Convergence Analysis of ADMM for a Family of Nonconvex Problems

Mingyi Hong* mingyi@iastate.edu
Zhi-Quan Luo† luozq@umn.edu
Meisam Razaviyayn‡ meisamr@stanford.edu

Abstract

In this paper, we analyze the behavior of the well-known alternating direction method of multipliers (ADMM), for solving a family of nonconvex problems. Our focus is given to the well-known *consensus* and *sharing* problems, both of which have wide applications in machine learning. We show that in the presence of nonconvex objective, the classical ADMM is able to reach the set of stationary solutions for these problems, if the stepsize is chosen large enough. An interesting consequence of our analysis is that the ADMM is convergent for a family of sharing problems, regardless of the number of blocks or the convexity of the objective function. Our analysis can be generalized to allow proximal update rules as well as other flexible block selection rules far beyond the traditional Gauss-Seidel rule.

1 Introduction

Consider the following linearly constrained (possibly nonsmooth or/and nonconvex) problem with K blocks of variables $\{x_k\}_{k=1}^K$:

$$\begin{aligned} \min \quad & f(x) := \sum_{k=1}^K g_k(x_k) + \ell(x_1, \cdots, x_K) \\ \text{s.t.} \quad & \sum_{k=1}^K A_k x_k = q, \ x_k \in X_k, \ \forall\, k = 1, \cdots, K \end{aligned} \tag{1.1}$$

where $A_k \in \mathbb{R}^{M \times N_k}$ and $q \in \mathbb{R}^M$; $X_k \in \mathbb{R}^{N_k}$ is a closed convex set; $\ell(\cdot)$ is a smooth (possibly nonconvex) function; each $g_k(\cdot)$ can be either a smooth function, or a convex nonsmooth function. The augmented Lagrangian for problem (1.1) is given by

$$L(x; y) = \sum_{k=1}^K g_k(x_k) + \ell(x_1, \cdots, x_K) + \langle y, q - Ax \rangle + \frac{\rho}{2}\|q - Ax\|^2, \tag{1.2}$$

where $\rho > 0$ is a constant representing the primal step-size.

To solve problem (1.1), consider the popular alternating direction method of multipliers (ADMM) displayed below:

*Industrial and Manufacturing Systems Engineering, Iowa State University
†Department of Electrical and Computer Engineering, University of Minnesota
‡Electrical Engineering Department, Stanford University

Algorithm 0. ADMM for Problem (1.1)

At each iteration $t+1$, update the primal variables:

$$x_k^{t+1} = \arg\min_{x_k \in X_k} L(x_1^{t+1}, \cdots, x_{k-1}^{t+1}, x_k, x_{k+1}^t, \cdots, x_K^t; y^t), \; \forall\, k.$$

Update the dual variable:

$$y^{t+1} = y^t + \rho(q - Ax^{t+1}).$$

The ADMM algorithm was originally introduced in early 1970s [1, 2], and has since been studied extensively [3–6]. Recently it has become popular in big data related problems arising in various engineering domains; see, e.g., [7–14]. There is a vast literature that applies the ADMM for all sorts of problems in the form of (1.1). Most of its convergence analysis is done for certain special form of problem (1.1) — the *two-block convex separable* problems, where $K = 2$, $\ell = 0$ and g_1, g_2 are both convex. In this case, ADMM is known to converge under very mild conditions; see [6] and [7]. Recent analysis on its rate of convergence can be found in [15–19].

Unlike the convex case, the behavior of the ADMM is rarely analyzed when it is applied to solve nonconvex problems. Nevertheless, it has been observed by many researchers that the ADMM works very well for various applications involving nonconvex objectives, such as the nonnegative matrix factorization, phase retrieval, distributed matrix factorization etc.; see [20–29] and the references therein. However, to the best of our knowledge, existing convergence analysis of ADMM for nonconvex problems is very limited — most of the known global convergence analysis needs to impose requirements on the sequence generated by the algorithm. Unfortunately these requirements are nonstandard and overly restrictive. Reference [30] analyzes a family of splitting algorithms (which includes ADMM as a special case) for certain nonconvex quadratic problem, and shows that they converge to the stationary solution when certain condition on the dual stepsize is met.

In this paper, we analyze the convergence of ADMM for two special types of nonconvex problems in the form of (1.1). Our focus is given to a family of nonconvex consensus and sharing problems, and show that ADMM converges without any assumptions on the iterates – as long as the problem (1.1) satisfies certain regularity conditions, and the stepsize ρ is chosen large enough (with computable bounds), then the algorithm is guaranteed to converge to the set of stationary solutions.

2 The Nonconvex Consensus Problem

Consider the following nonconvex consensus problem

$$\min \quad f(x) := \sum_{k=1}^{K} g_k(x) + h(x) \quad \text{s.t.} \quad x \in X \tag{2.3}$$

where each g_k is a smooth but possibly nonconvex functions; $h(x)$ is a convex possibly nonsmooth function. This problem is related to the convex consensus problem discussed in [7, Section 7], but with the important difference that g_k can be nonconvex.

In many practical applications, each g_k is handled by a single agent, such as a thread or processor. This motivates the following consensus formulation. Let us introduce a set of new variables $\{x_k\}_{k=1}^K$, and transform problem (2.3) equivalently to the following linearly constrained problem

$$\min \quad \sum_{k=1}^{K} g_k(x_k) + h(x) \quad \text{s.t.} \quad x_k = x, \; \forall\, k = 1, \cdots, K, \quad x \in X. \tag{2.4}$$

The augmented Lagrangian function is given by

$$L(\{x_k\}, x; y) = \sum_{k=1}^{K} g_k(x_k) + h(x) + \sum_{k=1}^{K} \langle y_k, x_k - x \rangle + \sum_{k=1}^{K} \frac{\rho_k}{2} \|x_k - x\|^2. \tag{2.5}$$

Problem (2.4) can be solved distributedly by applying the classical ADMM algorithm. The details are given in the table below.

Algorithm 1. The Classical ADMM for the Consensus Problem (2.4)
At each iteration $t+1$, compute:

$$x^{t+1} = \operatorname*{argmin}_{x\in X} L(\{x_k^t\}, x; y^t).$$

Each node k computes x_k in parallel, by solving:

$$x_k^{t+1} = \arg\min_{x_k} g_k(x_k) + \langle y_k^t, x_k - x^{t+1}\rangle + \frac{\rho_k}{2}\|x_k - x^{t+1}\|^2.$$

Each node k updates the dual variable:

$$y_k^{t+1} = y_k^t + \rho_k\left(x_k^{t+1} - x^{t+1}\right).$$

In Algorithm 1, the x update step can be expressed as

$$x^{t+1} = \text{prox}_{\iota(X)+h}\left[\frac{\sum_{k=1}^K \rho_k x_k^t + \sum_{k=1}^K y_k^t}{\sum_{k=1}^K \rho_k}\right] \tag{2.6}$$

where prox_p is the *proximity operator* of a convex function $p(\cdot)$ [31, Section 31]. Note that x can be viewed as the first block and $\{x_k\}_{k=1}^K$ together is the second block. Therefore the two primal blocks are updated in a sequential (i.e., Gauss-Seidel) manner. In this paper we will analyze a more general version, in which the blocks are updated in a *flexible* manner; see Algorithm 2.

We consider the following two types of variable block update order rules: let $k = 1, 2, ..., K$ be the indices for the primal variable blocks $x_1, x_2, ..., x_K$ and $k = 0$ be the index for primal variable block x, and let $\mathcal{C}^t \subseteq \{0, 1, \cdots, K\}$ denote the set of variables updated in iteration t, then

1. *Randomized update rule*: At each iteration $t+1$ the indices are chosen randomly and independently from the previous iterations, i.e.,
$$\Pr\left(k \in \mathcal{C}^{t+1} \mid x^t, y^t, \{x_k^t\}\right) = p_k^{t+1} \geq p_{\min} > 0. \tag{2.7}$$
2. *Essentially cyclic (EC) update rule*: There exists a given period $T \geq 1$ during which each index is updated at least once, i.e.,$\bigcup_{i=1}^T \mathcal{C}^{t+i} = \{0, 1, \cdots, K\},\ \forall\, t.$
We call this update rule a *period-T* EC rule.

Algorithm 2. The Flexible ADMM for the Consensus Problem (2.4)
At each iteration $t+1$, pick an index set $\mathcal{C}^{t+1} \subseteq \{0, \cdots, K\}$.
If $0 \in \mathcal{C}^{t+1}$, compute:

$$x^{t+1} = \operatorname*{argmin}_{x\in X} L(\{x_k^t\}, x; y^t). \tag{2.8}$$

Else $x^{t+1} = x^t$.
If $k \in \mathcal{C}^{t+1}$, node k computes x_k by solving:

$$x_k^{t+1} = \arg\min_{x_k} g_k(x_k) + \langle y_k^t, x_k - x^{t+1}\rangle + \frac{\rho_k}{2}\|x_k - x^{t+1}\|^2.$$

Update the dual variable:

$$y_k^{t+1} = y_k^t + \rho_k\left(x_k^{t+1} - x^{t+1}\right).$$

Else $x_k^{t+1} = x_k^t$, $y_k^{t+1} = y_k^t$.

Clearly Algorithm 1 is simply Algorithm 2 with period-1 EC rule. Therefore we will focus on analyzing Algorithm 2. To this end, we make the following assumption.

Assumption A.

A1. There exists a positive constant $L_k > 0$ such that
$$\|\nabla_k g_k(x_k) - \nabla_k g_k(z_k)\| \leq L_k\|x_k - z_k\|,\ \forall\, x_k, z_k,\ \forall\, k.$$
Moreover, h is convex (possible nonsmooth); X is a closed convex set.

A2. For all k, the stepsize ρ_k is chosen large enough such that:
   1. For all k, the x_k subproblem is strongly convex with modulus $\gamma_k(\rho_k)$;
   2. For all k, $\rho_k\gamma_k(\rho_k) > 2L_k^2$ and $\rho_k \geq L_k$.

A3. $f(x)$ is lower bounded for all $x \in X$.

Clearly, assumption A does not impose any restriction on the *iterates* generated by the algorithm. This is in contrast to the existing analysis of the nonconvex ADMM algorithms, such as those developed in [20, 26, 28].

Now we state the first main result of this paper. We briefly mention that the key of the proof is to use the *reduction of the augmented Lagrangian* to measure the progress of the algorithm.

Theorem 2.1 *Assume that Assumption A is satisfied. Then the following is true for Algorithm 2:*

1. $\lim_{t\to\infty} \|x_k^{t+1} - x^{t+1}\| = 0,\ \forall, k$, *deterministically for the EC rule and almost surely (a.s.) for randomized rule.*
2. *Let* $(\{x_k^*\}, x^*, y^*)$ *denote any limit point of the sequence* $\{\{x_k^{t+1}\}, x^{t+1}, y^{t+1}\}$ *generated by Algorithm 2. Then the following statement is true (deterministically for the EC rule and a.s. for the randomized update rule)*
$$0 = \nabla g_k(x_k^*) + y_k^*, \quad x_k^* = x^*,\ k = 1, \cdots, K,\ x^* \quad \in \arg\min_{x\in X} \quad h(x) + \sum_{k=1}^{K}\langle y_k^*, x_k^* - x\rangle$$
That is, any limit point of Algorithm 2 is a stationary solution of problem (2.4)*.*
3. *If X is a compact set, then Algorithm 2 converges to the set of stationary solutions of problem* (2.4)*.*

3 The Nonconvex Sharing Problem

Consider the following well-known sharing problem (see, e.g., [7, Section 7.3] for motivation)
$$\min \quad f(x_1, \cdots, x_K) := \sum_{k=1}^{K} g_k(x_k) + \ell\left(\sum_{k=1}^{K} A_k x_k\right), \text{ s.t. } \quad x_k \in X_k,\ k = 1, \cdots, K \tag{3.9}$$
where $x_k \in \mathbb{R}^{N_k}$ is the variable associated with a given agent k, and $A_k \in \mathbb{R}^{M\times N_k}$ is some data matrix. The variables are coupled through the function $\ell(\cdot)$.

To facilitate distributed computation, this problem can be equivalently formulated as:
$$\min \quad \sum_{k=1}^{K} g_k(x_k) + \ell(x) \text{ s.t. } \quad \sum_{k=1}^{K} A_k x_k = x, \quad x_k \in X_k,\ k = 1, \cdots, K. \tag{3.10}$$
We make the following assumptions.

Assumption B.

B1. There exists a positive constant $L > 0$ such that
$$\|\nabla\ell(x) - \nabla\ell(z)\| \leq L\|x - z\|,\ \forall\, x, z.$$
Moreover, X_k's are closed convex sets; each A_k is full column rank, $\rho_{\min}(A_k^T A_k) > 0$.

B2. The stepsize ρ is chosen large enough such that:
   (1) each x_k subproblem as well as the x subproblem is strongly convex, with modulus $\{\gamma_k(\rho)\}_{k=1}^K$ and $\gamma(\rho)$, respectively.
   (2) $\rho\gamma(\rho) > 2L^2$, and that $\rho \geq L$.

B3. $f(x_1, \cdots, x_K)$ is lower bounded for all $x_k \in X_k$ and all k.

B4. g_k is either smooth nonconvex or convex (possibly nonsmooth). For the former case, there exists $L_k > 0$ such that $\|g_k(x_k) - g_k(z_k)\| \leq L_k\|x_k - z_k\|$, $\forall\, x_k, z_k \in X_k$.

Again one can show that when Assumption B is satisfied, then the a flexible ADMM similar to Algorithm 2 will converge to the set of stationary solutions of problem (3.10). To conclude, we provide a remark on generalizing the flexible ADMM to include proximal steps.

Remark 3.1 *In certain applications it is beneficial to have cheap updates for the subproblems. The flexible ADMM can be further generalized to the case where the subproblems are not solved exactly – only a single proximal update is sufficient for each x_k subproblem.*

References

[1] R. Glowinski and A. Marroco. Sur l'approximation, par elements finis d'ordre un,et la resolution, par penalisation-dualite, d'une classe de problemes de dirichlet non lineares. *Revue Franqaise d'Automatique, Informatique et Recherche Opirationelle*, 9:41–76, 1975.

[2] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2:17–40, 1976.

[3] J. Eckstein. Splitting methods for monotone operators with applications to parallel optimization. 1989. Ph.D Thesis, Operations Research Center, MIT.

[4] J. Eckstein and D. P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992.

[5] R. Glowinski. *Numerical methods for nonlinear variational problems*. Springer-Verlag, New York, 1984.

[6] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods, 2nd ed*. Athena Scientific, Belmont, MA, 1997.

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

[8] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for l1-minimization with applications to compressed sensing. *SIAM Journal on Imgaging Science*, 1(1):143–168, March 2008.

[9] J. Yang, Y. Zhang, and W. Yin. An efficient TVL1 algorithm for deblurring multichannel images corrupted by impulsive noise. *SIAM Journal on Scientific Computing*, 31(4):2842–2865, 2009.

[10] X. Zhang, M. Burger, and S. Osher. A unified primal-dual algorithm framework based on Bregman iteration. *Journal of Scientific Computing*, 46(1):20–46, 2011.

[11] K. Scheinberg, S. Ma, and D. Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In *Twenty-Fourth Annual Conference on Neural Information Processing Systems (NIPS)*, 2010.

[12] I. Schizas, A. Ribeiro, and G. Giannakis. Consensus in ad hoc wsns with noisy links - part i: Distributed estimation of deterministic signals. *IEEE Transactions on Signal Processing*, 56(1):350C364, 2008.

[13] C. Feng, H. Xu, and B. Li. An alternating direction method approach to cloud traffic management. 2014. preprint.

[14] W.-C. Liao, M. Hong, Hamid Farmanbar, Xu Li, Z.-Q. Luo, and Hang Zhang. Min flow rate maximization for software defined radio access networks. *IEEE Journal on Selected Areas in Communication*, 32(6):1282–1294, 2014.

[15] B. He and X. Yuan. On the O(1/n) convergence rate of the Douglas-Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

[16] R. Monteiro and B. Svaiter. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization*, 23(1):475–507, 2013.

[17] T. Goldstein, B. O'Donoghue, and S. Setzer. Fast alternating direction optimization methods. *UCLA CAM technical report*, 2012.

[18] D. Goldfarb, S. Ma, and K. Scheinberg. Fast alternating linearization methods for minimizing the sum of two convex functions. *Mathematical Programming*, 141(1-2):349–382, 2012.

[19] W. Deng and W. Yin. On the global linear convergence of alternating direction methods. 2012. preprint.

[20] Y. Zhang. An alternating direction algorithm for nonnegative matrix factorization. 2010. Preprint.

[21] D. L. Sun and C. Fevotte. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In *the Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[22] Z. Wen, C. Yang, X. Liu, and S. Marchesini. Alternating direction methods for classical and ptychographic phase retrieval. *Inverse Problems*, 28(11):1–18, 2012.

[23] Q. Ling, Y. Xu, W. Yin, and Z. Wen. Decentralized low-rank matrix completion. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2925–2928, March 2012.

[24] P.A. Forero, A. Cano, and G.B. Giannakis. Distributed clustering using wireless sensor networks. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):707–724, Aug 2011.

[25] B. Ames and M. Hong. Alternating directions method of multipliers for l1-penalized zero variance discriminant analysis and principal component analysis. Preprint.

[26] B. Jiang, S. Ma, and S. Zhang. Alternating direction method of multipliers for real and complex polynomial optimization models. 2013. Preprint.

[27] Y. Shen, Z. Wen, and Y. Zhang. Augmented lagrangian alternating direction method for matrix separation based on low-rank factorization. *Optimization Methods Software*, 29(2):239–263, March 2014.

[28] Y. Xu, W. Yin, Z. Wen, and Y. Zhang. An alternating direction algorithm for matrix completion with nonnegative factors. *Journal of Frontiers of Mathematics in China, Special Issues on Computational Mathematics*, pages 365–384, 2011.

[29] Z. Wen, X. Peng, X. Liu, X. Bai, and X. Sun. Asset allocation under the basel accord risk measures. 2013. Preprint.

[30] Y. Zhang. Convergence of a class of stationary iterative methods for saddle point problems. 2010. Preprint.

[31] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.