A Continuous Limit Theory for Nonsmooth ADMM Variants

Huizhuo Yuan [*] Yuren Zhou [†] Chris Junchi Li [‡] Qingyun Sun [§]

November 29, 2019

Abstract

Recently, there has been a great deal of research attention on understanding the convergence behavior of first-order methods. One line of this research focuses on analyzing the convergence behavior of first-order methods using tools from continuous dynamical systems such as ordinary differential equations and differential inclusions. These research results shed lights on better understanding first-order methods from a non-optimization point of view. The alternating direction method of multipliers (ADMM) is a widely used first-order method for solving optimization problems arising from machine learning and statistics, and it is important to investigate its behavior using these new techniques from dynamical systems. Existing works along this line have been mainly focusing on problems with smooth objective functions, which exclude many important applications that are traditionally solved by ADMM variants. In this paper, we analyze some well-known and widely used ADMM variants for nonsmooth optimization problems using tools of differential inclusions. In particular, we analyze the convergence behavior of linearized ADMM, gradient-based ADMM, generalized ADMM and accelerated generalized ADMM for nonsmooth problems and show their connections with dynamical systems. We anticipate that these results will provide new insights on understanding ADMM for solving nonsmooth problems.

1 Introduction

Recently, there has been tremendous interests in using continuous-time dynamical system tools to analyze first-order optimization algorithms such as Nesterov's accelerated gradient method (AGM) (Nesterov, 1983) and its variants. In the seminal work Su et al. (2016), the authors designed a differential equation for modeling AGM, and analyzed the connection between the solution of the differential equation and the continuous limit of the iterates of AGM. Their work provided

[*]Peking University, Beijing 100871, China; email: yuanhz@pku.edu.cn

[†]Duke University, Durham, NC 27708, USA; email: yuren.zhou@duke.edu

[‡]Tencent AI Lab, Shennan Ave, Nanshan District, Shenzhen, Guangdong Province 518057, China; email: junchi.li.duke@gmail.com

[§]Stanford University, Stanford, California 94305, USA; email: qysun@stanford.ed

new insights on understanding the convergence behavior of AGM. Later investigations along this line mainly focused on analyzing AGM and its variants such as FISTA and heavy ball method using the tools of ordinary differential equations, differential inclusions, and more generally, continuous dynamical systems (see, e.g., Krichene et al. (2015); Wibisono et al. (2016); Wilson et al. (2016); Shi et al. (2018, 2019); An et al. (2018); Zhou et al. (2017)). Very recently, França et al. (2018, 2019) made a significant step towards understanding the alternating direction method of multipliers (ADMM) using the tools from continuous dynamical systems. ADMM is now a widely used algorithm for solving problems with separable structures, which include a lot of important applications arising from image processing, signal processing, machine learning, statistics, etc. It has a close connection with some classical operator-splitting methods in numerical PDEs such as Douglas-Rachford (Douglas & Rachford, 1956) and Peaceman-Rachford (Peaceman & Rachford, 1955) operator-splitting methods that dated back to the 1950s. These operator-splitting methods were later studied in Gabay & Mercier (1975); Glowinski & Marroco (1975); Gabay (1983); Fortin & Glowinski (1983); Glowinski & Le Tallec (1989); Eckstein & Bertsekas (1992). But the renaissance of ADMM was due to several works in 2007-2008 that introduced this algorithm to solving signal processing and image processing problems (Combettes & Pesquet, 2007; Goldstein & Osher, 2009; Wang et al., 2008; Mardani et al., 2018). Since then, ADMM was successfully used for solving important applications in many areas in science and engineering. The popularity and importance of ADMM has been partly demonstrated by the recognition of the highly influential survey paper Boyd et al. (2011). As a result, the works of França et al. (França et al., 2018, 2019) are very timely and important as they provided new tools for further understanding the convergence behavior of this influential algorithm.

However, one major drawback of França et al. (2018, 2019) is that they assume that the objective function is smooth, which in fact rules out most of the applications solved by ADMM and its variants. More specifically, França et al. (2018, 2019) consider the following problem:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x)+g(Ax), \tag{1.1}$$

where $A\in\mathbb{R}^{m\times d}$, $f:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$, $g:\mathbb{R}^m\to\mathbb{R}\cup\{+\infty\}$, and both f and g are convex and differentiable. We need to point out that under such assumptions, there exist many other efficient algorithms for solving (1.1) and ADMM may not be a good choice.

In this paper, we allow f and g to be nonsmooth functions in (1.1). To apply ADMM, one standard technique is to rewrite (1.1) as

$$\begin{aligned}\underset{x\in\mathbb{R}^d,z\in\mathbb{R}^m}{\text{minimize}}\quad & f(x)+g(z)\\ \text{subject to}\quad & Ax-z=0.\end{aligned} \tag{1.2}$$

Note that the classical setting of linear constraint $Ax+Bz=c$ can be reformulated as $z=Ax$ by a simple linear transformation operation when B is invertible.

One typical iteration of ADMM for solving (1.2) is

$$x_{k+1} := \operatorname*{argmin}_x \left[f(x) + \frac{\rho}{2}\|Ax - z_k + u_k\|_2^2 \right], \tag{1.3a}$$

$$z_{k+1} := \operatorname*{argmin}_z \left[g(z) + \frac{\rho}{2}\|Ax_{k+1} - z + u_k\|_2^2 \right], \tag{1.3b}$$

$$u_{k+1} := u_k + Ax_{k+1} - z_{k+1}, \tag{1.3c}$$

where u is the (scaled) Lagrange multiplier and $\rho > 0$ is a penalty parameter in the augmented Lagrangian function:

$$\mathcal{L}_\rho(x, z, u) = f(x) + g(z) + \rho u^\top (Ax - z) + \frac{\rho}{2}\|Ax - z\|_2^2. \tag{1.4}$$

Throughout this paper, we assume that both f and g are proper closed convex functions.

1.1 Our Contributions

We extend the analysis of França et al. (2018, 2019) to nonsmooth cases using a completely different tool: differential inclusion, which is motivated by the analysis of FISTA by Vassilis et al. (2018). More specifically, we analyze the convergence rate of continuous limit of two widely used ADMM variants and two generalizations for nonsmooth problems: linearized ADMM, gradient-based ADMM, generalized ADMM and accelerated generalized ADMM. We anticipate that these results will provide new insights on understanding ADMM for solving nonsmooth problems.

After the initial submission and the acceptance of the short version of this work in May 2019, the authors became aware of the concurrent work by Franca et al. (2019) has appeared online, who has extended the ADMM analysis of França et al. (2018, 2019) to the context of nonsmooth constrained optimization. Independent of our work, our analysis are substantially different in several aspects. Franca et al. (2019) studies the generalized ADMM and their Nesterov's and Polyak's accelerations to the nonsmooth cases and obtains convergence rates highly related to ours, but we concentrate more on linearized and gradient-based ADMM as practical ADMM variants. Furthermore, our analysis utilizes the approximate differential equation techniques while França et al. (2019) adopts a straightforward analysis. We believe both works serve as independent interests to the community.

1.2 Notations

Let $F, f, g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be function mappings from $\mathbb{R}^d$ to the extended real space $\mathbb{R} \cup \{+\infty\}$. Let $x_k \in \mathbb{R}^d, z_k, u_k, \hat{z}_k, \hat{u}_k \in \mathbb{R}^m$ be solution sequences of the ADMM iterates, while the capital letters $X(t), Z(t), U(t)$ denote the solutions of the continuous dynamical systems. Let $\langle \cdot, \cdot \rangle$ be the inner product, and let $\|\cdot\|_2$ and $\|\cdot\|_1$ be the L_2 and L_1 norms, respectively. Let τ_L and τ_G be parameters controlling the quadratic penalties in the linearized and gradient-based cases, respectively. We use k to denote the discrete timestep and t denotes the continuous time. Other

notations will be explained at their first entrances.

2 Continuous Limit of Linearized ADMM and Gradient-Based ADMM

We need to point out that the ADMM given in (1.3) is rarely used in practice, because for most applications, the x-subproblem does *not* have closed-form solutions and an iterative solver is still needed to solve it. Note that although the z-subproblem in (1.3) corresponds to the proximal mapping of function g, the x-subproblem does *not* correspond to the proximal mapping of f because of the presence of matrix A. Moreover, it is possible that in some applications, f does *not* have an easy proximal mapping. Two most commonly used nonsmooth ADMM variants in practice are linearized ADMM and gradient-based ADMM, and they are suitable for the following two cases, which cover most applications of ADMM:

Case (i): *Linearized ADMM* is suitable for the case where f is nonsmooth with easy proximal mappings; one representative application in this case is the Lasso problem where $f(x) = \|x\|_1$ and $g(z) = \frac{1}{2}\|z - b\|_2^2$ (Tibshirani, 1996).[1]

Case (ii): *Gradient-based ADMM* is suitable for the case where g is nonsmooth with easy proximal mapping, f is differentiable but does not have an easy proximal mapping; one representative application in this case is the sparse logistic regression problem where g is the ℓ_1 norm and f is the logistic loss function (Liu et al., 2009). Note that $A = I$ in this particular application.

We now provide more details about the applicability of linearized ADMM and gradient-based ADMM. In Case (i), where both f and g are nonsmooth with easy proximal mappings, the z-subproblem (1.3b) corresponds to the proximal mapping of g and is thus easy to solve; however the presence of matrix A brings difficulty to solving the x-subproblem (1.3a). Linearized ADMM addresses this issue by adding a suitably chosen proximal term $\frac{1}{2}\|x - x_k\|_{\tau_L I - \rho A^\top A}$ to the objective function of (1.3a),[2] which results in the following subproblem whose solution corresponds to the proximal mapping of f:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x) + \frac{\tau_L}{2}\left\|x - \left(x_k - \frac{\rho}{\tau_L}A^\top(Ax_k - z_k + u_k)\right)\right\|_2^2,$$

where $1/\tau_L$ can be viewed as the step size of the gradient step of the quadratic penalty. Notice that by the above equation we are making a first-order Taylor approximation to the second term of (1.3a) to avoid trouble caused by matrix A. Combining this subproblem with (1.3b) and (1.3c) yields the linearized ADMM. Here we consider a slightly more general version of linearized ADMM

[1] Both f and g can be nonsmooth in this case, so we could also write $f = \frac{1}{2}\|\cdot - b\|_2^2$ and $g = \|\cdot\|_1$. We represent f and g the other way around to follow the notations in Yang & Zhang (2011) as an illustration.

[2] The norm $\|x\|_{\tau_L I - \rho A^T A}$ is defined as $x^\top(\tau_L I - \rho A^T A)x$.

by adding a relaxation term to the intermediate residual $Ax_{k+1} - z_k$, which is summarized in (2.1):

$$x_{k+1} = \operatorname*{argmin}_x \left\{ f(x) + \frac{\tau_L}{2} \left\| x - \left(x_k - \frac{\rho}{\tau_L} A^\top (Ax_k - z_k + u_k) \right) \right\|_2^2 \right\}, \tag{2.1a}$$

$$z_{k+1} = \operatorname*{argmin}_z \left\{ g(z) + \frac{\rho}{2} \|\alpha A x_{k+1} + (1-\alpha) z_k - z + u_k\|_2^2 \right\}, \tag{2.1b}$$

$$u_{k+1} = u_k + (\alpha A x_{k+1} + (1-\alpha) z_k - z_{k+1}), \tag{2.1c}$$

where $\alpha \in (0, 2)$ is a relaxation parameter, and when $\alpha = 1$, it reduces to the classical linearized ADMM. As a widely used nonsmooth ADMM variant, linearized ADMM has been studied by many researchers, (see, e.g., Chen & Teboulle (1994); Eckstein (1994); He et al. (2002); Zhang et al. (2010); Yang & Zhang (2011); Lin et al. (2011); Ma (2016); Xu (2015); Yang & Yuan (2013); Ouyang et al. (2015)). The difference between (1.3b)-(1.3c) and (2.1b)-(2.1c) is that Ax_{k+1} is replaced by $\alpha Ax_{k+1} + (1-\alpha)z_k$. This is called *relaxation*, which has been suggested in many papers (see, e.g., Eckstein & Bertsekas (1992)) to provide more flexibility and potentially improve the convergence speed of the algorithm.

We now use the total variation minimization problem (Rudin et al., 1992) as an example to show how (2.1) works for a particular problem. The total variation minimization problem can be casted as the following form after variable splitting,

$$\begin{aligned} \underset{x,z}{\text{minimize}} \quad & \frac{1}{2}\|x - b\|_2^2 + \lambda \|z\|_1 \\ \text{subject to} \quad & z = Dx, \end{aligned} \tag{2.2}$$

which is in the form of (1.1) with $f(\cdot) = \lambda \|\cdot\|_1$ and $g(\cdot) = \frac{1}{2}\|\cdot - b\|^2$. When linearized ADMM (2.1) is applied to solve (2.2), the two subproblems (2.1a) and (2.1b) respectively correspond to the proximal mappings of $\|\cdot\|_1$ and $\frac{1}{2}\|\cdot - b\|_2^2$, which are both very easy to compute.

In Case (ii), gradient-based ADMM is suitable for the case where f does *not* have an easy proximal mapping. In this case, the linearized ADMM (2.1) is not a good choice, because the x-subproblem (2.1a) is still not easy to solve. As a result, the gradient-based ADMM is proposed to address this issue. A typical iteration of gradient-based ADMM is as follows:

$$x_{k+1} = x_k - \frac{1}{\tau_G} \left(\nabla f(x_k) + \rho A^\top (Ax_k - z_k + u_k) \right), \tag{2.3a}$$

$$z_{k+1} = \operatorname*{argmin}_z \left\{ g(z) + \frac{\rho}{2} \|\alpha A x_{k+1} + (1-\alpha) z_k - z + u_k\|_2^2 \right\}, \tag{2.3b}$$

$$u_{k+1} = u_k + (\alpha A x_{k+1} + (1-\alpha) z_k - z_{k+1}), \tag{2.3c}$$

where we again used a relaxation term to make the algorithm more general, and $1/\tau_G$ is the step size of the negative gradient step taken on the augmented Lagrangian function $\mathcal{L}_\rho(x_k, z_k, u_k)$. Note that since we assume that f is differentiable in this case, the objective function of the x-subproblem

in (1.3a) becomes a differentiable function. As a result, gradient-based ADMM suggests that a gradient step is taken instead of minimizing the augmented Lagrangian function directly, which results in the new x-subproblem in (2.3a). This gradient-based ADMM has been studied in the literature extensively, (see, e.g., Condat (2013); Vu (2013); Davis & Yin (2017); Lin et al. (2017)).

We now use sparse logistic regression (Koh et al., 2007) as an example to show how (2.3) works for a particular problem. The sparse logistic regression problem can be casted as

$$\underset{x}{\text{minimize}} \ \frac{1}{N}\sum_{i=1}^{N}\log(1+\exp(-b_i(a_i^\top x+v)))+\lambda\|x\|_1, \tag{2.4}$$

which is in the form of (1.1) with f being the logistic loss function, g being the ℓ_1 norm and $A=I$. Note that the logistic loss function f does not adopt an easy proximal mapping, but it is differentiable and thus a gradient step can be taken for the x-subproblem. When the gradient-based ADMM (2.3) is applied to solve (2.4), the two subproblems (2.3a) and (2.3b) are both easy to be implemented.

2.1 Main Results

In this subsection, we present the main results on the convergence of the continuous limit of the iterates of linearized ADMM (2.1) and gradient-based ADMM (2.3). We focus on the continuous approximation results when $t=\rho^{-1}k$ and $\rho\to\infty$, where t denotes the time and k is the iteration counter. The following definitions and assumption are needed for our results.

Definition 1. *(Nesterov, 2004). A vector v is called a subgradient of the function f at a point x_0 satisfying $f(x_0)<\infty$ if for any y satisfying $f(y)<\infty$, we have*

$$f(y)\geq f(x_0)+v^\top(y-x_0).$$

The set of all subgradients of f at x_0, $\partial f(x_0)$, is called the subdifferential of the function f at the point x_0.

Definition 2. *A function f defined on $\mathbb{R}^d$ and taking values in $\mathbb{R}\cup\{+\infty\}$ is called closed, if its epigraph*

$$\{(x,t)\in\mathbb{R}^d\times\mathbb{R}\mid f(x)\leq t\}$$

is a closed set in $\mathbb{R}^d\times\mathbb{R}$.

Definition 3. *A function f defined on $\mathbb{R}^d$ and taking values in $\mathbb{R}\cup\{+\infty\}$ is called convex, if for every $x,y\in\mathbb{R}^d$ and $\theta\in(0,1)$,*

$$f(\theta x+(1-\theta)y)\leq\theta f(x)+(1-\theta)f(y).$$

Assumption 1. *We assume that f and g are closed convex functions taking values in $\mathbb{R}\cup\{+\infty\}$ and properly defined over $\mathbb{R}^d$ (in the sense that $f(x)<+\infty$ for at least one x), and $F(x):=f(x)+g(Ax)$*

has a global minimum. We assume further that matrix A has full column rank and singular values $\sigma_1 \geq \cdots \geq \sigma_d > 0$.

Assumption 1 implies that f, g, F are proper, lower semicontinuous and subdifferentiable (Nesterov, 2004). For linearized ADMM, we have the following theorem for continuous-time approximation:

Theorem 1 (Linearized ADMM)**.** *Let Assumption 1 hold, and the relaxation parameter $\alpha \in (0, 2)$. Rescaling the time by setting $t = \rho^{-1}k$, the continuous-time limit of the iterates $\{x_k\}$ of linearized ADMM (2.1) as $\rho \to \infty$ and $\tau_L/\rho \to c \in (0, \infty)$ is given by the differential inclusion*

$$0 \in \partial F(X(t)) + \left(cI + \frac{1-\alpha}{\alpha} A^\top A\right) \dot{X}(t), \tag{2.5}$$

with initial value $X(0) = x_0$.

Analogously, we have the following Theorem 2 for the continuous limit of iterates of gradient-based ADMM (2.3):

Theorem 2 (Gradient-based ADMM)**.** *Let Assumption 1 hold, and the relaxation parameter $\alpha \in (0, 2)$. In addition, we assume that f is smooth. Rescaling the time by setting $t = \rho^{-1}k$, the continuous-time limit of the iterates $\{x_k\}$ of gradient-based ADMM (2.3) as $\rho \to \infty$ and $\tau_G/\rho \to c \in (0, \infty)$ is given by the same differential inclusion (2.5) in Theorem 1 with initial value $X(0) = x_0$.*

Remark 1. *Here we assume mildly that the solution of the differential inclusion (2.5) exists and is unique. The existence and uniqueness of a solution are usually dealt with classical monotone theory (Vassilis et al., 2018). However in some cases, the uniqueness of a solution does not necessarily hold. Theorem 1 in Aubin (1984) provides an additional assumption that can ensure the uniqueness of a solution within the context of this work.*[3]

The next theorem shows the convergence rate of the continuous-time limit of the iterates $\{x_k\}$ generated by linearized ADMM (2.1) and gradient-based ADMM (2.3) in the convex case. We use x^* to denote an arbitrary minimizer of F. We recall that σ_1 is the largest singular value of A and σ_d is the smallest singular value of A, and assume that $cI + \frac{1-\alpha}{\alpha} A^\top A$ is a well defined positive definite matrix. For simplicity of notations, we define κ_1 and κ_d the largest and smallest singular value of $cI + \frac{1-\alpha}{\alpha} A^T A$, respectively.

Theorem 3 (Convergence Rates)**.** *Let Assumption 1 hold. Assume that c and α are chosen such that the matrix $(cI + \frac{1-\alpha}{\alpha} A^\top A)$ is positive definite, with largest and smallest eigenvalues being κ_1^2, κ_d^2 ($\kappa_1, \kappa_d > 0$). Let $X(t)$ be any shock solution*[4] *of the differential inclusion (2.5) with convex objective function F and initial value $X(0) = x_0$. Then $X(t)$ has bounded trajectory and the function value*

[3] For conditions that guarantee the existence and uniqueness of the solution to differential inclusion, we refer the readers to Adly et al. (2006); Attouch et al. (2002); Paoli (2000).

[4] The concept of shock solution is given in Attouch et al. (2002); Paoli (2000); Vassilis et al. (2018). Since the goal of this paper is to illustrate the main idea, we include its definition and existence result in Appendix A for completeness.

gap has $\mathcal{O}(t^{-1})$ convergence rate almost everywhere, i.e., for a minimizer x^* and a.e. $t > 0$ it holds that $\|X(t) - x^*\|_2 \leq \frac{\kappa_1}{\kappa_d}\|x_0 - x^*\|_2$, and

$$F(X(t)) - F(x^*) \leq \frac{\kappa_1^2\|x_0 - x^*\|_2^2}{2t}.$$

Moreover, we have

$$\int_0^{+\infty} [F(X(t)) - F(x^*)] \leq \frac{\kappa_1^2}{2}\|x_0 - x^*\|_2^2,$$

and

$$\int_0^{+\infty} t\|\dot{X}(t)\|_2^2 \leq \frac{\kappa_1^2}{2\kappa_d^2}\|x_0 - x^*\|_2^2.$$

We notice that the convergence rate in continuous time depends heavily on α: when $0 < \alpha \leq 1$, $\kappa_1 = \sqrt{c + \frac{1-\alpha}{\alpha}\sigma_1^2}$ and $\kappa_d = \sqrt{c + \frac{1-\alpha}{\alpha}\sigma_d^2}$, and when $1 < \alpha < 2$, $\kappa_1 = \sqrt{c + \frac{1-\alpha}{\alpha}\sigma_d^2}$ and $\kappa_d = \sqrt{c + \frac{1-\alpha}{\alpha}\sigma_1^2}$, where σ_1, σ_d are singular value of matrix A.

3 Continuous Limit of Generalized ADMM

In this section, we study the continuous limit of the generalized ADMM (G-ADMM) proposed by Eckstein & Bertsekas (1992). We point out that this has been studied by França et al. (2018, 2019) for the cases where f and g are both smooth. We now extend the analysis to problems with nonsmooth f and g. G-ADMM allows more flexibility of ADMM by introducing a new relaxation parameter $\alpha \in (0, 2)$, and it updates the iterates as

$$x_{k+1} = \operatorname*{argmin}_x \left\{ f(x) + \frac{\rho}{2}\|Ax - z_k + u_k\|_2^2 \right\}, \tag{3.1a}$$

$$z_{k+1} = \operatorname*{argmin}_z \left\{ g(z) + \frac{\rho}{2}\left\|\alpha A x_{k+1} + (1-\alpha)z_k - z + u_k\right\|_2^2 \right\}, \tag{3.1b}$$

$$u_{k+1} = u_k + (\alpha A x_{k+1} + (1-\alpha)z_k - z_{k+1}). \tag{3.1c}$$

In the following, §3.1 studies the differential inclusion approximation for G-ADMM and §3.2 studies that for Accelerated G-ADMM.

3.1 Differential Inclusion for G-ADMM

Following França et al. (2018), we rescale the time by a factor of ρ^{-1}, i.e. let $t = \rho^{-1}k$, and obtain the continuous-time approximation. One can see from (3.1a) that larger parameter $\rho > 0$ results in smaller-pace updates of x_k. We study the limit of updates in the regime $\rho \to \infty$. Our main result on the differential inclusion approximation of G-ADMM is as follows:

Theorem 4. *Let Assumption 1 hold, and the relaxation parameter $\alpha \in (0,2)$. Rescale the time by setting $t = \rho^{-1}k$, the continuous limit of iterates of $\{x_k\}$ in Algorithm (B.1) as $\rho \to \infty$ is given by the following differential inclusion:*

$$\frac{1}{\alpha}(A^\top A)\dot{X}(t) + \partial F\left(X(t)\right) \ni 0, \tag{3.2}$$

with $X(0) = x_0$.

We move forward and analyze the convergence property of differential inclusion (3.2). Recall that σ_1 and σ_d are the largest and smallest singular values of A, respectively.

Theorem 5. *When the Assumption 1 holds, the shock solution $X(t)$ of differential inclusion (3.2) has bounded trajectory and $\mathcal{O}(t^{-1})$ convergence rate almost everywhere, i.e., for a.e. $t \geq 0$, $\|X(t) - x^*\|_2 \leq \frac{\sigma_1}{\sigma_d}\|x_0 - x^*\|_2$, and*

$$F(X(t)) - F(x^*) \leq \frac{\sigma_1^2\|x_0 - x^*\|_2^2}{2\alpha t}.$$

Moreover,

$$\int_0^{+\infty} [F(X(t)) - F(x^*)]dt \leq \frac{\sigma_1^2}{2\alpha}\|x_0 - x^*\|_2^2,$$

$$\int_0^{+\infty} t\|\dot{X}(t)\|_2^2 dt \leq \frac{\sigma_1^2}{2\sigma_d^2}\|x_0 - x^*\|_2^2.$$

Here, the key idea of the convergence analysis of differential inclusion (3.2) is to use a sequence of approximating differential equations (ADE) that approaches the differential inclusion.

3.2 Differential Inclusion for Accelerated G-ADMM

Goldstein et al. (2014) proposed an accelerated ADMM by incorporating Nesterov's extrapolation technique. (Nesterov, 1983) This method is generalized by França et al. (2018, 2019) which jointly consider relaxation and acceleration. The accelerated G-ADMM considered in França et al. (2018, 2019) uses $\gamma_{k+1} = \frac{k}{k+r}$ as the momentum coefficient (r being a positive constant), and is described as follows:

$$x_{k+1} = \operatorname*{argmin}_x \left\{ f(x) + \frac{\rho}{2}\|Ax - \hat{z}_k + \hat{u}_k\|_2^2 \right\}, \tag{3.3a}$$

$$z_{k+1} = \operatorname*{argmin}_z \left\{ g(z) + \frac{\rho}{2}\|\alpha A x_{k+1} + (1-\alpha)\hat{z}_k - z + \hat{u}_k\|_2^2 \right\}, \tag{3.3b}$$

$$u_{k+1} = \hat{u}_k + (\alpha A x_{k+1} + (1-\alpha)\hat{z}_k - z_{k+1}), \tag{3.3c}$$

$$\hat{u}_{k+1} = u_{k+1} + \gamma_{k+1}(u_{k+1} - u_k), \tag{3.3d}$$

$$\hat{z}_{k+1} = z_{k+1} + \gamma_{k+1}(z_{k+1} - z_k). \tag{3.3e}$$

Here we extend the results in França et al. (2018, 2019) to the case where f and g are nonsmooth functions. We present our first main result in the following theorem:

Theorem 6. *Let Assumption 1 hold, and let the relaxation parameter $\alpha \in (0, 2)$. Rescale the time by setting $t = \rho^{-1/2}k$, the continuous-time approximation of Algorithm (3.3) iteration as $\rho \to \infty$ is given by the differential inclusion*

$$\frac{1}{\alpha}(A^\top A)\left(\ddot{X}(t) + \frac{r}{t}\dot{X}(t)\right) + \partial F(X(t)) \ni 0, \tag{3.4}$$

with $X(t_0) = x_0$ and $\dot{X}(t_0) = 0$. Here t_0 is an arbitrary positive starting time.

Unlike all previous time rescaling scheme where $k = \rho t$, for accelerated G-ADMM the time rescaling is $k = \rho^{1/2}t$, which is in accordance with the idea of *acceleration*. Factor r/t is the *damping ratio* of differential inclusion (3.4). We refer to the case $r \geq 3$ as high friction case, $0 < r < 3$ as low friction case and provide two separate convergence theorems under these two cases. Moreover, we specifically point out that the convergence rate can be *sharper* when $r > 3$.

In the following, we define $\Delta_0^2 = \max\{t_0^2(F(x_0) - F(x^*)), \|x_0 - x^*\|_2^2\}$. We remark that all the factors C used in this subsection can possibly depend on friction parameter r, the relaxation parameter α, and singular values σ_1 and σ_d. Firstly, we show that $X(t)$ has almost surely bounded trajectory for $t \geq t_0$, and the convergence rate is $\mathcal{O}(t^{-2})$ in terms of both $F(X(t)) - F(x^*)$ and $\|\dot{X}(t)\|_2^2$.

Theorem 7 (High Friction)**.** *When $r \geq 3$, the shock solution $X(t)$ of differential inclusion (3.4) has bounded trajectory and $\mathcal{O}(t^{-2})$ convergence rate almost everywhere, i.e. there exists positive factors C_1, C_2, C_3 depending on r, α such that, for a.e. $t \geq t_0$, $\|X(t) - x^*\|_2 \leq C_1\Delta_0$, and*

$$F(X(t)) - F(x^*) \leq \frac{C_2\Delta_0^2}{t^2}, \qquad \|\dot{X}(t)\|_2 \leq \frac{C_3\Delta_0}{t}.$$

When $r > 3$, there exist positive factors C_4, C_5 depending on r, α such that

$$\int_{t_0}^{\infty} t(F(X(t)) - F(x^*))dt \leq C_4\Delta_0^2,$$

$$\int_{t_0}^{\infty} t\|\dot{X}(t)\|_2^2 dt \leq C_5\Delta_0^2.$$

Using the bounded integration result in Theorem 7 we can show that when $r > 3$, the convergence rate is in fact $o(t^{-2})$:

$$\lim_{t\to\infty} t^2(F(X(t)) - F(x^*)) = 0, \qquad \lim_{t\to\infty} t\|\dot{X}(t)\|_2 = 0.$$

The constant r has the so-called *magic number* 3, as discussed in details by Su et al. (2016), in the sense that an $\mathcal{O}(t^{-2})$ convergence rate could be guaranteed for AGM in high friction case $r \geq 3$, but not for the low friction case $0 < r < 3$. The follow-up works Attouch et al. (2017), Attouch et al.

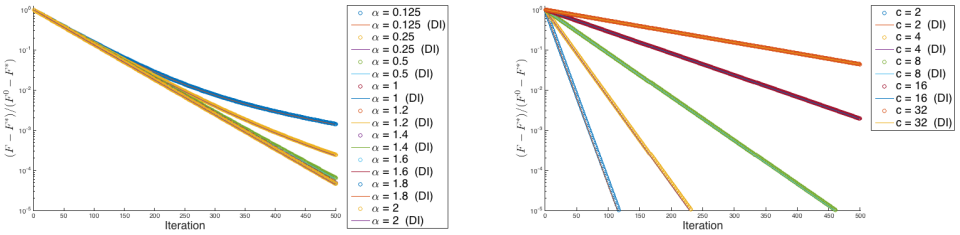

Figure 1: On total variation minimization problem, the plots are the trajectory of linearized ADMM with $\rho = 10$ and the corresponding differential inclusion, the first plot is for different α from 2^{-3} to 2 when $c = 10$, second plot is for different c from 1 to 32 when $\alpha = 1.6$.

(2018), Vassilis et al. (2018), França et al. (2018) derive the $\mathcal{O}(t^{-2r/3})$ convergence rate for AGM under low friction case and extend $\mathcal{O}(t^{-2})$ convergence rate from AGM to accelerated G-ADMM in high friction case, respectively.

For nonsmooth and low friction case, we show a $\mathcal{O}(t^{-2r/3})$ convergence rate for $F(X(t)) - F(x^*)$, and for $\|\dot{X}(t)\|_2^2$ as long as the trajectory is almost surely bounded:

Theorem 8 (Low Friction). *When $0 < r < 3$, the shock solution $X(t)$ of differential inclusion (3.4) has $\mathcal{O}(t^{-2r/3})$ convergence rate almost everywhere, i.e. there exists positive factor C_6 depending on r, α such that, for a.e. $t \geq t_0$,*

$$F(X(t)) - F(x^*) \leq \frac{C_6 t_0^{-2(3-r)/3} \Delta_0^2}{t^{2r/3}}.$$

If in addition the trajectory $\{X(t)\}_{t \geq t_0}$ is bounded almost everywhere for $t \geq t_0$, then there also exists some positive factor C_7 depending on r, α such that for a.e. $t \geq t_0$,

$$\|\dot{X}(t)\|_2 \leq \frac{C_7 t_0^{-(3-r)/3} \Delta_0}{t^{r/3}}.$$

Theorems 7 and 8 provide convergence results for accelerated G-ADMM in continuous-time scheme with the second-order differential inclusion and accompanying tools, which sheds light on the discrete-time algorithm (Su et al., 2016).

We highlight that Franca et al. (2019) has concurrently derived the differential inclusion approximation of generalized (relaxed) ADMM and its accelerated versions in the nonsmooth case and obtained convergence rates of similar forms of ours.

4 Numerical Experiments

According to the convergence theorems in previous sections, both the linearized ADMM and the gradient ADMM share the same differential inclusion (2.5). In the following two examples, we show

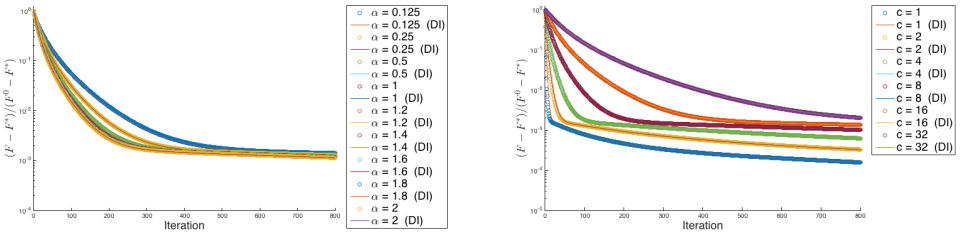

Figure 2: On sparse logistic regression, the plots are gradient ADMM and the differential inclusion when $\rho = 10$, first plot is for different α from 2^{-3} to 2 when $c = 10$, second plot is for different c from 1 to 32 when $\alpha = 1.6$.

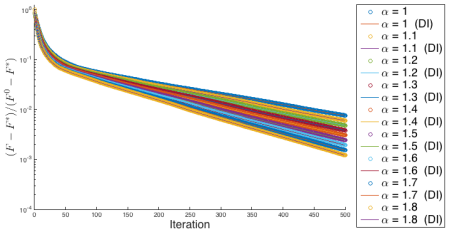

Figure 3: On Lasso problem, the plots are the trajectory of G-ADMM with $\rho = 50$ and the corresponding differential inclusion.

that for $\rho = 10$ the trajectory of the linearized or gradient-based ADMM algorithm is very close to the trajectory of the corresponding differential inclusion.

To plot the trajectory of the differential inclusion numerically, we use entropic approximation Teboulle (1992) of the nonsmooth objective to compute the trajectory of the differential inclusion, we remark that entropic approximation is parrellel in theory to Moreau-Yosida approximation, which use quadratic approximation. Entropic approximation would provide a smooth approximation of the sub-gradient for $\|z\|_1$ as $\tanh \beta z$, where we choose $\beta = 10^6$ in the following experiments, so that the numerical approximation error is of the order 10^{-6}.

4.1 Total Variation Minimization with Linearized ADMM

Consider the total variation minimization problem (2.2), see Rudin et al. (1992). Using variable splitting, we can write the problem as

$$\begin{aligned} \underset{x,z}{\text{minimize}} \quad & \frac{1}{2}\|x-b\|_2^2 + \lambda\|z\|_1 \\ \text{subject to} \quad & z = Dx, \end{aligned}$$

where D is the finite difference matrix, $x, z \in \mathbb{R}^n$. This problem fits to the general framework of ADMM with $A = D$, $f(x) = \frac{1}{2}\|x-b\|_2^2$ and $g(z) = \lambda\|z\|_1$, since f is quadratic with easy proximal and g is nonsmooth with easy proximal, namely, its proximal operator is the soft thresholding

operator $S_\lambda(z) = \text{sign}(z)(|z| - \lambda)_+$, we could use linearized ADMM to solve this problem.

We generated a problem instance, where the true signal x^{true} is a piece-wise constant signal, the observation is $b = x^{\text{true}} + n$, $n \sim \mathcal{N}(0, I)$. We solve the total variantion minimization with $\lambda = 0.01$ to recover the true signal x^{true} from b, following the MATLAB examples of the paper Boyd et al. (2011).

In the following, we run both the linearized ADMM algorithm with $\rho = 10$ and the differential inclusion. We demonstrate the trajetory of linearized ADMM algorithm and the differential inclusion for different parameter configuration α and c. First of all, we fix $c = 10$, and vary α from 2^{-3} to 2; then, we fix $\alpha = 1.6$, and vary c from 1 to 32. Figures 1 are the trajectory of linearized ADMM with $\rho = 10$ and the corresponding differential inclusion, with the x axis being the iteration count, the y axis being the relative error $\frac{F(x_k)-F(x^*)}{F(x_0)-F(x^*)}$, where $F(x^*)$ is the function value at optimal solution, $F(x_0)$ is the function value at initialization. We can see that the differential inclusion matches the trajectory of the linearized ADMM algorithm very closely for all parameter settings.

4.2 Sparse Logistic Regression with Gradient ADMM

Consider the sparse logistic regression problem (2.4)(see Koh et al. (2007) for more details). Using variable splitting, we can write the problem as

$$\begin{aligned} \underset{x,z}{\text{minimize}} \quad & \frac{1}{N}\sum_{i=1}^{N} \log(1 + \exp(-b_i(a_i^\top x + v)) + \lambda\|z\|_1 \\ \text{subject to} \quad & z = x, \end{aligned}$$

with variable $x \in \mathbb{R}^{d-1}, v \in \mathbb{R}$.

This problem fits to the general framework with variable $\bar{x} = (x, v)$, $A = I$, $f(\bar{x}) = \log(1 + \exp(-b_i(a_i^\top x + v))$ and $g(\bar{x}) = \lambda\|\bar{x}_{1:n}\|_1$, since f is differentiable but does not have an easy proximal mapping and g is nonsmooth with easy proximal, we could use gradient ADMM to solve this problem.

We generated a problem instance following the MATLAB examples of the paper Boyd et al. (2011). More specifically, we chose a true weight vector x^{true} sampled from Bernoulli-Gaussian distribution with mean 0, variance 1 and sparsity level 0.1, along with the true intercept v^{true} sampled from standard normal. Each feature vector a_i was generated from Bernoulli-Gaussian distribution at sparsity level 0.2. The labels were then generated using $b_i = \text{sign}(a_i^\top x^{\text{true}} + v^{\text{true}} + \nu_i)$, where $\nu_i \sim \mathcal{N}(0, 0.1)$. The regularization parameter is set to $\lambda = 0.1\lambda_{\max}$ according to Koh et al. (2007), where $\lambda_{\max} = \|A^\top \tilde{b}\|_\infty$ is the critical value above which the solution of the problem is $x = 0$, where $\tilde{b}$ is defined in page 93 of Boyd et al. (2011).

We run both the gradient ADMM algorithm with $\rho = 10$ and the differential inclusion. We demonstrate the trajetory of gradient ADMM algorithm and the differential inclusion for different

parameter configuration α and c. First of all, we fix $c = 10$, and vary α from 2^{-3} to 2; then, we fix $\alpha = 1.6$, and vary c from 1 to 32. Figures 2 are the trajectory of gradient ADMM with $\rho = 10$ and the corresponding differential inclusion, with the x axis being the iteration count, the y axis being the relative error $\frac{F(x_k)-F(x^*)}{F(x_0)-F(x^*)}$, where $F(x^*)$ is the function value at optimal solution, $F(x_0)$ is the function value at initialization. We can see that the differential inclusion matches the trajectory of the gradient ADMM algorithm very closely for all parameter settings.

4.3 Lasso with G-ADMM

Additionally, we show that the trajectory of G-ADMM algorithm with different α is close to the trajectory of the differential inclusion in Lasso example Tibshirani (1996).

The Lasso problem can be casted as

$$\begin{aligned} \underset{x,z}{\text{minimize}} \quad & \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|z\|_1 \\ \text{subject to} \quad & z = x, \end{aligned}$$

which fits to the general framework with $f(\cdot) = \frac{1}{2}\|A\cdot - b\|_2^2$ and $g(\cdot) = \lambda\|\cdot\|_1$. When G-ADMM (B.1) is applied to solve this problem, the two subproblems (3.1a) and (3.1b) respectively correspond to the proximal mappings of $\frac{1}{2}\|A\cdot - b\|_2^2$ and $\|\cdot\|_1$, which are both very easy to compute.

We generated a problem instance following the MATLAB examples of the paper Boyd et al. (2011). Specifically, we sample true sparse signal x_0 from Bernoulli-Gaussian distribution with mean 0, variance 1 and sparsity level 0.05, A is sampled from Gaussian random matrix of size 100 by 400 with columns norm normalized to one, $b = Ax_0 + v$, where $v \sim \mathcal{N}(0, 0.001)$. The regularization parameter is set to $\lambda = 0.1\lambda_{\max}$ according to Koh et al. (2007), where $\lambda_{\max} = \|A^\top b\|_\infty$ is the critical value above which the solution of the problem is $x = 0$.

We run both the G-ADMM algorithm with $\rho = 50$ and the differential inclusion. We vary α from 1 to 1.8 as the range that people typically use for G-ADMM(Boyd et al., 2011). Figure 3 is the trajectory of G-ADMM with $\rho = 50$ and the corresponding differential inclusion, with the x axis being the iteration count, the y axis being the relative error $\frac{F(x_k)-F(x^*)}{F(x_0)-F(x^*)}$, where $F(x^*)$ is the function value at optimal solution, $F(x_0)$ is the function value at initialization. We can see that the differential inclusion matches the trajectory of the linearized ADMM algorithm very closely for all parameter settings.

5 Conclusions

In this paper, we analyzed the convergence behavior of the continuous limits of some widely used nonsmooth ADMM variants: linearized ADMM, gradient-based ADMM, as well as G-ADMM and its Nesterov's acceleration. Such continuous limits are characterized by the tool of differential

inclusion and promote understandings of these ADMM variants from the angles of dynamical systems. Our novel continuous-time convergence theorems characterize these ADMM variants, which is further supported by experimental results. The differential inclusion for linearized and gradient ADMM variants suggests that we could choose the algorithmic parameters via a principled approach that uses the condition number of the matrix, which serves as a practical guidance learned from theoretical insights.

Acknowledgement CJL would like to thank Tencent Technology for supporting his research. The authors would like to thank Guilherme França for discussing in depth the work Franca et al. (2019). We also sincerely thank Wotao Yin for highly enlightening discussions while this work was in preparation.

References

Adly, S., Attouch, H., & Cabot, A. (2006). Finite time stabilization of nonlinear oscillators subject to dry friction. In *Nonsmooth mechanics and analysis* (pp. 289–304). Springer.

Afriat, S. (1971). Theory of maxima and the method of lagrange. *SIAM Journal on Applied Mathematics*, 20(3), 343–357.

An, W., Wang, H., Sun, Q., Xu, J., Dai, Q., & Zhang, L. (2018). A pid controller approach for stochastic optimization of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8522–8531).

Attouch, H., Cabot, A., & Redont, P. (2002). The dynamics of elastic shocks via epigraphical regularization of a differential inclusion barrier and penalty approximations. *Adv. Math. Sci. Appl.*, 12, 273–306.

Attouch, H., Chbani, Z., Peypouquet, J., & Redont, P. (2018). Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168(1-2), 123–175.

Attouch, H., Chbani, Z., & Riahi, H. (2017). Rate of convergence of the nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$. *arXiv preprint arXiv:1706.05671*.

Aubin, J.-P. (1984). Cellina. *Differential Inclusions*, 264.

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

Chen, G. & Teboulle, M. (1994). A proximal-based decomposition method for convex minimization problems. *Mathematical Programming*, 64, 81–101.

Combettes, P. L. & Pesquet, J.-C. (2007). A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1(4), 564–574.

Condat, L. (2013). A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *J. Optimization Theory and Applications*, 158(2), 460–479.

Davis, D. & Yin, W. (2017). A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis*, 25(4), 829–858.

Douglas, J. & Rachford, H. H. (1956). On the numerical solution of the heat conduction problem in 2 and 3 space variables. *Transactions of the American Mathematical Society*, 82, 421–439.

Eckstein, J. (1994). Some saddle-function splitting methods for convex programming. *Optimization Methods and Software*, 4(1), 75–83.

Eckstein, J. & Bertsekas, D. P. (1992). On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3), 293–318.

Fortin, M. & Glowinski, R. (1983). *Augmented Lagrangian methods: applications to the numerical solution of boundary-value problems*. North-Holland Pub. Co.

França, G., Robinson, D. P., & Vidal, R. (2018). ADMM and accelerated ADMM as continuous dynamical systems. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 1559–1567).

Franca, G., Robinson, D. P., & Vidal, R. (2019). A dynamical systems perspective on nonsmooth constrained optimization. *arXiv preprint arXiv:1808.04048*.

França, G., Robinson, D. P., & Vidal, R. (2019). Relax, and accelerate: A continuous perspective on ADMM. *Lehigh University ISE Technical Report 19T-021*.

Gabay, D. (1983). Applications of the method of multipliers to variational inequalities. In M. Fortin & R. Glowinski (Eds.), *Augmented Lagrangian Methods: Applications to the Solution of Boundary Value Problems*. Amsterdam: North-Hollan.

Gabay, D. & Mercier, B. (1975). *A dual algorithm for the solution of non linear variational problems via finite element approximation*. Institut de recherche d'informatique et d'automatique.

Glowinski, R. & Le Tallec, P. (1989). *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*. Philadelphia, Pennsylvania: SIAM.

Glowinski, R. & Marroco, A. (1975). Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(R2), 41–76.

Goldstein, T., O'Donoghue, B., Setzer, S., & Baraniuk, R. (2014). Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3), 1588–1623.

Goldstein, T. & Osher, S. (2009). The split Bregman method for L1-regularized problems. *SIAM J. Imaging Sci.*, 2, 323–343.

He, B. S., Liao, L., Han, D., & Yang, H. (2002). A new inexact alternating direction method for monotone variational inequalities. *Mathematical Programming*, 92, 103–118.

Koh, K., Kim, S.-J., & Boyd, S. (2007). An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine learning research*, 8(Jul), 1519–1555.

Krichene, W., Bayen, A., & Bartlett, P. L. (2015). Accelerated mirror descent in continuous and discrete time. In *NIPS*.

Lin, T., Ma, S., & Zhang, S. (2017). An extragradient-based alternating direction method for convex minimization. *Foundations of Computational Mathematics*, 17(1), 35–59.

Lin, Z., Liu, R., & Su, Z. (2011). Linearized alternating direction method with adaptive penalty for low-rank representation. In *Advances in neural information processing systems* (pp. 612–620).

Liu, J., Chen, J., & Ye, J. (2009). Large-scale sparse logistic regression. In *SIGKDD*.

Ma, S. (2016). Alternating proximal gradient method for convex minimization. *Journal of Scientific Computing*, 68(2), 546–572.

Mardani, M., Sun, Q., Donoho, D., Papyan, V., Monajemi, H., Vasanawala, S., & Pauly, J. (2018). Neural proximal gradient descent for compressive imaging. In *Advances in Neural Information Processing Systems* (pp. 9596–9606).

Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27 (pp. 372–376).

Nesterov, Y. (2004). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

Ouyang, Y., Chen, Y., Lan, G., & Pasiliao Jr, E. (2015). An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, 8(1), 644–681.

Paoli, L. A. (2000). An existence result for vibrations with unilateral constraints: case of a nonsmooth set of constraints. *Math. Models Methods Appl. Sci.*, 10, 815–831.

Peaceman, D. H. & Rachford, H. H. (1955). The numerical solution of parabolic elliptic differential equations. *SIAM Journal on Applied Mathematics*, 3, 28–41.

Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4), 259–268.

Shi, B., Du, S. S., Jordan, M. I., & Su, W. J. (2018). Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*.

Shi, B., Du, S. S., Su, W. J., & Jordan, M. I. (2019). Acceleration via symplectic discretization of high-resolution differential equations. *arXiv preprint arXiv:1902.03694*.

Su, W., Boyd, S., & Candes, E. J. (2016). A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Journal of Machine Learning Research*, 17(153), 1–43.

Takayama, A. (1985). *Mathematical economics*. Cambridge University Press.

Teboulle, M. (1992). Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research*, 17(3), 670–690.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 267–288).

Vassilis, A., Jean-François, A., & Charles, D. (2018). The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case $b \leq 3$. *SIAM Journal on Optimization*, 28(1), 551–574.

Vu, B. C. (2013). A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3), 667–681.

Wang, Y., Yang, J., Yin, W., & Zhang, Y. (2008). A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences*, 1(3), 248–272.

Wibisono, A., Wilson, A. C., & Jordan, M. I. (2016). A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, (pp. 201614734).

Wilson, A. C., Recht, B., & Jordan, M. I. (2016). A lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*.

Xu, Y. (2015). Alternating proximal gradient method for sparse nonnegative Tucker decomposition. *Mathematical Programming Computation*, 7(1), 39–70.

Yang, J. & Yuan, X. (2013). Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of Computation*, 82(281), 301–329.

Yang, J. & Zhang, Y. (2011). Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SIAM journal on scientific computing*, 33(1), 250–278.

Zhang, X., Burger, M., Bresson, X., & Osher, S. (2010). Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. *SIAM Journal on Imaging Science*, 3, 253–276.

Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S., & Glynn, P. (2017). Mirror descent in non-convex stochastic programming. *arXiv preprint arXiv:1706.05681*.

A Preliminaries of Differential Inclusion

Recall that we denote $F(x) = f(x) + g(Ax)$, and Assumption 1 holds. To transit from the smooth case to the nonsmooth case, we use the tool of differential inclusion to build the connection between subdifferentiable F and differentiable functions. One basic example of a differential inclusion takes the form of:

$$\dot{x}(t) \in \partial F(x(t))$$

To bridge the gap between differentiable objective functions and nondifferentiable objective functions, we follow Vassilis et al. (2018) and consider the Moreau-Yosida Approximation, which is a standard tool in convex analysis.

Definition 4 (Moreau-Yosida Approximation)**.** *Moreau-Yosida Approximation of a convex function F with parameter $\mu > 0$ is defined as*

$$F_\mu(x) := \inf_y \left\{ F(y) + \frac{1}{2\mu}\|y - x\|_2^2 \right\}$$

Use $J_\mu(x)$ to denote the unique point that achieves the infimum above, then $\nabla F_\mu(x) = \frac{1}{\mu}(x - J_\mu(x))$ by the Envelope Theorem (Afriat, 1971; Takayama, 1985). For any $\mu > 0$, F_μ is a convex, continuously differentiable function.

We take the Definition 3.1 in Vassilis et al. (2018) of a shock solution to define a solution of a differential inclusion. The existence of a shock solution are described in Section 3 of Vassilis et al. (2018). More specifically, we can build a sequence $x_\mu(t)$ such that its subsequence converges, where $x_\mu(t)$ are the solutions to the Approximate Differential Equation (ADE) defined below:

Approximate Differential Equation (ADE)

We consider the Moreau-Yosida approximation $F_\mu(x)$ of the objective $F(x)$ with $\mu > 0$. We consider the following approximating ODE:

$$\begin{cases} \dot{x}_\mu(t) + \nabla F_\mu(x_\mu(t)) = 0 \\ x_\mu(0) = x_0 \end{cases}$$

Here ∇F_μ can approximate ∂F and F_μ is differentiable as is shown in the theory of Moreau-Yosida approximation.

The convergence to a shock solution is described as the Approximation Scheme ($\mathcal{AS}$):

Approximation Scheme ($\mathcal{AS}$)

Let $\{F_\mu\}_{\mu>0}$ be a family of functions such that F_μ is the Moreau–Yosida approximation of F for all $\mu > 0$. Then there exists a subsequence $\{x_\mu\}_{\mu>0}$ of solutions of (ADE) that converges to a shock solution x of differential inclusion in the following sense:

- $x_\mu \to x$ uniformly on $[0,T]$ for all $T>0$ as $\mu \to 0$
- $\dot{x}_\mu \to \dot{x}$ in $L^p([0,T];\mathbb{R}^d)$ for all $p \in [1,\infty)$ for all $T>0$ as $\mu \to 0$
- $F_\mu(x_\mu) \to F(x)$in $L^p([0,T];\mathbb{R}^d)$ for all $p \in [1,\infty)$ for all $T>0$ as $\mu \to 0$

B Proofs of the Theorems Related to Linearized ADMM and Gradient-Based ADMM

In this following sections, we prove the main results provided in Section 2.1. Sections B.1, B.2 and B.3 prove Theorems 1, 2 and 3, respectively.

B.1 Proof of Theorem 1

Proof of Theorem 1. Due to the strong convexity of the optimization subproblems (2.1a) and (2.1b), it is easy to verify that the sequence $\{x_k, z_k, u_k\}$ is unique. We have from the first-order optimality conditions of (2.1a) and (2.1b) that

$$0 \in \partial f(x_{k+1}) + \tau_L \left[x_{k+1} - \left(x_k - \frac{\rho}{\tau_L} A^\top (Ax_k - z_k + u_k) \right) \right], \tag{B.1a}$$

$$0 \in \frac{1}{\rho} \partial g(z_{k+1}) - (\alpha A x_{k+1} + (1-\alpha) z_k - z_{k+1} + u_k). \tag{B.1b}$$

We detail the proof in the following:

(i) Adding up (B.1b) and (2.1c) eliminates the common term $(\alpha A x_{k+1} + (1-\alpha) z_k - z_{k+1} + u_k)$ and reduces to a simple u-update:

$$u_{k+1} \in \frac{1}{\rho} \partial g(z_{k+1}). \tag{B.2}$$

Taking the continuous limit $\rho \to \infty$ gives $U(t) = 0$, and hence $\dot{U}(t) = 0$.[5]

(ii) Reorganize (B.1a) into the following form:

$$0 \in \partial f(x_{k+1}) + \tau_L (x_{k+1} - x_k) + \rho A^\top (Ax_k - z_k + u_k). \tag{B.3}$$

Bringing (B.2) into (B.3) leads to:

$$0 \in \partial f(x_{k+1}) + A^\top \partial g(z_k) + \tau_L (x_{k+1} - x_k) + \rho A^\top (Ax_k - z_k). \tag{B.4}$$

[5] Although the continuous version of $U(t)$ is constantly zero, it is different with $u_k = 0$. One may regard u_k as an infinitesimal number that dynamically changes in the system.

Again from (2.1c),

$$\begin{aligned} u_{k+1} - u_k &= \alpha A x_{k+1} + (1-\alpha) z_k - z_{k+1} \\ &= \alpha A(x_{k+1} - x_k) - (z_{k+1} - z_k) + \alpha(A x_k - z_k), \end{aligned}$$

and hence

$$A x_k - z_k = \frac{1}{\alpha}[(u_{k+1} - u_k) + (z_{k+1} - z_k)] - A(x_{k+1} - x_k). \tag{B.5}$$

Plugging (B.5) into (B.4) gives

$$0 \in \partial f(x_{k+1}) + A^\top \partial g(z_k) + \tau_L(x_{k+1} - x_k) + \rho A^\top \left(\frac{1}{\alpha}[(u_{k+1} - u_k) + (z_{k+1} - z_k)] - A(x_{k+1} - x_k) \right). \tag{B.6}$$

Taking the limit $\rho \to \infty$ and letting $\tau_L/\rho \to c$, using the fact that $\dot{U}(t) = 0$, (B.6) reduces to

$$0 \in \partial f(X(t)) + A^\top \partial g(Z(t)) + \left(cI - A^\top A \right) \dot{X}(t) + \frac{1}{\alpha} A^\top \dot{Z}(t). \tag{B.7}$$

(iii) We directly take the $\rho \to \infty$ limit in (2.1c) with the fact that $u_{k_1} \to u_k$ and $z_{k+1} \to z_k$, we conclude

$$Z(t) = AX(t), \qquad \dot{Z}(t) = A\dot{X}(t).$$

It is straightforward to check that

$$\partial f(X(t)) + A^\top \partial g(Z(t)) \subseteq \partial F(X(t)) \tag{B.8}$$

Combining the above and (B.7) concludes

$$0 \in \partial F(X(t)) + \left(cI + \frac{1-\alpha}{\alpha} A^\top A \right) \dot{X}(t),$$

This completes the proof.

□

B.2 Proof of Theorem 2

Proof of Theorem 2. Again the sequence $\{x_k, z_k, u_k\}$ is unique due to the strong convexity of the optimization subproblem (2.1a) and (2.1b). It follows from the optimality conditions that

$$0 = \nabla f(x_k) + \rho A^T (A x_k - z_k + u_k) + \tau_G (x_{k+1} - x_k), \tag{B.9a}$$

$$0 \in \partial g(z_{k+1}) - \rho(\alpha A x_{k+1} + (1-\alpha) z_k - z_{k+1} + u_k), \tag{B.9b}$$

$$u_{k+1} = u_k + (\alpha A x_{k+1} + (1-\alpha) z_k - z_{k+1}). \tag{B.9c}$$

Seeing τ_L in the place of τ_G, (B.9b) and (B.9c) are identical to (B.1b) and (2.1c), while (B.9a) is identical to (B.1a) with $\partial f(x_{k+1})$ replaced by $\nabla f(x_k)$.

Carrying out the proof of Theorem 1 in §B.1 gives (B.6) with $\partial f(x_{k+1})$ replaced by $\nabla f(x_k)$, and hence taking corresponding limits gives differential inclusion (B.7) with $\partial f(X(t))$ replaced by $\nabla f(X(t))$. The rest of the proof follows in the same fashion as Part (iii) in the proof of Theorem 1.

□

B.3 Proof of Theorem 3

Proof of Theorem 3. For notation simplicity, we choose a matrix B such that $B^\top B = cI + \frac{1-\alpha}{\alpha} A^\top A$. Recall that the largest and smallest singular value of B are κ_1 and κ_d. Note that when $0 < \alpha \leq 1$, $\kappa_1 = \sqrt{c + \frac{1-\alpha}{\alpha}\sigma_1^2}$ and $\kappa_d = \sqrt{c + \frac{1-\alpha}{\alpha}\sigma_d^2}$, and when $1 < \alpha < 2$, $\kappa_1 = \sqrt{c + \frac{1-\alpha}{\alpha}\sigma_d^2}$ and $\kappa_d = \sqrt{c + \frac{1-\alpha}{\alpha}\sigma_1^2}$, where σ_1, σ_d are singular value of matrix A. Then the original differential inclusion becomes $0 \in \partial F(X(t)) + (B^\top B)\dot{X}(t)$. Because Moreau-Yosida approximation $F_\mu(X_\mu(t))$ is a continuously differentiable, convex function for all $\mu > 0$, we denote an arbitrary minimizer as x_μ^*.

For each $\mu > 0$, consider the energy functional of Moreau-Yosida approximation defined as

$$\mathcal{E}_\mu(t) = t(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) + \frac{\lambda}{2}\|B(X_\mu(t) - x_\mu^*)\|_2^2, \tag{B.10}$$

where λ is an arbitrary constant greater than or equal to 1. Because F_μ is a continuously differentiable function, we could write the time derivative of $\mathcal{E}_\mu(t)$ as

$$\dot{\mathcal{E}}_\mu(t) = (F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) + t\langle \nabla F_\mu(X_\mu(t)), \dot{X}_\mu(t)\rangle + \lambda\langle B^\top B(X_\mu(t) - x_\mu^*), \dot{X}_\mu(t)\rangle. \tag{B.11}$$

By substituting $B^\top B\dot{X}_\mu(t)$ by $-\nabla F_\mu(X_\mu(t))$ and $\nabla F_\mu(X_\mu(t))$ by $-B^\top B\dot{X}_\mu(t)$ according to (2.5) and the definition of the shock solution $X_\mu(t)$ in Appendix A, we have

$$\dot{\mathcal{E}}_\mu(t) = -t\|B\dot{X}_\mu(t)\|_2^2 + (F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) - \lambda\langle (X_\mu(t) - x_\mu^*), \nabla F_\mu(X_\mu(t))\rangle \leq 0, \tag{B.12}$$

where we used the convexity of F_μ and nonnegativity of $(F_\mu(X_\mu) - F_\mu(x_\mu^*))$, $\|B\dot{X}_\mu\|_2$ in the last inequality.

Similar to $\mathcal{E}_\mu(t)$, we define the energy functional for $F(X(t))$ as

$$\mathcal{E}(t) = t(F(X(t)) - F(x^*)) + \frac{\lambda}{2}\|B(X(t) - x^*)\|_2^2. \tag{B.13}$$

At time $t = 0$, there is an upper bound on $\mathcal{E}(0)$ as

$$\mathcal{E}(0) = \frac{\lambda}{2}\|B(x_0 - x^*)\|_2^2 \leq \frac{\lambda\kappa_1^2}{2}\|x_0 - x^*\|_2^2. \tag{B.14}$$

By applying the approximation scheme ($\mathcal{AS}$) argument (details as in Appendix A) as $\mu \to 0$, we have for a.e. $t \geq 0$ that $\mathcal{E}(t) \leq \mathcal{E}(0)$.

By non-negativity of $F(X) - F(x^*)$ in (B.13), we find

$$\frac{\lambda\kappa_d^2}{2}\|(X(t) - x^*)\|_2^2 \leq \mathcal{E}(0). \tag{B.15}$$

Combining with the upper bound of $\mathcal{E}(0)$ in (B.14), we derive for a.e. $t \geq 0$ that

$$\|X(t) - x^*\|_2 \leq \frac{\kappa_1}{\kappa_d}\|x_0 - x^*\|_2. \tag{B.16}$$

Using the nonnegativity of all terms in (B.13) and monotonicity of $\mathcal{E}(t)$ on a.e. $t \geq 0$, we have, for a.e. $t \geq 0$,

$$t(F(X(t)) - F(x^*)) \leq \mathcal{E}(t) \leq \mathcal{E}(0) \leq \frac{\lambda\kappa_1^2}{2}\|x_0 - x^*\|_2^2 \tag{B.17}$$

Choosing $\lambda = 1$, we have the following result, for a.e. $t \geq 0$,

$$F(X(t)) - F(x^*) \leq \mathcal{E}(t) \leq \mathcal{E}(0) \leq \frac{\kappa_1^2}{2t}\|x_0 - x^*\|_2^2 \tag{B.18}$$

By applying convexity of F_μ to (B.12), we have

$$\dot{\mathcal{E}}_\mu(t) \leq (1 - \lambda)(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) - t\|B\dot{X}_\mu(t)\|_2^2. \tag{B.19}$$

Notice that the two terms in (B.19) are all negative, we find

$$F_\mu(X_\mu(t)) - F_\mu(x_\mu^*) \leq \frac{-\dot{\mathcal{E}}_\mu(t)}{\lambda - 1} \qquad \text{and} \qquad t\|\dot{X}_\mu(t)\|_2^2 \leq -\frac{\dot{\mathcal{E}}_\mu(t)}{\kappa_d^2}. \tag{B.20}$$

By integrating over $(0, T)$, the inequalities above give for all $T > 0$ that

$$\int_0^T (F_\mu(X_\mu(t)) - F_\mu(x_\mu^*))dt \leq \frac{\mathcal{E}_\mu(0)}{\lambda - 1}, \quad \int_0^T t\|\dot{X}_\mu(t)\|_2^2 dt \leq \frac{\mathcal{E}_\mu(0)}{\kappa_d^2}. \tag{B.21}$$

By applying approximation scheme ($\mathcal{AS}$), taking limit $T \to \infty$, choosing $\lambda \to \infty$ and $\lambda = 1$

respectively, and plugging in (B.14), we have

$$\int_0^\infty (F(X_\mu(t)) - F(x^*))dt \leq \frac{\kappa_1^2}{2}\|x_0 - x^*\|_2^2, \quad \int_0^\infty t\|\dot{X}(t)\|_2^2 dt \leq \frac{\kappa_1^2}{2\kappa_d^2}\|x_0 - x^*\|_2^2. \tag{B.22}$$

□

C Proofs of the Theorems Related to G-ADMM and the Accelerated G-ADMM

C.1 Proof of Theorem 4

Proof of Theorem 4. Proof of Theorem 4 uses idea similar to the proof of Theorem 1 to analyze G-ADMM updates. By strong convexity of the optimization subproblems (3.1a) and (3.1b), we could verify that the sequence $\{x_k, z_k, u_k\}$ is unique. Together with (3.1c), we have from the first-order optimality conditions of (3.1a) and (3.1b) that

$$\partial f\left(x_{k+1}\right) + \rho A^T\left(Ax_{k+1} - z_k + u_k\right) \ni 0, \tag{C.1a}$$

$$\frac{1}{\rho}\partial g\left(z_{k+1}\right) - \left(\alpha Ax_{k+1} + (1-\alpha)z_k - z_{k+1} + u_k\right) \ni 0, \tag{C.1b}$$

$$u_{k+1} - \left(\alpha Ax_{k+1} + (1-\alpha)z_k - z_{k+1} + u_k\right) = 0. \tag{C.1c}$$

Adding up (C.1b) and (C.1c) eliminates the common term $(\alpha Ax_{k+1} + (1-\alpha)z_k - z_{k+1} + u_k)$ and reduces to a simple u-update:

$$u_{k+1} \in \frac{1}{\rho}\partial g(z_{k+1}). \tag{C.2}$$

Taking the continuous limit $\rho \to \infty$ gives $U(t) = 0$, and hence $\dot{U}(t) = 0$.

Bringing (C.2) into (C.1a) leads to:

$$0 \in \partial f(x_{k+1}) + A^\top \partial g(z_k) + \rho A^\top (Ax_{k+1} - z_k), \tag{C.3}$$

where again from (C.1c),

$$u_{k+1} - u_k = \alpha(Ax_{k+1} - z_k) - (z_{k+1} - z_k),$$

and hence

$$Ax_{k+1} - z_k = \frac{1}{\alpha}[(u_{k+1} - u_k) + (z_{k+1} - z_k)] \tag{C.4}$$

Plugging (C.4) into (C.3) gives

$$0 \in \partial f(x_{k+1}) + A^\top \partial g(z_k) + \rho A^\top \left(\frac{1}{\alpha}[(u_{k+1} - u_k) + (z_{k+1} - z_k)] \right). \tag{C.5}$$

Taking the limit $\rho \to \infty$, using the fact that $\dot{U}(t) = 0$, (C.5) reduces to

$$0 \in \partial f(X(t)) + A^T \partial g(Z(t)) + \frac{1}{\alpha} A^T(\dot{Z}(t)). \tag{C.6}$$

We directly take the $\rho \to \infty$ limit in (C.1c) and conclude

$$Z(t) = AX(t), \qquad \dot{Z}(t) = A\dot{X}(t).$$

Recalling (B.8) and combining the above with (C.6) concludes

$$0 \in \partial F(X(t)) + \left(\frac{1}{\alpha} A^\top A \right) \dot{X}(t),$$

Thus we complete the proof.

□

C.2 Proof of Theorem 5

Proof of Theorem 5. For each $\mu > 0$, consider the energy functional of Moreau-Yosida approximation defined as

$$\mathcal{E}_\mu(t) = \alpha t(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) + \frac{\lambda}{2}\|A(X_\mu(t) - x_\mu^*)\|_2^2, \tag{C.7}$$

where λ is an arbitrary constant chosen as $\lambda \geq 1$ and x_μ^* denotes the minimizer of F_μ. Because F_μ is a continuously differentiable function ,we could write the time derivative of $\mathcal{E}_\mu(t)$ as

$$\dot{\mathcal{E}}_\mu(t) = \alpha(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) + \alpha t\langle \nabla F_\mu(X_\mu(t)), \dot{X}_\mu(t)\rangle + \lambda\langle A^\top A(X_\mu(t) - x_\mu^*), \dot{X}_\mu(t).\rangle \tag{C.8}$$

By using the equality of $A^T A\dot{X}_\mu(t)$ and $-\alpha\nabla F_\mu(X_\mu(t))$, we have

$$\dot{\mathcal{E}}_\mu(t) = -t\|A\dot{X}_\mu(t)\|_2^2 + \alpha(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) - \lambda\alpha\langle (X_\mu(t) - x_\mu^*), \nabla F_\mu(X_\mu(t))\rangle \leq 0, \tag{C.9}$$

where we used the convexity of F_μ and nonnegativity of $(F_\mu(X_\mu(t)) - F(x_\mu^*))$, $\|A\dot{X}_\mu\|_2$ in the last inequality.

Similar to $\mathcal{E}_\mu(t)$, we define the energy functional for $F(X(t))$ as

$$\mathcal{E}(t) = \alpha t(F(X(t)) - F(x^*)) + \frac{\lambda}{2}\|A(X(t) - x^*)\|_2^2. \tag{C.10}$$

At time 0, there is an upper bound on $\mathcal{E}(0)$ as

$$\mathcal{E}(0) = \frac{\lambda}{2}\|A(X(0) - x^*)\|_2^2 \leq \frac{\lambda\sigma_1^2}{2}\|x_0 - x^*\|_2^2. \tag{C.11}$$

By applying the approximation scheme ($\mathcal{AS}$) argument (details as in Appendix A) as $\mu \to 0$ to equation (C.9), we have for a.e. $t \geq 0$, $\dot{\mathcal{E}}(t) \leq 0$ and that $\mathcal{E}(t) \leq \mathcal{E}(0)$.

In (C.10), by non-negativity of $F(X(t)) - F(x^*)$ and $\|X(t) - x^*\|_2^2$, we find

$$\frac{\lambda}{2}\|A(X(t) - x^*)\|_2^2 \leq \mathcal{E}(0). \tag{C.12}$$

Combining with the upper bound of $\mathcal{E}(0)$ in (C.11), and by taking $\lambda = 1$, we derive for a.e. $t \geq 0$ that

$$\|X(t) - x^*\|_2 \leq \frac{\sigma_1}{\sigma_d}\|x_0 - x^*\|_2. \tag{C.13}$$

Using the nonnegativity of all terms in (C.10) and monotonicity of $\mathcal{E}(t)$ on a.e. $t \geq 0$, we have

$$\alpha t(F(X(t)) - F(x^*)) \leq \mathcal{E}(t) \leq \mathcal{E}(0) \leq \frac{\lambda\sigma_1^2}{2}\|x_0 - x^*\|_2^2 \quad for \quad a.e. \quad t, \tag{C.14}$$

which is given by (C.11). Thus $(F(X(t)) - F(x^*)) \leq \frac{\sigma_1^2}{2\alpha t}\|x_0 - x^*\|_2^2$ by taking $\lambda = 1$.

From (C.9) and using the convexity of F_μ, we have

$$\dot{\mathcal{E}}_\mu(t) \leq \alpha(1 - \lambda)(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) - t\|A\dot{X}_\mu(t)\|_2^2. \tag{C.15}$$

Notice that the two terms in (C.15) are all negative, we find

$$F_\mu(X_\mu(t)) - F_\mu(x_\mu^*) \leq \frac{-\dot{\mathcal{E}}_\mu(t)}{\alpha(\lambda - 1)} \qquad \text{and} \qquad t\|A\dot{X}_\mu(t)\|_2^2 \leq -\dot{\mathcal{E}}_\mu(t). \tag{C.16}$$

By integrating over $(0, T)$, the inequalities above give

$$\int_0^T (F_\mu(X_\mu(t)) - F_\mu(x_\mu^*))dt \leq \frac{\mathcal{E}_\mu(0)}{\alpha(\lambda - 1)}, \quad \int_0^T t\|A\dot{X}_\mu(t)\|_2^2 dt \leq \mathcal{E}_\mu(0). \tag{C.17}$$

By applying approximation scheme ($\mathcal{AS}$) and plugging in (C.11), we have

$$\int_0^T (F_\mu(X_\mu(t)) - F_\mu(x_\mu^*))dt \leq \frac{\lambda\sigma_1^2}{2\alpha(\lambda - 1)}\|x_0 - x^*\|_2^2, \quad \int_0^T t\|A\dot{X}_\mu(t)\|_2^2 dt \leq \frac{\lambda\sigma_1^2}{2}\|x_0 - x^*\|_2^2. \tag{C.18}$$

Taking the limit when $\mu \to 0, T \to \infty$ and choosing $\lambda \to \infty$ and $\lambda = 1$ respectively, we get

$$\int_0^\infty (F(X(t)) - F(x^*))dt \leq \frac{\sigma_1^2}{2\alpha}\|x_0 - x^*\|_2^2, \quad \int_0^\infty t\|\dot{X}(t)\|_2^2 dt \leq \frac{\sigma_1^2}{2\sigma_d^2}\|x_0 - x^*\|_2^2. \tag{C.19}$$

This completes our proof.

□

C.3 Proof of Theorem 6

Proof of Theorem 6. To analyze Accelerated G-ADMM, we adopt idea similar to proof of Theorem 1. Using strong convexity of the optimization subproblems (3.3a) and (3.3b), we know that the sequence $\{x_k, z_k, u_k, \hat{u}_k, \hat{z}_k\}$ is unique. Together with (3.3c), we have from the first-order optimality conditions of (3.3a) and (3.3b) that

$$\partial f\left(x_{k+1}\right) + \rho A^T\left(Ax_{k+1} - \hat{z}_k + \hat{u}_k\right) \ni 0, \tag{C.20a}$$

$$\frac{1}{\rho}\partial g\left(z_{k+1}\right) - \left(\alpha A x_{k+1} + (1-\alpha)\hat{z}_k - z_{k+1} + \hat{u}_k\right) \ni 0, \tag{C.20b}$$

$$u_{k+1} - \left(\alpha A x_{k+1} + (1-\alpha)\hat{z}_k - z_{k+1} + \hat{u}_k\right) = 0. \tag{C.20c}$$

Adding up (C.20b) and (C.20c) eliminates the common term $-\left(\alpha A x_{k+1} + (1-\alpha)\hat{z}_k - z_{k+1} + \hat{u}_k\right)$ and reduces to a simple u-update:

$$u_{k+1} \in \frac{1}{\rho}\partial g(z_{k+1}). \tag{C.21}$$

Taking the continuous limit $\rho \to \infty$ gives $U(t) = 0$, and hence $\dot{U}(t) = 0$, $\ddot{U}(t) = 0$. The idea is similar to the proof of Theorem 1.

Bringing (C.21) and equation (3.3d) which is the definition of $\hat{u}$ into (C.20a) leads to:

$$\partial f(x_{k+1}) + A^T\partial g(z_k) + \rho A^T(Ax_{k+1} - \hat{z}_k) + \rho\gamma_{k+1}A^\top(u_{k+1} - u_k) \ni 0, \tag{C.22}$$

where again from (C.20c),

$$(Ax_{k+1} - \hat{z}_k) = \frac{1}{\alpha}[(u_{k+1} - \hat{u}_k) + (z_{k+1} - \hat{z}_k)]. \tag{C.23}$$

In addition, from equation (3.3d) and equation (3.3e), we find that $u_{k+1} - \hat{u}_k = u_{k+1} - (1 + \gamma_{k+1})u_k + \gamma_{k+1}u_{k-1}$ and $z_{k+1} - \hat{z}_k = z_{k+1} - (1 + \gamma_{k+1})z_k + \gamma_{k+1}z_{k-1}$. For $u_{k+1} - \hat{u}_k$, we add the term $u_k - u_k + u_{k-1} - u_{k-1}$ to the right hand side, the resulting equation is a combination of the

second order difference and first order difference of the sequence $\{u_k\}$:

$$u_{k+1} - \hat{u}_k = (u_{k+1} - 2u_k + u_{k-1}) + (1 - \gamma_{k+1})(u_k - u_{k-1}). \tag{C.24}$$

Similarily, the equation holds that:

$$z_{k+1} - \hat{z}_k = (z_{k+1} - 2z_k + z_{k-1}) + (1 - \gamma_{k+1})(z_k - z_{k-1}). \tag{C.25}$$

We note that $1 - \gamma_k = 1 - \frac{k}{k+r} = \frac{r}{\rho^{1/2}t+r}$. Taking the limit $\rho \to \infty$, under infinitesimal step sizes, using relationships (C.23), (C.24), (C.25) and the fact that $\dot{U}(t) = 0, \ddot{U}(t) = 0$, equation (C.22) becomes:

$$\partial f(X(t)) + A^T \partial g(Z(t)) + \frac{1}{\alpha} A^T (\frac{r}{t}\dot{Z}(t) + \ddot{Z}(t)) \ni 0. \tag{C.26}$$

We directly take the $\rho \to \infty$ limit in (C.20c) and conclude

$$Z(t) = AX(t), \qquad \dot{Z}(t) = A\dot{X}(t), \qquad \ddot{Z}(t) = A\ddot{X}(t).$$

Recalling (B.8) and combining the above with (C.26) concludes

$$0 \in \partial F(X(t)) + \left(\frac{1}{\alpha} A^\top A\right) (\ddot{X}(t) + \frac{r}{t}\dot{X}(t)).$$

□

C.4 Proof of Theorem 7

Proof of Theorem 7. Recall that x_μ^* is the minimizer of F_μ. For each $\mu > 0$, consider the energy functional of Moreau-Yosida approximation defined as

$$\mathcal{E}_\mu(t) = t^2(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) + \frac{1}{2\alpha} \left\| A\left(\lambda(X_\mu(t) - x_\mu^*) + t\dot{X}_\mu(t)\right)\right\|_2^2 + \frac{\lambda(r-\lambda-1)}{2\alpha} \|A(X_\mu(t) - x_\mu^*)\|_2^2 \tag{C.27}$$

where λ is a constant chosen within $2 \le \lambda \le r - 1$. Because F_μ is a continuously differentiable function, we could write the time derivative of $\mathcal{E}_\mu(t)$ as

$$\dot{\mathcal{E}}_\mu = 2t(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) + t^2 \nabla F_\mu(X_\mu(t))^\top \dot{X}_\mu + \left(\lambda(X_\mu - x_\mu^*) + t\dot{X}_\mu\right)^\top (\frac{1}{\alpha} A^\top A) \left((\lambda+1)\dot{X}_\mu + t\ddot{X}_\mu\right)$$
$$+ \lambda(r - \lambda - 1)(X_\mu - x_\mu^*)^\top (\frac{1}{\alpha} A^\top A)\dot{X}_\mu$$

By using the equality of $tA^\top A\ddot{X}_\mu$ and $-rA^\top A\dot{X}_\mu - \alpha t\nabla F_\mu(X_\mu(t))$, we have

$$\begin{aligned}\dot{\mathcal{E}}_\mu = &-\lambda t\left(F_\mu(x_\mu^*) - F_\mu(X_\mu(t)) - (x_\mu^* - X_\mu)^\top \nabla F_\mu(X_\mu(t))\right)\\ &- (\lambda-2)t(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) - \frac{(r-1-\lambda)t}{\alpha}\|A\dot{X}_\mu\|_2^2 \le 0\end{aligned} \tag{C.28}$$

where we used the convexity of F_μ and nonnegativity of $F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)$, $\|A\dot{X}_\mu\|_2$ in the last inequality.

Similar to $\mathcal{E}_\mu(t)$, we define the energy functional for $F(X(t))$ as

$$\mathcal{E}(t) = t^2(F(X(t)) - F(x^*)) + \frac{1}{2\alpha}\left\|A\left(\lambda(X(t)-x^*) + t\dot{X}(t)\right)\right\|_2^2 + \frac{\lambda(r-\lambda-1)}{2\alpha}\|A(X(t)-x^*)\|_2^2$$

At time t_0, there is an upper bound on $\mathcal{E}(t_0)$ as

$$\mathcal{E}(t_0) = t_0^2(F(x_0) - F(x^*)) + \frac{\lambda(r-1)}{2\alpha}\|A(x_0-x^*)\|_2^2 \le \frac{2\alpha + \lambda(r-1)\sigma_1^2}{2\alpha}\Delta_0^2 \tag{C.29}$$

By non-negativity of $F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)$, $\|X_\mu - x_\mu^*\|_2^2$ and $\|\dot{X}_\mu\|_2^2$, we find for all $r \ge 3$ and $t \ge t_0$ that

$$\frac{d}{dt}(t\|X_\mu - x_\mu^*\|_2^2) = \|X_\mu - x_\mu^*\|_2^2 + 2t(X_\mu - x_\mu^*)^\top \dot{X}_\mu \le \frac{1}{2}\|2(X_\mu - x_\mu^*) + t\dot{X}_\mu\|_2^2 \le \frac{\alpha\mathcal{E}_\mu}{\sigma_d^2} \le \frac{\alpha\mathcal{E}_\mu(t_0)}{\sigma_d^2}$$

By integrating over (t_0, t), this gives us

$$t\|X_\mu - x_\mu^*\|_2^2 - t_0\|x_0 - x_\mu^*\|_2^2 \le \frac{\alpha(t-t_0)}{\sigma_d^2}\mathcal{E}_\mu(t_0)$$

By applying the approximation scheme ($\mathcal{AS}$) argument (details as in Appendix A) as $\mu \to 0$, we have for a.e. $t \ge t_0$ that

$$\|X - x^*\|_2^2 \le \frac{\alpha\mathcal{E}(t_0)}{\sigma_d^2} + \|x_0 - x^*\|_2^2$$

Combining with the upper bound of $\mathcal{E}(t_0)$ in (C.29), we derive for a.e. $t \ge t_0$ that

$$\|X(t) - x^*\|_2 \le C_1\Delta_0 \tag{C.30}$$

with factor $C_1 = \sqrt{\frac{\alpha + (r-1)\sigma_1^2 + \sigma_d^2}{\sigma_d^2}}$. Here we choose $\lambda = 2$ to minimize C_1.

From (C.28), we know that $\mathcal{E}_\mu(t)$ is nonincreasing for $t \ge t_0$, for all $\mu > 0$. By applying ($\mathcal{AS}$) we find that $\mathcal{E}(t)$ is nonincreasing for a.e. $t \ge t_0$. Using the nonnegativity of all three terms in (C.27) and monotonicity of $\mathcal{E}(t)$ on a.e. $t \ge t_0$, we have for a.e. $t \ge t_0$ that

$$F(X(t)) - F(x^*) \le \frac{1}{t^2}\mathcal{E}(t) \le \frac{1}{t^2}\mathcal{E}(t_0) \le \frac{C_2}{t^2}\Delta_0^2$$

where factor $C_2 = 1 + (r-1)\sigma_1^2/\alpha$ is given by (C.29) with $\lambda = 2$, and

$$\|\lambda(X(t) - x^*) + t\dot{X}\|_2^2 \leq \frac{2\alpha}{\sigma_d^2}\mathcal{E}(t) \leq \frac{2\alpha}{\sigma_d^2}\mathcal{E}(t_0) \leq \frac{2\alpha + \lambda(r-1)\sigma_1^2}{\sigma_d^2}\Delta_0^2$$

Therefore, by triangle inequality and (C.30),

$$\|\dot{X}(t)\|_2 \leq \frac{1}{t}\|\lambda(X(t) - x^*) + t\dot{X}(t)\|_2 + \frac{1}{t}\lambda\|X(t) - x^*\|_2 \leq \frac{C_3}{t}\Delta_0$$

with factor $C_3 = \sqrt{\frac{2\alpha + 2(r-1)\sigma_1^2}{\sigma_d^2}} + 2C_1$. Here we choose $\lambda = 2$ to minimize C_3.

From (C.28), we have

$$\dot{\mathcal{E}}_\mu \leq -(\lambda - 2)t(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) - \frac{(r - 1 - \lambda)t}{\alpha}\|A\dot{X}_\mu\|_2^2$$

when $r = 3$, we could only choose $\lambda = 2 = r - 1$ and the right hand side of the inequality above is always zero. However, if we further assume $r > 3$, then we could choose $\lambda = r - 1$ and $\lambda = 2$ respectively, such that

$$t(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) \leq -\frac{1}{r-3}\dot{\mathcal{E}}_\mu \quad \text{and} \quad t\|\dot{X}_\mu\|_2^2 \leq -\frac{\alpha}{(r-3)\sigma_d^2}\dot{\mathcal{E}}_\mu$$

By integrating over (t_0, ∞), the inequalities above give

$$\int_{t_0}^{\infty} t(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*))dt \leq \frac{\mathcal{E}_\mu(t_0)}{r-3} \quad \text{and} \quad \int_{t_0}^{\infty} t\|\dot{X}_\mu(t)\|_2^2 dt \leq \frac{\alpha\mathcal{E}_\mu(t_0)}{(r-3)\sigma_d^2}$$

By applying ($\mathcal{AS}$) and plugging in (C.29), we have

$$\int_{t_0}^{\infty} t(F(X(t)) - F(x^*))dt \leq C_4\Delta_0^2 \quad \text{and} \quad \int_{t_0}^{\infty} t\|\dot{X}(t)\|_2^2 dt \leq C_5\Delta_0^2$$

with factors $C_4 = \frac{2\alpha + (r-1)^2\sigma_1^2}{2(r-3)\alpha}$ and $C_5 = \frac{\alpha + (r-1)\sigma_1^2}{(r-3)\sigma_d^2}$. □

C.5 Proof of Theorem 8

Proof of Theorem 8. The energy functional we used in Theorem 7 is no longer applicable, because we can not find λ satisfying $\lambda - 2 \geq 0$ and $r - 1 - \lambda \geq 0$ simultaneously when $0 < r < 3$. Here we consider a new energy functional for the Moreau-Yosida approximation

$$\mathcal{E}_\mu(t) = t^2(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) + \frac{1}{2\alpha}\left\|\frac{2r}{3}A(X_\mu(t) - x_\mu^*) + tA\dot{X}_\mu(t)\right\|_2^2 + \frac{r(3-r)}{9\alpha}\|A(X_\mu(t) - x_\mu^*)\|_2^2 \tag{C.31}$$

By taking its time derivative, we have

$$\begin{aligned}\dot{\mathcal{E}}_\mu =&2t(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) + t^2\nabla F_\mu(X_\mu(t))^\top \dot{X}_\mu \\ &+ \left(\frac{2r}{3}(X_\mu - x_\mu^*) + t\dot{X}_\mu\right)^\top (\frac{1}{\alpha}A^\top A)\left((\frac{2r}{3}+1)\dot{X}_\mu + t\ddot{X}_\mu\right) + \frac{2r(3-r)}{9}(X_\mu - x_\mu^*)^\top(\frac{1}{\alpha}A^\top A)\dot{X}_\mu\end{aligned}$$

By using the equality of $tA^\top A\ddot{X}_\mu$ and $-rA^\top A\dot{X}_\mu - \alpha t\nabla F_\mu(X_\mu)$ and applying the convexity of F_μ, we have

$$\dot{\mathcal{E}}_\mu \le \frac{2(3-r)}{3}t(F_\mu(X_\mu(t)) - F_\mu(x_\mu^*)) + \frac{4r(3-r)}{9\alpha}(X_\mu - x_\mu^*)^\top A^\top A\dot{X}_\mu + \frac{3-r}{3\alpha}t\|A\dot{X}_\mu\|_2^2$$

Although this energy functional does not have nonnegative derivative, there is a special relationship between it and its derivative. We notice that

$$\dot{\mathcal{E}}_\mu - \frac{2(3-r)}{3t}\mathcal{E}_\mu \le -\frac{2r(3-r)(3+r)}{27\alpha t}\|A(X_\mu - x_\mu^*)\|_2^2 \le 0$$

This implies that, for $\mathcal{H}_\mu(t) := t^{-\frac{2(3-r)}{3}}\mathcal{E}_\mu(t)$, for all $t \ge t_0$,

$$\dot{\mathcal{H}}_\mu = t^{-\frac{2(3-r)}{3}} \cdot (\dot{\mathcal{E}}_\mu - \frac{2(3-r)}{3t}\mathcal{E}_\mu) \le 0$$

Therefore, $\mathcal{H}_\mu(t)$ is nonincreasing over $t \ge t_0$, for all $\mu > 0$. By making similar definition as $\mathcal{H}(t) := t^{-\frac{2(3-r)}{3}}\mathcal{E}(t)$ and applying the approximation scheme, we have that $\mathcal{H}(t)$ is nonincreasing for a.e. $t \ge t_0$. At time t_0,

$$\mathcal{H}(t_0) \le t_0^{-\frac{2(3-r)}{3}} \cdot \left(1 + \frac{r(3+r)}{9\alpha}\sigma_1^2\right)\Delta_0^2$$

By the nonnegativity of all terms in (C.31) and the monotonicity of $\mathcal{H}(t)$, we have for a.e. $t \ge t_0$ that

$$F(X(t)) - F(x^*) \le \frac{1}{t^{\frac{2r}{3}}}\mathcal{H}(t) \le \frac{1}{t^{\frac{2r}{3}}}\mathcal{H}(t_0) \le \frac{C_6 t_0^{-\frac{2(3-r)}{3}}\Delta_0^2}{t^{\frac{2r}{3}}}$$

with factor $C_6 = 1 + \frac{r(3+r)\sigma_1^2}{9\alpha}$.

Similarly, we have for a.e. $t \ge t_0$ that

$$\left\|\frac{2r}{3}(X(t) - x^*) + t\dot{X}\right\|_2^2 \le \frac{2\alpha}{\sigma_d^2}t^{\frac{2(3-r)}{3}}\mathcal{H}(t) \le \frac{2\alpha}{\sigma_d^2}t^{\frac{2(3-r)}{3}}\mathcal{H}(t_0) \le \frac{2\alpha C_6 t_0^{-\frac{2(3-r)}{3}}}{\sigma_d^2}t^{\frac{2(3-r)}{3}}\Delta_0^2$$

If we also assume the trajectory $\{X(t)\}_{t\ge t_0}$ is bounded, then by adopting the same interpretation as in Theorem 7, there exists some positive factor C_0 such that, for a.e. $t \ge t_0$, $\|X(t) - x^*\|_2 \le C_0\Delta_0$.

Then triangle inequality gives us, for a.e. $t \geq t_0$, that

$$\|\dot{X}\|_2 \leq \frac{1}{t}\left\|\frac{2r}{3}(X(t)-x^*)+t\dot{X}\right\|_2 + \frac{2r}{3t}\|X(t)-x^*\|_2 \leq \frac{C_7 t_0^{-\frac{3-r}{3}}\Delta_0}{t^{\frac{r}{3}}}$$

with factor $C_7 = \sqrt{\frac{2\alpha C_6}{\sigma_d^2}} + \frac{2r}{3}C_0$. □