DISSERTATION | DOCTORAL THESIS

Titel | Title

Accelerated schemes via an asymptotically vanishing damping for convex optimization and monotone inclusions

verfasst von | submitted by

David Alexander Hulett

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Doktor der Naturwissenschaften (Dr.rer.nat.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 796 605 405
Dissertationsgebiet lt. Studienblatt | Field of study as it appears on the student record sheet:	Mathematik
Betreut von | Supervisor:	Univ.-Prof. Dr. Radu Ioan Bot Privatdoz.

Acknowledgements

To see a world in a grain of sand
And a heaven in a wild flower,
Hold infinity in the palm of your hand
And eternity in an hour.
- W. Blake.

After almost five years in this Faculty, the list of people I would like to acknowledge is long, so I will try to do it in order.

First of all, I would like to give my heartfelt thanks to my doctoral supervisor, Radu I. Boț, for his guidance and his foresight; not only is he an excellent mathematician, but an excellent person as well.

To Robert Csetnek, my thanks for always being there for a question, especially during the first months of my PhD. To Dang-Khoa Nguyen, my thanks for being one of my first collaborators and for having laid much of the foundation upon which this thesis stands. I want to especially thank Enis Chenchene, who rekindled my motivation to do math, and from whom I've learned so much: it was incredibly fun to work alongside you. To Bettina Hiebl, lifeline in all things administrative, my thanks for all the paperwork-related questions, matters in which I'm deeply illiterate.

To my office mates, who have come and gone over the years, my thanks for creating an excellent workplace. I wouldn't wish to work anywhere else.

I would like to thank the wonderful friends I made in the math department, for all the love and the shared moments, particularly Mischa, for his generosity and love for memes; Pilar and Bea, for all the lunches and the endless conversations in the common room; and Giancarlo, for always asking and caring, and for always going the extra mile.

Gracias a la universidad pública argentina, que me permitió estudiar sin pagar un peso. Gracias a mi directora de tesis de Licenciatura, Rocío Díaz Martín, por las largas horas de estudio a la par y por ser un antídoto contra el mal humor, y gracias a todos los profesores del FaMAF, lugar donde empezó la mejor parte de mi vida.

Gracias a mis amigos de siempre: a Agustín G., por enseñarme el nombre del viento; a Agustín B. y Magui, que con ellos el tiempo se detiene; a Zurdo (perdón, nunca me acuerdo el nombre) por el cine y tardes de verano en la maticueva; a Hilario, por el aguante y por contagiar el amor a la montaña; a Marie, que es como un rayo de sol en una mañana de escarcha; a Mateo, músico, carpintero de barcos, amante de los refrigeradores, por ser lo más cercano que tengo a un hermano.

A mis viejos, Claudia y Eddie, que me trajeron a este mundo, espero que este momento sirva para poder devolverles aunque sea un pedacito de todo el amor que me dieron. En algún pasillo de pueblo, todavía soy el nene que, con una sonrisa sin los dientes de adelante, acaba de aprender a andar en bici sin rueditas.

Hay una línea de Verlaine que ya no volveré a recordar.
Hay una calle próxima que está vedada a mis pasos,
Hay un espejo que me ha visto por última vez,
hay una puerta que he cerrado hasta el fin del mundo.
- J. L. Borges.

Abstract

In this work, we explore the asymptotic properties of certain dynamical systems (i.e., continuous-time methods or schemes) and of certain numerical algorithms (i.e., discrete-time methods or schemes) attached to convex optimization problems and to monotone inclusion problems. More precisely, in the framework of a real Hilbert space, we study: i) the link which exists between the Heavy Ball dynamics and the Nesterov accelerated gradient dynamics, both in the optimization setting and in the monotone operator setting; ii) a second-order system attached to an inclusion problem with a maximally monotone+cocoercive structure; iii) a second-order primal-dual dynamical system attached to a linearly constrained convex minimization problem; and iv) two algorithms devised to solve a hierarchical minimization problem, that is, minimizing and outer function subject to the set of minimizers of an inner function, where both inner and outer functions are convex and have a smooth+nonsmooth structure. In every case, we derive convergence rates for the appropriate objects (e.g., the norm of the governing operator along the generated trajectories or iterates in the monotone operator setting, and the functional values along the trajectories or iterates in the optimization setting) and establish weak convergence of the trajectories or iterates towards solution of the corresponding problem. Weak convergence is often the best we can expect, in the context of a (possibly) infinite-dimensional Hilbert space.

At the core of our work lies a so-called asymptotically vanishing coefficient which damps the velocity in our second-order continuous-time schemes. When such schemes are discretized, they result in numerical algorithms with momentum, that is, where we take into account the direction defined by the current and the previous iterate to define the new iterate. Ever since the seminal 1983 paper by Y. Nesterov, it has observed that this vanishing damping has accelerating effects not only in convex optimization, but also reaching into monotone operator theory. The systems and algorithms which we study add to the already considerable amount of evidence on the usefulness of considering such a damping.

Zusammenfassung

In dieser Arbeit untersuchen wir die asymptotischen Eigenschaften bestimmter dynamischer Systeme (d. h. zeitkontinuierlicher Methoden oder Verfahren) und bestimmter numerischer Algorithmen (d. h. zeitdiskreter Methoden oder Verfahren) im Zusammenhang mit konvexen Optimierungsproblemen und monotonen Inklusionsproblemen. Genauer gesagt untersuchen wir im Rahmen eines realen Hilbert-Raums: i) die Verbindung zwischen der Heavy-Ball-Dynamik und der beschleunigten Nesterov-Gradientendynamik sowohl im Optimierungs- als auch im monotonen Operator-Kontext; ii) ein System zweiter Ordnung im Zusammenhang mit einem Inklusionsproblem mit maximal monotoner+kokoerzitiver Struktur; iii) ein primalduales dynamisches System zweiter Ordnung im Zusammenhang mit einem linear beschränkten konvexen Minimierungsproblem; und iv) zwei Algorithmen zur Lösung eines hierarchischen Minimierungsproblems, d. h. der Minimierung einer äußeren Funktion in Abhängigkeit von der Menge der Minimierer einer inneren Funktion, wobei sowohl die innere als auch die äußere Funktion konvex sind und eine glatte+nichtglatte Struktur aufweisen. In jedem Fall leiten wir Konvergenzraten für die entsprechenden Objekte ab (z. B. die Norm des maßgeblichen Operators entlang der generierten Trajektorien oder Iterationen im monotonen Operator-Setting und die Funktionswerte entlang der Trajektorien oder Iterationen im Optimierungs-Setting) und stellen eine schwache Konvergenz der Trajektorien oder Iterationen in Richtung der Lösung des entsprechenden Problems fest. Schwache Konvergenz ist oft das Beste, was wir im Kontext eines (möglicherweise) unendlichdimensionalen Hilbert-Raums erwarten können.

Im Mittelpunkt unserer Arbeit steht ein sogenannter asymptotisch verschwindender Koeffizient, der die Geschwindigkeit in unseren zeitkontinuierlichen Verfahren zweiter Ordnung dämpft. Werden solche Verfahren diskretisiert, resultieren sie in numerischen Algorithmen mit Impuls, d. h., wir berücksichtigen die durch das aktuelle und das vorherige Iterationsmuster definierte Richtung bei der Definition des neuen Iterationsmusters. Seit der bahnbrechenden Arbeit von Y. Nesterov aus dem Jahr 1983 ist bekannt, dass diese verschwindende Dämpfung nicht nur in der konvexen Optimierung, sondern auch in der monotonen Operatortheorie beschleunigende Effekte hat. Die von uns untersuchten Systeme und Algorithmen ergänzen die bereits vorhandenen Belege für die Nützlichkeit der Berücksichtigung einer solchen Dämpfung.

Contents

Chapter 1

Introduction

1.1 The story so far

Mathematical optimization is a vast field which encompasses a plethora of subfields, each with its own history. This thesis touches topics both within optimization and monotone operator theory, two areas of mathematics which are tightly interlinked and which have seen tremendous interest and progress since the early 1960s, having found applications in a wide array of disciplines, including mechanics, economics, partial differential equations, information theory, approximation theory, signal and image processing, game theory, optimal transport theory, probability and statistics, and machine learning.

Specifically, this work explores continuous- and discrete-time methods (not to be confused with continuous and discrete optimization) suited for two classes of problems: i) convex optimization and ii) finding the roots or zeros of a maximally monotone operator (also known as a monotone inclusion problem). In the following, we will give an overview of the results this thesis builds and expands upon.

1.1.1 Unconstrained minimization

A good place to start is the age-old global minimization problem. In a real Hilbert space $\mathcal{H}$, for a convex and differentiable function $f : \mathcal{H} \to \mathbb{R}$, consider

$$\begin{aligned} \min \quad & f(x), \\ \text{subject to} \quad & x \in \mathcal{H}, \end{aligned} \tag{1.1}$$

which we assume to have a solution. Perhaps the most known approach to solve (1.1) is the gradient descent method, namely, starting from an initial point $x_0 \in \mathcal{H}$, we iteratively define, for every $k \geq 0$,

$$x_{k+1} = x_k - s\nabla f(x_k), \tag{1.2}$$

where $s > 0$ is some stepsize. It is known that if ∇f is L-Lipschitz continuous, with $L > 0$, then taking $s \in (0, 2/L)$ guarantees that $f(x_k) - \min f = o(1/k)$ and $\|\nabla f(x_k)\| = o(1/k)$ as $k \to +\infty$. Furthermore, x_k converges weakly to a solution to (1.1) as $k \to +\infty$. It is worth mentioning that there exist workarounds if the Lipschitz constant is unknown beforehand; namely, where the stepsize is defined adaptively. For such an algorithm, we refer the reader to the work by Traoré and Pauwels [97].

In a seminal paper in 1983, Nesterov proposed a simple but fundamental change in the algorithm, with far-reaching consequences. Instead of ditching the iterate x_{k-1} we keep it and go along the direction of $x_k - x_{k-1}$ before executing the gradient step: given initial points $x_0, x_1 \in \mathcal{H}$, one version of Nesterov's algorithm may be expressed, for every $k \geq 1$, as

$$\begin{cases} y_k &= x_k + \dfrac{k-1}{k+2}(x_k - x_{k-1}), \\ x_{k+1} &= y_k - s\nabla f(y_k). \end{cases} \tag{1.3}$$

At no extra computational complexity, choosing $s \in (0, 1/L]$ ensures the convergence rate $f(x_k) - \min f = \mathcal{O}(1/k^2)$ as $k \to +\infty$. The previous is not the original form of Nesterov's algorithm, which has a more involved coefficient accompanying $x_k - x_{k-1}$, but which asymptotically approaches $\frac{k-1}{k+2}$. The reason to have (1.6) is because of the connection with its corresponding continuous-time counterpart, which we will present shortly.

The gradient descent algorithm (1.2) may be regarded as a discrete-time version of a corresponding continuous-time dynamical system. Indeed, given some initial time t_0, for $t \geq t_0$ consider

$$\dot{x}(t) + \nabla f(x(t)) = 0. \tag{1.4}$$

A straightforward Eulerian discretization of (1.4), i.e., approximating $x(t) \approx x_k$ and $\dot{x}(t) \approx \frac{x_{k+1}-x_k}{s}$, will yield back (1.2). What is remarkable is that the dynamics (1.4) enjoy the same convergence properties as the gradient descent algorithm; namely, for any trajectory $t \mapsto x(t)$ solution to (1.4) it holds $f(x(t)) - \min f = o(1/t)$ and $\left\|\dot{x}(t)\right\| = \|\nabla f(x(t))\| = o(1/t)$ as $t \to +\infty$. Furthermore, $x(t)$ converges weakly to a solution to (1.1) as $t \to +\infty$.

It turns out that a similar statement may be said for Nesterov's accelerated gradient algorithm. In 2014, Su, Boyd and Candès [94] provided a continuous-time analog to (1.3). By assuming that the iterates x_k arise from a smooth curve $x(t)$ through the approximation $x_k \approx x(k\sqrt{s})$, the authors showed that a limit process as $s \to 0$ gives rise to the following second-order differential equation: for $t \geq t_0 > 0$,

$$\ddot{x}(t) + \frac{3}{t}\dot{x}(t) + \nabla f(x(t)) = 0. \tag{1.5}$$

The "magic" constant 3 in the coefficient $\frac{3}{t}$ comes from the fact that $\frac{k-1}{k+2} \approx 1 - \frac{3}{k}$ for large enough k. Again, the properties of the numerical algorithm carry over to the continuous-time setting: any solution $t \mapsto x(t)$ to (1.5) satisfies $f(x(t)) - \min f = \mathcal{O}(1/t^2)$ as $t \to +\infty$. In fact, they showed that the same rate holds for any solution to

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0, \tag{1.6}$$

where $\alpha \geq 3$. The weak convergence of the trajectories towards solutions to (1.1) when $\alpha > 3$ has been shown by Attouch, Chbani, Peypouquet and Redont [11] and May [76], together with the improved rate of convergence $f(x(t)) - \min f = o(1/t^2)$ and $\left\|\dot{x}(t)\right\| = o(1/t)$ as $t \to +\infty$. Meaningful convergence rates may also be derived in the subcritical case $\alpha < 3$, as shown by Attouch, Chbani and Riahi [16]. In addition to the viscous term $\frac{\alpha}{t}$ damping the velocity, the following system with Hessian-driven damping was considered by Attouch, Peypouquet and Redont in [26]

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \xi\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0, \tag{1.7}$$

where $\xi \geq 0$. While preserving the fast convergence properties of (1.6), the Hessian-driven damping term reduces in practice the oscillatory aspect of the trajectories when $\xi > 0$.

In [15, 14], Attouch, Chbani and Riahi analyzed a more general version of (1.6) which reads, for $t \geq t_0 > 0$,

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta(t)\nabla f(x(t)) = 0, \tag{1.8}$$

where $\beta : [t_0, +\infty) \to \mathbb{R}_{++}$ is differentiable and nondecreasing. They explain how the presence of β can be interpreted as the result of a time reparametrization of (1.6), which can yield faster convergence rates for the functional values provided β is correctly chosen. More specifically, if β satisfies the growth condition

$$\sup_{t \geq t_0} \frac{t\dot{\beta}(t)}{\beta(t)} \leq \alpha - 3, \tag{1.9}$$

then it enters the convergence rates of the functional values as $f(x(t)) - \min f = \mathcal{O}(1/t^2\beta(t))$ as $t \to +\infty$. The convergence rate is actually little-o fast if (1.9) is a strict inequality. It is worth pointing out that, while β can clearly accelerate the convergence rates in the continuous-time setting, the only explicit discretization of (1.8) we can have is for the case $\beta(t) \equiv 1$. The trade-off for benefiting from the accelerating effects of a parameter β_k which approaches infinity as $k \to +\infty$ is having to compute (often expensive) a proximal evaluation of f at each step, as opposed to directly evaluating ∇f.

Attouch, Chbani and Riahi [14] addressed the more general so-called inertial gradient system, for $t \geq t_0 > 0$,

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \beta(t)\nabla f(x(t)) = 0, \tag{1.10}$$

where convergence rates depending only on $\gamma(\cdot)$ and $\beta(\cdot)$ can be ensured for the functional values provided these parameters satisfy a growth condition akin to (1.9), but which also involves γ.

If we combine the ideas from (1.7) and (1.10) we obtain, for $t \geq t_0 > 0$,

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \xi(t)\nabla^2 f(x(t))\dot{x}(t) + \beta(t)\nabla f(x(t)) = 0, \tag{1.11}$$

which was studied by Attouch, Chbani, Fadili and Riahi [19]. Again, meaningful convergence rates can be ensured under a growth condition which only involves $\xi(\cdot)$ and $\beta(\cdot)$.

The so-called asymptotically vanishing damping term $\frac{\alpha}{t}$ is at the heart of this thesis. As we will see later, the presence of such a term has had tremendous success in producing dynamics with fast convergence properties, with far-reaching effects that go beyond the optimization setting and into monotone operator theory.

Before we move to constrained optimization, we mention another landmark in the chapter of second-order dynamics attached to (1.1). The Heavy Ball dynamics, which read, for $t \geq t_0 \geq 0$,

$$\ddot{x}(t) + \lambda\dot{x}(t) + \nabla f(x(t)) = 0, \tag{1.12}$$

were first introduced by Polyak [86, 85]. The name comes from the mechanical interpretation of this system: $x(t)$ describes the horizontal position of an object which moves alongside the graph of the function f in a medium with viscous friction coefficient $\lambda > 0$. For more details about this derivation, we refer the reader to [21]. Álvarez [2] first showed in 2000 that any trajectory $x(t)$ generated by the Heavy Ball dynamics satisfies $f(x(t)) \to \min f$ and converges weakly to a global minimizer of f as $t \to +\infty$. Later, it was proved that the functional values actually exhibit a rate $f(x(t)) - \min f = \mathcal{O}\left(\frac{1}{t}\right)$ as $t \to +\infty$. Again, we emphasize the accelerating effect of the vanishing damping $\frac{\alpha}{t}$ on the rates for the functional values, as opposed to a constant friction λ.

1.1.2 Linearly constrained minimization

So far, we have mentioned some fundamental contributions pertaining to the world of unconstrained optimization, i.e., where the feasible set is the entire Hilbert space $\mathcal{H}$. In the following, we will see how the accelerating effects of the damping $\frac{\alpha}{t}$ carry over to the constrained optimization setting. We start with the large class of problems

$$\begin{array}{rl} \min & f(x), \\ \text{subject to} & Ax = b, \end{array} \tag{1.13}$$

where $\mathcal{X}, \mathcal{Y}$ are real Hilbert spaces, $f : \mathcal{X} \to \mathbb{R}$ is convex and continuously differentiable, and $A : \mathcal{X} \to \mathcal{Y}$ is a continuous linear operator, with $b \in \mathcal{Y}$. This model formulation underlies many important applications in various areas, such as image recovery [58], machine learning [44, 73], the energy dispatch of power grids [99, 100], distributed optimization [74, 103] and network optimization [91, 102]. In order to approach constrained optimization problems, Augmented Lagrangian Method (ALM) [88] (for linearly constrained problems) and Alternating Direction Method of Multipliers (ADMM) [57, 44] (for problems with separable objectives and block variables linearly coupled in the constraints) and some of their variants have been shown to enjoy substantial success. Continuous-time approaches for structured convex minimization problems formulated in the spirit of the full splitting paradigm have been recently addressed in [38] and, closely connected to our approach, in [102, 63, 18, 43]. The temporal discretization resulting from these dynamics gives rise to the numerical algorithm with fast convergence rates [64, 65] and with a convergence guarantee for the generated iterate [39], without additional assumptions such as strong convexity.

We may approach (1.13) by means of the Lagrangian formulation of the problem. Attached to this problem, the function $\mathcal{L} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ given by

$$\mathcal{L}(x, \lambda) := f(x) + \langle \lambda, Ax - b \rangle$$

is convex on x and concave on λ, i.e., $\mathcal{L}(\cdot, \lambda)$ is convex for every $\lambda \in \mathcal{Y}$ and $\mathcal{L}(x, \cdot)$ is concave for every $x \in \mathcal{X}$. The set of saddle points of $\mathcal{L}$, which we denote by $\mathbb{S}$, consists of the pairs $(x^*, \lambda^*) \in \mathcal{X} \times \mathcal{Y}$ such that

$$\mathcal{L}(x^*, \lambda) \leq \mathcal{L}(x^*, \lambda^*) \leq \mathcal{L}(x, \lambda^*).$$

for every $(x, \lambda) \in \mathcal{X} \times \mathcal{Y}$. To be able to mention meaningful results, we of course need to assume that $\mathbb{S} \neq \emptyset$. Notice that

$$(x^*, \lambda^*) \in \mathbb{S} \Leftrightarrow \begin{cases} \nabla_x \mathcal{L}(x^*, \lambda^*) & = 0 \\ -\nabla_\lambda \mathcal{L}(x^*, \lambda^*) & = 0 \end{cases} \Leftrightarrow \begin{cases} \nabla f(x^*) + A^*\lambda^* & = 0 \\ b - Ax^* & = 0 \end{cases}, \tag{1.14}$$

where $A^* : \mathcal{Y} \to \mathcal{X}$ denotes the adjoint operator of A. In other words, $\mathbb{S}$ is precisely the set of primal-dual solutions to (1.13), i.e., pairs (x^*, λ^*) where x^* is an optimal solution to (1.13) and λ^* is a solution to the Lagrange dual problem. If x^* is an optimal solution to (1.13) and a suitable constraint qualification is fulfilled, then there exists a solution $\lambda^* \in \mathcal{Y}$ to the Lagrange dual problem such that (x^*, λ^*) is a primal-dual solution to (1.13). For details and insights into the topic of constraint qualifications for convex duality we refer to [28, 32]. We call $\mathbb{F} := \{x \in \mathcal{X} : Ax = b\}$ the set of feasible points of (1.13). For $\mu \geq 0$, we consider also the augmented Lagrangian $\mathcal{L}_\mu : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ associated with (1.13)

$$\mathcal{L}_\mu(x, \lambda) := \mathcal{L}(x, \lambda) + \frac{\mu}{2}\|Ax - b\|^2 = f(x) + \langle \lambda, Ax - b \rangle + \frac{\mu}{2}\|Ax - b\|^2. \tag{1.15}$$

For every $(x, \lambda) \in \mathbb{F} \times \mathcal{Y}$ it holds

$$f(x) = \mathcal{L}_\mu(x, \lambda) = \mathcal{L}(x, \lambda). \tag{1.16}$$

If $(x^*, \lambda^*) \in \mathbb{S}$, then for every $(x, \lambda) \in \mathcal{X} \times \mathcal{Y}$ we have

$$\mathcal{L}(x^*, \lambda) = \mathcal{L}_\mu(x^*, \lambda) \leq \mathcal{L}(x^*, \lambda^*) = \mathcal{L}_\mu(x^*, \lambda^*) \leq \mathcal{L}(x, \lambda^*) \leq \mathcal{L}_\mu(x, \lambda^*).$$

In addition, from (1.14) we have

$$\begin{aligned}
(x^*, \lambda^*) \in \mathbb{S} \ \Leftrightarrow \ & \begin{cases} \nabla f(x^*) + A^*\lambda^* & = 0, \\ Ax^* - b & = 0 \end{cases} \ \Leftrightarrow \ \begin{cases} \nabla f(x^*) + A^*\lambda^* + \mu A^*(Ax^* - b) & = 0, \\ Ax^* - b & = 0 \end{cases} \\
\Leftrightarrow \ & \begin{cases} \nabla_x \mathcal{L}_\mu(x^*, \lambda^*) & = 0, \\ \nabla_\lambda \mathcal{L}_\mu(x^*, \lambda^*) & = 0. \end{cases}
\end{aligned}$$

In other words, for any $\mu \geq 0$ the sets of saddle points of $\mathcal{L}$ and $\mathcal{L}_\mu$ are identical.

The previous considerations give rise to the following primal-dual dynamical system attached to (1.13), which for $t \geq t_0 > 0$ reads

$$\begin{cases} \ddot{x}(t) + \gamma(t)\dot{x}(t) + \beta(t)\nabla_x \mathcal{L}_\mu\Big(x(t), \lambda(t) + \theta(t)\dot{\lambda}(t)\Big) & = 0, \\ \ddot{\lambda}(t) + \gamma(t)\dot{\lambda}(t) - \beta(t)\nabla_\lambda \mathcal{L}_\mu\Big(x(t) + \theta(t)\dot{x}(t), \lambda(t)\Big) & = 0. \end{cases} \tag{1.17}$$

We mention now some notable instances of these dynamics which are relevant to our work. The case with general parameters $\gamma(\cdot), \beta(\cdot)$ and $\theta(\cdot)$ was addressed by Attouch, Chbani, Fadili, and Riahi [18] and He, Hu and Fang [63]. The case where $\gamma(t) = \frac{\alpha}{t^r}$ and $\theta(t) = \theta t^s$ was studied by He, Hu and Fang [66]. Of particular interest to us are the choices $\gamma(t) = \frac{\alpha}{t}, \theta(t) = \theta t$ and $\beta(t) \equiv 1$: this setting was studied by Zeng, Lei and Chen [102]. It was also addressed by Boţ and Nguyen [43], and to the best of our knowledge, they were the first to produce simultaneous fast convergence rates for the functional values, the primal-dual gap, and the feasibility measure of the trajectories: as $t \to +\infty$, they showed that

$$|f(x(t)) - f(x^*)| = \mathcal{O}\left(\frac{1}{t^2}\right), \quad \mathcal{L}(x(t), \lambda^*) - \mathcal{L}(x^*, \lambda(t)) = \mathcal{O}\left(\frac{1}{t^2}\right) \quad \text{and} \quad \|Ax(t) - b\| = \mathcal{O}\left(\frac{1}{t^2}\right),$$

as well as the weak convergence of $(x(t), \lambda(t))$ towards a primal-dual solution to (1.13). These rates evidence that the viscous damping $\frac{\alpha}{t}$ is also successful in producing fast convergent dynamics in linearly constrained optimization. The role of the extrapolation parameter $\theta(\cdot)$ is to induce more flexibility in the dynamical system and in the associated discrete schemes, as it has been recently noticed in [18, 13, 63, 102]. The time scaling function $\beta(\cdot)$ has the role to further improve the rates of convergence of the objective function value along the trajectory, as it was noticed in the context of unconstrained minimization problems in [19, 15, 14] and of linearly constrained minimization problems in [18, 66]. Again, just like in the unconstrained case, we emphasize that only the choice $\beta(t) \equiv 1$ will yield explicit meaningful numerical algorithms. Such an explicit algorithm was formulated in [39], where Boţ, Csetnek and Nguyen provided a discretization of the system in [43], with identical $\mathcal{O}(1/k^2)$ convergence rates of the functional values, primal-dual gap and feasibility measure along the iterates, as well as weak convergence of the iterates towards a primal-dual solution to (1.13).

1.1.3 Hierarchical optimization

Another large class of constrained optimization problems stems from having the feasible sets given by the minimizers of another convex function. This problem often goes by the name of simple bilevel optimization (to emphasize the difference with follower-leader bilevel optimization) or hierarchical optimization.

The problem can be formulated as

$$\begin{array}{rl} \min & H(x) := h(x) + \hat{h}(x), \\ \text{subject to} & x \in \operatorname{argmin}_{z \in \mathcal{H}} \ F(z) := f(z) + \hat{f}(z), \end{array} \tag{1.18}$$

where $f, h : \mathcal{H} \to \mathbb{R}$ are convex and continuously differentiable, and $\hat{f}, \hat{h} : \mathcal{H} \to \overline{\mathbb{R}}$ are proper, convex and lower semicontinuous. Bilevel optimization originated in economic game theory in the 50s [53], and found enormous application in modern applied mathematics, such as in hyperparameter optimization [70, 47], meta-learning [87], and reinforcement learning [31], see [53] and the references therein for a comprehensive account.

A classical approach to tackle (1.18) is by an approximation argument, the so-called viscosity method [7, 6]. This method consists of replacing (1.18) with the regularized optimization problem

$$\min_{x \in \mathcal{H}} \ F(x) + \varepsilon H(x), \quad \text{for } \varepsilon > 0, \tag{1.19}$$

and recovering a solution to (1.18) by considering (weak) cluster points of minimizers of (1.19) as $\varepsilon \to 0$. This classical approach appears across a wide range of mathematical disciplines, including the calculus of variations [55], numerical linear algebra [49], optimal transport [52], control theory and partial differential equations [50], among others. Its origins trace back to the work of Tikhonov [96] in the context of ill-posed linear problems, where the regularization term takes the form $H := \| \cdot \|^2$. While in the so-called exact penalization setting [56], solving (1.19) for ε small enough actually yields solutions to (1.18), such a threshold is usually not known a-priori. Thus, successive solutions of (1.19) with $\varepsilon \equiv \varepsilon_k$ and $\varepsilon_k \to 0$ as $k \to +\infty$ are needed to approach a solution to (1.18). Quite surprisingly, it turns out that we can replace these successive minimization problems by a single iteration of a minimization algorithm. This defines the so-called diagonal or Tikhonov methods.

The first breakthrough in this direction was proposed by Cabot in [45], who replaced the minimization oracle (1.19) with a proximal step on (1.19) and highlighted the role of a slow-vanishing behavior of $(\varepsilon_k)_{k \in \mathbb{N}}$. Later, Solodov provided the first gradient-type algorithm [92] designed for solving (1.18) with $\hat{h} = 0$ and $\hat{f} = \delta_D$, where $D \subseteq \mathcal{H}$ is a nonempty, convex and closed set and δ_D denotes its indicator function. In the smooth setting, i.e., $\hat{h} = \hat{f} = 0$, these methods can be understood as a discretization of the dynamical system, which for $t \geq t_0 \geq 0$ reads

$$\dot{x}(t) + \nabla f(x(t)) + \varepsilon(t) \nabla h(x(t)) = 0, \tag{1.20}$$

where $\varepsilon : [t_0, +\infty) \to \mathbb{R}_{++}$ is a vanishing regularization function. Observe that indeed (1.20) can be seen as a gradient flow applied to (1.19) with a vanishing parameter function $t \mapsto \varepsilon(t)$. The convergence of (1.20) has been first established by Attouch and Czarnecki in [20], who showed that, under a certain growth condition on f, $x(t)$ converges weakly to an optimal solution of (1.18) with a convergence rate for $f(x(t)) - \min f$ expressed in terms of $\varepsilon(t)$, cf. [20, Theorem 5.1]. More recently, by replacing ∇f with a general monotone and Lipschitz continuous operators and setting $h := \frac{1}{2} \| \cdot \|^2$, (1.20) and some of its discretizations have gained renewed interest for their ability to yield strong convergence toward the minimal norm solution, as well as their connection to Halpern-type methods [89, 72, 101], see also [33, 35].

Building on the work of Cabot and Solodov, Peypouquet proposed in [84] an adaptation of the Attouch–Czarnecki condition [20] for gradient systems, establishing weak convergence of the generated sequence to a solution of (1.18). In finite-dimensional settings, convergence results can still be obtained under milder assumptions, e.g., coercivity of H, even in the general framework of (1.18). For example, [77], which complements and extends [68], highlights an inherent trade-off between inner and outer function performance. Recently, backtracking strategies have been developed for functions with merely locally Lipschitz continuous gradients [71]. Additional notable extensions (without claiming to be exhaustive) include bundle methods [93], cutting plane approaches [46], and algorithms tailored for variational inequality inner problems [68].

Building on this rationale, Merchav, Sabach, and Teboulle [78] were the first to observe that applying a single iteration of Nesterov's fast gradient method, or FISTA in the composite (smooth + non-smooth) setting, to (1.19) can accelerate convergence, while still generating sequences that converge weakly to solutions of (1.18). In continuous time, these methods can be interpreted as discretizations of the dynamical system [94], which for $t \geq t_0 > 0$ reads

$$\ddot{x}(t) + \frac{\alpha}{t} \dot{x}(t) + \nabla f(x(t)) + \varepsilon(t) \nabla h(x(t)) = 0, \tag{1.21}$$

where $\alpha > 3$. This system is known to provide a theoretical foundation for analyzing corresponding discrete numerical algorithms, including those involving non-smooth proximal terms.

1.1.4 Finding zeros of maximally monotone operators

So far we have talked about methods designed to tackle problems in convex optimization. It turns out that many of the techniques that work in optimization directly carry over or at least have a close analog suited for the world of maximally monotone operators. As we remarked before, an especially important role is played by the viscous damping coefficient $\frac{\alpha}{t}$.

A good way to introduce monotone operators is to think about gradients: for any differentiable and convex function $f : \mathcal{H} \to \mathbb{R}$, it for every $x, y \in \mathcal{H}$ it holds

$$\langle x - y, \nabla f(x) - \nabla f(y) \rangle \geq 0.$$

Any map $M : \mathcal{H} \to \mathcal{H}$ which satisfies the above inequality is called a monotone operator. We also allow them to be multivalued, i.e., $M(x) \in 2^{\mathcal{H}}$ for every $x \in \mathcal{H}$, in which case the definition just generalizes to $\langle x - y, u - v \rangle \geq 0$ whenever $x, y \in \mathcal{H}$ and $u \in M(x), v \in M(y)$. This extension allows us to fit a further large class of operators into this definition: namely, if $f : \mathcal{H} \to \overline{\mathbb{R}}$ is proper, convex and lower semicontinuous, then its subgradient $\partial f : \mathcal{H} \to 2^{\mathcal{H}}$ is not only monotone, but maximally so, meaning it cannot be properly extended to a larger monotone operator. For this class of functions, Fermat's rule tells us that a point x^* is a global minimizer of f if and only if $0 \in \partial f(x^*)$.

We have already gathered some evidence as to the importance of addressing monotone inclusion problems, i.e., finding the zeros of a maximally monotone operator:

$$0 \in M(x). \tag{1.22}$$

However, (1.22) goes beyond the gradient or subgradient setting. Since we already discussed the class of linearly constrained minimization problems, we seize this opportunity to make the case for monotone operators which do not arise from a potential. We said that (under a mild constraint qualification) solving (1.13) was equivalent to finding the saddle points of the associated Lagrangian $\mathcal{L}$. This is in turn equivalent to finding a pair $(x^*, \lambda^*) \in \mathcal{X} \times \mathcal{Y}$ such that $0 = \nabla_x \mathcal{L}(x^*, \lambda^*)$ and $0 = -\nabla_\lambda \mathcal{L}(x^*, \lambda^*)$. In other words, we are interested in finding the zeros of the map $M : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$ given by

$$M(x, \lambda) = \Big(\nabla_x \mathcal{L}(x, \lambda),\ -\nabla_\lambda \mathcal{L}(x, \lambda) \Big). \tag{1.23}$$

In general, for a convex-concave function $\mathcal{L} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the saddle point problem

$$\min_{x \in \mathcal{X}} \max_{\lambda \in \mathcal{Y}} \mathcal{L}(x, \lambda) \tag{1.24}$$

has a corresponding attached saddle point operator (1.23) which does not come from a potential, i.e., there is no convex function $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that $\nabla F = M$. Thus, formulation (1.22) already encompasses a wider spectrum of problems than optimization. Indeed, both (1.22) and (1.24) are fundamental models in various fields such as optimization (as we have just previously remarked), economics, game theory and partial differential equations. They have recently regained significant attention, in particular in the machine learning and data science community, due to the fundamental role they play, for instance, in multi-agent reinforcement learning [81], robust adversarial learning [75] and generative adversarial networks (GANs) [30, 59].

Given a continuous and monotone operator $M : \mathcal{H} \to \mathcal{H}$, we may ask ourselves if the gradient flow dynamics (1.4) also work in the monotone operator world: attached to (1.22), consider

$$\dot{x}(t) + M(x(t)) = 0. \tag{1.25}$$

It turns out that the only way to ensure weak convergence of the trajectories generated by this system towards solutions to (1.25) is to ask that M fulfills a stronger assumption: namely, we ask M to be ρ-cocoercive, i.e., for $\rho > 0$ it holds $\rho \|M(x) - M(y)\|^2 \leq \langle x - y, M(x) - M(y) \rangle$ for every $x, y \in \mathcal{H}$. For a proof for this result, we refer the readers to the work by Attouch, Boț and Nguyen [9], where the authors additionally show that any solution $t \mapsto x(t)$ to (1.25) also fulfills $\|M(x(t))\| = o(1/\sqrt{t})$ as $t \to +\infty$. In the general case when M is not necessarily cocoercive, weak convergence only holds in an ergodic sense, i.e., it holds for the trajectory $\overline{x}(t) := \frac{1}{t}\int_0^t x(s)ds$. For a simple counterexample, consider the solutions to (1.25) for the case

$$M = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad \text{where one solution is given by} \quad x(t) = \begin{bmatrix} \sin(t) \\ \cos(t) \end{bmatrix},$$

which evidently does not weakly converge to a zero of M as $t \to +\infty$. For more information and counterexamples related to system (1.25), we refer the reader to [48].

Notably, as shown by Attouch and Svaiter in [27], by adding a correction term which is the time derivative of the operator along the trajectory, i.e., for $t \geq t_0 \geq 0$,

$$\lambda(t)\dot{x}(t) + \frac{d}{dt}M(x(t)) + M(x(t)) = 0, \tag{1.26}$$

where $\lambda(\cdot)$ is a positive function, then even in the case of a monotone operator M which is not necessarily cocoercive any trajectory $t \mapsto x(t)$ generated by this system satisfies $\|M(x(t))\| \to 0$ and weakly converges to a zero of M as $s \to +\infty$. For λ constant, the previous system can be rewritten equivalently as being of the form (1.25), but applied to the Moreau-Yosida envelope of M, a new operator given by $M_\rho := \frac{1}{\rho}\Big(\operatorname{Id} - (\operatorname{Id} + \rho M)^{-1}\Big)$. This operator is always ρ-cocoercive and its set of zeros is identical to that of M. We may also obtain convergence rates for $\|M(x(t))\|$ by means of a similar system to (1.26):

$$\dot{x}(t) + \mu(t)\frac{d}{dt}M(x(t)) + \gamma(t)M(x(t)) = 0, \tag{1.27}$$

where under suitable assumptions for $\mu(\cdot)$ and $\gamma(\cdot)$ we not only have weak convergence of trajectories towards zeros of M, but also, as $t \to +\infty$,

$$\|M(x(t))\| = o\left(\frac{1}{\sqrt{\int_{t_0}^t \gamma(s)\mu(s)ds}}\right).$$

There is particular case of the previous mentioned systems which merits being mentioned. If we either take $\lambda(t) \equiv 1$ in (1.26) or $\mu(t) = \gamma(t) \equiv 1$ in (1.27), we obtain the dynamics

$$\dot{x}(t) + \frac{d}{dt}M(x(t)) + M(x(t)) = 0. \tag{1.28}$$

Any solution $t \mapsto x(t)$ to (1.28) satisfies $\|M(x(t))\| = o(1/\sqrt{t})$ as $t \to +\infty$, and $x(t)$ converges weakly to a zero of M as $t \to +\infty$. System (1.28) can then be considered as the "proper" analog to gradient flow in the general monotone operator setting, and emphasizes the importance of the presence of the correcting term $\frac{d}{dt}M(x(t))$, which will appear many times throughout this work. This system is also significant in that it admits one particularly relevant discretization: the so-called Extragradient method, which for a monotone and L-Lipschitz continuous operator M reads, starting at some initial point $x_0 \in \mathcal{H}$,

$$\begin{cases} y_k &= x_k - sM(x_k), \\ x_{k+1} &= x_k - sM(y_k). \end{cases} \tag{1.29}$$

This algorithm was introduced in 1976 by Korpelevich [69] and Antipin [5]. If $0 < s < 1/L$, the generated sequence $(x_k)_{k\geq 1}$ converges weakly towards a zero of M. The question of a last iterate rate of convergence was answered only recently by Gorbunov, Loizou and Gidel [60], where they showed that $\|M(x_k)\| = \mathcal{O}(1/\sqrt{k})$ as $k \to +\infty$.

Let us remark the importance of this convergence rate: as in the monotone operator setting there is often no function giving rise to M (as we have just previously noted with the saddle point operator) one of the only measures of the performance of a method is to analyze the asymptotic behavior of $\|M(x(t))\|$ in the continuous-time case, or of $\|M(x_k)\|$ for the numerical algorithms.

We observed how in the optimization realm moving from first-order to second-order dynamical systems which feature a viscous term $\frac{\alpha}{t}$ damping the velocity was successful in producing faster converging methods. The natural question is to ask ourselves if something similar can be said for monotone operators. First, we mention some relevant systems which are formulated in the spirit of the Heavy Ball dynamics. Álvarez and Attouch [3] and Attouch and Maingé [24] studied the dynamics, for $t \geq t_0 \geq 0$,

$$\ddot{x}(t) + \mu\dot{x}(t) + M(x(t)) = 0,$$

where $M : \mathcal{H} \to \mathcal{H}$ is ρ-cocoercive operator. They were able to prove that the solutions of this system weakly converge to elements of $\operatorname{zer} M$ provided that the cocoercivity parameter ρ and the friction coefficient μ satisfy $\rho\mu^2 > 1$. In the same vein, Boţ and Csetnek [37] considered the system

$$\ddot{x}(t) + \mu(t)\dot{x}(t) + \nu(t)M(x(t)) = 0, \tag{1.30}$$

where M is again ρ-cocoercive. Under the assumption that $\mu(\cdot)$ and $\nu(\cdot)$ are locally absolutely continuous, $\dot{\mu}(t) \leq 0 \leq \dot{\nu}(t)$ for almost every $t \in [0, +\infty)$ and $\inf_{t\geq 0} \frac{\mu^2(t)}{\nu(t)} > \frac{1}{\rho}$, the authors were able to prove that the solutions to this system converge weakly to zeros of M.

Before moving on to accelerating these systems, we mention a few results pertaining to the problem of an additively structured monotone inclusion. Namely, given $A, B : \mathcal{H} \to 2^{\mathcal{H}}$ maximally monotone, consider the problem

$$0 \in (A + B)(x). \tag{1.31}$$

For $\rho > 0$ and a starting point $z_0 \in \mathcal{H}$, the celebrated Douglas-Rachford algorithm may be compactly written, for every $k \geq 0$, as

$$z_{k+1} = \left(\frac{\operatorname{Id} + R_{\rho A}R_{\rho B}}{2}\right)(z_k), \tag{1.32}$$

where R_A denotes the reflected resolvent of a monotone operator A. It is known that (cf. [28, Theorem 26.11]) if $(z_k)_{k\geq 0}$ is the sequence generated by (1.32), then z_k converges weakly to a fixed point of $R_{\rho A}R_{\rho B}$ as $k \to +\infty$ (say $\overline{z}$), and that $J_{\rho B}(z_k)$ converges weakly to $J_{\rho B}(\overline{z}) \in \operatorname{zer}(A + B)$ as $k \to +\infty$, where J_A denotes the resolvent of a monotone operator A. In [51] Csetnek, Malitsky and Tam addressed the problem (1.31), but where $B : \mathcal{H} \to \mathcal{H}$ is monotone and L-Lipschitz continuous. They formulate a dynamical system which, when discretized one way, will yield back the Douglas-Rachford algorithm (1.32), but when discretized in a slightly different way, gives rise to a new algorithm which evaluates A through a backward step, but B through a forward, explicit step. Namely, for $\rho > 0$ and starting points $x_0, x_1 \in \mathcal{H}$, their algorithm reads, for every $k \geq 1$,

$$x_{k+1} = J_{\rho A}(x_k - \rho B(x_k)) - \rho(B(x_k) - B(x_{k-1})). \tag{1.33}$$

The authors show that if $\rho \in (0, 3/L)$, then the sequence $(x_k)_{k\geq 0}$ generated by (1.33) converges weakly to a point $\overline{x} \in \operatorname{zer}(A + B)$; moreover, $B(x_k)$ converges weakly to $B(\overline{x})$ as $k \to +\infty$.

As far as we know, the first analog to the accelerated gradient flow dynamics (1.6) in the operator setting were introduced by Attouch and Peypouquet [25]. For a maximally monotone operator $M : \mathcal{H} \to 2^{\mathcal{H}}$ (i.e., not necessarily single-valued) they formulated a second-order system with asymptotically vanishing damping governed by the Moreau-Yosida approximation $M_{\rho(t)}$ of M. For $t \geq t_0 > 0$, it reads

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + M_{\rho(t)}(x(t)) = 0, \tag{1.34}$$

where $\alpha > 1$ and the time-dependent regularizing parameter $\rho(t)$ behaves like $\rho(t) \simeq t^2$. As well as ensuring the weak convergence of the trajectories towards elements of $\operatorname{zer} M$, choosing the regularizing parameter in such a fashion allowed the authors to obtain fast convergence of the velocities and accelerations towards zero; namely, they show that, as $t \to +\infty$,

$$\|\ddot{x}(t)\| = \mathcal{O}\left(\frac{1}{t^2}\right), \quad \|\dot{x}(t)\| = \mathcal{O}\left(\frac{1}{t}\right) \quad \text{and thus} \quad \|M_{\rho(t)}(x(t))\| = \mathcal{O}\left(\frac{1}{t^2}\right).$$

If M is single-valued and continuous, we would prefer a scheme that evaluates $M(x(t))$ directly, rather than through its Moreau-Yosida envelope, which is often very difficult to compute. Combining the ideas of having a correction term like in (1.26) together with second-order and asymptotic vanishing terms like in (1.34) and a time rescaling coefficient similar to that of (1.8) gives rise to the Fast OGDA dynamics, which for $t \geq t_0 > 0$ read

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta(t)\frac{d}{dt}M(x(t)) + \frac{1}{2}\left(\dot{\beta}(t) + \frac{\alpha}{t}\beta(t)\right)M(x(t)) = 0, \tag{1.35}$$

introduced and studied by Boţ, Csetnek and Nguyen [40]. Provided that $\beta(\cdot)$ fulfills a growth condition almost identical to (1.9) (but with $\alpha - 2$ instead of $\alpha - 3$), the authors show a rate of $o(1/t\beta(t))$ for $\|M(x(t))\|$ and the weak convergence of $x(t)$ towards a zero of M as $t \to +\infty$. A remarkably special case of this system is given when we choose $\beta(\cdot) \equiv 1$:

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \frac{d}{dt}M(x(t)) + \frac{\alpha}{2t}M(x(t)) = 0. \tag{1.36}$$

The authors produce an explicit discretization of (1.36) (i.e., involving only forward evaluations of M) which for starting points $x_0, x_1, y_0 \in \mathcal{H}$ and for every $k \geq 1$ reads

$$\begin{cases} y_k &= x_k + \left(1 - \dfrac{\alpha}{k+\alpha}\right)(x_k - x_{k-1}) - \dfrac{\alpha s}{2(k+\alpha)} M(y_{k-1}), \\ x_{k+1} &= y_k - \dfrac{s}{2}\left(1 + \dfrac{k}{k+\alpha}\right)(M(y_k) - M(y_{k-1})). \end{cases} \tag{1.37}$$

When M is L-Lipschitz continuous, $\alpha > 2$ and $0 < s < 1/2L$, the authors show that the sequence $(x_k)_{k\geq 2}$ produced by this algorithm converges weakly to a zero of M as $k \to +\infty$. Furthermore, it holds $\|M(x_k)\| = o(1/k)$ as $k \to +\infty$. This was a landmark result, as it was the first explicit algorithm to achieve the fastest provable convergence rates and convergence of iterates towards zeros of the operator, within the general class of monotone and Lipschitz continuous operators.

1.2 Structure of the the thesis

This thesis is the culmination of almost five years of work spread across five research papers, and it is organized as follows:

- **Chapter 2:** provides the preliminary definitions and results that are used throughout this work;
- **Chapter 3:** is based on my paper in collaboration with H. Attouch, R. I. Boț and D.-K. Nguyen [8]. We explore a link between the Heavy Ball and the Nesterov accelerated gradient dynamics, both in the optimization and monotone operator settings. A section of this chapter is devoted to mentioning the main results of my work with R. I. Boț and D.-K. Nguyen [42], where we are concerned with a second-order dynamical with a slow vanishing damping ($\frac{\alpha}{t^r}$, with $r \in [0, 1]$) attached to a monotone equation problem;
- **Chapter 4:** is based on my first paper, in collaboration with R. I. Boț [41], and it studies the asymptotic properties of a second-order system with asymptotically vanishing damping attached to a monotone inclusion problem of the form $0 \in (A + B)(x)$, where $A : \mathcal{H} \to 2^{\mathcal{H}}$ is maximally monotone and $B : \mathcal{H} \to \mathcal{H}$ is cocoercive;
- **Chapter 5:** is based on my work in collaboration with D.-K. Nguyen [67], where we formulate a second-order primal-dual dynamical system with asy vanishing damping and a time rescaling coefficient, attached to a linearly constrained minimization problem;
- **Chapter 6:** is based on my paper written with R. I. Boț, E. Chenchene and E. R. Csetnek [34], where we analyze a first- and a second-order dynamical system, and two numerical algorithms which arise from their respective discretizations, attached to a hierarchical or simple bilevel problem, i.e., where we minimize a convex outer function H subject to the set of minimizers of an inner convex function F. Both outer and inner levels present a composite smooth+nonsmooth structure;
- **Chapter 7:** contains some conclusive remarks and poses some open questions for future research.

Chapter 2

Preliminaries

In this chapter we will be laying out the theoretical foundations, i.e., the definitions, results and auxiliary lemmas we will be using throughout this thesis. We will be working within the framework of a real Hilbert space, often denoted by $\mathcal{H}$, with its inner product and canonical norm denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$, respectively.

2.1 Sets, functions and topology

The *extended real line*, which we denote by $\overline{\mathbb{R}}$, consists of $\mathbb{R}$ together with the elements $-\infty$ and $+\infty$, with the added arithmetic rules

$$\begin{gathered}(-\infty)+(+\infty)=(+\infty)+(-\infty)=+\infty,\\ a+(-\infty)=(-\infty)+a=-\infty,\quad a+(+\infty)=(+\infty)+a=+\infty\quad\forall a\in\mathbb{R},\\ 0(+\infty)=(+\infty)0=+\infty,\quad 0(-\infty)=(-\infty)0=0,\\ a(-\infty)=(-\infty)a=\begin{cases}+\infty & \text{if } a<0,\\ -\infty & \text{if } a>0,\end{cases}\quad a(+\infty)=(+\infty)a=\begin{cases}-\infty & \text{if } a<0,\\ +\infty & \text{if } a>0.\end{cases}\end{gathered}$$

We will denote $\mathbb{R}_+ := \{x\in\mathbb{R} : x\geq 0\}$ and $\mathbb{R}_{++} := \{x\in\mathbb{R} : x>0\}$.

Throughout this chapter, let $f, g, f_i : \mathcal{H}\to\overline{\mathbb{R}}$, $i\in I$, be functions. We define the *effective domain* and *epigraph* of f respectively by

$$\operatorname{dom} f := \{x\in\mathcal{H} : f(x)<+\infty\}\qquad\text{and}\qquad \operatorname{epi} f := \{(x,r)\in\mathcal{H}\times\mathbb{R} : f(x)\leq r\}.$$

We say that f is *proper* if $\operatorname{dom} f\neq\emptyset$ and $f(x)>-\infty$ for every $x\in\mathcal{H}$. For the sake of completeness, we allow functions to take the values $\pm\infty$, although throughout this work only functions taking values in $\mathbb{R}\cup\{+\infty\}$ will appear. We say that f is *convex* if for every $x,y\in\mathcal{H}$ and all $\lambda\in[0,1]$ it holds

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

Equivalently, f is convex if and only if its epigraph $\operatorname{epi} f$ is a convex set. Regarding operations which preserve convexity, we have that

- if f and g are convex and $\lambda\geq 0$, then $f+\lambda g$ is convex;
- [28, Proposition 8.16]: if f_i is convex for every $i\in I$, then the *pointwise supremum* $x\mapsto\sup_{i\in I} f_i(x)$ is convex;
- [28, Proposition 12.11]: if f and g are convex, their *infimal convolution* $(f\,\square\,g)(x) := \inf_{y\in\mathcal{H}}\{f(y)+g(x-y)\}$ is convex.

For a subset $C\subseteq\mathcal{H}$, its *indicator* and *support* functions are given respectively by

$$\delta_C(x) := \begin{cases}0 & \text{if } x\in C\\ +\infty & \text{if } x\notin C\end{cases}\qquad\text{and}\qquad \sigma_C(x) := \sup_{y\in C}\langle x,y\rangle.$$

The support function σ_C is always convex, since it is the pointwise supremum of convex functions of the form $\langle \cdot, y \rangle$. If C is a convex set, then its indicator function δ_C is convex.

f is *lower semicontinuous* if for every $x \in \mathcal{H}$ it holds

$$\liminf_{y \to x} f(y) := \sup_{V \in \mathcal{V}(x)} \inf_{y \in V} f(y) \geq f(x),$$

where $\mathcal{V}(x)$ is any basis of neighborhoods of x for the strong (i.e., induced by $\|\cdot\|$) topology on $\mathcal{H}$. In particular, we may choose $\mathcal{V}(x) = \{B(x,\delta) : \delta > 0\}$, i.e., the set of open balls centered at x.

The *weak topology* on $\mathcal{H}$ is the weakest (or coarsest) topology of $\mathcal{H}$ which makes the mappings $\langle \cdot, x \rangle$ continuous for every $x \in \mathcal{H}$. For $x \in \mathcal{H}$, a basis of neighborhoods of x for the weak topology is given by $\mathcal{V}_w(x) = \{V_{x_1,\ldots,x_n,\varepsilon}(x) : n \in \mathbb{N}, x_i \in \mathcal{H} \text{ for } i = 1,\ldots,n, \varepsilon > 0\}$, where each $V_{x_1,\ldots,x_n,\varepsilon}(x)$ is given by

$$V_{x_1,\ldots,x_n,\varepsilon}(x) := \{y \in \mathcal{H} : |\langle x_i, y - x \rangle| < \varepsilon \text{ for } i = 1,\ldots,n\}.$$

It follows that a net $(x_a)_{a \in A}$ of elements of $\mathcal{H}$ converges weakly to x, which we denote by $x_a \rightharpoonup x$, if and only if $(\langle y, x_a \rangle)_{a \in A}$ converges to $\langle y, x \rangle$ for every $y \in \mathcal{H}$. Now that we have introduced the weak topology on $\mathcal{H}$, we mention Opial's Lemma (cf. [82]), a cornerstone result which allows us to establish the weak convergence of curves and sequences in $\mathcal{H}$.

Lemma 2.1.1 (Opial's Lemma)**.** *Let S be a nonempty subset of $\mathcal{H}$ and let $(x_k)_{k \geq 0}$ be a sequence in $\mathcal{H}$. Assume that*

(i) *For every $x^* \in S$, $\lim_{k \to +\infty} \|x_k - x^*\|$ exists;*

(ii) *Every weak sequential cluster point of $(x_k)_{k \geq 0}$ belongs to S.*

Then, x_k converges weakly to an element of S as $k \to +\infty$.

The previous lemma has a continuous-time version as well.

Lemma 2.1.2 (Opial's Lemma, continuous-time version)**.** *Let S be a nonempty subset of $\mathcal{H}$ and let $x : [t_0, +\infty) \to \mathcal{H}$ be a map. Assume that*

(i) *for every $x^* \in S$, $\lim_{t \to +\infty} \|x(t) - x^*\|$ exists;*

(ii) *every weak sequential limit point of $t \mapsto x(t)$ belongs to S.*

Then, $x(t)$ converges weakly to an element of S as $t \to +\infty$.

Going back to functions, we say that f is *weakly lower semicontinuous* if for every $x \in \mathcal{H}$ it holds

$$\liminf_{y \rightharpoonup x} f(y) := \sup_{V \in \mathcal{V}_w(x)} \inf_{y \in V} f(y) \geq f(x).$$

In terms of nets, f is lower semicontinuous (resp. weakly lower semicontinuous) if and only if for every $x \in \mathcal{H}$ and every net $(x_a)_{a \in A}$ such that $x_a \to x$ (resp. $x_a \rightharpoonup x$) it holds

$$f(x) \leq \liminf f(x_a).$$

An important result (cf. [28, Theorem 9.1]) tells us that if f is convex, then f is lower semicontinuous if and only if it is weakly lower semicontinuous. Some operations which preserve lower semicontinuity are the following:

- if f and g are lower semicontinuous and $\lambda \geq 0$, then $f + \lambda g$ is lower semicontinuous;
- [28, Lemma 1.26]: if f_i is lower semicontinuous for every $i \in I$, then the pointwise supremum $\sup_{i \in I} f_i$ is lower semicontinuous.

If f and g are lower semicontinuous, it need not happen that $f \,\square\, g$ is lower semicontinuous (cf. [28, Example 12.13]).

The *(Fenchel) conjugate* of f is the function $f^* : \mathcal{H} \to \overline{\mathbb{R}}$ given by

$$f^*(x) := \sup_{y \in \mathcal{H}} \{\langle x, y \rangle - f(y)\}.$$

Since f^* is the pointwise supremum of the convex and continuous functions $\langle \cdot, y\rangle - f(y)$, we have that f^* is convex and lower semicontinuous. The *biconjugate* of f is the function $f^{**} := (f^*)^*$. For our work, the most relevant properties of the conjugate function are the following:

- [28, Example 13.3]: for a subset $C \subseteq \mathcal{H}$, it holds $\delta_C^* = \sigma_C$;
- [28, Proposition 13.23]: if $\lambda > 0$, then $(\lambda f)^*(\cdot) = \lambda f^*\left(\frac{\cdot}{\lambda}\right)$;
- [28, Examples 13.2 and 13.8]: if $\rho \in (1, +\infty)$ and $\rho^* = \frac{\rho}{\rho-1}$ is its dual exponent, then $\left(\frac{\|\cdot\|^\rho}{\rho}\right)^* = \frac{\|\cdot\|^{\rho^*}}{\rho^*}$;
- [28, Proposition 13.24]: it holds $(f \,\square\, g)^* = f^* + g^*$.

We say that f is *(Fréchet) differentiable* at $x \in \operatorname{int}(\operatorname{dom} f)$ if there exists a vector $\nabla f(x) \in \mathcal{H}$ such that

$$\lim_{y \to 0} \frac{f(x+y) - f(x) - \langle \nabla f(x), y\rangle}{\|y\|} = 0.$$

If $f : \mathcal{H} \to \mathbb{R}$ and the previous statement holds for every $x \in \mathcal{H}$, we say that f is differentiable. A fundamental result is the characterization of convexity for differentiable functions. If $f : \mathcal{H} \to \mathbb{R}$ is differentiable, then the following are equivalent (cf. [28, Proposition 17.7]):

- f is convex;
- (gradient inequality): $f(y) \geq \langle y - x, \nabla f(x)\rangle + f(x)$ for every $x, y \in \mathcal{H}$;
- (monotonicity of the gradient): $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \geq 0$ for every $x, y \in \mathcal{H}$.

When ∇f is L-Lipschitz continuous, there exists a stronger version of the gradient inequality. In fact, if $f : \mathcal{H} \to \mathbb{R}$ is convex and differentiable and $L > 0$, the following are equivalent (cf. [104, Lemma 4]):

- $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for every $x, y \in \mathcal{H}$, i.e., ∇f is L-Lipschitz continuous;
- $f(y) \leq \langle y - x, \nabla f(x)\rangle + f(x) + \frac{L}{2}\|x - y\|$ for every $x, y \in \mathcal{H}$;
- $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \leq L\|x - y\|^2$ for every $x, y \in \mathcal{H}$;
- $f(y) \geq \langle y - x, \nabla f(x)\rangle + f(x) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2$ for every $x, y \in \mathcal{H}$;
- $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$ for every $x, y \in \mathcal{H}$.

An more general definition of differentiability which captures a larger class of functions is that of the (convex) subdifferential. For $f : \mathcal{H} \to \overline{\mathbb{R}}$ and $x \in \mathcal{H}$ such that $f(x) \in \mathbb{R}$, the *(convex) subdifferential* of f at x is given by

$$\partial f(x) := \{u \in \mathcal{H} : f(y) \geq \langle y - x, u\rangle + f(x) \ \forall y \in \mathcal{H}\}.$$

Defining $\partial f(x) := \emptyset$ when $f(x) \notin \mathbb{R}$ gives rise to a (possibly) multivalued operator $\partial f : \mathcal{H} \to 2^{\mathcal{H}}$. Fermat's rule (cf. [28, Theorem 16.3]) states that if f is proper, then

$$x \in \operatorname{argmin} f \quad \Leftrightarrow \quad 0 \in \partial f(x).$$

Naturally, when $f : \mathcal{H} \to \mathbb{R}$ is convex and differentiable, then its subgradient is single-valued and coincides with its gradient, i.e., $\partial f(x) = \{\nabla f(x)\}$ for all $x \in \mathcal{H}$ (cf. [28, Proposition 17.31]). The Fenchel-Young inequality (cf. [28, Proposition 13.15]) tells us that for a proper convex function $f : \mathcal{H} \to \overline{\mathbb{R}}$, for every $x, u \in \mathcal{H}$ it holds

$$f(x) + f^*(u) \geq \langle x, u\rangle.$$

Furthermore, equality is achieved if and only if $u \in \partial f(x)$ (cf. [28, Proposition 16.10]).

The *Moreau envelope* of index $\lambda > 0$ of a function $f : \mathcal{H} \to \overline{\mathbb{R}}$ is given by

$$f_\lambda := f \,\square \left(\frac{1}{2\lambda}\|\cdot\|^2\right).$$

If f is proper, convex and lower semicontinuous, f_λ is convex, real-valued, continuous, and *exact*, meaning that the infimum is attained. Furthermore, this infimum is attained uniquely (cf. [28, Proposition

12.15]). This gives rise to another fundamental object in convex analysis: if f is proper, convex and lower semicontinuous, its *proximal point operator* is given by

$$\operatorname{prox}_{\lambda f}(x) := \operatorname{argmin}_{y\in\mathcal{H}} \left\{ f(y) + \frac{1}{2\lambda}\|x-y\|^2 \right\}.$$

Thus, for a proper, convex and lower semicontinuous function f and $\lambda > 0$, for every $x \in \mathcal{H}$ it holds

$$f_\lambda(x) = f(\operatorname{prox}_{\lambda f}(x)) + \frac{1}{2\lambda}\|x - \operatorname{prox}_{\lambda f}(x)\|^2.$$

Moreover, f_λ is differentiable and its gradient is given by

$$\nabla f_\lambda = \frac{1}{\lambda}(\operatorname{Id} - \operatorname{prox}_{\lambda f}).$$

We have the following generalization of the gradient inequality for a convex function which is the sum of a smooth and a nonsmooth part. For a proof for this result, see e.g. [94].

Lemma 2.1.3. *Let $\hat{\Phi} : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper, convex and lower semicontinuous function and $\Phi : \mathcal{H} \to \mathbb{R}$ be an L-smooth function. For $s > 0$ and $y \in \mathcal{H}$, define*

$$G(y) := \frac{1}{s}\Big[y - \operatorname{prox}_{s\hat{\Phi}}(y - s\nabla\Phi(y))\Big]. \tag{2.1}$$

Set $\Psi := \Phi + \hat{\Phi}$. Then, for every $x, y \in \mathcal{H}$ it holds

$$\Psi(y - sG(y)) \le \Psi(x) - \frac{s}{2}(2 - sL)\|G(y)\|^2 + \langle G(y), y - x\rangle - \frac{1}{2L}\|\nabla\Phi(x) - \nabla\Phi(y)\|^2. \tag{2.2}$$

2.2 Monotone operators

An *operator* on $\mathcal{H}$ is any map $A : \mathcal{H} \to 2^{\mathcal{H}}$, i.e., it takes values in the power set of $\mathcal{H}$. Any operator A is characterized by its *graph*

$$\operatorname{gra} A := \{(x, u) \in \mathcal{H} \times \mathcal{H} : y \in A(x)\}.$$

If C is a subset of $\mathcal{H}$, we denote $A(C) := \cup_{x\in C} A(x)$. Consequently, the *range* of A is $A(\mathcal{H})$. The *domain* of A is the set

$$\operatorname{dom} A := \{x \in \mathcal{H} : A(x) \neq \emptyset\}.$$

The *inverse* of A, denoted by A^{-1}, is given via its graph $\operatorname{gra} A^{-1} = \{(u, x) \in \mathcal{H} \times \mathcal{H} : (x, u) \in \operatorname{gra} A\}$. Thus, $u \in A(x)$ if and only if $x \in A^{-1}(u)$. The *set of zeros* of A is

$$\operatorname{zer} A := \{x \in \mathcal{H} : 0 \in A(x)\}.$$

If for every $x \in \operatorname{dom} A$ $A(x)$ is a singleton, we may write $A : \operatorname{dom} A \to \mathcal{H}$. When we write $A : \mathcal{H} \to \mathcal{H}$, we mean that A has full domain (i.e., $\operatorname{dom} A = \mathcal{H}$) and that it is single-valued. Let us reserve the letter $T : \mathcal{H} \to \mathcal{H}$ for a single-valued operator for now. We say T is *nonexpansive* if it is 1-Lipschitz continuous, i.e., $\|T(x) - T(y)\| \le \|x - y\|$ for every $x, y \in \mathcal{H}$. We say that T is *firmly nonexpansive* if for every $x, y \in \mathcal{H}$ it holds

$$\|T(x) - T(y)\|^2 + \|(\operatorname{Id} - T)(x) - (\operatorname{Id} - T)(y)\|^2 \le \|x - y\|^2.$$

An important result (cf. [28, Proposition 4.4]) is that the following are equivalent:

- T is firmly nonexpansive;
- $\operatorname{Id} - T$ is firmly nonexpansive;
- $2\operatorname{Id} - T$ is nonexpansive.

If $\rho > 0$, we say that T is *ρ-cocoercive* if for every $x, y \in \mathcal{H}$ it holds

$$\rho\|T(x) - T(y)\|^2 \le \langle x - y, T(x) - T(y)\rangle.$$

In particular, if T is ρ-cocoercive then it is automatically $1/\rho$-Lipschitz continuous. Another equivalence we have is that T is nonexpansive if and only if $\operatorname{Id} - T$ is $(1/2)$-cocoercive (cf. [28, Proposition 4.11]). If $\alpha \in (0,1)$, we say that T is *α-averaged* if there exists a nonexpansive operator $R : \mathcal{H} \to \mathcal{H}$ such that

$$T = (1-\alpha)\operatorname{Id} + \alpha R.$$

Related to this definition, we have that (cf. [28, Remark 4.34])

- if T is averaged, then it is nonexpansive;
- T is firmly nonexpansive if and only if it is $(1/2)$-averaged;
- T is ρ-cocoercive if and only if ρT is $(1/2)$-averaged.

Just as convex functions form a privileged class, we define the corresponding analog for operators. We say that $A : \mathcal{H} \to 2^{\mathcal{H}}$ is *monotone* if for every $(x,u),(y,v) \in \operatorname{gra} A$ it holds

$$\langle x-y, u-v\rangle \geq 0.$$

We say that A is *maximally monotone* if A is monotone and its graph is not properly contained in the graph of another monotone operator, i.e., every time that $B : \mathcal{H} \to 2^{\mathcal{H}}$ is monotone and $\operatorname{gra} A \subseteq \operatorname{gra} B$, it holds $A = B$. Some facts regarding maximal monotonicity are the following:

- [28, Proposition 20.22]: if A is maximally monotone and $\lambda \geq 0$, then A^{-1} and λA are maximally monotone;
- [28, Corollary 20.28]: if $T : \mathcal{H} \to \mathcal{H}$ is continuous and monotone, it is automatically maximally monotone. In particular, if T is cocoercive then it is maximally monotone;
- [28, Theorem 20.25] (Moreau's theorem): if $f : \mathcal{H} \to \overline{\mathbb{R}}$ is proper, convex and lower semicontinuous, then ∂f is maximally monotone.

A fundamental result (cf. [28, Proposition 20.37]) tells us that the graph of a maximally monotone operator is closed in $\mathcal{H}^{\text{weak}} \times \mathcal{H}^{\text{strong}}$: namely, if A is maximally monotone and $(x_b, u_b)_{b\in B}$ is a net in $\operatorname{gra} A$ such that $x_b \rightharpoonup x$ and $u_b \to u$, then $(x,u) \in \operatorname{gra} A$.

Given an operator $A : \mathcal{H} \to 2^{\mathcal{H}}$ and $\lambda > 0$, its *resolvent* and *Moreau-Yosida approximation* of index λ are given respectively by

$$J_{\lambda A} := (\operatorname{Id} + \lambda A)^{-1} \qquad \text{and} \qquad A_\lambda := \frac{1}{\lambda}(\operatorname{Id} - J_{\lambda A}).$$

If A is maximally monotone and $\lambda > 0$, the following hold (cf. [28, Corollary 23.11]):

- $J_{\lambda A}$ is single valued, and both $J_{\lambda A}$ and $\operatorname{Id} - J_{\lambda A}$ are firmly nonexpansive and maximally monotone;
- A_λ is λ-cocoercive, thus it is $(1/\lambda)$-Lipschitz continuous and maximally monotone.

Let $f : \mathcal{H} \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous and $\lambda > 0$. Then (cf. [28, Proposition 16.44])

$$p = \operatorname{prox}_{\lambda f}(x) \quad \Leftrightarrow \quad x - p \in \lambda \partial f(p).$$

In other words,

$$\operatorname{prox}_{\lambda f} = (\operatorname{Id} + \lambda\partial f)^{-1} = J_{\lambda\partial f}, \quad \text{and thus} \quad \nabla f_\lambda = \frac{1}{\lambda}(\operatorname{Id} - \operatorname{prox}_{\lambda f}) = \frac{1}{\lambda}(\operatorname{Id} - J_{\lambda\partial f}) = (\partial f)_\lambda.$$

It follows that the proximal operator $\operatorname{prox}_{\lambda f}$ is firmly nonexpansive and maximally monotone.

To close off this section, we mention a crucial result (cf. [28, Proposition 23.38]): if $A : \mathcal{H} \to 2^{\mathcal{H}}$ is monotone and $\lambda > 0$, then the set of fixed points of $J_{\lambda A}$ coincides with the set of zeros of A, which in turn coincides with the set of zeros of A_λ. Indeed,

$$0 \in A(x) \;\Leftrightarrow\; x \in x + \lambda A(x) \;\Leftrightarrow\; x = (\operatorname{Id} + \lambda A)^{-1}(x) \;\Leftrightarrow\; x = J_{\lambda A}(x) \;\Leftrightarrow\; 0 = \frac{1}{\lambda}(x - J_{\lambda A}(x)) = A_\lambda(x).$$

2.3 Auxiliary results

In this section, we collect some auxiliary results which, while not especially related with convex analysis or monotone operators, are nonetheless vital and required numerous times throughout this work. Every time we have a curve $x : [t_0, +\infty) \to \mathcal{H}$, its derivative at a point $t > t_0$ is defined as

$$\dot{x}(t) := \lim_{s \to 0} \frac{x(t+s) - x(t)}{s},$$

whenever said limit exists. The right-sided derivative at t_0 is defined similarly. In some cases, we will need to consider the integral $\int_{t_1}^{t_2} x(t)dt$, which we will always do in the *Bochner* sense, for which the expected properties of integration hold (e.g. linearity, triangular inequality).

The following two results can be found as [1, Lemma 5.1 and Lemma 5.2] respectively.

Lemma 2.3.1. *Let $t_0 > 0$. Suppose that $F\colon [\delta, +\infty) \to \mathbb{R}$ is locally absolutely continuous, bounded from below, and there exists $G \in \mathbb{L}^1([t_0, +\infty); \mathbb{R})$ such that for almost every $t \geq t_0$*

$$\frac{d}{dt} F(t) \leq G(t).$$

Then, limit $\lim_{t \to +\infty} F(t) \in \mathbb{R}$ exists.

Lemma 2.3.2. *Let $t_0 > 0$, $1 \leq p < +\infty$ and $1 \leq q \leq +\infty$. Suppose that $F \in \mathbb{L}^p([t_0, +\infty))$ is a locally absolutely continuous nonnegative function, $G \in \mathbb{L}^q([t_0, +\infty))$ and for almost every $t \geq t_0$*

$$\frac{d}{dt} F(t) \leq G(t)$$

Then, $\lim_{t \to +\infty} F(t) = 0$.

The following lemma appears as Lemma A.3 in [42] and generalizes Lemma A.2 from [26].

Lemma 2.3.3. *Let $a > 0$, $r \in [0, 1]$ and $\varphi\colon [t_0, +\infty) \to \mathbb{R}$ be a continuously differentiable function such that*

$$\lim_{t \to +\infty} \left(\varphi(t) + \frac{t^r}{a} \dot{\varphi}(t) \right) = L \in \mathbb{R}.$$

Then, $\lim_{t \to +\infty} \varphi(t) = L$.

The previous lemma has a discrete-time (and slightly more general) version as well. A special case (i.e., when $t_k = k$) can be found in the literature, for example, in [40, Lemma A.5]. However, that version requires the sequence $(\varphi_k)_{k \geq 0}$ to be bounded, and ours does not.

Lemma 2.3.4. *Let $(\varphi_k)_{k \geq 0}$ and $(t_k)_{k \geq 0}$ be sequences of real and positive numbers respectively, and such that $\sum_{k=0}^{+\infty} \frac{1}{t_k} = +\infty$. Suppose that*

$$\lim_{k \to +\infty} \left(\varphi_{k+1} + t_k(\varphi_{k+1} - \varphi_k) \right) = L \in \mathbb{R}. \tag{2.3}$$

Then, $\lim_{k \to +\infty} \varphi_k = L$.

Proof. For every $k \geq 0$, denote $\ell_k := \varphi_{k+1} + t_k(\varphi_{k+1} - \varphi_k)$ and suppose that $\ell_k \to L$ as $k \to +\infty$. Consider the positive sequence $(\mu_k)_{k \geq 0}$ defined recursively, for every $k \geq 0$, as

$$\mu_0 := \frac{1}{t_0} \quad \text{and} \quad \mu_{k+1} := \frac{\mu_k(1 + t_k)}{t_{k+1}}. \tag{2.4}$$

We denote by $P_k := \mu_k t_k$ for all $k \geq 0$. Then, $P_0 = 1$ and for all $k \geq 0$

$$P_k > 0, \quad \sum_{i=0}^{k} \mu_i = P_{k+1} - 1 \quad \text{and} \quad P_{k+1} = \prod_{j=0}^{k} \left(\frac{1}{t_j} + 1 \right) \geq 1 + \sum_{j=0}^{k} \frac{1}{t_j}, \quad \text{thus} \quad P_k \to +\infty, \tag{2.5}$$

where the first three statements follow from (2.4), and the fourth statement from the summability assumption on $(t_k)_{k\geq 0}$. Multiplying ℓ_k by μ_k yields

$$\begin{aligned}\mu_{k+1}t_{k+1}\varphi_{k+1} - \mu_k t_k \varphi_k &= (\mu_{k+1}t_{k+1} - \mu_k t_k)\varphi_{k+1} + \mu_k t_k(\varphi_{k+1} - \varphi_k) \\ &= \mu_k \varphi_{k+1} + \mu_k t_k (\varphi_{k+1} - \varphi_k) = \mu_k \ell_k.\end{aligned} \tag{2.6}$$

Therefore, by summing (2.6) from 0 to k, we obtain for all $k \geq 0$

$$\varphi_{k+1} = \frac{\varphi_0}{P_{k+1}} + \frac{1}{P_{k+1}} \sum_{i=0}^{k} \mu_i \ell_i = \frac{\varphi_0}{P_{k+1}} + \frac{\sum_{i=0}^{k} \mu_i \ell_i}{\sum_{i=0}^{k} \mu_i + 1}. \tag{2.7}$$

In light of (2.5), applying the Stolz-Cesáro Theorem leads to $\varphi_k \to L$ as $k \to +\infty$. $\square$

The proof of Lemma 2.3.5 is straightforward.

Lemma 2.3.5. *Let $A, B, C \in \mathbb{R}$ be such that $A \neq 0$ and $B^2 - AC \leq 0$. The following statements are true:*

(i) if $A > 0$, then for every $x, y \in \mathcal{H}$ it holds

$$A\|x\|^2 + 2B\langle x, y\rangle + C\|y\|^2 \geq 0;$$

(ii) if $A < 0$, then for every $x, y \in \mathcal{H}$ it holds

$$A\|x\|^2 + 2B\langle x, y\rangle + C\|y\|^2 \leq 0.$$

The proof for Lemma 2.3.6 can be found in [25].

Lemma 2.3.6. *Let $t_0 > 0$, and let $u : [t_0, +\infty) \to \mathbb{R}$ be a continuously differentiable function which is bounded from below. Given $\alpha > 1$, a nonnegative function $\theta : [t_0, +\infty) \to \mathbb{R}$ and a nonnegative function $k \in \mathbb{L}^1([t_0, +\infty))$, let us assume that*

$$t\ddot{u}(t) + \alpha\dot{u}(t) + \theta(t) \leq k(t)$$

for almost every $t \geq t_0$. Then, the positive part $[\dot{u}]_+$ of $\dot{u}$ belongs to $\mathbb{L}^1([t_0, +\infty))$ and $\lim_{t\to+\infty} u(t)$ exists. Moreover, we have $\int_{t_0}^{+\infty} \theta(t)dt < +\infty$.

A proof for the following lemma in the finite-dimensional case can be found in [64, Lemma 6]. The proof for the infinite-dimensional case is short and virtually identical, so we include it here for the sake of completeness.

Lemma 2.3.7. *Assume that $t_0 > 0$, $g : [t_0, +\infty) \to \mathcal{Y}$ is a continuous differentiable function, $a : [t_0, +\infty) \to [0, +\infty)$ is a continuous function, and $C \geq 0$. If for every $t \geq t_0$ it holds*

$$\left\| g(t) + \int_{t_0}^{t} a(s)g(s)ds \right\| \leq C, \tag{2.8}$$

then

$$\sup_{t\geq t_0} \|g(t)\| \leq 2C < +\infty.$$

Proof. Define, for every $t \geq t_0$,

$$G(t) := e^{\int_{t_0}^{t} a(s)ds} \int_{t_0}^{t} a(s)g(s)ds.$$

Fix $t \geq t_0$. The time derivative of G reads

$$\dot{G}(t) = a(t)e^{\int_{t_0}^{t} a(s)ds} \int_{t_0}^{t} a(s)g(s)ds + e^{\int_{t_0}^{t} a(s)ds} a(t)g(t) = a(t)e^{\int_{t_0}^{t} a(s)ds} \left[g(t) + \int_{t_0}^{t} a(s)g(s)ds \right],$$

so by using (2.8) and the previous equality we arrive at

$$\|\dot{G}(t)\| \leq Ca(t)e^{\int_{t_0}^t a(s)ds} = C\frac{d}{dt}\left(e^{\int_{t_0}^t a(s)ds}\right). \tag{2.9}$$

Since $G(t_0) = 0$, we have

$$G(t) = G(t) - G(t_0) = \int_{t_0}^t \dot{G}(s)ds,$$

so by employing (2.9) and the previous equality we obtain, for every $t \geq t_0$,

$$\begin{aligned} e^{\int_{t_0}^t a(s)ds}\left\|\int_{t_0}^t a(s)g(s)ds\right\| = \|G(t)\| \leq \int_{t_0}^t \|\dot{G}(s)\|ds \leq C\int_{t_0}^t \frac{d}{ds}\left(e^{\int_{t_0}^s a(\tau)d\tau}\right)ds \\ \leq C\left[e^{\int_{t_0}^t a(s)ds} - 1\right] \leq Ce^{\int_{t_0}^t a(s)ds}. \end{aligned}$$

Dividing both sides of the previous inequality by $e^{\int_{t_0}^t a(s)ds}$ gives us, for every $t \geq t_0$,

$$\left\|\int_{t_0}^t a(s)g(s)ds\right\| \leq C. \tag{2.10}$$

Now, by putting (2.8) and (2.10) together, for every $t \geq t_0$ we finally come to

$$\|g(t)\| \leq \left\|g(t) + \int_{t_0}^t a(s)g(s)ds\right\| + \left\|\int_{t_0}^t a(s)g(s)ds\right\| \leq 2C,$$

which leads to the desired statement. □

The proof for the following result can be found in [43, Lemma A.1].

Lemma 2.3.8. *Let $0 < t_0 \leq r \leq +\infty$ and $h : [t_0, +\infty) \to \mathbb{R}_+$ be a continuous function. For every $\alpha > 1$ it holds*

$$\int_{t_0}^r \frac{1}{t^\alpha}\left[\int_{t_0}^t s^{\alpha-1}h(s)\right]dt \leq \frac{1}{\alpha-1}\int_{t_0}^r h(t)dt.$$

If $r = +\infty$, then equality holds.

The following lemma is a slight variation of results already present in the literature. See, for example, [12, Lemma A.2].

Lemma 2.3.9. *Let $t_0 > 0$, $\alpha > 1$, and let $\phi : [t_0, +\infty) \to \mathbb{R}$ be a twice continuously differentiable function which is bounded from below. Furthermore, assume $w : [t_0, +\infty) \to \mathbb{R}_+$ to be a continuously differentiable function such that $t \mapsto tw(t)$ belongs to $\mathbb{L}^1([t_0, +\infty))$ and that for every $t \geq t_0$ it holds*

$$t\ddot{\phi}(t) + \alpha\dot{\phi}(t) + t^2\dot{w}(t) \leq 0.$$

Then, the positive part $[\dot{\phi}]_+$ of $\dot{\phi}$ belongs to $\mathbb{L}^1([t_0, +\infty))$ and the limit $\lim_{t\to+\infty}\phi(t)$ is a real number.

Proof. Fix $t \geq t_0$. Adding $(\alpha+1)tw(t)$ to both sides of the previous inequality and then multiplying it by $t^{\alpha-1}$ yields

$$\frac{d}{dt}\left(t^\alpha\dot{\phi}(t)\right) + \frac{d}{dt}\left(t^{\alpha+1}w(t)\right) \leq (\alpha+1)t^\alpha w(t).$$

Since the previous inequality holds for any $t \geq t_0$, we can integrate it from t_0 to $t \geq t_0$ to get

$$t^\alpha\dot{\phi}(t) - t_0^\alpha\dot{\phi}(t_0) + t^{\alpha+1}w(t) - t_0^{\alpha+1}w(t_0) \leq (\alpha+1)\int_{t_0}^t s^\alpha w(s)ds.$$

After dropping the nonnegative term $t^{\alpha+1}w(t)$ and dividing by t^α, for every $t \geq t_0$ we arrive at

$$\dot{\phi}(t) \leq \frac{\widetilde{C}}{t^\alpha} + \frac{\alpha+1}{t^\alpha}\int_{t_0}^t s^\alpha w(s)ds,$$

where
$$\widetilde{C} := t_0^{\alpha}\left|\dot{\phi}(t_0)\right| + t_0^{\alpha+1}w(t_0),$$
which further leads to, for every $t \geq t_0$,
$$[\dot{\phi}(t)]_+ \leq \frac{\widetilde{C}}{t^{\alpha}} + \frac{\alpha+1}{t^{\alpha}}\int_{t_0}^{t} s^{\alpha}w(s)ds.$$
Now, we integrate this inequality from t_0 to $r \geq t_0$ and we apply Lemma 2.3.8 with $h : [t_0, +\infty) \to \mathbb{R}_+$ given by $h(s) := sw(s)$ to obtain
$$\begin{aligned}\int_{t_0}^{r}[\dot{\phi}(t)]_+dt &\leq \widetilde{C}\int_{t_0}^{r}\frac{1}{t^{\alpha}}dt + (\alpha+1)\int_{t_0}^{r}\frac{1}{t^{\alpha}}\left[\int_{t_0}^{t} s^{\alpha-1}\cdot sw(s)ds\right]dt \\ &\leq \frac{\widetilde{C}}{1-\alpha}\left(\frac{1}{t_0^{\alpha-1}} - \frac{1}{r^{\alpha-1}}\right) + \frac{\alpha+1}{1-\alpha}\int_{t_0}^{r} tw(t)dt.\end{aligned}$$
By hypothesis, as $r \to +\infty$, the right hand side of the previous inequality is finite. In other words,
$$\int_{t_0}^{+\infty}[\dot{\phi}(t)]_+dt < +\infty.$$
The previous statement, together with the fact that we assumed that ϕ was bounded from below, allow us to deduce that the function $\psi : [t_0, +\infty) \to \mathbb{R}$ given by
$$\psi(t) := \phi(t) - \int_{t_0}^{t}[\dot{\phi}(s)]_+ds$$
is also bounded from below. An easy computation shows that $\dot{\psi}$ is nonpositive on $[t_0, +\infty)$, thus ψ is nonincreasing on $[t_0, +\infty)$. These facts imply that $\lim_{t\to+\infty}\psi(t)$ is a real number. Finally, we conclude that
$$\lim_{t\to+\infty}\phi(t) = \lim_{t\to+\infty}\psi(t) + \int_{t_0}^{+\infty}[\dot{\phi}(s)]_+ds \in \mathbb{R}.$$
□

The following result follows from simple integral estimates.

Lemma 2.3.10. *Let $r \in \mathbb{R}$. Then, for every $k_0 \geq 2$ and $k \geq k_0$, the following are true:*

- *If $r < 0$, then $\sum_{\ell=k_0}^{k}\frac{1}{\ell^r} \in \left[\frac{1}{1-r}\left(k^{1-r} - (k_0-1)^{1-r}\right), \frac{1}{1-r}\left((k+1)^{1-r} - k_0^{1-r}\right)\right]$;*
- *If $r = 1$, then $\sum_{\ell=k_0}^{k}\frac{1}{\ell} \in \left[\ln(k+1) - \ln(k_0), \ln(k) - \ln(k_0-1)\right]$;*
- *If $r \geq 0$ and $r \neq 1$, then $\sum_{\ell=k_0}^{k}\frac{1}{\ell^r} \in \left[\frac{1}{1-r}\left(\frac{1}{(k+1)^{r-1}} - \frac{1}{k_0^{r-1}}\right), \frac{1}{1-r}\left(\frac{1}{k^{r-1}} - \frac{1}{(k_0-1)^{r-1}}\right)\right]$;*
- *In particular, if $1-r > 0$, then $\sum_{k=k_0}^{\ell}\frac{1}{\ell^r} \leq \frac{1}{1-r}\left[(k+1)^{1-r} - (k_0-1)^{1-r}\right]$.*

Chapter 3

A link between Heavy Ball and Nesterov's accelerated gradient in continuous-time

3.1 Introduction

In this chapter, which is mostly based on our work [8], we explore the link which exists between the Heavy Ball dynamics and the accelerated gradient dynamics, both in the optimization and in the monotone operator settings, within a real Hilbert space $\mathcal{H}$.

The first question we address is as follows. Given a convex and differentiable function $f : \mathcal{H} \to \mathbb{R}$, attached to the unconstrained minimization problem

$$\begin{array}{rl} \min & f(x), \\ \text{subject to} & x \in \mathcal{H}, \end{array} \tag{3.1}$$

consider the following Heavy Ball dynamical system, for $t \geq t_0 \geq 0$:

$$\ddot{y}(t) + \lambda \dot{y}(t) + b(t)\nabla f(y(t)) = 0, \tag{3.2}$$

where $b : [t_0, +\infty) \to \mathbb{R}_{++}$ acts as a time rescaling parameter. We will show convergence rates for the functional values along the trajectories which depend on b, as well as weak convergence of the trajectories towards solutions to (3.1). One of the main points of this chapter is that for a particular choice for b, system (3.2) is equivalent to the Nesterov accelerated gradient dynamics, which for $s \geq s_0 > 0$ read

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \nabla f(x(s)) = 0. \tag{3.3}$$

To be more precise, for a certain choice for b, (3.2) and (3.3) are time reparametrizations of each other. The message we want to emphasize is therefore

> Nesterov accelerated gradient dynamics can be recovered from the Heavy Ball dynamics through a time rescaling process.

A similar statement can be said in the context of monotone operators. Let $M : \mathcal{H} \to \mathcal{H}$ be continuous and monotone. Attached to the monotone equation problem

$$\text{find } x \in \mathcal{H} \text{ such that } 0 = M(x), \tag{3.4}$$

consider the following Heavy Ball-like dynamical system, for $t \geq t_0 \geq 0$:

$$\ddot{y}(t) + \lambda\dot{y}(t) + \mu(t)\frac{d}{dt}M(y(t)) + \gamma(t)M(y(t)) = 0, \tag{3.5}$$

where $\mu, \gamma : [t_0, +\infty) \to \mathbb{R}_{++}$ are time rescaling parameters. Again, we show convergence rates of the norm of the operator along the trajectories which depend on μ, as well as weak convergence of the trajectories towards solutions to (3.4). The other main point of this chapter is that for a particular choice for μ and γ, system (3.5) is equivalent to the fast OGDA dynamics, which for $s \geq s_0 > 0$ read

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \frac{d}{ds}M(x(s)) + \frac{\alpha}{2s}M(x(s)) = 0. \tag{3.6}$$

To be more precise, for a certain choice for μ and γ, (3.5) and (3.6) are time reparametrizations of each other. Thus, the second message we wish to emphasize is that

> The Fast OGDA dynamics can be recovered from the Heavy Ball dynamics for monotone equations through a time rescaling process.

This chapter is organized as follows: in Section 3.2, we show the existence and uniqueness of solutions and the asymptotic properties (convergence rates, weak convergence) of (3.2); in Section 3.3, we recover (3.3) from (3.2); in Section 3.4, we show the existence and uniqueness of solutions and the asymptotic properties (convergence rates, weak convergence) of (3.5); in Section 3.5, we recover (3.6) from (3.5); in Section 3.6, we mention the main results of our paper [42], where also attached to (3.4) we analyze the following second-order system with slow vanishing damping, for $t \geq t_0 > 0$:

$$\ddot{z}(t) + \frac{\alpha}{t^r}\dot{z}(t) + \theta t^r \beta(t)\frac{d}{dt}M(z(t)) + \beta(t)M(z(t)) = 0,$$

where the slowness of the damping comes from the fact that we consider $r \in [0, 1]$. The form of the system was inspired by the fast OGDA dynamics [40]. The main difference lies in that the inclusion of the parameter r allows for a wider array of choices for the time rescaling parameter β.

3.2 Heavy Ball with friction dynamics from the time scaling perspective

For $\lambda > 0$ and $b : [t_0, +\infty) \to \mathbb{R}_{++}$ a differentiable and nondecreasing function, for $t \geq t_0 \geq 0$ we consider the following dynamical system:

$$\ddot{y}(t) + \lambda\dot{y}(t) + b(t)\nabla f(y(t)) = 0, \tag{HBF$_\lambda$}$$

with Cauchy data $y(t_0) := y_0$ and $\dot{y}(t_0) := y_1$, where $y_0, y_1 \in \mathcal{H}$. It will turn out that the friction parameter λ will play an important role when connecting (HBF$_\lambda$) to other dynamics.

We lay out this section as follows: in Subsection 3.2.1, we show the existence and uniqueness of solutions to (HBF$_\lambda$), and in Subsection 3.2.2 we discuss their asymptotic properties. We now introduce the energy function that will help in our analysis, and show an initial bound that will be needed later. Assume that $x^* \in \operatorname{argmin} f$, the set of minimizers of f (thus $f(x^*) = \min f$), and that $t \mapsto y(t)$ solves (HBF$_\lambda$) for $t \geq t_0$. For $0 \leq \eta \leq \lambda$ we define on $[t_0, +\infty)$

$$\mathcal{E}_\eta(t) := b(t)(f(y(t)) - \min f) + \frac{1}{2}\left\|\eta(y(t) - x^*) + \dot{y}(t)\right\|^2 + \frac{1}{2}\eta(\lambda - \eta)\|y(t) - x^*\|^2.$$

The time derivative of $\mathcal{E}_\eta$ at $t \geq t_0$ gives

$$\begin{aligned}\frac{d}{dt}\mathcal{E}_\eta(t) &= \dot{b}(t)(f(y(t)) - \min f) + b(t)\langle\nabla f(y(t)), \dot{y}(t)\rangle \\ &\quad + \left\langle \eta(y(t) - x^*) + \dot{y}(t), \eta\dot{y}(t) + \ddot{y}(t)\right\rangle + \eta(\lambda - \eta)\langle y(t) - x^*, \dot{y}(t)\rangle.\end{aligned} \tag{3.7}$$

According to the distributive property of the inner product and the equation (HBF$_\lambda$) we have

$$\begin{aligned}
&\Big\langle \eta(y(t)-x^*)+\dot y(t), \eta\dot y(t)+\ddot y(t)\Big\rangle \\
=& \Big\langle \eta(y(t)-x^*)+\dot y(t), \lambda\dot y(t)+\ddot y(t)\Big\rangle + (\eta-\lambda)\Big\langle \eta(y(t)-x^*)+\dot y(t), \dot y(t)\Big\rangle \\
=& -b(t)\Big\langle \eta(y(t)-x^*)+\dot y(t), \nabla f(y(t))\Big\rangle + \eta(\eta-\lambda)\big\langle y(t)-x^*, \dot y(t)\big\rangle + (\eta-\lambda)\big\|\dot y(t)\big\|^2.
\end{aligned}$$

By plugging this expression into (3.7), we deduce that for every $t \geq t_0$

$$\begin{aligned}
\frac{d}{dt}\mathcal{E}_\eta(t) &= \dot b(t)(f(y(t)) - \min f) - \eta b(t)\left\langle y(t)-x^*, \nabla f(y(t))\right\rangle + (\eta-\lambda)\big\|\dot y(t)\big\|^2 \\
&\leq (\dot b(t) - \eta b(t))(f(y(t)) - \min f) + (\eta-\lambda)\big\|\dot y(t)\big\|^2,
\end{aligned} \tag{3.8}$$

where the last inequality comes from the convexity of f.

3.2.1 Existence and uniqueness of global solutions

We will rewrite (HBF$_\lambda$) as a first order system in two variables. It is simple to check that

$$\begin{cases} \ddot y(t) + \lambda \dot y(t) + b(t)\nabla f(y(t)) = 0, \\ y(t_0) = y_0, \quad \dot y(t_0) = y_1 \end{cases} \quad \text{is equivalent to} \quad \begin{cases} \dot y(t) &= u(t) - \lambda y(t), \\ \dot u(t) &= -b(t)\nabla f(y(t)), \\ y(t_0) &= y_0, \\ u(t_0) &= \lambda y_0 + y_1, \end{cases} \tag{HBF$_\lambda$}$$

where $u(t) := \lambda y(t) + \dot y(t)$. We can write the system above in a more compact way, namely,

$$\begin{cases} (\dot y(t), \dot u(t)) &= G(t, y(t), u(t)), \\ (y(t_0), u(t_0)) &= \big(y_0,\ \lambda y_0 + y_1\big), \end{cases} \tag{3.9}$$

where $G : [t_0, +\infty) \times \mathcal{H} \times \mathcal{H} \to \mathcal{H} \times \mathcal{H}$ is given by

$$G(t, y, u) := \Big(u - \lambda y,\ -b(t)\nabla f(y)\Big).$$

To establish uniqueness and global existence of the solution trajectory, we require the following assumption:

Assumption 3.1. Suppose that $\lambda > 0$ and $b : [t_0, +\infty) \to \mathbb{R}_{++}$ additionally satisfy

$$\sup_{t \geq t_0} \frac{\dot b(t)}{b(t)} < \lambda. \tag{3.10}$$

Theorem 3.2.1. *Suppose further that ∇f is Lipschitz continuous on bounded sets, $b : [t_0, +\infty) \to \mathbb{R}_{++}$ is continuously differentiable, and that λ and $b(\cdot)$ satisfy Assumption 3.1. Then, the dynamical system (HBF$_\lambda$) admits a unique global solution $y : [t_0, +\infty) \to \mathcal{H}$.*

Proof. Since $b(\cdot)$ is continuously differentiable, it is Lipschitz continuous on the bounded subsets of $[t_0, +\infty)$. This, together with our other assumptions ensures that G is Lipschitz continuous on the bounded subsets of $[t_0, +\infty) \times \mathcal{H} \times \mathcal{H}$. According to [90, Theorems 46.2 and 46.3], the ordinary differential equation (3.9) admits a unique continuously differentiable solution $t \mapsto (y(t), u(t))$ defined on a maximal interval $[t_0, T_{\max})$, where

$$T_{\max} = +\infty \quad \text{or} \quad \Big\{T_{\max} < +\infty \quad \text{and} \quad \lim_{t \to T_{\max}} \|(y(t), u(t))\| = +\infty\Big\}.$$

According to Assumption 3.1, there exists $0 < \eta_0 < \lambda$ such that for every $t \geq t_0$ it holds

$$\frac{\dot b(t)}{b(t)} \leq \eta_0 < \lambda.$$

We now define the energy functional $\mathcal{E}_{\eta_0}(\cdot)$ restricted to the interval $[t_0, T_{\max})$. According to the previous inequality and (3.8), we have for every $t \in [t_0, T_{\max})$

$$\frac{d}{dt}\mathcal{E}_{\eta_0}(t) \leq \left(\sup_{t\geq t_0}\frac{\dot{b}(t)}{b(t)} - \eta_0\right) b(t)\Big(f(y(t)) - \min f\Big) + (\eta_0 - \lambda)\big\|\dot{y}(t)\big\|^2 \leq 0.$$

This means that $\mathcal{E}_{\eta_0}(\cdot)$ is nonincreasing on $[t_0, T_{\max})$. In particular, we obtain that for every $t \in [t_0, T_{\max})$

$$\frac{1}{2}\big\|\eta_0(y(t) - x^*) + \dot{y}(t)\big\|^2 + \frac{1}{2}\eta_0(\lambda - \eta_0)\|y(t) - x^*\|^2 \leq \mathcal{E}_{\eta_0}(t) \leq \mathcal{E}_{\eta_0}(t_0).$$

This immediately yields that $t \mapsto \|y(t) - x^*\|$ and thus $t \mapsto \|y(t)\|$ are bounded on $[t_0, T_{\max})$. Since we also obtain that $t \mapsto \|\eta_0(y(t) - x^*) + \dot{y}(t)\|$ is bounded on $[t_0, T_{\max})$, using the triangle inequality, we obtain that $t \mapsto \big\|\dot{y}(t)\big\|$ and thus $t \mapsto \|u(t)\|$ are bounded on $[t_0, T_{\max})$. So it must be $T_{\max} = +\infty$, which completes the proof of this theorem. □

3.2.2 Convergence rates for function values and weak convergence of trajectories

We recall that $f : \mathcal{H} \to \mathbb{R}$ is assumed to be convex and continuously differentiable function, $\lambda > 0$ and $b : [t_0, +\infty) \to \mathbb{R}_{++}$ is a differentiable and nondecreasing function. We will now assume that we have a global solution $y : [t_0, +\infty) \to \mathcal{H}$ to (HBF_λ).

Proposition 3.2.2. *Suppose further that λ and $b(\cdot)$ satisfy Assumption 3.1. Let $y : [t_0, +\infty) \to \mathcal{H}$ be a solution trajectory of (HBF_λ). Then, the following statements are true:*

(i) *It holds*

$$\int_{t_0}^{+\infty} b(t)\left(f(y(t)) - \min f\right) dt < +\infty \quad \textit{and} \quad \int_{t_0}^{+\infty} \|\dot{y}(t)\|^2 dt < +\infty;$$

(ii) *The limit $\lim_{t\to+\infty}\mathcal{E}_\eta(t) \in \mathbb{R}$ exists for every η satisfying $0 \leq \eta \leq \lambda$.*

Proof. (i) According to (3.1), there exists $\eta_0 > 0$ such that

$$\sup_{t\geq t_0}\frac{\dot{b}(t)}{b(t)} < \eta_0 < \lambda.$$

It follows from (3.8) that for every $t \geq t_0$

$$\frac{d}{dt}\mathcal{E}_{\eta_0}(t) \leq \left(\sup_{t\geq t_0}\frac{\dot{b}(t)}{b(t)} - \eta_0\right) b(t)\left(f(y(t)) - \min f\right) + (\eta_0 - \lambda)\|\dot{y}(t)\|^2. \tag{3.11}$$

The statement follows upon integration of this inequality.

(ii) Let $0 \leq \eta \leq \lambda$. From (3.8) we derive that for every $t \geq t_0$

$$\frac{d}{dt}\mathcal{E}_\eta(t) \leq (\lambda - \eta)b(t)\left(f(y(t)) - \min f\right).$$

The first statement in (i) ensures that the right hand side of this estimate belongs to $\mathbb{L}^1([t_0, +\infty); \mathbb{R})$. Hence, the conclusion follows from Lemma 2.3.1. □

For the purpose of studying the convergence and rate of convergence of (HBF_λ), we introduce the following function $W : [t_0, +\infty) \to \mathbb{R}_+$ defined for every $t \geq t_0$ as

$$W(t) := \left(f(y(t)) - \min f\right) + \frac{1}{2b(t)}\big\|\dot{y}(t)\big\|^2.$$

The function is nonincreasing, which plays a crucial role in establishing convergence rates (see [36]). Indeed, for every $t \geq t_0$ we have

$$\dot{W}(t) = \langle\nabla f(y(t)), \dot{y}(t)\rangle - \frac{\dot{b}(t)}{2b^2(t)}\big\|\dot{y}(t)\big\|^2 + \frac{1}{b(t)}\langle\dot{y}(t), \ddot{y}(t)\rangle = -\frac{1}{b(t)}\left(\frac{\dot{b}(t)}{2b(t)} + \lambda\right)\|\dot{y}(t)\|^2 \leq 0, \tag{3.12}$$

which holds due to the fact that $\dot{b}(t) \geq 0$ for every $t \geq t_0$. We are now ready for the main convergence results for (HBF$_\lambda$).

Theorem 3.2.3. *Suppose further that λ and $b(\cdot)$ satisfy Assumption 3.1. Let $y : [t_0, +\infty) \to \mathcal{H}$ be a solution trajectory of (HBF$_\lambda$). Then, the following statements are true:*

(i) *As $t \to +\infty$, it holds*

$$W(t) = o\left(\frac{1}{\int_{\frac{t+t_0}{2}}^{t} b(r)dr}\right).$$

In particular, as $t \to +\infty$ we have

$$f(y(t)) - \min f = o\left(\frac{1}{\int_{\frac{t+t_0}{2}}^{t} b(r)dr}\right) \quad \text{and} \quad \|\dot{y}(t)\| = o\left(\sqrt{\frac{b(t)}{\int_{\frac{t+t_0}{2}}^{t} b(r)dr}}\right);$$

(ii) *The solution trajectory $y(t)$ converges weakly to an element of* $\operatorname{argmin} f$ *as $t \to +\infty$.*

Proof. (i) From Proposition 3.2.2, we infer that $\int_{t_0}^{+\infty} b(t)W(t)dt < +\infty$. In particular, for $\psi : [t_0, +\infty) \to \mathbb{R}_+$ defined as $\psi(t) := \int_{t_0}^{t} b(r)W(r)dr$, the limit $\lim_{t\to+\infty} \psi(t) \in \mathbb{R}$ exists. Hence, for every $t \geq t_0$ and $r \in \left[\frac{t_0+t}{2}, t\right]$ we have, according to the decreasing property of $W(\cdot)$, that $W(t) \leq W(r)$ and thus

$$0 \leq W(t) \int_{\frac{t+t_0}{2}}^{t} b(r)dr \leq \int_{\frac{t+t_0}{2}}^{t} b(r)W(r)dr = \psi(t) - \psi\left(\frac{t+t_0}{2}\right) \to 0 \quad \text{as } t \to +\infty,$$

which gives the desired small o rate for $W(\cdot)$.

(ii) Let $0 < \eta_1 < \eta_2 < \lambda$ and $x^* \in \operatorname{argmin} f$. We have for every $t \geq t_0$

$$\begin{aligned}\mathcal{E}_{\eta_2}(t) - \mathcal{E}_{\eta_1}(t) &= \frac{1}{2}(\eta_2 - \eta_1)\lambda\|y(t) - x^*\|^2 + (\eta_2 - \eta_1)\langle y(t) - x^*, \dot{y}(t)\rangle \\ &= (\eta_2 - \eta_1)\left(\frac{1}{2}\lambda\|y(t) - x^*\|^2 + \frac{d}{dt}\left(\frac{1}{2}\|y(t) - x^*\|^2\right)\right).\end{aligned}$$

Proposition 3.2.2 (ii) guarantees that $\lim_{t\to+\infty}(\mathcal{E}_{\eta_2}(t) - \mathcal{E}_{\eta_1}(t)) \in \mathbb{R}$ exists. If we set $q(t) := \frac{1}{2}\|y(t) - x^*\|^2$, Lemma 2.3.3 ensures that $\lim_{t\to+\infty} \frac{1}{2}\|y(t) - x^*\|^2$ exists and is a real number. This shows that the first condition of Opial's Lemma (see Lemma 2.1.2) is verified.

For the second condition, let $\overline{y}$ be a weak sequential cluster point of $t \mapsto y(t)$, i.e., there exists a sequence $(t_n)_{n\in\mathbb{N}} \subseteq [t_0, +\infty)$ such that $t_n \to +\infty$ and $y(t_n) \rightharpoonup \overline{y}$ as $n \to +\infty$. Since we know that $f(y(t)) \to \min f$ as $t \to +\infty$ and thus $f(y(t_n)) \to \min f$ as $n \to +\infty$, and that f is weakly lower semicontinuous, we obtain that

$$f(\overline{y}) \leq \liminf_{n\to+\infty} f(y(t_n)) = \min f,$$

i.e., $\overline{y} \in \operatorname{argmin} f$. We have verified both conditions of Opial's Lemma and therefore concluded the proof for this theorem. $\square$

Let us illustrate the preceding results with two specific choices. For $\kappa > 0$ and $\rho \geq 0$, consider

$$b(t) := \kappa \exp(\rho t) \quad \text{and} \quad b(t) := \kappa t^\rho,$$

respectively. Note that in both cases the classical Heavy Ball Method with friction is recovered by setting $(\kappa, \rho) = (1, 0)$, yielding a convergence rate of $o(1/t)$ for the function value along the solution trajectory [10]. The $\mathcal{O}(1/k)$ convergence rate for the discrete counterpart can be found in [95]. For these choices of $b(\cdot)$ we respectively have, for $t \geq t_0$ sufficiently large,

$$\int_{\frac{t+t_0}{2}}^{t} b(r)dr = \frac{\kappa}{\rho}\left[\exp(\rho t) - \exp\left(\frac{\rho(t+t_0)}{2}\right)\right] = \frac{\kappa}{\rho}\exp(\rho t)\left[1 - \exp\left(\frac{\rho(t_0-t)}{2}\right)\right] \geq \frac{\kappa}{2\rho}\exp(\rho t),$$

$$\int_{\frac{t+t_0}{2}}^{t} b(r)dt = \frac{\kappa}{\rho+1}\left[t^{\rho+1} - \left(\frac{t+t_0}{2}\right)^{\rho+1}\right] = \frac{\kappa}{\rho+1}t^{\rho+1}\left[1 - \left(\frac{t+t_0}{2t}\right)^{\rho+1}\right] \geq \frac{\kappa}{4(\rho+1)}t^{\rho+1},$$

respectively. To comply with the growth condition (3.1) for these choices for $b(\cdot)$, we respectively ask that

$$\lambda > \rho > 0 \quad \text{and} \quad \lambda > \frac{\rho}{t_0} \geq 0.$$

The previous estimates lead to the following corollaries, respectively:

Corollary 3.2.4. *Let $y : [t_0, +\infty) \to \mathcal{H}$ be a solution trajectory of*

$$\ddot{y}(t) + \lambda\dot{y}(t) + \kappa \exp(\rho t) \nabla f(y(t)) = 0.$$

If $\lambda > \rho > 0$, then the following statements are true:

(i) *As $t \to +\infty$, it holds*

$$f(y(t)) - \min f = o\left(\frac{1}{\exp(\rho t)}\right) \quad \text{and} \quad \|\dot{y}(t)\| \to 0;$$

(ii) *The solution trajectory $y(t)$ converges weakly to an element of* $\operatorname{argmin} f$ *as $t \to +\infty$.*

Corollary 3.2.5. *Let $y : [t_0, +\infty) \to \mathcal{H}$ be a solution trajectory of*

$$\ddot{y}(t) + \lambda\dot{y}(t) + \kappa t^{\rho} \nabla f(y(t)) = 0 \quad \text{for } t \geq t_0 > 0.$$

If $\lambda > \frac{\rho}{t_0} \geq 0$, then the following statements are true:

(i) *As $t \to +\infty$, it holds*

$$f(y(t)) - \min f = o\left(\frac{1}{t^{\rho+1}}\right) \quad \text{and} \quad \|\dot{y}(t)\| = o\left(\frac{1}{\sqrt{t}}\right);$$

(ii) *The solution trajectory $y(t)$ converges weakly to an element of* $\operatorname{argmin} f$ *as $t \to +\infty$.*

Remark 3.2.6. Previous statements show that one can get faster rates from the Heavy Ball Method (even linearly, as in Corollary 3.2.4). Nevertheless, we emphasize that these dynamics are not only interesting in terms of their acceleration phenomenon but also because they have a connection with other dynamics in the literature, for a nonclassical choice of $b(\cdot)$. This was one of the main points made in the introduction, which we elaborate upon in the next section. Naturally, this framework is only truly meaningful in the continuous-time setting, since one cannot expect gradient-type algorithms with unbounded step sizes—arising as explicit discretizations of time-rescaled systems—to converge. In contrast, implicit discretizations, which give rise to proximal-type methods, inherit the convergence properties of the continuous-time systems.

3.3 Connection with the dynamical system with asymptotic vanishing damping

For $\alpha > 3$ and $s \geq s_0 > 0$, we study the convergence behavior of the second order dynamical system with asymptotic vanishing damping

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \nabla f(x(s)) = 0. \tag{AVD$_\alpha$}$$

In particular, we will show that (AVD$_\alpha$) can be derived from (HBF$_\lambda$) via a time rescaling argument. This connection allows us to transfer the convergence results established for (HBF$_\lambda$) to (AVD$_\alpha$).

3.3.1 Two equivalent dynamical systems through time rescaling

We start with a solution trajectory $y : [t_0, +\infty) \to \mathcal{H}$ of

$$\ddot{y}(t) + \lambda \dot{y}(t) + b(t)\nabla f(y(t)) = 0, \tag{3.13}$$

and define $x(s) := y(\tau(s))$, where $\tau : [s_0, +\infty) \to [t_0, +\infty)$ is a continuously differentiable function such that $\dot{\tau}(s) > 0$ for every $s \geq s_0 > 0$ and $\lim_{s\to+\infty} \tau(s) = +\infty$. We have

$$\dot{x}(s) = \dot{\tau}(s)\dot{y}(\tau(s)) \quad \text{and} \quad \ddot{x}(s) = \ddot{\tau}(s)\dot{y}(\tau(s)) + \big(\dot{\tau}(s)\big)^2 \ddot{y}(\tau(s)).$$

These expressions lead to

$$\dot{y}(\tau(s)) = \frac{1}{\dot{\tau}(s)}\dot{x}(s) \quad \text{and} \quad \ddot{y}(\tau(s)) = \frac{1}{\big(\dot{\tau}(s)\big)^2}\Big[\ddot{x}(s) - \ddot{\tau}(s)\dot{y}(\tau(s))\Big] = \frac{1}{\big(\dot{\tau}(s)\big)^2}\Big[\ddot{x}(s) - \frac{\ddot{\tau}(s)}{\dot{\tau}(s)}\dot{x}(s)\Big].$$

Now, plugging $t = \tau(s)$ in (3.13) for $s \geq s_0$ gives

$$\frac{1}{\big(\dot{\tau}(s)\big)^2}\left[\ddot{x}(s) - \frac{\ddot{\tau}(s)}{\dot{\tau}(s)}\dot{x}(s)\right] + \frac{\lambda}{\dot{\tau}(s)}\dot{x}(s) + b(\tau(s))\nabla f(x(s)) = 0$$

or, equivalently,

$$\ddot{x}(s) + \left[\lambda\dot{\tau}(s) - \frac{\ddot{\tau}(s)}{\dot{\tau}(s)}\right]\dot{x}(s) + \big(\dot{\tau}(s)\big)^2 b(\tau(s))\nabla f(x(s)) = 0. \tag{3.14}$$

Recall that the Su-Boyd-Candès dynamics [94] are given by

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \nabla f(x(s)) = 0.$$

Going back to (3.14), to match the asymptotic vanishing viscosity accompanying the velocity we need the following to hold

$$\begin{cases} \lambda\dot{\tau}(s) - \dfrac{\ddot{\tau}(s)}{\dot{\tau}(s)} & = \dfrac{\alpha}{s}, \\ \tau(s_0) & = t_0. \end{cases}$$

It is straightforward to check that

$$\tau(s) := \frac{\alpha - 1}{\lambda}\ln\left(\frac{s}{s_0}\right) + t_0$$

satisfies the previous differential equation, since

$$\dot{\tau}(s) = \frac{\alpha - 1}{\lambda s} \quad \text{and} \quad \ddot{\tau}(s) = -\frac{\alpha - 1}{\lambda s^2}.$$

We need to assume that $\alpha > 1$. Furthermore, we want the coefficient attached to $\nabla f(x(s))$ to be 1, i.e.,

$$\frac{(\alpha-1)^2}{\lambda^2 s^2} b(\tau(s)) = \big(\dot{\tau}(s)\big)^2 b(\tau(s)) = 1 \quad \Leftrightarrow \quad b\left(\frac{\alpha - 1}{\lambda}\ln\left(\frac{s}{s_0}\right) + t_0\right) = \frac{\lambda^2 s^2}{(\alpha-1)^2},$$

which is fulfilled if we choose

$$b(t) = \left(\frac{\lambda s_0}{\alpha - 1}\right)^2 \exp\left(\frac{2\lambda(t - t_0)}{\alpha - 1}\right).$$

Furthermore, we need $b(\cdot)$ to satisfy (3.1). Indeed, we have

$$\frac{\dot{b}(t)}{b(t)} = \frac{\left(\frac{2\lambda}{\alpha-1}\right)\exp\left(\frac{2\lambda(t-t_0)}{\alpha-1}\right)}{\exp\left(\frac{2\lambda(t-t_0)}{\alpha-1}\right)} = \frac{2\lambda}{\alpha - 1} < \lambda \quad \Leftrightarrow \quad \alpha > 3.$$

All in all, with these choices, the trajectory $x : [s_0, +\infty) \to \mathcal{H}$ fulfills

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \nabla f(x(s)) = 0.$$

Conversely, if for $\alpha > 3$, $x : [s_0, +\infty) \to \mathcal{H}$ is a solution to the previous system and we define $y(t) = x(\sigma(t))$, where $\sigma : [t_0, +\infty) \to [s_0, +\infty)$ is a continuously differentiable function such that $\dot{\sigma}(t) > 0$ for $t \geq t_0 > 0$ and $\lim_{t\to+\infty} \sigma(t) = +\infty$, arguing in a similar fashion as it was done previously we arrive at

$$\ddot{y}(t) + \left[\alpha \frac{\dot{\sigma}}{\sigma(t)} - \frac{\ddot{\sigma}(t)}{\dot{\sigma}(t)}\right] \dot{y}(t) + \big(\dot{\sigma}(t)\big)^2 \nabla f(y(t)) = 0.$$

We want the coefficient attached to $\dot{y}(t)$ to be λ, i.e., we want $\sigma(\cdot)$ to satisfy the differential equation

$$\begin{cases} \alpha \dfrac{\dot{\sigma}(t)}{\sigma(t)} - \dfrac{\ddot{\sigma}(t)}{\dot{\sigma}(t)} & = \lambda, \\ \sigma(t_0) & = s_0, \end{cases} \quad \text{which is fulfilled by} \quad \sigma(t) := s_0 \exp\left(\frac{\lambda(t - t_0)}{\alpha - 1}\right).$$

With this choice for σ, the resulting system reads, for every $t \geq t_0$,

$$\ddot{y}(t) + \lambda \dot{y}(t) + s_0^2 \exp\left(\frac{2\lambda(t - t_0)}{\alpha - 1}\right) \nabla f(y(t)) = 0.$$

We have thus showed the following statement.

Proposition 3.3.1. *Assume that $\alpha > 3$, $\lambda > 0$ and that $s_0 > 0, t_0 \geq 0$ are initial times. Consider the following second-order systems:*

$$\begin{cases} \ddot{y}(t) + \lambda \dot{y}(t) + s_0^2 \exp\left(\dfrac{2\lambda(t - t_0)}{\alpha - 1}\right) \nabla f(y(t)) = 0, \\ y(t_0) = y_0, \quad \dot{y}(t_0) = y_1, \end{cases} \tag{3.15}$$

and

$$\begin{cases} \ddot{x}(s) + \dfrac{\alpha}{s} \dot{x}(s) + \nabla f(x(s)) = 0, \\ x(s_0) = x_0, \quad \dot{x}(s_0) = x_1. \end{cases} \tag{3.16}$$

Then, the following statements are true:

(i) *If $y : [t_0, +\infty) \to \mathcal{H}$ is a solution trajectory of (3.15) and the function $\tau : [s_0, +\infty) \to [t_0, +\infty)$ is given by*

$$\tau(s) := \frac{\alpha - 1}{\lambda} \ln\left(\frac{s}{s_0}\right) + t_0,$$

then the reparametrized trajectory $x : [s_0, +\infty) \to \mathcal{H}$ given by $x(s) := y(\tau(s))$ is a solution of (3.16) for initial conditions

$$x(s_0) = y_0 \quad \text{and} \quad \dot{x}(s_0) = \frac{\alpha - 1}{\lambda s_0} y_1;$$

(ii) *If $x : [s_0, +\infty) \to \mathcal{H}$ is a solution trajectory to (3.16) and the function $\sigma : [t_0, +\infty) \to [s_0, +\infty)$ is given by*

$$\sigma(t) := s_0 \exp\left(\frac{\lambda(t - t_0)}{\alpha - 1}\right),$$

then the reparametrized trajectory $y : [t_0, +\infty) \to \mathcal{H}$ given by $y(t) := x(\sigma(t))$ is a solution of (3.15) for initial conditions

$$y(t_0) = x_0 \quad \text{and} \quad \dot{y}(t_0) = \frac{\lambda s_0}{\alpha - 1} x_1.$$

3.3.2 Transferring the convergence results to the (AVD$_\alpha$) framework

As a direct corollary of Theorem 3.2.3 and Proposition 3.3.1, we obtain the following theorem.

Theorem 3.3.2. *Let $\alpha > 3$ and $x : [s_0, +\infty) \to \mathcal{H}$ be a solution trajectory of*

$$\begin{cases} \ddot{x}(s) + \dfrac{\alpha}{s}\dot{x}(s) + \nabla f(x(s)) = 0, \\ x(s_0) = x_0, \quad \dot{x}(s_0) = x_1. \end{cases}$$

Then, as $s \to +\infty$ it holds

$$f(x(s)) - \min f = o\left(\frac{1}{s^2}\right) \quad \text{and} \quad \|\dot{x}(s)\| = o\left(\frac{1}{s}\right).$$

Furthermore, $x(s)$ converges weakly to an element of $\operatorname{argmin} f$ *as $s \to +\infty$.*

Proof. Let $\lambda > 0$ and $s_0 > 0$. As per Proposition 3.3.1, define $y(t) := x(\sigma(t))$. We know that $y : [t_0, +\infty) \to \mathcal{H}$ is a solution trajectory of (3.15). Taking into account the fact that $\sigma \circ \tau : [s_0, +\infty) \to [s_0, +\infty)$ is the identity function, we have $x(s) = y(\tau(s))$, and therefore $\frac{\lambda s}{\alpha - 1}\dot{x}(s) = \dot{y}(\tau(s))$. Thus, according to Theorem 3.2.3, we know that for $b(t) = s_0^2 \exp\left(\frac{2\lambda(t-t_0)}{\alpha-1}\right)$, as $s \to +\infty$ we have

$$f(x(s)) - \min f = o\left(\frac{1}{\int_{\frac{t_0+\tau(s)}{2}}^{\tau(s)} b(r)dr}\right) \quad \text{and} \quad \frac{\lambda s}{\alpha - 1}\|\dot{x}(s)\| = o\left(\sqrt{\frac{b(\tau(s))}{\int_{\frac{t_0+\tau(s)}{2}}^{\tau(s)} b(r)dr}}\right).$$

Notice that for every $s \geq s_0$ it holds

$$\begin{aligned} \int_{\frac{t_0+\tau(s)}{2}}^{\tau(s)} b(r)dr &= s_0^2 \int_{\frac{t_0+\tau(s)}{2}}^{\tau(s)} \exp\left(\frac{2\lambda(r - t_0)}{\alpha - 1}\right) dr \\ &= \frac{s_0^2(\alpha - 1)}{2\lambda}\left[\exp\left(\frac{2\lambda(\tau(s) - t_0)}{\alpha - 1}\right) - \exp\left(\frac{\lambda(\tau(s) - t_0)}{\alpha - 1}\right)\right] = \frac{s_0^2(\alpha - 1)}{2\lambda}\left[\left(\frac{s}{s_0}\right)^2 - \frac{s}{s_0}\right] \end{aligned}$$

and $b(\tau(s)) = s^2$. This gives $f(x(s)) - \min f = o\left(\frac{1}{s^2}\right)$ and $\frac{\lambda s}{\alpha-1}\|\dot{x}(s)\| \to 0$ as $s \to +\infty$, which verifies the rates we had claimed in the statement. Since $y(t)$ converges weakly to a global minimizer of f as $t \to +\infty$, so does $x(s) = y(\tau(s))$ as $s \to +\infty$. $\square$

3.4 Heavy Ball dynamics governed by a maximally monotone and continuous operator

In this section, we explore an analog to the Heavy Ball dynamics (HBF$_\lambda$), but tailored to solving equations governed by maximal monotone operators. For a monotone and continuous operator $M : \mathcal{H} \to \mathcal{H}$, we consider the equation

$$\text{Find } y \in \mathcal{H} \text{ such that } M(y) = 0. \tag{3.17}$$

Attached to this equation, for $\lambda > 0$, $\mu : [t_0, +\infty) \to \mathbb{R}_{++}$, a continuously differentiable and nondecreasing function, and $\gamma : [t_0, +\infty) \to \mathbb{R}_{++}$ a continuous function, we study the asymptotic properties of the solutions to the following dynamics, for $t \geq t_0 \geq 0$:

$$\ddot{y}(t) + \lambda\dot{y}(t) + \mu(t)\frac{d}{dt}M(y(t)) + \gamma(t)M(y(t)) = 0. \tag{3.18}$$

When M is not cocoercive, which is the case for example when it does not arise from a convex potential, the presence of the "Hessian" term $\frac{d}{dt}M(y(t))$ is needed to ensure convergence rates for the residual and weak convergence results for the trajectory solution.

This section is organized as follows: in Subsection 3.4.1, we address the existence and uniqueness of strong global solutions to our system. In Subsection 3.4.2, we analyze the asymptotic properties of the global solutions to (3.18).

3.4.1 Existence and uniqueness of strong global solutions

We employ similar arguments to those used for the existence and uniqueness result for (HBF$_\lambda$). First of all, we can rewrite (3.18) as a first order system in two variables. Indeed, it is straightforward to check that

$$\begin{cases} \ddot{y}(t) + \lambda\dot{y}(t) + \mu(t)\dfrac{d}{dt}M(y(t)) + \gamma(t)M(y(t)) = 0, \\ y(t_0) = y_0, \quad \dot{y}(t_0) = y_1 \end{cases} \quad \text{and} \quad \begin{cases} \dot{y}(t) &= u(t) - \lambda y(t) - \mu(t)M(y(t)), \\ \dot{u}(t) &= \big(\dot{\mu}(t) - \gamma(t)\big)M(y(t)), \\ y(t_0) &= y_0, \\ u(t_0) &= \lambda y_0 + y_1 + \mu(t_0)M(y_0) \end{cases} \tag{3.19}$$

are equivalent, where $u(t) := \lambda y(t) + \dot{y}(t) + \mu(t)M(y(t))$, or

$$\begin{cases} \big(\dot{y}(t), \dot{u}(t)\big) &= G(t, y(t), u(t)), \\ \big(y(t_0), u(t_0)\big) &= \big(y_0,\ \lambda y_0 + y_1 + \mu(t_0)M(y_0)\big), \end{cases} \tag{3.20}$$

where $G : [t_0, +\infty) \times \mathcal{H} \times \mathcal{H} \to \mathcal{H} \times \mathcal{H}$ is given by

$$G(t, y, u) := \Big(u - \lambda y - \mu(t)M(y),\ \big(\dot{\mu}(t) - \gamma(t)\big)M(y)\Big).$$

The existence and uniqueness of global solutions to (3.20) require strong assumptions, such as the Fréchet differentiability of M. It is more fitting to consider the existence and uniqueness of strong global solutions, which only require that M is Lipschitz continuous.

Definition 3.4.1. We call $y : [t_0, +\infty) \to \mathcal{H}$ a *strong global solution* of (3.19) (see, for example, [37, Definition 3]) if the following statements hold:

(i) $y, \dot{y} : [t_0, +\infty) \to \mathcal{H}$ are locally absolutely continuous, that is, absolutely continuous on each interval $[t_0, t_1]$;

(ii) $\ddot{y}(t) + \lambda\dot{y}(t) + \mu(t)\frac{d}{dt}M(y(t)) + \gamma(t)M(y(t)) = 0$ for almost every $t \in [t_0, +\infty)$;

(iii) $y(t_0) = y_0$ and $\dot{y}(t_0) = y_1$.

Theorem 3.4.2. *Suppose further that $M : \mathcal{H} \to \mathcal{H}$ is L-Lipschitz continuous. Then, for each $(y_0, y_1) \in \mathcal{H} \times \mathcal{H}$, (3.19) has a unique strong global solution $y : [t_0, +\infty) \to \mathcal{H}$.*

Proof. We use the first order reformulation (3.20), and show the existence of strong global solutions to this differential equation in the Hilbert space $\mathcal{H} \times \mathcal{H}$, which we endow with the inner product $\langle (y, u), (\overline{y}, \overline{u}) \rangle_{\mathcal{H}\times\mathcal{H}} := \langle y, \overline{y} \rangle + \langle u, \overline{u} \rangle$ and corresponding norm $\|(y, u)\|_{\mathcal{H}\times\mathcal{H}} := \sqrt{\|y\|^2 + \|u\|^2}$.

(a) First, we check the Lipschitz continuity of $G(t, \cdot, \cdot)$ for each $t \in [t_0, +\infty)$. For $(y, u), (\overline{y}, \overline{u}) \in \mathcal{H}\times\mathcal{H}$

$$\begin{aligned} &\big\|G(t, y, u) - G(t, \overline{y}, \overline{u})\big\|_{\mathcal{H}\times\mathcal{H}} \\ &\leq \Big\|(u - \overline{u}) - \lambda(y - \overline{y}) - \mu(t)(M(y) - M(\overline{y}))\Big\| + \Big\|(\dot{\mu}(t) - \gamma(t))(M(y) - M(\overline{y}))\Big\| \\ &\leq \|u - \overline{u}\| + \lambda\|y - \overline{y}\| + L|\mu(t)|\,\|y - \overline{y}\| + L\big|\dot{\mu}(t) - \gamma(t)\big|\|y - \overline{y}\| \\ &\leq \Big(1 + \lambda + L|\mu(t)| + L\big|\dot{\mu}(t) - \gamma(t)\big|\Big)\big\|(y, u) - (\overline{y}, \overline{u})\big\|_{\mathcal{H}\times\mathcal{H}}. \end{aligned}$$

By hypothesis (we are just using the smoothness of μ and the continuity of γ), the (dependent on t) Lipschitz constant attached to $G(t, \cdot, \cdot)$ is locally integrable.

(b) Now, we prove the local integrability of $G(\cdot, y, u)$ for each $y, u \in \mathcal{H}$. We have

$$\begin{aligned} \int_{t_0}^{t_1} \|G(s, y, u)\|_{\mathcal{H}\times\mathcal{H}} ds &\leq \int_{t_0}^{t_1} \Big(\Big\|u - \lambda y - \mu(s)M(y)\Big\| + \Big\|\big(\dot{\mu}(s) - \gamma(s)\big)M(y)\Big\|\Big) ds \\ &\leq \int_{t_0}^{t_1} \Big(\|u\| + \lambda\|y\| + |\mu(s)|\,\|M(y)\| + \big(|\dot{\mu}(s)| + |\gamma(s)|\big)\,\|M(y)\|\Big) ds, \end{aligned}$$

and by hypothesis the right-hand side is finite.

With these two conditions verified, we invoke the Cauchy-Lipschitz-Picard theorem (see, for example, [62, Proposition 6.2.1]) to ensure the existence and uniqueness of a strong global solution to (3.20). The statement follows from the equivalence between the first- and second-order formulations. $\square$

3.4.2 Convergence rates and weak convergence of the trajectories

In this subsection, we will assume that we have a global solution $y : [t_0, +\infty) \to \mathcal{H}$ of (3.18).

Before proceeding, we introduce the energy function that will help us in our analysis and show some estimates that we will need later. Suppose x^* is a zero of M, and $t \mapsto y(t)$ solves (3.18) for $t \geq t_0$. For $0 < \eta < \lambda$, on $[t_0, +\infty)$ we define

$$\begin{aligned}
\mathcal{E}_\eta(t) := \frac{1}{2}\Big\|2\eta(y(t) - x^*) + 2\dot{y}(t) + \mu(t)M(y(t))\Big\|^2 \quad & (\mathcal{E}_\eta^1(t)) \\
+ 2\eta(\lambda - \eta)\|y(t) - x^*\|^2 \quad & (\mathcal{E}_\eta^2(t)) \\
+ 2\eta\mu(t)\langle y(t) - x^*, M(y(t))\rangle \quad & (\mathcal{E}_\eta^3(t)) \\
+ \frac{1}{2}\mu^2(t)\|M(y(t))\|^2. \quad & (\mathcal{E}_\eta^4(t))
\end{aligned}$$

We now compute the time derivative of $\mathcal{E}_\eta$ at a point $t \geq t_0$:

$$\begin{aligned}
\frac{d}{dt}\mathcal{E}_\eta^1(t) &= \left\langle 2\eta(y(t) - x^*) + 2\dot{y}(t) + \mu(t)M(y(t)),\ 2\eta\dot{y}(t) + 2\ddot{y}(t) + \dot{\mu}(t)M(y(t)) + \mu(t)\frac{d}{dt}M(y(t)) \right\rangle \\
&= \Big\langle 2\eta(y(t) - x^*) + 2\dot{y}(t) + \mu(t)M(y(t)), \\
&\qquad\qquad 2(\eta - \lambda)\dot{y}(t) + \Big[\dot{\mu}(t) - 2\gamma(t)\Big]M(y(t)) - \mu(t)\frac{d}{dt}M(y(t)) \Big\rangle \\
&= 4\eta(\eta - \lambda)\langle y(t) - x^*, \dot{y}(t)\rangle + 2\eta\Big[\dot{\mu}(t) - 2\gamma(t)\Big]\langle y(t) - x^*, M(y(t))\rangle \\
&\quad - 2\eta\mu(t)\left\langle y(t) - x^*, \frac{d}{dt}M(y(t)) \right\rangle + 4(\eta - \lambda)\big\|\dot{y}(t)\big\|^2 \\
&\quad + \Big\{2\Big[\dot{\mu}(t) - 2\gamma(t)\Big] + 2(\eta - \lambda)\mu(t)\Big\}\langle \dot{y}(t), M(y(t))\rangle - 2\mu(t)\left\langle \dot{y}(t), \frac{d}{dt}M(y(t)) \right\rangle \\
&\quad + \mu(t)\Big[\dot{\mu}(t) - 2\gamma(t)\Big]\|M(y(t))\|^2 - \mu^2(t)\left\langle M(y(t)), \frac{d}{dt}M(y(t)) \right\rangle, \\
\frac{d}{dt}\mathcal{E}_\eta^2(t) &= 4\eta(\lambda - \eta)\langle y(t) - x^*, \dot{y}(t)\rangle, \\
\frac{d}{dt}\mathcal{E}_\eta^3(t) &= 2\eta\dot{\mu}(t)\langle y(t) - x^*, M(y(t))\rangle + 2\eta\mu(t)\langle \dot{y}(t), M(y(t))\rangle + 2\eta\mu(t)\left\langle y(t) - x^*, \frac{d}{dt}M(y(t)) \right\rangle, \\
\frac{d}{dt}\mathcal{E}_\eta^4(t) &= \mu(t)\dot{\mu}(t)\|M(y(t))\|^2 + \mu^2(t)\left\langle M(y(t)), \frac{d}{dt}M(y(t)) \right\rangle.
\end{aligned}$$

Putting everything together yields

$$\begin{aligned}
\frac{d}{dt}\mathcal{E}_\eta(t) &= \frac{d}{dt}\mathcal{E}_\eta^1(t) + \frac{d}{dt}\mathcal{E}_\eta^2(t) + \frac{d}{dt}\mathcal{E}_\eta^3(t) + \frac{d}{dt}\mathcal{E}_\eta^4(t) \\
&= \Big\{2\eta\Big[\dot{\mu}(t) - 2\gamma(t)\Big] + 2\eta\dot{\mu}(t)\Big\}\langle y(t) - x^*, M(y(t))\rangle + 4(\eta - \lambda)\big\|\dot{y}(t)\big\|^2 \\
&\quad + \Big\{2\Big[\dot{\mu}(t) - 2\gamma(t)\Big] + 2(\eta - \lambda)\mu(t) + 2\eta\mu(t)\Big\}\langle \dot{y}(t), M(y(t))\rangle - 2\mu(t)\left\langle \dot{y}(t), \frac{d}{dt}M(y(t)) \right\rangle \\
&\quad + \Big\{\mu(t)\Big[\dot{\mu}(t) - 2\gamma(t)\Big] + \mu(t)\dot{\mu}(t)\Big\}\|M(y(t))\|^2 \\
&= 4\eta\Big[\dot{\mu}(t) - \gamma(t)\Big]\langle y(t) - x^*, M(y(t))\rangle + 2\mu(t)\Big[\dot{\mu}(t) - \gamma(t)\Big]\|M(y(t))\|^2 \\
&\quad + 4(\eta - \lambda)\big\|\dot{y}(t)\big\|^2 - 2\mu(t)\left\langle \dot{y}(t), \frac{d}{dt}M(y(t)) \right\rangle \\
&\quad + 2\Big\{\Big[2(\eta - \lambda)\mu(t) + \lambda\mu(t) - 2\gamma(t)\Big] + \dot{\mu}(t)\Big\}\langle \dot{y}(t), M(y(t))\rangle
\end{aligned}$$

$$
\begin{aligned}
&= 4\eta\Big[\dot\mu(t)-\gamma(t)\Big]\langle y(t)-x^*, M(y(t))\rangle + \frac{2}{3}\mu(t)\Big[\dot\mu(t)-\gamma(t)\Big]\|M(y(t))\|^2\\
&\quad + (\eta-\lambda)\big\|\dot y(t)\big\|^2 - 2\mu(t)\left\langle \dot y(t), \frac{d}{dt}M(y(t))\right\rangle + \mathcal{R}(t),
\end{aligned}
\tag{3.21}
$$

where $\mathcal{R}(t)$ results from rearranging terms and will be made explicit momentarily. Our goal is to have $\frac{d}{dt}\mathcal{E}_\eta(t) \le 0$ for t large enough, which will be ensured under the following assumption:

Assumption 3.2. Suppose that $\lambda > 0$, $\mu : [t_0, +\infty) \to \mathbb{R}_{++}$ and $\gamma : [t_0, +\infty) \to \mathbb{R}_{++}$ satisfy further

$$
\lim_{t\to+\infty}\frac{\gamma(t)}{\mu(t)} =: L > 0, \quad \sup_{t\ge t_0}\frac{\dot\mu(t)}{\gamma(t)} < 1 \quad \text{and} \quad 2\lambda - 3L + \inf_{t\ge t_0}\frac{\dot\mu(t)}{\mu(t)} > 0. \tag{3.22}
$$

Indeed, the growth condition on $\mu(\cdot)$ and $\gamma(\cdot)$ and the positivity of $\mu(\cdot)$ ensure $\dot\mu(t) \le \gamma(t)$, which makes the coefficients accompanying $\langle y(t)-x^*, M(y(t))\rangle$ and $\|M(y(t))\|^2$ nonpositive; the fact that $\eta < \lambda$ also makes the term corresponding to $\big\|\dot y(t)\big\|^2$ nonpositive. Furthermore, the monotonicity of M yields

$$
\left\langle \dot y(t), \frac{d}{dt}M(y(t))\right\rangle = \lim_{s\to t}\frac{1}{(s-t)^2}\Big\langle y(s)-y(t), M(y(s))-M(y(t))\Big\rangle \ge 0,
$$

further multiplication by $-2\mu(t)$ makes this term nonpositive as well.

It only remains to check that $\mathcal{R}(t)$ is nonpositive as well, which requires some preliminary computations. Define, $\varepsilon := \lambda - \eta > 0$, so now $\mathcal{R}(t)$ reads

$$
\mathcal{R}(t) = -3\varepsilon\big\|\dot y(t)\big\|^2 + 2\Big\{\Big[-2\varepsilon\mu(t)+\lambda\mu(t)-2\gamma(t)\Big]+\dot\mu(t)\Big\}\langle \dot y(t), M(y(t))\rangle + \frac{4}{3}\mu(t)\Big[\dot\mu(t)-\gamma(t)\Big]\|M(y(t))\|^2. \tag{3.23}
$$

To establish that $\mathcal{R}(t)$ is nonpositive for large enough t, we make use of Lemma 2.3.5 and set $X := \dot y(t)$, $Y := M(y(t))$ and A, B, C as

$$
A := -3\varepsilon, \quad B := \big(-2\varepsilon\mu(t)+\lambda\mu(t)-2\gamma(t)\big)+\dot\mu(t), \quad C := \frac{4}{3}\mu(t)\Big[\dot\mu(t)-\gamma(t)\Big].
$$

We have

$$
\begin{aligned}
B^2 - AC &= \Big[\big(-2\varepsilon\mu(t)+\lambda\mu(t)-2\gamma(t)\big)+\dot\mu(t)\Big]^2 + 3\cdot\frac{4}{3}\varepsilon\mu(t)\Big[\dot\mu(t)-\gamma(t)\Big]\\
&= \Big[-2\varepsilon\mu(t)+\lambda\mu(t)-2\gamma(t)\Big]^2 - 4\varepsilon\mu(t)\dot\mu(t) + 2\big(\lambda\mu(t)-2\gamma(t)\big)\dot\mu(t) + \big(\dot\mu(t)\big)^2\\
&\quad + 4\varepsilon\mu(t)\dot\mu(t) - 4\varepsilon\mu(t)\gamma(t)\\
&= 4\varepsilon^2\mu^2(t) - 4\varepsilon\mu(t)\big(\lambda\mu(t)-2\gamma(t)\big) + \big(\lambda\mu(t)-2\gamma(t)\big)^2 + 2\big(\lambda\mu(t)-2\gamma(t)\big)\dot\mu(t)\\
&\quad + \big(\dot\mu(t)\big)^2 - 4\varepsilon\mu(t)\gamma(t)\\
&= \mu^2(t)\left\{4\varepsilon^2 - 4\left(\lambda - \frac{\gamma(t)}{\mu(t)}\right)\varepsilon + \left[\lambda - \frac{2\gamma(t)}{\mu(t)} + \frac{\dot\mu(t)}{\mu(t)}\right]^2\right\}.
\end{aligned}
\tag{3.24}
$$

Notice that $B^2 - AC$ is a quadratic expression in ε. Proposition 3.4.4, which is an intermediate result before the main theorem of this subsection, will show that there exists a suitable interval $I \subseteq (0, \lambda)$ with nonempty interior such that this parabola eventually becomes negative for every $\varepsilon \in I$. It also shows integral statements which will be needed when showing weak convergence of trajectories and little-o rates.

Remark 3.4.3. Assumption 3.2 implies in particular that $\lambda > L$, a fact that we will need later. Let us show this claim here. Since

$$
2\lambda > 3L - \inf_{t\ge t_0}\frac{\dot\mu(t)}{\mu(t)}, \text{ there exists } \delta > 0 \text{ such that } 2\lambda - \delta \ge 3L - \inf_{t\ge t_0}\frac{\dot\mu(t)}{\mu(t)}.
$$

Since $\lim_{t\to+\infty} \frac{\gamma(t)}{\mu(t)} = L$, for every given $0 < \varepsilon < \delta$ there exists $t_\varepsilon \geq t_0$ such that for every $t \geq t_\varepsilon$

$$L - \varepsilon \leq \frac{\gamma(t)}{\mu(t)} \leq L + \varepsilon, \text{ and in particular, } \frac{\dot{\mu}(t)}{\mu(t)} \leq (L+\varepsilon)\frac{\dot{\mu}(t)}{\gamma(t)} \leq L + \varepsilon,$$

where in the last estimate we used the assumption $\sup_{t\geq t_0} \frac{\dot{\mu}(t)}{\gamma(t)} < 1$ and the nondecreasing property of μ. Now, for every $t \geq t_\varepsilon$ we have

$$2\lambda - \delta \geq 3L - \inf_{t\geq t_0} \frac{\dot{\mu}(t)}{\mu(t)} \geq 3L - \frac{\dot{\mu}(t)}{\mu(t)} \geq 3L - L - \varepsilon = 2L - \varepsilon,$$

from which we deduce

$$2\lambda > 2L + \delta - \varepsilon > 2L \text{ and thus } \lambda > L.$$

Proposition 3.4.4. *Suppose further that λ, $\mu(\cdot)$ and $\gamma(\cdot)$ satisfy Assumption 3.2. Let x^* be a zero of M and $y : [t_0, +\infty) \to \mathcal{H}$ be a solution trajectory of (3.18). The following statements are true:*

(i) *The limit $\lim_{t\to+\infty} \mathcal{E}_{\lambda-\varepsilon}(t) \in \mathbb{R}$ exists for every ε in a given interval $I \subseteq (0, \lambda)$;*

(ii) *It holds*

$$\int_{t_0}^{+\infty} \mu(t)\langle y(t) - x^*, M(y(t))\rangle dt < +\infty, \quad \int_{t_0}^{+\infty} \|\dot{y}(t)\|^2 dt < +\infty$$
$$\textit{and} \quad \int_{t_0}^{+\infty} \mu^2(t)\|M(y(t))\|^2 dt < +\infty.$$

Proof. (i) Just as we mentioned before the previous remark, we want to find an interval I with nonempty interior, independent of time, such that $I \subseteq (0, \lambda)$ and such that for every $\varepsilon \in I$ the parabola (3.24) becomes negative for large enough t. We do this in two steps: first, we show that the discriminant, which depends on t, stays away from zero as $t \to +\infty$, and thus the (also dependent on t) roots are properly separated as $t \to +\infty$. Second, we show that the left and right roots stay respectively below and above certain thresholds as $t \to +\infty$. Combining these facts will yield the desired interval I.

To proceed, we first observe that, according to (3.2), and since $\lambda > L$ following Remark 3.4.3, we can choose δ_1, δ_2 such that

$$\max\left\{2 - \frac{\lambda}{L}, \sup_{t\geq t_0} \frac{\dot{\mu}(t)}{\gamma(t)}\right\} < \delta_1 < \delta_2 < 1. \tag{3.25}$$

After some algebraic manipulation, we deduce that for every $t \geq t_0$

$$\lambda - \frac{2\gamma(t)}{\mu(t)} + \frac{\dot{\mu}(t)}{\mu(t)} < \lambda + (\delta_1 - 2)\frac{\gamma(t)}{\mu(t)} < \lambda + (\delta_2 - 2)\frac{\gamma(t)}{\mu(t)} < \lambda - \frac{\gamma(t)}{\mu(t)}.$$

Moreover, we know that there exist $\tilde{\delta}_2 > \tilde{\delta}_1 > 0$ satisfying

$$0 < \lambda + (\delta_1 - 2)L < \tilde{\delta}_1 < \tilde{\delta}_2 < \lambda + (\delta_2 - 2)L. \tag{3.26}$$

Using $\lim_{t\to+\infty} \frac{\gamma(t)}{\mu(t)} = L > 0$, we know that there exist $t_1 \geq t_0$ such that for every $t \geq t_1$

$$\lambda - \frac{2\gamma(t)}{\mu(t)} + \frac{\dot{\mu}(t)}{\mu(t)} < \lambda + (\delta_1 - 2)\frac{\gamma(t)}{\mu(t)} \leq \tilde{\delta}_1 < \tilde{\delta}_2 \leq \lambda + (\delta_2 - 2)\frac{\gamma(t)}{\mu(t)} < \lambda - \frac{\gamma(t)}{\mu(t)}. \tag{3.27}$$

On the other hand, according to (3.2), we know that

$$L - \lambda < \lambda - 2L + \inf_{t\geq t_0} \frac{\dot{\mu}(t)}{\mu(t)}.$$

Again recalling that $\lambda > L$, we can construct $\tilde{\delta}_4 > \tilde{\delta}_3 > 0$ and $t_2 \geq t_1$ such that for every $t \geq t_2$

$$L - \lambda < -\tilde{\delta}_4 < -\tilde{\delta}_3 < \lambda - 2L + \inf_{t\geq t_0} \frac{\dot{\mu}(t)}{\mu(t)} \quad \text{and}$$

$$\frac{\gamma(t)}{\mu(t)} - \lambda \leq -\tilde{\delta}_4 < -\tilde{\delta}_3 \leq \lambda - \frac{2\gamma(t)}{\mu(t)} + \inf_{t \geq t_0} \frac{\dot{\mu}(t)}{\mu(t)} \leq \lambda - \frac{2\gamma(t)}{\mu(t)} + \frac{\dot{\mu}(t)}{\mu(t)}. \tag{3.28}$$

Combining (3.27) and (3.28) yields, for every $t \geq t_2$,

$$\frac{\gamma(t)}{\mu(t)} - \lambda \leq -\tilde{\delta}_4 < -\tilde{\delta}_3 \leq \lambda - \frac{2\gamma(t)}{\mu(t)} + \frac{\dot{\mu}(t)}{\mu(t)} \leq \tilde{\delta}_1 < \tilde{\delta}_2 \leq \lambda - \frac{\gamma(t)}{\mu(t)}.$$

Without loss of generality, since both cases can be handed identically, we may assume that

$$\max\left\{\tilde{\delta}_1, \tilde{\delta}_3\right\} = \tilde{\delta}_1.$$

Therefore, the reduced discriminant of (3.24) satisfies, for every $t \geq t_2$,

$$\Delta_t := 4\left[\left(\lambda - \frac{\gamma(t)}{\mu(t)}\right)^2 - \left(\lambda - \frac{2\gamma(t)}{\mu(t)} + \frac{\dot{\mu}(t)}{\mu(t)}\right)^2\right] \geq 4\left(\tilde{\delta}_2^2 - \tilde{\delta}_1^2\right) > 0, \tag{3.29}$$

and thus stays away from zero as $t \to +\infty$. Now, choose $\delta_3 > 0$ such that

$$\delta_3 < \min\left\{L,\ \lambda - L,\ \frac{1}{2}\sqrt{\tilde{\delta}_2^2 - \tilde{\delta}_1^2}\right\}.$$

Once again using the fact that $\lim_{t \to +\infty} \frac{\gamma(t)}{\mu(t)} = L$, there exists $t_3 \geq t_2$ such that, for every $t \geq t_3$,

$$\lambda - L - \delta_3 \leq \lambda - \frac{\gamma(t)}{\mu(t)} \leq \lambda - L + \delta_3. \tag{3.30}$$

The roots of the quadratic expression (in ε) in equation (3.24) are

$$\underline{\varepsilon}_t := \frac{1}{2}\left(\lambda - \frac{\gamma(t)}{\mu(t)}\right) - \frac{1}{4}\sqrt{\Delta_t} \quad \text{and} \quad \overline{\varepsilon}_t := \frac{1}{2}\left(\lambda - \frac{\gamma(t)}{\mu(t)}\right) + \frac{1}{4}\sqrt{\Delta_t}.$$

Now, using (3.29) and (3.30) we can deduce that for every $t \geq t_3$

$$\begin{aligned} \delta_3 < \frac{1}{2}\sqrt{\tilde{\delta}_2^2 - \tilde{\delta}_1^2} \quad &\Rightarrow \quad \lambda - L + \delta_3 - \sqrt{\tilde{\delta}_2^2 - \tilde{\delta}_1^2} < \lambda - L - \delta_3 \\ \Rightarrow \quad \underline{\varepsilon}_t = \frac{1}{2}\left(\lambda - \frac{\gamma(t)}{\mu(t)}\right) - \frac{1}{4}\sqrt{\Delta_t} &< \frac{1}{2}\left(\lambda - L + \delta_3 - \sqrt{\tilde{\delta}_2^2 - \tilde{\delta}_1^2}\right) < \frac{1}{2}\left(\lambda - L - \delta_3\right). \end{aligned}$$

Similarly, for every $t \geq t_3$ we obtain

$$\frac{1}{2}\left(\lambda - L + \delta_3\right) < \overline{\varepsilon}_t.$$

Thus, we define the interval I as

$$I := \left[\frac{1}{2}\left(\lambda - L - \delta_3\right), \frac{1}{2}\left(\lambda - L + \delta_3\right)\right].$$

The previous inequalities, together the way we chose δ_3, ensure that

$$I \subseteq (\underline{\varepsilon}_t, \overline{\varepsilon}_t) \subseteq (0, \lambda).$$

Thus, for every $\varepsilon \in I$, the term (3.24) is negative for $t \geq t_3$:

$$B^2 - AC < 0,$$

and we recall that according to Lemma 2.3.5, this in turn implies $\mathcal{R}(t) \leq 0$ for all $t \geq t_3$, which is what we wanted. Having shown (3.23), we go back to (3.21) and we drop the nonpositive terms corresponding to $\langle \dot{y}(t), \frac{d}{dt}M(y(t))\rangle$ and $\mathcal{R}(t)$. For every $\eta = \lambda - \varepsilon > 0$ with $\varepsilon \in I$, for every $t \geq t_3$ it holds

$$\frac{d}{dt}\mathcal{E}_\eta(t) \leq 4\eta\Big[\dot{\mu}(t) - \gamma(t)\Big]\langle y(t) - x^*, M(y(t))\rangle + (\eta - \lambda)\big\|\dot{y}(t)\big\|^2 + \frac{2}{3}\mu(t)\Big[\dot{\mu}(t) - \gamma(t)\Big]\|M(y(t))\|^2 \leq 0. \tag{3.31}$$

This means that for every $\eta = \lambda - \varepsilon$, where $\varepsilon \in I$, $\mathcal{E}_\eta$ is nonincreasing on $[t_3, +\infty)$, therefore, for every $t \geq t_3$ we have

$$0 \leq \mathcal{E}_\eta(t) \leq \mathcal{E}_\eta(t_3), \tag{3.32}$$

thus

$$\lim_{t\to+\infty} \mathcal{E}_\eta(t) \in \mathbb{R} \text{ exists.} \tag{3.33}$$

In particular, going back to ($\mathcal{E}_\eta^1(t)$)-($\mathcal{E}_\eta^4(t)$), we deduce that, for every $t \geq t_3$,

$$\|M(y(t))\| \leq \frac{\sqrt{2\mathcal{E}_\eta(t_3)}}{\mu(t)} \quad \text{and} \quad \langle y(t) - x^*, M(y(t))\rangle \leq \frac{\mathcal{E}_\eta(t_3)}{2\eta\mu(t)}. \tag{3.34}$$

(ii) From (3.25) and (3.30), for every $t \geq t_3$ we have

$$\begin{aligned}
&\frac{\dot\mu(t)}{\gamma(t)} \leq \delta_1 < 1 \quad \text{and} \quad L - \delta_3 \leq \frac{\gamma(t)}{\mu(t)} \leq L + \delta_3 \quad \Rightarrow \quad (1-\delta_1)\gamma(t) \leq \gamma(t) - \dot\mu(t) \\
&\qquad \Rightarrow \quad (1-\delta_1)(L-\delta_3)\mu(t) \leq (1-\delta_1)\gamma(t) \leq \gamma(t) - \dot\mu(t).
\end{aligned}$$

Going back to (3.31), we integrate the inequality from t_3 to $t \geq t_3$ and obtain

$$\begin{aligned}
&4\eta(L-\delta_3)(1-\delta_1)\int_{t_3}^t \mu(s)\langle y(s) - x^*, M(y(s))\rangle ds + (\lambda - \eta)\int_{t_3}^t \left\|\dot y(s)\right\|^2 ds \\
&+ \frac{2}{3}(L-\delta_3)(1-\delta_1)\int_{t_3}^t \mu^2(s)\|M(y(s))\|^2 ds \\
\leq\ & 4\eta \int_{t_3}^t \Big[\gamma(s) - \dot\mu(s)\Big]\langle y(s) - x^*, M(y(s))\rangle ds + (\lambda-\eta)\int_{t_3}^t \left\|\dot y(s)\right\|^2 ds \\
&+ \frac{2}{3}\int_{t_3}^t \mu(s)\Big[\gamma(s) - \dot\mu(s)\Big]\|M(y(s))\|^2 ds \\
\leq\ & \mathcal{E}_\eta(t_3) - \mathcal{E}_\eta(t),
\end{aligned}$$

which produces the required integrability results. □

We are now in a position to state and prove the main theorem of this section.

Theorem 3.4.5. *Suppose that λ, $\mu(\cdot)$ and $\gamma(\cdot)$ satisfy Assumption 3.2. Let x^* be a zero of M and $y : [t_0, +\infty) \to \mathcal{H}$ be a solution trajectory of (3.18). Then, the following statements are true:*

(i) *As $t \to +\infty$, it holds*

$$\|M(y(t))\| = o\left(\frac{1}{\mu(t)}\right), \quad \langle y(t) - x^*, M(y(t))\rangle = o\left(\frac{1}{\mu(t)}\right) \quad \text{and} \quad \left\|\dot y(t)\right\| \to 0; \tag{3.35}$$

(ii) *The solution trajectory $y(t)$ converges weakly to a zero of M as $t \to +\infty$.*

Proof. (i) Let $I \subseteq (0, \lambda)$ be the interval provided by Proposition 3.4.4 (i). Choose $\varepsilon_1, \varepsilon_2 \in I$ such that $\varepsilon_1 \neq \varepsilon_2$. Set $\eta_i := \lambda - \varepsilon_i$, $i = 1, 2$. We have

$$\begin{aligned}
&\mathcal{E}_{\eta_2}(t) - \mathcal{E}_{\eta_1}(t) \\
=\ & \frac{1}{2}\Big\|2\eta_2(y(t) - x^*) + 2\dot y(t) + \mu(t)M(y(t))\Big\|^2 - \frac{1}{2}\Big\|2\eta_1(y(t) - x^*) + 2\dot y(t) + \mu(t)M(y(t))\Big\|^2 \\
&+ 2\Big[\eta_2(\lambda - \eta_2) - \eta_1(\lambda - \eta_1)\Big]\|y(t) - x^*\|^2 + 2(\eta_2 - \eta_1)\mu(t)\langle y(t) - x^*, M(y(t))\rangle \\
=\ & 2\big(\eta_2^2 - \eta_1^2\big)\|y(t) - x^*\|^2 + 2(\eta_2 - \eta_1)\Big\langle y(t) - x^*, 2\dot y(t) + \mu(t)M(y(t))\Big\rangle \\
&+ \Big[2\lambda(\eta_2 - \eta_1) - 2\big(\eta_2^2 - \eta_1^2\big)\Big]\|y(t) - x^*\|^2 + 2(\eta_2 - \eta_1)\mu(t)\langle y(t) - x^*, M(y(t))\rangle
\end{aligned}$$

$$= 4(\eta_2 - \eta_1)\left[\frac{\lambda}{2}\|y(t) - x^*\|^2 + \Big\langle y(t) - x^*, \dot{y}(t) + \mu(t)M(y(t))\Big\rangle\right].$$

Define, for $t \geq t_0$,

$$p(t) := \frac{\lambda}{2}\|y(t) - x^*\|^2 + \Big\langle y(t) - x^*, \dot{y}(t) + \mu(t)M(y(t))\Big\rangle.$$

Because of (3.33), and since $\eta_2 - \eta_1 \neq 0$, we deduce that

$$\lim_{t\to+\infty} p(t) \in \mathbb{R} \text{ exists.}$$

With this at hand, we can now rewrite $\mathcal{E}_{\eta_1}(t)$ for every $t \geq t_0$ as

$$\begin{aligned}
\mathcal{E}_{\eta_1}(t) &= \frac{1}{2}\Big\|2\eta_1(y(t) - x^*) + 2\dot{y}(t) + \mu(t)M(y(t))\Big\|^2 + 2\eta_1(\lambda - \eta_1)\|y(t) - x^*\|^2 \\
&\quad + 2\eta_1\mu(t)\langle y(t) - x^*, M(y(t))\rangle + \frac{1}{2}\mu^2(t)\|M(y(t))\|^2 \\
&= \frac{1}{2}\left[4\eta_1^2\|y(t) - x^*\|^2 + 4\eta_1\Big\langle y(t) - x^*, 2\dot{y}(t) + \mu(t)M(y(t))\Big\rangle + \Big\|2\dot{y}(t) + \mu(t)M(y(t))\Big\|^2\right] \\
&\quad + 2\eta_1\lambda\|y(t) - x^*\|^2 - 2\eta_1^2\|y(t) - x^*\|^2 + 2\eta_1\mu(t)\langle y(t) - x^*, M(y(t))\rangle + \frac{1}{2}\mu^2(t)\|M(y(t))\|^2 \\
&= 4\eta_1 p(t) + \frac{1}{2}\Big\|2\dot{y}(t) + \mu(t)M(y(t))\Big\|^2 + \frac{1}{2}\mu^2(t)\|M(y(t))\|^2 \\
&= 4\eta_1 p(t) + \Big\|\dot{y}(t) + \mu(t)M(y(t))\Big\|^2 + \big\|\dot{y}(t)\big\|^2.
\end{aligned}$$

Define

$$h(t) := \Big\|\dot{y}(t) + \mu(t)M(y(t))\Big\|^2 + \big\|\dot{y}(t)\big\|^2.$$

Since both $\lim_{t\to+\infty}\mathcal{E}_{\eta_1}(t)$ and $\lim_{t\to+\infty} p(t)$ exist, so does $\lim_{t\to+\infty} h(t)$. Notice that

$$\int_{t_0}^{t} h(s)ds \leq 3\int_{t_0}^{t}\big\|\dot{y}(s)\big\|^2 ds + 2\int_{t_0}^{t}\mu^2(s)\|M(y(s))\|^2 ds,$$

and the integrals on the right-hand side remain finite as $t \to +\infty$ according to Proposition 3.4.4 (i) (ii), i.e., $\int_{t_0}^{+\infty} h(t)dt < +\infty$. Combining this with the existence of $\lim_{t\to+\infty} h(t)$ allows us to deduce

$$\lim_{t\to+\infty} h(t) = 0.$$

This implies

$$\lim_{t\to+\infty}\|\dot{y}(t)\| = 0,$$

combining this with $\lim_{t\to+\infty}\big\|\dot{y}(t) + \mu(t)M(y(t))\big\| = 0$ gives that, as $t \to +\infty$ it holds

$$\lim_{t\to+\infty}\mu(t)\|M(y(t))\| = 0, \quad \text{or} \quad \|M(y(t))\| = o\left(\frac{1}{\mu(t)}\right).$$

Using the boundedness of $t \mapsto \|y(t) - x^*\|$, we obtain that as $t \to +\infty$

$$\langle y(t) - x^*, M(y(t))\rangle = o\left(\frac{1}{\mu(t)}\right).$$

(ii) We will make use of Opial's Lemma. Define, for $t \geq t_0$,

$$q(t) := \frac{1}{2}\|y(t) - x^*\|^2 + \int_{t_0}^{t}\mu(s)\langle y(s) - x^*, M(y(s))\rangle ds.$$

Recalling the definition of $p(\cdot)$, we have

$$\lambda q(t) + \dot{q}(t) = \frac{\lambda}{2}\|y(t) - x^*\|^2 + \Big\langle y(t) - x^*, \dot{y}(t) + \mu(t)M(y(t))\Big\rangle + \lambda\int_{t_0}^{t}\mu(s)\langle y(s) - x^*, M(y(s))\rangle ds$$

$$= p(t) + \lambda \int_{t_0}^{t} \mu(s)\langle y(s) - x^*, M(y(s))\rangle ds.$$

According to Proposition 3.4.4 (ii), the integral in the previous sum converges as $t \to +\infty$. Since we already established that $\lim_{t\to+\infty} p(t)$ exists, we are lead to

$$\lim_{t\to+\infty} \big(\lambda q(t) + \dot{q}(t)\big) \in \mathbb{R} \text{ exists.}$$

Using Lemma 2.3.3, we obtain the existence of $\lim_{t\to+\infty} q(t)$. Using again that

$$\int_{t_0}^{+\infty} \mu(s)\langle y(s) - x^*, M(y(s))\rangle ds < +\infty,$$

we finally deduce that

$$\lim_{t\to+\infty} \|y(t) - x^*\| \text{ exists.}$$

Thus, the first condition of Opial's Lemma is met.

For the second condition, let $\overline{y}$ be a sequential cluster point of $t \mapsto y(t)$, i.e., there exists a sequence $(t_n)_{n\in\mathbb{N}} \subseteq [t_0, +\infty)$ such that $t_n \to +\infty$ and $y(t_n) \rightharpoonup \overline{y}$ as $n \to +\infty$. Since we already know that $\|M(y(t))\| = o\left(\frac{1}{\mu(t)}\right)$ as $t \to +\infty$ and $\mu(\cdot)$ is nondecreasing, we have $M(y(t_n)) \to 0$ as $n \to +\infty$. Since M is maximally monotone, its graph is closed in $\mathcal{H}^{\text{weak}} \times \mathcal{H}^{\text{strong}}$, thus

$$M(\overline{y}) = 0.$$

Now both conditions of Opial's Lemma (see Lemma 2.1.2) are fulfilled, from which we finally conclude the proof of this theorem. □

3.5 Connection with a system with asymptotic vanishing damping governed by a monotone and continuous operator

Here, as it was done before, we present two systems attached to (3.17): one has a fixed viscosity parameter, i.e., the Heavy Ball dynamics (3.18), while the other features an asymptotically vanishing viscosity coefficient accompanying the velocity; precisely speaking, these are exactly the Fast OGDA dynamics (see [40]) for the case $\beta(\cdot) \equiv 1$. While at first glance they might appear fundamentally different, we will see that one is a time-rescaled version of the other.

3.5.1 Two equivalent dynamical systems through time rescaling

Similar to what was done in Section 3.3, we start with a trajectory solution $y : [t_0, +\infty) \to \mathcal{H}$ of

$$\ddot{y}(t) + \lambda\dot{y}(t) + \mu(t)\frac{d}{dt}M(y(t)) + \gamma(t)M(y(t)) = 0, \tag{3.36}$$

and define $x(s) := y(\tau(s))$, where $\tau : [s_0, +\infty) \to [t_0, +\infty)$ is a continuously differentiable function such that $\dot{\tau}(s) > 0$ for every $s \geq s_0 > 0$ and $\lim_{s\to+\infty} \tau(s) = +\infty$. We have

$$\dot{x}(s) = \dot{\tau}(s)\dot{y}(\tau(s)) \quad \text{and} \quad \ddot{x}(s) = \ddot{\tau}(s)\dot{y}(\tau(s)) + \big(\dot{\tau}(s)\big)^2\ddot{y}(\tau(s)).$$

These expressions lead to

$$\dot{y}(\tau(s)) = \frac{1}{\dot{\tau}(s)}\dot{x}(s) \quad \text{and} \quad \ddot{y}(\tau(s)) = \frac{1}{\big(\dot{\tau}(s)\big)^2}\Big[\ddot{x}(s) - \ddot{\tau}(s)\dot{y}(\tau(s))\Big] = \frac{1}{\big(\dot{\tau}(s)\big)^2}\left[\ddot{x}(s) - \frac{\ddot{\tau}(s)}{\dot{\tau}(s)}\dot{x}(s)\right].$$

Moreover

$$\frac{d}{ds}\Big(M(y(\cdot)) \circ \tau\Big)(s) = \dot{\tau}(s)\frac{d}{dt}M(y(t))\Big|_{t=\tau(s)} \Rightarrow \frac{d}{dt}M(y(t))\Big|_{t=\tau(s)} = \frac{1}{\dot{\tau}(s)}\frac{d}{ds}M(x(s)).$$

Now, plugging $t = \tau(s)$ in (3.36) gives

$$\frac{1}{\big(\dot{\tau}(s)\big)^2}\left[\ddot{x}(s) - \frac{\ddot{\tau}(s)}{\dot{\tau}(s)}\dot{x}(s)\right] + \frac{\lambda}{\dot{\tau}(s)}\dot{x}(s) + \mu(\tau(s))\frac{1}{\dot{\tau}(s)}\frac{d}{ds}M(x(s)) + \gamma(\tau(s))M(x(s)) = 0$$

or, equivalently,

$$\ddot{x}(s) + \left[\lambda\dot{\tau}(s) - \frac{\ddot{\tau}(s)}{\dot{\tau}(s)}\right]\dot{x}(s) + \dot{\tau}(s)\mu(\tau(s))\frac{d}{ds}M(x(s)) + \left(\dot{\tau}(s)\right)^2\gamma(\tau(s))M(x(s)) = 0. \tag{3.37}$$

We briefly recall that the Fast OGDA dynamics [40], for constant $\beta(\cdot) \equiv 1$, are, for $s \geq s_0 > 0$,

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \frac{d}{ds}M(x(s)) + \frac{\alpha}{2s}M(x(s)) = 0. \tag{3.38}$$

Going back to (3.37), if we want an asymptotic vanishing viscosity coefficient accompanying the velocity, we need to ask for

$$\begin{cases} \lambda\dot{\tau}(s) - \dfrac{\ddot{\tau}(s)}{\dot{\tau}(s)} &= \dfrac{\alpha}{s} \\ \tau(s_0) &= t_0. \end{cases}$$

Since it was done before in Section 3.3, we don't repeat the derivation for finding the solution here. We know that for $s \geq s_0 > 0$, the function

$$\tau(s) := \frac{\alpha-1}{\lambda}\ln(s) + \left(-\frac{\alpha-1}{\lambda}\ln(s_0) + t_0\right) = \frac{\alpha-1}{\lambda}\ln\left(\frac{s}{s_0}\right) + t_0 \tag{3.39}$$

satisfies this differential equation. We have

$$\dot{\tau}(s) = \frac{\alpha-1}{\lambda s} \quad \Rightarrow \quad \left(\dot{\tau}(s)\right)^2 = \frac{(\alpha-1)^2}{\lambda^2 s^2}.$$

We need, of course, to assume that $\alpha > 1$. Furthermore, we want the coefficient attached to $M(x(s))$ to be $\frac{\alpha}{2s}$, i.e.,

$$\left(\dot{\tau}(s)\right)^2\gamma(\tau(s)) = \frac{\alpha}{2s} \quad \Leftrightarrow \quad \gamma\left(\frac{\alpha-1}{\lambda}\ln\left(\frac{s}{s_0}\right) + t_0\right) = \frac{\alpha}{2}\cdot\frac{\lambda^2 s}{(\alpha-1)^2},$$

which is fulfilled if we choose

$$\gamma(t) := \frac{\alpha}{2}\cdot\frac{\lambda^2 s_0}{(\alpha-1)^2}\exp\left(\frac{\lambda(t-t_0)}{\alpha-1}\right).$$

Set $\mu(\cdot) = \gamma(\cdot)$. Notice that for every $t \geq t_0$

$$\frac{\dot{\mu}(t)}{\gamma(t)} = \frac{\dot{\mu}(t)}{\mu(t)} = \frac{\left(\frac{\lambda s_0}{\alpha-1}\right)^2\cdot\frac{\lambda}{\alpha-1}\exp\left(\frac{\lambda(t-t_0)}{\alpha-1}\right)}{\left(\frac{\lambda s_0}{\alpha-1}\right)^2\exp\left(\frac{\lambda(t-t_0)}{\alpha-1}\right)} = \frac{\lambda}{\alpha-1}.$$

Thus, Assumption 3.2 is satisfied if and only if

$$\frac{\lambda}{\alpha-1} < 1 \quad \text{and} \quad 2\lambda - 3 + \frac{\lambda}{\alpha-1} > 0.$$

After rearranging terms, the two previous inequalities are equivalent to

$$\frac{3(\alpha-1)}{2\alpha-1} = \frac{3}{2+\frac{1}{\alpha-1}} < \lambda < \alpha - 1. \tag{3.40}$$

In particular,

$$\frac{3(\alpha-1)}{2\alpha-1} < \alpha - 1 \quad \Leftrightarrow \quad 2 < \alpha.$$

We now turn our attention to the coefficient attached to $\frac{d}{ds}M(x(s))$ in (3.37). Since we want to reach (3.38), we need

$$\frac{\alpha}{2}\cdot\frac{\lambda}{\alpha-1} = \dot{\tau}(s)\mu(\tau(s)) = 1 \quad \Leftrightarrow \quad \lambda = \frac{2(\alpha-1)}{\alpha}.$$

We must verify that this choice of λ satisfies inequality (3.40). Indeed,

$$\frac{2(\alpha-1)}{\alpha} < \alpha - 1 \quad \Leftrightarrow \quad 2 < \alpha \quad \text{and}$$

$$\frac{3(\alpha-1)}{2\alpha-1} < \frac{2(\alpha-1)}{\alpha} \quad \Leftrightarrow \quad 3\alpha < 4\alpha - 2 \quad \Leftrightarrow \quad 2 < \alpha.$$

All in all, $x : [s_0, +\infty) \to \mathcal{H}$ fulfills

$$\ddot{x}(s) + \frac{\alpha}{s}\dot{x}(s) + \frac{d}{ds}M(x(s)) + \frac{\alpha}{2s}M(x(s)) = 0.$$

Conversely, if for $\alpha > 2$, $x : [s_0, +\infty) \to \mathcal{H}$ is a trajectory solution of the previous system and we define $y(t) := x(\sigma(t))$, where $\sigma : [t_0, +\infty) \to [s_0, +\infty)$ is a continuously differentiable function such that $\dot{\sigma}(t) > 0$ for all $t \geq t_0 \geq 0$ and $\lim_{t\to+\infty} \sigma(t) = +\infty$, arguing in a similar fashion as it was done previously we arrive at

$$\ddot{y}(t) + \left[\alpha\frac{\dot{\sigma}(t)}{\sigma(t)} - \frac{\ddot{\sigma}(t)}{\dot{\sigma}(t)}\right]\dot{y}(t) + \dot{\sigma}(t)\frac{d}{dt}M(y(t)) + \frac{\alpha}{2}\cdot\frac{\left(\dot{\sigma}(t)\right)^2}{\sigma(t)}M(y(t)) = 0.$$

We want the coefficient attached to $\dot{y}(t)$ to be $\lambda = \frac{2(\alpha-1)}{\alpha}$. For this end, we need σ to satisfy the differential equation

$$\begin{cases} \alpha\dfrac{\dot{\sigma}(t)}{\sigma(t)} - \dfrac{\ddot{\sigma}(t)}{\dot{\sigma}(t)} &= \dfrac{2(\alpha-1)}{\alpha}, \\ \sigma(t_0) &= s_0, \end{cases} \quad \text{which is fulfilled by} \quad \sigma(t) := s_0 \exp\left(\frac{2(t-t_0)}{\alpha}\right).$$

With this choice for $\sigma(\cdot)$, the resulting system reads

$$\ddot{y}(t) + \frac{2(\alpha-1)}{\alpha}\dot{y}(t) + \frac{2s_0}{\alpha}\exp\left(\frac{2(t-t_0)}{\alpha}\right)\frac{d}{dt}M(y(t)) + \frac{2s_0}{\alpha}\exp\left(\frac{2(t-t_0)}{\alpha}\right)M(y(t)) = 0.$$

It is straightforward to check (we have actually done this already when going from the Heavy Ball system to the Fast OGDA dynamics) that

$$\lambda = \frac{2(\alpha-1)}{\alpha} \quad \text{and} \quad \mu(t) = \gamma(t) = \frac{2s_0}{\alpha}\exp\left(\frac{2(t-t_0)}{\alpha}\right)$$

satisfy Assumption 3.2, and thus all the results ensured by Theorem 3.4.5 hold.

We have essentially shown the following proposition.

Proposition 3.5.1. *Assume that $\alpha > 2$ and that $s_0 > 0, t_0 \geq 0$ are initial times. Consider the following second-order systems:*

$$\begin{cases} \ddot{y}(t) + \dfrac{2(\alpha-1)}{\alpha}\dot{y}(t) + \dfrac{2s_0}{\alpha}\exp\left(\dfrac{2(t-t_0)}{\alpha}\right)\dfrac{d}{dt}M(y(t)) + \dfrac{2s_0}{\alpha}\exp\left(\dfrac{2(t-t_0)}{\alpha}\right)M(y(t)) = 0, \\ y(t_0) = y_0, \quad \dot{y}(t_0) = y_1, \end{cases} \tag{3.41}$$

and

$$\begin{cases} \ddot{x}(s) + \dfrac{\alpha}{s}\dot{x}(s) + \dfrac{d}{ds}M(x(s)) + \dfrac{\alpha}{2s}M(x(s)) = 0, \\ x(s_0) = x_0, \quad \dot{x}(s_0) = x_1. \end{cases} \tag{3.42}$$

Then, the following statements are true:

(i) *If $y : [t_0, +\infty) \to \mathcal{H}$ is a solution trajectory to (3.41) and the function $\tau : [s_0, +\infty) \to [t_0, +\infty)$ is given by*

$$\tau(s) := \frac{\alpha}{2}\ln\left(\frac{s}{s_0}\right) + t_0,$$

then the reparametrized trajectory $x : [s_0, +\infty) \to \mathcal{H}$ given by $x(s) := y(\tau(s))$ is a solution to (3.42) for initial conditions

$$x(s_0) = y_0 \quad \text{and} \quad \dot{x}(s_0) = \frac{\alpha}{2s_0}y_1.$$

(ii) *If $x : [s_0, +\infty) \to \mathcal{H}$ is a solution trajectory to (3.42) and the function $\sigma : [t_0, +\infty) \to [s_0, +\infty)$ is given by*

$$\sigma(t) := s_0 \exp\left(\frac{2(t-t_0)}{\alpha}\right),$$

then the reparametrized trajectory $y : [t_0, +\infty) \to \mathcal{H}$ given by $y(t) := x(\sigma(t))$ is a solution to (3.41) for initial conditions

$$y(t_0) = x_0 \quad \text{and} \quad \dot{y}(t_0) = \frac{2s_0}{\alpha} x_1.$$

3.5.2 Transferring the rates to Fast OGDA

A direct consequence of Theorem 3.4.5 and Proposition 3.5.1 yields the following theorem.

Theorem 3.5.2. *Let $\alpha > 2$, $s_0 > 0$, and $x : [s_0, +\infty) \to \mathcal{H}$ be a solution trajectory of*

$$\begin{cases} \ddot{x}(s) + \dfrac{\alpha}{s}\dot{x}(s) + \dfrac{d}{ds}M(x(s)) + \dfrac{\alpha}{2s}M(x(s)) = 0, \\ x(s_0) = x_0, \quad \dot{x}(s_0) = x_1. \end{cases}$$

and let x^* be a zero of M. Then, it holds

$$\|M(x(s))\| = o\left(\frac{1}{s}\right), \quad \langle x(s) - x^*, M(x(s))\rangle = o\left(\frac{1}{s}\right), \quad \|\dot{x}(s)\| = o\left(\frac{1}{s}\right) \quad \text{as } s \to +\infty.$$

Furthermore, $x(s)$ converges weakly to a zero of M as $s \to +\infty$.

Proof. The proof is near identical to the one in the optimization setting. As per Proposition 3.5.1, define $y(t) := x(\sigma(t))$. We know that $y : [t_0, +\infty) \to \mathcal{H}$ is a solution to (3.41). Again, the function $\sigma \circ \tau : [s_0, +\infty) \to [s_0, +\infty)$ is the identity, which gives $x(s) = y(\tau(s))$ and $\frac{2s}{\alpha}\dot{x}(s) = \dot{y}(\tau(s))$ for every $s \geq s_0$. According to Theorem 3.4.5, if we set $\mu(t) = \gamma(t) = \frac{2s_0}{\alpha}\exp\left(\frac{2(t-t_0)}{\alpha}\right)$ for all $t \geq t_0$, then it holds, as $s \to +\infty$,

$$\|M(x(s))\| = o\left(\frac{1}{\mu(\tau(s))}\right), \quad \langle x(s) - x^*, M(x(s))\rangle = o\left(\frac{1}{\mu(\tau(s))}\right) \quad \text{and} \quad \frac{2s}{\alpha}\|\dot{x}(s)\| \to 0$$

We just need to compute $\mu(\tau(s))$. We readily see

$$\mu(\tau(s)) = \mu\left(\frac{\alpha}{2}\ln\left(\frac{s}{s_0}\right) + t_0\right) = \frac{2s_0}{\alpha}\exp\left(\frac{2}{\alpha}\cdot\frac{\alpha}{2}\ln\left(\frac{s}{s_0}\right)\right) = \frac{2s}{\alpha},$$

which immediately shows the rates claimed in the statement. Since $y(t)$ converges weakly to a zero of M as $t \to +\infty$, so does $x(s) = y(\tau(s))$ as $s \to +\infty$. $\square$

3.6 A second-order system with slow vanishing damping

In this section we mention the main results shown in our paper [42], which bear a connection with the ideas discussed in this chapter. More details and the proof may be found therein.

In the context of linearly constrained convex minimization, for a primal-dual system like the one studied by Zeng, Lei and Chen [102], Attouch, Chbani, Fadili and Riahi [18] considered a vanishing damping term of the form $\frac{\alpha}{t^r} + \frac{r}{t}$, while He, Hu and Fang [66] measured the effects of a term $\frac{\alpha}{t^r}$. In both cases, the choice $r < 1$ allowed the authors to derive exponential rates of convergence for the primal-dual gap. We wanted to see if this influence carries over to a more general setting of a continuous and monotone operator $M : \mathcal{H} \to \mathcal{H}$. For $t \geq t_0 > 0$, we study the asymptotic properties of the solutions to

$$\ddot{z}(t) + \frac{\alpha}{t^r}\dot{z}(t) + \theta t^r \beta(t) \frac{d}{dt} M(z(t)) + \beta(t) M(z(t)) = 0. \tag{3.43}$$

The following are the assumptions on the parameters involved in the system.

Assumption 3.3. $r \in [0,1]$, and $\alpha, \theta > 0$ are taken as follows:

- If $r \in [0,1)$, then $\frac{2}{\alpha} < \theta$;
- If $r = 1$, then $\frac{2}{\alpha+1} \leq \theta < \frac{1}{2}$.

$\beta : [t_0, +\infty) \to (0, +\infty)$ is a continuously differentiable and nondecreasing function which satisfies the following growth condition:

$$\sup_{t \geq t_0} t^r \left(\frac{\dot{\beta}(t)}{\beta(t)} + \frac{2r}{t} \right) < \frac{1}{\theta}. \tag{3.44}$$

Theorem 3.6.1. *Suppose that $\alpha, \theta > 0$, $r \in [0,1]$ and that $\beta : [t_0, +\infty) \to (0, +\infty)$ satisfy Assumption 3.3. Let $t \mapsto z(t)$ be a solution to (3.43) and let z^* be a zero of V. Consider the following convergence rates as $t \to +\infty$:*

$$\|M(z(t))\| = o\left(\frac{1}{t^{\rho+r}\beta(t)}\right), \quad \langle z(t) - z^*, M(z(t))\rangle = o\left(\frac{1}{t^{2\rho}\beta(t)}\right), \quad \|\dot{z}(t)\| = o\left(\frac{1}{t^\rho}\right). \tag{3.45}$$

The following statements are true:

(i) *If $r \in (0,1)$, the above rates hold for $\rho \in (0,r)$. Furthermore, if $t \mapsto z(t)$ is bounded, then these rates hold for $\rho = r$.*

(ii) *If $r = 0$ or $r = 1$, then $t \mapsto z(t)$ is bounded; moreover, the above rates hold for $\rho = 0$ (if $r = 0$) and for $\rho = 1$ (if $r = 1$).*

(iii) *If $r = 0$ or $r = 1$ or $\Big(r \in (0,1)$ and $t \mapsto z(t)$ is bounded$\Big)$, then $z(t)$ converges weakly to a zero of V as $t \to +\infty$.*

Remark 3.6.2. Notice that when $r = 0$, the growth condition given by Assumption 3.3 reads

$$\sup_{t \geq t_0} \frac{\dot{\beta}(t)}{\beta(t)} < \frac{1}{\theta},$$

which is very similar to the growth condition of Assumption 3.1 in the optimization case.

Remark 3.6.3. When $r < 1$, the growth condition in Assumption 3.3 holds if for some $0 < \delta < \frac{1}{\theta}$ the following differential equation is fulfilled by β:

$$t^r \left(\frac{\dot{\beta}(t)}{\beta(t)} + \frac{2r}{t} \right) = \frac{1}{\theta} - \delta \quad \forall t \geq t_0.$$

A solution to this equation is

$$\beta(t) = \frac{1}{t^{2r}} \cdot \exp\left(\left(\frac{1}{\theta} - \delta \right) \frac{t^{1-r}}{1-r} \right),$$

thus convergence rates of $o\left(\exp\left(-\left(\frac{1}{\theta} - \delta\right) \frac{t^{1-r}}{1-r} \right) \right)$ are attained for $\|V(z(t))\|$ and $\langle z(t) - z_*, V(z(t))\rangle$ as $t \to +\infty$.

A discretization of (3.43) yields an implicit algorithm, which has analogous convergence properties to its continuous counterpart. We will only consider $r \in (0, 1]$.

Algorithm 3.1:

Input: $z_0, z_1 \in \mathcal{H}$, $r \in (0,1]$, $\alpha, \theta > 0$, $(\beta_k)_{k \geq 0}$ a sequence of positive numbers;
for $k = 1, 2, \ldots$ **do**

$$z_{k+1} := z_k + \frac{k^r}{\alpha - rk^{r-1} + (k+1)^r}\left(z_k - z_{k-1}\right) - \frac{\theta k^{2r}\beta_{k-1}}{\alpha - rk^{r-1} + (k+1)^r}\Big[M(z_{k+1}) - M(z_k)\Big]$$
$$- \frac{\theta\Big\{\Big[(k+1)^{2r} - k^{2r}\Big] - 2rk^{2r-1}\Big\} + k^r}{\alpha - rk^{r-1} + (k+1)^r}\beta_k M(z_{k+1}).$$

The following are the assumptions for the parameters in the numerical algorithm. They differ from the continuous-time counterpart in that the interval is now smaller for the case $r \in (0, 1)$.

Assumption 3.4. $r \in (0, 1]$, $\alpha, \theta > 0$ are taken as follows:

- If $r \in (0, 1)$, then $\frac{2}{\alpha} < \theta$;
- If $r = 1$, then $\frac{2}{\alpha+1} \leq \theta < \frac{1}{4}$.

The sequence $(\beta_k)_{k \geq 0}$ is positive, nondecreasing, and it satisfies

$$\sup_{k \geq k_0} k^r \left(\frac{\beta_k - \beta_{k-1}}{\beta_k} + \frac{2r}{k}\right) < \frac{1}{2\theta} \tag{3.46}$$

for some positive integer k_0.

Theorem 3.6.4. *Suppose that $\alpha, \theta > 0$, $r \in (0, 1]$ and that $(\beta_k)_{k \geq 0}$ satisfy Assumption 3.4. Let $(z_k)_{k \geq 0}$ be the sequence generated by Algorithm 3.1 and let z^* be a zero of V. Consider the following convergence rates as $k \to +\infty$:*

$$\|M(z_k)\| = o\left(\frac{1}{k^{\rho+r}\beta_k}\right), \quad \langle z_k - z^*, M(z_k)\rangle = o\left(\frac{1}{k^{2\rho}\beta_k}\right) \quad \text{and} \quad \|z_k - z_{k-1}\| = o\left(\frac{1}{k^\rho}\right).$$

The following statements are true:

(i) *If $r \in (0, 1)$, the above rates hold for $\rho \in (0, r)$. Furthermore, if $(z^k)_{k \geq 0}$ is bounded, then these rates hold for $\rho = r$. Moreover, if $r = 1$, then $(z_k)_{k \geq 0}$ is bounded and the above rates hold for $\rho = r = 1$;*

(ii) *If $r \in (0, 1)$ and $(z_k)_{k \geq 0}$ or if $r = 1$, then z_k converges weakly to a zero of V as $k \to +\infty$.*

Remark 3.6.5. Similar to the continuous case, when $r \in (0, 1)$, it is easy to show that for some $0 < \delta < \frac{1}{2\theta}$ the sequence

$$\beta_k = \frac{1}{k^{2r}} \cdot e^{\left(\frac{1}{2\theta} - \delta\right)\frac{k^{1-r}}{1-r}}$$

fulfills the discrete growth condition in Assumption 3.4. Hence, a linear convergence rate of $o\left(e^{-\left(\frac{1}{2\theta} - \delta\right)\frac{k^{1-r}}{1-r}}\right)$ is exhibited by $\|M(z^k)\|$ and $\langle z^k - z_*, M(z^k)\rangle$ as $k \to +\infty$.

Chapter 4

A second-order system attached to an additively structured monotone inclusion

4.1 Introduction

In this chapter, which is essentially our work [41], we study the asymptotic properties of a dynamical system attached to a maximally monotone+cocoercive inclusion problem. Namely, in real a Hilbert space, given a maximally monotone operator $A : \mathcal{H} \to 2^{\mathcal{H}}$ and a β-cocoercive operator $B : \mathcal{H} \to \mathcal{H}$, we address the monotone inclusion problem

$$\text{find } x \in \mathcal{H} \text{ such that } 0 \in (A + B)(x), \tag{4.1}$$

which we naturally assume to have at least one solution. The dynamical system in question reads, for $t \geq t_0 > 0$,

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \xi \left(\frac{d}{dt} T_{\lambda(t),\gamma(t)}(x(t)) \right) + T_{\lambda(t),\gamma(t)}(x(t)) = 0, \tag{4.2}$$

where, for $\lambda, \gamma > 0$, the operator $T_{\lambda,\gamma} : \mathcal{H} \to \mathcal{H}$ is given by

$$T_{\lambda,\gamma} := \frac{1}{\lambda}\Big[\operatorname{Id} - J_{\gamma A}(\operatorname{Id} - \gamma B) \Big].$$

The motivation for studying dynamics governed by this particular operator $T_{\lambda,\gamma}$ comes from two facts: first, it evaluates A and B in a splitting fashion, i.e., though a backward step for A and a forward step for B; second, for $\lambda, \gamma > 0$ the operators $A + B$ and $T_{\lambda,\gamma}$ have the same set of zeros. Indeed,

$$\begin{aligned} 0 = T_{\lambda,\gamma}(x) = \frac{1}{\lambda}\Big[x - J_{\gamma A}(x - \gamma B(x)) \Big] \;\Leftrightarrow\; x = J_{\gamma A}(x - \gamma B(x)) \;\Leftrightarrow\; x - \gamma B(x) \in x + \gamma A(x) \\ \Leftrightarrow\; 0 \in (A + B)(x). \end{aligned}$$

We will show that, as $t \to +\infty$, it holds

$$\|T_{\lambda(t),\gamma(t)}(x(t))\| = o\left(\frac{1}{t^2} \right),$$

as well as the weak convergence of the trajectories towards solutions to (4.1).

This chapter is organized as follows: in Section 4.2, we show the existence and uniqueness of solutions to (4.2), as well as some properties of the operator $T_{\lambda,\gamma}$; in Section 4.3 we show the asymptotic properties (convergence rates, weak convergence) of (4.2); in Section 4.4, we particularize our system to the respective special cases $A = 0$ and $B = 0$, the latter being of particular interesting since it produces (arbitrarily) fast polynomial convergence rates for the norm of the governing operator along the trajectories (only depending on $\gamma(\cdot)$); in Section 4.5, we particularize our system to the convex optimization case, i.e., where $A = \partial f$ and $B = \nabla g$; in Section 4.6, we mention an implicit numerical algorithm that can be obtained from a discretization of (4.2).

Remark 4.1.1. It is sometimes useful to write $T_{\lambda,\gamma}$ as

$$T_{\lambda,\gamma} = \frac{\gamma}{\lambda}\Big[A_\gamma(\operatorname{Id} - \gamma B) + B\Big],$$

where we recall that the Moreau-Yosida envelope of index γ of A is given by

$$A_\gamma = \frac{1}{\gamma}(\operatorname{Id} - J_{\gamma A}).$$

4.2 Existence and uniqueness of trajectories

In this section, we show the existence and uniqueness of strong global solutions to (4.2). For the sake of clarity, first we recall the definition of a strong global solution, which was also given in Chapter 3.

Definition 4.2.1. We call $x : [t_0, +\infty) \to \mathcal{H}$ a *strong global solution* of (4.2) if the following statements hold

(i) $x, \dot{x} : [t_0, +\infty) \to \mathcal{H}$ are locally absolutely continuous, that is, absolutely continuous on each interval $[t_0, t_1]$;

(ii) $\ddot{x}(t) + \frac{\alpha}{t}\dot{x} + \xi\left(\frac{d}{dt}T_{\lambda(t),\gamma(t)}(x(t))\right) + T_{\lambda(t),\gamma(t)}(x(t)) = 0$ for almost every $t \in [t_0, +\infty)$;

(iii) $x(t_0) = x_0$ and $\dot{x}(t_0) = u_0$.

A classic solution is just a strong global solution which is $\mathcal{C}^2$. Sometimes we will mention the terms *strong global solution* or *classic global solution* without explicit mention of the Cauchy data.

The following lemma will be used to prove the existence of strong global solutions of our system, and we will need it in the proof of the main theorem as well.

Lemma 4.2.2. *Let $A : \mathcal{H} \to 2^{\mathcal{H}}$ be a maximally monotone operator and $B : \mathcal{H} \to \mathcal{H}$ a β-cocoercive operator for some $\beta > 0$. Then, the following statements hold:*

(i) *For $\lambda > 0$ and $\gamma \in (0, 2\beta)$, $T_{\lambda,\gamma}$ is a $\lambda\frac{4\beta-\gamma}{4\beta}$-cocoercive operator. In particular, this also implies that $T_{\lambda,\gamma}$ is $\frac{\lambda}{2}$-cocoercive;*

(ii) *Choose $\lambda_1, \lambda_2 > 0$, $\gamma_1, \gamma_2 \in (0, 2\beta)$, $x, y \in \mathcal{H}$ and $x^* \in \operatorname{zer}(A + B)$. Then, it holds*

$$\|\lambda_1 T_{\lambda_1,\gamma_1}(x) - \lambda_2 T_{\lambda_2,\gamma_2}(y)\| \le 4\|x - y\| + \frac{4\beta|\gamma_1 - \gamma_2|}{\gamma_1}\|B(x)\| + \frac{2|\gamma_1 - \gamma_2|}{\gamma_1}\|x - x^*\|,$$

$$\|T_{\lambda_1,\gamma_1}(x) - T_{\lambda_2,\gamma_2}(y)\| \le \frac{1}{\lambda_1}\left[4\|x - y\| + 4\beta\frac{|\gamma_1 - \gamma_2|}{\gamma_1}\|Bx\| + 2\frac{|\gamma_1 - \gamma_2|}{\gamma_1}\|x - x^*\|\right] + 2\frac{|\lambda_2 - \lambda_1|}{\lambda_1\lambda_2}\|y - x^*\|;$$

(iii) *If x is a classic global solution to (4.2) and $x^* \in \operatorname{zer}(A+B)$, then, for every $t \ge t_0$, we have*

$$\left\|\frac{d}{dt}\left(\lambda(t)T_{\lambda(t),\gamma(t)}(x(t))\right)\right\| \le 4\|\dot{x}(t)\| + 4\beta\frac{|\dot{\gamma}(t)|}{\gamma(t)}\|B(x(t))\| + 2\frac{|\dot{\gamma}(t)|}{\gamma(t)}\|x(t) - x^*\|.$$

Proof. (i) From [28, Proposition 26.1] we know that the operator $J_{\gamma A}(\operatorname{Id} - \gamma B)$ is $\alpha = \frac{2\beta}{4\beta-\gamma}$-averaged. From [28, Proposition 4.39], we obtain that $\operatorname{Id} - J_{\gamma A}(\operatorname{Id} - \gamma B)$ is $\frac{1}{2\alpha}$-cocoercive, namely, it is $\frac{4\beta-\gamma}{4\beta}$-cocoercive. Since $\gamma \in (0, 2\beta)$, we have $\frac{4\beta-\gamma}{4\beta} > \frac{2\beta}{4\beta} = \frac{1}{2}$, which implies that $\operatorname{Id} - J_{\gamma A}(\operatorname{Id} - \gamma B)$ is $\frac{1}{2}$-

cocoercive and thus

$$T_{\lambda,\gamma} \text{ is } \lambda\frac{4\beta-\gamma}{4\beta}\text{-cocoercive and } T_{\lambda,\gamma} \text{ is } \frac{\lambda}{2}\text{-cocoercive.}$$

(ii) We have

$$\begin{aligned}
\|\lambda_1 T_{\lambda_1,\gamma_1}(x) - \lambda_2 T_{\lambda_2,\gamma_2}(y)\| &\leq \|x-y\| + \|J_{\gamma_1 A}(x-\gamma_1 B(x)) - J_{\gamma_2 A}(y-\gamma_2 B(y))\| \\
&\leq \|x-y\| + \|J_{\gamma_1 A}(x-\gamma_1 B(x)) - J_{\gamma_2 A}(x-\gamma_1 B(x))\| \\
&\quad + \|J_{\gamma_2 A}(x-\gamma_1 B(x)) - J_{\gamma_2 A}(y-\gamma_2 B(y))\| \\
&\leq 2\|x-y\| + |\gamma_1-\gamma_2|\|A_{\gamma_1}(x-\gamma_1 B(x))\| + \|\gamma_1 B(x) - \gamma_2 B(y)\| \\
&\leq 2\|x-y\| + |\gamma_1-\gamma_2|\|A_{\gamma_1}(x-\gamma_1 B(x))\| \\
&\quad + \|\gamma_1 B(x) - \gamma_2 B(x)\| + \|\gamma_2 B(x) - \gamma_2 B(y)\| \\
&= 2\|x-y\| + |\gamma_1-\gamma_2|\|A_{\gamma_1}(x-\gamma_1 B(x))\| \\
&\quad + |\gamma_1-\gamma_2|\|B(x)\| + \gamma_2\|B(x)-B(y)\|.
\end{aligned}$$

Now, notice that

$$\begin{aligned}
A_{\gamma_1}(x-\gamma_1 B(x)) &= \frac{1}{\gamma_1}(\mathrm{Id} - J_{\gamma_1 A})(x-\gamma_1 B(x)) = -B(x) + \frac{1}{\gamma_1}(x - J_{\gamma_1 A}(x-\gamma_1 B(x))) \\
&= -B(x) + T_{\gamma_1,\gamma_1}(x),
\end{aligned}$$

so using (i) and the fact that $T_{\gamma_1,\gamma_2}(x^*) = 0$, we obtain

$$\begin{aligned}
\|A_{\gamma_1}(x-\gamma_1 B(x))\| = \|-B(x) + T_{\gamma_1,\gamma_2}(x)\| &\leq \|B(x)\| + \|T_{\gamma_1,\gamma_2}(x) - T_{\gamma_1,\gamma_2}(x^*)\| \\
&\leq \|B(x)\| + \frac{2}{\gamma_1}\|x-x^*\|.
\end{aligned} \tag{4.3}$$

Altogether, plugging (4.3) into our initial inequality yields

$$\begin{aligned}
\|\lambda_1 T_{\lambda_1,\gamma_2}(x) - \lambda_2 T_{\lambda_2,\gamma_2}(y)\| &\leq 2\|x-y\| + 2|\gamma_1-\gamma_2|\|B(x)\| + \frac{2|\gamma_1-\gamma_2|}{\gamma_1}\|x-x^*\| + \gamma_2\|B(x)-B(y)\| \\
&\leq 2\|x-y\| + \frac{4\beta|\gamma_1-\gamma_2|}{\gamma_1}\|B(x)\| + \frac{2|\gamma_1-\gamma_2|}{\gamma_1}\|x-x^*\| + 2\beta\left(\frac{1}{\beta}\right)\|x-y\|.
\end{aligned}$$

To show the second inequality, we use the previous one. We have

$$\begin{aligned}
\|T_{\lambda_1,\gamma_1}(x) - T_{\lambda_2,\gamma_2}(y)\| &= \frac{1}{\lambda_1}\|\lambda_1 T_{\lambda_1,\gamma_1}(x) - \lambda_2 T_{\lambda_2,\gamma_2}(y) + (\lambda_2-\lambda_1)T_{\lambda_2,\gamma_2}(y)\| \\
&\leq \frac{1}{\lambda_1}\left[4\|x-y\| + 4\beta\frac{|\gamma_1-\gamma_2|}{\gamma_1}\|Bx\| + 2\frac{|\gamma_1-\gamma_2|}{\gamma_1}\|x-x^*\|\right] + \frac{|\lambda_2-\lambda_1|}{\lambda_1}\|T_{\lambda_2,\gamma_2}(y)\| \\
&\leq \frac{1}{\lambda_1}\left[4\|x-y\| + 4\beta\frac{|\gamma_1-\gamma_2|}{\gamma_1}\|Bx\| + 2\frac{|\gamma_1-\gamma_2|}{\gamma_1}\|x-x^*\|\right] + 2\frac{|\lambda_2-\lambda_1|}{\lambda_1\lambda_2}\|y-x^*\|,
\end{aligned}$$

where the last line is a consequence of T_{λ_2,γ_2} being $\frac{\lambda_2}{2}$-cocoercive, and hence $\frac{2}{\lambda_2}$-Lipschitz continuous (see (i)).

(iii) For $t, s \geq t_0$ set

$$x = x(t),\ \ y = x(s),\ \ \lambda_1 = \lambda(t),\ \ \gamma_1 = \gamma(t),\ \ \lambda_2 = \lambda(s),\ \ \gamma_2 = \gamma(s)$$

and use (ii) to obtain, for every $t \geq t_0$,

$$\begin{aligned}
&\frac{\|\lambda(t)T_{\lambda(t),\gamma(t)}(x(t)) - \lambda(s)T_{\lambda(s),\gamma(s)}(x(s))\|}{|t-s|} \\
&\leq 4\frac{\|x(t)-x(s)\|}{|t-s|} + \frac{4\beta}{\gamma(t)}\frac{|\gamma(t)-\gamma(s)|}{|t-s|}\|B(x(t))\| + \frac{2}{\gamma(t)}\frac{|\gamma(t)-\gamma(s)|}{|t-s|}\|x(t)-x^*\|.
\end{aligned}$$

Hence, by taking the limit as $s \to t$ we get, for any $t \geq t_0$,

$$\left\|\frac{d}{dt}\lambda(t)T_{\lambda(t),\gamma(t)}(x(t))\right\| \leq 4\|\dot{x}(t)\| + 4\beta\frac{|\dot{\gamma}(t)|}{\gamma(t)}\|B(x(t))\| + 2\frac{|\dot{\gamma}(t)|}{\gamma(t)}\|x(t)-x^*\|.$$

□

The next theorem concerns the existence and uniqueness of strong global solutions to (4.2), which only requires mild assumptions on $\lambda(\cdot)$ and $\gamma(\cdot)$.

Theorem 4.2.3. *Assume that $\lambda, \gamma : [t_0, +\infty) \to \mathbb{R}_{++}$ are Lebesgue measurable functions and that $\inf_{t \geq t_0} \lambda(t) > 0$. Then, for any $(x_0, u_0) \in \mathcal{H} \times \mathcal{H}$ there exists a unique strong global solution $x : [t_0, +\infty) \to \mathcal{H}$ of the system (4.2) that satisfies $x(t_0) = x_0$ and $\dot{x}(t_0) = u_0$.*

Proof. We will rely on [62, Proposition 6.2.1] and distinguish between the cases $\xi > 0$ and $\xi = 0$. For each case, we will check that the conditions of the aforementioned proposition are fulfilled. We will be working in the real Hilbert space $\mathcal{H} \times \mathcal{H}$ endowed with the norm $\|(x, y)\| = \|x\| + \|y\|$. Let $x^* \in \operatorname{zer}(A+B)$ be fixed.

The case $\xi > 0$. First, it can be easily checked (see also [4, 22, 26]) that for all $t \geq t_0$ the following dynamical systems are equivalent:

- $\ddot{x}(t) + \dfrac{\alpha}{t}\dot{x}(t) + \xi\left(\dfrac{d}{dt}T_{\lambda(t),\gamma(t)}(x(t))\right) + T_{\lambda(t),\gamma(t)}(x(t)) = 0;$
- $\begin{cases} \dot{x}(t) + \xi T_{\lambda(t),\gamma(t)}(x(t)) - \left(\frac{1}{\xi} - \frac{\alpha}{t}\right)x(t) + \frac{1}{\xi}y(t) = 0, \\ \dot{y}(t) - \left(\frac{1}{\xi} - \frac{\alpha}{t} + \frac{\alpha\xi}{t^2}\right)x(t) + \frac{1}{\xi}y(t) = 0. \end{cases}$

In other words, (4.2) with Cauchy data $(x_0, u_0) = (x(t_0), \dot{x}(t_0))$ is equivalent to the first order system

$$\begin{cases} \dot{z}(t) = F(t, z(t)), \\ z(t_0) = (x_0, y_0), \end{cases}$$

where $z(t) = (x(t), y(t))$, F is given, for every $t \geq t_0$, by

$$F(t, (x, y)) = \left[-\xi T_{\lambda(t),\gamma(t)}(x) + \left(\frac{1}{\xi} - \frac{\alpha}{t}\right)x - \frac{1}{\xi}y,\ \left(\frac{1}{\xi} - \frac{\alpha}{t} + \frac{\alpha\xi}{t^2}\right)x - \frac{1}{\xi}y\right]$$

and the Cauchy data is $x_0 = x(t_0)$, $y_0 = -\xi\left(u_0 + \xi T_{\lambda(t_0),\gamma(t_0)}(x_0) - \left(\frac{1}{\xi} - \frac{\alpha}{t_0}\right)x_0\right)$.

(i) Let $t \in [t_0, +\infty)$ be fixed. We need to verify the Lipschitz continuity of F on the z variable. Set $z = (x, y)$, $w = (u, v)$. We have

$$\begin{aligned} \|F(t, z) - F(t, w)\| = {} & \left\|-\xi\left(T_{\lambda(t),\gamma(t)}(x) - T_{\lambda(t),\gamma(t)}(u) + \left(\frac{1}{\xi} - \frac{\alpha}{t}\right)(x - u) - \frac{1}{\xi}(y - v)\right)\right\| \\ & + \left\|\left(\frac{1}{\xi} - \frac{\alpha}{t} + \frac{\alpha\xi}{t^2}\right)(x - u) - \frac{1}{\xi}(y - v)\right\|. \end{aligned}$$

Set $\underline{\lambda} := \inf_{t \geq t_0} \lambda(t) > 0$. According to Lemma 4.2.2(i), the term involving the operator $T_{\lambda(t),\gamma(t)}$ satisfies

$$\left\|T_{\lambda(t),\gamma(t)}(x) - T_{\lambda(t),\gamma(t)}(u)\right\| \leq \frac{2}{\lambda(t)}\|x - u\| \leq \frac{2}{\underline{\lambda}}\|x - u\|.$$

It follows that, if for $t \geq t_0$ we take

$$K(t) := \max\left\{\frac{2\xi}{\underline{\lambda}} + \left|\frac{1}{\xi} - \frac{\alpha}{t}\right| + \left|\frac{1}{\xi} - \frac{\alpha}{t} + \frac{\alpha\xi}{t^2}\right|,\ \frac{2}{\xi}\right\},$$

then we have $K \in \mathbb{L}^1_{\text{loc}}([t_0, +\infty))$ and, for all $t \geq t_0$,

$$\|F(t, z) - F(t, w)\| \leq K(t)\|z - w\|.$$

(ii) Now, we claim that F fulfills a boundedness condition. For $t \geq t_0$ and $z = (x, y) \in \mathcal{H} \times \mathcal{H}$ we have

$$\|F(t, z)\| = \left\|-\xi T_{\lambda(t),\gamma(t)}(x) + \left(\frac{1}{\xi} - \frac{\alpha}{t}\right)x - \frac{1}{\xi}y\right\| + \left\|\left(\frac{1}{\xi} - \frac{\alpha}{t} + \frac{\alpha\xi}{t^2}\right)x - \frac{1}{\xi}y\right\|.$$

By Lemma 4.2.2(i), we have, for every $t \geq t_0$,

$$\left\|T_{\lambda(t),\gamma(t)}(x)\right\| = \left\|T_{\lambda(t),\gamma(t)}(x) - T_{\lambda(t),\gamma(t)}(x^*)\right\| \leq \frac{2}{\lambda(t)}\|x - x^*\|.$$

Hence, if for $t \geq t_0$ we take

$$P(t) = \max\left\{\frac{2\xi}{\lambda(t)} + \left|\frac{1}{\xi} - \frac{\alpha}{t}\right| + \left|\frac{1}{\xi} - \frac{\alpha}{t} + \frac{\alpha\xi}{t^2}\right|, \frac{2\xi}{\lambda(t)}, \frac{2}{\xi}\right\},$$

then we have $P \in \mathbb{L}^1_{\text{loc}}([t_0, +\infty))$ and

$$\|F(t,z)\| \leq P(t)(1 + \|z\|).$$

We have checked that the conditions of [62, Proposition 6.2.1] hold. Therefore, there exists a unique locally absolutely continuous solution $t \mapsto x(t)$ of (4.2) that satisfies $x(t_0) = x_0$ and $\dot{x}(t_0) = u_0$.

The case $\xi = 0$. Now, (4.2) is easily seen to be equivalent to

$$\begin{cases} \dot{z}(t) = F(t, z(t)), \\ z(t_0) = (x_0, u_0), \end{cases}$$

where $z(t) = (x(t), y(t))$ and F is given, for every $t \geq t_0$, by

$$F(t, (x,y)) = \left[y, \ -\frac{\alpha}{t}y - T_{\lambda(t),\gamma(t)}(x)\right].$$

Showing that F fulfills the required properties is straightforward. □

4.3 The convergence properties of the trajectories

In this section, we will study the asymptotic behavior of our system (4.2). We will show convergence rates for the norm of the operator along the trajectories, weak convergence of the trajectories towards elements of $\text{zer}(A + B)$, as well as the fast convergence of the velocities and accelerations to zero. Additionally, we will provide convergence rates for $\|T_{\lambda(t),\gamma(t)}(x(t))\|$ and $\left\|\frac{d}{dt}T_{\lambda(t),\gamma(t)}(x(t))\right\|$ as $t \to +\infty$. To avoid repetition of the statement "for almost every t", in the following theorem we will assume we are working with a classic global solution of our system. In order to show the convergence statements, we need a slighter stronger assumption on $\lambda(\cdot)$ and $\gamma(\cdot)$.

Assumption 4.1. Suppose that $\alpha > 1$, $\xi \geq 0$, and that $\gamma : [t_0, +\infty) \to (0, 2\beta)$ is a differentiable function. Moreover, assume that as $t \to +\infty$ it holds

$$\lambda(t) = \lambda t^2, \text{ where } \lambda > \frac{2}{(\alpha-1)^2}, \quad \text{and} \quad \frac{|\dot{\gamma}(t)|}{\gamma(t)} = \mathcal{O}\left(\frac{1}{t}\right).$$

Theorem 4.3.1. *Suppose that the parameters $\alpha, \xi, \lambda(\cdot)$ and $\gamma(\cdot)$ satisfy Assumption 4.1. Then, for a solution $x : [t_0, +\infty) \to \mathcal{H}$ to (4.2), the following are true.*

(i) *The following convergence rates hold, as $t \to +\infty$:*

$$\|\dot{x}(t)\| = o\left(\frac{1}{t^2}\right), \quad \|\ddot{x}(t)\| = \mathcal{O}\left(\frac{1}{t^2}\right) \quad \text{and} \quad \|T_{\lambda(t),\gamma(t)}(x(t))\| = o\left(\frac{1}{t^2}\right);$$

(ii) *If $0 < \inf_{t \geq t_0}\gamma(t) \leq \sup_{t \geq t_0}\gamma(t) < 2\beta$, then $x(t)$ converges weakly to an element of $\text{zer}(A + B)$ as $t \to +\infty$.*

Proof. **Integral estimates and rates.** To develop the analysis, we will fix $x^* \in \text{zer}(A+B)$ and make of use of the Lyapunov function $\mathcal{E} : [t_0, +\infty) \to \mathbb{R}_+$ given by

$$\mathcal{E}(t) := \frac{1}{2}\left\|\frac{\alpha-1}{2}(x(t) - x^*) + t(\dot{x}(t) + \xi\, T_{\lambda(t),\gamma(t)}(x(t)))\right\|^2 + \frac{(\alpha-1)^2}{8}\|x(t) - x^*\|^2. \tag{4.4}$$

Differentiation of $\mathcal{E}$ with respect to time yields, for every $t \geq t_0$,

$$\dot{\mathcal{E}}(t) = \Bigg\langle \frac{\alpha-1}{2}(x(t) - x^*) + t\left(\dot{x}(t) + \xi\, T_{\lambda(t),\lambda(t)}(x(t))\right),$$
$$\frac{\alpha+1}{2}\dot{x}(t) + \xi\, T_{\lambda(t),\gamma(t)}(x(t)) + t\left(\ddot{x}(t) + \xi\frac{d}{dt}\left(T_{\lambda(t),\gamma(t)}(x(t))\right)\right)\Bigg\rangle + \frac{(\alpha-1)^2}{4}\langle x(t) - x^*, \dot{x}(t)\rangle.$$

After reduction and employing (4.2), we get, for every $t \geq t_0$,

$$\dot{\mathcal{E}}(t) = \frac{(\alpha-1)(\xi-t)}{2}\left\langle x(t)-x^*, T_{\lambda(t),\gamma(t)}(x(t))\right\rangle + \frac{(1-\alpha)t}{2}\left\|\dot{x}(t)\right\|^2 + \left(-t^2+\frac{\xi(3-\alpha)t}{2}\right)\left\langle T_{\lambda(t),\gamma(t)}(x(t)), \dot{x}(t)\right\rangle + \xi(\xi-t)t\left\|T_{\lambda(t),\gamma(t)}(x(t))\right\|^2.$$

Now, by Lemma 4.2.2(i), we know that $T_{\lambda(t),\gamma(t)}$ is $\frac{\lambda(t)}{2}$-cocoercive for every $t \geq t_0$. Using this on the first summand of the right hand side of the previous inequality yields, for $t \geq t_1 = \max\{\xi, t_0\}$,

$$\dot{\mathcal{E}}(t) \leq \frac{(1-\alpha)t}{2}\|\dot{x}(t)\|^2 + \left(-t^2+\frac{\xi(3-\alpha)t}{2}\right)\left\langle T_{\lambda(t),\gamma(t)}(x(t)), \dot{x}(t)\right\rangle + \left(\frac{(\alpha-1)(\xi-t)\lambda(t)}{4}+\xi(\xi-t)t\right)\left\|T_{\lambda(t),\gamma(t)}(x(t))\right\|^2. \tag{4.5}$$

Now, since $\lambda > \frac{2}{(\alpha-1)^2}$, we can choose $\epsilon > 0$ such that

$$0 < \epsilon < \alpha - 1 - \sqrt{\frac{2}{\lambda}} < \alpha - 1. \tag{4.6}$$

From (4.5) we get, for every $t \geq t_1$,

$$\begin{aligned}\dot{\mathcal{E}}(t) &+ \frac{\epsilon}{2}t\|\dot{x}(t)\|^2 + \frac{\epsilon}{4}t\lambda(t)\left\|T_{\lambda(t),\gamma(t)}(x(t))\right\|^2 \\ &\leq \left(\frac{1-\alpha}{2}+\frac{\epsilon}{2}\right)t\|\dot{x}(t)\|^2 + \left(-t^2+\frac{\xi(3-\alpha)t}{2}\right)\left\langle T_{\lambda(t),\gamma(t)}(x(t)), \dot{x}(t)\right\rangle \\ &+ \left(\left(\frac{(\alpha-1)(\xi-t)}{2}+\frac{\epsilon}{2}t\right)\frac{\lambda(t)}{2}+\xi(\xi-t)t\right)\left\|T_{\lambda(t),\gamma(t)}(x(t))\right\|^2.\end{aligned} \tag{4.7}$$

By (4.6) and the definition of $\lambda(t)$, we know that $\frac{1-\alpha}{2}+\frac{\epsilon}{2} < 0$, and that as $t \to +\infty$

$$\left(\left(\frac{(\alpha-1)(\xi-t)}{2}+\frac{\epsilon}{2}t\right)\frac{\lambda(t)}{2}+\xi(\xi-t)t\right) = \underbrace{\left(\frac{1-\alpha}{2}+\frac{\epsilon}{2}\right)}_{<0}\frac{\lambda}{2}t^3 + \mathcal{O}(t^2),$$

so we can find $t_2 \geq t_1$ such that for every $t \geq t_2$ the previous expression becomes nonpositive. According to Lemma 2.3.5, the right hand side of (4.7) is nonpositive whenever

$$R(t) := \left(-t^2+\frac{\xi(3-\alpha)t}{2}\right)^2 - 4\left(\frac{1-\alpha}{2}+\frac{\epsilon}{2}\right)t\left(\left(\frac{(\alpha-1)(\xi-t)}{2}+\frac{\epsilon}{2}t\right)\frac{\lambda(t)}{2}+\xi(\xi-t)t\right) \leq 0.$$

Rewriting, we have that as $t \to +\infty$

$$R(t) = \left(1+4\left(\frac{1-\alpha}{2}+\frac{\epsilon}{2}\right)\left(\frac{\alpha-1}{2}-\frac{\epsilon}{2}\right)\frac{\lambda}{2}\right)t^4 + \mathcal{O}(t^3).$$

Since $\epsilon < \alpha - 1 - \sqrt{\frac{2}{\lambda}}$, we have $\frac{\lambda}{2} > \frac{1}{(\alpha-1-\epsilon)^2}$. Hence,

$$1+4\left(\frac{1-\alpha}{2}+\frac{\epsilon}{2}\right)\left(\frac{\alpha-1}{2}-\frac{\epsilon}{2}\right)\frac{\lambda}{2} = 1-(\alpha-1-\epsilon)^2\frac{\lambda}{2} < 0.$$

This means we can find $t_3 \geq t_2$ such that for every $t \geq t_3$ we have $R(t) \leq 0$, that is, for every $t \geq t_3$ we have

$$\dot{\mathcal{E}}(t) + \frac{\epsilon}{2}t\|\dot{x}(t)\|^2 + \frac{\epsilon}{4}t\lambda(t)\left\|T_{\lambda(t),\gamma(t)}(x(t))\right\|^2 \leq 0. \tag{4.8}$$

Now, integrating (4.8) from t_3 to t we obtain

$$\mathcal{E}(t) + \frac{\epsilon}{2}\int_{t_3}^{t} s\|\dot{x}(s)\|^2 ds + \frac{\epsilon}{4}\lambda\int_{t_3}^{t} s^3\left\|T_{\lambda(s),\gamma(s)}(x(s))\right\|^2 ds \leq \mathcal{E}(t_3). \tag{4.9}$$

From (4.8) and the form of $\mathcal{E}$ we immediately obtain

$$t \mapsto \|x(t) - x^*\| \text{ is bounded,} \tag{4.10}$$

$$\int_{t_0}^{+\infty} t\|\dot{x}(t)\|^2 dt < +\infty, \tag{4.11}$$

$$\int_{t_0}^{+\infty} t^3 \left\|T_{\lambda(t),\gamma(t)}(x(t))\right\|^2 dt < +\infty, \tag{4.12}$$

$$\sup_{t\geq t_0} \left\| \left(\frac{\alpha-1}{2}\right)(x(t) - x^*) + t\left(\dot{x}(t) + \xi\, T_{\lambda(t),\gamma(t)}(x(t))\right) \right\| < +\infty. \tag{4.13}$$

From Lemma 4.2.2(i), we know that for every $t \geq t_0$ the operator $T_{\lambda(t),\gamma(t)}$ is $\frac{2}{\lambda(t)}$-Lipschitz continuous, which gives, for every $t \geq t_0$,

$$\left\|T_{\lambda(t),\gamma(t)}(x(t))\right\| = \left\|T_{\lambda(t),\gamma(t)}(x(t)) - T_{\lambda(t),\gamma(t)}(x^*)\right\| \leq \frac{2}{\lambda(t)}\|x(t) - x^*\|.$$

Thus, from (4.10) and recalling that $\lambda(t) = \lambda t^2$ we deduce that, as $t \to +\infty$,

$$\left\|T_{\lambda(t),\gamma(t)}(x(t))\right\| = \mathcal{O}\left(\frac{1}{t^2}\right). \tag{4.14}$$

By combining (4.10), (4.13) and (4.14) we obtain $\sup_{t\geq t_0} t\|\dot{x}(t)\| < +\infty$ and therefore, as $t \to +\infty$,

$$\|\dot{x}(t)\| = \mathcal{O}\left(\frac{1}{t}\right). \tag{4.15}$$

From Lemma 4.2.2, (4.10), (4.15) and the fact that B is $\frac{1}{\beta}$-Lipschitz continuous we deduce that, as $t \to +\infty$,

$$\left\|\frac{d}{dt}\lambda(t)T_{\lambda(t),\gamma(t)}(x(t))\right\| \leq 4\|\dot{x}(t)\| + 4\beta\frac{|\dot{\gamma}(t)|}{\gamma(t)}\|B(x(t))\| + 2\frac{|\dot{\gamma}(t)|}{\gamma(t)}\|x(t) - x^*\| = \mathcal{O}\left(\frac{1}{t}\right). \tag{4.16}$$

On the other hand, for every $t \geq t_0$ we have

$$\left\|\frac{d}{dt}\lambda(t)T_{\lambda(t),\gamma(t)}(x(t))\right\| = \left\|\dot{\lambda}(t)T_{\lambda(t),\gamma(t)}(x(t)) + \lambda(t)\frac{d}{dt}T_{\lambda(t),\gamma(t)}(x(t))\right\|, \tag{4.17}$$

so by combining (4.14), (4.16), (4.17) and the fact that $\dot{\lambda}(t) = 2\lambda t$ we arrive at

$$\left\|\lambda(t)\frac{d}{dt}T_{\lambda(t),\gamma(t)}(x(t))\right\| \leq \underbrace{\left\|\frac{d}{dt}\lambda(t)T_{\lambda(t),\gamma(t)}(x(t))\right\|}_{\mathcal{O}(\frac{1}{t})} + \dot{\lambda}(t)\underbrace{\left\|T_{\lambda(t),\gamma(t)}(x(t))\right\|}_{\mathcal{O}(\frac{1}{t^2})} = \mathcal{O}\left(\frac{1}{t}\right)$$

as $t \to +\infty$, which yields

$$\left\|\frac{d}{dt}T_{\lambda(t),\gamma(t)}(x(t))\right\| = \frac{1}{\lambda(t)}\mathcal{O}\left(\frac{1}{t}\right) = \mathcal{O}\left(\frac{1}{t^3}\right). \tag{4.18}$$

as $t \to +\infty$. Let us now improve (4.14) and show that, as $t \to +\infty$,

$$\left\|T_{\lambda(t),\gamma(t)}(x(t))\right\| = o\left(\frac{1}{t^2}\right). \tag{4.19}$$

According to (4.14) and (4.16) there exists a constant $K > 0$ such that for every $t \geq t_0$ it holds

$$\begin{aligned}\left|\frac{d}{dt}\left\|\lambda(t)T_{\lambda(t),\gamma(t)}(x(t))\right\|^4\right| &= \left|4\left\|\lambda(t)T_{\lambda(t),\gamma(t)}(x(t))\right\|^2\left\langle \lambda(t)T_{\lambda(t),\gamma(t)}(x(t)), \frac{d}{dt}\lambda(t)T_{\lambda(t),\gamma(t)}(x(t))\right\rangle\right| \\ &\leq 4\left\|\lambda(t)T_{\lambda(t),\gamma(t)}(x(t))\right\|^2\left\|\lambda(t)T_{\lambda(t),\gamma(t)}(x(t))\right\|\left\|\frac{d}{dt}\lambda(t)T_{\lambda(t),\gamma(t)}(x(t))\right\|\end{aligned}$$

$$\leq \frac{4K}{t} \left\| \lambda(t) T_{\lambda(t),\gamma(t)}(x(t)) \right\|^2.$$

By (4.12), the right hand side belongs to $\mathbb{L}^1([t_0, +\infty))$, so we get

$$\frac{d}{dt} \left\| \lambda(t) T_{\lambda(t),\gamma(t)}(x(t)) \right\|^4 \in \mathbb{L}^1([t_0, +\infty)),$$

hence the limit (cf. Lemma 2.3.1)

$$\lim_{t \to +\infty} \left\| \lambda(t) T_{\lambda(t),\gamma(t)}(x(t)) \right\|^4$$

exists. Obviously, this implies the existence of $L := \lim_{t \to +\infty} \|\lambda(t) T_{\lambda(t),\gamma(t)}(x(t))\|^2$. By using (4.12) again we come to

$$\int_{t_0}^{+\infty} \frac{1}{t} \left\| \lambda(t) T_{\lambda(t),\gamma(t)}(x(t)) \right\|^2 dt = \lambda^2 \int_{t_0}^{+\infty} t^3 \left\| T_{\lambda(t),\gamma(t)}(x(t)) \right\|^2 dt < +\infty,$$

and so we must have $L = 0$, which gives, as $t \to +\infty$,

$$\left\| T_{\lambda(t),\gamma(t)}(x(t)) \right\| = o\left(\frac{1}{t^2} \right). \tag{4.20}$$

By combining (4.2), (4.14), (4.15) and (4.18) we obtain, as $t \to +\infty$,

$$\begin{aligned} \|\ddot{x}(t)\| &= \left\| -\frac{\alpha}{t} \dot{x}(t) - \xi \frac{d}{dt} T_{\lambda(t),\gamma(t)}(x(t)) - T_{\lambda(t),\gamma(t)}(x(t)) \right\| \\ &\leq \frac{\alpha}{t} \underbrace{\|\dot{x}(t)\|}_{\mathcal{O}\left(\frac{1}{t}\right)} + \xi \underbrace{\left\| \frac{d}{dt} T_{\lambda(t),\gamma(t)}(x(t)) \right\|}_{\mathcal{O}\left(\frac{1}{t^3}\right)} + \underbrace{\left\| T_{\lambda(t),\gamma(t)}(x(t)) \right\|}_{\mathcal{O}\left(\frac{1}{t^2}\right)} = \mathcal{O}\left(\frac{1}{t^2} \right). \end{aligned}$$

Moreover, by using the well-known inequality $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$ for every $a, b, c \in \mathcal{H}$, for every $t \geq t_0$ it holds

$$\begin{aligned} t^3 \|\ddot{x}(t)\|^2 &\leq t^3 \left\| -\frac{\alpha}{t} \dot{x}(t) - \xi \frac{d}{dt} T_{\lambda(t),\gamma(t)}(x(t)) - T_{\lambda(t),\gamma(t)}(x(t)) \right\|^2 \\ &\leq 3\alpha t \|\dot{x}(t)\|^2 + 3\xi^2 t^3 \left\| \frac{d}{dt} T_{\lambda(t),\gamma(t)}(x(t)) \right\|^2 + 3t^3 \left\| T_{\lambda(t),\gamma(t)}(x(t)) \right\|^2. \end{aligned}$$

From (4.11), (4.18) and (4.12) it follows

$$\int_{t_0}^{+\infty} t^3 \|\ddot{x}(t)\|^2 dt < +\infty. \tag{4.21}$$

To see that $\|\dot{x}(t)\| = o\left(\frac{1}{t}\right)$ as $t \to +\infty$, we write, for every $t \geq t_0$,

$$\frac{d}{dt} \left(t^2 \|\dot{x}(t)\|^2 \right) = 2t \|\dot{x}(t)\|^2 + 2t^2 \langle \dot{x}(t), \ddot{x}(t) \rangle \leq 3t \|\dot{x}(t)\|^2 + t^3 \|\ddot{x}(t)\|^2.$$

From (4.11) and (4.21) we deduce that the left hand side belongs to $\mathbb{L}^1([t_0, +\infty), \mathbb{R})$, from which we infer that limit $\lim_{t \to +\infty} t^2 \|\dot{x}(t)\|^2$ exists. Using (4.11) again, we get

$$\int_{t_0}^{+\infty} \frac{1}{t} \left(t^2 \|\dot{x}(t)\|^2 \right) dt = \int_{t_0}^{+\infty} t \|\dot{x}(t)\|^2 dt < +\infty,$$

from which we finally deduce $\lim_{t \to +\infty} t^2 \|\dot{x}(t)\|^2 = 0$, therefore

$$\|\dot{x}(t)\| = o\left(\frac{1}{t} \right). \tag{4.22}$$

as $t \to +\infty$. The fact that $\|\ddot{x}(t)\| = \mathcal{O}\left(\frac{1}{t^2}\right)$ as $t \to +\infty$ comes from (4.2), (4.22), (4.18) and (4.19).

Weak convergence of the trajectories. We will work with the energy function $h : [t_0, +\infty) \to \mathbb{R}$ given by

$$h(t) := \frac{1}{2}\|x(t) - x^*\|^2.$$

For every $t \geq t_0$, we have

$$\dot{h}(t) = \langle x(t) - x^*, \dot{x}(t)\rangle, \quad \ddot{h}(t) = \langle x(t) - x^*, \ddot{x}(t)\rangle + \|\dot{x}(t)\|^2. \tag{4.23}$$

Combining (4.2) and (4.23) gives us, for every $t \geq t_0$,

$$\ddot{h}(t) + \frac{\alpha}{t}\dot{h}(t) + \langle T_{\lambda(t),\gamma(t)}(x(t)), x(t) - x^*\rangle = \|\dot{x}(t)\|^2 + \left\langle -\xi\frac{d}{dt}T_{\lambda(t),\gamma(t)}(x(t)), x(t) - x^*\right\rangle.$$

By using the $\frac{\lambda(t)}{2}$-cocoercivity of $T_{\lambda(t),\gamma(t)}$ on the left hand side, Cauchy-Schwarz on the right hand side and multiplying both sides by t, the previous inequality entails, for every $t \geq t_0$,

$$t\ddot{h}(t) + \alpha\dot{h}(t) + t\frac{\lambda(t)}{2}\left\|T_{\lambda(t),\gamma(t)}(x(t))\right\| \leq t\|\dot{x}(t)\|^2 + \xi t\left\|\frac{d}{dt}T_{\lambda(t),\gamma(t)}(x(t))\right\|\|x(t) - x^*\|.$$

Now, putting (4.10), (4.11) and (4.18) together results in

$$k(t) := t\|\dot{x}(t)\|^2 + \xi t\left\|\frac{d}{dt}T_{\lambda(t),\gamma(t)}(x(t))\right\|\|x(t) - x^*\| \in \mathbb{L}^1([t_0, +\infty)).$$

Now apply Lemma 2.3.6 with $\theta(t) := t\frac{\lambda(t)}{2}\left\|T_{\lambda(t),\gamma(t)}(x(t))\right\|$ for every $t \geq t_0$ to deduce that the limit

$$\lim_{t\to+\infty} h(t)$$

exists, which fulfills the first condition of Opial's Lemma 2.1.2.

Let us now move on to the second condition. Suppose $\overline{x}$ is a weak sequential cluster point of $t \mapsto x(t)$, that is, there exists a sequence $(t_n)_{n\in\mathbb{N}} \subseteq [t_0, +\infty)$ such that $t_n \to +\infty$ and $x_n := x(t_n)$ converges weakly to $\overline{x}$ as $n \to +\infty$. Define

$$U_\gamma := \operatorname{Id} - J_{\gamma A}(\operatorname{Id} - \gamma B).$$

According to (4.20), we have $U_{\gamma(t)}(x(t)) = \lambda(t)T_{\lambda(t),\gamma(t)}(x(t)) \to 0$ as $t \to +\infty$. Now, since $\gamma(t) \in [\delta, 2\beta - \delta]$ for all $t \geq t_0$ for some $\delta > 0$, we can extract a subsequence $(\gamma(t_{n_k}))_{k\in\mathbb{N}}$ such that $\gamma(t_{n_k}) \to \overline{\gamma} \in (0, 2\beta)$ as $k \to +\infty$. We may assume without loss of generality then that $\gamma_n := \gamma(t_n) \to \overline{\gamma}$ as $n \to +\infty$. We now have for every $n \in \mathbb{N}$

$$\begin{aligned}
\|U_{\gamma_n}(x_n) - U_{\overline{\gamma}}(x_n)\| &= \|J_{\gamma_n A}(x_n - \gamma_n B(x_n)) - J_{\overline{\gamma}A}(x_n - \overline{\gamma}B(x_n))\| \\
&= \|J_{\gamma_n A}(x_n - \gamma_n B(x_n)) - J_{\gamma_n A}(x_n - \overline{\gamma}B(x_n))\| \\
&\quad + \|J_{\gamma_n A}(x_n - \overline{\gamma}B(x_n)) - J_{\overline{\gamma}A}(x_n - \overline{\gamma}B(x_n))\| \\
&\leq |\overline{\gamma} - \gamma_n|\|B(x_n)\| + |\overline{\gamma} - \gamma_n|\|A_{\overline{\gamma}}(x_n - \overline{\gamma}B(x_n))\|.
\end{aligned}$$

Now, since every weakly convergent sequence is bounded and the operators B and $A_{\overline{\gamma}}$ are Lipschitz-continuous we deduce that the right-hand side of the previous inequality approaches zero as $n \to +\infty$, therefore getting

$$U_{\overline{\gamma}}(x_n) = U_{\gamma_n}(x_n) + \big(U_{\overline{\gamma}}(x_n) - U_{\gamma_n}(x_n)\big) \to 0$$

as $n \to +\infty$. Now, from the proof of part (i) of Lemma 4.2.2, we know that $U_{\overline{\gamma}}$ is $\frac{4\beta-\overline{\gamma}}{4\beta}$-cocoercive, thus monotone and Lipschitz continuous and therefore maximally monotone. Summarizing, we have

(a) $U_{\overline{\gamma}}$ is maximally monotone and thus its graph is closed in $\mathcal{H}^{\text{weak}} \times \mathcal{H}^{\text{strong}}$ (see [28, Proposition 20.38(ii)]);

(b) x_n converges weakly to $\overline{x}$ and $U_{\overline{\gamma}}(x_n) \to 0$ as $n \to +\infty$,

which allows us to conclude that $U_{\overline{\gamma}}(\overline{x}) = 0$, and gives finally $\overline{x} \in \operatorname{zer}(A + B)$. Now we just invoke Opial's Lemma to conclude that $x(t)$ converges weakly to an element of $\operatorname{zer}(A + B)$ as $t \to +\infty$. $\square$

4.4 Particularizing to the respective special cases $A = 0$ and $B = 0$

In the following subsections, we explore the particular cases $B = 0$ and $A = 0$, and we will show improvements with respect to previous results from the literature addressing continuous time approaches to monotone inclusions.

4.4.1 The case $B = 0$

If we let $B = 0$ in (4.2), then, attached to the monotone inclusion problem

$$\text{find } x \in \mathcal{H} \text{ such that } 0 \in A(x),$$

we obtain the dynamics

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \xi \frac{d}{dt}\left(A_{\lambda(t),\gamma(t)}(x(t))\right) + A_{\lambda(t),\gamma(t)}(x(t)) = 0, \tag{4.24}$$

where

$$A_{\lambda,\gamma}(x) := \frac{1}{\lambda}(\operatorname{Id} - J_{\gamma A}).$$

We can weaken Assumption 4.1 in this case. Namely, $\gamma(\cdot)$ need not be upper bounded.

Assumption 4.2. Suppose that $\alpha > 1$, $\xi \geq 0$, and that $\gamma : [t_0, +\infty) \to \mathbb{R}_{++}$ is a differentiable function. Moreover, assume that as $t \to +\infty$ it holds

$$\lambda(t) = \lambda t^2, \text{ where } \lambda > \frac{1}{(\alpha - 1)^2}, \quad \text{and} \quad \frac{\left|\dot{\gamma}(t)\right|}{\gamma(t)} = \mathcal{O}\left(\frac{1}{t}\right).$$

Theorem 4.4.1. *Suppose that the parameters $\alpha, \xi, \lambda(\cdot)$ and $\gamma(\cdot)$ satisfy Assumption 4.2. Then, for a solution $x : [t_0, +\infty) \to \mathcal{H}$ to (4.24), the following are true.*

(i) *The following convergence rates hold, as $t \to +\infty$:*

$$\left\|\dot{x}(t)\right\| = o\left(\frac{1}{t}\right), \quad \left\|\ddot{x}(t)\right\| = \mathcal{O}\left(\frac{1}{t^2}\right) \quad \text{and} \quad \left\|A_{\gamma(t)}(x(t))\right\| = o\left(\frac{1}{\gamma(t)}\right);$$

(ii) *If $0 < \inf_{t \geq t_0} \gamma(t)$, then $x(t)$ converges weakly to an element of $\operatorname{zer} A$ as $t \to +\infty$.*

Proof. We now have $T_{\lambda,\gamma} = \frac{1}{\lambda}(\operatorname{Id} - J_{\gamma A}) = A_{\lambda,\gamma}$. Since $J_{\lambda A}$ is firmly nonexpansive, by [28, Proposition 4.4] so is $\operatorname{Id} - J_{\lambda A}$. In other words, $\operatorname{Id} - J_{\gamma A}$ is 1-cocoercive, therefore $A_{\lambda,\gamma} = \frac{1}{\lambda}(\operatorname{Id} - J_{\gamma A})$ is λ-cocoercive, so now the condition on λ becomes $\lambda > \frac{1}{(\alpha-1)^2}$. Intuitively, the operator $B = 0$ is β-cocoercive for every $\beta > 0$, so no matter how large $\gamma(t)$ is, the operator $T_{\lambda(t),\gamma(t)} = A_{\lambda(t),\gamma(t)}$ will always be $\lambda(t)\frac{4\beta - \gamma(t)}{4\beta}$-cocoercive (cf. Lemma 4.2.2) for β arbitrarily large, which means it is actually $\lambda(t)$-cocoercive. For deriving the convergence rates, we proceed exactly as in the proof for Theorem 4.3.1, using the fact that $A_{\lambda(t),\gamma(t)}$ is $\lambda(t)$-cocoercive, and the fact that

$$\left\|\frac{\gamma(t)}{\lambda(t)} A_{\gamma(t)}(x(t))\right\| = \left\|\frac{1}{\gamma(t)} \cdot \frac{\gamma(t)}{\lambda(t)}\Big(x(t) - J_{\lambda(t)A}(x(t))\Big)\right\| = \left\|T_{\lambda(t),\gamma(t)}(x(t))\right\| = o\left(\frac{1}{t^2}\right)$$

as $t \to +\infty$ implies that

$$\left\|A_{\gamma(t)}(x(t))\right\| = o\left(\frac{1}{\gamma(t)}\right).$$

as $t \to +\infty$.

The proof also changes when we verify the second part of the Opial's Lemma, to get weak convergence of the trajectories $t \mapsto x(t)$. This is in order to allow for $\gamma(\cdot)$ not to be necessarily bounded. We do need, however, the assumption $0 < \inf_{t \geq t_0} \gamma(t)$. Indeed, from $\|A_{\lambda(t),\gamma(t)}(x(t))\| = o\left(\frac{1}{t^2}\right)$ as $t \to +\infty$, we obtain

$$y(t) := x(t) - J_{\gamma(t)A}x(t) = \lambda(t)A_{\lambda(t),\gamma(t)}(x(t)) \to 0$$

as $t \to +\infty$. Using the definition of the resolvent, we come to

$$J_{\gamma(t)A}x(t) = x(t) - y(t) \Leftrightarrow y(t) \in \gamma(t)A(x(t) - y(t)) \Leftrightarrow \frac{1}{\gamma(t)}y(t) \in A(x(t) - y(t)).$$

for all $t \geq t_0$. If $(t_n)_{n\in\mathbb{N}} \subseteq [t_0, +\infty)$ is such that $t_n \to +\infty$ and $x(t_n)$ converges weakly to $\overline{x}$ as $n \to +\infty$, then the previous inclusion, together with the assumption on γ gives

$$x(t_n) - y(t_n) \text{ converges weakly to } \overline{x} \quad \text{and} \quad \frac{1}{\gamma(t)}y(t) \to 0$$

as $n \to +\infty$, and because of the closedness of the graph of A in $\mathcal{H}^{\text{weak}} \times \mathcal{H}^{\text{strong}}$, we deduce that $\overline{x} \in \operatorname{zer} A$. □

Remark 4.4.2. If $\gamma(\cdot)$ is a polynomial of degree n for some $n \in \mathbb{N}$, the conditions of Assumption 4.2 are fulfilled. Indeed, assume $\gamma(t) = a_n t^n + a_{n-1}t^{n-1} + \cdots + a_0$ for all $t \geq t_0$, for some $a_i \in \mathbb{R}$ for $i \in \{0, \ldots, n\}$ and $a_n > 0$. Then, we have

$$\begin{aligned} t \cdot \frac{\dot{\gamma}(t)}{\gamma(t)} &= t \cdot \frac{na_n t^{n-1} + (n-1)a_{n-1}t^{n-1} + \cdots + a_1}{a_n t^n + a_{n-1}t^{n-1} + \cdots + a_0} \\ &\to \frac{na_n}{a_n} = n \end{aligned}$$

as $t \to +\infty$, so $\frac{\dot{\gamma}(t)}{\gamma(t)} = \mathcal{O}\left(\frac{1}{t}\right)$ as $t \to +\infty$. Since we also have $\gamma(t) \to +\infty$ as $t \to +\infty$, the condition $\inf_{t\geq t_0} \gamma(t) > 0$ is fulfilled for large enough t_0.
In particular, we can choose $\gamma(t) = \lambda(t) = \lambda t^2$, which fulfills $\gamma(t) \geq \lambda t_0^2 > 0$ for any $t \geq t_0$ and any t_0. Since $A_{\lambda,\lambda} = A_\lambda$ for $\lambda > 0$, this choice of γ allows us to recover the (DIN-AVD) system studied by Attouch and László in [22].

4.4.2 The case $A = 0$

Now let $A = 0$ in (4.2), and for every $t \geq t_0$ take $\gamma(t) = \gamma \in (0, 2\beta)$ and $\eta(t) = \eta t^2$ with $\eta = \lambda/\gamma$. Then, associated to the problem

$$\text{find } x \in \mathcal{H} \text{ such that } B(x) = 0,$$

we obtain the system

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \xi\frac{d}{dt}\left(\frac{1}{\eta(t)}Bx(t)\right) + \frac{1}{\eta(t)}Bx(t) = 0. \tag{4.25}$$

The assumptions on the parameters now read

Assumption 4.3. Suppose that $\alpha > 1$, $\xi \geq 0$ and

$$\eta(t) = \eta t^2, \text{ where } \eta > \frac{1}{\beta(\alpha-1)^2}.$$

Theorem 4.4.3. *Suppose that the parameters α, ξ and $\eta(\cdot)$ satisfy Assumption 4.3. Then, for a solution $x : [t_0, +\infty) \to \mathcal{H}$ to (4.25), the following are true.*

(i) *The following statements hold, as* $t \to +\infty$*:*

$$\|\dot{x}(t)\| = o\left(\frac{1}{t}\right), \quad \|\ddot{x}(t)\| = \mathcal{O}\left(\frac{1}{t^2}\right) \quad \text{and} \quad \|B(x(t))\| \to 0;$$

(ii) $x(t)$ *converges weakly to an element of* $\operatorname{zer} B$ *as* $t \to +\infty$.

Proof. Since $\eta > \frac{1}{\beta(\alpha-1)^2}$, we can find $\epsilon \in (0, \beta)$ such that $\eta > \frac{1}{(\beta-\epsilon)(\alpha-1)^2}$, equivalently, $2(\beta-\epsilon)\eta > \frac{2}{(\alpha-1)^2}$. Since (4.25) is just (4.2) with $A = 0$ and parameters $\lambda = 2(\beta-\epsilon)\eta > \frac{1}{(\alpha-1)^2}$ and $\gamma(t) \equiv 2(\beta-\epsilon) \in (0, 2\beta)$, the conclusion follows from Theorem 4.3.1, and the fact that, as $t \to +\infty$,

$$\frac{2(\beta-\epsilon)}{\lambda t^2}\|B(x(t))\| = \|T_{\lambda(t),\gamma(t)}(x(t))\| = o\left(\frac{1}{t^2}\right) \quad \text{and thus} \quad \|B(x(t))\| \to 0.$$

□

Remark 4.4.4. (a) As we mentioned in the introduction, the dynamical system (4.25) provides a way of finding the zeros of a cocoercive operator directly through forward evaluations, instead of having to resort to its Moreau envelope when following the approach in [22].
(b) The dynamics (4.25) bear some resemblance to the system addressed by Boț and Csetnek (1.30) (cf. [37])

$$\ddot{x}(t) + \mu(t)\dot{x}(t) + \nu(t)B(x(t)) = 0,$$

with $\mu(t) = \frac{\alpha}{t}$ and $\nu(t) = \frac{1}{\eta(t)}$, with an additional Hessian-driven damping term. In our case, since $\eta > \frac{1}{\beta(\alpha-1)^2}$, the parameters satisfy

$$\dot{\mu}(t) = -\frac{\alpha}{t^2} \leq 0, \quad \frac{\mu^2(t)}{\nu(t)} = \frac{\alpha^2\eta t^2}{t^2} = \alpha^2\eta > \frac{1}{\beta}.$$

for every $t \geq t_0$. However, we have

$$\dot{\nu}(t) = -\frac{2}{\lambda t^3} \leq 0$$

for every $t \geq t_0$, so one of the hypotheses which is needed for (1.30) is not fulfilled, which shows that one cannot address the dynamical system (4.25) as a particular case of it; indeed, in [37] a vanishing damping is not allowed. With our system, we obtain convergence rates for $\|\dot{x}(t)\|$ and $\|\ddot{x}(t)\|$ as $t \to +\infty$, which are not obtained in [37].

4.5 Structured convex minimization

We can particularize the previous results to the case of convex minimization, and show additionally the convergence of functional values along the generated trajectories to the optimal objective value. Let $f : \mathcal{H} \to \overline{\mathbb{R}}$ be a proper, convex and lower semicontinuous function, and let $g : \mathcal{H} \to \mathbb{R}$ be a convex and differentiable function with L_g-Lipschitz continuous gradient. Consider the convex minimization problem

$$\begin{array}{rl} \min & f(x) + g(x), \\ \text{subject to} & x \in \mathcal{H}, \end{array} \tag{4.26}$$

which we assume to have at least one solution. Fermat's rule tells us that x is a global minimum of $f+g$ if and only if

$$0 \in \partial(f+g)(x) = \partial f(x) + \nabla g(x),$$

where the subgradient can be distributed over the sum since g is continuous everywhere, in particular on any point in $\operatorname{dom} f \neq \emptyset$. Therefore, solving (4.26) is equivalent to solving the monotone inclusion $0 \in (A+B)(x)$ addressed previously, with $A = \partial f$ and $B = \nabla g$. Moreover, recall that if ∇g is L_g-Lipschitz then it is $(1/L_g)$-cocoercive (Baillon-Haddad's Theorem, see [28, Corollary 18.17], also [104, Lemma 4]). Therefore, recalling the equivalent formula for $T_{\lambda,\gamma}$ given in Remark 4.1.1, as well as the formula

$$(\partial f)_\gamma = \nabla f_\lambda,$$

attached to the problem (4.26) we have the dynamics

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \xi \frac{d}{dt}\left(\frac{\gamma(t)}{\lambda(t)}\left(\nabla f_{\gamma(t)}(u(t)) + \nabla g(x(t))\right)\right) + \frac{\gamma(t)}{\lambda(t)}\left(\nabla f_{\gamma(t)}(u(t)) + \nabla g(x(t))\right) = 0, \tag{4.27}$$

where we have denoted $u(t) := x(t) - \gamma(t)\nabla g(x(t))$ for all $t \geq t_0$ for convenience.

To produce a meaningful statement regarding the convergence of the functional values along the trajectories, we restrict ourselves to the case $\gamma(t) \equiv \gamma \in \left(0, \frac{2}{L_g}\right)$.

Theorem 4.5.1. *Suppose that the parameters $\alpha, \xi, \lambda(\cdot)$ satisfy Assumption 4.1, and that $\gamma(t) \equiv \gamma \in \left(0, \frac{2}{L_g}\right)$. Then, for a solution $x : [t_0, +\infty) \to \mathcal{H}$ to (4.27), as $t \to +\infty$ it holds*

$$\|\dot{x}(t)\| = o\left(\frac{1}{t}\right), \quad \|\ddot{x}(t)\| = \mathcal{O}\left(\frac{1}{t^2}\right) \quad \textit{and} \quad \left\|\operatorname{prox}_{\gamma f}(u(t)) - x(t)\right\| \to 0,$$

where $u(t) = x(t) - \gamma(t)\nabla g(x(t))$. In particular, this implies that, as $t \to +\infty$,

$$f\left(\operatorname{prox}_{\gamma f}(u(t))\right) + g\left(\operatorname{prox}_{\gamma f}(u(t))\right) - \min(f+g) \to 0.$$

Moreover, $x(t)$ converges weakly to an element of $\operatorname{argmin}(f+g)$ as $t \to +\infty$.

Proof. The statements are a direct application of Theorem (4.3.1). We just need to clarify some facts. Recalling that $J_{\gamma\partial f} = \operatorname{prox}_\gamma f$, we get

$$T_{\lambda(t),\gamma}(x(t)) = \frac{1}{\lambda(t)}\Big[\operatorname{Id} - J_{\gamma\partial f}(\operatorname{Id} - \gamma\nabla g)\Big](x(t)) = \frac{1}{\lambda(t)}\Big[x(t) - \operatorname{prox}_{\gamma f}(u(t))\Big]. \tag{4.28}$$

According to Theorem 4.3.1, we have $\|T_{\lambda(t),\gamma}\| = o(1/t^2)$ as $t \to +\infty$, so formula (4.28) entails

$$\left\|\operatorname{prox}_{\gamma f}(u(t)) - x(t)\right\| \to 0 \tag{4.29}$$

as $t \to +\infty$. Now, let $x^* \in \operatorname{argmin}_{\mathcal{H}}(f+g)$. According to Lemma 2.1.3, and using the Cauchy-Schwarz inequality and (4.29), we obtain

$$\begin{aligned}
& f\left(\operatorname{prox}_{\gamma f}(u(t))\right) + g\left(\operatorname{prox}_{\gamma f}(u(t))\right) - (f(x^*) + g(x^*)) + \frac{2 - \gamma L_g}{2\gamma}\left\|\operatorname{prox}_{\gamma f}(u(t)) - x(t)\right\|^2 \\
\leq{} & \frac{1}{\gamma}\left\langle x(t) - x^*, x(t) - \operatorname{prox}_{\gamma f}(u(t))\right\rangle \\
\leq{} & \frac{1}{\gamma}\left\|x(t) - \operatorname{prox}_{\gamma f}(u(t))\right\|\|x(t) - x^*\| \to 0
\end{aligned}$$

as $t \to +\infty$, since $t \mapsto x(t)$ is bounded (recall (4.10)). This concludes the proof. $\square$

4.5.1 The case $g \equiv 0$

It is also worth mentioning the system we obtain in the case where $g \equiv 0$, since we also obtain improved rates for the objective functional values when we compare, for example (4.2) to (DIN-AVD) [22]. In this case, we have the system

$$\ddot{x}(t) + \frac{\alpha}{t} + \xi\frac{d}{dt}\left(\frac{\gamma(t)}{\lambda(t)}\nabla f_{\gamma(t)}(x(t))\right) + \frac{\gamma(t)}{\lambda(t)}\nabla f_{\gamma(t)}(x(t)) = 0 \tag{4.30}$$

attached to the convex optimization problem

$$\begin{aligned} \min \quad & f(x), \\ \text{subject to} \quad & x \in \mathcal{H}. \end{aligned}$$

We can state the following theorem.

Theorem 4.5.2. *Suppose that the parameters $\alpha, \xi, \lambda(\cdot)$ and $\gamma(\cdot)$ satisfy Assumption 4.2. Then, for a solution $x : [t_0, +\infty) \to \mathcal{H}$ to (4.30), the following are true.*

(i) *The following convergence rates hold, as $t \to +\infty$:*

$$\|\dot{x}(t)\| = o\left(\frac{1}{t}\right), \quad \|\ddot{x}(t)\| = \mathcal{O}\left(\frac{1}{t^2}\right) \quad \text{and} \quad \|\nabla f_{\gamma(t)}(x(t))\| = o\left(\frac{1}{\gamma(t)}\right).$$

In particular, this implies that, as $t \to +\infty$,

$$f_{\gamma(t)}(x(t)) - \min f = o\left(\frac{1}{\gamma(t)}\right) \quad \text{and} \quad f\left(\operatorname{prox}_{\gamma(t)f}(x(t))\right) - \min f = o\left(\frac{1}{\gamma(t)}\right).$$

(ii) *If $0 < \inf_{t \geq t_0} \gamma(t)$, then $x(t)$ converges weakly to an element of $\operatorname{argmin} f$ as $t \to +\infty$.*

Proof. This is a direct application of Theorem 4.4.3; just a few comments are in order. Let $x^* \in \mathcal{H}$ be a minimizer of f. We apply the gradient inequality to $f_{\gamma(t)}$, from which we obtain

$$\begin{aligned} f_{\gamma(t)}(x(t)) - \min\nolimits_{\mathcal{H}} f &= f_{\gamma(t)}(x(t)) - f(x^*) \leq f_{\gamma(t)}(x(t)) - f_{\lambda(t)}(x^*) \\ &\leq \left\langle \nabla f_{\gamma(t)}(x(t)), x(t) - x^* \right\rangle \leq \left\| \nabla f_{\gamma(t)}(x(t)) \right\| \|x(t) - x^*\| \\ &= o\left(\frac{1}{\gamma(t)}\right) \end{aligned}$$

as $t \to +\infty$, since $t \mapsto x(t)$ is bounded (see (4.10)) and Theorem 4.4.3 ensures that $\left\|\nabla f_{\gamma(t)}(x(t))\right\| = o\left(\frac{1}{\gamma(t)}\right)$ as $t \to +\infty$. Recalling the definition of the Moreau envelope of f, this finally gives, as $t \to +\infty$,

$$f\left(\operatorname{prox}_{\gamma(t)f}(x(t))\right) + \frac{1}{2\gamma(t)}\left\|\operatorname{prox}_{\gamma(t)f}(x(t)) - x(t)\right\|^2 - \min\nolimits_{\mathcal{H}} f = f_{\gamma(t)}(x(t)) - \min\nolimits_{\mathcal{H}} f = o\left(\frac{1}{\gamma(t)}\right).$$

□

Remark 4.5.3. As pointed out in Remark 4.4.2, we can choose $\gamma(t) = \lambda t^2$ for every $t \geq t_0$ and recover the (DIN-AVD) system for nonsmooth convex minimization problems studied in [22]. Moreover, we can also set $\gamma(t) = t^n$ for a natural number $n > 3$ and all $t \geq t_0$. Now, not only are the convergence rates for $\nabla f_{\gamma(t)}(x(t))$ and $\frac{d}{dt}\nabla f_{\gamma(t)}(x(t))$ as $t \to +\infty$ improved with respect to the system in [22], but (4.30) also provides a better rate for the convergence of $f_{\gamma(t)}(x(t))$ to $\min f$ as $t \to +\infty$.

4.6 A numerical algorithm

In the following we will derive via time discretization of (4.2) a numerical algorithm for solving the monotone inclusion problem (4.1). We perform an implicit discretization of (4.2) as follows:

$$\ddot{x}(t) \approx x_{k+1} - 2x_k + x_{k-1}, \qquad \frac{\alpha}{t}\dot{x}(t) \approx \frac{\alpha}{k}(x_k - x_{k-1}),$$

$$\frac{d}{dt}T_{\lambda(t),\gamma(t)}(x(t)) \approx T_{\lambda_k,\gamma_k}(x_k) - T_{\lambda_{k-1},\gamma_{k-1}}(x_{k-1}), \qquad T_{\lambda(t),\gamma(t)}(x(t)) \approx T_{\lambda_{k+1},\gamma_{k+1}}(x_{k+1}),$$

so we get, for every $k \geq 1$,

$$x_{k+1} - 2x_k - x_{k-1} + \frac{\alpha}{k}(x_k - x_{k-1}) + \xi\left(T_{\lambda_k,\gamma_k}(x_k) - T_{\lambda_{k-1},\gamma_{k-1}}(x_{k-1})\right) + T_{\lambda_{k+1},\gamma_{k+1}}(x_{k+1}) = 0. \quad (4.31)$$

After rearranging the terms of (4.31), for every $k \geq 1$ we obtain

$$x_{k+1} + T_{\lambda_{k+1},\gamma_{k+1}}(x_{k+1}) = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) - \xi\left(T_{\lambda_k,\gamma_k}(x_k) - T_{\lambda_{k-1},\gamma_{k-1}}(x_{k-1})\right). \quad (4.32)$$

In other words, after setting $\alpha_k = 1 - \frac{\alpha}{k}$ and denoting the right hand side of (4.32) by y_k for every $k \geq 1$, we obtain

Algorithm 4.1:

Input: $z_0, z_1 \in \mathcal{H}$, $\alpha > 0$, $\xi \geq 0$, $(\lambda_k)_{k\geq 0}$, $(\gamma_k)_{k\geq 0}$ sequences of positive numbers;
for $k = 1, 2, \ldots$ **do**

$$\alpha_k := 1 - \frac{\alpha}{k},$$
$$y_k := x_k + \alpha_k(x_k - x_{k-1}) - \xi\left(T_{\lambda_k,\gamma_k}(x_k) - T_{\lambda_{k-1},\gamma_{k-1}}(x_{k-1})\right),$$
$$x_{k+1} := \left(\mathrm{Id} + T_{\lambda_{k+1},\gamma_{k+1}}\right)^{-1}(y_k).$$

Observe that the third step of Algorithm 4.1 is always well-defined. Indeed, for $\lambda, \gamma > 0$, $T_{\lambda,\gamma}$ is $\frac{\lambda}{2}$-cocoercive, hence monotone (see Lemma 4.2.2(i)). This also implies that $T_{\lambda,\gamma}$ is $\frac{2}{\lambda}$-Lipschitz continuous, and a monotone and continuous operator is maximally monotone, according to [28, Corollary 20.28]. Hence, by Minty's Theorem (see [28, Theorem 21.1]), we know that $\mathrm{Id} + T_{\lambda,\gamma} : \mathcal{H} \to \mathcal{H}$ is surjective. The assumptions for the discrete-time versions of the parameters reads as follows.

Assumption 4.4. Suppose that $\alpha > 1$, $\xi \geq 0$, and that $(\lambda_k)_{k\geq 0}$ and $(\gamma_k)_{k\geq 0} \subseteq (0, 2\beta)$ are sequences which satisfy, as $k \to +\infty$,

$$\lambda_k = \lambda k^2, \text{ where } \lambda > \frac{4\xi + 2}{(\alpha - 1)^2}, \quad \text{and} \quad \frac{|\gamma_k - \gamma_{k-1}|}{\gamma_k} = \mathcal{O}\left(\frac{1}{k}\right).$$

Theorem 4.6.1. *Suppose that α, ξ and the sequences $(\lambda_k)_{k\geq 0}, (\gamma_k)_{k\geq 0}$ satisfy Assumption 4.4. Let $(x_k)_{k\geq 0}$ be the sequence generated by Algorithm 4.1. Then, the following are true.*

(i) *The following convergence rates hold, as $k \to +\infty$:*

$$\|x_{k+1} - x_k\| = \mathcal{O}\left(\frac{1}{k}\right) \quad \textit{and} \quad \|T_{\lambda_k,\gamma_k}(x_k)\| = o\left(\frac{1}{k^2}\right);$$

(ii) *If $0 < \inf_{k\geq 0} \gamma_k \leq \sup_{k\geq 0} \gamma_k < 2\beta$, then x_k converges weakly to an element of $\operatorname{zer}(A + B)$ as $k \to +\infty$. Furthermore, as $k \to +\infty$ it holds*

$$\|x_k - y_k\| = \mathcal{O}\left(\frac{1}{k}\right).$$

The proof can be done by transposing the techniques used in the continuous time case to the discrete time case. Algorithm (4.1) can be seen as a splitting version of the (PRINAM) algorithm studied by Attouch and László in [23].

The third step of Algorithm 4.1 can be quite complicated to compute. However, if $B = 0$, we can resort to the fact that $(A_{\lambda_1})_{\lambda_2} = A_{\lambda_1+\lambda_2}$ for $\lambda_1, \lambda_2 > 0$. We now have, for $\lambda, \gamma > 0$,

$$T_{\lambda,\gamma} = \frac{1}{\lambda}\Big[\mathrm{Id} - J_{\gamma A}\Big] = \frac{\gamma}{\lambda} A_\gamma,$$

which gives

$$(\mathrm{Id} + T_{\gamma,\lambda})^{-1} = J_{\frac{\gamma}{\lambda} A_\gamma} = -\frac{\gamma}{\lambda}(A_\lambda)_{\frac{\gamma}{\lambda}} + \mathrm{Id} = \mathrm{Id} - \frac{\gamma}{\lambda} A_{\lambda + \frac{\gamma}{\lambda}}.$$

It is now possible to write Algorithm 4.1 in terms of the resolvents of A. We have, for every $k \geq 1$,

$$\begin{aligned}
T_{\lambda_k,\gamma_k}(x_k) - T_{\lambda_{k-1},\gamma_{k-1}}(x_{k-1}) &= \frac{1}{\lambda_k}\Big[x_k - J_{\gamma_k A}(x_k)\Big] - \frac{1}{\lambda_{k-1}}\Big[x_{k-1} - J_{\gamma_{k-1} A}(x_{k-1})\Big] \\
&= \left(\frac{1}{\lambda_k} - \frac{1}{\lambda_{k-1}}\right) x_k + \frac{1}{\lambda_{k-1}}(x_k - x_{k-1}) \\
&\quad - \left(\frac{1}{\lambda_k} J_{\gamma_k A}(x_k) - \frac{1}{\lambda_{k-1}} J_{\gamma_{k-1} A}(x_{k-1})\right),
\end{aligned}$$

$$y_k - \frac{\gamma_{k+1}}{\lambda_{k+1}} A_{\lambda_{k+1} + \frac{\gamma_{k+1}}{\lambda_{k+1}}}(y_k) = y_k - \frac{\gamma_{k+1}}{\lambda_{k+1}} \frac{1}{\frac{\lambda_{k+1}^2 + \gamma_{k+1}}{\lambda_{k+1}}} \left[y_k - J_{\left(\lambda_{k+1} + \frac{\gamma_{k+1}}{\lambda_{k+1}}\right) A}(y_k)\right]$$

$$= \frac{\lambda_{k+1}^2}{\lambda_{k+1}^2 + \gamma_{k+1}} y_k + \frac{\gamma_k}{\lambda_{k+1}^2 + \gamma_{k+1}} J_{\left(\lambda_{k+1} + \frac{\gamma_{k+1}}{\lambda_{k+1}}\right) A}(y_k).$$

So now Algorithm 4.1 becomes

Algorithm 4.2:

Input: $x_0, x_1 \in \mathcal{H}$, $\alpha > 0$, $\xi \geq 0$, $(\lambda_k)_{k \geq 0}, (\gamma_k)_{k \geq 0}$ sequences of positive numbers;

for $k = 1, 2, \ldots$ **do**

$$\alpha_k := 1 - \frac{\alpha}{k},$$

$$y_k := \left(1 - \xi\left(\frac{1}{\lambda_k} - \frac{1}{\lambda_{k-1}}\right)\right) x_k + \left(\alpha_k - \frac{\xi}{\lambda_{k-1}}\right)(x_k - x_{k-1})$$
$$+ \xi \left(\frac{1}{\lambda_k} J_{\gamma_k A}(x_k) - \frac{1}{\lambda_{k-1}} J_{\gamma_{k-1} A}(x_{k-1})\right),$$

$$x_{k+1} := \frac{\lambda_{k+1}^2}{\lambda_{k+1}^2 + \gamma_{k+1}} y_k + \frac{\gamma_k}{\lambda_{k+1}^2 + \gamma_{k+1}} J_{\left(\lambda_{k+1} + \frac{\gamma_{k+1}}{\lambda_{k+1}}\right) A}(y_k).$$

The corresponding discrete-time analog of Assumption 4.2 now reads

Assumption 4.5. Suppose that $\alpha > 1$, $\xi \geq 0$, and that $(\lambda_k)_{k \geq 0}, (\gamma_k)_{k \geq 0}$ be sequences of positive numbers which satisfy, as $k \to +\infty$,

$$\lambda_k = \lambda k^2, \text{ where } \lambda > \frac{4\xi + 2}{(\alpha - 1)^2}, \quad \text{and} \quad \frac{|\gamma_k - \gamma_{k-1}|}{\gamma_k} = \mathcal{O}\left(\frac{1}{k}\right).$$

Theorem 4.6.2. *Suppose that α, ξ and the sequences $(\lambda_k)_{k \geq 0}$, $(\gamma_k)_{k \geq 0}$ satisfy Assumption 4.5. Let $(x_k)_{k \geq 0}$ be the sequence generated by Algorithm 4.2. Then, the following are true.*

(i) *The following convergence rates hold, as $k \to +\infty$:*

$$\|x_{k+1} - x_k\| = \mathcal{O}\left(\frac{1}{k}\right) \quad \textit{and} \quad \|A_{\gamma_k}(x_k)\| = o\left(\frac{1}{\gamma_k}\right);$$

(ii) *If $0 < \inf_{k \geq 0} \gamma_k$, then x_k converges weakly to an element of $\operatorname{zer} A$ as $k \to +\infty$. Furthermore, as $k \to +\infty$ it holds*

$$\|x_k - y_k\| = \mathcal{O}\left(\frac{1}{k}\right).$$

Remark 4.6.3. Notice that the condition required for $(\gamma_k)_{k \geq 0}$ is fulfilled in particular for $\gamma_k = k^n$ for every $k \geq 1$ and a natural number $n \geq 1$. Thus, by choosing large n, we obtain a fast convergence rate for $\|A_{\gamma_k}(x_k)\|$ as $k \to +\infty$.

Chapter 5

A second-order primal-dual system attached to a linearly constrained minimization problem

5.1 Introduction

In this chapter, which relies on our paper [67], we will consider the linearly constrained optimization problem

$$\begin{array}{rl} \min & f(x), \\ \text{subject to} & Ax = b, \end{array} \tag{5.1}$$

where

$$\begin{cases} \mathcal{X}, \mathcal{Y} \text{ are real Hilbert spaces;} \\ f : \mathcal{X} \to \mathbb{R} \text{ is a continuously differentiable convex function;} \\ A : \mathcal{X} \to \mathcal{Y} \text{ is a continuous linear operator and } b \in \mathcal{Y}; \\ \text{the set } \mathbb{S} \text{ of primal-dual optimal solutions of (5.1) is assumed to be nonempty.} \end{cases} \tag{5.2}$$

Additionally, we will take the following inner product and norm on $\mathcal{X} \times \mathcal{Y}$:

$$\big\langle (x, \lambda), (z, \mu) \big\rangle := \langle x, z \rangle + \langle \lambda, \mu \rangle \quad \text{and} \quad \|(x, \lambda)\| := \sqrt{\|x\|^2 + \|\lambda\|^2}.$$

In the introductory chapter, we mentioned that under a mild constraint qualification, solving (5.1) is equivalent to finding the saddle points of either $\mathcal{L}$ or $\mathcal{L}_\beta$, where the *Lagrangian* $\mathcal{L}$ and the *augmented Lagrangian* $\mathcal{L}_\beta$ are given respectively by

$$\begin{aligned} \mathcal{L}(x, \lambda) &:= f(x) + \langle \lambda, Ax - b \rangle, \\ \mathcal{L}_\beta(x, \lambda) &:= \mathcal{L}(x, \lambda) + \frac{\beta}{2}\|Ax - b\|^2 = f(x) + \langle \lambda, Ax - b \rangle + \frac{\beta}{2}\|Ax - b\|^2, \end{aligned}$$

where $\beta > 0$. We recall that the saddle point sets of these functions coincide, and we denote it by $\mathbb{S}$. We will denote the feasible set of (5.1) by $\mathbb{F}$, i.e., $\mathbb{F} := \{x \in \mathcal{X} : Ax = b\}$.

Attached to (5.1), we will analyze the asymptotic properties of

$$\begin{cases} \ddot{x}(t) + \dfrac{\alpha}{t}\dot{x}(t) + \delta(t)\nabla_x \mathcal{L}_\beta\Big(x(t), \lambda(t) + \theta t \dot{\lambda}(t)\Big) &= 0, \\ \ddot{\lambda}(t) + \dfrac{\alpha}{t}\dot{\lambda}(t) - \delta(t)\nabla_\lambda \mathcal{L}_\beta\Big(x(t) + \theta t \dot{x}(t), \lambda(t)\Big) &= 0, \end{cases} \tag{5.3}$$

where $t \geq t_0 > 0$, $\alpha > 0$, $\theta > 0$ and $\delta : [t_0, +\infty) \to \mathbb{R}_{++}$ is a time rescaling parameter. We will show fast rates of the functional values along the trajectories, as well as the fast convergence of the feasibility gap to zero. The function $\delta(\cdot)$ will enter these rates, thus having faster convergence rates for the appropriate choice for $\delta(\cdot)$; namely, as $t \to +\infty$ we will have

$$|f(x(t)) - f(x^*)| = \mathcal{O}\left(\frac{1}{t^2\delta(t)}\right) \quad \text{and} \quad \|Ax(t) - b\| = \mathcal{O}\left(\frac{1}{t^2\delta(t)}\right),$$

where x^* is an optimal solution to (5.1). Additionally, we will show the weak convergence of trajectories towards primal-dual solutions to (5.1).

This chapter is organized as follows: in Section 5.2, we will introduce the energy function which we will then use to show the convergence rates of our system; in Section 5.3, we will show the weak convergence of trajectories; in Section 5.4, we will illustrate our theoretical findings with some numerical experiments.

5.2 Faster convergence rates via time rescaling

Replacing the expressions of the partial gradients of $\mathcal{L}_\beta$ into the system leads to the following formulation for (5.3):

$$\begin{cases} \ddot{x}(t) + \dfrac{\alpha}{t}\dot{x}(t) + \delta(t)\nabla f(x(t)) + \delta(t)A^*\big(\lambda(t) + \theta t\dot{\lambda}(t)\big) + \delta(t)\beta A^*(Ax(t) - b) & = 0, \\ \ddot{\lambda}(t) + \dfrac{\alpha}{t}\dot{\lambda}(t) - \delta(t)\Big(A\big(x(t) + \theta t\dot{x}(t)\big) - b\Big) & = 0. \end{cases}$$

In this section we will derive fast convergence rates for the primal-dual gap, the feasibility measure, and the objective function value along the trajectories generated by the dynamical system (5.3). We will make the following assumptions on the parameters α, θ, β and the function δ throughout this section.

Assumption 5.1. Suppose that $\delta : [t_0, +\infty) \to \mathbb{R}_{++}$ is continuously differentiable and nondecreasing. Moreover, suppose that the parameters α, β, θ and the function δ satisfy

$$\alpha \geq 3, \quad \beta \geq 0, \quad \frac{1}{2} \geq \theta \geq \frac{1}{\alpha - 1} \quad \text{and} \quad \sup_{t \geq t_0} \frac{t\dot{\delta}(t)}{\delta(t)} \leq \frac{1 - 2\theta}{\theta}. \tag{5.4}$$

Besides the first three conditions that are known previously in [43], it is worth pointing out that we can deduce from the last one the following inequality for every $t \geq t_0$:

$$\frac{t\dot{\delta}(t)}{\delta(t)} \leq \frac{1 - 2\theta}{\theta} = \frac{1}{\theta} - 2 \leq \alpha - 3. \tag{5.5}$$

This gives a connection to the condition which appears in [15]. A few more comments regarding the function δ will come later, after the convergence rates statements.

5.2.1 The energy function

Let $(x, \lambda) : [t_0, +\infty) \to \mathcal{X} \times \mathcal{Y}$ be a solution of (5.3). Let $(x^*, \lambda^*) \in \mathbb{S}$ be fixed, we define the energy function $\mathcal{E} : [t_0, +\infty) \to \mathbb{R}$ by

$$\mathcal{E}(t) := \theta^2 t^2 \delta(t)(\mathcal{L}_\beta(x(t), \lambda^*) - \mathcal{L}_\beta(x^*, \lambda(t))) + \frac{1}{2}\|v(t)\|^2 + \frac{\xi}{2}\|\big(x(t), \lambda(t)\big) - (x^*, \lambda^*)\|^2, \tag{5.6}$$

where

$$v(t) := (x(t), \lambda(t)) - (x^*, \lambda^*) + \theta t\big(\dot{x}(t), \dot{\lambda}(t)\big), \tag{5.7}$$
$$\xi := \alpha\theta - \theta - 1 \geq 0. \tag{5.8}$$

Notice that for every $t \geq t_0$ we have

$$\begin{aligned} \mathcal{L}_\beta(x(t), \lambda^*) - \mathcal{L}_\beta(x^*, \lambda(t)) &= \mathcal{L}(x(t), \lambda^*) - \mathcal{L}(x^*, \lambda(t)) + \frac{\beta}{2}\|Ax(t) - b\|^2 && (5.9) \\ &= \mathcal{L}(x(t), \lambda^*) - f(x^*) + \frac{\beta}{2}\|Ax(t) - b\|^2 \\ &= f(x(t)) - f(x^*) + \langle \lambda^*, Ax(t) - b\rangle + \frac{\beta}{2}\|Ax(t) - b\|^2 \geq 0. && (5.10) \end{aligned}$$

In addition, due to (5.10), for every $t \geq t_0$ we have

$$\mathcal{E}(t) \geq 0. \tag{5.11}$$

The construction of $\mathcal{E}$ is inspired by [43]. However, one can notice that we only consider $\mathcal{E}$ defined with respect to a fixed primal-dual solution $(x^*, \lambda^*) \in \mathbb{S}$ rather than a family of energy functions, each defined with respect to a point $(z, \mu) \in \mathbb{F} \times \mathcal{Y}$. This gives simpler proofs for some results when compared to those in [43].

Assumption 5.1 implies the nonnegativity of the following quantity, which will appear many times in our analysis:

$$\sigma : [t_0, +\infty) \to \mathbb{R}_+, \quad \sigma(t) := \frac{1-2\theta}{\theta}\delta(t) - t\dot{\delta}(t). \tag{5.12}$$

The following lemma gives us the decreasing property of the energy function. As a consequence of this lemma, we obtain some integrability results which will be needed later.

Lemma 5.2.1. *Let $(x, \lambda) : [t_0, +\infty) \to \mathcal{X} \times \mathcal{Y}$ be a solution of (5.3) and $(x^*, \lambda^*) \in \mathbb{S}$. Then, for every $t \geq t_0$ it holds*

$$\begin{aligned} \frac{d}{dt}\mathcal{E}(t) \leq & -\theta^2 t\sigma(t) \left(\mathcal{L}_\beta(x(t), \lambda^*) - \mathcal{L}_\beta(x^*, \lambda(t))\right) - \frac{1}{2}\beta\theta t\delta(t)\|Ax(t) - b\|^2 \\ & - \xi\theta t\big\|\big(\dot{x}(t), \dot{\lambda}(t)\big)\big\|^2. \end{aligned}$$

Proof. Let $t \geq t_0$ be fixed. Since $x^* \in \mathbb{F}$, we have

$$\begin{aligned} \nabla_x \left(\mathcal{L}_\beta(x(t), \lambda^*) - \mathcal{L}_\beta(x^*, \lambda(t))\right) &= \nabla_x \mathcal{L}_\beta(x(t), \lambda^*) = \nabla f(x(t)) + A^*\lambda^* + \beta A^*\left(Ax(t) - b\right), \\ \nabla_\lambda \left(\mathcal{L}_\beta(x(t), \lambda^*) - \mathcal{L}_\beta(x^*, \lambda(t))\right) &= -\nabla_\lambda \mathcal{L}_\beta(x^*, \lambda(t)) = 0. \end{aligned}$$

Under these expressions, the system (5.3) can be equivalently written as

$$\begin{aligned} \big(\ddot{x}(t), \ddot{\lambda}(t)\big) = & -\frac{\alpha}{t}\big(\dot{x}(t), \dot{\lambda}(t)\big) - \delta(t)\big(\nabla_x \mathcal{L}_\beta(x(t), \lambda^*), 0\big) \\ & - \delta(t)\Big(A^*\big(\lambda(t) - \lambda^* + \theta t\dot{\lambda}(t)\big), -\Big(A\left(x(t) + \theta t\dot{x}(t)\right) - b\Big)\Big), \end{aligned}$$

which leads to

$$\begin{aligned} \dot{v}(t) &= (1+\theta)\big(\dot{x}(t), \dot{\lambda}(t)\big) + \theta t\big(\ddot{x}(t), \ddot{\lambda}(t)\big) \\ &= -\xi\big(\dot{x}(t), \dot{\lambda}(t)\big) - \theta t\delta(t)(t)\big(\nabla_x \mathcal{L}_\beta(x(t), \lambda^*), 0\big) \\ &\quad - \theta t\delta(t)\Big(A^*\big(\lambda(t) - \lambda^* + \theta t\dot{\lambda}(t)\big), -\Big(A\left(x(t) + \theta t\dot{x}(t)\right) - b\Big)\Big). \end{aligned}$$

We get from the distributive property of the inner product

$$\begin{aligned} & \langle v(t), \dot{v}(t)\rangle \\ = & -\xi\left\langle \big(x(t), \lambda(t)\big) - (x^*, \lambda^*), \big(\dot{x}(t), \dot{\lambda}(t)\big)\right\rangle - \xi\theta t\big\|\big(\dot{x}(t), \dot{\lambda}(t)\big)\big\|^2 \\ & - \theta t\delta(t)\left\langle \left(\nabla_x \mathcal{L}_\beta(x(t), \lambda^*), 0\right), \big(x(t), \lambda(t)\big) - (x^*, \lambda^*)\right\rangle \\ & - \theta^2 t^2\delta(t)\left\langle \left(\nabla_x \mathcal{L}_\beta(x(t), \lambda^*), 0\right), \big(\dot{x}(t), \dot{\lambda}(t)\big)\right\rangle \\ & - \theta t\delta(t)\left\langle \lambda(t) - \lambda^* + \theta t\dot{\lambda}(t), Ax(t) - Ax^*\right\rangle - \theta^2 t^2\delta(t)\left\langle \lambda(t) - \lambda^* + \theta t\dot{\lambda}(t), A\dot{x}(t)\right\rangle \\ & + \theta t\delta(t)\left\langle A\left(x(t) + \theta t\dot{x}(t)\right) - b, \lambda(t) - \lambda^*\right\rangle + \theta^2 t^2\delta(t)\left\langle A\left(x(t) + \theta t\dot{x}(t)\right) - b, \dot{\lambda}(t)\right\rangle. \end{aligned}$$

Since $x^* \in \mathbb{F}$, the last four terms in the above identity vanish. Indeed,

$$\begin{aligned} & -\left\langle \lambda(t) - \lambda^* + \theta t\dot{\lambda}(t), Ax(t) - Ax^*\right\rangle - \theta t\left\langle \lambda(t) - \lambda^* + \theta t\dot{\lambda}(t), A\dot{x}(t)\right\rangle \\ & + \left\langle A\left(x(t) + \theta t\dot{x}(t)\right) - b, \lambda(t) - \lambda^*\right\rangle + \theta t\left\langle A\left(x(t) + \theta t\dot{x}(t)\right) - b, \dot{\lambda}(t)\right\rangle \\ = & -\left\langle \lambda(t) - \lambda^* + \theta t\dot{\lambda}(t), Ax(t) - b\right\rangle - \theta t\left\langle \lambda(t) - \lambda^* + \theta t\dot{\lambda}(t), A\dot{x}(t)\right\rangle \\ & + \left\langle Ax(t) - b, \lambda(t) - \lambda^*\right\rangle + \theta t\left\langle A\dot{x}(t), \lambda(t) - \lambda^*\right\rangle \\ & + \theta t\left\langle Ax(t) - b, \dot{\lambda}(t)\right\rangle + \theta^2 t^2\left\langle A\dot{x}(t), \dot{\lambda}(t)\right\rangle \\ = & \ 0. \end{aligned}$$

Therefore, differentiating $\mathcal{E}$ with respect to t gives

$$\frac{d}{dt}\mathcal{E}(t) = \theta^2 t\left(2\delta(t) + t\dot{\delta}(t)\right)\left(\mathcal{L}_\beta(x(t), \lambda^*) - \mathcal{L}_\beta(x^*, \lambda(t))\right)$$

$$
\begin{aligned}
&+\theta^2 t^2 \delta(t)\left\langle\left(\nabla_x \mathcal{L}_\beta(x(t),\lambda^*),0\right),\left(\dot{x}(t),\dot{\lambda}(t)\right)\right\rangle \\
&+\langle v(t),\dot{v}(t)\rangle+\xi\left\langle\left(x(t),\lambda(t)\right)-(x^*,\lambda^*),\left(\dot{x}(t),\dot{\lambda}(t)\right)\right\rangle \\
=\,&\theta^2 t\left(2\delta(t)+t\dot{\delta}(t)\right)\left(\mathcal{L}_\beta(x(t),\lambda^*)-\mathcal{L}_\beta(x^*,\lambda(t))\right)-\xi\theta t\left\|\left(\dot{x}(t),\dot{\lambda}(t)\right)\right\|^2 \\
&-\theta t\delta(t)\left\langle\left(\nabla_x \mathcal{L}_\beta(x(t),\lambda^*),0\right),\left(x(t),\lambda(t)\right)-(x^*,\lambda^*)\right\rangle. \qquad (5.13)
\end{aligned}
$$

Furthermore, the convexity of f and the fact that $x^* \in \mathbb{F}$ guarantee

$$
\begin{aligned}
&-\left\langle\left(\nabla_x \mathcal{L}_\beta(x(t),\lambda^*),0\right),\left(x(t),\lambda(t)\right)-(x^*,\lambda^*)\right\rangle \\
=\,&\langle\nabla f(x(t)),x^*-x(t)\rangle+\langle A^*\lambda^*,x^*-x(t)\rangle+\beta\left\langle A^*\left(Ax(t)-b\right),x^*-x(t)\right\rangle \\
\leq\,&-\left(f(x(t))-f(x^*)\right)-\langle\lambda^*,Ax(t)-b\rangle-\beta\|Ax(t)-b\|^2 \qquad (5.14)\\
=\,&-\left(\mathcal{L}_\beta(x(t),\lambda^*)-\mathcal{L}_\beta(x^*,\lambda(t))\right)-\frac{\beta}{2}\|Ax(t)-b\|^2,
\end{aligned}
$$

By multiplying this inequality by $\theta t\delta(t)$ and combining it with (5.13), the coefficient attached to the primal-dual gap becomes

$$
\theta^2 t\left(2\delta(t)+t\dot{\delta}(t)\right)-\theta t\delta(t)=-\theta^2 t\left(\frac{1-2\theta}{\theta}\delta(t)-t\dot{\delta}(t)\right)=-\theta^2 t\sigma(t),
$$

which finally gives the desired statement. □

Remark 5.2.2. It is straightforward to show the existence of local solutions to (5.3), under the additional assumption that ∇f is Lipschitz continuous on every bounded subset of $\mathcal{X}$. First, notice that (5.3) can be rewritten as a first-order dynamical system. Indeed, $(x,\lambda):[t_0,+\infty)\to\mathcal{X}\times\mathcal{Y}$ is a solution to (5.3) with initial conditions $(x(t_0),\lambda(t_0))=(x_0,\lambda_0)$ and $\left(\dot{x}(t_0),\dot{\lambda}(t_0)\right)=\left(\dot{x}_0,\dot{\lambda}_0\right)$ if and only if $(x,\lambda,y,\nu):[t_0,+\infty)\to\mathcal{X}\times\mathcal{Y}\times\mathcal{X}\times\mathcal{Y}$ is a solution to

$$
\begin{cases}
\left(\dot{x}(t),\dot{\lambda}(t),\dot{y}(t),\dot{\nu}(t)\right) & =F(t,x(t),\lambda(t),y(t),\nu(t))\\
\left(x(t_0),\lambda(t_0),y(t_0),\nu(t_0)\right) & =\left(x_0,\lambda_0,\dot{x}_0,\dot{\lambda}_0\right),
\end{cases}
$$

where $F:[t_0,+\infty)\times\mathcal{X}\times\mathcal{Y}\times\mathcal{X}\times\mathcal{Y}\to\mathcal{X}\times\mathcal{Y}\times\mathcal{X}\times\mathcal{Y}$ is given by

$$
\begin{aligned}
&F(t,z,\mu,w,\eta)\\
:=&\left(w,\eta,-\frac{\alpha}{t}w-\delta(t)\left[\nabla f(z)+A^*\left(\mu+\theta t\eta\right)+\beta A^*\left(Az-b\right)\right],-\frac{\alpha}{t}\eta+\delta(t)\left[A\left(z+\theta tw\right)-b\right]\right).
\end{aligned}
$$

Here F is evidently continuous in t, and $F(t,\cdot)$ is Lipschitz continuous on every bounded subset, provided that the same property holds for ∇f. We can then employ a theorem such as that by Cauchy-Lipschitz to obtain the existence of a unique solution to the previous system, and thus a unique solution to (5.3), defined on a maximal interval $[t_0,T_{\max})$. To go further and show the existence and uniqueness of a global solution (that is, $T_{\max}=+\infty$) we use the estimate given by the previous lemma, in a very similar way to [19, 43], so we omit the details here.

Theorem 5.2.3. *Under Assumption 5.1, let $(x,\lambda):[t_0,+\infty)\to\mathcal{X}\times\mathcal{Y}$ be a solution of (5.3) and $(x^*,\lambda^*)\in\mathbb{S}$. Then, the following statements are true:*

(i) *It holds*

$$
\int_{t_0}^{+\infty} t\sigma(t)\Big[\mathcal{L}(x(t),\lambda^*)-\mathcal{L}(x^*,\lambda(t))\Big]dt\leq\ \mathcal{E}(t_0)\ <+\infty, \qquad (5.15)
$$

$$
\beta\int_{t_0}^{+\infty} t\delta(t)\|Ax(t)-b\|^2dt\leq\frac{2\mathcal{E}(t_0)}{\theta}<+\infty, \qquad (5.16)
$$

$$
\xi\int_{t_0}^{+\infty} t\left\|\left(\dot{x}(t),\dot{\lambda}(t)\right)\right\|^2\leq\frac{\mathcal{E}(t_0)}{\theta}\ <+\infty; \qquad (5.17)
$$

(ii) *If, in addition, $\alpha>3$ and $\frac{1}{2}\geq\theta>\frac{1}{\alpha-1}$, then the trajectory $(x,\lambda):[t_0,+\infty)\to\mathcal{X}\times\mathcal{Y}$ is* *bounded and the convergence rate of its velocity is, as $t \to +\infty$,*

$$\left\|\left(\dot{x}(t), \dot{\lambda}(t)\right)\right\| = \mathcal{O}\left(\frac{1}{t}\right).$$

Proof. (i) Recall that Assumption 5.1 implies $\sigma(t) \geq 0$ for all $t \geq t_0$ and $\xi \geq 0$. Moreover, $(x^*, \lambda^*) \in \mathbb{S}$ yields $x^* \in \mathbb{F}$. Therefore, we can apply (5.10) and Lemma 5.2.1 to obtain, for every $t \geq t_0$,

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(t) \leq & -\theta^2 t\sigma(t)\left(\mathcal{L}_\beta\left(x(t), \lambda^*\right) - \mathcal{L}_\beta(x^*, \lambda(t))\right) \\ & - \frac{1}{2}\beta\theta t\delta(t)\|Ax(t) - b\|^2 - \xi\theta t\left\|\left(\dot{x}(t), \dot{\lambda}(t)\right)\right\|^2 \leq 0. \end{aligned} \tag{5.18}$$

This means that $\mathcal{E}$ is nonincreasing on $[t_0, +\infty)$. For every $t \geq t_0$, by integrating (5.18) from t_0 to t, we obtain

$$\begin{aligned} & \theta^2 \int_{t_0}^t t\sigma(t)\Big(\mathcal{L}(x(s), \lambda^*) - \mathcal{L}(x^*, \lambda(s))\Big)ds \\ & \quad + \frac{\beta\theta}{2}\int_{t_0}^t s\delta(s)\|Ax(s) - b\|^2 ds + \xi\theta\int_{t_0}^t s\left\|\left(\dot{x}(s), \dot{\lambda(s)}\right)\right\|^2 ds \\ & \leq \mathcal{E}(t_0) - \mathcal{E}(t) \leq \mathcal{E}(t_0), \end{aligned}$$

where the last inequality follows from (5.11). Since all quantities inside the integrals are nonnegative, we obtain (5.15)-(5.17) by letting $t \to +\infty$.

(ii) Let $t \geq t_0$ be fixed. Inequality (5.18) tells us that $\mathcal{E}$ is nonincreasing on $[t_0, +\infty)$.

$$\theta^2 t^2 \delta(t)\Big(\mathcal{L}_\beta(x(t), \lambda^*) - \mathcal{L}_\beta(x^*, \lambda(t))\Big) + \|v(t)\|^2 + \frac{\xi}{2}\left\|\left(x(t), \lambda(t)\right) - (x^*, \lambda^*)\right\|^2 \leq \mathcal{E}(t_0). \tag{5.19}$$

Assuming $\alpha > 3$ and $\frac{1}{2} \geq \theta > \frac{1}{\alpha - 1}$, we immediately see $\xi > 0$. From (5.19) we obtain

$$\left\|\left(x(t), \lambda(t)\right) - (x^*, \lambda^*)\right\|^2 \leq \frac{2\mathcal{E}(t_0)}{\xi}, \tag{5.20}$$

and

$$\|v(t)\| = \left\|\left(x(t), \lambda(t)\right) - (x^*, \lambda^*) + \theta t\left(\dot{x}(t), \dot{\lambda(t)}\right)\right\| \leq \sqrt{2\mathcal{E}(t_0)}. \tag{5.21}$$

The estimate (5.20) leads to the boundedness of the trajectory. Moreover, applying the triangle inequality and (5.20)-(5.21), we obtain

$$\begin{aligned} t\left\|\left(\dot{x}(t), \dot{\lambda}(t)\right)\right\| & \leq \frac{1}{\theta}\Big(\left\|\left(x(t), \lambda(t)\right) - (x^*, \lambda^*)\right\| + \|v(t)\|\Big) \\ & \leq \frac{1}{\theta}\left(\sqrt{\frac{2\mathcal{E}(t_0)}{\xi}} + \sqrt{2\mathcal{E}(t_0)}\right) = \frac{1}{\theta}\left(\frac{1}{\sqrt{\xi}} + 1\right)\sqrt{2\mathcal{E}(t_0)}, \end{aligned} \tag{5.22}$$

which gives the desired convergence rate. □

5.2.2 Fast convergence rates for the primal-dual gap, the feasibility measure and the objective function value

The following are the main convergence rates results of the paper.

Theorem 5.2.4. *Under Assumption 5.1, let $(x, \lambda) : [t_0, +\infty) \to \mathcal{X} \times \mathcal{Y}$ be a solution of (5.3) and $(x^*, \lambda^*) \in \mathbb{S}$. Then, the following statements are true:*

(i) *For every $t \geq t_0$ it holds*

$$0 \leq \mathcal{L}(x(t), \lambda^*) - \mathcal{L}(x^*, \lambda(t)) \leq \frac{\mathcal{E}(t_0)}{\theta^2 t^2 \delta(t)}; \tag{5.23}$$

(ii) *For every $t \geq t_0$ it holds*

$$\|Ax(t) - b\| \leq \frac{2C_1}{t^2\delta(t)}, \tag{5.24}$$

where

$$C_1 := \sup_{t \geq t_0} t\|\dot{\lambda}(t)\| + (\alpha - 1)\sup_{t \geq t_0} \|\lambda(t) - \lambda^*\| + t_0^2\delta(t_0)\|Ax(t_0) - b\| + t_0\big\|\dot{\lambda}(t_0)\big\|;$$

(iii) *For every $t \geq t_0$ it holds*

$$|f(x(t)) - f(x^*)| \leq \left(\frac{\mathcal{E}(t_0)}{\theta^2} + 2C_1\|\lambda^*\|\right)\frac{1}{t^2\delta(t)}. \tag{5.25}$$

Proof. (i) We have already established that $\mathcal{E}$ is nonincreasing on $[t_0, +\infty)$. Therefore, from the expression for $\mathcal{E}$ and relation (5.9), for every $t \geq t_0$ we deduce

$$\theta^2 t^2\delta(t)\big[\mathcal{L}(x(t), \lambda^*) - \mathcal{L}(x^*, \lambda(t))\big] \leq \mathcal{E}(t_0), \tag{5.26}$$

and the first claim follows.

(ii) From the second line of (5.3), for every $t \geq t_0$ we have

$$t\ddot{\lambda}(t) + \alpha\dot{\lambda}(t) = t\delta(t)\Big(A\big(x(t) + \theta t\dot{x}(t)\big) - b\Big) = t\delta(t)\big(Ax(t) - b\big) + \theta t^2\delta(t)A\dot{x}(t). \tag{5.27}$$

Fix $t \geq t_0$. On the one hand, integration by parts yields

$$\begin{aligned}\int_{t_0}^t \big(s\ddot{\lambda}(s) + \alpha\dot{\lambda}(s)\big)ds &= t\dot{\lambda}(t) - t_0\dot{\lambda}(t_0) - \int_{t_0}^t \dot{\lambda}(s)ds + \alpha\int_{t_0}^t \dot{\lambda}(s)ds \\ &= t\dot{\lambda}(t) - t_0\dot{\lambda}(t_0) + (\alpha - 1)(\lambda(t) - \lambda(t_0)).\end{aligned} \tag{5.28}$$

On the other hand, again integrating by parts leads to

$$\begin{aligned}\int_{t_0}^t s^2\delta(s)A\dot{x}(s)ds &= t^2\delta(t)(Ax(t) - b) - t_0^2\delta(t_0)(Ax(t_0) - b) \\ &\quad - \int_{t_0}^t \big(2s\delta(s) + s^2\dot{\delta}(s)\big)(Ax(s) - b)ds.\end{aligned} \tag{5.29}$$

Now, integrating (5.27) from t_0 to t and using (5.28) and (5.29) gives us

$$\begin{aligned}&t\dot{\lambda}(t) - t_0\dot{\lambda}(t_0) + (\alpha - 1)(\lambda(t) - \lambda(t_0)) \\ &= \int_{t_0}^t s\delta(s)(Ax(s) - b)ds + \theta\int_{t_0}^t s^2\delta(s)A\dot{x}(s)ds \\ &= t^2\delta(t)(Ax(t) - b) - t_0^2\delta(t_0)(Ax(t_0) - b) + \int_{t_0}^t s\big[(1 - 2\theta)\delta(s) - \theta s\dot{\delta}(s)\big](Ax(s) - b)ds \\ &= t^2\delta(t)(Ax(t) - b) - t_0^2\delta(t_0)(Ax(t_0) - b) + \int_{t_0}^t \frac{(1 - 2\theta)\delta(s) - \theta s\dot{\delta}(s)}{s\delta(s)}s^2\delta(s)(Ax(s) - b)ds.\end{aligned} \tag{5.30}$$

It follows that, for every $t \geq t_0$, we have

$$\left\|t^2\delta(t)(Ax(t) - b) + \int_{t_0}^t \frac{(1 - 2\theta)\delta(s) - \theta s\dot{\delta}(s)}{s\delta(s)}s^2\delta(s)(Ax(s) - b)ds\right\| \leq C_1, \tag{5.31}$$

where

$$C_1 = \sup_{t \geq t_0} t\|\dot{\lambda}(t)\| + (\alpha - 1)\sup_{t \geq t_0}\|\lambda(t) - \lambda\,(t_0)\,\| + t_0^2\delta(t_0)\|Ax(t_0) - b\| + t_0\|\dot{\lambda}\,(t_0)\,\| < +\infty,$$

and this quantity is finite in light of (5.22) and (5.20). Now, for $t \geq t_0$ we set

$$g(t) := t^2\delta(t)\|Ax(t) - b\|, \qquad a(t) := \frac{(1 - 2\theta)\delta(t) - \theta t\dot{\delta}(t)}{t\delta(t)}$$

and we apply Lemma 2.3.7 to deduce that, for every $t \geq t_0$,

$$t^2\delta(t)\|Ax(t) - b\| \leq 2C_1. \tag{5.32}$$

(iii) For every $t \geq t_0$ we have

$$\mathcal{L}(x(t), \lambda^*) - \mathcal{L}(x^*, \lambda(t)) = f(x(t)) - f(x^*) + \langle \lambda^*, Ax(t) - b \rangle.$$

Therefore, from using (5.32) and (5.26) we obtain, for every $t \geq t_0$,

$$\begin{aligned} |f(x(t)) - f(x^*)| &\leq \mathcal{L}(x(t), \lambda^*) - \mathcal{L}(x^*, \lambda(t)) + \|\lambda^*\|\|Ax(t) - b\| \\ &\leq \left(\frac{\mathcal{E}(t_0)}{\theta^2} + 2C_1\|\lambda^*\| \right) \frac{1}{t^2\delta(t)}, \end{aligned}$$

which leads to the last statement. □

Some comments regarding the previous proof and results are in order.

Remark 5.2.5. The proof we provided here is significantly shorter than the one derived in [43] thanks to Lemma 2.3.7. This Lemma is the one used in [64] for showing the fast convergence to zero of the feasibility measure, although the authors study a different dynamical system. On the other hand, when $\delta(t) \equiv 1$, the result in [43] is more robust than the one we obtain here, as it gives the $\mathcal{O}\left(\frac{1}{t^2}\right)$ rates for the sum of primal-dual gap and feasibility measure, rather than each one individually. It also allows us to focus only on the energy function defined with respect to a primal-dual optimal solution $(x^*, \lambda^*) \in \mathbb{S}$, rather than on an arbitrary feasible point $(z, \mu) \in \mathbb{F} \times \mathcal{Y}$ as in [43].

Remark 5.2.6. Here are some remarks comparing our rates of convergence to those in [18, 66].

- *Primal-dual gap*: According to (5.23), the primal-dual gap exhibits the convergence rate

  $$\mathcal{L}\big(x(t), \lambda^*\big) - \mathcal{L}\big(x^*, \lambda(t)\big) = \mathcal{O}\left(\frac{1}{t^2\delta(t)}\right)$$

  as $t \to +\infty$, which coincides with the findings of [18, 66].

- *Feasibility measure*: According to (5.24), we have

  $$\|Ax(t) - b\| = \mathcal{O}\left(\frac{1}{t^2\delta(t)}\right)$$

  as $t \to +\infty$, which improves the rate $\mathcal{O}\left(\frac{1}{t\sqrt{\delta(t)}}\right)$ reported in [18, 66].

- *Functional values*: Relation (5.25) tells us that

  $$|f(x(t)) - f(x^*)| = \mathcal{O}\left(\frac{1}{t^2\delta(t)}\right)$$

  as $t \to +\infty$. In [18], only the upper bound presents this order of convergence. The lower bound obtained is of order $\mathcal{O}\left(\frac{1}{t\sqrt{\delta(t)}}\right)$ as $t \to +\infty$. In [66], there are no comments on the rate attained by the functional values in the case of a general time rescaling parameter.

- *Further comparisons with* [66]: in [66, Theorem 2.16], unlike the preceding result [66, Theorem 2.15], the authors produce a rate of $\mathcal{O}\left(\frac{1}{t^{1/\theta}}\right)$ as $t \to +\infty$ for $\|Ax(t) - b\|$ and $|f(x(t)) - f(x^*)|$, provided the time rescaling parameter is chosen to be $\delta(t) = \delta_0 t^{\frac{1}{\theta}-2}$, for some $\delta_0 > 0$. This choice comes from the solution to the differential equation

  $$\frac{t\dot{\delta}(t)}{\delta(t)} = \frac{1 - 2\theta}{\theta},$$

and thus is covered by our results when the growth condition (5.4) holds with equality. The rates are consequently

$$\mathcal{O}\left(\frac{1}{t^2 \cdot \delta_0 t^{\frac{1}{\theta}-2}}\right) = \mathcal{O}\left(\frac{1}{t^{\frac{1}{\theta}}}\right)$$

as $t \to +\infty$. In this setting, if we wish to obtain fast convergence rates, we need to choose a small θ. In light of Assumption 5.1, where we have $\frac{1}{2} \geq \theta \geq \frac{1}{\alpha-1}$, this can be achieved by taking α large enough. Such behavior can also be seen in [18] and in the unconstrained case [15, 19].

5.3 Weak convergence of the trajectory to a primal-dual solution

In this section we will show that the solutions to (5.3) weakly converge to an element of $\mathbb{S}$. The fact that $\delta(t)$ enters the convergence rate statement suggests that one can benefit from this time rescaling function when it is at least nondecreasing on $[t_0, +\infty)$. We are, in fact, going to need this condition when showing trajectory convergence.

Assumption 5.2. In (5.3), assume that ∇f is L-Lipschitz continuous for some $L > 0$ and that $\delta : [t_0, +\infty) \to \mathbb{R}_{++}$ is continuously differentiable and nondecreasing. Moreover, suppose that the parameters α, β, θ and the function δ satisfy

$$\alpha > 3, \quad \beta \geq 0, \quad \frac{1}{2} > \theta > \frac{1}{\alpha - 1} \quad \text{and} \quad \sup_{t \geq t_0} \frac{t\dot{\delta}(t)}{\delta(t)} < \frac{1-2\theta}{\theta}.$$

Assumption 5.2 entails the existence of $C_2 > 0$ such that, for every $t \geq t_0$,

$$\frac{t\dot{\delta}(t)}{\delta(t)} + C_2 \leq \frac{1-2\theta}{\theta}. \tag{5.33}$$

and therefore it follows further from the nondecreasing property of δ that for every $t \geq t_0$

$$0 < C_2\delta(t_0) \leq C_2\delta(t) \leq (1-2\theta)\delta(t) - \theta t\dot{\delta}(t). \tag{5.34}$$

Moreover, from (5.33), for every $t \geq t_0$, we have

$$0 < C_2 \leq \frac{1-2\theta}{\theta} - \frac{t\dot{\delta}(t)}{\delta(t)} = \frac{\sigma(t)}{\delta(t)},$$

which for every $t \geq t_0$ gives

$$\delta(t) \leq \frac{\sigma(t)}{C_2}. \tag{5.35}$$

The additional Lipschitz continuity condition of ∇f and the fact that δ is nondecreasing give rise to the following two essential integrability statements.

Proposition 5.3.1. *Under Assumption 5.2, let $(x, \lambda) : [t_0, +\infty) \to \mathcal{X} \times \mathcal{Y}$ be a solution of (5.3) and $(x^*, \lambda^*) \in \mathbb{S}$. Then, it holds*

$$\int_{t_0}^{+\infty} t\delta(t) \left\|\nabla f(x(t)) - \nabla f(x^*)\right\|^2 dt < +\infty \tag{5.36}$$

and

$$\int_{t_0}^{+\infty} t\delta(t)\|Ax(t) - b\|^2 dt < +\infty. \tag{5.37}$$

Proof. Thanks to ∇f being L-Lipschitz continuous, we can use the improved gradient inequality (see the preliminaries chapter, also [104, Lemma 4]) to refine relation (5.14) in the proof of Lemma 5.2.1

$$\begin{aligned}
&-\left\langle (\nabla_x \mathcal{L}_\beta(x(t),\lambda^*),0), \big(x(t),\lambda(t)\big)-(x^*,\lambda^*)\right\rangle \\
&= \langle \nabla f(x(t)), x^*-x(t)\rangle + \langle A^*\lambda^*, x^*-x(t)\rangle + \beta \langle A^*(Ax(t)-b), x^*-x(t)\rangle \\
&\leq -(f(x(t))-f(x^*)) - \frac{1}{2L}\|\nabla f(x(t))-\nabla f(x^*)\|^2 - \langle \lambda^*, Ax(t)-b\rangle - \beta\|Ax(t)-b\|^2 \\
&= -(\mathcal{L}_\beta(x(t),\lambda^*)-\mathcal{L}_\beta(x^*,\lambda(t))) - \frac{1}{2L}\|\nabla f(x(t))-\nabla f(x^*)\|^2 - \frac{\beta}{2}\|Ax(t)-b\|^2.
\end{aligned}$$

Consequently, combining this inequality with (5.13) yields, for every $t \geq t_0$

$$\begin{aligned}
\frac{d}{dt}\mathcal{E}(t) \leq &-\theta^2 t\sigma(t)\left(\mathcal{L}_\beta(x(t),\lambda^*)-\mathcal{L}_\beta(x^*,\lambda(t))\right) - \xi\theta t\|\left(\dot{x}(t),\dot{\lambda}(t)\right)\|^2 \\
&- \frac{\theta t\delta(t)}{2L}\|\nabla f(x(t))-\nabla f(x^*)\|^2 - \frac{\theta\beta t\delta(t)}{2}\|Ax(t)-b\|^2 \\
\leq &-\frac{\theta t\delta(t)}{2L}\|\nabla f(x(t))-\nabla f(x^*)\|^2.
\end{aligned}$$

Integration of this inequality produces (5.36).

The finiteness of the second integral is only entailed by (5.16) when $\beta > 0$. For the general case $\beta \geq 0$, we use (5.24) and the fact that δ is nondecreasing on $[t_0,+\infty)$ to obtain

$$\int_{t_0}^{+\infty} t\delta(t)\|Ax(t)-b\|^2 dt \leq 4C_1^2\int_{t_0}^{+\infty}\frac{1}{t^3\delta(t)}dt \leq \frac{4C_1^2}{\delta(t_0)}\int_{t_0}^{+\infty}\frac{1}{t^3}dt < +\infty,$$

and the proof is complete. ☐

Now, for a given primal-dual solution $(x^*,\lambda^*) \in \mathbb{S}$, we define the following mappings on $[t_0,+\infty)$:

$$W(t) := \delta(t)\big[\mathcal{L}_\beta(x(t),\lambda^*)-\mathcal{L}_\beta(x^*,\lambda(t))\big] + \frac{1}{2}\|\big(\dot{x}(t),\dot{\lambda}(t)\big)\|^2 \geq 0, \tag{5.38}$$

$$\varphi(t) := \frac{1}{2}\left\|\big(x(t),\lambda(t)\big)-(x^*,\lambda^*)\right\|^2 \geq 0. \tag{5.39}$$

The following are three technical lemmas that we will need in this section. Lemma 5.3.3 guarantees that the first condition of Opial's Lemma is met.

Lemma 5.3.2. *Under Assumption 5.2, let $(x,\lambda) : [t_0,+\infty) \to \mathcal{X}\times\mathcal{Y}$ a solution of (5.3) and $(x^*,\lambda^*) \in \mathbb{S}$. Then, the following inequality holds for every $t \geq t_0$:*

$$\ddot{\varphi}(t) + \frac{\alpha}{t}\dot{\varphi}(t) + \theta t\dot{W}(t) + \frac{\delta(t)}{2L}\|\nabla f(x(t))-\nabla f(x^*)\|^2 + \frac{\beta\delta(t)}{2}\|Ax(t)-b\|^2 \leq 0. \tag{5.40}$$

Proof. Let $t \geq t_0$ be fixed. Differentiating W with respect to time yields

$$\begin{aligned}
\dot{W}(t) = &\dot{\delta}(t)\Big[\mathcal{L}_\beta(x(t),\lambda^*)-\mathcal{L}_\beta(x^*,\lambda(t))\Big] \\
&+ \delta(t)\Big[\langle\nabla_x\mathcal{L}_\beta(x(t),\lambda^*),\dot{x}(t)\rangle - \big\langle\nabla_\lambda\mathcal{L}_\beta(x^*,\lambda(t)),\dot{\lambda}(t)\big\rangle\Big] + \langle\ddot{x}(t),\dot{x}(t)\rangle + \langle\ddot{\lambda}(t),\dot{\lambda}(t)\rangle.
\end{aligned}$$

Recall the formulas for the gradients of $\mathcal{L}$

$$\begin{aligned}
&\nabla_x\mathcal{L}_\beta(x(t),\lambda^*) = \nabla f(x(t)) + A^*\lambda^* + \beta A^*(Ax(t)-b), \\
&\nabla_\lambda\mathcal{L}_\beta(x^*,\lambda(t)) = Ax^*-b = 0,
\end{aligned}$$

since $x^* \in \mathbb{F}$. Plugging this into the expression for $\dot{W}(t)$ gives us

$$\begin{aligned}
\dot{W}(t) = &\dot{\delta}(t)\Big[\mathcal{L}_\beta(x(t),\lambda^*)-\mathcal{L}_\beta(x^*,\lambda(t))\Big] \\
&+ \delta(t)\left\langle\nabla_x\mathcal{L}_\beta(x(t),\lambda(t)+\theta t\dot{\lambda}(t)),\dot{x}(t)\right\rangle + \langle\ddot{x}(t),\dot{x}(t)\rangle
\end{aligned}$$

$$
\begin{aligned}
&- \delta(t) \left\langle \lambda(t) - \lambda^* + \theta t \dot{\lambda}(t), A\dot{x}(t) \right\rangle \\
&- \delta(t) \left\langle \nabla_\lambda \mathcal{L}_\beta(x(t) + \theta t \dot{x}(t), \lambda(t)), \dot{\lambda(t)} \right\rangle + \langle \ddot{\lambda}(t), \dot{\lambda}(t) \rangle \\
&+ \delta(t) \left\langle Ax(t) - b + \theta t A\dot{x}(t), \dot{\lambda}(t) \right\rangle .
\end{aligned}
$$

By regrouping and using (5.3), we arrive at

$$
\begin{aligned}
\dot{W}(t) = \dot{\delta}(t) \Big[\mathcal{L}_\beta(x(t), \lambda^*) - \mathcal{L}_\beta(x^*, \lambda(t)) \Big] - \frac{\alpha}{t} \|\dot{x}(t)\|^2 - \frac{\alpha}{t} \|\dot{\lambda}(t)\|^2 \\
- \delta(t) \left\langle \lambda(t) - \lambda^*, A\dot{x}(t) \right\rangle + \delta(t) \big\langle Ax(t) - b, \dot{\lambda}(t) \big\rangle.
\end{aligned} \tag{5.41}
$$

On the other hand, by the chain rule, we have

$$
\begin{aligned}
\dot{\varphi}(t) &= \langle x(t) - x^*, \dot{x}(t) \rangle + \big\langle \lambda(t) - \lambda^*, \dot{\lambda}(t) \big\rangle, \\
\ddot{\varphi}(t) &= \langle x(t) - x^*, \ddot{x}(t) \rangle + \|\dot{x}(t)\|^2 + \big\langle \lambda(t) - \lambda^*, \ddot{\lambda}(t) \big\rangle + \big\| \dot{\lambda}(t) \big\|^2.
\end{aligned}
$$

By combining these relations, (5.3) and the fact that $x^* \in \mathbb{F}$, we get

$$
\begin{aligned}
\ddot{\varphi}(t) + \frac{\alpha}{t} \dot{\varphi}(t) = &\left\langle x(t) - x^*, \ddot{x}(t) + \frac{\alpha}{t} \dot{x}(t) \right\rangle + \left\langle \lambda(t) - \lambda^*, \ddot{\lambda}(t) + \frac{\alpha}{t} \dot{\lambda}(t) \right\rangle + \|\dot{x}(t)\|^2 + \|\dot{\lambda}(t)\|^2 \\
= &- \big\langle x(t) - x^*, \delta(t) \nabla_x \mathcal{L}_\beta(x(t), \lambda(t) + \theta t \dot{\lambda}(t)) \big\rangle \\
&+ \big\langle \lambda(t) - \lambda^*, \delta(t) \nabla_\lambda \mathcal{L}_\beta(x(t) + \theta t \dot{\lambda}(t), \lambda(t)) \big\rangle + \|\dot{x}(t)\|^2 + \|\dot{\lambda}(t)\|^2 \\
= &- \big\langle x(t) - x^*, \delta(t) \nabla_x \mathcal{L}_\beta(x(t), \lambda^*) \big\rangle - \big\langle Ax(t) - b, \delta(t) \big(\lambda(t) - \lambda^* + \theta t \dot{\lambda}(t) \big) \big\rangle \\
&+ \big\langle \lambda(t) - \lambda^*, \delta(t) \big(Ax(t) - b + \theta t A\dot{x}(t) \big) \big\rangle + \|\dot{x}(t)\|^2 + \|\dot{\lambda}(t)\|^2 \\
= &- \delta(t) \big\langle x(t) - x^*, \nabla_x \mathcal{L}_\beta(x(t), \lambda^*) \big\rangle - \theta t \delta(t) \big\langle Ax(t) - b, \dot{\lambda}(t) \big\rangle \\
&+ \theta t \delta(t) \big\langle \lambda(t) - \lambda^*, A\dot{x}(t) \big\rangle + \|\dot{x}(t)\|^2 + \|\dot{\lambda}(t)\|^2
\end{aligned} \tag{5.42}
$$

The Lipschitz continuity of ∇f entails

$$
\begin{aligned}
&- \langle x(t) - x^*, \nabla_x \mathcal{L}_\beta(x(t), \lambda^*) \rangle \\
&= -\langle x(t) - x^*, \nabla f(x(t)) \rangle - \big\langle x(t) - x^*, A^* \lambda^* \big\rangle - \beta \|Ax(t) - b\|^2 \\
&\leq -(f(x(t)) - f(x^*)) - \frac{1}{2L} \|\nabla f(x(t)) - \nabla f(x^*)\|^2 - \langle \lambda^*, Ax(t) - b \rangle - \beta \|Ax(t) - b\|^2 \\
&= -\Big(\mathcal{L}_\beta(x(t), \lambda^*) - \mathcal{L}_\beta(x^*, \lambda(t)) \Big) - \frac{1}{2L} \|\nabla f(x(t)) - \nabla f(x^*)\|^2 - \frac{\beta}{2} \|Ax(t) - b\|^2.
\end{aligned}
$$

This, together with (5.42), implies

$$
\begin{aligned}
\ddot{\varphi}(t) + \frac{\alpha}{t} \dot{\varphi}(t) \leq &- \delta(t) \Big(\mathcal{L}_\beta(x(t), \lambda^*) - \mathcal{L}_\beta(x^*, \lambda(t)) \Big) - \theta t \delta(t) \big\langle Ax(t) - b, \dot{\lambda}(t) \big\rangle \\
&+ \theta t \delta(t) \big\langle \lambda(t) - \lambda^*, A\dot{x}(t) \big\rangle + \|\dot{x}(t)\|^2 + \|\dot{\lambda}(t)\|^2 \\
&- \frac{\delta(t)}{2L} \|\nabla f(x(t)) - \nabla f(x^*)\|^2 - \frac{\beta \delta(t)}{2} \|Ax(t) - b\|^2.
\end{aligned} \tag{5.43}
$$

Multiplying (5.41) by $\theta t > 0$ and then adding the result to (5.43) yields

$$
\begin{aligned}
\ddot{\varphi}(t) + \frac{\alpha}{t} \dot{\varphi}(t) + \theta t \dot{W}(t) = &- \big(\delta(t) - \theta t \dot{\delta}(t) \big) \Big(\mathcal{L}_\beta(x(t), \lambda^*) - \mathcal{L}_\beta(x^*, \lambda(t)) \Big) \\
&- \frac{\delta(t)}{2L} \|\nabla f(x(t)) - \nabla f(x^*)\|^2 - \frac{\beta \delta(t)}{2} \|Ax(t) - b\|^2 \\
&+ (1 - \theta \alpha) \| \big(\dot{x}(t), \dot{\lambda}(t) \big) \|^2 \\
\leq &- \frac{\delta(t)}{2L} \|\nabla f(x(t)) - \nabla f(x^*)\|^2 - \frac{\beta \delta(t)}{2} \|Ax(t) - b\|^2,
\end{aligned}
$$

where the last inequality follows from Assumption 5.2:

$$
\begin{gathered}
1 - \theta \alpha \leq -\theta < 0, \\
-\delta(t) + \theta t \dot{\delta}(t) \leq (2\theta - 1) \delta(t) + \theta t \dot{\delta}(t) \leq 0.
\end{gathered}
$$

The desired result then follows after some rearranging. □

Lemma 5.3.3. *Under Assumption 5.2, let $(x, \lambda) : [t_0, +\infty) \to \mathcal{X} \times \mathcal{Y}$ be a solution to (5.3) and $(x^*, \lambda^*) \in \mathbb{S}$. Then, the positive part $[\dot{\varphi}]_+$ of $\dot{\varphi}$ belongs to $\mathbb{L}^1([t_0, +\infty))$ and $\lim_{t\to+\infty} \varphi(t)$ exists.*

Proof. For any $t \geq t_0$, we multiply (5.40) by t and drop the last two norm squared terms to obtain

$$t\ddot{\varphi}(t) + \alpha\dot{\varphi}(t) + \theta t^2 \dot{W}(t) \leq 0.$$

Recall from (5.38) that for every $t \geq t_0$ we have

$$tW(t) = t\delta(t)\big[\mathcal{L}_\beta(x(t), \lambda^*) - \mathcal{L}_\beta(x^*, \lambda(t))\big] + \frac{t}{2}\|\big(\dot{x}(t), \dot{\lambda}(t)\big)\|^2. \tag{5.44}$$

On the one hand, according to (5.17), the second summand of the previous expression belongs to $\mathbb{L}^1[t_0, +\infty)$. On the other hand, using (5.35) and (5.15), we assert that

$$\int_{t_0}^{+\infty} t\delta(t)\big[\mathcal{L}_\beta(x(t), \lambda^*) - \mathcal{L}_\beta(x^*, \lambda(t))\big]dt \leq \frac{1}{C_2}\int_{t_0}^{+\infty} t\sigma(t)\big[\mathcal{L}_\beta(x(t), \lambda^*) - \mathcal{L}_\beta(x^*, \lambda(t))\big]dt < +\infty.$$

Hence, the first summand of (5.44) also belongs to $\mathbb{L}^1([t_0, +\infty))$, which implies that the mapping $t \mapsto tW(t)$ belongs to $\mathbb{L}^1[t_0, +\infty)$ as well. For achieving the desired conclusion, we make use of Lemma 2.3.9 with $\phi := \varphi$ and $w := \theta W$. $\square$

Lemma 5.3.4. *Under Assumption 5.2, let $(x, \lambda) : [t_0, +\infty) \to \mathcal{X} \times \mathcal{Y}$ be a solution to (5.3) and $(x^*, \lambda^*) \in \mathbb{S}$. Then, the following inequality holds for every $t \geq t_0$:*

$$\begin{aligned}
&\frac{\alpha}{t\delta(t)}\frac{d}{dt}\|\big(\dot{x}(t), \dot{\lambda}(t)\big)\|^2 + 2\left\langle \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t), A^*(\lambda(t) - \lambda^*)\right\rangle \\
&+ \theta\frac{d}{dt}\Big(t\delta(t)\,\|A^*(\lambda(t) - \lambda^*)\|^2\Big) + \big((1-\theta)\delta(t) - \theta t\dot{\delta}(t)\big)\,\|A^*(\lambda(t) - \lambda^*)\|^2 \\
&\leq \delta(t)\Big[2\|\nabla f(x(t)) - \nabla f(x^*)\|^2 + \big(2\beta^2\|A\|^2 + 1\big)\,\|Ax(t) - b\|^2\Big].
\end{aligned}$$

Proof. Let $t \geq t_0$ be fixed. From (5.3) and the fact that $A^*\lambda^* = -\nabla f(x^*)$, we have

$$\begin{aligned}
&\delta^2(t)\big\|\nabla f(x(t)) - \nabla f(x^*) + \beta A^*(Ax(t) - b)\big\|^2 \\
={}& \left\|\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \delta(t)A^*\big(\lambda(t) - \lambda^* + \theta t\dot{\lambda}(t)\big)\right\|^2 \\
={}& \left\|\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t)\right\|^2 + \delta^2(t)\big\|A^*\big(\lambda(t) - \lambda^* + \theta t\dot{\lambda}(t)\big)\big\|^2 + 2\delta(t)\left\langle \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t), A^*\big(\lambda(t) - \lambda^*\big)\right\rangle \\
&+ 2\theta t\delta(t)\big\langle \ddot{x}(t), A^*\dot{\lambda}(t)\big\rangle + 2\alpha\theta t\delta(t)\big\langle \dot{x}(t), A^*\dot{\lambda}(t)\big\rangle.
\end{aligned} \tag{5.45}$$

Again using (5.3) yields

$$\begin{aligned}
\delta^2(t)\|Ax\left((t)\right) - b\|^2 &= \left\|\ddot{\lambda}(t) + \frac{\alpha}{t}\dot{\lambda}(t) - \theta t\delta(t)A\dot{x}(t)\right\|^2 \\
&= \left\|\ddot{\lambda}(t) + \frac{\alpha}{t}\dot{\lambda}(t)\right\|^2 + \theta^2 t^2\delta^2(t)\|A\dot{x}(t)\|^2 \\
&\quad - 2\theta t\delta(t)\big\langle \ddot{\lambda}(t), A\dot{x}(t)\big\rangle - 2\alpha\theta\delta(t)\big\langle \dot{\lambda}(t), A\dot{x}(t)\big\rangle.
\end{aligned} \tag{5.46}$$

Adding (5.45) and (5.46) together produces

$$\begin{aligned}
&\delta^2(t)\big\|\nabla f(x(t)) - \nabla f(x^*) + \beta A^*(Ax(t) - b)\big\|^2 + \delta^2(t)\|Ax(t) - b\|^2 \\
={}& \left\|\big(\ddot{x}(t), \ddot{\lambda}(t)\big) + \frac{\alpha}{t}\big(\dot{x}(t), \dot{\lambda}(t)\big)\right\|^2 + \delta^2(t)\big\|A^*\big(\lambda(t) - \lambda^* + \theta t\dot{\lambda}(t)\big)\big\|^2 + \theta^2 t^2\delta^2(t)\|A\dot{x}(t)\|^2 \\
&+ 2\theta t\delta(t)\big\langle \ddot{x}(t), A^*\dot{\lambda}(t)\big\rangle - 2\theta t\delta(t)\big\langle \ddot{\lambda}(t), A\dot{x}(t)\big\rangle + 2\delta(t)\left\langle \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t), A^*(\lambda(t) - \lambda^*)\right\rangle.
\end{aligned} \tag{5.47}$$

On the one hand, we have

$$\theta^2 t^2\delta^2(t)\|A\dot{x}(t)\|^2 + 2\theta t\delta(t)\big\langle \ddot{x}(t), A^*\dot{\lambda}(t)\big\rangle - 2\theta t\delta(t)\big\langle \ddot{\lambda}(t), A\dot{x}(t)\big\rangle$$

$$
\begin{aligned}
&= \theta^2 t^2 \delta^2(t) \big\| \big(A^* \dot\lambda(t), -A\dot x(t)\big) \big\|^2 - \theta^2 t^2 \delta^2(t) \big\| A^* \dot\lambda(t) \big\|^2 + 2\theta t \delta(t) \big\langle \big(\ddot x(t), \ddot\lambda(t)\big), \big(A^*\dot\lambda(t), -A\dot x(t)\big) \big\rangle \\
&= \ - \big\| \big(\ddot x(t), \ddot\lambda(t)\big) \big\|^2 + \big\| \big(\ddot x(t), \ddot\lambda(t)\big) + \theta t \delta(t) A^* \big(A^* \dot\lambda(t), -A\dot x(t)\big) \big\|^2 - \theta^2 t^2 \delta^2(t) \big\| A^* \dot\lambda(t) \big\|^2 \\
&\geq \ - \big\| \big(\ddot x(t), \ddot\lambda(t)\big) \big\|^2 - \theta^2 t^2 \delta^2(t) \big\| A^* \dot\lambda(t) \big\|^2.
\end{aligned} \tag{5.48}
$$

On the other hand, it holds

$$
\begin{aligned}
&\Big\| \big(\ddot x(t), \ddot\lambda(t)\big) + \frac{\alpha}{t} \big(\dot x(t), \dot\lambda(t)\big) \Big\|^2 - \big\| \big(\ddot x(t), \ddot\lambda(t)\big) \big\|^2 \\
= \ &\frac{\alpha^2}{t^2} \big\| \big(\dot x(t), \dot\lambda(t)\big) \big\|^2 + 2\frac{\alpha}{t} \big\langle \big(\ddot x(t), \ddot\lambda(t)\big), \big(\dot x(t), \dot\lambda(t)\big) \big\rangle \geq \frac{\alpha}{t} \frac{d}{dt} \big\| \big(\dot x(t), \dot\lambda(t)\big) \big\|^2.
\end{aligned} \tag{5.49}
$$

Moreover,

$$
\begin{aligned}
&\delta(t) \big\| A^* \big(\lambda(t) - \lambda^* + \theta t \dot\lambda(t)\big) \big\|^2 - \theta^2 t^2 \delta(t) \big\| A^* \dot\lambda(t) \big\|^2 \\
= \ &\delta(t) \big\| A^* (\lambda(t) - \lambda^*) \big\|^2 + 2\theta t \delta(t) \big\langle AA^*(\lambda(t) - \lambda^*), \dot{\lambda(t)} \big\rangle \\
= \ &\big((1-\theta)\delta(t) - \theta t \dot\delta(t)\big) \big\| A^* \big(\lambda(t) - \lambda^*\big) \big\|^2 + \theta \delta(t) \big\| A^*(\lambda(t) - \lambda^*) \big\|^2 \\
&\quad + \theta t \dot\delta(t) \| A^* \left(\lambda(t) - \lambda^*\right) \|^2 + \theta \frac{d}{dt} \| A^* (\lambda(t) - \lambda^*) \|^2 \\
= \ &\big((1-\theta)\delta(t) - \theta t \dot\delta(t)\big) \| A^* \left(\lambda(t) - \lambda^*\right) \|^2 + \theta \frac{d}{dt} \Big(t\delta(t) \big\| A^*(\lambda(t) - \lambda^*) \big\|^2 \Big).
\end{aligned} \tag{5.50}
$$

Now, using (5.48), (5.49) and (5.50) in (5.47) yields

$$
\begin{aligned}
&\delta^2(t)(t) \big\| \nabla f(x(t)) - \nabla f(x^*) + \beta A^*(Ax(t) - b) \big\|^2 + \delta^2(t) \| Ax\left((t)\right) - b \|^2 \\
\geq \ &\Big\| \big(\ddot x(t), \ddot\lambda(t)\big) + \frac{\alpha}{t} \big(\dot x(t), \dot\lambda(t)\big) \Big\|^2 + \delta^2(t) \big\| A^* \big(\lambda(t) - \lambda^* + \theta t \dot\lambda(t)\big) \big\|^2 \\
&\quad - \big\| \big(\ddot x(t), \ddot\lambda(t)\big) \big\|^2 - \theta^2 t^2 \delta^2(t) \big\| A^* \dot\lambda(t) \big\|^2 + 2\delta(t) \left\langle \ddot x(t) + \frac{\alpha}{t} \dot x(t), A^*(\lambda(t) - \lambda^*) \right\rangle \\
\geq \ &\frac{\alpha}{t} \frac{d}{dt} \big\| \big(\dot x(t), \dot\lambda(t)\big) \big\|^2 + 2\delta(t) \left\langle \ddot x(t) + \frac{\alpha}{t} \dot x(t), A^*(\lambda(t) - \lambda^*) \right\rangle \\
&\quad + \delta(t) \left[\big((1-\theta)\delta(t) - \theta t \dot\delta(t)\big) \big\| A^*(\lambda(t) - \lambda^*) \big\|^2 + \theta \frac{d}{dt} \Big(t\delta(t) \| A^* \left(\lambda(t) - \lambda^*\right) \|^2 \Big) \right]
\end{aligned}
$$

Finally, since

$$
\begin{aligned}
&\big\| \nabla f(x(t)) - \nabla f(x^*) + \beta A^*(Ax(t) - b) \big\|^2 + \| Ax\left((t)\right) - b \|^2 \\
\leq \ &2 \| \nabla f(x(t)) - \nabla f(x^*) \|^2 + (2\beta \|A\|^2 + 1) \| Ax(t) - b \|^2,
\end{aligned}
$$

the conclusion follows after dividing the inequality by $\delta(t)$. □

Lemma 5.3.5. *Under Assumption 5.2, let $(x, \lambda) : [t_0, +\infty) \to \mathcal{X} \times \mathcal{Y}$ be a solution to (5.3) and $(x^*, \lambda^*) \in \mathbb{S}$. Then, for every $t \geq t_0$ it holds*

$$
\begin{aligned}
&\theta t^{\alpha+1} \delta(t) \| A^* \left(\lambda(t) - \lambda^*\right) \|^2 \\
\leq \ &- t^\alpha \dot\varphi(t) + \int_{t_0}^t s^\alpha V(s) ds + \int_{t_0}^t s^\alpha \Big[\big(\theta(\alpha+1) - 1\big)\delta(t) + \theta s \dot\delta(s) \Big] \| A^*(\lambda(s) - \lambda^*) \|^2 ds \\
&- 2t^\alpha \big\langle \dot x(t), A^*(\lambda(t) - \lambda^*) \big\rangle + C_5,
\end{aligned} \tag{5.51}
$$

where, for $t \geq t_0$,

$$
\begin{aligned}
V(t) := \theta(\alpha+1) W(t) + \left(\frac{\alpha(\alpha-1)}{t_0^2 \delta(_0)} + \|A\| \right) \big\| \big(\dot x(t), \dot\lambda(t)\big) \big\|^2 \\
+ C_3 \delta(t) \| \nabla f(x(t)) - \nabla f(x^*) \|^2 + C_4 \delta(t) \| Ax(t) - b \|^2,
\end{aligned}
$$

for certain nonnegative constants C_3, C_4 and C_5.

Proof. For every $t \geq t_0$, by summing up the two inequalities produced by Lemmas 5.3.2 and 5.3.4 we deduce that

$$\begin{aligned}
&\ddot{\varphi}(t) + \frac{\alpha}{t}\dot{\varphi}(t) + \theta t\dot{W}(t) + \frac{\alpha}{t\delta(t)}\left\|\left(\dot{x}(t), \dot{\lambda}(t)\right)\right\|^2 \\
&+ \theta\frac{d}{dt}\left(t\delta(t)\left\|A^*(\lambda(t)-\lambda^*)\right\|^2\right) + 2\left\langle \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t), A^*(\lambda(t)-\lambda^*)\right\rangle \\
\leq & -\left((1-\theta)\delta(t) - \theta t\dot{\delta}(t)\right)\left\|A^*(\lambda(t)-\lambda^*)\right\|^2 + \left(2 - \frac{1}{2L}\right)\delta(t)\|\nabla f(x(t)) - \nabla f(x^*)\|^2 \\
&+ \left(2\beta^2\|A\|^2 + 1 - \frac{\beta}{2}\right)\delta(t)\|Ax(t)-b\|^2 \\
\leq & -\left((1-\theta)\delta(t) - \theta t\dot{\delta}(t)\right)\left\|A^*(\lambda(t)-\lambda^*)\right\|^2 + C_3\delta(t)\|\nabla f(x(t)) - \nabla f(x^*)\|^2 \\
&+ C_4\delta(t)\|Ax(t)-b\|^2,
\end{aligned} \tag{5.52}$$

where

$$C_3 := \left[2 - \frac{1}{2L}\right]_+ \geq 0 \quad \text{and} \quad C_4 := \left[2\beta^2\|A\|^2 + 1 - \frac{\beta}{2}\right]_+ \geq 0.$$

Multiplying (5.52) by t^α and integrating from t_0 to t, we obtain

$$\begin{aligned}
&I_1(t) + \theta I_2(t) + \alpha I_3(t) + \theta I_4(t) + 2I_5(t) \\
\leq & \int_{t_0}^t s^\alpha\left((1-\theta)\delta(s) - \theta s\dot{\delta}(s)\right)\left\|A^*(\lambda(s)-\lambda^*)\right\|^2 ds + C_3\int_{t_0}^t s^\alpha\delta(s)\|\nabla f(x(s)) - \nabla f(x^*)\|^2 ds \\
&+ C_4\int_{t_0}^t s^\alpha\delta(s)\|Ax(s)-b\|^2 ds,
\end{aligned} \tag{5.53}$$

where

$$\begin{aligned}
I_1(t) &:= \int_{t_0}^t \left(s^\alpha\ddot{\varphi}(s) + \alpha s^{\alpha-1}\dot{\varphi}(s)\right)ds, \\
I_2(t) &:= \int_{t_0}^t s^{\alpha+1}\dot{W}(s)ds, \\
I_3(t) &:= \int_{t_0}^t \frac{s^{\alpha-1}}{\delta(s)}\frac{d}{ds}\left\|\left(\dot{x}(s), \dot{\lambda}(s)\right)\right\|^2 ds, \\
I_4(t) &:= \int_{t_0}^t s^\alpha\frac{d}{ds}\left(s\delta(s)\left\|A^*(\lambda(s)-\lambda^*)\right\|^2\right)ds, \\
I_5(t) &:= \int_{t_0}^t \left\langle s^\alpha\ddot{x}(s) + \alpha s^{\alpha-1}\dot{x}(s), A^*(\lambda(s)-\lambda^*)\right\rangle ds.
\end{aligned}$$

We will furnish five different inequalities from computing each of these integrals separately. Let $t \geq t_0$ be fixed.

- *The integral $I_1(t)$.* By the chain rule, for $s \geq t_0$ it holds

$$s^\alpha\ddot{\varphi}(s) + \alpha s^{\alpha-1}\dot{\varphi}(s) = \frac{d}{ds}\left(s^\alpha\dot{\varphi}(s)\right),$$

which leads to

$$0 = I_1(t) - t^\alpha\dot{\varphi}(t) + t_0^\alpha\dot{\varphi}(t_0) \leq I_1(t) - t^\alpha\dot{\varphi}(t) + |t_0^\alpha\dot{\varphi}(t_0)|. \tag{5.54}$$

- *The integrals $I_2(t)$ and $I_4(t)$.* Integration by parts gives

$$\begin{aligned}
I_2(t) + I_4(t) = & t^{\alpha+1}W(t) - t_0^{\alpha+1}W(t_0) - (\alpha+1)\int_{t_0}^t s^\alpha W(s)ds + t^{\alpha+1}\delta(t)\|A^*\left(\lambda(t)-\lambda^*\right)\|^2 \\
& - t_0^{\alpha+1}\delta(t_0)\left\|A^*(\lambda(t_0)-\lambda^*)\right\|^2 - \alpha\int_{t_0}^t s^\alpha\delta(s)\|A^*(\lambda(s)-\lambda^*)\|^2 ds,
\end{aligned}$$

and from here

$$t^{\alpha+1}\delta(t)\|A^*\left(\lambda(t)-\lambda^*\right)\|^2 \leq t^{\alpha+1}W(t) + t^{\alpha+1}\delta(t)\|A^*\left(\lambda(t)-\lambda^*\right)\|^2$$

$$= I_2(t) + I_4(t) + t_0^{\alpha+1}W(t_0) + t_0^{\alpha+1}\delta(t_0)\big\|A^*(\lambda(t_0)-\lambda^*)\big\|^2 + (\alpha+1)\int_{t_0}^t s^\alpha W(s)ds$$
$$+ \alpha\int_{t_0}^t s^\alpha\delta(s)\|A^*(\lambda(s)-\lambda^*)\|^2 ds. \tag{5.55}$$

- *The integral $I_3(t)$.* Again by integrating by parts, we get

$$I_3(t) = \frac{t^{\alpha-1}}{\delta(t)}\big\|\big(\dot{x}(t),\dot{\lambda}(t)\big)\big\|^2 - \frac{t_0^{\alpha-1}}{\delta(t_0)}\big\|\big(\dot{x}(t_0),\dot{\lambda}(t_0)\big)\big\|^2$$
$$- \int_{t_0}^t\left[\frac{(\alpha-1)s^{\alpha-2}\delta(s) - s^{\alpha-1}\dot{\delta}(s)}{\delta^2(s)}\right]\big\|\big(\dot{x}(s),\dot{\lambda}(s)\big)\big\|^2 ds.$$

For $s \geq t_0$, according to Assumption 5.2 we have $\dot{\delta}(s) \geq 0$, hence δ is monotonically increasing and therefore

$$\frac{(\alpha-1)s^{\alpha-2}\delta(s) - s^{\alpha-1}\delta(s)}{\delta^2(s)} \leq \frac{(\alpha-1)s^{\alpha-2}\delta(s)}{\delta^2(s)} \leq \frac{(\alpha-1)s^\alpha}{t_0^2\delta(t_0)}.$$

It follows that

$$0 \leq \frac{t^{\alpha-1}}{\delta(t)}\big\|\big(\dot{x}(t),\dot{\lambda}(t)\big)\big\|^2$$
$$= I_3(t) + \frac{t_0^{\alpha-1}}{\delta(t_0)}\big\|\big(\dot{x}(t_0),\dot{\lambda}(t_0)\big)\big\|^2 + \int_{t_0}^t\left[\frac{(\alpha-1)s^{\alpha-2}\delta(s) - s^{\alpha-1}\dot{\delta}(s)}{\delta^2(s)}\right]\big\|\big(\dot{x}(s),\dot{\lambda}(s)\big)\big\|^2 ds$$
$$\leq I_3(t) + \frac{t_0^{\alpha-1}}{\delta(t_0)}\big\|\big(\dot{x}(t_0),\dot{\lambda}(t_0)\big)\big\|^2 + \frac{\alpha-1}{t_0^2\delta(t_0)}\int_{t_0}^t s^\alpha\big\|\big(\dot{x}(s),\dot{\lambda}(s)\big)\big\|^2 ds. \tag{5.56}$$

- *The integral $I_5(t)$.* Integration by parts entails

$$I_5(t) = \int_{t_0}^t\left\langle\frac{d}{ds}\big(s^\alpha\dot{x}(s)\big), A^*(\lambda(s)-\lambda^*)\right\rangle ds$$
$$= t^\alpha\big\langle\dot{x}(t), A^*(\lambda(t)-\lambda^*)\big\rangle - t_0^\alpha\big\langle\dot{x}(t_0), A^*(\lambda(t_0)-\lambda^*)\big\rangle - \int_{t_0}^t s^\alpha\big\langle\dot{x}(s), A^*\dot{\lambda}(s)\big\rangle ds.$$

By the Cauchy-Schwarz inequality, we deduce that

$$\int_{t_0}^t s^\alpha\langle\dot{x}(s), A^*\dot{\lambda}(s)\rangle ds \leq \|A\|\int_{t_0}^t s^\alpha\|\dot{x}(s)\|\|\dot{\lambda}(s)\|ds \leq \frac{\|A\|}{2}\int_{t_0}^t s^\alpha\Big(\|\dot{x}(s)\|^2 + \|\dot{\lambda}(s)\|^2\Big)ds,$$

and thus

$$0 \leq I_5(t) - t^\alpha\big\langle\dot{x}(t), A^*(\lambda(t)-\lambda^*)\big\rangle + \big|t_0^\alpha\big\langle\dot{x}(t_0), A^*(\lambda(t_0)-\lambda^*)\big\rangle\big| + \frac{\|A\|}{2}\int_{t_0}^t s^\alpha\big\|\big(\dot{x}(s),\dot{\lambda}(s)\big)\big\|^2 ds. \tag{5.57}$$

Now, summing up the inequalities (5.54), (5.55), (5.56), and (5.57), then we proceed to employ (5.53) and obtain

$$\theta t^{\alpha+1}\delta(t)\|A^*\left(\lambda(t)-\lambda^*\right)\|^2$$
$$\leq I_1(t) + \theta I_2(t) + \alpha I_3(t) + \theta I_4(t) + 2I_5 - t^\alpha\dot{\varphi}(t)$$
$$+ \int_{t_0}^t s^\alpha\left[\theta(\alpha+1)W(s) + \left(\frac{\alpha(\alpha-1)}{t_0^2\delta(t_0)} + \|A\|\right)\big\|\big(\dot{x}(s),\dot{\lambda}(s)\big)\big\|^2\right]ds$$
$$+ \theta\alpha\int_{t_0}^t s^\alpha\delta(s)\|A^*(\lambda(s)-\lambda^*)\|^2 ds - 2t^\alpha\big\langle\dot{x}(t), A^*(\lambda(t)-\lambda^*)\big\rangle + C_5$$
$$\leq -t^\alpha\dot{\varphi}(t) + \int_{t_0}^t s^\alpha V(s)ds + \int_{t_0}^t s^\alpha\Big[\big(\theta(\alpha+1)-1\big)\delta(t) + \theta s\dot{\delta}(s)\Big]\|A^*(\lambda(s)-\lambda^*)\|^2 ds$$
$$- 2t^\alpha\big\langle\dot{x}(t), A^*(\lambda(t)-\lambda^*)\big\rangle + C_5,$$

where recall that V was given by

$$V(s) := \theta(\alpha+1)W(s) + \left(\frac{\alpha(\alpha-1)}{t_0^2\delta(_0)} + \|A\|\right)\big\|\big(\dot{x}(s),\dot{\lambda}(s)\big)\big\|^2$$

$$+ C_3\delta(s)\|\nabla f(x(s)) - \nabla f(x^*)\|^2 + C_4\delta(s)\|Ax(s) - b\|^2,$$

and the constant C_5 is given by

$$\begin{aligned} C_5 := t_0^\alpha|\dot{\varphi}(t_0)| + \theta t_0^{\alpha+1}W(t_0) + \alpha\frac{t_0^{\alpha-1}}{\delta(t_0)}\left\|\left(\dot{x}(t_0, \dot{\lambda}(t_0))\right)\right\|^2 \\ + \theta t_0^{\alpha+1}\delta(t_0)\left\|A^*(\lambda(t_0) - \lambda^*)\right\|^2 + 2t_0^\alpha\left|\left\langle\dot{x}(t_0), A^*(\lambda(t_0) - \lambda^*)\right\rangle\right| \geq 0. \end{aligned}$$

We come then to the desired result. □

The following proposition provides us with the main integrability result that will be used for verifying the second condition of Opial's Lemma.

Proposition 5.3.6. *Under Assumption 5.2, let $(x, \lambda) : [t_0, +\infty) \to \mathcal{X} \times \mathcal{Y}$ be a solution to (5.3) and $(x^*, \lambda^*) \in \mathbb{S}$. Then, it holds*

$$\int_{t_0}^{+\infty} t\delta(t)\|A^*\left(\lambda(t) - \lambda^*\right)\|^2 dt < +\infty. \tag{5.58}$$

Proof. We divide (5.51) by t^α, thus obtaining

$$\begin{aligned} \theta t\delta(t)\|A^*(\lambda(t) - \lambda^*)\|^2 \leq & -\dot{\varphi}(t) + \frac{1}{t^\alpha}\int_{t_0}^t s^\alpha V(s)ds \\ & + \frac{1}{t^\alpha}\int_{t_0}^t s^\alpha\left[\left(\theta(\alpha+1) - 1\right)\delta(s) + \theta s\dot{\delta}(s)\right]\|A^*(\lambda(s) - \lambda^*)\|^2 ds \\ & - 2\left\langle\dot{x}(t), A^*(\lambda(t) - \lambda^*)\right\rangle + \frac{C_5}{t^\alpha}. \end{aligned}$$

Now, we integrate this inequality from t_0 to r. We get

$$\begin{aligned} & \theta\int_{t_0}^r t\delta(t)\|A^*\left(\lambda(t) - \lambda^*\right)\|^2 dt \\ \leq & \ \varphi(t_0) - \varphi(r) + \int_{t_0}^r \frac{1}{t^\alpha}\left(\int_{t_0}^t s^\alpha V(s)ds\right)dt \\ & + \int_{t_0}^r \frac{1}{t^\alpha}\left(\int_{t_0}^t s^\alpha\left[\left(\theta(\alpha+1) - 1\right)\delta(s) + \theta s\dot{\delta}(s)\right]\|A^*(\lambda(s) - \lambda^*)\|^2 ds\right)dt \\ & - 2\int_{t_0}^r \langle A\dot{x}(t), \lambda(t) - \lambda^*\rangle dt + C_5\int_{t_0}^r t^\alpha dt. \end{aligned} \tag{5.59}$$

We now recall some important facts. First of all, we have

$$\int_{t_0}^r \frac{1}{t^\alpha}dt \leq \frac{1}{(\alpha-1)t_0^{\alpha-1}}. \tag{5.60}$$

In addition, according to Lemma 2.3.8, it holds

$$\int_{t_0}^r \frac{1}{t^\alpha}\left(\int_{t_0}^t s^\alpha V(s)ds\right)dt \leq \frac{1}{\alpha-1}\int_{t_0}^r tV(t)dt, \tag{5.61}$$

and

$$\begin{aligned} & \int_{t_0}^r \frac{1}{t^\alpha}\left(\int_{t_0}^t s^\alpha\left[\left(\theta(\alpha+1) - 1\right)\delta(s) + \theta s\dot{\delta}(s)\right]\|A^*(\lambda(s) - \lambda^*)\|^2 ds\right)dt \\ \leq & \ \frac{1}{\alpha-1}\int_{t_0}^r \left[\left(\theta(\alpha+1) - 1\right)\delta(t) + \theta t\dot{\delta}(t)\right]\|A^*(\lambda(t) - \lambda^*)\|^2 dt, \end{aligned} \tag{5.62}$$

respectively. Finally, integrating by parts leads to

$$-\int_{t_0}^r \langle A\dot{x}(t), \lambda(t) - \lambda^*\rangle dt$$

$$
\begin{aligned}
&= -\left\langle Ax(r)-b, \lambda(r)-\lambda^*\right\rangle + \left\langle Ax(t_0)-b, \lambda(t_0)-\lambda^*\right\rangle + \int_{t_0}^{r} \langle Ax(t)-b, \dot\lambda(t)\rangle dt \\
&\leq \|Ax(r)-b\|\|\lambda(r)-\lambda^*\| + \|Ax(t_0)-b\|\|\lambda(t_0)-\lambda^*\| + \int_{t_0}^{r} \langle Ax(t)-b, \dot\lambda(t)\rangle dt \\
&\leq \sup_{t\geq t_0}\{\|Ax(t)-b\|\|\lambda(t)-\lambda^*\|\} + \|Ax(t_0)-b\|\|\lambda(t_0-\lambda^*)\| \\
&\qquad + \frac{1}{2}\int_{t_0}^{r} \left(\|Ax(t)-b\|^2 + \|\dot\lambda(t)\|^2\right) dt.
\end{aligned} \tag{5.63}
$$

The supremum term is finite due to the boundedness of the trajectory. Now, by using the nonnegativity of φ and the facts (5.60), (5.61), (5.62) and (5.63) on (5.59), we come to

$$
\begin{aligned}
&\frac{\theta}{\alpha-1}\int_{t_0}^{r} t\sigma(t)\|A^*(\lambda(t)-\lambda^*)\|^2 dt \\
&= \int_{t_0}^{r} \left[\theta\delta(t) - \frac{\big(\theta(\alpha+1)-1\big)\delta(t)+\theta t\dot\delta(t)}{\alpha-1}\right] t\|A^*\left(\lambda(t)-\lambda^*\right)\|^2 dt \\
&\leq \frac{1}{\alpha-1}\int_{t_0}^{r} tV(t)dt + \int_{t_0}^{r} t\left(\|Ax(t)-b\|^2 + \left\|\dot\lambda(t)\right\|^2\right) dt + C_6,
\end{aligned} \tag{5.64}
$$

where

$$
C_6 := \varphi(t_0) + 2\sup_{t\geq t_0}\left\{\|Ax(t)-b\|\|\lambda(t)-\lambda^*\|\right\} + 2\|Ax(t_0)-b\|\|\lambda(t_0-\lambda^*)\| + \frac{C_5}{(\alpha-1)t_0^{\alpha-1}}.
$$

According to (5.16) and (5.17) in Theorem 5.2.3, as well as Lemma 5.3.3, we know that the mappings $t\mapsto tV(t)$ and $t\mapsto t\left(\|Ax(t)-b\|^2 + \left\|\dot\lambda(t)\right\|\right)$ belong to $\mathbb{L}^1[t_0,+\infty)$. Therefore, by taking the limit as $r\to+\infty$ in (5.64) we obtain

$$
\int_{t_0}^{+\infty} t\sigma(t)\|A^*(\lambda(t)-\lambda^*)\|^2 dt < +\infty.
$$

Again, from (5.35) we conclude that

$$
\int_{t_0}^{+\infty} t\delta(t)\|A^*(\lambda(t)-\lambda^*)\|^2 dt \leq \frac{1}{C_2}\int_{t_0}^{+\infty} t\sigma(t)\|A^*(\lambda(t)-\lambda^*)\|^2 dt < +\infty,
$$

which completes the proof. □

Theorem 5.3.7. *Under Assumption 5.2, let $(x,\lambda):[t_0,+\infty)\to\mathcal{X}\times\mathcal{Y}$ be a solution to (5.3) and $(x^*,\lambda^*)\in\mathbb{S}$. Then, as $t\to+\infty$ it holds*

$$
\|\nabla f(x(t))-\nabla f(x^*)\| = o\left(\frac{1}{\sqrt{t}\sqrt[4]{\delta(t)}}\right) \quad \text{and} \quad \|A^*(\lambda(t)-\lambda^*)\| = o\left(\frac{1}{\sqrt{t}\sqrt[4]{\delta(t)}}\right). \tag{5.65}
$$

Consequently, as $t\to+\infty$

$$
\|\nabla_x\mathcal{L}(x(t),\lambda(t))\| = \|\nabla f(x(t)) + A^*\lambda(t)\| = o\left(\frac{1}{\sqrt{t}\sqrt[4]{\delta(t)}}\right),
$$

while, as seen earlier,

$$
\|\nabla_\lambda\mathcal{L}(x(t),\lambda(t))\| = \|Ax(t)-b\| = \mathcal{O}\left(\frac{1}{t^2\delta(t)}\right).
$$

Proof. We first show the gradient rate. For $t\geq t_0$, it holds

$$
\frac{d}{dt}\left(t\sqrt{\delta(t)}\|\nabla f(x(t))-\nabla f(x^*)\|^2\right) = \left(\sqrt{\delta(t)} + \frac{t\dot\delta(t)}{2\sqrt{\delta(t)}}\right)\|\nabla f(x(t))-\nabla f(x^*)\|^2
$$

$$+ 2t\sqrt{\delta(t)} \left\langle \nabla f(x(t)) - \nabla f(x^*), \frac{d}{dt}\nabla f(x(t)) \right\rangle. \tag{5.66}$$

On the one hand, by Assumption 5.2, we can write

$$\left(\sqrt{\delta(t)} + \frac{t\dot{\delta}(t)}{2\sqrt{\delta(t)}} \right) = \left(\sqrt{\delta(t)} + \frac{\sqrt{\delta(t)}}{2} \cdot \frac{t\dot{\delta}(t)}{\delta(t)} \right) \le \left(1 + \frac{1-2\theta}{2\theta} \right) \sqrt{\delta(t)} = \frac{1}{2\theta}\sqrt{\delta(t)}. \tag{5.67}$$

Since δ is nondecreasing, for $t \ge t_0$ we have $\sqrt{\delta(t)} \ge \sqrt{\delta(t_0)} > 0$. Set $t_1 := \max\left\{ t_0, \frac{1}{\sqrt{t_0}} \right\}$. Therefore, for $t \ge t_1$ it holds

$$\frac{1}{\sqrt{\delta(t)}} \le \frac{1}{\sqrt{\delta(t_0)}} = t_1 \le t$$

and thus

$$\sqrt{\delta(t)} \le t\delta(t). \tag{5.68}$$

On the other hand, for every $t \ge t_1$ we deduce

$$\begin{aligned} 2t\sqrt{\delta(t)} \left\langle \nabla f(x(t)) - \nabla f(x^*), \frac{d}{dt}\nabla f(x(t)) \right\rangle &= 2t \left\langle \sqrt{\delta(t)}\left[\nabla f(x(t)) - \nabla f(x^*) \right], \frac{d}{dt}\nabla f(x(t)) \right\rangle \\ &\le t\delta(t)\|\nabla f(x(t)) - \nabla f(x^*)\|^2 + t\left\| \frac{d}{dt}\nabla f(x(t)) \right\|^2 \\ &\le t\delta(t)\|\nabla f(x(t)) - \nabla f(x^*)\|^2 + L^2 t\left\|\dot{x}(t)\right\|^2, \end{aligned} \tag{5.69}$$

where the last inequality is a consequence of the L-Lipschitz continuity of ∇f. By combining (5.67), (5.68) and (5.69), from (5.66) we assert that for every $t \ge t_1$

$$\frac{d}{dt}\left(t\sqrt{\delta(t)}\|\nabla f(x(t)) - \nabla f(x^*)\|^2 \right) \le \left(1 + \frac{1}{2\theta} \right) t\delta(t)\|\nabla f(x(t)) - \nabla f(x^*)\|^2 + L^2 t\left\|\dot{x}(t)\right\|^2.$$

The right hand side of the previous inequality belongs to $\mathbb{L}^1([t_1, +\infty))$, according to (5.17) and (5.36). Since δ is nondecreasing, for every $t \ge t_1$ we have

$$\sqrt{\delta(t)} = \sqrt{\delta(t)} \cdot \frac{\sqrt{\delta(t)}}{\sqrt{\delta(t)}} \le \frac{\delta(t)}{\sqrt{\delta(t_0)}},$$

and thus

$$\int_{t_1}^{+\infty} t\sqrt{\delta(t)}\|\nabla f(x(t)) - \nabla f(x^*)\|^2 dt \le \frac{1}{\sqrt{\delta(t_0)}} \int_{t_1}^{+\infty} t\delta(t)\|\nabla f(x(t)) - \nabla f(x^*)\|^2 dt < +\infty. \tag{5.70}$$

This means that the function being differentiated also belongs to $\mathbb{L}^1([t_1, +\infty))$. Therefore, Lemma 2.3.2 gives us, as $t \to +\infty$,

$$t\sqrt{\delta(t)}\|\nabla f(x(t)) - \nabla f(x^*)\|^2 \to 0.$$

Proceeding in the exact same way, for every $t \ge t_1$ we have

$$\begin{aligned} &\frac{d}{dt}\left(t\sqrt{\delta(t)}\left\| A^*(\lambda(t) - \lambda^*) \right\|^2 \right) \\ = &\left(\sqrt{\delta(t)} + \frac{t\dot{\delta}(t)}{2\sqrt{\delta(t)}} \right) \left\| A^*(\lambda(t) - \lambda^*) \right\|^2 + 2t\sqrt{\delta(t)}\langle AA^*(\lambda(t) - \lambda^*), \dot{\lambda}(t) \rangle \\ \le &\left(\frac{1}{2\theta} + \|A\|^2 \right) t\delta(t)\left\| A^*(\lambda(t) - \lambda^*) \right\|^2 + t\|\dot{\lambda}(t)\|^2. \end{aligned}$$

According to (5.17) and (5.58), the right hand side of the previous inequality belongs to $\mathbb{L}^1([t_1, +\infty))$. Arguing as in (5.70), we deduce that the function being differentiated also belongs to $\mathbb{L}^1[t_1, +\infty)$. Again applying Lemma 2.3.2, we come to

$$t\sqrt{\delta(t)}\left\| A^*(\lambda(t) - \lambda^*) \right\|^2 \to 0$$

as $t \to +\infty$. Finally, recalling that $A^*\lambda^* = -\nabla f(x^*)$, we deduce from the triangle inequality that

$$\begin{aligned}\|\nabla_x \mathcal{L}\big(x(t), \lambda(t)\big)\| &= \|\nabla f(x(t)) + A^*\lambda(t)\| \\ &\leq \|\nabla f(x(t)) - \nabla f(x^*)\| + \big\|A^*(\lambda(t) - \lambda^*)\big\| \\ &= o\left(\frac{1}{\sqrt{t}\sqrt[4]{\delta(t)}}\right)\end{aligned}$$

as $t \to +\infty$, and the third claim follows. □

Remark 5.3.8. The previous theorem also has its own interest. It tells us that the time rescaling parameter also plays a role in accelerating the rates of convergence for $\|\nabla f(x(t)) - \nabla f(x^*)\|$ and $\|A^*(\lambda(t) - \lambda^*)\|$ as $t \to +\infty$. Moreover, we deduce from (5.65) that the mapping $(x, \lambda) \mapsto (\nabla f(x), A^*\lambda)$ is constant along $\mathbb{S}$, as reported in [43, Proposition A.4].

We now come to the final step and show weak convergence of the trajectories of (5.3) to elements of $\mathbb{S}$.

Theorem 5.3.9. *Under Assumption 5.2, let $(x, \lambda) : [t_0, +\infty) \to \mathcal{X} \times \mathcal{Y}$ be a solution to (5.3) and $(x^*, \lambda^*) \in \mathbb{S}$. Then, $(x(t), \lambda(t))$ converges weakly to a primal-dual solution to (5.1) as $t \to +\infty$.*

Proof. For proving this theorem, we make use of Opial's Lemma (see Lemma 2.1.2). Lemma 5.3.3 tells us that $\lim_{t\to+\infty} \|(x(t), \lambda(t)) - (x^*, \lambda^*)\|$ exists for every $(x^*, \lambda^*) \in \mathbb{S}$, which proves condition (i) of Opial's Lemma. Now let $\big(\overline{x}, \overline{\lambda}\big)$ be any weak sequential cluster point of $(x(t), \lambda(t))$ as $t \to +\infty$, which means there exists a strictly increasing sequence $(t_n)_{n\in\mathbb{N}} \subseteq [t_0, +\infty)$ such that

$$(x(t_n), \lambda(t_n)) \rightharpoonup \big(\overline{x}, \overline{\lambda}\big)$$

as $n \to +\infty$. We want to show the remaining condition of Opial's Lemma, which asks us to check that $\big(\overline{x}, \overline{\lambda}\big) \in \mathbb{S}$. In other words, we must show that, for every $(y, \nu) \in \mathcal{X} \times \mathcal{Y}$ it holds

$$\mathcal{L}\big(\overline{x}, \nu\big) \leq \mathcal{L}\big(\overline{x}, \overline{\lambda}\big) \leq \mathcal{L}\big(y, \overline{\lambda}\big). \tag{5.71}$$

Let $(y, \nu) \in \mathcal{X} \times \mathcal{Y}$ and $(x^*, \lambda^*) \in \mathbb{S}$ be fixed. Notice that the functions

$$f(\cdot) + \big\langle \overline{\lambda}, A(\cdot) - Ay\big\rangle \quad \text{and} \quad \big\langle \cdot, b - Ay\big\rangle$$

are convex and continuous, therefore they are lower semicontinuous. According to a known result (see, for example, [28, Theorem 9.1]), they are also weakly lower semicontinuous. Therefore, we can derive that

$$\begin{aligned}\mathcal{L}\big(\overline{x}, \overline{\lambda}\big) - \mathcal{L}\big(y, \overline{\lambda}\big) &= f(\overline{x}) + \big\langle \overline{\lambda}, A\overline{x} - Ay\big\rangle - f(y) \leq \liminf_{n\to+\infty}\Big[f(x(t_n)) + \big\langle \overline{\lambda}, A(x(t_n)) - Ay\big\rangle\Big] - f(y) \\ &= f(x^*) + \big\langle \overline{\lambda}, b - Ay\big\rangle - f(y) \leq f(x^*) - f(y) + \liminf_{n\to+\infty}\big\langle A^*\lambda(t_n), x^* - y\big\rangle \\ &= f(x^*) - \big(f(y) + \langle \lambda^*, Ay - b\rangle\big) = \mathcal{L}(x^*, \lambda^*) - \mathcal{L}(y, \lambda^*) \leq 0,\end{aligned}$$

where in the second and third equalities we used the fact that, as $n \to +\infty$, we have $f(x(t_n)) \to f(x^*)$ and $Ax(t_n) \to b$ (Theorem 5.2.4), and $A^*\lambda(t_n) \to A^*\lambda^*$ (Theorem 5.3.7) as $n \to +\infty$. Similarly, the weak lower semicontinuity of the function $\big\langle \nu - \overline{\lambda}, A(\cdot) - b\big\rangle$ yields

$$\mathcal{L}\big(\overline{x}, \nu\big) - \mathcal{L}\big(\overline{x}, \overline{\lambda}\big) = \big\langle \nu - \overline{\lambda}, A\overline{x} - b\big\rangle \leq \liminf_{n\to+\infty}\big\langle \nu - \overline{\lambda}, Ax(t_n) - b\big\rangle = 0,$$

We have thus showed (5.71) and the proof is concluded. □

5.4 Numerical experiments

We will illustate the theoretical results by two numerical examples, with $\mathcal{X} = \mathbb{R}^4$ and $\mathcal{Y} = \mathbb{R}^2$. We will address two minimization problems with linear constraints; one with a strongly convex objective function and another with a convex objective function which is not strongly convex. In both cases, the linear constraints are dictated by

$$A = \begin{bmatrix} 1 & -1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Example 5.4.1. Consider the minimization problem

$$\begin{array}{rl} \min & f(x_1, x_2, x_3, x_4) := (x_1 - 1)^2 + (x_2 - 1)^2 + x_3^2 + x_4^2 \\ \text{subject to} & x_1 - x_2 - x_3 = 0 \\ & x_2 - x_4 = 0. \end{array}$$

The optimality conditions can be calculated and lead to the following primal-dual solution pair

$$x^* = \begin{bmatrix} 0.8 \\ 0.6 \\ 0.2 \\ 0.6 \end{bmatrix} \quad \text{and} \quad \lambda^* = \begin{bmatrix} 0.4 \\ 1.2 \end{bmatrix}.$$

Example 5.4.2. Consider the minimization problem

$$\begin{array}{rl} \min & f(x_1, x_2, x_3, x_4) := \log\left(1 + e^{-x_1 - x_2}\right) + x_3^2 + x_4^2 \\ \text{subject to} & x_1 - x_2 - x_3 = 0 \\ & x_2 - x_4 = 0. \end{array}$$

This problem is similar to the regularized logistic regression frequently used in machine learning. We cannot explicitly calculate the optimality conditions as in the previous case; instead, we use the last solution in the numerical experiment as the approximate solution.

To comply with Assumption 5.2, we choose $t_0 > 0$, $\alpha = 8$, $\beta = 10$, $\theta = \frac{1}{6}$, and we test four different choices for the rescaling parameter: $\delta(t) = 1$ (i.e., the (PD-AVD) dynamics in [102, 43]), $\delta(t) = t$, $\delta(t) = t^2$ and $\delta(t) = t^3$. In both examples, the initial conditions are

$$x(t_0) = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}, \quad \lambda(t_0) = \begin{bmatrix} 0.2 \\ 0.2 \end{bmatrix}, \quad \dot{x}(t_0) = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{bmatrix} \quad \text{and} \quad \dot{\lambda}(t_0) = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

For each choice of δ we plot, using a logarithmic scale, the primal-dual gap $\mathcal{L}(x(t), \lambda^*) - \mathcal{L}(x^*, \lambda(t))$, the feasibility measure $\|Ax(t) - b\|$ and the functional values $|f(x(t)) - f(x^*)|$, to highlight the theoretical result in Theorem 5.2.4. We also illustrate the findings from Theorem 5.3.7, namely, we plot the quantities $\|\nabla f(x(t)) - \nabla f(x^*)\|$ and $\|A^*(\lambda(t) - \lambda^*)\|$, as well as the velocity $\left\|(\dot{x}(t), \dot{\lambda}(t))\right\|$.

Next we use Example 5.4.2 to compare the convergence behavior of our system (5.3) with the one where the asymptotically vanishing damping term is chosen to be $\frac{\alpha}{t^r}$, for $r \in [0, 1]$. Notice that $r = 1$ gives our system (5.3). When $r = 0$, in the setting of [66, Theorem 2.2], we know that the primal-dual gap exhibits a convergence rate of $\mathcal{O}\left(\frac{1}{t\delta(t)}\right)$ as $t \to +\infty$. This is illustrated in Figure 5.3, were we plotted the combinations $(\delta(t) = t; r = 0)$, $(\delta(t) = t; r = 1)$, $\left(\delta(t) = t^2; r = 0\right)$, and $\left(\delta(t) = t^2; r = 1\right)$. In particular, observe that the rate predicted by [66, Theorem 2.2] for the primal-dual gap for the case $\left(\delta(t) = t^2; r = 0\right)$ reads $\mathcal{O}\left(\frac{1}{t \cdot t^2}\right)$, while the rate predicted by our Theorem 5.2.4 for the case $(\delta(t) = t; r = 1)$ reads $\mathcal{O}\left(\frac{1}{t^2 \cdot t}\right)$. It is no surprise then to see the combinations $\left(\delta(t) = t^2; r = 0\right)$ and $(\delta(t) = t; r = 1)$ exhibiting similar convergence behavior in Figure 5.3.

For better understanding, we run Example 5.4.2 once more to show the plots which result from fixing the time rescaling parameter $\delta(t) = t$ and varying $r \in \{0, 0.25, 0.5, 0.75, 1\}$. Notice how the convergence

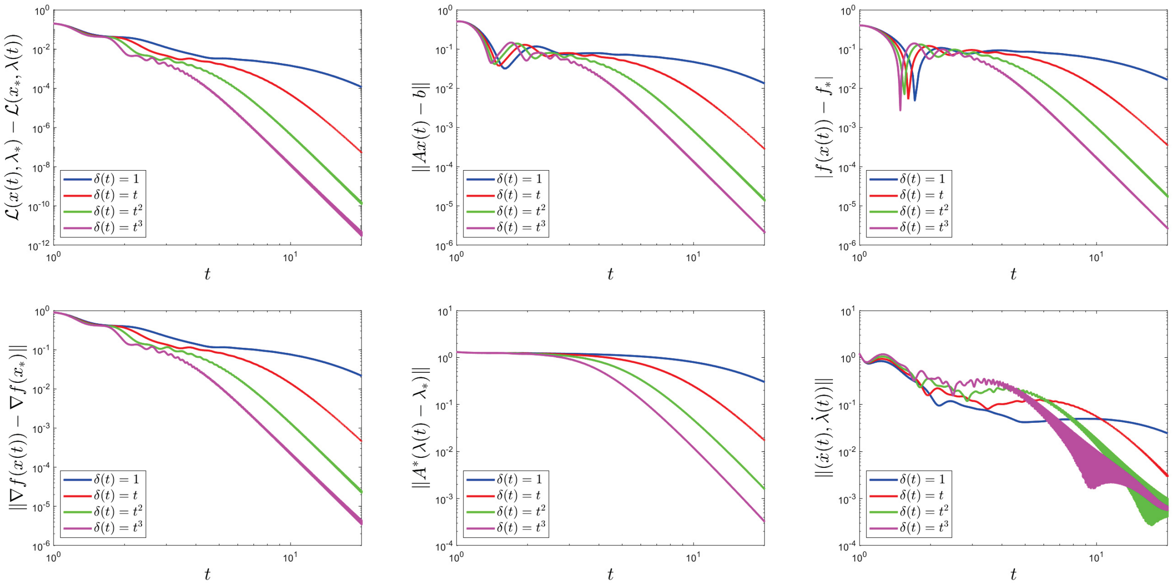


Figure 5.1: The function $\delta(\cdot)$ influences the convergence behavior in Example 5.4.1

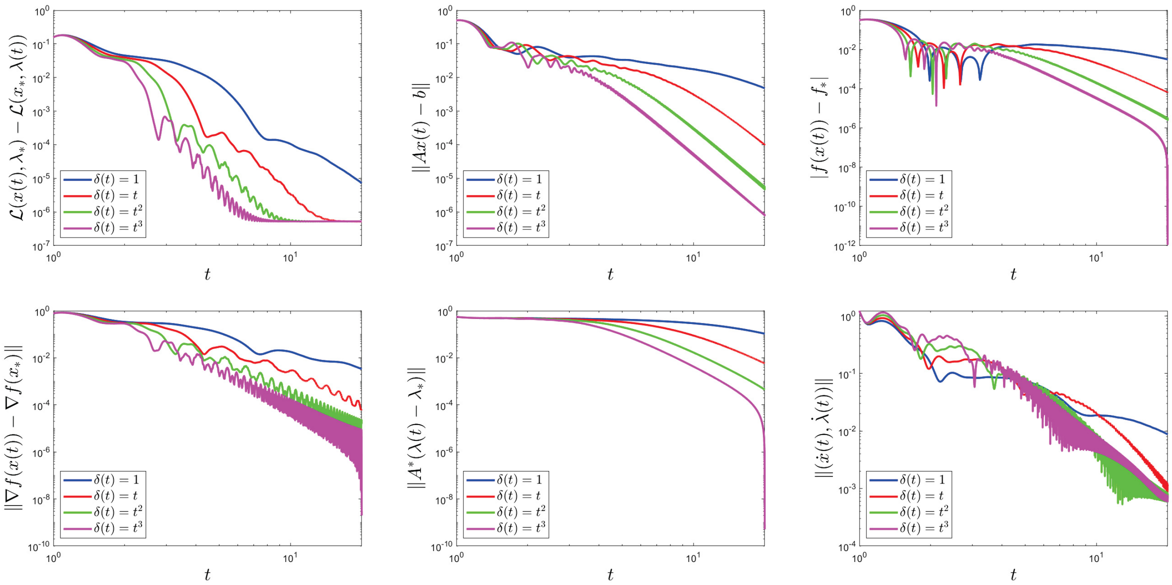


Figure 5.2: The function $\delta(t)$ influences convergence behavior in Example 5.4.2

improves as r approaches 1. As $t \to +\infty$, [66, Theorem 2.7] predicts convergence rates of $\mathcal{O}\left(\frac{1}{t^\tau \delta(t)}\right)$ for the primal-dual gap and of $\mathcal{O}\left(\frac{1}{t^{\tau/2}}\right)$ for the velocities, which is reflected in our plots.

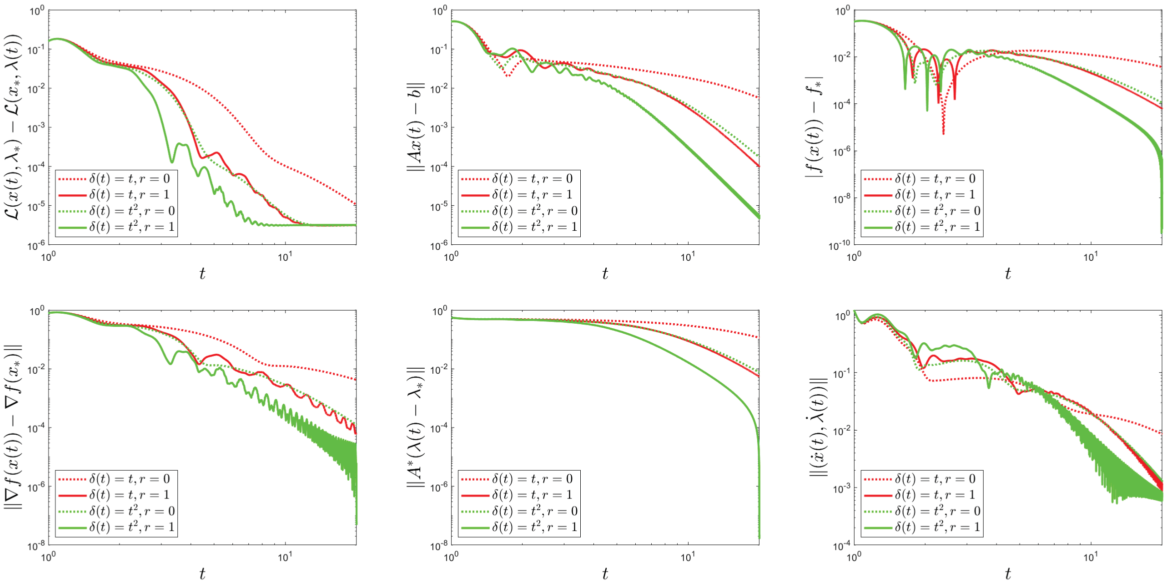


Figure 5.3: The function $\delta(t)$, as well as the parameter r, influence convergence behavior in Example 5.4.2

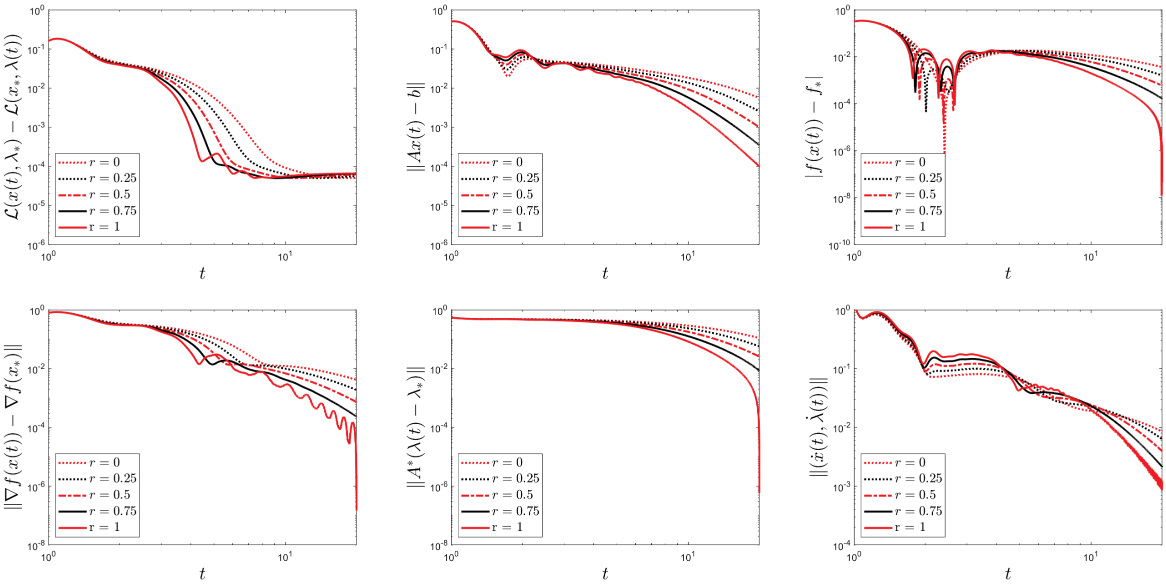


Figure 5.4: The parameter r influences convergence behavior in Example 5.4.2

Chapter 6

Algorithms with and without acceleration for hierarchical optimization

6.1 Introduction

In this chapter, which consists of our work [34], we explore the hierarchical (also called simple bilevel) optimization problem

$$\begin{aligned} &\min \ H(x) := h(x) + \hat{h}(x), \\ &\text{subject to } x \in \operatorname{argmin}_{z\in\mathcal{H}} \ F(z) := f(z) + \hat{f}(z), \end{aligned} \tag{6.1}$$

where

$$\begin{cases} \mathcal{H} \text{ is a real Hilbert space;} \\ f, h : \mathcal{H} \to \mathbb{R} \text{ are convex, differentiable, } \nabla f \text{ and } \nabla g \text{ are } L_f\text{- and } L_h\text{-Lipschitz continuous, respectively;} \\ \hat{f}, \hat{h} : \mathcal{H} \to \overline{\mathbb{R}} \text{ are proper, convex and lower semicontinuous;} \\ H \text{ is assumed to be lower bounded over } \mathcal{H}. \end{cases} \tag{6.2}$$

We assume (6.1) to have at least one solution. To address (6.1), we instead turn our attention to the formulation

$$\min_{x\in\mathcal{H}} \Psi_\varepsilon(x) := F(x) + \varepsilon H(x), \tag{6.3}$$

where $\varepsilon > 0$. Under mild conditions (cf. [56]), there exists a small enough $\overline{\varepsilon} > 0$ such that for $0 < \varepsilon \leq \overline{\varepsilon}$, the solution sets of (6.1) and (6.3) coincide. However, this fact serves only as a motivation: we see Ψ_ε as our objective function, but at each step of the (two) algorithms we will analyze, we will also update ε. In other words, we will be minimizing a dynamically updated objective function Ψ_{ε_k}, where $(\varepsilon_k)_{k\geq 0}$ will be a decreasing sequence of positive numbers which approach zero as $k \to +\infty$, but not too fast. For simplicity, we will restrict ourselves to regularization parameters of the form

$$\varepsilon_k := \frac{c}{(k+\beta)^\delta}, \tag{6.4}$$

where $\beta, c, \delta > 0$ and $k \geq 0$. Attached to (6.1), we will explore the asymptotic properties of two numerical algorithms:

Algorithm 6.1: Bilevel Proximal-Gradient method

Input: $x_0 \in \mathcal{H}$, $0 < \theta < \frac{2}{L_f}$ ▷ ε_k as in (6.4)

for $k = 1, 2, \ldots$ **do**

$\quad x_{k+1} := \operatorname{prox}_{\theta(\hat{f}+\varepsilon_k\hat{h})}\left(x_k - \theta(\nabla f(x_k) + \varepsilon_k \nabla h(x_k))\right)$

which at each step, it applies a proximal gradient step to the objective function Ψ_{ε_k}; and the second algorithm

Algorithm 6.2: Bilevel Fast Proximal-Gradient method

Input: $x_0, x_1 \in \mathcal{H}$, $\alpha > 3$, $\gamma \geq 0$, $0 < s < \frac{1}{L_f}$ $\qquad \triangleright\ \varepsilon_k$ as in (6.4)

for $k = 1, 2, \ldots$ **do**

$\quad \alpha_k := 1 - \dfrac{\alpha}{k + \gamma + 1}$

$\quad y_k := x_k + \alpha_k(x_k - x_{k-1})$

$\quad x_{k+1} := \operatorname{prox}_{s(\hat{f} + \varepsilon_k \hat{h})}\left(y_k - s(\nabla f(y_k) + \varepsilon_k \nabla h(y_k))\right)$

which can be seen as a version with Nesterov momentum of Algorithm 6.1. As we will discuss later, Algorithms 6.1 and 6.2 are tightly linked to the continuous-time systems

$$\dot{x}(t) + \nabla f(x(t)) + \varepsilon(t)\nabla h(x(t)) = 0 \quad \text{for } t \geq t_0, \tag{6.5}$$

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) + \varepsilon(t)\nabla h(x(t)) = 0 \quad \text{for } t \geq t_0. \tag{6.6}$$

Our main contribution is that, under a so-called *Hölderian error bound* assumption for the inner function F, we achieve fast, simultaneous and last-iterate convergence rates for $F(x_k) - \min F$ and $|H(x_k) - \min_{\operatorname{argmin} F} H|$. Namely, we will have, as $k \to +\infty$,

$$F(x_k) - \min F = o\left(\frac{1}{k}\right), \quad |H(x_k) - \min_{\operatorname{argmin} F} H| = o\left(\frac{1}{k^{1-\delta}}\right) \quad (\text{Algorithm 6.1}, \ \delta < 1),$$

$$F(x_k) - \min F = o\left(\frac{1}{k^2}\right), \quad |H(x_k) - \min_{\operatorname{argmin} F} H| = o\left(\frac{1}{k^{2-\delta}}\right) \quad (\text{Algorithm 6.2}, \ \delta < 2).$$

Additionally, we will show weak convergence of the sequences generated by Algorithms 6.1 and 6.2 towards solutions to (6.1). The little-o rates and the weak convergence in the framework of general real Hilbert spaces marks a first in the literature for this class of problems and under a Hölderian error bound assumption, to the best of our knowledge.

This chapter is organized as follows: in Section 6.2, we will lay out the main assumptions and the tools for our analysis of the two algorithms, which can be treated in a unified way to some extent; in Section 6.3 and Section 6.4 we study the asymptotic properties (i.e., convergence rates, weak convergence of iterates) of Algorithms 6.1 and 6.2 respectively; in Section 6.5, we discuss some facts regarding which statements can be said in the absence of a Hölderian error bound; in Section 6.6, we illustrate our theoretical findings with some numerical experiments; in Section 6.7, we mention the continuous-time analogs (6.5) and (6.6) of Algorithms 6.1 and 6.2 respectively, whose convergence analysis actually served as a guide for the convergence analysis of the numerical algorithms.

It is worth pointing out that, since many of our results are formulated in a general setting, we will be expressing many of our results in terms of a parameter η, where

$\eta = 1$ will refer to Algorithm 6.1, and $\eta = 2$ will refer to Algorithm 6.2.

Accompanied with Section 6.5, our comprehensive analysis complements and improves the recent works [77, 71, 78]. First, all results are formulated in a infinite dimensional Hilbert space setting, in line with [84, 20], where the case $\eta = 1$ and $\hat{h} = \hat{f} = 0$ is analyzed with a general regularization sequence/parameter. However, only weak convergence to an optimal solution of (6.1) is provided (in both continuous and discrete time), along with a convergence rate for the inner function (in continuous time only).

6.2 Principal assumptions and tools for the analysis

In our geometric setting, we assume a Hölderian error bound for the inner function, and a weaker summability condition due to Attouch and Czarnecki in [20]. We will now present both assumptions.

Assumption 6.1 (Hölderian growth condition)**.** A proper, convex and lower semicontinuous function $F : \mathcal{H} \to \overline{\mathbb{R}}$ with $\operatorname{argmin} F \neq \emptyset$ is said to satisfy a *Hölderian growth condition with exponent* $\rho \in (1, 2]$ if for some $\tau > 0$ we have, for every $x \in \mathcal{H}$,

$$\tau\rho^{-1} \operatorname{dist}(x, \operatorname{argmin} F)^{\rho} \leq F(x) - \min F. \tag{6.7}$$

Assumption 6.2 (Attouch and Czarnecki, [20])**.** A proper, convex and lower semicontinuous function $F : \mathcal{H} \to \overline{\mathbb{R}}$ with $\operatorname{argmin} F \neq \emptyset$ is said to satisfy the *Attouch-Czarnecki condition* with respect to $\eta \in \{1, 2\}$ if for any solution x^* to the bilevel problem (6.1), for all $p \in N_{\operatorname{argmin} F}(x^*)$, and $(\varepsilon_k)_{k\geq 0}$ defined as in (6.4) we have that

$$\sum_{k=0}^{+\infty} k^{\eta-1}\Big[(F - \min F)^* (\varepsilon_k p) - \sigma_{\operatorname{argmin} F} (\varepsilon_k p)\Big] < +\infty. \tag{6.8}$$

The Attouch-Czarnecki condition is known to be weaker than the Hölderian growth condition, since, whenever F satisfies Assumption 6.1, for every $z \in \mathcal{H}$ it holds

$$0 \leq (F - \min F)^*(z) - \sigma_{\operatorname{argmin} F}(z) \leq \tau^{1-\rho^\star} \frac{1}{\rho^\star}\|z\|^{\rho^\star}, \tag{6.9}$$

where the first inequality is a consequence of the Young-Fenchel inequality, and the second can be derived as in [84, Section 3.2]. Here, the dual conjugate exponent $\rho^\star$ is given by

$$\frac{1}{\rho} + \frac{1}{\rho^\star} = 1. \tag{6.10}$$

Therefore, as $\varepsilon_k \simeq k^{-\delta}$, Assumption 6.1 implies Assumption 6.2 when $\frac{\eta}{\rho^\star} < \delta$. A further mild technical assumption we will require is a qualification condition.

Assumption 6.3 (Qualification condition)**.** The functions H and F in (6.1) are such that

$$\partial(H + \delta_{\operatorname{argmin} F})(x^*) = \partial H(x^*) + N_{\operatorname{argmin} F}(x^*) \quad \text{for any solution } x^* \text{ to (6.1).} \tag{6.11}$$

The qualification condition is known to be fulfilled if either there exists $x' \in \operatorname{dom} H \cap \operatorname{argmin} F$ such that H is continuous at x' or if $0 \in \operatorname{int}(\operatorname{dom} H - \operatorname{argmin} F)$. This assumption proves to be crucial for controlling the (sign-less) residual of (6.3) with $\varepsilon > 0$ in terms of its value at an optimal solution to (6.1). Indeed, picking $p^* \in N_{\operatorname{argmin} F}(x^*)$ such that $-p^* \in \partial H(x^*)$ given by Assumption 6.3, we get that for every $x \in \mathcal{H}$

$$\begin{aligned} -\Big[F(x) - F(x^*) + \varepsilon(H(x) - H(x^*))\Big] &\leq -(F(x) - \min F) + \varepsilon\langle p^*, x - x^*\rangle \\ &= \Big[\langle \varepsilon p^*, x\rangle - (F - \min F)(x)\Big] - \langle \varepsilon p^*, x^*\rangle \\ &\leq (F - \min F)^*(\varepsilon p^*) - \sigma_{\operatorname{argmin} F}(\varepsilon p^*), \end{aligned} \tag{6.12}$$

where we used $\langle \varepsilon p^*, x^*\rangle = \sigma_{\operatorname{argmin} F}(\varepsilon p^*)$. Although the Attouch-Czarnecki condition in Assumption 6.2 might appear abstract, (6.12) shows that it comes naturally in the analysis of Tikhonov-type methods.

6.2.1 The link between first and second-order energy estimates

The analysis of the continuous-time systems (6.5) and (6.6) presented in Appendix 6.7 serves as a foundational guide for conducting the discrete-time Lyapunov analysis of the corresponding numerical schemes—Algorithm 6.1 and Algorithm 6.2. Inspired by [17], we introduce a discrete Lyapunov energy sequence $(E_k)_{k\geq 1}$, defined in (6.35) and (6.53), of the form

$$E_k := t_k^{\eta} \left(\Psi_{k-1}(x_k) - \Psi_{k-1}(x^*)\right) + V_k \tag{6.13}$$

for $k \geq 1$, where $\Psi_k(x) := F(x) + \varepsilon_k H(x)$ is the regularized function (6.3) with regularization parameter ε_k defined as in (6.4), $V_k \geq 0$ is specified in each case, x^* is a solution to (6.1), $t_k := \theta(k + \gamma)$ for

$\theta > 0$ and $\gamma \geq 0$, and $\eta = 1$ for Algorithm 6.1, and $\eta = 2$ for Algorithm 6.2. As in the continuous-time Lyapunov analysis, we arrive at the following dissipativity estimate

$$E_{k+1} - E_k + \zeta_{k,\delta}(H(x_k) - H(x^*)) + g_k \leq -C_2 k^{\eta-1}(F(x_k) - F(x^*)) \tag{6.14}$$

for k sufficiently large, where $g_k \geq 0$ and $\zeta_{k,\delta} \simeq k^{-\delta+\eta-1}$. The precise form of g_k and $\zeta_{k,\delta}$ is not important at this stage. As shown in Lemma 6.2.3, this Lyapunov inequality underpins much of our analysis, enabling the derivation of key convergence results, including little-o rates and weak convergence of the iterates.

6.2.2 The principal tools for the analysis

Here we include the auxiliary lemmas which are used to show the convergence rates for Algorithms 6.1 and 6.2. We feel these results were to specific to this particular paper to include them in the Preliminaries chapter.

The following straightforward result encodes an important inequality that is contained, e.g., in [77], and is used repeatedly in our paper.

Lemma 6.2.1. *Suppose Assumptions 6.1 and 6.3 hold. Then, for any solution $x^* \in \mathcal{H}$ to (6.1) and $p^* \in N_{\operatorname{argmin} F}(x^*)$ such that $-p^* \in \partial H(x^*)$, for every $x \in \mathcal{H}$ it holds*

$$H(x^*) - H(x) \leq \|p^*\| \left(\frac{F(x) - F(x^*)}{\tau\rho^{-1}} \right)^{\frac{1}{\rho}}. \tag{6.15}$$

Proof. Using the definitions of the convex subdifferential and of the normal cone, for all $x \in \mathcal{H}$ it holds

$$\begin{aligned} H(x^*) - H(x) &\leq -\langle p^*, x^* - x\rangle \leq -\big\langle p^*, P_{\operatorname{argmin} F}(x) - x\big\rangle \\ &\leq \|p^*\| \big\|P_{\operatorname{argmin} F}(x) - x\big\| = \|p^*\| \operatorname{dist}(x, \operatorname{argmin} F) \leq \|p^*\| \left(\frac{F(x) - F(x^*)}{\tau\rho^{-1}} \right)^{\frac{1}{\rho}}, \end{aligned} \tag{6.16}$$

which yields the claim. ☐

Lemma 6.2.2. *Assume that $t_k := \theta(k+\gamma)$, with $\theta > 0$ and $\gamma \geq 0$, and that $(F_k)_{k\geq 0}$ is a sequence of nonnegative numbers such that as $k \to +\infty$,*

$$F_k \leq C_E t_k^{-\eta} + C_{0,F} t_k^{-\delta} F_k^{\frac{1}{\rho}} \quad \text{and} \quad F_k = \mathcal{O}(k^{-\delta}), \tag{6.17}$$

where $\rho \in (1, 2]$, $\eta \in \{1, 2\}$, $\delta \in \left(\frac{\eta}{\rho^\star}, \eta\right)$ (where $\rho^\star$ is set as in (6.10)), and C_E and $C_{0,F}$ are positive constants. Then, as $k \to +\infty$ it holds

$$F_k = \mathcal{O}(k^{-\eta}) \quad \text{as } k \to +\infty. \tag{6.18}$$

Proof. Since $F_k = O(k^{-\delta})$ as $k \to +\infty$, there is a positive constant $C_{1,F}$ such that for all k large enough it holds

$$F_k \leq C_E t_k^{-\eta} + C_{0,F} t_k^{-\delta-\frac{1}{\rho}\delta} (t_k^\delta F_k)^{\frac{1}{\rho}} \leq C_E t_k^{-\eta} + C_{1,F} t_k^{-\delta-\frac{1}{\rho}\delta}. \tag{6.19}$$

If $\delta_1 := \delta + \frac{\delta}{\rho} \geq \eta$, we stop. Otherwise, the rate has improved to $F_k = \mathcal{O}(k^{-\delta_1})$ as $k \to +\infty$. We may repeat this first step again, but this time plugging $t_k^{\delta_1}$ instead of t_k^δ, possibly improving the rate to $F_k = \mathcal{O}(k^{-\delta_2})$, where $\delta_2 = \delta_1 + \frac{\delta_1}{\rho} = \delta(1 + \rho^{-1} + \rho^{-2})$. We may repeat this argument n times, as long as $\delta_{n+1} := \delta_n + \frac{\delta_n}{\rho} < \eta$, in order to obtain a positive constant $C_{n,F}$ such that for k large enough

$$F_k \leq C_E t_k^{-\eta} + C_{n,F} t_k^{-\delta_n}, \quad \text{where} \quad \delta_n := \delta(1 + \rho^{-1} + \cdots + \rho^{-n}). \tag{6.20}$$

Since $\delta_n \to \delta\rho^\star > \eta$ as $n \to +\infty$, we will stop at some n_0 such that $\delta_{n_0-1} < \eta$ and $\delta_{n_0} \geq \eta$, which yields the claim. ☐

Lemma 6.2.3. *Let $\hat{\delta} > 0$, $\eta \in \{1,2\}$, and set $\delta := \eta\hat{\delta}$. Consider a sequence $(E_k)_{k\geq 1}$ defined as*

$$E_k = t_k^{\eta}\Big(F_k + \varepsilon_{k-1}H_k\Big) + V_k,$$

with $t_k = \theta(k+\gamma)$, $(\varepsilon_k)_{k\geq 0}$ defined as in (6.4), and $F_k \geq 0$, $H_k \geq -H_*$, where $H_* \geq 0$, and $V_k \geq 0$ for every $k \geq 1$. Suppose that for k large enough

$$E_{k+1} - E_k + \zeta_{k,\delta}H_k + g_k \leq -C_2 k^{\eta-1}F_k, \tag{6.21}$$

where $(g_k)_{k\geq 1}$ is a nonnegative sequence, $C_2 \geq 0$, and $(\zeta_{k,\delta})_{k\geq 1}$ satisfies, for $C_{1,\zeta} > 0$, $C_{2,\zeta} > 0$ and k large enough

$$C_{1,\zeta}k^{\eta-1}\varepsilon_{k-1} \leq \zeta_{k,\delta} \leq C_{2,\zeta}k^{\eta-1}\varepsilon_k. \tag{6.22}$$

Consider the following conditions, which will only be assumed if we explicitly say so. For k large enough:

- $(C1)$ $\quad -H_k \leq C_3 F_k^{\frac{1}{\rho}}$*, for some constant $C_3 \geq 0$ and $1 < \rho \leq 2$. Set $\rho^\star$ as in (6.10),*
- $(C2)$ $\quad$ *For each $r > 0$, $-k^{\eta-1}\big(F_k + r\varepsilon_k H_k\big) \leq I_k^r$, where $I_k^r \geq 0$ and $(I_k^r)_{k\geq 1} \in \ell^1(\mathbb{N})$.*

Then, following statements are true:

- *Case $\hat{\delta} > 1$: It holds $F_k = \mathcal{O}(k^{-\eta})$ as $k \to +\infty$;*
- *Case $\hat{\delta} = 1$: If Assumption $(C2)$ holds and $C_2 > 0$ in (6.21), then*
  1. *The limit $\lim_{k\to+\infty} E_k$ exists;*
  2. *$F_k = \mathcal{O}(k^{-\eta})$ as $k \to +\infty$;*
  3. *The sequences $(k^{\eta-1}F_k)_{k\geq 1}$, $\big(k^{\eta-1}\varepsilon_{k-1}H_k\big)_{k\geq 1}$ and $(g_k)_{k\geq 1}$ are summable;*
- *Case $\frac{1}{\rho^\star} < \hat{\delta} < 1$: If Assumptions $(C1)$ and $(C2)$ hold and $C_2 > 0$ in (6.21), then*
  1. *$F_k = \mathcal{O}(k^{-\eta})$ as $k \to +\infty$;*
  2. *It holds $-C_H k^{-\delta+\frac{\eta}{\rho^\star}} \leq t_k^{\eta}\varepsilon_{k-1}H_k$ for some $C_H \geq 0$ and all $k \geq k_0$;*
  3. *The limit $\lim_{k\to+\infty} E_k$ exists;*
  4. *The sequences $(k^{\eta-1}F_k)_{k\geq 1}$, $\big(k^{\eta-1}\varepsilon_{k-1}H_k\big)_{k\geq 1}$ and $(g_k)_{k\geq 1}$ are summable.*

Proof. Let k_0 be a generic positive integer such that every instance of "for k sufficiently large" can be read as "for every $k \geq k_0$". First, in general, using $-H_* \leq H_k$, summing (6.21) from k_0 to $k \geq k_0$ and dropping the nonnegative terms V_k and g_k, we get

$$t_{k+1}^{\eta}F_{k+1} - \left(t_{k+1}^{\eta}\varepsilon_k + \sum_{\ell=k_0}^{k}\zeta_{\ell,\delta}\right)H_* \leq E_{k_0}. \tag{6.23}$$

Note that according to (6.22) and Lemma 2.3.10, we have, for some $C_{3,\zeta}, C_{4,\zeta} > 0$ large enough, for all $k \geq k_0$

$$\sum_{\ell=k_0}^{k}\zeta_{\ell,\delta} \leq C_{2,\zeta}\sum_{\ell=k_0}^{k}\ell^{\eta-1}\varepsilon_\ell \leq C_{3,\zeta}\sum_{\ell=k_0}^{k}\frac{1}{\ell^{\delta-\eta+1}} \leq \begin{cases} C_{4,\zeta} & \text{if } \delta > \eta, \quad \text{i.e., } \hat{\delta} > 1, \\ C_{4,\zeta}\ln(k) & \text{if } \delta = \eta, \quad \text{i.e., } \hat{\delta} = 1, \\ C_{4,\zeta}k^{\eta-\delta} & \text{if } \delta < \eta, \quad \text{i.e., } \hat{\delta} < 1. \end{cases} \tag{6.24}$$

Case $\hat{\delta} > 1$: Since $t_{k+1}^{\eta}\varepsilon_k = \mathcal{O}(k^{-(\delta-\eta)})$ vanishes and $\sum_{\ell=k_0}^{k}\zeta_{\ell,\delta}$ is bounded, (6.23) yields $F_k = \mathcal{O}(k^{-\eta})$ as $k \to +\infty$, which proves the convergence rate in this case.

Before proceeding with the two remaining cases, let us collect a few facts. First, notice that when $0 < \hat{\delta} \leq 1$, we have $t_{k+1}^{\eta}\varepsilon_k = \mathcal{O}(k^{\eta-\delta})$, thus from (6.23) and (6.24) we get, as $k \to +\infty$,

$$F_k = \mathcal{O}(\ln(k)k^{-\eta}), \quad \text{if } \delta = \eta, \quad \text{and} \quad F_k = \mathcal{O}(k^{-\delta}) \quad \text{if } \delta < \eta. \tag{6.25}$$

Furthermore, we need three inequalities which are a consequence of (6.21) when $C_2 > 0$. After dropping the nonnegative term g_k when required, we come for all $k \geq k_0$ to

$$E_{k+1} - E_k \leq E_{k+1} - E_k + \frac{C_2 k^{\eta-1}}{2} F_k + g_k \leq -\frac{C_2 k^{\eta-1}}{2} \left(F_k + \frac{2\zeta_{k,\delta}}{C_2 k^{\eta-1}} H_k \right), \tag{6.26}$$

$$E_{k+1} - E_k + \frac{\zeta_{k,\delta}}{2} H_k \leq -C_2 k^{\eta-1} \left(F_k + \frac{\zeta_{k,\delta}}{2C_2 k^{\eta-1}} H_k \right), \tag{6.27}$$

$$E_{k+1} - E_k - \zeta_{k,\delta} H_k \leq -C_2 k^{\eta-1} \left(F_k + \frac{2\zeta_{k,\delta}}{C_2 k^{\eta-1}} H_k \right). \tag{6.28}$$

Furthermore, since $F_k \geq 0$, for all $r > 0$ and all $k \geq k_0$ we have, using (6.22)

$$-k^{\eta-1} \left(F_k + r\zeta_{k,\delta} k^{1-\eta} H_k \right) \leq \begin{cases} 0 & \text{if } H_k > 0, \\ -k^{\eta-1} \left(F_k + rC_{2,\zeta} \varepsilon_k H_k \right) & \text{if } H_k \leq 0. \end{cases}$$

Therefore, when ($C2$) holds, for all $r > 0$ there exists $(I_k^{rC_{2,\zeta}})_{k\in\mathbb{N}} \in \ell^1(\mathbb{N})$ nonnegative such that, for every $k \geq k_0$,

$$-k^{\eta-1} \left(F_k + r\zeta_{k,\delta} k^{1-\eta} H_k \right) \leq I_k^{rC_{2,\zeta}}. \tag{6.29}$$

First, we establish that $\lim_{k\to+\infty} E_k$ exists, and then come to the summability statements.

Case $\hat{\delta} = 1$: Suppose ($C2$) holds. On the one hand, when $\hat{\delta} = 1$, we have $E_k \geq -cH_*$ for k large enough. On the other hand, according to (6.26) and (6.29), $(E_{k+1} - E_k)_{k\geq 1}$ is upper bounded by a summable sequence. Combining these facts yields the existence of the limit $\lim_{k\to+\infty} E_k$, i.e., Point 1. In particular, this also yields $F_k = \mathcal{O}(k^{-\eta})$ as $k \to +\infty$, i.e., Point 2.

Case $\frac{1}{\rho^\star} < \hat{\delta} < 1$: Suppose ($C1$) and ($C2$) hold. Using (6.26) and (6.29) we get that $(E_{k+1} - E_k)_{k\geq 1}$ is upper bounded by a summable sequence. Therefore, $E_k \leq C_E$ for some constant C_E and all $k \geq k_0$. In particular, using also ($C1$), this implies the existence of a constant $C_{0,F}$ such that for every $k \geq k_0$

$$t_k^\eta F_k \leq C_E - t_k^\eta \varepsilon_{k-1} H_k \leq C_E + C_{0,F} t_k^{\eta-\delta} F_k^{\frac{1}{\rho}}.$$

Since (6.25) holds, we can apply Lemma 6.2.2, which yields $F_k = \mathcal{O}(k^{-\eta})$ as $k \to +\infty$, that is, Point 1. With this result, using ($C1$), for some constant $C_H \geq 0$ it holds for every $k \geq k_0$

$$-t_k^\eta \varepsilon_{k-1} H_k \leq t_k^\eta \varepsilon_{k-1} C_3 F_k^{\frac{1}{\rho}} \leq C_H k^{\eta-\delta-\frac{\eta}{\rho}} = C_H k^{-(\delta-\frac{\eta}{\rho^\star})}. \tag{6.30}$$

This yields Point 2. Since $\delta = \eta\hat{\delta} > \frac{\eta}{\rho^\star}$, the right-hand side of the previous inequality vanishes as $k \to +\infty$, which, since $F_k, V_k \geq 0$, implies that $(E_k)_{k\geq 1}$ is lower bounded. As $(E_{k+1} - E_k)_{k\in\mathbb{N}}$ is upper bounded by a summable sequence, we obtain the existence of $\lim_{k\to+\infty} E_k$ and Point 3.

Let us now conclude with the summability statements. In both cases, we have shown that the limit $\lim_{k\to+\infty} E_k$ exists. In particular, $(E_k)_{k\geq 1}$ is bounded. Plugging this statement back into (6.26) and taking into account that $F_k, g_k \geq 0$ for every $k \geq k_0$, gives

$$\sum_{k=1}^{+\infty} k^{\eta-1} F_k < +\infty \quad \text{and} \quad \sum_{k=1}^{+\infty} g_k < +\infty. \tag{6.31}$$

Furthermore, relying once again on (6.29), from (6.27) and (6.28) we may write, for summable, nonnegative sequences $(C_{1,H}^k)_{k\geq 1}$, $(C_{3,H}^k)_{k\geq 1}$ and constants $C_{2,H}, C_{4,H} > 0$, and for every $k \geq k_0$,

$$\begin{aligned} \zeta_{k,\delta} H_k &\leq C_{1,H}^k - C_{2,H}(E_{k+1} - E_k) \leq C_{1,H}^k + C_{2,H}|E_{k+1} - E_k|, \\ -\zeta_{k,\delta} H_k &\leq C_{3,H}^k - C_{4,H}(E_{k+1} - E_k) \leq C_{3,H}^k + C_{4,H}|E_{k+1} - E_k|. \end{aligned} \tag{6.32}$$

Summing up these two inequalities and denoting by $C_{5,H} := C_{2,H} + C_{4,H} > 0$, we obtain for all $k \geq k_0$

$$\begin{aligned} &\sum_{\ell=k_0}^{k} \left| C_{1,H}^\ell + C_{3,H}^\ell - (C_{2,H} + C_{4,H})(E_{\ell+1} - E_\ell) \right| \\ &= \sum_{\ell=k_0}^{k} \left(C_{1,H}^\ell + C_{3,H}^\ell - C_{5,H}(E_{\ell+1} - E_\ell) \right) = \sum_{\ell=k_0}^{k} \left(C_{1,H}^\ell + C_{3,H}^\ell \right) - C_{5,H}(E_{\ell+1} - E_{k_0}). \end{aligned}$$

Since the right-hand side has a limit as $k \to +\infty$, and $(C_{1,H}^k)_{k\geq 1}$, and $(C_{3,H}^k)_{k\geq 1}$ are summable and $C_{5,H} > 0$, we obtain that $(|E_{k+1} - E_k|)_{k\geq 1}$ is summable as well. Therefore, from (6.32) and (6.22) we get

$$\sum_{k=1}^{+\infty} \zeta_{k,\delta}|H_k| < +\infty, \quad \text{thus} \quad \sum_{k=1}^{+\infty} k^{\eta-1}\varepsilon_{k-1}|H_k| < +\infty.$$

This, together with (6.31), concludes the summability statements, i.e., Point 3 of the case $\hat{\delta} = 1$ and Point 4 of the case $\frac{1}{\rho^\star} < \hat{\delta} < 1$ and hence finishes the proof. □

6.3 Bilevel Proximal-Gradient method

Consider the first-order continuous-time dynamics in (6.5). To handle the non-smoothness inherent in (6.1), we adopt a standard semi-implicit discretization of the monotone inclusion form of (6.5), treating the non-smooth components implicitly and the smooth ones explicitly: for every $k \geq 0$,

$$\frac{x_{k+1} - x_k}{\theta} + \partial \hat{f}(x_{k+1}) + \nabla f(x_k) + \varepsilon_k \partial \hat{h}(x_{k+1}) + \varepsilon_k \nabla h(x_k) \ni 0. \tag{6.33}$$

Rearranging suitably and recalling that $\text{prox}_{\tau f} = (I + \tau\partial f)^{-1}$, we obtain Algorithm 6.1. To study Algorithm 6.1, we will closely follow the analysis of (6.5), which is significantly simpler. Let us introduce the following notation, for $k \geq 0$:

$$\Phi_k(x) := f(x) + \varepsilon_k h(x), \quad \hat{\Phi}_k(x) := \hat{f}(x) + \varepsilon_k \hat{h}(x), \quad \Psi_k(x) := \Phi_k(x) + \hat{\Phi}_k(x). \tag{6.34}$$

Observe that Algorithm 6.1 can be understood as a standard proximal-gradient algorithm for minimizing the sum of the non-smooth function $\hat{\Phi}_k$ and the smooth function Φ_k. Therefore, the reader can expect that the step size θ will be picked to satisfy $\theta \in (0, 2/L_k)$, where $L_k := L_f + \varepsilon_k L_h$. Since $\varepsilon_k \to 0$, any choice $\theta \in (0, 2/L_f)$ will suffice to show the convergence of the algorithm.

For analyzing the asymptotic behavior of Algorithm 6.1, we make use of the following discrete Lyapunov energy function, defined every $k \geq 1$:

$$E_k^\lambda := t_k(\Psi_{k-1}(x_k) - \Psi_{k-1}(x^*)) + \frac{\lambda}{2}\|x_k - x^*\|^2, \tag{6.35}$$

where $t_k := \theta(k + \gamma)$, with $\gamma \geq 0$, represents the discrete time, x^* is a solution to (6.1), and $\lambda > 1$. We thus obtain the following statement.

Lemma 6.3.1. *Let $(x_k)_{k\geq 0}$ be the sequence generated by Algorithm 6.1 and $(E_k^\lambda)_{k\geq 1}$ be defined as in (6.35). Then, there exists $k_0 \geq 1$ such that for all $k \geq k_0$*

$$E_{k+1}^\lambda - E_k^\lambda + \zeta_{k,\delta}(H(x_k) - H(x^*)) \leq -\theta(\lambda - 1)(F(x_k) - F(x^*)), \tag{6.36}$$

where $(\zeta_{k,\delta})_{k\geq 1}$ is a sequence defined by

$$\zeta_{k,\delta} := \theta(\lambda - 1)\varepsilon_k - t_k(\varepsilon_k - \varepsilon_{k-1}).$$

Proof. Let $k \geq 1$. We have

$$\begin{aligned} E_{k+1}^\lambda - E_k^\lambda &= t_{k+1}(\Psi_k(x_{k+1}) - \Psi_k(x^*)) - t_k(\Psi_k(x_k) - \Psi_k(x^*)) & (6.37)\\ &\quad + t_k(\Psi_k(x_k) - \Psi_k(x^*)) - t_k(\Psi_{k-1}(x_k) - \Psi_{k-1}(x^*)) & (6.38)\\ &\quad + \frac{\lambda}{2}\|x_{k+1} - x^*\|^2 - \frac{\lambda}{2}\|x_k - x^*\|^2. & (6.39) \end{aligned}$$

We work with each line separately. For line (6.37), we use Lemma 2.1.3. The proximal step in Algorithm 6.1 can be written as

$$x_{k+1} = x_k - \theta G_k(x_k), \quad \text{where} \quad G_k(y) := \frac{1}{\theta}\Big[y - \text{prox}_{\theta\hat{\Phi}_k}(y - \theta\nabla\Phi_k(y))\Big]. \tag{6.40}$$

Since Φ_k is L_k-smooth, according to Lemma 2.1.3, for every $x, y \in \mathcal{H}$ it holds

$$\Psi_k(y - \theta G_k(y)) \leq \Psi_k(x) - \frac{\theta}{2}(2 - \theta L_k)\|G_k(y)\|^2 + \langle G_k(y), y - x\rangle.$$

We use this inequality twice. First, set $y = x := x_k$. We obtain

$$\Psi_k(x_{k+1}) \leq \Psi_k(x_k) - \frac{\theta}{2}(2 - \theta L_k)\|G_k(x_k)\|^2. \tag{6.41}$$

Now, set $y := x_k$ and $x := x^*$. This yields

$$\Psi_k(x_{k+1}) \leq \Psi_k(x^*) - \frac{\theta}{2}(2 - \theta L_k)\|G_k(x_k)\|^2 + \langle G_k(x_k), x_k - x^*\rangle. \tag{6.42}$$

For k_0 large enough and $k \geq k_0$, we multiply (6.41) by $t_{k+1} - \lambda\theta \geq 0$ and (6.42) by $\lambda\theta \geq 0$, add the resulting inequalities, sum $-t_{k+1}\Psi_k(x^*)$ to both sides and use that $t_{k+1} = t_k + \theta$ to obtain

$$\begin{aligned} t_{k+1}(\Psi_k(x_{k+1}) - \Psi_k(x^*)) \leq t_k(\Psi_k(x_k) - \Psi_k(x^*)) - \theta(\lambda - 1)(\Psi_k(x_k) - \Psi_k(x^*)) \\ - \frac{\theta t_{k+1}}{2}(2 - \theta L_k)\|G_k(x_k)\|^2 + \lambda\theta\langle G_k(x_k), x_k - x^*\rangle. \end{aligned} \tag{6.43}$$

According to (6.40), we have

$$\begin{aligned} &\lambda\theta\langle G_k(x_k), x_k - x^*\rangle = -\lambda\langle x_{k+1} - x_k, x_{k+1} - x^*\rangle + \lambda\|x_{k+1} - x_k\|^2, \\ &\|G_k(x_k)\|^2 = \frac{1}{\theta^2}\|x_{k+1} - x_k\|^2. \end{aligned} \tag{6.44}$$

Now, for line (6.38), we write

$$\begin{aligned} &t_k(\Psi_k(x_k) - \Psi_k(x^*)) - t_k(\Psi_{k-1}(x_k) - \Psi_{k-1}(x^*)) \\ &= t_k\Big[\Psi_k(x_k) - \Psi_{k-1}(x_k) - (\Psi_k(x^*) - \Psi_{k-1}(x^*))\Big] = t_k(\varepsilon_k - \varepsilon_{k-1})(H(x_k) - H(x^*)), \end{aligned} \tag{6.45}$$

and for line (6.39) we have

$$\frac{\lambda}{2}\|x_{k+1} - x^*\|^2 - \frac{\lambda}{2}\|x_k - x^*\|^2 = \lambda\langle x_{k+1} - x_k, x_{k+1} - x^*\rangle - \frac{\lambda}{2}\|x_{k+1} - x_k\|^2. \tag{6.46}$$

Putting (6.43), (6.44), (6.45) and (6.46) together yields for all $k \geq k_0$

$$\begin{aligned} E^\lambda_{k+1} - E^\lambda_k \leq &- \theta(\lambda - 1)(\Psi_k(x_k) - \Psi_k(x^*)) - \frac{\theta t_{k+1}}{2}(2 - \theta L_k)\frac{\|x_{k+1} - x_k\|^2}{\theta^2} \\ &- \lambda\langle x_{k+1} - x_k, x_{k+1} - x^*\rangle + \lambda\|x_{k+1} - x_k\|^2 \\ &+ t_k(\varepsilon_k - \varepsilon_{k-1})(H(x_k) - H(x^*)) + \lambda\langle x_{k+1} - x_k, x_{k+1} - x^*\rangle - \frac{\lambda}{2}\|x_{k+1} - x_k\|^2 \\ = &- \theta(\lambda - 1)(F(x_k) - F(x^*)) + \Big[t_k(\varepsilon_k - \varepsilon_{k-1}) - \theta(\lambda - 1)\varepsilon_k\Big](H(x_k) - H(x^*)) \\ &- \left[\frac{2 - \theta L_k}{2\theta}t_{k+1} - \frac{\lambda}{2}\right]\|x_{k+1} - x_k\|^2. \end{aligned}$$

The fact that $L_k \to L_f$ as $k \to +\infty$ ensures that the term with $\|x_{k+1} - x_k\|^2$ is eventually negative, so we drop it to obtain (6.36), after possibly increasing k_0. □

Remark 6.3.2. The proof of Lemma 6.3.1 is largely inspired by the continuous-time analysis of (6.5), which centers around the Lyapunov energy function

$$E^\lambda(t) := t(\Psi_t(x(t)) - \Psi_t(x^*)) + \frac{\lambda}{2}\|x(t) - x^*\|^2 \tag{6.47}$$

defined for $t \geq t_0$, where $\Psi_t(x) := f(x) + \varepsilon(t)h(x)$ is the regularized function (6.3) with $\varepsilon = \varepsilon(t)$. A similar inequality to (6.36) can easily be obtained for (6.47), see Section 6.7.

Theorem 6.3.3. *Let $(x_k)_{k\geq 0}$ be the sequence generated by Algorithm 6.1. Then, the following statements are true:*

- *Case $\delta > 1$: It holds $F(x_k) - \min F = \mathcal{O}(k^{-1})$ as $k \to +\infty$;*

Suppose Assumption 6.3 holds.

- *Case $\delta = 1$: If F satisfies Assumption 6.2 with $\eta = 1$, as $k \to +\infty$ it holds*

$$F(x_k) - \min F = o(k^{-1}) \quad \text{and} \quad H(x_k) \to \min_{\operatorname{argmin} F} H.$$

  Furthermore, x_k converges weakly towards a solution to the bilevel problem (6.1) as $k \to +\infty$;

- *Case $\frac{1}{\rho^\star} < \delta < 1$: If F satisfies Assumption 6.1 with exponent ρ, as $k \to +\infty$ it holds*

$$F(x_k) - F(x^*) = o(k^{-1}) \quad \text{and} \quad |H(x_k) - \min_{\operatorname{argmin} F} H| = o(k^{-1+\delta}).$$

  Furthermore, x_k converges weakly to a solution to the bilevel problem (6.1) as $k \to +\infty$.

Proof. Let x^* be a solution to the bilevel problem (6.1) and $\lambda > 1$. The first part of the proof is a consequence of Lemma 6.2.3 with $\hat{\delta} := \delta$, $\eta := 1$, $H_* := H(x^*) - \inf_{x \in \mathcal{H}} H(x)$, $C_2 := \theta(\lambda - 1) > 0$, and

$$F_k := F(x_k) - F(x^*), \quad H_k := H(x_k) - H(x^*), \quad g_k := 0, \quad V_k := \frac{\lambda}{2}\|x_k - x^*\|^2$$

for all $k \geq k_0$, where $k_0 \geq 1$ is given by Lemma 6.3.1. As ε_k is picked as in (6.4), $\zeta_{k,\delta} = \theta(\lambda-1)\varepsilon_k - t_k(\varepsilon_k - \varepsilon_{k-1})$ can be shown to satisfy (6.22). Let $k \geq 1$. Indeed, according to the mean value theorem, we have, for some $\xi_k \in (k + \beta - 1, k + \beta)$, that

$$\varepsilon_{k-1} - \varepsilon_k = \frac{c}{(k+\beta-1)^\delta} - \frac{c}{(k+\beta)^\delta} = \frac{c\delta}{\xi_k^{\delta+1}},$$

which gives, recalling that $t_k = \theta(k + \gamma)$,

$$\frac{\theta(k+\gamma)}{(k+\beta)^{\delta+1}} \leq t_k(\varepsilon_{k-1} - \varepsilon_k) \leq \frac{\theta(k+\gamma)}{(k+\beta-1)^{\delta+1}}.$$

Hence, by possibly increasing k_0, the positive constants $C_{1,\zeta}, C_{2,\zeta}$ required by (6.22) can be easily shown to exist.

The case $\delta > 1$ follows directly from Lemma 6.2.3. We treat the other two cases simultaneously. We first show that Assumption 6.2 implies condition ($C2$), and that Assumption 6.1 implies conditions ($C1$) + ($C2$) in Lemma 6.2.3. Using (6.12), we observe that Assumption 6.2 implies that condition ($C2$) is satisfied for

$$I_k^r := (F - \min F)^*(r\varepsilon_k p^*) - \sigma_{\operatorname{argmin} F}(r\varepsilon_k p^*),$$

for $k \geq 0$, where $p^* \in -\partial H(x^*)$ and $p^* \in N_{\operatorname{argmin} F}(x^*)$ is given by Assumption 6.3. For all $r > 0$, $rp^* \in N_{\operatorname{argmin} F}(x^*)$, and the sequence $(I_k^r)_{k \geq 0}$ is indeed summable. Using Lemma 6.2.1, Assumption 6.1 implies condition ($C1$) for $C_3 := \|p^*\|(\tau\rho^{-1})^{-\frac{1}{\rho}}$, and, using (6.9), also condition ($C2$). Lemma 6.2.3 ensures that, in both cases, the limit $\lim_{k\to+\infty} E_k^\lambda$ exists. Choosing $1 < \lambda_1 < \lambda_2$, for every $k \geq 1$ we have

$$E_k^{\lambda_2} - E_k^{\lambda_1} = \frac{\lambda_2 - \lambda_1}{2}\|x_k - x^*\|^2, \quad \text{thus} \quad \lim_{k\to+\infty} \|x_k - x^*\|^2 \quad \text{exists.} \tag{6.48}$$

Since x^* was an arbitrary solution to the bilevel problem (6.1), we have verified the first condition of Opial's Lemma (see Lemma 2.1.1). It follows, by definition of E_k^λ, that

$$\lim_{k\to+\infty} q_k := t_k(F(x_k) - F(x^*)) + t_k\varepsilon_{k-1}(H(x_k) - H(x^*)) \quad \text{exists.}$$

At the same time, both for the cases $\delta = 1$ and $\frac{1}{\rho^\star} < \delta < 1$, according to the summability statements given by Lemma 6.2.3 we have

$$\sum_{k=1}^{+\infty} \frac{|q_k|}{t_k} \leq \sum_{k=1}^{+\infty}(F(x_k) - F(x^*)) + \sum_{k=1}^{+\infty} \varepsilon_{k-1}|H(x_k) - H(x^*)| < +\infty. \tag{6.49}$$

Since $\lim_{k\to+\infty} |q_k|$ exists as well, it must be the case that $\lim_{k\to+\infty} q_k = 0$.

We now establish the weak convergence of $(x_k)_{k\geq 0}$, followed by the derivation of the little-o rates. From (6.48) we get that $(x_k)_{k\geq 0}$ is bounded. Let $x^\diamond$ be a weak sequential cluster point of this sequence and $x_{k_j} \rightharpoonup x^\diamond$ as $j \to +\infty$. From the convergence rate $F(x_k) - F(x^*) = \mathcal{O}(k^{-1})$ as $k \to +\infty$ given by Lemma 6.2.3, and the weak lower semicontinuity of F, we deduce that $x^\diamond \in \operatorname{argmin} F$. On the other hand, for every $j \geq 0$ we have

$$t_{k_j}\varepsilon_{k_j-1}\big(H(x_{k_j}) - H(x^*)\big) \leq t_{k_j}(F(x_{k_j}) - F(x^*)) + t_{k_j}\varepsilon_{k_j-1}\big(H(x_{k_j}) - H(x^*)\big) = q_{k_j}. \tag{6.50}$$

Thus, passing to the liminf, using that $(t_k\varepsilon_{k-1})_{k\geq 1}$ is bounded away from zero, and the weak lower semicontinuity of H, we deduce that $H(x^\diamond) \leq H(x^*)$. As $x^\diamond \in \operatorname{argmin} F$ and x^* is an arbitrary solution to (6.1), $x^\diamond$ is also a solution to (6.1). Opial's Lemma (see Lemma 2.1.1) yields that the whole sequence $(x_k)_{k\geq 0}$ converges weakly to a solution to the bilevel problem (6.1). Let us denote this solution by x^*.

To derive the little-o rates, we distinguish between the two cases. If $\delta = 1$, we use the subgradient inequality applied to H with $-p^* \in \partial H(x^*)$. If $\frac{1}{\rho^\star} < \delta < 1$, we use the lower bound for $t_k\varepsilon_{k-1}(H(x_k) - H(x^*))$ provided by Lemma 6.2.3. For all $k \geq 1$, we have, respectively,

$$t_k(F(x_k) - F(x^*)) - t_k\varepsilon_{k-1}\langle p^*, x_k - x^*\rangle \leq t_k(F(x_k) - F(x^*)) + t_k\varepsilon_{k-1}(H(x_k) - H(x^*)) = q_k, \tag{6.51}$$

$$t_k(F(x_k) - F(x^*)) - C_H k^{-\delta+\frac{1}{\rho^\star}} \leq t_k(F(x_k) - F(x^*)) + t_k\varepsilon_{k-1}(H(x_k) - H(x^*)) = q_k. \tag{6.52}$$

When $\delta = 1$, we have $t_k\varepsilon_{k-1} = \mathcal{O}(1)$ as $k \to +\infty$. Passing to the limit and using $x_k \rightharpoonup x^*$ or $k^{-\delta+\frac{1}{\rho^\star}} \to 0$, and $q_k \to 0$ as $k \to +\infty$, we get $\lim_{k\to+\infty} t_k(F(x_k) - F(x^*)) = 0$. Consequently, since $q_k \to 0$, also the limit of $t_k\varepsilon_{k-1}(H(x_k) - H(x^*))$ exists and must be zero in both cases. □

Remark 6.3.4. In Algorithm 6.1 (which coincides with the proximal-gradient method introduced in [71]) we provide a convergence analysis by choosing the regularization sequence $(\varepsilon_k)_{k\in\mathbb{N}}$ as in (6.4) and making the geometric assumptions on the inner function, but allowing larger step sizes. As discussed in Section 6.5, our analysis can be extended to accommodate milder assumptions (such as the coercivity of F or H) and more general parameter choices. However, without stronger assumptions, convergence rates are typically only available in the *ergodic* or *best-iterate* sense, and convergence to solutions to (6.1) can generally only be characterized by the vanishing distance to the solution set of (6.1), and stated in finite-dimensional settings; see Section 6.5 for further details. We obtain best-iterate results similar to [77, Theorem 3], but assuming coercivity on either function F or H instead of H, and for the method in Algorithm 6.1, instead of the iterative scheme in [77]. Furthermore, our analysis can be extended to accommodate general parameter sequences (see Section 6.5). However, we find little motivation to do so, and numerical results (e.g. those reported in [33]) seem to indicate that polynomial choices tend to give the best performance in Tikhonov methods.

6.4 Bilevel Fast Proximal-Gradient method

In this section, we study the asymptotic behavior of a semi-implicit discretization of the monotone inclusion form of (6.6). Following the same rationale of (6.33), for every $k \geq 1$ we consider

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{\theta^2} + \frac{\alpha}{t_{k+1}}\frac{x_k - x_{k-1}}{\theta} + \partial\hat{f}(x_{k+1}) + \nabla f(y_k) + \varepsilon_k\partial h(x_{k+1}) + \varepsilon_k\nabla h(y_k) \ni 0,$$

where we momentarily leave some freedom to define the variable y_k. After multiplication by $s := \theta^2$, we obtain, for every $k \geq 1$,

$$x_{k+1} + s\big(\partial\hat{f}(x_{k+1}) + \varepsilon_k\partial\hat{h}(x_{k+1})\big) \ni x_k + \left(1 - \frac{\alpha\theta}{t_{k+1}}\right)(x_k - x_{k-1}) - s\big(\nabla f(y_k) + \varepsilon_k\nabla h(y_k)\big),$$

which, setting $t_k := \theta(k+\gamma)$, denoting $\alpha_k := 1 - \frac{\alpha\theta}{t_{k+1}}$, and picking $y_k := x_k + \alpha_k(x_k - x_{k-1})$ leads us to Algorithm 6.2. Employing the same notation as in the first-order setting (6.34), we can interpret Algorithm 6.2 as the *Fast Proximal-Gradient algorithm* (also known as FISTA), proposed by Beck and Teboulle in [29] and built upon ideas of Nesterov [80] and Güler [61], applied to (6.3) with $\varepsilon_k \to 0$. As the latter requires that $s \in (0, \frac{1}{L_k}]$ and $\varepsilon_k \to 0$, here we will pick $s \in (0, \frac{1}{L_f})$.

For analyzing its asymptotic behavior, we make use of the following energy function:

$$E_k^\lambda := t_k^2(\Psi_{k-1}(x_k)-\Psi_{k-1}(x^*))+\frac{1}{2}\left\|\lambda(x_{k-1}-x^*)+t_k\frac{x_k-x_{k-1}}{\theta}\right\|^2+\frac{\lambda(\alpha-1-\lambda)}{2}\|x_{k-1}-x^*\|^2, \quad (6.53)$$

for $k \geq 1$, where x^* is a solution to (6.1), $\theta := \sqrt{s}$, and $\lambda \in (2, \alpha-1)$, which is nonempty since $\alpha > 3$. Following closely the Lyapunov analysis of its continuous-time counterpart (6.6), we arrive at the following estimate.

Lemma 6.4.1. *Let $(x_k)_{k\geq 0}$ be the sequence generated by Algorithm 6.2 and $(E_k^\lambda)_{k\geq 1}$ be defined as in (6.53). Then, there exists $k_0 \geq 1$ such that for every $k \geq k_0$*

$$\begin{aligned} E_{k+1}^\lambda - E_k^\lambda + (\alpha-1-\lambda)\theta t_{k+1}\alpha_k\left\|\frac{x_k-x_{k-1}}{\theta}\right\|^2 + \zeta_{k,\delta}(H(x_k)-H(x^*)) \\ \leq -\theta^2(\lambda-2)k(F(x_k)-F(x^*)), \end{aligned} \quad (6.54)$$

where $(\zeta_{k,\delta})_{k\geq 1}$ is a sequence defined by

$$\zeta_{k,\delta} := \theta\big((\lambda-2)t_k+\theta(\lambda-1)\big)\varepsilon_k - t_k^2(\varepsilon_k-\varepsilon_{k-1}).$$

Proof. Let $k \geq 1$. First, let us denote for brevity $v_k := \lambda(x_{k-1}-x^*)+t_k\frac{x_k-x_{k-1}}{\theta}$. We have

$$\begin{aligned} E_{k+1}^\lambda - E_k^\lambda &= t_{k+1}^2(\Psi_k(x_{k+1})-\Psi_k(x^*)) - t_k^2(\Psi_k(x_k)-\Psi_k(x^*)) && (6.55)\\ &+ t_k^2(\Psi_k(x_k)-\Psi_k(x^*)) - t_k^2(\Psi_{k-1}(x_k)-\Psi_{k-1}(x^*)) && (6.56)\\ &+ \frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 && (6.57)\\ &+ \frac{\lambda(\alpha-1-\lambda)}{2}\|x_k-x^*\|^2 - \frac{\lambda(\alpha-1-\lambda)}{2}\|x_{k-1}-x^*\|^2. && (6.58) \end{aligned}$$

We will work with each line separately. For line (6.55), we use Lemma 2.1.3. The proximal step in Algorithm 6.2 can be written as

$$x_{k+1} = y_k - sG_k(y_k), \quad \text{where} \quad G_k(y) := \frac{1}{s}\Big[y - \operatorname{prox}_{s\hat{\Phi}_k}(y - s\nabla\Phi_k(y))\Big]. \quad (6.59)$$

Since Φ_k is L_k-smooth, according to Lemma 2.1.3, for all $x, y \in \mathcal{H}$ it holds

$$\Psi_k(y - sG_k(y)) \leq \Psi_k(x) - \frac{s}{2}(2-sL_k)\|G_k(y)\|^2 + \langle G_k(y), y-x\rangle.$$

We apply the inequality twice. First, set $y := y_k$ and $x := x_k$. We obtain

$$\Psi_k(x_{k+1}) \leq \Psi_k(x_k) - \frac{s}{2}(2-sL_k)\|G_k(y_k)\|^2 + \langle G_k(y_k), y_k-x_k\rangle. \quad (6.60)$$

Now, we set $y := y_k$ and $x := x^*$. This yields

$$\Psi_k(x_{k+1}) \leq \Psi_k(x^*) - \frac{s}{2}(2-sL_k)\|G_k(y_k)\|^2 + \langle G_k(y_k), y_k-x^*\rangle. \quad (6.61)$$

From now on, we will use that $s = \theta^2$. For k_0 large enough and $k \geq k_0$, we multiply (6.60) by $t_{k+1}-\lambda\theta \geq 0$, and (6.61) by $\lambda\theta \geq 0$, add the resulting inequalities and sum $-t_{k+1}\Psi_k(x^*)$ to both sides to obtain

$$\begin{aligned} t_{k+1}(\Psi_k(x_{k+1})-\Psi_k(x^*)) &\leq (t_{k+1}-\lambda\theta)(\Psi_k(x_k)-\Psi_k(x^*)) \\ &+ \big\langle G_k(y_k), (t_{k+1}-\lambda\theta)(y_k-x_k)+\lambda\theta(y_k-x^*)\big\rangle - \frac{\theta^2 t_{k+1}}{2}(2-\theta^2L_k)\|G_k(y_k)\|^2. \end{aligned}$$

We multiply both sides of the previous inequality by t_{k+1}, and add and subtract $t_k^2(\Psi_k(x_k)-\Psi_k(x^*))$, which yields

$$t_{k+1}^2(\Psi_k(x_{k+1})-\Psi_k(x^*)) \leq \Big[t_{k+1}^2-\lambda\theta t_{k+1}-t_k^2\Big](\Psi_k(x_k)-\Psi_k(x^*)) + t_k^2(\Psi_k(x_k)-\Psi_k(x^*))$$

$$+ t_{k+1}\Big\langle G_k(y_k), (t_{k+1} - \lambda\theta)(y_k - x_k) + \lambda\theta(y_k - x^*)\Big\rangle - \frac{\theta^2 t_{k+1}^2}{2}(2 - \theta^2 L_k)\|G_k(y_k)\|^2. \tag{6.62}$$

Continuing with line (6.56), we have

$$t_k^2(\Psi_k(x_k) - \Psi_k(x^*)) - t_k^2(\Psi_{k-1}(x_k) - \Psi_{k-1}(x^*)) \tag{6.63}$$
$$= t_k^2\Big[\Psi_k(x_k) - \Psi_{k-1}(x_k) - (\Psi_k(x^*) - \Psi_{k-1}(x^*))\Big]$$
$$= t_k^2(\varepsilon_k - \varepsilon_{k-1})(H(x_k) - H(x^*)). \tag{6.64}$$

For line (6.57), from (6.59) first we notice that

$$\begin{aligned}
v_{k+1} - v_k &= \lambda(x_k - x^*) + \frac{t_{k+1}}{\theta}(x_{k+1} - x_k) - \lambda(x_{k-1} - x^*) - \frac{t_k}{\theta}(x_k - x_{k-1}) \\
&= \frac{t_{k+1}}{\theta}(x_{k+1} - x_k) - \left(\frac{t_k}{\theta} - \lambda\right)(x_k - x_{k-1}) \\
&= \frac{t_{k+1}}{\theta}(x_{k+1} - x_k) - \left[\frac{t_{k+1}\alpha_k}{\theta} + (\alpha - 1 - \lambda)\right](x_k - x_{k-1}) \\
&= \frac{t_{k+1}}{\theta}\Big[x_{k+1} - \Big(x_k + \alpha_k(x_k - x_{k-1})\Big)\Big] - (\alpha - 1 - \lambda)(x_k - x_{k-1}) \\
&= \frac{t_{k+1}}{\theta}(x_{k+1} - y_k) - (\alpha - 1 - \lambda)(x_k - x_{k-1}) \\
&= -\theta t_{k+1}G_k(y_k) - (\alpha - 1 - \lambda)(x_k - x_{k-1}).
\end{aligned}$$

Therefore,

$$\langle v_{k+1} - v_k, v_{k+1}\rangle = \Big\langle -\theta t_{k+1}G_k(y_k) - (\alpha - 1 - \lambda)(x_k - x_{k-1}), \lambda(x_k - x^*) + t_{k+1}\frac{x_{k+1} - x_k}{\theta}\Big\rangle$$
$$= -\theta t_{k+1}\left\langle G_k(y_k), \lambda(x_k - x^*) + t_{k+1}\frac{x_{k+1} - x_k}{\theta}\right\rangle$$
$$- (\alpha - 1 - \lambda)\left\langle x_k - x_{k-1}, \lambda(x_k - x^*) + t_{k+1}\frac{x_{k+1} - x_k}{\theta}\right\rangle. \tag{6.65}$$

From Algorithm 6.2 and (6.59), we know that

$$\frac{x_{k+1} - x_k}{\theta} = -\theta G_k(y_k) + \alpha_k\frac{x_k - x_{k-1}}{\theta}.$$

Using this equality, we can write line (6.65) as

$$\begin{aligned}
&- (\alpha - 1 - \lambda)\left\langle x_k - x_{k-1}, \lambda(x_k - x^*) + t_{k+1}\frac{x_{k+1} - x_k}{\theta}\right\rangle \\
=& - \lambda(\alpha - 1 - \lambda)\langle x_k - x_{k-1}, x_k - x^*\rangle + (\alpha - 1 - \lambda)\theta t_{k+1}\langle x_k - x_{k-1}, G_k(y_k)\rangle \\
&- (\alpha - 1 - \lambda)\theta t_{k+1}\alpha_k\left\|\frac{x_k - x_{k-1}}{\theta}\right\|^2.
\end{aligned}$$

We have

$$\begin{aligned}
&\|v_{k+1} - v_k\|^2 = \Big\|-\theta t_{k+1}G_k(y_k) - (\alpha - 1 - \lambda)(x_k - x_{k-1})\Big\|^2 \\
=& \theta^2 t_{k+1}^2\|G_k(y_k)\|^2 + 2\theta t_{k+1}(\alpha - 1 - \lambda)\langle G_k(y_k), x_k - x_{k-1}\rangle + (\alpha - 1 - \lambda)^2\|x_k - x_{k-1}\|^2.
\end{aligned}$$

Putting everything together gives

$$\begin{aligned}
&\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 = \langle v_{k+1} - v_k, v_{k+1}\rangle - \frac{1}{2}\|v_{k+1} - v_k\|^2 \\
=& - \theta t_{k+1}\left\langle G_k(y_k), \lambda(x_k - x^*) + t_{k+1}\frac{x_{k+1} - x_k}{\theta}\right\rangle \\
&- \lambda(\alpha - 1 - \lambda)\langle x_k - x_{k-1}, x_k - x^*\rangle + (\alpha - 1 - \lambda)\theta t_{k+1}\langle G_k(y_k), x_k - x_{k-1}\rangle
\end{aligned}$$

$$- (\alpha - 1 - \lambda)\theta t_{k+1}\alpha_k \left\| \frac{x_k - x_{k-1}}{\theta} \right\|^2 - \frac{\theta^2 t_{k+1}^2}{2} \|G_k(y_k)\|^2$$
$$- \theta t_{k+1}(\alpha - 1 - \lambda)\langle G_k(y_k), x_k - x_{k-1}\rangle - \frac{(\alpha - 1 - \lambda)^2}{2}\|x_k - x_{k-1}\|^2. \tag{6.66}$$

For line (6.58), it holds

$$\frac{\lambda(\alpha - 1 - \lambda)}{2}\|x_k - x^*\|^2 - \frac{\lambda(\alpha - 1 - \lambda)}{2}\|x_{k-1} - x^*\|^2$$
$$= \lambda(\alpha - 1 - \lambda)\langle x_k - x_{k-1}, x_k - x^*\rangle - \frac{\lambda(\alpha - 1 - \lambda)}{2}\|x_k - x_{k-1}\|^2. \tag{6.67}$$

We now combine (6.62), (6.64), (6.66) and (6.67). We cancel out the corresponding terms, drop some nonpositive terms, again take into account that $y_k - x_{k+1} = \theta^2 G_k(y_k)$ as per (6.59), and observe that, since $\lambda \geq 2$, simple algebraic computations yield, foe every $k \geq 1$,

$$t_{k+1}^2 - \lambda\theta t_{k+1} - t_k^2 = \theta\Big[(2 - \lambda)t_k + \theta(1 - \lambda)\Big] \leq 0. \tag{6.68}$$

This leads us for all $k \geq k_0$ to

$$\begin{aligned}
& E_{k+1}^\lambda - E_k^\lambda \\
& \leq \Big[t_{k+1}^2 - \lambda\theta t_{k+1} - t_k^2\Big](\Psi_k(x_k) - \Psi_k(x^*)) + t_{k+1}\Big\langle G_k(y_k), (t_{k+1} - \lambda\theta)(y_k - x_k) + \lambda\theta(y_k - x^*)\Big\rangle \\
& \quad - \frac{\theta^2 t_{k+1}^2}{2}(2 - \theta^2 L_k)\|G_k(y_k)\|^2 + t_k^2(\varepsilon_k - \varepsilon_{k-1})(H(x_k) - H(x^*)) \\
& \quad - \theta t_{k+1}\left\langle G_k(y_k), \lambda(x_k - x^*) + t_{k+1}\frac{x_{k+1} - x_k}{\theta}\right\rangle - \lambda(\alpha - 1 - \lambda)\langle x_k - x_{k-1}, x_k - x^*\rangle \\
& \quad + (\alpha - 1 - \lambda)\theta t_{k+1}\langle G_k(y_k), x_k - x_{k-1}\rangle - (\alpha - 1 - \lambda)\theta t_{k+1}\alpha_k\left\|\frac{x_k - x_{k-1}}{\theta}\right\|^2 \\
& \quad - \frac{\theta^2 t_{k+1}^2}{2}\|G_k(y_k)\|^2 - \theta t_{k+1}(\alpha - 1 - \lambda)\langle G_k(y_k), x_k - x_{k-1}\rangle - \frac{(\alpha - 1 - \lambda)^2}{2}\|x_k - x_{k-1}\|^2 \\
& \quad + \lambda(\alpha - 1 - \lambda)\langle x_k - x_{k-1}, x_k - x^*\rangle - \frac{\lambda(\alpha - 1 - \lambda)}{2}\|x_k - x_{k-1}\|^2 \\
& \leq \theta\Big[(2 - \lambda)t_k + \theta(1 - \lambda)\Big](\Psi_k(x_k) - \Psi_k(x^*)) + t_{k+1}\Big\langle G_k(y_k), t_{k+1}(y_k - x_{k+1})\Big\rangle \\
& \quad - \frac{\theta^2 t_{k+1}^2}{2}(3 - \theta^2 L_k)\|G_k(y_k)\|^2 + t_k^2(\varepsilon_k - \varepsilon_{k-1})(H(x_k) - H(x^*)) \\
& \quad - (\alpha - 1 - \lambda)\theta t_{k+1}\alpha_k\left\|\frac{x_k - x_{k-1}}{\theta}\right\|^2 \\
& \leq -\theta(\lambda - 2)t_k(F(x_k) - F(x^*)) - \frac{\theta^2 t_{k+1}^2}{2}(1 - \theta^2 L_k)\|G_k(y_k)\|^2 - (\alpha - 1 - \lambda)\theta t_{k+1}\alpha_k\left\|\frac{x_k - x_{k-1}}{\theta}\right\|^2 \\
& \quad + \theta\Big[((2 - \lambda)t_k + \theta(1 - \lambda))\varepsilon_k + t_k^2\frac{\varepsilon_k - \varepsilon_{k-1}}{\theta}\Big](H(x_k) - H(x^*)).
\end{aligned}$$

By dropping the term with $-(1 - \theta^2 L_k)$, which is eventually nonpositive, since $\theta^2 < \frac{1}{L_f}$ and $L_k \to L_f$ as $k \to +\infty$, and by taking into account that $t_k = \theta(k + \gamma) \geq \theta k$, this yields (6.54), after possibly increasing k_0. $\square$

Remark 6.4.2. Even in this setting, the proof of Lemma 6.4.1 is largely inspired from that of the continuous-time system (6.6), which can be studied with the Lyapunov energy function, defined for $t \geq t_0$,

$$E^\lambda(t) := t^2(\Psi_t(x(t)) - \Psi_t(x^*)) + \frac{1}{2}\|\lambda(x(t) - x^*) + t\dot{x}(t)\|^2 + \frac{\lambda(\alpha - 1 - \lambda)}{2}\|x(t) - x^*\|^2, \tag{6.69}$$

where we recall that $\Psi_t(x) = f(x) + \varepsilon(t)h(x)$ is the regularized function introduced in (6.3). Observe the striking similarity with its discrete-time counterpart (6.53). A similar result to Lemma 6.4.1 can be established for (6.69), see Section 6.7.

Theorem 6.4.3. *Let $(x_k)_{k\geq 0}$ be the sequence generated by Algorithm 6.2. Then, the following statements are true:*

- *Case $\delta > 2$: It holds $F(x_k) - \min F = \mathcal{O}(k^{-2})$ as $k \to +\infty$;*

Suppose Assumption 6.3 holds.

- *Case $\delta = 2$: If F satisfies Assumption 6.2 with $\eta = 2$, as $k \to +\infty$ it holds*

$$F(x_k) - \min F = o(k^{-2}) \quad \text{and} \quad H(x_k) \to \min\nolimits_{\operatorname{argmin} F} H.$$

Furthermore, x_k converges weakly towards a solution to the bilevel problem (6.1) as $k \to +\infty$;

- *Case $\frac{2}{\rho^*} < \delta < 2$: If F satisfies Assumption 6.1 with exponent ρ, as $k \to +\infty$ it holds*

$$F(x_k) - \min F = o(k^{-2}) \quad \text{and} \quad |H(x_k) - \min\nolimits_{\operatorname{argmin} F} H| = o(k^{-2+\delta}).$$

Furthermore, x_k converges weakly towards a solution to the bilevel problem (6.1) as $k \to +\infty$.

Proof. Let x^* be a solution to the bilevel problem (6.1) and $\lambda \in (2, \alpha - 1])$. The first part of the proof is a consequence of Lemma 6.2.3 with $\hat{\delta} := \frac{\delta}{2}$, $\eta = 2$, $\theta := \sqrt{s}$, $H_* := H(x^*) - \inf_{x\in\mathcal{H}} H(x)$, $C_2 := \theta^2(\lambda - 2) > 0$, and

$$F_k := F(x_k) - F(x^*), \quad H_k := H(x_k) - H(x^*), \quad g_k := (\alpha - 1 - \lambda)\theta t_{k+1}\alpha_k \left\| \frac{x_k - x_{k-1}}{\theta} \right\|^2,$$
$$V_k := \frac{1}{2}\|v_k\|^2 + \frac{\lambda(\alpha - 1 - \lambda)}{2}\|x_{k-1} - x^*\|^2$$

for $k \geq k_0$, where k_0 is given by Lemma 6.4.1. Arguing similarly as in the proof for Theorem 6.3.3, we can produce positive constants $C_{1,\varsigma}, C_{2,\varsigma}$ such that (6.22) holds.

The case $\delta > 2$ follows directly from Lemma 6.2.3. As in the proof of Theorem 6.3.3, we treat the other two cases simultaneously. Similar arguments show that Assumption 6.2 implies condition ($C2$), and that Assumption 6.1 implies conditions ($C1$) + ($C2$), even with $\eta = 2$. Lemma 6.2.3 ensures that, in both cases, the limit $\lim_{k\to+\infty} E_k^\lambda$ exists. Choosing $2 < \lambda_1 < \lambda_2 < \alpha - 1$, we have

$$\begin{aligned}
E_k^{\lambda_2} - E_k^{\lambda_1} &= \frac{1}{2}\left\| \lambda_2(x_{k-1} - x^*) + t_k \frac{x_k - x_{k-1}}{\theta} \right\|^2 + \frac{\lambda_2(\alpha - 1 - \lambda_2)}{2}\|x_{k-1} - x^*\|^2 \\
&\quad - \frac{1}{2}\left\| \lambda_1(x_{k-1} - x^*) + t_k \frac{x_k - x_{k-1}}{\theta} \right\|^2 - \frac{\lambda_1(\alpha - 1 - \lambda_1)}{2}\|x_{k-1} - x^*\|^2 \\
&= \frac{\lambda_2^2}{2}\|x_{k-1} - x^*\|^2 + \lambda_2 t_k \left\langle x_{k-1} - x^*, \frac{x_k - x_{k-1}}{\theta} \right\rangle + \frac{t_k^2}{2}\left\| \frac{x_k - x_{k-1}}{\theta} \right\|^2 \\
&\quad + \left[\frac{\lambda_2(\alpha - 1)}{2} - \frac{\lambda_2^2}{2} \right] \|x_{k-1} - x^*\|^2 - \frac{\lambda_1^2}{2}\|x_{k-1} - x^*\|^2 - \lambda_1 t_k \left\langle x_{k-1} - x^*, \frac{x_k - x_{k-1}}{\theta} \right\rangle \\
&\quad - \frac{t_k^2}{2}\left\| \frac{x_k - x_{k-1}}{\theta} \right\|^2 - \left[\frac{\lambda_1(\alpha - 1)}{2} - \frac{\lambda_1^2}{2} \right] \|x_{k-1} - x^*\|^2 \\
&= (\lambda_2 - \lambda_1) \underbrace{\left[\frac{\alpha - 1}{2}\|x_{k-1} - x^*\|^2 + t_k \left\langle x_{k-1} - x^*, \frac{x_k - x_{k-1}}{\theta} \right\rangle \right]}_{=: p_k} \quad \text{for all } k \geq 1.
\end{aligned}$$

Since $\lambda_2 - \lambda_1 \neq 0$, and $E_k^{\lambda_1}, E_k^{\lambda_2}$ admit a limit, also $\lim_{k\to+\infty} p_k$ exists. We now rewrite $E_k^{\lambda_1}$ as

$$\begin{aligned}
E_k^{\lambda_1} &= t_k^2(\Psi_{k-1}(x_k) - \Psi_{k-1}(x^*)) + \frac{\lambda_1^2}{2}\|x_{k-1} - x^*\|^2 + \lambda_1 t_k \left\langle x_{k-1} - x^*, \frac{x_{k-1} - x_k}{\theta} \right\rangle \\
&\quad + \frac{t_k^2}{2}\left\| \frac{x_{k-1} - x_k}{\theta} \right\|^2 + \left[\frac{\lambda_1(\alpha - 1)}{2} - \frac{\lambda_1^2}{2} \right] \|x_{k-1} - x^*\|^2
\end{aligned}$$

$$= t_k^2(\Psi_{k-1}(x_k) - \Psi_{k-1}(x^*)) + \frac{t_k^2}{2}\left\|\frac{x_k - x_{k-1}}{\theta}\right\|^2 + \lambda_1 p_k$$
$$= \underbrace{t_k^2(F(x_k) - F(x^*)) + t_k^2\varepsilon_{k-1}(H(x_k) - H(x^*)) + \frac{t_k^2}{2}\left\|\frac{x_k - x_{k-1}}{\theta}\right\|^2}_{=:q_k} + \lambda_1 p_k \quad \text{for all } k \geq 1.$$

Since both $\lim_{k\to+\infty} E_k^{\lambda_1}$ and $\lim_{k\to+\infty} p_k$ exist, so does $\lim_{k\to+\infty} q_k$. Both for the cases $\delta = 2$ and $\frac{2}{\rho^*} < \delta < 2$, according to the summability statements given by Lemma 6.2.3 together with the fact that $t_k\varepsilon_{k-1} = \mathcal{O}(k^{1-\delta})$ as $k \to +\infty$, we have

$$\sum_{k=1}^{+\infty} \frac{|q_k|}{t_k} \leq \sum_{k=1}^{+\infty} t_k(F(x_k) - F(x^*)) + \sum_{k=1}^{+\infty} t_k\varepsilon_{k-1}|H(x_k) - H(x^*)| + \frac{1}{2}\sum_{k=1}^{+\infty} t_k\left\|\frac{x_k - x_{k-1}}{\theta}\right\|^2 < +\infty.$$

Since $\lim_{k\to+\infty} |q_k|$ exists as well, it must be the case that $\lim_{k\to+\infty} q_k = 0$. Observe that, as $\left(t_k\|\frac{x_k - x_{k-1}}{\theta}\|^2\right)_{k\geq 1}$ is summable, its limit exists and is zero. If we set $\varphi_k := \frac{1}{2}\|x_{k-1} - x^*\|^2$, we have that for all $k \geq 1$

$$(\alpha - 1)\varphi_{k+1} + \frac{t_k - \theta(\alpha - 1)}{\theta}(\varphi_{k+1} - \varphi_k) = (\alpha - 1)\varphi_k + \frac{t_k}{\theta}(\varphi_{k+1} - \varphi_k)$$
$$= \frac{\alpha - 1}{2}\|x_{k-1} - x^*\|^2 + \frac{t_k}{\theta}\left[\langle x_k - x_{k-1}, x_k - x^*\rangle - \frac{1}{2}\|x_k - x_{k-1}\|^2\right]$$
$$= \frac{\alpha - 1}{2}\|x_{k-1} - x^*\|^2 + \frac{t_k}{\theta}\left[\langle x_k - x_{k-1}, x_{k-1} - x^*\rangle + \frac{1}{2}\|x_k - x_{k-1}\|^2\right] = p_k + \frac{t_k}{2\theta}\|x_k - x_{k-1}\|^2.$$

Therefore, according to Lemma 2.3.4, it must be the case that $\lim_{k\to+\infty} \varphi_k$ exists as well. As x^* is an arbitrary solution to (6.1), we have verified the first condition of Opial's Lemma (see Lemma 2.1.1).

We proceed to show weak convergence of $(x_k)_{k\geq 0}$ and then little-o rates. As $\lim_{k\to+\infty} \varphi_k \in \mathbb{R}$, we get that $(x_k)_{k\geq 0}$ is bounded. Let $x^\diamond$ be a weak sequential cluster point and $x_{k_j} \rightharpoonup x^\diamond$ as $j \to +\infty$. From the convergence rate $F(x_k) - F(x^*) = \mathcal{O}(k^{-2})$ as $k \to +\infty$ given by Lemma 6.2.3, and the weak lower semicontinuity of F, we deduce that $x^\diamond \in \operatorname{argmin} F$. On the other hand, denoting by $\hat{g}_k := \frac{t_k^2}{2\theta^2}\|x_k - x_{k-1}\|^2$, we have

$$t_{k_j}^2\varepsilon_{k_j-1}\big(H(x_{k_j}) + H(x^*)\big) \leq t_{k_j}^2(F(x_{k_j}) - F(x^*)) + t_{k_j}^2\varepsilon_{k_j-1}\big(H(x_{k_j}) + H(x^*)\big) + \hat{g}_{k_j} = q_{k_j} \quad \text{for all } j \geq 0. \tag{6.70}$$

Thus, passing to the liminf, using that $(t_k^2\varepsilon_{k-1})_{k\geq 1}$ is bounded away from zero, and the weak lower semicontinuity of H, we deduce that $H(x^\diamond) \leq H(x^*)$. As $x^\diamond \in \operatorname{argmin} F$ and x^* is an arbitrary solution to (6.1), $x^\diamond$ is also a solution to (6.1). Opial's Lemma (see Lemma 2.1.1) yields that the whole sequence $(x_k)_{k\geq 0}$ converges weakly to a solution to the bilevel problem (6.1). Let us denote this solution by x^* and consider the above estimates with respect to this specific solution.

To derive little-o rates, we proceed in a similar way as in the proof for Algorithm 6.1. If $\delta = 2$, we use the subgradient inequality applied to H with $-p^* \in \partial H(x^*)$. If $\frac{2}{\rho^*} < \delta < 2$, we use the lower bound for $t_k^2\varepsilon_{k-1}(H(x_k) - H(x^*))$ provided by Lemma 6.2.3. For all $k \geq 1$, we have, respectively,

$$t_k^2(F(x_k) - F(x^*)) + \hat{g}_k \leq q_k + t_k^2\varepsilon_{k-1}\langle p^*, x_k - x^*\rangle, \tag{6.71}$$
$$t_k^2(F(x_k) - F(x^*)) + \hat{g}_k \leq q_k + C_H k^{-\delta+\frac{2}{\rho^*}}. \tag{6.72}$$

When $\delta = 2$, we have $t_k^2\varepsilon_{k-1} = \mathcal{O}(1)$ as $k \to +\infty$. Passing to the limit and using $x_k \rightharpoonup x^*$ or $k^{-\delta+\frac{2}{\rho^*}} \to 0$, and $q_k \to 0$ as $k \to +\infty$, we get $\lim_{k\to+\infty} t_k^2(F(x_k) - F(x^*)) = 0$. Consequently, since $q_k \to 0$, also the limit of $t_k^2\varepsilon_{k-1}(H(x_k) - H(x^*))$ exists and must be zero in both cases. □

Remark 6.4.4. For Algorithm 6.2, the only method available for a fair comparison is that of [78]. Our results substantially extend those in [78] in several directions. First, we introduce a more general algorithmic framework. Second, we allow for ρ-Hölderian growth with any $\rho \in (1, 2]$, rather than restricting to the quadratic case $\rho = 2$. Third, our analysis is carried out in an infinite-dimensional Hilbert space setting. Fourth, for $\delta \in (\frac{2}{\rho^*}, 2)$, we refine the convergence rates established in [78], improving the simultaneous bounds from $\mathcal{O}(k^{-2})$ and $\mathcal{O}(k^{\delta-2})$ to $o(k^{-2})$ and $o(k^{\delta-2})$ as $k \to +\infty$, respectively, by building on techniques from [17]. In addition, we complement the results of [78] by treating the case $\delta = 2$ under the Attouch–Czarnecki assumption, which is weaker than the Hölderian growth condition, and by establishing weak convergence of the sequence$(x_k)_{k\in\mathbb{N}}$ in this setting. Finally, our analysis departs fundamentally from prior work in that all results are derived from the dissipation of a single energy functional: the so-called *Lyapunov function.*

6.5 Discussion

In Sections 6.3 and 6.4, we analyzed the asymptotic behavior of Algorithms 6.1 and 6.2 under certain geometric assumptions on F. In what follows, we present some additional results that arise from our analysis. Although these were not included in the main theorems, they are nonetheless noteworthy. For the sake of brevity, we focus each time on either Algorithm 6.2 or Algorithm 6.1, as similar arguments apply to both cases.

Ergodic convergence rates. Let us first drop all geometric assumptions and consider Algorithm 6.2. When $0 < \delta < 2$, starting from the dissipativity estimate in Lemma 6.4.1 and arguing as in (6.25), as $k \to +\infty$ we obtain

$$F(x_k) - \min F = \mathcal{O}(k^{-\delta}) \quad \text{and} \quad \min_{k_0 \le \ell \le k+1} \{H(x_\ell) - \min_{\operatorname{argmin} F} H\} = \mathcal{O}(k^{-2+\delta}), \tag{6.73}$$

which was reported in [78]. Alternatively, if $\delta \in (0, 2)$ and $\delta \neq 1$, we can obtain *simultaneous* (i.e., computed on the same sequence) ergodic convergence rates expressed in terms of

$$x_k^{\text{best}} := \operatorname{argmin} \{H(x_{k+1}), H(\overline{x}_{k,\zeta})\}, \quad \text{where} \quad \overline{x}_{k,\zeta} := \frac{1}{Z_{k,\delta}} \sum_{\ell=k_0}^{k} \zeta_{\ell,\delta} x_\ell, \quad Z_{k,\delta} := \sum_{\ell=k_0}^{k} \zeta_{\ell,\delta}. \tag{6.74}$$

Specifically, since $F(x_k) - \min F = \mathcal{O}(k^{-\delta})$, summing up from k_0 to k inequality (6.54), using Jensen's inequality, the convexity of F, and the growth $Z_{k,\delta} = \mathcal{O}(k^{-2+\delta})$ as $k \to +\infty$, as $k \to +\infty$ we obtain[1]

$$F(x_k^{\text{best}}) - \min F = \mathcal{O}(k^{-\delta}) \quad \text{and} \quad H(x_k^{\text{best}}) - \min_{\operatorname{argmin} F} H \le \mathcal{O}(k^{-2+\delta}). \tag{6.75}$$

Furthermore, if $\delta = 1$—which is the case considered in [78, Theorem 3]—the convergence rate for the inner function is weaker by a logarithmic factor than in (6.75), $F(x_k^{\text{best}}) - F(x^*) = \mathcal{O}(k^{-1}\ln(k))$ as $k \to +\infty$, and $(\zeta_{\ell,\delta})_{k\ge1}$ is bounded and bounded away from zero, so $\overline{x}_{k,\zeta}$ is actually closer to the usual ergodic sequence. It follows from (6.75), that if $(x_k^{\text{best}})_{k\ge1}$ is bounded (e.g., if F or H are coercive[2], once again from (6.75)), weak cluster points of $(x_k^{\text{best}})_{k\ge1}$ lie in the solution set of (6.1). Under this assumption, if (6.1) admits a unique solution, the whole sequence $(x_k^{\text{best}})_{k\ge1}$ converges weakly to it. Alternatively, without assuming uniqueness of solutions but $\dim \mathcal{H} < +\infty$, we immediately obtain that the distance from $(x^k)_{k\ge0}$ to the solutions set of the bilevel problem (6.1) converges to zero as $k \to +\infty$, which is in line with [77, 78, 71]. However, in infinite dimensions, no iterates convergence can be derived with this approach.

Weakening the Hölderian growth condition. When F is coercive, we can assume the Hölderian growth condition, i.e., Assumption 6.1, to hold only *locally*. Indeed, suppose that, for some $R > 0$,

$$\tau\rho^{-1} \operatorname{dist}(x, \operatorname{argmin} F)^\rho \le F(x) - \min F \quad \text{for all } x \in C_R := \operatorname{argmin} F + B_R(0). \tag{6.76}$$

Using (6.73), it is easy to see that $(x_k)_{k\ge0}$ eventually lies in the bounded set C_R. Therefore, up to replacing the non-smooth part of F by setting $\hat{f}_R := \hat{f} + \iota_{C_R}$, which does not change the algorithm, the same analysis assuming Assumption 6.1 globally on $\hat{f}_R$ yields Point 3 in Theorem 6.4.3. A similar remark can be done for Algorithm 6.1.

[1] Note that, strictly speaking, the bound on $H(x^k) - \min_{\operatorname{argmin} F} H$ is not a *convergence rate*, but only a *one-sided estimate*. Here, we use a terminology from [78]. We can provide a two-sided estimate, i.e., a convergence rate, only in the setting of Theorem 6.4.3.

[2] In [77], the coercivity of H is assumed to ensure the boundedness of $(x_k^{\text{best}})_{k\ge1}$. Here, we show that, alternatively, it suffices to assume coercivity of F instead.

General regularization sequence $(\varepsilon_k)_{k\in\mathbb{N}}$. In Lemma 6.3.1 and Lemma 6.4.1 we have intentionally made explicit the dependence on the regularization sequence $(\varepsilon_k)_{k\geq 0}$ and indeed the proof does not depend on the specific definition of ε_k in (6.4). Following the same arguments of Theorem 6.3.3 and Theorem 6.4.3 we can thus extend our analysis to more general parameter sequences, provided that $\zeta_{k,\delta} \geq 0$ for all $k \geq 1$. Let us illustrate one of such extensions for Algorithm 6.1.

Consider (6.36) and denote by $\zeta_{k,\varepsilon}$ the term $\zeta_{k,\delta}$ to emphasize that the term depends on $(\varepsilon_k)_{k\geq 0}$, which is not necessarily of polynomial decrease, but, rather, the discrete counterpart to $(\lambda-1)\varepsilon(t) - t\dot{\varepsilon}(t)$, see $\zeta(t)$ in Appendix 6.7.1. Then, for all $k \geq k_0$, applying the same argument as in (6.12), we obtain, setting $\bar{\zeta}_{k,\varepsilon} := \frac{\zeta_{k,\varepsilon}}{s(\lambda-1)}$,

$$\begin{aligned} E_{k+1}^{\lambda} - E_k^{\lambda} &\leq -s(\lambda-1)\big[\bar{\zeta}_{k,\varepsilon}(H(x_k) - H(x^*)) - (F(x_k) - F(x^*))\big] \\ &\leq s(\lambda-1)\big[(F - \min F)^*(\bar{\zeta}_{k,\varepsilon} p^*) - \sigma_C(\bar{\zeta}_{k,\varepsilon} p^*)\big]. \end{aligned} \tag{6.77}$$

In this context, the Attouch–Czarnecki condition, i.e., Assumption 6.2, would now require the summability of the right-hand side of (6.77). This yields, after suitably bounding E_k^{λ} from below, that the limit $\lim_{k\to+\infty} E_k^{\lambda}$ exists, and in particular, for some constant $C > 0$,

$$t_k(F(x_k) - \min F) \leq C + t_k\varepsilon_k \min_{\operatorname{argmin} F} H, \quad \text{i.e.,} \quad F(x_k) - \min F = \mathcal{O}(k^{-1} + \varepsilon_k) \tag{6.78}$$

as $k \to +\infty$. Plugging ε_k as in (6.4), (6.78) is in line with Point 2 of Theorem 6.3.3. The little-o rate necessitates a more careful analysis, which we believe can be performed using the same techniques of Theorem 6.3.3.

6.6 Numerical Experiments

In this section, we present our numerical experiments, which are performed in Python on a 12thGen. Intel(R) Core(TM) i7–1255U, 1.70–4.70 GHz laptop with 16 Gb of RAM and are available for reproducibility at https://github.com/echnen/fast-bilevel-methods.

6.6.1 Considered algorithms

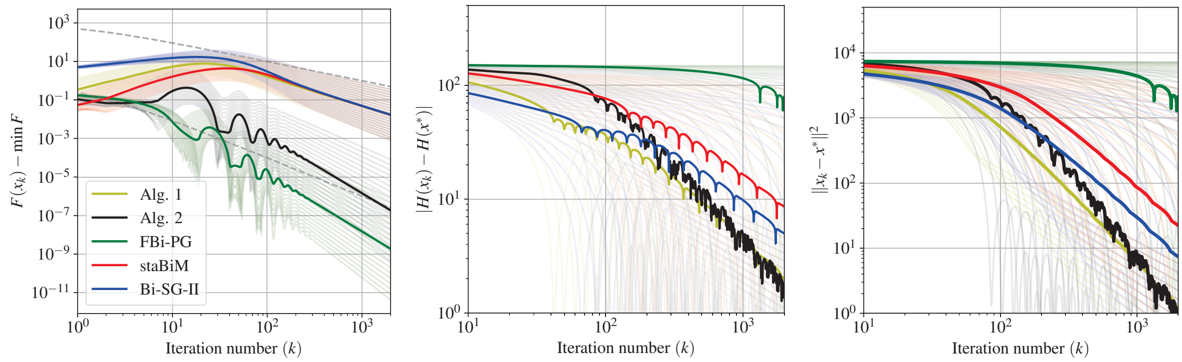
To begin, we briefly present the numerical algorithms utilized for comparison to test the performance of Algorithms 6.1 and 6.2. To the best of our knowledge, these are the only approaches that can cope with the generality of (6.1), so we refrain from discussing other variants requiring, e.g., strong convexity of the outer function.

Fast Bi-level Proximal Gradient. The *Fast Bi-level Proximal Gradient method* (FBi-PG) was recently introduced by Merchav, Sabach and Teboulle in [78]. It can be understood a special case of Algorithm 6.2 with $\gamma = \alpha - 1$, $c = 1$, and $\beta = \gamma$. Specifically, although the method is introduced in [78, (2.4)–(2.6)] with t_k picked as in the original Nesterov's method, the convergence results are only stated with $t_k := \frac{k+a}{a}$ for some $a \geq 2$. This choice yields a momentum parameter $\alpha_k := t_k^{-1}(t_{k-1} - 1)$ that can be obtained from our general choice in Algorithm 6.2 by setting $\gamma = \alpha - 1$ and $a := \alpha - 1$. Being a special case of Algorithm 6.2, its convergence properties are specified in Theorem 6.4.3.

Static Bilevel Method. The *Static Bilevel Method* (staBiM) was introduced by Latafat, Themelis, Villa and Patrinos in [71]. While the method is designed for an arbitrary slowly vanishing, i.e., not summable, regularization sequence $(\varepsilon_k)_{k\in\mathbb{N}}$, in these experiments we always consider ε_k as in (6.4) with $\delta \in (0, 1)$. For a fixed $\tilde{\theta} \in (0, 1)$ and $\eta_0 > 0$, the method iterates for all $k \geq 0$

$$x_{k+1} := \operatorname{prox}_{\theta_{k+1}(\hat{f}+\varepsilon_k\hat{h})}\big(x_k - \theta_{k+1}(\nabla f(x_k) + \varepsilon_k\nabla h(x_k))\big), \tag{6.79}$$

where $(\theta_k)_{k\geq 1}$ is the sequence defined by $\theta_{k+1} := \frac{\tilde{\theta}}{\eta_{k+1}L_h + L_f}$, where η_{k+1} can be picked arbitrarily in $\left[\frac{3}{4}\eta_k, \eta_k\right]$ for all $k \geq 0$. Observe the similarity with Algorithm 6.1: While allowing for general regularization sequences $(\varepsilon_k)_{k\geq 0}$, and originally designed in the more general context of locally Lipschitz gradients, it requires a non-stationary step size θ_k that, picking $\eta_{k+1} = \frac{3}{4}\eta_k$ for all $k \in \mathbb{N}$, asymptotically behaves like $\theta_k \simeq \frac{\tilde{\theta}}{L_f}$ with $\tilde{\theta} \in (0, 1)$. In contrast, our analysis suggests that indeed, we can choose i) stationary and ii) twice larger step sizes as $\theta_k = \frac{\tilde{\theta}}{L_f}$ and $\tilde{\theta} \in (0, 2)$ for every $k \geq 1$.



(a) Inner objective residual. (b) Outer objective residual. (c) Distance to solution.

Figure 6.1: Result of the experiment in Subsection 6.6.2: Comparing inner and outer function residuals as well as distance to solution for the methods listed in Subsection 6.6.1.

Bi-Sub-Gradient method. The *Bi-Sub-Gradient method* in its two different variants was introduced by Merchav and Sabach in [77]. In this paper, we focus on the so-called Variant II, (Bi-SG-II), which, for $\varepsilon_k := \frac{c}{(k+1)^\delta}$, $\delta \in (\frac{1}{2}, 1]$, $c \leq \min\{\frac{1}{L_h}, 1\}$, and $\theta \leq \frac{1}{L_f}$, iterates for all $k \geq 0$

$$\begin{cases} y_{k+1} := \operatorname{prox}_{\theta \hat{f}}(x_k - \theta \nabla f(x_k)), \\ x_{k+1} := \operatorname{prox}_{\theta \varepsilon_k \hat{h}}(y_{k+1} - \theta \varepsilon_k \nabla h(y_{k+1})). \end{cases} \tag{6.80}$$

The method alternates a proximal-gradient step in the inner problem (with step size θ) and in the outer problem (with step size $\theta\varepsilon_k$). The benefit of this formulation lies in the fact that it does not require computing the proximity operator of $\theta \hat{f} + \theta \varepsilon_k \hat{h}$, which for staBiM and Algorithm 6.1 requires a product-space trick as suggested in [54].

6.6.2 Linear inverse problem with simulated data

To test the dependency of the proposed method on the algorithm's parameter, we consider the following bilevel problem

$$\min_{x \in \mathbb{R}^d} \|x - \hat{x}\|_1 \quad \text{subject to } x \in \operatorname{argmin}_{z \in \mathbb{R}^d} f(z)\,, \tag{6.81}$$

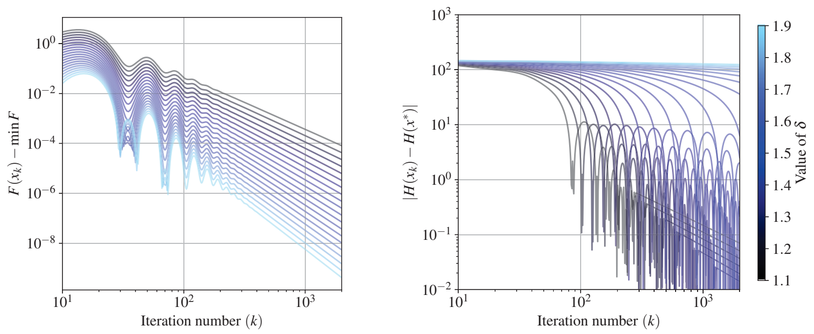
where $f : \mathbb{R}^d \to \mathbb{R}$ is the quadratic function defined, for $1 < J < d$, by

$$f(x) := \frac{1}{2}(x_1 - 1)^2 + \frac{1}{2}\sum_{j=2}^{J}(x_{j-1} - x_j)^2 \tag{6.82}$$

for every $x \in \mathbb{R}^d$, and $\hat{x} := (50, \ldots, 50)$. The inner function (6.82) is a standard benchmark to test first-order methods, usually revealing their sublinear rates of convergence [79]. In particular, it is not strongly convex, but satisfies Assumption 6.1 with $\rho = 2$. Since $J < d$, the set of minimizers of f is $\operatorname{argmin} f = \{x \in \mathbb{R}^d \,:\, x_1 = \cdots = x_J = 1\}$ and thus the unique solution to (6.81) is $x^* := (1, \ldots, 1, 50, \ldots, 50)$.

In our first experiment, we compare the performance of Algorithms 6.1 and 6.2 with $\beta = 10$, $\gamma = 20$, $c = 10$, and $\alpha = 4$ against the methods described in Subsection 6.6.1. For fair comparison, for each of them we set 95% of the largest possible step size, and consider the same initial point $x^0 := 0 \in \mathbb{R}^d$. Then, we let the algorithms run for 10^4 iterations with 20 different choices of $\delta \in (1, 2)$, sampled randomly—for Algorithm 6.1, staBiM and Bi-SG-II we use $\delta/2$ instead of δ to ensure convergence. We measure inner and outer residual values and show the results in Figure 6.1.

Although the objective of this paper is not to compete with existing algorithms, we can observe that the improved flexibility of Algorithm 6.2 plays a crucial role in performance, yielding results that clearly surpass those of FBi-PG. In this example, specifically from Figure 6.1a, we can also observe that the first-order methods Algorithm 6.1, staBiM and Bi-SG-II may indeed only achieve the rate $o(k^{-1})$ on the inner objective function decrease. On the other hand, the two second-order methods employed, i.e., Algorithm 6.2 and FBi-PG, show a much faster convergence behavior. Further, Figure 6.2 confirms the inherent trade-off between inner and outer function performance as δ varies, cf. Section 6.5.



(a) Inner objective residual

(b) Outer objective residual

Figure 6.2: Results of experiment in Section 6.6.2: Comparing residual decrease for Algorithm 6.2 as a function of the iteration number emphasizing the choice of $\delta \in (1, 2)$.

6.6.3 Logistic regression

In this experiment, we consider a copy of the UCI ML Breast Cancer Wisconsin (Diagnostic) dataset [98] freely available online [83]. It contains $n = 455$ instances with 30 features and a label y_i ($i = 1, \ldots, 455$) denoting weather the lesion is benign $y_i = 1$ or malign $y_i = 0$. We aim training a quadratic classifier, which for each sample $a \in \mathbb{R}^{30}$ yields 1 if $\mathfrak{s}(\hat{a}^T x) > \frac{1}{2}$ and 0 else, where $\hat{a} \in \mathbb{R}^d$ is the lifted feature vector of a that now belongs to a $d = 5456$ dimensional space, and $\mathfrak{s}(t) := (1 + \exp(-t))^{-1}$ is the standard sigmoid function. Due to overparametrization, among the optimal classifiers, we seek the one that minimizes the ℓ^1 norm by solving the following bilevel problem

$$\min_{x \in \mathbb{R}^d} \|x\|_1 \quad \text{subject to } x \in \operatorname{argmin}_{z \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i \log(\mathfrak{s}(\hat{a}_i^T z)) + (1 - y_i) \log(1 - \mathfrak{s}(\hat{a}_i^T z)) \Big). \tag{6.83}$$

To train the model, we run the three numerical algorithms presented in Subsection 6.6.1 and Algorithms 6.1 and 6.2 for $5 \cdot 10^4$ iterations with the $\delta = 1.9$ for the second-order methods, and $\delta/2 = 0.95$ for the first-order ones. Specifically, for Algorithm 6.2 we set $\gamma = \beta = 1$, $\alpha = 4$, $c = 10^2$, $s = \frac{0.95}{L_f}$. For FBi-PG, $\alpha = 4$, and $s = \frac{0.95}{L_f}$. For Algorithm 6.1 we set $c = 10^2$, $s = \frac{1.95}{L_f}$ and $\beta = \gamma = 1$, while for staBiM $\gamma = 1$, $c = 10^2$ and for Bi-SG-II, $s = \frac{1}{L_f}$, and $c = 10^2$. Along the iterations, we measure the inner and outer objective values and the norm of the difference of consecutive iterates and show the results in Figure 6.3. Since the optimal solution cannot be reached within a reasonable time, for the outer function we only show its function values, while for the inner we show the residual with the minimum value attained.

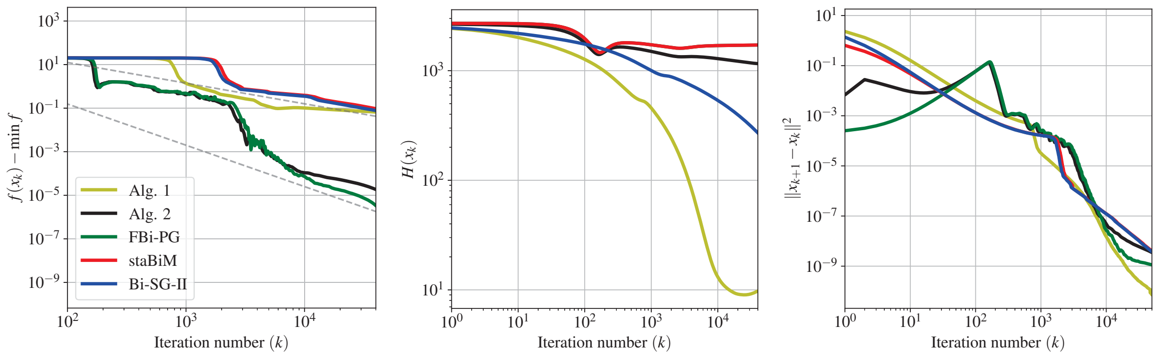
From Figure 6.3a, we observe that in this example, the convergence rate perfectly matches the one predicted by the theory, see Section 6.5. The fast schemes (i.e., Algorithm 6.2 and FBi-PG) exhibit a convergence rate of $o(k^{-\delta})$, outperforming the non-accelerated ones (i.e., Algorithm 6.1, staBiM and Bi-SG-II) that only exhibit a convergence rate of $o(k^{-\delta/2})$. Meanwhile, Figure 6.3b indicates that at least one of the considered methods may still be away from the optimal solution, and Figure 6.3c shows that the distance between consecutive iterates decreases similarly across all four algorithms.

6.7 Continuous-time analysis

Our convergence analysis of Algorithm 6.1 and Algorithm 6.2 is closely guided by the analysis of the continuous-time systems (6.5) and (6.6), even if $\hat{h} = \hat{f} = 0$. This analysis not only emerges as a foundational result with intrinsic theoretical interest, but also as an excellent assistant to derive convergence results for the corresponding numerical algorithms. In this section, we briefly illustrate and state the main results that can be obtained for (6.5) and (6.6), with $\varepsilon : [t_0, +\infty) \to \mathbb{R}_+$ defined by

$$\varepsilon(t) := \frac{c}{t^\delta}, \tag{6.84}$$

where $c, \delta > 0$. Observe first that Assumption 6.2 with $\eta \in \{1, 2\}$, originally introduced by Attouch and Czarnecki in [20] in the context of continuous-time systems such as (6.5), now reads as: For any solution



(a) Inner function residual. (b) Outer function. (c) Distance of consecutive iterates.

Figure 6.3: Results of experiment in Subsection 6.6.3. The dotted lines in Figure 6.3a represent the rates $\mathcal{O}(k^{-\delta/2})$ and $\mathcal{O}(k^{-\delta})$, with $\delta = 1.9$.

x^* to the bilevel problem (6.1) and all $p \in N_{\operatorname{argmin} f}(x^*)$ it holds

$$\int_{t_0}^{\infty} t^{\eta-1}\Big((f - \min f)^*(\varepsilon(t)p) - \sigma_{\operatorname{argmin} f}(\varepsilon(t)p)\Big)dt < +\infty. \tag{6.85}$$

Furthermore, all auxiliary results including Lemma 2.3.10, Lemma 2.3.4, Lemma 2.1.1, Lemma 6.2.2, and Lemma 6.2.3 admit continuous-time counterparts. Specifically, Lemma 2.3.10 can be stated with integrals instead of sums. The continuous-time counterpart of Lemma 2.3.4 can be found, e.g., in [40, Lemma A.4]. The continuous-time version of Opial's Lemma (Lemma 2.1.1) is well known, see, e.g., [40, Lemma A.2]. Lemma 6.2.2 can now be stated as follows: For a function $F : [t_0, +\infty) \to \mathbb{R}_+$, if $F(t) = \mathcal{O}(t^{-\delta})$ as $t \to +\infty$ and

$$F(t) \leq C_E t^{-\eta} + C_{0,F} t^{-\delta} F(t)^{\frac{1}{\rho}} \tag{6.86}$$

for every $t \geq t_0$, then $F(t) = \mathcal{O}(t^{-\eta})$ as $t \to +\infty$. This fact can be shown using the same argument of Lemma 6.2.2. Eventually, a unifying result similar to Lemma 6.2.3 can be established in continuous time by replacing differences of consecutive terms in a sequence by derivatives and sums by integrals. More precisely, we have:

Lemma 6.7.1. *Let $\hat{\delta} > 0$, $\eta \in \{1, 2\}$, and set $\delta := \eta\hat{\delta}$. Consider a function $E : [t_0, +\infty) \to \mathbb{R}$ defined as*

$$E(t) = t^{\eta}\Big(F(t) + \varepsilon(t)H(t)\Big) + V(t),$$

with $t \mapsto \varepsilon(t)$ as in (6.84) and $F(t) \geq 0$, $H(t) \geq -H_*$, where $H_* \geq 0$, and $V(t) \geq 0$ for every $t \geq t_0$. Suppose that, for every $t \geq t_0$,

$$\dot{E}(t) + \zeta(t)H(t) + g(t) \leq -C_2 t^{\eta-1} F(t), \tag{6.87}$$

where $g : [t_0, +\infty) \to \mathbb{R}_+$, $C_2 > 0$, and $\zeta : [t_0, +\infty) \to \mathbb{R}$ can be written as

$$\zeta(t) = C_\zeta t^{\eta-1}\varepsilon(t) \tag{6.88}$$

for some $C_\zeta > 0$. Consider the following conditions, which will only be assumed if we explicitly say so. For every $t \geq t_0$:

(C1) $\quad - H(t) \leq C_3 F(t)^{\frac{1}{\rho}}$, *for some constant $C_3 \geq 0$ and $1 < \rho \leq 2$. Set $\rho^\star$ as in (6.10),*

(C2) $\quad$ *For each $r > 0$, $-t^{\eta-1}\big(F(t) + r\varepsilon(t)H(t)\big) \leq I^r(t)$, where $I^r \geq 0$ and $I^r \in \mathbb{L}^1([t_0, +\infty))$.*

Then, following statements are true:

- *Case $\hat{\delta} > 1$: It holds $F(t) = \mathcal{O}(t^{-\eta})$ as $t \to +\infty$;*
- *Case $\hat{\delta} = 1$: If Assumption (C2) holds and $C_2 > 0$ in (6.87), then*
    1. *The limit $\lim_{t\to+\infty} E(t)$ exists;*

2. $F(t) = \mathcal{O}(t^{-\eta})$ *as* $t \to +\infty$;
3. *The functions* $t \mapsto t^{\eta-1}F(t)$, $t \mapsto t^{\eta-1}\varepsilon(t)H(t)$ *and* $t \mapsto g(t)$ *are integrable;*

- *Case* $\frac{1}{\rho^\star} < \hat{\delta} < 1$: *If Assumptions* $(C1)$ *and* $(C2)$ *hold and* $C_2 > 0$ *in* (6.87), *then*
  1. $F(t) = \mathcal{O}(t^{-\eta})$ *as* $t \to +\infty$;
  2. *It holds* $-C_H t^{-\delta+\frac{\eta}{\rho^\star}} \leq t^\eta \varepsilon(t)H(t)$ *for some* $C_H \geq 0$ *and all* $t \geq t_0$;
  3. *The limit* $\lim_{t\to+\infty} E(t)$ *exists;*
  4. *The functions* $\mapsto t^{\eta-1}F(t)$, $t \mapsto t^{\eta-1}\varepsilon(t)H(t)$ *and* $t \mapsto g(t)$ *are integrable.*

6.7.1 First-order system

Let us start with the first-order system (6.5) with starting point $x(t_0) = x_0 \in \mathcal{H}$. As anticipated in Remark 6.3.2, we consider the following Lyapunov energy, defined for $t \geq t_0$:

$$E^\lambda(t) := t(\Psi_t(x(t)) - \Psi_t(x^*)) + \frac{\lambda}{2}\|x(t) - x^*\|^2,$$

where $\Psi_t(x) := f(x) + \varepsilon(t)h(x)$ is the regularized function (6.3), x^* is a solution to the bilevel problem (6.1), and $\lambda > 1$. The analogous statement of Lemma 6.3.1 can be easily obtained by taking the time derivative of E_λ. We get for all $t \geq t_0$

$$\begin{aligned}
\dot{E}_\lambda(t) &= \Psi_t(x(t)) - \Psi_t(x^*) + t\Big(\dot{\varepsilon}(t)(h(x(t)) - h(x^*)) + \big\langle \dot{x}(t), \nabla f(x(t)) + \varepsilon(t)\nabla h(x(t))\big\rangle\Big) \\
&\qquad + \lambda\big\langle x(t) - x^*, \dot{x}(t)\big\rangle \\
&\leq \Psi_t(x(t)) - \Psi_t(x^*) - t\big\|\dot{x}(t)\big\|^2 + t\dot{\varepsilon}(t)(h(x(t)) - h(x^*)) - \lambda(\Psi_t(x(t)) - \Psi_t(x^*)) \\
&= -(\lambda - 1)(f(x(t)) - f(x^*)) - t\big\|\dot{x}(t)\big\|^2 + \big(t\dot{\varepsilon}(t) - (\lambda-1)\varepsilon(t)\big)(h(x(t)) - h(x^*)).
\end{aligned}$$

Rearranging the terms and dropping nonnegative ones, for all $t \geq t_0$ we arrive at

$$\dot{E}_\lambda(t) + \zeta(t)(h(x(t)) - h(x^*)) \leq -(\lambda - 1)(f(x(t)) - f(x^*)), \text{ where } \zeta(t) := (\lambda - 1)\varepsilon(t) - t\dot{\varepsilon}(t). \tag{6.89}$$

Applying Lemma 6.7.1 we obtain the following: for case i) $\delta > 1$, that $f(x(t)) - \min f = \mathcal{O}(t^{-1})$ as $t \to +\infty$; for both cases ii) $\delta = 1$ with (6.85) and $\eta = 1$, or iii) $\delta \in (\frac{1}{\rho^\star}, 1)$ with Assumption 6.1, that for all $\lambda \geq 1$,

1. The limit $\lim_{t\to+\infty} E^\lambda(t)$ exists;
2. $f(x(t)) - \min f = \mathcal{O}(t^{-1})$ as $t \to +\infty$;
3. In case iii), the following lower bound holds $t\varepsilon(t)(h(x(t) - h(x^*)) \geq -C_h t^{-(\delta - \frac{1}{\rho^\star})}$ for some $C_h \geq 0$ and all $t \geq t_0$;
4. The following integrability statements hold

$$\int_{t_0}^{\infty} (f(x(t)) - f(x^*))dt < +\infty, \quad \text{and} \quad \int_{t_0}^{\infty} t^{-\delta}|h(x(t)) - h(x^*)|dt < +\infty. \tag{6.90}$$

By taking $1 < \lambda_1 < \lambda_2$ and considering the difference $E_{\lambda_2}(t) - E_{\lambda_1}(t)$, we deduce that $\varphi(t) := \frac{1}{2}\|x(t) - x^*\|^2$ admits a limit as $t \to +\infty$ for all solutions x^* to the bilevel problem (6.1), and that

$$\lim_{t\to+\infty} q(t) := t(f(x(t)) - f(x^*)) + t\varepsilon(t)(h(x(t)) - h(x^*)) \quad \text{exists.} \tag{6.91}$$

From (6.91) and (6.90) we deduce, arguing as in (6.49), that $\lim_{t\to+\infty} q(t) = 0$. Proceeding as in (6.50) we deduce that $h(x(t)) - h(x^*)$ vanishes as $t \to +\infty$, and, therefore, that all weak sequential cluster points of $t \mapsto x(t)$ as $t \to +\infty$ lie in the solution set of the bilevel problem (6.1). Opial's Lemma in continuous time yields weak convergence of the full trajectory to some solution to the bilevel problem (6.1). To conclude, using a similar argument as in (6.51) and (6.52) yields the little-o rates. To summarize, we can show the following convergence result for trajectories generated by (6.5).

Theorem 6.7.2. *Let $x : [t_0, +\infty) \to \mathcal{H}$ be a solution trajectory to (6.5). Then, the following statements are true:*

- *Case $\delta > 1$: It holds $f(x(t)) - \min f = \mathcal{O}(t^{-1})$ as $t \to +\infty$;*

Suppose Assumption 6.3 holds.

- *Case $\delta = 1$: If f satisfies (6.85) with $\eta = 1$, as $t \to +\infty$ it holds*

$$f(x(t)) - \min f = o(t^{-1}) \quad \text{and} \quad h(x(t)) \to \min_{\operatorname{argmin} f} h.$$

Furthermore, $x(t)$ converges weakly towards a solution to the bilevel problem (6.1) as $t \to +\infty$;

- *Case $\frac{1}{\rho^\star} < \delta < 1$: If f satisfies Assumption 6.1 with exponent ρ, as $t \to +\infty$ it holds*

$$f(x(t)) - \min f = o(t^{-1}) \quad \text{and} \quad |h(x(t)) - \min_{\operatorname{argmin} f} h| = o(t^{-1+\delta}).$$

Furthermore, $x(t)$ converges weakly towards a solution to the bilevel problem (6.1) as $t \to +\infty$.

6.7.2 Second-order system

We now consider the second-order system (6.6) with starting position $x(t_0) = x_0 \in \mathcal{H}$ and starting velocity $\dot{x}(t_0) = v$. As anticipated in Remark 6.4.2, the continuous-time counterpart of the Lyapunov energy (6.53) is defined for $\alpha > 3$, $\lambda \in (2, \alpha - 1)$, and $t \geq t_0$ as

$$E_\lambda(t) := t^2(\Psi_t(x(t)) - \Psi_t(x^*)) + \frac{1}{2}\|\lambda(x(t) - x^*) + t\dot{x}(t)\|^2 + \frac{\lambda(\alpha - 1 - \lambda)}{2}\|x(t) - x^*\|^2, \tag{6.92}$$

where, once again, $\Psi_t(x) := f(x) + \varepsilon(t)h(x)$ and x^* is a solution to the bilevel problem (6.1). To derive the analogous result to Lemma 6.4.1, we denote by $v_\lambda(t) := \lambda(x(t) - x^*) + t\dot{x}(t)$ and take the time derivative of E^λ to get for every $t \geq t_0$

$$\begin{aligned}
\dot{E}_\lambda(t) &= 2t(\Psi_t(x(t)) - \Psi_t(x^*)) + t^2\Big(\dot{\varepsilon}(t)(h(x(t)) - h(x^*)) + \big\langle \dot{x}(t), \nabla f(x(t)) + \varepsilon(t)\nabla h(x(t))\big\rangle\Big) \\
&\quad + \big\langle v_\lambda(t), \dot{v}_\lambda(t)\big\rangle + \lambda(\alpha - 1 - \lambda)\big\langle x(t) - x^*, \dot{x}(t)\big\rangle \\
&= 2t(\Psi_t(x(t)) - \Psi_t(x^*)) + t^2\Big(\dot{\varepsilon}(t)(h(x(t)) - h(x^*)) + \big\langle \dot{x}(t), \nabla f(x(t)) + \varepsilon(t)\nabla h(x(t))\big\rangle\Big) \\
&\quad - \lambda(\alpha - 1 - \lambda)\big\langle x(t) - x^*, \dot{x}(t)\big\rangle - \lambda t\big\langle x(t) - x^*, \nabla f(x(t)) + \varepsilon(t)\nabla h(x(t))\big\rangle \\
&\quad - (\alpha - 1 - \lambda)t\big\|\dot{x}(t)\big\|^2 - t^2\big\langle \dot{x}(t), \nabla f(x(t)) + \varepsilon(t)\nabla h(x(t))\big\rangle + \lambda(\alpha - 1 - \lambda)\big\langle x(t) - x^*, \dot{x}(t)\big\rangle \\
&= 2t(\Psi_t(x(t)) - \Psi_t(x^*)) + t^2\dot{\varepsilon}(t)(h(x(t)) - h(x^*)) - (\alpha - 1 - \lambda)t\big\|\dot{x}(t)\big\|^2 \\
&\quad - \lambda t\big\langle x(t) - x^*, \nabla f(x(t)) + \varepsilon(t)\nabla h(x(t))\big\rangle \\
&\leq -t(\lambda - 2)(\Psi_t(x(t)) - \Psi_t(x^*)) + t^2\dot{\varepsilon}(t)(h(x(t)) - h(x^*)) + (\lambda + 1 - \alpha)t\big\|\dot{x}(t)\big\|^2.
\end{aligned}$$

Rearranging suitably and using the definition of Ψ_t, we obtain:

$$\dot{E}_\lambda(t) + \zeta(t)(h(x(t)) - h(x^*)) + (\alpha - 1 - \lambda)t\big\|\dot{x}(t)\big\|^2 \leq -t(\lambda - 2)(f(x(t)) - f(x^*)), \tag{6.93}$$

where $\zeta(t) := t(\lambda - 2)\varepsilon(t) - t^2\dot{\varepsilon}(t)$ for all $t \geq t_0$. We have obtained in (6.93) the continuous-time counterpart to the statement of Lemma 6.4.1. Using once again the continuous-time counterpart of Lemma 6.2.3 (i.e., Lemma 6.7.1) we obtain the following: for case i) $\delta > 2$, that $f(x(t)) - \min f = \mathcal{O}(t^{-2})$ as $t \to +\infty$; for both cases ii) $\delta = 2$ with (6.85) and $\eta = 2$, and iii) $\delta \in (\frac{2}{\rho^\star}, 2)$, with Assumption 6.1, that for all $\lambda \in (2, \alpha - 1)$,

1. The limit $\lim_{t \to +\infty} E^\lambda(t)$ exists;
2. $f(x(t)) - \min f = \mathcal{O}(t^{-2})$ as $t \to +\infty$;

3. In case iii), the following lower bound holds $t^2\varepsilon(t)(h(x(t) - h(x^*)) \geq -C_h t^{-(\delta - \frac{2}{\rho^\star})}$ for some $C_h \geq 0$ and all $t \geq t_0$;

4. The following integrability statements hold

$$\int_{t_0}^{\infty} t(f(x(t)) - f(x^*))dt < +\infty, \quad \text{and} \quad \int_{t_0}^{\infty} t^{1-\delta}|h(x(t)) - h(x^*)|dt < +\infty. \tag{6.94}$$

Arguing similarly as in the proof of Theorem 6.4.3 we obtain the following convergence result.

Theorem 6.7.3. *Let $x : [t_0, +\infty) \to \mathcal{H}$ be a solution trajectory to (6.6). Then, the following statements are true:*

- *Case $\delta > 2$: It holds $f(x(t)) - \min f = \mathcal{O}(t^{-2})$ as $t \to +\infty$;*

Suppose Assumption 6.3 holds.

- *Case $\delta = 2$: If f satisfies (6.85) with $\eta = 2$, as $t \to +\infty$ it holds*

$$f(x(t)) - \min f = o(t^{-2}) \quad \text{and} \quad h(x(t)) \to \min\nolimits_{\operatorname{argmin} f} h.$$

Furthermore, $x(t)$ converges weakly towards a solution to the bilevel problem (6.1) as $t \to +\infty$.

- *Case $\frac{2}{\rho^\star} < \delta < 2$: If f satisfies Assumption 6.1 with exponent ρ, as $t \to +\infty$ it holds*

$$f(x(t)) - \min f = o(t^{-2}) \quad \text{and} \quad |h(x(t)) - \min\nolimits_{\operatorname{argmin} f} h| = o(t^{-2+\delta}).$$

Furthermore, $x(t)$ converges weakly towards a solution to the bilevel problem (6.1) as $t \to +\infty$.

Chapter 7

Conclusions

In this thesis, we explored continuous- and discrete-time schemes attached to convex optimization and monotone inclusion problems. A fundamental role was played by the coefficient $\frac{\alpha}{t}$ damping the velocity in our second-order systems, which had the effect of producing dynamics with fast convergence properties, both in the convex optimization and monotone operator settings. Thus, we have contributed to the already considerable existing literature on second-order systems with asymptotic vanishing damping. As it was the case in our work, the analyses for the continuous-time systems, which are in and of themselves interesting, can also serve as guides for the analyses of the numerical algorithms which result from their discretizations.

In Chapter 3, we explored the equivalence which exists between the Heavy Ball dynamics and the Nesterov accelerated gradient dynamics, both in the optimization setting, and also in the monotone operator setting. By analyzing a Heavy Ball system with a time rescaling parameter $b(\cdot)$, we showed that for a particular choice for $b(\cdot)$, the Heavy Ball and the Nesterov systems are time reparametrizations of each other, and every statement about the Nesterov system can be recovered from a corresponding statement for the Heavy Ball dynamics. This was done in the optimization setting, i.e., where the governing operator is the gradient of a convex and smooth function, and we also produced analog systems and statements suited for the monotone operator setting.

In Chapter 4, we studied the asymptotic properties of a second-order dynamical system with asymptotically vanishing damping attached to a monotone+cocoercive inclusion problem. We derived fast convergence of the velocities and accelerations to zero, as well as a $o(1/t^2)$ convergence rate for the norm of the governing operator along the trajectories generated by the system. We also showed weak convergence of the trajectories towards solutions to the monotone inclusion problem.

In Chapter 5, we studied the asymptotic properties of a second-order primal-dual dynamical system attached to a linearly constrained convex optimization problem. The system features an asymptotically vanishing damping coefficient and a time rescaling parameter which has the effect of accelerating the convergence rates for the functional values along the trajectories and the feasibility gap, which measures how far the trajectory is to the feasible set of the problem. Indeed, for these quantities we obtained rates $\mathcal{O}(1/t^2 b(t))$. Additionally, we showed weak convergence of the trajectories towards primal-dual solutions to the problem.

In Chapter 6, we studied two numerical algorithms designed to solve a hierarchical optimization problem, that is, where we minimize an outer function subject to the set of minimizers of an inner function. Under a geometric assumption for the inner function, for the first algorithm we showed $o(1/t)$ and $o(1/t^{1-\delta})$ rates for the inner and outer functional values along the iterates respectively; and for the second algorithm we showed $o(1/t^2)$ and $o(1/t^{2-\delta})$ rates for the inner and outer functional values along the iterates respectively. We also showed weak convergence of the iterates generated by both algorithms to solutions to the problem. The little-o rates, together with the weak convergence of trajectories in the framework of a (possibly) infinite-dimensional Hilbert space marked a first in the literature for this class of problems.

One of the principal open questions we have, which is a future project, concerns our work presented in chapter 6. Namely, we are curious to see if our methods can be adapted when the inner level is a structured monotone inclusion; particularly, we are interested to see if we can formulate a numerical algorithm for such a problem. We suspect that, just as we had to assume a Hölderian error bound assumption for the inner function, we will have to assume an analogous sharpness condition for the inner operator.

Bibliography

[1] B. Abbas, H. Attouch, and B. F. Svaiter. “Newton-like dynamics and forward-backward methods for structured monotone inclusions in Hilbert spaces”. In: *Journal of Optimization Theory and Applications* 161.2 (2014), pp. 331–360.

[2] F. Álvarez. “On the minimizing property of a second-order dissipative system in Hilbert spaces”. In: *SIAM Journal on Control and Optimization* 38.4 (2000), pp. 1102–1119.

[3] F. Álvarez and H. Attouch. “An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping”. In: *Set-Valued Analysis* 9.1–2 (2001), pp. 3–11.

[4] F. Álvarez, H. Attouch, and P. Redont. “A second-order gradient-like dissipative dynamical system with Hessian-driven damping. Application to optimization and mechanics”. In: *Journal de Mathématiques Pures et Apliquées* 81.8 (2002), pp. 747–779.

[5] A. S. Antipin. “On a method for convex programs using a symmetrical modification of the Lagrange function”. In: *Ekonomika i Matematicheskie Metody* 12 (1976), pp. 1164–1173.

[6] G. Anzellotti and S. Baldo. “Asymptotic development by Γ-convergence”. In: *Applied Mathematics and Optimization* 27.2 (1993), pp. 105–123.

[7] H. Attouch. “Viscosity solutions of minimization problems”. In: *SIAM Journal on Optimization* 6.3 (1996), pp. 769–806.

[8] H. Attouch, R. I. Boț, D. A. Hulett, and D.-K. Nguyen. “Recovering Nesterov accelerated dynamics from Heavy Ball dynamics via time rescaling”. In: (2025). arXiv: 2504.15852.

[9] H. Attouch, R. I. Boț, and D.-K. Nguyen. “Fast convex optimization via time scale and averaging of the steepest descent”. In: *Mathematics of Operations Research* (2024).

[10] H. Attouch and A. Cabot. “Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity”. In: *Journal of Differential Equations* 263.9 (2017), pp. 5412–5458.

[11] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont. “Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity”. In: *Mathematical Programming* 168.1 (2018), pp. 123–175.

[12] H. Attouch, Z. Chbani, and H. Riahi. “Combining fast inertial dynamics for convex optimization with Tikhonov regularization”. In: *Journal of Mathematical Analysis and Applications* 457.2 (2018), pp. 1065–1094.

[13] H. Attouch, Z. Chbani, and H. Riahi. “Fast convex optimization via a third-order in time evolution equation”. In: *Optimization* 71.5 (2022), pp. 1275–1304.

[14] H. Attouch, Z. Chbani, and H. Riahi. “Fast convex optimization via time scaling of damped inertial gradient dynamics”. In: *Pure and Applied Functional Analysis* 6.6 (2021), pp. 1081–1117.

[15] H. Attouch, Z. Chbani, and H. Riahi. “Fast proximal methods via time scaling of damped inertial dynamics”. In: *SIAM Journal on Optimization* 29.3 (2019), pp. 2227–2256.

[16] H. Attouch, Z. Chbani, and H. Riahi. “Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$”. In: *ESAIM: Control, Optimization and Calculus of Variations* 25.2 (2019).

[17] H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. “Convergence of iterates for first-order optimization algorithms with inertia and Hessian driven damping”. In: *Optimization* 72.5 (2023), pp. 1199–1238.

[18] H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. "Fast convergence of dynamical ADMM via time scaling of damped inertial dynamics". In: *Journal of Optimization Theory and Applications* 193.1 (2022), pp. 704–736.

[19] H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. "First-order optimization algorithms via inertial systems with Hessian driven damping". In: *Mathematical Programming* 193.1 (2022), pp. 113–155.

[20] H. Attouch and M.-O. Czarnecki. "Asymptotic behavior of coupled dynamical systems with multiscale aspects". In: *Journal of Differential Equations* 248.6 (2010), pp. 1315–1344.

[21] H. Attouch, X. Goudou, and P. Redont. "The heavy ball with friction method. The continuous dynamical system, global exploration of the local minima of a real-valued function by asymptotical analysis of a dissipative dynamical system". In: *Communications in Contemporary Mathematics* 2.1 (2000), pp. 1–34.

[22] H. Attouch and S. C. László. "Continuous Newton-like inertial dynamics for monotone inclusions". In: *Set-Valued and Variational Analysis* 29 (2021), pp. 555–581.

[23] H. Attouch and S. C. László. "Newton-like inertial dynamics and proximal algorithms governed by maximally monotone operators". In: *SIAM Journal on Optimization* 30.4 (2020), pp. 3252–3283.

[24] H. Attouch and P. E. Maingé. "Asymptotic behavior of second order dissipative evolution equations combining potential with non-potential effects". In: *ESAIM Control Optimization and Calculus of Variations* 17.3 (2011), pp. 836–857.

[25] H. Attouch and J. Peypouquet. "Convergence of inertial dynamics and proximal algorithms governed by maximal monotone operators". In: *Mathematical Programming* 174.1–2 (2019), pp. 391–432.

[26] H. Attouch, J. Peypouquet, and P. Redont. "Fast convex minimization via inertial dynamics with Hessian driven damping". In: *Journal of Differential Equations* 210.10 (2016), pp. 5734–5783.

[27] H. Attouch and B. F. Svaiter. "A continuous dynamical Newton-like approach to solving monotone inclusions". In: *SIAM Journal on Control and Optimization* 49.2 (2011), pp. 574–598.

[28] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory*. CMS Books in Mathematics. Springer, 2017.

[29] A. Beck and M. Teboulle. "A Fast Iterative Shrinkage-Thresholding algorithm for linear inverse problems". In: *SIAM Journal on Imaging Sciences* 2.1 (2009), pp. 183–202.

[30] A. Böhm, M. Drauch, E. R. Csetnek, and R. I. Boț. "Two steps at a time–taking GAN training in stride with Tseng's method". In: *SIAM Journal on Mathematics of Data Science* 4.2 (2022), pp. 750–771.

[31] Z. Borsos, M. Mutny, and A. Krause. "Coresets via bilevel optimization for continual learning and streaming". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 14879–14890.

[32] R. I. Boț. *Conjugate Duality in Convex Optimization*. Lecture Notes in Economics and Mathematical Systems. Springer, 2010.

[33] R. I. Boț and E. Chenchene. "Extra-Gradient method with flexible anchoring: strong convergence and fast residual decay". In: (2024). arXiv: 2410.14369.

[34] R. I. Boț, E. Chenchene, R. E. Csetnek, and D. A. Hulett. "Accelerating diagonal methods for bilevel optimization: unified convergence via continuous-time dynamics". In: (2025). arXiv: 2505.14389.

[35] R. I. Boț, E. Chenchene, and J. M. Fadili. "Generalized Fast Krasnoselskii–Mann method with preconditioners". In: (2024). arXiv: 2411.18574.

[36] R. I. Boț and E. R. Csetnek. "A dynamical system associated with the fixed points set of a nonexpansive operator". In: *Journal of Dynamics and Differential Equations* 29 (2017), pp. 155–168.

[37] R. I. Boț and E. R. Csetnek. "Second order forward-backward dynamical systems for monotone inclusion problems". In: *SIAM Journal on Control and Optimization* 54.3 (2016), pp. 1423–1443.

[38] R. I. Boț, E. R. Csetnek, and S. C. László. "A primal-dual dynamical approach to structured convex minimization problems". In: *Journal of Differential Equations* 269.12 (2020), pp. 10717–10757.

[39] R. I. Boț, E. R. Csetnek, and D.-K. Nguyen. "Fast Augmented Lagrangian Method in the convex regime with convergence guarantees for the iterates". In: *Mathematical Programming* 200 (2023), pp. 147–197.

[40] R. I. Boț, E. R. Csetnek, and D.-K. Nguyen. "Fast Optimistic Gradient Descent Ascent (OGDA) Method in Continuous and Discrete Time". In: *Foundations of Computational Mathematics* 25.1 (2025), pp. 163–222.

[41] R. I. Boț and D. A. Hulett. "Second order splitting dynamics with vanishing damping for additively structured monotone inclusions". In: *Journal of Dynamics and Differential Equations* 36 (2024), pp. 727–756.

[42] R. I. Boț, D. A. Hulett, and D.-K. Nguyen. "Fast second-order dynamics with slow vanishing damping approaching the zeros of a monotone and continuous operator". In: *Journal of Convex Analysis* (2025).

[43] R. I. Boț and D.-K. Nguyen. "Improved convergence rates and trajectory convergence for primal-dual dynamical systems with vanishing damping". In: *Journal of Differential Equations* 303 (2021), pp. 369–406.

[44] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers". In: *Foundations and Trends® in Machine Learning* 3.1 (2010), pp. 1–122.

[45] A. Cabot. "Proximal point algorithm controlled by a slowly vanishing term: applications to hierarchical minimization". In: *SIAM Journal on Optimization* 15.2 (2005), pp. 555–572.

[46] J. Cao, R. Jiang, E. Y. Hamedani, and A. Mokhtari. "An accelerated gradient method for convex smooth simple bilevel optimization". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024.

[47] E. Chenchene, A. Hosseini, and K. Bredies. "A hybrid proximal generalized conditional gradient method and application to total variation parameter learning". In: *2023 European Control Conference (ECC)*. 2023, pp. 1–6.

[48] R. Cominetti, J. Peypouquet, and S. Sorin. "Strong asymptotic convergence of evolution equations governed by maximal monotone operators with Tikhonov regularization". In: *Journal of Differential Equations* 245 (2008), pp. 3753–3763.

[49] R. Cominetti and J. San Martín. "Asymptotic analysis of the exponential penalty trajectory in linear programming". In: *Mathematical Programming* 67.1 (1994), pp. 169–187.

[50] M. G. Crandall and P.-L. Lions. "Viscosity solutions of Hamilton–Jacobi equations". In: *Transactions of the American Mathematical Society* 277.1 (1983), pp. 1–42.

[51] R. E. Csetnek, Y. Malitsky, and M. K. Tam. "Shadow Douglas-Rachford splitting for monotone inclusions". In: *Applied Mathematics & Optimization* 80 (2019), pp. 665–6789.

[52] M. Cuturi. "Sinkhorn distances: lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc., 2013.

[53] S. Dempe and A. Zemkoho. *Springer Optimization and its Applications*. Vol. 161. Springer, 2020.

[54] L. Doron and S. Shtern. "Methodology and first-order algorithms for solving nonsmooth and non-strongly convex bilevel optimization problems". In: *Mathematical Programming* 201.1 (2023), pp. 521–558.

[55] I. Ekeland and R. Temam. *Analyse Convexe et Problèmes Variationnels*. Etudes Mathématiques. Dunod, 1974.

[56] M. P. Friedlander and P. Tseng. "Exact regularization of convex programs". In: *SIAM Journal on Optimization* 18.4 (2008), pp. 1326–1350.

[57] D. Gabay and B. Mercier. "A dual algorithm for the solution of nonlinear variational problems via finite element approximation". In: *Computers & Mathematics with Applications* 2.1 (1976), pp. 17–40.

[58] T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk. "Fast alternating direction optimization methods". In: *SIAM Journal on Imaging Sciences* 17.3 (2014), pp. 1588–1623.

[59] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial networks". In: *NeurIPS 2014: Advances in Neural Information Processing Systems*. Vol. 27. 2014, pp. 2672–2680.

[60] E. Gorbunov, N. Loizou, and G. Gidel. “Extragradient method: $\mathcal{O}(1/k)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity”. In: *AISTATS 2022: The 25th International Conference on Artificial Intelligence and Statistics*. 2022.

[61] O. Güler. “New proximal point algorithms for convex minimization”. In: *SIAM Journal on Optimization* 2.4 (1992), pp. 649–664.

[62] A. Haraux. *Systémes Dynamiques Dissipatifs et Applications*. Masson, 1991.

[63] X. He, R. Hu, and Y. P. Fang. “Convergence rates of inertial primal-dual dynamical methods for separable convex optimization problems”. In: *SIAM Journal on Control and Optimization* 59.5 (2021).

[64] X. He, R. Hu, and Y. P. Fang. “Fast primal–dual algorithm via dynamical system for a linearly constrained convex optimization problem”. In: *Automatica* 146 (2022), p. 110547.

[65] X. He, R. Hu, and Y. P. Fang. “Inertial accelerated primal-dual methods for linear equality constrained convex optimization problems”. In: *Numerical Algorithms* 90.4 (2022), pp. 1669–1690.

[66] X. He, R. Hu, and Y. P. Fang. “Inertial primal-dual dynamics with damping and scaling for linearly constrained convex optimization problems”. In: *Applicable Analysis* 102.15 (2022), pp. 4114–4139.

[67] D. A. Hulett and D.-K. Nguyen. “Time rescaling of a primal-dual dynamical dystem with asymptotically vanishing damping”. In: *Apllied Mathematics & Optimization* 88.27 (2023).

[68] H. D. Kaushik and F. Yousefian. “A method with convergence rates for optimization problems with variational inequality constraints”. In: *SIAM Journal on Optimization* 31.3 (2021), pp. 2171–2198.

[69] G. M. Korpelevich. “An extragradient method for finding saddle points and for other problems”. In: *Ekonomika i Matematicheskie Metody* 12.4 (1976), pp. 747–756.

[70] K. Kunisch and T. Pock. “A bilevel optimization approach for parameter learning in variational models”. In: *SIAM Journal on Imaging Sciences* 6.2 (2013), pp. 938–983.

[71] P. Latafat, A. Themelis, S. Villa, and P. Patrinos. “On the convergence of proximal gradient methods for convex simple bilevel optimization”. In: *Journal of Optimization Theory and Applications* 204.51 (2025).

[72] F. Lieder. “On the convergence rate of the Halpern-iteration”. In: *Optimization Letters* 15.2 (2021), pp. 405–418.

[73] Z. Lin, H. Li, and C. Fang. *Accelerated optimization for machine learning: first-order algorithms*. Springer, 2020.

[74] R. Madan and S. Lall. “Distributed algorithms for maximum lifetime routing in wireless sensor networks”. In: *IEEE Transactions on Wireless Communications* 5.8 (2006), pp. 2185–2193.

[75] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. “Towards deep learning models resistant to adversarial attacks”. In: *ICLR 2018: International Conference on Learning Representations*. 2018.

[76] R. May. “Asymptotic for a second order evolution equation with convex potential and vanishing damping term”. In: *Turkish Journal of Mathematics* 41.3 (2017), pp. 681–685.

[77] R. Merchav and S. Sabach. “Convex bi-level optimization problems with nonsmooth outer objective function”. In: *SIAM Journal on Optimization* 33.4 (2023), pp. 3114–3142.

[78] R. Merchav, S. Sabach, and M. Teboulle. “A fast algorithm for convex composite bi-level optimization”. In: (2024). arXiv: 2407.21221.

[79] A. S. Nemirovskij and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983.

[80] Y. Nesterov. “A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$”. In: *Doklady Akademii Nauk SSSR* 269.3 (1983), pp. 543–547.

[81] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian. “Deep decentralized multi-task multi-agent reinforcement learning under partial observability”. In: *The 34th International Conference on Machine Learning*. Vol. 70. 2017, pp. 2681–2690.

[82] Z. Opial. “Weak convergence of the sequence of successive approximations for nonexpansive mappings”. In: *Bulletin of the American Mathematical Society* 73.4 (1967), pp. 591–597.

[83] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. "Scikit-learn: machine learning in Python". In: *Journal of machine learning research* 12.Oct (2011), pp. 2825–2830.

[84] J. Peypouquet. "Coupling the gradient method with a general exterior penalization scheme for convex minimization". In: *Journal of Optimization Theory and Applications* 153 (2012), pp. 123–138.

[85] B. T. Polyak. *Introduction to optimization*. Translations series in mathematics and engineering. Optimization Software, Publications Division, 1987.

[86] B. T. Polyak. "Some methods of speeding up the convergence of iteration methods". In: *USSR Computational Mathematics and Mathematical Physics* 4.5 (1964), pp. 1–17.

[87] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. "Meta-learning with implicit gradients". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[88] R. T. Rockafellar. "Augmented lagrangians and applications of the proximal point algorithm in convex programming". In: *Mathematics of Operations Research* 1.2 (1976), pp. 97–116.

[89] S. Sabach and S. Shtern. "A first order method for solving convex bilevel optimization problems". In: *SIAM Journal on Optimization* 27.2 (2017), pp. 640–660.

[90] G. R. Sell and Y. You. *Dynamics of Evolutionary Equations*. Springer, 2002.

[91] G. Shi and K. H. Johansson. "Randomized optimal consensus of multi-agent systems". In: *Automatica* 48.12 (2012), pp. 3018–3030.

[92] M. Solodov. "An explicit descent method for bilevel convex optimization". In: *Journal of Convex Analysis* 14.2 (2007), p. 227.

[93] M. V. Solodov. "A bundle method for a class of bilevel nonsmooth convex minimization problems". In: *SIAM Journal on Optimization* 18.1 (2007), pp. 242–259.

[94] W. J. Su, S. Boyd, and E. J. Candès. "A differential equation for modeling Nesterov's accelerated gradient method: theory and insights". In: *Journal of Machine Learning Research* 17.153 (2016), pp. 1–43.

[95] T. Sun, P. Yin, D. Li, C. Huang, L. Guan, and H. Jiang. "Non-ergodic convergence analysis of heavy-ball algorithms". In: AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019.

[96] A. N. Tikhonov. "On the solution of incorrectly formulated problems and the regularization method". In: *Doklady Akademii Nauk SSSR* 151 (1963), pp. 501–504.

[97] C. Traoré and E. Pauwels. "Sequential convergence of AdaGrad algorithm for smooth convex optimization". In: *Operations Research Letters* 49.4 (2021), pp. 452–458.

[98] W. Wolberg, O. Mangasarian, N. Street, and W. Street. *Breast Cancer Wisconsin (Diagnostic)*. UCI Machine Learning Repository. 1993.

[99] P. Yi, Y. Hong, and F. Liu. "Distributed gradient algorithm for constrained optimization with application to load sharing in power systems". In: *Systems & Control Letters* 83 (2015), pp. 45–52.

[100] P. Yi, Y. Hong, and F. Liu. "Initialization-free distributed algorithms for optimal resource allocation with feasibility constraints and application to economic dispatch of power systems". In: *Automatica* 74 (2016), pp. 259–269.

[101] T. Yoon and E. K. Ryu. "Accelerated algorithms for smooth convex-concave minimax problems with $O(1/k^2)$ rate on squared gradient norm". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 12098–12109.

[102] X. Zeng, J. Lei, and J. Chen. "Dynamical primal-dual Nesterov accelerated method and its application to network optimization". In: *IEEE Transactions on Automatic Control* 68.3 (2023), pp. 1760–1767.

[103] X. Zeng, P. Yi, Y. Hong, and L. Xie. "Distributed continuous-time algorithms for nonsmooth extended monotropic optimization problems". In: *SIAM Journal on Control and Optimization* 56.6 (2018), pp. 3973–3993.

[104] X. Zhou. "On the Fenchel duality between strong convexity and Lipschitz continuous gradient". In: (2018). arXiv: 1803.06573.