ISE

Industrial and
Systems Engineering

A Nonsmooth Dynamical Systems Perspective on Accelerated Extensions of ADMM

Guilherme França[1], Daniel P. Robinson[2], and René Vidal[1]

[1]Mathematical Institute for Data Science, Johns Hopkins University

[2]Department of Industrial and Systems Engineering, Lehigh University

ISE Technical Report 19T-021

LEHIGH
UNIVERSITY.

A Nonsmooth Dynamical Systems Perspective on Accelerated Extensions of ADMM

Guilherme França,[a*] Daniel P. Robinson,[b] and René Vidal[a]

[a] *Mathematical Institute for Data Science, Johns Hopkins University*
[b] *Department of Industrial and Systems Engineering, Lehigh University*

Abstract

The acceleration technique introduced by Nesterov for gradient descent is widely used in optimization but its principles are not yet fully understood. Recently, significant progress has been made to close this understanding gap through a continuous-time dynamical systems perspective associated with gradient methods for smooth and unconstrained problems. Here we extend this perspective to nonsmooth and linearly constrained problems by deriving nonsmooth dynamical systems related to variants of the relaxed and accelerated alternating direction method of multipliers (ADMM). More specifically, we introduce two new ADMM variants, one based on Nesterov's acceleration and the other inspired by the heavy ball method, and derive differential inclusions that model these algorithms in the continuous-time limit. Through a nonsmooth Lyapunov analysis, we obtain rates of convergence for these dynamical systems in the convex and strongly convex settings that illustrate an interesting tradeoff between Nesterov and heavy ball types of acceleration.

1 Introduction

A popular approach to accelerate the convergence of the gradient descent method was proposed by Nesterov [1]. For convex functions, accelerated gradient descent attains a convergence rate of $\mathcal{O}(1/k^2)$ in terms of the error in the objective function value, with k denoting the iteration number. This rate is known to be optimal in the sense of worst case complexity [2]. Another accelerated variant of gradient descent was introduced by Polyak [3], called the heavy ball method, which is known to have a convergence rate of $\mathcal{O}(1/k)$ for convex functions and an exponential rate (also called linear convergence in the optimization literature) for strongly convex functions [4, 5]. Nonetheless, the mechanisms by which optimization algorithms are accelerated is still not very well understood.

Recently, there has been progress in understanding acceleration by analyzing a differential equation modeling the continuous-time limit of Nesterov's method [6]. Follow-up work has brought a larger class of accelerated methods into a Hamiltonian formalism [7] thus giving opportunities for analysis through the lens of continuous dynamical systems. For example, analyses based on Lyapunov's theory were explored in continuous and discrete settings [8–11]. However, such connections have been limited mostly to gradient methods, smooth functions, and unconstrained problems.

[*] guifranca@gmail.com

1.1 Motivating examples

A simple example illustrating the interplay between discrete and continuous approaches is gradient descent

$$x_{k+1} - x_k = -\epsilon \nabla \Phi(x_k), \tag{1.1}$$

which can be seen as a discretization of the gradient flow

$$\dot{X}(t) = -\nabla \Phi(X(t)), \tag{1.2}$$

where $\epsilon > 0$ is the discretization stepsize, $X(t)$ is a continuous function of time such that $x_k = X(t)$ with $t = \epsilon k$, and $\dot{X} \equiv \frac{dX}{dt}$. This connection has been known for a long time [12]. The differential equation (1.2) has a convergence rate of $\mathcal{O}(1/t)$, which matches that of gradient descent. A second example is the heavy ball method [3]

$$x_{k+1} - x_k - \epsilon_1 (x_k - x_{k-1}) = -\epsilon_2 \nabla \Phi(x_k) \tag{1.3}$$

for constants $\{\epsilon_1, \epsilon_2\} \subset (0, \infty)$, which is a discretization of

$$\ddot{X}(t) + a_1 \dot{X}(t) = -a_2 \nabla \Phi(X(t)) \tag{1.4}$$

for certain constants a_1 and a_2, and $\ddot{X} \equiv \frac{d^2 X}{dt^2}$. A third example is given by Nesterov's accelerated gradient descent [1]

$$x_{k+1} - \hat{x}_k = -\epsilon \nabla \Phi(\hat{x}_k), \qquad \hat{x}_{k+1} - x_k = \tfrac{k}{k+3}(x_{k+1} - x_k). \tag{1.5a}$$

Only recently has its continuous limit been obtained as the differential equation [6]

$$\ddot{X}(t) + \tfrac{3}{t}\dot{X}(t) = -\nabla \Phi(X(t)). \tag{1.6}$$

When Φ is convex, this differential equation has a convergence rate of $\mathcal{O}(1/t^2)$ [6], matching the optimal $\mathcal{O}(1/k^2)$ rate of its discrete counterpart (1.5) [1]. We refer the reader to [10] for additional convergence properties of (1.6) over Hilbert spaces.

1.2 Nonsmooth cases

Recently, a continuous-time perspective for nonsmooth optimization problems has started to emerge. For instance, convergence of the dynamical system (1.6) with $\nabla\Phi$ replaced by a regularized monotone operator was considered in [13]. Also, in the context of minimizing $f + g$, where f is differentiable but g can be nonsmooth, a forward-backward Euler discretization of the differential inclusion

$$\ddot{X}(t) + \tfrac{3}{t}\dot{X}(t) + \nabla f(X(t)) \in -\partial g(X(t)), \tag{1.7}$$

where ∂g is the subdifferential of g, leads to the accelerated proximal gradient method, i.e., a nonsmooth version of (1.5) [11, 14]. Convergence rates of forward-backward proximal algorithms have also been recently considered [15].

An important algorithm for nonsmooth problems is the *alternating direction method of multipliers* (ADMM) [16–18]. ADMM is known for its ease of implementation, scalability, and applicability in many important areas of machine learning and statistics. In the convex case, ADMM converges at a rate of $\mathcal{O}(1/k)$ [19,20], while in the strongly convex case, it converges exponentially [21]. Many variants of ADMM exist including one that uses a relaxation strategy that empirically improves convergence in some cases [22,23]. However, few theoretical results are known for relaxed ADMM, with one exception being that it has exponential convergence for strongly convex functions [24–27].

An accelerated version of ADMM was recently proposed in [28] and called fast ADMM (although herein we call it accelerated ADMM, or A-ADMM for short). For a composite objective $f+g$, with f and g both strongly convex and g quadratic, A-ADMM attains a convergence rate of $\mathcal{O}(1/k^2)$ [28]. While numerical experiments in [28] show that A-ADMM may outperform Nesterov's method for some problems, we are not aware of any other rate of convergence results for such a method.

Recently, we obtained a differential equation modeling the continuous limit of A-ADMM [29]. We also showed that the underlying dynamical system has a convergence rate of $\mathcal{O}(1/t^2)$ under a convexity assumption. The work in the current paper is a significant extension of our previous paper [29] in many ways that include the following: (i) In [29], all functions were assumed to be *differentiable*, whereas here we allow all functions to be *nonsmooth*; (ii) We propose new algorithm extensions of ADMM that include *relaxation*, which were not considered in [29]; (iii) Only the *convex* setting was considered in [29], whereas we consider *both* the convex and strongly convex settings; and (iv) In [29], Nesterov acceleration was considered, whereas here we analyze *both* Nesterov and heavy-ball acceleration in a *unified* manner.

1.3 Contributions

Consider the optimization problem[1]

$$\min_{x\in\mathbb{R}^n} \{\Phi(x) \equiv f(x) + g(\boldsymbol{A}x)\} \tag{1.8}$$

where both $f:\mathbb{R}^n \to \mathbb{R}$ and $g:\mathbb{R}^m \to \mathbb{R}$ can be nonsmooth convex functions, and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Our main contributions can be summarized as follows.

- We introduce a *Relaxed and Accelerated ADMM* (R-A-ADMM) approach based on Nesterov's acceleration technique, which can be seen as a relaxation of A-ADMM [28]. We also introduce the *Relaxed Heavy Ball ADMM* (R-HB-ADMM), which is inspired by the heavy ball method.
- We derive nonsmooth dynamical systems modeling the family of R-ADMM, R-A-ADMM, and R-HB-ADMM. The limit of R-ADMM (resp., R-A-ADMM and R-HB-ADMM) is a first-order (resp., second-order) differential inclusion.
- We provide a nonsmooth Lyapunov analysis using techniques from control theory to obtain several new convergence rates for these nonsmooth dynamical systems. These results are summarized in Table 1.

One motivation for our analysis is that the derivations of convergence rates for the differential inclusions are often simpler compared to their discrete-time counterpart. Thus, such an approach

[1] Our results can be extended to $\min_{x,z}\{f(x)+g(z)\,|\,\boldsymbol{A}x+\boldsymbol{B}z=c\}$ provided $\boldsymbol{B}$ is invertible; since $z-\boldsymbol{B}^{-1}c = -\boldsymbol{B}^{-1}\boldsymbol{A}x$, one can redefine $\boldsymbol{A}$ and translate z to obtain a form similar to (1.8).

algorithm	rate (convex)	rate (strongly convex)	proof
ADMM	t^{-1}	$e^{-\mu t/\sigma_1^2(\boldsymbol{A})}$	
A-ADMM†	t^{-2}	$t^{-2r/3}$	
R-ADMM†	t^{-1}	$e^{-\mu t/\left((2-\alpha)\sigma_1^2(\boldsymbol{A})\right)}$	Theorem 4.2
R-A-ADMM‡	t^{-2}	$t^{-2r/3}$	Theorem 4.3
R-HB-ADMM‡	t^{-1}	$e^{-\sqrt{\mu}t/\left(\sqrt{2-\alpha}\,\sigma_1(\boldsymbol{A})\right)}$	Theorem 4.4

Table 1: Convergence rates of the dynamical systems related to the ADMM variants proposed in this paper. Algorithms marked with † are known but do not have known convergence rates; e.g., for A-ADMM both rates are unknown while for R-ADMM the $\mathcal{O}(1/t)$ rate seems to be unknown. Those marked with ‡ are new families of algorithms. Here $\alpha \in (0,2)$ is the relaxation parameter, μ is the strong convexity parameter, and $\sigma_1(\boldsymbol{A})$ is the largest singular value of $\boldsymbol{A}$.

provides valuable insight on the behavior of these algorithms even though they do not formally establish discrete-time rates. We stress that most of the rates in Table 1 are unknown for the associated algorithms. Also, one can see that by adding relaxation, an improved constant in the complexity bound may be attained; e.g., for R-ADMM and R-HB-ADMM in the strongly convex case. Moreover, the proposed R-HB-ADMM attains exponential convergence in the strongly convex case, which contrasts R-A-ADMM. This shows an interesting tradeoff between Nesterov and heavy ball type of acceleration in the convex versus strongly convex settings.

We show that the differential inclusions associated to the accelerated algorithms can be obtained through an extension of the nonsmooth Hamiltonian formalism introduced by Rockafellar [30–32]. We construct two Hamiltonian representations for the dissipative dynamical systems associated to R-A-ADMM and R-HB-ADMM: one with an explicit time-dependent Hamiltonian obeying standard Hamilton's equations, while the other is based on a time-independent Hamiltonian with equations of motion in the form of conformal Hamiltonian systems [33], but which we generalize to nonsmooth cases. Our approach generalizes the variational perspective put forward by [7] and opens opportunities beyond the scope of this paper, such as constructing suitable discretizations to (nonsmooth) Hamiltonian systems.

2 Preliminaries

2.1 Notation

For $\{x, y\} \subset \mathbb{R}^n$, let $\|x\| = \sqrt{x^T x}$ be the ℓ_2 norm of x and $\langle x, y\rangle = x^T y$ the inner product between x and y. The ℓ_1 norm of x is denoted by $\|x\|_1$. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we denote its largest and smallest singular values by $\sigma_1(\boldsymbol{A})$ and $\sigma_n(\boldsymbol{A})$, respectively, and its condition number by $\kappa(\boldsymbol{A}) \equiv \sigma_1(\boldsymbol{A})/\sigma_n(\boldsymbol{A})$. The nuclear norm of $\boldsymbol{A}$ is denoted by $\|\boldsymbol{A}\|_* = \sum_i \sigma_i(\boldsymbol{A})$.

2.2 Subdifferentials, convexity, and strong convexity

Consider a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ with effective domain $\operatorname{dom} f \equiv \{x \,|\, f(x) < \infty\}$. Its subdifferential at the point $x \in \operatorname{dom} f$ is defined as [34]

$$\partial f(x) = \{\xi \in \mathbb{R}^n \,|\, f(y) - f(x) \geq \langle \xi, y - x\rangle \text{ for all } y\}\,. \tag{2.1}$$

The subdifferential set $\partial f(x)$ is always closed and convex, and if f is convex it is also nonempty. A strongly convex function [34] is defined as follows.

Definition 2.1 (Strongly convex function)**.** *We say that $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is strongly convex if and only if there exists a constant $\mu > 0$ such that*

$$h(y) \geq h(x) + \langle \xi, y - x\rangle + \tfrac{\mu}{2}\|y - x\|^2 \tag{2.2}$$

for all $x, y \in \operatorname{dom} h$ and all $\xi \in \partial h(x)$. If f is only convex, (2.2) *holds with $\mu = 0$.*

The following conditions are assumed throughout the paper.

Assumption 2.2. *f and g in problem* (1.8) *are proper, lower semicontinuous, convex, and satisfy $0 \in \operatorname{int}(\operatorname{dom} g - \boldsymbol{A} \operatorname{dom} f)$. The matrix $\boldsymbol{A}$ has full column rank.*

Lemma 2.3. *Under Assumption 2.2 the following hold (see [35, 36]):*

(i) $\partial\Phi$ is upper semicontinuous on $\operatorname{int}(\operatorname{dom}\Phi)$;

(ii) $\partial(f + g \circ \boldsymbol{A})(x) = \partial f(x) + \boldsymbol{A}^T \partial g(\boldsymbol{A}x)$.

2.3 Continuous limits and differential inclusions

Let us mention two basic relationships involving continuous limits. Let $X = X(t) \in \mathbb{R}^n$ where $t \geq 0$ is the time variable. The corresponding state of an algorithm at discrete-times $k = 0, 1, \ldots$ will be denoted by $x_k \in \mathbb{R}^n$, where $x_k = X(t)$, for $t = \epsilon k$ and for some small enough stepsize $\epsilon > 0$. In the limit $\epsilon \to 0$ it holds that [37–39]

$$\left(x_{k\pm1} - x_k\right)/\epsilon \to \pm\dot{X}(t), \qquad \left(x_{k+1} - 2x_k + x_{k-1}\right)/\epsilon^2 \to \ddot{X}(t). \tag{2.3}$$

Consider a differential inclusion [35]

$$\dot{X}(t) \in F(X(t)) \qquad \text{with} \qquad X(0) = x_0, \tag{2.4}$$

where $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a multi-valued map. By a solution or trajectory of (2.4) we mean a function $\varphi : \mathcal{I} \to \mathbb{R}^n$, with $\mathcal{I} \subseteq \mathbb{R}_+$, such that φ is absolutely continuous and $\dot{\varphi}(t) \in F(\varphi(t))$ for all $t \in \mathcal{I}$. In this paper we consider first- and second-order differential inclusions (higher-order systems can always be written in the first-order form (2.4)). If F is upper semicontinuous (or lower semicontinuous), nonempty, closed, and convex, then a unique solution to (2.4) exists [35]. When F is not closed, or is nonconvex, the existence of solutions may be a delicate matter [40–42]. In this paper, since $F = -\partial\Phi$ and Φ satisfies Assumption 2.2 (see Lemma 2.3), the differential inclusion (2.4) has a unique solution for any $x_0 \in \operatorname{dom}\Phi$ and all $t \geq 0$ [35, Chapter 3].

3 Variants of ADMM as Dynamical Systems

We first consider the family of R-ADMM algorithms and introduce a nonsmooth dynamical system modeling these algorithms in the continuous-time limit. Later we will consider accelerated extensions together with second-order nonsmooth dynamical systems. These algorithms are developed for problem (1.8) by introducing the variable $z = \boldsymbol{A}x$ and considering the (scaled) augmented Lagrangian $\mathcal{L}_\rho(x,z,u) = f(x) + g(z) + \rho\langle u, \boldsymbol{A}x - z\rangle + \frac{\rho}{2}\|\boldsymbol{A}x - z\|_2^2$, where $u \in \mathbb{R}^m$ is the (scaled) Lagrange multiplier vector and $\rho > 0$ [18].

3.1 Relaxed ADMM

We start with the R-ADMM framework [18]:

$$x_{k+1} \leftarrow \arg\min\nolimits_x \left\{ f(x) + \tfrac{\rho}{2}\|\boldsymbol{A}x - z_k + u_k\|^2 \right\}, \tag{3.1a}$$

$$z_{k+1} \leftarrow \arg\min\nolimits_z \big\{ g(z) + \tfrac{\rho}{2}\|\alpha\boldsymbol{A}x_{k+1} + (1-\alpha)z_k - z + u_k\|^2 \big\}, \tag{3.1b}$$

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

The relaxation parameter $\alpha \in (0,2)$ is introduced to speed up convergence [22, 23]. Note that the standard ADMM is recovered when $\alpha = 1$.

Theorem 3.1. *Consider* (3.1) *for solving problem* (1.8) *under Assumption 2.2. Its continuous limit, with time scale $t = \rho^{-1}k$, is given by the differential inclusion*

$$(2-\alpha)(\boldsymbol{A}^T\boldsymbol{A})\dot{X}(t) \in -\partial\Phi(X(t)) \tag{3.2}$$

with initial condition $X(0) = x_0 \in \operatorname{dom}(\Phi)$ (see also Remark 3.2).

Proof. With $\epsilon = \rho^{-1}$, the optimality conditions for (3.1a) and (3.1b) combine into

$$0 \in \partial f(x_{k+1}) + \boldsymbol{A}^T\partial g(z_{k+1}) + (1-\alpha)\boldsymbol{A}^T(\boldsymbol{A}x_{k+1} - z_k)/\epsilon + \boldsymbol{A}^T(z_{k+1} - z_k)/\epsilon. \tag{3.3}$$

Let $(x_k, z_k, u_k) = (X(t), Z(t), U(t))$ where $t = \epsilon k$. From the last update (3.1c), we have $U(t+\epsilon) = U(t) + \alpha\boldsymbol{A}X(t+\epsilon) + (1-\alpha)Z(t) - Z(t+\epsilon)$, which expanded yields $\epsilon\dot{U} = \alpha\big(\boldsymbol{A}X - Z\big) + \epsilon\big(\alpha\boldsymbol{A}\dot{X} - \dot{Z}\big) + \mathcal{O}(\epsilon^2)$, where we denote $U = U(t)$, $Z = Z(t)$ and $X = X(t)$ for simplicity. In the limit $\epsilon \to 0$ this implies

$$Z = \boldsymbol{A}X. \tag{3.4}$$

Using this equality in (3.3) we thus have

$$0 \in \partial f\big(X(t+\epsilon)\big) + \boldsymbol{A}^T\partial g\big(\boldsymbol{A}X(t+\epsilon)\big) + (2-\alpha)\boldsymbol{A}^T\boldsymbol{A}\big(X(t+\epsilon) - X(t)\big)/\epsilon. \tag{3.5}$$

Making use of (2.3), in the limit $\epsilon \to 0$ this becomes

$$0 \in \partial f(X) + \boldsymbol{A}^T\partial g(\boldsymbol{A}X) + (2-\alpha)\boldsymbol{A}^T\boldsymbol{A}\dot{X}. \tag{3.6}$$

From Lemma 2.3 (ii) this is equal to (3.2). This is a first-order system whose dynamics is specified by the initial condition $X(0) = x_0 \in \operatorname{dom}(\Phi)$. □

If f and g are differentiable (so that $\partial\Phi(x) = \nabla f(x) + \boldsymbol{A}^T\nabla g(\boldsymbol{A}x)$), $\boldsymbol{A} = I$ and $\alpha = 1$, then (3.2) reduces to the gradient flow (1.2). In general, however, even in the smooth case, the presence of $\boldsymbol{A}^T\boldsymbol{A}$ in (3.2) can make its stability properties different.

Remark 3.2 (Initial condition). We point out a sublety regarding the initial condition in Theorem 3.1, which will also apply to the other differential inclusions obtained in this paper. It is necessary that $X(0) = x_0$ matches the starting point of the algorithm. Since the optimality condition of (3.1b) combined with (3.1c) implies $u_{k+1} \in (1/\rho)\partial(z_{k+1})$, in the limit $\epsilon^{-1} \equiv \rho \to \infty$ we additionally have $U(t) = 0$ for all t. Hence, we assume that (3.1) is initialized with $u_{-1} = 0$ and some z_{-1}, and then $X(0) = x_0$ is obtained from the output of the first update (3.1a). This condition is also assumed in algorithms (3.7) and (3.16) with respect to $\hat{u}_{-1} = 0$ and $\hat{z}_{-1}$.

3.2 Relaxed and Accelerated ADMM

We introduce variables $\hat{u} \in \mathbb{R}^m$ and $\hat{z} \in \mathbb{R}^m$ to obtain an accelerated version of (3.1), called R-A-ADMM, given by

$$x_{k+1} \leftarrow \arg\min_x \left\{ f(x) + \tfrac{\rho}{2}\|\boldsymbol{A}x - \hat{z}_k + \hat{u}_k\|^2 \right\}, \tag{3.7a}$$

$$z_{k+1} \leftarrow \arg\min_z \big\{ g(z) + \tfrac{\rho}{2}\|\alpha\boldsymbol{A}x_{k+1} + (1-\alpha)\hat{z}_k - z + \hat{u}_k\|^2 \big\}, \tag{3.7b}$$

$$u_{k+1} \leftarrow \hat{u}_k + \alpha\boldsymbol{A}x_{k+1} + (1-\alpha)\hat{z}_k - z_{k+1}, \tag{3.7c}$$

$$\gamma_{k+1} \leftarrow k/(k+r), \tag{3.7d}$$

$$\hat{u}_{k+1} \leftarrow u_{k+1} + \gamma_{k+1}\left(u_{k+1} - u_k\right), \tag{3.7e}$$

$$\hat{z}_{k+1} \leftarrow z_{k+1} + \gamma_{k+1}\left(z_{k+1} - z_k\right). \tag{3.7f}$$

This algorithm, which is new to the literature, is a relaxation of A-ADMM [28] that is recovered by setting $\alpha = 1$ and $r = 3$. It is worth noting that even for R-ADMM (3.1), the existing theoretical results are sparse compared to standard ADMM. Regarding the continuous limit of updates (3.7) we obtain the following.

Theorem 3.3. *Consider* (3.7) *with* $r \geq 3$ *for solving* (1.8) *under Assumption 2.2. Its continuous limit, with time scale* $t = \rho^{-1/2}k$, *is given by the differential inclusion*

$$(2-\alpha)\boldsymbol{A}^T\boldsymbol{A}\big(\ddot{X}(t) + \tfrac{r}{t}\dot{X}(t)\big) \in -\partial\Phi(X(t)) \tag{3.8}$$

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

Proof. Choosing $\epsilon = \rho^{-1/2}$ and combining the optimality conditions of the problems in (3.7a) and (3.7b) yields

$$0 \in \partial f(x_{k+1}) + \boldsymbol{A}^T\partial g(z_{k+1}) + (1-\alpha)\boldsymbol{A}^T(\boldsymbol{A}x_{k+1} - \hat{z}_k)/\epsilon^2 + \boldsymbol{A}^T(z_{k+1} - \hat{z}_k)/\epsilon^2. \tag{3.9}$$

Let $(x_k, z_k, u_k, \hat{z}_k, \hat{u}_k) = (X(t), Z(t), U(t), \hat{Z}(t), \hat{U}(t))$ at $t = \epsilon k$, and we use the shorthand $X = X(t)$, $Z = Z(t)$, and so on. Let us consider the last two terms of (3.9) separately. From (3.7f) we have $z_{k+1} - \hat{z}_k = z_{k+1} - (1+\gamma_k)z_k + \gamma_k z_{k-1}$. Adding $z_k - z_k + z_{k-1} - z_{k-1} = 0$ and reorganizing yields

$$(z_{k+1} - \hat{z}_k)/\epsilon^2 = (z_{k+1} - 2z_k + z_{k-1})/\epsilon^2 + (1-\gamma_k)(z_k - z_{k-1})/\epsilon^2. \tag{3.10}$$

When $\epsilon \to 0$, according to (2.3) the first term yields $\ddot{Z}$ while the second term yields $(1-\gamma_k)(z_k - z_{k-1})/\epsilon^2 \to \frac{r}{t}\dot{Z}$, therefore

$$(z_{k+1} - \hat{z}_k)/\epsilon^2 \to \ddot{Z} + (r/t)\dot{Z}. \tag{3.11}$$

Let us now focus on the third term of (3.9). Note that

$$(\boldsymbol{A}x_{k+1} - \hat{z}_k)/\epsilon^2 = (\boldsymbol{A}x_{k+1} - z_{k+1})/\epsilon^2 + (z_{k+1} - \hat{z}_k)/\epsilon^2 \tag{3.12}$$

and the last term was already computed in (3.11). We now show that the first term above vanishes. Considering (3.7e), i.e. $\hat{U}(t+\epsilon) - U(t+\epsilon) - \frac{t}{t+\epsilon r}\big(U(t+\epsilon) - U(t)\big) = 0$, when $\epsilon \to 0$ this implies that $\hat{U}(t) = U(t)$ (a similar argument also shows that $\hat{Z}(t) = Z(t)$), which combined with (3.7c) allow us to conclude that $\boldsymbol{A}X(t) = Z(t)$, which in turn implies $\boldsymbol{A}\dot{X}(t) = \dot{Z}(t)$ and $\boldsymbol{A}\ddot{X}(t) = \ddot{Z}(t)$. Therefore,

$$(\boldsymbol{A}x_{k+1} - \hat{z}_k)/\epsilon^2 \to \ddot{Z} + (r/t)\dot{Z}. \tag{3.13}$$

From (3.9), (3.11), and (3.13) we thus obtain

$$(2-\alpha)\boldsymbol{A}^T\boldsymbol{A}\big(\ddot{X} + \tfrac{r}{t}\dot{X}\big) \in -\partial f(X) - \boldsymbol{A}^T\partial g(\boldsymbol{A}X), \tag{3.14}$$

which along with Lemma 2.3 (ii) yields (3.8).

For initial conditions, the first is $X(0) = x_0$ for a suitable x_0 (see Remark 3.2). Next, using the Mean Value Theorem, we have $\dot{X}_j(t) = \dot{X}_j(0) + t\ddot{X}_j(\xi)$ for some $\xi \in [0,t]$ and for all components $j = 1,\dots,n$. Combining this with (3.8) yields

$$\sum_j (\boldsymbol{A}^T\boldsymbol{A})_{ij}\left(\dot{X}_j(t) - \dot{X}_j(0) + r\dot{X}_j(\xi)\right) \in \frac{-t\partial_i\Phi(X(\xi))}{2-\alpha} \tag{3.15}$$

where we use $\partial_i\Phi$ to denote the ith component of $\partial\Phi$. Letting $t \downarrow 0$, which forces $\xi \downarrow 0$, we have $\sum_j (\boldsymbol{A}^T\boldsymbol{A})_{ij}\dot{X}_j(0) = 0$ since $\alpha \neq 2$ and $r \neq 0$. Since this holds for each $i = 1,\dots,n$, $(\boldsymbol{A}^T\boldsymbol{A})\dot{X}(0) = 0$ and, since $\boldsymbol{A}$ has full column rank, that $\dot{X}(0) = 0$. □

Remark 3.4. Theorem 3.3 still holds without assuming $0 \in \operatorname{int}(\operatorname{dom} g - \boldsymbol{A}\operatorname{dom} f)$ (see Assumption 2.2), although in this case the equality in Lemma 2.3(ii) becomes an inclusion, which results in $\partial f + \boldsymbol{A}^T\partial g \circ \boldsymbol{A}$ replacing the right-hand side of (3.8).

The differential inclusion (3.8) reduces to the differential equation (1.6) when $\boldsymbol{A} = \boldsymbol{I}$, $\alpha = 1$, $r = 3$, and both f and g are smooth. Thus, the results for (3.8) hold for (1.6) as a special case. We show in Section 4 that the relaxation parameter α and the matrix $\boldsymbol{A}$ in (3.8) allow for refined convergence results.

3.3 Relaxed Heavy Ball ADMM

Another acceleration scheme for gradient descent is the heavy ball method [3]. Motivated by this we introduce another accelerated variant of (3.1), called R-HB-ADMM. The updates are essentially the same as those in (3.7), but $\{\gamma_k\}$ is now constant, namely $\gamma = 1 - r/\sqrt{\rho}$ with $r > 0$ and thus

$$
\begin{aligned}
x_{k+1} &\leftarrow \arg\min_x \big\{ f(x) + \tfrac{\rho}{2}\|\boldsymbol{A}x - \hat{z}_k + \hat{u}_k\|^2 \big\}, && \text{(3.16a)} \\
z_{k+1} &\leftarrow \arg\min_z \big\{ g(z) + \tfrac{\rho}{2}\|\alpha\boldsymbol{A}x_{k+1} + (1-\alpha)\hat{z}_k - z + \hat{u}_k\|^2 \big\}, && \text{(3.16b)} \\
u_{k+1} &\leftarrow \hat{u}_k + \alpha\boldsymbol{A}x_{k+1} + (1-\alpha)\hat{z}_k - z_{k+1}, && \text{(3.16c)} \\
\hat{u}_{k+1} &\leftarrow u_{k+1} + \gamma\,(u_{k+1} - u_k)\,, && \text{(3.16d)} \\
\hat{z}_{k+1} &\leftarrow z_{k+1} + \gamma\,(z_{k+1} - z_k)\,. && \text{(3.16e)}
\end{aligned}
$$

Compared to (3.7) this method uses $\gamma_k = \gamma$ as a constant depending on ρ. This choice is inspired by the continuous limit described below, but is otherwise not obvious.

Theorem 3.5. *Consider* (3.16) *with $r > 0$ for solving problem* (1.8) *under Assumption 2.2. Its continuous limit, with time scale $t = \rho^{-1/2}k$, is given by*

$$(2-\alpha)\boldsymbol{A}^T\boldsymbol{A}\big(\ddot{X}(t) + r\dot{X}(t)\big) \in -\partial\Phi(X(t)) \tag{3.17}$$

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

Proof. Since obtaining the differential equation follows the same steps as in the proof of Theorem 3.3, its proof is omitted; the only changes are initial conditions. The first is $X(0) = x_0$ (see Remark 3.2). For the velocity, consider the optimality condition for (3.16a), i.e., $0 \in \partial f(x_0) + \rho\boldsymbol{A}^T\boldsymbol{A}x_0 + \rho\boldsymbol{A}^T(\hat{u}_{-1} - \hat{z}_{-1})$, where $\hat{z}_{-1}$ and $\hat{u}_{-1} = 0$ are inputs. Recall that $\epsilon^2 = 1/\rho$ and $\hat{Z} \to Z$ and $X \to \boldsymbol{A}Z$ as $\epsilon \to 0$, thus $\boldsymbol{A}^T\boldsymbol{A}\dot{X}(0) = 0$, implying $\dot{X}(0) = 0$ since $\boldsymbol{A}$ has full column rank. □

The comments in Remark 3.4 also apply to Theorem 3.5. Note that the differential inclusion (3.17) is closely related to (3.8). The key difference between the dynamical systems (3.8) and (3.17) is in the damping coefficient, which leads to different stability properties; see Table 1 and Section 4. Specifically, in (3.8) the damping vanishes asymptotically, thus for large time the system may exhibit strong oscillations, while in (3.17) the damping is constant, which helps to attenuate oscillations and improve stability. Although this has been observed empirically, it has also be shown that (3.17) is asymptotically stable for isolated minimizers while (3.8) is only stable [29].

Note that the dual variables u_k and $\hat{u}_k$ have no continuous counterparts in (3.2), (3.8), and (3.17). The reason is that these dynamical systems capture only the leading order behaviour of the discrete algorithm in the limit of large ρ, thus $U(t)$ and $\hat{U}(t)$ would potentially appear in higher-order corrections.

4 Convergence Rates of the Nonsmooth Dynamical Systems

In this section we derive convergence rates for the previous nonsmooth dynamical systems when Φ is convex or strongly convex; see Table 1 for a summary. These results are established through a *nonsmooth* Lyapunov analysis. We first introduce some basic concepts, and refer to [35, 41–43] for more details.

The directional derivative of a convex function f at a point $x \in \operatorname{dom} f$ in the direction $v \in \mathbb{R}^n$ is defined as

$$Df(x)(v) \equiv \lim_{\epsilon \downarrow 0} \frac{f(x+\epsilon v) - f(x)}{\epsilon}. \tag{4.1}$$

Consider the differential inclusion (2.4). Recall that with $F = -\partial\Phi$ and under Assumption 2.2, a unique solution exists. Our analysis depends on the time derivative of Φ along the trajectory, which is defined as

$$\dot{\Phi}(X(t)) \equiv D\Phi(X(t))\big(\dot{X}(t)\big) = \lim_{\epsilon \downarrow 0} \frac{\Phi(X(t+\epsilon)) - \Phi(X(t))}{\epsilon} \tag{4.2}$$

where the second equality can be proved using the convexity of Φ and the definition of the directional derivative in (4.1). The form of the time derivative is now proved.

Lemma 4.1. *Consider the differential inclusions* (3.2), (3.8), *and* (3.17), *with given initial conditions. The unique solution $X \equiv X(t)$ of such a differential inclusion obeys, for all $t \in \mathbb{R}_+$, the following relationships:*

$$\dot{\Phi}(X) = \begin{cases} -(2-\alpha)\big\|\boldsymbol{A}\dot{X}\big\|^2 & \textit{for } (3.2), \\ -(2-\alpha)\left(\big\langle \ddot{X}, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\big\rangle + \frac{r}{t}\big\|\boldsymbol{A}\dot{X}\big\|^2\right) & \textit{for } (3.8), \\ -(2-\alpha)\left(\big\langle \ddot{X}, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\big\rangle + r\big\|\boldsymbol{A}\dot{X}\big\|^2\right) & \textit{for } (3.17). \end{cases} \tag{4.3}$$

Proof. We start with the differential inclusion (3.2), and let

$$\eta_\epsilon \equiv -(2-\alpha)\boldsymbol{A}^T\boldsymbol{A}\dot{X}(t+\epsilon) \in \partial\Phi(X(t+\epsilon)) \qquad (\epsilon > 0). \tag{4.4}$$

It follows from (2.1) and convexity of Φ that

$$\Phi(X(t)) - \Phi(X(t+\epsilon)) \geq \langle \eta_\epsilon, X(t) - X(t+\epsilon)\rangle. \tag{4.5}$$

Under Assumption 2.2 the differential inclusion has a unique solution and $\dot{X}(t)$ exists for all $t \geq 0$. Dividing (4.5) by ϵ, and then taking the limit $\epsilon \downarrow 0$ and using (4.2) yields

$$\dot{\Phi}(X) \leq -(2-\alpha)\big\langle \boldsymbol{A}^T\boldsymbol{A}\dot{X}, \dot{X}\big\rangle = -(2-\alpha)\|\boldsymbol{A}\dot{X}\|^2. \tag{4.6}$$

Since Φ is lower semicontinuous, we also know that $\eta_0 \equiv \lim_{\epsilon\downarrow 0} \eta_\epsilon \in \partial\Phi(X(t))$. Combining this with (2.1) and convexity of Φ yields

$$\Phi(X(t+\epsilon)) - \Phi(X(t)) \geq \langle \eta_0, X(t+\epsilon) - X(t)\rangle, \tag{4.7}$$

and then using the same argument as above shows that

$$\dot{\Phi}(X) \geq \big\langle \eta_0, \dot{X}\big\rangle = -(2-\alpha)\|\boldsymbol{A}\dot{X}\|^2. \tag{4.8}$$

Thus, (4.6) together with (4.8) gives the desired result.

The same arguments can be applied to the second-order inclusion (3.8) with

$$\eta_\epsilon \equiv -(2-\alpha)\boldsymbol{A}^T\boldsymbol{A}\big(\ddot{X}(t+\epsilon) + \tfrac{r}{t+\epsilon}\dot{X}(t+\epsilon)\big) \in \partial\Phi(X(t+\epsilon)), \tag{4.9}$$

and also to the second-order inclusion (3.17) by replacing $\frac{r}{t+\epsilon} \mapsto r$ in (4.9) above. □

4.1 Convergence of Relaxed ADMM

We start with the first-order dynamical system (3.2) associated with R-ADMM.

Theorem 4.2. *Consider* (3.2) *and let* $x^\star \in \arg\min \Phi(x)$, $\Phi^\star \equiv \Phi(x^\star)$, *and* $X = X(t)$ *be the unique trajectory of the system from* $X(0) = x_0 \in \operatorname{dom}(\Phi)$. *If* Φ *is* convex, *then for all* $t \geq 0$ *it holds that*

$$\Phi(X(t)) - \Phi^\star \leq \frac{(2-\alpha)\|\boldsymbol{A}(x_0 - x^\star)\|^2}{2t}. \tag{4.10}$$

If Φ *is* μ-strongly convex, *then for all* $t \geq 0$, *it holds that*

$$\|X(t) - x^\star\|^2 \leq \frac{\|\boldsymbol{A}(x_0 - x^\star)\|^2 e^{-\eta t}}{\sigma_m^2(\boldsymbol{A})} \qquad \textit{where} \qquad \eta \equiv \frac{\mu}{(2-\alpha)\sigma_1^2(\boldsymbol{A})}. \tag{4.11}$$

Proof. For the first part consider

$$\mathcal{E}(X,t) \equiv \tfrac{t}{2-\alpha}\left(\Phi(X) - \Phi^\star\right) + \tfrac{1}{2}\left\|\boldsymbol{A}\left(X - x^\star\right)\right\|^2 \tag{4.12}$$

where $x^\star$ is any minimizer of Φ. Taking its time derivative,

$$\dot{\mathcal{E}} = \tfrac{t}{2-\alpha}\dot{\Phi}(X) + \tfrac{1}{2-\alpha}\left(\Phi(X) - \Phi^\star\right) + \big\langle X - x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\big\rangle. \tag{4.13}$$

Using (2.2) (with $\mu = 0$), $\alpha \in (0,2)$, and (3.2) one concludes that the second and third terms above are bounded above by zero. Hence, Lemma 4.1 implies that $\dot{\mathcal{E}} \leq -t\|\boldsymbol{A}\dot{X}\|^2 \leq 0$. Thus, $\mathcal{E}|_t \leq \mathcal{E}|_{t=0}$ and (4.12) yields $\Phi(X) - \Phi^\star \leq (2-\alpha)\mathcal{E}|_{t=0}\, t^{-1}$. Replacing $\mathcal{E}|_{t=0} = \frac{1}{2}\|\boldsymbol{A}(x_0 - x^\star)\|^2$ yields (4.10).

For the second part, from (2.2), for all $X \in \operatorname{dom}\Phi$ and every $\xi \in \partial\Phi(X)$, we have

$$\langle X - x^\star, \xi\rangle \geq \tfrac{\mu}{2}\|X - x^\star\|^2 + \Phi(X) - \Phi(x^\star) \geq \tfrac{\mu}{2\sigma_1^2(\boldsymbol{A})}\|\boldsymbol{A}(X - x^\star)\|^2 \tag{4.14}$$

where now $x^\star$ is the unique minimizer of Φ. Consider

$$\mathcal{E}(X) \equiv \tfrac{1}{2}\|\boldsymbol{A}(X - x^\star)\|^2. \tag{4.15}$$

Taking its total time derivative and using $-(2-\alpha)\boldsymbol{A}^T\boldsymbol{A}\dot{X} \in \partial\Phi(X)$ together with (4.14) yields $\dot{\mathcal{E}} \leq -(\eta/2)\left\|\boldsymbol{A}(X - x^\star)\right\|^2$, with η as in (4.11). From the definition (4.15) we can write this as $\frac{d}{dt}\|\boldsymbol{A}(X(t) - x^\star)\|^2 \leq -\eta\|\boldsymbol{A}(X(t) - X^\star)\|^2$. Grönwall's inequality thus implies $\|\boldsymbol{A}(X(t) - x^\star)\|^2 \leq \|\boldsymbol{A}(x_0 - x^\star)\|^2 e^{-\eta t}$, and after using norm inequalities we finally get $\sigma_m^2(\boldsymbol{A})\|X(t) - x^\star\|^2 \leq \|\boldsymbol{A}(x_0 - x^\star)\|^2 e^{-\eta t}$. □

Some remarks are appropriate:

- The rate (4.10) matches the $\mathcal{O}(1/k)$ rate of standard (non-relaxed) ADMM in the convex case [19, 20]. We believe the analogous result for *relaxed* ADMM (4.10) is new in the sense that this rate is unknown in the discrete case.
- The exponential rate in (4.11) is consistent with the known rate for relaxed ADMM when Φ is strongly convex [27].

- Results (4.10) and (4.11) suggest better performance with $\alpha \in (1,2)$. This is prominent in the strongly convex case since α is under an exponential.
- It may seem desirable to choose $\alpha \approx 2$. However, one must avoid divergence in the Lyapunov functions used to establish Theorem 4.2 (e.g., see (4.12)). In the extreme case, $\alpha = 2$, there is simply no dynamics in (3.2). These observations are consistent with the empirical guideline $\alpha \in (1.5, 1.8)$ suggested by [22, 23]. The choice $\alpha \geq 2$ should be avoided because (3.2) would follow the (sub)gradient *ascent* direction, which is consistent with the results of [27].

4.2 Convergence of Relaxed and Accelerated ADMM

The proof strategy for the dynamical system (3.8) is similar to Theorem 4.2, but more involved.

Theorem 4.3. *Consider* (3.8) *with $r \geq 3$. Let $x^\star \in \arg\min \Phi(x)$, $\Phi^\star \equiv \Phi(x^\star)$, and $X = X(t)$ be the unique trajectory of the system from $X(0) = x_0 \in \operatorname{dom}\Phi$ and $\dot{X}(0) = 0$. If Φ is* convex, *then for all $t \geq 0$ we have*

$$\Phi(X(t)) - \Phi^\star \leq \frac{(2-\alpha)(r-1)^2 \|\boldsymbol{A}(x_0 - x^\star)\|^2}{2t^2}. \tag{4.16}$$

*If Φ is μ-*strongly convex, *then for all $t \geq t_0 \equiv \frac{2}{3}\sigma_1(\boldsymbol{A})\sqrt{r(r-3)(2-\alpha)\mu^{-1}}$ we have*

$$\|X(t)) - x^\star\|^2 \leq 4(2-\alpha)\, c\mu^{-1} t^{-2r/3}, \tag{4.17}$$

where $c \equiv \frac{t_0^\lambda}{2}\left(\frac{2}{2-\alpha}(\Phi_0 - \Phi^\star) + \frac{\lambda^2}{t_0^2}\|\boldsymbol{A}(X_0 - x^\star)\|^2 + \|\boldsymbol{A}\dot{X}_0\|^2\right)$ with $\lambda \equiv 2r/3$, $X_0 \equiv X(t_0)$, $\dot{X}_0 \equiv \dot{X}(t_0)$, and $\Phi_0 \equiv \Phi(X_0)$.

Proof. For the first part, define

$$\mathcal{E}(X,t) \equiv \tfrac{t^2}{(r-1)^2(2-\alpha)}\left(\Phi(X) - \Phi^\star\right) + \tfrac{1}{2}\big\|\boldsymbol{A}\big(X - x^\star + \tfrac{t}{r-1}\dot{X}\big)\big\|^2 \tag{4.18}$$

where $x^\star$ is a minimizer of Φ. The total time derivative gives

$$\begin{aligned}\dot{\mathcal{E}} = {} & \tfrac{2t}{(r-1)^2(2-\alpha)}\left(\Phi(X) - \Phi^\star\right) + \tfrac{t^2}{(r-1)^2(2-\alpha)}\dot{\Phi}(X) \\ & + \tfrac{1}{r-1}\big\langle X - x^\star + \tfrac{t}{r-1}\dot{X}, \boldsymbol{A}^T\boldsymbol{A}\big(r\dot{X} + t\ddot{X}\big)\big\rangle.\end{aligned} \tag{4.19}$$

Using Lemma 4.1 and (3.8) we can simplify this to

$$\dot{\mathcal{E}} = \tfrac{2t}{(r-1)(2-\alpha)}\Big(\tfrac{1}{r-1}\left(\Phi(X) - \Phi^\star\right) - \tfrac{1}{2}\langle X - x^\star, \xi\rangle\Big) \tag{4.20}$$

where $\xi \equiv \eta_0$ is given by (4.9). From convexity of Φ we have $\Phi(X) - \Phi^\star \leq \langle \xi, X - x^\star\rangle$ for all X and any minimizer $x^\star$, which combined with (4.20) yields

$$\dot{\mathcal{E}} \leq -\tfrac{t(r-3)}{(r-1)^2(2-\alpha)}\left(\Phi(X) - \Phi^\star\right) \leq 0 \tag{4.21}$$

since $r \geq 3$. Thus, $\mathcal{E}|_t \leq \mathcal{E}|_{t=0}$ or $\Phi(X) - \Phi^\star \leq (r-1)^2(2-\alpha)\mathcal{E}|_{t=0}\, t^{-2}$, which gives (4.16) because $\mathcal{E}|_{t=0} = \frac{1}{2}\|\boldsymbol{A}(x_0 - x^\star)\|^2$.

The proof of the strongly convex case takes several steps. First, define

$$\mathcal{E}(X,t) \equiv \tfrac{t^\lambda}{2-\alpha}\left(\Phi(X)-\Phi^\star\right) + \tfrac{t^{\lambda-2}}{2}\|\boldsymbol{A}(\lambda(X-x^\star)+t\dot{X})\|^2 \tag{4.22}$$

where $\lambda \equiv 2r/3 \geq 2$ (since $r \geq 3$) and $x^\star$ is the unique minimizer of Φ (thus $0 \in \partial\Phi(x^\star)$). The total time derivative is given by

$$\begin{aligned}\dot{\mathcal{E}} = \tfrac{\lambda t^{\lambda-1}}{2-\alpha}\left(\Phi(X)-\Phi^\star\right) + \tfrac{t^\lambda}{2-\alpha}\dot{\Phi}(X) + \tfrac{1}{2}(\lambda-2)t^{\lambda-3}\big\|\boldsymbol{A}\big(\lambda(X-x^\star)+t\dot{X}\big)\big\|^2 \\ + t^{\lambda-2}\big\langle \lambda(X-x^\star)+t\dot{X}, \boldsymbol{A}^T\boldsymbol{A}\big((\lambda+1)\dot{X}+t\ddot{X}\big)\big\rangle.\end{aligned} \tag{4.23}$$

Using Lemma 4.1 in the second term and noting that it cancels with part of the last term, after simplifying the resulting equation one obtains

$$\begin{aligned}\dot{\mathcal{E}} = \tfrac{\lambda t^{\lambda-1}}{2-\alpha}\left(\Phi(X)-\Phi^\star\right) + \tfrac{1}{2}\lambda^2(\lambda-2)t^{\lambda-3}\big\|\boldsymbol{A}(X-x^\star)\big\|^2 \\ + \lambda\left(\tfrac{\lambda}{2}-1\right)t^{\lambda-2}\big\langle X-x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\big\rangle - \tfrac{\lambda t^{\lambda-1}}{2-\alpha}\big\langle X-x^\star, \xi\big\rangle\end{aligned} \tag{4.24}$$

for the same $\xi \equiv \eta_0 \in \partial\Phi(X(t))$ as given earlier in this proof. From (2.2) we have

$$\langle X-x^\star, \xi\rangle \geq \Phi(X)-\Phi^\star + \tfrac{\mu}{2\sigma_1^2(\boldsymbol{A})}\|\boldsymbol{A}(X-x^\star)\|^2. \tag{4.25}$$

Using the two previous bounds and simplifying yields

$$\begin{aligned}\dot{\mathcal{E}} \leq \tfrac{\lambda t^{\lambda-3}}{2}\left(\lambda(\lambda-2) - \tfrac{\mu t^2}{(2-\alpha)\sigma_1^2(\boldsymbol{A})}\right)\big\|\boldsymbol{A}(X-x^\star)\big\|^2 \\ + \lambda\left(\tfrac{\lambda}{2}-1\right)t^{\lambda-2}\big\langle X-x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\big\rangle.\end{aligned} \tag{4.26}$$

Note that t_0 with $\lambda = 2r/3$ can be written as $t_0 = \sigma_1(\boldsymbol{A})\sqrt{\lambda(\lambda-2)(2-\alpha)\mu^{-1}}$, so that the first term in (4.26) is nonpositive for all $t \geq t_0$, hence

$$\dot{\mathcal{E}} \leq \tfrac{\lambda}{2}\left(\tfrac{\lambda}{2}-1\right)t^{\lambda-2}\tfrac{d}{dt}\|\boldsymbol{A}(X-x^\star)\|^2 \tag{4.27}$$

for all $t \geq t_0$. The strategy is to integrate (4.27) to obtain an upper bound on $\mathcal{E}|_t$, which by the form of (4.22) provides a bound on $\Phi(X)-\Phi^\star$. Thus, integrating (4.27) from t_0 to t, and using integration by parts on the right-hand side gives

$$\begin{aligned}\mathcal{E}|_t - \mathcal{E}|_{t_0} + \tfrac{\lambda}{2}\left(\tfrac{\lambda}{2}-1\right)\Big\{t_0^{\lambda-2}\big\|\boldsymbol{A}\big(X(t_0)-x^\star\big)\big\|^2 \\ + (\lambda-2)\int_{t_0}^t s^{\lambda-3}\big\|\boldsymbol{A}\big(X(s)-x^\star\big)\big\|^2 ds\Big\} \leq \tfrac{\lambda}{2}\left(\tfrac{\lambda}{2}-1\right)t^{\lambda-2}\big\|\boldsymbol{A}\big(X(t)-x^\star\big)\big\|^2.\end{aligned} \tag{4.28}$$

By dropping the two positive terms on the left side of (4.28) (recall $\lambda \geq 2$) we get

$$\mathcal{E}|_t \leq \mathcal{E}|_{t_0} + \tfrac{\lambda}{2}\left(\tfrac{\lambda}{2}-1\right)t^{\lambda-2}\left\|\boldsymbol{A}(X(t)-x^\star)\right\|^2. \tag{4.29}$$

Combining (4.22) (ignore the positive quadratic) with (4.29) we get for $t \geq t_0$ that

$$\frac{\Phi(X)-\Phi^\star}{2-\alpha} \leq \frac{\mathcal{E}|_{t_0}}{t^\lambda} + \frac{\lambda(\lambda-2)}{4t_0^2}\left\|\boldsymbol{A}(X-x^\star)\right\|^2. \tag{4.30}$$

Since Φ is strongly convex and $0 \in \partial\Phi(x^\star)$, it follows using a similar argument as that used to obtain (4.25) that

$$\|\boldsymbol{A}(X - x^\star)\|^2 \le \tfrac{2\sigma_1^2(\boldsymbol{A})}{\mu}\left(\Phi(X) - \Phi^\star\right). \tag{4.31}$$

Using this inequality in the last term of (4.30), recalling the definition of t_0, and using $\lambda \equiv 2r/3$, we obtain

$$\Phi(X(t)) - \Phi^\star \le 2(2-\alpha)\mathcal{E}|_{t_0}\, t^{-\lambda} \tag{4.32}$$

for all $t \ge t_0$. Strong convexity of Φ implies $\|X - x^\star\|^2 \le (2/\mu)\left(\Phi(X) - \Phi^\star\right)$, which combined with (4.32) shows that

$$\|X(t) - x^\star\|^2 \le (2/\mu)\big(\Phi(X) - \Phi^\star\big) \le 4(2-\alpha)\mu^{-1}\mathcal{E}|_{t_0}\, t^{-2r/3}. \tag{4.33}$$

Combing this inequality with the observation from (4.22) yields

$$\mathcal{E}|_{t_0} = \tfrac{t_0^\lambda}{2-\alpha}\left(\Phi_0 - \Phi^\star\right) + \tfrac{\lambda^2 t_0^{\lambda-2}}{2}\|\boldsymbol{A}(X_0 - x^\star)\|^2 + \tfrac{t_0^\lambda}{2}\|\boldsymbol{A}\dot{X}_0\|^2, \tag{4.34}$$

where $X_0 \equiv X(t_0)$, $\Phi_0 \equiv \Phi(X_0)$, and $\dot{X}_0 \equiv \dot{X}(t_0)$ yields (4.17). □

Note that α in (4.16) can improve the constant in the $\mathcal{O}(1/t^2)$ bound. Also, it is possible that the dynamical system (3.8) may not attain linear convergence in the strongly convex case according to the result (4.17).

4.3 Convergence of Relaxed Heavy Ball ADMM

We now turn to the dynamical system (3.17) associated with the R-HB-ADMM updates (3.16). The reader should compare the next result with that of Theorem 4.3.

Theorem 4.4. *Consider* (3.17) *with $r > 0$. Let $x^\star \in \arg\min \Phi(x)$, $\Phi^\star \equiv \Phi(x^\star)$, and $X = X(t)$ be the unique trajectory of the system from $X(0) = x_0$ and $\dot{X}(0) = 0$. If Φ is* convex*, then for all $t \ge t_0 \equiv 1/r$, and with the constant c defined by* (4.43)*,*

$$\Phi(X(t)) - \Phi^\star \le \frac{(2-\alpha)c}{t}. \tag{4.35}$$

*If Φ is μ-*strongly convex*, then for all $t \ge 0$, and with $r \le \bar{r} \equiv \frac{3}{2}\sigma_1^{-1}(\boldsymbol{A})\sqrt{\mu/(2-\alpha)}$,*

$$\|X(t) - x^\star\|^2 \le \tfrac{6}{\mu}\big(\Phi(x_0) - \Phi^\star\big)e^{-2rt/3}. \tag{4.36}$$

Proof. First, consider

$$\mathcal{E}(X,t) \equiv \tfrac{t}{2-\alpha}\left(\Phi(X) - \Phi^\star\right) + \tfrac{r}{2}\|\boldsymbol{A}(X - x^\star)\|^2 + \tfrac{t}{2}\|\boldsymbol{A}\dot{X}\|^2 + \langle X - x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\rangle \tag{4.37}$$

where $x^\star$ is some minimizer of Φ. The total time derivative is

$$\begin{aligned}\dot{\mathcal{E}} = {} & \tfrac{1}{2-\alpha}\left(\Phi(X) - \Phi^\star\right) + \tfrac{t}{2-\alpha}\dot{\Phi}(X) + r\big\langle X - x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\big\rangle \\ & + t\big\langle \dot{X}, \boldsymbol{A}^T\boldsymbol{A}\ddot{X}\big\rangle + \big\langle X - x^\star, \boldsymbol{A}^T\boldsymbol{A}\ddot{X}\big\rangle + \|\boldsymbol{A}\dot{X}\|^2.\end{aligned} \tag{4.38}$$

Using Lemma 4.1 for the second term and (3.17) we obtain

$$\dot{\mathcal{E}} = \tfrac{1}{2-\alpha}\big(\Phi(X) - \Phi^\star - \big\langle X - x^\star, \xi\big\rangle\big) + (1 - tr)\|\boldsymbol{A}\dot{X}\|^2 \tag{4.39}$$

where $\xi \equiv -(2-\alpha)\boldsymbol{A}^T\boldsymbol{A}\big(\ddot{X}(t) + r\dot{X}(t)\big) \in \partial\Phi(X(t))$. From convexity of Φ (i.e., the relation (2.2) with $\mu = 0$), it follows that the first term above is negative. Since the second term is nonpositive for all $t \ge t_0 = 1/r$, we conclude that $\dot{\mathcal{E}} \le 0$ for all $t \ge t_0$. We now proceed to show that $\mathcal{E} \ge 0$. To this end, we can write (4.37) in the form

$$\mathcal{E}(X) = \tfrac{t}{2-\alpha}\left(\Phi(X) - \Phi^\star\right) + \widetilde{\mathcal{E}}(X), \tag{4.40}$$

$$\widetilde{\mathcal{E}}(X) \equiv \tfrac{r}{2}\|\boldsymbol{A}(X - x^\star)\|^2 + \big\langle X - x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\big\rangle + \tfrac{t}{2}\|\boldsymbol{A}\dot{X}\|^2. \tag{4.41}$$

Now, for all $t \ge t_0 = 1/r$, it follows that

$$\begin{aligned} \widetilde{\mathcal{E}}(X) &\ge \tfrac{r}{2}\|\boldsymbol{A}(X - x^\star)\|^2 + \big\langle X - x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\big\rangle + \tfrac{1}{2r}\|\boldsymbol{A}\dot{X}\|^2 \\ &= \tfrac{1}{2}\|(1/\sqrt{r})\boldsymbol{A}\dot{X} + \sqrt{r}\boldsymbol{A}(X - x^\star)\|^2, \end{aligned} \tag{4.42}$$

thus $\mathcal{E} \ge 0$. Since $\dot{\mathcal{E}} \le 0$, we know that $\mathcal{E}|_t \le \mathcal{E}|_{t_0}$ for all $t \ge t_0$. Therefore, after dropping the nonnegative term $\widetilde{\mathcal{E}}$ in (4.40), we obtain $\Phi(X(t)) - \Phi^\star \le (2-\alpha)\mathcal{E}|_{t_0}\, t^{-1}$. This gives (4.35) once we observe from (4.37) that

$$c \equiv \mathcal{E}_{t_0} = \tfrac{1}{r(2-\alpha)}\left(\Phi(X_0) - \Phi^\star\right) + \tfrac{1}{2}\|\sqrt{r}\boldsymbol{A}(X_0 - x^\star) + \tfrac{1}{\sqrt{r}}\boldsymbol{A}\dot{X}_0\|^2 \tag{4.43}$$

where $X_0 \equiv X(t_0)$ and $\dot{X}_0 \equiv \dot{X}(t_0)$.

Second, let $r \in (0, \bar{r}]$ and define

$$\mathcal{E}(X, t) \equiv e^{2rt/3}\Big\{\tfrac{\Phi(X) - \Phi^\star}{2-\alpha} + \tfrac{r^2}{9}\|\boldsymbol{A}(X - x^\star)\|^2 + \tfrac{1}{2}\|\boldsymbol{A}\dot{X}\|^2 + \tfrac{2r}{3}\langle\boldsymbol{A}(X - x^\star), \boldsymbol{A}\dot{X}\rangle\Big\} \tag{4.44}$$

where $x^\star$ is the unique minimizer of Φ. Taking its total time derivative, using (3.17), and Lemma 4.1 one obtains

$$\dot{\mathcal{E}} \le \tfrac{2r}{3(2-\alpha)}e^{2rt/3}\left(\Phi(X) - \Phi^\star - \big\langle X - X^\star, \xi\big\rangle\right) + \tfrac{2r^3}{27}e^{2rt/3}\|\boldsymbol{A}(X - X^\star)\|^2, \tag{4.45}$$

for the same $\xi \in \partial\Phi(X(t))$ defined earlier in this proof. From (2.2) we have the inequality (4.25), which applied to the first term of (4.45) results in

$$\dot{\mathcal{E}} \le e^{2rt/3}\tfrac{r}{3}\left(\tfrac{2r^2}{9} - \tfrac{\mu}{(2-\alpha)\sigma_1^2(\boldsymbol{A})}\right)\|\boldsymbol{A}(X - x^\star)\|^2. \tag{4.46}$$

It follows that $\dot{\mathcal{E}} \le 0$. Next, notice that $\mathcal{E}$, as defined in (4.44), can be written as

$$\mathcal{E} = \tfrac{1}{2(2-\alpha)}e^{2rt/3}(\Phi(X) - \Phi^*) + \widetilde{\mathcal{E}}, \tag{4.47}$$

$$\widetilde{\mathcal{E}} \equiv e^{2rt/3}\Big\{\tfrac{\Phi(X) - \Phi^\star}{2(2-\alpha)} + \tfrac{r^2}{9}\|\boldsymbol{A}(X - x^\star)\|^2 + \tfrac{1}{2}\|\boldsymbol{A}\dot{X}\|^2 + \tfrac{2r}{3}\big\langle\boldsymbol{A}(X - x^\star), \boldsymbol{A}\dot{X}\big\rangle\Big\}. \tag{4.48}$$

Note that (4.31) holds under our current assumptions, thus by defining

$$c_1 \equiv \tfrac{\mu}{4(2-\alpha)\sigma_1^2(\boldsymbol{A})} + \tfrac{r^2}{9}, \qquad c_2 \equiv \tfrac{r}{3}, \qquad a \equiv \|\boldsymbol{A}(X - x^\star)\|, \qquad b \equiv \|\boldsymbol{A}\dot{X}\|, \tag{4.49}$$

we have $\widetilde{\mathcal{E}} \geq e^{2rt/3}\big\{c_1 a^2 + \frac{1}{2}b^2 - 2c_2 ab\big\}$, i.e.,

$$\widetilde{\mathcal{E}} \geq e^{2rt/3}\left((\sqrt{c_1}a - c_2/\sqrt{c_1}b)^2 + \left(1/2 - c_2^2/c_1\right) b^2\right). \tag{4.50}$$

We conclude that $\widetilde{\mathcal{E}} \geq 0$ since $c_2^2 \leq c_1/2$ due to (4.49) and the definition of $\bar{r}$.

Since $\dot{\mathcal{E}} \leq 0$ for all $t \geq 0$, we have $\mathcal{E}|_t \leq \mathcal{E}|_{t=0}$. Using (4.47) together with $\widetilde{\mathcal{E}} \geq 0$,

$$\Phi(X(t)) - \Phi^\star \leq 2(2-\alpha)\mathcal{E}|_{t=0}\, e^{-2rt/3}. \tag{4.51}$$

Strong convexity of Φ implies $\|X - x^\star\|^2 \leq (2/\mu)\left(\Phi(X) - \Phi^\star\right)$. Using (4.51) yields

$$\|X(t) - x^\star\|^2 \leq 4(2-\alpha)\mu^{-1}\mathcal{E}|_{t_0}\, e^{-2rt/3}. \tag{4.52}$$

Finally, (4.47), (4.48), the initial condition $\dot{X}(0) = 0$, (4.31), and $r \leq \bar{r}$ yield

$$\begin{aligned}\mathcal{E}|_{t_0} &= \tfrac{1}{2-\alpha}\left(\Phi(x_0) - \Phi^\star\right) + \tfrac{r^2}{9}\|\boldsymbol{A}(x_0 - x^\star)\|^2 \\ &\leq \left(\tfrac{1}{2-\alpha} + \tfrac{2r^2\sigma_1^2(A)}{9\mu}\right)\left(\Phi(x_0) - \Phi^\star\right) \leq \tfrac{3}{2(2-\alpha)}\left(\Phi(x_0) - \Phi^\star\right).\end{aligned} \tag{4.53}$$

Combining this inequality with (4.52) gives (4.36). □

We mention some noteworthy remarks regarding Theorem 4.3 and Theorem 4.4.

- Over-relaxation (i.e., $\alpha \in (1,2)$) can improve convergence in some cases more than others. For instance, in the convex case (4.16) the improvement in the constant is linear in α, while in the strongly convex case (4.36) α appears inside an exponential. An optimal choice of α depends on the problem and the values of the other parameters.

- Although R-HB-ADMM attains an exponential convergence rate in (4.36), the damping constant r must be below a certain threshold. In contrast, R-A-ADMM achieves only a sublinear rate in (4.17), but r is unconstrained. In our experiments we observed that R-HB-ADMM outperforms R-A-ADMM in most cases for appropriate values of r in the strongly convex setting.

- Theorem 4.4 applies to the heavy ball method (1.4) as a special case for which an $\mathcal{O}(1/k)$ rate was obtained in a Cesàro average sense [4]. In the strongly convex case, exponential convergence of heavy ball is known [4,5] but not in the form (4.36).

- Comparing Theorem 4.3 and Theorem 4.4 we see an interesting tradeoff between the two types of acceleration—Nesterov versus heavy ball—in convex versus strongly convex settings. Nesterov's acceleration achieves the optimal $\mathcal{O}(1/t^2)$ rate in the convex setting, but only the power law $\mathcal{O}(t^{-2r/3})$ in the strongly convex setting. On the other hand, the heavy ball type of acceleration has $\mathcal{O}(1/t)$ for the convex case but can recover exponential convergence $\mathcal{O}(e^{-2rt/3})$ for the strongly convex case. This suggests that in settings of higher curvature of Φ, heavy ball type of acceleration might be preferable.

5 Nonsmooth Hamiltonian Formalism

Hamiltonian systems are ubiquitous and have a special mathematical structure. Thus, we cannot omit the fact that both (3.8) and (3.17) are (nonsmooth) Hamiltonian systems. Actually, these systems admit two different Hamiltonian formalisms. In the standard Hamiltonian formalism, the system is described by a Hamiltonian $H = H(X, P; t)$, where $(X, P) \in \mathbb{R}^{2n}$ are the canonical position and momentum variables, and the equations of motion are

$$\dot{X} = \nabla_P H, \qquad \dot{P} = -\nabla_X H. \tag{5.1}$$

A generalization of the Hamiltonian formalism to nonsmooth systems was proposed by Rockafellar [30–32] (see also [44]), which allows us to replace the equalities in (5.1) by inclusions and gradients by subdifferentials. Thus, consider the Hamiltonian

$$H \equiv \tfrac{1}{2} e^{-\eta(t)} \big\langle P, \boldsymbol{M}^{-1} P \big\rangle + \lambda e^{\eta(t)} \Phi(X), \tag{5.2}$$

where $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ is a "mass matrix" and λ is a "coupling constant" that measures the strength of the potential Φ. The Hamiltonian (5.2) is a generalization of the one associated to the damped harmonic oscillator [45, 46]. Therefore, using (5.2) and the nonsmooth version of (5.1) we obtain

$$\dot{X} = e^{-\eta} \boldsymbol{M}^{-1} P, \qquad \dot{P} \in -e^{\eta} \lambda \partial \Phi(X). \tag{5.3}$$

This pair is equivalent to the second-order system

$$\lambda^{-1} \boldsymbol{M} \big(\ddot{X} + \dot{\eta} \dot{X} \big) \in -\partial \Phi(X). \tag{5.4}$$

Note that (3.8) and (3.17) are particular cases of (5.4) with $\boldsymbol{M} = \boldsymbol{A}^T \boldsymbol{A}$, $\lambda = (2 - \alpha)^{-1}$, and $\eta = r \log t$ for (3.8), while $\eta = rt$ for (3.17). Physically, the entries in $\boldsymbol{A}^T \boldsymbol{A}$ play the role of particle masses and the relaxation parameter α controls the coupling constant. It is interesting to see how these parameters have an intuitive interpretation. A large entry in $\boldsymbol{A}^T \boldsymbol{A}$ would correspond to a heavy particle whose inertia would make it difficult to be slowed down close to a minimum (thus creating oscillations), or difficult to accelerate in a flat region. The relaxation parameter α affects convergence since it can strengthen or weaken the amount of movement in the direction of subgradients.

A second Hamiltonian representation can be obtained from a conformal Hamiltonian formalism [33]. We use a time-independent Hamiltonian and introduce dissipation by directly modifying Hamilton's equations (5.1) with the addition of a linear term in the momentum. Taking into account the nonsmoothness of Φ we thus have

$$\dot{X} = \nabla_P H, \qquad \dot{P} \in -\partial_X H - \dot{\eta} P. \tag{5.5}$$

Consider the Hamiltonian

$$H \equiv \tfrac{1}{2} \big\langle P, \boldsymbol{M}^{-1} P \big\rangle + \lambda \Phi(X) \tag{5.6}$$

It follows from (5.5) that

$$\dot{X} = \boldsymbol{M}^{-1} P, \qquad \dot{P} \in -\lambda \partial \Phi(X) - \dot{\eta} P. \tag{5.7}$$

This pair can be equivalently written as the second-order system (5.4), which means that it is also equivalent to (3.8) and (3.17). Above, we allowed the damping term in (5.7) to depend on time.

The conformal case, however, requires $\dot{\eta}$ to be constant [33]. Thus, the heavy ball case, $\eta = rt$, is of particular interest.[2] Conformal Hamiltonian systems are interesting since they have a well-defined symplectic structure and discretizations can preserve the phase portrait of the flow map; see our recent paper [48] for more details. An interesting problem that we leave for future work is to consider discretizations of the equations of motion (5.3) or (5.7). Since these equations are nonsmooth, one can impose discontinuous forces on the system.

6 Numerical Experiments

We consider two numerical experiments to verify the behavior of the variants of ADMM previously introduced. We also want to compare the performance of R-HB-ADMM, which in the continuous limit has a constant damping, versus R-A-ADMM whose continuous limit has an asymptotically vanishing damping. Note that A-ADMM (with $\alpha = 1$) was already proposed in [28]. However, R-HB-ADMM is a completely new method in the literature.

6.1 Trend filtering with ℓ_1-regularization

One can estimate piecewise linear trends in time series data by solving [49]

$$\min_{x,z} \tfrac{1}{2}\|y - x\|^2 + \lambda\|z\|_1 \quad \text{subject to} \quad z = \boldsymbol{D}x \tag{6.1}$$

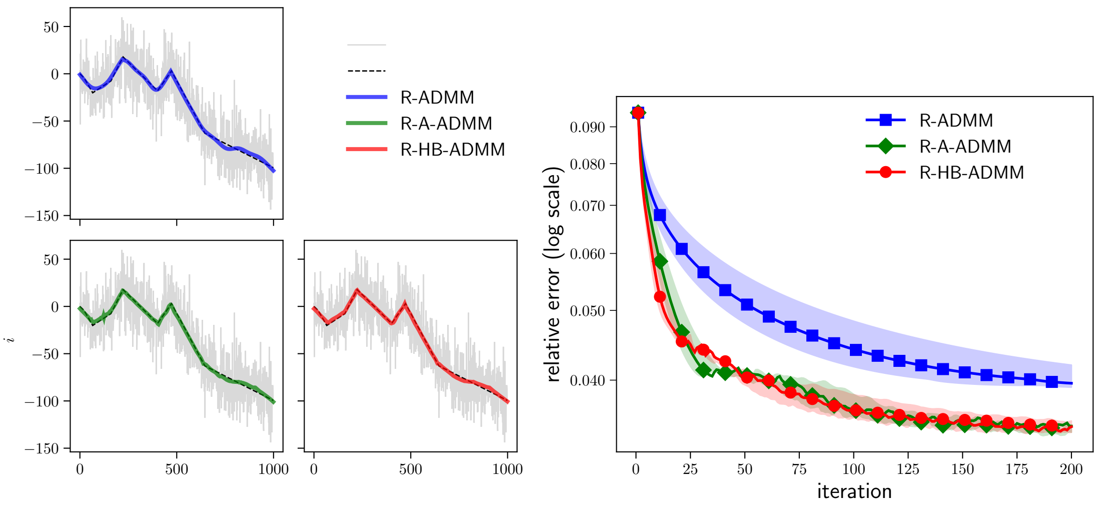
where $y \in \mathbb{R}^n$ is a given signal, $\boldsymbol{D} \in \mathbb{R}^{(n-2)\times n}$ is a Toeplitz matrix with first row $(1, -2, 1, 0, \dots, 0)$, and $\lambda > 0$ is the regularization parameter. (For this type of problem λ has to be quite large [49]). The problem is well-suited to the previous ADMM variants, where the proximal operators of $f(x)$ and $g(z)$ are well-known [18].

Let us consider the same example as in [49], namely

$$y_i = x_i + \xi_i, \quad x_{i+1} = x_i + v_i, \quad x_0 = 0, \tag{6.2}$$

for $i = 1, \dots, n$, where x_i is the true underlying trend that is superimposed with noise $\xi_i \sim \mathcal{N}(0, \sigma^2)$, and the slopes are generated by a Markov process where $v_{i+1} = v_i$ with probability p, and $v_i \sim \mathcal{U}(-b, b)$ with probability $(1 - p)$ for some b. The goal is to recover x given the observed signal y. We consider one sample of this model with $n = 1000$, $p = 0.99$, $\sigma = 20$, and $b = 0.5$. The trends recovered by variants of ADMM are shown in Fig. 1a, and their respective convergence rates are shown in Fig. 1b. Note that both accelerated variants R-A-ADMM and R-HB-ADMM are faster than R-ADMM. Since (6.1) is a convex problem, based only on Theorem 4.3 and Theorem 4.4, one would expect R-A-ADMM to be faster than R-HB-ADMM. However, in practice, they perform similarly. From Fig. 1b we also see that the effect of α on the convergence rate is consistent with the theoretical predictions.

[2] Recently, conformal Hamiltonian formulations were studied in *smooth* optimization [47, 48].



(a) (b)

Figure 1: (a) Piecewise linear trends in a time series (6.2) by solving (6.1) with variants of ADMM. We choose $\lambda = 2500$, $\rho = 5 \times 10^2$ and $\alpha = 1.35$. For R-A-ADMM we use the standard $r = 3$, and for HB-ADMM $r = 1.5$. We run each method for 200 iterations. Note how the accelerated variants provide more accurate recovery. (b) We show the relative error $\|\hat{x}_k - x\|/\|x\|$ where $\hat{x}_k$ is the estimate of the true trend x at iteration k. R-ADMM achieved 0.039, and both R-A-ADMM and R-HB-ADMM achieved 0.034. As a comparison, CVXPY with its default setup (using an interior point method) achieved 0.038 in 250 iterations. The solid lines corresponds to $\alpha = 1$ and the shaded areas correspond to a choice of $\alpha \in [0.65, 1.35]$ ($\alpha > 1$ is faster).

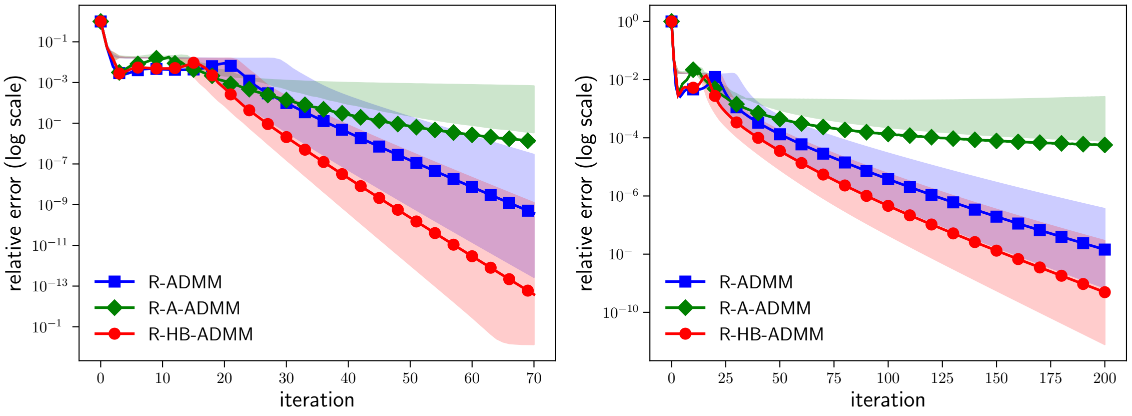
6.2 Robust principal component analysis

Suppose we are given a matrix $\boldsymbol{M} = \boldsymbol{X}^\star + \boldsymbol{Z}^\star \in \mathbb{R}^{n\times m}$ where $\boldsymbol{X}^\star$ has low rank and $\boldsymbol{Z}^\star$ is sparse. Under certain rank and sparsity conditions, it is possible to exactly recover both components $\boldsymbol{X}^\star$ and $\boldsymbol{Z}^\star$ from observation of $\boldsymbol{M}$ alone. This is achieved by solving the convex problem [50]

$$\min \|\boldsymbol{X}\|_* + \lambda \|\boldsymbol{Z}\|_1 \quad \text{s.t.} \quad \boldsymbol{X} + \boldsymbol{Z} = \boldsymbol{M}, \tag{6.3}$$

where $\lambda = 1/\max\{n, m\}$, and $\|\boldsymbol{X}\|_* = \sum_i \sigma_i(\boldsymbol{X})$ is the nuclear norm. This problem is known as *Robust Principal Component Analysis* (robust PCA) and can be seen as an idealized version of PCA for highly corrupted data. PCA is arguably one of the most used techniques for dimensionality reduction. Robust PCA also finds important applications in statistics, signal processing, and machine learning.

It has been noted [50] that standard ADMM ($\alpha = 1$) with $\rho = 1$ is very effective in solving (6.3), being faster and more accurate than several methods that include nonsmooth extensions of Nesterov's method (1.5). We wish to verify whether the accelerated variants of ADMM are able to improve over standard ADMM. The proximal operators for $\|\boldsymbol{X}\|_*$ and $\|\boldsymbol{Z}\|_1$ have well-known closed form expressions [50]. The results in a setting where exact recovery is possible are shown in Fig. 2a. A more challenging case that is close to the phase transition boundary where exact recovery fails is shown in Fig. 2b. Interestingly, our R-HB-ADMM improves over R-ADMM, whereas R-A-



(a) (b)

Figure 2: (a) Robust PCA (6.3) where $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is generated with $\boldsymbol{X}^\star = \boldsymbol{M}_1 \boldsymbol{M}_2^T$, where $\boldsymbol{M}_{1,2} \sim \mathcal{N}(0, 1/n)$ are $n \times q$ matrices and $\boldsymbol{Z}^\star \in \{-1, 0, 1\}^{n\times n}$ has Bernoulli ± 1 entries and support of size s chosen uniformly at random. We report the relative error $\|\boldsymbol{X}_k + \boldsymbol{Z}_k - \boldsymbol{M}\|/\|\boldsymbol{M}\|$ versus k. We choose $q = 0.05 \times n$ and $s = 0.1 \times n^2$ for $n = 1000$ [50]. We fix $\rho = 1$. For R-A-ADMM, $r = 3$ which is standard, and for R-HB-ADMM, $r = 0.75$. Solid lines correspond to $\alpha = 1$ and the shaded areas $\alpha \in [0.7, 1.3]$. For R-A-ADMM, different choices of α did not improve convergence. (b) Same setting and parameters but with $q = 0.2 \times n$ and $s = 0.1 \times n^2$ for $n = 1000$. This is close to the phase transition where exact recovery is impossible [50].

ADMM does not. The improvement of R-ADMM and R-HB-ADMM with $\alpha > 1$ agree with the previous theoretical predictions for the continuous systems. In these examples, choosing $\alpha > 1$ for R-A-ADMM did not improve over $\alpha = 1$.

Although (6.3) is a convex optimization problem, the plots in Figs. 2a and 2b show that R-ADMM exhibits linear convergence for sufficiently large k. Based on Theorem 4.2 this suggests that there is a region in which the objective function behaves as a strongly convex function. Hence, R-HB-ADMM also attains linear convergence as predicted in Theorem 4.4. On the other hand, Theorem 4.3 tells us that R-A-ADMM is unlikely to attain linear convergence, as also reflected in Fig. 2a. Therefore, these empirical results are consistent with our theoretical predictions.

7 Conclusions

We introduced two new families of *relaxed* and *accelerated* ADMM algorithms. The first follows Nesterov's type of acceleration and is given by the updates in (3.7). The second is inspired by Polyak's heavy ball method and is given by the updates in (3.16). We then derived differential inclusions (nonsmooth dynamical systems) that model the leading order behavior of these algorithms in the continuous-time limit. This extends prior work by accounting for nonsmooth problems and allowing for linear constraints (see Theorems 3.1, 3.3 and 3.5). Moreover, we obtained rates of convergence for the continuous dynamical systems in convex and strongly convex settings through a nonsmooth Lyapunov analysis; see Theorems 4.2, 4.3 and 4.4. The complexity results obtained in this paper are summarized in Table 1, most of which are new to the best of our knowledge. The convergence analysis for the nonsmooth dynamical systems considered in this paper can serve as a useful guide in studying algorithms arising as discretizations.

These results strengthen the connections between optimization and continuous dynamical systems. The proof techniques in continuous-time may shed light in obtaining analogous rates in discrete-time. The tradeoff between Nesterov and heavy ball acceleration also deserves additional investigation. In many numerical experiments the latter seems to have benefits. Once provided with appropriate continuous systems, a natural direction is to consider other discretizations, and symplectic integrators offer a promising potential alternative.

Acknowledgements

This work is supported by grants ARO MURI W911NF-17-1-0304 and NSF 1704458.

References

[1] Y. Nesterov, "A Method of Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$," *Soviet Mathematics Doklady* **27** no. 2, (1983) 372–376.

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

[3] B. T. Polyak, "Some Methods of Speeding Up the Convergence of Iteration Methods," *USSR Computational Mathematics and Mathematical Physics* **4** no. 5, (1964) 1–17.

[4] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson, "Global Convergence of the Heavy-Ball Method for Convex Optimization," in *2015 European Control Conference (ECC)*, pp. 310–315. 2015.

[5] B. Polyak and P. Shcherbakov, "Lyapunov Functions: An Optimization Theory Perspective," *IFAC-PapersOnLine* **50** no. 1, (2017) 7456–7461. 20th IFAC World Congress.

[6] W. Su, S. Boyd, and E. J. Candès, "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights," *Journal of Machine Learning Research* **17** no. 153, (2016) 1–43.

[7] A. Wibisono, A. C. Wilson, and M. I. Jordan, "A Variational Perspective on Accelerated Methods in Optimization," *Proceedings of the National Academy of Sciences* **113** no. 47, (2016) E7351–E7358.

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

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

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

[11] H. Attouch and J. Peypouquet, "The Rate of Convergence of Nesterov's Accelerated Forward-Backward Method is Actually Faster Than $1/k^2$," *SIAM J. Optim.* **26** no. 3, (2016) 1824–1834.

[12] A. Cauchy, "Méthode générale pour la résolution des systèmes d'équations simultanées," *C. R. Acad. Sci. Paris* **25** (1847) 536–538.

[13] H. Attouch and A. Cabot, "Convergence of damped inertial dynamics governed by regularized maximally monotone operators," *J. Differential Equations* **264** no. 12, (2018) .

[14] R. May, "Asymptotic for a Second-Order Evolution Equation with Convex Potential and Vanishing Damping Term," *Turkish Journal of Mathematics* **41** (2016) 681–785.

[15] H. Attouch and A. Cabot, "Convergence Rates of Inertial Forward-Backward Algorithms," *SIAM J. Optim.* **28** no. 1, (2018) 849–874.

[16] D. Gabay and B. Mercier, "A Dual Algorithm for the Solution of Nonlinear Variational Problems via Finite Element Approximations," *Computers and Mathematics with Applications* **2** no. 1, (1976) 17–40.

[17] R. Glowinski and A. Marroco, "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** no. R2, (1975) 41–76.

[18] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning* **3** no. 1, (2011) 1–122.

[19] B. He and X. Yuan, "On the $O(1/n)$ Convergence Rate of the DouglasRachford Alternating Direction Method," *SIAM Journal on Numerical Analysis* **50** no. 2, (2012) 700–709.

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

[21] W. Deng and W. Yin, "On the Global and Linear Convergence of the Generalized Alternating Direction Method of Multipliers," *Journal of Scientific Computing* **66** no. 3, (2016) 889–916.

[22] J. Eckstein, "Parallel Alternating Direction Multiplier Decomposition of Convex Programs," *Journal of Optimization Theory and Applications* **80** no. 1, (1994) 39–62.

[23] J. Eckstein and M. C. Ferris, "Operator-Splitting Methods for Monotone Affine Variational Inequalities, with a Paralell Application to Optimal Control," *INFORMS Journal on Computing* **10** (1998) 218–235.

[24] D. Davis and W. Yin, "Faster Convergence Rates of Relaxed Peaceman-Rachford and ADMM Under Regularity Assumptions," *Mathematics of Operations Research* **42** no. 3, (2017) 783–805.

[25] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. I. Jordan, "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.

[26] P. Giselsson and S. Boyd, "Diagonal scaling in Douglas-Rachford splitting and ADMM," in *53rd IEEE Conference on Decision and Control*, pp. 5033–5039. 2014.

[27] G. França and J. Bento, "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.

[28] T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk, "Fast Alternating Direction Optimization Methods," *SIAM Journal on Imaging Sciences* **7** no. 3, (2014) 1588–1623.

[29] G. França, D. P. Robinson, and R. Vidal, "ADMM and Accelerated ADMM as Continuous Dynamical Systems," *International Conference on Machine Learning* (2018) . arXiv:1805.06579 [math.OC].

[30] R. T. Rockafellar, "Generalized Hamiltonian Equations for Convex Problems of Lagrange," *Pacific J. Math.* **33** (1970) 411–428.

[31] P. Lowen and R. T. Rockafellar, "The Adjoint Arc in Nonsmooth Optimization," *Trans. Amer. Math. Soc.* **325** (1991) 39–72.

[32] P. Lowen and R. T. Rockafellar, "Optimal Control of Unbounded Differential Inclusions," *SIAM J. Control Opt.* **32** (1994) 442–470.

[33] R. McLachlan and M. Perlmutter, "Conformal Hamiltonian Systems," *J. Geometry and Physics* **39** (2001) 276–300.

[34] R. T. Rockafellar, *Convex Analysis.* Princeton University Press, 1996.

[35] J.-P. Aubin and A. Cellina, *Differential Inclusions.* Springer-Verlag, 1984.

[36] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization.* Springer, 2000.

[37] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations.* John Wiley & Sons, 2008.

[38] A. Dontchev and F. Lempio, "Difference Methods for Differential Inclusions: A Survey," *SIAM Review* **34** (1992) 263–294.

[39] G. Grammel, "Towards Fully Discretized Differential Inclusions," *Set-Valued Analysis* **11** (2003) 1–8.

[40] T. Taniguchi, "Global Existence of Solutions of Differential Inclusions," *Journal of Mathematical Analysis and Applications* **166** (1992) 41–51.

[41] A. Cellina and A. Ornelas, "Existence of Solutions to Differential Inclusions and to Time Optimal Control Problems in the Autonomous Case," *SIAM J. Control Optim.* **42** (2003) 260–265.

[42] F. H. Clarke, *Nonsmooth Analysis and Control Theory.* Springer, 2013.

[43] A. Bacciotti and F. Ceragioli, "Stability and Stabilization of Discontinuous Systems and Nonsmooth Lyapunov Functions," *ESAIM: Control, Optimisation and Calculus of Variations* **4** (1999) 361–376.

[44] A. Ioffe, "Euler-Lagrange and Hamiltonian Formalisms in Dynamic Optimization," *Trans. Amer. Math. Soc.* **349** (1997) 2871–2900.

[45] H. Bateman, "On Dissipative Systems and Related Variational Principles," *Physical Review* **38** no. 10, (1931) 815–819.

[46] N. A. Lemos, "Canonical Approach to the Damped Harmonic Oscillator," *American Journal of Physics* **47** no. 10, (1979) 857–858.

[47] C. J. Maddison, D. Paulin, Y. W. Teh, B. O'Donoghue, and A. Doucet, "Hamiltonian Descent Methods." arXiv:1809.05042 [math.OC], 2018.

[48] G. França, J. Sulam, D. P. Robinson, and R. Vidal, "Conformal Symplectic and Relativistic Optimization." arXiv:1903.04100 [math.OC], 2019.

[49] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky, "ℓ_1 Trend Filtering," *SIAM Review* **51** no. 2, (2009) 339–360.

[50] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust Principal Component Analysis?," *Journal of the ACM* **58** no. 11, (2011) .