Numerical splitting methods for nonsmooth convex optimization problems

von der Fakultät für Mathematik
der Technischen Universität Chemnitz
genehmigte

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt von

Sandy Bitterlich, M. Sc.
geboren am 16.02.1991 in Chemnitz

Tag der Einreichung: 29.11.2022

Betreuer: Prof. Dr. Dr. h.c. (NUM) Gert Wanka, Technische Universität Chemnitz

Gutachter: Prof. Dr. Dr. h.c. (NUM) Gert Wanka, Technische Universität Chemnitz
Prof. Dr. László Szilárd Csaba, Technical University of Cluj-Napoca
Assoc. Prof. Dr. Mathias Staudigl, Maastricht University

Tag der Verteidigung: 10.5.2023

Veröffentlichungsjahr: 2023

https://nbn-resolving.org/urn:nbn:de:bsz:ch1-qucosa2-883961

Bibliographical description

Sandy Bitterlich

Numerical splitting methods for nonsmooth convex optimization problems

Dissertation, **126** pages, Chemnitz University of Technology, Faculty of Mathematics, 2022

Report

In this thesis, we develop and investigate numerical methods for solving nonsmooth convex optimization problems in real Hilbert spaces. We construct algorithms, such that they handle the terms in the objective function and constraints of the minimization problems separately, which makes these methods simpler to compute. In the first part of the thesis, we extend the well known AMA method from Tseng to the Proximal AMA algorithm by introducing variable metrics in the subproblems of the primal-dual algorithm. For a special choice of metrics, the subproblems become proximal steps. Thus, for objectives in a lot of important applications, such as signal and image processing, machine learning or statistics, the iteration process consists of expressions in closed form that are easy to calculate. In the further course of the thesis, we intensify the investigation on this algorithm by considering and studying a dynamical system. Through explicit time discretization of this system, we obtain Proximal AMA. We show the existence and uniqueness of strong global solutions of the dynamical system and prove that its trajectories converge to the primal-dual solution of the considered optimization problem. In the last part of this thesis, we minimize a sum of finitely many nonsmooth convex functions (each can be composed by a linear operator) over a nonempty, closed and convex set by smoothing these functions. We consider a stochastic algorithm in which we take gradient steps of the smoothed functions (which are proximal steps if we smooth by Moreau envelope), and use a mirror map to "mirror" the iterates onto the feasible set. In applications, we compare them to similar methods and discuss the advantages and practical usability of these new algorithms.

Keywords

conjugate functions; convex optimization; structured convex minimization; dynamical system; Lyapunov analysis; Proximal AMA; primal dual algorithm; Lagrangian; saddle points; (convex) subdifferential; duality; Nesterov smoothing; Moreau envelope; mirror descent; proximal algorithm; Support Vector Machines

Acknowledgments

I want to express my thanks to the people who have greatly contributed to the realization of this dissertation.

First of all, I express my deepest gratitude to my supervisor, Professor Gert Wanka, for his outstanding supervision and support, both from a mathematical and a directional point of view.

I would like to direct special thanks to Professor Radu Ioan Boţ for his comprehensive support of my research. Especially, my research stays in his group at the Faculty of Mathematics in Vienna were important and valuable for the development of this dissertation.

Likewise, I would like to thank Professor Sorin-Mihai Grad and Dr. Ernö Robert Csetnek for their help, advice, and discussions on the research topic.

Furthermore, I would like to thank my current colleagues from the MRZ, Dr. Roman Unger, Dr. Elke Luschnat and Margit Matt, for their support and encouragement during the time of work on completing the dissertation thesis. I also address my thanks to Dr. Franziska Nestler and Jens Pönisch for valuable discussions.

I would like to thank Professor Mathias Staudigl and Professor László Szilárd Csaba for reviewing my thesis and for their valuable and helpful remarks and suggestions.

Many thanks to my family and friends for supporting me in many ways, encouraging me for my academic work and enriching my life.

Last, but not least, my sincere thanks to my fiancé Clemens for love, patience, and supporting me in countless ways, both in academic and personal matters.

I would also like to thank the DFG (Deutsche Forschungsgemeinschaft), who financially supported my research as the project WA922/9-1.

Contents

Chapter 1

Introduction

Nonsmooth convex optimization problems appear in a variety of different applications, such as in signal and image processing, machine learning or statistics. A lot of problems are structured in such a way that they contain smooth and nonsmooth terms, which are possibly composed by linear and continuous operators. Such problems arise, for example, in the following image deblurring and denoising problem:

$$\inf_{x\in\mathbb{R}^n} \left\{ f(Ax) + g(Lx) \right\},$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is defined as $f(x) = \frac{1}{2}\|x-b\|^2$, $g : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is given by $g(y,z) = \lambda\|(y,z)\|_1$, with a specific linear operator L, and $\lambda > 0$ is the regularization parameter. The function $g \circ L$ represents the anisotropic total variation functional, where $|\cdot|_1$ denotes the ℓ^1-norm. Furthermore, $A \in \mathbb{R}^{n\times n}$ is a blur operator and $b \in \mathbb{R}^n$ is the given blurred and noisy image.

To solve such problems, proximal methods of full splitting type have gained particular importance in the last years, which treat the nonsmooth terms of the objective function via proximal operators, the smooth terms via gradients and handle the linear and continuous operators separately. In a lot of applications, proximal operators have a simple and closed form such that in these cases, these algorithms are easy to implement and show good numerical performance. For an overview of more recent development of gradient-based optimization methods, proximal gradient methods and their acceleration, and proximal versions of primal-dual schemes, see [89], [21] and [60].

In this thesis we develop and extend algorithms that can be formulated as proximal splitting methods.

After giving some basic notations, definitions, and results of convex analysis in Chapter 2, we propose in **Chapter 3**, which is based on paper [28], a proximal version of the Alternating Minimization Algorithm of Tseng (see [110]) for solving convex optimization problems with two-block separable linear constraints and objectives, whereby one of the components of the latter is assumed to be strongly convex. For example, the Fenchel dual problem of the problem above can be written as such a structured optimization problem:

$$\inf_{p\in\mathbb{R}^n, q\in\mathbb{R}^n\times\mathbb{R}^n} \left\{ f^*(p) + g^*(q) \right\}, \text{ s.t. } A^*p + L^*q = 0,$$

where f^* and g^* are the convex conjugate of f and g, respectively. Due to the differentiability of f, we have that f^* is strongly convex and since f and g have full domains, strong duality holds. In the algorithm of Tseng, which is a primal-dual algorithm, the subproblems to be solved within an iteration do not usually correspond to the calculation of a proximal operator through a closed formula. This can make it computationally expensive and hard to implement, as it would be the case for the optimization problem above. In our algorithm, called Proximal AMA, we consider the same optimization problem, but we allow in each block of the objective a further smooth convex function. In the iteration process, we add variable metrics to the subproblems, as it was done in [19], where the authors created a proximal version of ADMM. If we choose these metrics in a suitable way, the iterative scheme can be reduced to the computation of proximal operators. In many applications (as in the optimization problem above), this means that each iteration can be carried out without solving an optimization subproblem in each step. This increases the attractiveness for implementation. We investigate the convergence of this algorithm in a real Hilbert space setting, and we illustrate its numerical performances on two applications: the first one in an image deblurring and denoising problem, introduced above, and the second one in machine learning.

In **Chapter 4**, based on paper [29], we propose a dynamical system related to the same optimization problem mentioned above. By time discretization, its trajectories can be seen as a continuous version of the Proximal AMA algorithm and the AMA numerical method. We show the existence and uniqueness of strong global solutions of the dynamical system, and prove that its trajectories asymptotically converge to a saddle point of the Lagrangian of the convex optimization problem. Our methods are based on [38], where a dynamical system related to the Proximal ADMM algorithm was studied.

In **Chapter 5**, based on paper [30] and the preprint [31], we develop incremental stochastic mirror descent algorithms to minimize a sum of finitely many nonsmooth convex functions over a nonempty, closed and convex set in the Euclidean space. Thus, we investigate the incremental mirror descent subgradient algorithm with random sweeping and proximal step, which can be found in the work of Boţ and Böhm in [35] and is based on the mirror descents algorithms of Beck and Teboulle (see [23]). Instead of evaluating the functions over their subgradients, which requires Lipschitz continuity of the functions, as in the work of Beck and Teboulle, we approximate the component functions using Nesterov's smoothing technique and use the gradients of the smoothed functions. For this, we require the weaker condition of closedness of the domains of their conjugates. Since the number of summands of the objective function can be very large, the gradient of a single component smoothed function is evaluated in each iteration step and is mirrored back to the feasible domain. This makes the computation of the iterations very cheap. Boţ and Böhm proposed this approach in their algorithms too, but they used subgradients of the nonsmooth component functions. The Moreau envelope is a special case of Nesterov's smoothing technique. As a result, our algorithms can also be formulated with proximal steps and can be seen as an extension of a proximal algorithm. We prove convergence order of $O(1/\sqrt{k})$ in expectation for the kth best objective function value and compare the numerical performance to the algorithms in [35] in three applications, the first in logistics, the second in medical imaging and the third in machine learning.

In the final chapter of this thesis, we provide conclusions and discuss potential future research directions. Additionally, we include an appendix that offers a brief overview and explanation of the SVM classification models, which were considered in applications in Chapter 3 and Chapter 5.

Chapter 2

Preliminaries

In this chapter, we give some basic notation and definitions from convex analysis and provide important results and properties which are used throughout this thesis. Our main source for these notions is [20], which we refer to for further information.

In the following, let $\mathcal{H}$ be a real Hilbert space with the corresponding inner product $\langle \cdot, \cdot \rangle$ and the associated norm $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$. The *identity operator* is defined as $\operatorname{Id} : \mathcal{H} \to \mathcal{H},\ \operatorname{Id}(x) = x$ for all $x \in \mathcal{H}$. Further, we denote $\mathbb{R}$ as the set of real numbers and define $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ as the extended real line. We write $\mathbb{N} := \{1, 2, \dots\}$ for the set of natural numbers.

A subset $C \subseteq \mathcal{H}$ is *convex* , if for all $\lambda \in [0, 1]$ and all $x, y \in C$ it holds that $\lambda x + (1 - \lambda) y \in C$. Let $C \subseteq \mathcal{H}$ be a convex and closed set. The strong quasi-relative interior of C is given by

$$\operatorname{sqri}(C) = \left\{ x \in C : \bigcup_{\lambda > 0} \lambda(C - x) \text{ is a closed linear subspace of } \mathcal{H} \right\}.$$

We have the inclusion $\operatorname{int}(C) \subseteq \operatorname{sqri}(C)$, meaning that the interior of a set C is always contained in its strong quasi-relative interior. If $\mathcal{H}$ is finite-dimensional, then $\operatorname{sqri}(C) = \operatorname{ri}(C)$, where $\operatorname{ri}(C)$ denotes the *relative interior* of C and represents the interior of C relative to its affine hull. Furthermore, we denote $\operatorname{cl}(C)$ as the closure of the set C.

Let $\mathcal{X}$ be a metric space with distance d. The *metric topology* of $\mathcal{X}$ is the topology which admits the family of all open balls $\mathcal{B}(x, \rho) = \{y \in X : d(x, y) < \rho\}$ for all $\rho > 0$ as a base. Let d be the canonical metric induced by the inner product of a Hilbert space $\mathcal{H}$. We call the metric topology of $(\mathcal{H}, d)$ the *strong topology*. We say that a sequence $(x_n)_{n \in \mathbb{N}}$ in $\mathcal{H}$ *converges strongly* to $x \in \mathcal{H}$ if it converges in the strong topology. Thus, a sequence converges strongly to $x \in \mathcal{H}$ if and only if it holds that $\lim\limits_{n \to +\infty} \|x_n - x\| = 0$. Then, we write $x_n \to x$. The *weak topology* of $\mathcal{H}$ is the family of all finite intersections of open half-spaces of $\mathcal{H}$. We say that a sequence $(x_n)_{n \in \mathbb{N}}$ in $\mathcal{H}$ *converges weakly* to $x \in \mathcal{H}$ if it converges in the weak topology. Thus, a sequence converges weakly to $x \in \mathcal{H}$ if and only if it holds that $\lim\limits_{n \to +\infty} \langle x_n, u \rangle = \langle x, u \rangle\ \forall u \in \mathcal{H}$. Then, we write $x_n \rightharpoonup x$.

If a sequence $(x_n)_{n \in \mathbb{N}}$ in $\mathcal{H}$ has a subsequence that converges weakly to a point $x \in \mathcal{H}$, then x is called a *weak sequential cluster point* of $(x_n)_{n \in \mathbb{N}}$.

In the following, we give two important propositions concerning convergent sequences.

Proposition 2.1. *(Banach-Alaoglu-Theorem) Every bounded sequence in a Hilbert space has a weakly convergent subsequence.*

Proof. See [[20], Lemma 2.45]. □

Proposition 2.2. *Let $(x_n)_{n\in\mathbb{N}}$ and $(u_n)_{n\in\mathbb{N}}$ be sequences in $\mathcal{H}$ and let $x, u \in \mathcal{H}$. Assume that $x_n \rightharpoonup x$ and $u_n \to u$. Then, $\langle x_n, u_n\rangle \to \langle x, u\rangle$.*

Proof. See [[20], Lemma 2.51 (iii)]. □

2.1 Convex functions

Now we define some essential properties of functions that are required in this thesis, and provide some important results.

Definition 2.3. We say that the function $f : \mathcal{H} \to \overline{\mathbb{R}}$ is

- *proper*, if $\operatorname{dom}(f) := \{x \in \mathcal{H} : f(x) < +\infty\} \neq \emptyset$ and $f(x) > -\infty$ for all $x \in \mathcal{H}$,
- *convex*, if for all $x, y \in \operatorname{dom}(f)$ and for all $\lambda \in [0, 1]$ it follows that

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

- *σ-strongly convex* for a $\sigma > 0$, if $f - (\sigma/2)\|\cdot\|^2$ is convex,
- *lower semicontinuous*, if $\operatorname{epi}(f) := \{(x, \xi) \in \mathcal{H} \times \mathbb{R} : f(x) \leq \xi\}$ is closed in $\mathcal{H} \times \mathbb{R}$.

For a function $f : \mathcal{H} \to \overline{\mathbb{R}}$ we denote the *image* of a function as

$$\operatorname{Im}(f) = \{f(x) : x \in \mathcal{H}\}$$

and the *graph* as

$$\operatorname{graph}(f) = \{(x, \xi) \in \mathcal{H} \times \mathbb{R} : f(x) = \xi\}.$$

Definition 2.4. The *(convex) subdifferential* of a proper function $f : \mathcal{H} \to \overline{\mathbb{R}}$ at a point $x \in \mathcal{H}$ is defined as

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

if $f(x) \in \mathbb{R}$ and as $\partial f(x) = \emptyset$, otherwise.

The global minimizers of a proper function f, which we denote by $\operatorname{argmin} f := \{x^* \in \mathcal{H} : \min_{x\in\mathcal{H}} f(x) = f(x^*)\}$, can be characterized using the subdifferential by Fermat's rule:

Theorem 2.5 (Fermat's rule). *Let $f : \mathcal{H} \to \overline{\mathbb{R}}$ be proper. Then, it holds*

$$\operatorname{argmin} f = \{x \in \mathcal{H} : 0 \in \partial f(x)\}.$$

Proof. See [[20], Theorem 16.3]. □

Proposition 2.6. *Let $f, g : \mathcal{H} \to \overline{\mathbb{R}}$ be proper.*

(i) For all $x \in \operatorname{dom}(f) \cap \operatorname{dom}(g)$ it holds

$$\partial f(x) + \partial g(x) \subseteq \partial(f+g)(x).$$

(ii) If f, g are convex and lower semicontinuous, and one of the following conditions is fulfilled

(i) $0 \in \operatorname{sqri}(\operatorname{dom}(f) - \operatorname{dom}(g))$,

(ii) $\operatorname{dom}(f) \cap \operatorname{int}(\operatorname{dom}(g)) \neq \emptyset$,

(iii) $\operatorname{dom}(g) = \mathcal{H}$,

(iv) $\mathcal{H}$ is finite-dimensional and $\operatorname{ri}(\operatorname{dom}(f)) \cap \operatorname{ri}(\operatorname{dom}(g)) \neq \emptyset$,

then it holds for all $x \in \mathcal{H}$ that

$$\partial f(x) + \partial g(x) = \partial(f+g)(x).$$

Proof. For the first part, see [[20] Proposition 16.6] and for the second part, see [[20], Corollary 16.48]. □

Note that the subdifferential of a function f, defined as

$$\partial f : \mathcal{H} \to 2^{\mathcal{H}}, \quad x \to \partial f(x),$$

is a set-valued operator. For a set-valued operator $A : \mathcal{H} \to 2^{\mathcal{H}}$, we denote the *graph* as

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

The inverse of A is denoted by $A^{-1} : \mathcal{H} \to 2^{\mathcal{H}}$ which is the mapping such that for all $x, u \in \mathcal{H}$ it holds that $x \in A^{-1}u$ if and only if $u \in Ax$.

Proposition 2.7. *Let $\mathcal{H}^{strong}$ be the strong topology of $\mathcal{H}$, $\mathcal{H}^{weak}$ the weak topology of $\mathcal{H}$, and $f : \mathcal{H} \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous. Then, the following holds:*

(i) graph(∂f) is sequentially closed in $\mathcal{H}^{strong} \times \mathcal{H}^{weak}$, i.e., for every sequence $(x_n, u_n)_{n\in\mathbb{N}}$ in graph(∂f) and every $(x,u) \in \mathcal{H} \times \mathcal{H}$, if $x_n \to x$ and $u_n \rightharpoonup u$, then $(x,u) \in$ graph(∂f).

(ii) graph(∂f) is sequentially closed in $\mathcal{H}^{weak} \times \mathcal{H}^{strong}$, i.e., for every sequence $(x_n, u_n)_{n\in\mathbb{N}}$ in graph(∂f) and every $(x,u) \in \mathcal{H} \times \mathcal{H}$, if $x_n \rightharpoonup x$ and $u_n \to u$, then $(x,u) \in$ graph(∂f)

Proof. See [[20], Proposition 16.36]. □

Definition 2.8. Let $f : \mathcal{H} \to \overline{\mathbb{R}}$ be a function. The *(Fenchel) conjugate function $f^* : \mathcal{H} \to \overline{\mathbb{R}}$* of f is defined as

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

Proposition 2.9. *Let $f : \mathcal{H} \to \overline{\mathbb{R}}$ be a proper function.*

(i) (Fenchel–Young inequality) For all $x, u \in \mathcal{H}$ it holds

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

Furthermore, this inequality becomes an equality if and only if $u \in \partial f(x)$. In this case, it follows that $x \in \partial f^*(u)$.

(ii) (Fenchel-Moreau-Theorem) The function f is convex and lower semicontinuous if and only if $f^{**} = f$, where f^{**} is the conjugate function of f^*. In this case, f^* is also a proper function.

Proof. For the first part, see [[20], Proposition 13.15 and Proposition 16.10] and for the second part, see [[20], Theorem 13.37]. □

Proposition 2.10. *Let f be proper, convex and lower semicontinuous.*

(i) Let $u \in \mathcal{H}$. Then, $(x, u) \in graph(\partial f) \Leftrightarrow (u, x) \in graph(\partial f^*)$.

(ii) It holds $\partial f^* = (\partial f)^{-1}$.

Proof. For both parts, see[[20], Theorem 16.29 and Corollary 16.30]. □

Definition 2.11. Let $\mathcal{B}$ be a Banach space, let $x \in \mathcal{H}$, let C be a open subset of $\mathcal{H}$ such that $x \in C$, and let $T : C \to \mathcal{B}$. Then, T is called *Fréchet differentiable* at x, if there exists a linear and continuous operator $DT(x) : \mathcal{H} \to \mathcal{B}$, called the *Fréchet derivative* of T at x such that

$$\lim_{\|y\|\to 0} \frac{\|T(x+y) - T(x) - DT(x)y\|}{\|y\|} = 0.$$

The *Fréchet gradient* of a Fréchet differentiable function $f : C \to \mathbb{R}$ at x is the unique vector $\nabla f(x) \in \mathcal{H}$ such that for all $y \in \mathcal{H}$ it holds

$$Df(x)y = \langle y, \nabla f(x) \rangle.$$

Note, that it follows from [[20], (iii) and (vi) from Theorem 18.15], that if $f : \mathcal{H} \to \overline{\mathbb{R}}$ is Fréchet differentiable, then f is σ-strongly convex if and only if for all $x, y \in \mathcal{H}$ it holds

$$\frac{\sigma}{2}\|x - y\|^2 \leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle. \tag{2. 1}$$

Definition 2.12. The *infimal convolution* of two proper functions $f, g : \mathcal{H} \to \overline{\mathbb{R}}$ is the function $f \Box g : \mathcal{H} \to \overline{\mathbb{R}}$, defined by

$$(f \Box g)(x) = \inf_{y \in \mathcal{H}} \{f(y) + g(x - y)\}.$$

Proposition 2.13. *Let $f, g : \mathcal{H} \to \overline{\mathbb{R}}$. Then, it holds*

(i) $(f \Box g)^* = f^* + g^*$.

(ii) If f and g are proper, convex and lower semicontinuous and $\operatorname{dom} g = \mathcal{H}$, *then*

$$(f + g)^* = f^* \Box g^*.$$

Proof. For the first part, see [[20], Proposition 13.24 (i)] and for the second part, see [[20], Proposition 15.2]. □

Definition 2.14. The *Moreau envelope* of a proper, convex and lower semicontinuous function $f : \mathcal{H} \to \overline{\mathbb{R}}$ with coefficient $\gamma > 0$ is defined as

$$f^{\gamma}(x) = \inf_{y \in \mathcal{H}} \left\{ f(y) + \frac{1}{2\gamma}\|y - x\|^2 \right\}, \quad \forall x \in \mathcal{H} \tag{2. 2}$$

and the *proximal point* of coefficient γ of the function f at the point $x \in \mathcal{H}$ is the unique optimal solution of the minimization problem above (the solution is unique, due to the strong convexity of the problem):

$$\text{Prox}_{\gamma f}(x) = \underset{y \in \mathcal{H}}{\text{argmin}} \left\{ \gamma f(y) + \frac{1}{2}\|y - x\|^2 \right\}.$$

Proposition 2.15. *Let $f : \mathcal{H} \to \overline{\mathbb{R}}$ proper, convex and lower semicontinuous and $\gamma > 0$.*

(i) (Moreau's decomposition formula). It holds

$$\text{Prox}_{\gamma f}(x) + \gamma \,\text{Prox}_{(1/\gamma)f^*}(\gamma^{-1}x) = x, \ \ \forall x \in \mathcal{H}. \tag{2. 3}$$

(ii) Let $x, p \in \mathcal{H}$. Then,

$$p = \text{Prox}_f(x) \Leftrightarrow x - p \in \partial f(p).$$

So $\text{Prox}_f = (\text{Id} + \partial f)^{-1}$.

(iii) The Moreau envelope $f^{\gamma} : \mathcal{H} \to \mathbb{R}$ is Fréchet differentiable on $\mathcal{H}$. Furthermore, its gradient

$$\nabla f^{\gamma} = \frac{1}{\gamma}(\text{Id} - \text{Prox}_{\gamma f})$$

is $(1/\gamma)$-Lipschitz continuous.

Proof. For (i), see [[20], Theorem 14.3 (ii)], for (ii), see [[20], Proposition 16.44] and for (iii), see [[20], Proposition 12.30]. □

Definition 2.16. The *indicator function* $\iota_C : \mathcal{H} \to \overline{\mathbb{R}}$ of a nonempty, closed and convex set $C \subset \mathcal{H}$ is defined as

$$\iota_C(x) = \begin{cases} 0, & \text{if } x \in C \\ +\infty, & \text{otherwise} \end{cases}$$

and the *projection operator* $\mathcal{P}_C : \mathcal{H} \to \mathcal{H}$ onto C as

$$\mathcal{P}_C(x) = \underset{y \in C}{\text{argmin}}\, \|y - x\|.$$

Note, that the projection operator at the point x is the proximal point of the indicator function at x.

We set

$$S_+(\mathcal{H}) = \{M : \mathcal{H} \to \mathcal{H} : M \text{ is linear, continuous, self-adjoint and positive semidefinite}\}.$$

For $M \in S_+(\mathcal{H})$, we define the seminorm $\|\cdot\|_M : \mathcal{H} \to [0, +\infty)$, $\|x\|_M = \sqrt{\langle x, Mx \rangle}$. We consider the *Loewner partial ordering* on $S_+(\mathcal{H})$, defined for $M_1, M_2 \in \mathcal{S}_+(\mathcal{H})$ by

$$M_1 \succcurlyeq M_2 \Leftrightarrow \|x\|^2_{M_1} \geq \|x\|^2_{M_2} \, \forall x \in \mathcal{H}. \tag{2. 4}$$

Definition 2.17. An operator sequence $(M_k)_{k\in\mathbb{N}} \in S_+(\mathcal{H})$ is said to be

- *monotonically decreasing*, if for all $k \in \mathbb{N}$ it holds $M_k \succcurlyeq M_{k+1}$ and
- *monotonically increasing*, if for all $k \in \mathbb{N}$ it holds $M_{k+1} \succcurlyeq M_k$.

Furthermore, we define for $\alpha > 0$

$$\mathcal{P}_\alpha(\mathcal{H}) := \{M \in \mathcal{S}_+(\mathcal{H}) : M \succcurlyeq \alpha \operatorname{Id}\}.$$

Let $\mathcal{G}$ be a Hilbert space and let $A : \mathcal{H} \to \mathcal{G}$ be a linear continuous operator. The operator $A^* : \mathcal{G} \to \mathcal{H}$, fulfilling $\langle A^*y, x\rangle = \langle y, Ax\rangle$ for all $x \in \mathcal{H}$ and $y \in \mathcal{G}$, denotes the *adjoint* operator of A, while $\|A\| := \sup\{\|Ax\| : \|x\| \leq 1\}$ denotes its *operator norm* .

Proposition 2.18. *Let $\alpha > 0$, let $(\eta_k)_{k\in\mathbb{N}}$ such that $\sum_{k=1}^{\infty} \eta_k < +\infty$ and let $(M_k)_{k\in\mathbb{N}}$ be a sequence in $\mathcal{P}_\alpha(\mathcal{H})$ such that $\sup_{k\in\mathbb{N}} \|M_k\| < +\infty$. Assume, that one of the following holds true:*

(i) for all $k \in \mathbb{N}$ it holds $(1 + \eta_k)M_k \succcurlyeq M_{k+1}$,

(ii) for all $k \in \mathbb{N}$ it holds $(1 + \eta_k)M_{k+1} \succcurlyeq M_k$,

then there exists $M \in \mathcal{P}_\alpha(\mathcal{H})$ such that M_k converges pointwise to M.

Proof. See [[51], Lemma 2.3]. □

Definition 2.19. We say, that the mapping $A : \mathcal{H} \to \mathcal{H}$ is

- *Lipschitz continuous* with Lipschitz constant $L > 0$ (or L-Lipschitz continuous), if for every $x, y \in \mathcal{H}$
$$\|Ax - Ay\| \leq L\|x - y\|,$$
- *nonexpansive*, if it is Lipschitz continuous with constant 1,
- *β-cocoercive* for a $\beta > 0$, if for every $x, y \in \mathcal{H}$
$$\langle x - y, Ax - Ay\rangle \geq \beta\|Ax - Ay\|^2.$$

Theorem 2.20. *(Baillon–Haddad) Let $f : \mathcal{H} \to \overline{\mathbb{R}}$ be Fréchet differentiable and convex and let $\beta > 0$. Then, ∇f is β-Lipschitz continuous if and only if ∇f is $(1/\beta)$-cocoercive.*

Proof. See [[17], Corollaire 10]. □

2.2 Monotone operators

In the following, we will look at some definitions and results concerning set-valued operators.

Definition 2.21. A set-valued operator $A : \mathcal{H} \to 2^{\mathcal{H}}$ is said to be

- *monotone* , if for all $(x, u), (y, v) \in \operatorname{graph}(A)$ it holds that
$$\langle x - y, u - v\rangle \geq 0,$$
- *maximally monotone* , if A is monotone and there exists no monotone operator $B : \mathcal{H} \to 2^{\mathcal{H}}$ such that $\operatorname{graph}(A) \subsetneq \operatorname{graph}(B)$, i.e. for every $(x, u) \in \mathcal{H} \times \mathcal{H}$ it holds that
$$(x, u) \in \operatorname{graph}(A) \quad \Leftrightarrow \quad \langle x - y, u - v\rangle \geq 0 \text{ for all } (y, v) \in \operatorname{graph}(A),$$

- *α-strongly monotone*, if for an $\alpha > 0$ and all $(x, u), (y, v) \in \text{graph}(A)$ it holds that

$$\langle x - y, u - v \rangle \geq \alpha \|x - y\|^2.$$

Proposition 2.22. *Let $f : \mathcal{H} \to \overline{\mathbb{R}}$ be a proper, convex and lower semicontinuous function.*

(i) The subdifferential ∂f is a maximally monotone operator.

(ii) If f is α-strongly convex with $\alpha > 0$, then ∂f is α-strongly monotone.

Proof. For the first part, see [[84], Proposition 12.b.] and for the second part, see [[84], Example 22.4 (iv)]. □

Chapter 3

The Proximal Alternating Minimization Algorithm (Proximal AMA)

This chapter is based on the paper [28].

Tseng introduced in [110] the so-called Alternating Minimization Algorithm (AMA) to solve optimization problems with two-block separable linear constraints and two nonsmooth convex objective functions, one of them assumed to be strongly convex. The numerical scheme consists in each iteration of two minimization subproblems, each involving one of the two objective functions, and of an update of the dual sequence which approaches asymptotically a Lagrange multiplier of the dual problem.

The strong convexity of one of the objective functions allows to reduce the corresponding minimization subproblem to the calculation of the proximal operator of a proper, convex and lower semicontinuous function. This is for the second minimization problem in general not the case, thus, with the exception of some very particular cases, one has to use a subroutine in order to compute the corresponding iterate. This may have a negative influence on the convergence behavior of the algorithm and affects its computational tractability. One possibility to avoid this is to properly modify this subproblem with the aim of transforming it into a proximal step, and, of course, without losing the convergence properties of the algorithm. The papers [24] and [50] provide convincing evidences for the efficiency and versatility of proximal point algorithms for solving nonsmooth convex optimization problems; we also refer to [49] for a block coordinate variable metric forward-backward method.

We address in a real Hilbert space setting a more involved two-block separable optimization problem, which is obtained by adding in each block of the objective a further smooth convex function. To solve this problem, we propose a so-called Proximal Alternating Minimization Algorithm (Proximal AMA), which is obtained by inducing in each of the minimization subproblems additional proximal terms defined by means of positively semidefinite operators. The two smooth convex functions in the objective are evaluated via gradient steps. For appropriate choices of these operators, we show that the minimization subproblems turn into proximal steps and the algorithm becomes an iterative scheme formulated in the spirit of the full splitting paradigm. We show that the generated sequence converges weakly to a saddle point of the Lagrangian associated with the optimization problem under investigation. The numerical performances of Proximal AMA are illustrated in particular in comparison to AMA for two applications in image processing and machine learning.

A similarity of AMA to the classical Alternating Direction Method of Multipliers (ADMM) algorithm, introduced by Gabay and Mercier in [63], is obvious. In [14], [61] and [102] (see also [19] and [39]) proximal versions of the ADMM algorithm have been proposed and proved to provide a unifying framework for primal-dual algorithms for convex optimization. Parts of the convergence analysis for the Proximal AMA are carried out in a similar spirit to the convergence proofs in these papers.

3.1 The Alternating Minimization Algorithm

The convex optimization problems addressed in [110] is of the form

$$\inf_{x\in R^n, z\in\mathbb{R}^m} f(x)+g(z) \quad \text{s.t.} \quad Ax+Bz=b, \tag{3. 1}$$

where $f:\mathbb{R}^n\to\overline{\mathbb{R}}$ is a proper, γ-strongly convex with $\gamma>0$ and lower semicontinuous function, $g:\mathbb{R}^m\to\overline{\mathbb{R}}$ is a proper, convex and lower semicontinuous function, $A\in\mathbb{R}^{r\times n}, B\in\mathbb{R}^{r\times m}$ and $b\in\mathbb{R}^r$.

The Lagrangian associated with problem (3. 1) is

$$L:\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^r\to\overline{\mathbb{R}}, \quad L(x,z,y)=f(x)+g(z)+\langle y,b-Ax-Bz\rangle.$$

For $c>0$, the augmented Lagrangian associated with problem (3. 1), $L_c:\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^r\to\overline{\mathbb{R}}$ reads

$$L_c(x,z,y)=f(x)+g(z)+\langle y,b-Ax-Bz\rangle+\frac{c}{2}\|Ax+Bz-b\|^2.$$

Tseng proposed in [110] the following so-called Alternating Minimization Algorithm (AMA) for solving (3. 1):

Algorithm 3.1 Alternating Minimization Algorithm (AMA)

Choose $y^0\in\mathbb{R}^r$ and a sequence of strictly positive stepsizes $(c_k)_{k\geq 0}$. For all $k\geq 0$ set:

$$x^k=\operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{f(x)-\langle y^k,Ax\rangle\right\} \tag{3. 2}$$

$$z^k\in\operatorname*{argmin}_{z\in\mathbb{R}^m}\left\{g(z)-\langle y^k,Bz\rangle+\frac{c_k}{2}\|Ax^k+Bz-b\|^2\right\} \tag{3. 3}$$

$$y^{k+1}=y^k+c_k(b-Ax^k-Bz^k). \tag{3. 4}$$

The main convergence properties of this numerical algorithm are summarized in the theorem below (see [110]).

Theorem 3.2. *Let $A\neq 0$ and $(x,z)\in\operatorname{ri}(\operatorname{dom} f)\times\operatorname{ri}(\operatorname{dom} g)$ be such that the equality $Ax+Bz=b$ holds. Assume that the sequence of stepsizes $(c_k)_{k\geq 0}$ satisfies*

$$\epsilon\leq c_k\leq\frac{2\gamma}{\|A\|^2}-\epsilon \quad \forall k\geq 0,$$

where $0<\epsilon<\frac{\gamma}{\|A\|^2}$. Let $(x^k,z^k,y^k)_{k\geq 0}$ be the sequence generated by Algorithm 3.1. Then, there exist $x^*\in\mathbb{R}^n$ and an optimal Lagrange multiplier $y^*\in\mathbb{R}^r$ associated with the constraint $Ax+Bz=b$ such that

$$x^k\to x^*, Bz^k\to b-Ax^*, y^k\to y^*(k\to+\infty).$$

If the function $z\mapsto g(z)+\|Bz\|^2$ has bounded level sets, then $(z^k)_{k\geq 0}$ is bounded and any of its cluster points z^* provides with (x^*,z^*) an optimal solution of (3. 1).

It is the aim of this chapter to propose a proximal variant of this algorithm, called Proximal AMA, which overcomes its drawbacks, and to investigate its convergence properties.

3.2 Problem formulation

The two-block separable optimization problem we are going to investigate in this and the next chapter has the following formulation.

Problem 3.3. *Let $\mathcal{H}$, $\mathcal{G}$ and $\mathcal{K}$ be real Hilbert spaces, $f : \mathcal{H} \to \overline{\mathbb{R}}$ a proper, lower semicontiuous and γ-strongly convex function with $\gamma > 0$, $g : \mathcal{G} \to \overline{\mathbb{R}}$ a proper, convex and lower semicontinuous function, $h_1 : \mathcal{H} \to \mathbb{R}$ a convex and Fréchet differentiable function with L_1-Lipschitz continuous gradient with $L_1 \geq 0$, $h_2 : \mathcal{G} \to \mathbb{R}$ a convex and Fréchet differentiable functions with L_2-Lipschitz continuous gradient with $L_2 \geq 0$, $A : \mathcal{H} \to \mathcal{K}$ and $B : \mathcal{G} \to \mathcal{K}$ linear continuous operators such that $A \neq 0$ and $b \in \mathcal{K}$. Consider the following optimization problem with two-block separable objective function and linear constraints*

$$\min_{x\in\mathcal{H},z\in\mathcal{G}} f(x) + h_1(x) + g(z) + h_2(z) \quad s.t. \quad Ax + Bz = b. \tag{3. 5}$$

We allow the Lipschitz constants of the gradients of the functions h_1 and h_2 to be zero. In this case, the functions are affine.

We can write the optimization problem (3. 5) as

$$\min_{(x,z)\in\mathcal{H}\times\mathcal{G}} F(x,z) + G(L(x,z)), \tag{3. 6}$$

where $F(x,z) = f(x) + h_1(x) + g(z) + h_2(z)$, $G : \mathcal{K} \to \overline{\mathbb{R}}$ is defined as $G(x) = \iota_{\{b\}}(x)$ and $L : \mathcal{H} \times \mathcal{G} \to \mathcal{K}$ is a linear continuous operator such that $L(x,z) = Ax + Bz$.

The Fenchel dual problem of the optimization problem (3. 6) is defined as

$$\sup_{y\in\mathcal{K}}\{-F^*(-L^*y - G^*(y)\} \tag{3. 7}$$

(see [[34], chapter 2]), which can be written for the optimization problem (3. 5) as (note Proposition 2.13)

$$\sup_{y\in\mathcal{K}}\{-(f^*\Box h_1^*)(-A^*y) - (g^*\Box h_2^*)(-B^*y) - \langle y,b\rangle\}. \tag{3. 8}$$

The Lagrangian associated with the optimization problem (3. 5) is defined by

$$L : \mathcal{H} \times \mathcal{G} \times \mathcal{K} \to \overline{\mathbb{R}},$$

which can be written as

$$L(x,z,y) = f(x) + h_1(x) + g(z) + h_2(z) + \langle y, b - Ax - Bz\rangle.$$

Definition 3.4. We say that $(x^*, z^*, y^*) \in \mathcal{H} \times \mathcal{G} \times \mathcal{K}$ is a *saddle point* of the Lagrangian L, if

$$L(x^*, z^*, y) \leq L(x^*, z^*, y^*) \leq L(x, z, y^*) \quad \forall(x,z,y) \in \mathcal{H} \times \mathcal{G} \times \mathcal{K}.$$

In the following, we assume that a qualification condition, like for instance the Attouch-Brézis-type condition (other qualification conditions can be found in [40], page 76f.)

$$b \in \operatorname{sqri}(A(\operatorname{dom} f) + B(\operatorname{dom} g)), \tag{3. 9}$$

holds. In the finite-dimensional setting, this asks for the existence of $x \in \operatorname{ri}(\operatorname{dom} f)$ and $z \in \operatorname{ri}(\operatorname{dom} g)$ satisfying $Ax + Bz = b$ and coincides with the assumption used by Tseng in [110]. It follows that the dual problem (3. 8) has an optimal solution (see [[40], Theorem 3.2.11]).

Moreover, it holds according to [40] (Theorem 3.3.4. and Remark 3.3.3) that if (3. 9) is fulfilled, then for any optimal solution (x^*, z^*) of (3. 5) there exists an optimal solution y^* of the dual problem (3. 8) such that the optimality conditions

$$A^*y^* - \nabla h_1(x^*) \in \partial f(x^*),\ B^*y^* - \nabla h_2(z^*) \in \partial g(z^*) \text{ and } Ax^* + Bz^* = b \tag{3. 10}$$

are fulfilled. Conversely, if the optimality conditions (3. 10) are fulfilled for (x^*, z^*, y^*), then (x^*, z^*) is an optimal solution for the primal problem (3. 5), y^* is an optimal solution for the dual problem (3. 8) and the objective values of (3. 5) and (3. 8) coincide. However, the validity of (3. 10) doesn't necessarily imply that (3. 9) holds. Furthermore, we have that the point (x^*, z^*, y^*) is a saddle point of the Lagrangian L if and only if (x^*, z^*) is an optimal solution of (3. 5), y^* is an optimal solution of its Fenchel dual problem (3. 8) and the optimal objective values of (3. 5) and (3. 8) coincide (see [[40], Theorem 3.3.5.]).

As a consequence, it follows that $(x^*, z^*, y^*) \in \mathcal{H} \times \mathcal{G} \times \mathcal{K}$ is a saddle point of the Lagrangian L if and only if (3. 10) holds.

Remark 3.5. If (x_1^*, z_1^*, y_1^*) and (x_2^*, z_2^*, y_2^*) are two saddle points of the Lagrangian L, then $x_1^* = x_2^*$.

This follows from (3. 10) and the strong convexity of f with $\gamma > 0$. So we have for two saddle points (x_1^*, z_1^*, y_1^*) and (x_2^*, z_2^*, y_2^*), that $(x_1^*, A^*y_1^*)$, $(x_2^*, A^*y_2^*) \in \operatorname{graph}(\partial(f + h_1))$ and $(z_1^*, B^*y_1^*)$, $(z_2^*, B^*y_2^*) \in \operatorname{graph}(\partial(g + h_2))$ and by using the γ-strong monotonicity of ∂f and the monotonicity of ∂g from Proposition 2.22 we have

$$\langle x_1^* - x_2^*, A^*y_1^* - A^*y_2^* \rangle \geq \gamma \|x_1^* - x_2^*\|^2 \quad \text{and} \quad \langle z_1^* - z_2^*, B^*y_1^* - B^*y_2^* \rangle \geq 0,$$

which is equivalent to

$$\langle Ax_1^* - Ax_2^*, y_1^* - y_2^* \rangle \geq \gamma \|x_1^* - x_2^*\|^2 \quad \text{and} \quad \langle Bz_1^* - Bz_2^*, y_1^* - y_2^* \rangle \geq 0.$$

Adding this two inequalities and using $Ax_1^* + Bz_1^* = Ax_2^* + Bz_2^*$ from (3. 10), we obtain

$$0 \geq \gamma \|x_1^* - x_2^*\|^2,$$

which implies that $x_1^* = x_2^*$.

For more on the AMA algorithm introduced by Tseng and motivation for considering this setting, we refer the reader to [110, 65].

3.3 The Proximal Alternating Minimization Algorithm

In the following, we formulate the Proximal Alternating Minimization Algorithm to solve the optimization problem (3. 5). To this end, we modify Tseng's AMA by evaluating in each of the two subproblems the functions h_1 and h_2 via gradient steps, respectively, and by introducing proximal terms defined through two sequences of positively semidefinite operators $(M_1^k)_{k\geq 0}$ and $(M_2^k)_{k\geq 0}$.

Algorithm 3.6 Proximal Alternating Minimization Algorithm (Proximal AMA)

Let $(M_1^k)_{k\geq 0} \subseteq \mathcal{S}_+(\mathcal{H})$ and $(M_2^k)_{k\geq 0} \subseteq \mathcal{S}_+(\mathcal{G})$. Choose $(x^0, z^0, y^0) \in \mathcal{H} \times \mathcal{G} \times \mathcal{K}$ and a sequence of stepsizes $(c_k)_{k\geq 0} \subseteq (0, +\infty)$. For all $k \geq 0$ set:

$$x^{k+1} = \operatorname*{argmin}_{x\in\mathcal{H}} \left\{ f(x) - \langle y^k, Ax\rangle + \langle x - x^k, \nabla h_1(x^k)\rangle + \frac{1}{2}\|x - x^k\|^2_{M_1^k} \right\} \tag{3. 11}$$

$$\begin{aligned} z^{k+1} \in \operatorname*{argmin}_{z\in\mathcal{G}} \Big\{ & g(z) - \langle y^k, Bz\rangle + \frac{c_k}{2}\|Ax^{k+1} + Bz - b\|^2 \\ & + \langle z - z^k, \nabla h_2(z^k)\rangle + \frac{1}{2}\|z - z^k\|^2_{M_2^k} \Big\} \end{aligned} \tag{3. 12}$$

$$y^{k+1} = y^k + c_k(b - Ax^{k+1} - Bz^{k+1}). \tag{3. 13}$$

Remark 3.7. The sequence $(z^k)_{k\geq 0}$ is uniquely determined if there exists $\alpha_k > 0$ such that $c_k B^*B + M_2^k \in \mathcal{P}_{\alpha_k}(\mathcal{G})$ for all $k \geq 0$. This actually ensures that the objective function in the subproblem (3. 12) is strongly convex.

Remark 3.8. Let $k \geq 0$ be fixed and $M_2^k := \frac{1}{\sigma_k}\operatorname{Id} - c_k B^*B$, where $\sigma_k > 0$ and $\sigma_k c_k \|B\|^2 \leq 1$. Then, M_2^k is positive semidefinite and the update of z^{k+1} in the Proximal AMA method becomes a proximal step. Indeed, (3. 12) holds according to Fermat's rule 2.5 and Proposition 2.6 (ii) if and only if

$$0 \in \partial g(z^{k+1}) + (c_k B^*B + M_2^k)z^{k+1} + c_k B^*(Ax^{k+1} - b) - M_2^k z^k + \nabla h_2(z^k) - B^*y^k$$

or, equivalently,

$$0 \in \partial g(z^{k+1}) + \frac{1}{\sigma_k} z^{k+1} - \left(\frac{1}{\sigma_k}\operatorname{Id} - c_k B^*B\right) z^k + \nabla h_2(z^k) + c_k B^*(Ax^{k+1} - b) - B^*y^k.$$

But this is nothing else than

$$\begin{aligned} z^{k+1} &= \operatorname*{argmin}_{z\in\mathcal{G}} \left\{ g(z) + \frac{1}{2\sigma_k}\left\| z - \left(z^k - \sigma_k \nabla h_2(z^k) + \sigma_k c_k B^*(b - Ax^{k+1} - Bz^k) + \sigma_k B^* y^k\right)\right\|^2 \right\} \\ &= \operatorname{Prox}_{\sigma_k g}\left(z^k - \sigma_k \nabla h_2(z^k) + \sigma_k c_k B^*(b - Ax^{k+1} - Bz^k) + \sigma_k B^* y^k\right). \end{aligned}$$

The convergence of the Proximal AMA method is addressed in the next theorem.

Theorem 3.9. *In the setting of Problem 3.3, let the set of the saddle points of the Lagrangian L be nonempty. We assume that* $M_1^k - \frac{L_1}{2}\operatorname{Id} \in \mathcal{S}_+(\mathcal{H}), M_1^k \succcurlyeq M_1^{k+1}$, $M_2^k - \frac{L_2}{2}\operatorname{Id} \in \mathcal{S}_+(\mathcal{G}), M_2^k \succcurlyeq M_2^{k+1}$ *for all* $k \geq 0$ *and that* $(c_k)_{k\geq 0}$ *is a monotonically decreasing sequence satisfying*

$$\epsilon \leq c_k \leq \frac{2\gamma}{\|A\|^2} - \epsilon \quad \forall k \geq 0, \tag{3. 14}$$

where $0 < \epsilon < \frac{\gamma}{\|A\|^2}$. *If one of the following assumptions:*

(i) there exists $\alpha > 0$ *such that* $M_2^k - \frac{L_2}{2}\operatorname{Id} \in \mathcal{P}_\alpha(\mathcal{G})$ *for all* $k \geq 0$;

(ii) there exists $\beta > 0$ *such that* $B^*B \in \mathcal{P}_\beta(\mathcal{G})$;

holds true, then the sequence $(x^k, z^k, y^k)_{k\geq 0}$ *generated by Algorithm 3.1 converges weakly to a saddle point of the Lagrangian* L.

Proof. Let (x^*, z^*, y^*) be a fixed saddle point of the Lagrangian L. This means that it fulfills the system of optimality conditions

$$A^*y^* - \nabla h_1(x^*) \in \partial f(x^*) \tag{3. 15}$$
$$B^*y^* - \nabla h_2(z^*) \in \partial g(z^*) \tag{3. 16}$$
$$Ax^* + Bz^* = b. \tag{3. 17}$$

We start by proving that

$$\sum_{k\geq 0} \|x^{k+1} - x^*\|^2 < +\infty, \sum_{k\geq 0} \|Bz^{k+1} - Bz^*\|^2 < +\infty, \sum_{k\geq 0} \|z^{k+1} - z^k\|^2_{M_2^k - \frac{L_2}{2}\operatorname{Id}} < +\infty$$

and that the sequences $(z^k)_{k\geq 0}$ and $(y^k)_{k\geq 0}$ are bounded.

Assume that $L_1 > 0$ and $L_2 > 0$. Let $k \geq 0$ be fixed. Writing the optimality conditions for the subproblems (3. 11) and (3. 12), we obtain

$$A^*y^k - \nabla h_1(x^k) + M_1^k(x^k - x^{k+1}) \in \partial f(x^{k+1}) \tag{3. 18}$$

and

$$B^*y^k - \nabla h_2(z^k) + c_k B^*(-Ax^{k+1} - Bz^{k+1} + b) + M_2^k(z^k - z^{k+1}) \in \partial g(z^{k+1}), \tag{3. 19}$$

respectively. Combining (3. 15) - (3. 19) with the strong monotonicity of ∂f and the monotonicity of ∂g (see Proposition 2.22), it yields

$$\langle A^*(y^k - y^*) - \nabla h_1(x^k) + \nabla h_1(x^*) + M_1^k(x^k - x^{k+1}), x^{k+1} - x^*\rangle \geq \gamma\|x^{k+1} - x^*\|^2$$

and

$$\begin{aligned}\langle B^*(y^k - y^*) - \nabla h_2(z^k) + \nabla h_2(z^*) + c_k B^*(-Ax^{k+1} - Bz^{k+1} + b)&\\ +M_2^k(z^k - z^{k+1}), z^{k+1} - z^*\rangle &\geq 0,\end{aligned}$$

which after summation leads to

$$\begin{aligned}\langle y^k - y^*, Ax^{k+1} - Ax^*\rangle + \langle y^k - y^*, Bz^{k+1} - Bz^*\rangle&\\ +\langle c_k(-Ax^{k+1} - Bz^{k+1} + b), Bz^{k+1} - Bz^*\rangle&\\ -\langle \nabla h_1(x^k) - \nabla h_1(x^*), x^{k+1} - x^*\rangle - \langle \nabla h_2(z^k) - \nabla h_2(z^*), z^{k+1} - z^*\rangle&\\ +\langle M_1^k(x^k - x^{k+1}), x^{k+1} - x^*\rangle + \langle M_2^k(z^k - z^{k+1}), z^{k+1} - z^*\rangle &\geq \gamma\|x^{k+1} - x^*\|^2.\end{aligned} \tag{3. 20}$$

According to the Baillon-Haddad-Theorem 2.20, the gradients of h_1 and h_2 are $\frac{1}{L_1}$ and $\frac{1}{L_2}$-cocoercive, respectively, thus

$$\langle \nabla h_1(x^*) - \nabla h_1(x^k), x^* - x^k\rangle \geq \frac{1}{L_1}\|\nabla h_1(x^*) - \nabla h_1(x^k)\|^2$$
$$\langle \nabla h_2(z^*) - \nabla h_2(z^k), z^* - z^k\rangle \geq \frac{1}{L_2}\|\nabla h_2(z^*) - \nabla h_2(z^k)\|^2.$$

On the other hand, by taking into account (3. 13) and (3. 17), it holds

$$\begin{aligned}\langle y^k - y^*, Ax^{k+1} - Ax^* \rangle + \langle y^k - y^*, Bz^{k+1} - Bz^* \rangle &= \langle y^k - y^*, Ax^{k+1} + Bz^{k+1} - b \rangle \\ &= \frac{1}{c_k} \langle y^k - y^*, y^k - y^{k+1} \rangle.\end{aligned}$$

By employing the last three relations in (3. 20), it yields

$$\begin{aligned}&\frac{1}{c_k}\langle y^k - y^*, y^k - y^{k+1}\rangle + c_k\langle -Ax^{k+1} - Bz^{k+1} + b, Bz^{k+1} - Bz^*\rangle \\ &\quad + \langle M_1^k(x^k - x^{k+1}), x^{k+1} - x^*\rangle + \langle M_2^k(z^k - z^{k+1}), z^{k+1} - z^*\rangle \\ &+ \langle \nabla h_1(x^*) - \nabla h_1(x^k), x^{k+1} - x^*\rangle + \langle \nabla h_1(x^*) - \nabla h_1(x^k), x^* - x^k\rangle \\ &- \frac{1}{L_1}\|\nabla h_1(x^*) - \nabla h_1(x^k)\|^2 + \langle \nabla h_2(z^*) - \nabla h_2(z^k), z^{k+1} - z^*\rangle \\ &+ \langle \nabla h_2(z^*) - \nabla h_2(z^k), z^* - z^k\rangle - \frac{1}{L_2}\|\nabla h_2(z^*) - \nabla h_2(z^k)\|^2 \geq \gamma\|x^{k+1} - x^*\|^2,\end{aligned}$$

which, after expressing the inner products by means of norms, becomes

$$\begin{aligned}&\frac{1}{2c_k}\left(\|y^k - y^*\|^2 + \|y^k - y^{k+1}\|^2 - \|y^{k+1} - y^*\|^2\right) \\ &+ \frac{c_k}{2}\left(\|Ax^* - Ax^{k+1}\|^2 - \|b - Ax^{k+1} - Bz^{k+1}\|^2 - \|Ax^* + Bz^{k+1} - b\|^2\right) \\ &+ \frac{1}{2}\left(\|x^k - x^*\|^2_{M_1^k} - \|x^k - x^{k+1}\|^2_{M_1^k} - \|x^{k+1} - x^*\|^2_{M_1^k}\right) \\ &+ \frac{1}{2}\left(\|z^k - z^*\|^2_{M_2^k} - \|z^k - z^{k+1}\|^2_{M_2^k} - \|z^{k+1} - z^*\|^2_{M_2^k}\right) \\ &+ \langle \nabla h_1(x^*) - \nabla h_1(x^k), x^{k+1} - x^k\rangle - \frac{1}{L_1}\|\nabla h_1(x^*) - \nabla h_1(x^k)\|^2 \\ &+ \langle \nabla h_2(z^*) - \nabla h_2(z^k), z^{k+1} - z^k\rangle - \frac{1}{L_2}\|\nabla h_2(z^*) - \nabla h_2(z^k)\|^2 \geq \gamma\|x^{k+1} - x^*\|^2.\end{aligned}$$

Using again (3. 13), the inequality $\|Ax^* - Ax^{k+1}\|^2 \leq \|A\|^2\|x^* - x^{k+1}\|^2$ and the following expressions

$$\begin{aligned}&\langle \nabla h_1(x^*) - \nabla h_1(x^k), x^{k+1} - x^k\rangle - \frac{1}{L_1}\|\nabla h_1(x^*) - \nabla h_1(x^k)\|^2 \\ &\qquad = -L_1\left\|\frac{1}{L_1}(\nabla h_1(x^*) - \nabla h_1(x^k)) + \frac{1}{2}(x^k - x^{k+1})\right\|^2 + \frac{L_1}{4}\|x^k - x^{k+1}\|^2,\end{aligned}$$

and

$$\begin{aligned}&\langle \nabla h_2(x^*) - \nabla h_2(z^k), z^{k+1} - z^k\rangle - \frac{1}{L_2}\|\nabla h_2(z^*) - \nabla h_2(z^k)\|^2 \\ &\qquad = -L_2\left\|\frac{1}{L_2}(\nabla h_2(z^*) - \nabla h_2(z^k)) + \frac{1}{2}(z^k - z^{k+1})\right\|^2 + \frac{L_2}{4}\|z^k - z^{k+1}\|^2,\end{aligned}$$

it yields

$$\begin{aligned}
&\frac{1}{2}\|x^{k+1}-x^*\|^2_{M_1^k}+\frac{1}{2c_k}\|y^{k+1}-y^*\|^2+\frac{1}{2}\|z^{k+1}-z^*\|^2_{M_2^k}\\
&\quad\leq \frac{1}{2}\|x^k-x^*\|^2_{M_1^k}+\frac{1}{2c_k}\|y^k-y^*\|^2+\frac{1}{2}\|z^k-z^*\|^2_{M_2^k}-\frac{c_k}{2}\|Ax^*+Bz^{k+1}-b\|^2\\
&\quad-\frac{1}{2}\|z^k-z^{k+1}\|^2_{M_2^k}-\left(\gamma-\frac{c_k}{2}\|A\|^2\right)\|x^{k+1}-x^*\|^2-\frac{1}{2}\|x^k-x^{k+1}\|^2_{M_1^k}\\
&\quad-L_1\left\|\frac{1}{L_1}(\nabla h_1(x^*)-\nabla h_1(x^k))+\frac{1}{2}(x^k-x^{k+1})\right\|^2+\frac{L_1}{4}\|x^k-x^{k+1}\|^2\\
&\quad-L_2\left\|\frac{1}{L_2}(\nabla h_2(z^*)-\nabla h_2(z^k))+\frac{1}{2}(z^k-z^{k+1})\right\|^2+\frac{L_2}{4}\|z^k-z^{k+1}\|^2.
\end{aligned}$$

Finally, by using the monotonicity of $(M_1^k)_{k\geq 0}$, $(M_2^k)_{k\geq 0}$ and of $(c_k)_{k\geq 0}$, we obtain

$$\begin{aligned}
c_{k+1}\|x^{k+1}-x^*\|^2_{M_1^{k+1}}+\|y^{k+1}-y^*\|^2+c_{k+1}\|z^{k+1}-z^*\|^2_{M_2^{k+1}}&\\
\leq c_k\|x^k-x^*\|^2_{M_1^k}+\|y^k-y^*\|^2+c_k\|z^k-z^*\|^2_{M_2^k}-R_k,&\qquad (3.\ 21)
\end{aligned}$$

where

$$\begin{aligned}
R_k:=\ &c_k\left(2\gamma-c_k\|A\|^2\right)\|x^{k+1}-x^*\|^2+c_k^2\|Bz^{k+1}-Bz^*\|^2\\
&+c_k\|z^k-z^{k+1}\|^2_{M_2^k-\frac{L_2}{2}\operatorname{Id}}+c_k\|x^k-x^{k+1}\|^2_{M_1^k-\frac{L_1}{2}\operatorname{Id}}\\
&+\ 2c_kL_1\left\|\frac{1}{L_1}(\nabla h_1(x^*)-\nabla h_1(x^k))+\frac{1}{2}(x^k-x^{k+1})\right\|^2\\
&+\ 2c_kL_2\left\|\frac{1}{L_2}(\nabla h_2(z^*)-\nabla h_2(z^k))+\frac{1}{2}(z^k-z^{k+1})\right\|^2.
\end{aligned}$$

If $L_1=0$ (and, consequently, ∇h_1 is constant) and $L_2>0$, then, by using the same arguments, we obtain again (3. 21), but with

$$\begin{aligned}
R_k:=\ &c_k\left(2\gamma-c_k\|A\|^2\right)\|x^{k+1}-x^*\|^2+c_k^2\|Bz^{k+1}-Bz^*\|^2\\
&+\ c_k\|z^k-z^{k+1}\|^2_{M_2^k-\frac{L_2}{2}\operatorname{Id}}+c_k\|x^k-x^{k+1}\|^2_{M_1^k}\\
&+\ 2c_kL_2\left\|\frac{1}{L_2}(\nabla h_2(z^*)-\nabla h_2(z^k))+\frac{1}{2}(z^k-z^{k+1})\right\|^2
\end{aligned}$$

and analogously, if $L_2=0$ and $L_1>0$

$$\begin{aligned}
R_k:=\ &c_k\left(2\gamma-c_k\|A\|^2\right)\|x^{k+1}-x^*\|^2+c_k^2\|Bz^{k+1}-Bz^*\|^2\\
&+\ c_k\|z^k-z^{k+1}\|^2_{M_2^k}+c_k\|x^k-x^{k+1}\|^2_{M_1^k-\frac{L_1}{2}\operatorname{Id}}\\
&+\ 2c_kL_1\left\|\frac{1}{L_1}(\nabla h_1(x^*)-\nabla h_1(x^k))+\frac{1}{2}(x^k-x^{k+1})\right\|^2.
\end{aligned}$$

Relation (3. 21) follows even if $L_1=L_2=0$, but with

$$R_k:=c_k\left(2\gamma-c_k\|A\|^2\right)\|x^{k+1}-x^*\|^2+c_k^2\|Bz^{k+1}-Bz^*\|^2+c_k\|z^k-z^{k+1}\|^2_{M_2^k}+c_k\|x^k-x^{k+1}\|^2_{M_1^k}.$$

Notice that, due to $M_1^k - \frac{L_1}{2}\operatorname{Id} \in \mathcal{S}_+(\mathcal{H})$ and $M_2^k - \frac{L_2}{2}\operatorname{Id} \in \mathcal{S}_+(\mathcal{G})$, all summands in R_k are nonnegative.

Let $N \geq 0$ be fixed. By summing the inequality in (3. 21) for $k = 0, ..., N$ and using telescoping arguments, we obtain

$$\begin{aligned} c_{N+1}\|x^{N+1} - &x^*\|^2_{M_1^{N+1}} + \|y^{N+1} - y^*\|^2 + c_{N+1}\|z^{N+1} - z^*\|^2_{M_2^{N+1}} \\ &\leq c_0\|x^0 - x^*\|^2_{M_1^0} + \|y^0 - y^*\|^2 + c_0\|z^0 - z^*\|_{M_2^0} - \sum_{k=0}^{N} R_k. \end{aligned}$$

On the other hand, from (3. 21), we derive

$$\exists \lim_{k\to\infty}\left(c_k\|x^k - x^*\|^2_{M_1^k} + \|y^k - y^*\|^2 + c_k\|z^k - z^*\|^2_{M_2^k}\right). \tag{3. 22}$$

Thus, $(y^k)_{k\geq 0}$ is bounded and $\sum_{k\geq 0} R_k < +\infty$.

Taking (3. 14) into account, we have $c_k(2\gamma - c_k\|A\|^2) \geq \varepsilon^2\|A\|^2$ for all $k \geq 0$. It follows that

$$\sum_{k\geq 0}\|x^{k+1} - x^*\|^2 < +\infty, \quad \sum_{k\geq 0}\|Bz^{k+1} - Bz^*\|^2 < +\infty \tag{3. 23}$$

and

$$\sum_{k\geq 0}\|z^{k+1} - z^k\|^2_{M_2^k - \frac{L_2}{2}\operatorname{Id}} < +\infty. \tag{3. 24}$$

From here we obtain

$$x^k \to x^*, \quad Bz^k \to Bz^* \ (k \to +\infty), \tag{3. 25}$$

which, by using (3. 13) and (3. 17), leads to

$$y^k - y^{k+1} \to 0 \ (k \to +\infty). \tag{3. 26}$$

Taking into account the monotonicity properties of $(c_k)_{k\geq 0}$ and $(M_1^k)_{k\geq 0}$, a direct implication of (3. 22) and (3. 25) is

$$\exists \lim_{k\to\infty}\left(\|y^k - y^*\|^2 + c_k\|z^k - z^*\|^2_{M_2^k}\right). \tag{3. 27}$$

Suppose that assumption (i) holds true, namely, that there exists $\alpha > 0$ such that $M_2^k - \frac{L_2}{2}\operatorname{Id} \in \mathcal{P}_\alpha(\mathcal{G})$ for all $k \geq 0$. From (3. 27), we can conclude that $(z^k)_{k\geq 0}$ is bounded, while (3. 24) ensures that

$$z^{k+1} - z^k \to 0 \ (k \to +\infty). \tag{3. 28}$$

In the following, let us prove that each weak sequential cluster point of $(x^k, z^k, y^k)_{k\geq 0}$ (notice that the sequence is bounded) is a saddle point of L. Let be $(\bar{z}, \bar{y}) \in \mathcal{G} \times \mathcal{K}$ such that the subsequence $(x^{k_j}, z^{k_j}, y^{k_j})_{j\geq 0}$ converges weakly to $(x^*, \bar{z}, \bar{y})$ as $j \to +\infty$, according to Proposition 2.1 (Notice that x^k converges strongly to x^*). From (3. 18), we have

$$A^*y^{k_j} - \nabla h_1(x^{k_j}) + M_1^{k_j}(x^{k_j} - x^{k_j+1}) \in \partial f(x^{k_j+1}) \ \forall j \geq 1.$$

Due to the fact that x^{k_j} converges strongly to x^* and y^{k_j} converges weakly to a $\bar{y}$ as $j \to +\infty$, using the continuity of ∇h_1 and the fact that the graph of the convex subdifferential of f is sequentially closed in the strong-weak topology according to Proposition 2.7, it follows

$$A^*\bar{y} - \nabla h_1(x^*) \in \partial f(x^*).$$

From (3. 19), we have for all $j \geq 0$

$$B^*y^{k_j} - \nabla h_2(z^{k_j}) + c_{k_j}B^*(-Ax^{k_j+1} - Bz^{k_j+1} + b) + M_2^{k_j}(z^{k_j} - z^{k_j+1}) \in \partial g(z^{k_j+1}),$$

which is equivalent to

$$B^*y^{k_j} + \nabla h_2(z^{k_j+1}) - \nabla h_2(z^{k_j}) + c_{k_j}B^*(-Ax^{k_j+1} - Bz^{k_j+1} + b) + M_2^{k_j}(z^{k_j} - z^{k_j+1}) \in \partial(g + h_2)(z^{k_j+1})$$

and further, due to Proposition (2.10), to

$$\begin{aligned} z^{k_j+1} \in \partial(g+h_2)^* \Big(& B^*y^{k_j} + \nabla h_2(z^{k_j+1}) - \nabla h_2(z^{k_j}) + c_{k_j}B^*(-Ax^{k_j+1} - Bz^{k_j+1} + b) \\ & + M_2^{k_j}(z^{k_j} - z^{k_j+1}) \Big). \end{aligned} \tag{3. 29}$$

By denoting for all $j \geq 0$

$$\begin{aligned} & v^j := z^{k_j+1}, u^j := y^{k_j}, \\ & w^j := \nabla h_2(z^{k_j+1}) - \nabla h_2(z^{k_j}) + c_{k_j}B^*(-Ax^{k_j+1} - Bz^{k_j+1} + b) + M_2^{k_j}(z^{k_j} - z^{k_j+1}), \end{aligned}$$

(3. 29) reads

$$v^j \in \partial(g+h_2)^*(B^*u^j + w^j) \; \forall j \geq 0.$$

According to (3. 28), we have $v^j \rightharpoonup \bar{z}, u^j \rightharpoonup \bar{y}$ as $j \to +\infty$ thus, by taking into account (3. 25), $Bv^j \to B\bar{z} = Bz^*$ as $j \to +\infty$. Considering the Lipschitz continuity of ∇h_2, (3. 26), (3. 28) and (3. 13), one can easily see that $w^j \to 0$ as $j \to +\infty$. Due to the monotonicity of the subdifferential, we have that for all (u, v) in the graph of $\partial(g+h_2)^*$ and for all $j \geq 0$

$$\langle v^j - v, B^*u^j + w^j - u \rangle \geq 0,$$

which is equivalent to

$$\langle Bv^j - Bv, u^j \rangle + \langle v^j - v, w^j - u \rangle \geq 0.$$

For $j \to +\infty$ we obtain according to Proposition 2.2

$$\langle B\bar{z} - Bv, \bar{y} \rangle + \langle \bar{z} - v, -u \rangle \geq 0 \;\; \forall (u, v) \text{ in the graph of } \partial(g+h_2)^*,$$

which is the same as

$$\langle \bar{z} - v, B^*\bar{y} - u \rangle \geq 0 \;\; \forall (u, v) \text{ in the graph of } \partial(g+h_2)^*.$$

The maximal monotonicity of the convex subdifferential of $(g+h_2)^*$ ensures that $\bar{z} \in \partial(g+h_2)^*(B^*\bar{y})$, which is the same as $B^*\bar{y} \in \partial(g+h_2)(\bar{z})$. In other words, $B^*\bar{y} - \nabla h_2(\bar{z}) \in \partial g(\bar{z})$. Finally, by combining (3. 13) and (3. 26), the equality $Ax^* + B\bar{z} = b$ follows. In conclusion, $(x^*, \bar{z}, \bar{y})$ is a saddle point of the Lagrangian L.

In the following, we show that sequence $(x^k, z^k, y^k)_{k\geq 0}$ converges weakly. To this end, we consider two sequential cluster points (x^*, z_1, y_1) and (x^*, z_2, y_2). Consequently, according to Proposition 2.1, there exists $(k_s)_{s\geq 0}$, $k_s \to +\infty$ as $s \to +\infty$ such that the subsequence $(x^{k_s}, z^{k_s}, y^{k_s})_{s\geq 0}$ converges weakly to (x^*, z_1, y_1) as $s \to +\infty$. Furthermore, there exists $(k_t)_{t\geq 0}$, $k_t \to +\infty$ as $t \to +\infty$ such that a subsequence $(x^{k_t}, z^{k_t}, y^{k_t})_{t\geq 0}$ converges weakly to (x^*, z_2, y_2) as $t \to +\infty$. As seen before, (x^*, z_1, y_1) and (x^*, z_2, y_2) are both saddle points of the Lagrangian L.

From (3. 27), which is fulfilled for every saddle point of the Lagrangian L, we obtain

$$\exists \lim_{k\to+\infty}(\|y^k-y_1\|^2-\|y^k-y_2\|^2+c_k\|z^k-z_1\|^2_{M_2^k}-c_k\|z^k-z_2\|^2_{M_2^k}) =: T. \tag{3. 30}$$

For all $k\geq 0$ we have

$$\begin{aligned}\|y^k-y_1\|^2-&\|y^k-y_2\|^2+c_k\|z^k-z_1\|^2_{M_2^k}-c_k\|z^k-z_2\|^2_{M_2^k}\\ &=\|y_2-y_1\|^2+2\langle y_k-y_2,y_2-y_1\rangle+c_k\|z_2-z_1\|^2_{M_2^k}+2c_k\langle z_k-z_2,z_2-z_1\rangle_{M_2^k}.\end{aligned}$$

Since $M_2^k\geq\left(\alpha+\frac{L_2}{2}\right)\mathrm{Id}$, for all $k\geq 0$, and $(M_2^k)_{k\geq 0}$ is a nonincreasing sequence of symmetric operators in the sense of the Loewner partial ordering, there exists a symmetric operator $M\geq\left(\alpha+\frac{L_2}{2}\right)\mathrm{Id}$ such that $(M_2^k)_{k\geq 0}$ converges pointwise to M with respect to the strong topology of $\mathcal{G}$ as $k\to+\infty$ (see Proposition 2.18). Furthermore, let $c:=\lim_{k\to+\infty}c_k>0$. Taking the limits in (3. 30) along the subsequences $(k_s)_{s\geq 0}$ and $(k_t)_{t\geq 0}$, it yields

$$T=-\|y_2-y_1\|^2-c\|z_2-z_1\|^2_M=\|y_2-y_1\|^2+c\|z_2-z_1\|^2_M,$$

thus

$$\|y_2-y_1\|^2+c\|z_2-z_1\|^2_M=0.$$

It follows that $y_1=y_2$ and $z_1=z_2$, thus $(x^k,z^k,y^k)_{k\geq 0}$ converges weakly to a saddle point of the Lagrangian L.

Assume now that condition (ii) holds, namely, that there exists $\beta>0$ such that $B^*B\in\mathcal{P}_\beta(\mathcal{H})$. Then, it holds that $\beta\|z_k-z^*\|^2\leq\|Bz_k-Bz^*\|^2$ for $k\geq 0$ and it follows from (3. 25) and (3. 26) that

$$x^k\to x^*,\quad z^k\to z^*,\quad y^k-y^{k+1}\to 0\ (k\to+\infty). \tag{3. 31}$$

The remainder of the proof follows in analogy to the one given under assumption (i). The only difference is when we show that the sequence $(x^k,z^k,y^k)_{k\geq 0}$ converges weakly. For this, we have the two sequential cluster point (x^*,z^*,y_1) and (x^*,z^*,y_2), so that (3. 30) reduces to

$$\exists \lim_{k\to+\infty}(\|y^k-y_1\|^2-\|y^k-y_2\|^2) := T. \tag{3. 32}$$

Then, we can use the same argumentation as above, concerning y_1 and y_2. □

Remark 3.10. If $h_1=0$ and $h_2=0$, and $M_1^k=0$ and $M_2^k=0$ for all $k\geq 0$, then the Proximal AMA method becomes the AMA method, as it has been proposed by Tseng in [110]. According to Theorem 3.9 (for $L_1=L_2=0$), the generated sequence converges weakly to a saddle point of the Lagrangian, if there exists $\beta>0$ such that $B^*B\in\mathcal{P}_\beta(\mathcal{G})$. In finite-dimensional spaces, this condition reduces to the assumption that B is injective.

3.4 Numerical experiments

In this section, we compare the numerical performances of AMA and Proximal AMA on two applications in image processing and machine learning. The numerical experiments were performed on a computer with an Intel Core i5-3470 CPU and 8 GB DDR3 RAM. The first application is an extension of the approach considered in [[69], section 4.1.1] but without the blur operator A, the application to machine learning was also observed in [[69], section 4.4]. We used the source codes as a basis for our own programs to compare the performance of Proximal AMA with that of AMA.

3.4.1 Image denoising and deblurring

We addressed an image denoising and deblurring problem formulated as a nonsmooth convex optimization problem (see [41], [69] and [100]).

$$\inf_{x\in\mathbb{R}^n}\left\{\frac{1}{2}\|Ax-b\|^2+\lambda\mathrm{TV}(x)\right\}, \tag{3. 33}$$

where $A\in\mathbb{R}^{n\times n}$ represents a blur operator, $b\in\mathbb{R}^n$ is a given blurred and noisy image, $\lambda>0$ is a regularization parameter and $\mathrm{TV}:\mathbb{R}^n\to\mathbb{R}$ is a discrete total variation functional. The vector $x\in\mathbb{R}^n$ is the vectorized image $X\in\mathbb{R}^{M\times N}$, where $n=MN$ and $x_{i,j}:=X_{i,j}$ stands for the normalized value of the pixel in the i-th row and the j-th column, for $1\le i\le M, 1\le j\le N$.

Two choices have been considered for the discrete total variation, namely, the isotropic total variation $\mathrm{TV}_{\mathrm{iso}}:\mathbb{R}^n\to\mathbb{R}$,

$$\mathrm{TV}_{\mathrm{iso}}(x)=\sum_{i=1}^{M-1}\sum_{j=1}^{N-1}\sqrt{(x_{i+1,j}-x_{i,j})^2+(x_{i,j+1}-x_{i,j})^2}+\sum_{i=1}^{M-1}|x_{i+1,N}-x_{i,j}|+\sum_{j=1}^{N-1}|x_{M,j+1}-x_{M,j}|,$$

and the anisotropic total variation $\mathrm{TV}_{\mathrm{aniso}}:\mathbb{R}^n\to\mathbb{R}$,

$$\mathrm{TV}_{\mathrm{aniso}}(x)=\sum_{i=1}^{M-1}\sum_{j=1}^{N-1}|x_{i+1,j}-x_{i,j}|+|x_{i,j+1}-x_{i,j}|+\sum_{i=1}^{M-1}|x_{i+1,N}-x_{i,j}|+\sum_{j=1}^{N-1}|x_{M,j+1}-x_{M,j}|.$$

Consider the linear operator $L:\mathbb{R}^n\to\mathbb{R}^n\times\mathbb{R}^n, x_{i,j}\mapsto\big(L_1x_{i,j},L_2x_{i,j}\big)$, where

$$L_1x_{i,j}=\begin{cases}x_{i+1,j}-x_{i,j}, & \text{if } i<M\\ 0, & \text{if } i=M\end{cases}\ \text{and}\ L_2x_{i,j}=\begin{cases}x_{i,j+1}-x_{i,j}, & \text{if } j<N\\ 0, & \text{if } j=N\end{cases}$$

One can easily see that $\|L\|^2\le 8$. The optimization problem (3. 33) can be written as

$$\inf_{x\in\mathbb{R}^n}\left\{f(Ax)+g(Lx)\right\}, \tag{3. 34}$$

where $f:\mathbb{R}^n\to\mathbb{R}, f(x)=\frac{1}{2}\|x-b\|^2$, and $g:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$ is defined by $g(y,z)=\lambda\|(y,z)\|_1$ for the anisotropic total variation, and by
$g(y,z)=\lambda\|(y,z)\|_\times:=\lambda\sum_{i=1}^M\sum_{j=1}^N\sqrt{y_{i,j}^2+z_{i,j}^2}$ for the isotropic total variation.

We solved the Fenchel dual problem of (3. 34) by AMA and Proximal AMA and determined in this way an optimal solution of the primal problem, too. The reason for this strategy was that the Fenchel dual problem of (3. 34) is a convex optimization problem with two-block separable linear constraints and objective function.

Indeed, the Fenchel dual problem of (3. 34) reads (see [20] and [34])

$$\inf_{p\in\mathbb{R}^n,q\in\mathbb{R}^n\times\mathbb{R}^n}\left\{f^*(p)+g^*(q)\right\},\ \text{s.t. } A^*p+L^*q=0. \tag{3. 35}$$

Since f and g have full domains, strong duality for (3. 34)-(3. 35) holds.

As $f^*(p)=\frac{1}{2}\|p\|^2+\langle p,b\rangle$ for all $p\in\mathbb{R}^n$, f^* is 1-strongly convex. We chose $M_1^k=0$ and $M_2^k=\frac{1}{\sigma_k}\mathrm{I}-c_kL^*L$ (see Remark 3.8) and obtained for Proximal AMA the iterative scheme which reads for every $k\ge 0$:

$$\begin{aligned}p^{k+1}&=Ax^k-b\\ q^{k+1}&=\mathrm{Prox}_{\sigma_kg^*}\left(q^k+\sigma_kc_kL(-A^*p^{k+1}-L^*q^k)+\sigma_kL(x^k)\right)\\ x^{k+1}&=x^k+c_k(-A^*p^{k+1}-L^*q^{k+1}).\end{aligned}$$

In the case of the anisotropic total variation, the conjugate of g is the indicator function of the set $[-\lambda,\lambda]^n \times [-\lambda,\lambda]^n$, thus $\text{Prox}_{\sigma_k g^*}$ is the projection operator $\mathcal{P}_{[-\lambda,\lambda]^n\times[-\lambda,\lambda]^n}$ on the set $[-\lambda,\lambda]^n \times [-\lambda,\lambda]^n$. The iterative scheme reads for all $k \geq 0$:

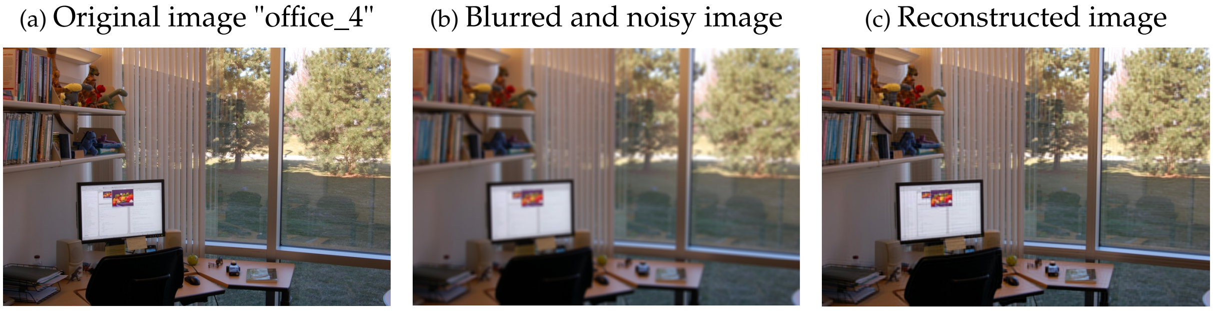
$$\begin{aligned} p^{k+1} &= Ax^k - b \\ (q_1^{k+1}, q_2^{k+1}) &= \mathcal{P}_{[-\lambda,\lambda]^n\times[-\lambda,\lambda]^n}\left((q_1^k, q_2^k) + c_k\sigma_k(-LA^*p^{k+1} - LL^*(q_1^k, q_2^k)) + \sigma_k Lx^k\right) \\ x^{k+1} &= x^k + c_k\left(-A^*p^{k+1} - L^*(q_1^{k+1}, q_2^{k+1})\right). \end{aligned}$$

In the case of the isotropic total variation, the conjugate of g is the indicator function of the set $S := \left\{(v,w) \in \mathbb{R}^n \times \mathbb{R}^n : \max_{1\leq i\leq n} \sqrt{v_i^2 + w_i^2} \leq \lambda\right\}$, thus $\text{Prox}_{\sigma_k g^*}$ is the projection operator $\mathcal{P}_S : \mathbb{R}^n \times \mathbb{R}^n \to S$ on S, defined as

$$(v_i, w_i) \mapsto \lambda \frac{(v_i, w_i)}{\max\left\{\lambda, \sqrt{v_i^2 + w_i^2}\right\}}, \quad i = 1, ..., n.$$

The iterative scheme reads for all $k \geq 0$:

$$\begin{aligned} p^{k+1} &= Ax^k - b \\ (q_1^{k+1}, q_2^{k+1}) &= P_S\left((q_1^k, q_2^k) + c_k\sigma_k(-LA^*p^{k+1} - LL^*(q_1^k, q_2^k)) + \sigma_k Lx^k\right) \\ x^{k+1} &= x^k + c_k\left(-A^*p^{k+1} - L^*(q_1^{k+1}, q_2^{k+1})\right). \end{aligned}$$


(a) Original image "office_4"

(b) Blurred and noisy image

(c) Reconstructed image

Figure 3.1: The original image, the blurred and noisy image and the reconstructed image after 50 seconds cpu time.

We compared the Proximal AMA method with Tseng's AMA method. While in Proximal AMA a closed formula is available for the computation of $(q_1^{k+1}, q_2^{k+1})_{k\geq 0}$, in AMA we solved the resulting optimization subproblem

$$(q_1^{k+1}, q_2^{k+1}) = \underset{q_1,q_2}{\text{argmin}} \left\{ g^*(q_1, q_2) - \langle x^{k+1}, L^*(q_1, q_2)\rangle + \frac{1}{2}c_k \|A^*p^{k+1} + L^*(q_1, q_2)\|^2 \right\}$$

in every iteration $k \geq 0$ by making some steps of the FISTA method [24].

We used in our experiments a Gaussian blur of size 9×9 and standard deviation 4, which led to an operator A with $\|A\|^2 = 1$ and $A^* = A$. Furthermore, we added Gaussian white noise with standard deviation 10^{-3}. We used for both algorithms a constant sequence of stepsizes $c_k = 2 - 10^{-7}$ for all $k \geq 0$. One can notice that $(c_k)_{k\geq 0}$ fulfills (3. 14). For Proximal AMA we

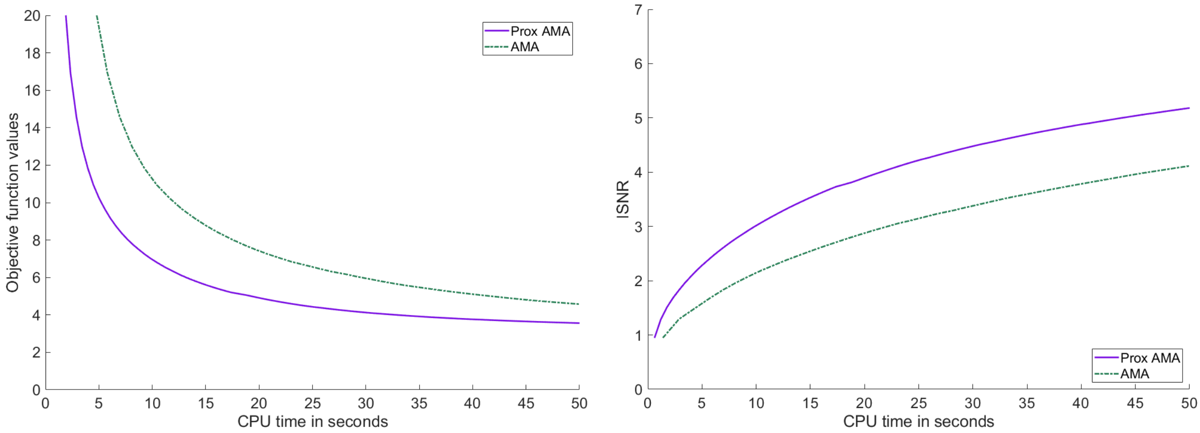


Figure 3.2: The objective function values and the ISNR values for the anisotropic TV and $\lambda = 5 \cdot 10^{-5}$.

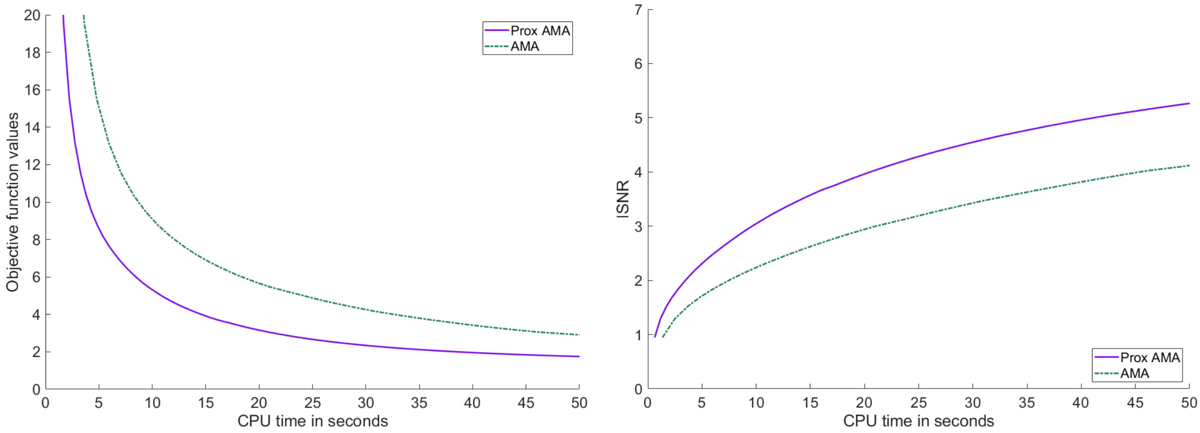


Figure 3.3: The objective function values and the ISNR values for the anisotropic TV and $\lambda = 10^{-5}$.

considered $\sigma_k = \frac{1}{8.00001 \cdot c_k}$ for all $k \geq 0$, which ensured that every matrix $M_2^k = \frac{1}{\sigma_k}\mathrm{I} - c_k L^* L$ is positively definite for all $k \geq 0$. This is actually the case, if $\sigma_k c_k \|L\|^2 < 1$ for all $k \geq 0$. In other words, assumption (i) in Theorem 3.9 was verified.

In the Figures 3.2 - 3.5, we show how Proximal AMA and AMA perform when reconstructing the blurred and noisy colored MATLAB test image "office_ 4" of 600×903 pixels for different choices for the regularization parameter λ and by considering both the anisotropic and isotropic total variation as regularization functionals. In all considered instances that Proximal AMA outperformed AMA from the point of view of both the convergence behavior of the sequence of the function values and of the sequence of ISNR (Improvement in Signal-to-Noise Ratio) values, which is defined at the iteration $k \in \mathbb{N}$ as

$$\mathrm{ISNR}_k = 10 \log_{10} \frac{\|x - b\|^2}{\|x - x_k\|^2} \quad ,$$

where x is the original, b the observed blurred and noisy image, and x_k is reconstructed image (see [47]).

An explanation could be that the number of iterations Proximal AMA makes in a certain amount of time is more than double the number of outer iterations performed by AMA.

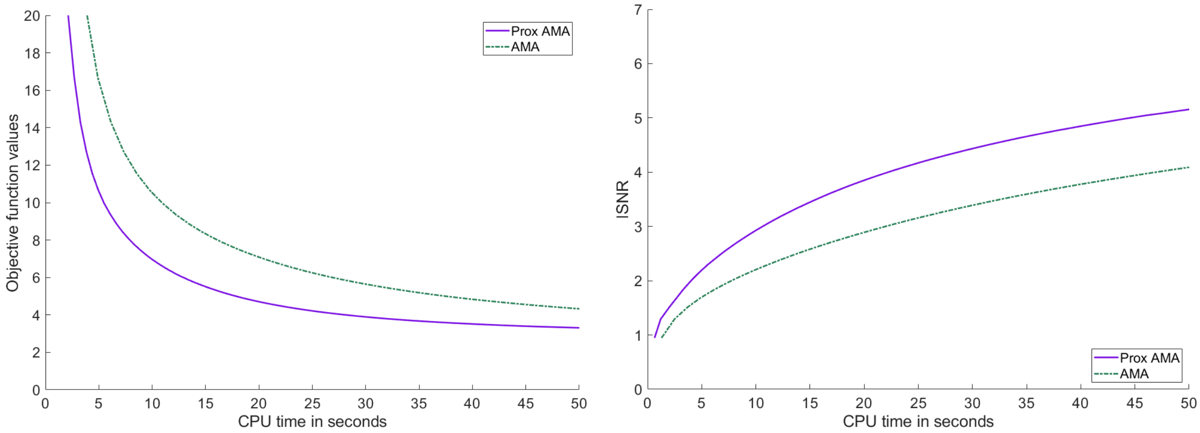


Figure 3.4: The objective function values and the ISNR values for the isotropic TV and $\lambda = 5 \cdot 10^{-5}$.

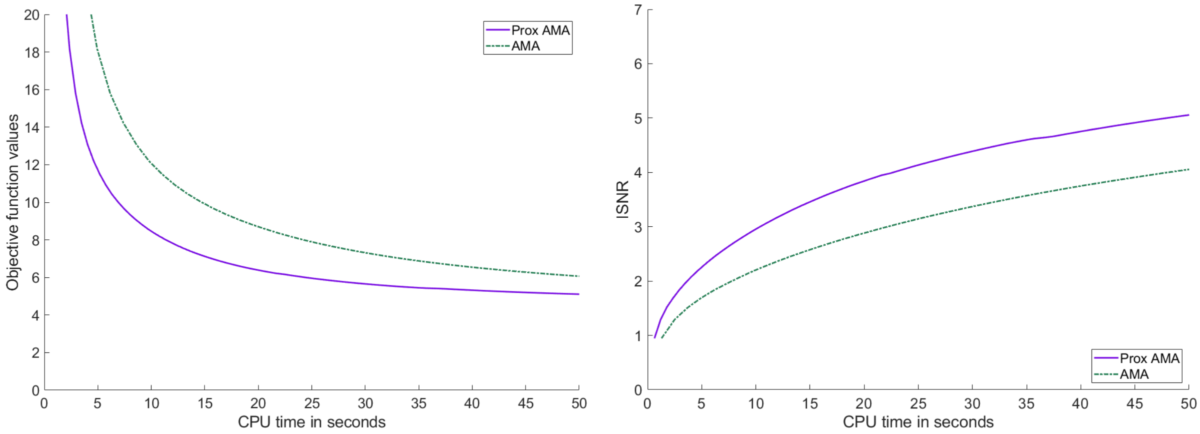


Figure 3.5: The objective function values and the ISNR values for the isotropic TV and $\lambda = 10^{-4}$.

3.4.2 Kernel based machine learning

In this subsection we will describe the numerical experiments we carried out in the context of classifying images via support vector machines. For more information regarding this nonlinear SVM model we refer the reader to Appendix A.2.

The given data set consisting of 5570 training images and 1850 test images of size 28×28 was taken from `http://www.cs.nyu.edu/~roweis/data.html`. The problem we considered was to determine a decision function based on a pool of handwritten digits showing either the number five or the number six, labeled by $+1$ and -1, respectively (see Figure 3.6). To evaluate the quality of the decision function we computed the percentage of misclassified images of the test data set.

In order to describe the approach we used, we denote by

$$\mathcal{Z} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \subseteq \mathbb{R}^d \times \{+1, -1\},$$

the given training data set. The decision functional `f` was assumed to be an element of the Reproducing Kernel Hilbert Space (RHKS) $\mathcal{H}_\kappa$, induced by the symmetric and finitely positive

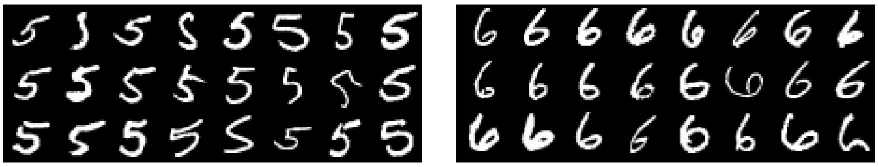


Figure 3.6: A sample of images belonging to the classes $+1$ and -1, respectively.

definite Gaussian kernel function

$$\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R},\ \kappa(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right).$$

By $K \in \mathbb{R}^{n\times n}$ we denoted the Gram matrix with respect to the training data set $\mathcal{Z}$, namely, the symmetric and positive definite matrix with entries $K_{ij} = \kappa(X_i, X_j)$ for $i,j = 1,\ldots,n$. To penalize the deviation between the predicted value $\mathtt{f}(x)$ and the true value $y \in \{+1,-1\}$ we used the hinge loss functional $(x,y) \mapsto \max\{1-xy, 0\}$. The decision function $\mathtt{f}$, which we want to obtain, is the optimal solution of

$$\min_{\mathtt{f}\in\mathcal{H}_\kappa} \frac{1}{2}\|\mathtt{f}\|^2_{\mathcal{H}_\kappa} + C\sum_{i=1}^{n}\max\{1 - Y_i\mathtt{f}(X_i), 0\}. \tag{3. 36}$$

Here, $C > 0$ denotes the regularization parameter controlling the tradeoff between the loss function and the regularization term.

According to the Representer Theorem, the decision function $\mathtt{f}$ can be expressed as a kernel expansion in terms of the training data, in other words $\mathtt{f}(\cdot) = \sum_{i=1}^{n} x_i\kappa(\cdot, X_i)$, where $x = (x_1,\ldots,x_n) \in \mathbb{R}^n$ is the optimal solution of the optimization problem

$$\min_{x\in\mathbb{R}^n}\left\{\frac{1}{2}x^TKx + C\sum_{i=1}^{n}\max\{1-(Kx)_iY_i, 0\}\right\}. \tag{3. 37}$$

Hence, in order to determine the decision function we solved the convex optimization problem (3. 37), which can be written as

$$\min_{x\in\mathbb{R}^n}\{f(x) + g(Kx)\}$$

or, equivalently,

$$\min_{x\in\mathbb{R}^n, z\in\mathbb{R}^n}\{f(x)+g(z)\},\ \text{s.t. } Kx - z = 0$$

where $f : \mathbb{R}^n \to \mathbb{R}, f(x) = \frac{1}{2}x^TKx$, and $g : \mathbb{R}^n \to \mathbb{R}$ is defined by $g(z) = C\sum_{i=1}^{n}\max\{1 - z_iY_i, 0\}$.

Since the Gram matrix K is positively definite, the function f is $\lambda_{\min}(K)$-strongly convex, where $\lambda_{\min}(K)$ denotes the minimal eigenvalue of K, and differentiable, and it holds $\nabla f(x) = Kx$ for all $x \in \mathbb{R}^n$. For an element of the form $p = (p_1, ..., p_n) \in \mathbb{R}^n$, it holds

$$g^*(p) = \begin{cases} \sum_{i=1}^{n} p_iY_i, & \text{if } p_iY_i \in [-C, 0], i = 1,\ldots,n, \\ +\infty, & \text{otherwise.} \end{cases}$$

Consequently, for every $\mu > 0$ and $p = (p_1, ..., p_n) \in \mathbb{R}^n$, it holds

$$\text{Prox}_{\mu g^*}(x) = \left(\mathcal{P}_{Y_1[-C,0]}(p_1 - \sigma Y_1), \ldots, \mathcal{P}_{Y_n[-C,0]}(p_n - \sigma Y_n)\right),$$

where $\mathcal{P}_{Y_i[-C,0]}$ denotes the projection operator on the set $Y_i[-C, 0], i = 1, ..., n$.

We implemented Proximal AMA for $M_2^k = 0$ for all $k \geq 0$ and different choices for the sequence $(M_1^k)_{k\geq 0}$. This resulted in an iterative scheme which reads for all $k \geq 0$:

$$x^{k+1} = \underset{x\in\mathbb{R}^n}{\text{argmin}} \left\{ f(x) - \langle p^k, Kx\rangle + \frac{1}{2}\|x - x^k\|^2_{M_1^k} \right\} = (K + M_1^k)^{-1}(Kp^k + M_1^k x^k) \tag{3. 38}$$

$$z^{k+1} = \text{Prox}_{\frac{1}{c_k}g}\left(Kx^{k+1} - \frac{1}{c^k}p^k\right) = \left(Kx^{k+1} - \frac{1}{c^k}p^k\right) - \frac{1}{c_k}\text{Prox}_{c_k g^*}\left(c_k K x^{k+1} - p^k\right) \tag{3. 39}$$

$$p^{k+1} = p^k + c_k(-Kx^{k+1} + z^{k+1}).$$

We would like to emphasize that the AMA method updates the sequence $(z^{k+1})_{k\geq 0}$ also via (3. 39), while the sequence $(x^{k+1})_{k\geq 0}$, as $M_1^k = 0$, is updated via $x^{k+1} = p^k$ for all $k \geq 0$. However, it turned out that the Proximal AMA where $M_1^k = \tau_k K$, for $\tau_k > 0$ and all $k \geq 0$, performs better than the version with $M_1^k = 0$ for all $k \geq 0$, which actually corresponds to the AMA method. In this case (3. 38) becomes $x^{k+1} = \frac{1}{1+\tau_k}(p^k + \tau_k x^k)$ for all $k \geq 0$.

We used for both algorithms a constant sequence of stepsizes given by $c_k = 2 \cdot \frac{\lambda_{\min}(K)}{\|K\|^2} - 10^{-8}$ for all $k \geq 0$. The tables below show for $C = 1$ and different values of the kernel parameter σ that Proximal AMA outperforms AMA in what concerns the time and the number of iterates needed to achieve a certain value for a given fixed misclassification rate (which proved to be the best one among several obtained by varying C and σ) and for the RMSE (Root-Mean-Square-Error) for the sequence of primal iterates. The RMSE is defined at the iteration $k \in \mathbb{N}$ as

$$\text{RMSE}_k = \frac{\|x_k - x^*\|}{\sqrt{n}},$$

where $x_k \in \mathbb{R}^n$ is the iterate at iteration k and x^* is the unique optimizer.

Table 3.1: Performance evaluation of Proximal AMA (with $\tau_k = 10$ for all $k \geq 0$) and AMA for the classification problem with $C = 1$ and $\sigma = 0.2$. The entries refer to the CPU times in seconds and the number of iterations.

Algorithm	misclassification rate at 0.7027 %	RMSE $\leq 10^{-3}$
Proximal AMA	8.18s (145)	23.44s (416)
AMA	8.65s (153)	26.64s (474)

Table 3.2: Performance evaluation of Proximal AMA (with $\tau_k = 102$ for all $k \geq 0$) and AMA for the classification problem with $C = 1$ and $\sigma = 0.25$. The entries refer to the CPU times in seconds and the number of iterations.

Algorithm	misclassification rate at 0.7027 %	RMSE $\leq 10^{-3}$
Proximal AMA	141.78s (2448)	629.52s (10940)
AMA	147.99s (2574)	652.61s (11368)

Chapter 4

Dynamical system of Proximal AMA

In this chapter, we investigate a dynamical system which concerns the same optimization problem as in section 3.2 of the previous chapter. This chapter is based on the paper [29]. In addition, we have included some further exposition on certain concepts employed therein, in particular some background on Lyapunov analysis.

Since the seventies of the last century, the investigation of dynamical systems approaching monotone inclusions and optimization problems gained a lot of attention (see Brézis, Baillon and Bruck, Crandall and Pazy [44, 16, 45, 53]). This is due to their intrinsic importance in areas like differential equations and applied functional analysis, and also since they have been recognized as a valuable tool for deriving and investigating numerical schemes for optimization problems obtained by time discretization of the continuous dynamics. The dynamic approach to iterative methods in optimization can furnish deep insights into the expected behavior of the method and the techniques used in the continuous case can be adapted to obtain results for the discrete algorithm. We invite the reader to consult [97] and [54] for more insights into the relations between the continuous and discrete dynamics.

This research area continuously attracts the attention of the community. There are several works in the last years concerning dynamical systems, which have a connection to numerical algorithms. Motivated by the applications in optimization where nonsmooth functions are involved, many authors consider dynamical systems defined via proximal evaluations. Through explicit time discretization they transform into relaxed versions of proximal point algorithms. For example, in [1] Abbas and Attouch proposed a dynamical system which is a continuous version of the forward backward algorithm (we mention here also the works of Bolte [32] and Antipin [6]), in [18] an implicit forward-backward-forward dynamical system was introduced, and in [55] a dynamical system of Douglas-Rachford type was proposed. In [42] Tikhonov regularized dynamical systems of Krasnoselskiǐ-Mann type were investigated and even strong convergence of the trajectories towards the minimum norm solution of the underlying monotone inclusion problem were shown. Acceleration of the dynamics in terms of function values along the trajectories can be achieved by considering second order differential equations/inclusions where again resolvents and proximal operators are involved in the description of the systems (see for example [37] and the works of Attouch and his co-authors [10, 11, 13, 9], see also [12] for the discrete counterpart of [11]). This is a flourishing area in the continuous setting since the work of Su-Boyd-Candès [107], where a second-order ordinary differential equation was proposed as the limit of Nesterov's accelerated gradient method which involves inertial type schemes.

Let us underline that approaching optimization problems where compositions with linear operators are involved by means of differential equations/inclusions is relatively new in the literature (and this is the focus also in this chapter). We mention here [38] (which is related to continuous counterparts of primal-dual algorithms, Proximal ADMM and the linearized proximal method of multipliers) and also the contribution of Attouch [8] (related to some fast inertial Proximal ADMM schemes).

4.1 Dynamical system

The dynamical system we propose and investigate in this chapter is:

$$\begin{cases} \dot{x}(t) + x(t) \in (\partial f + M_1(t))^{-1} \left[M_1(t)x(t) + A^*y(t) - \nabla h_1(x(t)) \right] \\ \dot{z}(t) + z(t) \in (\partial g + c(t)B^*B + M_2(t))^{-1} \left[M_2(t)z(t) + B^*y(t) - c(t)B^*A(\dot{x}(t) + x(t)) \right. \\ \qquad\qquad \left. + c(t)B^*b - \nabla h_2(z(t)) \right] \\ \dot{y}(t) = c(t) \left(b - A(x(t) + \dot{x}(t)) - B(z(t) + \dot{z}(t)) \right) \\ x(0) = x^0 \in \mathcal{H}, z(0) = z^0 \in \mathcal{G}, y(0) = y^0 \in \mathcal{K}, \end{cases} \tag{4. 1}$$

where $c : [0 + \infty) \to (0 + \infty)$, $M_1 : [0, +\infty) \to S_+(\mathcal{H})$ and $M_2 : [0, +\infty) \to S_+(\mathcal{G})$.

In the next section, we will see that the dynamical system leads through explicit time discretization to the proximal AMA algorithm 3.6 and the AMA numerical scheme [110]. Furthermore, we underline the role of the operators M_1 and M_2, namely for a special choice of the linear maps M_1 and M_2 we obtain a dynamical system of primal-dual type which is a full splitting scheme. For this, we consider a numerical example in order to show how the parameters for these particular linear maps can be chosen and influence the convergence of the trajectories.

We continue with the existence and uniqueness of strong global solutions of the dynamical system proposed above. The study relies on classical semigroup theory, showing that the system corresponds in fact to a Cauchy-Lipschitz system in a product space. This is far from being trivial and requires several technical prerequisites which are described in detail.

The last section is devoted to the asymptotic analysis of the trajectories and the connection to the optimization problems (3. 5) and (3. 8). The analysis relies on Lyapunov theory where the derivation of an appropriate energy functional plays a central role. The way the Lyapunov functional is obtained is quite involved and technical issues have to be investigated in order to achieve this goal (see the proof of Theorem 4.22 and (4. 33)). Finally, we prove that the trajectories converge weakly to a saddle point of the Lagrangian L.

The analysis used in this chapter relies on similar tools considered in [38]. Let us underline some differences in comparison to [38]. First of all, our optimization problem (3. 5) has a different structure, with two linear operators involved in the constrained set. Secondly, our dynamical system is related to the Proximal AMA algorithm [28], the AMA numerical scheme [110] and primal dual-type algorithms obtained in [28]. The one in [38] is related to the Proximal ADMM [19],the classical ADMM and primal-dual type algorithms. Moreover, notice that in our case f is strongly convex which has an influence in the investigations performed here (and in particular the inclusion corresponding to f has a more tractable form). Additionally, in our analysis we have an additional parameter $c(t)$, which is time varying, and this makes the investigation more involved (taking variable $c(t)$ is motivated by [110], where the numerical scheme AMA also involves a variable parameter).

4.2 Solution concept, discretizations

We need the following definition before we specify what we mean by a solution of (4. 1). Let $\mathcal{B}$ be a Banach space. The real vector space

$$\mathcal{L}(\mathcal{H}) := \{A : \mathcal{H} \to \mathcal{H} : A \text{ is linear and continuous}\}$$

is endowed with the norm $\|A\| = \sup_{\|x\|\leq 1} \|Ax\|$.

Definition 4.1. The map $M : [0, +\infty) \to \mathcal{B}$ is said to be differentiable at $t_0 \in [0, +\infty)$ if the limit

$$\lim_{h\to 0} \frac{M(t_0 + h) - M(t_0)}{h}, \tag{4. 2}$$

taken with respect to the norm topology of $\mathcal{B}$, exists. When this is the case, we denote by $\dot{M}(t_0) \in \mathcal{B}$ the value of the limit.

Definition 4.2. A map $u : [0, +\infty) \to \mathcal{B}$ is called *absolutely continuous on a bounded interval* $[0, T], T > 0$ if for every $\varepsilon > 0$ there exists $\eta > 0$ such that for any $N \in \mathbb{N}$ and any finite family of intervals $I_k = (a_k, b_k)_{k=1}^{N}$ such that $I_k \subseteq [0, T]$ the following property holds:

$$\text{for any subfamily of disjoint intervals } I_j \text{ with } \sum_j |b_j - a_j| < \eta \text{ it holds}$$

$$\sum_j \|u(b_j) - u(a_j)\| < \varepsilon.$$

If u is absolutely continuous on every interval $[0, T]$, then u is called *locally absolutely continuous*.

Remark 4.3. 1. The integral in the following proposition is the Bochner integral, which is the vector-valued extension of the Lebesgue integral. For deeper insight in the theory of the Bochner integral, see [[73], page 13 ff].

2. Let $\mathcal{I} \subseteq [0, +\infty)$ be a (bounded or unbounded) interval and let $1 \leq p < +\infty$. We define $L^p(\mathcal{I}, \mathcal{B})$ as the linear space of all strongly measurable functions $u : \mathcal{I} \to \mathcal{B}$ such that

$$\int_{\mathcal{I}} \|u\|^p \mathrm{d}t < +\infty.$$

Furthermore, we define $L^\infty(\mathcal{I}, \mathcal{B})$ as the linear space of all strongly measurable functions $u : \mathcal{I} \to \mathcal{B}$ for which there exists a $r \geq 0$ such that $\lambda(\{t \in \mathcal{I} : f(t) > r\}) = 0$. Here, λ denotes the Lebesgue measure. Endowed with the norms, for $1 \leq p < +\infty$

$$\|u\|_{L^p(\mathcal{I},\mathcal{B})} := \left(\int_{\mathcal{I}} \|u\|^p \mathrm{d}t\right)^{\frac{1}{p}}$$

and

$$\|u\|_{L^\infty(\mathcal{I},\mathcal{B})} := \text{ess sup}_{t\geq 0}(\|u\|) := \inf\{r \geq 0 : \lambda(\{t \in \mathcal{I} : f(t) > r\}) = 0\}$$

the spaces $L^p(\mathcal{I}, \mathcal{B})$, $1 \leq p \leq +\infty$ are Banach spaces. We denote $\text{ess sup} \|u\|$ as the *essential supremum* of u.

3. In the following, $L^1_{loc}([0,+\infty),\mathcal{B})$ denotes the space of $f:[0,+\infty)\to\mathcal{B}$ which are locally integrable . If $\mathcal{B}=\mathbb{R}$ we write $L^1_{loc}([0,+\infty))$. A function f is locally integrable on $[0,+\infty)$ if its Bochner integrable on every compact subset of $[0,+\infty)$.

Proposition 4.4. *Let $T>0$, $\mathcal{H}$ be a Hilbert space and $u:[0,T]\to\mathcal{H}$. Then, the following statements are equivalent:*

(i) The function u is absolutely continuous.

(ii) There exists an integrable function $v:[0,T]\to\mathcal{H}$ such that

$$u(t)=u(0)+\int_0^t v(r)dr \quad \forall t\in[0,T].$$

Proof. This follows from Propositions 2.5.9 and Theorem 2.5.12 in [73]. □

Any function $u:[0,T]\to\mathcal{H}$ satisfying the statement (ii) in the preceding Proposition is differentiable almost everywhere and it holds that $\dot{u}=v$ almost everywhere where v is as in statement (ii).

Remark 4.5. Let $\mathcal{B}$ be a Banach space and $u:[0,T]\to\mathcal{B}$ be absolutely continuous. For arbitrary Banach spaces, this does no longer imply that statement (ii) from Proposition 4.4 holds for u. In particular, it may no longer be the case that u is differentiable almost everywhere. While it is true that an analogue of Proposition 4.4 continues to hold if $\mathcal{B}$ posesses the Radon-Nikodym property, this property is not guaranteed to hold for the Banach space $\mathcal{L}(\mathcal{H})$ which we shall consider in this section regarding the operators M_1 and M_2.

We are now ready to consider the following solution concept.

Definition 4.6. Let $u^0\in\mathcal{H}$ and $f:[0,+\infty)\times\mathcal{H}\to 2^{\mathcal{H}}$. The function $u:[0,+\infty)\to\mathcal{H}$ is called a strong global solution of the dynamical system

$$\begin{aligned}\dot{u}(t)&\in f(t,u(t))\\ u(0)&=u^0,\end{aligned}$$

if $u(t)$ is locally absolutely continuous and verifies this system for almost every $t\in[0,+\infty)$ with initial value u^0.

Remark 4.7. Let us consider a discretization of the considered dynamical system. The first two inclusions in (4. 1) can be written in an equivalent way as

$$0\in\partial f(\dot{x}(t)+x(t))+M_1(t)\dot{x}(t)-(A^*y(t)-\nabla h_1(x(t))), \tag{4. 3}$$

$$\begin{aligned}0\in{}&\partial g(\dot{z}(t)+z(t))+c(t)B^*B(\dot{z}(t)+z(t))+M_2(t)\dot{z}(t)\\ &-(B^*y(t)-c(t)B^*A(\dot{x}(t)+x(t))+c(t)B^*b-\nabla h_2(z(t))),\end{aligned} \tag{4. 4}$$

where $t\in[0,+\infty)$. Through explicit discretization with respect to the time variable t and constant step size $h_k=1$ (i.e. $x(t)\approx x^k$ and $\dot{x}(t)\approx x^{k+1}-x^k$), we obtain for all $k\geq 0$ the inclusions:

$$\begin{aligned}&0\in\partial f(x^{k+1})+M_1^k(x^{k+1}-x^k)-A^*y^k+\nabla h_1(x^k),\\ &0\in\partial g(z^{k+1})+c_kB^*B(z^{k+1})+M_2^k(z^{k+1}-z^k)-B^*y^k+c_kB^*A(x^{k+1})-c_kB^*b+\nabla h_2(z^k).\end{aligned}$$

Furthermore, using convex subdifferential calculus this can be written equivalently for all $k \geq 0$ as

$$0 \in \partial \left(f + \langle \cdot - x^k, \nabla h_1(x^k) \rangle - \langle y^k, A \cdot \rangle + \frac{1}{2} \| \cdot - x^k \|^2_{M_1^k} \right) (x^{k+1})$$

$$0 \in \partial \left(g + \langle \cdot - z^k, \nabla h_2(z^k) \rangle - \langle y^k, B \cdot \rangle + \frac{c_k}{2} \| Ax^{k+1} + B \cdot - b \|^2 + \frac{1}{2} \| \cdot - z^k \|^2_{M_2^k} \right) (z^{k+1})$$

Hence, the dynamical system (4. 1) provides through explicit time discretization the Proximal AMA algorithm 3.6:

Let $M_1^k \in \mathcal{S}_+(\mathcal{H})$ and $M_2^k \in \mathcal{S}_+(\mathcal{G})$. Choose $(x^0, z^0, y^0) \in \mathcal{H} \times \mathcal{G} \times \mathcal{K}$ and $(c_k)_{k \geq 0} > 0$. For all $k \geq 0$ generate the sequence $(x^k, z^k, y^k)_{k \geq 0}$ as follows:

$$\begin{cases} x^{k+1} = \operatorname{argmin}_{x \in \mathcal{H}} \left\{ f(x) - \langle y^k, Ax \rangle + \langle x - x^k, \nabla h_1(x^k) \rangle + \frac{1}{2} \| x - x^k \|^2_{M_1^k} \right\} \\ z^{k+1} \in \operatorname{argmin}_{z \in \mathcal{G}} \left\{ g(z) - \langle y^k, Bz \rangle + \frac{1}{2} c_k \| Ax^{k+1} + Bz - b \|^2 + \langle z - z^k, \nabla h_2(z^k) \rangle + \frac{1}{2} \| z - z^k \|^2_{M_2^k} \right\}, \\ y^{k+1} = y^k + c_k(b - Ax^{k+1} - Bz^{k+1}). \end{cases}$$

In the particular case $M_1^k = M_2^k = 0$ and $h_1 = h_2 = 0$, the numerical scheme is the AMA algorithm introduced by Tseng in [110].

Remark 4.8. Let us show now that an appropriate choice of M_2 leads (both in continuous and discrete case) to an implementable proximal step in the second inclusion. This is crucial for numerical results in applications, see also [19]. For every $t \in [0, +\infty)$, we define

$$M_2(t) = \frac{1}{\tau(t)} \operatorname{Id} - c(t) B^* B,$$

where $\tau(t) > 0$ and $\tau(t) c(t) \|B\|^2 \leq 1$.

Let $t \in [0, +\infty)$ be fixed. Then, $M_2(t)$ is positive semidefinite, and the second relation in the dynamical system (4. 1) becomes a proximal step. Indeed, under the given conditions, one can see that (4. 4) is equivalent to

$$\left(\frac{1}{\tau(t)} \operatorname{Id} - c(t) B^* B \right) z(t) + B^* y(t) - c(t) B^* A(\dot{x}(t) + x(t)) + c(t) B^* b - \nabla h_2(z(t)) \in \frac{1}{\tau(t)} \dot{z}(t) + \frac{1}{\tau(t)} z(t) + \partial g(\dot{z}(t) + z(t)).$$

It follows that

$$\dot{z}(t) + z(t) = (\operatorname{Id} + \tau(t) \partial g)^{-1} ((\operatorname{Id} - \tau(t) c(t) B^* B) z(t) + \tau(t) B^* y(t) - c(t) \tau(t) B^* A(\dot{x}(t) + x(t)) + c(t) \tau(t) B^* b - \tau(t) \nabla h_2(z(t))),$$

which is the same as

$$\dot{z}(t) + z(t) = \operatorname{Prox}_{\tau(t) g} ((\operatorname{Id} - \tau(t) c(t) B^* B) z(t) + \tau(t) B^* y(t) - c(t) \tau(t) B^* A(\dot{x}(t) + x(t)) + c(t) \tau(t) B^* b - \tau(t) \nabla h_2(z(t))).$$

If we choose furthermore $M_1(t) = 0$, our dynamical system (4. 1) can be written in this particular setting equivalently as

$$\begin{cases} \dot{x}(t) + x(t) \in (\partial f)^{-1}\left[A^*y(t) - \nabla h_1(x(t))\right] \\ \dot{z}(t) + z(t) = \operatorname{Prox}_{\tau(t)g}((\operatorname{Id} - \tau(t)c(t)B^*B)z(t) + \tau(t)B^*y(t) - c(t)\tau(t)B^*A(\dot{x}(t) + x(t)) \\ \qquad\qquad + c(t)\tau(t)B^*b - \tau(t)\nabla h_2(z(t))). \\ \dot{y}(t) = c(t)\left(b - A(x(t) + \dot{x}(t)) - B(z(t) + \dot{z}(t))\right) \\ x(0) = x^0 \in \mathcal{H}, z(0) = z^0 \in \mathcal{G}, y(0) = y^0 \in \mathcal{K}, \end{cases} \tag{4. 5}$$

where $c : [0 + \infty) \to (0 + \infty)$. This can be seen as the continuous counterpart with proximal step of the Proximal AMA scheme.

Remark 4.9. In this chapter, we will often use the following equivalent formulation of the dynamical system (4. 1). For $U(t) = (x(t), z(t), y(t))$, (4. 1) can be written as

$$\begin{cases} \dot{U}(t) = \Gamma(t, U(t)), \\ U(0) = (x^0, z^0, y^0), \end{cases}$$

where

$$\Gamma : [0, +\infty) \times \mathcal{H} \times \mathcal{G} \times \mathcal{K} \longrightarrow \mathcal{H} \times \mathcal{G} \times \mathcal{K}, \quad \Gamma(t, x, z, y) = (u, v, w),$$

is defined as

$$\begin{cases} u = u(t, x, z, y) = \underset{p \in \mathcal{H}}{\operatorname{argmin}} \left\{ F(t, p) + \frac{1}{2}\|p - (M_1(t)x + A^*y - \nabla h_1(x))\|^2 \right\} - x \\ v = v(t, x, z, y) \in \underset{q \in \mathcal{G}}{\operatorname{argmin}} \left\{ G(t, q) + \frac{c(t)}{2} \left\| q - \left(\frac{1}{c(t)} M_2(t)z + \frac{1}{c(t)} B^*y \right.\right.\right. \\ \qquad\qquad \left.\left.\left. - B^*A(u + x) + B^*b - \frac{1}{c(t)}\nabla h_2(z) \right) \right\|^2 \right\} - z \\ w = w(t, x, z, y) = c(t)(b - A(x + u) - B(z + v)) \end{cases} \tag{4. 6}$$

with

$$F : [0, +\infty) \times \mathcal{H} \to \overline{\mathbb{R}}, \quad F(t, p) = f(p) - \frac{1}{2}\|p\|^2 + \frac{1}{2}\|p\|^2_{M_1(t)}$$

and

$$G : [0, +\infty) \times \mathcal{G} \to \overline{\mathbb{R}}, \quad G(t, q) = g(q) + \frac{c(t)}{2}\left(\|Bq\|^2 - \|q\|^2\right) + \frac{1}{2}\|q\|^2_{M_2(t)}.$$

In the following, we will make the assumption that for every $(x, z, y) \in \mathcal{H} \times \mathcal{G} \times \mathcal{K}$, the function $\Gamma(\cdot, x, z, y)$ is strongly measurable. Giving precise conditions on when this occurs is a nontrivial problem since the component functions u, v are obtained as the solutions of optimization problems.

For a fixed $t \in [0, +\infty)$, the functions $F(t, \cdot)$ and $G(t, \cdot)$ are proper and lower semicontinuous. Since $M_1(t) \in S_+(\mathcal{H})$, f is strongly convex and

$$F(t,p) + \frac{1}{2}\|p-u\|^2 = f(p) + \frac{1}{2}\|p\|^2_{M_1(t)} - \langle p, u\rangle + \frac{1}{2}\|u\|^2,$$

the function

$$p \mapsto F(t,p) + \frac{1}{2}\|p-u\|^2$$

is proper, strongly convex, and lower semicontinuous for every $u \in \mathcal{H}$. Further, we have

$$G(t,q) + \frac{1}{2}\|q-v\|^2 = g(q) + \frac{1}{2}\|q\|^2_{M_2(t)+c(t)B^*B} - \langle q, v\rangle + \frac{1}{2}\|v\|^2,$$

so if the assumption

$$(\mathcal{P}) \quad \text{there exists } \beta > 0 \text{ such that } c(t)B^*B + M_2(t) \in \mathcal{P}_\beta(\mathcal{G}) \quad \forall t \in [0, +\infty)$$

holds, then, additionally,

$$q \mapsto G(t,q) + \frac{1}{2}\|q-v\|^2$$

is proper, strongly convex and lower semicontinuous for every $v \in \mathcal{G}$. Therefore, we have that in (4. 6) u and v are uniquely defined.

Example 4.10. We consider the following optimization problem

$$\inf_{x\in\mathbb{R}^2, z\in\mathbb{R}^2} \frac{1}{2}\|x-d\|^2 + \|z\|_1, \quad \text{s.t.} \quad Ax + Bz = 0 \tag{4. 7}$$

with

$$A = \frac{1}{\sqrt{8}}\begin{pmatrix} 2 & 1 \\ -2 & 1 \end{pmatrix} \quad , \quad B = \frac{1}{5}\begin{pmatrix} -3 & 0 \\ 4 & 0 \end{pmatrix} \quad \text{and} \quad d = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

which is problem (3. 5) with $\mathcal{H} = \mathcal{G} = \mathbb{R}^2$ and

$$\begin{aligned} f, g, h_1, h_2 &: \mathbb{R}^2 \to \mathbb{R}, \\ f(x) &= \frac{1}{2}\|x-d\|^2, \\ g(z) &= \|z\|_1, \\ h_1(x) &= h_2(z) = 0, \end{aligned}$$

for every $x \in \mathbb{R}^2$ and $z \in \mathbb{R}^2$. One can verify that (4. 7) has a unique optimal solution, which is $x^* = (0,0)$ and $z^* = (0,0)$. The Fenchel-Rockafellar dual problem of (4. 7) is

$$\sup_{y\in\mathbb{R}^2}\{-f^*(A^*y) - g^*(B^*y)\},$$

which is equivalent to

$$-\inf_{y\in\mathbb{R}^2}\{f^*(A^*y) + g^*(B^*y)\}.$$

and

$$-\inf_{\|B^*y\|_\infty \le 1}\{\frac{1}{2}\|A^*y\|^2 + \langle A^*y, d\rangle\}, \tag{4. 8}$$

where the unique optimal solution is $y^* = (-0.7071, 0.7071)$.
For

$$\begin{aligned} U(t) &= (x(t), z(t), y(t)), \\ M_1(t) &= 0, \\ M_2(t) &= \frac{1}{\tau(t)} \operatorname{Id} - c(t) B^* B, \end{aligned}$$

we can write the dynamical system for this problem (4. 1) similarly as in Remark 4.9 (see also (4. 5))

$$\begin{cases} \dot{U}(t) = \Gamma(t, U(t)) \\ U(0) = (x^0, z^0, y^0), \end{cases}$$

where

$$\Gamma : [0, +\infty) \times \mathcal{H} \times \mathcal{G} \times \mathcal{K} \longrightarrow \mathcal{H} \times \mathcal{G} \times \mathcal{K}, \quad \Gamma(t, x, z, y) = (u, v, w),$$

is defined as

$$\begin{cases} u = \underset{p \in \mathcal{H}}{\operatorname{argmin}} \left\{ f(p) - \frac{1}{2}\|p\|^2 + \frac{1}{2}\|p - A^* y\|^2 \right\} - x \\ \quad = A^* y + d - x \\ v = \operatorname{Prox}_{\tau(t) g} \left((\operatorname{Id} - \tau(t) c(t) B^* B) z + \tau(t) B^* y - c(t) \tau(t) B^* A(x + u) \right) - z \\ w = c(t)(-A(x + u) - B(z + v)). \end{cases}$$

We solved the dynamical system with the initial values $x^0 = (-8, 8), z^0 = (-8, 8)$ and $y^0 = (-8, 8)$ in the case when $c(t) > 0$ and $\tau(t) > 0$ and used the Matlab function `ode15s`. The source code is based on the code used in [38].
Note, that

$$\operatorname{Prox}_{\tau(t) g}(x) = x - \tau(t) \mathcal{P}_{[-1,1]^2} \left(\frac{1}{\tau(t)} x \right),$$

where $\mathcal{P}_C$ is the projection operator on a convex and closed set $C \subseteq \mathcal{H}$. To assure the convergence of the algorithm, we will prove later in Theorem 4.22 that it has to be fulfilled for an $\epsilon > 0$ that

$$c(t) < \frac{\sigma}{\|A\|^2} - \epsilon$$

for all t, where σ is the strong convexity parameter of $f(x)$ (here $\sigma = 1$), and that $c(t)$ is monotonically decreasing and Lipschitz continuous. For $c(t) = c$ constant, we can choose c such that

$$c < \frac{2\sigma}{\|A\|^2} - \epsilon.$$

Besides, it has to be fulfilled that $M_2(t)$ is monotonically decreasing, locally absolutely continuous, positive definite, and $\sup_{t \geq 0} \|\dot{M}_2(t)\| < +\infty$ (we are in setting 1. of Theorem 4.22, see also Corollary 4.24 and Remark 4.25).

To guarantee that $M_2(t)$ is positive definite, we have to choose $\tau(t)$ such that $\tau(t) c(t) \|B\|^2 < 1$. Since $\|A\|^2 = 1$ and $\|B\|^2 = 1$, we can choose $c(t) \in (\epsilon, 1 - \epsilon)$ and for c constant $c \in (\epsilon, 2 - \epsilon)$ and $\tau(t) c(t) < 1$. In Figure 4.1, we chose

$$c(t) = c_3(t) = \frac{1}{\sqrt{t + 1.1}} + 0.01$$

and considered for $\tau(t)c(t)$ three different choices, namely, 0.25, 0.5 and 0.99. In Figure 4.2 and 4.3, we chose $\tau(t)c(t) = 0.99$ and varied the parameter $c(t)$ with the three constant choices, namely, $c(t) = 0.25$, $c(t) = 0.5$ and $c(t) = 1.0$ (Figure 4.2) and three variable choices (Figure 4.3)

$$\begin{aligned} c_1(t) &= \frac{1}{t^2 + 1.1} + 0.01, \\ c_2(t) &= \frac{1}{t + 1.1} + 0.01, \\ c_3(t) &= \frac{1}{\sqrt{t + 1.1}} + 0.01. \end{aligned}$$

These parameters fulfill the conditions above.

For $c(t) = c_3(t)$, all three trajectories converge faster for a greater value for $\tau(t)c(t)$ (see Figure 4.1). We could observe this fact as well for the other choices of $c(t)$. In Figure 4.2, we can see for the convergence of the z-trajectory that the convergence against the first coordinate of the optimal solution (blue curve) is faster for greater constant values of $c(t)$ and the convergence against the second coordinate of the optimal solution (red curve) is faster for smaller constant values. There are no big differences concerning the convergence of the x- and y-trajectories, but we can observe that the convergence of the y-trajectory is for both coordinates faster for greater constant values. For the three constant values in Figure 4.2, we have for $c(t) = 0.5$ the best convergence result as a compromise of these facts. In Figure 4.3, we can observe the same facts, but for variable choices for $c(t)$. So, the convergence of the z-trajectory against the first coordinate of the optimal solution (blue curve) is faster for those variable $c(t)$ that are slower monotonically decreasing and the convergence against the second coordinate of the optimal solution (red curve) is faster for those variable $c(t)$ that are faster monotonically decreasing. The convergence of the x- and y-trajectories is for both coordinates faster for those variable $c(t)$ that are slower monotonically decreasing. For the three variable choices for $c(t)$ in Figure 4.3, we have for $c_3(t)$ the best convergence result as a compromise of these facts.

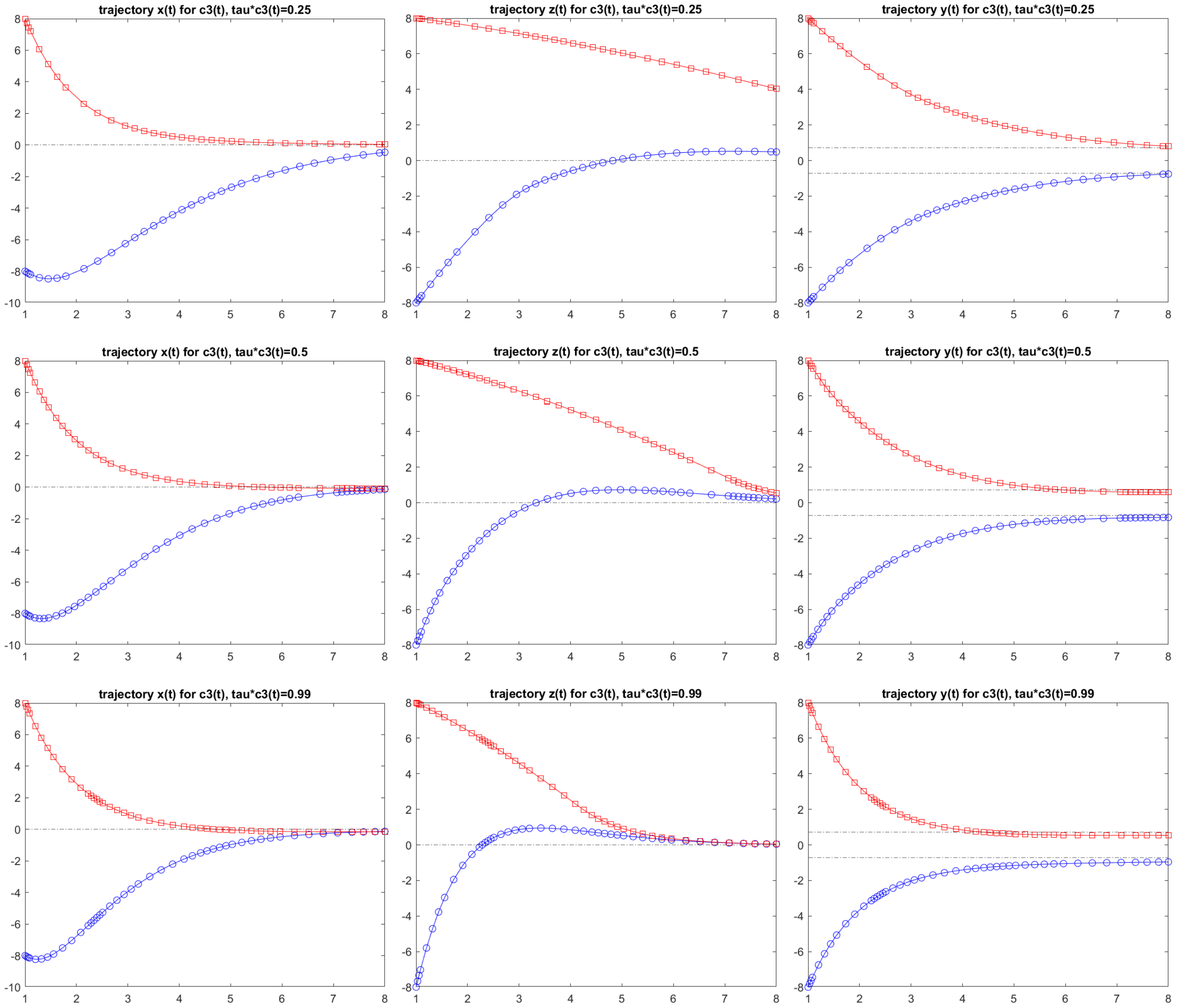


Figure 4.1: First and second column: the primal trajectories $x(t)$ and $z(t)$ converge to the primal optimal solution $(0, 0)$ for $c_3(t)$ and initial value $(-8, 8)$. Third column: the dual trajectory $y(t)$ converges to the dual optimal solution $(-0.7071, 0.7071)$ for $c_3(t)$ and initial value $(-8, 8)$.

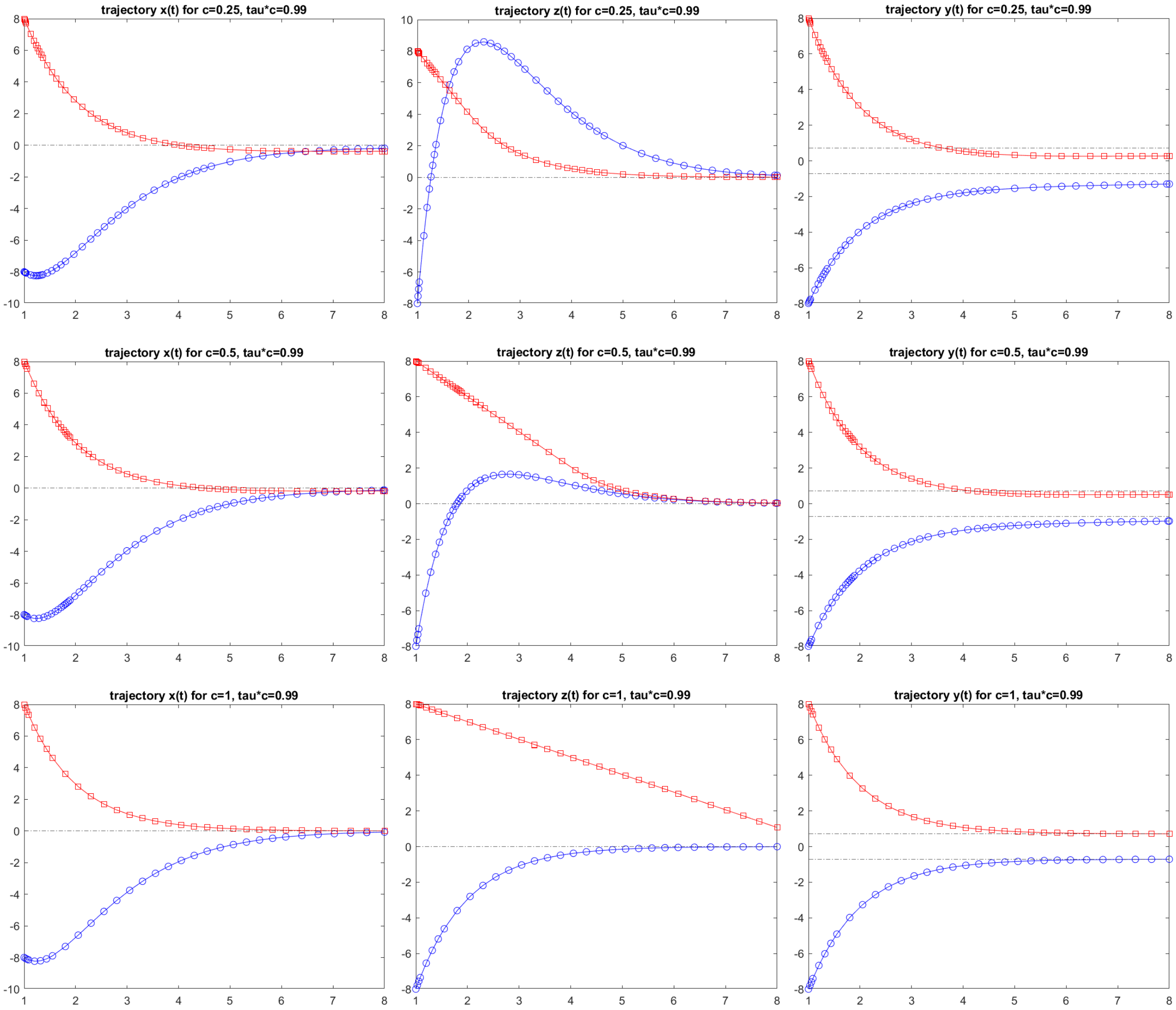
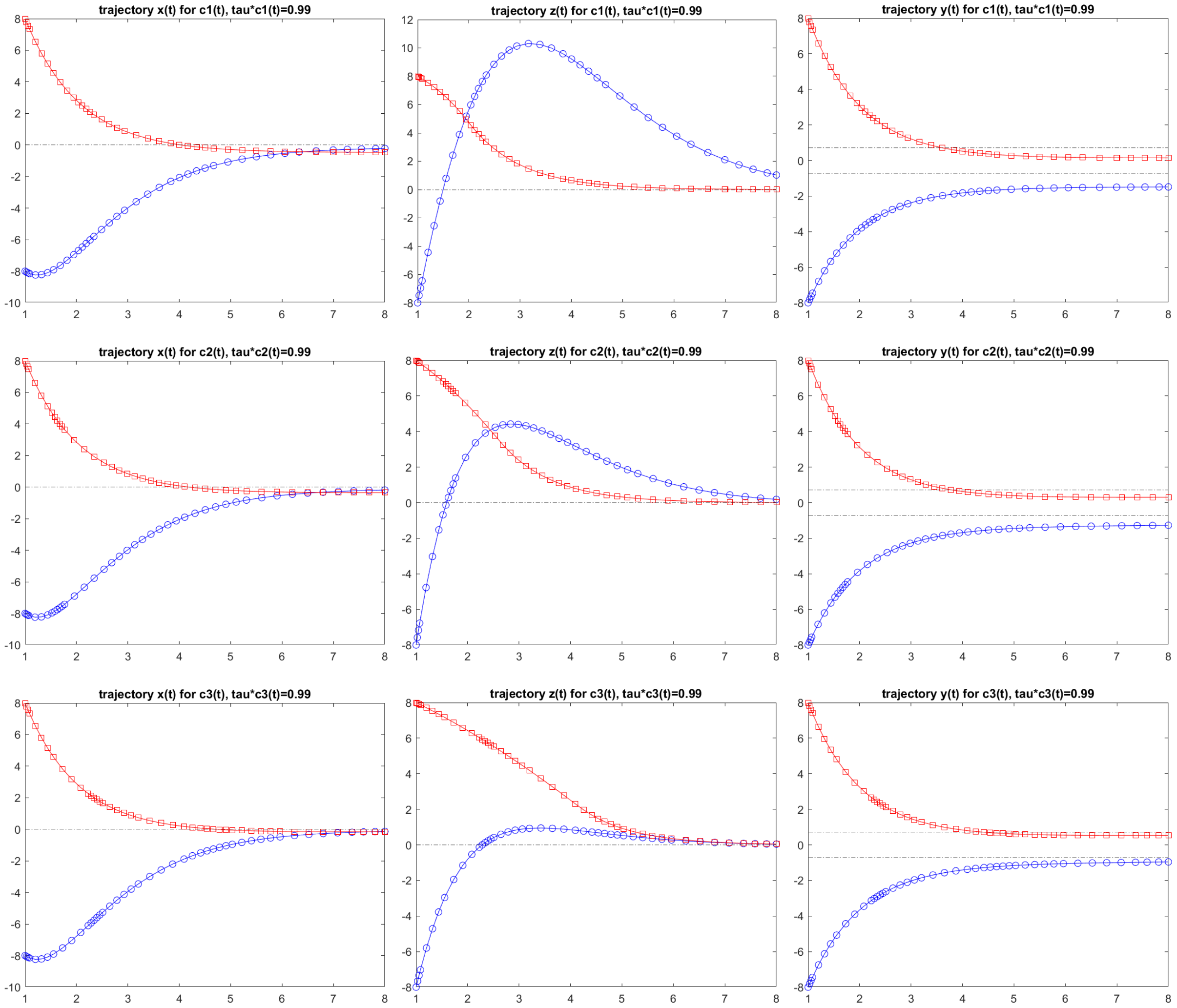


Figure 4.2: First and second column: the primal trajectories $x(t)$ and $z(t)$ converge to the primal optimal solution $(0,0)$ for constant $c(t)$ and initial value $(-8,8)$. Third column: the dual trajectory $y(t)$ converges to the dual optimal solution $(-0.7071, 0.7071)$ for constant $c(t)$ and initial value $(-8,8)$.


trajectory x(t) for c1(t), tau*c1(t)=0.99

trajectory z(t) for c1(t), tau*c1(t)=0.99

trajectory y(t) for c1(t), tau*c1(t)=0.99

trajectory x(t) for c2(t), tau*c2(t)=0.99

trajectory z(t) for c2(t), tau*c2(t)=0.99

trajectory y(t) for c2(t), tau*c2(t)=0.99

trajectory x(t) for c3(t), tau*c3(t)=0.99

trajectory z(t) for c3(t), tau*c3(t)=0.99

trajectory y(t) for c3(t), tau*c3(t)=0.99

Figure 4.3: First and second column: the primal trajectories $x(t)$ and $z(t)$ converge to the primal optimal solution $(0, 0)$ for variable $c(t)$ and initial value $(-8, 8)$. Third column: the dual trajectory $y(t)$ converges to the dual optimal solution $(-0.7071, 0.7071)$ for variable $c(t)$ and initial value $(-8, 8)$.

4.3 Existence and uniqueness of the trajectories

In this section, we will investigate the existence and uniqueness of the trajectories generated by the dynamical system (4. 1). For showing this, we will apply the following Theorem:

Theorem 4.11. *(Cauchy-Lipschitz-Picard Theorem) Let $\mathcal{H}$ be a Hilbert space and let the function $f : [0, +\infty) \times \mathcal{H} \to \mathcal{H}$ fulfill the following properties:*

(i) for all $x \in \mathcal{H}$ it holds $f(\cdot, x) \in L^1_{loc}([0, +\infty), \mathcal{H})$,

(ii) for almost all $t \geq 0$, $f(t, \cdot) : \mathcal{H} \to \mathcal{H}$ is continuous and there exists a function $L : [0, +\infty) \times [0, +\infty) \to [0, +\infty)$ such that it holds for all $x, y \in \mathcal{H}$:

$$\|f(t, x) - f(t, y)\| \leq L(t, \|x\| + \|y\|)\|x - y\|$$

with $L(\cdot, r) \in L^1_{loc}([0,+\infty)), \forall r \in [0,+\infty)$.

(iii) there exists a $P \in L^1_{loc}([0,+\infty))$ *such that for almost all* $t \geq 0$ *it holds for all* $x \in \mathcal{H}$:

$$\|f(t,x)\| \leq P(t)(1+\|x\|).$$

Then for all $u^0 \in \mathcal{H}$, *there exists a unique strong global solution* $u : [0,+\infty) \to \mathcal{H}$ *for the following dynamical system:*

$$\begin{aligned} \dot{u}(t) &= f(t,u(t)) \\ u(0) &= u^0. \end{aligned}$$

Proof. See [68, Proposition 6.2.1]. □

Remark 4.12. If in the property (ii) of the Theorem above the constant L depends only on t, then it follows

$$\begin{aligned} \|f(t,x)\| \leq \|f(t,x) - f(t,0)\| + \|f(t,0)\| &\leq L(t)\|x\| + \|f(t,0)\| \\ &\leq (L(t) + \|f(t,0)\|)\|x\| + L(t) + \|f(t,0)\| \\ &= (L(t) + \|f(t,0)\|)(1+\|x\|) \end{aligned}$$

and we can define $P(t) := (L(t) + \|f(t,0)\|) \in L^1_{loc}([0,+\infty))$. So, we have that (iii) follows from (ii).

First, we need several preparatory results in order to show that we are in the setting of the Cauchy-Lipschitz-Picard Theorem.

Lemma 4.13. *Let assumption* ($\mathcal{P}$) *hold true and* $t \in [0,+\infty)$. *Then, the operator*

$$K_t : \mathcal{H} \to \mathcal{H}, \quad K_t(u) = \underset{x \in \mathcal{H}}{\operatorname{argmin}} \left(F(t,x) + \frac{1}{2}\|x-u\|^2 \right)$$

is $\frac{1}{\sigma}$*-Lipschitz continuous and the operator*

$$J_t : \mathcal{G} \to \mathcal{G}, \quad J_t(v) = \underset{z \in \mathcal{G}}{\operatorname{argmin}} \left(G(t,z) + \frac{c(t)}{2}\|z-v\|^2 \right)$$

is $\frac{c(t)}{\beta}$*-Lipschitz continuous.*

Proof. Let $t \in [0,+\infty)$ be fixed. Then, we have, due to Fermat's rule 2.5,

$$0 \in \partial \left(f(\cdot) + \frac{1}{2}(\|\cdot\|^2_{M_1(t)} - \|\cdot\|^2) + \frac{1}{2}\|\cdot - u\|^2 \right)(K_t(u)),$$

which is equivalent to

$$0 \in \partial \left(f(\cdot) + \frac{1}{2}\|\cdot\|^2_{M_1(t)} - \langle\ \cdot\ , u\rangle + \frac{1}{2}\|u\|^2 \right)(K_t(u)).$$

Using Proposition 2.6 (ii), we obtain for all $u, v \in \mathcal{H}$

$$u \in \partial f(K_t(u)) + M_1(t)(K_t(u))$$

and

$$v \in \partial f(K_t(v)) + M_1(t)(K_t(v)).$$

Due to the σ-strong convexity of f and $M_1(t) \in S_+(\mathcal{H})$, it follows from Proposition 2.22 (ii) that $\partial f + M_1(t)$ is σ-strongly monotone and we get

$$\sigma\|K_t u - K_t v\|^2 \leq \langle u - v, K_t(u) - K_t(v)\rangle.$$

Using the Cauchy-Schwarz inequality, it follows

$$\|K_t u - K_t v\| \leq \frac{1}{\sigma}\|u - v\|,$$

which means that K_t is $\frac{1}{\sigma}$-Lipschitz continuous.

For $t \in [0, +\infty)$ fixed, we have, due to Fermat's rule 2.5,

$$0 \in \partial\left(g(\cdot) + \frac{c(t)}{2}(\|B\cdot\|^2 - \|\cdot\|^2) + \frac{1}{2}\|\cdot\|^2_{M_2(t)} + \frac{c(t)}{2}\|\cdot - u\|^2\right)(J_t(u)),$$

which is equivalent to

$$0 \in \partial\left(g(\cdot) + \frac{c(t)}{2}\|B\cdot\|^2 + \frac{1}{2}\|\cdot\|^2_{M_2(t)} - c(t)\langle\ \cdot\ , u\rangle + \frac{c(t)}{2}\|u\|^2\right)(J_t(u)).$$

Using Proposition 2.6 (ii), we obtain for all $u, v \in \mathcal{G}$

$$c(t)u \in \partial g(J_t(u)) + (c(t)B^*B + M_2(t))(J_t(u))$$

and

$$c(t)v \in \partial g(J_t(v)) + (c(t)B^*B + M_2(t))(J_t(v)).$$

Because of assumption $(\mathcal{P})$ and from Proposition 2.22 (ii), we have that $\partial g + c(t)B^*B + M_2(t)$ is β-strongly monotone and we get

$$\beta\|J_t u - J_t v\|^2 \leq c(t)\langle u - v, J_t(u) - J_t(v)\rangle.$$

Using the Cauchy-Schwarz inequality, it follows

$$\|J_t u - J_t v\| \leq \frac{c(t)}{\beta}\|u - v\|,$$

which means that J_t is $\frac{c(t)}{\beta}$-Lipschitz continuous. □

Lemma 4.14. *Let assumption $(\mathcal{P})$ hold true, $(x, z, y) \in \mathcal{H} \times \mathcal{G} \times \mathcal{K}$ and consider the maps* $R_{(x,z,y)} : [0, +\infty) \to \mathcal{H}$,

$$R_{(x,z,y)}(t) = \underset{u \in \mathcal{H}}{\operatorname{argmin}}\left\{F(t, u) + \frac{1}{2}\|u - (M_1(t)x + A^*y - \nabla h_1(x))\|^2\right\} - x,$$

$Q_{(x,z,y)} : [0, +\infty) \to \mathcal{G}$,

$$Q_{(x,z,y)}(t) = \underset{v \in \mathcal{G}}{\operatorname{argmin}}\left\{G(t, v) + \frac{c(t)}{2}\left\|v - \left(\frac{1}{c(t)}M_2(t)z + \frac{1}{c(t)}B^*y - B^*A(R_{(x,z,y)}(t) + x)\right.\right.\right.$$
$$\left.\left.\left. + B^*b - \frac{1}{c(t)}\nabla h_2(z)\right)\right\|^2\right\} - z,$$

and $P_{(x,z,y)} : [0,+\infty) \to \mathcal{K}$*,*

$$P_{(x,z,y)}(t) = c(t)(b - A(R_{(x,z,y)}(t) + x) - B(Q_{(x,z,y)}(t) + z)).$$

Then, the following holds for every $t, r \in [0,+\infty)$ *:*

$$\begin{aligned}
&\textit{(i)} \quad \|R_{(x,z,y)}(t) - R_{(x,z,y)}(r)\| \leq \frac{\|R_{(x,z,y)}(r)\|}{\sigma}\|M_1(t) - M_1(r)\| \\
&\textit{(ii)} \quad \|Q_{(x,z,y)}(t) - Q_{(x,z,y)}(r)\| \leq \frac{c(t)\|A\|\|B\|\|R_{(x,z,y)}(r)\|}{\sigma\beta}\|M_1(t) - M_1(r)\| \\
&\qquad + \frac{\|Q_{(x,z,y)}(r)\|}{\beta}\|M_2(t) - M_2(r)\| + \frac{\|P_{(x,z,y)}(r)\| \cdot \|B\|}{\beta c(r)}|c(t) - c(r)|.
\end{aligned}$$

Proof. Let $t, r \in [0,+\infty)$ be fixed.

(i) From the definition of $R_{(x,z,y)}$, Fermat's rule 2.5, and Proposition 2.6 (ii), we have

$$M_1(t)x + A^*y - \nabla h_1(x) \in \partial f(R_{(x,z,y)}(t) + x) + M_1(t)(R_{(x,z,y)}(t) + x) \tag{4. 9}$$

and

$$M_1(r)x + A^*y - \nabla h_1(x) \in \partial f(R_{(x,z,y)}(r) + x) + M_1(r)(R_{(x,z,y)}(r) + x).$$

If we add $M_1(t)(R_{(x,z,y)}(r) + x)$ on both sides of the relation above, we obtain

$$\begin{aligned}
M_1(t)(R_{(x,z,y)}(r) + x) - M_1(r)(R_{(x,z,y)}(r)) + A^*y - \nabla h_1(x) \in \\
\partial f(R_{(x,z,y)}(r) + x) + M_1(t)(R_{(x,z,y)}(r) + x).
\end{aligned} \tag{4. 10}$$

From (4. 9) and (4. 10), and using that $\partial f + M_1(t)$ is σ-strongly monotone according to Proposition 2.22 (ii), we have

$$\langle M_1(t)R_{(x,z,y)}(r) - M_1(r)R_{(x,z,y)}(r), R_{(x,z,y)}(r) - R_{(x,z,y)}(t)\rangle \geq \sigma\|R_{(x,z,y)}(r) - R_{(x,z,y)}(t)\|^2.$$

The result follows from the Cauchy-Schwarz inequality.

(ii) Using the definition of $Q_{(x,z,y)}$, Fermat's rule 2.5, and Proposition 2.6 (ii) we have

$$\begin{aligned}
M_2(t)z + B^*y - c(t)B^*A(R_{(x,z,y)}(t) + x) + c(t)B^*b - \nabla h_2(z) \in \\
\partial g(Q_{(x,z,y)}(t) + z) + (c(t)B^*B + M_2(t))(Q_{(x,z,y)}(t) + z)
\end{aligned} \tag{4. 11}$$

and

$$\begin{aligned}
M_2(r)z + B^*y - c(r)B^*A(R_{(x,z,y)}(r) + x) + c(r)B^*b - \nabla h_2(z) \in \\
\partial g(Q_{(x,z,y)}(r) + z) + (c(r)B^*B + M_2(r))(Q_{(x,z,y)}(r) + z).
\end{aligned}$$

If we add $(c(t) - c(r))B^*B(Q_{(x,z,y)}(r) + z) + M_2(t)(Q_{(x,z,y)}(r) + z)$ on both sides of the relation above, we obtain

$$\begin{aligned}
(c(t) - c(r))B^*B(Q_{(x,z,y)}(r) + z) - M_2(r)Q_{(x,z,y)}(r) + M_2(t)(Q_{(x,z,y)}(r) + z) + B^*y \\
-c(r)B^*A(R_{(x,z,y)}(r) + x) + c(r)B^*b - \nabla h_2(z) \in \\
\partial g(Q_{(x,z,y)}(r) + z) + (c(t)B^*B + M_2(t))(Q_{(x,z,y)}(r) + z).
\end{aligned} \tag{4. 12}$$

From (4. 11) and (4. 12), and using that $\partial g + c(t)B^*B + M_2(t)$ is β-strongly monotone, we have

$$\begin{aligned}\langle (c(t)-c(r))B^*B(Q_{(x,z,y)}(r)+z) + (M_2(t)-M_2(r))Q_{(x,z,y)}(r) + c(t)B^*AR_{(x,z,y)}(t)\\ -c(r)B^*AR_{(x,z,y)}(r) + (c(t)-c(r))B^*Ax - (c(t)-c(r))B^*b, Q_{(x,z,y)}(r) - Q_{(x,z,y)}(t)\rangle\\ \geq \beta\|Q_{(x,z,y)}(r) - Q_{(x,z,y)}(t)\|^2.\end{aligned}$$

Using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\|Q_{(x,z,y)}(r) - Q_{(x,z,y)}(t)\|\\ &\quad\leq \frac{1}{\beta}\|(c(t)-c(r))B^*B(Q_{(x,z,y)}(r)+z) + (M_2(t)-M_2(r))Q_{(x,z,y)}(r)\\ &\qquad + c(t)B^*AR_{(x,z,y)}(t) - c(r)B^*AR_{(x,z,y)}(r) + (c(t)-c(r))B^*Ax - (c(t)-c(r))B^*b\|\\ &\quad= \frac{1}{\beta}\|(c(t)-c(r))B^*BQ_{(x,z,y)}(r) + c(t)B^*AR_{(x,z,y)}(t) - c(r)B^*AR_{(x,z,y)}(r)\\ &\qquad + (c(t)-c(r))B^*(Ax+Bz-b) + (M_2(t)-M_2(r))Q_{(x,z,y)}(r)\|\end{aligned}$$

and because of the definition of $P_{(x,z,y)}$ and (i), it follows from the inequality above

$$\begin{aligned}&\|Q_{(x,z,y)}(r) - Q_{(x,z,y)}(t)\|\\ &\quad\leq \frac{1}{\beta}\|(c(t)-c(r))B^*BQ_{(x,z,y)}(r) + c(t)B^*AR_{(x,z,y)}(t) - c(r)B^*AR_{(x,z,y)}(r)\\ &\qquad + (c(t)-c(r))B^*(-BQ_{(x,z,y)}(r) - AR_{(x,z,y)}(r) - \frac{1}{c(r)}P_{(x,z,y)}(r))\\ &\qquad + (M_2(t)-M_2(r))Q_{(x,z,y)}(r)\|\\ &\quad= \frac{1}{\beta}\left\|c(t)B^*AR_{(x,z,y)}(t) - c(t)B^*AR_{(x,z,y)}(r) - \frac{(c(t)-c(r))}{c(r)}B^*P_{(x,z,y)}(r)\right.\\ &\qquad\left. +(M_2(t)-M_2(r))Q_{(x,z,y)}(r)\right\|\\ &\quad\leq \frac{1}{\beta}\left(c(t)\|A\|\|B\|\|R_{(x,z,y)}(t) - R_{(x,z,y)}(r)\| + \frac{|c(t)-c(r)|}{c(r)}\|B\|\|P_{(x,z,y)}(r)\|\right.\\ &\qquad\left. +\|M_2(t)-M_2(r)\|\|Q_{(x,z,y)}(r)\|\right).\end{aligned}$$

Finally, considering (i), we obtain from the inequality above

$$\begin{aligned}&\|Q_{(x,z,y)}(r) - Q_{(x,z,y)}(t)\|\\ &\quad\leq \frac{1}{\beta}\left(\frac{c(t)\|A\|\|B\|\|R_{(x,z,y)}(r)\|}{\sigma}\|M_1(t)-M_1(r)\| + \frac{|c(t)-c(r)|}{c(r)}\|B\|\|P_{(x,z,y)}(r)\|\right.\\ &\qquad\left. +\|M_2(t)-M_2(r)\|\|Q_{(x,z,y)}(r)\|\right).\end{aligned}$$

□

Having now all these estimations at our disposal, we are now ready to prove the existence and uniqueness of the trajectories.

Theorem 4.15. *Let assumption ($\mathcal{P}$) hold true and let $\|M_1\|, \|M_2\| \in L^1_{loc}([0,+\infty))$. Furthermore, we assume that $0 < \inf_{t\geq 0} c(t) \leq \sup_{t\geq 0} c(t) < +\infty$ and that for every $(x,z,y) \in \mathcal{H}\times\mathcal{G}\times\mathcal{K}$, the function $\Gamma(\cdot,x,z,y)$ is strongly measurable. Then for every initial value $(x^0,z^0,y^0) \to \mathcal{H}\times\mathcal{G}\times\mathcal{K}$, the dynamical system* (4. 1) *has a unique strong global solution $(x,z,y) : [0,+\infty) \to \mathcal{H}\times\mathcal{G}\times\mathcal{K}$.*

Proof. In the following, we use the equivalent formulation of the dynamical system described in Remark 4.9. We show the existence and uniqueness of a strong global solution, using the Cauchy-Lipschitz-Picard Theorem 4.11 to this end.

In the first part, we have to show that $\Gamma(t,\cdot,\cdot,\cdot)$ is $L(t)$-Lipschitz continuous for every $t \in [0,+\infty)$ and that the Lipschitz constant as a function of time fulfills $L(\cdot) \in L^1_{loc}([0,+\infty))$. In the second part, we will prove that $\Gamma(\cdot,x,z,y) \in L^1_{loc}([0,+\infty),\mathcal{H}\times\mathcal{G}\times\mathcal{K})$ for every $(x,z,y) \in \mathcal{H}\times\mathcal{G}\times\mathcal{K}$.

(1) For a fixed $t \in [0,+\infty)$ and for $(x,z,y),(\overline{x},\overline{z},\overline{y}) \in \mathcal{H}\times\mathcal{G}\times\mathcal{K}$, we have

$$\|\Gamma(t,x,z,y) - \Gamma(t,\overline{x},\overline{z},\overline{y})\| = \sqrt{\|u-\overline{u}\|^2 + \|v-\overline{v}\|^2 + \|w-\overline{w}\|^2},$$

where (taking into account the definition of K_t in Lemma 4.13)

$$\begin{aligned} u-\overline{u} &= \operatorname*{argmin}_{p\in\mathcal{H}} \left\{ F(t,p) + \frac{1}{2}\|p - (M_1(t)x + A^*y - \nabla h_1(x))\|^2 \right\} \\ &\quad - \operatorname*{argmin}_{p\in\mathcal{H}} \left\{ F(t,p) + \frac{1}{2}\|p - (M_1(t)\overline{x} + A^*\overline{y} - \nabla h_1(\overline{x}))\|^2 \right\} + \overline{x} - x \\ &= K_t(M_1(t)x + A^*y - \nabla h_1(x)) - K_t(M_1(t)\overline{x} + A^*\overline{y} - \nabla h_1(\overline{x})) + \overline{x} - x. \end{aligned}$$

Therefore,

$$\|u-\overline{u}\|^2 \leq 2\|K_t(M_1(t)x + A^*y - \nabla h_1(x)) - K_t(M_1(t)\overline{x} + A^*\overline{y} - \nabla h_1(\overline{x}))\|^2 + 2\|\overline{x} - x\|^2.$$

From Lemma 4.13, we know that K_t is $\frac{1}{\sigma}$-Lipschitz-continuous. Using this, the fact that $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$ $\forall x,y \in \mathcal{H}$ and the L_{h_1}-Lipschitz continuity of h_1 we have:

$$\begin{aligned} \|u-\overline{u}\|^2 &\leq \frac{2}{\sigma^2}\|M_1(t)(x-\overline{x}) + A^*(y-\overline{y}) - (\nabla h_1(x) - \nabla h_1(\overline{x}))\|^2 + 2\|\overline{x}-x\|^2 \\ &\leq \frac{2}{\sigma^2}(2\|M_1(t)(x-\overline{x}) + A^*(y-\overline{y})\|^2 + 2\|\nabla h_1(x) - \nabla h_1(\overline{x})\|^2) + 2\|\overline{x}-x\|^2 \\ &\leq \frac{2}{\sigma^2}(4\|M_1(t)\|^2\|x-\overline{x}\|^2 + 4\|A\|^2\|y-\overline{y}\|^2 + 2\|\nabla h_1(x) - \nabla h_1(\overline{x})\|^2) + 2\|\overline{x}-x\|^2 \\ &\leq 2\left(\frac{4\|M_1(t)\|^2 + 2L_{h_1}^2}{\sigma^2} + 1\right)\|x-\overline{x}\|^2 + \frac{8\|A\|^2}{\sigma^2}\|y-\overline{y}\|^2. \end{aligned}$$

Furthermore, by taking into account the definition of J_t in Lemma 4.13, we have

$$\begin{aligned} &v-\overline{v} \\ &= \operatorname*{argmin}_{q\in\mathcal{G}} \left\{ G(t,q) + \frac{c(t)}{2}\left\| q - \left(\frac{1}{c(t)}M_2(t)z + \frac{1}{c(t)}B^*y - B^*A(u+x) + B^*b - \frac{1}{c(t)}\nabla h_2(z)\right)\right\|^2 \right\} \\ &\quad - \operatorname*{argmin}_{q\in\mathcal{G}} \left\{ G(t,q) + \frac{c(t)}{2}\left\| q - \left(\frac{1}{c(t)}M_2(t)\overline{z} + \frac{1}{c(t)}B^*\overline{y} - B^*A(\overline{u}+\overline{x}) + B^*b - \frac{1}{c(t)}\nabla h_2(\overline{z})\right)\right\|^2 \right\} \\ &\quad + \overline{z} - z \\ &= J_t\left(\frac{1}{c(t)}M_2(t)z + \frac{1}{c(t)}B^*y - B^*A(u+x) + B^*b - \frac{1}{c(t)}\nabla h_2(z)\right) \\ &\quad - J_t\left(\frac{1}{c(t)}M_2(t)\overline{z} + \frac{1}{c(t)}B^*\overline{y} - B^*A(\overline{u}+\overline{x}) + B^*b - \frac{1}{c(t)}\nabla h_2(\overline{z})\right) + \overline{z} - z. \end{aligned}$$

According to Lemma 4.13 and assumption $(\mathcal{P})$, we have that J_t is $\frac{c(t)}{\beta}$-Lipschitz-continuous and so it follows

$$\begin{aligned}
\|v-\overline{v}\|^2 \leq\ & 2\left\|J_t\left(\frac{1}{c(t)}M_2(t)z+\frac{1}{c(t)}B^*y-B^*A(u+x)+B^*b-\frac{1}{c(t)}\nabla h_2(z))\right)-\right.\\
&\left.J_t\left(\frac{1}{c(t)}M_2(t)\overline{z}+\frac{1}{c(t)}B^*\overline{y}-B^*A(\overline{u}+\overline{x})+B^*b-\frac{1}{c(t)}\nabla h_2(\overline{z}))\right)\right\|^2+2\|\overline{z}-z\|^2\\
\leq\ & \frac{2c^2(t)}{\beta^2}\left\|\frac{1}{c(t)}M_2(t)(z-\overline{z})+\frac{1}{c(t)}B^*(y-\overline{y})-B^*A(u-\overline{u}+x-\overline{x})\right.\\
&\left.-\frac{1}{c(t)}(\nabla h_2(z)-\nabla h_2(\overline{z}))\right\|^2+2\|z-\overline{z}\|^2\\
\leq\ & \frac{2c^2(t)}{\beta^2}\left(\frac{4}{c^2(t)}\|M_2(t)\|^2\|z-\overline{z}\|^2+\frac{4}{c^2(t)}\|B\|^2\|y-\overline{y}\|^2\right.\\
&\left.+4\|B\|^2\|A\|^2\|u-\overline{u}+x-\overline{x}\|^2+\frac{4}{c^2(t)}\|\nabla h_2(z)-\nabla h_2(\overline{z})\|^2\right)+2\|z-\overline{z}\|^2\\
\leq\ & \frac{8}{\beta^2}\|M_2(t)\|^2\|z-\overline{z}\|^2+\frac{8}{\beta^2}\|B\|^2\|y-\overline{y}\|^2+\frac{16c^2(t)}{\beta^2}\|B\|^2\|A\|^2\|u-\overline{u}\|^2\\
&+\frac{16c^2(t)}{\beta^2}\|B\|^2\|A\|^2\|x-\overline{x}\|^2+\frac{8}{\beta^2}\|\nabla h_2(z)-\nabla h_2(\overline{z})\|^2+2\|z-\overline{z}\|^2.
\end{aligned}$$

Using the L_{h_2}-Lipschitz continuity of h_2, we derive from the inequality above:

$$\begin{aligned}
\|v-\overline{v}\|^2 \leq\ & \left(\frac{8\|M_2(t)\|^2+8L_{h_2}^2}{\beta^2}+2\right)\|z-\overline{z}\|^2+\frac{8}{\beta^2}\|B\|^2\|y-\overline{y}\|^2+\frac{16c^2(t)}{\beta^2}\|B\|^2\|A\|^2\|u-\overline{u}\|^2\\
&+\frac{16c^2(t)}{\beta^2}\|B\|^2\|A\|^2\|x-\overline{x}\|^2\\
\leq\ & \left(\frac{8\|M_2(t)\|^2+8L_{h_2}^2}{\beta^2}+2\right)\|z-\overline{z}\|^2+\frac{8}{\beta^2}\|B\|^2\|y-\overline{y}\|^2\\
&+\frac{16c^2(t)}{\beta^2}\|B\|^2\|A\|^2\left(\left(\frac{8\|M_1(t)\|^2+4L_{h_1}^2}{\sigma^2}+3\right)\|x-\overline{x}\|^2+\frac{8\|A\|^2}{\sigma^2}\|y-\overline{y}\|^2\right).
\end{aligned}$$

So, we have

$$\begin{aligned}
\|v-\overline{v}\|^2 \leq\ & \frac{16c^2(t)}{\beta^2}\|A\|^2\|B\|^2\left(\frac{8\|M_1(t)\|^2+4L_{h_1}^2}{\sigma^2}+3\right)\|x-\overline{x}\|^2\\
&+\frac{8}{\beta^2}\|B\|^2\left(1+\frac{16c^2(t)}{\sigma^2}\|A\|^4\right)\|y-\overline{y}\|^2+\left(\frac{8\|M_2(t)\|^2+8L_{h_2}^2}{\beta^2}+2\right)\|z-\overline{z}\|^2.
\end{aligned}$$

Finally, using the inequalities from above, we get

$$
\begin{aligned}
&\|w-\overline{w}\|^2 \\
&\quad = \| -c(t)(A(u-\overline{u}+x-\overline{x})+B(v-\overline{v}+z-\overline{z}))\|^2 \\
&\quad \leq 4c^2(t)\|A\|^2\|u-\overline{u}\|^2+4c^2(t)\|A\|^2\|x-\overline{x}\|^2+4c^2(t)\|B\|^2\|v-\overline{v}\|^2+4c^2(t)\|B\|^2\|z-\overline{z}\|^2 \\
&\quad \leq 4c^2(t)\|A\|^2\left(3+\frac{8\|M_1(t)\|^2+4L_{h_1}^2}{\sigma^2}+\frac{16c^2(t)}{\beta^2}\|B\|^4\left(\frac{8\|M_1(t)\|^2+4L_{h_1}^2}{\sigma^2}+3\right)\right)\|x-\overline{x}\|^2 \\
&\qquad +4c^2(t)\|B\|^2\left(\frac{8\|M_2(t)\|^2+8L_{h_2}^2}{\beta^2}+3\right)\|z-\overline{z}\|^2 \\
&\qquad +32c^2(t)\left(\frac{\|B\|^4}{\beta^2}+\frac{16c^2(t)}{\sigma^2\beta^2}\|A\|^4\|B\|^4+\frac{\|A\|^4}{\sigma^2}\right)\|y-\overline{y}\|^2.
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
\|\Gamma(t,x,z,y)-\Gamma(t,\overline{x},\overline{z},\overline{y})\| &\leq \sqrt{L_1(t)\|x-\overline{x}\|^2+L_2(t)\|z-\overline{z}\|^2+L_3(t)\|y-\overline{y}\|^2} \\
&\leq \sqrt{L_1(t)+L_2(t)+L_3(t)}\sqrt{\|x-\overline{x}\|^2+\|z-\overline{z}\|^2+\|y-\overline{y}\|^2} \\
&= L(t)\|(x,z,y)-(\overline{x},\overline{z},\overline{y})\|,
\end{aligned}
$$

with

$$L(t)=\sqrt{L_1(t)+L_2(t)+L_3(t)}$$

and

$$
\begin{aligned}
L_1(t) &= \left(2+\frac{32}{\beta^2}\|A\|^2\|B\|^2+8c^2(t)\|A\|^2\left(1+\frac{16c^2(t)}{\beta^2}\|B\|^4\right)\right)\left(\frac{4\|M_1(t)\|^2+2L_{h_1}^2}{\sigma^2}+1\right) \\
L_2(t) &= 2+\frac{8\|M_2(t)\|^2+8L_{h_2}^2}{\beta^2}+4c^2(t)\|B\|^2\left(\frac{8\|M_2(t)\|^2+8L_{h_2}^2}{\beta^2}+3\right) \\
L_3(t) &= \frac{8\|A\|^2}{\sigma^2}+\frac{8}{\beta^2}\|B\|^2\left(1+\frac{16c^2(t)}{\sigma^2}\|A\|^4\right)+32c^2(t)\left(\frac{\|B\|^4}{\beta^2}+\frac{16c^2(t)}{\sigma^2\beta^2}\|A\|^4\|B\|^4+\frac{\|A\|^4}{\sigma^2}\right).
\end{aligned}
$$

So, it follows that $\Gamma(t,\cdot,\cdot,\cdot)$ is $L(t)$-Lipschitz continuous. Since $\|M_1\|,\|M_2\|\in L^1_{loc}([0,+\infty))$ and $c(t)$ is bounded, it follows that $L(\cdot)\in L^1_{loc}([0,+\infty))$.

(2) In this second part, we will prove that $\Gamma(\cdot,x,z,y)\in L^1_{loc}([0,+\infty),\mathcal{H}\times\mathcal{G}\times\mathcal{K})$ for every $(x,z,y)\in\mathcal{H}\times\mathcal{G}\times\mathcal{K}$. For a fixed $(x,z,y)\in\mathcal{H}\times\mathcal{G}\times\mathcal{K}$ and $T>0$, we have

$$\int_0^T\|\Gamma(t,x,z,y)\|\mathrm{d}t=\int_0^T\sqrt{\|u(t,x,z,y)\|^2+\|v(t,x,z,y)\|^2+\|w(t,x,z,y)\|^2}\mathrm{d}t.$$

From Lemma 4.14 and the fact that $\sigma>0$, for all $t\in[0,+\infty)$, we have

$$
\begin{aligned}
\|u(t,x,z,y)\|^2 &= \|u(t,x,z,y)-u(0,x,z,y)+u(0,x,z,y)\|^2 \\
&\leq 2\|u(t,x,z,y)-u(0,x,z,y)\|^2+2\|u(0,x,z,y)\|^2 \\
&\leq \frac{2\|u(0,x,z,y)\|^2}{\sigma^2}\|M_1(t)-M_1(0)\|^2+2\|u(0,x,z,y)\|^2.
\end{aligned}
$$

From Lemma 4.14, the facts that $\sigma > 0$ and $\beta > 0$, and $(x+y+z)^2 \leq 3x^2 + 3y^2 + 3z^2$ $\forall x, z, y \in \overline{\mathbb{R}}$, for all $t \in [0, +\infty)$, we have

$$\begin{aligned}
\|v(t,x,z,y)\|^2 &\leq 2\|v(t,x,z,y) - v(0,x,z,y)\|^2 + 2\|v(0,x,z,y)\|^2 \\
&\leq 2\left(\frac{c(t)\|A\|\|B\|\|u(0,x,z,y)\|}{\sigma\beta}\|M_1(t) - M_1(0)\|\right. \\
&\quad \left. + \frac{\|v(0,x,z,y)\|}{\beta}\|M_2(t) - M_2(0)\| + \frac{\|w(0,x,z,y)\| \cdot \|B\|}{c_0\beta}|c(t) - c_0|\right)^2 \\
&\quad + 2\|v(0,x,z,y)\|^2 \\
&\leq \frac{6c^2(t)\|A\|^2\|B\|^2\|u(0,x,z,y)\|^2}{\sigma^2\beta^2}\|M_1(t) - M_1(0)\|^2 \\
&\quad + \frac{6\|v(0,x,z,y)\|^2}{\beta^2}\|M_2(t) - M_2(0)\|^2 + \frac{6\|w(0,x,z,y)\|^2 \cdot \|B\|^2}{c_0^2\beta^2}|c(t) - c_0|^2 \\
&\quad + 2\|v(0,x,z,y)\|^2,
\end{aligned}$$

(where $c_0 = c(0)$). Again using Lemma 4.14 and the inequalities above, we obtain

$$\begin{aligned}
\|w(t,x,z,y)\|^2 &\leq c^2(t)\|b - A(u(t,x,z,y) + x) - B(v(t,x,z,y) + z))\|^2 \\
&\leq 4c^2(t)(\|b\|^2 + \|A\|^2\|u(t,x,z,y)\|^2 + \|B\|^2\|v(t,x,z,y)\|^2 + \|Ax - Bz\|^2). \\
&\leq 4c^2(t)\left(\|b\|^2 + \frac{2\|A\|^2\|u(0,x,z,y)\|^2}{\sigma^2}\|M_1(t) - M_1(0)\|^2\right. \\
&\quad +2\|A\|^2\|u(0,x,z,y)\|^2 + \frac{6c^2(t)\|A\|^2\|B\|^4\|u(0,x,z,y)\|^2}{\sigma^2\beta^2}\|M_1(t) - M_1(0)\|^2 \\
&\quad +\frac{6\|B\|^2\|v(0,x,z,y)\|^2}{\beta^2}\|M_2(t) - M_2(0)\|^2 + \frac{6\|B\|^4\|w(0,x,z,y)\|^2}{c_0^2\beta^2}|c(t) - c_0|^2 \\
&\quad \left. +2\|B\|^2\|v(0,x,z,y)\|^2 + \|Ax - Bz\|^2\right).
\end{aligned}$$

Because $\|M_1\|, \|M_2\| \in L^1_{loc}([0, +\infty))$ and $c(t)$ is bounded, the integral

$$\int_0^T \|\Gamma(t,x,z,y)\| \mathrm{d}t$$

exists and it is finite. So, we have, according to Proposition 1.2.2. in [73], that $\Gamma(\cdot, x, z, y) \in L^1_{loc}([0, +\infty), \mathcal{H} \times \mathcal{G} \times \mathcal{K})$. The conclusion follows. □

4.4 Convergence of the trajectories

In the beginning of this section, we will give some results, which we will use then to prove the convergence of the trajectories of the dynamical system (4. 1).

Lemma 4.16. *Let $M : [0, +\infty) \to \mathcal{L}(\mathcal{H})$, $t \to M(t)$, be differentiable at $t_0 \in [0, +\infty)$ and $x, y : [0, +\infty) \to \mathcal{H}$ be also differentiable at t_0. Then, the real function $t \to \langle M(t)x(t), y(t)\rangle$ is differentiable at t_0 and it yields*

$$\frac{d}{dt}\langle M(t)x(t), y(t)\rangle|_{t=t_0} = \langle \dot{M}(t_0)x(t_0), y(t_0)\rangle + \langle M(t_0)\dot{x}(t_0), y(t_0)\rangle + \langle M(t_0)x(t_0), \dot{y}(t_0)\rangle. \tag{4. 13}$$

Proof. See [38, Lemma 4]. □

We start with a result where we show that under appropriate conditions the second derivatives of the trajectories exist almost everywhere and give also an upper bound on their norms. This will be used in the proof of the main result Theorem 4.22. Notice two differences with respect to [38, Lemma 5]: in our case we do not evaluate $\|\ddot{y}\|$ on the left hand side of the inequality and secondly, we have to take into account the time varying parameter c.

Lemma 4.17. *Let assumption ($\mathcal{P}$) hold true and the maps $M_1 : [0, +\infty) \to \mathcal{L}(\mathcal{H})$ and $M_2 : [0, +\infty) \to \mathcal{L}(\mathcal{G})$ be locally absolutely continuous and differentiable almost everywhere. Furthermore, we assume that c is locally absolutely continuous and*

$$0 < \inf_{t\geq 0} c(t) \leq \sup_{t\geq 0} c(t) < +\infty.$$

For a given initial value $(x^0, z^0, y^0) \in \mathcal{H} \times \mathcal{G} \times \mathcal{K}$, let $(x, z, y) : [0, +\infty) \to \mathcal{H} \times \mathcal{G} \times \mathcal{K}$ be the unique strong global solution of the dynamical system (4. 1). Then, $(\ddot{x}(t), \ddot{z}(t), \ddot{y}(t))$ exists for almost every $t \in [0, +\infty)$. If we assume additionally that

$$\sup_{t\geq 0} \|M_1(t)\| < +\infty \quad \textit{and} \quad \sup_{t\geq 0} \|M_2(t)\| < +\infty,$$

then there exists $L > 0$ such that

$$\|\ddot{x}(t)\| + \|\ddot{z}(t)\| \leq L(\|\dot{x}(t)\| + \|\dot{z}(t)\| + \|\dot{y}(t)\| + \|\dot{M}_1(t)\|\|\dot{x}(t)\| + \|\dot{M}_2(t)\|\|\dot{z}(t)\| + |\dot{c}(t)|\|\dot{y}(t)\|)$$

for almost every $t \in [0, +\infty)$.

Proof. In the following, we use the notation (4. 6) again. Let $T > 0$ and $t, r \in [0, T]$ be fixed. So

$$\begin{aligned}
&\|\Gamma(t, U(t)) - \Gamma(r, U(r))\| \leq \|\Gamma(t, U(t)) - \Gamma(t, U(r))\| + \|\Gamma(t, U(r)) + \Gamma(r, U(r))\| \\
&\quad \leq \|u(t, x(t), z(t), y(t)) - u(t, x(r), z(r), y(r))\| + \|v(t, x(t), z(t), y(t)) - v(t, x(r), z(r), y(r))\| \\
&\qquad + \|w(t, x(t), z(t), y(t)) - w(t, x(r), z(r), y(r))\| + \|u(t, x(r), z(r), y(r) - u(r, x(r), z(r), y(r))\| \\
&\qquad + \|v(t, x(r), z(r), y(r)) - v(r, x(r), z(r), y(r))\| + \|w(t, x(r), z(r), y(r)) - w(r, x(r), z(r), y(r))\|.
\end{aligned}$$

We have

$$\begin{aligned}
u(t, x(t), z(t), y(t)) - u(t, x(r), z(r), y(r)) =& K_t(M_1(t)x(t) + A^*y(t) - \nabla h_1(x(t))) \\
&- K_t(M_1(t)x(r) + A^*y(r) - \nabla h_1(x(r))) - x(t) + x(r)
\end{aligned}$$

and, due to Lemma 4.13 and the $L_{h_1]}$-Lipschitz continuity of ∇h_1, we get

$$\begin{aligned}
&\|u(t, x(t), z(t), y(t)) - u(t, x(r), z(r), y(r))\| \\
&\quad \leq \frac{1}{\sigma}\|M_1(t)(x(t) - x(r)) + A^*(y(t) - y(r)) - (\nabla h_1(x(t)) - \nabla h_1(x(r)))\| + \|x(t) - x(r)\| \\
&\quad \leq \left(\frac{\|M_1(t)\|}{\sigma} + \frac{L_{h_1}}{\sigma} + 1\right) \|x(t) - x(r)\| + \frac{\|A\|}{\sigma}\|y(t) - y(r)\|.
\end{aligned}$$

Because $t \to \|M_1(t)\|$ is bounded on $[0, T]$, there exists $L_1 := L_1(T) > 0$ such that

$$\|u(t, x(t), z(t), y(t)) - u(t, x(r), z(r), y(r))\| \leq L_1(\|x(t) - x(r)\| + \|y(t) - y(r)\|). \tag{4. 14}$$

Analogously, we have

$$\begin{aligned}
&v(t,x(t),z(t),y(t)) - v(t,x(r),z(r),y(r))\\
&\quad = J_t\left(\frac{1}{c(t)}M_2(t)z(t) + \frac{1}{c(t)}B^*y(t) - B^*A(u(t,x(t),z(t),y(t)) + x(t)) + B^*b - \frac{1}{c(t)}\nabla h_2(z(t))\right)\\
&\quad\quad - J_t\left(\frac{1}{c(t)}M_2(t)z(r) + \frac{1}{c(t)}B^*y(r) - B^*A(u(t,x(r),z(r),y(r)) + x(r)) + B^*b - \frac{1}{c(t)}\nabla h_2(z(r))\right)\\
&\quad\quad - z(t) + z(r),
\end{aligned}$$

and again, according to Lemma 4.13 , the L_{h_2}-Lipschitz continuity of ∇h_2 and (4. 14), we get

$$\begin{aligned}
&\|v(t,x(t),z(t),y(t)) - v(t,x(r),z(r),y(r))\|\\
&\quad \leq \frac{c(t)}{\beta}\left\|\frac{1}{c(t)}M_2(t)(z(t)-z(r)) + \frac{1}{c(t)}B^*(y(t)-y(r)) - B^*A(u(t,x(t),z(t),y(t))\right.\\
&\quad\quad \left. - u(t,x(r),z(r),y(r)) + x(t) - x(r)) - \frac{1}{c(t)}(\nabla h_2(z(t)) - \nabla h_2(z(r)))\right\| + \|z(t)-z(r)\|\\
&\quad \leq \frac{c(t)}{\beta}\|A\|\|B\|(L_1+1)\|x(t)-x(r)\| + \left(\frac{\|M_2(t)\|}{\beta} + \frac{L_{h_2}}{\beta} + 1\right)\|z(t)-z(r)\|\\
&\quad\quad + \frac{\|B\|}{\beta}(1 + c(t)\|A\|L_1)\|y(t)-y(r)\|.
\end{aligned}$$

Because $t \to \|M_2(t)\|$ is bounded on $[0,T]$, there exists $L_2 := L_2(T) > 0$ such that

$$\|v(t,x(t),z(t),y(t)) - v(t,x(r),z(r),y(r))\| \leq L_2(\|x(t)-x(r)\| + \|z(t)-z(r)\| + \|y(t)-y(r)\|). \tag{4. 15}$$

Using (4. 14) and (4. 15), we obtain

$$\begin{aligned}
&\|w(t,x(t),z(t),y(t)) - w(t,x(r),z(r),y(r))\|\\
&\quad \leq c(t)\|A\|\|u(t,x(t),z(t),y(t)) - u(t,x(r),z(r),y(r) + x(t) - x(r))\|\\
&\quad\quad + c(t)\|B\|\|v(t,x(t),z(t),y(t)) - v(t,x(r),z(r),y(r)) + z(t) - z(r)\|\\
&\quad \leq c(t)(\|A\|L_1 + \|B\|L_2 + \|A\|)\|x(t)-x(r)\| + c(t)(\|B\|L_2 + \|B\|)\|z(t)-z(r)\|\\
&\quad\quad + c(t)(\|A\|L_1 + \|B\|L_2)\|y(t)-y(r)\|.
\end{aligned}$$

So, there exists $L_3(T) := L_3 := \sup_{t\in[0,T]} c(t)(\|A\|L_1 + \|B\|L_2 + \|A\| + \|B\|) > 0$ such that

$$\|w(t,x(t),z(t),y(t)) - w(t,x(r),z(r),y(r))\| \leq L_3(\|x(t)-x(r)\| + \|z(t)-z(r)\| + \|y(t)-y(r)\|). \tag{4. 16}$$

Using Lemma 4.14(i), we obtain

$$\begin{aligned}
\|u(t,x(r),z(r),y(r)) - u(r,x(r),z(r),y(r))\| &= \|R_{(x(r),z(r),y(r))}(t) - R_{(x(r),z(r),y(r))}(r)\|\\
&\leq \frac{\|R_{(x(r),z(r),y(r))}(r)\|}{\sigma}\|M_1(t) - M_1(r)\|.
\end{aligned} \tag{4. 17}$$

Due to the Lipschitz continuity of K_r and ∇h_1 (see Lemma 4.13) and the fact that x, z, y and M_1 are absolutely continuous on $[0,T]$, the map

$$r \mapsto R_{(x(r),z(r),y(r))}(r) = K_r(M_1(r)x(r) + A^*y(r) - \nabla h_1(x(r))) - x(r)$$

is bounded in $[0,T]$. Therefore, there exists $L_4 := L_4(T) > 0$ such that

$$\|u(t,x(r),z(r),y(r)) - u(r,x(r),z(r),y(r))\| \leq L_4\|M_1(t) - M_1(r)\|. \tag{4. 18}$$

In an analog way, using Lemma 4.14 (ii), we get

$$\begin{aligned}
&\|v(t,x(r),z(r),y(r)) - v(r,x(r),z(r),y(r))\| = \|Q_{(x(r),z(r),y(r))}(t) - Q_{(x(r),z(r),y(r))}(r)\| \\
&\quad\leq \frac{c(t)\|A\|\|B\|\|R_{(x(r),z(r),y(r))}(r)\|}{\sigma\beta}\|M_1(t) - M_1(r)\| + \frac{\|Q_{(x(r),z(r),y(r))}(r)\|}{\beta}\|M_2(t) - M_2(r)\| \\
&\qquad + \frac{\|P_{(x(r),z(r),y(r))}(r)\| \cdot \|B\|}{c(r)\beta}|c(t) - c(r)|.
\end{aligned} \tag{4. 19}$$

Using the same arguments about the Lipschitz continuity of J_r and ∇h_2 as above (see again Lemma 4.13) and the fact that x, z, y and M_2 are absolutely continuous on $[0,T]$ and c is bounded, the maps

$$\begin{aligned}
r \mapsto Q_{(x(r),z(r),y(r))}(r) =&J_r\left(\frac{M_2(r)}{c(r)}z(r) + \frac{B^*}{c(r)}y(r) - B^*A(u(r,x(r),z(r),y(r) + x(r))\right. \\
&\left. +B^*b - \frac{1}{c(r)}\nabla h_2(z(r))\right) - z(r)
\end{aligned}$$

and

$$r \mapsto P_{(x(r),z(r),y(r))}(r) = c(r)\left(b - A(R_{(x(r),z(r),y(r))}(r) + x(r)) - B(Q_{(x(r),z(r),y(r))}(r) + z(r)\right)$$

are bounded in $[0,T]$. Therefore, there exists $L_5 := L_5(T) > 0$ such that

$$\begin{aligned}
\|v(t,x(r),z(r),y(r)) - v(r,x(r),z(r),y(r))\| \leq L_5(&\|M_1(t) - M_1(r)\| + \|M_2(t) - M_2(r)\| \\
&+ |c(t) - c(r)|).
\end{aligned} \tag{4. 20}$$

Furthermore, we have

$$\begin{aligned}
&\|w(t,x(r),z(r),y(r)) - w(r,x(r),z(r),y(r))\| \\
&\quad\leq \|(c(t) - c(r))b - (c(t) - c(r))Ax(r) - (c(t) - c(r))Bz(r) - A(c(t)(u(t,x(r),z(r),y(r)) \\
&\qquad - c(r)(u(r,x(r),z(r),y(r)))) - B(c(t)(v(t,x(r),z(r),y(r)) - c(r)(v(r,x(r),z(r),y(r))))\| \\
&\quad\leq \|b\||c(t) - c(r)| + \|A\|\|x(r)\||c(t) - c(r)| + \|B\|\|z(r)\||c(t) - c(r)| \\
&\qquad + \|A(c(t)u(t,x(r),z(r),y(r) - c(r)u(r,x(r),z(r),y(r)))\| \\
&\qquad + \|B(c(t)v(t,x(r),z(r),y(r) - c(r)v(r,x(r),z(r),y(r)))\|
\end{aligned}$$

and futher

$$\begin{aligned}
&\|w(t,x(r),z(r),y(r)) - w(r,x(r),z(r),y(r))\| \\
&\quad\leq (\|b\| + \|A\|\|x(r)\| + \|B\|\|z(r)\|)|c(t) - c(r)| \\
&\qquad + \|A\|(\|c(t)u(t,x(r),z(r),y(r) - c(r)u(t,x(r),z(r),y(r))\| \\
&\qquad + \|A\|\|c(r)u(t,x(r),z(r),y(r) - c(r)u(r,x(r),z(r),y(r))\|) \\
&\qquad + \|B\|\|c(t)v(t,x(r),z(r),y(r) - c(r)v(t,x(r),z(r),y(r))\| \\
&\qquad + \|B\|\|c(r)v(t,x(r),z(r),y(r) - c(r)v(r,x(r),z(r),y(r))\| \\
&\quad\leq (\|b\| + \|A\|\|x(r)\| + \|B\|\|z(r)\| + \|A\|\|(u(t,x(r),z(r),y(r))\| \\
&\qquad + \|B\|\|v(t,x(r),z(r),y(r))\|)|c(t) - c(r)| \\
&\qquad + c(r)\|A\|\|u(t,x(r),z(r),y(r) - u(r,x(r),z(r),y(r))\| \\
&\qquad + c(r)\|B\|\|v(t,x(r),z(r),y(r) - v(r,x(r),z(r),y(r))\|.
\end{aligned}$$

Using (4. 17) and (4. 19), we obtain

$$\begin{aligned}&\|w(t,x(r),z(r),y(r))-w(r,x(r),z(r),y(r))\|\\ &\quad\le (\|b\|+\|A\|\|x(r)\|+\|B\|\|z(r)\|+\|A\|\|(u(t,x(r),z(r),y(r))\|\\ &\qquad+\|B\|\|v(t,x(r),z(r),y(r))\|+c(r)\|B\|L_5)|c(t)-c(r)|\\ &\qquad+c(r)(\|A\|L_4+\|B\|L_5)\|M_1(t)-M_1(r)\|+c(r)\|B\|L_5\|M_2(t)-M_2(r)\|.\end{aligned}$$

So, there exists $L_6 := L_6(T) = \sup_{r\in[0,T]}(\|b\|+\|A\|\|x(r)\|+\|B\|\|z(r)\|+\|A\|\|(u(t,x(r),z(r),y(r))\|+\|B\|\|v(t,x(r),z(r),y(r))\|+c(r)\|A\|L_4+c(r)\|B\|L_5)>0$ such that

$$\begin{aligned}&\|w(t,x(r),z(r),y(r))-w(r,x(r),z(r),y(r))\|\\ &\qquad\le L_6(\|M_1(t)-M_1(r)\|+\|M_2(t)-M_2(r)\|+|c(t)-c(r)|).\end{aligned} \tag{4. 21}$$

When we sum the relations (4. 14), (4. 15), (4. 16), (4. 18), (4. 20), and (4. 21), we get that there exists $L_7 := L_7(T) > 0$ such that

$$\begin{aligned}\|\Gamma(t,U(t))-\Gamma(r,U(r))\| \le L_7(&\|x(t)-x(r)\|+\|z(t)-z(r)\|+\|y(t)-y(r)\|\\ &+\|M_1(t)-M_1(r)\|+\|M_2(t)-M_2(r)\|+|c(t)-c(r)|).\end{aligned}$$

Let $\epsilon > 0$. Due to the absolute continuity of the maps x, z, y, M_1, M_2 and c on $[0,T]$, there exists $\eta > 0$ such that for any finite family of intervals $I_k = (a_k, b_k) \subseteq [0,T]$ and for any subfamily of disjoint intervals I_j with $\sum_j |b_j - a_j| < \eta$ holds

$$\sum_j \|x(b_j)-x(a_j)\| < \frac{\epsilon}{6L_7},\quad \sum_j \|z(b_j)-z(a_j)\| < \frac{\epsilon}{6L_7},\quad \sum_j \|y(b_j)-y(a_j)\| < \frac{\epsilon}{6L_7},$$

$$\sum_j \|M_1(b_j)-M_1(a_j)\| < \frac{\epsilon}{6L_7},\quad \sum_j \|M_2(b_j)-M_2(a_j)\| < \frac{\epsilon}{6L_7} \text{ and } \sum_j |c(b_j)-c(a_j)| < \frac{\epsilon}{6L_7}.$$

So, we have

$$\sum_j \|\Gamma(b_j,U(b_j))-\Gamma(a_j,U(a_j))\| < \epsilon,$$

and therefore $t \to \Gamma(t,U(t))$ is absolutely continuous on $[0,T]$. Since $\dot{U}(t) = \Gamma(U(t),t)$ for almost every $t \in [0,+\infty)$, we can extend $\dot{U}(t)$ to a locally absolutely continuous function on $[0,+\infty)$ by setting $\dot{U}(t) = \Gamma(U(t),t)$ for $t \in [0,+\infty)$. It follows that the second order derivatives $\ddot{x}, \ddot{z}, \ddot{y}$ exist almost everywhere on $[0,+\infty)$.

To prove the second statement, we assume that

$$\sup_{t\ge 0}\|M_1(t)\| < +\infty \quad\text{and}\quad \sup_{t\ge 0}\|M_2(t)\| < +\infty.$$

Note that $c(t)$ is bounded for all $t \in [0,+\infty)$. Then, L_1, L_2, and L_3 can be taken as being global constants, so that (4. 14), (4. 15) and (4. 16) hold for every $t, r \in [0,+\infty)$.

Because $R_{(x(r),z(r),y(r)}(r) = \dot{x}(r)$, $Q_{(x(r),z(r),y(r)}(r) = \dot{z}(r)$, and $P_{(x(r),z(r),y(r)}(r) = \dot{y}(r)$ for every $r \in [0,+\infty)$ and taking into account (4. 17) and (4. 19), we get

$$\|u(t,x(r),z(r),y(r))-u(r,x(r),z(r),y(r))\| \le \frac{\|\dot{x}(r)\|}{\sigma}\|M_1(t)-M_1(r)\|. \tag{4. 22}$$

and

$$\begin{aligned}\|v(t,x(r),z(r),y(r)) - v(r,x(r),z(r),y(r))\| \\ &\le \frac{c(t)\|A\|\|B\|\|\dot{x}(r)\|}{\sigma\beta}\|M_1(t)-M_1(r)\| \\ &\quad + \frac{\|\dot{z}(r)\|}{\beta}\|M_2(t)-M_2(r)\| + \frac{\|\dot{y}(r)\|\cdot\|B\|}{\beta c(r)}|c(t)-c(r)|\end{aligned} \tag{4. 23}$$

for every $t, r \in [0,+\infty)$. It holds

$$\begin{aligned}&\|\dot{x}(t)-\dot{x}(r)\| + \|\dot{z}(t)-\dot{z}(r)\| \\ &\quad = \|u(t,x(t),z(t),y(t)) - u(r,x(r),z(r),y(r))\| + \|v(t,x(t),z(t),y(t)) - v(r,x(r),z(r),y(r))\| \\ &\quad \le \|u(t,x(t),z(t),y(t)) - u(t,x(r),z(r),y(r))\| + \|u(t,x(r),z(r),y(r)) - u(r,x(r),z(r),y(r))\| \\ &\qquad + \|v(t,x(t),z(t),y(t)) - v(t,x(r),z(r),y(r))\| + \|v(t,x(r),z(r),y(r)) - v(r,x(r),z(r),y(r))\|.\end{aligned}$$

So, it follows from (4. 14), (4. 15), (4. 22) and (4. 23) that there exists $L > 0$ such that

$$\begin{aligned}\|\dot{x}(t)-\dot{x}(r)\| + \|\dot{z}(t)-\dot{z}(r)\| \le & L(\|x(t)-x(r)\| + \|z(t)-z(r)\| + \|y(t)-y(r)\| \\ & + \|\dot{x}(r)\|\|M_1(t)-M_1(r)\| + \|\dot{z}(r)\|\|M_2(t)-M_2(r)\| \\ & + \|\dot{y}(r)\||c(t)-c(r)|)\end{aligned}$$

for every $t, r \in [0,+\infty)$. Now, we fix $r \in [0,+\infty)$ such that $\ddot{x}(r), \ddot{z}(r), \dot{M}_1(r), \dot{M}_2(r)$ exist and consider the above inequality for $t = r + h$ for some $h > 0$ and obtain

$$\begin{aligned}&\|\dot{x}(r+h)-\dot{x}(r)\| + \|\dot{z}(r+h)-\dot{z}(r)\| \\ &\quad \le L(\|x(r+h)-x(r)\| + \|z(r+h)-z(r)\| + \|y(r+h)-y(r)\|) \\ &\qquad + L(\|\dot{x}(r)\|\|M_1(r+h)-M_1(r)\| + \|\dot{z}(r)\|\|M_2(r+h)-M_2(r)\| \\ &\qquad + \|\dot{y}(r)\||c(r+h)-c(r)|).\end{aligned}$$

Finally, we divide the inequality above by h and let $h \to 0$. We get

$$\|\ddot{x}(r)\| + \|\ddot{z}(r)\| \le L(\|\dot{x}(r)\| + \|\dot{z}(r)\| + \|\dot{y}(r)\| + \|\dot{x}(r)\|\|\dot{M}_1(r)\| + \|\dot{z}(r)\|\|\dot{M}_2(r)\| + \|\dot{y}(r)\||\dot{c}(r)|)$$

and the proof is complete. □

In the following, we recall two results which we need for the asymptotic analysis (see [2, Lemma 5.1] and [2, Lemma 5.2]).

Lemma 4.18. *Assume that $u : [0,+\infty) \to \mathbb{R}$ is locally absolutely continuous and bounded from below and that there exists $v \in L^1([0,+\infty),\mathbb{R})$ with the property that for almost every $t \in [0,+\infty)$*

$$\frac{d}{dt}u(t) \le v(t).$$

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

Lemma 4.19. *Assume that $1 \le p < \infty$, $1 \le r \le \infty$, $u : [0,+\infty) \to [0,+\infty)$ is locally absolutely continuous, $u \in L^p([0,+\infty),\mathbb{R})$, $v : [0,+\infty) \to \mathbb{R}$, $v \in L^r([0,+\infty),\mathbb{R})$ and for almost every $t \in [0,+\infty)$*

$$\frac{d}{dt}u(t) \le v(t).$$

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

Definition 4.20. The map $M : [0, +\infty) \to S_+(\mathcal{H})$ is said to be *monotonically decreasing* if it satisfies $M(t_1) \succcurlyeq M(t_2)$ for every $t_1, t_2 \in [0, +\infty)$ with $t_1 \leq t_2$.

Lemma 4.21. *For a $\alpha > 0$ let the map $M : [0, +\infty) \to \mathcal{P}_\alpha(\mathcal{H})$ be monotonically decreasing. Then, there exists $M \in \mathcal{P}_\alpha(\mathcal{H})$ such that for every $z \in \mathcal{H}$ it holds that*

$$\lim_{t \to +\infty} M(t)z = Mz.$$

Proof. See [[38], proof of Lemma 8]. □

Theorem 4.22. *In the context of optimization problem* (3. 5)*, suppose that the set of saddle points of the Lagrangian L is nonempty. Let the maps maps $M_1 : [0, +\infty) \to \mathcal{L}(\mathcal{H})$ and $M_2 : [0, +\infty) \to \mathcal{L}(\mathcal{G})$ be locally absolutely continuous, differentiable almost everywhere and monotonically decreasing in the sense of the Loewner partial ordering defined in* (2. 4)*,*

$$M_1(t) - \frac{L_{h_1}}{4} I \in S_+(\mathcal{H}),$$

$$M_2(t) - \frac{L_{h_2}}{4} I \in S_+(\mathcal{G}) \quad \forall t \in [0, +\infty),$$

and

$$\text{ess sup}_{t \geq 0} \|\dot{M}_1(t)\| < +\infty \textit{ and } \text{ess sup}_{t \geq 0} \|\dot{M}_2(t)\| < +\infty.$$

Furthermore, we assume that for $0 < \epsilon < \frac{\sigma}{2\|A\|^2}$, the function

$$c : [0, +\infty) \to \left[\epsilon, \frac{\sigma}{\|A\|^2} - \epsilon\right]$$

is monotonically decreasing and Lipschitz continuous. If $c(t)$ is a constant function, namely $c(t) = c$ for all $t \in [0, +\infty)$, then it is enough to assume that $\epsilon \leq c \leq \frac{2\sigma}{\|A\|^2} - \epsilon$. For an arbitrary initial value $(x^0, z^0, y^0) \in \mathcal{H} \times \mathcal{G} \times \mathcal{K}$, let $(x, z, y) : [0, +\infty) \to \mathcal{H} \times \mathcal{G} \times \mathcal{K}$ be the unique strong global solution of the dynamical system (4. 1)*. If one of the following assumptions is satisfied:*

1. *there exists $\alpha > 0$ such that $M_2(t) - \frac{L_{h_2}}{4} I \in \mathcal{P}_\alpha(\mathcal{G})$ for every $t \in [0, +\infty)$*
2. *there exists $\beta > 0$ such that $B^*B \in \mathcal{P}_\beta(\mathcal{G})$;*

then for $t \to +\infty$, the trajectory $(x(t), z(t), y(t))$ converges weakly to a saddle point of L.

In the proof of the theorem, we will construct a Lyapunov function to show the convergence of the trajectories. Lyapunov functions are used to gain knowledge about the stability of dynamical systems. For deeper insight in the Lyapunov analysis, we refer the reader to [94]. In the following, we will give the definition of the Lyapunov function. For this, we consider the dynamical system of the form

$$\begin{aligned} \dot{u}(t) &\in \Gamma(t, u(t)) \\ u(s) &= u^s \in \mathcal{X} \end{aligned} \tag{4. 24}$$

with $\mathcal{X} = \mathcal{H} \times \mathcal{G} \times \mathcal{K}, s \in [0, +\infty), u : [s, +\infty) \to \mathcal{X}$ and Γ defined as in Remark 4.9. Given s, u^s and $u(t)$ such that (4. 24) holds, we write

$$S(t, s, u^s) = u(t).$$

It is clear by Theorem 4.15 that $S(t, s, u^s)$ is uniquely determined. We call

$$S : \{(t, s) \in [0, \infty)^2 : t \geq s\} \times \mathcal{X} \to \mathcal{X}$$

the *propagator* associated to (4. 24).

Definition 4.23. Let $v_0 : [0, +\infty) \to [0, +\infty)$ and $v_1 : [0, +\infty) \to [0, +\infty)$ be continuous, strictly increasing functions such that $v_0(0) = 0$ and $v_1(0) = 0$, respectively. Let $u^* \in \mathcal{X}$. A *strict Lyapunov function* of the dynamical system (4. 24) is a continuous function $V : \mathcal{X} \times \mathbb{R} \to [0, +\infty)$, which is

- strictly positive definite , i.e. it holds that $V(u,t) \geq v_0(\|u - u^*\|) \geq 0$ for all $(u,t) \in \mathcal{H} \times \mathbb{R}$,
- decrescent , i.e. it holds that $V(u,t) \leq v_1(\|u - u^*\|)$ for all $(u,t) \in \mathcal{H} \times \mathbb{R}$,

and, furthermore, it holds that

$$\dot{V}(u,t) := \limsup_{\tau \to 0+} \frac{V(S(t+\tau, t, u), t+\tau) - V(u,t)}{\tau}$$

is negative definite, i.e.,

$$-\dot{V}(u,t) \geq v(\|u - u^*\|) \geq 0, \tag{4. 25}$$

where $v : [0, +\infty) \to [0, +\infty)$ is a continuous and positive definite function, i.e., $v(0) = 0$, and $v(s) > 0$ for all $s > 0$. If $v : [0, +\infty) \to [0, +\infty)$ is only continuous and positive semi-definite, i.e., $v(s) \geq 0$ for all $s \geq 0$ the Lyapunov function $V(u,t)$ is said to be *non-strict*.

In the theorem above, we have a result which states the asymptotic convergence of the trajectories generated by the dynamical system (4. 1) to a saddle point of the Lagrangian of the problem (3. 5). To show this convergence, we will use a Lyapunov function $V(x,z,y,t)$, which is given by

$$\begin{aligned} V(x,z,y,t) =& (2\sigma c(t) - c^2(t)\|A\|^2)\|x - x^*\|^2 + \|x - x^*\|^2_{c(t)M_1(t)} \\ &+ \|z - z^*\|^2_{c(t)M_2(t) + c^2(t)B^*B} + \|y - y^*\|^2, \end{aligned}$$

which fulfills Definition 4.23: If $c(t)$ is not constant, we have that

$$\epsilon(\sigma + \epsilon\|A\|^2) \leq c(t)(2\sigma - c(t)\|A\|^2) \leq (\sigma/\|A\|^2 - \epsilon)(2\sigma - \epsilon\|A\|^2) \quad \forall t \in [0, +\infty).$$

So, we can choose, if $c(t)$ is not constant,

$$v_1(r) = ((\sigma/\|A\|^2 - \epsilon)(2\sigma - \epsilon\|A\|^2) + c(0)\|M_1(0)\| + c(0)\|M_2(0)\| + c^2(0)\|B\|^2 + 1)r^2$$

and, if $c(t)$ is a constant function $c(t) = c$ for all $t \in [0, +\infty)$, we choose

$$v_1(r) = (2\sigma c - c^2\|A\|^2 + c\|M_1(0)\| + c\|M_2(0)\| + c\|B\|^2 + 1)r^2$$

Further, if assumption 1 of Theorem 4.22 is not fulfilled, we set $\alpha = 0$, if assumption 2 is not fulfilled, we set $\beta = 0$. Then, we can choose, if $c(t)$ is not constant

$$v_0(r) = (\epsilon(\sigma + \epsilon\|A\|^2) + \alpha + \beta + 1)r^2$$

and, if $c(t)$ is a constant function $c(t) = c$ for all $t \in [0, +\infty)$, we choose

$$v_0(r) = ((2\sigma c - c^2\|A\|^2) + \alpha + \beta + 1)r^2.$$

We have from inequality (4. 33), if $L_{h_1} > 0$ and $L_{h_2} > 0$, from inequality (4. 39), if $L_{h_1} = 0$ and $L_{h_2} > 0$, from inequality (4. 40), if $L_{h_1} > 0$ and $L_{h_2} = 0$, and from inequality (4. 41), if $L_{h_1} = 0$ and $L_{h_2} = 0$, that

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big((2\sigma c(t) - c^2(t)\|A\|^2)\|x(t) - x^*\|^2 + \|x(t) - x^*\|^2_{c(t)M_1(t)} \\ + \|z(t) - z^*\|^2_{c(t)M_2(t)+c^2(t)B^*B} + \|y(t) - y^*\|^2\Big) \leq 0,$$

which means that $\dot{V}(x,t)$ is negative semi-definite. So, we have that $V(x,z,y,t)$ is a non-strict Lyapunov function. Now, we proof Theorem 4.22.

Proof of Theorem 4.22. We need an appropriate energy functional in order to conclude. This will be accomplished in (4. 33) below. Let $(x^*, z^*, y^*) \in \mathcal{H} \times \mathcal{G} \times \mathcal{K}$ be a saddle point of the Lagrangian L. Then, it fulfills the system of the optimality conditions

$$\begin{cases} A^*y^* - \nabla h_1(x^*) \in \partial f(x^*) \\ B^*y^* - \nabla h_2(z^*) \in \partial g(z^*) \\ Ax^* + Bz^* = b. \end{cases}$$

From (4. 3), we have for almost every $t \in [0, +\infty)$

$$-M_1(t)\dot{x}(t) + A^*y(t) - \nabla h_1(x(t)) \in \partial f(\dot{x}(t) + x(t)),$$

and, due to the strong monotonicity of ∂f, which follows from Proposition 2.22 (ii), we have

$$\langle -M_1(t)\dot{x}(t) + A^*(y(t) - y^*) - (\nabla h_1(x(t)) - \nabla h_1(x^*)), \dot{x}(t) + x(t) - x^*\rangle \geq \sigma\|\dot{x}(t) + x(t) - x^*\|^2. \tag{4. 26}$$

In an analog way, according to (4. 4), we have for almost every $t \in [0, +\infty)$

$$- c(t)B^*B(\dot{z}(t) + z(t)) - M_2(t)(\dot{z}(t)) + B^*y(t) - c(t)B^*A(\dot{x}(t) + x(t)) + c(t)B^*b - \nabla h_2(z(t)) \\ \in \partial g(\dot{z}(t) + z(t)),$$

and, by taking into account the monotonicity of ∂g, we have

$$\langle -c(t)B^*B(\dot{z}(t) + z(t)) - M_2(t)(\dot{z}(t)) + B^*(y(t) - y^*) - c(t)B^*A(\dot{x}(t) + x(t)) + c(t)B^*b \\ - (\nabla h_2(z(t)) - \nabla h_2(z^*)), \dot{z}(t) + z(t) - z^*\rangle \geq 0. \tag{4. 27}$$

We use the optimality condition $Ax^* + Bz^* = b$ and the last equation of (4. 1) to obtain for almost every $t \in [0, +\infty)$

$$\begin{aligned} &\langle A^*(y(t) - y^*), \dot{x}(t) + x(t) - x^*\rangle + \langle B^*(y(t) - y^*), \dot{z}(t) + z(t) - z^*\rangle \\ &\quad = -\langle y(t) - y^*, -A(\dot{x}(t) + x(t)) + Ax^* - B(\dot{z}(t) + z(t)) + Bz^*\rangle \\ &\quad = -\langle y(t) - y^*, -A(\dot{x}(t) + x(t)) - B(\dot{z}(t) + z(t)) + b\rangle \\ &\quad = -\frac{1}{c(t)}\langle y(t) - y^*, \dot{y}(t)\rangle = -\frac{1}{2c(t)}\frac{\mathrm{d}}{\mathrm{d}t}\|y(t) - y^*\|^2. \end{aligned} \tag{4. 28}$$

Suppose that $L_{h_1} > 0$ and $L_{h_2} > 0$. Due to the Baillon-Haddad Theorem 2.20, we know that the gradients of h_1 and h_2 are L_1^{-1}- and L_2^{-1}-cocoercive, respectively, we have for almost every

$t \in [0, +\infty)$

$$\begin{aligned}
&\langle -(\nabla h_1(x(t)) - \nabla h_1(x^*)), \dot{x}(t) + x(t) - x^* \rangle \\
&\quad = -\langle \nabla h_1(x(t)) - \nabla h_1(x^*), x(t) - x^* \rangle - \langle \nabla h_1(x(t)) - \nabla h_1(x^*), \dot{x}(t) \rangle \\
&\quad \le -\frac{1}{L_{h_1}} \|\nabla h_1(x(t)) - \nabla h_1(x^*)\|^2 - \langle \nabla h_1(x(t)) - \nabla h_1(x^*), \dot{x}(t) \rangle \\
&\quad = -\frac{1}{L_{h_1}} \left(\left\| \nabla h_1(x(t)) - \nabla h_1(x^*) + \frac{L_{h_1}}{2} \dot{x}(t) \right\|^2 - \frac{L_{h_1}^2}{4} \|\dot{x}(t)\|^2 \right)
\end{aligned} \tag{4. 29}$$

and, respectively

$$\begin{aligned}
&\langle -(\nabla h_2(z(t)) - \nabla h_2(z^*)), \dot{z}(t) + z(t) - z^* \rangle \\
&\quad = -\frac{1}{L_{h_2}} \left(\left\| \nabla h_2(z(t)) - \nabla h_2(z^*) + \frac{L_{h_2}}{2} \dot{z}(t) \right\|^2 - \frac{L_{h_2}^2}{4} \|\dot{z}(t)\|^2 \right).
\end{aligned} \tag{4. 30}$$

If we sum (4. 26) and (4. 27) and take into account (4. 28), (4. 29), and (4. 30), we get for almost every $t \in [0, +\infty)$

$$\begin{aligned}
0 \le & \langle -M_1(t)\dot{x}(t), \dot{x}(t) + x(t) - x^* \rangle + \langle -c(t)B^*B(\dot{z}(t) + z(t)) - M_2(t)(\dot{z}(t)) \\
& - c(t)B^*A(\dot{x}(t) + x(t)) + c(t)B^*b, \dot{z}(t) + z(t) - z^* \rangle - \frac{1}{2c(t)} \frac{\mathrm{d}}{\mathrm{d}t} \|y(t) - y^*\|^2 \\
& - \frac{1}{L_{h_1}} \left(\left\| \nabla h_1(x(t)) - \nabla h_1(x^*) + \frac{L_{h_1}}{2} \dot{x}(t) \right\|^2 - \frac{L_{h_1}^2}{4} \|\dot{x}(t)\|^2 \right) \\
& - \frac{1}{L_{h_2}} \left(\left\| \nabla h_2(z(t)) - \nabla h_2(z^*) + \frac{L_{h_2}}{2} \dot{z}(t) \right\|^2 - \frac{L_{h_2}^2}{4} \|\dot{z}(t)\|^2 \right) - \sigma \|\dot{x}(t) + x(t) - x^*\|^2.
\end{aligned} \tag{4. 31}$$

Furthermore, we have for almost every $t \in [0, +\infty)$ (use also the last equality for $\dot{y}$ in (4. 1)):

$$\begin{aligned}
&\langle -c(t)B^*B(\dot{z}(t) + z(t)) - c(t)B^*A(\dot{x}(t) + x(t)) + c(t)B^*b, \dot{z}(t) + z(t) - z^* \rangle \\
&\quad = -\frac{1}{c(t)} \langle \dot{y}(t), -c(t)B(\dot{z}(t) + z(t) - z^*) \rangle \\
&\quad = -\frac{1}{c(t)} \left[\frac{1}{2} \|\dot{y}(t)\|^2 + \frac{1}{2} \|c(t)B(\dot{z}(t) + z(t) - z^*)\|^2 - \frac{1}{2} \|\dot{y}(t) + c(t)B(\dot{z}(t) + z(t) - z^*)\|^2 \right] \\
&\quad = -\frac{1}{c(t)} \left[\frac{1}{2} \|\dot{y}(t)\|^2 + \frac{1}{2} c^2(t) \left[\|B(z(t) - z^*)\|^2 + \|B\dot{z}(t)\|^2 + 2\langle \dot{z}(t), B^*B(z(t) - z^*) \rangle \right] \right. \\
&\qquad \left. - \frac{1}{2} \|c(t)(b - A(x(t) + \dot{x}(t)) - Bz^*)\|^2 \right] \\
&\quad = -\frac{1}{c(t)} \left[\frac{1}{2} \|\dot{y}(t)\|^2 + \frac{1}{2} c^2(t) \left[\|Bz(t) - Bz^*\|^2 + \|B\dot{z}(t)\|^2 + \frac{\mathrm{d}}{\mathrm{d}t} \|Bz(t) - Bz^*\|^2 \right] \right. \\
&\qquad \left. - \frac{1}{2} \|c(t)(-A(x(t) + \dot{x}(t)) + Ax^*)\|^2 \right],
\end{aligned}$$

Using the last equation, we obtain for almost every $t \in [0, +\infty)$

$$\begin{aligned}
&\langle -c(t)B^*B(\dot{z}(t)+z(t)) - c(t)B^*A(\dot{x}(t)+x(t)) + c(t)B^*b, \dot{z}(t)+z(t)-z^*\rangle - \sigma\|\dot{x}(t)+x(t)-x^*\|^2 \\
&\quad \leq -\frac{1}{c(t)}\left[\frac{1}{2}\|\dot{y}(t)\|^2 + \frac{1}{2}c^2(t)\left[\|Bz(t)-Bz^*\|^2 + \|B\dot{z}(t)\|^2 + \frac{\mathrm{d}}{\mathrm{d}t}\|Bz(t)-Bz^*\|^2\right]\right] \\
&\qquad + \left(\frac{1}{2}c(t)\|A\|^2 - \sigma\right)\|\dot{x}(t)+x(t)-x^*\|^2 \\
&\quad \leq -\frac{1}{2c(t)}\|\dot{y}(t)\|^2 - \frac{c(t)}{2}\|Bz(t)-Bz^*\|^2 - \frac{c(t)}{2}\|B\dot{z}(t)\|^2 - \frac{c(t)}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|Bz(t)-Bz^*\|^2 \\
&\qquad + \left(\frac{1}{2}c(t)\|A\|^2 - \sigma\right)\left(\|x(t)-x^*\|^2 + \|\dot{x}(t)\|^2 + \frac{\mathrm{d}}{\mathrm{d}t}\|x(t)-x^*\|^2\right).
\end{aligned} \tag{4. 32}$$

We use Lemma 4.16 to observe that for almost every $t \in [0, +\infty)$, it holds

$$\begin{aligned}
&\langle -M_1(t)\dot{x}(t), \dot{x}(t)+x(t)-x^*\rangle \\
&\qquad = -\|\dot{x}(t)\|^2_{M_1(t)} - \langle M_1(t)\dot{x}(t), x(t)-x^*\rangle \\
&\qquad = -\|\dot{x}(t)\|^2_{M_1(t)} - \frac{1}{2}\langle M_1(t)\dot{x}(t), x(t)-x^*\rangle - \frac{1}{2}\langle \dot{x}(t), M_1(t)(x(t)-x^*)\rangle \\
&\qquad = -\|\dot{x}(t)\|^2_{M_1(t)} + \frac{1}{2}\langle \dot{M}_1(t)(x(t)-x^*), x(t)-x^*\rangle - \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|x(t)-x^*\|^2_{M_1(t)}
\end{aligned}$$

and

$$\begin{aligned}
&\langle -M_2(t)\dot{z}(t), \dot{z}(t)+z(t)-z^*\rangle \\
&\qquad = -\|\dot{z}(t)\|^2_{M_2(t)} - \langle M_2(t)\dot{z}(t), z(t)-z^*\rangle \\
&\qquad = -\|\dot{z}(t)\|^2_{M_2(t)} - \frac{1}{2}\langle M_2(t)\dot{z}(t), z(t)-z^*\rangle - \frac{1}{2}\langle \dot{z}(t), M_2(t)(z(t)-z^*)\rangle \\
&\qquad = -\|\dot{z}(t)\|^2_{M_2(t)} + \frac{1}{2}\langle \dot{M}_2(t)(z(t)-z^*), z(t)-z^* - \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|z(t)-z^*\|^2_{M_2(t)}.
\end{aligned}$$

By inserting the last two identities and (4. 32) to (4. 31), we get for almost every $t \in [0, +\infty)$

$$\begin{aligned}
0 \leq & -\frac{1}{2c(t)}\|\dot{y}(t)\|^2 - \frac{c(t)}{2}\|B\dot{z}(t)\|^2 - \left(\sigma - \frac{1}{2}c(t)\|A\|^2\right)(\|x(t)-x^*\|^2 + \|\dot{x}(t)\|^2) \\
& - \frac{c(t)}{2}\|Bz(t)-Bz^*\|^2 - \frac{1}{2}\Bigg((2\sigma - c(t)\|A\|^2)\frac{\mathrm{d}}{\mathrm{d}t}\|x(t)-x^*\|^2 + \frac{\mathrm{d}}{\mathrm{d}t}\|x(t)-x^*\|^2_{M_1(t)} \\
& + c(t)\frac{\mathrm{d}}{\mathrm{d}t}\|Bz(t)-Bz^*\|^2 + \frac{\mathrm{d}}{\mathrm{d}t}\|z(t)-z^*\|^2_{M_2(t)} + \frac{1}{c(t)}\frac{\mathrm{d}}{\mathrm{d}t}\|y(t)-y^*\|^2\Bigg) - \|\dot{x}(t)\|^2_{M_1(t)} \\
& + \frac{1}{2}\langle \dot{M}_1(t)(x(t)-x^*), x(t)-x^*\rangle - \|\dot{z}(t)\|^2_{M_2(t)} + \frac{1}{2}\langle \dot{M}_2(t)(z(t)-z^*), z(t)-z^*\rangle \\
& - \frac{1}{L_{h_1}}\left\|\nabla h_1(x(t)) - \nabla h_1(x^*) + \frac{L_{h_1}}{2}\dot{x}(t)\right\|^2 + \frac{L_{h_1}}{4}\|\dot{x}(t)\|^2 \\
& - \frac{1}{L_{h_2}}\left\|\nabla h_2(z(t)) - \nabla h_2(z^*) + \frac{L_{h_2}}{2}\dot{z}(t)\right\|^2 + \frac{L_{h_2}}{4}\|\dot{z}(t)\|^2.
\end{aligned}$$

Taking into account that

$$\begin{aligned}
&-\frac{1}{2}\Big((2\sigma - c(t)\|A\|^2)\frac{\mathrm{d}}{\mathrm{d}t}\|x(t)-x^*\|^2 + \frac{\mathrm{d}}{\mathrm{d}t}\|x(t)-x^*\|^2_{M_1(t)} + c(t)\frac{\mathrm{d}}{\mathrm{d}t}\|Bz(t)-Bz^*\|^2 \\
&\quad + \frac{\mathrm{d}}{\mathrm{d}t}\|z(t)-z^*\|^2_{M_2(t)} + \frac{1}{c(t)}\frac{\mathrm{d}}{\mathrm{d}t}\|y(t)-y^*\|^2\Big) \\
&= -\frac{1}{2c(t)}\Big(\frac{\mathrm{d}}{\mathrm{d}t}((2\sigma c(t) - c^2(t)\|A\|^2)\|x(t)-x^*\|^2) - (2\dot{c}(t)\sigma - 2c(t)\dot{c}(t)\|A\|^2)\|x(t)-x^*\|^2 \\
&\quad + \frac{\mathrm{d}}{\mathrm{d}t}(c(t)\|x(t)-x^*\|^2_{M_1(t)}) - \dot{c}(t)\|x(t)-x^*\|^2_{M_1(t)} + \frac{\mathrm{d}}{\mathrm{d}t}(c^2(t)\|Bz(t)-Bz^*\|^2) \\
&\quad -2c(t)\dot{c}(t)\|Bz(t)-Bz^*\|^2 + \frac{\mathrm{d}}{\mathrm{d}t}(c(t)\|z(t)-z^*\|^2_{M_2(t)}) - \dot{c}(t)\|z(t)-z^*\|^2_{M_2(t)} + \frac{\mathrm{d}}{\mathrm{d}t}\|y(t)-y^*\|^2\Big) \\
&= -\frac{1}{2c(t)}\frac{\mathrm{d}}{\mathrm{d}t}\Big((2\sigma c(t) - c^2(t)\|A\|^2)\|x(t)-x^*\|^2 + \|x(t)-x^*\|^2_{c(t)M_1(t)} \\
&\quad + \|z(t)-z^*\|^2_{c(t)M_2(t)+c^2(t)B^*B} + \|y(t)-y^*\|^2\Big) + \frac{\dot{c}(t)}{2c(t)}\Big(2(\sigma - c(t)\|A\|^2)\|x(t)-x^*\|^2 \\
&\quad + \|x(t)-x^*\|^2_{M_1(t)} + 2c(t)\|Bz(t)-Bz^*\|^2 + \|z(t)-z^*\|^2_{M_2(t)}\Big),
\end{aligned}$$

we obtain for almost every $t \in [0, +\infty)$ that

$$\begin{aligned}
0 \leq &-\frac{1}{2c(t)}\frac{\mathrm{d}}{\mathrm{d}t}\Big((2\sigma c(t) - c^2(t)\|A\|^2)\|x(t)-x^*\|^2 + \|x(t)-x^*\|^2_{c(t)M_1(t)} \\
&+ \|z(t)-z^*\|^2_{c(t)M_2(t)+c^2(t)B^*B} + \|y(t)-y^*\|^2\Big) + \frac{\dot{c}(t)}{2c(t)}\Big(2(\sigma - c(t)\|A\|^2)\|x(t)-x^*\|^2 \\
&+ \|x(t)-x^*\|^2_{M_1(t)} + 2c(t)\|Bz(t)-Bz^*\|^2 + \|z(t)-z^*\|^2_{M_2(t)}\Big) \\
&- \frac{1}{2c(t)}\|\dot{y}(t)\|^2 - \frac{c(t)}{2}\|B\dot{z}(t)\|^2 - \Big(\sigma - \frac{1}{2}c(t)\|A\|^2\Big)(\|x(t)-x^*\|^2 + \|\dot{x}(t)\|^2) \\
&- \frac{c(t)}{2}\|Bz(t)-Bz^*\|^2 - \|\dot{x}(t)\|^2_{M_1(t)} + \frac{1}{2}\langle \dot{M}_1(t)(x(t)-x^*), x(t)-x^*\rangle \\
&- \|\dot{z}(t)\|^2_{M_2(t)} + \frac{1}{2}\langle \dot{M}_2(t)(z(t)-z^*), z(t)-z^*\rangle - \frac{1}{L_{h_1}}\left\|\nabla h_1(x(t)) - \nabla h_1(x^*) + \frac{L_{h_1}}{2}\dot{x}(t)\right\|^2 \\
&+ \frac{L_{h_1}}{4}\|\dot{x}(t)\|^2 - \frac{1}{L_{h_2}}\left\|\nabla h_2(z(t)) - \nabla h_2(z^*) + \frac{L_{h_2}}{2}\dot{z}(t)\right\|^2 + \frac{L_{h_2}}{4}\|\dot{z}(t)\|^2.
\end{aligned}$$

Since $\dot{c}(t) \leq 0$, $0 < c(t) \leq \frac{\sigma}{\|A\|^2}$, if $c(t)$ is not constant (if $c(t)$ is constant we have $\dot{c}(t) = 0$) and $\langle \dot{M}_1(t)(x(t)-x^*), x(t)-x^*\rangle \leq 0$ and $\langle \dot{M}_2(t)(z(t)-z^*), z(t)-z^*\rangle \leq 0$ (which follows easily from Definition 4.1 and the decreasing property of M_1 and M_2), we have for almost every

$t \in [0, +\infty)$

$$\begin{aligned}
0 \geq \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}&\Big((2\sigma c(t) - c^2(t)\|A\|^2)\|x(t) - x^*\|^2 + \|x(t) - x^*\|^2_{c(t)M_1(t)} \\
&+\|z(t) - z^*\|^2_{c(t)M_2(t)+c^2(t)B^*B} + \|y(t) - y^*\|^2\Big) \\
&+ c(t)\|\dot{x}(t)\|^2_{M_1(t)-\frac{L_{h_1}}{4}I} + c(t)\left(\sigma - \frac{1}{2}c(t)\|A\|^2\right)\|\dot{x}(t)\|^2 + c(t)\|\dot{z}(t)\|^2_{M_2(t)+\frac{c(t)}{2}B^*B-\frac{L_{h_2}}{4}I} \\
&+ \frac{1}{2}\|\dot{y}(t)\|^2 + c(t)\left(\sigma - \frac{1}{2}c(t)\|A\|^2\right)\|x(t) - x^*\|^2 + \frac{c^2(t)}{2}\|Bz(t) - Bz^*\|^2 \\
&+ \frac{c(t)}{L_{h_1}}\left\|\nabla h_1(x(t)) - \nabla h_1(x^*) + \frac{L_{h_1}}{2}\dot{x}(t)\right\|^2 + \frac{c(t)}{L_{h_2}}\left\|\nabla h_2(z(t)) - \nabla h_2(z^*) + \frac{L_{h_2}}{2}\dot{z}(t)\right\|^2.
\end{aligned}$$

For $\underline{c} := \epsilon$ and $\overline{c} := \frac{2\sigma}{\|A\|^2} - \epsilon$, we have

$$\begin{aligned}
0 \geq \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}&\Big((2\sigma c(t) - c^2(t)\|A\|^2)\|x(t) - x^*\|^2 + \|x(t) - x^*\|^2_{c(t)M_1(t)} \\
&+\|z(t) - z^*\|^2_{c(t)M_2(t)+c^2(t)B^*B} + \|y(t) - y^*\|^2\Big) \\
&+ \underline{c}\|\dot{x}(t)\|^2_{M_1(t)-\frac{L_{h_1}}{4}I} + \underline{c}\left(\sigma - \frac{1}{2}\overline{c}\|A\|^2\right)\|\dot{x}(t)\|^2 + \underline{c}\|\dot{z}(t)\|^2_{M_2(t)+\frac{c(t)}{2}B^*B-\frac{L_{h_2}}{4}I} + \frac{1}{2}\|\dot{y}(t)\|^2 \\
&+ \underline{c}\left(\sigma - \frac{1}{2}\overline{c}\|A\|^2\right)\|x(t) - x^*\|^2 + \frac{\underline{c}^2}{2}\|Bz(t) - Bz^*\|^2 \\
&+ \frac{\underline{c}}{L_{h_1}}\left\|\nabla h_1(x(t)) - \nabla h_1(x^*) + \frac{L_{h_1}}{2}\dot{x}(t)\right\|^2 + \frac{\underline{c}}{L_{h_2}}\left\|\nabla h_2(z(t)) - \nabla h_2(z^*) + \frac{L_{h_2}}{2}\dot{z}(t)\right\|^2.
\end{aligned} \tag{4. 33}$$

From Lemma 4.18, we have

$$\begin{aligned}
\exists \lim_{t\to+\infty}&\Big((2\sigma c(t) - c^2(t)\|A\|^2)\|x(t) - x^*\|^2 + \|x(t) - x^*\|^2_{c(t)M_1(t)} \\
&+\|z(t) - z^*\|^2_{c(t)M_2(t)+c^2(t)B^*B} + \|y(t) - y^*\|^2\Big) \in \mathbb{R}.
\end{aligned} \tag{4. 34}$$

Let $T > 0$. If we integrate (4. 33) on the interval $[0, T]$, we get

$$\begin{aligned}
\frac{1}{2}&\Big((2\sigma c(T) - c^2(T)\|A\|^2)\|x(T) - x^*\|^2 + \|x(T) - x^*\|^2_{c(T)M_1(T)} \\
&+\|z(T) - z^*\|^2_{c(T)M_2(T)+c^2(T)B^*B} + \|y(T) - y^*\|^2\Big) \\
&+ \underline{c}\int_0^T \|\dot{x}(t)\|^2_{M_1(t)-\frac{L_{h_1}}{4}I}\mathrm{d}t + \underline{c}\left(\sigma - \frac{1}{2}\overline{c}\|A\|^2\right)\int_0^T \|\dot{x}(t)\|^2\mathrm{d}t + \underline{c}\int_0^T \|\dot{z}(t)\|^2_{M_2(t)+\frac{c(t)}{2}B^*B-\frac{L_{h_2}}{4}I}\mathrm{d}t \\
&+ \frac{1}{2}\int_0^T \|\dot{y}(t)\|^2\mathrm{d}t + \underline{c}\left(\sigma - \frac{1}{2}\overline{c}\|A\|^2\right)\int_0^T \|x(t) - x^*\|^2\mathrm{d}t + \frac{\underline{c}^2}{2}\int_0^T \|Bz(t) - Bz^*\|^2\mathrm{d}t \\
&+ \frac{\underline{c}}{L_{h_1}}\int_0^T \left\|\nabla h_1(x(t)) - \nabla h_1(x^*) + \frac{L_{h_1}}{2}\dot{x}(t)\right\|^2\mathrm{d}t + \frac{\underline{c}}{L_{h_2}}\int_0^T \left\|\nabla h_2(z(t)) + \nabla h_2(z^*) + \frac{L_{h_2}}{2}\dot{z}(t)\right\|^2\mathrm{d}t \\
\leq \frac{1}{2}&\Big((2\sigma c(0) - c^2(0)\|A\|^2)\|x_0 - x^*\|^2 + \|x_0 - x^*\|^2_{c(0)M_1(0)} \\
&+\|z_0 - z^*\|^2_{c(0)M_2(0)+c^2(0)B^*B} + \|y_0 - y^*\|^2\Big).
\end{aligned}$$

Letting T converge to $+\infty$, we have

$$\|\dot{x}(\cdot)\|^2_{M_1(\cdot)-\frac{L_{h_1}}{4}I} \in L^1([0,+\infty),\mathbb{R}), \quad \|\dot{x}(\cdot)\|^2 \in L^1([0,+\infty),\mathbb{R}), \tag{4. 35}$$

$$\|\dot{z}(\cdot)\|^2_{M_2(\cdot)+\frac{c(\cdot)}{2}B^*B-\frac{L_{h_2}}{4}I} \in L^1([0,+\infty),\mathbb{R}), \tag{4. 36}$$

$$\dot{y}(\cdot) \in L^2([0,+\infty),\mathcal{K}), \tag{4. 37}$$

$$x(\cdot)-x^* \in L^2([0,+\infty),\mathcal{H}), \quad Bz(\cdot)-Bz^* \in L^2([0,+\infty),\mathcal{H}). \tag{4. 38}$$

In the case, when $L_{h_1}=0$ and $L_{h_2}>0$, we have that ∇h_1 is constant and, instead of (4. 33), we have for almost every $t \in [0,+\infty)$

$$\begin{aligned}0 \geq {} & \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\Big((2\sigma c(t)-c^2(t)\|A\|^2)\|x(t)-x^*\|^2+\|x(t)-x^*\|^2_{c(t)M_1(t)}\\ & +\|z(t)-z^*\|^2_{c(t)M_2(t)+c^2(t)B^*B}+\|y(t)-y^*\|^2\Big)\\ & +\underline{c}\|\dot{x}(t)\|^2_{M_1(t)}+\underline{c}\left(\sigma-\frac{1}{2}\bar{c}\|A\|^2\right)\|\dot{x}(t)\|^2+\underline{c}\|\dot{z}(t)\|^2_{M_2(t)+\frac{c(t)}{2}B^*B-\frac{L_{h_2}}{4}I}+\frac{1}{2}\|\dot{y}(t)\|^2\\ & +\underline{c}\left(\sigma-\frac{1}{2}\bar{c}\|A\|^2\right)\|x(t)-x^*\|^2+\frac{\underline{c}^2}{2}\|Bz(t)-Bz^*\|^2\\ & +\frac{\underline{c}}{L_{h_2}}\left\|\nabla h_2(z(t))-\nabla h_2(z^*)+\frac{L_{h_2}}{2}\dot{z}(t)\right\|^2.\end{aligned} \tag{4. 39}$$

Similarly, in the case, when $L_{h_1}>0$ and $L_{h_2}=0$, we get for almost every $t \in [0,+\infty)$

$$\begin{aligned}0 \geq {} & \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\Big((2\sigma c(t)-c^2(t)\|A\|^2)\|x(t)-x^*\|^2+\|x(t)-x^*\|^2_{c(t)M_1(t)}\\ & +\|z(t)-z^*\|^2_{c(t)M_2(t)+c^2(t)B^*B}+\|y(t)-y^*\|^2\Big)\\ & +\underline{c}\|\dot{x}(t)\|^2_{M_1(t)-\frac{L_{h_1}}{4}I}+\underline{c}\left(\sigma-\frac{1}{2}\bar{c}\|A\|^2\right)\|\dot{x}(t)\|^2+\underline{c}\|\dot{z}(t)\|^2_{M_2(t)+\frac{c(t)}{2}B^*B}+\frac{1}{2}\|\dot{y}(t)\|^2\\ & +\underline{c}\left(\sigma-\frac{1}{2}\bar{c}\|A\|^2\right)\|x(t)-x^*\|^2+\frac{\underline{c}^2}{2}\|Bz(t)-Bz^*\|^2\\ & +\frac{\underline{c}}{L_{h_1}}\left\|\nabla h_1(x(t))-\nabla h_1(x^*)+\frac{L_{h_1}}{2}\dot{x}(t)\right\|^2\end{aligned} \tag{4. 40}$$

and in the case, when $L_{h_1}=0$ and $L_{h_2}=0$, we obtain for almost every $t \in [0,+\infty)$

$$\begin{aligned}0 \geq {} & \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\Big((2\sigma c(t)-c^2(t)\|A\|^2)\|x(t)-x^*\|^2+\|x(t)-x^*\|^2_{c(t)M_1(t)}\\ & +\|z(t)-z^*\|^2_{c(t)M_2(t)+c^2(t)B^*B}+\|y(t)-y^*\|^2\Big)\\ & +\underline{c}\|\dot{x}(t)\|^2_{M_1(t)}+\underline{c}\left(\sigma-\frac{1}{2}\bar{c}\|A\|^2\right)\|\dot{x}(t)\|^2+\underline{c}\|\dot{z}(t)\|^2_{M_2(t)+\frac{c(t)}{2}B^*B}+\frac{1}{2}\|\dot{y}(t)\|^2\\ & +\underline{c}\left(\sigma-\frac{1}{2}\bar{c}\|A\|^2\right)\|x(t)-x^*\|^2+\frac{\underline{c}^2}{2}\|Bz(t)-Bz^*\|^2.\end{aligned} \tag{4. 41}$$

By arguing as above, we obtain also in these three cases that (4. 34) and (4. 35)-(4. 38) hold.

We can easily see that, if assumptions 1. or 2. from the theorem hold true, then we have $\dot{z}(\cdot) \in L^2([0,+\infty),\mathcal{G})$. Further, taking into account the hypotheses concerning $\dot{M}_1, \dot{M}_2$ and c, we can easily derive from Lemma 4.17 that

$$\ddot{x}(\cdot) \in L^2([0,+\infty),\mathcal{H}) \text{ and } \ddot{z}(\cdot) \in L^2([0,+\infty),\mathcal{G}).$$

It follows, for almost every $t \in [0,+\infty)$

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\dot{x}(t)\|^2 = 2\langle \ddot{x}(t), \dot{x}(t)\rangle \leq (\|\ddot{x}(t)\|^2 + \|\dot{x}(t)\|^2)$$

and the right-hand side is a function in $L^1([0,+\infty),\mathbb{R})$. By Lemma 4.19, we have

$$\lim_{t\to+\infty} \dot{x}(t) = 0. \tag{4. 42}$$

Analogously, we get that

$$\lim_{t\to+\infty} \dot{z}(t) = 0 \quad \lim_{t\to+\infty} (x(t) - x^*) = 0 \text{ and } \lim_{t\to+\infty} (Bz(t) - Bz^*) = 0. \tag{4. 43}$$

Because of

$$\lim_{t\to+\infty} \frac{1}{c(t)}\dot{y}(t) = \lim_{t\to+\infty} (b - A(x(t)+\dot{x}(t)) - B(z(t)+\dot{z}(t))) = b - Ax^* - Bz^*$$

and the optimality condition $Ax^* + Bz^* = b$, we have

$$\lim_{t\to+\infty} \dot{y}(t) = 0. \tag{4. 44}$$

In the following, let us prove that each weak sequential cluster point of $(x(t), z(t), y(t))$, $t \in [0,+\infty)$ is a saddle point of L (notice that the trajectories are bounded according to (4. 34)). Let $(x^*, \overline{z}, \overline{y})$ be such a weak sequential cluster point. So according to Proposition 2.1 there exists a sequence $(s_n)_{n\geq 0}$ with $s_n \to +\infty$ such that $(x(s_n), z(s_n), y(s_n))$ converges to $(x^*, \overline{z}, \overline{y})$ as $n \to +\infty$ in the weak topology of $\mathcal{H} \times \mathcal{G} \times \mathcal{K}$ (notice that the trajectory $x(t)$ converges to x^* strongly).

From (4. 3), we have for all $n \in [0,+\infty)$

$$-M_1(s_n)\dot{x}(s_n) + A^*y(s_n) - \nabla h_1(x(s_n)) \in \partial f(\dot{x}(s_n) + x(s_n)).$$

Since $(M_1(s_n))_{n\geq 0}$ is bounded, ∇h_1 is continuous, $(y(s_n))_{n\geq 0}$ converges weakly to $\overline{y}$, $\lim\limits_{t\to+\infty} \dot{x}(t) = 0$ and $\lim\limits_{t\to+\infty} x(t) = x^*$, it follows from Proposition 2.7(i)

$$A^*\overline{y} - \nabla h_1(x^*) \in \partial f(x^*).$$

From (4. 4), we obtain for every $n \geq 0$

$$B^*\dot{y}(s_n) - M_2(s_n)\dot{z}(s_n) + B^*y(s_n) - \nabla h_2(z(s_n)) + \nabla h_2(\dot{z}(s_n) + z(s_n)) \in \partial(g+h_2)(\dot{z}(s_n) + z(s_n)),$$

which is equivalent to

$$\dot{z}(s_n) + z(s_n) \in \partial(g+h_2)^*(B^*(\dot{y}(s_n) + y(s_n)) - M_2(s_n)\dot{z}(s_n) - \nabla h_2(z(s_n)) + \nabla h_2(\dot{z}(s_n) + z(s_n))).$$

By denoting for all $n \geq 0$

$$\begin{aligned}
&v_n := \dot{z}(s_n) + z(s_n),\ u_n := \dot{y}(s_n) + y(s_n)\\
&w_n := -M_2(s_n)\dot{z}(s_n) - \nabla h_2(z(s_n)) + \nabla h_2(\dot{z}(s_n) + z(s_n)),
\end{aligned}$$

we obtain

$$v_n \in \partial(g + h_2)^*(B^* u_n + w_n).$$

Since ∇h_2 is Lipschitz continuous, we get

$$\nabla h_2(\dot{z}(s_n) + z(s_n)) - \nabla h_2(z(s_n)) \to 0\ (n \to +\infty).$$

According to this fact, (4. 42), (4. 43), and (4. 44), we have $v_n \rightharpoonup \overline{z}$, $u_n \rightharpoonup \overline{y}$, $Bv_n \to B\overline{z} = Bz^*$ and $w_n \to 0$ as $n \to +\infty$. Due to the monotonicity of the subdifferential, we have for all (u, v) in the graph of $\partial(g + h_2)^*$ and for all $n \geq 0$

$$\langle v_n - v, B^* u_n + w_n - u\rangle \geq 0,$$

which is the same according to Proposition 2.2 as

$$\langle Bv_n - Bv, u_n\rangle + \langle v_n - v, w_n - u\rangle \geq 0.$$

We let $n \to +\infty$ and obtain

$$\langle B\overline{z} - Bv, \overline{y}\rangle + \langle \overline{z} - v, -u\rangle \geq 0,$$

which is equivalent to

$$\langle \overline{z} - v, B^*\overline{y} - u\rangle \geq 0\ \ \forall (u, v) \text{ in the graph of } \partial(g + h_2)^*.$$

The maximal monotonicity of the convex subdifferential of $\partial(g + h_2)^*$ ensures that $\overline{z} \in \partial(g + h_2)^*(B^*\overline{y})$, which is equivalent to $B^*\overline{y} \in \partial(g + h_2)(\overline{z})$. So we have $B^*\overline{y} - \nabla h_2(\overline{z}) \in \partial g(\overline{z})$. From (4. 1) and (4. 44) we have

$$b - A(\dot{x}(s_n) + x(s_n)) - B(\dot{z}(s_n) + z(s_n)) = \frac{1}{c(s_n)}\dot{y}(s_n) \to 0\ \ (n \to \infty)$$

and so it follows that $A\overline{x} + B\overline{z} = b$. In conclusion, $(x^*, \overline{z}, \overline{y})$ is a saddle point of the Lagrangian L.

In the following, we show that $(x(t), z(t), y(t))$, $t \in [0, +\infty)$ converges weakly. So, we consider two sequential cluster points (x^*, z_1, y_1) and (x^*, z_2, y_2). Consequently, there exists $(k_n)_{n\geq 0}$ and $(l_n)_{n\geq 0}$ such that the subsequence $(x(k_n), z(k_n), y(k_n))$ converges weakly to (x^*, z_1, y_1) as $n \to +\infty$ and $(x(l_n), z(l_n), y(l_n))$ converges weakly to (x^*, z_2, y_2) as $n \to +\infty$, respectively. As seen before, (x^*, z_1, y_1) and (x^*, z_2, y_2) are both saddle points of the Lagrangian L. From (4. 34), which is fulfilled for every saddle point of the Lagrangian L, we obtain

$$\begin{aligned}
\exists \lim_{t\to+\infty} \Big(&\|z(t) - z_1\|^2_{c(t)M_2(t)+c^2(t)B^*B} - \|z(t) - z_2\|^2_{c(t)M_2(t)+c^2(t)B^*B}\\
&+\|y(t) - y_1\|^2 - \|y(t) - y_2\|^2\Big) =: T. \qquad (4.\ 45)
\end{aligned}$$

For $t \in [0, +\infty)$, we have

$$\begin{aligned}
&\|z(t) - z_1\|^2_{c(t)M_2(t)+c^2(t)B^*B} - \|z(t) - z_2\|^2_{c(t)M_2(t)+c^2(t)B^*B} + \|y(t) - y_1\|^2 - \|y(t) - y_2\|^2\\
&\quad = \|z_2 - z_1\|^2_{c(t)M_2(t)+c^2(t)B^*B} + 2\langle z(t) - z_2, z_2 - z_1\rangle_{c(t)M_2(t)+c^2(t)B^*B} + \|y_2 - y_1\|^2\\
&\qquad + \langle y(t) - y_2, y_2 - y_1\rangle.
\end{aligned}$$

Since $c(t)M_2(t) + c^2(t)B^*B$ is monotonically decreasing and positive definite, there exists, according to Lemma 4.21, a positive definite operator M such that $c(t)M_2(t) + c^2(t)B^*B$ converges to M in the strong topology as $t \to +\infty$. Furthermore, let $c := \lim\limits_{t\to+\infty} c(t) > 0$. Taking the limits in (4. 45) along the subsequences $(k_n)_{n\geq 0}$ and $(l_n)_{n\geq 0}$, it yields

$$T = -\|z_2 - z_1\|_M^2 - \|y_2 - y_1\|^2 = \|z_2 - z_1\|_M^2 + \|y_2 - y_1\|^2,$$

so that

$$\|z_2 - z_1\|_M^2 + \|y_2 - y_1\|^2 = 0.$$

It follows that $z_1 = z_2$ and $y_1 = y_2$. In consequence, $(x(t), z(t), y(t))$ converges weakly to a saddle point of the Lagrangian L.

□

In the following corollary, we set for every $t \in [0, +\infty)$

$$M_1(t) = 0 \quad \text{and} \quad M_2(t) = \frac{1}{\tau(t)} \operatorname{Id} - c(t)B^*B,$$

where $\tau(t) > 0$ and $\tau(t)c(t)\|B\|^2 \leq 1$, like in Remark 4.8. Then, we get the following convergence result for the trajectory $(x(t), z(t), y(t))$ of the dynamical system (4. 5) as a special case of Theorem 4.22:

Corollary 4.24. *In the context of optimization problem* (3. 5)*, suppose that the set of saddle points of the Lagrangian L is nonempty, the map $\tau : [0, +\infty) \to (0, +\infty)$ is locally absolutely continuous, monotonically increasing, and fulfills $\sup_{t\geq 0} \frac{\dot{\tau}(t)}{\tau(t)^2} < \infty$. Furthermore, we assume that for an $\epsilon > 0$ the map*

$$c : [0, +\infty) \to \left[\epsilon, \frac{\sigma}{\|A\|^2} - \epsilon\right]$$

is monotonically decreasing and Lipschitz continuous. If $c(t)$ is a constant function, namely $c(t) = c$ for all $t \in [0, +\infty)$, then its enough to assume that $\epsilon \leq c \leq \frac{2\sigma}{\|A\|^2} - \epsilon$. Let

$$c(t)\tau(t)\|B\|^2 \leq 1 - \frac{\tau(t)}{4} L_{h_2}, \quad -\dot{c}(t)\|B\|^2 \leq \frac{\dot{\tau}(t)}{\tau(t)^2} \tag{4. 46}$$

for almost all $t \in [0, +\infty)$. For an arbitrary initial value $(x^0, z^0, y^0) \in \mathcal{H} \times \mathcal{G} \times \mathcal{K}$, let $(x, z, y) : [0, +\infty) \to \mathcal{H} \times \mathcal{G} \times \mathcal{K}$ be the unique strong global solution of the dynamical system (4. 5)*. If one of the following assumptions is satisfied:*

1. *$c(t)\tau(t)\|B\|^2 < 1 - \frac{\tau(t)}{4} L_{h_2}$ for all $t \in [0, +\infty)$*
2. *there exists $\beta > 0$ such that $B^*B \in P_\beta(\mathcal{G})$;*

then for $t \to +\infty$ the trajectory $(x(t), z(t), y(t))$ converges weakly to a saddle point of L.

Remark 4.25. An appropriate choice for $\tau(t)$ to fulfill the assumptions (4. 46) is, for example, $\tau(t) = \frac{a}{c(t)}$, where $0 < a \leq \frac{1}{\|B\|^2}$. That's how we set it in Example 4.10.

If $h_1 = 0$ and $h_2 = 0$, and $M_1(t) = 0$ and $M_2(t) = 0$ for all $t \geq 0$, then the dynamical system (4. 1) becomes a continuous version of the AMA method proposed by Tseng in [110] which can be written as

$$\begin{cases} \dot{x}(t)+x(t)=\operatorname{argmin}_{x\in\mathcal{H}}\left\{f(x)-\langle y(t),Ax(t)\rangle\right\} \\ \dot{z}(t)+z(t)\in\operatorname{argmin}_{z\in\mathcal{G}}\{g(z)-\langle y(t),Bz\rangle+\frac{c(t)}{2}\|A(x(t)+\dot{x}(t))+Bz-b\|^2\} \\ \dot{y}(t)=c(t)\left(b-A(x(t)+\dot{x}(t))-B(z(t)+\dot{z}(t))\right) \\ x(0)=x^0\in\mathcal{H},z(0)=z^0\in\mathcal{G},y(0)=y^0\in\mathcal{K}, \end{cases}$$

where $c(t)>0$ for all $t\in[0,+\infty)$.

According to Theorem 4.22 (for $L_{h_1}=L_{h_2}=0$), the generated trajectories converge weakly to a saddle point of the Lagrangian, if we choose the map $c(t)$ as in this theorem and if there exists $\beta>0$ such that $B^*B\in\mathcal{P}_\beta(\mathcal{G})$.

Chapter 5

Stochastic incremental mirror descent algorithms with Nesterov smoothing

This chapter is based on the paper [30] and the preprint [31].

The original mirror descent method was introduced by Nemirovski in [90] (see also [91]) as a non-Euclidean extension of the subgradient method for solving unconstrained convex optimization problems. Since then, it has been subject to various developments and employment in different areas (such as machine learning [80, 78, 101, 71], signal and image processing [5, 72, 22, 26], location research [104, 108, 105], network optimization [74], system identification [33, 58], optimal control [85]), enjoying an increasing popularity. Over these four decades, it was noticed that it is strongly connected to other iterative methods for solving various classes of optimization problems, such as FTRL (follow the regularized leader) [82], proximal gradient [118], conditional gradient [15, 96], AdaBoost [62], or dual averaging (also called *lazy mirror descent*) [75], being seen as a generalization of the proximal point algorithm with a nonlinear distance function (that could be a Bregman type one or, for instance, the Fenchel coupling [116, 117]) and an optimal stepsize (see [79]) and as a dual approach to gradient descent (see [4]). Due to their convergence properties, mirror descent algorithms proved to be especially suitable for large-scale optimization problems. In [67], two main streams of current work on mirror descent methods are identified, namely accelerating deterministic mirror descent (see, for instance, [104, 70, 4, 72]) and stochastic mirror descent with access to noised gradient oracle (like in [5, 58, 108, 80, 83, 67, 116, 117, 86, 33]).

Mirror descent type algorithms are usually employed for minimizing a single function. However, in works such as [35, 15, 96, 72, 78, 59, 58, 57, 108, 67, 70, 81, 114, 26], these methods were used for minimizing sums of (convex) functions by considering splitting techniques. This approach is used to solve problems arising in various applications from fields like machine learning or imaging. A specific feature of mirror descent type algorithms is that the convergence statements are provided in terms of values of objective functions. However, in papers like [57, 104, 106, 86, 114], the convergence of the generated iterative sequence is also investigated.

In this chapter, we propose a stochastic incremental mirror descent algorithm with Nesterov smoothing to minimize a sum of finitely many proper, convex and lower semicontinuous functions over a given nonempty closed convex set in an Euclidean space. We employ smooth approximations (via the Nesterov smoothing from [87]) of the involved functions. To the best of our knowledge, smoothing methods for the involved functions have been considered in connection to mirror descent algorithms only in [71, 72] (see also [62] for objective functions somewhat similar to the ones considered in our work), in contexts only vaguely related to our study. Then, we show that the algorithm can be modified in order to minimize over a given nonempty closed convex set in an Euclidean space a sum of finitely many proper, convex and lower semicontinuous functions composed with linear operators mapping between two Euclidean spaces. Adding to the sum a further proper, convex and lower semicontinuous function that is prox-friendly requires modifications to the previously mentioned method. The resulting algorithm becomes a stochastic incremental mirror descent Bregman-proximal scheme with Nesterov smoothing, which is further modified in order to minimize the sum over a given nonempty closed convex set in an Euclidean space of finitely many proper, convex and lower semicontinuous functions composed with linear operators, and the mentioned prox-friendly proper, convex and lower semicontinuous function. Different to the previous contributions from the literature on designing mirror descent methods for minimizing sums of functions mentioned above (in particular [35, 72, 78, 59, 57, 67, 70]), the functions we consider need not be (Lipschitz) continuous or differentiable. Moreover, our approach does not require knowledge of the Lipschitz constants or the subgradients of the involved functions, which can sometimes be computationally expensive to determine. Additionally, we show that our methods can be easily combined with the ones proposed in [35], on which we base our study. In [36], one can find a variable smoothing approach to minimize convex optimization problems with stochastic gradients, so that large scale problems can be addressed, where, different to our work, the Moreau-envelope, a special case of Nesterov smoothing, is used. In order to illustrate our theoretical achievements, we consider applications in logistics (location optimization), medical imaging (tomography), and machine learning (Support Vector Machines) modeled as optimization problems that are iteratively solved via the algorithms we propose in this work.

The mirror descent algorithm , on which we build our study, was considered in [88], called *dual averaging*. It addresses the problem of minimizing a proper and convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ over a nonempty, convex and closed set $C \subseteq \mathbb{R}^n$, involving a proper, lower semicontinuous and σ-strongly convex function (where $\sigma > 0$) $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ such that $C = \operatorname{cl}(\operatorname{dom} H)$ and $\operatorname{Im} \nabla H^*$ is a subset of the interior of the domain of f. The iterative scheme is as follows (where x_0 lies in the interior of the domain of f, $y_0 \in \mathbb{R}^n$ and $t_k > 0$, $k \geq 0$, are positive stepsizes)

$$(\forall k \geq 0) \ \begin{cases} y_{k+1} = y_k - t_k f'(x_k), \\ x_{k+1} = \nabla H^*(y_{k+1}), \end{cases}$$

where $f'(x_k)$ is a subgradient of f at x_k. As noted in [35], this scheme generalizes the classical subgradient method and is close to the subgradient projection algorithm.

5.1 A stochastic incremental mirror descent algorithm with Nesterov smoothing

Problem 5.1. *We consider the convex optimization problem*

$$\min_{x \in C} \left\{ \sum_{i=1}^{m} f_i(x) \right\}, \tag{5. 1}$$

where $C \subseteq \mathbb{R}^n$ is a nonempty, convex and closed set and for all $i = 1, \ldots, m$, ($m \in \mathbb{N}$) $f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ fulfills

$$f_i(x) = \max_{u \in U_i} \{ \langle A_i x, u \rangle - \phi_i(u) \}, \ x \in \operatorname{dom} f_i, \tag{5. 2}$$

where $U_i \subseteq \mathbb{R}^p$ is compact and convex, $A_i : \mathbb{R}^n \to \mathbb{R}^p$ is linear and $\phi_i : \mathbb{R}^p \to \overline{\mathbb{R}}$ a proper, lower semicontinuous and convex function. We assume that $C \cap (\cap_{i=1}^m \operatorname{dom} f_i) \neq \emptyset$.

Furthermore, let $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, lower semicontinuous and σ-strongly convex function (for $\sigma > 0$) such that $C = \operatorname{cl}(\operatorname{dom} H)$ and $\operatorname{Im} \nabla H^* \subseteq \cap_{i=1}^m \operatorname{dom} f_i$.

Due to the fact that H is a proper, lower semicontinuous and σ-strongly convex function, its conjugate function H^* is Fréchet differentiable and its gradient ∇H^* is *σ-cocoercive* and ∇H^* is $(1/\sigma)$-Lipschitz continuous. In the algorithms we propose in this chapter, we have the map ∇H^* as mirror map , which is induced by the function H. This map mirrors each iterate onto the feasible set C. So we can choose $H(x) = \frac{1}{2}\|x\|^2$, for $x \in C$ and $H(x) = +\infty$, otherwise, to obtain for the mirror map ∇H^* the orthogonal projection onto C. When $C = \mathbb{R}^n$, the map ∇H^* reduces to the identity operator. However, one can choose other mirror maps depending and taking advantage on the structure of C and the considered optimization problem. For further examples of mirror maps for corresponding sets C, please refer to the applications presented in section 5.4.

Remark 5.2. The construction (5. 2) guarantees that the functions f_i, $i = 1, \dots, m$, are proper, convex and lower semicontinuous. From the Fenchel-Moreau-Theorem (see Proposition 2.9 (ii)), it follows that for every proper, lower semicontinuous and convex function $f : \mathbb{R}^p \to \overline{\mathbb{R}}$, one has

$$f \circ A(\cdot) = \sup_{u \in \operatorname{dom} f^*} \{\langle A\cdot, u\rangle - f^*(u)\},$$

where $A : \mathbb{R}^n \to \mathbb{R}^p$ is a linear operator. A maximum, as in (5. 2), can be guaranteed, for instance, when $\operatorname{dom} f^*$ is bounded. This is fulfilled, for example, when f is Lipschitz continuous, while the opposite implication is not known to hold. For deeper insights and examples of this construction, we refer the reader to [87, 89]. Moreover, in works such as [7, 115, 25], this approach is employed to design algorithms for solving various classes of optimization problems, some of which stemming from concrete applications.

To minimize the sum of the nonsmooth convex functions f_i $(i = 1, \dots, m)$ in Problem 5.1, at first we approximate them by smooth functions. For this, we use the Nesterov smoothing technique (see [87], also employed in works like [109, 7, 99, 115]). We choose to employ a mirror descent type technique due to the known qualities of these methods, the suitability of the considered functions to our approach, and the fact that in many applications only certain convergence properties of the values of the objective function, best obtained via mirror descent, are relevant.

Definition 5.3. For $i = 1, \dots, m$, a function $b_{U_i} : \mathbb{R}^p \to \overline{\mathbb{R}}$ is called *prox-function* of the compact set $U_i \subseteq \mathbb{R}^p$, if $U_i \subset \operatorname{dom}(b_{U_i})$ and if b_{U_i} is continuous and β-strongly convex over the set U_i $(\beta > 0)$. Its *prox-center* is denoted by $u_i^c = \operatorname{argmin}_{u \in U_i} b_{U_i}(u)$ and its *prox-diameter* by
$D_{U_i} = \sup_{u \in U_i} b_{U_i}(u)$.

Without loss of generality, we set in the following $\beta = 1$ and assume that for all $i = 1, \dots, m$, $b_{U_i}(u_i^c) = 0$ and therefore $b_{U_i}(u) \geq 0$ for all $u \in U_i$.

Next, we approximate the functions f_i $(i = 1, \dots, m)$ by the smooth functions $f_i^\gamma : \mathbb{R}^n \to \mathbb{R}$

$$f_i^\gamma(x) = \max_{u \in U_i}\{\langle A_i x, u\rangle - \phi_i(u) - \gamma b_{U_i}(u)\}, \tag{5. 3}$$

where $\gamma > 0$ is the *smoothing parameter*. This procedure originates from [87] (see also [89]) and is called *Nesterov smoothing*. We define

$$u_i^\gamma(x) = \operatorname*{argmax}_{u \in U_i}\{\langle A_i x, u\rangle - \phi_i(u) - \gamma b_{U_i}(u)\}.$$

Furthermore, it holds

$$f_i^\gamma(x) \le f_i(x) \le f_i^\gamma(x) + \gamma D_{U_i} \quad \forall x \in \operatorname{dom} f_i. \tag{5. 4}$$

Lemma 5.4. *The functions f_i^γ, $i = 1, \dots, m$, defined as above, are well defined, convex and continuously differentiable at every $x \in U_i$. Furthermore, $\nabla f_i^\gamma = A_i^* u_i^\gamma$ which is $\|A_i\|^2/\gamma$-Lipschitz continuous, and it holds*

$$\|\nabla f_i^\gamma(x)\|^2 \le 2\|A_i\|^2(2D_{U_i} + \|u_i^c\|^2) \quad \forall x \in \mathbb{R}^n.$$

Proof. For the first part of the proof see [87, Theorem 1], where the continuity and finiteness of f_i, $i = 1, \dots, m$, imposed in the hypothesis, were not employed. It remains only the inequality to be shown.
For $i \in \{1, \dots, m\}$, and $x \in \mathbb{R}^n$ it holds

$$\|\nabla f_i^\gamma(x)\|^2 \le \|A_i\|^2 \|u_i^\gamma(x)\|^2 \le \|A_i\|^2 \left(2\|u_i^\gamma(x) - u_i^c\|^2 + 2\|u_i^c\|^2\right).$$

Due to the 1-strong convexity of b_{U_i} and the inequality (2. 1), we have

$$\|u_i^\gamma(x) - u_i^c\|^2 \le 2b_{U_i}(u_i^\gamma(x)) - 2b_{U_i}(u_i^c) - 2\nabla b_{U_i}(u_i^c)(u_i^\gamma(x) - u_i^c)$$

and, taking into consideration, that $b_{U_i}(u_i^c) = 0$ and that $\nabla b_{U_i}(u_i^c) = 0$, it follows from this inequality that

$$\|u_i^\gamma(x) - u_i^c\|^2 \le 2b_{U_i}(u_i^\gamma(x)) \le 2D_{U_i}.$$

Hence

$$\|\nabla f_i^\gamma(x)\|^2 \le 2\|A_i\|^2(2D_{U_i} + \|u_i^c\|^2).$$

□

Remark 5.5. Notice that for $i = 1, \dots, m$, $g_i : \mathbb{R}^p \to \overline{\mathbb{R}}$, $\phi_i = g_i^*, b_{U_i} = (1/2)\|\cdot\|^2$ and $U_i = \operatorname{dom} g_i^*$ is compact and convex for a $\gamma > 0$, the function

$$f_i^\gamma(x) = (g_i \Box (1/(2\gamma))\|\cdot\|^2)(A_i x)$$

is the Moreau-envelope of $g_i \circ A_i$ and

$$\nabla f_i^\gamma(x) = (1/\gamma) A_i^* \left(A_i x - \operatorname{Prox}_{\gamma g_i}(A_i x)\right),$$

$\forall x \in \mathbb{R}^n$. In this case $u_i^c = 0$.

Remark 5.6. Other smoothing methods, such as the general one presented in [25], could also be employed in the framework we consider in this chapter, as long as they guarantee the last result from Lemma 5.4. This lemma states that the norms of the gradients of the smooth approximations of the considered functions are bounded.

For the convergence analysis of the following algorithms, we use two measures of distance in the sense of Bregman.

Definition 5.7. Let $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper and convex function. The *Bregman-distance-like function* of H is denoted as

$$d_H : \mathbb{R}^n \times \operatorname{dom} H \times \mathbb{R}^n \to \overline{\mathbb{R}}, \quad d_H(x, y, z) := H(x) - H(y) - \langle z, x - y \rangle.$$

Due to the subgradient inequality, it holds that $d_H(x, y, z) \ge 0$ for every
$(x, y) \in \mathbb{R}^n \times \operatorname{dom} H$ and all $z \in \partial H(y)$.

Definition 5.8. The *Bregman distance* associated to a proper and convex function $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ fulfilling $\operatorname{dom} \nabla H := \{x \in \mathbb{R}^n : H \text{ is differentiable at } x\} \neq \emptyset$ is defined as

$$D_H : \mathbb{R}^n \times \operatorname{dom} \nabla H \to \overline{\mathbb{R}}, \quad D_H(x,y) := H(x) - H(y) - \langle \nabla H(y), x - y \rangle,$$

which was introduced by Bregman in 1967 (see [43]). The following algorithm relies on the stochastic incremental mirror descent approach of [35, Algorithm 3.2], but instead of using subgradients of the functions f_i, we smooth them by the Nesterov smoothing approach (5. 3) and employ the gradients of the smooth functions, provided by Lemma 5.4. We choose the smoothing parameters $\gamma_k = \frac{\delta}{\sigma} t_k$, where t_k is the step size, σ is the strongly convex parameter of H, and $\delta > 0$ is a constant parameter .

Algorithm 5.9

```
Choose $x_0 \in \bigcap_{i=1}^{m} \operatorname{dom} f_i \cap C$, $y_{m,-1} \in \mathbb{R}^n$, the parameter $\delta > 0$ and the stepsizes $t_k > 0, k \geq 0$:
for all $k \geq 0$ do
  $\psi_{0,k} := x_k$
  $y_{0,k} := y_{m,k-1}$
  for all $i := 1, \ldots, m$ do
    $y_{i,k} := y_{i-1,k} - \epsilon_{i,k} \frac{t_k}{p_i} \nabla f_i^{t_k \delta / \sigma}(\psi_{i-1,k})$
    $\psi_{i,k} := \nabla H^*(y_{i,k})$
  end for
  $x_{k+1} := \psi_{m,k}$
end for,
where $\epsilon_{i,k} \in \{0,1\}$ is a random variable independent of $\psi_{i-1,k}$ and $\mathbb{P}(\epsilon_{i,k} = 1) = p_i$ for all $1 \leq i \leq m$ and $k \geq 0$.
```

Remark 5.10. The hypothesis $\operatorname{Im} \nabla H^* \subseteq \cap_{i=1}^m \operatorname{dom} f_i$ guarantees that the sequence $\{x_k\}_k$ generated by Algorithm 5.9 contains only elements that lie in the intersection of the domains of the functions f_i, $i = 1, \ldots, m$.

Theorem 5.11. *For Problem 5.1, let the sequence $\{x_k\}_k$ generated by Algorithm 5.9. Then for all $N \geq 1$ and $y \in \mathbb{R}^n$, it holds*

$$\mathbb{E}\left(\min_{0 \leq k \leq N-1} \sum_{i=1}^m f_i(x_k) - \sum_{i=1}^m f_i(y)\right)$$
$$\leq \frac{d_H(y, x_0, y_{0,0}) + \frac{1}{\sigma}\left(\delta \sum_{i=1}^m D_{U_i} + 2\left(\sum_{i=1}^m \|A_i\| \sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\right) \sum_{k=0}^{N-1} t_k^2}{\sum_{k=0}^{N-1} t_k}.$$

Proof. Arguing as in in the proof of [35, Theorem 3.3] and replacing the estimates $\|f_i'(\psi_{i-1,k})\|^2 \leq L_{f_i}^2$ with $\|\nabla f_i^{t_k\delta/\sigma}\|^2 \leq 2\|A_i\|^2(2D_{U_i} + \|u_i^c\|^2)$, we arrive at

$$\mathbb{E}(d_H(y, \psi_{m,k}, y_{m,k})) \leq \mathbb{E}(d_H(y, x_k, y_{0,k})) + t_k \mathbb{E}\left(\sum_{i=1}^m f_i^{t_k\delta/\sigma}(y) - \sum_{i=1}^m f_i^{t_k\delta/\sigma}(x_k)\right)$$

$$+\frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2)}\right)^2\left(\sum_{i=1}^m\frac{1}{p_i^2}\right)^{\frac{1}{2}} - \mathbb{E}\left(\sum_{i=1}^m \frac{1}{2} d_H(\psi_{i,k},\psi_{i-1,k},y_{i-1,k})\right)$$
$$+\mathbb{E}\left(t_k\sum_{i=1}^m(f_i^{t_k\delta/\sigma}(x_k)-f_i^{t_k\delta/\sigma}(\psi_{i-1,k}))\right), \qquad (5.\ 5)$$

which is a modification of [(6),[35]].

We have for every $k \geq 0$

$$\begin{aligned}\sum_{i=1}^m(f_i^{t_k\delta/\sigma}(x_k)-f_i^{t_k\delta/\sigma}(\psi_{i-1,k})) &\leq \sum_{i=2}^m\sum_{j=1}^{i-1}(f_i^{t_k\delta/\sigma}(\psi_{j-1,k})-f_i^{t_k\delta/\sigma}(\psi_{j,k}))\\ &\leq \sum_{i=2}^m\sum_{j=1}^{i-1}\langle\nabla f_i^{t_k\delta/\sigma}(\psi_{j-1,k}),\psi_{j-1,k}-\psi_{j,k}\rangle\\ &\leq \sum_{i=2}^m\sum_{j=1}^{i-1}\|\nabla f_i^{t_k\delta/\sigma}(\psi_{j-1,k})\|\|\psi_{j-1,k}-\psi_{j,k}\|.\end{aligned}$$

Using Lemma 5.4, we obtain from the inequality above for every $k \geq 0$

$$\begin{aligned}\sum_{i=1}^m(f_i^{t_k\delta/\sigma}(x_k)-f_i^{t_k\delta/\sigma}(\psi_{i-1,k})) &\leq \sum_{i=2}^m\sum_{j=1}^{i-1}\|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}\|\psi_{j-1,k}-\psi_{j,k}\|\\ &\leq \sum_{l=1}^m\|A_l\|\sqrt{4D_{U_l}+2\|u_l^c\|^2}\sum_{i=2}^m\|\psi_{i-1,k}-\psi_{i,k}\|.\end{aligned}$$

Furthermore, using the Lipschitz continuity of ∇H^*, it yields for every $2 \leq i \leq m$ and for every $k \geq 0$

$$\|\psi_{i-1,k}-\psi_{i,k}\| = \|\nabla H^*(y_{i-1,k})-\nabla H^*(y_{i,k})\| \leq \frac{1}{\sigma}\|y_{i-1,k}-y_{i,k}\| = \frac{1}{\sigma}\left\|\epsilon_{i,k}\frac{t_k}{p_i}\nabla f_i^{t_k\delta/\sigma}(\psi_{i-1,k})\right\|.$$

Inserting the last inequality in the inequality above, we obtain for every $k \geq 0$

$$\begin{aligned}\sum_{i=1}^m(f_i^{t_k\delta/\sigma}(x_k)-f_i^{t_k\delta/\sigma}(\psi_{i-1,k})) &\leq \frac{1}{\sigma}\sum_{l=1}^m\|A_l\|\sqrt{4D_{U_l}+2\|u_l^c\|^2}\sum_{i=2}^m\left\|\epsilon_{i,k}\frac{t_k}{p_i}\nabla f_i^{t_k\delta/\sigma}(\psi_{i-1,k})\right\|\\ &\leq \frac{1}{\sigma}t_k\sum_{l=1}^m\|A_l\|\sqrt{4D_{U_l}+2\|u_l^c\|^2}\sum_{i=1}^m\frac{\epsilon_{i,k}}{p_i}\|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}.\end{aligned}$$

So, for every $k \geq 0$ it holds

$$\begin{aligned}&\mathbb{E}\left(t_k\sum_{i=1}^m(f_i^{t_k\delta/\sigma}(x_k)-f_i^{t_k\delta/\sigma}(\psi_{i-1,k}))\right)\\ &\leq \frac{1}{\sigma}t_k^2\left(\sum_{l=1}^m\|A_l\|\sqrt{4D_{U_l}+2\|u_l^c\|^2)}\right)\mathbb{E}\left(\sum_{i=1}^m\frac{\epsilon_{i,k}}{p_i}\|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}\right)\\ &\leq \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m\|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}\right)^2. \qquad (5.\ 6)\end{aligned}$$

Inequality (5. 4) yields for every $k \geq 0$

$$t_k \mathbb{E}\left(\sum_{i=1}^{m} f_i^{t_k\delta/\sigma}(y) - \sum_{i=1}^{m} f_i^{t_k\delta/\sigma}(x_k)\right) \leq t_k \left(\mathbb{E}\left(\sum_{i=1}^{m} f_i(y) - \sum_{i=1}^{m} f_i(x_k)\right) + \frac{t_k\delta}{\sigma}\sum_{i=1}^{m} D_{U_i}\right). \qquad (5.\ 7)$$

Combining (5. 5) with (5. 6) and (5. 7) gives for every $k \geq 0$

$$\begin{aligned}
&\mathbb{E}(d_H(y, \psi_{m,k}, y_{m,k}))\\
&\quad \leq \mathbb{E}(d_H(y, x_k, y_{0,k})) + t_k \left(\mathbb{E}\left(\sum_{i=1}^{m} f_i(y) - \sum_{i=1}^{m} f_i(x_k)\right) + \frac{t_k\delta}{\sigma}\sum_{i=1}^{m} D_{U_i}\right)\\
&\quad\quad + \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 \left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} - \mathbb{E}\left(\sum_{i=1}^{m} \frac{1}{2} d_H(\psi_{i,k}, \psi_{i-1,k}, y_{i-1,k})\right)\\
&\quad\quad + \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 .
\end{aligned}$$

Because $\psi_{m,k} = x_{k+1}, y_{m,k} = y_{0,k+1}$ and $d_H(\psi_{i,k}, \psi_{i-1,k}, y_{i-1,k}) \geq 0$ as $y_{i-1,k} \in \partial H(\psi_{i-1,k})$, it holds for every $k \geq 0$

$$\begin{aligned}
\mathbb{E}(d_H(y, x_{k+1}, y_{0,k+1})) \leq &\mathbb{E}(d_H(y, x_k, y_{0,k})) + t_k \mathbb{E}\left(\sum_{i=1}^{m} f_i(y) - \sum_{i=1}^{m} f_i(x_k)\right) + \frac{\delta}{\sigma} t_k^2 \sum_{i=1}^{m} D_{U_i}\\
&+ \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right). \qquad (5.\ 8)
\end{aligned}$$

Summing up the inequality from $k = 0$ to $N - 1$, where $N \geq 1$, we obtain

$$\begin{aligned}
&\sum_{k=0}^{N-1} t_k \mathbb{E}\left(\sum_{i=1}^{m} f_i(x_k) - \sum_{i=1}^{m} f_i(y)\right) + \mathbb{E}(d_H(y, x_N, y_{0,N}))\\
&\quad \leq \mathbb{E}(d_H(y, x_0, y_{0,0})) + \frac{\delta}{\sigma}\sum_{i=1}^{m} D_{U_i} \sum_{k=0}^{N-1} t_k^2\\
&\quad\quad + \frac{1}{\sigma}\left(\sum_{i=1}^{m} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) \sum_{k=0}^{N-1} t_k^2.
\end{aligned}$$

Since $d_H(y, x_N, y_{0,N}) \geq 0$, as $y_{0,N} \in \partial H(x_N)$, we get

$$\begin{aligned}
&\mathbb{E}\left(\min_{0 \leq k \leq N-1} \sum_{i=1}^{m} f_i(x_k) - \sum_{i=1}^{m} f_i(y)\right)\\
&\leq \frac{d_H(y, x_0, y_{0,0}) + \frac{1}{\sigma}\left(\delta \sum_{i=1}^{m} D_{U_i} + 2\left(\sum_{i=1}^{m} \|A_i\| \sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\right) \sum_{k=0}^{N-1} t_k^2}{\sum_{k=0}^{N-1} t_k}.
\end{aligned}$$

□

In the following corollary, we give the optimal stepsize choice for Algorithm 5.9, which follows from [23, Proposition 4.1].

Corollary 5.12. *Let $x^* \in \operatorname{dom} H$ be an optimal solution to* (5. 1) *and a constant $\delta > 0$. Then, the optimal stepsize for the algorithm above is given by*

$$t_k := \sqrt{\frac{\sigma d_H(x^*, x_0, y_{0,0})}{\delta \sum\limits_{i=1}^m D_{U_i} + 2\left(\sum\limits_{i=1}^m \|A_i\|\sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum\limits_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)}} \cdot \frac{1}{\sqrt{k}} \quad \forall k \geq 0,$$

which yields for every $N \geq 1$

$$\mathbb{E}\left(\min_{0\leq k\leq N-1} \sum_{i=1}^m f_i(x_k) - \sum_{i=1}^m f_i(x^*)\right)$$
$$\leq \sqrt{\frac{d_H(x^*, x_0, y_{0,0})\left(\delta \sum\limits_{i=1}^m D_{U_i} + 2\left(\sum\limits_{i=1}^m \|A_i\|\sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum\limits_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\right)}{\sigma}} \cdot \frac{2}{\sqrt{N}}.$$

Let us consider now the following optimization problem consisting in minimizing a sum of functions fulfilling (5. 2) composed with linear operators. Such problems can be seen both as special cases and generalizations of Problem 5.1, as mentioned in remark 5.2. Taking this remark into consideration, the maximum in the construction (5. 2) only needs to be attained in the case of such compositions when the involved functions are proper, convex and semicontinuous, and the operators are linear. In this case, we say that they fulfill the property (5.2′). Unlike the construction proposed in [35], our approach is flexible enough to allow modifying Algorithm 5.9 in order to solve such problems as well.

Problem 5.13. *We consider the convex optimization problem*

$$\min_{x\in C}\left\{\sum_{i=1}^m f_i(A_i x)\right\}, \tag{5. 9}$$

where $C \subseteq \mathbb{R}^n$ is a nonempty, convex and closed set, $f_i : \mathbb{R}^p \to \overline{\mathbb{R}}$, $i = 1, \ldots, m$, are proper, convex and semicontiuous functions, and $A_i : \mathbb{R}^n \to \mathbb{R}^p$ are linear operators such that (5.2′) *holds for them, and $C \cap (\cap_{i=1}^m \operatorname{dom}(f_i \circ A_i)) \neq \emptyset$.*

For $i = 1, \ldots, m$, we smooth the functions $f_i \circ A_i$ via the Moreau envelope, which is a special case of Nesterov smoothing as mentioned above. This results in

$$(f_i^\gamma \circ A_i)(x) = (f_i \square (1/2\gamma)\|\cdot\|^2)(A_i x)$$

with the gradients

$$\nabla(f_i^\gamma \circ A_i)(x) = (1/\gamma)A_i^*(A_i x - \operatorname{Prox}_{\gamma f_i}(A_i x))$$

for all $x \in \mathbb{R}^n$, where $\gamma > 0$.
Then, we obtain the following mirror descent proximal point algorithm.

Algorithm 5.14

```
Choose $x_0 \in \bigcap_{i=1}^{m} \operatorname{dom}(f_i \circ A_i) \cap C$, $y_{m,-1} \in \mathbb{R}^n$, the parameter $\delta > 0$, and the stepsizes $t_k > 0$, $k \geq 0$:
for all $k \geq 0$ do
  $\psi_{0,k} := x_k$
  $y_{0,k} := y_{m,k-1}$
  for all $i := 1, \ldots, m$ do
    $y_{i,k} := y_{i-1,k} - \epsilon_{i,k} \frac{\sigma}{\delta p_i} A_i^* \left( A_i \psi_{i-1,k} - \operatorname{Prox}_{t_k \frac{\delta}{\sigma} f_i} (A_i \psi_{i-1,k}) \right)$
    $\psi_{i,k} := \nabla H^*(y_{i,k})$
  end for
  $x_{k+1} := \psi_{m,k}$
end for,
where $\epsilon_{i,k} \in \{0,1\}$ is a random variable independent of $\psi_{i-1,k}$ and $\mathbb{P}(\epsilon_{i,k} = 1) = p_i$ for all $1 \leq i \leq m$ and $k \geq 0$.
```

Because this algorithm is derived from Algorithm 5.9, the convergence result of Theorem 5.11 is also valid, where $D_{U_i} = D_{\operatorname{dom} f_i^*} = \sup_{u \in \operatorname{dom} f_i^*} \frac{1}{2}\|u\|^2$ and $\|u_i^c\| = 0$:

Theorem 5.15. *For Problem 5.13, let the sequence $\{x_k\}_k$ generated by Algorithm 5.14 and a constant $\delta > 0$, $k \geq 0$. Then for all $N \geq 1$ and $y \in \mathbb{R}^n$, it holds*

$$\mathbb{E}\left(\min_{0 \leq k \leq N-1} \sum_{i=1}^{m} f_i(x_k) - \sum_{i=1}^{m} f_i(y) \right)$$
$$\leq \frac{d_H(y, x_0, y_{0,0}) + \frac{1}{\sigma}\left(\delta \sum_{i=1}^{m} D_{dom f_i^*} + 4 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{D_{dom f_i^*}} \right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2} \right)^{\frac{1}{2}} + 1 \right) \right) \sum_{k=0}^{N-1} t_k^2}{\sum_{k=0}^{N-1} t_k}.$$

So, the optimal stepsize choice for Algorithm 5.14 is given by the following corollary:

Corollary 5.16. *Let $x^* \in \operatorname{dom} H$ be an optimal solution to* (5. 9) *and a constant $\delta > 0$, $k \geq 0$. Then, the optimal stepsize for Algorithm 5.14 above is given by*

$$t_k := \sqrt{\frac{\sigma d_H(x^*, x_0, y_{0,0})}{\delta \sum_{i=1}^{m} D_{dom f_i^*} + 4 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{D_{dom f_i^*}} \right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2} \right)^{\frac{1}{2}} + 1 \right)}} \cdot \frac{1}{\sqrt{k}} \quad \forall k \geq 0,$$

which yields for every $N \geq 1$

$$\mathbb{E}\left(\min_{0 \leq k \leq N-1} \sum_{i=1}^{m} f_i(A_i x_k) - \sum_{i=1}^{m} f_i(A_i x^*) \right)$$
$$\leq 2\sqrt{\frac{d_H(x^*, x_0, y_{0,0}) \left(\delta \sum_{i=1}^{m} D_{dom f_i^*} + 4 \left(\sum_{i=1}^{m} \|A_i\| \sqrt{D_{dom f_i^*}} \right)^2 \left(\left(\sum_{i=1}^{m} \frac{1}{p_i^2} \right)^{\frac{1}{2}} + 1 \right) \right)}{\sigma}} \cdot \frac{1}{\sqrt{N}}.$$

Remark 5.17. The difference between Algorithm 5.9 and its counterpart in [35, Algorithm 2.2] is that we do not need to know the Lipschitz constants or the subgradients of the functions f_i ($i = 1, \dots, m$), which sometimes can be computationally expensive to determine (cf. [3, 46, 64, 95]). Instead, our approach relies on their proximal point mappings (in particular for its special case Algorithm 5.14). For many functions, including those commonly encountered in applications in fields like image deblurring and denoising or machine learning, these mappings are already known. A further advantage of our method is that we do not need to impose the Lipschitz continuity of the gradients of the objective functions, as the gradients of their Nesterov smooth approximations satisfy this hypothesis by construction. Instead, we ask the weaker condition of closedness of the domains of their conjugates. Note also that, by employing the parameter $\delta > 0$, $k \geq 0$, Algorithm 5.9 presents additional flexibility in comparison to its mentioned counterpart.

Remark 5.18. Moreover, assuming Lipschitz continuity for the functions f_i, $i = 1, \dots, m$, does not make Algorithm 5.9 collapse to [35, Algorithm 3.2]. Similarly, the assertion of Theorem 5.11 does not rediscover its counterpart [35, Theorem 3.3] due to the distinct constructions. This has motivated us to include in our study the results in Section 5.3, where we present combinations of these algorithms.

5.2 Incremental mirror descent Bregman-prox-scheme with Nesterov smoothing

In this section, we consider an extension of the optimization problem (5. 1) by adding another nonsmooth function to its objective function. The iterative scheme we propose for solving is an extension of Algorithm 5.9. But instead of smoothing the new function, we evaluate it using a proximal step of Bregman type. For this, we require additional differentiability assumptions on the function that induces the mirror map.

Problem 5.19. *We consider the convex optimization problem*

$$\min_{x \in C} \left\{ \sum_{i=1}^{m} f_i(x) + g(x) \right\}, \tag{5. 10}$$

where $C \subseteq \mathbb{R}^n$ is a nonempty, convex and closed set. For $i = 1, \dots, m$, the functions $f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ are defined like in Problem 5.1 and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper, convex and lower semicontinuous function such that $C \cap (\cap_{i=1}^m \operatorname{dom} f_i \cap \operatorname{dom} g) \neq \emptyset$. Furthermore, let $H : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, lower semicontinuous and σ-strongly convex function (for $\sigma > 0$) such that $C = \operatorname{cl}(\operatorname{dom} H)$, let H be continuously differentiable on $\operatorname{int}(\operatorname{dom} H)$, $\operatorname{Im} \nabla H^* \subseteq (\cap_{i=1}^m \operatorname{dom} f_i) \cap \operatorname{int}(\operatorname{dom} H)$ *and* $\operatorname{int}(\operatorname{dom} H) \cap \operatorname{dom} g \neq \emptyset$.

Definition 5.20. Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper, convex and lower semicontinuous function. The *Bregman-proximal operator* of h with respect to the proper, lower semicontinuous and σ-strongly convex function H is defined as

$$\operatorname{Prox}_h^H : \operatorname{dom} \nabla H \to \mathbb{R}^n, \quad \operatorname{Prox}_h^H(x) := \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \{h(u) + D_H(u, x)\}.$$

Because H is σ-strongly convex, the Bregman-proximal operator is well defined. For $H = (1/2)\| \cdot \|^2$ the Bregman-proximal operator is the classical proximity operator.
We propose the following algorithm for solving the optimization problem (5. 10).

Algorithm 5.21

```
Choose $x_0 \in \operatorname{Im} \nabla H^* \cap C$, the parameter $\delta > 0$ and the stepsizes $t_k > 0$, $k \geq 0$:
for all $k \geq 0$ do
    $\psi_{0,k} := x_k$
    for all $i := 1, \ldots, m$ do
        $\psi_{i,k} := \nabla H^*(\nabla H(\psi_{i-1,k}) - \epsilon_{i,k} \frac{t_k}{p_i} \nabla f_i^{t_k \delta/\sigma}(\psi_{i-1,k}))$
    end for
    $x_{k+1} := \operatorname{Prox}_{t_k g}^H(\psi_{m,k})$
end for,
where $\epsilon_{i,k} \in \{0,1\}$ is a random variable independent of $\psi_{i-1,k}$ and $\mathbb{P}(\epsilon_{i,k} = 1) = p_i$ for all $1 \leq i \leq m$ and $k \geq 0$.
```

Remark 5.22. Note that when $g = 0$, Algorithm 5.21 corresponds essentialy to Algorithm 5.9. But even for this case, the constants obtained in the convergence result given below and in Theorem 5.11 are not the same, due to the construction of the algorithms (note, for instance, that Algorithm 5.9 requires an additional starting point) and therefore there are some main differences in the proofs.

Theorem 5.23. *Let the sequence $\{x_k\}_k$ generated by Algorithm 5.21 and a constant $\delta > 0$. Then for all $N \geq 1$ and all $y \in \mathbb{R}^n$, one has*

$$\mathbb{E}\left(\min_{0\leq k\leq N-1}\left(\sum_{i=1}^m f_i + g\right)(x_{k+1}) - \left(\sum_{i=1}^m f_i + g\right)(y)\right)$$
$$\leq \frac{D_H(y,x_0) + \frac{1}{\sigma}\left(\delta \sum\limits_{i=1}^m D_{U_i} + 2\left(\sum\limits_{i=1}^m \|A_i\|\sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum\limits_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + \frac{3}{2} + m\right)\right)\sum\limits_{k=0}^{N-1} t_k^2}{\sum\limits_{k=0}^{N-1} t_k}.$$

Proof. When $y \notin \cap_{i=1}^m \operatorname{dom} f_i \cap \operatorname{dom} g$, the assertion follows automatically, so we will now consider the case when $y \in \cap_{i=1}^m \operatorname{dom} f_i \cap \operatorname{dom} g$. We start the proof with inequalities (5. 5) and (5. 6) from Theorem 5.11 and use instead of the Bregman distance like functions the Bregman distance to obtain

$$\mathbb{E}(D_H(y,\psi_{m,k})) \leq \mathbb{E}(D_H(y,x_k)) + t_k \mathbb{E}\left(\sum_{i=1}^m f_i^{\delta/\sigma t_k}(y) - \sum_{i=1}^m f_i^{\delta/\sigma t_k}(x_k)\right)$$

$$+ \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) - \mathbb{E}\left(\sum_{i=1}^m \frac{1}{2} D_H(\psi_{i,k}, \psi_{i-1,k})\right). \quad (5.\ 11)$$

Like in [(12),[35]] we get for every $k \geq 0$

$$t_k \mathbb{E}((g(x_{k+1}) - g(y))) + \mathbb{E}(D_H(y,x_{k+1})) \leq \mathbb{E}(D_H(y,\psi_{m,k})) - \mathbb{E}(D_H(x_{k+1},\psi_{m,k})). \quad (5.\ 12)$$

By combining (5. 11) and (5. 12), we obtain for every $k \geq 0$,

$$\begin{aligned}
&t_k\mathbb{E}((g(x_{k+1}) - g(y))) + t_k\mathbb{E}\left(\sum_{i=1}^m f_i^{\delta/\sigma t_k}(x_k) - \sum_{i=1}^m f_i^{\delta/\sigma t_k}(y)\right) + \mathbb{E}(D_H(y, x_{k+1})) \\
&\leq \mathbb{E}(D_H(y, x_k)) + \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2\left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) \\
&\quad - \mathbb{E}(D_H(x_{k+1}, \psi_{m,k})) - \sum_{i=1}^m \frac{1}{2}\mathbb{E}(D_H(\psi_{i,k}, \psi_{i-1,k})).
\end{aligned}$$

We add and subtract $t_k\mathbb{E}(\sum_{i=1}^m f_i^{\delta/\sigma t_k}(x_{k+1}))$ to get

$$\begin{aligned}
&t_k\mathbb{E}\left(\left(\sum_{i=1}^m f_i^{\delta/\sigma t_k} + g\right)(x_{k+1}) - \left(\sum_{i=1}^m f_i^{\delta/\sigma t_k} + g\right)(y)\right) \\
&\quad + t_k\mathbb{E}\left(\sum_{i=1}^m f_i^{\delta/\sigma t_k}(x_k) - \sum_{i=1}^m f_i^{\delta/\sigma t_k}(x_{k+1})\right) + \mathbb{E}(D_H(y, x_{k+1})) \\
&\quad \leq \mathbb{E}(D_H(y, x_k)) + \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2\left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) \\
&\qquad - \mathbb{E}(D_H(x_{k+1}, \psi_{m,k})) - \sum_{i=1}^m \frac{1}{2}\mathbb{E}(D_H(\psi_{i,k}, \psi_{i-1,k})).
\end{aligned}$$

Due to the differentiability and convexity of $f_i^{\delta/\sigma t_k}$ for $i = 1, \ldots, m$ and for all $k \geq 0$, along with Lemma 5.4, it follows that

$$\begin{aligned}
-t_k\mathbb{E}\left(\sum_{i=1}^m f_i^{\delta/\sigma t_k}(x_{k+1}) - \sum_{i=1}^m f_i^{\delta/\sigma t_k}(x_k)\right) &\geq -t_k\mathbb{E}\left(\left\|\sum_{i=1}^m \nabla f_i^{\delta/\sigma t_k}(x_{k+1})\right\|\|x_k - x_{k+1}\|\right) \\
&\geq -t_k\mathbb{E}\left(\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\|x_k - x_{k+1}\|\right) \\
&\geq -t_k\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\mathbb{E}\left(\|x_k - x_{k+1}\|\right)
\end{aligned} \tag{5. 13}$$

and from (5. 4), we have that

$$\begin{aligned}
&t_k\left(\mathbb{E}\left(\left(\sum_{i=1}^m f_i + g\right)(x_{k+1}) - \left(\sum_{i=1}^m f_i + g\right)(y)\right) - \frac{\delta}{\sigma}t_k\sum_{i=1}^m D_{U_i}\right) \\
&\quad - t_k\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\mathbb{E}\left(\|x_k - x_{k+1}\|\right) + \mathbb{E}(D_H(y, x_{k+1})) \\
&\quad \leq \mathbb{E}(D_H(y, x_k)) + \frac{1}{\sigma}t_k^2\left(\sum_{i=1}^m \|A_i\|\sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2\left(\left(\sum_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) \\
&\qquad - \mathbb{E}(D_H(x_{k+1}, \psi_{m,k})) - \sum_{i=1}^m \frac{1}{2}\mathbb{E}(D_H(\psi_{i,k}, \psi_{i-1,k})).
\end{aligned} \tag{5. 14}$$

By the triangle inequality, we get for every $k \geq 0$

$$t_k \sum_{i=1}^m \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \mathbb{E}\left(\|x_k - x_{k+1}\|\right)$$
$$\leq t_k \sum_{i=1}^m \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \mathbb{E}\left(\|x_k - \psi_{m,k}\|\right) + t_k \sum_{i=1}^m \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \mathbb{E}\left(\|\psi_{m,k} - x_{k+1}\|\right). \tag{5. 15}$$

Using Young's inequality and the strong convexity of H, we have

$$\begin{aligned} t_k \sum_{i=1}^m \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \mathbb{E}\left(\|\psi_{m,k} - x_{k+1}\|\right) \\ \leq \frac{1}{2\sigma} t_k^2 \left(\sum_{i=1}^m \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 + \frac{\sigma}{2} \mathbb{E}\left(\|\psi_{m,k} - x_{k+1}\|\right)^2 \\ \leq \frac{1}{2\sigma} t_k^2 \left(\sum_{i=1}^m \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 + \mathbb{E}(H(x_{k+1}) - H(\psi_{m,k}) - \langle \nabla H(x_{k+1}), x_{k+1} - \psi_{m,k}\rangle) \\ = \frac{1}{2\sigma} t_k^2 \left(\sum_{i=1}^m \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 + \mathbb{E}(D_H(x_{k+1}, \psi_{m,k})), \end{aligned}$$

and since

$$\|x_k - \psi_{m,k}\| = \left\| \sum_{i=1}^m (\psi_{i-1,k} - \psi_{i,k}) \right\| \leq \sum_{i=1}^m \|\psi_{i-1,k} - \psi_{i,k}\|,$$

the inequality (5. 15) becomes

$$t_k \sum_{i=1}^m \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \mathbb{E}\left(\|x_k - x_{k+1}\|\right) \leq \frac{1}{2\sigma} t_k^2 \left(\sum_{i=1}^m \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2$$
$$+\mathbb{E}(D_H(x_{k+1}, \psi_{m,k})) \quad + t_k \sum_{i=1}^m \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \mathbb{E}\left(\sum_{i=1}^m \|\psi_{i-1,k} - \psi_{i,k}\|\right).$$

Using again Young's inequality and the strong convexity of H, we get for every $i = 1, \ldots, m$, and every $k \geq 0$

$$t_k \sum_{i=1}^m \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \|\psi_{i-1,k} - \psi_{i,k}\| \leq \frac{1}{\sigma} t_k^2 \left(\sum_{j=1}^m \|A_j\| \sqrt{4D_{U_j} + 2\|u_j^c\|^2}\right)^2$$
$$+\frac{\sigma}{4} \|\psi_{i-1,k} - \psi_{i,k}\| \leq \frac{1}{\sigma} t_k^2 \left(\sum_{j=1}^m \|A_j\| \sqrt{4D_{U_j} + 2\|u_j^c\|^2}\right)^2 + \frac{1}{2} D_H(\psi_{i,k}, \psi_{i-1,k}).$$

So, we have

$$t_k \sum_{i=1}^m \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \mathbb{E}\left(\|x_k - x_{k+1}\|\right) \leq \frac{1}{2\sigma} t_k^2 \left(\sum_{i=1}^m \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2$$

$$+\,\mathbb{E}(D_H(x_{k+1},\psi_{m,k}))+\frac{1}{\sigma}mt_k^2\left(\sum_{j=1}^{m}\|A_j\|\sqrt{4D_{U_j}+2\|u_j^c\|^2}\right)^2+\sum_{i=1}^{m}\frac{1}{2}D_H(\psi_{i,k},\psi_{i-1,k}). \qquad (5.\ 16)$$

Combining (5. 16) and (5. 14), we obtain

$$t_k\mathbb{E}\left(\left(\sum_{i=1}^{m}f_i+g\right)(x_{k+1})-\left(\sum_{i=1}^{m}f_i+g\right)(y)\right)+\mathbb{E}(D_H(y,x_{k+1}))\leq\mathbb{E}(D_H(y,x_k))$$
$$+\frac{1}{\sigma}t_k^2\left(\sum_{i=1}^{m}\|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}\right)^2\left(\left(\sum_{i=1}^{m}\frac{1}{p_i^2}\right)^{\frac{1}{2}}+\frac{3}{2}+m\right)+t_k^2\frac{\delta}{\sigma}\sum_{i=1}^{m}D_{U_i}.$$

Summing up this inequality from $k=0$ to $N-1$, for $N\geq 1$, we get

$$\sum_{k=0}^{N-1}t_k\mathbb{E}\left(\left(\sum_{i=1}^{m}f_i+g\right)(x_{k+1})-\left(\sum_{i=1}^{m}f_i+g\right)(y)\right)+\mathbb{E}(D_H(y,x_N))\leq\mathbb{E}(D_H(y,x_0))$$
$$+\frac{1}{\sigma}\left(\sum_{i=1}^{m}\|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}\right)^2\left(\left(\sum_{i=1}^{m}\frac{1}{p_i^2}\right)^{\frac{1}{2}}+\frac{3}{2}+m\right)\sum_{k=0}^{N-1}t_k^2+\sum_{i=1}^{m}D_{U_i}\sum_{k=0}^{N-1}t_k^2\frac{\delta}{\sigma}.$$

Since $\mathbb{E}(D_H(y,x_N))\geq 0$, we obtain

$$\mathbb{E}\left(\min_{0\leq k\leq N-1}\left(\sum_{i=1}^{m}f_i+g\right)(x_{k+1})-\left(\sum_{i=1}^{m}f_i+g\right)(y)\right)$$
$$\leq\frac{D_H(y,x_0)+\frac{1}{\sigma}\left(\delta\sum\limits_{i=1}^{m}D_{U_i}+2\left(\sum\limits_{i=1}^{m}\|A_i\|\sqrt{2D_{U_i}+\|u_i^c\|^2}\right)^2\left(\left(\sum\limits_{i=1}^{m}\frac{1}{p_i^2}\right)^{\frac{1}{2}}+\frac{3}{2}+m\right)\right)\sum\limits_{k=0}^{N-1}t_k^2}{\sum\limits_{k=0}^{N-1}t_k}.$$

□

In the following corollary, we give the optimal stepsize choice for Algorithm 5.21, which follows from [23, Proposition 4.1].

Corollary 5.24. *Let $x^*\in\operatorname{dom}H$ be an optimal solution to* (5. 10) *and a constant $\delta>0$. Then, the optimal stepsize for Algorithm 5.21 is given by*

$$t_k:=\sqrt{\frac{\sigma D_H(x^*,x_0)}{\delta\sum\limits_{i=1}^{m}D_{U_i}+2\left(\sum\limits_{i=1}^{m}\|A_i\|\sqrt{2D_{U_i}+\|u_i^c\|^2}\right)^2\left(\left(\sum\limits_{i=1}^{m}\frac{1}{p_i^2}\right)^{\frac{1}{2}}+\frac{3}{2}+m\right)}}\cdot\frac{1}{\sqrt{k}}\quad\forall k\geq 0,$$

which yields for every $N\geq 1$

$$\mathbb{E}\left(\min_{0\leq k\leq N-1}\left(\sum_{i=1}^{m}f_i+g\right)(x_k)-\left(\sum_{i=1}^{m}f_i+g\right)(x^*)\right)$$
$$\leq 2\sqrt{\frac{D_H(x^*,x_0)\left(\delta\sum\limits_{i=1}^{m}D_{U_i}+2\left(\sum\limits_{i=1}^{m}\|A_i\|\sqrt{2D_{U_i}+\|u_i^c\|^2}\right)^2\left(\left(\sum\limits_{i=1}^{m}\frac{1}{p_i^2}\right)^{\frac{1}{2}}+\frac{3}{2}+m\right)\right)}{\sigma}}\cdot\frac{1}{\sqrt{N}}.$$

Similar to the previous section, we modify the considered problem by composing the smoothed functions with the linear operators used in their construction.

Problem 5.25. *We consider the convex optimization problem*

$$\min_{x\in C}\left\{\sum_{i=1}^{m} f_i(A_i x) + g(x)\right\},$$

where $C \subseteq \mathbb{R}^n$ is a nonempty, convex and closed set, $f_i : \mathbb{R}^p \to \overline{\mathbb{R}}$, for $i = 1,\dots,m$, are proper, convex and lower semicontinuous functions, $A_i : \mathbb{R}^n \to \mathbb{R}^p$ linear operators such that (5.2′) holds for them and
$g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper, convex and lower semicontinuous function such that
$C \cap (\cap_{i=1}^m \operatorname{dom}(f_i \circ A_i)) \cap \operatorname{dom} g \neq \emptyset$. Furthermore, let the function H defined as in Problem 5.19.

By smoothing the functions $f_i \ \ (i = 1,\dots,m)$ via the Moreau envelope, we obtain

$$(f_i^\gamma \circ A_i)(x) = (f_i \Box (1/2\gamma)\|\cdot\|^2)(A_i x) \quad \forall x \in \mathbb{R}^n$$

with the gradients

$$\nabla(f_i^\gamma \circ A_i)(x) = (1/\gamma)A_i^*(A_i x - \operatorname{Prox}_{\gamma f_i}(A_i x)) \quad \forall x \in \mathbb{R}^n,$$

as in the previous section. Then, we obtain from Algorithm 5.21 the following mirror descent proximal point algorithm for solving Problem 5.25.

Algorithm 5.26

```
Choose $x_0 \in \operatorname{Im}\nabla H^* \cap C$, the parameter $\delta > 0$ and the stepsizes $t_k > 0$, $k \geq 0$:
for all $k \geq 0$ do
    $\psi_{0,k} := x_k$
    for all $i := 1,\dots,m$ do
        $\psi_{i,k} := \nabla H^*(\nabla H(\psi_{i-1,k}) - \epsilon_{i,k}\frac{\sigma}{\delta p_i}A_i^*\left(A_i\psi_{i-1,k} - \operatorname{Prox}_{t_k\frac{\delta}{\sigma}f_i}(A_i\psi_{i-1,k})\right)$
    end for
    $x_{k+1} := \operatorname{Prox}^H_{t_k g}(\psi_{m,k})$,
end for
where $\epsilon_{i,k} \in \{0,1\}$ is a random variable independent of $\psi_{i-1,k}$ and $\mathbb{P}(\epsilon_{i,k} = 1) = p_i$ for all $1 \leq i \leq m$ and $k \geq 0$.
```

Because this algorithm is derived from Algorithm 5.21, the convergence result follows directly from Theorem 5.23, where $D_{U_i} = D_{\operatorname{dom} f_i^*} = \sup_{u\in\operatorname{dom} f_i^*}\frac{1}{2}\|u\|^2$ and $\|u_i^c\| = 0$:

Theorem 5.27. *Let the sequence $\{x_k\}_k$ generated by Algorithm 5.26 and a constant $\delta > 0$. Then for all $N \geq 1$ and all $y \in \mathbb{R}^n$, one has*

$$\mathbb{E}\left(\min_{0\leq k\leq N-1}\left(\sum_{i=1}^m f_i\circ A_i + g\right)(x_{k+1}) - \left(\sum_{i=1}^m f_i\circ A_i + g\right)(y)\right)$$

$$\leq \frac{D_H(y,x_0) + \frac{1}{\sigma}\left(\delta\sum\limits_{i=1}^m D_{dom f_i^*} + 4\left(\sum\limits_{i=1}^m \|A_i\|\sqrt{D_{dom f_i^*}}\right)^2\left(\left(\sum\limits_{i=1}^m \frac{1}{p_i^2}\right)^{\frac{1}{2}} + \frac{3}{2} + m\right)\right)\sum\limits_{k=0}^{N-1} t_k^2}{\sum\limits_{k=0}^{N-1} t_k}.$$

So, the optimal stepsize choice for Algorithm 5.26 is given by the following corollary:

Corollary 5.28. *Let $x^* \in \operatorname{dom} H$ be an optimal solution to Problem 5.25 and a constant $\delta > 0$. Then, the optimal stepsize for Algorithm 5.26 is given by*

$$t_k := \sqrt{\frac{\sigma D_H(y, x_0)}{\delta \sum\limits_{i=1}^{m} D_{domf_i^*} + 4\left(\sum\limits_{i=1}^{m} \|A_i\| \sqrt{D_{domf_i^*}}\right)^2 \left(\left(\sum\limits_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + \frac{3}{2} + m\right)}} \cdot \frac{1}{\sqrt{k}} \quad \forall k \geq 0,$$

which yields for every $N \geq 1$

$$\mathbb{E}\left(\min_{0 \leq k \leq N-1}\left(\sum_{i=1}^{m} f_i \circ A_i + g\right)(x_k) - \left(\sum_{i=1}^{m} f_i \circ A_i + g\right)(x^*)\right)$$

$$\leq 2\sqrt{\frac{D_H(y, x_0)\left(\delta \sum\limits_{i=1}^{m} D_{domf_i^*} + 4\left(\sum\limits_{i=1}^{m} \|A_i\| \sqrt{D_{domf_i^*}}\right)^2 \left(\left(\sum\limits_{i=1}^{m} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + \frac{3}{2} + m\right)\right)}{\sigma}} \cdot \frac{1}{\sqrt{N}}.$$

Remark 5.29. Analogous to [35, Remark 3.7 and Remark 4.7], one can establish corresponding statements within the framework presented in this work. We leave them to the interested reader.

5.3 Stochastic incremental mirror descent algorithms with subgradient and Nesterov smoothing

In the following, we combine the mirror descent algorithms proposed above, which use the Nesterov smoothing approach, and the mirror descent algorithms proposed in [35], which rely on the subgradients of the objective functions for minimization.

Problem 5.30. *We consider the convex optimization problem*

$$\min_{x \in C}\left\{\sum_{i=1}^{m_1} f_i(x) + \sum_{i=m_1+1}^{m_2} f_i(x)\right\}, \tag{5. 17}$$

where $C \subseteq \mathbb{R}^n$ is a nonempty, convex and closed set such that $C \cap (\cap_{i=1}^{m_2} \operatorname{dom} f_i) \neq \emptyset$, for all $i = 1, \ldots, m_1$ ($m_1 \in \mathbb{N}$), the functions $f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ are proper, convex and L_{f_i}-Lipschitz continuous on $\operatorname{Im} \nabla H^*$, where H is defined as in Problem 5.19, and for all $i = m_1 + 1, \ldots, m_2$ ($m_1 \leq m_2 \in \mathbb{N}$), the functions $f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ fulfill $f_i(x) = \max_{u \in U_i}\{\langle A_i x, u\rangle - \phi_i(u)\}$ for $x \in \operatorname{dom} f_i$, where $U_i \subseteq \mathbb{R}^p$ is a compact and convex set, $A_i : \mathbb{R}^n \to \mathbb{R}^p$ are linear operators and $\phi_i : \mathbb{R}^p \to \overline{\mathbb{R}}$ are proper, lower semicontinuous and convex functions.

For the following algorithm, we use the subgradients of the first m_1 functions f_i and the gradients of the smooth functions $f_i^{t_k\delta/\sigma}$ for $i = m_1 + 1, \ldots, m_2$.

Algorithm 5.31

```
Choose $x_0 \in \bigcap_{i=1}^{m_2} \operatorname{dom} f_i \cap C$, $y_{m_2,-1} \in \mathbb{R}^n$, the parameter $\delta > 0$ and the stepsizes $t_k > 0, k \geq 0$:
for all $k \geq 0$ do
  $\psi_{0,k} := x_k$
  $y_{0,k} := y_{m_2,k-1}$
  for all $i := 1, \dots, m_1$ do
    $y_{i,k} := y_{i-1,k} - \epsilon_{i,k}\frac{t_k}{p_i} f_i'(\psi_{i-1,k})$
    $\psi_{i,k} := \nabla H^*(y_{i,k})$
  end for
  for all $i := m_1 + 1, \dots, m_2$ do
    $y_{i,k} := y_{i-1,k} - \epsilon_{i,k}\frac{t_k}{p_i} \nabla f_i^{t_k\delta/\sigma}(\psi_{i-1,k})$
    $\psi_{i,k} := \nabla H^*(y_{i,k})$
  end for
  $x_{k+1} := \psi_{m_2,k}$
end for,
where $\epsilon_{i,k} \in \{0,1\}$ is a random variable independent of $\psi_{i-1,k}$ and $\mathbb{P}(\epsilon_{i,k} = 1) = p_i$ for all $1 \leq i \leq m_2$ and $k \geq 0$.
```

In the following statement, we give the convergence result for this algorithm. The proof is basically a combination of the ones of Theorem 5.11 and [35, Theorem 3.3].

Theorem 5.32. *For Problem 5.30, let the sequence $\{x_k\}_k$ generated by the algorithm above and $\delta > 0$. Then for all $N \geq 1$ and $y \in \mathbb{R}^n$, it holds*

$$\mathbb{E}\left(\min_{0\leq k\leq N-1} \sum_{i=1}^{m_2} f_i(x_k) - \sum_{i=1}^{m_2} f_i(y)\right) \leq \frac{d_H(y, x_0, y_{0,0}) + \frac{1}{\sigma} C \sum_{k=0}^{N-1} t_k^2}{\sum_{k=0}^{N-1} t_k},$$

where

$$\begin{aligned} C &= \delta \sum_{i=m_1+1}^{m_2} D_{U_i} + \left(\sum_{i=1}^{m_1} L_{f_i}\right)^2 \left(\left(\sum_{i=1}^{m_1} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) \\ &\quad + 2\left(\sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=m_1+1}^{m_2} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right). \end{aligned}$$

Proof. For the functions f_i for $i = 1, \dots m_1$, we have from the proof of [[35], Theorem 3.3] the inequality (8)

$$\begin{aligned}
&\mathbb{E}(d_H(y, \psi_{m_1,k}, y_{m_1,k})) \\
&\quad \leq \mathbb{E}(d_H(y, x_k, y_{0,k})) + t_k \mathbb{E}\left(\sum_{i=1}^{m_1} f_i(y) - \sum_{i=1}^{m_1} f_i(x_k)\right) + \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^{m_1} L_{f_i}\right)^2 \left(\sum_{i=1}^{m_1} \frac{1}{p_i^2}\right)^{\frac{1}{2}} \\
&\quad - \mathbb{E}\left(\sum_{i=1}^{m_1} \frac{1}{2} d_H(\psi_{i,k}, \psi_{i-1,k}, y_{i-1,k})\right) + \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^{m_1} L_{f_i}\right)^2,
\end{aligned}$$

and obtain

$$\begin{aligned}
&\mathbb{E}(d_H(y, \psi_{m_1,k}, y_{m_1,k})) \\
&\quad \leq \mathbb{E}(d_H(y, x_k, y_{0,k})) + t_k \mathbb{E}\left(\sum_{i=1}^{m_1} f_i(y) - \sum_{i=1}^{m_1} f_i(x_k)\right) + \frac{1}{\sigma} t_k^2 \left(\sum_{i=1}^{m_1} L_{f_i}\right)^2 \left(\left(\sum_{i=1}^{m_1} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right).
\end{aligned} \tag{5. 18}$$

Furthermore, we have from equation (5. 8) in the proof of Theorem 5.11 for the functions f_i for $i = m_1 + 1, \dots m_2$ (note that that $\psi_{m_2,k} = x_{k+1}, y_{m_2,k} = y_{0,k+1}$)

$$\begin{aligned}
&\mathbb{E}(d_H(y, x_{k+1}, y_{0,k+1})) \\
&\quad \leq \mathbb{E}(d_H(y, \psi_{m_1,k}, y_{m_1,k})) + t_k \mathbb{E}\left(\sum_{i=m_1+1}^{m_2} f_i(y) - \sum_{i=m_1+1}^{m_2} f_i(x_k)\right) + t_k^2 \frac{\delta}{\sigma} \sum_{i=m_1+1}^{m_2} D_{U_i} \\
&\quad + \frac{1}{\sigma} t_k^2 \left(\sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=m_1+1}^{m_2} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right).
\end{aligned} \tag{5. 19}$$

Inserting (5. 18) in (5. 19), we obtain

$$\begin{aligned}
&\mathbb{E}(d_H(y, x_{k+1}, y_{0,k+1})) \\
&\quad \leq \mathbb{E}(d_H(y, x_k, y_{0,k})) + t_k \mathbb{E}\left(\sum_{i=1}^{m_2} f_i(y) - \sum_{i=1}^{m_2} f_i(x_k)\right) + t_k^2 \frac{\delta}{\sigma} \sum_{i=m_1+1}^{m_2} D_{U_i} \\
&\quad + \frac{t_k^2}{\sigma} \left(\left(\sum_{i=1}^{m_1} L_{f_i}\right)^2 \left(\left(\sum_{i=1}^{m_1} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\right. \\
&\quad \left. + \left(\sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=m_1+1}^{m_2} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\right).
\end{aligned}$$

Summing up this inequality from $k = 0$ to $N - 1$, where $N \geq 1$, we obtain

$$\begin{aligned}\sum_{k=0}^{N-1} t_k \mathbb{E}\left(\sum_{i=1}^{m_2} f_i(x_k) - \sum_{i=1}^{m_2} f_i(y)\right) + \mathbb{E}(d_H(y, x_N, y_{0,N}))\\ \leq \mathbb{E}(d_H(y, x_0, y_{0,0})) + \sum_{i=m_1+1}^{m_2} \frac{\delta}{\sigma} D_{U_i} \sum_{k=0}^{N-1} t_k^2 + \frac{1}{\sigma} \sum_{k=0}^{N-1} t_k^2 \left(\left(\sum_{i=1}^{m_1} L_{f_i}\right)^2 \left(\left(\sum_{i=1}^{m_1} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\right.\\ \left. +2\left(\sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=m_1+1}^{m_2} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\right).\end{aligned}$$

Since $d_H(y, x_N, y_{0,N}) \geq 0$, as $y_{0,N} \in \partial H(x_N)$, we obtain

$$\mathbb{E}\left(\min_{0 \leq k \leq N-1} \sum_{i=1}^{m_2} f_i(x_k) - \sum_{i=1}^{m_2} f_i(y)\right) \leq \frac{d_H(y, x_0, y_{0,0}) + \frac{1}{\sigma} C \sum_{k=0}^{N-1} t_k^2}{\sum_{k=0}^{N-1} t_k},$$

where

$$\begin{aligned} C &= \delta \sum_{i=m_1+1}^{m_2} D_{U_i} + \left(\sum_{i=1}^{m_1} L_{f_i}\right)^2 \left(\left(\sum_{i=1}^{m_1} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\\ &\quad +2\left(\sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=m_1+1}^{m_2} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right).\end{aligned}$$

□

The optimal stepsize choice for Algorithm 5.31 can be deduced from [23, Proposition 4.1].

Corollary 5.33. *Let $x^* \in \operatorname{dom} H$ be an optimal solution to (5. 20), $\delta > 0$, $k \geq 0$, and*

$$\begin{aligned} P &:= \delta \sum_{i=m_1+1}^{m_2} D_{U_i} + \left(\sum_{i=1}^{m_1} L_{f_i}\right)^2 \left(\left(\sum_{i=1}^{m_1} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)\\ &\quad +2\left(\sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{2D_{U_i} + \|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=m_1+1}^{m_2} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right).\end{aligned}$$

Then, the optimal stepsize for the algorithm above is given by

$$t_k := \sqrt{\frac{\sigma d_H(x^*, x_0, y_{0,0})}{P}} \cdot \frac{1}{\sqrt{k}},$$

for all $k \geq 0$, which yields for every $N \geq 1$

$$\mathbb{E}\left(\min_{0 \leq k \leq N-1} \sum_{i=1}^{m_2} f_i(x_k) - \sum_{i=1}^{m_2} f_i(x^*)\right) \leq 2\sqrt{\frac{d_H(x^*, x_0, y_{0,0}) P}{\sigma}} \cdot \frac{1}{\sqrt{N}}.$$

Adding another nonsmooth function to the objective function of Problem 5.30 brings into attention the following problem, which can be solved by the algorithm below it.

Problem 5.34. *We consider the convex optimization problem*

$$\min_{x\in C}\left\{\sum_{i=1}^{m_1} f_i(x)+\sum_{i=m_1+1}^{m_2} f_i(x)+g(x)\right\}, \tag{5. 20}$$

where ($m_1+1<m_2\in\mathbb{N}$), $C\subseteq\mathbb{R}^n$ is a nonempty, convex and closed set, the functions $f_i:\mathbb{R}^n\to\overline{\mathbb{R}}$ (for $i=1,\dots m_1$) and $f_i:\mathbb{R}^n\to\overline{\mathbb{R}}$ (for $i=m_1+1,\dots m_2$) are defined like in Problem 5.30 and $g:\mathbb{R}^n\to\overline{\mathbb{R}}$ is a proper, convex and lower semicontinuous function such that $C\cap(\cap_{i=1}^{m_2}\operatorname{dom} f_i\cap\operatorname{dom} g)\neq\emptyset$. Let $H:\mathbb{R}^n\to\overline{\mathbb{R}}$ be defined like in Problem 5.19.

Algorithm 5.35

```
Choose $x_0 \in \operatorname{Im}\nabla H^* \cap C$, $y_{m_2,-1}\in\mathbb{R}^n$, the parameter $\delta>0$, the stepsizes $t_k>0, k\geq 0$:
for all $k\geq 0$ do
  $\psi_{0,k} := x_k$
  $y_{0,k} := y_{m_2,k-1}$
  for all $i := 1,\dots,m_1$ do
    $\psi_{i,k} := \nabla H^*\left(\nabla H(\psi_{i-1,k}) - \epsilon_{i,k}\frac{t_k}{p_i} f_i'(\psi_{i-1,k})\right)$
  end for
  for all $i := m_1+1,\dots,m_2$ do
    $\psi_{i,k} := \nabla H^*\left(\nabla H(\psi_{i-1,k}) - \epsilon_{i,k}\frac{t_k}{p_i}\nabla f_i^{t_k\delta/\sigma}(\psi_{i-1,k})\right)$
  end for
  $x_{k+1} := \operatorname{Prox}^H_{t_k g}(\psi_{m_2,k})$.
end for,
where $\epsilon_{i,k}\in\{0,1\}$ is random variable independent of $\psi_{i-1,k}$ and let $\mathbb{P}(\epsilon_{i,k}=1)=p_i$ for all $1\leq i\leq m_2$ and $k\geq 0$.
```

The convergence result and the optimal stepsize t_k, $k\geq 0$, for this algorithm are derivable via Theorem 5.23 and [35, Theorem 4.5], and [23, Proposition 4.1], respectively.

Theorem 5.36. *Let the sequence $\{x_k\}_k$ generated by Algorithm 5.35 and $\delta>0$. Then for all $N\geq 1$,*

$$\mathbb{E}\left(\min_{0\leq k\leq N-1}\left(\sum_{i=1}^{m_2} f_i+g\right)(x_{k+1})-\left(\sum_{i=1}^{m_2} f_i+g\right)(y)\right)\leq\frac{D_H(y,x_0)+\frac{1}{\sigma}C\sum\limits_{k=0}^{N-1}t_k^2}{\sum\limits_{k=0}^{N-1}t_k},$$

where

$$P=\ \delta\sum_{i=m_1+1}^{m_2} D_{U_i}+\left(\sum_{i=m_1+1}^{m_2}\|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}\right)^2\left(\left(\sum_{i=m_1+1}^{m_2}\frac{1}{p_i^2}\right)^{\frac{1}{2}}+1\right)$$
$$+\left(\sum_{i=1}^{m_1}L_{f_i}\right)^2\left(\left(\sum_{i=1}^{m_1}\frac{1}{p_i^2}\right)^{\frac{1}{2}}+1\right)+\frac{3}{2}\left(\left(\sum_{i=1}^{m_1}L_{f_i}\right)+\sum_{i=m_1+1}^{m_2}\|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}\right)^2.$$

Proof. Combining (5. 18), (5. 5), (5. 6) and (5. 12) (for $\psi_{m_2,k}$ instead of $\psi_{m,k}$), we get

$$\begin{aligned}
&t_k\mathbb{E}((g(x_{k+1})-g(y))) + t_k\mathbb{E}\left(\sum_{i=1}^{m_1} f_i(x_k) - \sum_{i=1}^{m_1} f_i(y) + \sum_{i=m_1+1}^{m_2} f_i^{t_k\delta/\sigma}(x_k) - \sum_{i=m_1+1}^{m_2} f_i^{t_k\delta/\sigma}(y)\right)\\
&\quad + \mathbb{E}(D_H(y,x_{k+1}))\\
&\quad \le \mathbb{E}(D_H(y,x_k)) + \frac{1}{\sigma}t_k^2\left(\sum_{i=m_1+1}^{m_2} \|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}\right)^2\left(\left(\sum_{i={}_1+1}^{m_2}\frac{1}{p_i^2}\right)^{\frac{1}{2}}+1\right).\\
&\quad + \left(\sum_{i=1}^{m_1} L_{f_i}\right)^2\left(\left(\sum_{i=1}^{m_1}\frac{1}{p_i^2}\right)^{\frac{1}{2}}+1\right)\Bigg) - \mathbb{E}(D_H(x_{k+1},\psi_{m_2,k})) - \mathbb{E}\left(\sum_{i=m_1+1}^{m_2}\frac{1}{2}D_H(\psi_{i,k},\psi_{i-1,k})\right).
\end{aligned}$$

By adding and subtracting $t_k\mathbb{E}(\sum_{i=1}^{m_1} f_i(x_{k+1}))$ and $t_k\mathbb{E}(\sum_{i=m_1+1}^{m_2} f_i^{t_k\delta/\sigma}(x_{k+1}))$, and considering (5. 13), the Lipschitz continuity of $\sum_{i=1}^{m_1} f_i$, and (5. 4), we get

$$\begin{aligned}
&t_k\left(\mathbb{E}\left(\left(\sum_{i=1}^{m_2} f_i + g\right)(x_{k+1}) - \left(\sum_{i=1}^{m_2} f_i + g\right)(y)\right) - t_k\frac{\delta}{\sigma}\sum_{i=m_1+1}^{m_2} D_{U_i}\right)\\
&\quad - t_k\sum_{i=m_1+1}^{m_2}\|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}\mathbb{E}\left(\|x_k-x_{k+1}\|\right) - t_k\sum_{i=1}^{m_1} L_{f_i}\mathbb{E}\left(\|x_k-x_{k+1}\|\right)\\
&\quad + \mathbb{E}(D_H(y,x_{k+1}))\\
&\quad \le \mathbb{E}(D_H(y,x_k)) + \frac{1}{\sigma}t_k^2\left(\sum_{i=m_1+1}^{m_2}\|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}\right)^2\left(\left(\sum_{i={}_1+1}^{m_2}\frac{1}{p_i^2}\right)^{\frac{1}{2}}+1\right).\\
&\quad + \left(\sum_{i=1}^{m_1} L_{f_i}\right)^2\left(\left(\sum_{i=1}^{m_1}\frac{1}{p_i^2}\right)^{\frac{1}{2}}+1\right)\Bigg) - \mathbb{E}(D_H(x_{k+1},\psi_{m_2,k})) - \sum_{i=m_1+1}^{m_2}\frac{1}{2}\mathbb{E}(D_H(\psi_{i,k},\psi_{i-1,k})).
\end{aligned} \tag{5. 21}$$

Due to the triangle inequality, we get for every $k \ge 0$

$$\begin{aligned}
&t_k\left(\sum_{i=1}^{m_1} L_{f_i} + \sum_{i=m_1+1}^{m_2}\|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}\right)\mathbb{E}\left(\|x_k-x_{k+1}\|\right)\\
&\quad \le t_k\left(\sum_{i=1}^{m_1} L_{f_i} + \sum_{i=m_1+1}^{m_2}\|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}\right)\left(\mathbb{E}\left(\|x_k-\psi_{m_2,k}\|\right) + \mathbb{E}\left(\|\psi_{m_2,k}-x_{k+1}\|\right)\right).
\end{aligned} \tag{5. 22}$$

Because of the Young's inequality and the strong convexity of H, we have

$$\begin{aligned}
&t_k\left(\sum_{i=1}^{m_1} L_{f_i} + \sum_{i=m_1+1}^{m_2}\|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}\right)\mathbb{E}\left(\|\psi_{m_2,k}-x_{k+1}\|\right)\\
&\quad \le \frac{1}{2\sigma}t_k^2\left(\sum_{i=1}^{m_1} L_{f_i} + \sum_{i=m_1+1}^{m_2}\|A_i\|\sqrt{4D_{U_i}+2\|u_i^c\|^2}\right)^2 + \mathbb{E}(D_H(x_{k+1},\psi_{m_2,k})),
\end{aligned}$$

and similarly, using the same arguments as above and additionally (5.2), we have

$$\begin{aligned}
&t_k \left(\left(\sum_{i=1}^{m_1} L_{f_i} \right) + \sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \right) \mathbb{E}\left(\|x_k - \psi_{m_2,k}\| \right) \\
&\quad \leq \frac{1}{\sigma} t_k^2 \left(\left(\sum_{i=1}^{m_1} L_{f_i} \right) + \sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \right)^2 + \frac{1}{2} \mathbb{E}\left(\sum_{i=m_1+1}^{m_2} D_H(\psi_{i,k}, \psi_{i-1,k}) \right).
\end{aligned}$$

So we get for (5. 22)

$$\begin{aligned}
&t_k \left(\left(\sum_{i=1}^{m_1} L_{f_i} \right) + \sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \right) \mathbb{E}\left(\|x_k - x_{k+1}\| \right) \\
&\quad \leq \frac{3}{2\sigma} t_k^2 \left(\left(\sum_{i=1}^{m_1} L_{f_i} \right) + \sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \right)^2 + \mathbb{E}(D_H(x_{k+1}, \psi_{m_2,k})) \\
&\quad\quad + \frac{1}{2} \mathbb{E}\left(\sum_{i=m_1+1}^{m_2} D_H(\psi_{i,k}, \psi_{i-1,k}) \right).
\end{aligned}$$

Inserting this inequality to (5. 21), we obtain

$$\begin{aligned}
&t_k \mathbb{E}\left(\left(\sum_{i=1}^{m_2} f_i + g \right)(x_{k+1}) - \left(\sum_{i=1}^{m_2} f_i + g \right)(y) \right) + \mathbb{E}(D_H(y, x_{k+1})) \\
&\quad \leq \mathbb{E}(D_H(y, x_k)) + \frac{1}{\sigma} t_k^2 \left(\left(\sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \right)^2 \left(\left(\sum_{i=m_1+1}^{m_2} \frac{1}{p_i^2} \right)^{\frac{1}{2}} + 1 \right) \right. \\
&\quad\quad \left. + \left(\sum_{i=1}^{m_1} L_{f_i} \right)^2 \left(\left(\sum_{i=1}^{m_1} \frac{1}{p_i^2} \right)^{\frac{1}{2}} + 1 \right) \right) + t_k^2 \frac{\delta}{\sigma} \sum_{i=m_1+1}^{m_2} D_{U_i} \\
&\quad\quad + \frac{3}{2\sigma} t_k^2 \left(\left(\sum_{i=1}^{m_1} L_{f_i} \right) + \sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \right)^2.
\end{aligned}$$

Summing up this inequality from $k = 0$ to $N - 1$, for $N \geq 1$, we get

$$\begin{aligned}
&\sum_{k=0}^{N-1} t_k \mathbb{E}\left(\left(\sum_{i=1}^{m_2} f_i + g \right)(x_{k+1}) - \left(\sum_{i=1}^{m_2} f_i + g \right)(y) \right) + \mathbb{E}(D_H(y, x_N)) \\
&\quad \leq \mathbb{E}(D_H(y, x_0))) + \frac{1}{\sigma} \left(\left(\sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \right)^2 \left(\left(\sum_{i=m_1+1}^{m_2} \frac{1}{p_i^2} \right)^{\frac{1}{2}} + 1 \right) \right. \\
&\quad\quad + \left(\sum_{i=1}^{m_1} L_{f_i} \right)^2 \left(\left(\sum_{i=1}^{m_1} \frac{1}{p_i^2} \right)^{\frac{1}{2}} + 1 \right) + \frac{3}{2} \left(\left(\sum_{i=1}^{m_1} L_{f_i} \right) \right. \\
&\quad\quad \left. \left. + \sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2} \right)^2 \right) \sum_{k=0}^{N-1} t_k^2 + \frac{\delta}{\sigma} \sum_{i=m_1+1}^{m_2} D_{U_i} \sum_{k=0}^{N-1} t_k^2.
\end{aligned}$$

Since $\mathbb{E}(D_H(y, x_N)) \geq 0$, we obtain

$$\mathbb{E}\left(\min_{0\leq k\leq N-1}\left(\sum_{i=1}^{m_2} f_i + g\right)(x_{k+1}) - \left(\sum_{i=1}^{m_2} f_i + g\right)(y)\right) \leq \frac{D_H(y, x_0) + \frac{1}{\sigma} C \sum_{k=0}^{N-1} t_k^2}{\sum_{k=0}^{N-1} t_k},$$

where

$$C = \delta \sum_{i=m_1+1}^{m_2} D_{U_i} + \left(\sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=m_1+1}^{m_2} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)$$
$$+ \left(\sum_{i=1}^{m_1} L_{f_i}\right)^2 \left(\left(\sum_{i=1}^{m_1} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) + \frac{3}{2}\left(\left(\sum_{i=1}^{m_1} L_{f_i}\right) + \sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2$$

□

Corollary 5.37. *Let $x^* \in \operatorname{dom} H$ be an optimal solution to* (5. 20), $\delta > 0$ *and*

$$P = \delta \sum_{i=m_1+1}^{m_2} D_{U_i} + \left(\sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2 \left(\left(\sum_{i=m_1+1}^{m_2} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right)$$
$$+ \left(\sum_{i=1}^{m_1} L_{f_i}\right)^2 \left(\left(\sum_{i=1}^{m_1} \frac{1}{p_i^2}\right)^{\frac{1}{2}} + 1\right) + \frac{3}{2}\left(\left(\sum_{i=1}^{m_1} L_{f_i}\right) + \sum_{i=m_1+1}^{m_2} \|A_i\| \sqrt{4D_{U_i} + 2\|u_i^c\|^2}\right)^2.$$

Then, the optimal stepsize for Algorithm 5.35 is given by

$$t_k := \sqrt{\frac{\sigma D_H(x^*, x_0)}{P}} \cdot \frac{1}{\sqrt{k}},$$

for all $k \geq 0$, which yields for every $N \geq 1$

$$\mathbb{E}\left(\min_{0\leq k\leq N-1}\left(\sum_{i=1}^{m_2} f_i + g\right)(x_k) - \left(\sum_{i=1}^{m_2} f_i + g\right)(x^*)\right) \leq 2\sqrt{\frac{D_H(x^*, x_0)P}{\sigma}} \cdot \frac{1}{\sqrt{N}}.$$

5.4 Applications

We consider three applications that can be modeled as optimization problems of the format considered in this work. The first of them stems from logistics and was modeled in [87] as a continuous location optimization problem. We compare the performance of our algorithm with those of three versions of the method proposed in [35]. The other two applications, one in medical imaging (more precisely in tomography) and one in machine learning (Support Vector Machines), were discussed in [35], too. We compare the performance of our algorithm to the stochastic version of the method introduced there. We use the proximal points of the smoothed objective functions instead of their subgradients, motivated also by the fact (noted, for instance, in [66]) that proximal point algorithms tend to solve certain optimization problems faster and cheaper than subgradient methods. To this end, we smooth the involved functions in the second and third application with the Moreau envelope, in the first application with Nesterov's smoothing approach. The experiments were carried out for one run of the algorithms and then averaged over 10 runs (and 100 runs for the first application) of the algorithms, as the stochastic methods perform slightly differently on each run due to the stochastic component. For the applications in tomography and SVM, our source code is built upon the code utilized in [38].

5.4.1 Continuous location problem

We consider the following location problem: given m locations placed at points $c_i \in \mathbb{R}^2$, each of them weighted with a parameter $w_i > 0$, $i = 1, \dots, m$, find a position $x \in \mathbb{R}^2$ for a service center so that the sum of the distances from it to the m locations is minimized under the restriction that the distance from the service center to the origin is less than or equal to a given radius $r > 0$. We can formulate this problem as

$$\min_{x \in S} \sum_{i=1}^{m} w_i \|x - c_i\|,$$

where $S = \{x \in \mathbb{R}^2 : \|x\| \leq r\}$.

We can write [92, Example 2.22]

$$f_i(x) = w_i \|x - c_i\| = \sup_{y \in \mathbb{B}} \{\langle w_i x, y \rangle - \langle w_i c_i, y \rangle\}$$

and according to [92, Corollary 2.20] (note, that for our setting, we have $A = w_i$, $b = w_i c_i$, $y_0 = 0$, $\mu = \gamma_k$ and $Q = \mathbb{B}$) with $b_{\mathbb{B}}(y) = \frac{1}{2}\|y\|^2$, using Nesterov's smoothing approach, we obtain the smooth approximations

$$f_i^{\gamma_k}(x) = w_i^2 \frac{\|x - c_i\|^2}{2\gamma_k} - \frac{\gamma_k}{2} \left[d\left(\frac{w_i(x - c_i)}{\gamma_k}, \mathbb{B} \right) \right]^2,$$

where $\mathbb{B}$ is the closed unit ball of $\mathbb{R}$ and $d(x, \mathbb{B})$ is the Euclidean distance from x to $\mathbb{B}$. Then, the gradients $\nabla f_i^{\gamma_k}$ can be written in terms of the projection operator $\mathcal{P}_{\mathbb{B}}$ on $\mathbb{B}$ as

$$\nabla f_i^{\gamma_k} = w_i \mathcal{P}_{\mathbb{B}} \left(\gamma_k^{-1} w_i (x - c_i) \right).$$

We choose $H(x) = \frac{1}{2}\|x\|^2$ for $x \in S$ and $H(x) = +\infty$ otherwise, so that we obtain for the mirror map the orthogonal projection onto the set S. We have for our smoothing parameters $\gamma_k = 2r\delta t_k$ for a $\delta > 0$.

In our numerical experiments, we chose $m = 1000000$, $r = 0.3$ and the m locations such that $c_i \in [-1, 1] \times [-1, 1]$. The weights $w_i \in (0, 1)$ are beta randomly distributed. A histogram for the the number of locations for the different weights is presented in the left picture of Figure 5.1. The positions of the m locations are indicated by blue dots in the right image of Figure 5.1. The greater the weight of the respective location, the greater the point. The red circle with radius r represents the permissible set for the position of the service center. The calculated position of the service center is shown as the red dot.

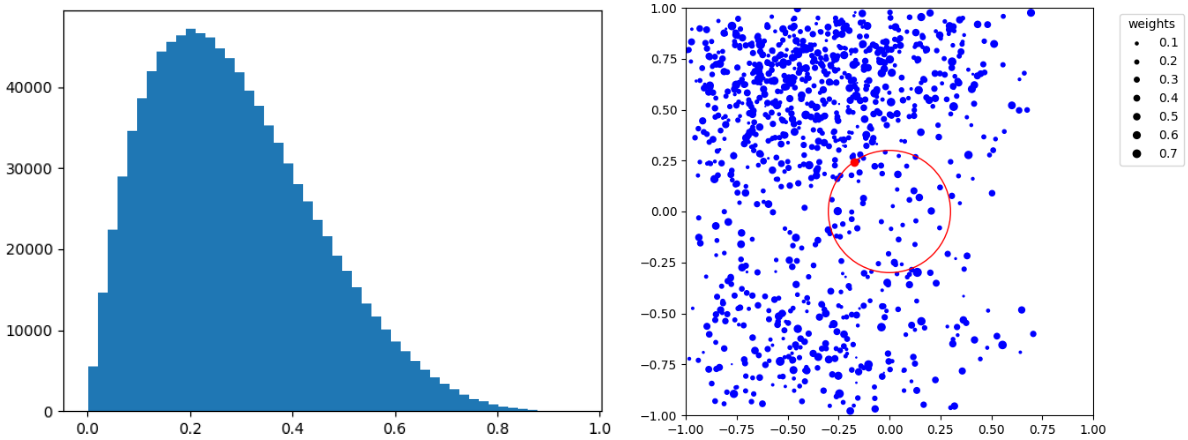


Figure 5.1: The left picture shows the histogram for the number of locations (y-axis) for the different weights (x-axis). The right image displays the positions of 1000 locations randomly selected from the set of the m locations as blue dots, along with the service center calculated using Algorithm 5.14 represented as a red dot. The red circle denotes the restricted set.

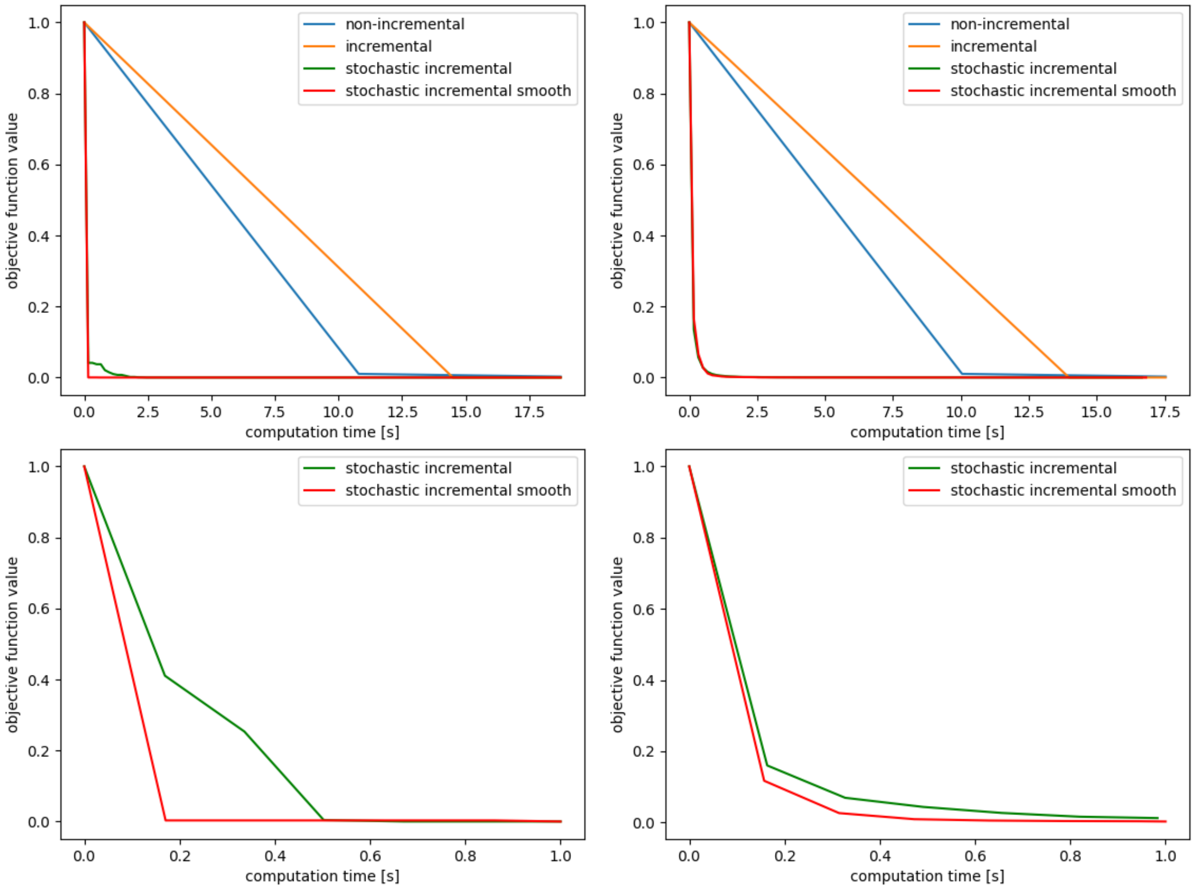


Figure 5.2: The plots show $(f_N - f(x_{\text{best}})/(f(x_0) - f(x_{\text{best}}))$, where $f_N := \min_{0 \le k \le N} f(x_k)$, as a function of time, so x_k is the last iterate before a given point in time. In the first row, we see the results after 17.5*s* for a single run left and for 100 runs right. In the second row, the left image displays the results after 1 second for a single run, while the right image shows the results after 100 runs.

We compared our algorithm 5.9 (*stochastic incremental smooth*) to three versions of the algorithms described in [35]. The *stochastic incremental version* is the algorithm, which is exactly given in [35]. The *non-incremental version* takes a full subgradient step of the objective function $f(x)$ in each iteration instead of the single components $f_i(x)$, so basically its the stochastic incremental version with $m = 1$ and $\epsilon_{1,k} = 1$ for every $k \geq 0$. The *incremental version* is the same as stochastic incremental version, if we choose $\epsilon_{i,k} = 1$ for every $i = 1, \ldots, m$ and every $k \geq 0$, so that we use the subgradient of all single components instead of a random choice. We chose $p_i = 0.000001$ for every $i = 1, \ldots, m$ for the stochastic algorithms. In Figure 5.2 in the first row we can see the comparison of all four algorithms after one run in the left and 100 runs in the right respectively and see that the stochastic algorithms outperform the non-stochastic versions clearly. In the second row we compared only the stochastic algorithms to have a better look after 1s CPU time for one run and 100 runs. Here we can see that our algorithm is slightly better than the stochastic incremental.

5.4.2 Tomography

We consider the following optimization problem, which was proposed in [26]

$$\min_{x\in\Delta} \left\{ -\sum_{i=1}^{m} y_i \log\left(\sum_{j=1}^{n} r_{ij}x_j\right)\right\},$$

where $\Delta := \{x \in \mathbb{R}^n : \sum_{j=1}^n x_j = 1, x \geq 0\}$ and $r_{ij} > 0$ is for $i = 1, \ldots, m$, and $j = 1, \ldots, n$, the entry of the i-th row and the j-th column of the matrix $R \in \mathbb{R}^{m\times n}$. Furthermore, y_i represents the positive number of photons measured in the i-th bin, where $i = 1, \ldots, m$. As mirror map we choose

$$H(x) = \begin{cases} \sum_{i=1}^n x_i \log(x_i), & \text{for } x \in \Delta \\ +\infty & \text{otherwise.} \end{cases}$$

The function

$$f_i(x) := -y_i \log\left(\sum_{j=1}^{n} r_{ij}x_j\right)$$

is Lipschitz continuous for all $i = 1, \ldots, m$, and so it follows that $\operatorname{dom} f_i^*$ is bounded. So, we can apply Algorithm 5.14. The proximal point mapping of the function f_i can be deduced from Lemma 6.5 and Theorem 6.15 in [21] and is given by

$$\operatorname{Prox}_{\gamma_k f_i}(v) = v + \frac{1}{\alpha} R_i^\top \frac{\sqrt{(R_i v)^2 + 4\gamma_k \alpha y_i} - R_i v}{2},$$

where

$$R_i = (r_{i1}\, r_{i2}\, \ldots\, r_{in}), \quad \alpha = \sum_{j=1}^{n} r_{ij}^2,$$

$R_i^\top$ is the transposed vector of R_i and the smoothing parameters $\gamma_k = \delta t_k$ for a $\delta > 0$.

We can see in Figure 5.3, that both algorithms have similar numerical performances, with the one proposed in this work reaching slightly lower objective function values.

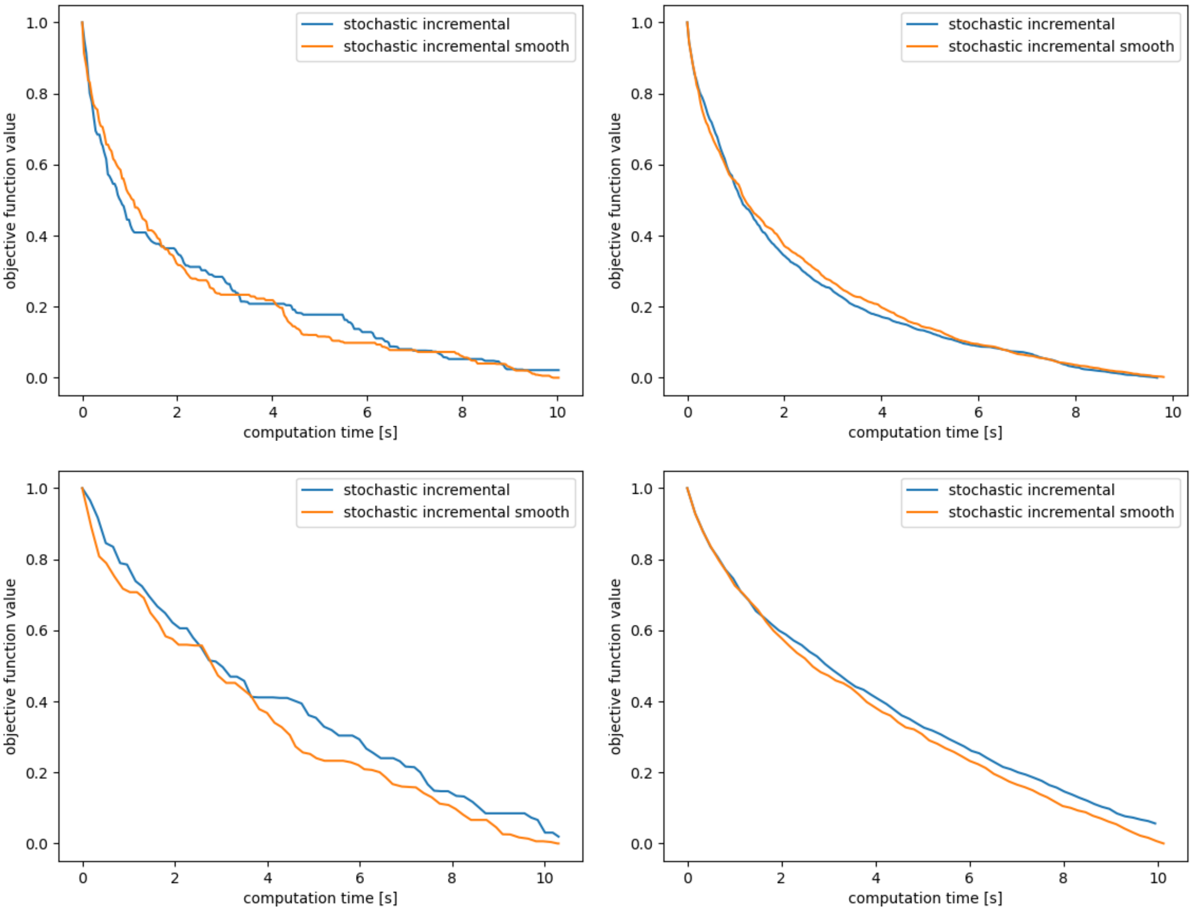


Figure 5.3: The plots show $(f_N - f(x_{\text{best}})/(f(x_0) - f(x_{\text{best}})$, where $f_N := \min_{0 \leq k \leq N} f(x_k)$, as a function of time, so x_k is the last iterate before a given point in time. In the first row, we see the results for $n = 1000$ and $m = 6000$ for a single run in the left plot and the average values of 10 runs in the right plot (with $p_i = 0.01667\ \forall i$). In the second row, we see the results for $n = 5000$ and $m = 15000$ for a single run and the average values of 10 runs , respectively (with $p_i = 0,0066\ \forall i$).

5.4.3 Linear SVM

In this subsection, we consider an optimization problem of classifying images via binary linear support vector machines with 1-norm. For an introduction to the linear SVM model, see Appendix A.1.

The given data set for classification consists of 11339 training images and 1850 test images of size 28 × 28 of handwritten digits on a gray-scale pixel map. The data set was taken from [93]. In the following optimization problem, we search a weight w for a decision function $f(\cdot) = \langle w, \cdot \rangle$ to classify the numbers 5 and 6 to the class with label $+1$ and -1, respectively

$$\min_{w \in \mathbb{R}^d} \left\{ \sum_{i=1}^{m} \max\{1 - y_i \langle w, x_i \rangle, 0\} + \lambda \|w\|_1 \right\}, \tag{5. 23}$$

$\{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq \mathbb{R}^d \times \{+1, -1\}$ is the given training data set with the training images

x_i and the labels y_i (here $d = 28 \cdot 28 = 784$ and $m = 11339$). The 1-norm is a regularization term with the regularization parameter $\lambda > 0$.

We set $H = \frac{1}{2}\|\cdot\|^2$, as done in [35], to obtain the identity as mirror map, considering that this problem is unconstrained. We can write the optimization problem above as

$$\min_{w \in \mathbb{R}^d} \left\{ \sum_{i=1}^{m} f_i(w) + g(w) \right\},$$

where

$$f_i(w) = \max\{1 - Y_i\langle w, x_i\rangle, 0\} \quad \text{and} \quad g(w) = \lambda\|w\|_1.$$

The function f_i is Lipschitz continuous for all $i = 1, \dots, m$, and so it follows that $\operatorname{dom} f_i^*$ is bounded. It follows that we can apply our algorithm. The proximal point mapping of the function f_i can be found in [56, Appendix A] and is given by

$$\operatorname{Prox}_{\gamma_k f_i}(v) = v + \begin{cases} \gamma_k y_i x_i, & s_i \in [\gamma_k\|x_i\|^2, +\infty) \\ \frac{s_i y_i}{\|x_i\|^2} x_i, & s_i \in (0, \gamma_k\|x_i\|^2) \\ 0 & \text{otherwise,} \end{cases}$$

where

$$s_i = 1 - y_i\langle v, x_i\rangle$$

and $\gamma_k = 2\delta t_k$ for a $\delta > 0$.

The algorithms show also for this application similar numerical performance, with a slightly improved classification by the method proposed in this work, see Figure 5.4 for the corresponding plots.

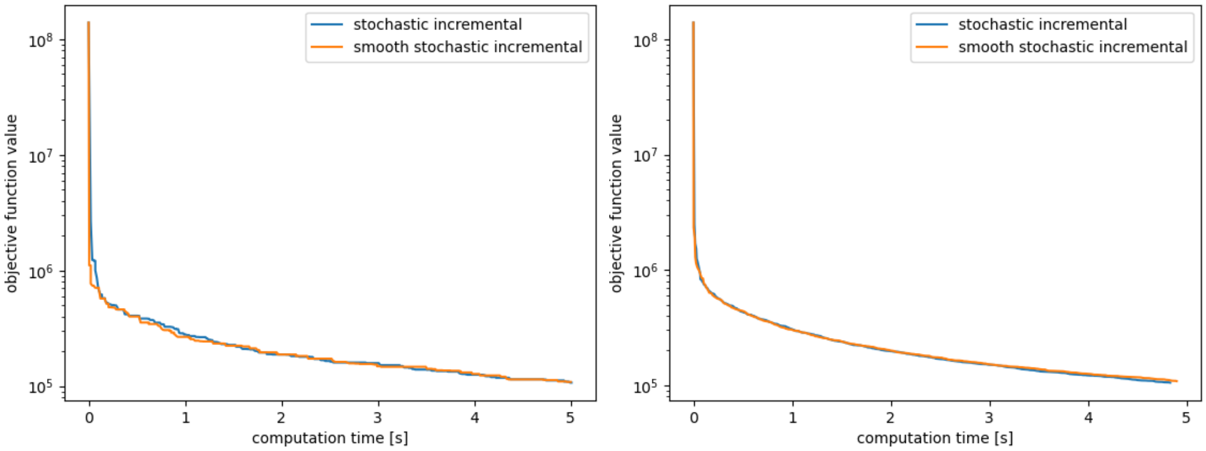


Figure 5.4: The plots show $f := \min_{0 \le k \le N} f(w_k)$ as a function of time, so w_k is the last iterate before a given point in time. We see the results for $\lambda = 0.001$ and $p_i = 0,0082\ \forall i$ for a single run in the left plot and the average values of 10 experiments in the right plot.

Table 5.1: Numerical results for the SVM problem for stochastic incremental algorithm (Algorithm 4.2 [35]) (SI) and stochastic incremental smoothing algorithm (Algorithm 5.26) (SIS). The results are shown for a single run, with the values in brackets representing the results over 10 runs for $p_i = 0,0082$ for all i.

regularization parameter	algorithm	decrease obj.function value	misclassified in %
$\lambda = 0.01$	SI	99.928 (99.923)	2.595 (2.595)
	SIS	99.929 (99.924)	2.324 (2.654)
$\lambda = 0.001$	SI	99.923 (99.927)	3.027 (2.605)
	SIS	99.922 (99.923)	2.432 (2.568)

Chapter 6

Conclusion and perspective

In this thesis, we used proximal splitting and smoothing techniques to solve nonsmooth, convex optimization problems. We showed the convergence of these schemes to an optimal solution of the respective minimization problem being considered . Furthermore, we discussed the importance of these methods and compared them to other numerical schemes in various applications.

In Chapter 3, we solved structured convex optimization problems with linear constraints by developing the Proximal AMA method as an extension of the AMA algorithm. We added variable metrics to the subproblems in the iteration process. As a result, the advantage over the classical AMA method is that, as long as the sequence of variable metrics is chosen appropriately, we can perform proximal steps to calculate new iterates. In this way, Proximal AMA avoids the need to use minimization subroutines in every iteration. In addition, it handles properly smooth and convex functions which might appear in the objective. The sequences of generated iterates converge to a primal–dual solution in the same setting as for the classical AMA method. The fact that instead of solving of minimization subproblems one has only to make proximal steps may lead to better numerical performances, as we showed in the experiments on image processing and support vector machines classification.

In Chapter 4, we introduced and investigated a dynamical system which generates three trajectories in order to approach the set of saddle points of the Lagrangian associated with the same structured convex optimization problem discussed in the previous chapter. After proving the existence and uniqueness of a solution of this system, we showed in the framework of Lyapunov analysis the convergence of the trajectories to the optimal solution of the optimization problem. In a numerical example, we demonstrated the impact of various parameter choices on the convergence behavior of the trajectories. The discretization of the considered dynamical system is related to the Proximal AMA method, introduced in the previous chapter, and the AMA scheme [110].

For Chapter 3 and Chapter 4, we would like to present some open questions regarding further research. It might be interesting to investigate convergence rates for the iterates and objective function values of Proximal AMA as well as for the trajectories and for the function values along the orbits regarding the dynamical system. For the AMA algorithm in [110], there are some results related to rates. Another suggestion is to consider second order dynamical systems in order to accelerate the convergence of the trajectories. This would induce inertial terms in the discretized counterparts of the dynamics. For an accelerated AMA numerical scheme, we refer to [65]. Further research could also involve embedding the investigations regarding Proximal AMA in the more general framework of monotone inclusion problems, similar to the approach taken in [39] starting from the Proximal ADMM algorithm.

In Chapter 5, we presented incremental stochastic mirror descent algorithms with Nesterov smoothing meant to minimize a sum of finitely many nonsmooth convex functions over a convex set. In contrast to the related algorithm from [35], we use the gradients of the smoothed summands of the objective function of the problem instead of their subgradients. To achieve this, we used the Nesterov smoothing technique, but since the Moreau envelope is a special case of this smoothing technique, these algorithms can also be formulated with proximal steps, too. We managed to obtain the same convergence order $\mathcal{O}(1/\sqrt{k})$ in expectation for the kth best objective function value and could show in three applications similar numerical performance as in [35], with slight improvements. Due to the fact that we do not need to calculate subgradients, we have more variations of the proposed algorithms. This allows us to choose the most suitable smoothing method depending on the structure of the considered optimization problem. If we use the Moreau envelope, we have uniquely defined proximal points, which have closed formulae for a variety of commonly used functions, instead of subgradients which one would have to pick from the subdifferentials of the involved functions at the given points. Moreover, the involved functions are not required to be (Lipschitz) continuous or differentiable, as they are usually taken in the literature on mirror descent methods. To improve convergence order, we are interested in accelerating the proposed algorithms as in [77], where the authors combined Nesterov's accelerated method and Nemirivski's mirror descent method both in continuous and discrete time.

Appendix A

Support Vector Machines

In this appendix, we give a brief overview on SVM, mainly to understand the SVM models considered in applications in section 5.4.3 and section 3.4.2. Support Vector Machine (SVM) for binary classification problems is a machine learning model that separates two given classified data sets through a decision boundary. The decision boundary in the linear model (linear SVM) is a hyperplane, chosen in such a way that the margin, defined as the distance between the decision boundary and the closest of the data points, is maximized. Data, for which the class is still unknown, is now classified by the decision boundary, depending on whether they are in the positive or negative half space of the hyperplane. Because the decision boundary is uniquely determined by only a few vectors closest to it (support vectors), this model is called Support Vector Machine.

Data sets cannot always be separated by a hyperplane. In this case, one can consider SVM with soft margin, which softens a hard separation through slack variables and thus allows outliers.

For some data sets, a separation by a hypersurface may be more suitable than by a hyperplane. In this case, the vectors are mapped into another (often higher dimensional) space, in which they can be separated by a hyperplane. Through a so-called kernel trick, the vectors of the other space only appear in a inner product of a chosen kernel function, which means that the optimization problem itself remains in the original dimension. This method is called nonlinear Support Vector Machine.

For further information on linear SVM, we refer the reader to [27], which is the basis for section A.1. The section on nonlinear SVM is based on [48, Section 2.3] and [103, Chapter 4].

A.1 Linear SVM

Fix an *input set* $X \subset \mathbb{R}^n$, which includes data points, and an *output set* $Y = \{-1, 1\}$, which contains class labels. For $N \geq 1$, we denote by

$$\mathcal{Z} = \{(x_1, y_1), \ldots, (x_N, y_N)\} \subseteq (X \times Y)^N$$

the given *training data set*, where x_k is called a *training vector* and y_k is called the corresponding *class* ($k = 1, \ldots, N$).

Our objective is to find a hyperplane in $\mathbb{R}^n$, which separates the training data such that the data points with label -1 lie in the negative half space and those with label $+1$ lie in the positive half space. At first, we present the method for separating training data without errors

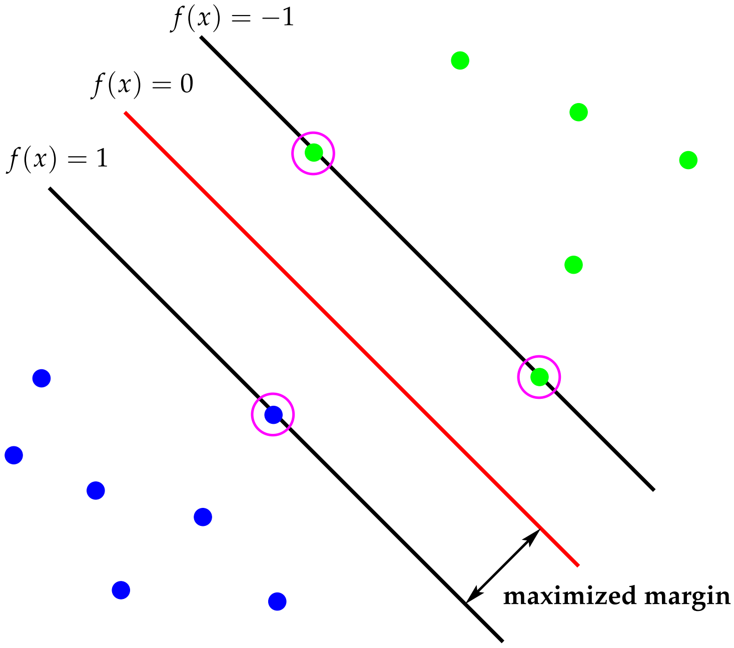


Figure A.1: In the figure, one can see a hyperplane, which is chosen such that the margin, defined as the orthogonal distance between the hyperplane and the closest of the data points, is maximized. The hyperplane is determined solely by a subset of data points, referred to as support vectors, which are circled in this illustration.

proposed by Vapnik in 1982 [111]. Therefore, we assume that our data set is linearly separable, which means that there exists such a separating hyperplane.

We search $w \in \mathbb{R}^n \setminus \{0\}$, a so called *weight vector*, and $b \in \mathbb{R}$, denoted as *bias*, for an affine function

$$\begin{aligned} f : X &\to \mathbb{R}, \\ x &\mapsto \langle w, x \rangle + b\,, \end{aligned}$$

such that $y_k f(x_k) > 0$ for all $1 \leq k \leq N$. In this section, $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product on $\mathbb{R}^n$ and $\| \cdot \|$ the associated norm.

In the context of linear SVM, the aim is to choose w and b associated to the hyperplane $H_{w,b} := \{x \in \mathbb{R}^n : \langle w, x \rangle + b = 0\}$ in such a way that the smallest distance between the training vectors from the training data set to the hyperplane is as large as possible. This means that we have to maximize the so called *margin*

$$M_{w,b} = \min\{\|x_k - x\| : \; k = 1, \ldots, N, \; x \in H_{w,b}\}$$

with respect to w and b. Due to the fact that $y_k f(x_k) > 0$ for all $1 \leq k \leq N$, the orthogonal distance of x_k to the hyperplane $H_{w,b}$ is given by $\frac{|f(x_k)|}{\|w\|} = \frac{y_k(\langle w, x_k \rangle + b)}{\|w\|}$, where $w \in \mathbb{R}^n \setminus \{0\}$ and $b \in \mathbb{R}$. To maximize the margin $M_{w,b}$, we can write

$$\max M_{w,b} = \max_{w,b} \left\{ \frac{1}{\|w\|} \min_k y_k(\langle w, x_k \rangle + b) \right\}. \tag{A. 1}$$

Let us assume that this optimization problem has a solution. If w and b is such a solution, then hw and hb for $h > 0$ solves this problem as well. Thus, w.l.o.g, we set

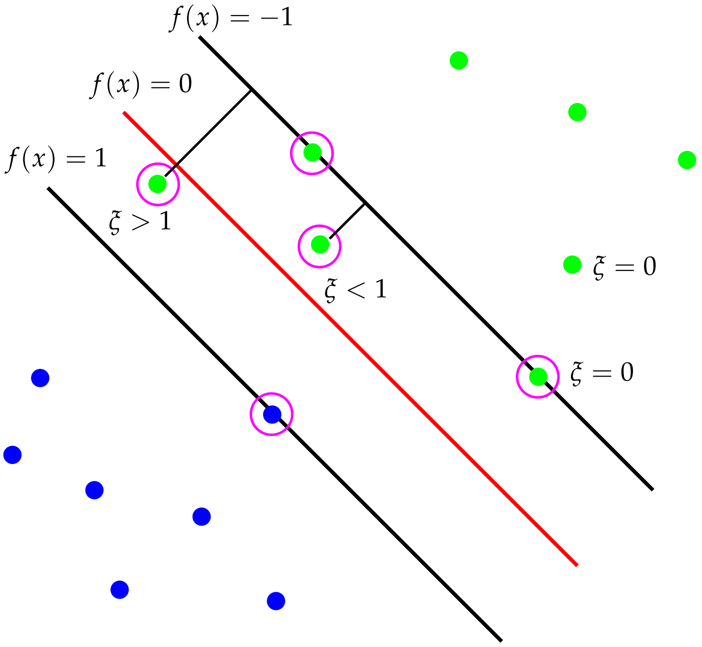


Figure A.2: Illustration of slack variables $\xi \geq 0$ corresponding to several data points. Data points with circles around them represent support vectors.

$\min_k y_k(\langle w, x_k \rangle + b) = 1$. Then, it follows for all $1 \leq k \leq N$

$$y_k(\langle w, x_k \rangle + b) \geq 1.$$

Since minimizing $|w|^2$ is equivalent to maximizing $|w|^{-1}$, we are led to consider the problem

$$\begin{aligned} &\min_{w,b} \frac{1}{2} \|w\|^2 \\ &\text{s.t. } y_k(\langle w, x_k \rangle + b) \geq 1 \quad k = 1, \dots, N. \end{aligned} \tag{A. 2}$$

It is clear that this optimization problem has a unique solution. With the above normalization, it is also a solution of (A. 1).

Often, the data sets are distributed in such a way that no linear separation is possible and the optimization problem above has no solution. This can be the case if some training vectors contain measurement errors or "outliers" are present in the data. Even in the case, where a hard linear separation is possible, it may sometimes be more suitable to have a decision function that misclassifies some training data points in order to achieve a larger margin. Therefore, we use slack variables $\xi_k \geq 0$ that describe the violation of the constraints of (A. 2). More precisely, we define

$$\xi_k = \max\{1 - y_k f(x_k), 0\}.$$

Thus, $\xi_k > 0$ is fulfilled if and only if the constraints of (A. 2) are not fulfilled for x_k. If $0 < \xi_k \leq 1$, then the corresponding data point is correctly classified but lies within the margin. If $\xi_k > 1$, then the data point x_k is on the wrong side of the hyperplane and is misclassified. For correctly classified data that are located outside or on the margin, we have $\xi_k = 0$ (see Figure A.2). This violation of the constraints is penalized by augmenting the objective function with the sum of the ξ_k multiplied by a constant $C \geq 0$, which allows us to control the level of penalization. Thus, for $C > 0$, we achieve a "softer" separation compared to the previous model described.

This approach, known as *soft margin SVM*, was introduced by Cortes and Vapnik in 1995 (see [52]).

We obtain the new unconstrained optimization problem

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{k=1}^{N} \max\{1 - y_k\left(\langle w, x_k\rangle + b\right), 0\}, \tag{A. 3}$$

where the function $C\sum_{k=1}^{N} \max\{1 - y_k\langle w, x_k\rangle + b, 0\}$ is called *loss function* . We can also choose the 1-norm of w instead of $\|\cdot\|^2$, resulting in the following optimization problem as an alternative to the previous one:

$$\min_{w,b} \frac{1}{2}\|w\|_1 + C\sum_{k=1}^{N} \max\{1 - y_k\left(\langle w, x_k\rangle + b\right), 0\}, \tag{A. 4}$$

where $\|w\|_1 := |w_1| + \cdots + |w_n|$.

Depending on the input data and our chosen algorithm to solve the problem, we can select either the 2-norm SVM (A. 3) or the 1-norm SVM (A. 4) as the appropriate model.

Remark A.1. In section 5.4.3, we considered the 1-norm SVM (A. 4) for the given application. In this context, we introduce a regularization parameter $\lambda = \frac{1}{C}$ for this particular problem. Note that in this model, the bias term b has been chosen to be 0 for the sake of simplicity. As a result, the model slightly differs from the SVM model original introduced by Cortes and Vapnik in this appendix. For more information and a detailed comparison between these two models, please refer to [98].

One can show that the vector w can be expressed as a linear combination of the training vectors $\{x_1, \ldots, x_N\}$ as follows

$$w = \sum_{k=1}^{N} a_k x_k, \tag{A. 5}$$

where $a_k \in \mathbb{R}$ (see [52] appendix A.2.). We call x_k *support vector*, if the corresponding $a_k \neq 0$. Thus, the optimal hyperplane is determined by only a few data points.

An affine function $\hat{f}(x) = \langle \hat{w}, x\rangle + \hat{b}$, where $(\hat{w}, \hat{b})$ is an optimal solution of SVM, is called *decision function*. We can classify data points with unknown labels using the function

$$\begin{aligned} g : X &\to Y, \\ x &\mapsto \begin{cases} +1, & \text{if } \hat{f}(x) \geq 0, \\ -1, & \text{if } \hat{f}(x) < 0. \end{cases} \end{aligned}$$

In order to evaluate the quality of our optimal hyperplane $H_{\hat{w},\hat{b}}$, resulting from the optimization problem of SVM, we classify data from a test dataset using the function g. Since the correct classification for the test data is already known, we can determine the misclassification rate, which serves as a quality measure for our hyperplane and model.

A.2 Nonlinear SVM

Not all data sets can be linearly separated, and a linear separation using soft margin is not always suitable. In this case, we map the data points into another space (in general into a higher dimensional space) by a non-linear feature function $\phi : X \to \mathcal{H}$, where $X \subset \mathbb{R}^n$ is the original space of our data points, referred to as the *input space*, and $\mathcal{H}$ denotes a Hilbert space and is known as the *feature space*. In this Hilbert space, we search for a separating hyperplane as the decision boundary according to the principles of linear SVM. Due to the non-linearity of the feature function, this model is referred to as a *non-linear SVM*. In the following, for the sake of simplicity, we assume that the bias b is set to 0. However, it is worth noting that the following calculations can also be applied for $b \neq 0$ (see [48], Appendix 2.B). The decision function is then given by

$$\begin{aligned} f : X &\to \mathbb{R}, \\ x &\mapsto \langle u, \phi(x) \rangle. \end{aligned}$$

The corresponding hypersurface is defined as

$$H_u := \{x \in X : 0 = \langle u, \phi(x) \rangle\},$$

where $u \in \mathcal{H}$ is determined by solving the optimization problem

$$\min_{u \in \mathcal{H}} \frac{1}{2}\|u\|^2 + C \sum_{i=1}^{N} \max\{1 - y_i \langle u, \phi(x_i) \rangle, 0\} \tag{A. 6}$$

with $C > 0$. In this section, $\langle \cdot, \cdot \rangle$ denotes the inner product on $\mathcal{H}$, while $\| \cdot \|$ represents the norm associated with this inner product. Mapping the problem into the feature space using the feature function ϕ can make the problem challenging to solve. Additionally, in certain cases, it may even be impossible due to the possibility of mapping into an infinite dimensional space. Therefore, we have to avoid the direct computation of $\phi(x)$. This can be done using the technique known as the "kernel trick". For this we need some definitions and properties, which can be found in chapter 4 in [103]:

In the following, let $X \subseteq R^n$.

Definition A.2. Let $X \neq \emptyset$. Then, the function $\kappa : X \times X \to \mathbb{R}$ is called *kernel function* (or kernel) if and only if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : X \to \mathcal{H}$ such that for all $x, y \in X$, it holds that

$$\kappa(x, y) = \langle \phi(x), \phi(y) \rangle.$$

The Hilbert space $\mathcal{H}$ is called feature space of the feature function ϕ.

Kernels can be constructed by utilizing a feature function $\phi : X \to \mathcal{H}$ and applying the definition mentioned above. It is also possible to determine whether a function κ is a kernel function without knowing the corresponding feature function ϕ by employing the following definition and theorem:

Definition A.3. The function $\kappa : X \times X \to \mathbb{R}$ is called *positive definite*, if for all $N \geq 1$, for all $(a_1, \ldots, a_N) \in \mathbb{R}^N$, and for all $(x_1, \ldots, x_N) \in X^n$, it holds

$$\sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j \kappa(x_i, x_j) \geq 0. \tag{A. 7}$$

Furthermore, κ is said to be *symmetric*, if $\kappa(x, y) = \kappa(y, x)\ \forall x, y \in X$.

Definition A.4. For fixed $x_1, \ldots, x_n \in X$ and a function $\kappa : X \times X \to \mathbb{R}$, we define the *Gram matrix* (also called kernel matrix) as $K = (\kappa(x_i, x_j))_{i,j=1}^{N}$.

Note that the Gram matrix is for all $(x_1, \ldots, x_N) \in X^n$ positive semi-definite if and only if its underlying kernel is positive definite.

Theorem A.5. *A function $\kappa : X \times X \to \mathbb{R}$ is a kernel if and only if the function κ is symmetric and positive definite.*

Proof. See [103, Theorem 4.16]. □

Note that for a given kernel, neither the feature function ϕ nor the feature space $\mathcal{H}$ are uniquely determined. The following definition is about a canonical feature space consisting of functions, referred to as the reproducing kernel Hilbert space, which is uniquely associated with a kernel, and vice versa:

Definition A.6. Let $X \neq \emptyset$ and $\mathcal{H}$ be a Hilbert function space over X, i.e., a Hilbert space that consists of functions mapping from X to $\mathbb{R}$.

- A function $\kappa : X \times X \to \mathbb{R}$ is said to be a *reproducing kernel* of $\mathcal{H}$, if it satisfies the following properties: $\kappa(\cdot, x) \in \mathcal{H}$ for all $x \in X$, and for all $f \in \mathcal{H}$ and $x \in X$, the reproducing property
$$f(x) = \langle f, \kappa(\cdot, x) \rangle$$
holds.

- The space $\mathcal{H}$ is called a *reproducing kernel Hilbert space* (RKHS) over X, if for all $x \in X$, the Dirac functional $\delta_x : \mathcal{H} \to \mathbb{R}$ defined by
$$\delta_x(f) := f(x), \quad f \in \mathcal{H},$$
is continuous.

According to the following lemma, reproducing kernels are indeed kernels:

Lemma A.7. *Let $\mathcal{H}$ be a Hilbert function space over X which has a reproducing kernel κ. Then, $\mathcal{H}$ is an RKHS and $\mathcal{H}$ is also a feature space of ϕ, where the feature map $\phi : X \to \mathcal{H}$ is given by*
$$\phi(x) = \kappa(\cdot, x), \quad x \in X.$$
We call ϕ the canonical feature map.

Proof. See [103, Lemma 4.19]. □

Theorem A.8. *Every RKHS has a unique reproducing kernel and every kernel has a unique RKHS.*

Proof. See [103, Theorem 4.20 and Theorem 4.21]. □

The RKHS $\mathcal{H}$ is often denoted as $\mathcal{H}_\kappa$, where κ represents the associated reproducing kernel. Then, the inner product on $\mathcal{H}_\kappa$ is written as $\langle \cdot, \cdot \rangle_{\mathcal{H}_\kappa}$ and the associated norm is denoted as $\| \cdot \|_{\mathcal{H}_\kappa}$.

Having these definitions and properties at our disposal, we consider in the following a kernel function $\kappa : X \times X \to \mathbb{R}$ and its associated RKHS $\mathcal{H}_\kappa$. For $f \in \mathcal{H}_\kappa$, according to to Lemma A.7 and the reproducing property, we have $\langle f, \phi(x_i) \rangle = \langle f, \kappa(\cdot, x_i) \rangle = f(x_i)$. Then, we can write the optimization problem (A. 6) as:

$$\min_{f \in \mathcal{H}_\kappa} \frac{1}{2} \|f\|^2_{\mathcal{H}_\kappa} + C \sum_{i=1}^{N} \max\{1 - y_i f(x_i), 0\}. \tag{A. 8}$$

Kernels, their RKHS, and the reproducing property play a crucial role in proving the following important theorem, originally introduced by Kimeldorf and Wahba in 1970 (see [76]) in the setting of Chebyshev splines and generalized to RKHS by Wahba in 1990 (see [112] and also [113]). We require this theorem for the "kernel trick" and the derivation of our nonlinear SVM model:

Theorem A.9 (Representer Theorem)**.** *The solution of the optimization problem* (A. 8)*, $\hat{f}$, can be expressed as a linear combination of kernel functions evaluated at the training data $\{x_1, \dots, x_N\}$ in the following form:*

$$\hat{f}(x) = \sum_{i=1}^{N} \beta_i \kappa(x_i, x),$$

where $\beta_i \in \mathbb{R}$ for all $1 \leq i \leq N$.

Proof. See [48, section 2.3]. □

Now, we can apply the “kernel trick” and express the optimization problem (A. 8) in terms of β_i using the Representer Theorem:

$$\min_{\beta_1, \dots \beta_N \in \mathbb{R}} \frac{1}{2} \sum_{i,j=1}^{N} \beta_i \beta_j \kappa(x_i, x_j) + C \sum_{i=1}^{N} \max\{1 - y_i \sum_{j=1}^{N} \beta_j \kappa(x_i, x_j), 0\}, \tag{A. 9}$$

where we utilized the kernel reproducing property for

$$\|f\|_{\mathcal{H}_\kappa}^2 = \sum_{i,j=1}^{N} \beta_i \beta_j \langle \kappa(x_i, \cdot), \kappa(\cdot, x_j) \rangle = \sum_{i,j=1}^{N} \beta_i \beta_j \kappa(x_i, x_j).$$

Using the kernel matrix K with $K_{ij} = \kappa(x_i, x_j)$ and K_i as the i-th column of K, we can write this problem as follows:

$$\min_{\beta := (\beta_1, \dots, \beta_N) \in \mathbb{R}^N} \frac{1}{2} \beta^T K \beta + C \sum_{i=1}^{N} \max\{1 - y_i K_i \beta, 0\}. \tag{A. 10}$$

Remark A.10. The simplest example of a kernel function is the *linear kernel*, where $\phi(x) = x$ and $\kappa(x, y) = \langle x, y \rangle$. In section 3.4.2, we considered the nonlinear SVM (A. 10) with the symmetric and finitely positive definite *Gaussian kernel* (for $\sigma > 0$)

$$\kappa : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \ \kappa(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2} \right),$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^n$.

Index of notation

$\mathcal{H}$	a real Hilbert space $\mathcal{H}$
$\mathcal{H} \times \mathcal{G}$	the Cartesian product of two Hilbert spaces $\mathcal{H}$ and $\mathcal{G}$
$\mathcal{H}^{\text{strong}}$	the strong topology of the Hilbert space $\mathcal{H}$
$\mathcal{H}^{\text{weak}}$	the weak topology of the Hilbert space $\mathcal{H}$
$\langle \cdot, \cdot \rangle$	a inner product
$\|\cdot\|$	a norm
$\|\cdot\|_1$	the 1-norm
$\mathbb{R}$	the set of real numbers
$+\infty$	plus infinity
$\overline{\mathbb{R}}$	the extended set of real numbers, $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$
$\mathbb{N}$	the set of natural numbers, $\mathbb{N} := \{1, 2, \dots\}$
Id	the identity operator
$C \subseteq D$	C is a subset of D
$C + D$	the Minkowski sum of two sets C and D
αC	the scaled set C by a constant α
$x + C$	the translation of a set C by a vector x
$\operatorname{int}(C)$	the interior of a set C
$\operatorname{ri}(C)$	the relative interior of a set C
$\operatorname{sqri}(C)$	the strong quasi-relative interior of a set C
$\operatorname{cl}(C)$	the closure of the set C
$\emptyset$	the empty set
$(x_n)_{n \in \mathbb{N}}$	a sequence of vectors starting with x_1
$\rightharpoonup$	weak convergence
$\to$	strong convergence
$\operatorname{dom}(f)$	the domain of a function f
$\operatorname{Im} f$	the image of a function f
$\operatorname{epi}(f)$	the epigraph of a function f
$\operatorname{graph}(f)$	the graph of a function f
∂f	the subdifferential of a function f
∇f	the gradient of a function f
f'	a subgradient of a function f
$\operatorname{argmin} f$	the set of minimizers of a function f
$\dot{f}$	the derivative of a function f
f^*	conjugate of a function f
f^{**}	biconjugate of a function f
$\inf, \min$	the infimum and minimum, respectively
$\sup, \max$	the supremum and maximum, respectively
$\operatorname{ess\,sup}$	the essential supremum

$f\Box g$	the infimal convolution of two functions f and g
$\text{Prox}_{\gamma f}$	the proximal point of a coefficient γ of a function f
ι_C	the indicator function of a set C
δ_x	the Dirac functional for a point x
$\mathcal{P}_C$	the projection onto a nonempty closed convex set C
$\mathcal{B}$	the closed unit ball of $\mathbb{R}$
$d(x, C)$	the Euclidean distance from a point x to a closed and convex set C
$S_+(\mathcal{H})$	the set of linear, continuous, self-adjoint and positive semidefinite operators $M : \mathcal{H} \to \mathcal{H}$
$\\| \cdot \\|_M$	the seminorm induced by an operator M
$\succcurlyeq$	the Loewner partial ordering
$(M_k)_{k\in\mathbb{N}}$	a sequence of operators starting with M_1
$\mathcal{P}_\alpha(\mathcal{H})$	the set of linear, continuous, self-adjoint and α-positive definite operators $M : \mathcal{H} \to \mathcal{H}$
A^{-1}	the inverse of a operator A
A^*	the adjoint of a linear operator A
$\\|A\\|$	the norm of a linear operator A
$2^{\mathcal{H}}$	the power set of $\mathcal{H}$
$\text{graph}(A)$	the graph of an operator A
X^*	the topological dual space of the topological vector space X
ISNR	improvement in signal-to-noise ratio
RMSE	root-mean-square error
$L^p([0, +\infty), \mathcal{B})$	the space of p-integrable functions $f : [0, \infty) \to \mathcal{B}$ where $\mathcal{B}$ is a Banach space
L^1_{loc}	the space of functions which are locally integrable
$\mathcal{L}(\mathcal{H})$	the set containing all linear and continuous operators $A : \mathcal{H} \to \mathcal{H}$
$H_{w,b}$	hyperplane or hypersurface respectively for a weight vector w and a bias b
SVM	support vector machine
RKHS	reproducing kernel Hilbert space

Bibliography

[1] B. Abbas, H. Attouch: *Dynamical systems and forward-backward algorithms associated with the sum of a convex subdifferential and a monotone cocoercive operator*, Optimization 64(10), 2223–2252 (2015)

[2] B. Abbas, H. Attouch, B.F. Svaiter: *Newton-like dynamics and forward-backward methods for structured monotone inclusions in Hilbert spaces*, Journal of Optimization Theory and its Applications 161(2), 331–360 (2014)

[3] M. Ahookhosh: *Optimal subgradient methods: computational properties for large-scale linear inverse problems*, Optimization and Engineering 19, 815–844 (2018)

[4] Z. Allen-Zhu, L. Orecchia: *Linear coupling: An ultimate unification of gradient and mirror descent*, in: C.H. Papadimitrou (ed.),8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Leibniz International Proceedings in Informatics (LIPIcs), 3:1–3:22 (2017)

[5] M. Amini, F. Yousefian: *An iterative regularized mirror descent method for ill-posed nondifferetiable stochastic optimization*, arXiv:1901.09506 (2019)

[6] A.S. Antipin: *Minimization of convex functions on convex sets by means of differential equations*, (Russian) Differentsial'nye Uravneniya 30(9): 1475–1486, 1994. Translation in Differential Equations 30(9), 1365–1375 (1994)

[7] A. Asl, M.L. Overton: *Behavior of limited memory BFGS when applied to nonsmooth functions and their Nesterov smoothings*, In: M. Al-Baali, A. Purnama, L. Grandinetti (eds) Numerical Analysis and Optimization, NAO 2020, Springer Proceedings in Mathematics and Statistics 354 (2021)

[8] H. Attouch: *Fast inertial proximal ADMM algorithms for convex structured optimization with linear constraint*, Minimax Theory and its Applications 06(1), 001–024 (2021)

[9] H. Attouch, R.I. Boţ, E.R. Csetnek: *Fast optimization via inertial dynamics with closed-loop damping*, Journal of the European Mathematical Society (2022)

[10] H. Attouch, A. Cabot: *Convergence of a relaxed inertial forward-backward algorithm for structured monotone inclusions*, Applied Mathematics and Optimization 80(3), 547–598 (2019)

[11] H. Attouch, S.C. László: *Continuous Newton-like inertial dynamics for monotone inclusions*, Set-Valued and Variational Analysis 29(3), 555–581 (2021)

[12] H. Attouch, S.C. László: *Newton-like inertial dynamics and proximal algorithms governed by maximally monotone operators*, SIAM Journal on Optimization 30(4), 3252–3283 (2021)

[13] H. Attouch, J. Peypouquet: *Convergence of inertial dynamics and proximal algorithms governed by maximally monotone operators*, Mathematical Programming 174, 391–432 (2019)

[14] H. Attouch, M. Soueycatt: *Augmented Lagrangian and proximal alternating direction methods of multipliers in Hilbert spaces. Applications to games, PDE's and control*, Pacific Journal of Optimization 5(1), 17–37 (2009)

[15] F. Bach: *Duality between subgradient and conditional gradient methods*, SIAM Journal on Optimization 25(1), 115–129 (2015)

[16] J.B. Baillon, H. Brézis: *Une remarque sur le comportement asymptotique des semigroupes non linéaires*, Houston Journal of Mathematics 2(1), 5–7 (1976)

[17] J.B. Baillon, G. Haddad: *Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones*, Israel Journal of Mathematics 26, 137–150 (1977)

[18] S. Banert, R.I. Boţ: *A forward-backward-forward differential equation and its asymptotic properties*, Journal of Convex Analysis 25(2), 371–388 (2018)

[19] S. Banert, R.I. Boţ, E.R. Csetnek: *Fixing and extending some recent results on the ADMM algorithm*, Numerical Algorithms 86, 1303–1325 (2021)

[20] H.H. Bauschke, P.L. Combettes: *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed. CMS Books in Mathematics, Springer, New York (2017)

[21] A. Beck: *First Order Methods in Optimization*, MOS-SIAM Series on Optimization, SIAM (2017)

[22] A. Beck, A. Ben-Tal, N. Guttmann-Beck, L. Tetruashvili: *The comirror algorithm for solving nonsmooth constrained convex problems*, Operations Research Letters 38(6), 493–498 (2010)

[23] A. Beck, M. Teboulle: *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters 31, 167–175 (2003)

[24] A. Beck, M. Teboulle: *A Fast Iterative Shrinkage- Thresholding Algorithm for Linear Inverse Problems*, SIAM Journal on Imaging Sciences 2(1), 183–202 (2009)

[25] A. Beck, M. Teboulle: *Smoothing and first order methods: a unified framework*, SIAM Journal on Optimization 22(2), 557–580 (2012)

[26] A. Ben-Tal, T. Margalit, A. Nemirovski: *The ordered subsets mirrordescent optimization method with applications to tomography*, SIAM Journal on Optimization 12(1), 79–108 (2001)

[27] C. M. Bishop: *Pattern Recognition and Machine Learning*, Springer-Verlag Science and Business Media (2006)

[28] S. Bitterlich, R.I. Boţ, E.R. Csetnek, G. Wanka: *The Proximal Alternating Minimization Algorithm for Two-Block Separable Convex Optimization Problems with Linear Constraints*, Journal of Optimization Theory and its Applications 182, 110–132 (2019). https://doi.org/10.1007/s10957-018-01454-y

[29] S. Bitterlich, E.R. Csetnek, G. Wanka: *A Dynamical Approach to Two-Block Seperable Convex Optimization Problems with Linear Constraints*, Numerical Functional Analysis and Optimization 42(1), 1–38 (2021). https://doi.org/10.1080/01630563.2020.1845730

[30] S. Bitterlich, S.-M. Grad: *Stochastic incremental mirror descent algorithms with Nesterov smoothing*, Numerical Algorithms, 1–32 (2023). https://doi.org/10.1007/s11075-023-01574-1

[31] S. Bitterlich, S.-M. Grad: *Stochastic incremental mirror descent algorithms with Nesterov smoothing*, hal-03428808 (2021)

[32] J. Bolte: *Continuous gradient projection method in Hilbert spaces*, Journal of Optimization Theory and its Applications 119(2), 235–259 (2003)

[33] A. Borovykh, N. Kantas, P. Parpas, G.A. Pavliotis: *On stochastic mirror descent with interacting particles: convergence properties and variance reduction*, Physica D 418, 132844 (2021)

[34] R.I. Boţ: *Conjugate Duality in Convex Optimization*, Lecture Notes in Economics and Mathematical Systems, Vol. 637, Springer, Berlin Heidelberg (2009)

[35] R.I. Boţ, A. Böhm: *An incremental mirror descent subgradient algorithm with random sweeping and proximal step*, Optimization 68(1), 33–50 (2019)

[36] R.I. Boţ, A. Böhm: *Variable smoothing for convex optimization problems using stochastic gradients*, Journal of Scientific Computing 85(2), 33 (2020)

[37] R.I. Boţ, E.R. Csetnek: *Second order forward-backward dynamical systems for monotone inclusion problems*, SIAM Journal on Control and Optimization 54(3), 1423–1443 (2016)

[38] R.I. Boţ, E.R. Csetnek, S.C. László: *A primal-dual dynamical approach to structured convex minimization problems*, Journal of Differential Equations 269(5), 10717–10757 (2020)

[39] R.I. Boţ, E.R. Csetnek: *ADMM for monotone operators: convergence analysis and rates*, Advances in Computational Mathematics 45, 327–359 (2019)

[40] R.I. Boţ, S.M. Grad, G. Wanka: *Duality in Vector Optimization*, Springer, Berlin Heidelberg (2009)

[41] R.I. Boţ, C. Hendrich: *Convergence analysis for a primal-dual monotone + skew splitting algorithm with applications to total variation minimization*, Journal of Mathematical Imaging and Vision 49(3), 551–568 (2014)

[42] R.I. Boţ, S.-M. Grad, D. Meier, M. Staudigl: *Inducing strong convergence of trajectories in dynamical systems associated to monotone inclusions with composite structure*, Advances in Nonlinear Analysis 10(1), 450–476 (2021)

[43] L. M. Bregman: *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR computational mathematics and mathematical physics 7(3), 200–217 (1967)

[44] H. Brézis: *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland Mathematics Studies No. 5, Notas de Matemática (50), North-Holland/Elsevier, New York (1973)

[45] R.E. Bruck: *Asymptotic convergence of nonlinear contraction semigroups in Hilbertspaces*, Journal of Functional Analysis 18, 15–26 (1975)

[46] J.-P. Calliess: *Lipschitz optimisation for Lipschitz interpolation*, in: 2017 American Control Conference (ACC2017), 17000349 (2017)

[47] G. Chantas, N. Galatsanos, A. Likas and M. Saunders: *Variational bayesian image restoration based on a product of t-distributions image prior*, IEEE transactions on image processing 17(10), 1795–1805 (2008)

[48] O. Chapelle: *Training a support vector machine in the primal*, Neural Computation 19(5), 1155–1178 (2007)

[49] E. Chouzenoux, J.-C. Pesquet, A. Repetti: *A block coordinate variable metric forward-backward algorithm*, Journal of Global Optimization 66(3), 457–485 (2016)

[50] P.L. Combettes, J.-C. Pesquet: *Proximal Splitting Methods in Signal Processing*, In: H.H. Bauschke, R. Burachik, P.L. Combettes, V. Elser, D.R. Luke, H. Wolkowiczm, (eds): *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, Springer Optimization and Its Applications, vol 49. Springer, New York (2011)

[51] P.L. Combettes, B.C. Vũ: *Variable metric quasi-Fejér monotonicity*, Nonlinear Analysis: Theory, Methods and Applications 78, 17—31 (2013)

[52] C. Cortes, V. Vapnik: *Support-vector networks*, Machine Learning 20(3), 273–297 (1995)

[53] M.G Crandall, A. Pazy: *Semi-groups of nonlinear contractions and dissipativesets*, Journal of Functional Analysis 3, 376–418 (1969)

[54] E.R. Csetnek: *Continuous dynamics related to monotone inclusions and non-smooth optimization problems*, Set-Valued and Variational Analysis 28(4), 611–642 (2020)

[55] E.R. Csetnek, Y. Malitsky, M.K. Tam: *Shadow Douglas-Rachford Splitting for Monotone Inclusions*, Applied Mathematics and Optimization 80(3), 665–678 (2019)

[56] A. Defazio: *A simple practical accelerated method for finite sums*, in: D.D. Lee, U. von Luxburg, R. Garnett, M. Sugiyama and I.M.Guyon (eds.), Advances in neural information processing systems 29, NIPS (2016)

[57] T.T. Doan, S. Bose, D.H. Nguyen, C.L. Beck: *Convergence of the iterates in mirror descent methods*, IEEE control systems letters 3(1), 114–119 (2019)

[58] J.C. Duchi, A. Agarwal, M. Johansson, M.I. Jordan: *Ergodic mirror descent*, SIAM Journal on Optimization 22(4), 1549–1578 (2012)

[59] J.C. Duchi, S. Shalev-Shwartz, Y. Singer, A. Tewari: *Composite objective mirror descent*, in: A.T. Kalai and M. Mohri (eds.), COLT 2010 - The 23rd Conference on Learning Theory, 14–26 (2010)

[60] P. Dvurechensky, S. Shtern, M. Staudigl, M.: *First-order methods for convex optimization*, EURO Journal on Computational Optimization 9, 100015 (2021)

[61] M. Fazel, T.K. Pong, D. Sun, P. Tseng: *Hankel matrix rank minimization with applications in system identification and realization*, SIAM Journal on Matrix Analysis and Applications 34, 946–977 (2013)

[62] R.M. Freund, P. Grigas, R. Mazumder: *AdaBoost and forward stagewise regression are first-order convex optimization methods*, arXiv:1307.1192 (2013)

[63] D. Gabay, B. Mercier: *A Dual Algorithm for the Solution of Nonlinear Variational Problems via Finite Element Approximations*, Computers & mathematics with applications 2, 17–40 (1976)

[64] G. Goh: *Optimization with Costly Subgradients*, ProQuest Dissertations Publishing, 2017.10685037 (2017)

[65] T. Goldstein, B. O'Donoghue, S. Setzer, R. Baraniuk: *Fast alternating direction optimization methods*, SIAM Journal on Imaging Sciences 7(3), 1588–1623 (2014)

[66] S.-M. Grad, O. Wilfer: *A proximal method for solving nonlinear minmax location problems with perturbed minimal time functions via conjugate duality*, Journal of Global Optimization 74, 121–160 (2019)

[67] F. Hanzely, P. Richtárik: *Fastest rates for stochastic mirror descent methods*, Computational Optimization and Applications 79(3), 717–766 (2021)

[68] A. Haraux: *Systèmes Dynamiques Dissipatifs et Applications*, Recherches en Mathématiques Appliquéées 17, Masson, Paris (1991)

[69] C. Hendrich: *Proximal Splitting Methods in Nonsmooth Convex Optimization*, PhD Thesis, Technical University of Technology, Chemnitz (2014)

[70] L.T.K. Hien, N. Gillis, P. Patrinos: *Inertial block mirror descent method for non-convex non-smooth optimization*, arXiv:1903.01818 (2019)

[71] L.T.K. Hien, C.V. Nguyen, H. Xu, C. Lu, J. Feng: *Accelerated randomized mirror descent algorithms for composite non-strongly convex optimization*, Journal of Optimization Theory and Applications 181(2), 541–566 (2019)

[72] V. Hovhannisyan, P. Parpas, S. Zafeiriou: *MAGMA - multilevel accelerated gradient mirror descent algorithm for large-scale convex composite minimization*, SIAM Journal on Imaging Sciences 9(4), 1829–1857 (2016)

[73] T. Hytönen, J. van Neerven, M. Veraar, L. Weis: *Analysis in Banach spaces, Volume I: Martingales and Littlewood-Paley Theory*, Springer, Cham (2016)

[74] A. Ivanova, F. Stonyakin, D. Pasechnyuk, E. Vorontsova, A. Gasnikov: *Adaptive mirror descent for the network utility maximization problem*, IFAC-PapersOnLine 53, 7851–7856 (2020)

[75] A. Juditsky, J. Kwon, É. Moulines: *Unifying mirror descent and dual averaging*, Mathematical Programming, 1–38 (2022)

[76] G.S. Kimeldorf, G. Wahba: *A correspondence between bayesian estimation on stochastic processes and smoothing by splines*, Annals of Mathematical Statistics 41, 495–502 (1970)

[77] W. Krichene, A. M. Bayen, P. L. Bartlett: *Accelerated mirror descent in continuous and discrete time*, NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems 2, 2845–2853 (2015)

[78] J. Li, G. Li, Z. Wu, C. Wu: *Stochastic mirror descent method for distributed multi-agent optimization*, Optimization Letters 12(6), 1179–1197 (2018)

[79] D.V.N. Luong, P. Parpas, D. Rueckert, B. Rustem: *Solving MRF minimization by mirror descent*, in: G. Bebis et al. (eds.), Advances in Visual Computing (ISVC 2012), Lecture Notes in Computer Science 7431, Springer, 587–598 (2012)

[80] H. Lu: *Relative continuity for non-Lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent*, INFORMS Journal on Optimization 1(4), 288–303 (2019)

[81] S. Mahadevan, B. Liu, P. Thomas, W. Dabney, S. Giguere, N. Jacek, I. Gemp, J. Liu: *Proximal reinforcement learning: a new theory of sequential decision making in primal-dual spaces*, arXiv:1405.6757 (2014)

[82] H.B. McMahan: *A unified view of regularized dual averaging and mirror descent with implicit updates*, arXiv:1009.3240v2 (2011)

[83] P. Mertikopoulos, M. Staudigl: *Stochastic mirror descent dynamics and their convergence in monotone variational inequalities*, Journal of optimization theory and applications 179(3), 838–867 (2018)

[84] J.J. Moreau: *Proximité et dualité dans un espace hilbertien*, Bulletin de la Société mathématique de France 93, 273–299 (1965)

[85] A.V. Nazin, S. Anulova, A. Tremba: *Application of the mirror descent method to minimize average losses coming by a Poisson flow*, in: Proceedings of the European Control Conference (ECC14), 2194–2197 (2014)

[86] A. Nedić, S. Lee: *On stochastic subgradient mirror-descent algorithm with weighted averaging*, SIAM Journal on Optimization 24(1), 84–107 (2014)

[87] Y. Nesterov: *Smooth minimization of non-smooth functions*, Mathematical programming 103(1), 127–152 (2005)

[88] Y. Nesterov: *Primal-dual subgradient methods for convex problems*, Mathematical programming, 120(1), 221–259 (2009)

[89] Y. Nesterov: *Lectures on Convex Optimization*, Springer Optimization and Its Applications 137, Springer (2018)

[90] A. Nemirovski: *Efficient methods for large-scale convex optimization problems*, Ekonomika i Matematicheskie Metody 15, 135–152 (1979) (in Russian)

[91] A. Nemirovski, D.B. Yudin: *Problem Complexity and Method Efficiency in Optimization*, J. Wiley & Sons, New York (1983)

[92] M.N. Nguyen, T.H.A. Le, D. Giles, T.A. Nguyen: *Smoothing techniques and difference of convex functions algorithms for image reconstruction*, Optimization 69(7–8), 1601–1633 (2019)

[93] S. Roweis: *Data for* MATLAB *hackers*, http://www.cs.nyu.edu/~roweis/data.html

[94] Y. Orlov: *Nonsmooth Lyapunov Analysis in Finite and Infinite Dimensions*, Communications and Control Engineering, Springer (2020)

[95] R. Paulavičius, J Žilinskas: *Analysis of different norms and corresponding Lipschitz constants for global optimization*, Technological and Economic Development of Economy 12(4), 301–306 (2006)

[96] J.F. Peña: *Generalized conditional subgradient and generalized mirror descent: duality, convergence, and symmetry*, arXiv:1903.00459v2 (2019)

[97] J, Peypouquet, S. Sorin: *Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time*, Journal of Convex Analysis 17(3–4): 1113–1163 (2010)

[98] T. Poggio, S. Mukherjee, R. Rifkin, A. Raklin, A. Verri: *b*, in: Uncertainty in geometric computations vol. 704, Kluwer Springer International Series in Engineering and Computer Science, 131–141 (2002)

[99] T.-D. Quoc: *Adaptive smoothing algorithms for nonsmooth composite convex minimization*, Computational Optimization and Applications 66(3), 425–451 (2017)

[100] L.I. Rudin, S. Osher, E. Fatemi: *Nonlinear total-variation-based noise removal algorithms*, Physica D. Nonlinear Phenomena 60 (1–4), 259–268 (1992)

[101] V.V. Semenov: *A version of the mirror descent method to solve variational inequalities*, Cybernetics and Systems Analysis 53(2), 234–243 (2017)

[102] R. Shefi, M. Teboulle: *Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization*, SIAM Journal on Optimization 24, 269–297 (2014)

[103] I. Steinwart, A. Christmann: *Support Vector Machines (Information Science and Statistics)*, Springer Science and Business Media (2008)

[104] F.S. Stonyakin, M. Alkousa, A.N. Stepanov, A.A. Titov: *Adaptive mirror descent algorithms for convex and strongly convex optimization problems with functional constraints*, Journal of Applied and Industrial Mathematics 13(3), 557–574 (2019)

[105] F. Stonyakin, A. Stepanov, A. Titov, A. Gasnikov: *Mirror descent for constrained optimization problems with large subgradient values of functional constraints*, Computer research and modeling 12(2), 301–317 (2020)

[106] F.S. Stonyakin, A.A. Titov: *One mirror descent algorithm for convex constrained optimization problems with non-standard growth properties*, in: S. Belim et al. (eds.): Proceedings of the School-Seminar on Optimization Problems and Their Applications (OPTA-SCL 2018), CEUR Workshop Proceedings-Series 2098, 372–384 (2018)

[107] W. Su, S. Boyd, E.J. Candès: *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, Journal of Machine Learning Research 17(153), 1–43 (2016)

[108] A. Titov, F. Stonyakin, M. Alkousa, S. Ablaev, A. Gasnikov: *Analogues of switching subgradient schemes for relatively Lipschitz-continuous convex programming problems*, in: MOTOR 2020, 133–149 (2020)

[109] Q. Tran-Dinh: *Adaptive smoothing algorithms for nonsmooth composite convex minimization*, Computational Optimization and Applications 66(3), 425–451 (2017)

[110] P. Tseng: *Applications of a Splitting Algorithm to Decomposition in Convex Programming and Variational Inequalities*, SIAM Journal on Control and Optimization 29(1), 119–138 (1991)

[111] V.N. Vapnik: *Estimation of Dependences Based on Empirical Data*, Addendum 1, New York, Springer-Verlag (1982)

[112] G. Wahba: *Spline Models for Observational Data*, SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics 59 (1990)

[113] G. Wahba: *Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV*, in: B. Schölkopf, C. Burges and A. Smola (eds.): Advances in Kernel Methods-Support Vector Learning, 69–88 (1999)

[114] S. Zhang, N. He: *On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization*, arXiv:1806.04781 (2018)

[115] X. Zhou, C. Du, X. Cai: *An efficient smoothing proximal gradient algorithm for convex clustering*, arXiv:2006.12592 (2020)

[116] Z. Zhou, P. Mertikopoulos, N. Bambos, S.P. Boyd, P.W. Glynn: *On the convergence of mirror descent beyond stochastic convex programming*, SIAM Journal on Optimization, 30(1), 687–716 (2020)

[117] Z. Zhou, P. Mertikopoulos, N. Bambos, S.P. Boyd, P.W. Glynn: *Stochastic mirror descent in variationally coherent optimization problems*, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (eds.), Advances in Neural Information Processing Systems 30 (NIPS 2017), 7040–7049 (2017)

[118] Y. Zhou, Y. Liang, L. Shen: *A Unified Approach to Proximal Algorithms using Bregman Distance*, Technical Report, Syracuse University, Syracuse, NY (2016)

Index