ISE

Industrial and
Systems Engineering

A Dynamical Systems Perspective on Non-smooth Constrained Optimization

Guilherme Franca[1], Daniel P. Robinson[2], and Rene Vidal[1]

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

[2]Industrial and Systems Engineering, Lehigh University

ISE Technical Report 19T-021

LEHIGH
UNIVERSITY

A Dynamical Systems Perspective on Nonsmooth Constrained Optimization

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

Mathematical Institute for Data Science,
Johns Hopkins University, Baltimore MD 21218, USA
and
Industrial and Systems Engineering,
Lehigh University, Bethlehem PA 18015, USA

Abstract

The acceleration technique introduced by Nesterov for gradient descent is widely used in machine learning 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). We introduce two new ADMM variants, one based on Nesterov's acceleration and the other inspired by Polyak's heavy ball method, and derive differential inclusions modelling these algorithms in the continuous-time limit. Using a nonsmooth Lyapunov analysis, we obtain rate-of-convergence results for these dynamical systems in the convex and strongly convex setting that illustrate an interesting tradeoff between Nesterov and heavy ball acceleration.

1 Introduction

Accelerated gradient based methods A popular method to accelerate the convergence of gradient descent was proposed in the seminal paper [1]. In the convex case, accelerated gradient descent attains a convergence rate of $O(1/k^2)$ 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

[*] guifranca@jhu.edu

introduced in [3], called the heavy ball method, which is known to have a convergence rate of $O(1/k)$ for convex functions and linear convergence for strongly convex functions [4, 5]. Nonetheless, the mechanisms by which momentum speedup optimization algorithms are still not very well-understood.

Recently, there has been significant progress in better understanding acceleration by analyzing a differential equation modeling the continuous-time limit of Nesterov's method [6]. Followup 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 for both continuous and discrete settings [8–11]. However, such connections have been limited mostly to gradient descent based methods for minimizing unconstrained differentiable functions.

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

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

which can be seen as a discretization of the gradient flow

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

where $\epsilon > 0$ is the discretization stepsize, $X(t)$ is a continuous function of time such that $x_k = X(t)$ with $t = k\epsilon$, and $\dot{X} \equiv \frac{dX}{dt}$. Interestingly, this connection was known to Cauchy [12] since a long time ago. It is not hard to show that the differential equation (2) has a convergence rate of $O(1/t)$, which matches that of gradient descent (1).

A second example is the *heavy ball* method [3],

$$x_{k+1} - x_k - \epsilon_1 (x_k - x_{k-1}) = -\epsilon_2 \nabla \Phi(x_k), \tag{3}$$

which is a discretization of

$$\ddot{X}(t) + a_1 \dot{X}(t) = -a_2 \nabla \Phi(X(t)), \tag{4}$$

where $\epsilon_1, \epsilon_2 > 0$, a_1 and a_2 are constants, and $\ddot{X} \equiv \frac{d^2X}{dt^2}$.

A third example is Nesterov's *accelerated gradient descent* [1], whose updates are given by

$$x_{k+1} - \hat{x}_k = -\epsilon \nabla \Phi(\hat{x}_k), \tag{5a}$$

$$\hat{x}_{k+1} - x_k = \tfrac{k}{k+3}(x_{k+1} - x_k). \tag{5b}$$

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

$$\ddot{X}(t) + \tfrac{3}{t}\dot{X}(t) = -\nabla \Phi(X(t)). \tag{6}$$

This differential equation has a convergence rate of $O(1/t^2)$ for a convex function Φ [6], which matches the optimal $O(1/k^2)$ rate of its discrete counterpart (5). Further convergence properties of (6) over Hilbert spaces were considered in [10].

Recently, some extensions of this continuous-time perspective to nonsmooth optimization problems started to emerge. For instance, convergence of the dynamical system (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 non-smooth, consider the following differential inclusion

$$\ddot{X}(t) + \tfrac{3}{t}\dot{X}(t) + \nabla f(X(t)) \in -\partial g(X(t)), \tag{7}$$

where ∂g is the subdifferential of g. A forward-backward Euler discretization of (7) leads to the accelerated proximal gradient method, which is a proximal version of Nesterov's method (5) [11,14]. Convergence rates of forward-backward proximal algorithms were also considered in [15].

Accelerated ADMM The *alternating direction method of multipliers* (ADMM) [16–18] is an important algorithm for linearly constrained problems, which is well-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 $O(1/k)$ [19, 20], while in the strongly convex case, it converges linearly [21]. Many variants of ADMM exist including one that uses a relaxation strategy, which empirically is known to improve convergence in some cases [22, 23]. However, few theoretical results are known for relaxed ADMM when compared to vanilla ADMM, except for the fact that it has linear convergence for strongly convex functions [24–27].

The first accelerated version of ADMM was proposed by [28], called fast ADMM (here we call it *accelerated ADMM*, or A-ADMM for short). For a composite objective function $f(x)+g(x)$ with f and g both strongly convex, and g quadratic, it was shown that A-ADMM attains a convergence rate of $O(1/k^2)$ [28]; we are not aware of any other convergence rates. Numerical experiments [28] show that A-ADMM may outperform Nesterov's method (5) in some cases.

Very recently, the continuous limit of A-ADMM was considered [29], which generalizes previous results such as (6). It was also shown that the corresponding dynamical system has a convergence rate of $O(1/t^2)$ under a mere convexity assumption. However, both f and g were assumed to be differentiable and the constraint matrix $\boldsymbol{A}$ was assumed to have full column rank (see (8)). In this paper, we analyze a more general ADMM framework in the fully nonsmooth case and provide analyses that complement the results of [29] in several aspects.

	convex	strongly convex
ADMM	$O\left(\frac{\sigma_1^2(\boldsymbol{A})}{t}\right)$	$O\left(\kappa(\boldsymbol{A})e^{-\mu t/(2\sigma_1^2(\boldsymbol{A}))}\right)$
A-ADMM†	$O\left(\frac{(r-1)^2\sigma_1^2(\boldsymbol{A})}{t^2}\right)$	$O\left(\frac{(r\sigma_1(\boldsymbol{A}))^{2r/3}}{\mu^{r/3}}\frac{1}{t^{2r/3}}\right)$
R-ADMM†	$O\left(\frac{\sigma_1^2(\boldsymbol{A})}{\alpha t}\right)$	$O\left(\kappa(\boldsymbol{A})e^{-\mu\alpha t/(2\sigma_1^2(\boldsymbol{A}))}\right)$
R-A-ADMM‡	$O\left(\frac{(r-1)^2\sigma_1^2(\boldsymbol{A})}{\alpha t^2}\right)$	$O\left(\frac{(r\sigma_1(\boldsymbol{A}))^{2r/3}}{\mu^{r/3}\alpha^{r/3}}\frac{1}{t^{2r/3}}\right)$
R-HB-ADMM‡	$O\left(\frac{r\sigma_1^2(\boldsymbol{A})}{\alpha t}\right)$	$O\left(\alpha^{-1}r^2\sigma_1^2(\boldsymbol{A})e^{-2rt/3}\right)$

Table 1: Convergence rates of the dynamical systems related to relaxed and accelerated variants of ADMM proposed in this paper; see (16)/(17), (18)/(19) and (20)/(21), which apply to solving problem (8). The relaxation parameter is $\alpha \in (0, 2)$, μ is the strong convexity constant (see (11)), and $r > 0$ is a damping constant. Algorithms marked with † are known, but with previously unknown convergence rates, e.g., for A-ADMM both rates (convex/strongly convex) are unknown while for R-ADMM the $O(1/t)$ rate for the convex case seems to be unknown. Those marked with ‡ indicate a new family of algorithms.

Paper contributions We propose new variants of accelerated ADMM for solving the general problem[1]

$$\min_{x\in\mathbb{R}^n}\left\{\Phi(x) \equiv f(x) + g(\boldsymbol{A}x)\right\}, \tag{8}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$ can *both* be nonsmooth functions, and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $m \geq n$. We consider the known family of *relaxed ADMM* (R-ADMM) algorithms, and introduce two accelerated variants to the R-ADMM scheme: one follows Nesterov's approach, which we refer to as *relaxed and accelerated ADMM* (R-A-ADMM), and the other is closer to Polyak's heavy ball method, which we call *relaxed heavy ball ADMM* (R-HB-ADMM). To the best of our knowledge, this is the first time that acceleration and relaxation are considered jointly.

After introducing these new families of algorithms, we turn our attention to deriving their continuous limits. Since f and g in (8) are nonsmooth, we obtain differential inclusions instead of differential equations. We then obtain rates of convergence for these nonsmooth dynamical systems by constructing appropriate Lyapunov functions in both the convex and strongly convex settings; our results are summarized in Table 1. The results of [29] only consider the smooth and convex setting (we consider the nonsmooth, convex and strongly

[1] Our results can be extended to the common formulation $\min_{x,z}\left\{f(x) + g(z) \,|\, \boldsymbol{A}x + \boldsymbol{B}z = c\right\}$ provided $\boldsymbol{B}$ is invertible [20, 25, 27]. Since $z - \boldsymbol{B}^{-1}c = -\boldsymbol{B}^{-1}\boldsymbol{A}x$ one can easily redefine $\boldsymbol{A}$ to cast the problem into a similar form as (8).

convex settings) and their results correspond to one particular instance of our framework (specifically, the first column for ADMM and A-ADMM in Table 1). Our results in Table 1 show that by adding relaxation an improved constant in the complexity bound is attained. Also, the proposed R-HB-ADMM recovers linear convergence in the strongly convex case, which contrasts with R-A-ADMM. We thus see an interesting tradeoff between Nesterov and heavy ball type of acceleration in the convex versus strongly convex settings. Note also that the linear constraint matrix $\boldsymbol{A}$ is reflected in the convergence rates of Table 1, which may improve convergence depending on its singular values. We note that these rates are attained for the continuous dynamical systems, which suggests that the same rates hold for the discrete algorithms although we do not formally establish them. Nevertheless, most of the rates shown in Table 1 are new and will serve as a guide for establishing the same rates in the discrete case in future research.

Preliminaries and notation Given $x, y \in \mathbb{R}^n$ we let $\|x\| = \sqrt{x^T x}$ denote the norm of x and $\langle x, y\rangle = x^T y$ denote the inner product between x and y. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we denote the largest and smallest singular values of $\boldsymbol{A}$ by $\sigma_1(\boldsymbol{A})$ and $\sigma_n(\boldsymbol{A})$, respectively. The induced matrix norm of $\boldsymbol{A}$ is denoted as $\|\boldsymbol{A}\| = \sigma_1(\boldsymbol{A})$ and the condition number of $\boldsymbol{A}$ is written as $\kappa(\boldsymbol{A}) \equiv \sigma_1(\boldsymbol{A})/\sigma_n(\boldsymbol{A})$.

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 [30] as

$$\partial f(x) = \{\xi \in \mathbb{R}^n \,|\, f(y) - f(x) \ge \langle \xi, y - x\rangle \ \forall y\}\,. \tag{9}$$

The subdifferential set $\partial f(x)$ is always closed and convex, and if f is convex it is also nonempty. Convex and strongly convex functions [30] are defined as follows.

Definition 1 (Convex function)**.** *We say that the function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex if and only if*

$$h(y) \ge h(x) + \langle \xi, y - x\rangle \tag{10}$$

for all $x, y \in \operatorname{dom} h$ and all $\xi \in \partial h(x)$.

Definition 2 (Strongly convex function)**.** *We say that the function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is strongly convex if and only if there exists a constant $\mu > 0$ such that*

$$h(y) \ge h(x) + \langle \xi, y - x\rangle + \frac{\mu}{2}\|y - x\|^2 \tag{11}$$

for all $x, y \in \operatorname{dom} h$ and all $\xi \in \partial h(x)$.

The following assumption is used in this paper.

Assumption 3. *The function $\Phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ in* (8) *is proper, lower semicontinuous, convex, and satisfies* $0 \in \operatorname{int}(\operatorname{dom} g - \boldsymbol{A} \operatorname{dom} f)$.

An important consequence of Assumption 3 is that the subdifferential of two functions f and g composed with a linear map $\boldsymbol{A}$ satisfies ([31, Theorem 3.3.5])

$$\partial(f + g \circ \boldsymbol{A})(x) = \partial f(x) + \boldsymbol{A}^T \partial g(\boldsymbol{A}x), \tag{12}$$

because otherwise it would only hold as an inclusion.

Let us mention two basic relations involving the continuous limit. Let $X = X(t) \in \mathbb{R}^n$ where $t \geq 0$ denotes the continuous-time variable. The corresponding state of an algorithm at discrete-time $k = 0, 1, \ldots$ will be denoted by $x_k \in \mathbb{R}^n$. Let $x_k = X(k\epsilon)$ for some small enough $\epsilon > 0$. Then, in the limit $\epsilon \to 0$, it holds that [32–34]

$$\left(x_{k\pm1} - x_k\right)/\epsilon \to \pm\dot{X}(t), \tag{13}$$

$$\left(x_{k+1} - 2x_k + x_{k-1}\right)/\epsilon^2 \to +\ddot{X}(t). \tag{14}$$

Finally, consider a first order differential inclusion[2]

$$\dot{X}(t) \in F(X(t); t) \tag{15}$$

where $F : \mathbb{R}^n \times \mathbb{R} \rightrightarrows \mathbb{R}^n$ is a multi-valued map. By a solution or a trajectory of (15) we mean a function $\varphi : \mathcal{I} \to \mathbb{R}^n$ with $\mathcal{I} \subseteq \mathbb{R}_+$, such that φ is absolutely continuous and $\dot{\varphi}(t) \in F(\varphi(t); t)$ for almost every $t \in \mathcal{I}$ in the Lesbegue measure sense. If F is lower semicontinuous, closed and convex, then existence of at least one continuously differentiable solution $\varphi : \mathcal{I} \to \mathbb{R}^n$ is guaranteed [35]. In more general cases, for instance if F is unbounded or nonconvex, existence of solutions of differential inclusions can be a delicate issue; see, e.g., [36–38] and references therein. Fortunately, in this paper where $F = -\partial\Phi$ and under Assumption 3, the differential inclusion (15) has a unique solution [35].

2 Variants of ADMM as Dynamical Systems

We first consider the family of R-ADMM algorithms along with two new accelerated variants. We then present nonsmooth dynamical systems that model these algorithms in the continuous limit. The family of ADMM algorithms is developed for problem (8) by introducing the variable $z = \boldsymbol{A}x$ and considering the (scaled) augmented Lagrangian $\mathcal{L}_\rho(x, z, u) = f(x) + g(z) + \rho\langle u, Ax - z\rangle + \frac{\rho}{2}\|Ax - z\|_2^2$, where $u \in \mathbb{R}^m$ is the Lagrange multiplier vector and $\rho > 0$.

[2] Every higher order system can be written in first-order form.

2.1 Relaxed ADMM

Let us start with the R-ADMM framework [18] for problem (8) whose updates are given by

$$x_{k+1} \leftarrow \arg\min_x \left\{ f(x) + \tfrac{\rho}{2}\|\boldsymbol{A}x - z_k + u_k\|^2 \right\}, \tag{16a}$$

$$z_{k+1} \leftarrow \arg\min_z \big\{ g(z) + \tfrac{\rho}{2}\|\alpha \boldsymbol{A}x_{k+1} + (1-\alpha)z_k - z + u_k\|^2 \big\}, \tag{16b}$$

$$u_{k+1} \leftarrow u_k + \alpha \boldsymbol{A}x_{k+1} + (1-\alpha)z_k - z_{k+1}. \tag{16c}$$

The relaxation parameter $\alpha \in (0,2)$ is introduced to speedup convergence [22, 23], and the standard ADMM is recovered when $\alpha = 1$. We now state the continuous limit of the updates (16), and note that the derivation is shown in Appendix A.

Theorem 4. *Consider the R-ADMM updates* (16) *for solving problem* (8) *under Assumption 3. Then, the continuous limit of such updates, with time scale $t = \rho^{-1}k$, is given by the differential inclusion*

$$\alpha^{-1}(\boldsymbol{A}^T\boldsymbol{A})\dot{X}(t) \in -\partial\Phi(X(t)) \tag{17}$$

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

Note that if both functions f and g are differentiable, thus $\partial\Phi(x) = \nabla f(x) + \boldsymbol{A}^T\nabla g(\boldsymbol{A}x)$, $\boldsymbol{A} = I$, and $\alpha = 1$, then (17) reduces to the differential equation (2). In general, however, even in the smooth case, the presence of $\boldsymbol{A}^T\boldsymbol{A}$ in (17) can make the stability properties and rate of convergence of this system different from those of (2).

2.2 Relaxed and Accelerated ADMM

Motivated by [1] we introduce new variables $\hat{u} \in \mathbb{R}^m$ and $\hat{z} \in \mathbb{R}^m$ to obtain an accelerated version of R-ADMM. The resulting algorithm, called R-A-ADMM, is defined by the following updates:

$$x_{k+1} \leftarrow \arg\min_x \left\{ f(x) + \tfrac{\rho}{2}\|\boldsymbol{A}x - \hat{z}_k + \hat{u}_k\|^2 \right\}, \tag{18a}$$

$$z_{k+1} \leftarrow \arg\min_z \big\{ g(z) + \tfrac{\rho}{2}\|\alpha \boldsymbol{A}x_{k+1} + (1-\alpha)\hat{z}_k - z + \hat{u}_k\|^2 \big\}, \tag{18b}$$

$$u_{k+1} \leftarrow \hat{u}_k + \alpha \boldsymbol{A}x_{k+1} + (1-\alpha)\hat{z}_k - z_{k+1}, \tag{18c}$$

$$\hat{u}_{k+1} \leftarrow u_{k+1} + \gamma_{k+1}\left(u_{k+1} - u_k\right), \tag{18d}$$

$$\hat{z}_{k+1} \leftarrow z_{k+1} + \gamma_{k+1}\left(z_{k+1} - z_k\right), \tag{18e}$$

where $\gamma_{k+1} \leftarrow k/(k+r)$ with $r \geq 3$. The above algorithm has not been considered in the literature and is a relaxation of A-ADMM [28],[3] which is recovered by setting $\alpha = 1$ and $r = 3$.

[3] Strictly speaking, [28] and [1] uses the parametrization $\gamma_{k+1} = \theta_{k-1}/\theta_{k+1}$ with the recursion $\theta_0 = 0$ and $\theta_{k+1} = (1 + \sqrt{1 + 4\theta_k^2})/2$. However, asymptotically this is the same as $\gamma_{k+1} = k/(k+r)$ with $r = 3$.

It is also worth noting that even for relaxed ADMM (16) (without acceleration) the existing theoretical results are sparse compared to standard ADMM. Regarding the continuous limit of updates (18), we obtain the following result, whose proof is in Appendix A.

Theorem 5. *Consider the R-A-ADMM updates* (18) *for solving problem* (8) *under Assumption 3. Then, the continuous limit of such updates, with time scale $t = \rho^{-1/2}k$, is given by the differential inclusion*

$$\alpha^{-1}\boldsymbol{A}^T\boldsymbol{A}\left(\ddot{X}(t) + \frac{r}{t}\dot{X}(t)\right) \in -\partial\Phi(X(t)) \tag{19}$$

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

Remark 6. *Theorem 5 still holds without assuming that $0 \in \operatorname{int}(\operatorname{dom} g - \boldsymbol{A}\operatorname{dom} f)$ (see Assumption 3), although in this case* (12) *becomes an inclusion, which results in $\partial f + \boldsymbol{A}^T\partial g \circ \boldsymbol{A}$ replacing the right-hand side of* (19).

The differential inclusion (19) reduces to the differential equation (6) when $\boldsymbol{A} = \boldsymbol{I}$, $\alpha = 1$, $r = 3$, and both f and g are smooth. Thus, (19) is more general and results related to (17) automatically hold for (6) as a special case. We show in Section 3 that the relaxation parameter α and the matrix $\boldsymbol{A}$ in (19) allow for refined convergence results.

2.3 Relaxed Heavy Ball ADMM

Another acceleration scheme for gradient descent is the heavy ball method introduced by [3]. Motivated by this approach we now introduce another accelerated variant of relaxed ADMM that we call relaxed heavy ball ADMM (R-HB-ADMM). The updates are essentially the same as those in (18), which we repeat below for convenience, except for the choice of γ_k. Specifically, we have:

$$x_{k+1} \leftarrow \arg\min_x\big\{f(x) + \tfrac{\rho}{2}\|\boldsymbol{A}x - \hat{z}_k + \hat{u}_k\|^2\big\}, \tag{20a}$$

$$z_{k+1} \leftarrow \arg\min_z\big\{g(z) + \tfrac{\rho}{2}\|\alpha\boldsymbol{A}x_{k+1} + (1-\alpha)\hat{z}_k - z + \hat{u}_k\|^2\big\}, \tag{20b}$$

$$u_{k+1} \leftarrow \hat{u}_k + \alpha\boldsymbol{A}x_{k+1} + (1-\alpha)\hat{z}_k - z_{k+1}, \tag{20c}$$

$$\hat{u}_{k+1} \leftarrow u_{k+1} + \gamma\left(u_{k+1} - u_k\right), \tag{20d}$$

$$\hat{z}_{k+1} \leftarrow z_{k+1} + \gamma\left(z_{k+1} - z_k\right), \tag{20e}$$

where $\gamma = 1 - r/\sqrt{\rho}$, with $r > 0$. Compared to (18) the only difference is that $\gamma_k = \gamma$ is a constant that depends on the penalty parameter ρ. This choice is inspired by the continuous limit but is otherwise not obvious.

Theorem 7. *Consider the R-HB-ADMM updates* (20) *for solving problem* (8) *under Assumption 3. Then, the continuous limit of such updates, with time scale $t = \rho^{-1/2}k$, is given by the differential inclusion*

$$\alpha^{-1}\boldsymbol{A}^T\boldsymbol{A}\left(\ddot{X}(t) + r\dot{X}(t)\right) \in -\partial\Phi(X(t)). \tag{21}$$

Assuming the algorithm is initialized with $\hat{z}_0 = z_0$, $\hat{u}_0 = u_0$ and $z_0 = \boldsymbol{A}x_0$ the initial conditions are $X(0) = x_0$ and $\boldsymbol{A}^T\boldsymbol{A}\dot{X}(0) \in \sqrt{\rho}\boldsymbol{A}^T u_0 - (1/\sqrt{\rho})\partial f(x_1)$.

The comments in Remark 6 also apply to Theorem 7.

The differential inclusion (21) is closely related to (19). The key difference is that different dissipation terms lead to different stability properties in the dynamical systems, which are reflected in the different behavior observed between algorithms (18) and (20); see Table 1 and Section 5.

The difference between the dynamical systems (19) and (21) lies in the type of dissipation. In the former, the damping vanishes asymptotically, thus for large time the system may exhibit strong oscillations, while in the latter the damping is constant, which helps attenuate oscillations and improve stability. This has been observed empirically, and it can also be shown that (21) is asymptotically stable for isolated minimizers while (19) is only stable.

Note that the dual variables u_k and $\hat{u}_k$ have no continuous counterpart in the dynamical systems (17), (19) and (21). The reason is that these dynamical systems capture only the leading order behaviour of the discrete algorithm in the limit of large ρ. The corresponding $U(t)$ and $\hat{U}(t)$ appear only in higher-order corrections to these differential inclusions.

3 Convergence Rates of the Nonsmooth Dynamical Systems

We provide convergence rates for the previous dynamical systems when Φ is convex or strongly convex (see Table 1 for a summary). These results are established through a nonsmooth Lyapunov analysis [35, 37–39]. In what follows, we provide the statements of the results and refer the reader to Appendix B for formal proofs.

3.1 Convergence of Relaxed ADMM

We first consider the dynamical system (17) associated with R-ADMM given by (16). We obtain the following.

Theorem 8. *Consider the dynamical system* (17). *Let* $x^\star \in \arg\min \Phi(x) \neq \emptyset$, $\Phi^\star \equiv \Phi(x^\star)$, *and* $X(t)$ *be a trajectory of the system with initial condition* $X(0) = x_0$.

(i) If Φ *is convex, then for all* $t > 0$ *it holds that*

$$\Phi(X(t)) - \Phi^\star \le \frac{\sigma_1^2(\boldsymbol{A})}{2\alpha t}\|x_0 - x^\star\|^2. \tag{22}$$

(ii) If Φ *is* μ*-strongly convex, then with* $\eta \equiv \frac{\mu\alpha}{2\sigma_1^2(\boldsymbol{A})}$ *and for all* $t > 0$ *it holds that*

$$\|X(t) - x^\star\| \le \kappa(\boldsymbol{A})\|x_0 - x^\star\|e^{-\eta t}. \tag{23}$$

Some remarks are appropriate:

- The rate (22) matches the $O(1/k)$ rate of standard (non relaxed) ADMM in the convex case [19,20]. We believe that the analogous result for *relaxed* ADMM presented in (22) is new in the sense that this rate is unknown in the discrete case.
- The exponential rate in (23) is consistent with the linear convergence of relaxed ADMM when Φ is strongly convex [27].
- It is interesting, although not surprising, that the relaxation parameter α appears in the convergence rates. The results (22) and (23) suggest an improved performance when over-relaxation is used (i.e., when $\alpha \in (1, 2)$). This is especially prominent in the strongly convex setting since α is under an exponential. These observations are consistent with the empirical guideline $\alpha \in (1.5, 1.8)$ suggested by [22, 23].

3.2 Convergence of Relaxed and Accelerated ADMM

For the dynamical system (19) related to R-A-ADMM as given by (18), we obtain the following rates.

Theorem 9. *Consider the dynamical system* (19). *Let* $x^\star \in \arg\min \Phi(x) \neq \emptyset$, $\Phi^\star = \Phi(x^\star)$, *and* $X(t)$ *be a trajectory of the system with conditions* $X(0) = x_0$ *and* $\dot{X}(0) = 0$.

(i) If Φ *is convex and* $r \ge 3$, *then for all* $t > 0$ *we have*

$$\Phi(X(t)) - \Phi^\star \le \frac{(r-1)^2\sigma_1^2(\boldsymbol{A})}{2\alpha t^2}\|x_0 - x^\star\|^2. \tag{24}$$

(ii) If Φ is μ-strongly convex, then there exists $C > 0$, independent of parameters, such that for all $t > t_0$,

$$\|X(t)) - x^\star\|^2 \le \frac{(r\sigma_1(\boldsymbol{A}))^{2r/3}}{\mu^{1+r/3}\alpha^{r/3}} \frac{C}{t^{2r/3}} \tag{25}$$

where $t_0 \ge \frac{\sigma_1(\boldsymbol{A})}{3}\sqrt{\frac{2r(2r-6)}{\mu\alpha}}$.

Note that α and $\boldsymbol{A}$ in (24) can improve the constant in the $O(1/t^2)$ bound. Another important observation is that the dynamical system (19) may not attain linear convergence in the strongly convex case according to (25).

3.3 Convergence of Relaxed Heavy Ball ADMM

We now turn to the dynamical system (21) associated with the R-HB-ADMM updates (20). The reader should compare the rates of the next theorem with those of Theorem 9.

Theorem 10. *Consider the dynamical system* (21)*. Let $x^\star \in \arg\min \Phi \neq \emptyset$, $\Phi^\star \equiv \Phi(x^\star)$, and $X(t)$ be a trajectory of the system with initial conditions $X(0) = x_0$ and $\dot{X}(0) = 0$. Then, there exists a constant $C > 0$, that is independent of parameters, such that the following holds:*

(i) If Φ is convex, then for all $t > 1/r$ we have

$$\Phi(X(t)) - \Phi^\star \le r\alpha^{-1}\sigma_1^2(\boldsymbol{A})\,\frac{C}{t}. \tag{26}$$

(ii) If Φ is μ-strongly convex, then with $r \le \frac{3}{2\sigma_1(\boldsymbol{A})}\sqrt{\mu\alpha}$ and for all $t > 0$ we have

$$\|X(t) - x^\star\|^2 \le r^2\alpha^{-1}\sigma_1^2(\boldsymbol{A})\mu^{-1}Ce^{-2rt/3}. \tag{27}$$

Some remarks regarding Theorems 9 and 10 are noteworthy.

- Over-relaxation $\alpha \in (1, 2)$ improves convergence in some cases more than others. In the convex cases (24) and (26) the improvement is linear in α, while in the strongly convex cases (25) and (27) α appears raised to a power and inside an exponential, respectively.

- Although the term $t^{-2r/3}$ in (25) seems to indicate faster convergence for larger values of r, one should remember that the constant grows as $r^{2r/3}$.

- Although R-HB-ADMM attains an exponential convergence rate in (27), the damping constant r must be below a certain threshold. In contrast, R-A-ADMM achieves only a sublinear rate in (25), 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 10 applies to the heavy ball method (4) as a special case for which an $O(1/k)$ rate was obtained in a Cesàro average sense [4], which contrasts (26). In the strongly convex case, linear convergence of heavy ball is known [4, 5] but not in the form (27).

- Comparing Theorem 9 and Theorem 10 we see an interesting tradeoff between the two types of acceleration, Nesterov versus heavy ball, in convex versus strongly convex settings. Nesterov's acceleration can achieve the optimal $O(1/t^2)$ rate in convex settings, however only the power law $O(t^{-2r/3})$ for strongly convex settings. On the other hand, heavy ball type of acceleration has $O(1/t)$ for convex but recovers linear convergence $O(e^{-2rt/3})$ in strongly convex settings. This suggests that in settings of higher curvature of Φ, heavy ball type of acceleration might be preferable.

4 Nonsmooth Hamiltonian Formulation

Both (19) and (21) admit (at least two) Hamiltonian representations. This is important because there are specific discretization schemes suited to Hamiltonian systems, such as symplectic integrators [40] which preserve some geometric structures of a Hamiltonian system. In the next section, we will use symplectic integrators to simulate the dynamical systems and compare with the associated variant of ADMM.

In this formalism, the system is described by a Hamiltonian function $H = H(X, P; t)$, where $(X, P) \in \mathbb{R}^{2n}$ are the canonical position and momentum variables. The equations of motion are given by [41]

$$\dot{X} = \nabla_P H, \qquad \dot{P} = -\nabla_X H. \tag{28}$$

A generalization of the Hamiltonian formalism to nonsmooth differential inclusions was proposed in [42–45], which allows us to replace the equalities in (28) by inclusions, and the gradients by subdifferentials. We use this approach to formulate a Hamiltonian representation of (19) and (21).

Let us consider the (explicit time dependent) Hamiltonian

$$H \equiv \tfrac{1}{2} e^{-\eta(t)} \big\langle P, \boldsymbol{M}^{-1} P \big\rangle + e^{\eta(t)} \lambda \Phi(X), \tag{29}$$

where $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is a "mass matrix" and λ is a "coupling constant" measuring the strength of the potential Φ. The Hamiltonian (29) is a generalization of the one associated to the damped harmonic oscillator that dates back to [46, 47]. Therefore, using (29) and taking into account the nonsmooth version of (28) we obtain

$$\dot{X} = e^{-\eta} \boldsymbol{M}^{-1} P, \qquad \dot{P} \in -e^{\eta} \lambda \partial \Phi(X). \tag{30}$$

After differentiating $P \equiv e^{\eta} \boldsymbol{M} \dot{X}$, the two first-order differential equations are equivalent to the second-order system

$$\tfrac{1}{\lambda} \boldsymbol{M} \left(\ddot{X} + \dot{\eta} \dot{X} \right) \in -\partial \Phi(X) \tag{31}$$

Notice that (19) and (21) are particular cases of (31) with $\boldsymbol{M} = \boldsymbol{A}^T \boldsymbol{A}$, $\lambda = \alpha$, and[4]

$$\eta = \begin{cases} r \log t & \text{for (19),} \\ rt & \text{for (21).} \end{cases} \tag{32}$$

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 through the dynamical system (29)/(31). A large entry in $\boldsymbol{A}^T \boldsymbol{A}$ would be like 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 with friction is described by the equations [48]

$$\dot{X} = \nabla_P H, \qquad \dot{P} \in -\partial_X H - \dot{\eta} P. \tag{33}$$

Here, consider the (explicit time independent) Hamiltonian

$$H \equiv \tfrac{1}{2} \langle P, \boldsymbol{M}^{-1} P \rangle + \lambda \Phi(X). \tag{34}$$

This is the classical Hamiltonian of the (nonlinear) harmonic oscillator [49]. From the definition in (34), it now follows from (33) that

$$\dot{X} = \boldsymbol{M}^{-1} P, \qquad \dot{P} \in -\lambda \partial \Phi(X) - \dot{\eta} P. \tag{35}$$

This pair of equations may be written equivalently as the second-order system (31), which means that it is also equivalent to (19) and (21) under the identification (32). Of particular interest is the case $\eta = rt$ in (32) (i.e when heavy ball acceleration is used) so that the system is *conformal Hamiltonian*[5] [48], that is to say that the Hamiltonian is explicit time independent and the friction term in (35) is a constant times P. Conformal Hamiltonian systems are interesting since they have a well-defined symplectic structure and a (conformal) symplectic integrator preserves the phase portrait of the flow map. Thus, such a discretization preserves the dissipative properties of the system exactly.

[4] Only in this section, we assume that $\boldsymbol{M} = \boldsymbol{A}^T \boldsymbol{A}$ is invertible.

[5] Recently, this formulation was considered in the context of optimization [50].

With representations (30) and (33) in hand, we can apply a symplectic integrator to obtain discretizations [40]. For instance, a dissipative version of the symplectic Euler method when applied to (33) (the derivation is provided in Appendix C) takes the form

$$p_{k+1/2} \leftarrow e^{-\Delta\eta(t_k)} p_k - \epsilon\lambda\partial\Phi(x_k), \tag{36a}$$

$$x_{k+1} \leftarrow x_k + \epsilon \boldsymbol{M}^{-1} p_{k+1/2}, \tag{36b}$$

where $\Delta\eta(t_k) = \eta(t_{k+1}) - \eta(t_k)$ with η given by (32), and recall that $t_{k+1} = t_k + \epsilon$ with $t_k = k\epsilon$. Note that one subgradient computation is needed per iteration. We note that any symplectic method can also be applied to (30). In our numerical experiments (see Section 5) we verified that both approaches give similar results, although (36) has a simpler form. Our goal in introducing (36) is to compare numerical solutions to the continuous dynamical systems with the variants of ADMM algorithms.

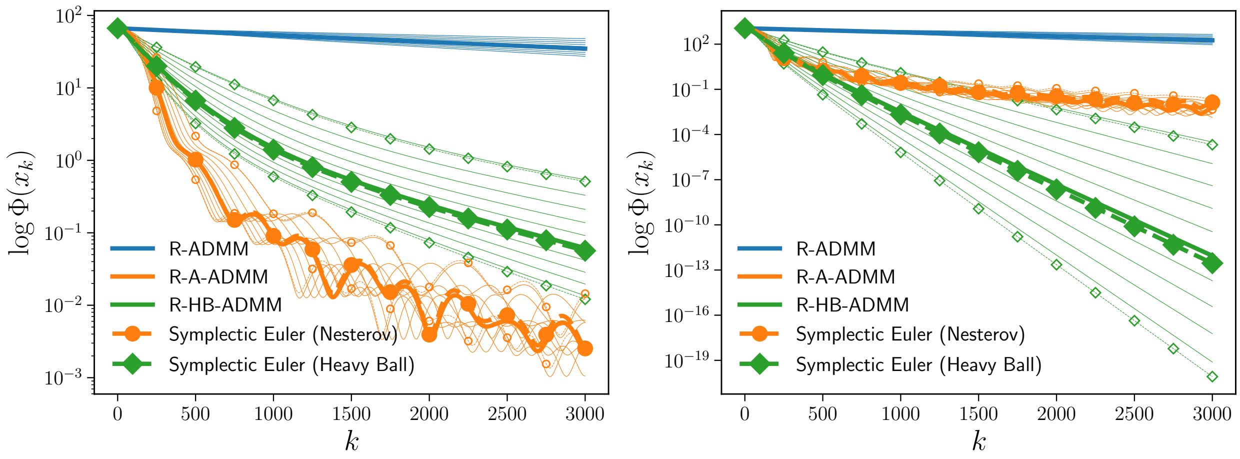
5 Numerical Experiments

The sole purpose of this section is to verify numerically if the continuous dynamical systems are able to approximately model the behaviour of the discrete ADMM algorithms. To this end we choose a large penalty parameter ρ and use the integrator (36) to reproduce the continuous-time dynamics.

First, consider a quadratic problem

$$\min_x \tfrac{1}{2}\langle x, \boldsymbol{Q}x\rangle \quad \text{subject to} \quad z = \boldsymbol{A}x \tag{37}$$

where $\boldsymbol{A} \in \mathbb{R}^{60\times 60} \overset{iid}{\sim} \mathcal{N}(0,1)$ with $\sigma_1(\boldsymbol{A}) = 50$ and $\sigma_{60}(\boldsymbol{A}) = 1$, and $\boldsymbol{Q} \in \mathbb{R}^{60\times 60}$ is a random symmetric and positive-semidefinite matrix. Note that $g(z) = 0$ and $\boldsymbol{A}$ still enters the updates. In Figure 1 we solve (37) with R-ADMM (16), R-A-ADMM (18) and R-HB-ADMM (20), as well as the dynamical systems (19) and (21) through the symplectic Euler method (36) (dashed lines with markers). For the convex case (Figure 1a) we choose eigenvalues $\lambda_i(\boldsymbol{Q}) = 0$ for $i = 1, \ldots, 50$ and $\lambda_i(\boldsymbol{Q}) \sim \text{unif}([0,1])$ otherwise. For the strongly convex case (Figure 1b) we choose $\lambda_i(\boldsymbol{Q}) \sim \text{unif}([1,2])$ for $i = 1, \ldots, 60$. We fix $\rho = 5 \times 10^3$, $r = 3$, $\alpha \in [0.5, 1.5]$ (thin lines, with thick lines corresponding to $\alpha = 1$). For (36) we use a stepsize $\epsilon = 1/\sqrt{\rho}$. The lines corresponding to $\alpha > 1$ are below the solid line and improve convergence, while lines corresponding to $\alpha < 1$ are above. We thus verify the predictions of Theorems 8, 9, and 10, and in particular the tradeoff between Nesterov and heavy ball acceleration. Note also the agreement between the algorithms and the dynamical systems simulations. The reason R-ADMM is converging very slowly is because of the large penalty parameter.



(a)

(b)

Figure 1: Quadratic problem (37) (where $g = 0$). (a) Convex problem with $\boldsymbol{Q} \succcurlyeq 0$. (b) Strongly convex problem with $\boldsymbol{Q} \succ 0$. R-ADMM converges slowly due to large $\rho = 5 \cdot 10^3$. Note the perfect agreement between ODEs solutions and ADMM algorithms.

Next, consider a linear regression problem with elastic net regularization [51, 52]:

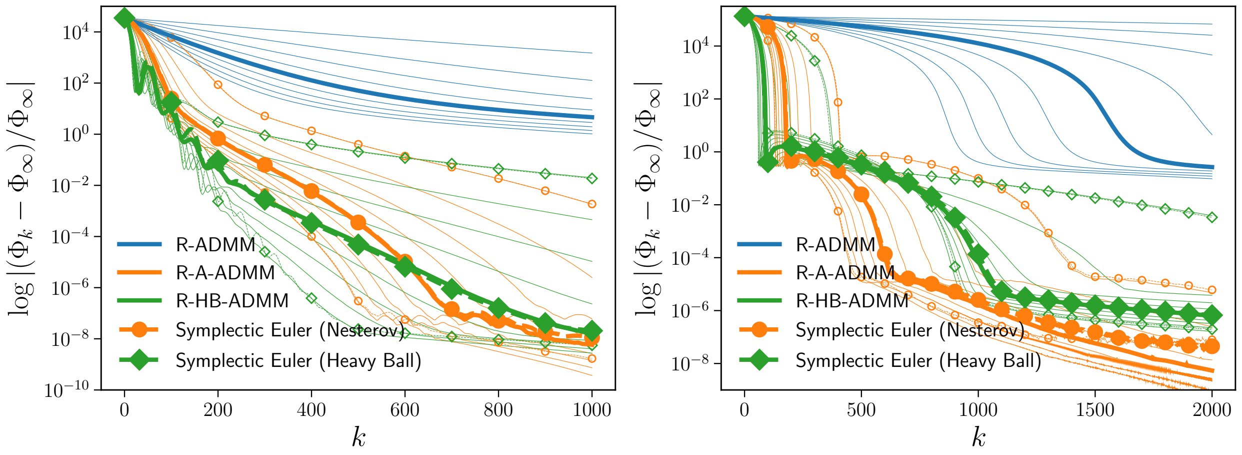
$$\min_x \tfrac{1}{2}\|y - \boldsymbol{M}x\|^2 + \|x\|_1 + \tfrac{1}{2}\|x\|_2^2. \tag{38}$$

We use the same setup of [51], which was also used by [28]. We thus have $y = \boldsymbol{M}q + \mathcal{N}(0, 0.1\boldsymbol{I})$, where $q \in \mathbb{R}^{40}$ with entries $q_i = 3$ for $i = 1, \dots, 15$ and $q_i = 0$ otherwise. We sample three normal vectors v_1, v_2 and v_3 in $\mathbb{R}^{50}$ to form the matrix $\boldsymbol{M} \in \mathbb{R}^{50\times 40}$ with columns $M_i = v_1 + \mathcal{N}(0, 0.1\boldsymbol{I})$ for $i = 1, \dots, 5$, $M_i = v_2 + \mathcal{N}(0, 0.1\boldsymbol{I})$ for $i = 6, \dots, 10$, $M_i = v_3 + \mathcal{N}(0, 0.1\boldsymbol{I})$ for $i = 11, \dots, 15$, and $\boldsymbol{M}_i \sim \mathcal{N}(0, \boldsymbol{I})$ for $i = 16, \dots, 40$. The results are shown in Figure 2a, with parameters $\rho = 10^5$, $\epsilon = 1/\sqrt{\rho}$, $r = 10$, and $\alpha \in [0.2, 2]$. Again, note the agreement between algorithms and their associated dynamical systems simulations.

Finally, consider the sparse logistic regression problem

$$\min_{x\in\mathbb{R}^{100}} \sum_{i=1}^{50} \log\left(1 + e^{-y_i\langle x, d_i\rangle}\right) + \|x\|_1 \tag{39}$$

for data points d_i sampled from a 100 dimensional Guassian mixture, $d_i \sim (1/2)\mathcal{N}(-1, \boldsymbol{I}) + (1/2)\mathcal{N}(+1, \boldsymbol{I})$, with respective labels $y_i \in \{-1, +1\}$. Solutions to (39) are in Figure 2b, where $\rho = 5 \cdot 10^4$, $\epsilon = 1/\sqrt{\rho}$, $r = 10$ and $\alpha \in [0.2, 2]$. Again, we see a good agreement between ADMM algorithms and numerical solutions of the differential equations. However, we can see a slight discrepancy for large k, especially for R-A-ADMM. This indicates that the ADMM discretizations, compared to the symplectic method (36), introduces some spurious



(a)

(b)

Figure 2: Regression problems. (a) Elastic net regression problem (38). (b) Logistic regression problem (39). In both cases the convergence of R-ADMM is slow due to a large penalty parameter ρ. Note the agreement between algorithms and dynamical systems, however the ADMM algorithms introduce an extra damping.

damping that is not present in the continuous-time dynamical system. This may be beneficial for some optimization problems, such as when the minimum lies on a flat region. This extra damping can make the algorithm converge faster when close to the minimum.

6 Conclusion

We introduced two new families of *relaxed* and *accelerated* ADMM algorithms. The first follows Nesterov's acceleration approach and is given by the updates in (18). The second is inspired by Polyak's heavy ball method and is given by updates in (20). We then derived differential inclusions (nonsmooth dynamical systems) modelling the leading order behaviour of these algorithms in the continuous-time limit. This extends prior work by accounting for nonsmooth problems and allowing for linear constraints (see Theorems 4, 5 and 7). Moreover, we obtained rates of convergence for the continuous dynamical systems in convex and strongly convex settings through a nonsmooth Lyapunov analysis; see Theorems 8, 9 and 10. The complexity results obtained in this paper are summarized in Table 1, most of which are new to the best of our knowledge, and all proofs are presented in the Appendix. To verify numerically our theoretical results we introduced a (conformal) Hamiltonian formulation for the second-order dynamical systems (see Section 4) that allowed us to construct a symplectic discretization. The proposed algorithms were compared to simulations of their associated continuous dynamical systems, showing an excellent agreement for small stepsize (large penalty parameter).

These results strengthen the connections between optimization and (nonsmooth) continuous dynamical systems. The proof techniques in continuous-time may shed light in obtaining analogous rates in discrete-time, which is a more challenging problem. The tradeoff between Nesterov versus heavy ball acceleration also deserves a more thorough 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, where our nonsmooth Hamiltonian formulation can be useful. Symplectic integrators offer a potential alternative. Finally, an interesting extension would be to consider continuous dynamical systems that include more general nonlinear constraints where, again, our nonsmooth Hamiltonian formulation can be a good starting point.

Note added: After this manuscript was posted on arXiv we became aware of Ref. [53] that also considers ADMM and differential inclusions.

Acknowledgements

This work was supported by grants ARO MURI W911NF-17-1-0304 and NSF 1447822. We thank Junchi Li for letting us know about their recent work [53] and also for pointing out a small mistake in a previous version of this manuscript.

A Continuous Limit

In this section we derive the differential inclusions modelling the variants of ADMM considered in this paper. The derivation of (16) is considerably simpler compared to that of (19), thus we provide a short proof of Theorem 4 and in the sequence we provide a detailed proof of Theorem 5.

Proof of Theorem 4. Analogously to (51) the optimality conditions for the proximal operators in the updates (16) can be combined into

$$\partial f\left(x_{k+1}\right)+\boldsymbol{A}^T \partial g\left(z_{k+1}\right)+\rho \boldsymbol{A}^T\big((1-\alpha) \boldsymbol{A} x_{k+1}+z_{k+1}-(2-\alpha) z_k\big) \ni 0. \tag{40}$$

Let $(x_k, z_k, u_k)=(X(t), Z(t), U(t))$ where $t=k\epsilon$. Choosing $\epsilon=\rho^{-1}$, from (40) we have

$$\partial f(x_{k+1})+\boldsymbol{A}^T \partial g(z_{k+1})+(1-\alpha) \boldsymbol{A}^T \frac{\boldsymbol{A} x_{k+1}-z_k}{\epsilon}+\boldsymbol{A}^T \frac{z_{k+1}-z_k}{\epsilon} \ni 0. \tag{41}$$

From the last update in (16) we have $U(t+\epsilon) = U(t) + \alpha \boldsymbol{A} X(t+\epsilon) + (1-\alpha)Z(t) - Z(t+\epsilon)$. Using the Mean Value theorem on each component of this equation allows us to conclude that

$$\epsilon \dot{U}(t) = \alpha \left(\boldsymbol{A}X(t) - Z(t)\right) + \epsilon \left(\boldsymbol{A}\dot{X}(t) - \dot{Z}(t)\right) + O(\epsilon^2). \tag{42}$$

Thus, in the limit $\epsilon \to 0$ we obtain

$$\boldsymbol{A}X(t) = Z(t), \qquad \boldsymbol{A}\dot{X}(t) = \dot{Z}(t). \tag{43}$$

Moreover, taking terms of $O(\epsilon)$ in (42) and using the above result we conclude that

$$\dot{U}(t) = \boldsymbol{A}\dot{X}(t) - \dot{Z}(t) = 0. \tag{44}$$

Now, from the last update (16) we have

$$\boldsymbol{A}x_{k+1} - z_k = \alpha^{-1}(u_{k+1} - u_k) + \alpha^{-1}(z_{k+1} - z_k). \tag{45}$$

Replacing this into (41) and taking the limit $\epsilon \to 0$ we obtain

$$\partial f(X(t)) + \boldsymbol{A}^T \partial g(Z(t)) + (1-\alpha)\alpha^{-1}\boldsymbol{A}^T \left(\dot{U}(t) + \dot{Z}(t)\right) + \boldsymbol{A}^T \dot{Z}(t) \ni 0. \tag{46}$$

Using (43) and (44) we finally obtain

$$\alpha^{-1}\boldsymbol{A}^T\boldsymbol{A}\dot{X}(t) \in -\partial f(X(t)) - \boldsymbol{A}^T \partial g(Z(t)). \tag{47}$$

By assumption, $0 \in \operatorname{int}(\operatorname{dom} g - \boldsymbol{A} \operatorname{dom} f)$, so that relation (12) holds with equality and we thus obtain (17). Note that this is a first order system whose dynamics is specified by one initial condition such as $X(0) = x_0$. □

Proof of Theorem 5. The respective optimality conditions for the subproblems (18a) and (18b) are given by

$$\partial f(x_{k+1}) + \rho \boldsymbol{A}^T \left(\boldsymbol{A}x_{k+1} - \hat{z}_k + \hat{u}_k\right) \ni 0, \tag{48a}$$

$$\partial g(z_{k+1}) - \rho\big(\alpha \boldsymbol{A}x_{k+1} + (1-\alpha)\hat{z}_k - z_{k+1} + \hat{u}_k\big) \ni 0. \tag{48b}$$

These two equations can be further combined into

$$\partial f(x_{k+1}) + \boldsymbol{A}^T \partial g(z_{k+1}) + \rho(1-\alpha)\boldsymbol{A}^T \left(\boldsymbol{A}x_{k+1} - \hat{z}_k\right) + \rho \boldsymbol{A}^T \left(z_{k+1} - \hat{z}_k\right) \ni 0. \tag{49}$$

Using (18c), i.e.

$$\boldsymbol{A}x_{k+1} - \hat{z}_k = \alpha^{-1}(u_{k+1} - \hat{u}_k) + \alpha^{-1}(z_{k+1} - \hat{z}_k) \tag{50}$$

we can further write (49) as

$$\partial f(x_{k+1}) + \boldsymbol{A}^T \partial g(z_{k+1}) + \rho(1-\alpha)\alpha^{-1}\boldsymbol{A}^T \left(u_{k+1} - \hat{u}_k\right) + \rho\alpha^{-1}\boldsymbol{A}^T \left(z_{k+1} - \hat{z}_k\right) \ni 0. \tag{51}$$

From (18e) we have $z_{k+1} - \hat{z}_k = z_{k+1} - (1+\gamma_k)z_k + \gamma_k z_{k-1}$. By adding $0 = z_k - z_k + z_{k-1} - z_{k-1}$ to the right hand side, and then reorganizing, we obtain

$$z_{k+1} - \hat{z}_k = (z_{k+1} - 2z_k + z_{k-1}) + (1-\gamma_k)(z_k - z_{k-1}). \tag{52}$$

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 $t = k\epsilon$. Furthermore, let us choose $\rho = 1/\epsilon^2$. According to (14), ρ times the first term on the right hand side of (52) gives $\rho\,(z_{k+1} - 2z_k + z_{k-1}) \to \ddot{Z}(t)$ as $\epsilon \to 0$, while according to (13), ρ times the second term on the right hand side of (52) satisfies

$$\frac{1-\gamma_k}{\epsilon}\,\frac{z_k - z_{k-1}}{\epsilon} = \frac{r}{t+\epsilon(r-1)}\,\frac{z_k - z_{k-1}}{\epsilon} \to \frac{r}{t}\dot{Z}(t) \tag{53}$$

as $\epsilon \to 0$. Therefore,

$$\rho(z_{k+1} - \hat{z}_k) \to \ddot{Z}(t) + \frac{r}{t}\dot{Z}(t) \qquad (\epsilon \to 0). \tag{54}$$

Analogously, we immediately have that $\rho(u_{k+1} - \hat{u}_k) \to \ddot{U}(t) + \frac{r}{t}\dot{U}(t)$. However, as we will show below, $\dot{U}(t)$ and higher order derivatives vanish so the U variable will not contribute in the continuous limit of (51), which is therefore

$$\partial f(X(t)) + \boldsymbol{A}^T \partial g(Z(t)) + \alpha^{-1}\boldsymbol{A}^T\left(\ddot{Z} + \frac{r}{t}\dot{Z}(t)\right) \ni 0. \tag{55}$$

Let us now consider (18d), and note that $\gamma_{k+1} = \frac{k}{k+r} = \frac{\epsilon k}{\epsilon k + \epsilon r}$, thus it follows that

$$\hat{u}_{k+1} - u_{k+1} - \gamma_{k+1}(u_{k+1} - u_k) = \hat{U}(t+\epsilon) - U(t+\epsilon) - \frac{t}{t+\epsilon r}\left(U(t+\epsilon) - U(t)\right) = 0. \tag{56}$$

In the limit $\epsilon \to 0$ this implies that $\hat{U}(t) = U(t)$. Using this together with (18c), by the same argument used in (42) we conclude that (43) and (44) still hold true. Hence, we also have $\dot{U}(t) = \dot{\hat{U}}(t) = 0$, as previously mentioned in obtaining (55). Finally, replacing $\boldsymbol{A}X(t) = Z(t)$, $\boldsymbol{A}\dot{X}(t) = \dot{Z}(t)$ and $\boldsymbol{A}\ddot{X}(t) = \ddot{Z}(t)$ we obtain

$$\alpha^{-1}\boldsymbol{A}^T\boldsymbol{A}\left(\ddot{X}(t) + \frac{r}{t}\dot{X}(t)\right) \in -\partial f(X(t)) - \boldsymbol{A}^T\partial g(\boldsymbol{A}X(t)). \tag{57}$$

Recalling that $0 \in \operatorname{int}(\operatorname{dom} g - \boldsymbol{A}\operatorname{dom} f)$ by assumption, so that relation (12) holds with equality we finally obtain the differential inclusion (19), as claimed.

For the initial conditions, one can choose $X(0) = x_0$ where x_0 is the initial estimate of a solution to (8). 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 (19) yields

$$\alpha^{-1}\sum_j (\boldsymbol{A}^T\boldsymbol{A})_{ij}\left(\dot{X}_j(t) - \dot{X}_j(0) + r\dot{X}_j(\xi)\right) \in -t\partial_i \Phi(X(\xi)) \tag{58}$$

where we denote by $\partial_i \Phi$ the ith component of each vector in the subdifferential set $\partial \Phi$. Letting $t \downarrow 0$, which also forces $\xi \downarrow 0$, we have $\sum_j (\boldsymbol{A}^T \boldsymbol{A})_{ij} \dot{X}_j(0) = 0$ since $r \neq 0$. Note that this holds for each component $i = 1, \ldots, n$, therefore $(\boldsymbol{A}^T \boldsymbol{A})\dot{X}(0) = 0$. □

Proof of Theorem 7. Obtaining the differential equation follows exactly the same steps as the proof of Theorem 5 above, and is therefore ommited. The only change is in the initial conditions.

The first is obviously $X(0) = x_0$. For the velocity, consider (48b), i.e. $\partial f(x_1) + \rho \boldsymbol{A}^T \boldsymbol{A} x_1 + \rho \boldsymbol{A}^T (\hat{u}_0 - \hat{z}_0) \ni 0$. Assuming the algorithm is initialized such that $\hat{z}_0 = z_0$, $\hat{u}_0 = u_0$ and $z_0 = \boldsymbol{A} x_0$, and recalling that $\epsilon^2 = 1/\rho$ is the stepsize, we have

$$\boldsymbol{A}^T \boldsymbol{A} \left(\frac{x_1 - x_0}{\epsilon} \right) + \epsilon^{-1} \boldsymbol{A}^T u_0 \in -\epsilon \partial f(x_1), \tag{59}$$

thus $\boldsymbol{A}^T \boldsymbol{A} \dot{X}(0) \in \sqrt{\rho} \boldsymbol{A}^T u_0 - (1/\sqrt{\rho}) \partial f(x_1)$. □

B Convergence Rates

In this section we derive the convergence rates summarized in Table 1. We first introduce some basic concepts; we refer the reader to [35, 38, 39] for more details.

For a nondifferentiable function f, the directional derivative of f at point $x \in \operatorname{dom} f$ in the direction $v \in \mathbb{R}^n$ is defined as [31]

$$Df(x)(v) \equiv \lim_{\epsilon \downarrow 0} \frac{f(x + \epsilon v) - f(x)}{\epsilon}. \tag{60}$$

We have the following result connecting the directional derivative to the subdifferential set defined in (9).

Theorem 11 (Max formula [31])**.** *If f is convex, then for any $x \in \operatorname{dom} f$ and any $v \in \mathbb{R}^n$ we have*

$$Df(x)(v) = \max_{\xi \in \partial f(x)} \langle \xi, v \rangle. \tag{61}$$

Note that if f is differentiable then $\partial f(x) = \{\nabla f(x)\}$ and (61) becomes $Df(x)(v) = \langle \nabla f(x), v \rangle$, which is the projection of the gradient in the direction v.

Consider the (autonomous) differential inclusion

$$\dot{X}(t) \in F(X(t)), \qquad X(0) = x_0, \tag{62}$$

where we recall that by a solution we mean a function $X(t)$ that is absolutely continuous and obeys (15) for almost every t. We recall that with $F = -\partial\Phi$ (and under Assumption 3) then $\dot{X}(t)$ is also continuous and a unique solution to (62) exists [35]. Solutions to (62) obeying

$$\mathcal{E}(X(t)) - \mathcal{E}(X(\xi)) + \int_{\xi}^{t} W(X(s), \dot{X}(s)) ds \le 0, \tag{63}$$

for $t \ge \xi \ge 0$, are called *monotone trajectories*, where $\mathcal{E} : \operatorname{dom} F \to \mathbb{R}_+$ and $W : \operatorname{graph} F \to \mathbb{R}_+$. The function $\mathcal{E}$ is said to be a Lyapunov function with respect to W if for all $x \in \operatorname{dom} F$ there exists some $v \in F(x)$ such that

$$D\mathcal{E}(x)(v) + W(x, v) \le 0, \tag{64}$$

where $D\mathcal{E}$ is the directional derivative (60)[6]. If $W = 0$ then $\mathcal{E}$ is simply called a Lyapunov function.

Theorem 12 (see [35])**.** *Let* $\operatorname{dom} F \subset \mathbb{R}^n$ *be compact and* $F : \operatorname{dom} F \to \mathbb{R}^n$ *an upper semicontinuous map with a nonempty and convex image. Let* $W : \operatorname{graph} F \to \mathbb{R}_+$ *be continuous and a convex function of* v*. Let* $\mathcal{E} : \operatorname{dom} F \to \mathbb{R}_+$ *be a continuous Lyapunov function with respect to* W*, defined by* (64)*. Then, for every* $x_0 \in \mathbb{R}^n$*, there exists* $T > 0$ *and a monotone trajectory* $X : [0, T) \to \mathbb{R}^n$ *of the differential inclusion* (62)*. If* $\operatorname{dom} F$ *is closed and* F *bounded then we can take* $T = +\infty$*.*

This theorem implies that the existence of a Lyapunov function is sufficient to guarantee monotone trajectories. In the case where $F = -\partial\Phi$ (under Assumption (3)) the conditions for F in Theorem 12 are obeyed. From (61) we have

$$D\mathcal{E}(x)(v) = \max_{\xi \in \partial\mathcal{E}(x)} \langle \xi, v \rangle. \tag{65}$$

If $\mathcal{E}$ is differentiable, choosing $x = X(t)$ and $v = \dot{X}(t)$ this reduces to $\dot{\mathcal{E}} \equiv \frac{d}{dt}\mathcal{E} = \langle \nabla\mathcal{E}(X), \dot{X} \rangle$, i.e., the familiar notion of time derivative. In this section we will abuse notation and denote by $\dot{\mathcal{E}}$ the generalization of time derivative to nonsmooth cases as well, i.e., $\dot{\mathcal{E}} \equiv D\mathcal{E}(X(t))(\dot{X}(t))$ for all $\dot{X}(t) \in F(X(t))$. All the results mentioned above naturally extend to nonautonomous systems as well [35]. For the Lyapunov functions considered below, the only source of nonsmootheness in our derivations comes from Φ in problem (8). To account for this we use the following.

[6] All the statements mentioned in this section hold for more general notions of directional derivatives (which are meaningful even for nonconvex functions). However, for a function which is lower semicontinuous and convex, such as in Assumption 3, these more general directional derivatives coincide with (60).

Lemma 13. *Let $F = -\partial\Phi$ in the differential inclusion* (62) *where Φ is lower semicontinuous, proper and convex. Then, for almost every $t \in \mathbb{R}_+$, it holds that*

$$\frac{d}{dt}(\Phi \circ X)(t) = \langle \xi, \dot{X}(t) \rangle \quad \text{for all } \xi \in \partial\Phi(X(t)). \tag{66}$$

Proof. By definition of subdifferential (9) we have

$$\Phi(X(t)) - \Phi(X(t+\epsilon)) \geq \langle \xi, X(t) - X(t+\epsilon) \rangle \tag{67}$$

for all $\xi \in \partial\Phi(X(t+\epsilon))$. By assumption, Φ is convex thus it is locally Lipschitz. Moreover, a solution $X(t)$ of (62) is absolutely continuous. Thus, from Rademacher's theorem $\frac{d}{dt}(\Phi \circ X)(t)$ exists for almost every t. Under the assumptions on Φ the differential inclusion (62) also have a (unique) solution so that $\dot{X}(t)$ exists for almost every t as well. Hence, dividing (67) by ϵ and taking the limit $\epsilon \downarrow 0$ we obtain

$$\frac{d}{dt}\Phi(X(t)) \leq \big\langle \xi, \dot{X}(t) \big\rangle \tag{68}$$

for all $\xi \in \partial\Phi(X(t))$. By definition of subdifferential we also have that

$$\Phi(X(t+\epsilon)) - \Phi(X(t)) \geq \langle \xi, X(t+\epsilon) - X(t) \rangle \tag{69}$$

for all $\xi \in \partial\Phi(X(t))$. Hence, by the same argument above we conclude that

$$\frac{d}{dt}\Phi(X(t)) \geq \big\langle \xi, \dot{X}(t) \big\rangle. \tag{70}$$

Thus, (68) and (70) imply (66). □

In the following proofs we consider a Lyapunov function $\mathcal{E}$ (with $W = 0$ in (64)) and we use (66) to account for the nonsmoothness of Φ. For simplicity of notation we also ommit the explicit time dependence on $X = X(t)$ and other quantities.

B.1 Proof of Theorem 8

Theorem 8 (i). Consider

$$\mathcal{E} \equiv t\alpha \left(\Phi(X) - \Phi(x^\star)\right) + \frac{1}{2} \left\| \boldsymbol{A} \left(X - x^\star\right) \right\|^2 \tag{71}$$

where $X = X(t)$ is a trajectory of (17). Note that $\mathcal{E} \geq 0$ for any $X \in \mathbb{R}^n$. Differentiating (71) with respect to time and noticing that $\frac{d}{dt}\Phi(X(t))$ exists for almost every t,

$$\dot{\mathcal{E}} = t\alpha \frac{d}{dt}\Phi(X) + \alpha \left(\Phi(X) - \Phi(x^\star)\right) + \big\langle X - x^\star, \boldsymbol{A}^T \boldsymbol{A} \dot{X} \big\rangle. \tag{72}$$

Using (17) and the convexity relation (10), the second and third terms above yields

$$\alpha\left(\Phi(X) - \Phi(x^\star) - \langle X - x^\star, \xi\rangle\right) \le 0 \tag{73}$$

for all $\xi \in \partial\Phi(X)$. Therefore, using (66), we conclude that

$$\dot{\mathcal{E}} = t\alpha \frac{d}{dt}\Phi(X) \le t\alpha\langle \xi, \dot{X}\rangle \tag{74}$$

for all $\xi \in \partial\Phi(X)$. Using (17) once again we conclude that

$$\dot{\mathcal{E}} \le -t\|\boldsymbol{A}\dot{X}\|^2 \le 0. \tag{75}$$

Thus, $\mathcal{E}|_t \le \mathcal{E}|_{t=0}$. Combining this with the definition (71) we obtain the upper bound

$$\begin{aligned}\Phi(X) - \Phi(x^\star) &= \frac{1}{\alpha t}\left(\mathcal{E}|_t - \tfrac{1}{2}\|\boldsymbol{A}(X - x^\star)\|^2\right) \\ &\le \frac{\mathcal{E}|_{t=0}}{\alpha t}.\end{aligned} \tag{76}$$

Note that $\mathcal{E}|_{t=0} = \frac{1}{2}\|\boldsymbol{A}(x_0 - x^\star)\|^2 \le \frac{1}{2}\sigma_1^2(\boldsymbol{A})\|x_0 - x^\star\|^2$, finally yielding (22). □

Theorem 8 (ii). Since Φ is strongly convex, let us use (11) to conclude that, for every $\xi \in \partial\Phi(X)$, it holds that

$$\begin{aligned}\langle X - x^\star, \xi\rangle &\ge \frac{\mu}{2}\|X - x^\star\|^2 + \Phi(X) - \Phi(x^\star) \\ &\ge \frac{\mu}{2}\|X - x^\star\|^2 \\ &\ge \frac{\mu\|\boldsymbol{A}(X - x^\star)\|^2}{2\|\boldsymbol{A}\|^2}\end{aligned} \tag{77}$$

for all $X \in \mathbb{R}^n$, and where $x^\star$ the unique minimizer of Φ. Consider

$$\mathcal{E} \equiv \tfrac{1}{2}\|\boldsymbol{A}(X - x^\star)\|^2 \ge 0. \tag{78}$$

Taking its total time derivative, and letting $X = X(t)$ be a trajectory of the dynamical system (17), thus $\dot{X}$ exists for almost every t, we have that for every $\xi \in \partial\Phi(X)$,

$$\begin{aligned}\dot{\mathcal{E}} &= \left\langle X - x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\right\rangle \\ &= -\alpha\left\langle X - x^\star, \xi\right\rangle \\ &\le -\frac{\mu\alpha\left\|\boldsymbol{A}(X - x^\star)\right\|^2}{2\|\boldsymbol{A}\|^2},\end{aligned} \tag{79}$$

where we used (77) in the last passage. From (78) we can write this last relation as

$$\frac{d}{dt}\|\boldsymbol{A}(X(t) - x^\star)\|^2 \le -\frac{2\alpha\mu\|\boldsymbol{A}(X(t) - X^\star)\|^2}{2\|\boldsymbol{A}\|^2}. \tag{80}$$

From Grönwall's inequality we thus conclude that $\|\boldsymbol{A}(X(t) - x^\star)\|^2 \le \|\boldsymbol{A}(x_0 - x^\star)\|^2 e^{-2\eta t}$, where $\eta \equiv \frac{\mu\alpha}{2\|\boldsymbol{A}\|^2}$. After taking the square root of this last inequality we also have

$$\sigma_m(\boldsymbol{A})\|X(t) - x^\star\| \le \sigma_1(\boldsymbol{A})\|x_0 - x^\star\| e^{-\eta t} \tag{81}$$

where $\sigma_m(\boldsymbol{A})$ and $\sigma_1(\boldsymbol{A})$ are the smallest and largest singular values of $\boldsymbol{A}$, respectively. To obtain (23), recall that $\kappa(\boldsymbol{A}) \equiv \frac{\sigma_1(\boldsymbol{A})}{\sigma_m(\boldsymbol{A})}$ is the condition number of $\boldsymbol{A}$. □

B.2 Proof of Theorem 9

Theorem 9 (i). Consider

$$\mathcal{E}(X, \dot{X}, t) \equiv \left(\frac{t}{r-1}\right)^2 \alpha\left(\Phi(X) - \Phi(x^\star)\right) + \frac{1}{2}\left\|\boldsymbol{A}\Big(X - x^\star + \frac{t}{r-1}\dot{X}\Big)\right\|^2 \tag{82}$$

over a trajectory $X = X(t)$ of (19). Taking the total time derivative we have

$$\begin{aligned} \dot{\mathcal{E}} = {} & \frac{2\alpha t}{(r-1)^2}\left(\Phi(X) - \Phi(x^\star)\right) + \left(\frac{t}{r-1}\right)^2 \alpha \frac{d}{dt}\Phi(X) \\ & + \left\langle X - x^\star + \frac{t}{r-1}\dot{X}, \boldsymbol{A}^T\boldsymbol{A}\left(\frac{r}{r-1}\dot{X} + \frac{t}{r-1}\ddot{X}\right)\right\rangle. \end{aligned} \tag{83}$$

Using (19) in the last term and noticing that relation (66) holds for any $\xi \in \partial\Phi(X)$ we can simplify the above expression into

$$\dot{\mathcal{E}} = \frac{2\alpha t}{(r-1)^2}\Big(\Phi(X) - \Phi(x^\star) - \frac{r-1}{2}\langle X - x^\star, \xi\rangle\Big). \tag{84}$$

Since Φ is convex, $\Phi(X) - \Phi(x^\star) \le \langle \xi, X - x^\star\rangle$ for all $X \in \mathbb{R}^n$ and for all $\xi \in \partial\Phi(X)$, thus

$$\dot{\mathcal{E}} \le -\frac{t\alpha(r-3)}{(r-1)^2}\left(\Phi(X) - \Phi(x^\star)\right) \le 0, \tag{85}$$

provided $r \ge 3$. To conclude the proof, note that (85) implies that $\mathcal{E}|_t \le \mathcal{E}|_{t=0}$, therefore

$$\left(\frac{t}{r-1}\right)^2 \alpha\left(\Phi(X) - \Phi(x^\star)\right) \le \mathcal{E}|_{t=0}, \tag{86}$$

which gives (24) uppon replacing $\mathcal{E}|_{t=0} = \frac{1}{2}\|\boldsymbol{A}(x_0 - x^\star)\|^2 \le \frac{1}{2}\sigma_1^2(\boldsymbol{A})\|x_0 - x^\star\|^2$. □

Theorem 9 (ii). Consider

$$\mathcal{E} \equiv t^\lambda \alpha\left(\Phi(X) - \Phi(x^\star)\right) + \frac{t^{\lambda-2}}{2}\|\boldsymbol{A}(\lambda(X - x^\star) + t\dot{X})\|^2 \tag{87}$$

where $\lambda = 2r/3$ and $x^\star$ is the unique minimizer of Φ. Let $X = X(t)$ be a trajectory of (19). Note also that $\mathcal{E} \geq 0$ for all $t \geq 0$. Taking the total time derivative of (87), and noting that $\frac{d}{dt}\Phi(X(t))$ exists for almost every t, we obtain

$$\begin{aligned}\dot{\mathcal{E}} = \lambda\alpha t^{\lambda-1}\left(\Phi(X) - \Phi(x^\star)\right) + t^\lambda \alpha \frac{d}{dt}\Phi(X) + \frac{(\lambda-2)t^{\lambda-3}}{2}\big\|\boldsymbol{A}\big(\lambda(X - x^\star) + t\dot{X}\big)\big\|^2 \\ + t^{\lambda-2}\big\langle \lambda(X - x^\star) + t\dot{X}, \boldsymbol{A}^T\boldsymbol{A}\big((\lambda+1)\dot{X} + t\ddot{X}\big)\big\rangle.\end{aligned} \tag{88}$$

Using (66) in the second term and (19) in the last term we can write

$$\begin{aligned}\dot{\mathcal{E}} = \lambda t^{\lambda-1}\alpha\left(\Phi(X) - \Phi(x^\star)\right) + t^\lambda \alpha \langle \xi, \dot{X}\rangle + \frac{(\lambda-2)t^{\lambda-3}}{2}\big\|\boldsymbol{A}\big(\lambda(X - x^\star) + t\dot{X}\big)\big\|^2 \\ + t^{\lambda-2}\Big\langle \lambda(X - x^\star) + t\dot{X}, (\lambda + 1 - r)\boldsymbol{A}^T\boldsymbol{A}\dot{X} - t\alpha\xi\Big\rangle\end{aligned} \tag{89}$$

for all $\xi \in \partial\Phi(X)$. Note that the second term and the part of the last one cancels out, thus

$$\begin{aligned}\dot{\mathcal{E}} = \lambda t^{\lambda-1}\alpha\left(\Phi(X) - \Phi(x^\star)\right) + \frac{(\lambda-2)t^{\lambda-3}}{2}\big\|\boldsymbol{A}\big(\lambda(X - x^\star) + t\dot{X}\big)\big\|^2 \\ + \lambda(\lambda + 1 - r)t^{\lambda-2}\big\langle X - x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\big\rangle - \lambda t^{\lambda-1}\alpha\big\langle X - x^\star, \xi\big\rangle + (\lambda + 1 - r)t^{\lambda-1}\big\|\boldsymbol{A}\dot{X}\big\|^2.\end{aligned} \tag{90}$$

Expanding the second term above and simplifying,

$$\begin{aligned}\dot{\mathcal{E}} &= \lambda t^{\lambda-1}\alpha\left(\Phi(X) - \Phi(x^\star)\right) + \frac{\lambda^2(\lambda-2)t^{\lambda-3}}{2}\big\|\boldsymbol{A}(X - x^\star)\big\|^2 \\ &\quad + \lambda(2\lambda - 1 - r)t^{\lambda-2}\big\langle X - x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\big\rangle - \lambda t^{\lambda-1}\alpha\big\langle X - x^\star, \xi\big\rangle \\ &\quad + \left(\lambda + 1 - r + \frac{\lambda-2}{2}\right)t^{\lambda-1}\big\|\boldsymbol{A}\dot{X}\big\|^2.\end{aligned} \tag{91}$$

Note that the coefficient in the last term vanishes since $\lambda \equiv 2r/3$, thus

$$\begin{aligned}\dot{\mathcal{E}} &= \lambda t^{\lambda-1}\alpha\left(\Phi(X) - \Phi(x^\star)\right) + \frac{\lambda^2(\lambda-2)t^{\lambda-3}}{2}\big\|\boldsymbol{A}(X - x^\star)\big\|^2 \\ &\quad + \lambda\left(\frac{\lambda}{2} - 1\right)t^{\lambda-2}\big\langle X - x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\big\rangle - \lambda t^{\lambda-1}\alpha\big\langle X - x^\star, \xi\big\rangle.\end{aligned} \tag{92}$$

Since Φ is strongly convex, according to (11) we have

$$\begin{aligned}\langle X - x^\star, \xi\rangle &\geq \Phi(X) - \Phi(x^\star) + \frac{\mu}{2}\|X - x^\star\|^2 \\ &\geq \Phi(X) - \Phi(x^\star) + \frac{\mu}{2}\frac{\|\boldsymbol{A}(X - x^\star)\|^2}{\|\boldsymbol{A}\|^2}.\end{aligned} \tag{93}$$

Using (93) in the last term of (92) and simplifying,

$$\begin{aligned}\dot{\mathcal{E}} &\le \frac{\lambda t^{\lambda-3}}{2}\left(\lambda(\lambda-2) - \frac{\mu t^2 \alpha}{\|\boldsymbol{A}\|^2}\right)\big\|\boldsymbol{A}(X-x^\star)\big\|^2 \\ &\quad + \lambda\left(\tfrac{\lambda}{2}-1\right)t^{\lambda-2}\big\langle X-x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\big\rangle.\end{aligned} \tag{94}$$

Finally, defining

$$t_0 \equiv \sqrt{\frac{\lambda(\lambda-2)\|\boldsymbol{A}\|^2}{\mu\alpha}} \tag{95}$$

the first term in (94) is negative for all $t \ge t_0$, therefore

$$\dot{\mathcal{E}} \le \frac{\lambda}{2}\left(\frac{\lambda}{2}-1\right)t^{\lambda-2}\frac{d}{dt}\|\boldsymbol{A}(X-x^\star)\|^2 \tag{96}$$

for all $t \ge t_0$.

The strategy now is to integrate (96) directly to obtain an upper bound on $\mathcal{E}|_t$, which by the form of (87) automatically provides a bound on $\Phi(X) - \Phi(x^\star)$. Thus, integrating (96) from t_0 to t, and using integration by parts on the right hand side, i.e.

$$\begin{aligned}\int_{t_0}^t s^{\lambda-2}\frac{d}{ds}\big\|\boldsymbol{A}\big(X(s)-x^\star\big)\big\|^2 ds &= t^{\lambda-2}\big\|\boldsymbol{A}\big(X(t)-x^\star\big)\big\|^2 - t_0^{\lambda-2}\big\|\boldsymbol{A}\big(X(t_0)-x^\star\big)\big\|^2 \\ &\quad - (\lambda-2)\int_{t_0}^t s^{\lambda-3}\big\|\boldsymbol{A}\big(X(s)-x^\star\big)\big\|^2 ds,\end{aligned} \tag{97}$$

we obtain

$$\begin{aligned}&\mathcal{E}|_t - \mathcal{E}|_{t_0} + \tfrac{\lambda}{2}\left(\tfrac{\lambda}{2}-1\right)\Big\{t_0^{\lambda-2}\big\|\boldsymbol{A}\big(X(t_0)-x^\star\big)\big\|^2 \\ &\qquad + (\lambda-2)\int_{t_0}^t s^{\lambda-3}\big\|\boldsymbol{A}\big(X(s)-x^\star\big)\big\|^2 ds\Big\} \le \tfrac{\lambda}{2}\left(\tfrac{\lambda}{2}-1\right)t^{\lambda-2}\big\|\boldsymbol{A}\big(X(t)-x^\star\big)\big\|^2.\end{aligned} \tag{98}$$

Dropping two positive terms on the left hand side of (98) we can also write

$$\mathcal{E}|_t \le \mathcal{E}|_{t_0} + \tfrac{\lambda}{2}\left(\tfrac{\lambda}{2}-1\right)\|\boldsymbol{A}(X(t)-x^\star)\|^2. \tag{99}$$

From the definition (87), and neglecting the positive quadratic term, we thus find

$$\begin{aligned}\alpha\left(\Phi(X(t)) - \Phi(x^\star)\right) &\le \frac{\mathcal{E}|_{t_0}}{t^\lambda} + \frac{\lambda}{2}\left(\frac{\lambda}{2}-1\right)\frac{1}{t^2}\|\boldsymbol{A}(X(t)-x^\star)\|^2 \\ &\le \frac{\mathcal{E}|_{t_0}}{t^\lambda} + \frac{\lambda}{2}\left(\frac{\lambda}{2}-1\right)\frac{1}{t_0^2}\|\boldsymbol{A}(X(t)-x^\star)\|^2\end{aligned} \tag{100}$$

where in the last passage we used the fact that $t \ge t_0$. From the strong convexity relation (93), with X and $x^\star$ interchanged, we also have that

$$\|\boldsymbol{A}(X-x^\star)\|^2 \le \frac{2\|\boldsymbol{A}\|^2}{\mu}\left(\Phi(X) - \Phi(x^\star)\right), \tag{101}$$

which replaced in the last term of (100) yields

$$\left(\alpha - \frac{\lambda}{2}\left(\frac{\lambda}{2}-1\right)\frac{2\|\boldsymbol{A}\|^2}{\mu t_0^2}\right)\left(\Phi(X(t)) - \Phi(x^\star)\right) \le \frac{\mathcal{E}|_{t_0}}{t^\lambda}. \tag{102}$$

Recalling the definition of t_0 in (95), and also that $\lambda = 2r/3$, we finally obtain

$$\Phi(X(t)) - \Phi(x^\star) \le \frac{2\mathcal{E}|_{t_0}}{\alpha t^\lambda} \tag{103}$$

for all $t \ge t_0$.

The above result (103) already shows that the convergence rate of the objective function is $O\left(t^{-2r/3}\right)$. However, there are dependencies on the parameters r, μ and $\|\boldsymbol{A}\|$ inside the constant $\mathcal{E}|_{t_0}$. To extract these dependencies we now bound $\mathcal{E}|_{t_0}$. From the definition (87) we have

$$\mathcal{E}|_{t_0} \le t_0^\lambda \alpha\left(\Phi(X(t_0)) - \Phi(x^\star)\right) + \frac{\lambda^2 t_0^{\lambda-2}}{2}\|\boldsymbol{A}(X(t_0)-x^\star)\|^2 + \frac{t_0^\lambda}{2}\|\boldsymbol{A}\dot{X}(t_0)\|^2. \tag{104}$$

Using (101) in the second term above we obtain

$$\mathcal{E}|_{t_0} \le t_0^\lambda\left(\alpha + \frac{\lambda^2\|\boldsymbol{A}\|^2}{\mu t_0^2}\right)\left(\Phi(X(t_0)) - \Phi(x^\star)\right) + \frac{t_0^\lambda}{2}\|\boldsymbol{A}\dot{X}(t_0)\|^2. \tag{105}$$

Using the definition (95) this simplifies to

$$\mathcal{E}|_{t_0} \le t_0^\lambda \alpha \left\{\left(1 + \frac{\lambda}{\lambda-2}\right)\left(\Phi(X(t_0)) - \Phi(x^\star)\right) + \frac{1}{2\alpha}\|\boldsymbol{A}\dot{X}(t_0)\|^2\right\}. \tag{106}$$

Note that $\frac{\lambda}{\lambda-2} = \left(1-\frac{2}{\lambda}\right)^{-1} = 1 + O\left(\frac{1}{\lambda}\right)$, therefore

$$\mathcal{E}|_{t_0} \le t_0^\lambda \alpha \left\{2\left(\Phi(X(t_0)) - \Phi(x^\star)\right) + \|\boldsymbol{A}\dot{X}(t_0)\|^2\right\}. \tag{107}$$

This implies that there is some constant $C > 0$, independent of r, μ or α, such that (103) becomes

$$\Phi(X) - \Phi(x^\star) \le \frac{t_0^\lambda C}{t^\lambda} \sim \frac{\lambda^\lambda \|\boldsymbol{A}\|^\lambda}{\mu^{\lambda/2}\alpha^{\lambda/2}t^\lambda}, \tag{108}$$

which implies (25) upon replacing $\lambda = 2r/3$. Moreover, from strong convexity we have that $\|X - x^\star\|^2 \le \frac{2}{\mu}\left(\Phi(X) - \Phi(x^\star)\right)$, thus we conclude convergence of the trajectories as well,

$$\|X - x^\star\|^2 \sim \frac{\lambda^\lambda\|\boldsymbol{A}\|^\lambda}{\mu^{1+\lambda/2}\alpha^{\lambda/2}t^\lambda}. \tag{109}$$

□

B.3 Proof of Theorem 10

Theorem 10 (i). Consider

$$\mathcal{E} \equiv t\alpha\left(\Phi(X) - \Phi(x^\star)\right) + \frac{r}{2}\left\|\boldsymbol{A}(X - x^\star)\right\|^2 + \frac{t}{2}\|\boldsymbol{A}\dot{X}\|^2 + \langle X - x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\rangle. \tag{110}$$

where $x^\star$ is a minimizer of the convex Φ, and let $X = X(t)$ be a trajectory of (21). Taking its total time derivative,

$$\begin{aligned}\dot{\mathcal{E}} = {} & \alpha\left(\Phi(X) - \Phi(x^\star)\right) + t\alpha\frac{d}{dt}\Phi(X) + r\big\langle X - x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\big\rangle + t\big\langle \dot{X}, \boldsymbol{A}^T\boldsymbol{A}\ddot{X}\big\rangle \\ & + \big\langle X - x^\star, \boldsymbol{A}^T\boldsymbol{A}\ddot{X}\big\rangle + \|\boldsymbol{A}\dot{X}\|^2.\end{aligned} \tag{111}$$

Using (21) in the terms containing $\ddot{X}$ and also (66) in the second term we can simplify this expression to obtain

$$\dot{\mathcal{E}} = \alpha\left(\Phi(X) - \Phi(x^\star) - \big\langle X - x^\star, \xi\big\rangle\right) + (1 - tr)\|\boldsymbol{A}\dot{X}\|^2. \tag{112}$$

From convexity of Φ, i.e. relation (10), the first term in parenthesis is negative. Regarding the second term, we define $t_0 \equiv 1/r$ to conclude that $\dot{\mathcal{E}} \le 0$ for all $t \ge t_0$.

It remains to show that $\mathcal{E} \ge 0$. To this end, consider writing (110) in the following form:

$$\mathcal{E} = \alpha t\left(\Phi(X) - \Phi(x^\star)\right) + \widetilde{\mathcal{E}} \tag{113}$$

where

$$\widetilde{\mathcal{E}} \equiv \frac{r}{2}\|\boldsymbol{A}(X - x^\star)\|^2 + \big\langle X - x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}\big\rangle + \frac{t}{2}\|\boldsymbol{A}\dot{X}\|^2. \tag{114}$$

Note that we can further write $\widetilde{\mathcal{E}}$ in matrix form,

$$\widetilde{\mathcal{E}} = \tfrac{1}{2}v^T\boldsymbol{M}v, \quad \boldsymbol{M} = \boldsymbol{M}_1 \otimes \boldsymbol{I}_{2m}, \quad \boldsymbol{M}_1 \equiv \begin{bmatrix} r & 1 \\ 1 & t\end{bmatrix}, \quad v = \begin{bmatrix}\boldsymbol{A}(X - x^\star) \\ \boldsymbol{A}\dot{X}\end{bmatrix}, \tag{115}$$

where $\boldsymbol{I}_{2m}$ is the $2m \times 2m$ identity matrix and $\otimes$ denotes the Kronecker product. It is well-known that for arbitrary matrices $\boldsymbol{A} \in \mathbb{R}^{n_A \times n_A}$ and $\boldsymbol{B} \in \mathbb{R}^{n_B \times n_B}$ the eigenvalues of $\boldsymbol{A} \otimes \boldsymbol{B}$ have the form $\lambda_i(\boldsymbol{A})\lambda_j(\boldsymbol{B})$ for $i = 1, \dots, n_A$ and $j = 1, \dots, n_B$. In our case, $\lambda_j(\boldsymbol{B}) = \lambda_j(\boldsymbol{I}_{2m}) = 1$ for all $j = 1, \dots, 2m$, thus the matrix $\boldsymbol{M}$ have only positive eigenvalues if and only if the eigenvalues of $\boldsymbol{M}_1$ are posivite. In other words, it is sufficient to show that $\boldsymbol{M}_1$ is positive semidefinite, which can be done by showing that there exists a matrix $\boldsymbol{N}_1$ such that $\boldsymbol{M}_1 = \boldsymbol{N}_1^T\boldsymbol{N}_1$. This is indeed the case which can be verified by direct computation with

$$\boldsymbol{N}_1 = \begin{bmatrix} t^{-1/2} & t^{1/2} \\ (r-1)^{1/2}t^{-1/2} & 0\end{bmatrix}. \tag{116}$$

Therefore, $\widetilde{\mathcal{E}} \geq 0$ for all $t \geq 0$, which also implies that $\mathcal{E} \geq 0$ for all $t \geq 0$. Since $\dot{\mathcal{E}} \leq 0$ for all $t \geq t_0$, where we recall that $t_0 \equiv 1/r$, it implies that $\mathcal{E}|_t \leq \mathcal{E}|_{t_0}$ for all $t \geq t_0$. Thus, after dropping the positive term $\widetilde{\mathcal{E}}$ in (113), we conclude that

$$\Phi(X) - \Phi(x^\star) \leq \frac{\mathcal{E}|_{t_0}}{\alpha t}. \tag{117}$$

To remove the dependency on r and $\|\boldsymbol{A}\|$ from the constant $\mathcal{E}|_{t_0}$ note that

$$\begin{aligned}\mathcal{E}_{t_0} = r\left\|\boldsymbol{A}\left(X(t-0)\right) - x^\star)\right\|^2 + \frac{1}{2r}\|\boldsymbol{A}\dot{X}(t_0)\|^2 + \frac{\alpha}{r}\left(\Phi(X(t_0)) - \Phi(x^\star)\right) \\ + \big\langle X(t_0) - x^\star, \boldsymbol{A}^T\boldsymbol{A}\dot{X}(t_0)\big\rangle.\end{aligned} \tag{118}$$

The only term that grows with r is the first term. The remaining ones decay or remain constant. This allows us to conclude that there exists some constant $C > 0$, which does not depend on r or other parameters, such that

$$\Phi(X) - \Phi(x^\star) \leq \frac{r\sigma_1^2(\boldsymbol{A})}{\alpha t}\, C, \tag{119}$$

which finishes the proof. □

Theorem 10 (ii). Consider

$$\mathcal{E} \equiv e^{2rt/3}\Big\{\alpha\left(\Phi(X) - \Phi(x^\star)\right) + \frac{r^2}{9}\|\boldsymbol{A}(X - x^\star)\|^2 + \frac{1}{2}\|\boldsymbol{A}\dot{X}\|^2 + \frac{2r}{3}\langle \boldsymbol{A}(X - x^\star), \boldsymbol{A}\dot{X}\rangle\Big\} \tag{120}$$

where $x^\star$ is the unique minimizer of Φ and we denote $X = X(t)$ a trajectory of the dynamical system (21). Taking the total time derivative of (120), replacing (21) in terms containing $\ddot{X}$, and using (66), after straightforward simplifications we obtain

$$\dot{\mathcal{E}} \leq \frac{2re^{2rt/3}\alpha}{3}\Big(\Phi(X) - \Phi(x^\star) - \big\langle X - X^\star, \xi\big\rangle\Big) + \frac{2r^3e^{2rt/3}}{27}\|\boldsymbol{A}(X - X^\star)\|^2, \tag{121}$$

where $\xi \in \partial\Phi(X)$. Since Φ is strongly convex, from (11) we have

$$\Phi(X) - \Phi(x^\star) - \langle \xi, X - x^\star\rangle \leq -\frac{\mu}{2}\|X - x^\star\|^2 \leq -\frac{\mu\|\boldsymbol{A}(X - x^\star)\|^2}{2\|\boldsymbol{A}\|^2}. \tag{122}$$

Replacing (122) into the first term of (121) we conclude that

$$\dot{\mathcal{E}} \leq \frac{re^{2rt/3}}{3}\left(\frac{2r^2}{9} - \frac{\mu\alpha}{\|\boldsymbol{A}\|^2}\right)\|\boldsymbol{A}(X - x^\star)\|^2. \tag{123}$$

Defining

$$r_\mu \equiv \sqrt{\frac{9\mu\alpha}{2\|\boldsymbol{A}\|^2}} \tag{124}$$

we thus have $\dot{\mathcal{E}} \le 0$, provided $r \le r_\mu$.

We still need to show that $\mathcal{E} \ge 0$. Using (122) with X and $x^\star$ interchanged, so that $0 \in \partial\Phi(x^\star) = 0$, it implies that (120) is lower bounded as

$$\mathcal{E} \ge e^{2rt/3}\Big\{ c_1 \|\boldsymbol{A}(X - x^\star)\|^2 + \tfrac{1}{2}\|\boldsymbol{A}\dot{X}\|^2 + 2c_2 \big\langle \boldsymbol{A}(X - x^\star), \boldsymbol{A}\dot{X} \big\rangle \Big\} \tag{125}$$

where

$$c_1 = \left(\frac{\mu\alpha}{2\|\boldsymbol{A}\|^2} + \frac{r^2}{9} \right), \qquad c_2 = \frac{r}{3}. \tag{126}$$

We can re-write (125) in matrix form,

$$e^{-2rt/3}\mathcal{E} \ge v^T \boldsymbol{M} v, \quad \boldsymbol{M} = \boldsymbol{M}_1 \otimes \boldsymbol{I}_{2m}, \quad \boldsymbol{M}_1 = \begin{bmatrix} c_1 & c_2 \\ c_2 & \frac{1}{2} \end{bmatrix}, \quad v \equiv \begin{bmatrix} \boldsymbol{A}(X - x^\star) \\ \boldsymbol{A}\dot{X} \end{bmatrix}. \tag{127}$$

By the same argument following (115), it suffices to show that $\boldsymbol{M}_1$ is positive semidefinite, which can be done by finding an $\boldsymbol{N}_1$ such that $\boldsymbol{N}_1^T \boldsymbol{N}_1 = \boldsymbol{M}_1$. Such a matrix indeed exists, as one can check by direct computation with

$$\boldsymbol{N}_1 = \begin{bmatrix} \dfrac{r_\mu}{3} & \dfrac{r}{2r_\mu + \sqrt{2}\sqrt{r_\mu^2 + r^2}} \\ \dfrac{r}{3} & \dfrac{2r^2 + r_\mu\sqrt{r_\mu^2 - r^2}}{2(r_\mu^2 + r^2)} \end{bmatrix}, \tag{128}$$

where we recall that r_μ is defined in (124).

Since $\dot{\mathcal{E}} \le 0$ for all $t \ge 0$ we have $\mathcal{E}|_t \le \mathcal{E}|_{t=0}$, and therefore every term in inside the square brackets of (120) is decaying as $e^{-2rt/3}$. However, in this form we cannot isolate the convergence rate of the objective function in a convenient manner. A simple trick, which enforces a slightly stronger constraint, allows us to circumvent this. Consider splitting the term $\Phi(X) - \Phi(X^\star) = \left(\frac{1}{2} + \frac{1}{2}\right)(\Phi(X) - \Phi(X^\star))$ in (120). Then, we obtain the equivalent of inequality (125) as

$$e^{-2rt/3}\mathcal{E} \ge \frac{\alpha}{2}\left(\Phi(X) - \Phi(x^\star)\right) + \Big(\widetilde{c}_1 \|\boldsymbol{A}(X - x^\star)\|^2 + \tfrac{1}{2}\|\boldsymbol{A}\dot{X}\|^2 + 2c_2 \big\langle \boldsymbol{A}(X - x^\star), \boldsymbol{A}\dot{X} \big\rangle \Big) \tag{129}$$

where

$$\widetilde{c}_1 = \left(\frac{\mu\alpha}{4\|\boldsymbol{A}\|^2} + \frac{r^2}{9} \right). \tag{130}$$

Thus, exactly the same argument used in (127)–(128) applies to the second term in square brackets of (129), allowing us to conclude that such a term is positive if we replace $r_\mu^2 \mapsto \widetilde{r}_\mu^2 \equiv \frac{1}{2} r_\mu^2$. Therefore, for $r \le \widetilde{r}_\mu$, we conclude that

$$e^{2rt/3}\frac{\alpha}{2}\left(\Phi(X) - \Phi(x^\star)\right) \le \mathcal{E}|_t \le \mathcal{E}|_{t=0} = \alpha\left(\Phi(x_0) - \Phi(x^\star)\right) + \frac{r^2}{9}\|\boldsymbol{A}(x_0 - x^\star)\|^2, \tag{131}$$

finally yielding

$$\Phi(X) - \Phi(x^\star) \le \alpha^{-1} r^2 e^{-2rt/3} C \tag{132}$$

for some constant C independent of α or r. Note that this result combined with the fact that Φ is strongly convex, thus $\|X - x^\star\|^2 \le \frac{2}{\mu}\left(\Phi(X) - \Phi(x^\star)\right)$, automatically implies exponential convergence of the trajectories as well. □

C Symplectic Integration

We provide a derivation of the discretization method (36) used to simulate the dynamical systems (19) and (21). Note that for any $\tilde{\nabla} f(X(t)) \in \partial\Phi(X(t))$ we can write Hamilton's equations (33) in the form

$$\begin{aligned}\begin{pmatrix} \dot{X} \\ \dot{P} \end{pmatrix} &= \begin{pmatrix} 0 & \boldsymbol{M}^{-1} \\ -\lambda\tilde{\nabla}\Phi & -\dot{\eta} \end{pmatrix} \begin{pmatrix} X \\ P \end{pmatrix} \\ &= \left[\begin{pmatrix} 0 & \boldsymbol{M}^{-1} \\ -\lambda\tilde{\nabla}\Phi & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & -\dot{\eta} \end{pmatrix} \right] \begin{pmatrix} X \\ P \end{pmatrix}. \end{aligned} \tag{133}$$

Starting at instant s, the above equation defines a flow map $\Psi_t : (X(s), P(s)) \mapsto (X(s + t), P(s + t))$. Note, however, that we cannot integrate (133) directly. Thus, the idea is to approximate Ψ_t by splitting (133) into two separate parts:

$$\Psi_t^{(1)} : \quad \dot{X} = \boldsymbol{M}^{-1} P, \qquad \dot{P} = -\lambda\tilde{\nabla}\Phi(X) \tag{134}$$

and

$$\Psi_t^{(2)} : \quad \dot{X} = 0, \qquad \dot{P} = -\dot{\eta} P. \tag{135}$$

Note that the flow $\Psi_t^{(2)}$ can be integrated exactly, i.e.

$$\Psi_t^{(2)} \begin{pmatrix} X(s) \\ P(s) \end{pmatrix} = \begin{pmatrix} X(s) \\ e^{-\eta|_s^t} P(s) \end{pmatrix}. \tag{136}$$

For the flow $\Psi_t^{(1)}$, note that this is the Hamiltonian systems of a conservative system. Thus we can immediately apply e.g. the implicit Euler method to this part, which is a symplectic integrator [40]. Then, we consider the composition

$$\hat{\Psi}_\epsilon \equiv \Psi_\epsilon^{(1)} \circ \Psi_\epsilon^{(2)}. \tag{137}$$

It is possible to show that (137) is a conformal symplectic integrator. Moreover, this method is first order accurate. Now, setting $s = t_k$, where $t_k = k\epsilon$, we obtain

$$p_{k+1/2} \leftarrow e^{-\Delta\eta(t_k)} p_k - \epsilon\lambda\tilde{\nabla}\Phi(x_k), \tag{138a}$$

$$x_{k+1} \leftarrow x_k + \epsilon \boldsymbol{M}^{-1} p_{k+1/2}. \tag{138b}$$

Note that according to (32) we have $\Delta\eta(t_k) = \eta(t_{k+1}) - \eta(t_k) = r\epsilon$ for the dynamical system (21), while $\Delta\eta(t_k) = r\log(t_{k+1}/t_k) = r\log(1+\epsilon/t_k)$ for the dynamical system (19). Note that the above updates require only one computation of the subdifferential per iteration, since the last one can be reused in the next iteration. Note also these updates are explicit and thus cheap.

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, *Convex Analysis.* Princeton University Press, 1996.

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

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

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

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

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

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

[37] 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.

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

[39] 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.

[40] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration.* Springer, 2006.

[41] L. D. Landau and E. M. Lifshitz, *Mechanics.* Butterworth-Heinemann, 1976.

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

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

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

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

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

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

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

[49] H. Goldstein, C. P. Poole, and J. Safko, *Classical Mechanics*. Pearson Education, 2011.

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

[51] H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society* **67** (2005) 301–320.

[52] R. Tibshirani, "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society* **58** (1996) 267–288.

[53] H. Yuan, Y. Zhou, C. J. Li, and Q. Sun, "Differential Inclusions for Modeling Nonsmooth ADMM Variants: A Continuous Limit Theory," *Int. Conf. Mach. Learning* (2019) .