Convergence Analysis of Inexact Over-relaxed ADMM via Dissipativity Theory

Qiang Zhou ZHOUQIANG@U.NUS.EDU
Southeast University, Nanjing, Jiangsu 211189, China
Purple Mountain Laboratories, Nanjing, Jiangsu 211111, China

Sinno Jialin Pan SINNOPAN@CUHK.EDU.HK
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

We present a new convergence analysis for the over-relaxed alternating direction method of multipliers (ADMM) when the subproblem cannot be exactly solved, *i.e.*, inexact over-relaxed ADMM. Our method builds on (Hu and Lessard, 2017) that relates the convergence analysis of optimization algorithms to the stability of a discrete-time linear dynamic system. By expressing the inexact over-relaxed ADMM as a discrete-time linear dynamic system, we show that both the linear and sublinear convergence of inexact over-relaxed ADMM can be obtained by solving or verifying the feasibility of a small semidefinite program (SDP). More importantly, we prove that the associated SDP has an analytical solution for various parameters. We demonstrate the theoretical result by applying the inexact over-relaxed ADMM to solve a distributed ℓ_1-norm regularized logistic regression problem.

Keywords: Over-relaxed ADMM; Inexact Solution, Dissipativity Theory

1. Introduction

The alternating direction method of multipliers (ADMM) is a classic optimization algorithm and traces back to (Gabay and Mercier, 1976; Lions and Mercier, 1979). In the last decade, the alternating direction method of multipliers (ADMM) (Boyd et al., 2011; Eckstein and Yao, 2015; Xu et al., 2017; França et al., 2018) has garnered attention and interest in the machine learning community because of its simplicity and applicability to a wide variety of large-scale distributed convex optimization problems (Deng and Yin, 2016; Nishihara et al., 2015; Zhang et al., 2012; Wang et al., 2013; Zhang and Kwok, 2014; Xu et al., 2017; Liu et al., 2021; Zhou and Li, 2023). Specifically, it solves the problems in the form of

$$\min_{\mathbf{x},\mathbf{z}} f(\mathbf{x}) + g(\mathbf{z}), \quad \text{s.t. } \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{z} \in \mathbb{R}^q$, $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times q}$ and $\mathbf{c} \in \mathbb{R}^p$. In general, $f(\mathbf{x})$ is a loss function (e.g., logistic loss), and $g(\mathbf{z})$ is some regularization function (e.g., ℓ_1-norm) which is typically used to promote a certain structure (e.g., sparsity) in the optimal solution. Many convex optimization problems in machine learning and statistics (Bubeck, 2015; Sra et al., 2012; Boyd et al., 2011) can be formulated in the form of problem (1).

In order to solve the problem (1) via ADMM, we first introduce Lagrange multiplier $\rho\mathbf{u} \in \mathbb{R}^p$ for (1) where $\rho > 0$ is the penalty parameter. Then the augmented Lagrangian $L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{u})$ is defined as following

$$L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{u}) = f(\mathbf{x}) + g(\mathbf{z}) - \rho\langle \mathbf{u}, \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\rangle + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\|^2. \tag{2}$$

Then, standard ADMM (Boyd et al., 2011; Gabay and Mercier, 1976; Eckstein and Bertsekas, 1992) performs following three updates at each iteration:

$$\mathbf{x}_{k+1} \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{x}} L_\rho(\mathbf{x}, \mathbf{z}_k, \mathbf{u}_k), \tag{3a}$$

$$\mathbf{z}_{k+1} \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{z}} L_\rho(\mathbf{x}_{k+1}, \mathbf{z}, \mathbf{u}_k), \tag{3b}$$

$$\mathbf{u}_{k+1} \stackrel{\text{def}}{=} \mathbf{u}_k - \big(\mathbf{A}\mathbf{x}_{k+1} + \mathbf{B}\mathbf{z}_{k+1} - \mathbf{c}\big), \tag{3c}$$

In this work, we consider a popular variant of standard ADMM, named over-relaxed ADMM, which introduces a relaxation parameter to improve the convergence of ADMM (Eckstein and Bertsekas, 1992). Specifically, the over-relaxed ADMM consists of the iterations

$$\mathbf{x}_{k+1} \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{x}} \Big\{ f(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}_k - \mathbf{c} - \mathbf{u}_k\|^2 \Big\}, \tag{4a}$$

$$\mathbf{z}_{k+1} \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{z}} \Big\{ g(\mathbf{z}) + \frac{\rho}{2}\|\alpha\mathbf{A}\mathbf{x}_{k+1} - (1-\alpha)\mathbf{B}\mathbf{z}_k + \mathbf{B}\mathbf{z} - \alpha\mathbf{c} - \mathbf{u}_k\|^2 \Big\}, \tag{4b}$$

$$\mathbf{u}_{k+1} \stackrel{\text{def}}{=} \mathbf{u}_k - \big(\alpha\mathbf{A}\mathbf{x}_{k+1} - (1-\alpha)\mathbf{B}\mathbf{z}_k + \mathbf{B}\mathbf{z}_{k+1} - \alpha\mathbf{c}\big), \tag{4c}$$

where the relaxation parameter α is typically chosen from $\alpha \in (0, 2]$. If $\alpha = 1$, the algorithm reduces to standard ADMM. It is called over-relaxed ADMM when $\alpha > 1$ (Boyd et al., 2011). Generally, the over-relaxed ADMM can improve the convergence of ADMM (Eckstein and Bertsekas, 1992; Eckstein, 1994). In this work, our analysis covers all choices $\alpha \in (0, 2]$ even in a larger range of α.

ADMM is an appealing approach for solving distributed machine learning problems in which the datasets are stored locally on different workers (Boyd et al., 2011; Nishihara et al., 2015; Zhang and Kwok, 2014; Xu et al., 2017; Zhou and Li, 2023). It is straightforward to show that (4a) can be decomposed into several independent problems that can be simultaneously and independently solved by only accessing the local dataset on each particular worker. This is highly desirable in the era of big data. More importantly, a larger number of practical experiences suggest that ADMM often quickly converges to modest accuracy solutions, which is particularly useful in machine learning applications (Boyd et al., 2011).

While the ADDM has shown great success in practical applications, the theoretical understanding of over-relaxed ADMM remains limited. Most of the works in the literature analyze the convergence of ADMM by assuming (4a) and (4b) must be exactly solved at every iteration (Nishihara et al., 2015; Monteiro and Svaiter, 2013; Goldstein et al., 2014; Xu et al., 2017; França et al., 2018). Indeed, we can efficiently solve them to obtain the exact solutions $\mathbf{x}_{k+1}$ and $\mathbf{z}_{k+1}$ for several notable choices of $f(\mathbf{x})$ and $g(\mathbf{z})$, e.g., $g(\mathbf{z}) = \|\mathbf{z}\|_1$ (Tibshirani, 1996). However, in many important scenarios (4a) and (4b) may not admit an analytic solution, or it may be computationally expensive to exactly solve them. Examples include sparse logistic regression (Boyd et al., 2011), total-variation (Fadili and Peyré, 2011; Ng et al., 2010; Wang et al., 2008) and overlapping group lasso (Yuan et al., 2011). Therefore, it is important to study the convergence of ADMM when (4a) and (4b) are not exactly solved (Gol'shtein and Tret'yakov, 1979; Eckstein and Bertsekas, 1992).

When it is difficult to exactly solve (4a) and (4b), existing works mainly present two different methods to obtain approximate solutions $\mathbf{x}_{k+1}$ and $\mathbf{z}_{k+1}$.

- One method is to find an approximate solution of (4a) and (4b) by using a generic solver (e.g., L-BFGS, first-order gradient method) to solve (4a) and (4b) or their dual problems (Boyd et al., 2011; Fadili and Peyré, 2011; Ng et al., 2010).

- The other method seeks to construct an upper bound problem for (4a) and (4b) in which the upper bound admits an analytical solution (Zhang et al., 2010, 2011; He and Yuan, 2012; Lin et al., 2011). By doing that, one can easily obtain approximate solutions $\mathbf{x}_{k+1}$ and $\mathbf{z}_{k+1}$ by solving the upper bound problem since they admit an analytical solution.

Nevertheless, the assumption of the optimality of $\mathbf{x}_{k+1}$ and $\mathbf{z}_{k+1}$ for convergence analysis is invalid in both methods. It has been empirically observed that the ADMM still converges even when the (4a) and (4b) are approximately but carefully solved (Boyd et al., 2011). Therefore, it is interesting to theoretically study the convergence of inexact over-relaxed ADMM (He et al., 2002; Eckstein and Bertsekas, 1992; Gol'shtein and Tret'yakov, 1979).

Motivated by these recent advances (Lessard et al., 2016; Hu and Lessard, 2017) on convergence analysis of first-order methods, we present a new general convergence analysis for over-relaxed ADMM by allowing (4a) and (4b) to be approximately solved. Specifically, our analysis is based on representing an optimization algorithm as a discrete-time dynamical system (Hu and Lessard, 2017; Lessard et al., 2016). By expressing the inexact over-relaxed ADMM as a discrete-time linear dynamic system, we show that both the linear and sublinear convergence of inexact over-relaxed ADMM can be obtained by solving or verifying the feasibility of a small semidefinite program (SDP). More importantly, we further prove that the associated SDP has an analytical solution for various parameters.

1.1. Related Work

In the seminal work, Lessard et al. (2016) present the integral quadratic constraints (IQC) framework to analyze optimization algorithms by casting them as discrete-time linear dynamical systems. In this way, the convergence analysis of optimization algorithms is converted into the stability analysis of the associated dynamic systems, which is equivalent to verifying the feasibility of a SDP. One limitation of the IQC framework is the SDP generally does not admit an analytical solution, thus the converge rates of optimization algorithms can only be obtained by numerically solving the SDP.

Later, Hu and Lessard (2017) improve the IQC framework by drawing a connection between IQC with the discretization approach (Wilson et al., 2021). Specifically, they use the concept of dissipativity theory from control to provide a natural understanding of Nesterov's accelerated method. Consequently, the method based on dissipativity theory leads to a simpler SDP than the one in (Lessard et al., 2016) and it can be solved analytically. Furthermore, Hu and Lessard (2017) show that the dissipativity theory can be used to obtain the sublinear rate for algorithms with general convex objectives while the IQC is tailored for analyzing linear rates of algorithms that require the objective to be strongly convex.

Both Lessard et al. (2016) and Hu and Lessard (2017) focus on the convergence analysis of first-order methods and Nesterov's accelerated method. Inspired by (Lessard et al., 2016), Nishihara et al. (2015) also present a convergence analysis for over-relaxed ADMM by relating it to verify the stability of a dynamic system. They show that the convergence rate of over-relaxed ADMM can be obtained by numerically solving a SDP. However, their analysis still exists following two limitations:

- Nishihara et al. (2015) assume subproblems (4a) and (4b) must be exactly solved that is invalid in many real applications. For example, for large-scale distributed optimization, the subproblem (4a) is often distributively solved by different workers that have various computational abilities. Thus, assuming (4a) has been exactly solved may lead to a significant delay due to the existing stragglers in distributed optimization.

- The method presented by Nishihara et al. (2015) employ a dynamic system with *time-invariant* state transition matrices to represent the over-relaxed ADMM algorithm. Thus, their method only works for the analysis of linear convergence rates that require the objective to be strongly convex. On the other hand, many machine learning problems are general but not strongly convex (Bach et al., 2012).

Both these two assumptions are too restrictive in many machine learning applications. In contrast, we present a new convergence analysis for over-relaxed ADMM by removing the above two assumptions. Specifically, our method allows the (4a) and (4b) to be approximately solved in some extent. Furthermore, the proposed method can also be extended to obtain the sublinear convergence rate for general convex objectives.

1.2. Organization and Contributions of This Paper

We first introduce the notation and preliminaries in Section 2. We then present convergence analysis for inexact over-relaxed ADMM with strongly convex objectives by using the dissipativity theory in Section 3, and extend the proposed method extends the method to the case of generally convex objectives in Section 4. Finally, we demonstrate our analysis by experiments in Section 5. In summary, this work makes the following contributions.

- We provide a new convergence analysis for over-relaxed ADMM by allowing the subproblems (4a) and (4b) to be approximately solved. Specifically, Section 3.1 shows that the inexact over-relaxed ADMM can be represented as a discrete-time linear dynamical system. We then prove the linear convergence rate of inexact over-relaxed ADMM in Section 3.2 by applying dissipativity theory to the dynamic system. In this way, the proposed method provides a natural understanding of inexact over-relaxed ADMM by adapting the intuitive notion of energy dissipation. Unlike (Nishihara et al., 2015), our analysis leads to a simpler SDP that can be solved analytically.

- In Section 4, we extend our analysis to the case of $f(\mathbf{x})$ being generally convex by choosing an appropriate supply rate to define the dynamical system. In contrast, the method of (Nishihara et al., 2015) is only tailored for strongly convex $f(\mathbf{x})$. In this case, we prove that the sublinear convergence rate of inexact over-relaxed ADMM can be obtained by verifying the feasibility of an SDP. More importantly, we further prove that the SDP is feasible for a wide range of α and ρ.

- In Section 5, we demonstrate our theoretical results by applying inexact over-relaxed ADMM to solve ℓ_1-norm regularized logistic regression. It is worth noting that the analysis of (Nishihara et al., 2015) is not applicable to this case as neither the optimality of $\mathbf{x}_{k+1}$ nor strong convexity of $f(\mathbf{x})$ is valid. In contrast, our Theorem 9 provides a convergence guarantee for this case.

2. Preliminaries and Notation

Throughout this paper, vectors and matrices are denoted by lower- and upper-case boldface characters (e.g., $\mathbf{a}$ and $\mathbf{A}$), respectively. We use a_i and A_{ij} to denote the i-th entry of $\mathbf{a}$ and the (i,j)-th entry of $\mathbf{A}$, respectively. Let $\mathbf{0}$ denote a vector of zeros. Let $\mathbf{I}_p$ and $\mathbf{0}_p$ denote the $p \times p$ identity and zero matrices, respectively. For $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{p\times q}$, their Kronecker product is denoted by $\mathbf{A} \otimes \mathbf{B}$ and satisfies the properties $(\mathbf{A} \otimes \mathbf{B})^\top = \mathbf{A}^\top \otimes \mathbf{B}^\top$ and $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$ when the matrices have compatible dimensions. For matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, the largest and smallest singular values are denoted by $\sigma_{\mathbf{A}}^{\max}$ and $\sigma_{\mathbf{A}}^{\min}$, respectively. Then, we use $\kappa_{\mathbf{A}} \stackrel{\text{def}}{=} \sigma_{\mathbf{A}}^{\max}/\sigma_{\mathbf{A}}^{\min}$ to denote the condition number of $\mathbf{A}$. We use $\mathbb{S}_{++}^p$ to denote the set of symmetric positive definite $p \times p$ matrices.

For a generic norm $\|\cdot\|$, we say that function f is m-strongly convex w.r.t. $\|\cdot\|$ if

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{m}{2}\|\mathbf{x}-\mathbf{y}\|^2, \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^p. \tag{5}$$

We say that function f is L-smooth w.r.t. $\|\cdot\|$ if

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2, \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^p. \tag{6}$$

For $0 < m \leq L$, let $\mathcal{S}_p(m, L)$ denote the set of differentiable convex functions $f : \mathbb{R}^p \to \mathbb{R}$ that are m-strongly convex and L-smooth w.r.t. $\|\cdot\|$. If $f \in \mathcal{S}_p(m, L)$, the condition number of f is denoted by $\kappa_f \stackrel{\text{def}}{=} m/L$. We use $\mathcal{S}_p(0,\infty)$ to denote the set of convex functions. For any $f \in \mathcal{S}_p(0,\infty)$, let ∂f denote the subdifferential of f.

Lemma 1 *For any $f \in \mathcal{S}_p(0,\infty)$, the subdifferential ∂f is a monotone operator, i.e., if $\mathbf{u} \in \partial f(\mathbf{x})$ and $\mathbf{v} \in \partial f(\mathbf{y})$, it holds*

$$\langle \mathbf{x}-\mathbf{y}, \mathbf{u}-\mathbf{v}\rangle \geq 0, \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^p. \tag{7}$$

Lemma 1 can be easily proved by applying the definition of subdifferential.

Lemma 2 *(Theorem 2.1.11 of (Nesterov, 2013)) For any $f \in \mathcal{S}_p(m, L)$ where $0 < m \leq L < \infty$, it holds*

$$\langle \mathbf{x}-\mathbf{y}, \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) \geq \frac{mL}{m+L}\|\mathbf{x}-\mathbf{y}\|^2 + \frac{1}{m+L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2, \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^p. \tag{8}$$

Algorithm 1 Inexact Over-relaxed Alternating Direction Method of Multipliers (ADMM)

Input: penalty parameter ρ, relaxation parameter α

Initialize $\mathbf{x}_0, \mathbf{z}_0, \mathbf{u}_0$

for $k = 0, \ldots, T-1$ **do**

  Set $\mathbf{x}_{k+1} \approx \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ f(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}_k - \mathbf{c} - \mathbf{u}_k\|^2 \right\}$ s.t. (9a) is satisfied

  Set $\mathbf{z}_{k+1} \approx \underset{\mathbf{z}}{\operatorname{argmin}} \left\{ g(\mathbf{z}) + \frac{\rho}{2}\|\alpha\mathbf{A}\mathbf{x}_{k+1} - (1-\alpha)\mathbf{B}\mathbf{z}_k + \mathbf{B}\mathbf{z} - \alpha\mathbf{c} - \mathbf{u}_k\|^2 \right\}$ s.t. (9b) is satisfied

  Set $\mathbf{u}_{k+1} = \mathbf{u}_k - \left(\alpha\mathbf{A}\mathbf{x}_{k+1} - (1-\alpha)\mathbf{B}\mathbf{z}_k + \mathbf{B}\mathbf{z}_{k+1} - \alpha\mathbf{c}\right)$

end for

Output: $\mathbf{x}_T, \mathbf{y}_T, \mathbf{u}_T$

3. Dissipativity for Inexact ADMM with Strongly Convex Objectives

In this section, we present convergence analysis for over-relaxed ADMM with inexact subproblem solutions (*i.e.*, $\mathbf{x}_{k+1}$ and $\mathbf{z}_{k+1}$) for (4a) and (4b), respectively. Our method builds on the dissipativity framework (Hu and Lessard, 2017), which provides rigorous convergence analysis for Nesterov's accelerated method from an energy dissipation perspective.

3.1. Inexact ADMM as a Dynamical System

We first consider interpreting the over-relaxed ADMM with inexact subproblem solutions (*i.e.*, $\mathbf{x}_{k+1}$ and $\mathbf{z}_{k+1}$) as a discrete-time linear dynamical system. In this way, we can analyze its convergence by applying dissipativity theory to the dynamical system (Hu and Lessard, 2017). To this end, we formally introduce the notion of inexact solutions for subproblems (4a) and (4b). For convenience, we denote the objectives of (4a) and (4b) by $\widetilde{f}_\rho(\mathbf{x}) \overset{\text{def}}{=} f(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}_k - \mathbf{c} + \mathbf{u}_k\|^2$ and $\widetilde{g}_\rho(\mathbf{z}) \overset{\text{def}}{=} g(\mathbf{z}) + \frac{\rho}{2}\|\alpha\mathbf{A}\mathbf{x}_{k+1} - (1-\alpha)\mathbf{B}\mathbf{z}_k + \mathbf{B}\mathbf{z} - \alpha\mathbf{c} - \mathbf{u}_k\|^2$, respectively. For non-negative scalars ϵ_{k+1} and δ_{k+1}, we say $\mathbf{x}_{k+1}$ and $\mathbf{z}_{k+1}$ are ϵ_{k+1}-suboptimal and δ_{k+1}-suboptimal solutions to (4a) and (4b), respectively, if they satisfy For non-negative scalars ϵ_{k+1} and δ_{k+1}, we say $\mathbf{x}_{k+1}$ and $\mathbf{z}_{k+1}$ are ϵ_{k+1}-suboptimal and δ_{k+1}-suboptimal solutions to (4a) and (4b), respectively, if they satisfy

$$\widetilde{f}_\rho(\mathbf{x}_{k+1}) - \min_{\mathbf{x}} \widetilde{f}_\rho(\mathbf{x}) \le \epsilon_{k+1}, \tag{9a}$$

$$\widetilde{g}_\rho(\mathbf{z}_{k+1}) - \min_{\mathbf{z}} \widetilde{g}_\rho(\mathbf{z}) \le \delta_{k+1}. \tag{9b}$$

It is easy to see that the accuracy of suboptimal solutions $\mathbf{x}_{k+1}$ and $\mathbf{z}_{k+1}$ are denoted by ϵ_{k+1} and δ_{k+1}, respectively. If $\epsilon_{k+1} = \delta_{k+1} = 0$, $\mathbf{x}_{k+1}$ and $\mathbf{z}_{k+1}$ are the exact optimums, otherwise they are just an approximate solutions to (4a) and (4b), respectively. Algorithm 1 presents the inexact ADMM with over-relaxation.

To represent the inexact over-relaxed ADMM as a dynamic system, same as in (Nishihara et al., 2015), we make the following assumption.

Assumption 1 *For m and L with $0 < m \le L < \infty$, we assume that $f \in \mathcal{S}_p(m, L)$ and $g \in \mathcal{S}_p(0, \infty)$. In addition, we assume that $\mathbf{A}$ is invertible and $\mathbf{B}$ has full column rank.*

For any $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{z} \in \mathbb{R}^q$, we define new variables $\mathbf{w} \overset{\text{def}}{=} \mathbf{A}\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{s} \overset{\text{def}}{=} \mathbf{B}\mathbf{z} \in \mathbb{R}^p$. Given functions f and g, we define $\widehat{f}_\rho,\ \widehat{g}_\rho : \mathbb{R}^p \to \mathbb{R}$ as

$$\widehat{f}_\rho(\mathbf{w}) \overset{\text{def}}{=} \rho^{-1} f(\mathbf{A}^{-1}\mathbf{w}) \ \text{ and } \ \widehat{g}_\rho(\mathbf{s}) \overset{\text{def}}{=} \rho^{-1} g(\mathbf{B}^{+}\mathbf{s}) + \mathbb{I}_{\text{im}\mathbf{B}}(\mathbf{s}), \tag{10}$$

where $\mathbf{B}^+$ is any left inverse of $\mathbf{B}$ and $\mathbb{I}_{\text{im}\mathbf{B}}$ is the $\{0,\infty\}$-indicator function of the image of $\mathbf{B}$. It is easy to see that $\widehat{f}_\rho$ is $(\widehat{m}/\rho)$-strongly convex and $(\widehat{L}/\rho)$-smooth w.r.t. $\|\cdot\|$, where $\widehat{m} \overset{\text{def}}{=} m/(\sigma_{\mathbf{A}}^{\max})^2$ and $\widehat{L} \overset{\text{def}}{=} L/(\sigma_{\mathbf{A}}^{\min})^2$. By applying the definitions of $\widehat{f}_\rho$ and $\widehat{g}_\rho$, we can rewrite (2) as

$$L_\rho(\mathbf{w}, \mathbf{s}, \mathbf{u}) = \widehat{f}_\rho(\mathbf{w}) + \widehat{g}_\rho(\mathbf{s}) - \langle \mathbf{u}, \mathbf{w} + \mathbf{s} - \mathbf{c}\rangle. \tag{11}$$

Then, the inexact over-relaxed ADMM updates can be rewritten as

$$\mathbf{w}_{k+1} \approx \underset{\mathbf{w}}{\operatorname{argmin}} \Big\{ \widehat{f}_\rho(\mathbf{w}) + \frac{1}{2}\|\mathbf{w} + \mathbf{s}_k - \mathbf{c} - \mathbf{u}_k\|^2 \Big\}, \tag{12a}$$

$$\mathbf{z}_{k+1} \approx \underset{\mathbf{s}}{\operatorname{argmin}} \Big\{ \widehat{g}_\rho(\mathbf{s}) + \frac{1}{2}\|\alpha\mathbf{w}_{k+1} - (1-\alpha)\mathbf{s}_k + \mathbf{s} - \alpha\mathbf{c} - \mathbf{u}_k\|^2 \Big\}, \tag{12b}$$

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \big(\alpha\mathbf{w}_{k+1} - (1-\alpha)\mathbf{s}_k + \mathbf{s}_{k+1} - \alpha\mathbf{c}\big). \tag{12c}$$

The key of our method is to represent the inexact over-relaxed ADMM Algortihm 1 as a discrete-time linear dynamical system, so we can analyze its convergence behavior by using an energy dissipation perspective. To this end, we define the discrete-time linear dynamical system with the states $\{\boldsymbol{\xi}_k\}_{k\geq 0}$, inputs $\{\mathbf{v}_k\}_{k\geq 0}$, and outputs $\{\mathbf{y}_k^1\}_{k\geq 0}, \{\mathbf{y}_k^2\}_{k\geq 0}$ defined as

$$\boldsymbol{\xi}_k \overset{\text{def}}{=} \begin{bmatrix} \mathbf{s}_k \\ \mathbf{u}_k \end{bmatrix}, \quad \mathbf{v}_k \overset{\text{def}}{=} \begin{bmatrix} \boldsymbol{\beta}_{k+1} + \boldsymbol{\eta}_{k+1} \\ \boldsymbol{\gamma}_{k+1} + \boldsymbol{\zeta}_{k+1} \end{bmatrix}, \quad \mathbf{y}_k^1 \overset{\text{def}}{=} \begin{bmatrix} \mathbf{w}_{k+1} - \mathbf{c} \\ \boldsymbol{\beta}_{k+1} + \boldsymbol{\eta}_{k+1} \end{bmatrix}, \quad \mathbf{y}_k^2 \overset{\text{def}}{=} \begin{bmatrix} \mathbf{s}_{k+1} \\ \boldsymbol{\gamma}_{k+1} + \boldsymbol{\zeta}_{k+1} \end{bmatrix}, \tag{13}$$

where $\boldsymbol{\beta}_{k+1} \overset{\text{def}}{=} \nabla \widehat{f}_\rho(\mathbf{w}_{k+1}), \boldsymbol{\gamma}_{k+1} \in \partial\widehat{g}_\rho(\mathbf{s}_{k+1}), \|\boldsymbol{\eta}_{k+1}\|^2 \leq 2(\rho + \widehat{m})\epsilon_{k+1}/\rho^2$ and $\|\boldsymbol{\zeta}_{k+1}\|^2 \leq 2\delta_{k+1}/\rho$.

Lemma 3 *Suppose that Assumption 1 holds and Algorithm 1 is run with penalty parameter ρ and over-relaxation parameter α. Then inexact over-relaxed ADMM Algorithm 1 can be expressed as the following discrete-time linear dynamical system*

$$\boldsymbol{\xi}_{k+1} = (\widehat{\mathbf{A}} \otimes \mathbf{I}_p)\boldsymbol{\xi}_k + (\widehat{\mathbf{B}} \otimes \mathbf{I}_p)\mathbf{v}_k, \tag{14a}$$

$$\mathbf{y}_k^1 = (\widehat{\mathbf{C}}^1 \otimes \mathbf{I}_p)\boldsymbol{\xi}_k + (\widehat{\mathbf{D}}^1 \otimes \mathbf{I}_p)\mathbf{v}_k, \tag{14b}$$

$$\mathbf{y}_k^2 = (\widehat{\mathbf{C}}^2 \otimes \mathbf{I}_p)\boldsymbol{\xi}_k + (\widehat{\mathbf{D}}^2 \otimes \mathbf{I}_p)\mathbf{v}_k, \tag{14c}$$

where $\widehat{\mathbf{A}}$, $\widehat{\mathbf{B}}$, $\widehat{\mathbf{C}}^1$, $\widehat{\mathbf{C}}^2$, $\widehat{\mathbf{D}}^1$ and $\widehat{\mathbf{D}}^2$ are defined as

$$\widehat{\mathbf{A}} \overset{\text{def}}{=} \begin{bmatrix} 1 & 1-\alpha \\ 0 & 0 \end{bmatrix}, \widehat{\mathbf{B}} \overset{\text{def}}{=} \begin{bmatrix} \alpha & -1 \\ 0 & 1 \end{bmatrix}, \widehat{\mathbf{C}}^1 \overset{\text{def}}{=} \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix}, \widehat{\mathbf{D}}^1 \overset{\text{def}}{=} \begin{bmatrix} -1 & 0 \\ 1 & 0 \end{bmatrix}, \widehat{\mathbf{C}}^2 \overset{\text{def}}{=} \begin{bmatrix} 1 & 1-\alpha \\ 0 & 0 \end{bmatrix}, \widehat{\mathbf{D}}^2 \overset{\text{def}}{=} \begin{bmatrix} \alpha & -1 \\ 0 & 1 \end{bmatrix}. \tag{15}$$

Due to the limit in space, we provide all proofs in Appendix. By applying the convexity of $f(\mathbf{x})$ and $g(\mathbf{z})$, it can be shown that the solution to (1) is a fixed point of (14) as it satisfies the KKT conditions of (1).

3.2. Convergence Analysis via Dissipativity Theory

Let $\mathbf{x}_\star, \mathbf{z}_\star$ and $\mathbf{u}_\star$ be the primal and dual optimal solutions of (1), respectively. We define $\mathbf{w}_\star \overset{\text{def}}{=} \mathbf{A}\mathbf{x}_\star, \mathbf{s}_\star \overset{\text{def}}{=} \mathbf{B}\mathbf{z}_\star, \boldsymbol{\beta}_\star \overset{\text{def}}{=} \nabla \widehat{f}_\rho(\mathbf{w}_\star), \boldsymbol{\gamma}_\star \in \partial \widehat{g}_\rho(\mathbf{s}_\star)$ and

$$\boldsymbol{\xi}_\star \overset{\text{def}}{=} \begin{bmatrix} \mathbf{s}_\star \\ \mathbf{u}_\star \end{bmatrix}, \quad \mathbf{v}_\star \overset{\text{def}}{=} \begin{bmatrix} \boldsymbol{\beta}_\star \\ \boldsymbol{\gamma}_\star \end{bmatrix}, \quad \mathbf{y}_\star^1 \overset{\text{def}}{=} \begin{bmatrix} \mathbf{w}_\star - \mathbf{c} \\ \boldsymbol{\beta}_\star \end{bmatrix}, \quad \mathbf{y}_\star^2 \overset{\text{def}}{=} \begin{bmatrix} \mathbf{s}_\star \\ \boldsymbol{\gamma}_\star \end{bmatrix},$$

such that $(\boldsymbol{\xi}_\star, \mathbf{v}_\star, \mathbf{y}_\star^1, \mathbf{y}_\star^2)$ is a fixed point of the dynamical system (14). We define the following Lyapunov function as

$$V_{\mathbf{P}}(\boldsymbol{\xi}) \overset{\text{def}}{=} (\boldsymbol{\xi} - \boldsymbol{\xi}_\star)^\top \mathbf{P} (\boldsymbol{\xi} - \boldsymbol{\xi}_\star), \tag{16}$$

where $\mathbf{P} \overset{\text{def}}{=} \widehat{\mathbf{P}} \otimes \mathbf{I}_p$ and $\widehat{\mathbf{P}} \in \mathbb{S}^2_{++}$. We analyze the convergence of inexact over-relaxed ADMM by showing that the associated dynamical system (14) is dissipative with respect to a supply rate $S(\boldsymbol{\xi}_k, \mathbf{v}_k)$ (Hu and Lessard, 2017). Specifically, we seek to construct the following dissipation inequality

$$V_{\mathbf{P}}(\boldsymbol{\xi}_{k+1}) - \tau^2 V_{\mathbf{P}}(\boldsymbol{\xi}_k) \leq S(\boldsymbol{\xi}_k, \mathbf{v}_k), \forall k \geq 0, \tag{17}$$

where $\tau \in (0,1)$. This inequality states that every step of the dynamical system dissipates at least $(1-\tau^2)$ of the internal energy. If (17) is constructed, we can easily bound $V_{\mathbf{P}}(\boldsymbol{\xi}_{k+1})$ by recursively applying it. Therefore, the stability of the dynamical system implies the convergence of the optimization algorithm as the Lyapunov function $V_{\mathbf{P}}(\boldsymbol{\xi}_k)$ is convergent.

In our case, we define the supply rate function $S(\boldsymbol{\xi}_k, \mathbf{v}_k)$ as

$$S(\boldsymbol{\xi}_k, \mathbf{v}_k) \overset{\text{def}}{=} \begin{bmatrix} \boldsymbol{\xi}_k - \boldsymbol{\xi}_\star \\ \mathbf{v}_k - \mathbf{v}_\star \end{bmatrix}^\top \left(\left(\begin{bmatrix} \widehat{\mathbf{C}}^1 & \widehat{\mathbf{D}}^1 \\ \widehat{\mathbf{C}}^2 & \widehat{\mathbf{D}}^2 \end{bmatrix}^\top \begin{bmatrix} \lambda^1 \widehat{\mathbf{M}}^1 & \mathbf{0} \\ \mathbf{0} & \lambda^2 \widehat{\mathbf{M}}^2 \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{C}}^1 & \widehat{\mathbf{D}}^1 \\ \widehat{\mathbf{C}}^2 & \widehat{\mathbf{D}}^2 \end{bmatrix} \right) \otimes \mathbf{I}_p \right) \begin{bmatrix} \boldsymbol{\xi}_k - \boldsymbol{\xi}_\star \\ \mathbf{v}_k - \mathbf{v}_\star \end{bmatrix}, \tag{18}$$

where $\lambda^1, \lambda^2 \geq 0$ and $\widehat{\mathbf{M}}^1, \widehat{\mathbf{M}}^2$ are defined as

$$\widehat{\mathbf{M}}^1 \overset{\text{def}}{=} \begin{bmatrix} 2\rho^{-2}\widehat{m}\widehat{L} & -\rho^{-1}(\widehat{m}+\widehat{L}) \\ -\rho^{-1}(\widehat{m}+\widehat{L}) & 2 \end{bmatrix}, \quad \widehat{\mathbf{M}}^2 \overset{\text{def}}{=} \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}.$$

To construct (17), we need to provide an appropriate $\mathbf{P}$ such that the dissipation inequality holds. Next, we show that $\mathbf{P}$ can be obtained by solving a 4×4 semidefinite program.

Lemma 4 *Suppose that Assumption 1 holds and Algorithm 1 is run with penalty parameter ρ and over-relaxation parameter α. If there exists $\tau \in (0,1)$ and $\widehat{\mathbf{P}} \in \mathbb{S}^2_{++}$ such that*

$$\begin{bmatrix} \widehat{\mathbf{A}}^\top \widehat{\mathbf{P}} \widehat{\mathbf{A}} - \tau^2 \widehat{\mathbf{P}} & \widehat{\mathbf{A}}^\top \widehat{\mathbf{P}} \widehat{\mathbf{B}} \\ \widehat{\mathbf{B}}^\top \widehat{\mathbf{P}} \widehat{\mathbf{A}} & \widehat{\mathbf{B}}^\top \widehat{\mathbf{P}} \widehat{\mathbf{B}} \end{bmatrix} - \begin{bmatrix} \widehat{\mathbf{C}}^1 & \widehat{\mathbf{D}}^1 \\ \widehat{\mathbf{C}}^2 & \widehat{\mathbf{D}}^2 \end{bmatrix}^\top \begin{bmatrix} \lambda^1 \widehat{\mathbf{M}}^1 & \mathbf{0}_2 \\ \mathbf{0}_2 & \lambda^2 \widehat{\mathbf{M}}^2 \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{C}}^1 & \widehat{\mathbf{D}}^1 \\ \widehat{\mathbf{C}}^2 & \widehat{\mathbf{D}}^2 \end{bmatrix} \preceq \mathbf{0}, \tag{19}$$

then $\forall k \geq 0$:

$$V_{\mathbf{P}}(\boldsymbol{\xi}_{k+1}) - \tau^2 V_{\mathbf{P}}(\boldsymbol{\xi}_k) \leq \widetilde{\theta}\epsilon_{k+1} + \widehat{\theta}\|\boldsymbol{\xi}_k - \boldsymbol{\xi}_\star\|\sqrt{\epsilon_{k+1}} + \bar{\theta}\|\boldsymbol{\xi}_{k+1} - \boldsymbol{\xi}_\star\|\sqrt{\delta_{k+1}}, \tag{20}$$

where $\widetilde{\theta}, \widehat{\theta}$ and $\bar{\theta}$ are defined as

$$\widetilde{\theta} \overset{\text{def}}{=} \frac{4\lambda^1(\rho + \widehat{m})\max(3\rho, \rho + \widehat{m} + \widehat{L})}{\rho^3}, \quad \widehat{\theta} \overset{\text{def}}{=} \frac{2\sqrt{2}\lambda^1\sqrt{\rho+\widehat{m}}\max(2\rho, \widehat{m}+\widehat{L})}{\rho^2}, \quad \bar{\theta} \overset{\text{def}}{=} 2\lambda^2\sqrt{\frac{2}{\rho}}.$$

Remark 5 *For first-order methods and Nesterov's method, Hu and Lessard (2017) show one can choose a non-positive supply rate,* i.e., $S(\boldsymbol{\xi}_k, \mathbf{v}_k) \le 0$*, so that the internal energy* $V(\boldsymbol{\xi}_k)$ *converges to the minimum value no slower than a linear rate* ρ^2*. In contrast, (20) in Lemma 4 implies the supply rate* $S(\boldsymbol{\xi}_k, \mathbf{v}_k)$ *for inexact over-relaxed ADMM is generally non-negative, except the subproblems (4a) and (4b) are exactly solved at every iterations,* i.e., $\epsilon_k = \delta_k = 0, \forall k \ge 1$*. From an energy dissipation perspective, this is due to the driving force* $\mathbf{v}_k$ *continues to inject some energy into the dynamic system due to the inexactness of* $\mathbf{x}_{k+1}$ *and* $\mathbf{z}_{k+1}$*.*

As we will see in Theorem 7, τ is the linear convergence rate of the over-relaxed ADMM. Therefore, it is important to find the smallest rate τ such that there exists a $\widehat{\mathbf{P}}$ to satisfy (19). To this end, Nishihara et al. (2015) suggest performing a binary search over τ and numerically solving (19) for each τ to find the smallest τ such that (19) is feasible. Although the SDP is small, it is still computationally expensive as one needs to solve it for many values of τ. In contrast, França and Bento (2016) show that there is an exact analytical solution to (19). For convenience, we assume the value of ρ is set to $\rho = \rho_0\sqrt{\widehat{m}\widehat{L}}$ where $\rho_0 > 0$ and definite $\kappa \overset{\text{def}}{=} \kappa_f \kappa_{\mathbf{A}}^2$.

Lemma 6 *(Theorem 3 of (França and Bento, 2016)) For ADMM (i.e.,* $\epsilon_k = \delta_k = 0, \forall k \ge 1$*) with over-relaxation* $\alpha \in (0,2)$*, if* $\kappa > 1$ *and* $\rho_0 > 0$*, the following is an explicit point of (19) with* $\lambda_1, \lambda_2 \ge 0$ *and* $\widehat{\mathbf{P}} \succ \mathbf{0}$ *and* $\tau \in (0,1)$*:*

$$\widehat{\mathbf{P}} = \begin{bmatrix} 1 & \zeta \\ \zeta & 1 \end{bmatrix}, \quad \zeta = -1 + \frac{\alpha\big(\chi(\rho_0)\sqrt{\kappa} - 1\big)}{1 - \alpha + \chi(\rho_0)\sqrt{\kappa}}, \quad \lambda_1 = \frac{\alpha\rho_0\sqrt{\kappa}\big(1 - \alpha + \chi(\rho_0)\sqrt{\kappa}\big)}{(\kappa - 1)\big(1 + \chi(\rho_0)\sqrt{\kappa}\big)}, \quad \lambda_2 = 1 + \zeta$$

with $\tau = 1 - \frac{\alpha}{1+\chi(\rho_0)\sqrt{\kappa}}$ *where* $\chi(x) \overset{\text{def}}{=} \max(x, x^{-1}) \ge 1, \forall x > 0$*.*

If (4a) and (4b) are exactly solved at every iteration (i.e., $\epsilon_k = \delta_k = 0, \forall k \ge 1$), the dissipation inequality (20) becomes

$$V_{\mathbf{P}}(\boldsymbol{\xi}_{k+1}) - \tau^2 V_{\mathbf{P}}(\boldsymbol{\xi}_k) \le 0. \tag{21}$$

Comparing it with (17), it is equivalent to stating $S(\boldsymbol{\xi}_k, \mathbf{v}_k) \le 0, \forall k \ge 0$. Then, $V_{\mathbf{P}}(\boldsymbol{\xi}_{k+1})$ can be easily bounded by applying (21). In a general case, inexact $\mathbf{x}_{k+1}$ and $\mathbf{z}_{k+1}$ leads to $S(\boldsymbol{\xi}_k, \mathbf{v}_k) > 0$ that needs to be first bounded. The next theorem shows the convergence of inexact over-relaxed ADMM.

Theorem 7 *Suppose that Assumption 1 holds. Let the sequences* $\{\mathbf{x}_k\}_{k\ge1}, \{\mathbf{z}_k\}_{k\ge1}$ *and* $\{\mathbf{u}_k\}_{k\ge1}$ *generated by running Algorithm 1 with penalty parameter* ρ *and over-relaxation parameter* α*. We define*

$$\boldsymbol{\varphi}_k \overset{\text{def}}{=} \big[(\mathbf{B}\mathbf{z}_k)^\top \quad \mathbf{u}_k^\top \quad \big]^\top, \quad \boldsymbol{\varphi}_\star \overset{\text{def}}{=} \big[(\mathbf{B}\mathbf{z}_\star)^\top \quad \mathbf{u}_\star^\top\big]^\top.$$

If there exists $\tau \in (0,1)$ *and* $\widehat{\mathbf{P}} \in \mathbb{S}^2_{++}$ *such that the linear matrix inequality (19) is satisfied, then for* $\forall T \ge 1$*:*

$$\|\boldsymbol{\varphi}_T - \boldsymbol{\varphi}_\star\| \le \tau^T \left(\sqrt{\kappa_{\mathbf{P}}}\|\boldsymbol{\varphi}_0 - \boldsymbol{\varphi}_\star\| + \left(\sqrt{\frac{\widetilde{\theta}}{\sigma_{\mathbf{P}}^{min}}} + \frac{\widehat{\theta}}{\tau\sigma_{\mathbf{P}}^{min}}\right)\sum_{k=1}^{T}\tau^{-k}\sqrt{\epsilon_k} + \frac{\bar{\theta}}{\sigma_{\mathbf{P}}^{min}}\sum_{k=1}^{T}\tau^{-k}\sqrt{\delta_k}\right). \tag{22}$$

Theorem 7 suggests that we obtain a linear convergence rate for inexact over-relaxed ADMM, provided that $\sqrt{\epsilon_k}$ and $\sqrt{\delta_k}$ decrease linearly to 0. Suppose $\sqrt{\epsilon_k}$ and $\sqrt{\delta_k}$ decrease linearly at a rate of $\widehat{\tau} \in (0,1)$. If $\widehat{\tau} < \tau$, then we can show inexact over-relaxed ADMM linearly converges at a rate τ that is the same as the exact ADMM. If $\widehat{\tau} > \tau$, then inexact over-relaxed ADMM achieve linear convergence rate $\widehat{\tau}$. Theorem 7 implies that (4a) and (4b) are allowed to be approximately solved as long as they become more accurate as the iterations progress.

4. Dissipativity for Inexact ADMM with Generally Convex Objectives

In the previous section, our convergence analysis for inexact over-relaxed ADMM assumes the objective to be strongly convex. However, this is not the case for many machine learning problems. Nishihara et al. (2015) suggest to add a term $\frac{\delta}{2}\|\mathbf{x}\|^2$ to the objective such that $f(\mathbf{x}) + \frac{\delta}{2}\|\mathbf{x}\|^2$ is strongly convex. This trick is less appealing as it modifies the objective. Furthermore, it is difficult to tune the value for δ in practice. In order to address this problem, we now extend our analysis to the case of generally convex objectives. We first relax Assumption 1 by allowing f to be generally convex.

Assumption 2 *For $0 < L < \infty$, we assume that $f \in \mathcal{S}_p(0, L)$ and $g \in \mathcal{S}_p(0,\infty)$. In addition, we assume that $\mathbf{A}$ is invertible and $\mathbf{B}$ has full column rank.*

Similar to the case of strongly convex objectives, we first represent the inexact over-relaxed ADMM as a discrete-time linear dynamical system. The states $\{\boldsymbol{\xi}_k\}_{k\geq 0}$, inputs $\{\mathbf{v}_k\}_{k\geq 0}$ and outputs $\{\mathbf{y}_k^1\}_{k\geq 0}, \{\mathbf{y}_k^2\}_{k\geq 0}$ are the same as (13) except $\|\boldsymbol{\eta}_{k+1}\|^2 \leq 2\epsilon_{k+1}/\rho$. The proof is similar to that of Lemma 3.

Lemma 1 *Suppose that Assumption 2 holds and Algorithm 1 is run with penalty parameter ρ and over-relaxation parameter α. Then, the inexact over-relaxed ADMM can be expressed as the same discrete-time linear dynamical system defined as (14) except $\|\boldsymbol{\eta}_{k+1}\|^2 \leq 2\epsilon_{k+1}/\rho$.*

By the variational inequality reformulation (He and Yuan, 2012), the optimal condition of (1) is characterized as following

$$f(\mathbf{x}) - f(\mathbf{x}_\star) + g(\mathbf{z}) - g(\mathbf{z}_\star) + \begin{bmatrix} \mathbf{x} - \mathbf{x}_\star \\ \mathbf{z} - \mathbf{z}_\star \\ \rho(\mathbf{u} - \mathbf{u}_\star) \end{bmatrix}^\top \begin{bmatrix} -\rho\mathbf{A}^\top \mathbf{u}_\star \\ -\rho\mathbf{B}^\top \mathbf{u}_\star \\ \mathbf{A}\mathbf{x}_\star + \mathbf{B}\mathbf{z}_\star - \mathbf{c} \end{bmatrix} \geq 0, \forall \mathbf{x}, \mathbf{y}, \mathbf{u}, \qquad (23)$$

where $\mathbf{x}_\star, \mathbf{z}_\star$ and $\mathbf{u}_\star$ are the primal and dual optimal solutions of (1), respectively. Next, we show that the convergence of inexact over-relaxed ADMM for generally convex objectives can be proved by showing that the left-hand side of (23) is convergent. The key idea is to choose an appropriate supply rate $S(\boldsymbol{\xi}_k, \mathbf{v}_k)$. Specifically, we choose the following supply rate

$$S(\boldsymbol{\xi}_k, \mathbf{v}_k) \stackrel{\text{def}}{=} \begin{bmatrix} \boldsymbol{\xi}_k - \boldsymbol{\xi}_\star \\ \mathbf{v}_k - \mathbf{v}_\star \end{bmatrix}^\top \left(\left(\begin{bmatrix} \widehat{\mathbf{C}}^1 & \widehat{\mathbf{D}}^1 \\ \widehat{\mathbf{C}}^2 & \widehat{\mathbf{D}}^2 \end{bmatrix}^\top \begin{bmatrix} \widehat{\mathbf{M}}^1 & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{M}}^2 \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{C}}^1 & \widehat{\mathbf{D}}^1 \\ \widehat{\mathbf{C}}^2 & \widehat{\mathbf{D}}^2 \end{bmatrix} \right) \otimes \mathbf{I}_p \right) \begin{bmatrix} \boldsymbol{\xi}_k - \boldsymbol{\xi}_\star \\ \mathbf{v}_k - \mathbf{v}_\star \end{bmatrix}, \qquad (24)$$

where $\widehat{\mathbf{M}}^1$ and $\widehat{\mathbf{M}}^2$ are

$$\widehat{\mathbf{M}}^1 \stackrel{\text{def}}{=} \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & \frac{\rho}{2\widehat{L}} \end{bmatrix}, \quad \widehat{\mathbf{M}}^2 \stackrel{\text{def}}{=} \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix}.$$

The next lemma provides an upper bound for $S(\boldsymbol{\xi}_k, \mathbf{v}_k)$.

Lemma 8 *Suppose that Assumption 2 holds and Algorithm 1 is run with penalty parameter ρ and over-relaxation parameter α. Then for $\forall k \geq 0$:*

$$\begin{aligned} S(\boldsymbol{\xi}_k, \mathbf{v}_k) \leq & \widehat{f}_\rho(\mathbf{w}_\star) - \widehat{f}_\rho(\mathbf{w}_{k+1}) + \widehat{g}_\rho(\mathbf{s}_\star) - \widehat{g}_\rho(\mathbf{s}_{k+1}) - \begin{bmatrix} \mathbf{w}_{k+1} - \mathbf{w}_\star \\ \mathbf{s}_{k+1} - \mathbf{s}_\star \\ \mathbf{u}_{k+1} - \mathbf{u}_\star \end{bmatrix}^\top \begin{bmatrix} -\mathbf{u}_\star \\ -\mathbf{u}_\star \\ \mathbf{w}_\star + \mathbf{s}_\star - \mathbf{c} \end{bmatrix} \\ & + \left\langle \boldsymbol{\eta}_{k+1}, \frac{\rho}{\widehat{L}}(\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_\star) - (\mathbf{w}_{k+1} - \mathbf{w}_\star) \right\rangle + \frac{\rho}{2\widehat{L}} \|\boldsymbol{\eta}_{k+1}\|^2 - \langle \mathbf{s}_{k+1} - \mathbf{s}_\star, \boldsymbol{\zeta}_{k+1} \rangle. \end{aligned} \tag{25}$$

In the case of generally convex objectives, the ADMM can only achieve a sublinear instead of a linear convergence rate. Thus, we need to change the dissipation inequality (17). Specifically, we now seek to construct the following dissipation inequality

$$V_{\mathbf{P}}(\boldsymbol{\xi}_{k+1}) - V_{\mathbf{P}}(\boldsymbol{\xi}_k) \leq S(\boldsymbol{\xi}_k, \mathbf{v}_k), \forall k \geq 0. \tag{26}$$

It can be constructed by solving a small semidefinite program. The next theorem presents the sublinear convergence of inexact over-relaxed ADMM for generally convex objectives.

Theorem 9 *Suppose that Assumption 2 holds. Let the sequences $\{\mathbf{x}_k\}_{k\geq 1}, \{\mathbf{z}_k\}_{k\geq 1}$ and $\{\mathbf{u}_k\}_{k\geq 1}$ generated by running Algorithm 1 and with penalty parameter ρ and over-relaxation parameter α. If there exists $\widehat{\mathbf{P}} \in \mathbb{S}^2_{++}$ such that*

$$\begin{bmatrix} \widehat{\mathbf{A}}^\top \widehat{\mathbf{P}} \widehat{\mathbf{A}} - \widehat{\mathbf{P}} & \widehat{\mathbf{A}}^\top \widehat{\mathbf{P}} \widehat{\mathbf{B}} \\ \widehat{\mathbf{B}}^\top \widehat{\mathbf{P}} \widehat{\mathbf{A}} & \widehat{\mathbf{B}}^\top \widehat{\mathbf{P}} \widehat{\mathbf{B}} \end{bmatrix} - \begin{bmatrix} \widehat{\mathbf{C}}^1 & \widehat{\mathbf{D}}^1 \\ \widehat{\mathbf{C}}^2 & \widehat{\mathbf{D}}^2 \end{bmatrix}^\top \begin{bmatrix} \widehat{\mathbf{M}}^1 & \mathbf{0}_2 \\ \mathbf{0}_2 & \widehat{\mathbf{M}}^2 \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{C}}^1 & \widehat{\mathbf{D}}^1 \\ \widehat{\mathbf{C}}^2 & \widehat{\mathbf{D}}^2 \end{bmatrix} \preceq \mathbf{0}, \tag{27}$$

then we have a sublinear convergence rate for inexact over-relaxed ADMM, i.e.,

$$\begin{aligned} & f(\bar{\mathbf{x}}_T) - f(\mathbf{x}_\star) + g(\bar{\mathbf{z}}_T) - g(\mathbf{z}_\star) + \begin{bmatrix} \bar{\mathbf{x}}_T - \mathbf{x}_\star \\ \bar{\mathbf{z}}_T - \mathbf{z}_\star \\ \rho(\bar{\mathbf{u}}_T - \mathbf{u}_\star) \end{bmatrix}^\top \begin{bmatrix} -\rho \mathbf{A}^\top \mathbf{u}_\star \\ -\rho \mathbf{B}^\top \mathbf{u}_\star \\ \mathbf{A}\mathbf{x}_\star + \mathbf{B}\mathbf{z}_\star - \mathbf{c} \end{bmatrix} \\ \leq & \frac{\rho}{T} \left(\sqrt{V_{\mathbf{P}}(\boldsymbol{\xi}_0)} + \left(\sqrt{\frac{3\rho + 2\widehat{L}}{\rho \widehat{L}}} + \frac{\rho + \widehat{L}}{\widehat{L}} \sqrt{\frac{2}{\rho \sigma_{\mathbf{P}}^{min}}} \right) \sum_{k=1}^T \sqrt{\epsilon_k} + \sqrt{\frac{2}{\rho \sigma_{\mathbf{P}}^{min}}} \sum_{k=1}^T \sqrt{\delta_k} \right)^2, \end{aligned} \tag{28}$$

where $\bar{\mathbf{x}}_T \stackrel{\text{def}}{=} \frac{1}{T} \sum_{k=1}^T \mathbf{x}_k, \bar{\mathbf{z}}_T \stackrel{\text{def}}{=} \frac{1}{T} \sum_{k=1}^T \mathbf{z}_k$ and $\bar{\mathbf{u}}_T \stackrel{\text{def}}{=} \frac{1}{T} \sum_{k=1}^T \mathbf{u}_k$.

Remark 10 *When the $\mathbf{x}$- and $\mathbf{z}$-subproblems are exactly solved at every iteration (i.e., $\epsilon_k = 0$ and $\delta_k = 0, \forall k \geq 1$), then Theorem 9 provides the same sublinear convergence rate of ADMM for generally convex objectives provided by He and Yuan (2012). However, their result is only applicable for standard ADMM (i.e., $\alpha = 1$). In contrast, the result of Theorem 9 holds for a wide range of α.*

Theorem 9 suggests that inexact over-relaxed ADMM achieves sublinear convergence rate $O(1/T)$ if both $\{\sqrt{\epsilon_k}\}$ and $\{\sqrt{\delta_k}\}$ are summable. A sufficient condition for achieving $O(1/T)$ convergence rate is that $\sqrt{\epsilon_k}$ and $\sqrt{\delta_k}$ decreases at a rate of $O(1/k^r)$ for any $r > 1$. It is worth noting that the convergence rate of inexact over-relaxed ADMM is still $O(1/T)$ but yielding a better constant factor for a larger value of r when $r > 1$. In addition, the sum of error sequences $\{\sqrt{\epsilon_k}\}$ and $\{\sqrt{\delta_k}\}$ in $O(\sqrt{k})$ need to be in $o(1/\sqrt{k})$ for ensuring inexact over-relaxed ADMM to converge. This is a considerably weak condition. Thus, inexact over-relaxed ADMM for convex objectives still converges even the (4a) and (4a) are not exactly solved by carefully controlling the error sequences $\{\sqrt{\epsilon_k}\}$ and $\{\sqrt{\delta_k}\}$. This explains the practical performances of inexact over-relaxed ADMM.

In the case of strongly convex objectives, the explicit convergence rate τ of inexact over-relaxed ADMM is determined by the solution of (19). In contrast, the convergence rate of inexact over-relaxed ADMM for generally convex objectives is always $O(1/T)$ if (27) is satisfied. Thus, if we can prove (27) is feasible, then we do not need to numerically solve it in this case. The next theorem shows (27) is feasible by providing an explicit solution.

Theorem 11 *Suppose that Assumption 2 holds. For inexact over-relaxed ADMM with penalty parameter ρ and over-relaxation parameter α, then $\widehat{\mathbf{P}} = 0.5\mathbf{I}_2$ satisfies the linear matrix inequality (27).*

Theorem 11 suggests that (27) is definitely satisfied. Thus, we do not need to numerically solve or verify the SDP in practice. This makes Theorem 9 is more useful in real applications. It is worth noting that Theorem 11 requires no assumptions on the value of α. Specifically, Theorem 11 also suggests that the sublinear convergence of over-relaxed ADMM is even for $\alpha \geq 2$. In contrast, prior works assume either $\alpha = 1$ or $\alpha \in (0, 2)$.

5. Experiments

In this section, we apply inexact over-relaxed ADMM to solve a distributed ℓ_1-norm regularized logistic regression (Koh et al., 2007). The problem can be written as

$$\min_{\{\mathbf{x}_j, v_j\}_{j=1}^J, \mathbf{z}} \quad \sum_{j=1}^{J} \sum_{i=1}^{n_j} f_i^j(\mathbf{x}_j) + \lambda \|\mathbf{z}\|_1 \quad \text{s.t.} \quad \mathbf{x}_j - \mathbf{z} = \mathbf{0}, \forall j = 1, \ldots, J, \tag{29}$$

where $f_i^j(\mathbf{x}_j) = \log\big(1 + \exp\big(-b_i^j\big(\langle \mathbf{a}_i^j, \mathbf{x}_j \rangle + v_j\big)\big)\big)$, J is the number of workers, n_j is the number of samples on the jth worker, $\mathbf{a}_i^j$ and b_i^j are the input and output of the ith sample on the jth worker, respectively.

It is easy to verify that the objective is generally convex instead of strongly convex. In this case, the convergence rate of inexact over-relaxed ADMM can be obtained from Theorem 9 as it satisfies Assumption 2. In this problem, the $\mathbf{x}$-minimization (4a) can be divided into J independent problems that can be solved in parallel. Note that the (4a) does not admit an analytical solution due to the logistic loss. Thus, we employ the L-BFGS to solve it at every iteration. We test two different ways to terminate the L-BFGS solver:

Table 1: Statistics of the datasets.

Dataset	MDS	RCV1
# Samples	63,904	20,242
# Features	10,000	47,236
Sparsity (%)	0.9	0.2

- Running the L-BFGS solver such that ϵ_k decreasing at the rate $O(1/k^c)$ with $c = 1, 2, 3$.
- Running the L-BFGS solver up to a fixed number of iterations $c = 1, 3, 5, 7$.

In our experiments, we use two dastasets: *MDS* (Blitzer et al., 2007) and *RCV1* (Lewis et al., 2004). Table 1 shows the statistics of two datasets. As observed, these two datasets

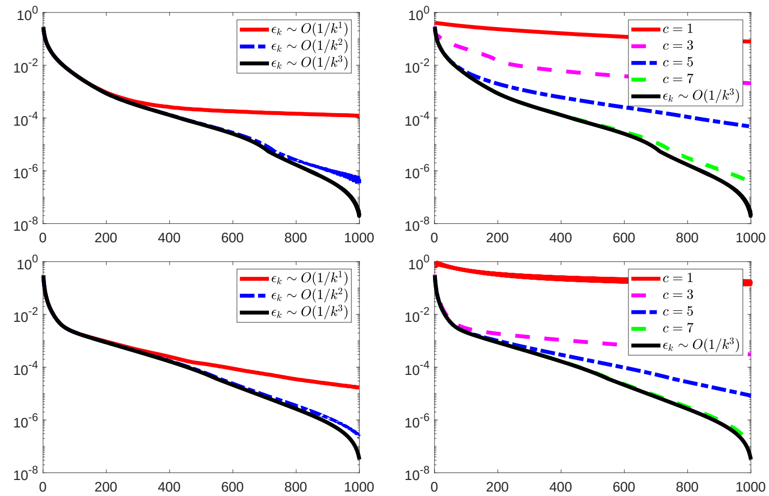

Figure 1: Result of inexact over-relaxed ADMM for solving (29). From top to bottom: *MDS* and *RCV1*. Horizontal axis denotes the number of inexact over-relaxed ADMM iterations and vertical axis is $(f(\mathbf{x}_k) - f(\mathbf{x}_\star) + g(\mathbf{z}_k) - g(\mathbf{z}_\star))/(f(\mathbf{x}_\star) + g(\mathbf{z}_\star))$.

are highly sparse. Specifically, we randomly and averagely divide each dataset into J parts. We set $J = 10$ and $J = 4$ for *MDS* and *RCV1*, respectively. In our experiments, we set $\rho = 10$ and $\alpha = 1.8$ for both datasets.

Figure 1 shows the results of inexact over-relaxed ADMM on the two datasets. As observed, both strategies for terminating L-BFGS solvers lead to global convergence by carefully choosing a reasonable value for c. As observed on both datasets, the choice of $\epsilon_k \sim O(1/k^3)$ achieves the fastest convergence rate and the choice of $\epsilon_k \sim O(1/k^2)$ achieves very similar performance as $\epsilon_k \sim O(1/k^3)$. In addition, running the L-BFGS solver with fixed 7 iterations also provides considerably good performance. This empirically shows that ADMM is not highly sensitive to inexact solutions of subproblems which is consistent with our theoretical results.

6. Conclusion

In this work, we present a new convergence analysis for over-relaxed ADMM. Our work is inspired by (Nishihara et al., 2015) but overcomes its two limitations. First, our method allows the subproblems of ADMM can be inexactly solved. Second, we further prove the sublinear convergence rate of inexact over-relaxed ADMM for a generally convex objective. Both these two improvements are important for applying ADMM to machine learning problems in practice. Our analysis works for a wide range of parameters of over-relaxed ADMM.

Acknowledgments

Qiang Zhou is supported by National Science Foundation of China under Grant 62106045 and the Start-up Research Fund of Southeast University under Grant RF1028623070. Sinno J.Pan thanks the support of the Hong Kong Jockey Club Charities Trust to the JC STEM Lab of Integration of Machine Learning and Symbolic Reasoning.

References

Francis R. Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Proceedings of Annual Conference on Neural Information Processing Systems (NIPS)*, 2007.

Stephen P. Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

Wei Deng and Wotao Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916, 2016.

Jonathan Eckstein. Parallel alternating direction multiplier decomposition of convex programs. *Journal of Optimization Theory and Applications*, 80(1):39–62, 1994.

Jonathan Eckstein and Dimitri P. Bertsekas. On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992.

Jonathan Eckstein and Wang Yao. Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives. *Pacific Journal of Optimization*, 11(4):619–644, 2015.

Jalal Fadili and Gabriel Peyré. Total variation projection with first order schemes. *IEEE Transactions on Image Processing (TIP)*, 20(3):657–669, 2011.

Guilherme França and José Bento. An explicit rate bound for over-relaxed ADMM. In *Proceedings of IEEE International Symposium on Information Theory (ISIT)*, 2016.

Guilherme França, Daniel P. Robinson, and René Vidal. ADMM and accelerated ADMM as continuous dynamical systems. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1554–1562, 2018.

Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

Tom Goldstein, Brendan O'Donoghue, Simon Setzer, and Richard G. Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences (SIIMS)*, 7(3):1588–1623, 2014.

E. Gi Gol'shtein and NV Tret'yakov. Modified lagrangians in convex programming and their generalizations. *Point-to-Set Maps and Mathematical Programming*, 55:1979, 1979.

Bingsheng He and Xiaoming Yuan. On the O(1/n) convergence rate of the douglas-rachford alternating direction method. *SIAM Journal on Numerical Analysis (SINUM)*, 50(2): 700–709, 2012.

Bingsheng He, Li-Zhi Liao, Deren Han, and Hai Yang. A new inexact alternating directions method for monotone variational inequalities. *Mathematical Programming*, 92(1):103–118, 2002.

Bin Hu and Laurent Lessard. Dissipativity theory for Nesterov's accelerated method. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1549–1557, 2017.

Kwangmoo Koh, Seung-Jean Kim, and Stephen P. Boyd. An interior-point method for large-scale l_1-regularized logistic regression. *Journal of Machine Learning Research (JMLR)*, 8: 1519–1555, 2007.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research (JMLR)*, 5:361–397, 2004.

Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Proceedings of Annual Conference on Neural Information Processing Systems (NIPS)*, 2011.

Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

Yuanyuan Liu, Fanhua Shang, Hongying Liu, Lin Kong, Licheng Jiao, and Zhouchen Lin. Accelerated variance reduction stochastic ADMM for large-scale machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 43(12):4242–4255, 2021.

Renato D. C. Monteiro and Benar Fux Svaiter. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization (SIOPT)*, 23(1):475–507, 2013.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013.

Michael K. Ng, Pierre Weiss, and Xiaoming Yuan. Solving constrained total-variation image restoration and reconstruction problems via alternating direction methods. *SIAM Journal on Scientific Computing (SISC)*, 32(5):2710–2736, 2010.

Robert Nishihara, Laurent Lessard, Benjamin Recht, Andrew K. Packard, and Michael I. Jordan. A general analysis of the convergence of ADMM. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 343–352, 2015.

Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright. *Optimization for Machine Learning*. MIT Press, 2012.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

Huahua Wang, Arindam Banerjee, Cho-Jui Hsieh, Pradeep Ravikumar, and Inderjit S. Dhillon. Large scale distributed sparse precision estimation. In *Proceedings of Annual Conference on Neural Information Processing Systems (NIPS)*, pages 584–592, 2013.

Yilun Wang, Junfeng Yang, Wotao Yin, and Yin Zhang. A new alternating minimization algorithm for total variation image econstruction. *SIAM Journal on Imaging Sciences (SIIMS)*, 1(3):248–272, 2008.

Ashia C. Wilson, Ben Recht, and Michael I. Jordan. A lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research (JMLR)*, 22:113:1–113:34, 2021.

Zheng Xu, Gavin Taylor, Hao Li, Mário A. T. Figueiredo, Xiaoming Yuan, and Tom Goldstein. Adaptive consensus ADMM for distributed optimization. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 3841–3850, 2017.

Lei Yuan, Jun Liu, and Jieping Ye. Efficient methods for overlapping group lasso. In *Proceedings of Annual Conference on Neural Information Processing Systems (NIPS)*, pages 352–360, 2011.

Caoxie Zhang, Honglak Lee, and Kang G. Shin. Efficient distributed linear classification algorithms via the alternating direction method of multipliers. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1398–1406, 2012.

Ruiliang Zhang and James T. Kwok. Asynchronous distributed ADMM for consensus optimization. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1701–1709, 2014.

Xiaoqun Zhang, Martin Burger, Xavier Bresson, and Stanley J. Osher. Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. *SIAM Journal on Imaging Sciences (SIIMS)*, 3(3):253–276, 2010.

Xiaoqun Zhang, Martin Burger, and Stanley J. Osher. A unified primal-dual algorithm framework based on bregman iteration. *Journal of Scientific Computing*, 46(1):20–46, 2011.

Shenglong Zhou and Geoffrey Ye Li. Federated learning via inexact ADMM. *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 45(8):9699–9708, 2023.