ADMM and Accelerated ADMM as Continuous Dynamical Systems

Guilherme França [1] **Daniel P. Robinson** [1] **René Vidal** [1]

Abstract

Recently, there has been an increasing interest in using tools from dynamical systems to analyze the behavior of simple optimization algorithms such as gradient descent and accelerated variants. This paper strengthens such connections by deriving the differential equations that model the continuous limit of the sequence of iterates generated by the alternating direction method of multipliers, as well as an accelerated variant. We employ the direct method of Lyapunov to analyze the stability of critical points of the dynamical systems and to obtain associated convergence rates.

1. Introduction

The theory of dynamical systems has been extensively developed since its origins by Poincaré in the late 19th century. For example, the work of Lyapunov on stability is commonly used in physics, control systems, and other branches of applied mathematics. However, the connection between dynamical systems and optimization algorithms has only recently been studied. The basic idea is that tools from dynamical systems can be used to analyze the stability and convergence rates of the continuous limit of the sequence of iterates generated by optimization algorithms. Prior work has established these connections for simple optimization algorithms such as gradient descent (GD) and accelerated gradient descent (A-GD). This paper improves upon prior work by deriving the dynamical systems associated with the continuous limit of two commonly used optimization methods: the alternating direction method of multipliers (ADMM) and an accelerated version of ADMM (A-ADMM). Moreover, this paper analyzes the stability properties of the resulting dynamical systems and derives their convergence rates, which for ADMM matches the known rate of its discrete counterpart, and for A-ADMM provides a new result.

[1]Mathematical Institute for Data Science, Johns Hopkins University, Baltimore MD 21218, USA. Correspondence to: Guilherme França <guifranca@jhu.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

1.1. Related work

Perhaps the simplest connection between an optimization algorithm and a continuous dynamical system is exhibited by the GD method. The GD algorithm aims to minimize a function $f : \mathbb{R}^n \to \mathbb{R}$ via the update

$$x_{k+1} = x_k - \eta \nabla f(x_k), \tag{1}$$

where x_k denotes the kth solution estimate and $\eta > 0$ is the step size. It is immediate that a continuous limit of the iterate update (1) leads to the gradient flow

$$\dot{X} = -\nabla f(X), \tag{2}$$

where $X = X(t)$ is the continuous limit of x_k and $\dot{X} \equiv \frac{dX}{dt}$ denotes its time derivative. It is known that GD has a convergence rate of $O(1/k)$ for convex functions (Nesterov, 2004). Interestingly, the differential equation (2) also has a convergence rate of $O(1/t)$, which is consistent with GD.

Nesterov (1983) proposed an accelerated GD algorithm, henceforth referred to as A-GD, by adding momentum to the x variables. The update for A-GD may be written as

$$x_{k+1} = \hat{x}_k - \eta \nabla f(\hat{x}_k), \tag{3a}$$

$$\hat{x}_{k+1} = x_k + \frac{k}{k+r}(x_{k+1} - x_k), \tag{3b}$$

where $r \geq 3$ ($r = 3$ is the standard choice) and $\hat{x}_k$ denotes the kth accelerated vector. It is known that A-GD has a convergence rate of $O(1/k^2)$ for convex functions, which is known to be optimal in the sense of worst-case complexity (Nesterov, 2004). Recently, the continuous limit of A-GD was computed by Su et al. (2016) to be

$$\ddot{X} + \frac{r}{t}\dot{X} + \nabla f(X) = 0, \tag{4}$$

where $\ddot{X} \equiv \frac{d^2X}{dt^2}$ is the acceleration. By using a Lyapunov function, it was shown that the convergence rate associated with (4) is $O(1/t^2)$ when f is convex (Su et al., 2016), which matches the known rate of $O(1/k^2)$ for A-GD.

To better understand the acceleration mechanism, a variational approach was proposed (Wibisono et al., 2016) for which the continuous limit of a class of accelerated methods was obtained using the Bregman Lagrangian; the class of methods includes A-GD, its non-Euclidean extension, and

accelerated higher-order gradient methods. Also, a differential equation modeling the continuous limit of accelerated mirror descent was obtained (Krichene et al., 2015).

While Su et al. (2016) focused on the discrete to continuous limit, Wibisono et al. (2016); Krichene et al. (2015) stress the converse, by which one starts with a second-order differential equation and then constructs a discretization with a matching convergence rate. This approach can lead to new accelerated algorithms. Indeed, Krichene et al. (2015) introduced a family of first-order accelerated methods and established their convergence rates by using a discrete Lyapunov function, which is analogous to its continuous counterpart. In related work, Wilson et al. (2016) proposed continuous and discrete time Lyapunov frameworks for A-GD based methods that built additional connections between rates of convergence and the choice of discretization. In particular, they showed that a naive discretization can produce iterates that do not match the convergence rate of the differential equation and proposed rate-matching algorithms (Wibisono et al., 2016; Krichene et al., 2015). However, such rate-matching algorithms introduce extra conditions such as an intermediate sequence or a new function that must obey certain constraints, which is in stark contrast to A-GD.

An important algorithm commonly used in machine learning and statistics is ADMM (Gabay & Mercier, 1976; Glowinski & Marroco, 1975; Boyd et al., 2011; Eckstein & Yao, 2015), which is often more easily distributed when compared to its competitors and hence appealing for large scale applications. There are some interesting relations between ADMM and *discrete* dynamical systems. For instance, the formalism of integral quadratic constraints (Lessard et al., 2016) was applied to ADMM (Nishihara et al., 2015) under the assumption of strong convexity. Based on this, França & Bento (2016) establish an explicit upper bound on the convergence rate of ADMM in terms of the algorithm parameters and the condition number of the problem by analytically solving the semi-definite program introduced by Nishihara et al. (2015). Moreover, for a class of convex quadratic consensus problems defined over a graph, ADMM can be viewed as a lifted Markov chain (França & Bento, 2017a;b) that exhibits a significant speedup in the mixing time compared to GD, which corresponds to the base Markov chain. For convex functions, ADMM has an $O(1/k)$ convergence rate (Eckstein & Yao, 2015). Recently, using the ideas of Nesterov (1983), Goldstein et al. (2014) proposed an accelerated version of ADMM (henceforth called A-ADMM) and established a convergence rate of $O(1/k^2)$ when both f and g are *strongly* convex functions, and moreover g is a quadratic function.

Although there is extensive literature on ADMM and its variants, connections between their continuous limit and differential equations is unknown. This paper is a first step in establishing such connections in the context of the problem[1]

$$\min_{x,z} \{V(x,z) = f(x) + g(z)\} \text{ subject to } z = Ax, \quad (5)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$ are continuously differentiable convex functions, $x \in \mathbb{R}^n$, $z \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$ and $m \geq n$. The problem formulation (5) covers many interesting applications in machine learning and statistics. With that said, we also recognize that many important problems do not fall within the framework (5) due to the assumption of differentiability, especially of g which is usually a regularization term. Such an assumption is a theoretical necessity to allow connections to differential equations[2].

1.2. Paper contributions

Our first contribution is to show in Theorem 2 that the dynamical system that is the continuous limit of ADMM when applied to (5) is given by the ADMM flow

$$\left(A^T A\right)\dot{X} + \nabla V(X) = 0. \quad (6)$$

Note that when $A = I$ we obtain the dynamical system (1) (i.e., the continuous limit of GD), which can be thought of as an unconstrained formulation of (5). Our second contribution is to show in Theorem 3 that the dynamical system that is the continuous limit of A-ADMM is the A-ADMM flow

$$\left(A^T A\right)\left(\ddot{X} + \frac{r}{t}\dot{X}\right) + \nabla V(X) = 0. \quad (7)$$

Here, the dynamical system (4) that is the continuous limit of A-GD is a particular case obtained when $A = I$. We then employ the direct method of Lyapunov to study the stability properties of both dynamical systems (6) and (7). We show that under reasonable assumptions on f and g, these dynamical systems are asymptotically stable. Also, we prove that (6) has a convergence rate of $O(1/t)$, whereas (7) has a convergence rate of $O(1/t^2)$.

1.3. Notation

We use $\|\cdot\|$ to denote the Euclidean two norm, and $\langle u, v\rangle = u^T v$ to denote the inner product of $u, v \in \mathbb{R}^n$. For our analysis, it is convenient to define the function

$$V(x) = f(x) + g(Ax), \quad (8)$$

which is closely related to the function $V(x, z)$ that defines the objective function in (5). In particular, for all (x, z) satisfying $z = Ax$, the relationship $V(x, z) = V(x)$ holds. Throughout the paper, we make the following assumption.

Assumption 1. *The functions f and g in (5) are continuously differentiable and convex, and A has full column rank.*

[1] The standard problem $\min_{x,z} f(x) + g(z)$ subject to $Ax + Bz = c$ can be recovered by redefining A when B is invertible.

[2] Perhaps the differentiability assumption may be relaxed by using subdifferential calculus and differential inclusions, but this is beyond the scope of this work.

2. Continuous Dynamical Systems

In this section, we show that the continuous limits of the ADMM and A-ADMM algorithms are first- and second-order differential equations, respectively.

2.1. ADMM

The scaled form of ADMM is given by (Boyd et al., 2011)

$$x_{k+1} = \underset{x\in\mathbb{R}^n}{\arg\min}\ f(x) + \tfrac{\rho}{2}\|Ax - z_k + u_k\|^2, \tag{9a}$$

$$z_{k+1} = \underset{z\in\mathbb{R}^m}{\arg\min}\ g(z) + \tfrac{\rho}{2}\|Ax_{k+1} - z + u_k\|^2, \tag{9b}$$

$$u_{k+1} = u_k + Ax_{k+1} - z_{k+1}, \tag{9c}$$

where $\rho > 0$ is a penalty parameter and $u_k \in \mathbb{R}^m$ is the kth Lagrange multiplier estimate for the constraint $z = Ax$. Our next result shows how a continuous limit of the ADMM updates leads to a particular first-order differential equation.

Theorem 2. *Consider the optimization problem (5) and the associated function $V(\cdot)$ in (8). The continuous limit associated with the ADMM updates in (9), with time scale $t = k/\rho$, corresponds to the initial value problem*

$$\dot{X} + \left(A^TA\right)^{-1}\nabla V(X) = 0 \quad \textit{with } X(0) = x_0. \tag{10}$$

Proof. Since f and g are convex and A has full column rank (see Assumption 1), the optimization problems in (9a) and (9b) are strongly convex so that $(x_{k+1}, z_{k+1}, u_{k+1})$ is unique. It follows from the optimality conditions for problems (9a) and (9b) that $(x_{k+1}, z_{k+1}, u_{k+1})$ satisfies

$$\nabla f\left(x_{k+1}\right) + \rho A^T\left(Ax_{k+1} - z_k + u_k\right) = 0, \tag{11a}$$

$$\nabla g\left(z_{k+1}\right) - \rho\left(Ax_{k+1} - z_{k+1} + u_k\right) = 0, \tag{11b}$$

$$u_{k+1} - u_k - Ax_{k+1} + z_{k+1} = 0, \tag{11c}$$

which can be combined to obtain

$$\nabla f\left(x_{k+1}\right) + A^T\nabla g\left(z_{k+1}\right) + \rho A^T\left(z_{k+1} - z_k\right) = 0. \tag{12}$$

Let $t = \delta k$ and $x_k = X(t)$, with a similar notation for z_k and u_k. Using the Mean Value Theorem on the ith component of z_{k+1} we have that $z_{k+1,i} = Z_i(t+\delta) = Z_i(t) + \delta\dot{Z}_i(t + \lambda_i\delta)$ for some $\lambda_i \in [0,1]$. Therefore,

$$\lim_{\delta\to 0}\frac{z_{k+1,i} - z_{k,i}}{\delta} = \lim_{\delta\to 0}\dot{Z}_i(t+\lambda_i\delta) = \dot{Z}_i(t). \tag{13}$$

Since this holds for every component $i = 1,\dots,m$ we see that, in the limit $\delta \to 0$, the third term in (12) is exactly equal to the vector $\dot{Z}(t)$, provided we choose $\rho = 1/\delta$. For the first two terms of (12), note that

$$\begin{aligned}\nabla f(x_{k+1}) + A^T\nabla g(z_{k+1}) &\\ = \nabla f(X(t+\delta)) + A^T\nabla g(Z(t+\delta)) &\\ \to \nabla f(X(t)) + A^T\nabla g(Z(t)) &\end{aligned} \tag{14}$$

as $\delta \to 0$. Thus, taking the limit $\delta \to 0$ in (12) and substituting (13) and (14) yields

$$\nabla f(X(t)) + A^T\nabla g(Z(t)) + A^T\dot{Z}(t) = 0. \tag{15}$$

Let us now consider the ith component of (11c). By the Mean Value Theorem there exists $\lambda_i \in [0,1]$ such that

$$\begin{aligned}0 &= U_i(t+\delta) - U_i(t) - (AX)_i(t+\delta) + Z_i(t+\delta)\\ &= \delta\dot{U}_i(t+\lambda_i\delta) - (AX)_i(t+\delta) + Z_i(t+\delta)\\ &\to Z_i(t) - (AX)_i(t)\end{aligned} \tag{16}$$

as $\delta \to 0$. Since this holds for every $i = 1,\dots,m$ we conclude that $Z(t) = AX(t)$ and $\dot{Z}(t) = A\dot{X}(t)$. Moreover, recalling the definition (8), note that

$$\begin{aligned}\nabla f(X) + A^T\nabla g(Z) &= \nabla f(X) + A^T\nabla g(AX)\\ &= \nabla V(X).\end{aligned} \tag{17}$$

Therefore, (15) becomes

$$\nabla V(X(t)) + A^TA\dot{X}(t) = 0, \tag{18}$$

which is equivalent to (10) since A has full column rank.

Finally, since (10) is a first-order differential equation, the dynamics is specified by the initial condition $X(0) = x_0$, where x_0 is an initial solution estimate to (5). □

We remark that the continuous limit of ADMM (see (10)) and GD (see (2)) are similar—first-order gradient systems—with the only difference being the additional $(A^TA)^{-1}$ term. Thus, in the special case $A = I$, i.e., the unconstrained case, the differential equation (10) reduces to (2).

2.2. A-ADMM

We now consider an accelerated version of ADMM that was originally proposed by Goldstein et al. (2014), which follows the same idea introduced by Nesterov (1983) to accelerate GD. The scaled A-ADMM method for solving problem (5) can be written as follows:

$$x_{k+1} = \underset{x\in\mathbb{R}^n}{\arg\min}\ f(x) + \tfrac{\rho}{2}\|Ax - \hat{z}_k + \hat{u}_k\|^2, \tag{19a}$$

$$z_{k+1} = \underset{z\in\mathbb{R}^m}{\arg\min}\ g(z) + \tfrac{\rho}{2}\|Ax_{k+1} - z + \hat{u}_k\|^2, \tag{19b}$$

$$u_{k+1} = \hat{u}_k + Ax_{k+1} - z_{k+1}, \tag{19c}$$

$$\hat{u}_{k+1} = u_{k+1} + \gamma_{k+1}\left(u_{k+1} - u_k\right), \tag{19d}$$

$$\hat{z}_{k+1} = z_{k+1} + \gamma_{k+1}\left(z_{k+1} - z_k\right), \tag{19e}$$

where $\hat{u}$ and $\hat{z}$ are the "accelerated variables" and

$$\gamma_{k+1} = k/(k+r) \tag{20}$$

with $r \geq 3$. We remark that the particular choice $r = 3$ produces the same asymptotic behavior as the parameter choice in Goldstein et al. (2014); Nesterov (1983). Our next result shows how a continuous limit of the A-ADMM updates is a second-order differential equation.

Theorem 3. *Consider the optimization problem (5) and the associated function $V(\cdot)$ in (8). The continuous limit associated with the A-ADMM updates in (19), with time scale $t = k/\sqrt{\rho}$, corresponds to the initial value problem*

$$\ddot{X} + \frac{r}{t}\dot{X} + \left(A^T A\right)^{-1}\nabla V(X) = 0 \tag{21}$$

with $X(0) = x_0$ and $\dot{X}(0) = 0$.

Proof. According to Assumption 1, the functions f and g are convex and A has full column rank, therefore the optimization problems (19a) and (19b) are strongly convex, making $(x_{k+1}, z_{k+1}, u_{k+1}, \hat{u}_{k+1}, \hat{z}_{k+1})$ unique. It thus follows from the optimality conditions that

$$\nabla f(x_{k+1}) + A^T\nabla g(z_{k+1}) + \rho A^T(z_{k+1} - \hat{z}_k) = 0, \tag{22a}$$
$$u_{k+1} - \hat{u}_k - Ax_{k+1} + z_{k+1} = 0, \tag{22b}$$
$$\hat{u}_{k+1} - u_{k+1} - \gamma_{k+1}(u_{k+1} - u_k) = 0, \tag{22c}$$
$$\hat{z}_{k+1} - z_{k+1} - \gamma_{k+1}(z_{k+1} - z_k) = 0. \tag{22d}$$

Let $t = \delta k$, $x_k = X(t)$ and similarly for z_k, u_k, $\hat{z}_k$ and $\hat{u}_k$. Consider Taylor's Theorem for the ith component of $z_{k\pm1}$:

$$\begin{aligned} z_{k\pm1,i} &= Z_i(t \pm \delta) \\ &= Z_i(t) \pm \delta\dot{Z}_i(t) + \tfrac{1}{2}\delta^2\ddot{Z}_i(t \pm \lambda_i^{\pm}\delta) \end{aligned} \tag{23}$$

for some $\lambda_i^{\pm} \in [0, 1]$. Hence, from (22d) we have

$$\begin{aligned} z_{k+1,i} - \hat{z}_{k,i} &= z_{k+1,i} - z_{k,i} - \gamma_k(z_{k,i} - z_{k-1,i}) \\ &= (1-\gamma_k)\delta\dot{Z}_i(t) + \tfrac{1}{2}\delta^2\ddot{Z}_i(t + \lambda_i^+\delta) \\ &\quad + \tfrac{1}{2}\gamma_k\delta^2\ddot{Z}_i(t - \lambda_i^-\delta). \end{aligned} \tag{24}$$

From the definition (20), and $t = \delta k$, we have

$$\begin{aligned} \gamma_k &= \frac{k-1}{k+r-1} = 1 - \frac{r}{k+r-1} = 1 - \frac{\delta r}{t + \delta(r-1)} \\ &= 1 - \frac{\delta r}{t} + O(\delta^2). \end{aligned} \tag{25}$$

Replacing this into (24) we obtain

$$\begin{aligned} \frac{z_{k+1,i} - \hat{z}_{k,i}}{\delta^2} &= \frac{r}{t}\dot{Z}_i(t) + \tfrac{1}{2}\ddot{Z}_i(t + \lambda_i^+\delta) \\ &\quad + \tfrac{1}{2}\ddot{Z}_i(t - \lambda_i^-\delta) + O(\delta) \\ &\to \frac{r}{t}\dot{Z}_i(t) + \ddot{Z}_i(t) \end{aligned} \tag{26}$$

as $\delta \to 0$. Hence, if we choose $\rho = 1/\delta^2$, then the limit of the third term in (22a) is equal to $\frac{r}{t}A\dot{Z}(t) + A\ddot{Z}(t)$. Recalling (14), the limit of (22a) as $\delta \to 0$ is thus given by

$$\nabla f(X) + A^T\nabla g(Z) + A^T\left(\frac{r}{t}\dot{Z} + \ddot{Z}\right) = 0. \tag{27}$$

Next, using (25) into the ith component of (22c) we obtain

$$\begin{aligned} 0 &= \hat{u}_{k,i} - u_{k,i} - \gamma_k(u_{k,i} - u_{k-1,i}) \\ &= \hat{U}_i(t) - U_i(t) - (1 - O(\delta))\,\delta\,\dot{U}_i(t - \lambda_i^-\delta) \\ &\to \hat{U}_i(t) - U_i(t) \end{aligned} \tag{28}$$

as $\delta \to 0$. Since this holds for every component $i = 1, \ldots, n$ it follows that $\hat{U}(t) = U(t)$. Substituting this into (22b) implies that $Z(t) = AX(t)$, which in turn implies $\dot{Z}(t) = A\dot{X}(t)$ and $\ddot{Z}(t) = A\ddot{X}(t)$. Using these, and also (17), we obtain from (22a) the differential equation

$$\nabla V(X) + A^T A\left(\ddot{X} + \frac{r}{t}\dot{X}\right) = 0. \tag{29}$$

Since A has full column rank, so that $A^T A$ is invertible, we see that (29) is equivalent to (21).

It remains to consider the initial conditions. The first condition is $X(0) = x_0$, where x_0 is an initial estimate of a solution to (5). Next, using the Mean Value Theorem, we have $\dot{X}_i(t) = \dot{X}_i(0) + t\ddot{X}_i(\xi_i)$ for some $\xi_i \in [0, t]$ and $i = 1, \ldots, n$. Combining this with (21) yields

$$\dot{X}_i(t) = \dot{X}_i(0) - r\dot{X}_i(\xi_i) - t\left[\left(A^T A\right)^{-1}\nabla V(X(\xi_i))\right]_i \tag{30}$$

for all $i = 1, \ldots, n$. Letting $t \to 0^+$, which also forces $\xi_i \to 0^+$, we have that $\dot{X}_i(0) = (1 - r)\dot{X}_i(0)$ for each $i = 1, \ldots, n$. Since $r \neq 1$ by the choice of γ_k in (20), it follows that $\dot{X}(0) = 0$, as claimed. □

We remark that the continuous limit of A-ADMM (see (21)) and A-GD (see (4)) are similar—second-order dynamical systems—with the only difference being the additional $(A^T A)^{-1}$ term. Therefore, in the special case $A = I$, i.e., the unconstrained case, (21) reduces to (4).

We close this section by noting one interesting difference between the derivations for the dynamical systems associated with ADMM and A-ADMM. Namely, the derivation for ADMM required choosing $\delta = 1/\rho$, whereas the derivation for A-ADMM made the choice $\delta = 1/\sqrt{\rho}$. Since the relationship $t = \delta k$ holds, we see that for fixed k and $\rho > 1$, the time elapsed for A-ADMM is larger than that for ADMM, which highlights the acceleration achieved by A-ADMM.

3. A Review of Lyapunov Stability

In the next section, we will use a Lyapunov stability approach to analyze the dynamical systems established in the previous section for ADMM and A-ADMM, namely (10) and (21), respectively. In this section, we give the required background material.

For generality, consider the first-order dynamical system

$$\dot{Y} = F(Y, t) \quad \text{with } Y(t_0) = Y_0, \tag{31}$$

where $F : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}^p$, $Y = Y(t) \in \mathbb{R}^p$, and $Y_0 \in \mathbb{R}^p$. When F is Lipschitz continuous this initial value problem is well-posed. Indeed, let $\Omega \subseteq \mathbb{R}^p \times \mathbb{R}$ and suppose that F is continuously differentiable on Ω. Let $(Y_0, t_0) \in \Omega$. The Cauchy-Lipschitz theorem assures that (31) has a unique solution $Y(t)$ on an open interval around t_0 such that $Y(t_0) = Y_0$. This solution may be extended throughout Ω. Moreover, the solution is a continuous function of the initial condition (Y_0, t_0), and if F depends continuously on some set of parameters, then it is also a continuous function of those parameters (Hirsch et al., 2004).

For a dynamical system in the form (31) we have the following three basic types of stability.

Definition 4 (Stability (Hirsch et al., 2004))**.** *A point $Y^\star$ such that $F(Y^\star, t) = 0$ for all $t \geq t_0$ is called a critical point of the dynamical system (31). We say the following:*

(i) $Y^\star$ *is* stable *if for every neighborhood $\mathcal{O} \subseteq \mathbb{R}^p$ of $Y^\star$, there exists a neighborhood $\bar{\mathcal{O}} \subseteq \mathcal{O}$ of $Y^\star$ such that every solution $Y(t)$ with initial condition $Y(t_0) = Y_0 \in \bar{\mathcal{O}}$ is defined and remains in $\mathcal{O}$ for all $t > t_0$;*

(ii) $Y^\star$ *is* asymptotically stable *if it is stable and, additionally, satisfies $\lim_{t\to\infty} Y(t) = Y^\star$ for all $Y_0 \in \bar{\mathcal{O}}$;*

(iii) $Y^\star$ *is* unstable *if it is not stable.*

Stability implies the existence of a region around $Y^\star$, i.e., the basin of attraction, in which solutions to the differential equation remain in such a region provided the initial condition Y_0 is sufficiently close to $Y^\star$. Asymptotic stability is stronger, further requiring that trajectories converge to $Y^\star$. We note that convergence of the trajectory alone does not imply stability.

Lyapunov formulated a strategy that enables one to conclude stability without integrating the equations of motion.

Theorem 5 (Lyapunov (Hirsch et al., 2004))**.** *Let $Y^\star$ be a critical point of the dynamical system (31). Also, let $\mathcal{O} \subseteq \mathbb{R}^n$ be an open set containing $Y^\star$ and $\mathcal{E} : \mathcal{O} \to \mathbb{R}$ be a continuously differentiable function. We have the following:*

(i) if $\mathcal{E}(\cdot)$ satisfies

$$\mathcal{E}(Y^\star) = 0, \tag{32}$$

$$\mathcal{E}(Y) > 0 \text{ for all } Y \in \mathcal{O} \setminus Y^\star, \tag{33}$$

$$\dot{\mathcal{E}}(Y) \leq 0 \text{ for all } Y \in \mathcal{O} \setminus Y^\star, \tag{34}$$

then $Y^\star$ is stable and $\mathcal{E}$ is called a Lyapunov function;

(ii) if instead of (34) we have the strict inequality

$$\dot{\mathcal{E}}(Y) < 0 \text{ for all } Y \in \mathcal{O} \setminus Y^\star, \tag{35}$$

then $Y^\star$ is asymptotically stable and $\mathcal{E}$ is called a strict Lyapunov function.

The drawback of Lyapunov's approach is that it requires knowing an appropriate $\mathcal{E}(\cdot)$; unfortunately, there is no systematic procedure for constructing such a function. Also, note that Lyapunov's criteria are sufficient but not necessary.

4. Stability and Convergence Analysis

In this section we analyze the stability properties and rates of convergence of the dynamical systems associated with both ADMM and A-ADMM.

4.1. Analysis of the Dynamical System for ADMM

ASYMPTOTIC STABILITY

The asymptotic stability of the ADMM flow (10) follows from Theorem 5 with an appropriately chosen Lyapunov function.

Theorem 6. *Let $X^\star$ be a strict local minimizer and an isolated stationary point of $V(\cdot)$, i.e., there exists $\mathcal{O} \subseteq \mathbb{R}^n$ such that $X^\star \in \mathcal{O}$, $\nabla V(X) \neq 0$ for all $X \in \mathcal{O} \setminus X^\star$, and*

$$V(X) > V(X^\star) \text{ for all } X \in \mathcal{O} \setminus X^\star. \tag{36}$$

Then, it follows that $X^\star$ is an asymptotically stable critical point of the ADMM flow (10).

Proof. Since $X^\star$ is a minimizer of $V(\cdot)$, it follows from first-order optimality conditions that $\nabla V(X^\star) = 0$. Combining this fact with Definition 4 shows that $X^\star$ is a critical point of the dynamical system (10). To prove that $X^\star$ is asymptotically stable, let us define

$$\mathcal{E}(X) \equiv V(X) - V(X^\star) \tag{37}$$

and observe from (36) that (32) and (33) hold. Then, taking the total time derivative of $\mathcal{E}$ and using (10) we have

$$\dot{\mathcal{E}}(X) = \left\langle \nabla V(X), \dot{X} \right\rangle = -\|A\dot{X}\|^2. \tag{38}$$

Since $X^\star$ is assumed to be an isolated critical point, we know that if $X \in \mathcal{O} \setminus X^\star$, then $\nabla V(X) \neq 0$, which in light of (10) and Assumption 1 means that $\dot{X} \neq 0$. Combining this conclusion with (38) and Assumption 1 shows that if $X \in \mathcal{O} \setminus X^\star$, then $\dot{\mathcal{E}}(X) < 0$, i.e., (35) holds. Therefore, it follows from Theorem 5 that $X^\star$ is an asymptotically stable critical point of the dynamical system (10). □

Some remarks concerning Theorem 6 are appropriate.

- If $V(\cdot)$ is strongly convex, then it has a unique minimizer. Moreover, that unique minimizer will satisfy the assumptions of Theorem 6 with $\mathcal{O} = \mathbb{R}^n$. Strong convexity of $V(\cdot)$ holds, for example, when either f or g is convex and the other is strongly convex (recall that A has full column rank by assumption). Similar remarks also hold when $V(\cdot)$ is merely strictly convex.

- If $X^\star$ satisfies the second-order sufficient optimality conditions for minimizing $V(\cdot)$, i.e., $\nabla V(X^\star) = 0$ and $\nabla^2 V(X^\star)$ is positive definite, then the assumptions of Theorem 6 will hold at $X^\star$ for all sufficiently small neighborhoods $\mathcal{O}$ of $X^\star$. Note that in this case, the function $V(\cdot)$ need not be convex.

- It follows from (38) that $\dot{\mathcal{E}}(X) \leq 0$ for all X. Thus, $X^\star$ will be stable (not necessarily asymptotically stable) without having to assume that $X^\star$ is an isolated stationary point of $V(\cdot)$.

CONVERGENCE RATE

For the dynamical system governing ADMM we are able to establish a convergence rate for how fast the objective function converges to its optimal value.

Theorem 7. *Let $X(t)$ be a trajectory of the ADMM flow (10), with initial condition $X(t_0) = x_0$. Assume that $\arg\min V \neq \emptyset$ and denote $V^\star \equiv \min_x V(x)$. Then, there is a constant $C > 0$ such that*

$$V(X(t)) - V^\star \leq \frac{C}{t}. \tag{39}$$

Proof. Let $X^\star \in \arg\min V$ and consider

$$\mathcal{E}(X,t) \equiv t\,[V(X) - V(X^\star)] + \tfrac{1}{2}\left\|A\,(X - X^\star)\,\right\|^2. \tag{40}$$

By taking the total time derivative of $\mathcal{E}$, using (10), and then the convexity of $V(\cdot)$, we find that

$$\begin{aligned}\dot{\mathcal{E}} &= t\left\langle \nabla V(X), \dot{X}\right\rangle + V(X) - V(X^\star) + \left\langle X - X^\star, A^T A\,\dot{X}\right\rangle \\ &= -t\left\|A\dot{X}\right\|^2 + V(X) - V(X^\star) + \left\langle X^\star - X, \nabla V(X)\right\rangle \\ &\leq 0,\end{aligned} \tag{41}$$

from which we may conclude that $\mathcal{E}(X,t) \leq \mathcal{E}(X_0,t_0)$ for all $t \geq t_0$. Combining this with the definition of $\mathcal{E}$ gives

$$\begin{aligned}V(X) - V(X^\star) &= \frac{1}{t}\mathcal{E}(X,t) - \frac{1}{2t}\|A\,(X - X^\star)\,\|^2 \\ &\leq \frac{\mathcal{E}(X_0,t_0)}{t},\end{aligned} \tag{42}$$

where we note that $\mathcal{E}(X_0,t_0) \geq 0$ since V is convex. □

Some remarks concerning Theorem 7 are warranted.

- Theorem 7 holds under the assumption that f and g are convex. This is a strength compared with Theorem 6, which has to make relatively strong assumptions about the critical point. Under those stronger assumptions, however, Theorem 6 gives a convergence result for the state X, whereas Theorem 7 only guarantees convergence of the objective value.

- The $O(1/t)$ rate promised by Theorem 7 for the dynamical system (10) associated with ADMM agrees with the rate $O(1/k)$ of ADMM when $V(\cdot)$ is assumed to be convex (Eckstein & Yao, 2015; He & Yuan, 2012).

4.2. Analysis of the Dynamical System for A-ADMM

STABILITY

A stability result for the A-ADMM flow (21) can be established by combining Theorem 5 with an appropriately chosen Lyapunov function.

Let $Y_1 = X$ and $Y_2 = \dot{X}$, and denote $Y = (Y_1, Y_2)$. Thus, we are able to write the second-order dynamical system (21) as the following system of first-order differential equations:

$$\frac{d}{dt}\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} Y_2 \\ -\frac{r}{t}Y_2 - (A^TA)^{-1}\nabla V(Y_1) \end{pmatrix}. \tag{43}$$

We can now give conditions on minimizers $X^\star$ of $V(\cdot)$ that ensure that $Y^\star = (X^\star, 0)$ is a stable critical point for (43).

Theorem 8. *If $X^\star$ be a strict local minimizer of $V(\cdot)$, i.e., there exists $\mathcal{O} \subseteq \mathbb{R}^n$ such that $X^\star \in \mathcal{O}$ and*

$$V(X) > V(X^\star) \text{ for all } X \in \mathcal{O} \setminus X^\star, \tag{44}$$

then $Y^\star = (X^\star, 0)$ is a stable critical point of (43), which is equivalent to the A-ADMM flow (21).

Proof. Since $X^\star$ is a minimizer of $V(\cdot)$, it follows from first-order optimality conditions that $\nabla V(X^\star) = 0$. Combining this with Definition 4 shows that $Y^\star = (X^\star, 0)$ is a critical point of the first-order dynamical system (43).

Next, we prove that $Y^\star = (X^\star, 0)$ is stable. Let $\mathcal{O} \subseteq \mathbb{R}^n$ and define $\mathcal{E} : \mathcal{O} \to \mathbb{R}$ as

$$\mathcal{E}(Y) = \tfrac{1}{2}\|AY_2\|^2 + V(Y_1) - V(X^\star). \tag{45}$$

Note that $\mathcal{E}(Y^\star) = 0$, i.e., condition (32) holds. Also, since $X^\star$ is isolated, $\mathcal{E}(Y) > 0$ for all $Y \neq Y^\star$, so that (33) holds. If we take the total time derivative of (45) we obtain

$$\begin{aligned}\dot{\mathcal{E}} &= \left\langle \nabla_{Y_1}\mathcal{E}, \dot{Y}_1\right\rangle + \left\langle \nabla_{Y_2}\mathcal{E}, \dot{Y}_2\right\rangle \\ &= \left\langle \nabla V(Y_1), Y_2\right\rangle - \left\langle A^TAY_2, \frac{r}{t}Y_2 + (A^TA)^{-1}\nabla V(Y_1)\right\rangle \\ &= -\tfrac{r}{t}\|AY_2\|^2.\end{aligned} \tag{46}$$

Thus, $\dot{\mathcal{E}}(Y) \leq 0$ for all $Y \in \mathcal{O}$, i.e., (34) holds. This implies that $Y^\star = (X^\star, 0)$ is stable, as claimed. □

We remark that the discussions in the first two bullet points of the subsection "Asymptotic stability" in Section 4.1 also apply to Theorem 8. We do not repeat them for brevity. We also note that the stability of system (4) was not considered by Su et al. (2016); Wibisono et al. (2016); Krichene et al.

(2015); Wilson et al. (2016). In contrast, Theorems 6 and 8 provide a simple argument for the stability of (10) and (21), respectively, based only on Theorem 5. However, contrary to the first-order system (10) associated to ADMM, it is not obvious how to apply Theorem 5 to system (21) associated to A-ADMM to obtain asymptotic stability without further assumptions on the critical point. However, we give a case where asymptotic stability holds in the end of this section.

Let us briefly mention existing results regarding the convergence of trajectories of the system (4) when $X(t)$ is an element of a general Hilbert space. Convergence of trajectories for convex and even some particular cases of non-convex $f(\cdot)$ was studied by Cabot et al. (2009). If $r > 3$ and $\arg\min f \neq \emptyset$, then the trajectory of the system weakly converges to some minimizer of $f(\cdot)$, even in the presence of small perturbations (Attouch et al., 2016). These results should extend naturally to (21), but we avoided diving in this direction since it would deviate from the main goal. It is important to note, however, that convergence of the trajectories do not necessarily imply stability.

Convergence rate

We now consider the convergence rate of the dynamical system (21) associated to A-ADMM.

Theorem 9. *Let $X(t)$ be a trajectory of the A-ADMM flow (21), with initial conditions $X(t_0) = x_0$ and $\dot{X}(t_0) = 0$. Assume that $\arg\min V \neq \emptyset$ and denote $V^\star \equiv \min_x V(x)$. If $r \geq 3$, then there is some constant $C > 0$ such that*

$$V(X(t)) - V^\star \leq \frac{C}{t^2}. \tag{47}$$

Proof. Following Su et al. (2016); Wibisono et al. (2016), we define $\eta : [t_0, \infty) \to \mathbb{R}$ as $\eta(t) = 2\log(t/(r-1))$ and

$$\mathcal{E}(Y,t) = e^{\eta}\big[V(Y_1) - V(X^\star)\big] + \tfrac{1}{2}\big\|A\big(Y_1 - X^\star + e^{\eta/2}Y_2\big)\big\|^2$$

where $X^\star$ is any minimizer of $V(\cdot)$. Note that $\mathcal{E} \geq 0$ and its total time derivative is given by

$$\begin{aligned}\dot{\mathcal{E}} &= \langle \nabla_{Y_1}\mathcal{E}, \dot{Y}_1\rangle + \langle \nabla_{Y_2}\mathcal{E}, \dot{Y}_2\rangle + \partial_t\mathcal{E} \\ &= \big\langle \nabla_{Y_1}\mathcal{E} - \tfrac{r}{t}\nabla_{Y_2}\mathcal{E}, Y_2\big\rangle - \langle \nabla_{Y_2}\mathcal{E}, (A^TA)^{-1}\nabla_{Y_1}V\rangle + \partial_t\mathcal{E}\end{aligned}$$

where we made use of (43). Observe that

$$\nabla_{Y_1}\mathcal{E} = e^{\eta}\nabla_{Y_1}V + A^TA\big(Y_1 - X^\star + e^{\eta/2}Y_2\big), \tag{48}$$

$$\nabla_{Y_2}\mathcal{E} = e^{\eta/2}A^TA\big(Y_1 - X^\star + e^{\eta/2}Y_2\big), \tag{49}$$

$$\begin{aligned}\partial_t\mathcal{E} &= \dot{\eta}e^{\eta}\big(V(Y_1) - V(X^\star)\big) \\ &\quad + \tfrac{\dot{\eta}}{2}e^{\eta/2}\big\langle Y_1 - X^\star + e^{\eta/2}Y_2, A^TAY_2\big\rangle,\end{aligned} \tag{50}$$

and also that

$$e^{-\eta/2} + \tfrac{1}{2}\dot{\eta} = \tfrac{r}{t}. \tag{51}$$

Therefore, using the convexity of $V(\cdot)$ we obtain

$$\begin{aligned}\dot{\mathcal{E}} &= \dot{\eta}e^{\eta}\big(V(Y_1) - V(X^\star)\big) - e^{\eta/2}\langle Y_1 - X^\star, \nabla V\rangle \\ &\leq -e^{\eta/2}(1 - \dot{\eta}e^{\eta/2})\,(V(Y_1) - V(X^\star)) \\ &= -\frac{t(r-3)}{(r-1)^2}\big(V(Y_1) - V(X^\star)\big),\end{aligned} \tag{52}$$

so that $\dot{\mathcal{E}} \leq 0$, implying $\mathcal{E}(Y,t) \leq \mathcal{E}((X(t_0), \dot{X}(t_0)), t_0)$. By the definition of $\mathcal{E}(\cdot)$ we thus have

$$\begin{aligned}V(Y_1) - V(X^\star) &\leq e^{-\eta}\mathcal{E}(Y,t) \\ &\leq e^{-\eta}\mathcal{E}(X(t_0), \dot{X}(t_0), t_0).\end{aligned} \tag{53}$$

To conclude the proof, observe that $e^{\eta} = t^2/(r-1)^2$. □

Theorem 9 suggests that A-ADMM has a convergence rate of $O\big(1/k^2\big)$ for convex functions. This agrees with the result by Goldstein et al. (2014), which assumes strong convexity of both f and g, and also that g is quadratic; see (5). Moreover, Goldstein et al. (2014) do not bound the objective function as in (47) but the combined residuals. To the best of our knowledge, there is no $\mathcal{O}(1/k^2)$ convergence proof for A-ADMM assuming only convexity. It would be interesting to consider the convergence rate of A-ADMM directly through a discrete analog of the Lyapunov function used in the above theorem, in the same spirit as Su et al. (2016) considered for A-GD and more recently Attouch et al. (2016) considered for a perturbed version of A-GD.

Asymptotic stability

Under stronger conditions than in Theorem 6, we have asymptotic stability of the dynamical system (21).

Theorem 10. *Let $X^\star \in \mathcal{O}$, for some $\mathcal{O} \subseteq \mathbb{R}^n$, be a local minimizer of $V(\cdot)$ satisfying*

$$V(X) - V(X^\star) \geq \phi(\|X - X^\star\|) \quad \textit{for all } X \in \mathcal{O}, \tag{54}$$

where $\phi : [0,\infty) \to [0,\infty)$ is a forcing function (Ortega & Rheinboldt, 2000) such that for any $\{\xi_k\} \subset [0,\infty)$, $\lim_{k\to\infty}\phi(\xi_k) = 0$ implies $\lim_{k\to\infty}\xi_k = 0$. Moreover, suppose that the conditions of Theorem 9 hold over $\mathcal{O}$. Then, it follows that $Y^\star = (X^\star, 0)$ is an asymptotically stable critical point of the dynamical system (43), which is equivalent to the A-ADMM flow (21).

Proof. Consider (54) over a trajectory $X = X(t)$. Using (47) and (54) we have $\lim_{t\to\infty}\phi(\|X(t) - X^\star\|) = 0$, which combined with the properties of the forcing function gives

$$\lim_{t\to\infty}\|X(t) - X^\star\| = 0. \tag{55}$$

Denote $Y(t) = (Y_1(t), Y_2(t)) = (X(t), \dot{X}(t))$. From the proof of Theorem 9, i.e., the definition of $\mathcal{E}$ and $\dot{\mathcal{E}} \leq 0$,

we also have that $\|A(Y_1 - X^\star) + e^{\eta/2}Y_2\|^2 \leq C$, where $C = \mathcal{E}(X(t_0), \dot{X}(t_0))$, hence

$$e^{\eta(t)}\|Y_2(t)\|^2 \leq C + \|A(Y_1(t) - X^\star)\|^2. \tag{56}$$

This implies that $\lim_{t\to\infty} \|Y_2(t)\| = 0$ upon using (55).

We showed that $\lim_{t\to\infty} Y(t) = (X^\star, 0)$. From Theorem 6 we already know that $Y^\star = (X^\star, 0)$ is a stable critical point of (43). From these two facts, and Definition 4, we thus conclude that $Y^\star$ is asymptotically stable, as claimed. □

Condition (54) holds, for instance, for both uniformly convex functions and strongly convex functions.

5. A Numerical Example

We numerically verify that the differential equations (10) and (21) accurately model ADMM and A-ADMM, respectively, when ρ is large as needed to derive the continuous limit. The numerical integration of the first-order system (10) is straightforward; we use a 4th order Runge-Kutta method (an explicit Euler method could also be employed). The numerical integration of (21) is more challenging due to strong oscillations. To obtain a faithful discretization of the continuous dynamical system (21), i.e., one that preserves its properties, a standard approach is to use a Hamiltonian symplectic integrator, which is designed to preserve the phase-space volume. Consider the Hamiltonian

$$\mathcal{H} \equiv \tfrac{1}{2}e^{-\xi(t)}\big\langle P, (A^TA)^{-1}P\big\rangle + e^{\xi(t)}V(X), \tag{57}$$

where $\xi(t) \equiv r\log t$ and $P = e^{\xi}(A^TA)\dot{X}$ is the canonical momentum. Hamilton's equations are given by

$$\dot{X} = \nabla_P \mathcal{H} \quad \text{and} \quad \dot{P} = -\nabla_X \mathcal{H}. \tag{58}$$

One can check that (58) together with (57) is equivalent to (21). The simplest scheme is the symplectic Euler method, which for equations (58) with (57) is given explicitly as

$$p_{k+1} = p_k - he^{\xi(t_k)}\nabla V(x_k), \tag{59a}$$
$$x_{k+1} = x_k + he^{-\xi(t_k)}\big(A^TA\big)^{-1}p_{k+1}, \tag{59b}$$
$$t_{k+1} = t_k + h, \tag{59c}$$

where $h > 0$ is the step size. Thus, we compare the iterates (59) with the A-ADMM algorithm. A simple example is provided in Figure 1, which illustrate our theoretical results.

6. Conclusions

Previous work considered dynamical systems for continuous limits of gradient-based methods for unconstrained optimization (Su et al., 2016; Wibisono et al., 2016; Krichene et al., 2015). Our paper builds upon these results by showing

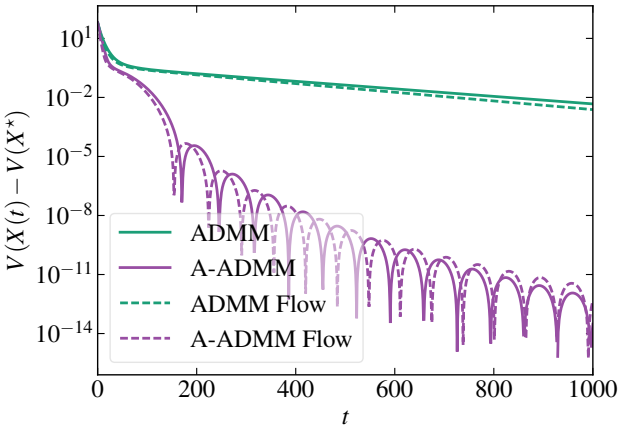

Figure 1. $\min_x V(x)$ such that $z = Ax$ with $V(x) = \frac{1}{2}\langle x, Mx\rangle$, where $M \in \mathbb{R}^{60\times 60}$ is a random matrix with 40 zero eigenvalues and the remaining ones are uniformly distributed on $[0, 10]$, and A is a full column random matrix with condition number 100. Comparison of ADMM versus ADMM flow (10) through 4th order Runge-Kutta, and A-ADMM versus A-ADMM flow (21) through symplectic Euler (59). We choose $r = 10$ and $\rho = 50$. The initial conditions are $X(0) = x_0 = 5(1, 1, \ldots, 1)^T$ and $\dot{X}(0) = 0$. The curves are close and the rates (42) and (47) hold.

that the continuous limits of ADMM and A-ADMM correspond to first- and second-order dynamical systems, respectively; see Theorems 2 and 3. Next, using a Lyapunov stability analysis, we presented conditions that ensure stability and asymptotic stability of the dynamical systems; see Theorems 6, 8 and 10. Furthermore, in Theorem 7 we obtained a convergence rate of $O(1/t)$ for the dynamical system related to ADMM, which is consistent with the known $O(1/k)$ convergence rate of the discrete-time ADMM, whereas in Theorem 9 we obtained a convergence rate of $O(1/t^2)$ for the dynamical system related to A-ADMM, which is a new result since this rate is unknown for discrete-time A-ADMM. We also showed that the dynamical system associated to A-ADMM is a Hamiltonian system, and by employing a simple symplectic integrator verified numerically the agreement between discrete- and continuous-time dynamics.

The results presented in this paper may be useful for understanding the behavior of ADMM and A-ADMM for non-convex problems as well. For instance, following ideas from Jin et al. (2017) and Lee et al. (2017) an analysis of the center manifold of the dynamical systems (10) and (21) can provide valuable insights on the stability of saddle points, which is considered a major issue in non-convex optimization. Also, ADMM is well-suited to large-scale problems in statistics and machine learning, being equivalent to Douglas-Rachford splitting and closely related to other algorithms such as augmented Lagrangian methods, dual decomposition, and Dykstra's alternating projections. Therefore, our results may give new insights into these methods as well.

Acknowledgements

This work was supported by grants ARO MURI W911NF-17-1-0304 and NSF 1447822.

References

Attouch, H., Chbani, Z., Peypouquet, J., and Redont, P. Fast Convergence of Inertial Dynamics and Algorithms with Asymptotic Vanishing Viscosity. *Mathematical Programming*, pp. 1–53, March 2016.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Cabot, A., Engler, H., and Gadat, S. On the Long Time Behaviour of Second Order Differential Equations with Asymptotically Small Dissipation. *Transactions of the American Mathematical Society*, 361(11):5983–6017, 2009.

Eckstein, J. and Yao, W. Understanding the Convergence of the Alternating Direction Method of Multipliers: Theoretical and Computational Perspectives. 2015.

França, G. and Bento, J. An Explicit Rate Bound for Over-Relaxed ADMM. In *IEEE International Symposium on Information Theory, ISIT 2016, Barcelona, Spain, July 10-15*, pp. 2104–2108, 2016.

França, G. and Bento, J. Markov Chain Lifting and Distributed ADMM. *IEEE Signal Processing Letters*, 24: 294–298, 2017a.

França, G. and Bento, J. How is Distributed ADMM Affected by Network Topology? arXiv:1710.00889 [stat.ML], 2017b.

Gabay, D. and Mercier, B. A Dual Algorithm for the Solution of Nonlinear Variational Problems via Finite Element Approximations. *Computers and Mathematics with Applications*, 2(1):17–40, 1976.

Glowinski, R. and Marroco, A. Sur l'approximation, par él'ements finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de probèmes de Dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 9(R2):41–76, 1975.

Goldstein, T., O'Donoghue, B., Setzer, S., and Baraniuk, R. Fast Alternating Direction Optimization Methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014.

He, B. and Yuan, X. On the $O(1/n)$ Convergence Rate of the DouglasRachford Alternating Direction Method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

Hirsch, M. W., Smale, S., and Devaney, R. L. *Differential Equations, Dynamical Systems, and An introduction to Chaos*. Academic Press, 2004.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to Escape Saddle Points Efficiently. arXiv:1703.00887 [cs.LG], 2017.

Krichene, W., Bayen, A., and Bartlett, P. L. Accelerated mirror descent in continuous and discrete time. *Advances in Neural Information Processing Systems 28*, pp. 2845–2853, 2015.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order Methods Almost Always Avoid Saddle Points. arXiv:1710.07406 [stat.ML], 2017.

Lessard, L., Recht, B., and Packard, A. Analysis and Design of Optimization Algorithms via Integral Quadratic Constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Nesterov, Y. A Method of Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

Nishihara, R., Lessard, L., Recht, B., Packard, A., and Jordan, M. I. A General Analysis of the Convergence of ADMM. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 343–352, 2015.

Ortega, J. M. and Rheinboldt, W. C. *Iterative Solution of Nonlinear Equations in Several Variables*. SIAM, 2000.

Su, W., Boyd, S., and Candès, E. J. A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

Wibisono, A., Wilson, A. C., and Jordan, M. I. A Variational Perspective on Accelerated Methods in Optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

Wilson, A. C., Recht, B., and Jordan, M. I. A Lyapunov Analysis of Momentum Methods in Optimization. arXiv:1611.02635v3 [math.OC], 2016.