (Nos. 61300086, 61432003), the Fundamental Research Funds for the Central Universities (DUT15QY15) and the Hong Kong Scholar Program (No. XJ2015008). Zhixun Su is supported by National Natural Science Foundation of China (Nos. 61173103, 61572099, 61320106008, 91230103) and National Science and Technology Major Project (Nos. 2013ZX04005021, 2014ZX04001011).

References

Attouch, H.; Bolte, J.; Redont, P.; and Soubeyran, A. 2010. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the kurdyka-lojasiewicz inequality. *Mathematics of Operations Research* 35:438–457.

Bach, F.; Jenatton, R.; Mairal, J.; and Obozinski, G. 2012. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning* 4:1–106.

Bao, C.; Ji, H.; Quan, Y.; and Shen, Z. 2014. ℓ_0 norm based dictionary learning by proximal methods with global convergence. In *CVPR*.

Bolte, J.; Sabach, S.; and Teboulle, M. 2014. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming* 146:459–494.

Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; and Eckstein, J. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* 3:1–122.

Candès, E. J., and Wakin, M. B. 2008. An introduction to compressive sampling. *IEEE Signal Processing Magazine* 25:21–30.

Chen, S., and Donoho, D. 1994. Basis pursuit. In *Conference on Signals, Systems and Computers*.

Chen, C.; He, B.; Ye, Y.; and Yuan, X. 2014. The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming* 1–23.

Goldstein, T.; O'Donoghue, B.; Setzer, S.; and Baraniuk, R. 2014. Fast alternating direction optimization methods. *SIAM J. Imaging Sciences* 7:1588–1623.

Gregor, K., and LeCun, Y. 2010. Learning fast approximations of sparse coding. In *ICML*.

Li, Y., and Osher, S. 2009. Coordinate descent optimization for ℓ_1 minimization with application to compressed sensing; a greedy algorithm. *Inverse Problems and Imaging* 3:487–503.

Liavas, A. P., and Sidiropoulos, N. D. 2014. Parallel algorithms for constrained tensor factorization via the alternating direction method of multipliers. *arXiv preprint*.

Lin, Z.; Liu, R.; and Li, H. 2013. Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning. *Machine Learning* 99:287–325.

Lin, Z.; Liu, R.; and Su, Z. 2011. Linearized alternating direction method with adaptive penalty for low-rank representation. In *NIPS*, 612–620.

Liu, R.; Lin, Z.; Su, Z.; and Gao, J. 2014. Linear time principal component pursuit and its extensions using ℓ_1 filtering. *Neurocomputing* 142:529–541.

Mallat, S. G., and Zhang, Z. 1993. Matching pursuits with time-frequency dictionaries. *IEEE TIP* 41:3397–3415.

Nocedal, J., and Wright, S. 2006. *Numerical Optimization*. Springer Science & Business Media.

Temlyakov, V. 2011. *Greedy Approximation*. Cambridge University Press.

Tropp, J., and Gilbert, A. C. 2007. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory* 53:4655–4666.

Tseng, P. 1991. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Control and Optimization* 29:119–138.

Wang, Y.; Liu, R.; Song, X.; and Su, Z. 2014. Saliency detection via nonlocal ℓ_0 minimization. In *ACCV*. 521–535.

Wang, F.; Xu, Z.; and Xu, H. K. 2014. Convergence of bregman alternating direction method with multipliers for nonconvex composite problems. *arXiv preprint*.

Wright, J.; Ma, Y.; Mairal, J.; Sapiro, G.; Huang, T. S.; and Yan, S. 2010. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE* 98:1031–1044.

Xu, Y., and Yin, W. 2014. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *arXiv preprint*.

Xu, Y.; Yin, W.; Wen, Z.; and Zhang, Y. 2012. An alternating direction algorithm for matrix completion with nonnegative factors. *technical report, Shanghai Jiaotong University*.

Yang, J.; Yu, K.; Gong, Y.; and Huang, T. 2009. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*.

Yin, W. 2010. Analysis and generalizations of the linearized bregman method. *SIAM J. Imaging Sciences* 3:856–877.

Yuan, G., and Ghanem, B. 2013. $\ell_0$tv: A new method for image restoration in the presence of impulse noise. In *CVPR*.

Zuo, W., and Lin, Z. 2011. A generalized accelerated proximal gradient approach for total-variation-based image restoration. *IEEE TIP* 20:2748–2759.