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 $\{x_k\}_{k=1}^K$:

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

(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

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

(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:
Algorithm 0. ADMM for Problem (1.1)

At each iteration \( t + 1 \), update the primal variables:

\[
x_{k}^{t+1} = \arg \min_{x_k \in X_k} \mathcal{L}(x_1^{t+1}, \ldots, x_{k-1}^{t+1}, x_k, x_{k+1}^{t}, \ldots, x_K^{t}; y^t), \quad \forall \ k.
\]

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 \( \rho \) 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

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

(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

\[
\min \sum_{k=1}^{K} g_k(x_k) + h(x) \quad \text{s.t.} \quad x_k = x, \quad \forall \ k = 1, \ldots, K, \quad x \in X.
\]  

(2.4)

The augmented Lagrangian function is given by

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

(2.5)

Problem (2.4) can be solved distributedly by applying the classical ADMM algorithm. The details are given in the table below.
where prox $p$ Assumption A.

To this end, we make the following assumption.

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 (possibly nonsmooth); $X$ is a closed convex set.
A2. For all $k$, the stepsize $\rho_k$ 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 $k$, $k \gamma_k(\rho_k) > 2L_k^2$ 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 \( \{ x_k^0, x^*, y^* \} \) 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)
   \[ 0 = \nabla g_k(x_k^*) + y^*, \quad x_k^* = x^*, \quad k = 1, \ldots, K, \quad x^* \in \arg \min_{x \in X} h(x) + \sum_{k=1}^{K} \langle y_k^*, x_k^* - x \rangle \]
   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)

\[
\min_{x_1, \ldots, x_K} f(x_1, \ldots, x_K) := \sum_{k=1}^{K} g_k(x_k) + \ell \left( \sum_{k=1}^{K} A_k x_k \right), \quad \text{s.t.} \quad x_k \in X_k, \quad k = 1, \ldots, K \quad (3.9)
\]

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:

\[
\min_{x_1, \ldots, x_K} \sum_{k=1}^{K} g_k(x_k) + \ell(x) \quad \text{s.t.} \quad \sum_{k=1}^{K} A_k x_k = x, \quad x_k \in X_k, \quad k = 1, \ldots, K. \quad (3.10)
\]

We make the following assumptions.

**Assumption B.**

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

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

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

B4. $g_k(x_k)$ 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


