Brief paper

Analytical convergence regions of accelerated gradient descent in nonconvex optimization under Regularity Condition ☆

Huaqing Xiong [a,1], Yuejie Chi [b,2], Bin Hu [c], Wei Zhang [d,*,1]

[a] *Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210, USA*
[b] *Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[c] *Department of Electrical and Computer Engineering and Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*
[d] *Department of Mechanical and Energy Engineering, Southern University of Science and Technology (SUSTech), Shenzhen, Guangdong 518055, China*

ARTICLE INFO

ABSTRACT

There is a growing interest in using robust control theory to analyze and design optimization and machine learning algorithms. This paper studies a class of nonconvex optimization problems whose cost functions satisfy the so-called Regularity Condition (RC). Empirical studies show that accelerated gradient descent (AGD) algorithms (e.g. Nesterov's acceleration and Heavy-ball) with proper initializations often work well in practice. However, the convergence of such AGD algorithms is largely unknown in the literature. The main contribution of this paper is the analytical characterization of the convergence regions of AGD under RC via robust control tools. Since such optimization problems arise frequently in many applications such as phase retrieval, training of neural networks and matrix sensing, our result shows promise of robust control theory in these areas.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

Recently, control theoretic tools have gained popularity in analyzing optimization and machine learning algorithms (Cherukuri, Mallada, Low, & Cortés, 2017; Fazlyab, Ribeiro, Morari, & Preciado, 2018; Hu & Lessard, 2017b; Hu, Seiler, & Lessard, 2017; Hu, Seiler, & Rantzer, 2017; Hu, Wright, & Lessard, 2018; Lessard, Recht, & Packard, 2016; Nishihara, Lessard, Recht, Packard, & Jordan, 2015; Sundararajan, Hu, & Lessard, 2017; Wilson, Recht, & Jordan, 2016). Typically, convergence analysis in optimization is performed in a case-by-case manner and the corresponding techniques highly depend on the structure of algorithms and assumptions of objective functions. However, by representing iterative algorithms and prior information of objective functions as feedback dynamical systems, we can apply tools from control theory to carry out the convergence analysis in a more systematic way.

Such a framework is pioneered by Lessard et al. (2016), where the authors used semidefinite programming to analyze the convergence of a class of optimization algorithms including gradient descent (GD), Heavy-ball (HB) and Nesterov's accelerated gradient (NAG), by assuming the gradient of the loss function satisfies some integral quadratic constraints (IQCs). Indeed, the standard smoothness and convexity assumptions can be well rephrased as IQCs. Afterwards, similar approaches have been developed to analyze various optimization algorithms such as ADMM (Nishihara et al., 2015), distributed methods (Sundararajan et al., 2017), proximal algorithms (Fazlyab et al., 2018), and stochastic finite-sum methods (Hu, Seiler, & Rantzer, 2017; Hu et al., 2018). Exploiting the connection between control and optimization also provides new insights into momentum methods (Hu & Lessard, 2017a, 2017b; Wilson et al., 2016) and facilitates the design of new algorithms (Cyrus, Hu, Van Scoy, & Lessard, 2018; Dhingra, Khong, & Jovanovic, 2019; Kolarijani, Esfahani, & Keviczky, 2018; Van Scoy, Freeman, & Lynch, 2018). Moreover, control tools are also useful in analyzing the robustness of algorithms against computation inexactness (Aybat, Fallah, Gurbuzbalaban, & Ozdaglar, 2018; Cherukuri et al., 2017; Hu, Seiler, & Lessard, 2017; Lessard et al., 2016).

This paper considers a class of nonconvex optimization problems whose objective functions satisfy the so-called Regularity

Condition (RC) (Candes, Li, & Soltanolkotabi, 2015; Chi, Lu, & Chen, 2019), which is a geometric condition characterizing the curvatures of nonconvex functions. Such a condition has appeared in many important machine learning and signal processing applications, including phase retrieval (Candes et al., 2015; Chen & Candès, 2017; Wang, Giannakis, & Eldar, 2017; Zhang, Chi, & Liang, 2016; Zhang, Liang, & Chi, 2017), deep linear neural networks (Zhou & Liang, 2017), shallow nonlinear networks (Li & Yuan, 2017), matrix sensing (Li, Chi, Zhang, & Liang, 2019; Tu, Boczar, Simchowitz, Soltanolkotabi, & Recht, 2016), to name a few. While it is straightforward to show that GD converges linearly for problems satisfying RC (Candes et al., 2015), however, the behavior of AGD under RC remains poorly understood in theory, despite its empirical success (Pauwels, Beck, Eldar, & Sabach, 2017).

Our work aims to deepen the understanding of convergence of AGD under RC. The main contribution of this paper lies in the theoretical convergence guarantee of AGD algorithms under RC. In particular, we provide an analytical characterization of hyperparameter choices that ensure linear convergence of AGD for all functions that satisfy RC. In addition, the framework and tools developed herein may inspire more research on analyzing the convergence of sophisticated algorithms under nonconvex settings. Specifically, the analysis for momentum methods typically involves subtle constructions of Hamiltonians (or Lyapunov functions). Our frequency domain approach can implicitly ensure the existence of such Lyapunov functions without explicit constructions. This sheds new light on the analysis of momentum methods for nonconvex optimization problems.

It is worth noting that, RC is essentially the same as the sector bound condition in Lessard et al. (2016) when it holds globally. In Lessard et al. (2016, Sections 4.5 and 4.6), LMI conditions have been derived to analyze momentum methods under the sector bound condition. The results, however, are numerical validations of convergence for given hyperparameters by running a small semidefinite program, and such validation needs to be done whenever the hyperparameters are changed. Built on this prior work, we focus on how to obtain analytical convergence regions from these LMIs without solving semidefinite programs. Our analytical results for momentum methods under RC provide deeper understanding on the connection between control theory and nonconvex optimization.

The rest of the paper is organized as follows. In Section 2, we state the problem and introduce the AGD methods of interest. Section 3 presents how to incorporate the algorithms and RC into a dynamical system and transfer the convergence analysis into the stability analysis. Section 4 derives the analytical convergence conditions of AGD methods under RC via a frequency domain approach by applying the KYP lemma. In Section 5, we discuss how to extend the results to the general case where RC only holds locally.

2. Problem background

A general optimization problem can be described as

$$\underset{z \in \mathbb{R}^n}{\text{minimize}} \quad f(z), \tag{1}$$

where $f(z)$ may be both nonconvex and nonsmooth.

2.1. Regularity condition

This paper focuses on a special yet important case of nonconvex optimization, where the objective function $f(\cdot)$ satisfies the Regularity Condition (RC) (Candes et al., 2015), defined as follows.

Definition 1 (*Regularity Condition*)**.** A function $f(\cdot)$ is said to satisfy the Regularity Condition RC(μ, λ, ϵ) with positive constants μ, λ and ϵ, if

$$\langle \nabla f(z), z - x^\star \rangle \geq \frac{\mu}{2} \|\nabla f(z)\|^2 + \frac{\lambda}{2} \|z - x^\star\|^2 \tag{2}$$

for all $z \in \mathcal{N}_\epsilon(x^\star) := \{z : \|z - x^\star\| \leq \epsilon \|x^\star\|\}$, where $x^\star$ is a local minimizer of $f(z)$.

It is straightforward to check that one must have $\mu\lambda \leq 1$ by Cauchy–Schwartz inequality. RC can be regarded as a combination of one-point strong convexity and smoothness (Chi et al., 2019), and does not require the function $f(\cdot)$ to be convex. RC has appeared in a wide range of applications, a partial list of which includes phase retrieval (Candes et al., 2015), deep linear neural networks (Zhou & Liang, 2017), shallow nonlinear neural networks (Li & Yuan, 2017) and matrix sensing (Li et al., 2019; Tu et al., 2016).

Our framework can handle the general case where RC only holds locally as defined. To simplify the presentation, we will first assume RC holds globally, i.e. $\epsilon = \infty$ in Definition 1, in which case $x^\star$ becomes the global minimizer correspondingly. It turns out that our main results can be directly applied to the case when RC holds locally, using proper initializations, which will be explained in detail in Section 5. Without ambiguity, we will omit the neighborhood radius ϵ and use the notation $RC(\mu, \lambda)$ to denote the global RC in the derivation of the main results.

2.2. Accelerated gradient descent methods

In practice, AGD methods are widely adopted for its ability to accelerate the convergence. Two widely-used acceleration schemes include Nesterov's accelerated gradient (NAG) method (Nesterov, 2003), given as

$$y_k = (1 + \beta)z_k - \beta z_{k-1},$$
$$z_{k+1} = y_k - \alpha \nabla f(y_k), \quad k = 0, 1, \dots, \tag{3}$$

where $\alpha > 0$ is the step size, $0 \leq \beta < 1$ is the momentum parameter; and Heavy-Ball (HB) method (Polyak, 1964), given as

$$y_k = (1 + \beta)z_k - \beta z_{k-1},$$
$$z_{k+1} = y_k - \alpha \nabla f(z_k), \quad k = 0, 1, \dots, \tag{4}$$

where $\alpha > 0$ is the step size, $0 \leq \beta < 1$ is the momentum parameter. In fact, we can describe a general AGD method that subsumes HB and NAG as special cases:

$$y_k = (1 + \beta_2)z_k - \beta_2 z_{k-1},$$
$$z_{k+1} = (1 + \beta_1)z_k - \beta_1 z_{k-1} - \alpha \nabla f(y_k). \tag{5}$$

Despite the empirical success, the convergence of AGD in the nonconvex setting remains unclear to a large extent. For example, it is not known whether AGD converges under RC, whether it converges linearly if it does and how to set the step size and the momentum parameters to guarantee its convergence. These challenging questions motivate us to look for new tools to better understand AGD for nonconvex optimization.

3. A control view on the convergence analysis of AGD under RC

Robust control theory has been tailored to the convergence analysis of optimization methods (Hu & Lessard, 2017b; Hu, Seiler, & Rantzer, 2017; Lessard et al., 2016; Nishihara et al., 2015). The proofs of our main theorems also rely on such techniques. In this section, we will discuss how to transform the convergence analysis of AGD under RC to the robust stability analysis of dynamical systems and derive LMI conditions to guarantee the convergence.
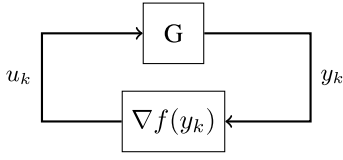
Fig. 1. The dynamical system representation of first-order methods.

3.1. AGD as feedback dynamical system

Observe that a general AGD method (5) can be viewed as a linear dynamical system subject to nonlinear feedback:

$$
\begin{aligned}
z_{k+1}^{(1)} &= (1 + \beta_1)z_k^{(1)} - \beta_1 z_k^{(2)} - \alpha u_k, \\
z_{k+1}^{(2)} &= z_k^{(1)}, \\
y_k &= (1 + \beta_2)z_k^{(1)} - \beta_2 z_k^{(2)}, \\
u_k &= \nabla f(y_k).
\end{aligned}
\tag{6}
$$

To see this, let $z_k^{(1)} = z_k$, and $z_k^{(2)} = z_{k-1}$. Then it can be easily verified that (6) represents HB when $(\beta_1, \beta_2) = (\beta, 0)$, and NAG when $(\beta_1, \beta_2) = (\beta, \beta)$.

Let $\otimes$ denote the Kronecker product. We use the notation $G(A, B, C, D)$ to denote a dynamical system G governed by the following iterative state-space model:

$$
\begin{aligned}
\phi_{k+1} &= (A \otimes I_n)\phi_k + (B \otimes I_n)u_k, \\
y_k &= (C \otimes I_n)\phi_k + (D \otimes I_n)u_k,
\end{aligned}
$$

where I_n is the identity matrix of size n. If we define $\phi_k = \begin{bmatrix} z_k^{(1)} \\ z_k^{(2)} \end{bmatrix}$ as the state, u_k as the input and y_k as the output, then (6) can be regarded as a dynamical system shown in Fig. 1, where the feedback $\nabla f(y_k)$ is a static nonlinearity that depends on the gradient of the loss function, and $G(A, B, C, D)$ is a linear system specified by

$$
\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right] = \left[\begin{array}{cc|c} 1 + \beta_1 & -\beta_1 & -\alpha \\ 1 & 0 & 0 \\ \hline 1 + \beta_2 & -\beta_2 & 0 \end{array} \right].
\tag{7}
$$

3.2. Convergence analysis of AGD under RC via stability analysis of a feedback system

In the following, we will illustrate how to show the convergence of AGD. First, define $\phi_* = \begin{bmatrix} x^\star \\ x^\star \end{bmatrix}$ as the equilibrium of the dynamical system (6). If the system is (globally) asymptotically stable, then $\phi_k \xrightarrow{k \to \infty} \phi_*$. It further implies $z_k \xrightarrow{k \to \infty} x^\star$. In other words, the asymptotic stability of the dynamical system can indicate the convergence of the iterates to a fixed point. From now on, we focus on the stability analysis of the feedback control system (6).

The stability analysis of the feedback control system (6) can be carried out using robust control theory. The main challenge is on the nonlinear feedback term $u_k = \nabla f(y_k)$. Our key observation is that RC can be rewritten as the following quadratic constraint:

$$
\begin{bmatrix} y_k - y_* \\ u_k - u_* \end{bmatrix}^T \begin{bmatrix} -\lambda I_n & I_n \\ I_n & -\mu I_n \end{bmatrix} \begin{bmatrix} y_k - y_* \\ u_k - u_* \end{bmatrix} \geq 0,
\tag{8}
$$

where $y_* = x^\star$ and $u_* = \nabla f(y_*) = 0$. Applying the quadratic constraint framework in Lessard et al. (2016), we can derive an LMI as a sufficient stability condition as stated in the following proposition. A formal proof of this result can be found in the Appendix.

Proposition 2. *Let $x^\star \in \mathbb{R}^n$ be the global minimizer of the loss function $f(\cdot)$ which satisfies $RC(\mu, \lambda)$. For a given first-order method characterized by $G(A, B, C, D)$, if there exist a matrix $P \succ 0$ and $\rho \in (0, 1)$ such that the following linear matrix inequality (LMI) (9) holds,*

$$
\begin{bmatrix} A^T PA - \rho^2 P & A^T PB \\ B^T PA & B^T PB \end{bmatrix} +
$$
$$
\begin{bmatrix} C & D \\ 0_{1\times 2} & 1 \end{bmatrix}^T \begin{bmatrix} -\lambda & 1 \\ 1 & -\mu \end{bmatrix} \begin{bmatrix} C & D \\ 0_{1\times 2} & 1 \end{bmatrix} \preceq 0,
\tag{9}
$$

then the state ϕ_k generated by the first-order algorithm $G(A, B, C, D)$ converges to the fixed point ϕ_* linearly, i.e.,

$$
\|\phi_k - \phi_*\| \leq \sqrt{\text{cond}(P)}\rho^k \|\phi_0 - \phi_*\| \text{ for all } k,
\tag{10}
$$

where $\text{cond}(P)$ is the condition number of P.

Remark 3. For fixed (A, B, C, D) and ρ, the LMI (9) is linear in P and hence an LMI. The size of this LMI is 3×3, and the decision variable P is a 2×2 matrix. The size of the LMI (9) is independent of the state dimension n.

Remark 4. The LMI (9) is similar to the one derived in Lessard et al. (2016) under the so-called sector bound condition. The relationship between RC and the sector bound is discussed in detail in the Appendix. Different from Lessard et al. (2016), we focus on deriving analytical convergence regions (see Section 4), in contrast to verifying convergence numerically for specific parameters, offering deeper insight regarding the convergence behavior of AGD methods under RC. In addition, we also extend the results to the case where RC holds only locally around the fixed point (see Section 5).

4. Convergence conditions of AGD

In this section, we focus on how to obtain analytical convergence conditions of AGD under RC based on (9).

Analytically solving (9) is challenging since one typically needs to express P explicitly as a function of (A, B, C, D) and (λ, μ). Our main idea is to transform the LMI (9) to equivalent frequency domain inequalities (FDIs) which can reduce unknown parameters using the classical KYP lemma (Rantzer, 1996). Then we can derive the main convergence results by solving the FDIs analytically.

4.1. The Kalman–Yakubovich–Popov (KYP) lemma

We first introduce the KYP lemma and the reader is referred to Rantzer (1996) for an elegant proof.

Lemma 5 (*Rantzer, 1996, Theorem 2*)**.** *Given A, B, M, with $\det(e^{j\omega}I - A) \neq 0$ for $\omega \in \mathbb{R}$, the following two statements are equivalent:*

(1) $\forall \omega \in \mathbb{R}$,

$$
\begin{bmatrix} (e^{j\omega}I - A)^{-1}B \\ I \end{bmatrix}^* M \begin{bmatrix} (e^{j\omega}I - A)^{-1}B \\ I \end{bmatrix} \prec 0.
\tag{11}
$$

(2) There exists a matrix $P \in \mathbb{R}^{n\times n}$ such that $P = P^T$ and

$$
M + \begin{bmatrix} A^T PA - P & A^T PB \\ B^T PA & B^T PB \end{bmatrix} \prec 0.
\tag{12}
$$

The general KYP lemma only asks P to be symmetric instead of being positive definite (PD) as in our problem. To ensure that the KYP lemma can be applied to solve (9), some adjustments of the lemma are necessary. In fact, we observe that if A of the dynamical system is Schur stable and the upper left corner of M,

denoted as M_{11}, is positive semidefinite (PSD), then by checking the principal minor $M_{11} + A^T PA - P \prec 0$, we know P satisfying (12) must be PD. We define these conditions on A and M as KYP Conditions, which are restrictions to make sure that all solutions of symmetric P for (12) are PD.

Definition 6 (*KYP Conditions*)**.** The KYPC(A, M) are listed as:

(1) $\det(e^{j\omega}I - A) \neq 0$ for $\omega \in \mathbb{R}$;
(2) A is Schur stable;
(3) The left upper corner of M in (12) is PSD.

Thus we can conclude the following corollary.

Corollary 7. *Under KYPC(A, M), the following two statements are equivalent:*

(1) $\forall \omega \in \mathbb{R}$,

$$
\begin{bmatrix} (e^{j\omega}I - A)^{-1}B \\ I \end{bmatrix}^* M \begin{bmatrix} (e^{j\omega}I - A)^{-1}B \\ I \end{bmatrix} \prec 0. \tag{13}
$$

(2) There exists a matrix $P \in \mathbb{R}^{n \times n}$ such that $P \succ 0$ and

$$
M + \begin{bmatrix} A^T PA - P & A^T PB \\ B^T PA & B^T PB \end{bmatrix} \prec 0. \tag{14}
$$

One can easily check, however, that A and M of a general AGD in (9) do not satisfy the KYPC(A, M). Therefore, we need to rewrite the dynamical system (6) in a different way to satisfy the KYPC(A, M), so that its stability analysis can be done by combining Proposition 2 and Corollary 7. In the following, we first introduce a way to rewrite the dynamical system to satisfy the KYPC(A, M).

4.2. How to satisfy the KYP conditions?

Recall that a general AGD can be written as (5). Here we introduce a slack variable δ to rewrite the algorithm:

$$
\begin{aligned}
z_{k+1} &= (1 + \delta + \beta_1 + \delta\beta_2) z_k - (\beta_1 + \delta\beta_2) z_{k-1} \\
&\quad - \alpha \nabla f(y_k) - \delta y_k, \\
y_k &= (1 + \beta_2) z_k - \beta_2 z_{k-1}.
\end{aligned} \tag{15}
$$

Observe that for any value of δ, (15) provides the same update rule as (5). It can be viewed as a generalized representation of the dynamical systems corresponding to the targeted AGD. Similar to (6), we rewrite (15) as a dynamical system $G(A', B', C', D')$:

$$
\begin{aligned}
z_{k+1}^{(1)} &= (1 + \beta_1 + \delta + \delta\beta_2) z_k^{(1)} - (\beta_1 + \delta\beta_2) z_k^{(2)} + u_k, \\
z_{k+1}^{(2)} &= z_k^{(1)}, \\
y_k &= (1 + \beta_2) z_k^{(1)} - \beta_2 z_k^{(2)}, \\
u_k &= -\alpha \nabla f(y_k) - \delta y_k.
\end{aligned} \tag{16}
$$

Correspondingly,

$$
\left[\begin{array}{c|c} A' & B' \\ \hline C' & D' \end{array} \right] = \left[\begin{array}{cc|c} 1 + \beta_1 + \delta + \delta\beta_2 & -(\beta_1 + \delta\beta_2) & 1 \\ 1 & 0 & 0 \\ \hline 1 + \beta_2 & -\beta_2 & 0 \end{array} \right].
$$

In addition to the adjustment of the dynamics, the feedback of $G(A', B', C', D')$ also differs from that in (6). As a consequence, the quadratic bound for the new feedback $u_k = -\alpha \nabla f(y_k) - \delta y_k$ is shifted as stated in the following lemma.

Lemma 8. *Let f be a loss function which satisfies $RC(\mu, \lambda)$ and $y_* = x^*$ be a minimizer. If $u_k = -\alpha \nabla f(y_k) - \delta y_k$, then y_k and u_k can be quadratically bounded as*

$$
\begin{bmatrix} y_k - y_* \\ u_k - u_* \end{bmatrix}^T M' \begin{bmatrix} y_k - y_* \\ u_k - u_* \end{bmatrix} \geq 0. \tag{17}
$$

where $M' = \begin{bmatrix} -(2\alpha\delta + \lambda\alpha^2 + \mu\delta^2) & -\alpha - \mu\delta \\ -\alpha - \mu\delta & -\mu \end{bmatrix}$.

Now we have general representations of A', M' with one unknown parameter δ. We need to certify the region of $(\alpha, \beta_1, \beta_2)$ such that its corresponding (A', M') has at least one δ satisfying the KYPC(A', M').

Lemma 9. *Let f be a loss function which satisfies $RC(\mu, \lambda)$. There is at least one representation of the dynamical system (16) satisfying KYPC(A', M'), if and only if the step size α and the momentum parameters β_1, β_2 obey the following restriction:*

$$
0 < \alpha < \frac{2(1 + \beta_1)(1 + \sqrt{1 - \mu\lambda})}{\lambda(1 + 2\beta_2)}. \tag{18}
$$

By Lemma 9, if the parameters of a fixed AGD with $(\alpha, \beta_1, \beta_2)$ satisfy (18), then all feasible symmetric P's for (14) can be guaranteed to be PD. Now we are ready to use the KYP lemma to complete the convergence analysis of an accelerated algorithm.

4.3. Stability region of AGD under RC

By Proposition 2, we can solve the stability of the new system (16) by finding some feasible $P \succ 0$ to the key LMI (9) with respect to the corresponding (A', B', C', D', M') for some rate $0 < \rho < 1$. We are interested in obtaining the analytical region of $(\alpha, \beta_1, \beta_2)$ that guarantees the linear convergence of AGD under RC. We use the following strict matrix inequality without caring about a specific rate ρ,

$$
\begin{bmatrix} A'^T PA' - P & A'^T PB' \\ B'^T PA' & B'^T PB' \end{bmatrix} + \begin{bmatrix} C' & 0 \\ 0_{1\times 2} & 1 \end{bmatrix}^T M' \begin{bmatrix} C' & 0 \\ 0_{1\times 2} & 1 \end{bmatrix} \prec 0. \tag{19}
$$

Remark 10. Our arguments can also be modified to derive the parameter region which guarantees the convergence with a fixed rate ρ. For such an analysis, we can modify the LMI (19) by rescaling the matrices (A', B') as $\tilde{A} = A'/\rho$ and $\tilde{B} = B'/\rho$. Then the resultant LMI can be converted to an FDI by the KYP lemma, and a similar analysis can be carried forward. Such analytical analysis of the convergence rate is even more difficult to interpret due to the presence of ρ in $(\tilde{A}, \tilde{B})$. For simplicity, this paper focuses on the derivation of stability regions.

Observe that now (19) is of the same form as (14). By the KYP lemma (Corollary 7), under KYPC(A', M') (19) can be equivalently solved by studying the following FDI:

$$
\begin{bmatrix} (e^{j\omega}I - A')^{-1}B' \\ I \end{bmatrix}^* \begin{bmatrix} C' & 0 \\ 0_{1\times 2} & 1 \end{bmatrix}^T M' \begin{bmatrix} C' & 0 \\ 0_{1\times 2} & 1 \end{bmatrix} \begin{bmatrix} (e^{j\omega}I - A')^{-1}B' \\ I \end{bmatrix} < 0, \quad \forall \omega \in \mathbb{R}. \tag{20}
$$

By simplifying (20) we observe that all uncertain terms can be canceled out and then conclude the following lemma.

Lemma 11. *To find the stability region of a general AGD method under $RC(\mu, \lambda)$, or equivalently, to find the region of $(\alpha, \beta_1, \beta_2)$ such that there exists a feasible $P \succ 0$ satisfying (19), it is equivalent to find $(\alpha, \beta_1, \beta_2)$ which simultaneously obeys (18) and guarantees the following FDI:*

$$
\begin{aligned}
&4(\alpha\beta_2 - \mu\beta_1)\cos^2\omega + 2[\mu(1+\beta_1)^2 + \lambda\alpha^2\beta_2(1+\beta_2) \\
&\quad - \alpha(1+\beta_1)(1+2\beta_2)]\cos\omega + 2\alpha(1+\beta_1+2\beta_1\beta_2) \\
&\quad - 2\mu(1+\beta_1^2) - \lambda\alpha^2[\beta_2^2 + (1+\beta_2)^2] < 0, \quad \forall \omega \in \mathbb{R}.
\end{aligned} \tag{21}
$$

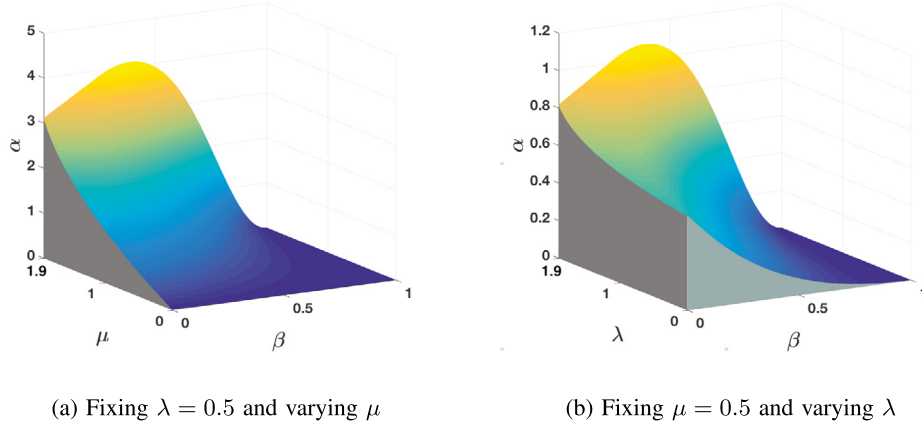(a) Fixing $\lambda = 0.5$ and varying μ (b) Fixing $\mu = 0.5$ and varying λ

Fig. 2. Visualization of the convergence regions of HB when perturbing the RC parameters.

We omit the proof of Lemma 11 since it follows easily from Corollary 7 and some simple calculations to simplify (20).

By setting different β_1, β_2 in (21), we can obtain the convergence condition of a general AGD method using the KYP lemma. In the following, we focus on the two most important cases: HB and NAG and other cases can be discussed in a similar way. The stability regions of HB and NAG can be obtained by letting $\beta_2 = 0$ and $\beta_1 = \beta_2 = \beta$ in (21), respectively. Then we can obtain Theorems 12 and 13.

Theorem 12. *Let $x^\star \in \mathbb{R}^n$ be the global minimizer of the loss function $f(\cdot)$ which satisfies $RC(\mu, \lambda)$. For any step size $\alpha > 0$ and momentum parameter $\beta \in (0, 1)$ lying in the region:*

$$\left\{ (\alpha, \beta) : H_1(\beta) \leq \alpha \leq \frac{2(\beta + 1)(1 - \sqrt{1 - \mu\lambda})}{\lambda} \right\}$$
$$\cup \left\{ (\alpha, \beta) : 0 < \alpha \leq \min\{H_1(\beta), H_2(\beta)\} \right\}.$$

where $H_1(\beta) = \frac{\mu\beta^2 + 6\mu\beta + \mu}{\beta + 1}$ and

$$H_2(\beta) = \frac{P_2(\beta) - \sqrt{P_2(\beta)^2 - 4P_1(\beta)P_3(\beta)}}{2P_1(\beta)}$$

with $P_1(\beta) = 4\mu\lambda\beta - \beta^2 - 1 - 2\beta$, $P_2(\beta) = 2\mu\beta + 2\mu\beta^2 - 2\mu\beta^3 - 2\mu$, and $P_3(\beta) = 4\mu^2\beta^3 + 4\mu^2\beta - 6\mu^2\beta^2 - \mu^2\beta^4 - \mu^2$, the iterates z_k generated by HB (4) converge linearly to $x^\star$ as $k \to \infty$.

Theorem 13. *Let $x^\star \in \mathbb{R}^n$ be bluethe global minimizer of the loss function $f(\cdot)$ which satisfies $RC(\mu, \lambda)$. For any step size $\alpha > 0$ and momentum parameter $\beta \in (0, 1)$ lying in the region:*

$$\left\{ (\alpha, \beta) : N_1(\beta) \leq \alpha < \frac{2(\beta + 1)(1 - \sqrt{1 - \mu\lambda})}{\lambda(1 + 2\beta)} \right\}$$
$$\cup \left\{ (\alpha, \beta) : 0 < \alpha \leq \min\{N_1(\beta), N_2(\beta)\} \right\}.$$

where

$$N_1(\beta) = \frac{Q_1(\beta) - \sqrt{Q_1(\beta)^2 - (1 + 6\beta + \beta^2)Q_2(\beta)}}{2\lambda\beta(\beta + 1)},$$

$$N_2(\beta) = \left\{ \beta : \frac{Q_3(\beta) - \sqrt{Q_3(\beta)^2 - (1 - \beta)^2 Q_2(\beta)}}{2\lambda\beta(\beta + 1)} \leq \alpha, \right.$$

$$\left. g\left(\frac{(\mu - \alpha)(1 + \beta)^2 + (\lambda\alpha^2 - \alpha)(\beta + \beta^2)}{4\mu\beta - 4\alpha\beta} \right) = 0 \right\},$$

$Q_1(\beta) = 1 + 7\beta + 2\beta^2$, $Q_2(\beta) = 4\mu\lambda\beta(1 + \beta)$, $Q_3(\beta) = 1 - \beta + 2\beta^2$, the iterates z_k generated by NAG (3) converge linearly to $x^\star$ as $k \to \infty$.

Remark 14. The bound $N_2(\beta)$ is an implicit function of β. It is hard to derive an explicit expression since $g(\cdot) = 0$ is a 4th-order equation of β. The function $g(\eta)$ is:

$$g(\eta) := 4\mu\beta\eta^2 - 2(2\mu\beta + \mu\beta^2 - \alpha\beta + \mu - \alpha)\eta$$
$$+ 2\mu + 2\mu\beta^2 - 2\alpha - 2\alpha\beta + \lambda\alpha^2.$$

The analytical results stated in the above theorems can provide rich insights on the convergence behaviors of AGD. Take the convergence region of HB as an example. In Fig. 2(a), we fix the RC parameter $\lambda = 0.5$ and vary μ within [0.01, 1.9], while in Fig. 2(b), we fix $\mu = 0.5$ and vary the value of λ within [0.01, 1.9]. Observe that when we fix one of the RC parameters and increase the other, the stability region of (α, β) gets larger.

Notice that μ plays a role similar to the inverse of the smoothness parameter, and therefore, it dominantly determines the step size, which is clearly demonstrated in Fig. 2(a). In addition, when we fix the values of a pair of (μ, λ) (e.g. Fig. 3), we can see that even when α exceeds the value of the bound of GD (the maximal feasible α when $\beta = 0$), the Heavy-ball method can still ensure convergence when we choose β properly. This property has not been discussed in the literature.

We emphasize that our theoretic analysis is a complement rather than a replacement for the numerical LMI approach in Lessard et al. (2016). Our closed-form expressions for the stability region do provide some complementary benefits to the numerical approach in Lessard et al. (2016). First, from our closed-form formulas, one can tell that the stability region of HB is well described by the feasible set of some relatively simple quadratic inequalities while the characterization of the stability region boundary of NAG partially involves a fourth-order polynomial. Such a difference is not directly reflected by the numerical LMI approach in Lessard et al. (2016). Actually our closed-form expression for the stability region of HB is quite simple. Our study on HB and NAG just illustrates that the interpretability of the analytical formulas for the stability region depends on the specific momentum method being analyzed. Second, the stability region is easier and faster to visualize from analytical forms than numerical results. When given a pair of (μ, λ), one needs to make a small grid of $(\alpha, \beta_1, \beta_2)$ and solve an LMI for each single pair, which is computationally complex but can be avoided if we have closed-form analytical results. More importantly, the LMI conditions in Lessard et al. (2016) can only certify convergence numerically for fixed $(\alpha, \beta_1, \beta_2)$ under a given pair of RC parameter (μ, λ). However, our analytical results provide continuous stability regions with respect to (μ, λ), which is hard to achieve using numerical results.
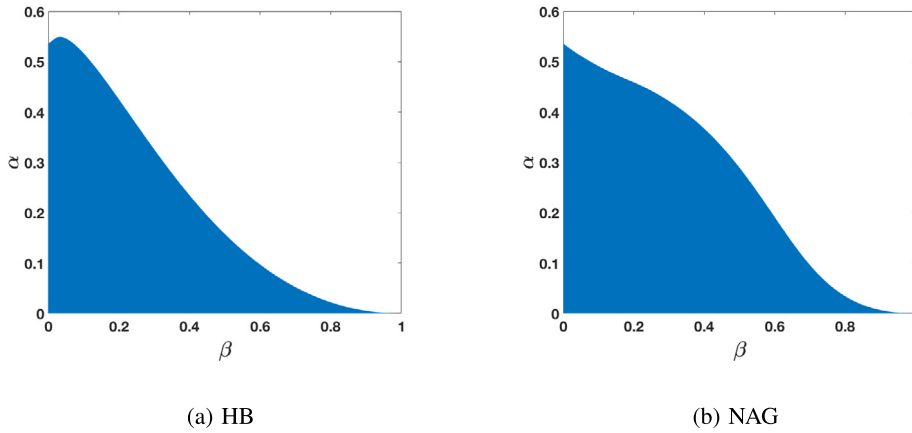
(a) HB   (b) NAG

Fig. 3. Visualization of the convergence regions of two AGD methods taking RC parameters as $\mu = 0.5, \lambda = 0.5$.

4.4. Numerical example

In this subsection, we will use a simple example satisfying RC globally to show how our results help to choose parameters of different first-order methods in practice.

Consider a loss function as shown in Fig. 4(a) with an expression as

$$f(x) = \begin{cases} x^2, & x \in [-6, 6] \\ x^2 + 1.5|x| \left(\cos(|x| - 6) - 1 \right), & \text{otherwise.} \end{cases}$$

This nonconvex loss function was also discussed in Chen and Candès (2017) and Chi et al. (2019). One can check that f satisfies $RC(0.5, 0.5)$.

We initialize at $x_0 = x_1 = 24$ and choose $\alpha = 0.1$. By Theorem 12, HB can converge when $\beta < 0.5942$. For NAG, we choose $\beta < 0.6950$ according to Theorem 13. Furthermore, it is common to choose the hyper-parameter β as large as possible to obtain a better performance. Therefore, the corresponding β's to HB and NAG are chosen as 0.59 and 0.69, respectively. In Fig. 4(b), we see that all the three algorithms can converge and the two accelerated methods HB and NAG obviously outperform GD.

5. Local regularity condition

So far, all the above derivations assume RC holds globally. In addition, the existing control framework for optimization methods (Cyrus et al., 2018; Dhingra et al., 2019; Fazlyab et al., 2018; Hu & Lessard, 2017b; Hu, Seiler, & Rantzer, 2017; Lessard et al., 2016; Van Scoy et al., 2018) all require global constraints. In certain problems, however, RC may only hold locally around the fixed point. In this section, we explain how our framework can accommodate such cases, as long as the algorithms are initialized properly as stated in the following theorem whose proof can be found in the Appendix.

Theorem 15. *Let $x^\star \in \mathbb{R}^n$ be a local minimizer of the loss function $f(\cdot)$ which satisfies $RC(\mu, \lambda, \epsilon)$ with some positive constants μ, λ, ϵ. Assume that $P \succ 0$ is a feasible solution of the LMI (9). If the first two iterates initialized properly according to $z_{-1}, z_0 \in \mathcal{N}_{\epsilon/\sqrt{10\text{cond}(P)}}(x^\star)$, then $y_k \in \mathcal{N}_\epsilon(x^\star), \forall k$.*

Theorem 15 ensures all the following iterates will not exceed the local neighborhood satisfying RC since $y_k \in \mathcal{N}_\epsilon(x^\star)$ for all k, so that we can still transfer RC to a quadratic bound at each iteration, and thus all the previous results still hold for the convergence analysis of AGD under a general setting where RC only holds locally.

In practice, spectral methods can be used as an initialization scheme to locate an initial estimate in a desired neighborhood of the fixed point. For example, we consider a popular inverse problem in signal processing called phase retrieval, where the goal is to recover a signal $x^\star \in \mathbb{R}^n$ from the magnitudes of its linear measurements, $y_r = |a_r^T x^\star|^2$, $r = 1, \ldots, m$, where $a_r \in \mathbb{R}^n$ is the rth sampling vector, and m is the number of samples. If a_r's are drawn with i.i.d. standard Gaussian entries, it is shown that the loss function $f(z) = \frac{1}{2m} \sum_{r=1}^m \left(y_r - |a_r^T z|^2 \right)^2$ satisfies RC locally in $\mathcal{N}_\epsilon(x^\star)$ (ignoring the sign ambiguity in identifying $x^\star$), where ϵ is a small constant (e.g. 1/10) (Candes et al., 2015). On the other end, the spectral method proposed in Chen and Candès (2017) returns an initial estimate $\mathcal{N}_\epsilon(x^\star)$ as soon as the sample complexity m is above the order of $O(n/\epsilon)$. Therefore, as long as $m = O(n)$, the spectral method can successfully land an initialization in the region satisfying RC. In addition, the quality of the initialization also impacts the iteration complexity logarithmically as suggested by (10). We refer the readers to Candes et al. (2015) and Keshavan, Montanari, and Oh (2010) for more details of initialization techniques.

6. Conclusions

In this paper, we apply control tools to analyze the convergence of AGD (including HB and NAG) under the Regularity Condition. Our main contribution lies in the *analytical* characterization of the convergence regions in terms of the algorithm parameters (α, β) and the RC parameters (μ, λ). Such convergence results do not exist in the current literature and offer useful insights in the analysis and design of AGD for a class of nonconvex optimization problems.

Appendix A. RC vs sector bound

Recall the quadratic bound for $RC(\mu, \lambda)$ has the following form:

$$\begin{bmatrix} y_k - y_* \\ u_k - u_* \end{bmatrix}^T \begin{bmatrix} -\lambda I_n & I_n \\ I_n & -\mu I_n \end{bmatrix} \begin{bmatrix} y_k - y_* \\ u_k - u_* \end{bmatrix} \geq 0. \tag{A.1}$$

The sector bound condition in Lessard et al. (2016, Lemma 6) is described by the following quadratic constraint:

$$\begin{bmatrix} y_k - y_* \\ u_k - u_* \end{bmatrix}^T \begin{bmatrix} -2mLI_n & (L+m)I_n \\ (L+m)I_n & -2I_n \end{bmatrix} \begin{bmatrix} y_k - y_* \\ u_k - u_* \end{bmatrix} \geq 0, \tag{A.2}$$

where m and L are the slopes of the lines forming the sector. For simplicity, assume $m \leq L$. By comparing (A.1) and (A.2), we can

(a) Loss function f
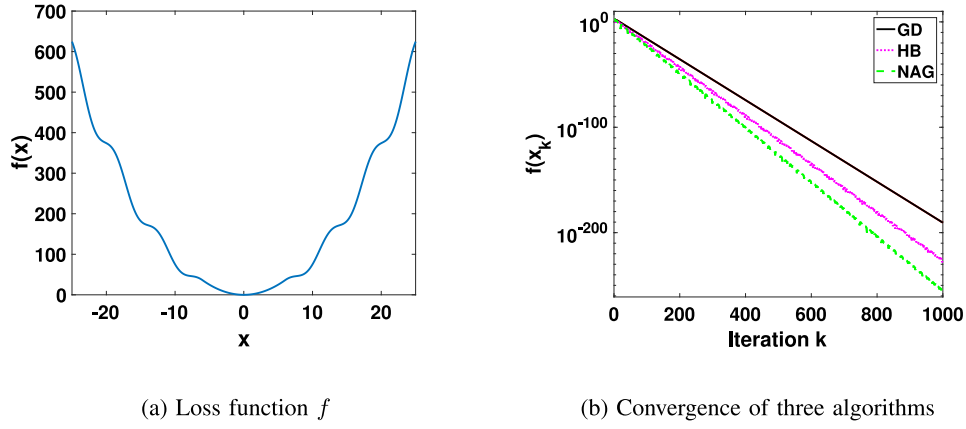

(b) Convergence of three algorithms

Fig. 4. An example satisfying RC and numerical experiments.

find that the quadratic bounds corresponding to RC and the sector bound are essentially the same. Specifically, given the sector bound (A.2), we can set $\lambda = \frac{2mL}{m+L}$ and $\mu = \frac{2}{m+L}$, which leads to RC in (A.1). Similarly, given RC in (A.1), we can immediately obtain an equivalent sector bound condition by setting $m = \frac{1-\sqrt{1-\lambda\mu}}{\mu}$ and $L = \frac{1+\sqrt{1-\lambda\mu}}{\mu}$. One special situation is when RC holds only locally, as is the case in most applications, which is handled carefully in this paper.

Appendix B. Proof details

B.1. Proof of Proposition 2

Assume the key LMI holds with some $P \succ 0$, i.e.,

$$
\begin{bmatrix} A^T P A - \rho^2 P & A^T P B \\ B^T P A & B^T P B \end{bmatrix} +
$$
$$
\begin{bmatrix} C & D \\ \mathbf{0}_{1\times 2} & 1 \end{bmatrix}^T \begin{bmatrix} -\lambda & 1 \\ 1 & -\mu \end{bmatrix} \begin{bmatrix} C & D \\ \mathbf{0}_{1\times 2} & 1 \end{bmatrix} \preceq 0.
\tag{B.1}
$$

Multiplying (B.1) by $\begin{bmatrix} \phi_k - \phi_* \\ u_k - u_* \end{bmatrix}^T$ from the left and $\begin{bmatrix} \phi_k - \phi_* \\ u_k - u_* \end{bmatrix}$ from the right respectively, and inserting the Kronecker product, we obtain

$$
\begin{bmatrix} \phi_k - \phi_* \\ u_k - u_* \end{bmatrix}^T \left(\begin{bmatrix} A^T P A - \rho^2 P & A^T P B \\ B^T P A & B^T P B \end{bmatrix} \otimes I_n \right) \begin{bmatrix} \phi_k - \phi_* \\ u_k - u_* \end{bmatrix}
$$
$$
+ \begin{bmatrix} y_k - y_* \\ u_k - u_* \end{bmatrix}^T \begin{bmatrix} -\lambda I_n & I_n \\ I_n & -\mu I_n \end{bmatrix} \begin{bmatrix} y_k - y_* \\ u_k - u_* \end{bmatrix} \leq 0.
$$

We have that RC can be equivalently represented as a quadratic bound of the feedback term $u_k = \nabla f(y_k)$ as:

$$
\begin{bmatrix} y_k - y_* \\ u_k - u_* \end{bmatrix}^T \begin{bmatrix} -\lambda I_n & I_n \\ I_n & -\mu I_n \end{bmatrix} \begin{bmatrix} y_k - y_* \\ u_k - u_* \end{bmatrix} \geq 0,
\tag{B.2}
$$

which further implies that

$$
\begin{bmatrix} \phi_k - \phi_* \\ u_k - u_* \end{bmatrix}^T \left(\begin{bmatrix} A^T P A - \rho^2 P & A^T P B \\ B^T P A & B^T P B \end{bmatrix} \otimes I_n \right)
$$
$$
\cdot \begin{bmatrix} \phi_k - \phi_* \\ u_k - u_* \end{bmatrix} \leq 0.
\tag{B.3}
$$

Observe $\phi_{k+1} = (A \otimes I_n)\phi_k + (B \otimes I_n)u_k$. Hence we can further rearrange and simplify (B.3) as

$$
(\phi_{k+1} - \phi_*)^T (P \otimes I_n)(\phi_{k+1} - \phi_*)
$$
$$
\leq \rho^2 (\phi_k - \phi_*)^T (P \otimes I_n)(\phi_k - \phi_*).
$$

Such exponential decay with $P \succ 0$ for all k can conclude Proposition 2.

B.2. Proof of Lemma 8

We check the inner product of the input u_k and output y_k (recall that $y_* = x^\star$) of the dynamical system (16):

$$
\langle u_k - u_*, y_k - y_* \rangle
$$
$$
= \langle -\alpha \nabla f(y_k) - \delta(y_k - y_*), y_k - y_* \rangle
$$
$$
= -\delta \|y_k - y_*\|^2 - \alpha \langle \nabla f(y_k), y_k - y_* \rangle
$$
$$
\leq -\delta \|y_k - y_*\|^2 - \frac{\alpha\mu}{2}\|\nabla f(y_k)\|^2 - \frac{\alpha\lambda}{2}\|y_k - y_*\|^2
$$
$$
= -\frac{\mu}{2\alpha}\|u_k - u_*\|^2 - \frac{\delta\mu}{\alpha}\langle u_k - u_*, y_k - y_* \rangle
$$
$$
- \left(\delta + \frac{\alpha\lambda}{2} + \frac{\delta^2\mu}{2\alpha} \right) \|y_k - y_*\|^2.
$$

By rearrangement, we have

$$
- \left(2\alpha\delta + \lambda\alpha^2 + \mu\delta^2 \right) \|y_k - y_*\|^2
$$
$$
- 2(\alpha + \mu\delta)\langle u_k - u_*, y_k - y_* \rangle - \mu\|u_k - u_*\|^2 \geq 0,
$$

and thus conclude (17).

B.3. Proof of Lemma 9

We check the KYPC(A', M') as listed for AGD (16) characterized by α, β_1, β_2:

(1) $\det(e^{j\omega}I - A') \neq 0$ for $\omega \in \mathbb{R}$;
(2) A' is Schur stable;
(3) The left upper corner of M' in (17) is PSD.

Condition 1. Write

$$
\det(e^{j\omega}I - A') = \begin{vmatrix} e^{j\omega} - (1 + \beta_1 + \delta(1 + \beta_2)) & \beta_1 + \delta\beta_2 \\ -1 & e^{j\omega} \end{vmatrix}
$$
$$
= \cos^2\omega - \sin^2\omega - (1 + \beta_1 + \delta(1 + \beta_2))\cos\omega + \beta_1
$$
$$
+ \delta\beta_2 + j(2\sin\omega\cos\omega - (1 + \beta_1 + \delta(1 + \beta_2))\sin\omega).
$$

By means of the opposite direction, let $\det(e^{j\omega}I - A') = 0$. Then we have

$$
\begin{cases} \cos^2\omega - \sin^2\omega - (1 + \beta_1 + \delta(1 + \beta_2))\cos\omega \\ \qquad + \beta_1 + \delta\beta_2 = 0, \\ 2\sin\omega\cos\omega - (1 + \beta_1 + \delta(1 + \beta_2))\sin\omega = 0. \end{cases}
$$

From the second equality, we have $\sin\omega = 0$ or $(1 + \beta_1 + \delta(1 + \beta_2)) = 2\cos\omega$, which we discuss separately. (a) If $\sin\omega$

$= 0$, $\cos\omega = \pm 1$. Then the first equality becomes $(1 + \beta_1 + \delta$ $(1 + \beta_2)) = \pm(1 + \beta_1 + \delta\beta_2)$; (b) if $(1 + \beta_1 + \delta(1 + \beta_2)) = 2\cos\omega$, then from the first equality we need $\cos^2\omega - \sin^2\omega - (1 + \beta_1 + \delta(1 + \beta_2))\cos\omega + \beta_1 + \delta\beta_2 = -1 + \beta_1 + \delta\beta_2 = 0$.

To conclude, Condition 1 is satisfied if and only if:

$$(1+\beta_1+\delta(1+\beta_2)) \neq \pm(1+\beta_1+\delta\beta_2), \beta_1+\delta\beta_2 \neq 1. \tag{B.4}$$

Condition 2. We want

$$A' = \begin{bmatrix} (1 + \beta_1 + \delta(1 + \beta_2)) & -(\beta_1 + \delta\beta_2) \\ 1 & 0 \end{bmatrix}$$

to be Schur stable, for which it suffices to check its eigenvalues are bounded by 1 in magnitude. We start by writing out the characteristic equation of A':

$$|\lambda I - A'| = \begin{vmatrix} \lambda - (1 + \beta_1 + \delta(1 + \beta_2)) & \beta_1 + \delta\beta_2 \\ -1 & \lambda \end{vmatrix}$$
$$= \lambda^2 - \lambda(1 + \beta_1 + \delta(1 + \beta_2)) + \beta_1 + \delta\beta_2 = 0.$$

The eigenvalues of A' are given as the two roots of the above polynomial:

$$\lambda_{1,2} = \frac{(1 + \beta_1 + \delta(1 + \beta_2))}{2}$$
$$\pm \frac{\sqrt{(1 + \beta_1 + \delta(1 + \beta_2))^2 - 4(\beta_1 + \delta\beta_2)}}{2}.$$

We need to make sure $|\lambda_{1,2}| < 1$. When the eigenvalues are complex-valued, i.e. $(1 + \beta_1 + \delta(1 + \beta_2))^2 < 4(\beta_1 + \delta\beta_2)$, $|\lambda_1| = |\lambda_2| = \beta_1 + \delta\beta_2$, and thus we need $\beta_1 + \delta\beta_2 < 1$ in this case. When the eigenvalues are real-valued, i.e. $(1 + \beta_1 + \delta(1 + \beta_2))^2 \geq 4(\beta_1 + \delta\beta_2)$, we have

$$4(\beta_1 + \delta\beta_2) \leq (1 + \beta_1 + \delta(1 + \beta_2))^2 < (1 + \beta_1 + \delta\beta_2)^2.$$

To conclude, Condition 2 is satisfied if and only if

$$-\frac{2(1 + \beta_1)}{1 + 2\beta_2} < \delta < 0. \tag{B.5}$$

Condition 3. We want the left upper corner of M' in (17) to be PSD, which means

$$-\left(2\alpha\delta + \lambda\alpha^2 + \mu\delta^2\right) \geq 0.$$

Then Condition 3 is satisfied if and only if

$$\frac{-\delta - |\delta|\sqrt{1 - \mu\lambda}}{\lambda} \leq \alpha \leq \frac{-\delta + |\delta|\sqrt{1 - \mu\lambda}}{\lambda}. \tag{B.6}$$

To conclude, by unifying all conditions (B.4) (B.5) (B.6), we conclude that the parameters of AGD need to satisfy

$$0 < \alpha < \frac{2(1 + \beta_1)(1 + \sqrt{1 - \mu\lambda})}{\lambda(1 + 2\beta_2)}. \tag{B.7}$$

B.4. Proof of Theorem 12

For HB, we set $\beta_1 = \beta$, $\beta_2 = 0$ and denote $u := \cos\omega$, then (21) can be rewritten as:

$$h(u) := 4\mu\beta u^2 - 2(2\mu\beta + \mu\beta^2 - \alpha\beta + \mu - \alpha)u$$
$$+ 2\mu + 2\mu\beta^2 - 2\alpha - 2\alpha\beta + \lambda\alpha^2 \geq 0, \quad \forall u \in [-1, 1]. \tag{B.8}$$

Observe that $h(u)$ is a quadratic function depending on u. We can check the minimal value of $h(\cdot)$ on $[-1, 1]$ by discussing its axis of symmetry denoted as $S = \frac{2\mu\beta + \mu\beta^2 - \alpha\beta + \mu - \alpha}{4\mu\beta}$.

(1) When $S \geq 1$, $\alpha \leq \frac{\mu(1-\beta)^2}{1+\beta}$. Then

$$h(u)_{\min} = h(1) = \lambda\alpha^2 > 0.$$

Thus the feasible region in this case is:

$$\left\{(\alpha, \beta) : \alpha \leq \frac{\mu(\beta - 1)^2}{\beta + 1}, 0 < \beta < 1\right\}. \tag{B.9}$$

(2) When $S \leq -1$, $\alpha \geq \frac{\mu\beta^2 + 6\mu\beta + \mu}{1 + \beta}$. We need

$$h(u)_{\min} = h(-1) = \lambda\alpha^2 - 4(1 + \beta)\alpha + 4\mu(1 + \beta)^2 \geq 0.$$

Thus the feasible region in this case is:

$$\left\{(\alpha, \beta) : \alpha \geq \frac{2(\beta+1)(1+\sqrt{1-\mu\lambda})}{\lambda}, 0 < \beta < 1\right\}$$
$$\cup \left\{(\alpha, \beta) : \frac{\mu\beta^2 + 6\mu\beta + \mu}{\beta + 1} \leq \alpha \right. \tag{B.10}$$
$$\left. \leq \frac{2(\beta + 1)(1 - \sqrt{1 - \mu\lambda})}{\lambda}, 0 < \beta < 1\right\}.$$

(3) When $-1 < S < 1$, $\frac{\mu\beta^2 + 6\mu\beta + \mu}{1 + \beta} < \alpha < \frac{\mu(1-\beta)^2}{1+\beta}$. We want $h(u)_{\min} = h\left(\frac{2\mu\beta + \mu\beta^2 - \alpha\beta + \mu - \alpha}{4\mu\beta}\right) \geq 0$. It is equivalent to solve

$$(4\mu\lambda\beta - \beta^2 - 1 - 2\beta)\alpha^2 - \left(2\mu\beta + 2\mu\beta^2 - 2\mu\beta^3 \right.$$
$$\left. - 2\mu\right)\alpha + 4\mu^2\beta^3 + 4\mu^2\beta - 6\mu^2\beta^2 - \mu^2\beta^4 - \mu^2 \geq 0.$$

Since $4\mu\lambda\beta - \beta^2 - 1 - 2\beta < 4\beta - \beta^2 - 1 - 2\beta \leq 0$, Thus the feasible region in this case is:

$$\left\{(\alpha, \beta) : \frac{\mu(\beta - 1)^2}{\beta + 1} \leq \alpha \leq R, 0 < \beta < 1\right\}$$
$$\cap \left\{(\alpha, \beta) : \alpha \leq \frac{\mu\beta^2 + 6\mu\beta + \mu}{\beta + 1}, 0 < \beta < 1\right\}, \tag{B.11}$$

where $R = \frac{P_2 - \sqrt{P_2^2 - 4P_1 P_3}}{2P_1}$, $P_1 = 4\mu\lambda\beta - \beta^2 - 1 - 2\beta$, $P_2 = 2\mu\beta + 2\mu\beta^2 - 2\mu\beta^3 - 2\mu$, $P_3 = 4\mu^2\beta^3 + 4\mu^2\beta - 6\mu^2\beta^2 - \mu^2\beta^4 - \mu^2$.

Taking the union of (B.9) (B.10) (B.11) gives the result of the FDI (B.8). Further intersecting with the condition (18) obtained in Lemma 9 leads to the final region.

B.5. Proof of Theorem 13

Similar with the proof of Theorem 12, for NAG, we set $\beta_1 = \beta_2 = \beta$ and denote $u := \cos\omega$, then (21) can be rewritten as:

$$h(u) := (-4\mu\beta + 4\alpha\beta)u^2$$
$$+ \left[2(\mu - \alpha)(1 + \beta)^2 + 2(\lambda\alpha - 1)\alpha(1 + \beta)\beta\right]u$$
$$- 2\mu(1 + \beta^2) + 2\alpha(1 + \beta)^2 - \lambda\alpha^2\beta^2 \tag{B.12}$$
$$- 2\alpha(1 - \beta)\beta - \lambda\alpha^2(1 + \beta)^2 \leq 0, \quad \forall u \in [-1, 1].$$

We check the maximal value of the quadratic function $h(\cdot)$ on $[-1, 1]$.

(1) When $\alpha = \mu$, $h(u)_{\max} = f(-1) \leq 0$. Thus the feasible region in this case is:

$$\left\{(\alpha, \beta) : \alpha = \mu, 0 < \beta \leq \frac{-1 + \mu\lambda + \sqrt{1 - \mu\lambda}}{2(1 - \mu\lambda)}\right\}. \tag{B.13}$$

(2) When $\alpha > \mu$, we need to let $h(1) \leq 0$, $h(-1) \leq 0$. Since $h(1) = -\lambda\alpha^2 < 0$, we only need to check $h(-1)$. Thus the feasible region in this case is:

$$\{(\alpha, \beta) : \alpha \geq L_1, 0 < \beta < 1\} \cup$$
$$\{(\alpha, \beta) : \mu < \alpha \leq R_1, 0 < \beta < 1\}, \tag{B.14}$$

where $L_1 = \frac{2(\beta+1)(1+\sqrt{1-\mu\lambda})}{\lambda(1+2\beta)}$, $R_1 = \frac{2(\beta+1)(1-\sqrt{1-\mu\lambda})}{\lambda(1+2\beta)}$.

(3) When $\alpha < \mu$, we can check the maximal value of the quadratic function $h(u)$ by discussing the axis of symmetry $S = \frac{(\mu-\alpha)(1+\beta)^2+(\lambda\alpha-1)\alpha(1+\beta)\beta}{4\mu\beta-4\alpha\beta}$.

(a) When $S \geq 1$, $h(u)_{max} = -\lambda\alpha^2 < 0$. Thus the feasible region in this case is:

$$\left\{ (\alpha,\beta): \alpha \leq \frac{B_1-\sqrt{B_1^2-C_1}}{2\lambda\beta(\beta+1)}, 0<\beta<1 \right\}, \qquad (B.15)$$

where $B_1 = 1-\beta+2\beta^2$, $C_1 = 4\mu\lambda\beta(1+\beta)(1-\beta)^2$.

(b) When $S \leq -1$, $h(u)_{max} = h(-1) \leq 0$. Thus the feasible region in this case is:

$$\left\{ (\alpha,\beta): \frac{B_2-\sqrt{B_2^2-C_2}}{2\lambda\beta(\beta+1)} \leq \alpha < \mu, 0<\beta<1 \right\}, \quad (B.16)$$

where $B2 = 1+7\beta+2\beta^2$, $C_2 = 4\mu\lambda\beta(1+\beta)(1+6\beta+\beta^2)$.

(c) When $-1 < S < 1$, $h(u)_{max} = h(S) \leq 0$. Thus the feasible region in this case is:

$$\{(\alpha,\beta): L<\alpha<R, 0<\beta<1, g(S)\leq 0\}, \qquad (B.17)$$

where $L = \frac{1-\beta+2\beta^2-\sqrt{(1-\beta+2\beta^2)^2-4\mu\lambda\beta(1+\beta)(1-\beta)^2}}{2\lambda\beta(\beta+1)}$, $R = \frac{1+7\beta+2\beta^2-\sqrt{(1+7\beta+2\beta^2)^2-4\mu\lambda\beta(1+\beta)(1+6\beta+\beta^2)}}{2\lambda\beta(\beta+1)}$. and $g(S)$ is noted in Remark 14, that is,

$$g(S) := 4\mu\beta S^2 - 2(2\mu\beta + \mu\beta^2 - \alpha\beta + \mu - \alpha)S + 2\mu + 2\mu\beta^2 - 2\alpha - 2\alpha\beta + \lambda\alpha^2.$$

The result of the FDI (B.12) is the union of all the above regions (B.13) (B.14) (B.15) (B.16) (B.17).

Further intersecting with the condition (18) obtained in Lemma 9 leads to the final region.

B.6. Proof of Theorem 15

Since $z_{-1}, z_0 \in \mathcal{N}_{\epsilon/\sqrt{10\text{cond}(P)}}(x^\star)$, we have $\|\phi_0-\phi_*\| < \frac{\epsilon}{\sqrt{5\text{cond}(P)}}$. The exponential decay with $P \succ 0$: $(\phi_{k+1} - \phi_*)^T P(\phi_{k+1} - \phi_*) \leq \rho^2(\phi_k-\phi_*)^T P(\phi_k-\phi_*)$ implies that $\|\phi_k-\phi_*\| \leq \sqrt{\text{cond}(P)}\rho^k\|\phi_0-\phi_*\|$, which we have argued in Section 3.2. Therefore,

$$\begin{aligned}
\|\phi_k - \phi_*\| &\leq \sqrt{\text{cond}(P)}\rho^k\|\phi_0 - \phi_*\| \\
&< \sqrt{\text{cond}(P)} \cdot \|\phi_0 - \phi_*\| \\
&< \sqrt{\text{cond}(P)} \cdot \frac{\epsilon}{\sqrt{5\text{cond}(P)}} \\
&< \epsilon/\sqrt{5}.
\end{aligned}$$

As a consequence,

$$\begin{aligned}
\|y_k - x^\star\| &= \left\| (1+\beta_2)z_k^{(1)} - \beta_2 z_k^{(2)} - x^\star \right\| \\
&= \|C(\phi_k - \phi_*)\| \\
&\leq \|C^T\| \|\phi_k - \phi_*\| \\
&< \sqrt{(1+\beta_2)^2 + \beta_2^2} \cdot \left(\epsilon/\sqrt{5}\right) \\
&< \epsilon,
\end{aligned}$$

where we recall $C = [1+\beta_2, -\beta_2]$, and the last line used $\beta_2 < 1$.
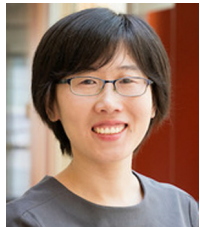
References

Aybat, Necdet Serhat, Fallah, Alireza, Gurbuzbalaban, Mert, & Ozdaglar, Asuman (2018). Robust accelerated gradient methods for smooth strongly convex functions. arXiv preprint arXiv:1805.10579.

Candes, Emmanuel J., Li, Xiaodong, & Soltanolkotabi, Mahdi (2015). Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4), 1985–2007.

Chen, Yuxin, & Candès, Emmanuel J. (2017). Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on Pure and Applied Mathematics*, 70(5), 822–883,.

Cherukuri, Ashish, Mallada, Enrique, Low, Steven, & Cortés, Jorge (2017). The role of convexity in saddle-point dynamics: Lyapunov function and robustness. *IEEE Transactions on Automatic Control*, 63(8), 2449–2464.

Chi, Yuejie, Lu, Yue M., & Chen, Yuxin (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20), 5239–5269.

Cyrus, Saman, Hu, Bin, Van Scoy, Bryan, & Lessard, Laurent (2018). A robust accelerated optimization algorithm for strongly convex functions. In *2018 annual American control conference, ACC* (pp. 1376–1381). IEEE.

Dhingra, Neil K., Khong, Sei Zhen, & Jovanovic, Mihailo R. (2019). The proximal augmented Lagrangian method for nonsmooth composite optimization. *IEEE Transactions on Automatic Control*, 64(7), 2861–2868.

Fazlyab, Mahyar, Ribeiro, Alejandro, Morari, Manfred, & Preciado, Victor M. (2018). Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. *SIAM Journal on Optimization*, 28(3), 2654–2689.

Hu, Bin, & Lessard, Laurent (2017a). Control interpretations for first-order optimization methods. In *2017 American control conference, ACC* (pp. 3114–3119). IEEE.

Hu, Bin, & Lessard, Laurent (2017b). Dissipativity theory for Nesterov's accelerated method. In *Proceedings of the 34th international conference on machine learning (vol. 70)* (pp. 1549–1557).

Hu, Bin, Seiler, Peter, & Lessard, Laurent (2017). Analysis of approximate stochastic gradient using quadratic constraints and sequential semidefinite programs. arXiv preprint arXiv:1711.00987.

Hu, Bin, Seiler, Peter, & Rantzer, Anders (2017). A unified analysis of stochastic optimization methods using jump system theory and quadratic constraints. In *Proceedings of the 2017 conference on learning theory (vol. 65)* (pp. 1157–1189).

Hu, Bin, Wright, Stephen, & Lessard, Laurent (2018). Dissipativity theory for accelerating stochastic variance reduction: A unified analysis of SVRG and Katyusha using semidefinite programs. In *Proceedings of the 35th international conference on machine learning (vol. 80)* (pp. 2038–2047).

Keshavan, Raghunandan H., Montanari, Andrea, & Oh, Sewoong (2010). Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6), 2980–2998.

Kolarijani, Arman Sharifi, Esfahani, Peyman Mohajerin, & Keviczky, Tamas (2018). Fast gradient-based methods with exponential rate: A hybrid control framework. In *Proceedings of the 35th international conference on machine learning (vol. 80)* (pp. 2728–2736).

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

Li, Yuanxin, Chi, Yuejie, Zhang, Huishuai, & Liang, Yingbin (2019). Non-convex low-rank matrix recovery with arbitrary outliers via median-truncated gradient descent. *Information and Inference: A Journal of the IMA*.

Li, Yuanzhi, & Yuan, Yang (2017). Convergence analysis of two-layer neural networks with ReLU activation. In *Advances in neural information processing systems* (pp. 597–607).

Nesterov, Yurii (2003). *Introductory lectures on convex optimization: A basic course (vol. 87)*. Springer Science & Business Media.

Nishihara, Robert, Lessard, Laurent, Recht, Benjamin, Packard, Andrew, & Jordan, Michael I. (2015). A general analysis of the convergence of ADMM. In *proceedings of the 32nd international conference on international conference on machine learning (vol. 37)* (pp. 343–352).

Pauwels, Edouard Jean Robert, Beck, Amir, Eldar, Yonina C, & Sabach, Shoham (2017). On fienup methods for sparse phase retrieval. *IEEE Transactions on Signal Processing*, 66(4), 982–991.

Polyak, Boris T. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5), 1–17.

Rantzer, Anders (1996). On the Kalman–Yakubovich–Popov lemma. *Systems & Control Letters*, 28(1), 7–10.

Sundararajan, Akhil, Hu, Bin, & Lessard, Laurent (2017). Robust convergence analysis of distributed optimization algorithms. In *2017 55th annual allerton conference on communication, control, and computing, Allerton* (pp. 1206–1212). IEEE.

Tu, Stephen, Boczar, Ross, Simchowitz, Max, Soltanolkotabi, Mahdi, & Recht, Ben (2016). Low-rank solutions of linear matrix equations via procrustes flow. In *International conference on machine learning* (pp. 964–973).

Van Scoy, Bryan, Freeman, Randy A., & Lynch, Kevin M. (2018). The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1), 49–54.

Wang, Gang, Giannakis, Georgios B., & Eldar, Yonina C. (2017). Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2), 773–794.

Wilson, Ashia C., Recht, Benjamin, & Jordan, Michael I. (2016). A Lyapunov analysis of momentum methods in optimization. arXiv preprint arXiv:1611.02635.

Zhang, Huishuai, Chi, Yuejie, & Liang, Yingbin (2016). Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow. *International conference on machine learning* (pp. 1022–1031).

Zhang, Huishuai, Liang, Yingbin, & Chi, Yuejie (2017). A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms. *Journal of Machine Learning Research (JMLR)*, 18(141), 1–35.

Zhou, Yi, & Liang, Yingbin (2017). Characterization of gradient dominance and regularity conditions for neural networks. arXiv preprint arXiv:1710.06910.

Huaqing Xiong received his B.E. in Control Science and Engineering from Zhejiang University in 2016, and is pursuing a Ph.D. degree in Electrical and Computer Engineering at the Ohio State University, where his research focuses on deep learning theory, reinforcement learning theory and nonconvex optimization.
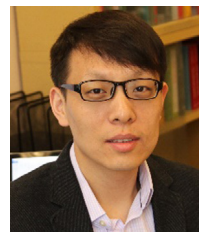
Yuejie Chi received the Ph.D. degree in Electrical Engineering from Princeton University in 2012, and the B.E. (Hon.) degree in Electrical Engineering from Tsinghua University, Beijing, China, in 2007. She was with The Ohio State University from 2012 to 2017. Since 2018, she is an Associate Professor with the department of Electrical and Computer Engineering at Carnegie Mellon University, where she holds the Robert E. Doherty Early Career Professorship. Her research interests include signal processing, statistical inference, machine learning, large-scale optimization, and their applications in data science, inverse problems, imaging, and sensing systems. She is a recipient of the PECASE Award, NSF CAREER Award, AFOSR and ONR Young Investigator Program Awards, Ralph E. Powe Junior Faculty Enhancement Award, Google Faculty Research Award, IEEE Signal Processing Society Young Author Best Paper Award and the Best Paper Award at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). She has served as an Elected Member of the SPTM, SAM and MLSP Technical Committees of the IEEE Signal Processing Society. She currently serves as an Associate Editor of IEEE Trans. on Signal Processing.

Bin Hu is currently an assistant professor in the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign and affiliated with the Coordinated Science Laboratory. He received the Ph.D. degree in Aerospace Engineering and Mechanics at the University of Minnesota in 2016. Between July 2016 and July 2018, he was a postdoctoral researcher in the Wisconsin Institute for Discovery at the University of Wisconsin-Madison. His current research focuses on the interplay between machine learning and control.

Wei Zhang is currently a Professor in the Department of Mechanical and Energy Engineering at the Southern University of Science and Technology (SUSTech), Shenzhen, China. He received the B.S. in Automatic Control from the University of Science and Technology of China in 2003, the M.S. in Electrical and Computer Engineering from the University of Kentucky in 2005, and the M.S. in Statistics and the Ph.D. in Electrical Engineering both from Purdue University in 2009. From January 2010 to August 2011, he was a Postdoctoral Researcher in the EECS Department at the University of California, Berkeley. Before joining SUSTech, he served as an Assistant Professor and then an Associate Professor (with tenure) of Electrical and Computer Engineering at the Ohio State University. His research interests include hybrid systems, control and learning theory, game theory, stochastic systems, and their applications in robotics, power systems, and intelligent transportations. He is a recipient of the NSF CAREER award and the Lumley Research Award at the Ohio State University.