KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Distributed Proximal Algorithms for Large-Scale Structured Optimization

Puya Latafat

Supervisors:
Prof. Panagiotis Patrinos
Prof. Alberto Bemporad

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

July 2020

Distributed Proximal Algorithms for Large-Scale Structured Optimization

Puya LATAFAT

Examination committee:
Prof. Panagiotis Patrinos, supervisor
Prof. Alberto Bemporad, supervisor
Prof. Alexander Bertrand
Prof. Goele Pipeleers
Prof. Mario Zanon
Prof. Mikael Johansson
(KTH Royal Institute of Technology)
Prof. Volkan Cevher
(École polytechnique fédérale de Lausanne)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

July 2020

To Nadira, Hamid and Hamed.

To Giovanna and Andreas.

Preface

I wish to express my deepest appreciation to my advisor Prof. Panagiotis Patrinos without whom this thesis would never have been possible. Panos introduced me to a fascinating area of research I knew nothing about and guided me patiently every step of the way. I cannot thank him enough for everything that I have learned from him, for the amazingly friendly environment and the perfect balance between freedom in research and guidance. It has been a great privilege and a delight to study with him.

I am deeply grateful to Prof. Alberto Bemporad for accepting me in his research group, the exceptional working environment, and all the support throughout the past five years.

I wish to thank the members of my examination committee, Prof. Volkan Cevher, Prof. Mikael Johansson, Prof. Alexander Bertrand, Prof. Goele Pipeleers and Prof. Mario Zanon for their efforts in reviewing this thesis and their insightful comments that have helped improve the quality of this work.

I would like to thank IMT Lucca, KU Leuven, and the Research Foundation Flanders (FWO) for supporting my studies in Italy and Belgium during the past five years. I am grateful to the staff for their relentless support. My special thanks to Daniela Giorgetti, Sara Olsun, Maria Mateos Irigoyen and Serena Argentieri in Lucca; Ida Tassens, Aldona Niemiro Sznajder, Elsy Vermoesen and Graziella Del Savio at KU Leuven.

I am especially thankful to Nick, for his encouragements and sharp observations, Lorenzo, whose many developed packages I have used throughout my PhD and learned a lot from, and Andreas, for volunteering his time, math and TEX skills without hesitation.

I also wish to thank all my friends and colleagues from IMT Lucca with whom I have shared many ping pong matches, Turkish baths, road trips, cycling lessons, ice creams, and stories of statues coming to life. Thank you Vasilis, Valerio, Veronique G., Vigneswarangovindarajan, Valentina C., Double Kumar, Dilip Mejari, Kulynovic, Carollo, Ilkay, Unal, Domagoj, Haneen, Van Tien, Anita, Soumali, Vincenzo, Paolo, Mogens, Alessandro, Luca and somebody random.

I have also shared many fond memories with friends and colleagues in Leuven: Bottegaal, Federica, Leen, Qinghua, Zahra, Reza, Kim, Koen, Riccardo, Bertrand, Mathjis, Arun, Joachim, Michael, Peter, Olga, Rocco, Carlos, Orest, Daniele, Ilyar, Oscar, Yinxue, Yingyi, Angela, Marcin, Manish, and Prof. Borgioli. I specially thank Silvio who has made everything less bad by holding out his hand.

Even though continents apart, I thank Ilyar for sharing precious life lessons on how digesting stones can repulse policemen, and how certain sweets age well, Ehsan, for his battle with concupiscence and apparent childhood osteoporosis, Gaban, for his stable and reassuring rate of mysterious disappearances, Sasha, for being a role model of tranquility, Crispy, for making me live through the eighteenth century culture, Cansu, for being so taarof, Ali and Sina, for their expertise in forgery related sciences, Peyman, for having a dream, and Rinel, for still waiting for his colors.

I had the original idea of writing a six page long acknowledgment mostly about having a low SES, and with secret messages embedded inside the paragraphs. But then I realized that I was beaten to it by certain individuals. Besides, I cannot possibly express the true magnitude of my gratitude to Andreas for all the soothing whistles, educational videos on wildlife, and most importantly, for introducing me to the godly words of Mosconi. Neither can I really articulate the degree of disgust and horror that I have been subjected to by the many graphic tales casually told by Veronique. How can I truly express the sempiternal effects of Pantelis' whimsy, the importance of helmets among shepherds of Veneto, the cultural significance of buffoons in Hindi language and civilization, or how impressed I was with the sheer amount of wealth that a single Shabi could amass. How could I do justice to the kindness of Masoud and Julia and their companion Suzan, to the most mighty and kind President, Daniele, or to the promising young graphic designer Sureman. On a more serious note, I thank Hothouse co. for their essential role in my battle against fungi in Leuven.

I wish to sincerely thank Anna, Maria Pia, Lala, Franca and D'Inverno family who have been so kind to me, and have helped me in my hour of need; Grazie per la vostra gentilezza, per essermi state vicino ed avermi aiutato nei momenti in cui ne ho avuto bisogno.

Finally, I would not be here if it was not for my parents Nadira and Hamid, and their thoughtful support throughout my life. I am specially indebted to my brother Hamed for giving me the opportunity to learn from his mistakes. I am also grateful to my grandma and my aunts for their encouragements throughout my childhood.

Puya Latafat

Leuven, June 2020

Contents

Preface vii

Abstract xiii

Index of Symbols xv

Index of Abbreviations xvii

Vita & publications xix

1 Introduction 1

1.1 Overview of the thesis 6

1.2 Preliminary material 9

1.2.1 Vector notation 9

1.2.2 Sequences 9

1.2.3 Functions and operators 10

1.2.4 Subgradients 12

1.2.5 Conjugate functions and infimal convolution 13

1.2.6 Proximal mappings and Moreau envelopes 14

2 Asymmetric forward-backward-adjoint splitting 15

2.1 Introduction 15

2.2 Asymmetric forward-backward-adjoint splitting 17

2.2.1 Convergence analysis 18

2.3 Operator splitting schemes as special cases 30

2.3.1 Forward-backward splitting 30

2.3.2 Solodov and Tseng 31

2.3.3 Forward-backward-forward splitting 31
2.3.4 Douglas-Rachford type with a forward term 32
2.4 Conclusions 40

3 A unifying framework for primal-dual proximal algorithms 41
3.1 Introduction 41
3.2 A simple framework for primal-dual algorithms 42
3.2.1 Primal-dual algorithms with constant stepsize 44
3.2.2 Primal-dual algorithms with a dynamically updated stepsize . . 49
3.3 Simplified asymmetric forward-backward-adjoint splitting 51
3.4 A unified convergence analysis for primal-dual algorithms 57
3.4.1 Linear convergence 59
3.5 Stepsize selection and simulations 64
3.6 Conclusions 69

4 A randomized block-coordinate primal-dual proximal algorithm 70
4.1 Introduction 70
4.2 Triangularly preconditioned primal-dual algorithm 74
4.2.1 Related primal-dual algorithms 80
4.3 A randomized block-coordinate algorithm 82
4.4 Linear convergence 87
4.5 Distributed optimization 92
4.6 Application: formation control 97
4.7 Conclusions 101

5 Plug and play distributed model predictive control 102
5.1 Introduction 102
5.2 Problem setup 104

5.3 A primal-dual algorithm for DMPC . . . 106
5.4 Deriving the algorithm and convergence results . . . 109
5.5 Numerical simulations . . . 112
5.6 Conclusions . . . 115

6 A primal-dual proximal algorithm for distributed optimization over graphs 116

6.1 Introduction . . . 116
6.2 Problem formulation . . . 118
6.3 Distributed primal-dual algorithms . . . 120
6.3.1 Special case . . . 124
6.4 Numerical simulations . . . 124

7 Multi-agent structured optimization with bounded communication delays 127

7.1 Introduction . . . 127
7.1.1 Motivating examples . . . 131
7.1.2 Main contributions . . . 132
7.2 Problem setup . . . 133
7.3 Notation and preliminary results . . . 136
7.4 The case of partial coupling . . . 138
7.5 The case of total coupling . . . 143
7.5.1 An AHU-type primal-dual algorithm . . . 144
7.5.2 Randomized variant . . . 148
7.6 Numerical simulations . . . 154
7.7 Conclusions . . . 155

8 Block-coordinate and incremental aggregated proximal gradient methods 157

8.1 Introduction . . . 157
8.1.1 The main block-coordinate algorithm . . . 159
8.1.2 Contributions . . . 160
8.1.3 Organization . . . 161
8.2 Convergence analysis . . . 161
8.2.1 The forward-backward envelope . . . 162
8.2.2 A sure descent lemma . . . 165
8.2.3 Randomized sampling . . . 167
8.2.4 Cyclic, shuffled and essentially cyclic samplings . . . 170
8.2.5 Global and linear convergence with KL inequality . . . 175
8.3 Nonconvex finite sum problems: the Finito/MISO algorithm . . . 177
8.3.1 A low memory variant . . . 180
8.4 Nonconvex sharing problem . . . 182
8.5 Numerical simulations for the regularized finite sum problem . . . 184
8.6 Conclusions . . . 186
8.7 The key tool: the forward-backward envelope . . . 187
8.7.1 Further results . . . 188

Conclusions **192**

Future directions . . . 193

Bibliography **194**

Abstract

Efficient first-order algorithms for large-scale distributed optimization is the main subject of investigation in this thesis. The algorithms considered cover a wide array of applications in machine learning, signal processing and control.

In recent years, a large number of algorithms have been introduced that rely on (possibly a reformulation of) one of the classical splitting algorithms, specifically forward-backward, Douglas-Rachford and forward-backward-forward splittings. In this thesis a new three term splitting technique is developed that recovers forward-backward and Douglas-Rachford splittings as special cases. In the context of structured optimization, this splitting is leveraged to develop a framework for a large class of primal-dual algorithms providing a unified convergence analysis for many seemingly unrelated algorithms. Moreover, linear convergence is established for all such algorithms under mild regularity conditions for the cost functions.

As another notable contribution we propose a randomized block-coordinate primal-dual algorithm that leads to a fully distributed asynchronous algorithm in a multi-agent model. Moreover, when specializing to multi-agent structured optimization over graphs, novel algorithms are proposed. In addition, it is shown that in a multi-agent model bounded communication delays are tolerated by primal-dual algorithms provided that certain strong convexity assumptions hold.

In the final chapter we depart from convex analysis and consider a fully nonconvex block-coordinate proximal gradient algorithm and show that it leads to nonconvex incremental aggregated algorithms for regularized finite sum and sharing problems with very general sampling strategies.

Index of Symbols

$\mathbb{N}$	Natural numbers $\{0, 1, 2, \dots\}$
$\mathbb{R}$	Real numbers
$\overline{\mathbb{R}}$	Extended real numbers $\mathbb{R} \cup \{\infty\}$
$\mathbb{R}^{n \times m}$	Space of real $n \times m$ matrices
$\mathbb{S}^n$	Space of real $n \times n$ symmetric matrices
$\mathbb{S}^n_+$	Space of real $n \times n$ symmetric positive semidefinite matrices
$\mathbb{S}^n_{++}$	Space of real $n \times n$ symmetric positive definite matrices
$\emptyset$	Empty set
$\mathcal{H}, \mathcal{G}$	Real Hilbert spaces
$\mathscr{B}(\mathcal{H}, \mathcal{G})$	Space of bounded linear operators from $\mathcal{H}$ to $\mathcal{G}$
$\mathscr{B}(\mathcal{H})$	Space $\mathscr{B}(\mathcal{H}, \mathcal{H})$
$\Gamma_0(\mathcal{H})$	Set of proper lsc convex functions from $\mathcal{H}$ to $\overline{\mathbb{R}}$
$(\Omega, \mathcal{F}, \mathbb{P})$	Probability space
$w^k \rightharpoonup w$	Weak convergence
$w^k \to w$	Strong convergence
$M^\top$	Transpose of matrix M
$\operatorname{ran} M$	Range of matrix M
blkdiag	Creates a block diagonal matrix
I_n	Identity $n \times n$ matrix
0_n	Zero $n \times n$ matrix
id	Identity mapping
$F : A \rightrightarrows B$	A set-valued operator
$\operatorname{dom} F$	Domain of operator F
$\operatorname{gra} F$	Graph of operator F
$\operatorname{zer} F$	Set of zeros of operator F
$\operatorname{fix} F$	Set of fixed points of operator F
F^{-1}	Inverse of operator F
J_F	Resolvent of operator F
N_X	Normal cone operator of set X
$\operatorname{dom} f$	Domain of extended-real-valued function f
$\operatorname{epi} f$	Epigraph of function f
$\operatorname{lev}_{\leq \alpha}$	α-(sub)level set of f
f^*	Fenchel conjugate
$f \,\square\, g$	Infimal convolution of functions f and g
δ_X	Indicator function of set X
$f : A \to B$	Single-valued function
$\operatorname{ri} C$	Relative interior
$[r]_+$	Positive part of r: $\max\{0, r\}$
$\mathcal{P}^V_X$	Projection onto set X wrt $\|\cdot\|_V$
prox^V_f	V-Proximal mapping of f
$\operatorname{dist}_V(x, X)$	Distance of x from set X wrt $\|\cdot\|_V$
$\hat{\partial} f$	Regular subdifferential of f

∂f	(Limiting) subdifferential of f
$\partial_B f$	Bouligand subdifferential of f
∇f	Gradient of f
$\nabla^2 f$	Hessian of f
$O(\cdot)$	Big-O Bachmann–Landau notation
$o(\cdot)$	Little-O Bachmann–Landau notation

Index of Abbreviations

a.s.	Almost surely
ADMM	Alternating direction method of multipliers
AFBA	Asymmetric forward-backward-adjoint splitting
DMPC	Distributed model predictive control
DRS	Douglas-Rachford splitting
ERM	Empirical risk minimization
FBFS	Forward-backward-forward splitting
FBE	Forward-backward envelope
FBS	Forward-backward splitting
FNE	Firmly nonexpansive
iff	If and only if
i.i.d.	Independent and identically distributed
lsc	Lower semicontinuous
LP	Linear program
MPC	Model predictive control
osc	Outer semicontinuous
PLQ	Piecewise linear-quadratic
QP	Quadratic program
SVM	Support vector machine
wrt	With respect to

Vita

Sept. 23, 1989	Born, Tabriz, Iran
2007-2011	B.Sc. in Electrical Engineering-Control Systems University of Tabriz, Iran
2012-2014	M.Sc. Mathematical Modeling in Engineering (MathMods, Erasmus Mundus Master Program) Final mark: 110/110 cum laude University of L'Aquila, Italy - University of Hamburg, Germany
Since 2014	Ph.D. in Control Systems IMT School for Advanced Studies Lucca, Italy
2015-2016	Visiting student KU Leuven, Belgium ESAT - Departmet of Electrical Engineering
Since 2016	Ph.D. student jointly at KU Leuven, Belgium ESAT - Departmet of Electrical Engineering
Since 2017	FWO Ph.D. fellow, Belgium

Publications

Journal papers

[101] **P. Latafat**, A. Themelis and P. Patrinos.
Block-coordinate and incremental aggregated proximal gradient methods for nonsmooth nonconvex problems.
arXiv:1906.10053 (submitted 2019).
https://arxiv.org/abs/1906.10053

[99] **P. Latafat** and P. Patrinos.
Primal-dual algorithms for multi-agent structured optimization over message-passing architectures with bounded communication delays.
arXiv:1809.07199 (submitted 2019).
https://arxiv.org/abs/1809.07199

[95] **P. Latafat**, N. M. Freris and P. Patrinos.
A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization.
IEEE Transactions on Automatic Control, vol. 64, no. 10, Oct. 2019, pp. 4050-4065
https://ieeexplore.ieee.org/document/8673559

[96] **P. Latafat** and P. Patrinos.
Asymmetric forward-backward-adjoint splitting for solving monotone inclusions involving three operators.
Computational Optimization and Applications, vol. 68, no. 1, Sep. 2017, pp. 57-93.
https://doi.org/10.1007/s10589-017-9909-6

Contributions to books

[98] **P. Latafat** and P. Patrinos.
Primal-dual proximal algorithms for structured convex optimization: a unifying framework.
In Chapter 5 of Large-Scale and Distributed Optimization, (Giselsson P., and Rantzer A., eds.), vol. 2227 of Lecture Notes in Mathematics, Springer International Publishing, 2018, pp. 97-120.
https://link.springer.com/chapter/10.1007/978-3-319-97478-1_5

Conference papers

[100] **P. Latafat**, L. Stella and P. Patrinos.
New primal-dual proximal algorithms for distributed optimization.
In Proc. of the 2016 IEEE 55th Conference on Decision and Control (CDC 2016), Las Vegas, USA, Dec. 2016, pp. 1959-1964.
https://ieeexplore.ieee.org/document/7798551

[94] **P. Latafat**, A. Bemporad and P. Patrinos.
Plug and play distributed model predictive control with dynamic coupling: a randomized primal-dual proximal algorithm.
In Proc. of the European Control Conference (ECC), Limassol, Cyprus, Jun. 2018, pp. 1160-1165.
https://ieeexplore.ieee.org/document/8550270

[97] **P. Latafat** and P. Patrinos.
Multi-agent structured optimization over message-passing architectures with bounded communication delays.
In Proc. of the 57th IEEE Conference on Decision and Control (CDC), Miami Beach, Florida, Dec. 2018, pp. 1688-1693.
https://ieeexplore.ieee.org/document/8619278

Chapter 1

Introduction

One of the main goals of current distributed optimization research is to develop easy-to-implement iterative schemes for structured problems. To this end, proximal algorithms have become the standard tool; they are suitable for large-scale, nonsmooth problems with different computational models. These methods operate by *splitting* the original problem into simpler subproblems that involve one function at a time and can often be solved efficiently. The most widely used proximal algorithms are the proximal gradient and Douglas-Rachford methods that are based on the classical two term splittings, forward-backward splitting (FBS) and Douglas-Rachford splitting (DRS).

Consider the structured convex optimization problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \;\; \varphi(x) = f(x) + g(x) + h(Lx), \tag{1.1}$$

where L is a linear mapping, g and h are extended-real-valued nonsmooth functions, and f is continuously differentiable with Lipschitz continuous gradient. The proximal gradient method solves (1.1) when $h \circ L \equiv 0$, and the Douglas-Rachford algorithm solves it when $f \equiv 0$ and L is the identity. The idea here is that in many applications (1.1) is formulated such that the proximal mappings (*cf.* to §1.2.6) of g and h are easy to compute (refer to [14, §6 and §7], [49, 130] for extensive lists of common *proximable* functions) but this is not the case for $g + h \circ L$ or $f + h \circ L$. In fact, in general even the proximal mapping of $h \circ L$ cannot be efficiently computed based on that of h. Therefore, having algorithms for the sum of more terms would allow us to effectively tackle larger classes of applications.

A recent trend for solving problem (1.1), possibly with the smooth term $f \equiv 0$ or the nonsmooth term $g \equiv 0$, is to solve the monotone inclusion defined by

the primal-dual optimality conditions [38, 67, 31, 50, 53, 170, 62, 96, 32]. The popularity of this approach is mainly due to the fact that it results in *fully split* primal-dual algorithms, in the sense that the proximal mappings of g and h, the gradient of f, the linear mapping L and its adjoint are evaluated individually. In particular, there are no matrix inversions or inner loops involved.

Different convergence analysis techniques have been proposed in the literature for primal-dual algorithms. Some can be viewed as intelligent applications of classical splitting methods such as forward-backward splitting (FBS), Douglas-Rachford splitting (DRS) and forward-backward-forward splitting (FBFS), see for example [170, 53, 31, 27, 50], while others employ different tools to show convergence [38, 84, 62, 39]. Convergence rates of primal-dual schemes have also been analyzed using different approaches, see for example [104, 55, 113, 39].

Our approach here is a systematic one and relies on first introducing a new three term operator splitting method, *asymmetric forward-backward-adjoint* (AFBA) splitting that is designed for solving monotone inclusions involving the sum of three terms, a maximally monotone, a cocoercive and a bounded linear operator. While AFBA cannot be recovered from existing operator splitting methods, classical splittings DRS and FBS are its special cases. These are discussed in detail in Section 2.3. In Chapter 3 we develop a simple primal-dual framework (*cf.* Alg. 3.1) for problems of the form (1.1) by solving the primal-dual optimality conditions using the splitting method *AFBA*. Based on this approach one can obtain a wide range of algorithms by selecting different values for two scalar parameters θ and μ (*cf.* Alg. 3.1). Many of the resulting algorithms are new, while some extend previously proposed algorithms and/or result in less conservative stepsize conditions. Figure 1.1 provides an overview of several prominent special cases. The function l, stepsizes and other parameters are defined in Section 3.2 for the more general problem (3.1). These special cases are implemented in the open-source Julia Package *ProximalAlgorithms*[1] as primal-dual AFBA solver.

Next, let us consider some motivating examples most of which are revisited throughout this thesis. Many machine learning applications involve solving empirical risk minimizations (ERM) of the form

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \ \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(b_i,\langle a_i,x\rangle)+\lambda\Omega(x), \tag{1.2}$$

where Ω is a regularizer (*e.g.*, ℓ_1 norm or ℓ_2 norm, elastic net, etc), λ is a positive constant and the pair (a_i, b_i) represents the ith data. The loss function denoted by $\mathcal{L}$ measures the mismatch of the model and the observed data.

[1] https://github.com/kul-forbes/ProximalAlgorithms.jl

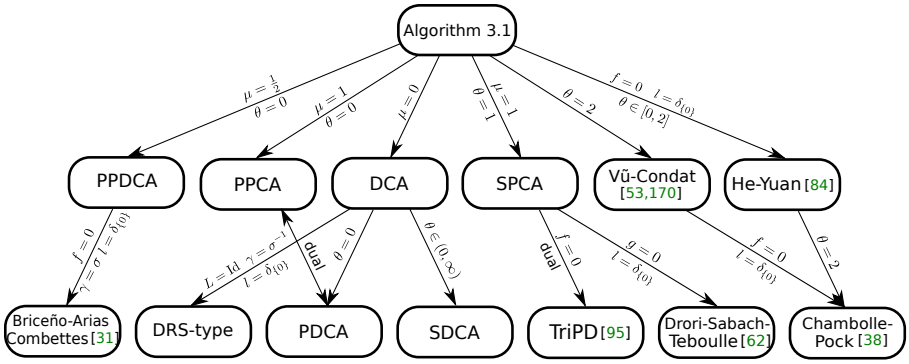

Figure 1.1: *Algorithm 3.1 and its special cases*

This formulation includes many important problems such as LASSO, logistic regression, SVM [83, 152]. Clearly, the ERM problem (1.2) can be written in the form of (1.1) by setting $g(x) = \lambda\Omega(x)$, $h(u) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(Y_i, u_i)$ (where $u = (u_1, u_2, \ldots, u_N)$) representing the loss function with the rows of L consisting of $a_i^\top$, and $f \equiv 0$. When the loss function is smooth one may choose to represent it using f and setting $h \circ L = 0$.

Another popular example in machine learning is the dual support vector machine problem [83]

$$\underset{\alpha_1,\ldots,\alpha_N}{\text{minimize}} \quad \tfrac{1}{2}\|\sum_{i=1}^{N}\alpha_i b_i a_i\|^2 - \sum_{i=1}^{N}\alpha_i \tag{1.3a}$$

$$\text{subject to} \quad 0 \le \alpha_i \le C, \quad i = 1, \ldots, N \tag{1.3b}$$

$$\sum_{i=1}^{N}\alpha_i b_i = 0. \tag{1.3c}$$

This problem can be written in the form of (1.1) by letting f represent the quadratic cost, g the indicator of box $[0, C]^N$, h the indicator of zero, and $L = b^\top$ where $b = (b_1, \ldots, b_N)$. This problem is used in Section 3.5 to compare the performance of several primal-dual algorithms.

Total variation denoising is another popular problem encountered in image processing applications [145, 37, 54]:

$$\begin{aligned} &\underset{x\in\mathbb{R}^n}{\text{minimize}} && \tfrac{1}{2}\|Ax-b\|^2 + \lambda\|Dx\|_{1,2} \\ &\text{subject to} && x \in C \end{aligned}$$

where D is the discrete gradient operator, $\|\cdot\|_{1,2}$ is the $\ell_{1,2}$ norm, and C enforces prior information on the target image, *e.g.*, the pixels being in the range $[0, 255]$. This problem is written in the form of (1.1) by setting $f(x) = \frac{1}{2}\|Ax-b\|^2$, g the indicator function of the set C, $h(x) = \|\cdot\|_{1,2}$, and $L = D$.

Model predictive control (MPC) is another application that can be formulated as in (1.1); the quadratic cost function may be represented by the smooth term f, the input and state constraints by g as the indicator function of the corresponding set, and the linear dynamics by $h \circ L$. Therefore, the resulting algorithm would involve simple matrix-vector products, and projection onto boxes and points.

Note that in many of the applications considered in this thesis we are interested in distributed algorithms in a *multi-agent* setting where a group of "agents" solve a minimization problem cooperatively. For example, in distributed MPC one may have several physically separate systems/agents each with its own dynamics who share a common goal, or in an ERM problem the data may be stored across multiple machines. A distributed algorithm would entail local computations by the systems/agents, and exchange of information with other agents.

Distributed algorithms are often derived by simply formulating a given problem in the form of (1.1) in such a way that solving it using a proximal algorithm (*e.g.*, one of the special cases of Algorithm 3.1) results in a desired distributed implementation. Let us clarify by considering as an example the problem of minimizing a finite sum problem:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^{N} f_i(x),$$

where f_i are convex continuously differentiable functions with Lipschitz-continuous gradient. In order to solve this problem in a distributed way over a network of agents, one may consider the equivalent problem

$$\underset{x_1,\ldots,x_N\in\mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{N} f_i(x_i), \quad \text{subject to} \;\; x_i = x_j \quad (i,j) \in \mathcal{E},$$

where $\mathcal{E}$ denotes the edge set of an underlying graph structure. We have effectively introduced slack variables for each agent in order to decouple the cost. The two problems are equivalent as long as the graph is connected. This problem can be written in the form of (1.1) with $f(x) = \sum_{i=1}^{N} f_i(x_i)$, and $g \equiv 0$, h the indicator of vector of zeros (of appropriate dimension), and $L = B^\top \otimes I_n$ with B representing the oriented node-arc incidence matrix (*cf.* §6.2 for details). Solving this problem with any of the special cases of Algorithm 3.1 results in schemes where the ith agent performs gradient operations on f_i, as well as some consensus-type updates using variables that it exchanges with relevant agents (neighbors). In such an algorithm, f_i (which may represent private data) does not need to be shared with other agents. This type of problem is studied in detail in Chapter 6.

Our focus so far has been on methods that split the cost function as the sum of several functions, with iterations that involve gradient or proximal operations on each function separately. However, in many large-scale applications updating the variables only after a full gradient or proximal update may be too costly, slow or physically infeasible. A simple and powerful idea is instead to update a subset of coordinates at every iteration. These methods are hereafter broadly referred to as block coordinate (BC) methods. BC methods have a long history and have been studied under various settings [19, 163, 166, 125, 15, 25, 106, 51, 22, 14].

In the context of multi-agent optimization, randomized BC methods admit updates that involve random activation of (subsets of) agents to perform local updates. These are sometimes referred to as asynchronous [88, 22, 134]. In this sense, occasionally, we also use the terminology synchronous to emphasize that at each clock tick all agents must perform their tasks and the iteration cannot proceed if any agent fails to do so. Note that this notion of asynchrony is quite different from the notions of *partial* and *total asynchrony* introduced in [20], since the information used by each agent must be up to date. We defer further details about notions of asynchrony and common computational models to Section 7.1.

Another interesting application of the BC framework is studied in Chapter 8 in the context of nonconvex optimization. It is shown that BC proximal gradient updates with a nonseparable nonsmooth term lead to stochastic and incremental methods for regularized finite sum and sharing problems. The analysis of this chapter is a departure from previous chapters where monotone operator theory and Fejér monotonicity made up most of the narrative. The Lyapunov function typically used in the convergence analysis of the nonconvex proximal gradient method is the cost function; however, in the BC setting with nonseparable nonsmooth term even in expectation it does not necessarily decrease along the trajectories. Instead we show that the forward-backward envelope (FBE) [132, 158] is a suitable Lyapunov function.

1.1 Overview of the thesis

We briefly describe the structure and content of the chapters. In Chapter 2 a three term operator splitting method, *asymmetric forward-backward-adjoint splitting* (AFBA) is introduced. Its convergence rate is studied under different assumptions and some prominent special cases are discussed. As depicted in Figure 1.2 this splitting is instrumental in the developments of the subsequent five chapters. In particular, as discussed above, in Chapter 3 a general primal-dual framework is developed (see Fig. 1.1) that relies on solving the primal-dual optimality conditions using AFBA. Moreover, in Section 3.4.1 linear convergence is established for all the special cases under mild regularity assumptions for the cost functions.

Chapters 2 and 3 are based on:

> **Latafat, P.**, and Patrinos, P. *Asymmetric forward–backward–adjoint splitting for solving monotone inclusions involving three operators.* Computational Optimization and Applications 68, 1 (Sep 2017), 57–93.
>
> **Latafat P.**, Patrinos P., *Primal-Dual Proximal Algorithms for Structured Convex Optimization: A Unifying Framework*, in Chapter 5 of Large-Scale and Distributed Optimization, (Giselsson P., and Rantzer A., eds.), vol. 2227 of Lecture Notes in Mathematics, Springer International Publishing, 2018, pp. 97-120.

In Chapter 4 a randomized block-coordinate primal-dual algorithm is introduced. The proposed algorithm features linear convergence rate when the functions involved are either piecewise linear-quadratic, or when they satisfy certain quadratic growth conditions. The developed algorithm is applied to the problem of multi-agent optimization on a graph, resulting in novel synchronous and asynchronous distributed methods. The proposed algorithms are fully distributed in the sense that the updates and the stepsizes of each agent only depend on local information. In fact, no (prior) global coordination is required. We showcase an application of our algorithm in distributed formation control. Moreover, as another application in Chapter 5 the problem of distributed model predictive control (DMPC) with coupling in the dynamics of the systems is considered. The resulting scheme does not require strong convexity, involves one round of communication at every iteration and allows a plug-and-play implementation where addition or removal of a subsystem only affects the neighboring nodes without the need for global coordination.

Chapter 4 is based on:

> **Latafat, P.**, Freris, N. M., and Patrinos, P. *A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization.* IEEE Transactions

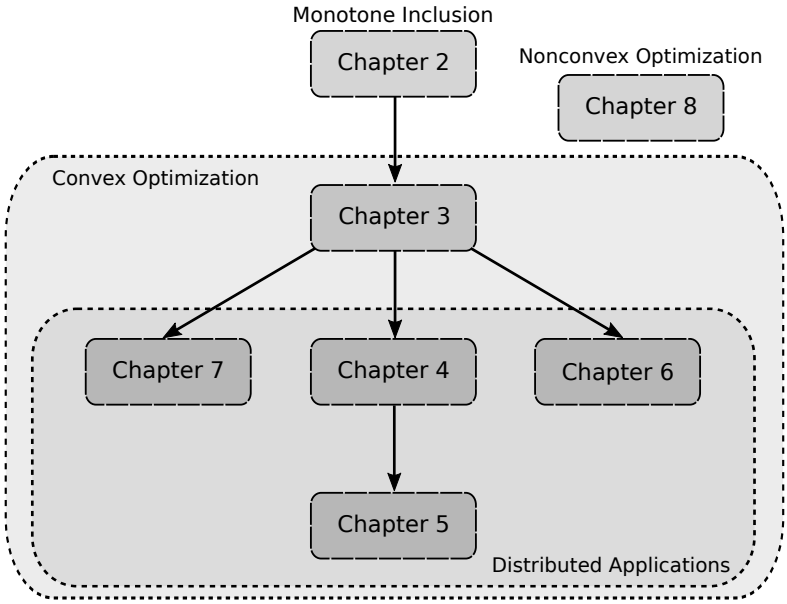

Figure 1.2: *Overview of the chapters*

on Automatic Control 64, 10 (10 2019), 4050–4065.

Chapter 5 is based on:

> **Latafat, P.**, Bemporad, A., and Patrinos, P. *Plug and play distributed model predictive control with dynamic coupling: A randomized primal-dual proximal algorithm.* In European Control Conference (ECC) (June 2018), pp. 1160–1165.

Chapter 6 considers a network of agents, each with its own private cost consisting of the sum of two possibly nonsmooth convex functions, one of which is composed with a linear operator. At every iteration each agent performs local calculations and can only communicate with its neighbors. The goal is to minimize the aggregate of the private cost functions and reach a consensus over a graph. A special case of AFBA is used to develop a primal-dual algorithm for solving this minimization over the communication graph. It is demonstrated through computational experiments how suitably selecting the parameters of our algorithm can lead to larger stepsizes and yield better performance.

Chapter 6 is based on:

> **Latafat, P.**, Stella, L., and Patrinos, P. *New primal-dual proximal algorithm for distributed optimization.* In 55th IEEE Conference on Decision and Control

(CDC) (Dec 2016), pp. 1959–1964.

Chapter 7 considers primal-dual algorithms over message-passing (multi-agent) architectures with communication delays. It is assumed that the delay with respect to each neighbor is bounded but otherwise arbitrary. The global optimization problem is the aggregate of the local cost functions and a common Lipschitz differentiable function. When the coupling between agents is represented only through the common function, the primal-dual algorithm proposed by Vũ and Condat [170, 53] is employed. In the case when the linear maps introduce additional couplings between agents a new algorithm is developed. Moreover, a randomized variant of this algorithm is presented that allows the agents to wake up at random and independently from one another. The convergence of the proposed algorithms is established under strong convexity assumptions.

Chapter 7 is based on:

> **Latafat P.** and Patrinos. P. *Primal-dual algorithms for multi-agent structured optimization over message-passing architectures with bounded communication delays* (submitted 2019).

Chapter 8 deals with block-coordinate proximal gradient methods for minimizing the sum of a separable smooth function and a (nonseparable) nonsmooth function, both of which are allowed to be nonconvex. The main tool in our analysis is the forward-backward envelope (FBE), which serves as a particularly suitable continuous and real-valued Lyapunov function. Global and linear convergence results are established when the cost function satisfies the Kurdyka-Łojasiewicz property without imposing convexity requirements on the smooth function. Two prominent special cases of the investigated setting are regularized finite sum minimization and the sharing problem; in particular, an immediate byproduct of our analysis leads to novel convergence results and rates for the popular Finito/MISO algorithm in the nonsmooth and nonconvex setting with very general sampling strategies.

Chapter 8 is based on:

> **Latafat P.**, Themelis A. and Patrinos P. *Block-coordinate and incremental aggregated proximal gradient methods for nonsmooth nonconvex problems.* arXiv:1906.10053 (submitted 2019).

1.2 Preliminary material

In this section we recap some standard definitions and results that are used throughout the thesis [143, 13, 144, 18, 14].

The set of real numbers is denoted by $\mathbb{R}$. The set of extended real numbers is defined as $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. We denote $(a, b) \coloneqq \{x \mid a < x < b\}$ where a, b can be taken to be $\pm\infty$, and square brackets $[a, b]$ when equality is allowed in the definition. Intervals $(a, b]$ and $[a, b)$ are defined accordingly.

1.2.1 Vector notation

We denote by $\mathbb{R}^n$ the standard n-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. For $w = (w_1, \ldots, w_N) \in \mathbb{R}^n$, $w_i \in \mathbb{R}^{n_i}$ is used to denote its i-th (block) coordinate.

The sets of symmetric, symmetric positive semi-definite and symmetric positive definite n-by-n matrices are denoted by $\mathbb{S}^n$, $\mathbb{S}^n_+$ and $\mathbb{S}^n_{++}$, respectively. We also write $P \succeq 0$ and $P \succ 0$ for $P \in \mathbb{S}^n_+$ and $P \in \mathbb{S}^n_{++}$, respectively. For $P \in \mathbb{S}^n_{++}$ we define the scalar product $\langle x, y \rangle_P = \langle x, Py \rangle$ and the induced norm $\|x\|_P = \sqrt{\langle x, x \rangle_P}$. The identity matrix is denoted by $\mathrm{I}_n \in \mathbb{R}^{n \times n}$; we write I when no ambiguity occurs.

1.2.2 Sequences

We use the notation $(w^k)_{k \in I}$ to denote a sequence with indices in the set $I \subseteq \mathbb{N}$. Occasionally, when dealing with scalar sequences we use the subscript notation $(\gamma_k)_{k \in I}$.

Definition 1.1 (Fejér monotonicity)**.** *A sequence $(w^k)_{k \in \mathbb{N}}$ is said to be* Fejér *monotone with respect to a nonempty set $\mathcal{U} \subseteq \mathbb{R}^n$ if for all $v \in \mathcal{U}$ and all $k \in \mathbb{N}$*

$$\|w^{k+1} - v\| \leq \|w^k - v\|.$$

It is said to be quasi-Fejér *monotone with respect to $\mathcal{U} \subseteq \mathbb{R}^n$ if for all $v \in \mathcal{U}$, there exists a summable nonnegative sequence $(\varepsilon_k)_{k \in \mathbb{N}}$ such that for all $k \in \mathbb{N}$*

$$\|w^{k+1} - v\|^2 \leq \|w^k - v\|^2 + \varepsilon_k.$$

Moreover, given $S \in \mathbb{S}^n_{++}$, we say that a sequence is S-(quasi)-Fejér monotone with respect to $\mathcal{U} \subseteq \mathbb{R}^n$ if it is (quasi)-Fejér monotone with respect to $\mathcal{U}$ in the space equipped with $\langle \cdot, \cdot \rangle_S$.

We use the following notions of linear convergence:

- A sequence $(w^k)_{k\in\mathbb{N}}$ is said to converge to a point $w^\star$ (at least) *Q-linearly* (with quotient rate) with Q-factor given by $\sigma \in (0,1)$, if there exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$,

$$\|w^{k+1} - w^\star\| \leq \sigma \|w^k - x^\star\|.$$

- A sequence $(w^k)_{k\in\mathbb{N}}$ is said to converge to a point $w^\star$ (at least) *R-linearly* (with root rate) if there exists a sequence of nonnegative scalars $(v^k)_{k\in\mathbb{N}}$ such that $\|w^k - w^\star\| \leq v^k$ and $(v^k)_{k\in\mathbb{N}}$ converges Q-linearly to zero.

1.2.3 Functions and operators

An operator (or set-valued mapping) $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^d$ maps each point $x \in \mathbb{R}^n$ to a subset Fx of $\mathbb{R}^d$. We denote the domain of F by

$$\operatorname{dom} F \coloneqq \{x \in \mathbb{R}^n \mid Fx \neq \emptyset\},$$

its graph by

$$\operatorname{gra} F \coloneqq \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^d \mid y \in Fx\},$$

the set of its zeros by $\operatorname{zer} F \coloneqq \{x \in \mathbb{R}^n \mid 0 \in Fx\}$, and the set of its fixed points by $\operatorname{fix} F \coloneqq \{x \mid x \in Fx\}$. The inverse of F is defined through its graph: $\operatorname{gra} F^{-1} \coloneqq \{(y, x) \mid (x, y) \in \operatorname{gra} F\}$. The *resolvent* of F is defined by $\mathrm{J}_F \coloneqq (\mathrm{id} + F)^{-1}$, where id denotes the identity operator. The mapping F is called *monotone* if for all $(x, y), (x', y') \in \operatorname{gra} F$

$$0 \leq \langle x - x', y - y' \rangle,$$

and is said to be maximally monotone if its graph is not strictly contained in the graph of another monotone operator. The mapping F is said to be *nonexpansive* if for all $(x, y), (x', y') \in \operatorname{gra} F$

$$\|y - y'\| \leq \|x - x'\|,$$

and *firmly nonexpansive* if

$$\|y - y'\|^2 \leq \|x - x'\|^2 - \|(x - y) - (x' - y')\|^2.$$

The resolvent J_F is firmly nonexpansive (with $\operatorname{dom} \mathrm{J}_F = \mathbb{R}^n$) if and only if F is (maximally) monotone [13, Prop. 23.8].

Operator F is *outer semicontinuous (osc)* at $\bar{x} \in \operatorname{dom} F$ if

$$\limsup_{x \to \bar{x}} Fx := \{y \mid \exists x^k \to \bar{x}, \exists y^k \to y \text{ with } y^k \in Fx^k\} \subseteq F\bar{x}. \tag{1.4}$$

Operator F is osc everywhere if and only if its graph is closed in $\mathbb{R}^n \times \mathbb{R}^d$.

Next, let us define the notion of *metric subregularity* which is used throughout the thesis for establishing linear convergence. Metric subregularity is a "one-point" version of metric regularity. We refer the interested reader to [61, §3] and [144, §9] for further discussion.

Definition 1.2 (metric subregularity)**.** *A set-valued mapping $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^d$ is* metrically subregular *at $\bar{x}$ for $\bar{y}$ if $(\bar{x}, \bar{y}) \in \operatorname{gra} F$ and there exists a positive constant η together with a* neighborhood of subregularity *$\mathcal{U}$ of $\bar{x}$ such that*

$$\operatorname{dist}(x, F^{-1}\bar{y}) \leq \eta \operatorname{dist}(\bar{y}, Fx) \ \ \forall x \in \mathcal{U}. \tag{1.5}$$

If the following stronger condition holds

$$\|x - \bar{x}\| \leq \eta \operatorname{dist}(\bar{y}, Fx) \ \ \forall x \in \mathcal{U}, \tag{1.6}$$

then F is said to be strongly subregular *at $\bar{x}$ for $\bar{y}$.*

Moreover, we say that F is globally (strongly) subregular at $\bar{x}$ for $\bar{y}$ if (strong) subregularity holds with $\mathcal{U} = \mathbb{R}^n$.

Definition 1.3 (cocoercivity)**.** *An operator $C : \mathbb{R}^p \to \mathbb{R}^p$ is said to be cocoercive with respect to $\|\cdot\|_V$ with $V \in \mathbb{S}^p_{++}$ if for all $z, z' \in \mathbb{R}^p$*

$$\langle Cz - Cz', z - z' \rangle \geq \|Cz - Cz'\|^2_{V^{-1}}. \tag{1.7}$$

For an extended-real-valued function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, its *domain* is the set

$$\operatorname{dom} f := \{x \in \mathbb{R}^n \mid f(x) < \infty\}.$$

Function f is said to be proper if its domain is a nonempty set. The *epigraph* of f is

$$\operatorname{epi} f := \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) < \alpha\},$$

and f is said to be closed if $\operatorname{epi} f$ is a closed set in $\mathbb{R}^{n+1}$. For any $\alpha \in \mathbb{R}$, *α-(sub)level set* of f is

$$\operatorname{lev}_{\leq \alpha} f := \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}.$$

Function f is called *lower semicontinuous (lsc)* at $\bar{x} \in \mathbb{R}^n$ if

$$f(\bar{x}) \leq \liminf_{x \to \bar{x}} f(x).$$

It is called lsc if it is lsc at all points $\mathbb{R}^n$. For extended-real-valued functions, lower semicontinuity, closedness, and all level sets being closed are equivalent [144, Thm 1.6].

The *indicator function* of a set $X \subseteq \mathbb{R}^n$ is given by

$$\delta_X(x) := \begin{cases} 0 & \text{if } x \in X \\ +\infty & \text{if } x \notin X. \end{cases}$$

The indicator function δ_X is closed if and only if X is a closed set. The projection onto and the distance from X with respect to $\|\cdot\|_V$ are denoted by

$$\mathcal{P}_X^V(z) := \operatorname*{arg\,min}_{w \in X} \{\|w - x\|_V\}, \quad \operatorname{dist}_V(z, X) := \inf_{x \in X} \{\|z - x\|_V\},$$

respectively. The absence of super/subscript V implies the same definitions with respect to the canonical norm.

A set $X \subseteq \mathbb{R}^n$ is said to be *polyhedral* if it can be expressed as the intersection of finitely many closed half-spaces and/or hyperplanes.

An important class of functions prevalent in optimization is the class of *piecewise linear-quadratic* (PLQ) functions, which is closed under scalar multiplication, addition, conjugation and Moreau envelope [144].

Definition 1.4 (piecewise linear-quadratic)**.** *A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called piecewise linear-quadratic (PLQ) if its domain can be represented as the union of finitely many polyhedral sets, and in each such set $f(x)$ is given by an expression of the form $\frac{1}{2}\langle x, Qx\rangle + \langle d, x\rangle + c$, for some $c \in \mathbb{R}$, $d \in \mathbb{R}^n$, and $Q \in \mathbb{S}^n$.*

1.2.4 Subgradients

Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper lsc function. Then, $v \in \mathbb{R}^n$ is a *regular subgradient* of f at $\bar{x}$ if

$$\liminf_{\substack{x \to \bar{x} \\ x \neq \bar{x}}} \frac{f(x) - f(\bar{x}) - \langle v, x - \bar{x}\rangle}{\|x - \bar{x}\|} \geq 0.$$

The set of all regular subgradient of f at $\bar{x}$ is called the subdifferential of f at $\bar{x}$ and is denoted by $\hat{\partial} f(\bar{x})$. The regular subdifferential is closed and convex.

A vector $v \in \mathbb{R}^n$ is a *(limiting) subgradient* of f at $\bar{x} \in \operatorname{dom} f$ if there exists sequences $x^k \to \bar{x}$ with $f(x^k) \to f(\bar{x})$, and $v^k \to v$ such that $v^k \in \hat{\partial} f(x^k)$. The set of all subgradients of f at $\bar{x}$ is called the limiting subdifferential of f at $\bar{x}$ and is denoted $\partial f(\bar{x})$. Local minimizers of f are characterized by the following extension of Fermat's rule [144, Thm. 10.1].

Lemma 1.5. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper lsc function. If f has a local minimum at $\bar{x}$, then $0 \in \hat{\partial} f(\bar{x})$. If f is also convex, then this condition is equivalent to $\bar{x} \in \arg\min f$.*

The regular and limiting subdifferentials coincide with the usual notion of subdifferential for convex functions [144, Prop. 8.12].

Lemma 1.6. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper lsc convex function. For any $x \in \operatorname{dom} f$*

$$\partial f(x) = \hat{\partial} f(x) = \{y \in \mathbb{R}^n \mid \forall z \in \mathbb{R}^n,\, \langle z - x, y\rangle + f(x) \le f(z)\}.$$

We conclude by noting that if f is a proper lsc convex function, then ∂f is maximally monotone [143, Cor. 31.5.2].

1.2.5 Conjugate functions and infimal convolution

The *Fenchel conjugate* of an extended-real-valued function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, denoted f^*, is defined by

$$f^*(v) := \sup_{x \in \mathbb{R}^n} \{\langle v, x\rangle - f(x)\}.$$

The conjugate function f^* is lsc and convex. The *Fenchel-Young* inequality states that for all $x, u \in \mathbb{R}^n$

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

In the special case when $f = \frac{1}{2}\|\cdot\|_V^2$ for some symmetric positive definite matrix V, this gives:

$$\langle x, u\rangle \le \tfrac{1}{2}\|x\|_V^2 + \tfrac{1}{2}\|u\|_{V^{-1}}^2. \tag{1.8}$$

The *infimal convolution* of two proper extended-real-valued functions $f, g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is defined by

$$(f \,\square\, g)(x) := \inf_{z \in \mathbb{R}^n} \{f(z) + g(x - z)\},$$

and the conjugate of the infimal convolution is given by

$$(f \,\square\, g)^* = f^* + g^*.$$

1.2.6 Proximal mappings and Moreau envelopes

For a proper lsc function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $V \in \mathbb{S}^n_{++}$, the *V-proximal mapping* of f is defined as the following set-valued mapping

$$\operatorname{prox}_f^V(x) \coloneqq \operatorname*{arg\,min}_{w \in \mathbb{R}^n} \{f(w) + \tfrac{1}{2}\|x - w\|_V^2\}.$$

When $V = \gamma^{-1}\mathrm{I}_n$ is a multiple of the identity matrix I_n, the notation $\operatorname{prox}_{\gamma f}$ is typically used and γ is referred to as a stepsize. The value function associated to the minimization defining the proximal mapping is the *Moreau envelope*

$$f^V(x) \coloneqq \min_{w \in \mathbb{R}^n} \left\{f(w) + \tfrac{1}{2}\|w - x\|_V^2\right\}.$$

Function f is called *prox-bounded* if $f + \frac{1}{2\gamma}\|\cdot\|^2$ is lower bounded for some $\gamma > 0$. The supremum of the set of all such γ is the threshold γ_f of prox-boundedness for f.

If f is a proper lsc convex function, then the V-proximal mapping is uniquely determined by the resolvent of $V^{-1}\partial f$:

$$\operatorname{prox}_f^V(x) = (\mathrm{id} + V^{-1}\partial f)^{-1}x,$$

and the Moreau decomposition is given by

$$\operatorname{prox}_f^{V^{-1}}(x) + V \operatorname{prox}_{f^*}^V(V^{-1}x) = x.$$

Chapter 2

Asymmetric forward-backward-adjoint splitting

This chapter is based on:

Latafat, P., and Patrinos, P. *Asymmetric forward–backward–adjoint splitting for solving monotone inclusions involving three operators.* Computational Optimization and Applications 68, 1 (Sep 2017), 57–93.

2.1 Introduction

The focus of this chapter is on solving monotone inclusion problems of the form

$$0 \in Ax + Mx + Cx, \tag{2.1}$$

where A is a maximally monotone operator, M is a bounded linear operator and C is cocoercive. The most well known algorithms for solving monotone inclusion problems are forward-backward (FBS), Douglas-Rachford (DRS) and forward-backward-forward (FBFS) splittings [118, 108, 49, 130, 29, 162]. The operator splitting schemes FBS and DRS are not well suited to handle (2.1) since they are designed for monotone inclusions involving the sum of two operators. The FBFS can solve (2.1) by considering $M+C$ as one Lipschitz continuous operator. However, being blind to the fact that C is cocoercive, it would require two evaluations of C per iteration. Many other variations of the three main splittings have been proposed over time that can be seen as intelligent applications of these classical methods (see for example [30, 38, 31, 170, 53]).

In this chapter a new algorithm called *asymmetric-forward-backward-adjoint splitting* (AFBA) to solve the monotone inclusion (2.1), without resorting to any kind of reformulation of the problem. One important property of AFBA

is that it includes asymmetric preconditioning. This gives great flexibility to the algorithm, and indeed it is the key for recovering and unifying existing primal-dual proximal splitting schemes for convex optimization and devising new ones. More importantly, it can deal with problems involving three operators, one of which is cocoercive. It is observed that FBS, DRS, and the Proximal Point Algorithm (PPA) can be derived as special cases of our method. Another notable special case is the method proposed by Solodov and Tseng for variational inequalities in [151, Alg. 2.1]. Moreover, when the cocoercive term C is absent in (2.1), in a yet another special case it coincides with the FBFS when its Lipschitz operator is skew-adjoint. Recently, a new splitting scheme was proposed in [58] for solving monotone inclusions involving the sum of three operators, one of which is cocoercive. This method can be seen as Douglas-Rachford splitting with an extra forward step for the cocoercive operator. As a special case of our scheme, we propose an algorithm that also bears heavy resemblance to the classic Douglas-Rachford splitting with an extra forward step (see Algorithm 2.2). The proposed algorithm differs from that of [58], in that the forward step precedes the two backward updates.

As another contribution, big-$O(1/(k+1))$ and little-$o(1/(k+1))$ convergence rates are derived for AFBA (see Theorem 2.3). It is observed that in many cases these convergence rates are guaranteed under mild conditions. In addition, under metric subregularity of the underlying operator, linear convergence is guaranteed without restrictions on the parameters (see Theorem 2.4). Given that AFBA generalizes a wide range of algorithms, this analysis provides a systematic way to deduce convergence rates for many algorithms.

Notation: In this section, we consider real Hilbert spaces. We denote the scalar product and the induced norm of a Hilbert space by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ respectively. Let $\mathcal{H}$ and $\mathcal{G}$ be real Hilbert spaces. We denote by $\mathscr{B}(\mathcal{H}, \mathcal{G})$ the space of bounded linear operators from $\mathcal{H}$ to $\mathcal{G}$ and set $\mathscr{B}(\mathcal{H}) = \mathscr{B}(\mathcal{H}, \mathcal{H})$. The space of self-adjoint operators is denoted by $\mathcal{S}(\mathcal{H}) = \{L \in \mathscr{B}(\mathcal{H}) | L = L^*\}$, where L^* denotes the adjoint of L. The *Loewner partial ordering* on $\mathcal{S}(\mathcal{H})$ is denoted by $\succeq$. Let $\tau \in]0, +\infty[$ and define the space of τ-strongly positive self-adjoint operators by $\mathcal{S}_\tau(\mathcal{H}) = \{U \in \mathcal{S}(\mathcal{H}) | U \succeq \tau \mathrm{id}\}$. For $U \in \mathcal{S}_\tau(\mathcal{H})$, define the scalar product and the norm by $\langle x, y \rangle_U = \langle x, Uy \rangle$, and $\|x\|_U = \sqrt{\langle x, Ux \rangle}$. We also define the Hilbert space $\mathcal{H}_U$ by endowing $\mathcal{H}$ with the scalar product $\langle x, y \rangle_U$.

All other preliminary definitons and notations in Section 1.2 extend directly to real Hilbert spaces considered in this chapter.

2.2 Asymmetric forward-backward-adjoint splitting

Let $\mathcal{H}$ be a real Hilbert space and consider the problem of finding $z \in \mathcal{H}$ such that

$$0 \in Tz \quad \text{where} \quad T \coloneqq A + M + C, \tag{2.2}$$

where operators A, M, C satisfy the following assumption:

Assumption 2.I. Throughout the chapter the following hold:

(*i*) Operator $A : \mathcal{H} \rightrightarrows \mathcal{H}$ is maximally monotone and $M \in \mathscr{B}(\mathcal{H})$ is monotone.

(*ii*) Operator $C : \mathcal{H} \to \mathcal{H}$ is β-cocoercive with respect to $\|\cdot\|_P$, where $\beta \in]1/4, +\infty)$ and $P \in \mathscr{S}_\rho(\mathcal{H})$ for some $\rho \in (0, \infty)$, *i.e.*,

$$(\forall z \in \mathcal{H})(\forall z' \in \mathcal{H}) \quad \langle Cz - Cz', z - z' \rangle \geq \beta \|Cz - Cz'\|_{P^{-1}}^2.$$

It is important to notice that the freedom in choosing P is a crucial part of our method. In Assumption 2.I(*ii*) we consider cocoercivity with respect to $\|\cdot\|_P$ with $\beta \in]1/4, +\infty)$. However, this is by no means a restriction of our setting; another approach would have been to consider cocoercivity with respect to the canonical norm $\|\cdot\|$ with $\beta \in]0, +\infty)$ but this would lead to statements involving $\|P\|$ and $\|P^{-1}\|$. Indeed convergence with respect to $\|\cdot\|$ and $\|\cdot\|_P$ are equivalent but in using $\|\cdot\|_P$ we simplify the notation substantially.

In addition, let S be a strongly positive, self-adjoint operator, $K \in \mathscr{B}(\mathcal{H})$ a skew-adjoint operator, *i.e.*, $K^* = -K$, $H = P + K$ and $(\lambda_k)_{k \in \mathbb{N}}$ is a sequence satisfying (2.5). Then, the algorithm for solving the monotone inclusion described above is as follows:

Algorithm 2.1 Asymmetric Forward-Backward-Adjoint Splitting (AFBA)

Inputs: $z^0 \in \mathcal{H}$

for $k = 0, 1, \ldots,$ **do**

$$\bar{z}^k = (H + A)^{-1}(H - M - C)z^k$$
$$\tilde{z}^k = \bar{z}^k - z^k$$
$$\alpha_k = \frac{\lambda_k \|\tilde{z}^k\|_P^2}{\|(H + M^*)\tilde{z}^k\|_{S^{-1}}^2}$$
$$z^{k+1} = z^k + \alpha_k S^{-1}(H + M^*)\tilde{z}^k$$

Before proceeding with the convergence analysis let us define

$$D = (H + M^*)^* S^{-1}(H + M^*). \tag{2.3}$$

Since $P \in \mathscr{S}_\rho(\mathcal{H})$ for some $\rho \in (0,\infty)$, K is skew-adjoint, and $M \in \mathscr{B}(\mathcal{H})$ is monotone, it follows that $\langle (H+M^*)z, z\rangle \geq \rho\|z\|^2$ for all $z \in \mathcal{H}$, and we have

$$(\forall z \in \mathcal{H}) \quad \langle z, Dz\rangle = \|(H+M^*)z\|^2_{S^{-1}} \geq \rho^2\|S\|^{-1}\|z\|^2. \tag{2.4}$$

Hence, $D \in \mathscr{S}_\nu(\mathcal{H})$ with $\nu = \rho^2\|S\|^{-1}$. Notice that the denominator of α_k in Algorithm 2.1 is equal to the left hand side of (2.4) for $z = \tilde{z}^k$ and thus it is bounded below by $\rho^2\|S\|^{-1}\|\tilde{z}^k\|^2$.

2.2.1 Convergence analysis

In this section we analyze convergence and rate of convergence of Algorithm 2.1. We also consider a special case of the algorithm in which it is possible to relax strong positivity of P to positivity. We begin by stating our main convergence result. The proof relies on showing that the sequence $(z^k)_{k\in\mathbb{N}}$ is Fejér monotone with respect to $\operatorname{zer}(A+M+C)$ in the Hilbert space $\mathcal{H}$ equipped with the scalar product $\langle\cdot,\cdot\rangle_S$.

Theorem 2.1. *Consider Algorithm 2.1 under Assumption 2.1 and assume that* $\operatorname{zer}(T) \neq \emptyset$ *where* $T = A + M + C$. *Let* $\sigma \in (0,\infty)$, $S \in \mathscr{S}_\sigma(\mathcal{H})$, $K \in \mathscr{B}(\mathcal{H})$ *a skew-adjoint operator, and* $H = P + K$. *Let* $(\lambda_k)_{k\in\mathbb{N}}$ *be a sequence such that*

$$(\lambda_k)_{k\in\mathbb{N}} \subseteq [0,\delta] \quad \textit{with} \quad \delta = 2 - \frac{1}{2\beta},\ \delta > 0,\ \liminf_{k\to\infty} \lambda_k(\delta-\lambda_k) > 0. \tag{2.5}$$

Then the following hold:

(i) $(z^k)_{k\in\mathbb{N}}$ *is* S*-Fejér monotone with respect to* $\operatorname{zer}(T)$.

(ii) $(\tilde{z}^k)_{k\in\mathbb{N}}$ *converges strongly to zero.*

(iii) $(z^k)_{k\in\mathbb{N}}$ *converges weakly to a point in* $\operatorname{zer}(T)$.

Furthermore, when $C \equiv 0$ *all of the above statements hold with* $\delta = 2$.

Proof. The operators $\tilde{A} = P^{-1}(A+K)$ and $\tilde{B} = P^{-1}(M+C-K)$ are monotone in the Hilbert space $\mathcal{H}_P$. We observe that

$$\bar{z}^k = (H+A)^{-1}(H-M-C)z^k = (\mathrm{id}+\tilde{A})^{-1}(\mathrm{id}-\tilde{B})z^k.$$

Therefore $z^k - \tilde{B}z^k \in \bar{z}^k + \tilde{A}\bar{z}^k$, or $-\tilde{z}^k - \tilde{B}z^k \in \tilde{A}\bar{z}^k$. Since $-\tilde{B}z^\star \in \tilde{A}z^\star$ for $z^\star \in \operatorname{zer}(T)$ by monotonicity of $\tilde{A}$ on $\mathcal{H}_P$ we have

$$\langle \tilde{B}z^k - \tilde{B}z^\star + \tilde{z}^k, z^\star - \bar{z}^k\rangle_P \geq 0.$$

Then,

$$\begin{aligned}0 &\leq \langle \tilde{B}z^k - \tilde{B}z^\star + \tilde{z}^k, z^\star - \bar{z}^k\rangle_P \\ &= \langle P^{-1}(M+C-K)z^k - P^{-1}(M+C-K)z^\star + \tilde{z}^k, z^\star - \bar{z}^k\rangle_P. \\ &= \langle (M-K)(z^k - z^\star) + Cz^k - Cz^\star + P\tilde{z}^k, z^\star - \bar{z}^k\rangle. \end{aligned} \tag{2.6}$$

On the other hand

$$\begin{aligned}\langle Cz^k - Cz^\star, z^\star - \bar{z}^k\rangle &= \langle Cz^k - Cz^\star, z^k - \bar{z}^k\rangle + \langle Cz^k - Cz^\star, z^\star - z^k\rangle \\ &\leq \frac{\epsilon}{2}\|\tilde{z}^k\|_P^2 + \frac{1}{2\epsilon}\|Cz^k - Cz^\star\|_{P^{-1}}^2 \\ &\quad + \langle Cz^k - Cz^\star, z^\star - z^k\rangle \\ &\leq \frac{\epsilon}{2}\|\tilde{z}^k\|_P^2 + \left(1 - \frac{1}{2\epsilon\beta}\right)\langle Cz^k - Cz^\star, z^\star - z^k\rangle.\end{aligned}$$

The first inequality follows from Fenchel-Young inequality for $\frac{\epsilon}{2}\|\cdot\|_P^2$, while the second from β-cocoercivity of C with respect to $\|\cdot\|_P$. Set $\epsilon := \frac{1}{2\beta}$ so that

$$\langle Cz^k - Cz^\star, z^\star - \bar{z}^k\rangle \leq \frac{1}{4\beta}\|\tilde{z}^k\|_P^2. \tag{2.7}$$

In turn, (2.6), (2.7) and monotonicity of $M-K$, yield

$$\begin{aligned}0 &\leq \langle (M-K)(z^k - z^\star) + Cz^k - Cz^\star + P\tilde{z}^k, z^\star - \bar{z}^k\rangle \\ &\leq \langle (M-K)(z^k - z^\star), z^\star - z^k\rangle + \langle (M-K)(z^k - z^\star), z^k - \bar{z}^k\rangle \\ &+ \frac{1}{4\beta}\|\tilde{z}^k\|_P^2 + \langle P\tilde{z}^k, z^\star - z^k\rangle + \langle P\tilde{z}^k, z^k - \bar{z}^k\rangle \\ &\leq \langle z^k - z^\star, (M^* + K)(z^k - \bar{z}^k)\rangle + \frac{1}{4\beta}\|\tilde{z}^k\|_P^2 + \langle P\tilde{z}^k, z^\star - z^k\rangle - \|\tilde{z}^k\|_P^2 \\ &= \langle z^k - z^\star, -(M^* + H)\tilde{z}^k\rangle - \left(1 - \frac{1}{4\beta}\right)\|\tilde{z}^k\|_P^2,\end{aligned}$$

or equivalently

$$\langle z^k - z^\star, (M^* + H)\tilde{z}^k\rangle \leq -\left(1 - \frac{1}{4\beta}\right)\|\tilde{z}^k\|_P^2. \tag{2.8}$$

For notational convenience define $\delta := 2 - \frac{1}{2\beta}$. We show that $\|z^k - z^\star\|_S^2$ is decreasing using (2.8) together with step 3 and 4 of Algorithm 2.1:

$$\begin{aligned}
\|z^{k+1}-z^\star\|_S^2 &= \|z^k - z^\star + \alpha_k S^{-1}(H+M^*)\tilde{z}^k\|_S^2 \\
&= \|z^k - z^\star\|_S^2 + 2\alpha_k\langle z^k - z^\star, (H+M^*)\tilde{z}^k\rangle + \alpha_k^2\|(H+M^*)\tilde{z}^k\|_{S^{-1}}^2 \\
&\leq \|z^k - z^\star\|_S^2 - \alpha_k\delta\|\tilde{z}^k\|_P^2 + \alpha_k^2\|(H+M^*)\tilde{z}^k\|_{S^{-1}}^2 \\
&= \|z^k - z^\star\|_S^2 - \delta\lambda_k\frac{\|\tilde{z}^k\|_P^4}{\|(H+M^*)\tilde{z}^k\|_{S^{-1}}^2} + \lambda_k^2\frac{\|\tilde{z}^k\|_P^4}{\|(H+M^*)\tilde{z}^k\|_{S^{-1}}^2} \\
&= \|z^k - z^\star\|_S^2 - \lambda_k(\delta-\lambda_k)\|(H+M^*)\tilde{z}^k\|_{S^{-1}}^{-2}\|\tilde{z}^k\|_P^4 \qquad (2.9) \\
&= \|z^k - z^\star\|_S^2 - \lambda_k(\delta-\lambda_k)\|P^{-1/2}S^{-1/2}(H+M^*)\tilde{z}^k\|_P^{-2}\|\tilde{z}^k\|_P^4 \\
&\leq \|z^k - z^\star\|_S^2 - \lambda_k(\delta-\lambda_k)\|P^{-1/2}S^{-1/2}(H+M^*)\|_P^{-2}\|\tilde{z}^k\|_P^2 \\
&= \|z^k - z^\star\|_S^2 - \lambda_k(\delta-\lambda_k)\|S^{-1/2}(H+M^*)P^{-1/2}\|^{-2}\|\tilde{z}^k\|_P^2. \qquad (2.10)
\end{aligned}$$

Furthermore, when $C \equiv 0$ all the above analysis holds with $\delta = 2$.

(*i*): Inequality (2.10) and $(\lambda_k)_{k\in\mathbb{N}} \subseteq [0,\delta]$ show that $(z^k)_{k\in\mathbb{N}}$ is S-Fejér monotone with respect to $\operatorname{zer}(T)$.

(*ii*): From (2.10) and $\liminf_{k\to\infty} \lambda_k(\delta-\lambda_k) > 0$, it follows that $\tilde{z}^k \to 0$.

(*iii*): Define

$$w^k := -(H-M)\tilde{z}^k + C\bar{z}^k - Cz^k. \qquad (2.11)$$

It follows from (2.11), linearity of $H-M$, cocoercivity of C and (*ii*) that

$$w^k \to 0. \qquad (2.12)$$

By step 1 of Algorithm 2.1 we have $(H-M-C)z^k \in (H+A)\bar{z}^k$, which together with (2.11) yields

$$w^k \in T\bar{z}^k. \qquad (2.13)$$

Now let z be a weak sequential cluster point of $(z^k)_{k\in\mathbb{N}}$, say $z_{k_q} \rightharpoonup z$. It follows from (*ii*) that $\bar{z}_{k_q} \rightharpoonup z$, and from (2.12) that $w_{k_q} \to 0$. Altogether, by (2.13), the members of the sequence $(\bar{z}_{k_q}, w_{k_q})_{q\in\mathbb{N}}$ belong to $\operatorname{gra}(T)$. Additionally, by [13, Ex. 20.31, 20.34 and Cor. 25.5(i)], T is maximally monotone. Then, an appeal to [13, Prop. 20.38(ii)] yields $(z,0) \in \operatorname{gra}(T)$. This together with (*i*) and [13, Thm. 5.5] completes the proof. □

Remark 2.2. It can be shown based on (2.8) that for a constant $\lambda_n = \lambda \in]0, \delta]$, the fixed-point mapping behind Algorithm 2.1 is $\frac{\delta-\lambda}{\lambda}$-*strongly quasi-nonexpansive* (SQNE), in the sense of [35, Def. 2.1.38]. It is well known that an averaged operator is also SQNE. In fact if an operator is $\frac{\lambda}{\delta}$-averaged then it is $\frac{\delta-\lambda}{\lambda}$-SQNE. But the converse is not true in general. See [35, Fig. 2.10] for an overview of the relation between algorithmic operators. □

Equation (2.10) implies that the sequence $(\min_{i=1\dots k} \|\tilde{z}^i\|_P^2)_{k\in\mathbb{N}}$, the cumulative minimum of $(\|\tilde{z}^k\|_P^2)_{k\in\mathbb{N}}$, converges sublinearly. Next, we derive big-$O(1/(k+1))$ and little-$o(1/(k+1))$ convergence rates for the sequence itself. This is established below, under further restrictions on $(\lambda_k)_{k\in\mathbb{N}}$, by showing that the sequence $\left(\|\tilde{z}^k\|_D^2\right)_{k\in\mathbb{N}}$ is monotonically nonincreasing and summable.

Theorem 2.3 (convergence rates)**.** *Consider Algorithm 2.1 under the assumptions of Theorem 2.1. Let c_1 and c_2 be two positive constants satisfying*

$$c_1 P \preceq D \preceq c_2 P, \tag{2.14}$$

with D defined in (2.3). Assume

$$(\lambda_k)_{k\in\mathbb{N}} \subseteq]0, c_1\delta/c_2], \tag{2.15}$$

where δ is defined in (2.5). Suppose that $\tau := \inf_{k\in\mathbb{N}} \lambda_k(\delta - \lambda_k) > 0$. Then, the following convergence estimates hold:

$$\|\tilde{z}^k\|_D^2 \leq \frac{c_2^2}{\tau(k+1)}\|z^0 - z^\star\|_S^2 \quad \textit{and} \quad \|\tilde{z}^k\|_D^2 = o(1/(k+1)).$$

Proof. Using the monotonicity of A and Step 1 of Algorithm 2.1

$$0 \leq \langle (H-M)(z^k - z^{k+1}) - H(\bar{z}^k - \bar{z}^{k+1}) + Cz^{k+1} - Cz^k, \bar{z}^k - \bar{z}^{k+1}\rangle. \tag{2.16}$$

On the other hand we have

$$\begin{aligned}
\langle Cz^{k+1} - Cz^k, \bar{z}^k - \bar{z}^{k+1}\rangle &= \langle Cz^{k+1} - Cz^k, \tilde{z}^k - \tilde{z}^{k+1}\rangle \\
&\quad + \langle Cz^{k+1} - Cz^k, z^k - z^{k+1}\rangle \\
&\leq \tfrac{\epsilon}{2}\|\tilde{z}^k - \tilde{z}^{k+1}\|_P^2 + \tfrac{1}{2\epsilon}\|Cz^{k+1} - Cz^k\|_{P^{-1}}^2 \\
&\quad + \langle Cz^{k+1} - Cz^k, z^k - z^{k+1}\rangle \\
&\leq \tfrac{\epsilon}{2}\|\tilde{z}^k - \tilde{z}^{k+1}\|_P^2
\end{aligned}$$

$$+ \left(1 - \tfrac{1}{2\epsilon\beta}\right)\langle Cz^{k+1} - Cz^k, z^k - z^{k+1}\rangle.$$

The first inequality follows from the Fenchel-Young inequality for $\frac{\epsilon}{2}\|\cdot\|_P^2$, and the second inequality follows from β-cocoercivity of C with respect to $\|\cdot\|_P$. Set $\epsilon = \frac{1}{2\beta}$ so that

$$\langle Cz^{k+1} - Cz^k, \bar{z}^k - \bar{z}^{k+1}\rangle \leq \tfrac{1}{4\beta}\|\tilde{z}^k - \tilde{z}^{k+1}\|_P^2. \tag{2.17}$$

Using (2.16), (2.17) and monotonicity of M we have

$$\begin{aligned}
0 &\leq \tfrac{1}{4\beta}\|\tilde{z}^k - \tilde{z}^{k+1}\|_P^2 + \langle -M(z^k - z^{k+1}) - H(\tilde{z}^k - \tilde{z}^{k+1}), \bar{z}^k - \bar{z}^{k+1}\rangle \\
&= \tfrac{1}{4\beta}\|\tilde{z}^k - \tilde{z}^{k+1}\|_P^2 + \langle -M(z^k - z^{k+1}) - H(\tilde{z}^k - \tilde{z}^{k+1}), \tilde{z}^k - \tilde{z}^{k+1}\rangle \\
&+ \langle -M(z^k - z^{k+1}) - H(\tilde{z}^k - \tilde{z}^{k+1}), z^k - z^{k+1}\rangle \\
&\leq \tfrac{1}{4\beta}\|\tilde{z}^k - \tilde{z}^{k+1}\|_P^2 + \langle -M(z^k - z^{k+1}) - H(\tilde{z}^k - \tilde{z}^{k+1}), \tilde{z}^k - \tilde{z}^{k+1}\rangle \\
&+ \langle -H(\tilde{z}^k - \tilde{z}^{k+1}), z^k - z^{k+1}\rangle \qquad (2.18) \\
&= -\left(1 - \tfrac{1}{4\beta}\right)\|\tilde{z}^k - \tilde{z}^{k+1}\|_P^2 - \langle (M + H^*)(z^k - z^{k+1}), \tilde{z}^k - \tilde{z}^{k+1}\rangle. \qquad (2.19)
\end{aligned}$$

It follows from (2.19) and Step 4 of Algorithm 2.1 that

$$\begin{aligned}
\left(1 - \tfrac{1}{4\beta}\right)\|\tilde{z}^k - \tilde{z}^{k+1}\|_P^2 &\leq \langle -(M + H^*)(z^k - z^{k+1}), \tilde{z}^k - \tilde{z}^{k+1}\rangle \\
&= \langle \alpha_k (H + M^*)^* S^{-1}(H + M^*)\tilde{z}^k, \tilde{z}^k - \tilde{z}^{k+1}\rangle \\
&\leq \langle \alpha_k D\tilde{z}^k, \tilde{z}^k - \tilde{z}^{k+1}\rangle. \qquad (2.20)
\end{aligned}$$

Let us show that $(\|\tilde{z}^k\|_D^2)_{k\in\mathbb{N}}$ is monotonically nonincreasing. Using the identity

$$\|a\|_D^2 - \|b\|_D^2 = 2\langle Da, a - b\rangle - \|a - b\|_D^2, \tag{2.21}$$

we have

$$\begin{aligned}
\|\tilde{z}^k\|_D^2 - \|\tilde{z}^{k+1}\|_D^2 &= 2\langle D\tilde{z}^k, \tilde{z}^k - \tilde{z}^{k+1}\rangle - \|\tilde{z}^k - \tilde{z}^{k+1}\|_D^2 \\
&\geq \tfrac{2}{\alpha_k}(1 - \tfrac{1}{4\beta})\|\tilde{z}^k - \tilde{z}^{k+1}\|_P^2 - \|\tilde{z}^k - \tilde{z}^{k+1}\|_D^2 \\
&\geq \tfrac{c_1\delta}{\lambda_k}\|\tilde{z}^k - \tilde{z}^{k+1}\|_P^2 - \|\tilde{z}^k - \tilde{z}^{k+1}\|_D^2
\end{aligned}$$

$$\geq \left(\tfrac{c_1\delta}{c_2\lambda_k} - 1\right)\|\tilde{z}^k - \tilde{z}^{k+1}\|_D^2$$

where the inequalities follow from (2.20), the definition of α_k and (2.14). Therefore $\|\tilde{z}^k\|_D^2$ is nonincreasing as long as (2.15) is satisfied. It follows from (2.9) and (2.14) that

$$\begin{aligned}\|z^{k+1} - z^\star\|_S^2 &\leq \|z^k - z^\star\|_S^2 - \lambda_k(\delta - \lambda_k)\|(H + M^*)\tilde{z}^k\|_{S^{-1}}^{-2}\|\tilde{z}^k\|_P^4 \\ &= \|z^k - z^\star\|_S^2 - \lambda_k(\delta - \lambda_k)\|\tilde{z}^k\|_D^{-2}\|\tilde{z}^k\|_P^4 \\ &\leq \|z^k - z^\star\|_S^2 - c_2^{-2}\lambda_k(\delta - \lambda_k)\|\tilde{z}^k\|_D^2.\end{aligned}$$

Summing over k yields $\sum_{i=0}^{\infty}\lambda^i(\delta - \lambda^i)\|\tilde{z}^i\|_D^2 \leq c_2^2\|z^0 - z^\star\|_S^2$. Therefore we have

$$\textstyle\sum_{i=0}^{\infty}\|\tilde{z}^i\|_D^2 \leq \frac{c_2^2}{\tau}\|z^0 - z^\star\|_S^2. \tag{2.22}$$

On the other hand, since $\|\tilde{z}^k\|_D^2$ is nonincreasing, it follows that $\|\tilde{z}^k\|_D^2 \leq \frac{1}{k+1}\sum_{i=0}^{k}\|\tilde{z}^i\|_D^2$. Combining this with (2.22) establishes the big-O convergence. The little-o convergence follows from [57, Lem. 3-(1a)]. □

Recall the notion of *metric subregulariy* in Definition 1.2. In Theorem 2.4, we derive linear convergence rates when the operator $T = A + M + C$ is metrically subregular at all $z^\star \in \operatorname{zer}(T)$ for 0. Metric subregularity is used in [104] to show linear convergence of Krasnosel'skiĭ-Mann iterations for finding a fixed point of a nonexpansive mapping (see Lemma 4.12 and the preceding discussion).

Theorem 2.4 (linear convergence)**.** *Consider Algorithm 2.1 under the assumptions of Theorem 2.1. Suppose that T is metrically subregular at all $z^\star \in \operatorname{zer}(T)$ for 0, cf. (1.5). If either $\mathcal{H}$ is finite-dimensional or $\mathcal{U} = \mathcal{H}$, then $(\operatorname{dist}_S(z^k, \operatorname{zer}(T)))_{k\in\mathbb{N}}$ converges Q-linearly to zero, $(z^k)_{k\in\mathbb{N}}$ and $(\|\tilde{z}^k\|_P)_{k\in\mathbb{N}}$ converge R-linearly to some $z^\star \in \operatorname{zer}(T)$ and zero, respectively.*

Proof. It follows from metric subregularity of T at all $z^\star \in \operatorname{zer}(T)$ for 0 that

$$\operatorname{dist}(x, \operatorname{zer}(T)) \leq \eta\|y\| \quad \forall x \in \mathcal{U} \text{ and } y = Tx \text{ with } \|y\| \leq \nu, \tag{2.23}$$

for some $\nu \in (0, \infty)$ and $\eta \in [0, \infty)$ and a neighborhood $\mathcal{U}$ of $\operatorname{zer}(T)$. Consider w^k defined in (2.11). It was shown in (2.12) that $w^k \to 0$ and if $\mathcal{H}$ is a finite-dimensional Hilbert space, Theorem 2.1(*ii*)-(*iii*) yield that $\bar{z}^k$ converges to a point in $\operatorname{zer}(T)$. Then there exists $\bar{k} \in \mathbb{N}$ such that for $k > \bar{k}$ we have $\|w^k\| \leq \nu$ and a neighborhood $\mathcal{U}$ of $\operatorname{zer}(T)$ exists with $\bar{z}^k \in \mathcal{U}$ (This holds trivially when $\mathcal{U} = \mathcal{H}$). Consequently (2.23) yields $\operatorname{dist}(\bar{z}^k, \operatorname{zer}(T)) \leq \eta\|w^k\|$. In addition,

triangle inequality and Lipschitz continuity of C yield

$$\|w^k\| = \|H\tilde{z}^k - M\tilde{z}^k - C\bar{z}^k + Cz^k\| \leq \|(H-M)\tilde{z}^k\| + \|C\bar{z}^k - Cz^k\|$$
$$\leq \left(\|H-M\| + \tfrac{1}{\beta}\|P\|\right)\|\tilde{z}^k\|.$$

Consider the projection of $\bar{z}^k$ onto $\operatorname{zer}(T)$, $\mathcal{P}_{\operatorname{zer}(T)}(\bar{z}^k)$. By definition $\|\bar{z}^k - \mathcal{P}_{\operatorname{zer}(T)}(\bar{z}^k)\| = \operatorname{dist}(\bar{z}^k, \operatorname{zer}(T))$ (the minimum is attained since T is maximally monotone [13, Proposition 23.39]), and we have

$$\|z^k - \mathcal{P}_{\operatorname{zer}(T)}(\bar{z}^k)\| \leq \|\bar{z}^k - \mathcal{P}_{\operatorname{zer}(T)}(\bar{z}^k)\| + \|\tilde{z}^k\| = \operatorname{dist}(\bar{z}^k, \operatorname{zer}(T)) + \|\tilde{z}^k\|$$
$$\leq \xi\eta\|\tilde{z}^k\| + \|\tilde{z}^k\| \leq (\xi\eta+1)\|P^{-1}\|^{1/2}\|\tilde{z}^k\|_P, \tag{2.24}$$

where $\xi = \|H-M\| + \frac{1}{\beta}\|P\|$. It follows from (2.24) that

$$\operatorname{dist}_S^2(z^k, \operatorname{zer}(T)) \leq \|z^k - \mathcal{P}_{\operatorname{zer}(T)}(\bar{z}^k)\|_S^2 \leq (\xi\eta+1)^2\|P^{-1}\|\|S\|\|\tilde{z}^k\|_P^2. \tag{2.25}$$

By definition we have $\|z^k - \mathcal{P}^S_{\operatorname{zer}(T)}(z^k)\|_S = \operatorname{dist}_S(z^k, \operatorname{zer}(T))$, and since inequality (2.9) holds for all $z^\star \in \operatorname{zer}(T)$, it follows that

$$\operatorname{dist}_S^2(z^{k+1}, \operatorname{zer}(T)) \leq \|z^{k+1} - \mathcal{P}^S_{\operatorname{zer}(T)}(z^k)\|_S^2$$
$$\leq \|z^k - \mathcal{P}^S_{\operatorname{zer}(T)}(z^k)\|_S^2 - \lambda_k(\delta-\lambda_k)\|(H+M^*)\tilde{z}^k\|_{S^{-1}}^{-2}\|\tilde{z}^k\|_P^4$$
$$= \operatorname{dist}_S^2(z^k, \operatorname{zer}(T)) - \lambda_k(\delta-\lambda_k)\|(H+M^*)\tilde{z}^k\|_{S^{-1}}^{-2}\|\tilde{z}^k\|_P^4 \tag{2.26}$$
$$\leq \operatorname{dist}_S^2(z^k, \operatorname{zer}(T)) - \lambda_k(\delta-\lambda_k)\|S^{-1/2}(H+M^*)P^{-1/2}\|^{-2}\|\tilde{z}^k\|_P^2 \tag{2.27}$$
$$\leq \operatorname{dist}_S^2(z^k, \operatorname{zer}(T)) - \tfrac{\lambda_k(\delta-\lambda_k)}{\varsigma}\operatorname{dist}_S^2(z^k, \operatorname{zer}(T)),$$

where $\varsigma = (\xi\eta+1)^2\|P^{-1}\|\|S\|\|S^{-1/2}(H+M^*)P^{-1/2}\|^2$ and in the last inequality we used (2.25). It follows from (2.5) that there exists $\tilde{k} \in \mathbb{N}$ such that $(\lambda_k(\delta-\lambda_k))_{k>\tilde{k}} \subseteq [\bar{\tau}, \infty)$ for some $\bar{\tau} > 0$. Hence, the sequence $(\operatorname{dist}_S(z^k, \operatorname{zer}(T)))_{k\in\mathbb{N}}$ converges Q-linearly to zero. Thus, R-linear convergence of $\left(\|\tilde{z}^k\|_P\right)_{k\in\mathbb{N}}$ follows from (2.27).

Step 4 of Algorithm 2.1 and (2.26) yield

$$\|z^{k+1} - z^k\|_S^2 = \lambda_k^2\|(H+M^*)\tilde{z}^k\|_{S^{-1}}^{-2}\|\tilde{z}^k\|_P^4$$

$$\leq \frac{\delta^2}{\bar{\tau}}\Big(\mathrm{dist}_S^2(z^k, \mathrm{zer}(T)) - \mathrm{dist}_S^2(z^{k+1}, \mathrm{zer}(T))\Big).$$

Therefore, $\big(\|z^{k+1} - z^k\|_S\big)_{k\in\mathbb{N}}$ converges R-linearly to zero. This is equivalent to saying that there exists $c \in (0,1)$, $\kappa \in (0,\infty)$, $\underline{k} \in \mathbb{N}$ such that for all $k \geq \underline{k}$, $\|z^{k+1} - z^k\|_S \leq \kappa c^k$ holds. Thus, for any $j > k \geq \underline{k}$ we have

$$\|z^j - z^k\|_S \leq \textstyle\sum_{i=k}^{j-1}\|z^{i+1} - z^i\|_S \leq \sum_{i=k}^{j-1}\kappa c^i \leq \sum_{i=k}^{\infty}\kappa c^i = \frac{\kappa}{1-c}c^k. \tag{2.28}$$

Hence, the sequence $(z^k)_{k\in\mathbb{N}}$ is a Cauchy sequence, and therefore converges to some $z \in \mathcal{H}$. From uniqueness of weak limit and Theorem 2.1(*iii*) we have $z \in \mathrm{zer}(T)$. Let $j \to \infty$ in (2.28) to obtain R-linear convergence of $(z^k)_{k\in\mathbb{N}}$. □

In the special case when $C \equiv 0$, M is skew-adjoint , $K = M$ and $S = P$, the operator $P \in \mathscr{B}(\mathcal{H})$ can be a self-adjoint, positive operator rather than a strongly positive operator. Under these assumptions AFBA simplifies to the following iteration:

$$\bar{z}^k = (H + A)^{-1}Pz^k \tag{2.29a}$$

$$z^{k+1} = z^k + \lambda_k(\bar{z}^k - z^k). \tag{2.29b}$$

Notice that if P was strongly positive, this could simply be seen as proximal point algorithm in a different metric applied to the operator $A+M$, but we have relaxed this assumption and only require P to be positive. Before providing convergence results for this algorithm we begin with the following lemma, showing that the mapping $(H + A)^{-1}$ has full domain and is continuous when H has a block triangular structure with strongly positive diagonal blocks, even though its symmetric part, P, might not be strongly positive. This lemma motivates the assumption on continuity of $(H + A)^{-1}P$ in Theorem 2.6. As an application of this theorem in Proposition 2.7(*iii*), when P is positive with a two-by-two block structure (see (2.54a) in the limiting case $\theta = 2$), DRS is recovered.

Lemma 2.5. *Let $\mathcal{H} = \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_N$, where $\mathcal{H}_1, \cdots, \mathcal{H}_N$ are real Hilbert spaces. Suppose that A is block separable and H has a conformable lower (upper) triangular partitioning, i.e.,*

$$A : z \mapsto (A_1 z_1, \cdots, A_N z_N),$$

$$H : z \mapsto (H_{11}z_1, H_{21}z_1 + H_{22}z_2, \cdots, \sum_{j=1}^{N} H_{Nj}z_j), \tag{2.30}$$

where $z_i \in \mathcal{H}_i$ *for* $i = 1, \cdots, N$, $z = (z_1, \cdots, z_N) \in \mathcal{H}$, *and* $H_{ij} \in \mathscr{B}(\mathcal{H}_j, \mathcal{H}_i)$ *for* $i, j = 1, \cdots, N$. *For* $i = 1, \cdots, N$, *assume that* A_i *is maximally monotone, and* $H_{ii} \in \mathscr{S}_{\tau_i}(\mathcal{H}_i)$ *with* $\tau_i \in]0, \infty]$. *Then, the mapping* $(H + A)^{-1}$ *is continuous and has full domain, i.e.,* $\operatorname{dom}((H + A)^{-1}) = \mathcal{H}$. *Furthermore, the update* $\bar{z} = (H + A)^{-1} z$ *is carried out using*

$$\bar{z}_i = \begin{cases} (H_{11} + A_1)^{-1} z_1, & i = 1; \\ (H_{ii} + A_i)^{-1} \Big(z_i - \sum_{j=1}^{i-1} H_{ij} \bar{z}_j \Big), & i = 2, \cdots, N, \end{cases} \tag{2.31}$$

where $\bar{z} = (\bar{z}_1, \cdots, \bar{z}_N) \in \mathcal{H}$.

Proof. We consider a block lower triangular H as in (2.30), the analysis for upper triangular case is identical. The goal is to consider A_i's separately. Let $\bar{z} = (H + A)^{-1} z$ with $\bar{z} = (\bar{z}_1, \cdots, \bar{z}_N)$. The block triangular structure of H in (2.30) yields the equivalent inclusion $z_i \in A_i \bar{z}_i + \sum_{j=1}^{i} H_{ij} \bar{z}_j$, for $i = 1, \cdots, N$. This is equivalent to (2.31), in which, each $\bar{z}_i$ is evaluated using z_i and $\bar{z}_j$ for $j < i$. For the first block we have $\bar{z}_1 = (H_{11} + A_1)^{-1} z_1$. Since A_1 is monotone and H_{11} is strongly monotone, it follows that $H_{11} + A_1$ is strongly monotone, which in turn implies that $(H_{11} + A_1)^{-1}$ is cocoercive and, as such, at most single-valued and continuous. Since A_1 is maximally monotone and H_{11} is strongly positive we have

$$\operatorname{dom}\big((H_{11} + A_1)^{-1}\big) = \operatorname{ran}(H_{11} + A_1) = \operatorname{ran}(\mathrm{id} + A_1 H_{11}^{-1}) = \mathcal{H}_1,$$

where the last equality follows from maximal monotonicity of $A_1 H_{11}^{-1}$ in the Hilbert space defined by endowing $\mathcal{H}_1$ with the scalar product $\langle \cdot, \cdot \rangle_{H_{11}^{-1}}$, and Minty's theorem [13, Thm. 21.1]. For the second block in (2.31) we have $\bar{z}_2 = (H_{22} + A_2)^{-1}(z_2 - H_{21} \bar{z}_1)$. Hence, by the same argument used for previous block, $(H_{22} + A_2)^{-1}$ is continuous and has full domain. Since $(z_2 - H_{21} \bar{z}_1)$ and $(H_{22} + A_2)^{-1}$ are continuous, so is their composition $(H_{22} + A_2)^{-1}(z_2 - H_{21} \bar{z}_1)$. Follow the same argument for the remaining blocks in (2.31) to conclude that $(H + A)^{-1}$ is continuous and has full domain. □

The next theorem provides convergence and rate of convergence results for algorithm (2.29a)-(2.29b) in finite-dimensions by employing the same idea used in [53, Thm. 3.3]. The idea is to consider the operator $R = P + \mathrm{id} - Q$, where Q is the orthogonal projection onto $\operatorname{ran}(P)$. The proof presented here is for a general P and it coincides with the one of Condat [53, Thm. 3.3] for a special choice $P \in \mathscr{B}(\mathcal{K}) : (x, y) \mapsto \big(\gamma^{-1} x - y, -x + \gamma y\big)$.

Theorem 2.6. *Suppose that* $\mathcal{H}$ *is finite-dimensional. Let* $P \in \mathscr{S}(\mathcal{H})$, $P \succeq 0$, $M \in \mathscr{B}(\mathcal{H})$ *a skew-adjoint operator and* $H = P + M$. *Consider the iteration*

(2.29a)-(2.29b) *and assume* $\operatorname{zer}(T) \neq \emptyset$ *where* $T = A+M$*. Furthermore, assume that* $(H+A)^{-1}P$ *is continuous. Let* $(\lambda_k)_{k\in\mathbb{N}}$ *be uniformly bounded in the interval* $(0, 2)$*. Then,*

(i) $(z^k)_{k\in\mathbb{N}}$ *converges to a point in* $\operatorname{zer}(T)$*.*

(ii) Let Q *be the orthogonal projection onto* $\operatorname{ran}(P)$*, and* $R = P + \mathrm{id} - Q$*. The following convergence estimates hold:*

$$\|Pz^{k+1} - Pz^k\|^2 \leq \tfrac{\|P\|}{\tau(k+1)}\|Qz^0 - Qz^\star\|_R^2, \tag{2.32}$$

for some constant $\tau > 0$*, and* $\|Pz^{k+1} - Pz^k\|^2 = o(1/(k+1))$*.*

Proof. (*i*): Since P is not strongly positive, it does not define a valid inner product. Consider $R \coloneqq P + \mathrm{id} - Q$, where Q is the orthogonal projection onto $\operatorname{ran}(P)$. We show that by construction R is strongly positive. By the spectral theorem we can write $Pz_1 = U\Lambda U^* z_1$, where U is an orthonormal basis consisting of eigenvectors of A. Consider two sets: $s_1 = \{i|\lambda^i \neq 0\}$ and $s_2 = \{i|\lambda^i = 0\}$. Denote by U_1 the orthonormal basis made up of u_i for $i \in s_1$. Then, $\operatorname{ran}(P) = \operatorname{ran}(U_1)$ and we have $Q = U_1U_1^*$. For any $z \in \mathcal{H}$, $z = z_1 + z_2$ where $z_1 = Qz$ and $z_2 = (\mathrm{id} - Q)z$. Then, $Rz = Pz + z - Qz = Pz_1 + z_2$ and $\langle Rz, z\rangle = \langle Pz_1 + z_2, z\rangle = \langle Pz_1, z_1\rangle + \|z_2\|^2$. If $z_1 = 0$ then $\langle Rz, z\rangle = \|z_2\|^2 = \|z\|^2$. Suppose that $z_2 \neq 0$ and $z_1 \neq 0$. Denote by $\lambda_{\min}$ the smallest non zero eigenvalue of P. We have

$$\begin{aligned}
\langle Rz, z\rangle &= \langle Pz_1, z_1\rangle + \|z_2\|^2 = \langle U\Lambda U^* z_1, z_1\rangle + \|z_2\|^2 \\
&= \langle \Lambda U^* z_1, U^* z_1\rangle + \|z_2\|^2 \geq \lambda_{\min}\|U_1^* z_1\|^2 + \|z_2\|^2 \\
&= \lambda_{\min}\langle z_1, Qz_1\rangle + \|z_2\|^2 = \lambda_{\min}\|z_1\|^2 + \|z_2\|^2 \\
&\geq \min\{1, \lambda_{\min}\}\|z\|^2.
\end{aligned}$$

If $z_2 = 0$ the above analysis holds with $z = z_1$ and result in strong positivity parameter equal to $\lambda_{\min}$.

We continue by noting that by definition we have $Q \circ P = P$, and symmetry of P yields $P \circ Q = P$. Therefore, $R \circ Q = P$ and for $z \in \mathcal{H}$ we have

$$\langle Pz, z\rangle = \langle QPz, z\rangle = \langle Pz, Qz\rangle = \langle RQz, Qz\rangle, \tag{2.33}$$

which will be used throughout this proof. Observe now that for $z^\star \in \operatorname{zer}(T) \neq \emptyset$ we have $-Mz^\star \in Az^\star$. By monotonicity of A and (2.29a) we have $\langle M\bar{z}^k -$

$Mz^\star + P\tilde{z}^k, z^\star - \bar{z}^k\rangle \geq 0$. Then

$$\begin{aligned} 0 &\leq \langle M\tilde{z}^k - Mz^\star + P\tilde{z}^k, z^\star - \bar{z}^k\rangle = \langle P\tilde{z}^k, z^\star - \bar{z}^k\rangle \\ &= \langle P\tilde{z}^k, Qz^\star - Q\bar{z}^k\rangle + \langle RQ\tilde{z}^k, Q\tilde{z}^k\rangle - \langle RQ\tilde{z}^k, Q\tilde{z}^k\rangle \\ &= \langle P\tilde{z}^k, Qz^\star - Qz^k\rangle - \|Q\tilde{z}^k\|_R^2, \end{aligned} \tag{2.34}$$

where the equalities follows from skew-symmetricity of M and (2.33). We show that $\|Qz^k - Qz^\star\|_R$ is decreasing using (2.34):

$$\begin{aligned} \|Qz^{k+1} - Qz^\star\|_R^2 &= \|Qz^k - Qz^\star + \lambda_k Q\tilde{z}^k\|_R^2 \\ &= \|Qz^k - Qz^\star\|_R^2 + 2\lambda_k\langle Qz^k - Qz^\star, P\tilde{z}^k\rangle + \lambda_k^2\|Q\tilde{z}^k\|_R^2 \\ &\leq \|Qz^k - Qz^\star\|_R^2 - \lambda_k(2-\lambda_k)\|Q\tilde{z}^k\|_R^2. \end{aligned} \tag{2.35}$$

Let us define the sequence $x^k = Qz^k$ and $\bar{x}^k = Q\bar{z}^k$ for every $k \in \mathbb{N}$. Then, since $P \circ Q = P$ the iteration for $x^k = Qz^k$ is written as

$$\begin{aligned} \bar{x}^k &= Q(H+A)^{-1}P(x^k) \\ x^{k+1} &= x^k + \lambda_k(\bar{x}^k - x^k). \end{aligned} \tag{2.36}$$

Let $G = (H+A)^{-1}P$ and $G' = Q \circ G$. It follows from $H = P + M$ and (2.29a) that

$$0 \in T(G(z)) + PG(z) - Pz. \tag{2.37}$$

Use (2.37) and monotonicity of T at $z^\star \in \operatorname{zer}(T)$ and $G(z^\star)$ to derive

$$0 \leq \langle G(z^\star) - z^\star, Pz^\star - PG(z^\star)\rangle. \tag{2.38}$$

In view of (2.38) and positivity of P, we have $\langle G(z^\star) - z^\star, PG(z^\star) - Pz^\star\rangle = 0$, and by [13, Cor. 18.18], $PG(z^\star) - Pz^\star = 0$. Hence, since $R \circ Q = P$, we have $RQG(z^\star) - RQz^\star = 0$, and strong positivity of R implies $Qz^\star = QG(z^\star) = QG(Qz^\star)$, where the last equality is due to $G \circ Q = G$. Thus, $Qz^\star$ is a fixed point of $G' = QG$. We showed that if $z^\star \in \operatorname{zer}(T)$ then $Qz^\star \in \operatorname{fix}(G')$, *i.e.*,

$$Q\operatorname{zer}(T) \subseteq \operatorname{fix}(G'). \tag{2.39}$$

Furthermore, for any $x^\star \in \operatorname{fix}(G')$ we have $Px^\star = PG'(x^\star) = PQG(x^\star) = PG(x^\star)$. Combine this with (2.37) to derive $G(x^\star) \in \operatorname{zer}(T)$. Therefore, $x^\star = G'(x^\star) = QG(x^\star) \in Q\operatorname{zer}(T)$. This shows that if $x^\star \in \operatorname{fix}(G')$, then $x^\star \in Q\operatorname{zer}(T)$, *i.e.*, $\operatorname{fix}(G') \subseteq Q\operatorname{zer}(T)$. Combine this with (2.39) to conclude that the two sets $\operatorname{fix}(G')$ and $Q\operatorname{zer}(T)$ are the same. On the other hand, we rewrite (2.35)

for x^k and $\bar{x}^k$:

$$\|x^{k+1} - Qz^\star\|_R^2 \leq \|x^k - Qz^\star\|_R^2 - \lambda_k(2-\lambda_k)\|\bar{x}^k - x^k\|_R^2. \tag{2.40}$$

Therefore, $(x^k)_{k\in\mathbb{N}}$ is R-Fejér monotone with respect to $\operatorname{fix}(G')$. Since $(\lambda_k)_{k\in\mathbb{N}}$ is uniformly bounded in $(0,2)$, it follows that

$$G'(x^k) - x^k = \bar{x}^k - x^k \to 0. \tag{2.41}$$

Let x be a sequential cluster point of $(x^k)_{k\in\mathbb{N}}$, say $x_{k_q} \to x$. G' is continuous since $G' = Q \circ G$ and G is assumed to be continuous. Thus, it follows from (2.41) that $G'(x) - x = 0$, *i.e.*, $x \in \operatorname{fix}(G')$. This together with Fejér monotonicity of x^k with respect to $\operatorname{fix}(G')$ and [13, Thm. 5.5] yields $x^k \to x \in \operatorname{fix}(G')$.

The proof is completed by first using $G \circ Q = G$ and continuity of G to deduce that $\bar{z}^k = G(z^k) = G(x^k)$ converges to $G(x^\star) \in \operatorname{zer}(T)$, and then arguing for convergence of z^k. We skip the details here because they are identical to the last part of the proof in [53, Thm. 3.3]).

(*ii*): Follow the procedure in the proof of Theorem 2.3 to derive (2.18), except that in this case the cocoercive term is absent. This yields

$$0 \leq \langle -M(z^k - z^{k+1}) - H(\tilde{z}^k - \tilde{z}^{k+1}), \tilde{z}^k - \tilde{z}^{k+1}\rangle - \langle H(\tilde{z}^k - \tilde{z}^{k+1}), z^k - z^{k+1}\rangle. \tag{2.42}$$

Since $H = P + M$ and M is skew-symmetric, (2.42) simplifies to

$$\begin{aligned} 0 &\leq \langle -P(\tilde{z}^k - \tilde{z}^{k+1}), \tilde{z}^k - \tilde{z}^{k+1}\rangle + \langle -P(z^k - z^{k+1}), \tilde{z}^k - \tilde{z}^{k+1}\rangle \\ &= -\|Q\tilde{z}^k - Q\tilde{z}^{k+1}\|_R^2 + \lambda_k\langle P\tilde{z}^k, \tilde{z}^k - \tilde{z}^{k+1}\rangle, \end{aligned} \tag{2.43}$$

where we used (2.33) and (2.29b). Using identity (2.21), we derive

$$\begin{aligned} \|Q\tilde{z}^k\|_R^2 - \|Q\tilde{z}^{k+1}\|_R^2 &= 2\langle RQ\tilde{z}^k, Q\tilde{z}^k - Q\tilde{z}^{k+1}\rangle - \|Q\tilde{z}^k - Q\tilde{z}^{k+1}\|_R^2 \\ &= 2\langle P\tilde{z}^k, \tilde{z}^k - \tilde{z}^{k+1}\rangle - \|Q\tilde{z}^k - Q\tilde{z}^{k+1}\|_R^2 \\ &\geq \left(\tfrac{2}{\lambda_k} - 1\right)\|Q\tilde{z}^k - Q\tilde{z}^{k+1}\|_R^2, \end{aligned} \tag{2.44}$$

where we made use of (2.43). Consider (2.35) and sum over k to derive

$$\sum_{i=0}^{\infty} \lambda^i(2-\lambda^i)\|Q\tilde{z}^i\|_R^2 \leq \|Qz^0 - Qz^\star\|_R^2. \tag{2.45}$$

Inequality (2.44) shows that $\|Q\tilde{z}^k\|_R^2$ is monotonically nonincreasing. Combined with (2.45) and uniform boundedness of λ_k, *i.e.*, $(\lambda_k)_{k\in\mathbb{N}} \subseteq [\epsilon, 2-\epsilon]$ for some $\epsilon > 0$,

$$\|Q\tilde{z}^k\|_R^2 \leq \frac{1}{(k+1)\epsilon^2}\|Qz^0 - Qz^\star\|_R^2. \tag{2.46}$$

Furthermore, it follows from (2.36) and definition of $x^k, \bar{x}^k$ that

$$\|x^{k+1} - x^k\|_R^2 = \lambda_k^2\|Q\tilde{z}^k\|_R^2 \leq (2-\epsilon)^2\|Q\tilde{z}^k\|_R^2. \tag{2.47}$$

Combine (2.47) and (2.46) to derive

$$\|x^k - x^{k+1}\|_R^2 \leq \frac{(2-\epsilon)^2}{(k+1)\epsilon^2}\|Qz^0 - Qz^\star\|_R^2. \tag{2.48}$$

This establishes big-O convergence for $(x^k)_{k\in\mathbb{N}}$. The little-o convergence of $\|Q\tilde{z}^k\|_R^2$ and subsequently $\|x^k - x^{k+1}\|_R^2$ follows from (2.44), (2.45) and [57, Lem. 3-(1a)]. We derive from (2.33) that $\|x^k - x^{k+1}\|_R^2 = \langle z^k - z^{k+1}, P(z^k - z^{k+1})\rangle$. Then, it follows from [13, Cor. 18.18] that

$$\|Pz^k - Pz^{k+1}\|^2 \leq \|P\|\|x^k - x^{k+1}\|_R^2. \tag{2.49}$$

Set $\tau = \frac{\epsilon^2}{(2-\epsilon)^2}$, and combine (2.49) with (2.48) to yield big-O convergence for the sequence $(Pz^k)_{k\in\mathbb{N}}$. Similarly, little-o convergence follows from that property of $\|x^k - x^{k+1}\|_R^2$. □

2.3 Operator splitting schemes as special cases

We are ready to consider some important special cases to illustrate the importance of parameters S, P and K. Further discussion on other special choices for the parameters appear in Chapter 3 in the framework of convex optimization with the understanding that it is straightforward to adapt the same analysis for the corresponding monotone inclusion problem.

2.3.1 Forward-backward splitting

When $H = \gamma^{-1}\mathrm{id}$, $S = \mathrm{id}$ and $M \equiv 0$, Algorithm 2.1 reduces to forward-backward splitting (FBS):

$$\bar{z}^k = (\mathrm{id} + \gamma A)^{-1}(\mathrm{id} - \gamma C)z^k$$

$$z^{k+1} = z^k + \lambda_k\big(\bar{z}^k - z^k\big).$$

Let β be the cocoercivity constant of C with respect to the canonical norm $\|\cdot\|$, then β/γ is the cocoercivity constant with respect to the P norm and condition (2.5) of Theorem 2.1 becomes

$$(\lambda_k)_{k\in\mathbb{N}} \subseteq [0,\delta] \quad \text{with} \quad \delta = 2 - \frac{\gamma}{2\beta},\ \gamma \in (0,4\beta),\ \liminf_{k\to\infty} \lambda_k(\delta - \lambda_k) > 0. \tag{2.50}$$

This allows a wider range of parameters than the standard ones found in the literature. The standard convergence results for FBS are based on the theory of averaged operators (see [52] and the references therein) and yield the same conditions as in (2.50) but with $\gamma \in (0, 2\beta]$ (see also [53, Lem. 4.4] and [13, Thm. 26.14]). Additionally, if $C \equiv 0$, FBS reduces to the classical PPA.

The convergence rate for FBS follows directly from Theorem 2.3. Since $D = \gamma^{-2}\mathrm{id}$ and $P = \gamma^{-1}\mathrm{id}$, (2.14) holds with $c_1 = c_2 = \gamma^{-1}$. Consequently, if $(\lambda_k(\delta - \lambda_k))_{k\in\mathbb{N}} \subseteq [\tau, \infty)$ for some $\tau > 0$, we have

$$\|\tilde{z}^k\|^2 \leq \frac{1}{\tau(k+1)}\|z^0 - z^\star\|^2,$$

and $\|\tilde{z}^k\|^2 = o(1/(k+1))$.

2.3.2 Solodov and Tseng

In Algorithm 2.1, set $C \equiv 0$, $H = \mathrm{id}$ and $A = N_X$ where X is a nonempty closed convex set in $\mathcal{H}$. Then, Algorithm 2.1 reduces to

$$\tilde{z}^k = \mathcal{P}_X(z^k - Mz^k) - z$$

$$z^{k+1} = z^k + \alpha_k S^{-1}(\mathrm{id} + M^*)\tilde{z}^k, \quad \alpha_k = {}^{\lambda_k\|\tilde{z}^k\|^2}\!/_{\|(\mathrm{id}+M^*)\tilde{z}^k\|^2_{S^{-1}}},$$

recovering the scheme proposed by Solodov and Tseng [151, Alg. 2.1].

2.3.3 Forward-backward-forward splitting

Consider Algorithm 2.1 when M is skew-adjoint and set $H = \gamma^{-1}\mathrm{id}$, $S = \mathrm{id}$. We can enforce $\alpha_k = \gamma$ by choosing $\lambda_k = (\gamma\|(\gamma^{-1}\mathrm{id} + M^*)\tilde{z}^k\|/\|\tilde{z}^k\|)^2$. It remains to show that the sequence $(\lambda_k)_{k\in\mathbb{N}}$ satisfies the conditions of Theorem 2.1. Since M is skew-adjoint, we have $\lambda_k = 1 + (\gamma\|M\tilde{z}^k\|/\|\tilde{z}^k\|)^2$, and if the stepsize satisfies $\gamma \in (0, \|M\|^{-1}\sqrt{1 - 1/(2\beta)})$, then $(\lambda_k)_{k\in\mathbb{N}}$ is uniformly bounded between 0 and

δ (in fact it is larger than 1) and thus satisfies (2.5). Under these assumptions Algorithm 2.1 simplifies to

$$\bar{z}^k = (\mathrm{id} + \gamma A)^{-1}(\mathrm{id} - \gamma M - \gamma C)z^k$$

$$z^{k+1} = \bar{z}^k - \gamma M\big(\bar{z}^k - z^k\big).$$

This algorithm resembles the FBFS [162]. Indeed, if $C \equiv 0$, then the range for the stepsize simplifies to $\gamma \in (0, \|M\|^{-1})$ and yields the FBFS when its Lipschitz operator is the skew-adjoint operator M.

2.3.4 Douglas-Rachford type with a forward term

We now focus our attention on a choice for P, K and S that lead to a new Douglas-Rachford type splitting with a forward term. Consider the problem of finding $x \in \mathcal{H}$ such that

$$0 \in Dx + Ex + Fx, \tag{2.51}$$

together with the dual inclusion problem of finding $y \in \mathcal{H}$ such that there exists $x \in \mathcal{H}$,

$$\begin{cases} 0 \in Dx + Fx + y \\ 0 \in E^{-1}y - x. \end{cases} \tag{2.52}$$

where $D : \mathcal{H} \rightrightarrows \mathcal{H}$, $E : \mathcal{H} \rightrightarrows \mathcal{H}$ are maximally monotone and $F : \mathcal{H} \to \mathcal{H}$ is η-cocoercive with respect to the canonical norm. Let $\mathcal{K}$ be the Hilbert direct sum $\mathcal{K} = \mathcal{H} \oplus \mathcal{H}$. The pair $(x^\star, y^\star) \in \mathcal{K}$ is called a primal-dual solution to (2.51) if it satisfies (2.52). Let $(x^\star, y^\star) \in \mathcal{K}$ be a primal-dual solution, then $x^\star$ solves the primal problem (2.51) and $y^\star$ the dual (2.52). In this section, we assume that there exists $x^\star$ such that $x^\star \in \operatorname{zer}(D + E + F)$. This assumption yields that the set of primal-dual solutions is nonempty (see [50, 26] and the references therein for more discussion).
Reformulate (2.52) in the form of (2.2) by defining

$$A : \mathcal{K} \rightrightarrows \mathcal{K} : (x, y) \mapsto (Dx, E^{-1}y), \tag{2.53a}$$

$$M \in \mathscr{B}(\mathcal{K}) : (x, y) \mapsto (y, -x), \tag{2.53b}$$

$$C : \mathcal{K} \to \mathcal{K} : (x, y) \mapsto (Fx, 0). \tag{2.53c}$$

The operators A and M are maximally monotone [13, Prop. 20.23 and Ex. 20.35]. It is easy to verify, by definition of cocoercivity, that C is cocoercive. Let $\gamma > 0$, $\theta \in [0, 2]$, (the case of $\theta = 2$ can only be considered in the absence of the

cocoercive term and results in classic DR, see Proposition 2.7). Set

$$P \in \mathscr{B}(\mathcal{K}) : (x, y) \mapsto \big(\gamma^{-1}x - \tfrac{1}{2}\theta y, -\tfrac{1}{2}\theta x + \gamma y\big), \tag{2.54a}$$

$$K \in \mathscr{B}(\mathcal{K}) : (x, y) \mapsto \big(\tfrac{1}{2}\theta y, -\tfrac{1}{2}\theta x\big), \tag{2.54b}$$

$$S \in \mathscr{B}(\mathcal{K}) : (x, y) \mapsto \big((3-\theta)\gamma^{-1}x - y, -x + \gamma y\big). \tag{2.54c}$$

The operators S and P are strongly positive for $\theta \in [0, 2)$ [96, Lem. 5.1, 5.3]. The operator $H = P + K$ is given by

$$H \in \mathscr{B}(\mathcal{K}) : (x, y) \mapsto (\gamma^{-1}x, -\theta x + \gamma y). \tag{2.55}$$

Notice that H has the block triangular structure described in Lemma 2.5. By using this structure as in (2.31) and substituting (2.54), (2.55) in Algorithm 2.1, after some algebraic manipulations involving Moreau's identity as well as a change of variables $s^k := x^k - \gamma y^k$ (see proof of Proposition 2.7 for details), we derive the following algorithm:

Algorithm 2.2 Douglas-Rachford Type with a Forward Term

Inputs: $x_0 \in \mathcal{H}$, $s_0 \in \mathcal{H}$

for $k = 0, 1, \ldots$ **do**

$$\bar{x}^k = \mathrm{J}_{\gamma D}(s^k - \gamma F x^k)$$
$$r^k = \mathrm{J}_{\gamma E}(\theta \bar{x}^k + (2-\theta)x^k - s^k)$$
$$s^{k+1} = s^k + \rho_k(r^k - \bar{x}^k)$$
$$x^{k+1} = x^k + \rho_k(\bar{x}^k - x^k)$$

In the special case when $\rho_k = 1$, the last line in Algorithm 2.2 becomes obsolete and $\bar{x}^k$ can be replaced with x^{k+1}. The next proposition provides the convergence properties for Algorithm 2.2.

Proposition 2.7. *Consider the sequences $(x^k)_{k\in\mathbb{N}}$ and $(s^k)_{k\in\mathbb{N}}$ generated by Algorithm 2.2. Let $\eta \in (0, +\infty)$ be the cocoercivity constant of F. Suppose that one of the following holds:*

(i) $\theta \in [0, 2)$, $\gamma \in (0, \eta(4-\theta^2))$ and the sequence of relaxation parameters $(\rho_k)_{k\in\mathbb{N}}$ is uniformly bounded in the interval

$$(\rho_k)_{k\in\mathbb{N}} \subseteq \left(0, \frac{4 - \theta^2 - \gamma/\eta}{(2-\theta)(2+\sqrt{2-\theta})}\right). \tag{2.56}$$

(ii) $F \equiv 0$, $\theta \in [0, 2)$, $\gamma \in (0, \infty)$, *and sequence* $(\rho_k)_{k\in\mathbb{N}}$ *uniformly bounded in the interval*

$$(\rho_k)_{k\in\mathbb{N}} \subseteq \Big(0, 2 - \sqrt{2-\theta}\Big).$$

(iii) $F \equiv 0$, $\theta = 2$, $\gamma \in (0, \infty)$, $(\rho_k)_{k\in\mathbb{N}}$ *uniformly bounded in the interval* $(0, 2)$, *and* $\mathcal{K}$ *is finite-dimensional.*

Then, there exists a pair of solutions $(x^\star, y^\star) \in \mathcal{K}$ *to* (2.52) *such that the sequences* $(x^k)_{k\in\mathbb{N}}$ *and* $(s^k)_{k\in\mathbb{N}}$ *converge weakly to* $x^\star$ *and* $x^\star - \gamma y^\star$, *respectively.*

Proof. (*i*): Let us start by noting that P and S defined in (2.54) are strongly positive if $\theta \in [0, 2)$ [96, Lem. 5.1, 5.3].

Next, consider step 3 of Algorithm 2.1 and Substitute the parameters defined in (2.54), to derive

$$\frac{\alpha_k}{\lambda_k} = \frac{\|\tilde{z}^k\|_P^2}{\|\tilde{z}^k\|_D^2} = \frac{\gamma^{-1}\|\tilde{x}^k\|^2 + \gamma\|\tilde{y}^k\|^2 - \theta\langle\tilde{x}^k, \tilde{y}^k\rangle}{\gamma^{-1}(\theta^2 - 3\theta + 3)\|\tilde{x}^k\|^2 + \gamma\|\tilde{y}^k\|^2 + 2(1-\theta)\langle\tilde{x}^k, \tilde{y}^k\rangle}, \tag{2.57}$$

where D is defined in (2.3) and by construction is strongly positive. Let λ_k be equal to the inverse of the right hand side in (2.57) multiplied by ρ_k where $(\rho_k)_{k\in\mathbb{N}} \subseteq (0, 2)$. This would result in $\alpha_k = \rho_k$ and simplify the iterations. Consider C defined in (2.53c), we have

$$\|Cz - Cz'\|_{P^{-1}}^2 = \gamma\big(1 - \tfrac{1}{4}\theta^2\big)^{-1}\|Cz - Cz'\|^2.$$

From here it is easy to see that C is β-cocoercive with respect to P norm, where $\beta = \eta\gamma^{-1}(1 - \frac{1}{4}\theta^2)$. In order to apply Theorem 2.1, condition (2.5) must hold. From strong positivity of D, boundedness of P and the fact that ρ_k is uniformly bounded above 0 it follows that $(\lambda_k)_{k\in\mathbb{N}} \subseteq (\nu_1, \infty)$ for some positive ν_1. Let ν_2 be a positive parameter such that

$$\nu_2 < 2 - \frac{1}{2\beta} - \frac{2\rho_k(2 + \sqrt{2-\theta})}{2+\theta}, \tag{2.58}$$

holds for all $k \in \mathbb{N}$. It's easy to verify that such ν_2 exists as long as ρ_k is uniformly bounded in the interval (2.56). For brevity, we define β' such that $\frac{1}{2\beta'} = \frac{1}{2\beta} + \nu_2$. Additionally, introduce the notation $\upsilon = (2 - \frac{1}{2\beta'})/\rho_k$, $\omega_1 = \upsilon\theta + 2(1-\theta)$, $\omega_2 = \theta^2 - 3\theta + 3$. We proceed by showing that λ_k is smaller than $2 - \frac{1}{2\beta} - \nu_2$. A sufficient condition for this to hold is

$$\xi = \gamma^{-1}(\upsilon - \omega_2)\|\tilde{x}^k\|^2 + \gamma(\upsilon - 1)\|\tilde{y}^k\|^2 - \omega_1\langle\tilde{x}^k, \tilde{y}^k\rangle > 0. \tag{2.59}$$

Apply the Fenchel-Young inequality for $\frac{\epsilon}{2}\|\cdot\|^2$ to lower bound ξ:

$$\xi \geq \left(\gamma^{-1}(\upsilon-\omega_2)-|\omega_1|\tfrac{\epsilon}{2}\right)\|\tilde{x}^k\|^2+\left(\gamma(\upsilon-1)-|\omega_1|\tfrac{1}{2\epsilon}\right)\|\tilde{y}^k\|^2, \tag{2.60}$$

where $|\cdot|$ is the absolute value. It follows from (2.58) that $\upsilon > 1$ for all $k \in \mathbb{N}$. Let $\epsilon = \frac{|\omega_1|}{2\gamma(\upsilon-1)}$ so that the term involving $\|\tilde{y}^k\|^2$ in (2.60) disappears. We obtain

$$\gamma^{-1}\left(\upsilon-\omega_2-\tfrac{\omega_1^2}{4(\upsilon-1)}\right)\|\tilde{x}^k\|^2>0, \tag{2.61}$$

which is sufficient for (2.59) to hold. By substituting ω_1 and ω_2 and after some algebraic manipulations we find that the condition (2.58) is sufficient for the right hand side in (2.61) to be positive. Consequently, Theorem 2.1 completes the proof of convergence. We showed that we can set $\alpha_k = \rho_k$ by choosing λ_k appropriately. Algorithm 2.2 follows by setting $\alpha_k = \rho_k$, a change of variables $s^k = x^k - \gamma y^k$, substituting x^{k+1} and application of Moreau's identity.

(*ii*): Mimic the proof of (*i*), but use Theorem 2.1 with $C \equiv 0$, by showing that λ_k is uniformly bounded between 0 and 2.

(*iii*): When $\theta = 2$, we have $P \in \mathscr{S}(\mathcal{K})$, $P \succeq 0$. It follows from (2.53a), (2.55) and Lemma 2.5 that $(H+A)^{-1}P$ is continuous. Therefore, since $F \equiv 0$, by appealing to Theorem 2.6 and following the same change of variables as in previous parts the assertion is proved. □

The next proposition provides convergence rate results for Algorithm 2.2 when $\theta = 2$, based on Theorem 2.6. Similarly, for the case when $\theta \in [0, 2)$, convergence rates can be deduced based on Theorem 2.3. Furthermore, when metric subregulariy assumption in Theorem 2.4 holds, linear convergence follows without any additional assumptions.

Proposition 2.8 (convergence rate)**.** *Let $\mathcal{K}$ be finite-dimensional. Consider the sequences $(x^k)_{k\in\mathbb{N}}$ and $(s^k)_{k\in\mathbb{N}}$ generated by Algorithm 2.2. Let $F \equiv 0$, $\theta = 2$, $\gamma \in (0,\infty)$, and $(\rho_k)_{k\in\mathbb{N}}$ be uniformly bounded in the interval $(0, 2)$. Then*

$$\|s^{k+1}-s^k\|^2 \leq \tfrac{\gamma}{\tau(k+1)}\|Qz^0-Qz^\star\|_R^2,$$

and $\|s^{k+1}-s^k\|^2 = o(1/(k+1))$ for some constant $\tau > 0$, where Q is the orthogonal projection onto $\operatorname{ran}(P)$, *$R = P+\mathrm{id}-Q$, and $z^k = \left(x^k, \gamma^{-1}(x^k-s^k)\right)$.*

Proof. Following the argument in proof of Proposition 2.7(*iii*) and Theorem 2.6(*ii*) yields

$$\|Pz^{k+1}-Pz^k\|^2 \leq \tfrac{\|P\|}{\tau(k+1)}\|Qz^0-Qz^\star\|_R^2, \tag{2.62}$$

and $\|Pz^{k+1} - Pz^k\|^2 = o(1/(k+1))$, where $z^k = \big(x^k, \gamma^{-1}(x^k - s^k)\big)$. Combine this with definition of P, (2.54a) with $\theta = 2$, to derive

$$\|Pz^{k+1} - Pz^k\|^2 = (1+\gamma^{-2})\|s^{k+1} - s^k\|^2. \tag{2.63}$$

Furthermore, simple calculation shows that $P^2 = (\gamma + \gamma^{-1})P$. Hence for all $z \in \mathcal{K}$

$$\begin{aligned}\|Pz\|^2 &= (\gamma+\gamma^{-1})\langle z, Pz\rangle \\ &= (\gamma+\gamma^{-1})\big(\gamma^{-1}\|x\|^2 + \gamma\|y\|^2 - 2\langle x, y\rangle\big) \\ &\leq (\gamma+\gamma^{-1})^2\|z\|^2,\end{aligned} \tag{2.64}$$

where we used Fenchel-Young inequality with $\epsilon = \gamma$. It follows from (2.64) that $\|P\| \leq (\gamma + \gamma^{-1})$. Combining this with (2.62) and (2.63) completes the proof. □

Remark 2.9. Recently, another three operator splitting algorithm was proposed in [58] which can also be seen as a generalization of Douglas-Rachford method and allows a third cocoercive operator. In the aforementioned paper, the forward step takes place after the first backward update, while in Algorithm 2.2 it precedes the backward update. The parameter range prescribed in [58, Thm. 3.1] is simply $\gamma \in (0, 2\eta)$ and $(\rho_k)_{k\in\mathbb{N}} \subseteq (0, 2 - \frac{\gamma}{2\eta})$, while for Algorithm 2.2 it consists of $\gamma \in (0, \eta(4-\theta^2))$ with relaxation parameter uniformly bounded in the interval (2.56). For $\theta \in [0, \sqrt{2})$, Algorithm 2.2 can have larger stepsize but it is important to notice that this might not necessarily be advantageous in practice because the upper bound for the relaxation parameter in (2.56) decreases as we reduce θ. For example if we fix $\rho_k = 1$, conditions of Proposition 2.7 become $\theta \in (1, 2)$ and $\gamma/\eta \in (0, (2-\theta)(\theta - \sqrt{2-\theta}))$. This stepsize is always smaller that the one of [58]. However, if the relaxation parameter ρ_k is selected to be small enough then γ can take values larger than the one allowed in [58]. In Section 3.5 numerical simulations are performed for the two algorithms which indicate that on dual support vector machine problem Algorithm 2.2 is slower presumably due to smaller stepsize.

□

Remark 2.10. In Algorithm 2.2 the case $\theta = 2$, $\rho_k = 1$ with $F \equiv 0$ (see Proposition 2.7(*iii*)) yields the classical DRS [108]. This choice of P is precisely the one considered in [53, §3.1.1]. □

ADMM form

Consider the following problem

$$\underset{x_1,x_2,x_3}{\text{minimize}} \quad f_1(x_1) + f_2(x_2) + f_3(x_3) \tag{2.65a}$$

$$\text{subject to} \quad L_1x_1 + L_2x_2 + L_3x_3 = b, \tag{2.65b}$$

where $f_i \in \Gamma_0(\mathcal{H}_i)$, $L_i \in \mathscr{B}(\mathcal{H}_i, \mathcal{H}_4)$, $i = 1, 2, 3$, and $b \in \mathcal{H}_4$ with $\mathcal{H}_1, \ldots, \mathcal{H}_4$ denoting real Hilbert spaces. Additionally, f_1 is ξ-strongly convex for some $\xi \in]0, +\infty[$. It is well known that the classic Alternating Direction Method of Multipliers (ADMM) is equivalent to Douglas-Rachford algorithm applied to the dual problem (see [29] and the references therein). We derive a new 3-block ADMM iteration in a similar way. Consider the dual problem

$$\underset{y}{\text{minimize}} \quad d_1(y) + d_2(y) + d_3(y), \tag{2.66}$$

where $d_1(y) = f_1^*(-L_1^*y)$, $d_2(y) = f_2^*(-L_2^*y)$ and $d_3(y) = f_3^*(-L_3^*y) + \langle y, b\rangle$. By strong convexity of f_1 we have that d_1 has a $\xi^{-1}\|L_1\|^2$-Lipschitz gradient. Form the augmented Lagrangian

$$\mathcal{L}_\gamma(x_1, x_2, x_3, y) = \sum_{i=1}^{3} f_i(x_i) + \left\langle y, \sum_{i=1}^{3} L_ix_i - b \right\rangle + \frac{\gamma}{2}\left\| \sum_{i=1}^{3} L_ix_i - b \right\|^2.$$

We apply Algorithm 2.2 with $\rho_k = 1$ to the monotone inclusion associated with (2.66) and after a change of order and some algebraic manipulations we derive Algorithm 2.3 (the procedure is similar to the one found in [66, §3.5.6] for the classic ADMM). Our 3-block ADMM can be written as

Algorithm 2.3 3-block ADMM

Inputs: $(x_1^0, x_2^0, x_3^0) \in \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{H}_3$, $(y^0, y^1) \in \mathcal{H}_4 \times \mathcal{H}_4$

for $k = 0, 1, \ldots$ **do**

$\bar{y}^k = (\theta - 1)y^k + (2 - \theta)y^{k-1}$

$x_1^{k+1} = \arg\min_{x_1} \mathcal{L}_0(x_1, x_2^k, x_3^k, y^k)$

$x_2^{k+1} = \arg\min_{x_2} \mathcal{L}_\gamma(x_1^k, x_2, x_3^k, \bar{y}^k)$

$x_3^{k+1} = \arg\min_{x_3} \mathcal{L}_\gamma(x_1^{k+1}, x_2^{k+1}, x_3, \bar{y}^k)$

$y^{k+1} = \bar{y}^k + \gamma(L_1x_1^{k+1} + L_2x_2^{k+1} + L_3x_3^{k+1} - b)$

Proposition 2.11. *Let $\mathcal{H}_1, \ldots, \mathcal{H}_4$ be finite-dimensional, $f_i \in \Gamma_0(\mathcal{H}_i)$, $b \in \mathcal{H}_4$,*

$L_i \in \mathscr{B}(\mathcal{H}_i, \mathcal{H}_4)$ for $i = 1, 2, 3$ and $\ker(L_2) = \{0\}$, $\ker(L_3) = \{0\}$. Let $\xi \in]0, +\infty[$ be the strong convexity constant of f_1. Assume that the set of saddle points of (2.65), denoted by Σ, is nonempty. Let $\theta \in]1, 2[$ and

$$\gamma < \xi(2-\theta)(\theta - \sqrt{2-\theta})/\|L_1\|^2. \tag{2.67}$$

Then the sequence $(x_1^k, x_2^k, x_3^k, y^k)_{k\in\mathbb{N}}$ generated by Algorithm 2.3, converges to some $(x_1^\star, x_2^\star, x_3^\star, y^\star) \in \Sigma$.

Proof. Let $(x_1^\star, x_2^\star, x_3^\star, y^\star)$ denote a KKT point of (2.65), *i.e.*,

$$\begin{cases} 0 \in \partial f_i(x_i^\star) + L_i^* y^\star, & \text{for } i = 1, 2, 3 \\ 0 = b - L_1 x_1^\star - L_2 x_2^\star - L_3 x_3^\star. \end{cases} \tag{2.68}$$

Algorithm 2.3 is an implementation of Algorithm 2.2 for solving (2.66). Hence, Proposition 2.7 yields $y^k \to y$, where y is a solution to (2.66). Let x_1 be a point satisfying $-L_1^* y \in \partial f_1(x_1)$. From strong convexity of f_1 at x_1 and x_1^{k+1}, we have

$$(\forall u \in \partial f_1(x_1^{k+1}))(\forall v \in \partial f_1(x_1)) \quad \xi\|x_1^{k+1} - x_1\|^2 \leq \langle u - v, x_1^{k+1} - x_1\rangle.$$

It follows from the optimality condition for the x_1^{k+1} update and the definition of x_1 that

$$\xi\|x_1^{k+1} - x_1\|^2 \leq \langle -L_1^* y^k + L_1^* y, x_1^{k+1} - x_1\rangle \leq \|L_1\|\|x_1^{k+1} - x_1\|\|y^k - y\|.$$

Combine this with the convergence of $(y^k)_{k\in\mathbb{N}}$ to derive $x_1^k \to x_1$. From the change of variables to derive Algorithm 2.3 we have

$$s^k - y^k = -\gamma L_1 x_1^k - \gamma L_3 x_3^k, \tag{2.69}$$

which together with the convergence of $(s^k)_{k\in\mathbb{N}}, (y^k)_{k\in\mathbb{N}}$ (see Proposition 2.7) and $(x_{1,n})_{k\in\mathbb{N}}$ imply that $(L_3 x_3^k)_{k\in\mathbb{N}}$ converges to a point. Since $\ker(L_3) = \{0\}$, it follows that $(x_{3,n})_{k\in\mathbb{N}}$ converges to some x_3. From the optimality condition for the x_3^{k+1} update and the last step in Algorithm 2.3, we have

$$-L_3^* y^{k+1} = -L_3^* \bar{y}^k - \gamma L_3^*(L_1 x_1^{k+1} + L_2 x_2^{k+1} + L_3 x_3^{k+1} - b) \in \partial f_3(x_3^{k+1}).$$

Taking the limit and using [13, Prop. 20.38(iii)], we have $-L_3^* y \in \partial f_3(x_3)$. On the other hand, Theorem 2.1(*ii*) and the last line of Algorithm 2.3 yield

$$y^{k+1} - y^k \to 0, \quad \text{and} \quad L_1 x_1^k + L_2 x_2^k + L_3 x_3^k \to b. \tag{2.70}$$

It follows from (2.69), (2.70), and the convergence of $(s^k)_{k\in\mathbb{N}}, (y^k)_{k\in\mathbb{N}}$ that

$(L_2 x_2^k)_{k\in\mathbb{N}}$ converges to a point. We can now argue almost exactly as we did for $(L_3 x_3^k)_{k\in\mathbb{N}}$. Since $\ker(L_2) = \{0\}$, we deduce that $(x_{2,n})_{k\in\mathbb{N}}$ converges to some x_2. Combine the optimality condition for the x_2^{k+1} update and the last step in Algorithm 2.3 with the convergence of y^k, to derive $-L_2^* y \in \partial f_2(x_2)$. Altogether, we showed that the limit points (x_1, x_2, x_3, y), are jointly optimal by the KKT condition (2.68). □

Remark 2.12. The convergence rate of Algorithm 2.3 can be deduced similar to Algorithm 2.2 from Theorems 2.3, 2.4 and 2.6 with $\rho_k = 1$. However, we do not consider it here. □

Remark 2.13. In the case when $f_1 \equiv 0$ we can choose the limiting value $\theta = 2$ and recover the classical ADMM (see Proposition 2.7(*iii*)). On the other hand if f_2 or f_3 vanish, the Alternating Minimization Method (AMM) [160] is recovered. Finally, when both f_2 and f_3 vanish then the dual ascent method is recovered. □

Remark 2.14. In [58, Alg. 8], another 3-block ADMM formulation is presented by following similar algebraic manipulations (It is derived by applying their Algorithm 7 to the dual). It should be noted that they do not require rank assumptions on L_2, L_3. In contrast to that work in our version $(x_1^k)_{k\in\mathbb{N}}$ and $(x_2^k)_{k\in\mathbb{N}}$ are updated in parallel which corresponds to the fact that in Algorithm 2.2 the forward step precedes the first prox step. Furthermore, in our algorithm, $(x_2^k)_{k\in\mathbb{N}}$ and $(x_3^k)_{k\in\mathbb{N}}$ are updated using the augmented Lagrangian at $(\theta-1)y^k + (2-\theta)y^{k-1}$ rather than y^k. Moreover, the stepsize in [58, Thm. 2.1] has to satisfy $\gamma < 2\xi/\|L_1\|^2$. This is always larger than the stepsize in (2.67). Refer to Remark 2.9 for further discussion, noting that Algorithm 2.3 is derived by setting the relaxation parameter, ρ_k, equal to one. □

Remark 2.15. Some of the other recent attempts to directly generalize ADMM for 3 blocks include [33, 42, 81, 103, 107]. In [42], it was shown through a counterexample that a direct extension of ADMM to more than 2 blocks is not convergent in general. In order to ensure convergence, additional assumptions on strong convexity of the functions or rank of L_i's are needed. In [33] the authors require one function to be strongly convex and L_2 and L_3 to have full column rank, while [103] modify the steps with regularization terms and [107] solves a perturbed problem (see [107] and the references therein for further discussion). In contrast to these papers, the first minimization step of Algorithm 2.3 consists of minimizing a normal Lagrangian rather than an augmented one (therefore it can be trivially executed in a distributed fashion in the case where f_1 is block-separable) and it can be performed in parallel to the second step. □

2.4 Conclusions

In this chapter the operator splitting technique asymmetric-forward-backward-adjoint splitting (AFBA) was introduced for solving monotone inclusions involving three terms. We discussed how it relates to, unifies and extends classical splitting methods. Asymmetric preconditioning is the main feature of AFBA that can lead to several extensions and new algorithmic schemes. We make extensive use of the results of this chapter in the next five chapters where we study primal-dual proximal algorithms for distributed applications.

We conclude by noting some recent developments. In Section 2.3.3 it was shown that a special case of AFBA coincides with the classical forward-backward-forward splitting (FBFS) when the Lipschitz operator in FBFS is skew-adjoint. Interesting recent work [78] proposes a four operator splitting that recovers AFBA as well as the forward-backward-forward splitting (FBFS). Moreover, several new splittings have been proposed recently that involve Lipschitz continuous operators but unlike FBFS require only one evaluation of the Lipschitz operator per iteration [116, 36, 146, 140].

Chapter 3

A unifying framework for primal-dual proximal algorithms

This chapter is based on:

Latafat, P., and Patrinos, P. *Asymmetric forward–backward–adjoint splitting for solving monotone inclusions involving three operators.* Computational Optimization and Applications 68, 1 (Sep 2017), 57–93.

Latafat P., Patrinos P., *Primal-dual proximal algorithms for structured convex optimization: a unifying framework*, in Chapter 5 of Large-Scale and Distributed Optimization, (Giselsson P., and Rantzer A., eds.), vol. 2227 of Lecture Notes in Mathematics, Springer International Publishing, 2018, pp. 97-120.

3.1 Introduction

In this chapter we revisit the convex optimization problem (1.1). As discussed in Chapter 1 this model is quite rich and captures a plethora of problems arising in machine learning, signal processing and control [49, 152, 92].

Here, we present a simple primal-dual framework that relies on the splitting method introduced in Chapter 2. Recall that this splitting involves a stepsize parameter α_k (*cf.* Alg. 2.1), which is dynamically computed at each iteration. We also consider a variant with constant stepsize that serves to simplify the analysis for primal-dual algorithms discussed in Section 3.2.1. We provide a general and easy-to-check convergence condition for the stepsizes in Assumption 3.II and Assumption 3.III (for the case with dynamically computed stepsize parameter). Furthermore, we discuss four mild regularity assumptions on the

functions involved in (1.1) that are sufficient for metric subregularity of the operator defining the primal-dual optimality conditions (*cf.* Lem.s 3.8 and 3.11). Linear convergence rate is then deduced based on the results developed for AFBA (*cf.* Thm. 3.5). These results do not impose additional restrictions on the stepsizes of the algorithms. It is important to note that the provided conditions are much weaker than strong convexity and in many cases do not imply a unique primal or dual solution.

It is worth mentioning that another class of primal-dual algorithms was introduced recently that rely on iterative projections onto half-spaces containing the set of solutions [4, 48]. This class of algorithms is not covered by the analysis of this chapter.

3.2 A simple framework for primal-dual algorithms

In this section we present a simple framework for primal-dual algorithms. For this purpose we consider the following extension of (1.1)

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \varphi(x) = f(x) + g(x) + (h \,\square\, l)(Lx), \tag{3.1}$$

where l is a strongly convex function. Notice that when $l = \delta_{\{0\}}$, the infimal convolution $h \,\square\, l$ reduces to h and problem (1.1) is recovered. As it will become apparent later in the chapter, the framework developed here is not symmetric with respect to the primal and dual variables. Some known algorithms are recovered by applying it to the dual problem (where l^* plays the role of the smooth term, see (3.2)).

Throughout this chapter the following assumptions hold for (3.1).

Assumption 3.I.

(i) $g : \mathbb{R}^n \to \overline{\mathbb{R}}$, $h : \mathbb{R}^r \to \overline{\mathbb{R}}$ are proper closed convex functions, and $L : \mathbb{R}^n \to \mathbb{R}^r$ is a linear mapping.

(ii) $f : \mathbb{R}^n \to \mathbb{R}$ is convex, continuously differentiable, and for some $\beta_f \in [0, \infty)$, ∇f is β_f-Lipschitz continuous with respect to the metric induced by some $Q \succ 0$, *i.e.*, for all $x, y \in \mathbb{R}^n$:

$$\|\nabla f(x) - \nabla f(y)\|_{Q^{-1}} \leq \beta_f \|x - y\|_Q,$$

(iii) $l : \mathbb{R}^r \to \overline{\mathbb{R}}$ is proper closed convex, its conjugate l^* is continuously differentiable, and for some $\beta_l \in [0, \infty)$, ∇l^* is β_l-Lipschitz continuous with respect to the metric induced by some $R \succ 0$.

(iv) The set of solutions to (3.1), denoted by $X^\star$, is nonempty.

(v) (Constraint qualification) There exists $x \in \operatorname{ri}\operatorname{dom} g$ such that $Lx \in \operatorname{ri}\operatorname{dom} h + \operatorname{ri}\operatorname{dom} l$.

In Assumption 3.I*(ii)* the constant β_f is not absorbed into the metric Q in order to be able to treat the case when ∇f is constant in a uniform fashion by setting $\beta_f = 0$. The same reasoning applies to Assumption 3.I*(iii)*.

The dual problem is given by

$$\underset{u\in\mathbb{R}^r}{\text{minimize}}\ (g^* \,\square\, f^*)(-L^\top u) + h^*(u) + l^*(u). \tag{3.2}$$

Notice the similar structure of the dual problem in which l, f and h, g have swapped roles. A well-established approach for solving (3.1) is to consider the associated convex-concave saddle point problem given by

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \underset{u\in\mathbb{R}^r}{\text{maximize}}\ \mathcal{L}(x,u) \coloneqq f(x) + g(x) + \langle Lx, u\rangle - h^*(u) - l^*(u). \tag{3.3}$$

The primal-dual optimality conditions are

$$\begin{cases} 0 \in \partial g(x) + \nabla f(x) + L^\top u, \\ 0 \in \partial h^*(u) + \nabla l^*(u) - Lx. \end{cases} \tag{3.4}$$

Under the constraint qualification condition, the set of solutions for the dual problem denoted by $U^\star$ is nonempty, a saddle point exists, and the duality gap is zero. In fact for any $x^\star \in X^\star$ and $u^\star \in U^\star$, the point $(x^\star, u^\star)$ is a primal-dual solution, see [143, Cor. 31.2.1] and [13, Thm. 19.1].

The right-hand side of the optimality conditions in (3.4) can be split as the sum of three operators:

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \underbrace{\begin{pmatrix} \partial g(x) \\ \partial h^*(u) \end{pmatrix}}_{Az} + \underbrace{\begin{pmatrix} \nabla f(x) \\ \nabla l^*(u) \end{pmatrix}}_{Cz} + \underbrace{\begin{pmatrix} 0 & L^\top \\ -L & 0 \end{pmatrix}}_{M} \underbrace{\begin{pmatrix} x \\ u \end{pmatrix}}_{z}. \tag{3.5}$$

Operator A defined above, is maximally monotone [13, Thm. 21.2, Prop. 20.23], while operator C, being the gradient of $\tilde{f}(x,u) = f(x) + l^*(u)$, is cocoercive and M is skew-symmetric and as such monotone.

Throughout this section we use T to denote the operator above, *i.e.*,

$$0 \in Tz \coloneqq Az + Cz + Mz. \tag{3.6}$$

3.2.1 Primal-dual algorithms with constant stepsize

Algorithm 3.1 describes the proposed primal-dual framework for solving (3.1). This framework is the result of solving the monotone inclusion (3.5) using the three term splitting AFBA with constant stepsize as described in Section 3.3. We defer the derivation and convergence analysis of Algorithm 3.1 to Section 3.4.

The proposed framework involves two scalar parameters $\theta \in [0, \infty)$ and $\mu \in [0, 1]$. Different primal-dual algorithms correspond to different values for these parameters. The iterates in Algorithm 3.1 consist of two proximal updates followed by two correction steps that may or may not be performed depending on the parameters μ and θ. Below we discuss some values for these parameters that are most interesting.

Notice that Algorithm 3.1 is not symmetric with respect to the primal and dual variables; another variant may be obtained by switching their roles (equivalently by applying the algorithm to the dual problem (3.2)).

Algorithm 3.1 A simple framework for primal-dual algorithms

Require: $x^0 \in \mathbb{R}^n$, $u^0 \in \mathbb{R}^r$, algorithm parameters $\mu \in [0, 1]$, $\theta \in [0, \infty)$.

Initialize: Σ, Γ and λ based on Assumptions 3.II*(ii)* and 3.II*(iii)*.

for $k = 0, 1, \ldots$ **do**

$$\bar{x}^k = \operatorname{prox}_g^{\Gamma^{-1}}(x^k - \Gamma L^\top u^k - \Gamma \nabla f(x^k))$$
$$\bar{u}^k = \operatorname{prox}_{h^*}^{\Sigma^{-1}}(u^k + \Sigma L((1-\theta)x^k + \theta \bar{x}^k) - \Sigma \nabla l^*(u^k))$$
$$\tilde{x}^k = \bar{x}^k - x^k,\ \tilde{u}^k = \bar{u}^k - u^k$$
$$x^{k+1} = x^k + \lambda(\tilde{x}^k - \mu(2-\theta)\Gamma L^\top \tilde{u}^k)$$
$$u^{k+1} = u^k + \lambda(\tilde{u}^k + (1-\mu)(2-\theta)\Sigma L \tilde{x}^k)$$

In Section 3.4 we show that the sequence $(x^k, u^k)_{k \in \mathbb{N}}$ generated by Algorithm 3.1 converges to a primal-dual solution if Assumption 3.II holds. Practical rules of thumb for selecting the stepsizes are presented in Section 3.5.

Assumption 3.II (convergence condition for Algorithm 3.1)**.**

(i) (Algorithm parameters) $\theta \in [0, \infty)$, $\mu \in [0, 1]$.

(ii) (Stepsizes) $\Gamma \in \mathbb{S}^n_{++}$, $\Sigma \in \mathbb{S}^r_{++}$ and (relaxation parameter) $\lambda \in (0, 2)$

(iii) The following condition holds

$$\begin{pmatrix} \mathbb{A} & \mathbb{B} \\ \mathbb{B}^\top & \mathbb{C} \end{pmatrix} \succ 0, \tag{3.7}$$

where

$$\mathbb{A} = (\tfrac{2}{\lambda} - 1)\Gamma^{-1} - (1-\mu)(1-\theta)(2-\theta)L^\top \Sigma L - \tfrac{\beta_f}{2\lambda} Q,$$

$$\mathbb{B} = \big(\mu - (1-\mu)(1-\theta) - \tfrac{\theta}{\lambda}\big) L^\top,$$

$$\mathbb{C} = (\tfrac{2}{\lambda} - 1)\Sigma^{-1} - \mu(2-\theta) L \Gamma L^\top - \tfrac{\beta_l}{2\lambda} R.$$

In Algorithm 3.1 the linear mappings L and $L^\top$ must be evaluated twice at every iteration. In the special cases when $\mu = 0, 1$ they may be evaluated only once per iteration by keeping track of the value computed in the previous iteration. As is evident from (3.7), the cases where both the algorithm and the convergence condition simplify are combinations of $\mu = 0, 1$, $\theta = 0, 1, 2$. Here we briefly discuss some of these special cases to demonstrate how (3.7) leads to simple conditions that are often less conservative than the conditions found in the literature. We have dubbed the algorithms based on whether the two proximal updates can be evaluated in parallel and if a primal or a dual correction step is performed.

The first algorithm is the result of setting $\theta = 2$ (regardless of μ) and leads to the algorithm of Condat and Vũ [53, 170] which itself is a generalization of the Chambolle-Pock algorithm [38]. With this choice of θ, the two proximal updates are performed sequentially while no correction step is required. In the box below a general condition is given for its convergence.

SNCA (Sequential No Corrector Algorithm): $\theta = 2$

Substituting $\theta = 2$ in (3.7) and dividing by $\frac{2}{\lambda} - 1$ yields

$$\begin{pmatrix} \Gamma^{-1} - \frac{\beta_f}{2(2-\lambda)} Q & -L^\top \\ -L & \Sigma^{-1} - \frac{\beta_l}{2(2-\lambda)} R \end{pmatrix} \succ 0. \tag{3.8}$$

If $Q, R = \mathrm{id}$ and $\Gamma = \gamma \mathrm{id}$, $\Sigma = \sigma \mathrm{id}$ for some scalars γ, σ, then the following sufficient condition may be used

$$\sigma\gamma \|L\|^2 < \Big(1 - \tfrac{\gamma\beta_f}{2(2-\lambda)}\Big)\Big(1 - \tfrac{\sigma\beta_l}{2(2-\lambda)}\Big), \quad \lambda \in (0, 2). \tag{3.9}$$

Notice that this condition is less conservative than the condition of [53, Thm. 3.1] (see [96, Rem. 5.6]).

In the next algorithm proximal updates are evaluated sequentially, followed by a correction step for the primal variable, hence the name SPCA. Most notable property of this algorithm is that the generated sequence is S-Fejér monotone where S is block diagonal. The algorithm introduced in [95] (discussed in Chapter 4) can be seen as an application of SPCA to the dual problem when the smooth term is zero. Note that SPCA can be transformed into SNCA (when $\lambda = 1$ and $f \equiv 0$) by a change of order, eliminating the primal variable x^k and keeping the auxiliary primal variable $\bar{x}^k$. Below we comment on algorithms proposed in [62, 95] emphasizing the close relation between them.

SPCA (Sequential Primal Corrector Algorithm): $\theta = 1$, $\mu = 1$, $\lambda = 1$

In this case the left-hand side in (3.7) is block diagonal. Therefore the convergence condition simplifies to

$$\Gamma^{-1} - \tfrac{\beta_f}{2} Q \succ 0, \quad \Sigma^{-1} - L\Gamma L^\top - \tfrac{\beta_l}{2} R \succ 0.$$

If $Q = \mathrm{id}$, $R = \mathrm{id}$, $\Gamma = \gamma \mathrm{id}$, $\Sigma = \sigma \mathrm{id}$ for some scalars γ, σ, then it is sufficient to have

$$\gamma\beta_f < 2, \quad \sigma\gamma\|L\|^2 < 1 - \tfrac{\sigma\beta_l}{2}. \tag{3.10}$$

This special case generalizes the recent algorithm proposed in [62]. In particular we allow a third nonsmooth function g as well as the strongly convex function l. In addition to this improvement our convergence condition with $l^* \equiv 0$ (set $\beta_l = 0$ in (3.10)) is less restrictive and doubles the range of acceptable stepsize γ. The convergence condition in that work is given in our notation as $\gamma\beta_f < 1$ and $\sigma\gamma\|L\|^2 < 1$ [62, Cor. 3.2].

The next algorithm features sequential proximal updates that are followed by a correction step for the dual variable, for all values of $\theta \in (0, \infty)$. The parallel variant of this algorithm, referred to as PDCA, is discussed in a separate box below. We have observed that selecting θ so as to maximize the stepsizes, *i.e.*, $\theta = 1.5$, leads to faster convergence [100]. Moreover, if we set $\Sigma = \Gamma^{-1}$, this choice of μ leads to a three-block ADMM or equivalently a generalization of DRS to include a third cocoercive operator (see Section 2.3.4).

SDCA (Sequential Dual Corrector Algorithm): $\theta \in (0,\infty)$, $\mu = 0$, $\lambda = 1$

In this case the convergence condition simplifies to

$$\begin{pmatrix} \Gamma^{-1} - (1-\theta)(2-\theta)L^\top \Sigma L - \frac{\beta_f}{2} Q & -L^\top \\ -L & \Sigma^{-1} - \frac{\beta_l}{2} R \end{pmatrix} \succ 0.$$

If $l = \delta_{\{0\}}$ then $\beta_l = 0$ and using Schur complement we derive the following condition

$$\Gamma^{-1} - \big(\theta^2 - 3\theta + 3\big) L^\top \Sigma L - \tfrac{\beta_f}{2} Q \succ 0.$$

If in addition $Q = \mathrm{id}$, and $\Gamma = \gamma \mathrm{id}$, $\Sigma = \sigma \mathrm{id}$ for some scalars γ, σ, then we have the following sufficient condition

$$(\theta^2 - 3\theta + 3)\sigma\gamma \|L\|^2 < 1 - \tfrac{\gamma \beta_f}{2}.$$

The next algorithm appears to be new and involves parallel proximal updates followed by a primal correction step.

PPCA (Parallel Primal Corrector Algorithm): $\theta = 0$, $\mu = 1$

The convergence condition is given by

$$\begin{pmatrix} (\frac{2}{\lambda} - 1)\Gamma^{-1} - \frac{\beta_f}{2\lambda} Q & L^\top \\ L & (\frac{2}{\lambda} - 1)\Sigma^{-1} - 2L\Gamma L^\top - \frac{\beta_l}{2\lambda} R \end{pmatrix} \succ 0.$$

If $f \equiv 0$ (set $\beta_f = 0$) and $\lambda = 1$, using Schur complement yields the following condition

$$\Sigma^{-1} - 3L\Gamma L^\top - \tfrac{\beta_l}{2} R \succ 0.$$

If in addition $R = \mathrm{id}$ and $\Gamma = \gamma \mathrm{id}$, $\Sigma = \sigma \mathrm{id}$ for some scalars γ, σ, then the following sufficient condition may be used

$$3\sigma\gamma \|L\|^2 < 1 - \tfrac{\sigma \beta_l}{2}.$$

The parallel variant of SDCA is considered below. Interestingly, by switching the order of the proximal updates (since $\theta = 0$), PDCA may be seen as PPCA applied to the dual problem (3.2).

PDCA (Parallel Dual Corrector Algorithm): $\theta = 0$, $\mu = 0$

In this case the convergence condition simplifies to

$$\begin{pmatrix} (\frac{2}{\lambda} - 1)\Gamma^{-1} - 2L^\top \Sigma L - \frac{\beta_f}{2\lambda} Q & -L^\top \\ -L & (\frac{2}{\lambda} - 1)\Sigma^{-1} - \frac{\beta_l}{2\lambda} R \end{pmatrix} \succ 0.$$

If $l = \delta_{\{0\}}$ (set $\beta_l = 0$) and $\lambda = 1$, using Schur complement yields the following condition

$$\Gamma^{-1} - 3L^\top \Sigma L - \tfrac{\beta_f}{2} Q \succ 0.$$

If in addition $Q = \mathrm{id}$, and $\Gamma = \gamma \mathrm{id}$, $\Sigma = \sigma \mathrm{id}$ for some scalars γ, σ, then the following sufficient condition may be used

$$3\sigma\gamma \|L\|^2 < 1 - \tfrac{\gamma\beta_f}{2}.$$

The last special case considered here involves sequential proximal updates followed by correction steps for the primal and dual variables. As noted before for this choice of μ, the linear mappings L and $L^\top$ must be evaluated twice at every iteration.

PPDCA (Parallel Primal and Dual Corrector Algorithm): $\theta = 0$, $\mu = 0.5$

In this case condition (3.7) reduces to:

$$\Gamma^{-1} - \tfrac{\lambda}{(2-\lambda)} L^\top \Sigma L - \tfrac{\beta_f}{2(2-\lambda)} Q \succ 0, \quad \Sigma^{-1} - \tfrac{\lambda}{(2-\lambda)} L \Gamma L^\top - \tfrac{\beta_l}{2(2-\lambda)} R \succ 0.$$

If $Q = \mathrm{id}$, $R = \mathrm{id}$, $\lambda = 1$ and $\Gamma = \gamma \mathrm{id}$, $\Sigma = \sigma \mathrm{id}$ for some scalars γ, σ, the following sufficient condition may be used

$$\sigma\gamma \|L\|^2 < \min\{1 - \tfrac{\gamma\beta_f}{2}, 1 - \tfrac{\sigma\beta_l}{2}\}.$$

This special case generalizes [31, Alg. (4.8)] with the addition of the smooth function f and the strongly convex function l.

Remark 3.1. In [180] a new primal-dual algorithm is proposed for solving (3.1) that recovers the three term splitting of [58] when $L = \mathrm{id}$. It is interesting to note the relation to our framework. This algorithm is given by [180, Eq. 4a-4c]:

$$u^{k+1} = \mathrm{prox}_{\sigma h^*}\left(u^k + \sigma L \bar{x}^k - \sigma \nabla l^*(u^k)\right) \tag{3.11a}$$

$$x^{k+1} = \mathrm{prox}_{\gamma g}\left(x^k - \gamma L^\top u^{k+1} - \gamma \nabla f(x^k)\right) \tag{3.11b}$$

$$\bar{x}^{k+1} = 2x^{k+1} - x^k + \gamma\nabla f(x^k) - \gamma\nabla f(x^{k+1}). \tag{3.11c}$$

Let us consider SPCA, a special case of Algorithm 3.1 where $\theta = 1$, $\mu = 1$, with $\lambda = 1$, $\Sigma = \sigma\mathrm{id}$, $\Gamma = \gamma\mathrm{id}$. Change the order of the updates starting from the dual update to obtain:

$$u^{k+1} = \operatorname{prox}_{\sigma h^*}\left(u^k + \sigma L\bar{x}^k - \sigma\nabla l^*(u^k)\right) \tag{3.12a}$$

$$x^{k+1} = \bar{x}^k - \gamma L^\top(u^{k+1} - u^k) \tag{3.12b}$$

$$\bar{x}^{k+1} = \operatorname{prox}_{\gamma g}\left(x^{k+1} - \gamma L^\top u^{k+1} - \gamma\nabla f(x^{k+1})\right). \tag{3.12c}$$

In order to be able to compare the stepsizes of the two algorithms without algebraic difficulty (see [180, Assumption 1 and Thm. 1]) set $l = \delta_{\{0\}}$ and $Q = \mathrm{id}$. Using the above parameters in Assumption 3.II*(iii)* we obtain the following sufficient condition for (3.12):

$$\gamma\beta_f < 2, \quad \gamma\sigma\|L\|^2 < 1. \tag{3.13}$$

This stepsize condition is the same as that of [180]. However, the two algorithms are quite different in the way the primal variables are updated. In Section 3.5, simulations are performed for solving dual support vector machine problems, and the results indicate very similar convergence speed for the two algorithms, see Figure 3.2. □

3.2.2 Primal-dual algorithms with a dynamically updated stepsize

In this section a variant of Algorithm 3.1 is discussed that involves a stepsize parameter that is computed based on the primal and dual variables at every iteration, leading to potentially larger stepsizes. This algorithm is the result of applying AFBA Algorithm 2.1 to solve the monotone inclusion (3.5). The algorithm converges to a primal-dual solution provided that Assumption 3.III holds. The analysis is detailed in Section 3.4. As remarked in Section 3.2.1, Algorithm 3.1 is also not symmetric with respect to the primal and dual variables, and a variant may be obtained by switching their roles.

Algorithm 3.2 A simple framework for primal-dual algorithms

Require: $x^0 \in \mathbb{R}^n$, $u^0 \in \mathbb{R}^r$, algorithm parameters $\mu \in [0,1]$, $\theta \in [0,\infty)$.

Initialize: Σ, Γ and λ based on Assumption 3.III.

for $k = 0, 1, \ldots$ **do**

$\bar{x}^k = \operatorname{prox}_g^{\Gamma^{-1}}(x^k - \Gamma L^\top u^k - \Gamma \nabla f(x^k))$

$\bar{u}^k = \operatorname{prox}_{h^*}^{\Sigma^{-1}}(u^k + \Sigma L((1-\theta)x^k + \theta \bar{x}^k) - \Sigma \nabla l^*(u^k))$

$\tilde{x}^k = \bar{x}^k - x^k$, $\tilde{u}^k = \bar{u}^k - u^k$

Compute α_k according to (3.14)

$x^{k+1} = x^k + \alpha_k(\tilde{x}^k - \mu(2-\theta)\Gamma L^\top \tilde{u}^k)$

$u^{k+1} = u^k + \alpha_k(\tilde{u}^k + (1-\mu)(2-\theta)\Sigma L \tilde{x}^k)$

In Algorithm 3.2

$$\alpha_k = \lambda \frac{\|\tilde{x}^k\|^2_{\Gamma^{-1}} + \|\tilde{u}^k\|^2_{\Sigma^{-1}} - \theta\langle \tilde{x}^k, L^\top \tilde{u}^k\rangle}{V(\tilde{x}^k, \tilde{u}^k)}, \tag{3.14}$$

where

$$\begin{aligned} V(\tilde{x}^k, \tilde{u}^k) = {} & \|\tilde{x}^k\|^2_{\Gamma^{-1}} + \|\tilde{u}^k\|^2_{\Sigma^{-1}} + (1-\mu)(1-\theta)(2-\theta)\|L\tilde{x}^k\|^2_\Sigma \\ & + \mu(2-\theta)\|L^\top \tilde{u}^k\|^2_\Gamma + 2((1-\mu)(1-\theta) - \mu)\langle \tilde{x}^k, L^\top \tilde{u}^k\rangle. \end{aligned} \tag{3.15}$$

Note that in many distributed applications the computation for the stepsize in (3.14) is disadvantageous since it would entail global coordination. Nevertheless, if this is not a concern, it can lead to larger stepsizes.

In Section 3.4 it is shown that the sequence $(x^k, u^k)_{k\in\mathbb{N}}$ generated by Algorithm 3.1 (resp. Algorithm 3.2) converges to a primal-dual solution if Assumption 3.II (resp. Assumption 3.III) holds. Moreover, linear convergence rates are established if either one of four mild regularity assumptions hold for functions f, g, l and h (*cf.* Cor. 3.12).

Assumption 3.III (convergence condition for Algorithm 3.2)**.** Additionally to Assumptions 3.II*(i)* and 3.II*(ii)* suppose that the following holds

$$\begin{pmatrix} \Gamma^{-1} - \frac{1}{2(2-\lambda)}\beta_f Q & -\frac{\theta}{2}L^\top \\ -\frac{\theta}{2}L & \Sigma^{-1} - \frac{1}{2(2-\lambda)}\beta_l R \end{pmatrix} \succ 0. \tag{3.16}$$

3.3 Simplified asymmetric forward-backward-adjoint splitting

A new three term splitting technique was introduced in Chapter 2 for the problem of finding $z \in \mathbb{R}^p$ such that

$$0 \in Tz := Az + Cz + Mz, \tag{3.17}$$

where A is maximally monotone, C is cocoercive and M is a monotone linear mapping. AFBA in its original form includes the stepsize α_k, (*cf.* Alg. 2.1). Here we simplify the algorithm by considering a constant stepsize (*cf.* Alg. 3.3). This variant of AFBA is particularly advantageous in distributed applications where global coordination may be infeasible. Furthermore, unlike Algorithm 2.1, cocoercivity of the operator C is considered with respect to some norm independent of the parameters of the algorithm, and the convergence condition is derived in terms of a matrix inequality. These changes simplify the analysis for the primal-dual algorithms discussed in Section 3.2. We remind the reader that Algorithm 3.1 is the result of solving the primal-dual optimality conditions using AFBA. We defer the derivation and convergence analysis of Algorithm 3.1 until Section 3.4. A basic key inequality that we use is the following.

Lemma 3.2 (three-point inequality)**.** *Suppose that $C : \mathbb{R}^p \to \mathbb{R}^p$ is cocoercive with respect to $\|\cdot\|_U$ with $U \in \mathbb{S}^p_{++}$, and let $V : \mathbb{R}^p \to \mathbb{R}^p$ be a linear mapping such that $V \circ C = C$ (identity is the trivial choice). Then, for any three points $z, z', z'' \in \mathbb{R}^p$ we have*

$$\langle Cz - Cz', z' - z''\rangle \leq \tfrac{1}{4}\|V^\top(z - z'')\|_U^2. \tag{3.18}$$

Proof. Use the inequality, valid for any $a, b \in \mathbb{R}^p$,

$$\langle a, b\rangle = 2\langle \tfrac{1}{2}U^{\frac{1}{2}}a, U^{-\frac{1}{2}}b\rangle \leq \tfrac{1}{4}\|a\|_U^2 + \|b\|_{U^{-1}}^2, \tag{3.19}$$

together with (1.7) and $V \circ C = C$ to derive

$$\begin{aligned}
\langle Cz - Cz', z' - z''\rangle &= \langle V(Cz - Cz'), z - z''\rangle + \langle Cz - Cz', z' - z\rangle \\
&= \langle Cz - Cz', V^\top(z - z'')\rangle + \langle Cz - Cz', z' - z\rangle \\
&\overset{3.19}{\leq} \tfrac{1}{4}\|V^\top(z - z'')\|_U^2 + \|Cz - Cz'\|_{U^{-1}}^2 \\
&\quad + \langle Cz - Cz', z' - z\rangle \\
&\overset{1.7}{\leq} \tfrac{1}{4}\|V^\top(z - z'')\|_U^2.
\end{aligned}$$

□

The main reason for considering V is to avoid conservative bounds in (3.18). For example, assume that the space is partitioned into two blocks $z = (z_1, z_2) \in \mathbb{R}^p$ with $z_1 \in \mathbb{R}^{p_1}$ and $z_2 \in \mathbb{R}^{p_2}$, and $C : \mathbb{R}^p \to \mathbb{R}^p$ is given by $Cz = (C_1 z_1, 0)$ where $C_1 : \mathbb{R}^{p_1} \to \mathbb{R}^{p_1}$ is cocoercive. Using inequality (3.19) without taking into account the structure of C, *i.e.*, that $V \circ C = C$ for $V = \mathrm{blkdiag}(\mathrm{I}_{p_1}, 0_{p_2})$, would result in the whole vector appearing in the upper bound in (3.18).

Algorithm 3.3 involves two matrices H and S that are instrumental to its flexibility. In Section 3.4 we discuss a choice for H and S and demonstrate how Algorithm 3.1 is derived. Below we summarize the assumptions for the monotone inclusion (3.17) and the convergence conditions for Algorithm 3.3 (*cf.* Thm. 3.4).

Assumption 3.IV.

(i) Assumptions for the monotone inclusion (3.17):

1) The operator $A : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ is maximally monotone.
2) The linear mapping $M : \mathbb{R}^p \to \mathbb{R}^p$ is monotone.
3) The operator $C : \mathbb{R}^p \to \mathbb{R}^p$ is cocoercive with respect to $\|\cdot\|_U$ with $U \in \mathbb{S}^p_{++}$. In addition, $V : \mathbb{R}^p \to \mathbb{R}^p$ is a linear mapping such that $V \circ C = C$ (identity is the trivial choice).

(ii) Convergence conditions for Algorithm 3.3:

1) The matrix $H := P + K$, where $P \in \mathbb{S}^p_{++}$ and K is a skew-symmetric matrix.
2) The matrix $S \in \mathbb{S}^p_{++}$ and the following holds

$$2P - \tfrac{1}{2} V U V^\top - D \succ 0, \tag{3.20}$$

where

$$D := (H + M^\top)^\top S^{-1} (H + M^\top). \tag{3.21}$$

Algorithm 3.3 AFBA with constant stepsize

Require: $z^0 \in \mathbb{R}^p$

Initialize: set S and H according to Assumption 3.IV *(ii)*.

for $k = 0, 1, \ldots$ **do**

$\bar{z}^k = (H + A)^{-1}(H - M - C) z^k$

$z^{k+1} = z^k + S^{-1}(H + M^\top)(\bar{z}^k - z^k)$

Lemma 3.3. *Let Assumption 3.IV hold. Consider the update for $\bar{z}$ in Algorithm 3.3*

$$\bar{z} = (H + A)^{-1}(H - M - C)z. \tag{3.22}$$

For all $z^\star \in \operatorname{zer} T$ the following holds

$$\langle z - z^\star, (H + M^\top)(\bar{z} - z)\rangle \leq \tfrac{1}{4}\|V^\top(z - \bar{z})\|_U^2 - \|z - \bar{z}\|_P^2. \tag{3.23}$$

Proof. Use (3.22) and the fact that $z^\star \in \operatorname{zer} T$, together with monotonicity of A at $z^\star$ and $\bar{z}$ to derive

$$0 \leq \langle -Mz^\star - Cz^\star + Mz + Cz + H(\bar{z} - z), z^\star - \bar{z}\rangle. \tag{3.24}$$

In Lemma 3.2 set $z' = z^\star$ and $z'' = \bar{z}$

$$\langle Cz - Cz^\star, z^\star - \bar{z}\rangle \leq \tfrac{1}{4}\|V^\top(z - \bar{z})\|_U^2. \tag{3.25}$$

For the remaining terms in (3.24) use skew-symmetry of K (twice) and monotonicity of M:

$$\begin{aligned}
&\langle -Mz^\star + Mz + H(\bar{z} - z), z^\star - \bar{z}\rangle \\
=&\langle -Mz^\star + Mz + P(\bar{z} - z) + K(\bar{z} - z) + K(z^\star - \bar{z}), z^\star - \bar{z}\rangle \\
=&\langle (M - K)(z - z^\star) + P(\bar{z} - z), z^\star - \bar{z}\rangle \\
=&\langle (M - K)(z - z^\star) + P(\bar{z} - z), z^\star - z\rangle \\
&+ \langle (M - K)(z - z^\star) + P(\bar{z} - z), z - \bar{z}\rangle \\
\leq&\langle P(\bar{z} - z), z^\star - z\rangle + \langle (M - K)(z - z^\star), z - \bar{z}\rangle - \|\bar{z} - z\|_P^2 \\
\leq&\langle z - z^\star, (M^\top + H)(z - \bar{z})\rangle - \|\bar{z} - z\|_P^2.
\end{aligned}$$

Combining this with (3.24) and (3.25) completes the proof. □

Theorem 3.4 (convergence)**.** *Let Assumption 3.IV hold. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 3.3. Then, the following inequality holds for all $k \in \mathbb{N}$*

$$\|z^{k+1} - z^\star\|_S^2 \leq \|z^k - z^\star\|_S^2 - \|z^k - \bar{z}^k\|^2_{2P - \frac{1}{2}VUV^\top - D}, \tag{3.26}$$

and $(z^k)_{k\in\mathbb{N}}$ converges to a point $z^\star \in \operatorname{zer} T$.

Proof. We show that the generated sequence is S-Fejér monotone with respect to $\operatorname{zer} T$. For any $z^\star \in \operatorname{zer} T$ using the z^{k+1} update in Algorithm 3.3 we have

$$\begin{aligned}\|z^{k+1} - z^\star\|_S^2 =& \|z^k - z^\star\|_S^2 + \|S^{-1}(H + M^\top)(\bar{z}^k - z^k)\|_S^2 \\ &+ 2\langle z^k - z^\star, (H + M^\top)(\bar{z}^k - z^k)\rangle \\ \overset{(3.23)}{\leq}& \|z^k - z^\star\|_S^2 + \|S^{-1}(H + M^\top)(\bar{z}^k - z^k)\|_S^2 \\ &+ \tfrac{1}{2}\|V^\top(z^k - \bar{z}^k)\|_U^2 - 2\|z^k - \bar{z}^k\|_P^2 \\ \overset{(3.21)}{\leq}& \|z^k - z^\star\|_S^2 - \|z^k - \bar{z}^k\|^2_{2P - \frac{1}{2}VUV^\top - D}.\end{aligned}$$

Therefore, the sequence $(z^k - \bar{z}^k)_{k\in\mathbb{N}}$ converges to zero. Convergence of $(z^k)_{k\in\mathbb{N}}$ to a point in $\operatorname{zer} T$ follows by standard arguments; see the last part of the proof of Theorem 2.1. □

It is shown in Theorem 2.4 that under a metric subregularity assumption for operator T, (3.17), the sequence generated by AFBA (*cf.* Alg. 2.1) converges R-linearly. In the next theorem we show that Algorithm 3.3 also enjoys linear convergence if either (i) operator T is metrically subregular at all $z^\star \in \operatorname{zer} T$ for 0 or (ii) when the operator $\mathrm{id} - T_{\mathrm{AFBA}}$ has this property, where T_{AFBA} denotes the operator that maps z^k to z^{k+1} in Algorithm 3.3. In addition, we show that the two conditions are equivalent. Note that although not stated, this equivalence also holds for T_{AFBA} corresponding to Algorithm 2.1 in Chapter 2. In Section 3.4 we exploit the first condition in order to establish linear convergence based on the properties of the cost functions involved.

Theorem 3.5 (metric subregularity equivalence and linear convergence)**.** *Let Assumption 3.IV hold. The following subregularity assumptions are equivalent.*

(i) $T = A + M + C$ *is metrically subregular at all* $z^\star \in \operatorname{zer} T$ *for* 0*.*

(ii) Let T_{AFBA} *denote the operator that maps* z^k *to* z^{k+1} *in Algorithm 3.3, i.e.,* $z^{k+1} = T_{\mathrm{AFBA}}(z^k)$*. The operator* $\mathrm{id} - T_{\mathrm{AFBA}}$ *is metrically subregular at all* $z^\star \in \operatorname{zer} T$ *for* 0*.*

Moreover, if either condition holds then $(z^k)_{k\in\mathbb{N}}$ *generated by Algorithm 3.3 converges R-linearly to some* $z^\star \in \operatorname{zer} T$ *and* $(\operatorname{dist}_S(z^k, \operatorname{zer} T))_{k\in\mathbb{N}}$ *converges Q-linearly to zero.*

Proof. (ii) ⇒ (i): assume that $\mathcal{R} \coloneqq \mathrm{id} - T_{\mathrm{AFBA}}$ is metrically subregular at $z^\star$ for 0. Then, there exists $\eta > 0$ and a neighborhood $\mathcal{U}$ of $z^\star$ such that

$$\operatorname{dist}(z, \mathcal{R}^{-1}0) \leq \eta \operatorname{dist}(0, \mathcal{R}z) \quad \forall z \in \mathcal{U}. \tag{3.27}$$

Using the definition of T_{AFBA} we have that $z \in \operatorname{fix} T_{\mathrm{AFBA}}$ if and only if $(H + A)^{-1}(H - M - C)z = z$ which is equivalent to $z \in \operatorname{zer} T$. Therefore, the sets $\mathcal{R}^{-1}0$ and $T^{-1}0$ are equal. In what follows, we upper bound $\operatorname{dist}(0, \mathcal{R}z)$ by $\operatorname{dist}(0, Tz)$. Let $w \in Tz = Az + Mz + Cz$ and consider $\bar{z}$ as in (3.22). We have

$$Hz - Mz - Cz - H\bar{z} \in A\bar{z}.$$

Using this together with the monotonicity of A at z and $\bar{z}$, we obtain:

$$\begin{aligned} 0 \leq & \langle z - \bar{z}, (w - Mz - Cz) - (Hz - Mz - Cz - H\bar{z}) \rangle \\ = & \langle z - \bar{z}, w - Hz + H\bar{z} \rangle = \langle z - \bar{z}, w \rangle - \|\bar{z} - z\|_P^2, \end{aligned}$$

where in the last equality we have used the fact that $H = P + K$ and K is skew-symmetric.

By the Cauchy–Schwarz inequality

$$\|\bar{z} - z\|_P^2 \leq \langle z - \bar{z}, w \rangle \leq \|\bar{z} - z\|_P \|w\|_{P^{-1}},$$

therefore

$$\|\bar{z} - z\|_P \leq \|w\|_{P^{-1}}. \tag{3.28}$$

On the other hand since $T_{\mathrm{AFBA}} z = z + S^{-1}(H + M^\top)(\bar{z} - z)$,

$$\|\mathcal{R}z\| \leq \|S^{-1}(H + M^\top)P^{-1/2}\| \|\bar{z} - z\|_P.$$

Combine this with (3.27) and (3.28) to obtain

$$\begin{aligned} \operatorname{dist}(z, T^{-1}0) = \operatorname{dist}(z, \mathcal{R}^{-1}0) \leq & \eta \|\mathcal{R}z\| \\ \leq & \eta \|S^{-1}(H + M^\top)P^{-1/2}\| \|P^{-1}\|^{1/2} \|w\|. \end{aligned}$$

Since $w \in Tz$ was arbitrary, we conclude that T is metrically subregular at $z^\star$ for 0 (possibly with a different subregularity modulus).

(i) ⇒ (ii): assume that T is metrically subregular at $z^\star$ for 0, *i.e.*, there exists $\eta > 0$ and neighborhood $\mathcal{U}$ of $z^\star$ such that

$$\operatorname{dist}(z, T^{-1}0) \leq \eta \operatorname{dist}(0, Tz) \quad \forall z \in \mathcal{U}. \tag{3.29}$$

By (3.23) and the Cauchy–Schwarz inequality we infer that

$$\|\bar{z} - z\| \leq c\|z - z^\star\|,$$

for some positive constant c. Hence, there exists a neighborhood $\bar{\mathcal{U}} \subset \mathcal{U}$ of $z^\star$ such that if $z \in \bar{\mathcal{U}}$ then $\bar{z} \in \mathcal{U}$. Fix a point $z \in \bar{\mathcal{U}}$ so that $\bar{z} \in \mathcal{U}$. By (3.29) it holds that:

$$\operatorname{dist}(\bar{z}, T^{-1}0) \leq \eta \operatorname{dist}(0, T\bar{z}). \tag{3.30}$$

Define

$$v \coloneqq -(H - M)(\bar{z} - z) + C\bar{z} - Cz.$$

By the triangle inequality and Lipschitz continuity of C

$$\begin{aligned}\|v\| &= \|(H - M)(\bar{z} - z) - C\bar{z} + Cz\| \\ &\leq \|(H - M)(\bar{z} - z)\| + \|C\bar{z} - Cz\| \leq \xi\|\bar{z} - z\|,\end{aligned} \tag{3.31}$$

for some positive ξ. Moreover, noting that $v \in T\bar{z}$, it follows from (3.30) that

$$\operatorname{dist}(\bar{z}, T^{-1}0) \leq \eta\|v\| \leq \eta\xi\|\bar{z} - z\|, \tag{3.32}$$

where we used (3.31) in the second inequality. Invoking triangle inequality we have

$$\begin{aligned}\operatorname{dist}(z, \mathcal{R}^{-1}0) = \operatorname{dist}(z, T^{-1}0) &\leq \operatorname{dist}(\bar{z}, T^{-1}0) + \|\bar{z} - z\| \\ &\leq (1 + \eta\xi)\|\bar{z} - z\|.\end{aligned} \tag{3.33}$$

On the other hand it holds that

$$\|\bar{z} - z\| \leq \|(H + M^\top)^{-1}S\|\|\mathcal{R}z\|.$$

Combining this with (3.33) yields

$$\operatorname{dist}(z, \mathcal{R}^{-1}0) \leq (1 + \eta\xi)\|(H + M^\top)^{-1}S\|\|\mathcal{R}z\| \quad \forall z \in \bar{\mathcal{U}},$$

i.e., that $\mathcal{R}$ is metrically subregular at $z^\star$ for 0. This completes the proof of equivalence of (i) and (ii).

The proof of linear convergence based on (i) is identical to Theorem 2.4 and is therefore omitted. □

3.4 A unified convergence analysis for primal-dual algorithms

Our goal in this section is to describe how Algorithms 3.1 and 3.2 are derived and to establish their convergence. The idea is to solve the monotone inclusion corresponding to the primal-dual optimality conditions using *asymmetric forward-backward-adjoint* splitting (AFBA) described in Section 3.3. In order to recover Algorithm 3.1 simply apply Algorithm 3.3 to this monotone inclusion with the following parameters: let $\theta \in [0, \infty)$ and set $H = P + K$ with

$$P = \begin{pmatrix} \Gamma^{-1} & -\frac{\theta}{2}L^\top \\ -\frac{\theta}{2}L & \Sigma^{-1} \end{pmatrix}, \quad K = \begin{pmatrix} 0 & \frac{\theta}{2}L^\top \\ -\frac{\theta}{2}L & 0 \end{pmatrix}, \tag{3.34}$$

and $S = \left(\lambda\mu S_1^{-1} + \lambda(1-\mu)S_2^{-1}\right)^{-1}$ where $\mu \in [0,1]$, $\lambda \in (0,2)$ with

$$S_1 = \begin{pmatrix} \Gamma^{-1} & (1-\theta)L^\top \\ (1-\theta)L & \Sigma^{-1} + (1-\theta)(2-\theta)L\Gamma L^\top \end{pmatrix}, \; S_2 = \begin{pmatrix} \Gamma^{-1} + (2-\theta)L^\top \Sigma L & -L^\top \\ -L & \Sigma^{-1} \end{pmatrix}.$$

Notice that with P and K set as in (3.34), H has a lower (block) triangular structure. Therefore the backward step $(H + A)^{-1}$ in Algorithm 3.3 can be carried out sequentially, see Lemma 2.5. Algorithm 3.1 is derived by noting this and substituting S and H defined above. Algorithm 3.2 is obtained similarly by using Algorithm 2.1 with P, K, M and S as defined in this chapter. Therefore, convergence of both algorithms follow directly from that of the two variants of AFBA. This is summarized in the next theorem.

Theorem 3.6 (convergence of Algorithms 3.1 and 3.2)**.** *Let Assumption 3.I hold. Consider the sequence $(z^k)_{k\in\mathbb{N}} = (x^k, u^k)_{k\in\mathbb{N}}$ generated by Algorithm 3.1 (resp. Algorithm 3.2). Suppose that Assumption 3.II (resp. Assumption 3.III) holds. Then, $(z^k)_{k\in\mathbb{N}}$ converges to some point $z^\star \in \operatorname{zer} T$.*

Proof. We begin by establishing the claim for Algorithm 3.3. The goal is to verify that Assumptions 3.I and 3.II are sufficient for Assumption 3.IV to hold. As noted in Section 3.2, operator A is maximally monotone [13, Thm. 21.2, Prop. 20.33], and the linear mapping M is skew-adjoint and as such monotone. The operator C is cocoercive with respect to the metric induced by $U = \operatorname{blkdiag}(\beta_f Q, \beta_l R)$. In Assumption 3.IV we use the linear mapping V in order to avoid conservative requirements. The special cases when $f \equiv 0$ (or $l^* \equiv 0$) are captured by setting $V = \operatorname{blkdiag}(0_n, \mathrm{I}_r)$ (or $V = \operatorname{blkdiag}(\mathrm{I}_n, 0_r)$). It remains to verify Assumption 3.IV *(ii)*. Evaluating D according to (3.21)

yields the following $D = \lambda\mu D_1 + \lambda(1-\mu)D_2$ where

$$D_1 = \begin{pmatrix} \Gamma^{-1} & -L^\top \\ -L & \Sigma^{-1} + (2-\theta)L\Gamma L^\top \end{pmatrix},$$

$$D_2 = \begin{pmatrix} \Gamma^{-1} + (1-\theta)(2-\theta)L^\top \Sigma L & (1-\theta)L^\top \\ (1-\theta)L & \Sigma^{-1} \end{pmatrix}.$$

Noting that $\Gamma, \Sigma \succ 0$, and using Schur complement for D_1 and P defined in (3.34) we have

$$D_1 \succ 0 \;\Leftrightarrow\; \Sigma^{-1} + (1-\theta)L\Gamma L^\top \succ 0, \quad \Sigma^{-1} - \tfrac{\theta^2}{4} L\Gamma L^\top \succ 0 \;\Leftrightarrow\; P \succ 0.$$

Thus, since $1-\theta \geq -\frac{\theta^2}{4}$ for all θ, we have that $D_1 \succ 0$ if $P \succ 0$. It can be shown that the same argument applies for S_1, S_2 and D_2. The sum of two positive definite matrices is also positive definite, therefore $S, D \succ 0$ if $P \succ 0$. Matrix P is symmetric positive definite if (3.20) holds. The convergence conditions in Assumption 3.II are the result of replacing D, P, U and V in (3.20).

In order to establish convergence of Algorithm 3.2 we start by adapting the convergence analysis of Algorithm 2.1 to suit the notation of this chapter. This amounts to the following condition:

$$(2-\lambda)P - \tfrac{1}{2}VUV^\top \succ 0, \quad \lambda \in (0,2). \tag{3.35}$$

To see this note that using Lemma 3.3, inequality (2.10) is updated as follows:

$$\begin{aligned}
\|z^{k+1} - z^\star\|_S^2 \leq{}& \|z^k - z^\star\|_S^2 + \lambda^2 \|(H+M^*)\tilde{z}^k\|_{S^{-1}}^{-2} \|\tilde{z}^k\|_P^4 \\
&+ \lambda \|(H+M^*)\tilde{z}^k\|_{S^{-1}}^{-2} \|\tilde{z}^k\|_P^2 \left(\tfrac{1}{2}\|V^\top(z^k - \bar{z}^k)\|_U^2 - 2\|z^k - \bar{z}^k\|_P^2 \right) \\
={}& \|z^k - z^\star\|_S^2 - \|(H+M^*)\tilde{z}^k\|_{S^{-1}}^{-2} \|\tilde{z}^k\|_P^2 \left(\lambda \|\tilde{z}\|_{2P - \frac{1}{2}VUV^\top}^2 - \lambda^2 \|\tilde{z}^k\|_P^2 \right) \\
={}& \|z^k - z^\star\|_S^2 - \lambda \|P^{-1/2}S^{-1/2}(H+M^*)\tilde{z}^k\|_P^{-2} \|\tilde{z}^k\|_P^2 \left(\|\tilde{z}\|_{(2-\lambda)P - \frac{1}{2}VUV^\top}^2 \right) \\
\leq{}& \|z^k - z^\star\|_S^2 - \lambda \|P^{-1/2}S^{-1/2}(H+M^*)\|_P^{-2} \|\tilde{z}\|_{(2-\lambda)P - \frac{1}{2}VUV^\top}^2
\end{aligned}$$

Therefore, using standard arguments we conclude that if (3.35) holds then the sequence $(z^k)_{k\in\mathbb{N}}$ converges to a primal-dual solution. It is straightforward to see that Assumption 3.III is a restatement of (3.35) with P, U, V defined here. This completes the proof. □

3.4.1 Linear convergence

In this section we explore sufficient conditions for f, g, h and l under which Algorithm 3.1 and Algorithm 3.2 achieve linear convergence rates. In Lemma 3.8 and Lemma 3.11 we provide four regularity assumptions under which T defining the primal-dual optimality conditions is metrically subregular at $z^\star \in \operatorname{zer} T$ for 0. The linear convergence rate results are then deduced in Corollary 3.12.

Let us first recall the notion of quadratic growth: a proper closed convex function g is said to have *quadratic growth* at $\bar{x}$ for 0 with $0 \in \partial g(\bar{x})$ if there exists a neighborhood $\mathcal{U}$ of $\bar{x}$ such that

$$g(x) \geq \inf g + c \operatorname{dist}^2(x, \partial g^{-1}(0)), \quad \forall x \in \mathcal{U}. \tag{3.36}$$

Metric subregularity of the subdifferential operator and the quadratic growth condition are known to be equivalent [6, 63]. In particular, ∂g is metrically subregular at $\bar{x}$ for $\bar{u}$ with $\bar{u} \in \partial g(\bar{x})$ if and only if the quadratic growth condition (3.36) holds for $g(\cdot) - \langle \bar{u}, \cdot \rangle$, *i.e.*, there exists a positive constant c and a neighborhood $\mathcal{U}$ of $\bar{x}$ such that [6, Thm. 3.3]

$$g(x) \geq g(\bar{x}) + \langle \bar{u}, x - \bar{x} \rangle + c \operatorname{dist}^2(x, \partial g^{-1}(\bar{u})), \quad \forall x \in \mathcal{U}. \tag{3.37}$$

Strong subregularity has a similar characterization [6, Thm. 3.5]

$$g(x) \geq g(\bar{x}) + \langle \bar{u}, x - \bar{x} \rangle + c\|x - \bar{x}\|^2, \quad \forall x \in \mathcal{U}. \tag{3.38}$$

Next, let us define the following general growth condition.

Definition 3.7 (quadratic growth relative to a set)**.** *Consider a proper closed convex function g and a pair $(\bar{x}, \bar{u}) \in \operatorname{gra} \partial g$. We say that g has quadratic growth at $\bar{x}$ for $\bar{u}$ relative to a nonempty closed convex set X containing $\bar{x}$, if there exists a positive constant c and a neighborhood $\mathcal{U}$ of $\bar{x}$ such that*

$$g(x) \geq g(\bar{x}) + \langle \bar{u}, x - \bar{x} \rangle + c \operatorname{dist}^2(x, X), \quad \forall x \in \mathcal{U}. \tag{3.39}$$

From the above definition it is evident that metric subregularity and strong subregularity characterized in (3.37) and (3.38) are recovered when $X = \partial g^{-1}(\bar{u})$ and $X = \{\bar{x}\}$, respectively.

Another regularity assumption used in Lemma 3.8 is the notion of local strong convexity: a proper closed convex function g is said to be locally strongly convex in a neighborhood of $\bar{x}$, denoted by $\mathcal{U}$, if there exists a positive constant c such

that

$$g(x') \geq g(x) + \langle v, x' - x\rangle + \tfrac{c}{2}\|x' - x\|^2, \quad \forall x, x' \in \mathcal{U},\ \forall v \in \partial g(x).$$

Notice that local strong convexity in a neighborhood of $\bar{x}$ implies (3.38), but (3.38) is *much weaker* than local strong convexity since it holds only at $\bar{x}$ and only for $\bar{u} \in \partial g(\bar{x})$.

In the next lemma we provide three different regularity assumptions that are sufficient for metric subregularity of the operator defining the primal-dual optimality conditions. In Lemma 3.8*(i)* (or Lemma 3.8*(ii)*) we use local strong convexity, as well as the quadratic growth condition (3.39) relative to the set of primal solutions (or dual solutions). Interestingly, this regularity assumption does not entail a unique primal-dual solution.

Lemma 3.8. *Let Assumption 3.I hold. The operator T defining the primal-dual optimality conditions, cf. (3.6), is metrically subregular at $z^\star = (x^\star, u^\star)$ for 0 with $0 \in Tz^\star$ if one of the following assumptions holds:*

(i) *$f + g$ has quadratic growth at $x^\star$ for $-L^\top u^\star$ relative to the set of primal solutions $X^\star$, and $h^* + l^*$ is locally strongly convex in a neighborhood of $u^\star$. In this case the set of dual solutions is a singleton, $U^\star = \{u^\star\}$.*

(ii) *$f + g$ is locally strongly convex in a neighborhood of $x^\star$, and $h^* + l^*$ has quadratic growth at $u^\star$ for $Lx^\star$ relative to the set of dual solutions $U^\star$. In this case the set of primal solutions is a singleton, $X^\star = \{x^\star\}$.*

(iii) *$\nabla f + \partial g$ is strongly subregular at $x^\star$ for $-L^\top u^\star$ and $\partial h^* + \nabla l^*$ is strongly subregular at $u^\star$ for $Lx^\star$. In this case the set of primal-dual solutions is a singleton, $\operatorname{zer} T = \{(x^\star, u^\star)\}$.*

Proof. 3.8*(i)*- Consider the point $z^\star = (x^\star, u^\star)$. By definition of quadratic growth there exists a neighborhood $\mathcal{U}_{x^\star}$ and a positive constant c_1 such that

$$(f+g)(x) \geq (f+g)(x^\star) + \langle -L^\top u^\star, x - x^\star\rangle + c_1 \operatorname{dist}^2(x, X^\star), \quad \forall x \in \mathcal{U}_{x^\star}. \tag{3.40}$$

Let $\mathcal{U}_{u^\star}$ denote the neighborhood of $u^\star$ in which the local strong convexity of $h^* + l^*$ holds. Fix a point $z = (x, u) \in \mathcal{Z}_{z^\star} := \mathcal{U}_{x^\star} \times \mathcal{U}_{u^\star}$. Now take $v = (v_1, v_2) \in Tz = Az + Mz + Cz$, *i.e.*,

$$\begin{cases} v_1 \in \partial g(x) + \nabla f(x) + L^\top u, \\ v_2 \in \partial h^*(u) + \nabla l^*(u) - Lx. \end{cases} \tag{3.41}$$

Let $z_0 = (x_0, u_0)$ denote the projection of z onto the set of solutions, $\operatorname{zer} T$. The subgradient inequality for $f + g$ at x using (3.41) gives

$$\langle v_1, x - x_0 \rangle \geq (f + g)(x) - (f + g)(x_0) + \langle L^\top u, x - x_0 \rangle. \tag{3.42}$$

Noting that $0 \in Tz_0$, by the subgradient inequality for $h^* + l^*$ at u_0 we have

$$(h^* + l^*)(u) \geq (h^* + l^*)(u_0) + \langle Lx_0, u - u_0 \rangle. \tag{3.43}$$

Summing (3.42) and (3.43) yields

$$\langle v_1, x - x_0 \rangle \geq \mathcal{L}(x, u) - \mathcal{L}(x_0, u_0) = \mathcal{L}(x, u) - \mathcal{L}(x^\star, u^\star), \tag{3.44}$$

where $\mathcal{L}$ is the Lagrangian defined in (3.3). By local strong convexity of $h^* + l^*$ at $u \in \mathcal{U}_{u^\star}$ (for some strong convexity parameter c_2):

$$(h^* + l^*)(u^\star) \geq (h^* + l^*)(u) + \langle v_2 + Lx, u^\star - u \rangle + \tfrac{c_2}{2} \|u^\star - u\|^2.$$

Sum this inequality with (3.40) to obtain that for $z \in \mathcal{Z}_{z^\star}$

$$\mathcal{L}(x, u) - \mathcal{L}(x^\star, u^\star) \geq c_1 \operatorname{dist}^2(x, X^\star) + \langle v_2, u^\star - u \rangle + \tfrac{c_2}{2} \|u^\star - u\|^2. \tag{3.45}$$

It follows from (3.44) and (3.45) that

$$\begin{aligned} \langle v_2, u - u^\star \rangle + \langle v_1, x - x_0 \rangle &\geq \tfrac{c_2}{2} \|u^\star - u\|^2 + c_1 \operatorname{dist}^2(x, X^\star) \\ &= \tfrac{c_2}{2} \|u^\star - u\|^2 + c_1 \|x - x_0\|^2 \\ &\geq c\big(\|u^\star - u\|^2 + \|x - x_0\|^2\big), \end{aligned} \tag{3.46}$$

where $c = \min\{c_1, \frac{c_2}{2}\}$. By the Cauchy-Schwarz inequality

$$\langle v_1, x - x_0 \rangle + \langle v_2, u - u^\star \rangle \leq \|v\| \big(\|u - u^\star\|^2 + \|x - x_0\|^2\big)^{\frac{1}{2}}. \tag{3.47}$$

Combining (3.46) and (3.47) yields

$$\|v\| \geq c\big(\|u - u^\star\|^2 + \|x - x_0\|^2\big)^{\frac{1}{2}} \geq c\|z - z_0\| = cd(z, T^{-1}0).$$

Since $v \in Tz$ was selected arbitrarily we have that

$$d(z, T^{-1}0) \leq \tfrac{1}{c} d(Tz, 0), \quad \forall z \in \mathcal{Z}_{z^\star}.$$

This completes the first claim. Next, consider $\bar{z}^\star = (\bar{x}^\star, \bar{u}^\star) \in \operatorname{zer} T$ such that $\bar{z}^\star \in \mathcal{Z}_{z^\star}$. Setting $z = \bar{z}^\star$ in (3.45) yields

$$0 = \mathcal{L}(\bar{x}^\star, \bar{u}^\star) - \mathcal{L}(x^\star, u^\star) \geq \tfrac{c_2}{2}\|u^\star - \bar{u}^\star\|^2.$$

Therefore, $\bar{u}^\star = u^\star$ and since $\operatorname{zer} T$ is convex we conclude that $U^\star = \{u^\star\}$.

3.8*(ii)*- The proof of the second part is similar to part 3.8*(i)*. Therefore, we outline the proof. Let $\mathcal{U}_{x^\star}, \mathcal{U}_{u^\star}$ be the neighborhoods in the definition of local strong convexity of $f + g$ and quadratic growth of $h^* + l^*$, respectively. Fix $z \in \mathcal{Z}_{z^\star} = \mathcal{U}_{x^\star} \times \mathcal{U}_{u^\star}$. Take $v = (v_1, v_2)$ as in (3.41), and let z_0 denote the projection of z onto $\operatorname{zer} T$. In contrast to the previous part, sum the subgradient inequalities for $f + g$ at x_0 and for $h^* + l^*$ at u to derive

$$\langle v_2, u - u_0\rangle \geq \mathcal{L}(x_0, u_0) - \mathcal{L}(x, u) = \mathcal{L}(x^\star, u^\star) - \mathcal{L}(x, u). \tag{3.48}$$

Use local strong convexity of $f + g$ at x, and the quadratic growth of $h^* + l^*$ at $u^\star$ for $Lx^\star$ relative to $U^\star$ to derive

$$\mathcal{L}(x^\star, u^\star) - \mathcal{L}(x, u) \geq c_1 d(u, U^\star) + \langle v_1, x^\star - x\rangle + \tfrac{c_2}{2}\|x - x^\star\|^2.$$

Combining this with (3.48) and arguing as in the previous part completes the proof.

3.8*(iii)*- The proof of this part is slightly different. Let $\mathcal{U}_{x^\star}, \mathcal{U}_{u^\star}$ be the neighborhoods in the definitions of the two strong subregularity assumptions. Fix $z \in \mathcal{Z}_{z^\star} = \mathcal{U}_{x^\star} \times \mathcal{U}_{u^\star}$. Sum the subgradient inequality for $f + g$ at x and for $h^* + l^*$ at u to derive

$$\langle v, z - z^\star\rangle = \langle v_2, u - u^\star\rangle + \langle v_1, x - x^\star\rangle \geq \mathcal{L}(x, u^\star) - \mathcal{L}(x^\star, u) \tag{3.49}$$

On the other hand by [6, Thm. 3.5], $f + g$ has quadratic growth at $x^\star$ for $-L^\top u^\star$ relative to $\{x^\star\}$ and $h^* + l^*$ has quadratic growth at $u^\star$ for $Lx^\star$ relative to $\{u^\star\}$. Summing the two yields

$$\mathcal{L}(x, u^\star) - \mathcal{L}(x^\star, u) \geq c_2\|u - u^\star\|^2 + c_1\|x - x^\star\|^2 \tag{3.50}$$

Combining this inequality with (3.49) and using the Cauchy-Schwartz inequality as in previous parts completes the proof. Uniqueness of the solution follows from (3.50) by setting $z = \bar{z}^\star \in \operatorname{zer} T$ such that $\bar{z}^\star \in \mathcal{Z}_{z^\star}$ and using the convexity of $\operatorname{zer} T$. □

The assumptions of Lemma 3.8 are much weaker than strong convexity and do not always imply a unique primal-dual solution. Here we present two simple examples for demonstration. Notice that in the next example the assumption of

Lemma 3.8*(i)* that $f + g$ has quadratic growth with respect to the set of primal solutions is equivalent to the metric subregularity assumption.

Example 3.9. Consider the problem

$$\underset{x\in\mathbb{R}}{\text{minimize}}\ \ g(x) + h(x) = \max\{1 - x, 0\} + \tfrac{x}{2}\min\{x, 0\}$$

The solution to this problem is not unique and any $x^\star \in [1, \infty)$ solves this problem. The dual problem is given by

$$\underset{u\in\mathbb{R}}{\text{minimize}}\ \ g^*(-u) + h^*(u), \tag{3.51}$$

where $g^*(u) = u + \delta_{[-1,0]}(u)$ and $h^*(u) = \frac{1}{2}u^2 + \delta_{\leq 0}(u)$. It is evident that the dual problem has the unique solution $u^\star = 0$. It is easy to verify that g has quadratic growth at all the points $x^\star \in [1, \infty)$ for 0 with respect to $X^\star = [1, \infty)$. Moreover, we have $\partial g^{-1}(0) = [1, \infty)$, *i.e.*, $X^\star = \partial g^{-1}(0)$. In other words in this case the assumption of Lemma 3.8*(i)* for g is equivalent to the metric subregularity of ∂g at $x^\star$ for 0. Notice that ∂g is not strongly subregular at any point in $[1, \infty)$ for 0. Furthermore, h^* is globally strongly convex given that ∇h is Lipschitz. Therefore, according to Lemma 3.8*(i)* one would expect a unique dual solution but not necessarily a unique primal solution, which is indeed the case. □

Example 3.10. Let $c \in [-1, 1]$ and consider

$$\underset{x\in\mathbb{R}}{\text{minimize}}\ \ g(x) + h(x) = |x| + cx.$$

When $c \in (-1, 1)$ the problem attains a unique minimum at $x^\star = 0$. When $c = 1$ (or $c = -1$) all $x^\star \in (-\infty, 0]$ (or $x^\star \in [0, \infty)$) solves the problem. The dual problem is given by (3.51) with $g^*(u) = \delta_{[-1,1]}(u)$ and $h^*(u) = \delta_{\{c\}}(u)$. The unique dual solution is $u^\star = c$. Furthermore, ∂h^* is strongly subregular at $u^\star = c$ for all $x^\star$ given that $x^\star \in \partial h^*(u^\star)$. It is easy to verify that ∂g is metrically subregular at $x^\star = 0$ for $u^\star \in [-1, 1]$ but is only strongly subregular at $x^\star = 0$ for $u^\star \in (-1, 1)$. Notice that $u^\star = c$, therefore by Lemma 3.8*(iii)* one would expect a unique primal-dual solution when $u^\star = c \in (-1, 1)$ which is indeed the case. □

A wide range of functions used in optimization applications belong to the class of *piecewise linear-quadratic* (PLQ) functions, Definition 1.4. Some notable examples include: affine functions, quadratic functions, indicators of polyhedral sets, polyhedral norms such as ℓ_1, and regularizers such as elastic net, Huber loss, hinge loss, and many more [144, 7]. For a proper closed convex PLQ function g, ∂g is piecewise polyhedral, and therefore metrically subregular at any z for any z' provided that $z' \in \partial g(z)$ [144, 61].

Lemma 3.11. *Let Assumption 3.I hold. In addition, assume that f, g, l, and h are piecewise linear-quadratic. Then, the operator T defining the primal-dual optimality conditions, cf. (3.6), is metrically subregular with the same constant at any z for any z' provided that $z' \in Tz$.*

Proof. The subdifferential ∂g, ∇f, ∂h^* and ∇l^* are piecewise polyhedral [144, Prop. 12.30, Thm. 11.14]. Therefore, A and C are piecewise polyhedral. Furthermore, the graph of M is polyhedral since M is linear. Therefore, the graph of $T = A + M + C$ is also piecewise polyhedral. The inverse of a piecewise polyhedral mapping is also piecewise polyhedral. Therefore by [61, Prop. 3H.1, 3H.3] the mapping T is metrically subregular at z for z' whenever $(z, z') \in \operatorname{gra} T$. □

In the next corollary linear convergence results based on the two previous lemmas and Theorem 3.5 are presented.

Corollary 3.12 (linear convergence for Algorithms 3.1 and 3.2)**.** *Let Assumption 3.I hold. Consider the sequence $(z^k)_{k\in\mathbb{N}} = (x^k, u^k)_{k\in\mathbb{N}}$ generated by Algorithm 3.1 (resp. Algorithm 3.2). Suppose that Assumption 3.II (resp. Assumption 3.III) holds. In addition, suppose that one of the following assumptions holds:*

(i) f, g, l, and h are piecewise linear-quadratic.

(ii) f, g, l, and h satisfy at least one of the conditions of Lemma 3.8 at every $z^\star \in \operatorname{zer} T$ (not necessarily the same condition at all the points).

Then, the sequence $(z^k)_{k\in\mathbb{N}}$ converges R-linearly to some $z^\star \in \operatorname{zer} T$, and $(\operatorname{dist}_S(z^k, \operatorname{zer} T))_{k\in\mathbb{N}}$ converges Q-linearly to zero.

3.5 Stepsize selection and simulations

In this section we present some rules of thumb for selecting the stepsizes of the algorithms introduced in Section 3.2. Note that in applications where function g (or h) is separable, *i.e.*, $g(x) = \sum_{i=1}^N g_i(x_i)$ with $x = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^n$, $x_i \in \mathbb{R}^{n_i}$, using a diagonal matrix $\Sigma = \operatorname{blkdiag}(\sigma_1 \mathrm{I}_{n_N}, \sigma_2 \mathrm{I}_{n_2}, \ldots, \sigma_N \mathrm{I}_{n_N})$ results in the following decomposition of the proximal mapping of g

$$\operatorname{prox}_g^{\Sigma^{-1}}(x) = \big(\operatorname{prox}_{\sigma_1 g_1}(x_1), \operatorname{prox}_{\sigma_2 g_2}(x_2), \ldots, \operatorname{prox}_{\sigma_N g_N}(x_N)\big),$$

and leads to larger stepsizes than if we had used $\Sigma = \sigma \mathrm{I}_n$ (this would effectively correspond to $\sigma = \min_i \sigma_i$). Below, after discussing stepsize selection and

practical aspects of termination criteria, numerical simulations are performed comparing the performance of the algorithms presented in this chapter.

Stepsize selection: For simplicity we discuss some rules of thumb for stepsize selection when $\Sigma = \sigma\mathrm{id}$, $\Gamma = \gamma\mathrm{id}$, $R, Q = \mathrm{id}$ and the relaxation parameter $\lambda = 1$. Similar rules can be obtained following the same reasoning for more general cases such as when diagonal stepsize matrices are used.

While the choice of the stepsizes is application dependent, the choices discussed below perform reasonably well in our experiments. These are also used as default parameters in the Julia package *ProximalAlgorithms*[1] for the primal-dual AFBA solver.

Consider the stepsize condition for the Vũ-Condat algorithm (see (3.9)):

$$\sigma\gamma\|L\|^2 < \left(1 - \tfrac{\gamma\beta_f}{2}\right)\left(1 - \tfrac{\sigma\beta_l}{2}\right). \tag{3.52}$$

This inequality is always satisfied for

$$\gamma = \frac{1}{\frac{\beta_f}{2} + \frac{\|L\|}{\nu}}, \quad \sigma = \frac{0.99}{\frac{\beta_l}{2} + \nu\|L\|}, \quad \nu \in (0, \infty). \tag{3.53}$$

In our experiments, setting

$$\nu = \begin{cases} \xi_2\|L\|/\beta_f & \text{if } \beta_f > \xi_1\beta_l \;\&\; \xi_1\beta_f > \|L\| \\ \beta_l/\xi_2\|L\| & \text{if } \beta_l > \xi_1\beta_f \;\&\; \xi_1\beta_l > \|L\| \\ 1 & \text{otherwise,} \end{cases} \tag{3.54}$$

with $\xi_1 = 5$ and $\xi_2 = 100$ seems to perform best. The idea here is that if β_f is much larger that β_l, it is best to select γ as large as possible. If β_l and β_f are of the same order, or if $\|L\|$ is much larger than both of them, then given the symmetry in (3.53), stepsizes γ and σ of the same order are used.

In the case of SPCA, the condition is as follows:

$$\gamma\sigma\|L\|^2 < (1 - \tfrac{\sigma\beta_l}{2}), \quad \gamma\beta_f < 2.$$

In the simulations we set $\gamma = 1.99/\beta_f$ and $\sigma = 0.99/\gamma\|L\|^2$ (since $\beta_l = 0$).

In the case of SDCA, PDCA, PPDCA when $\beta_l = 0$ (as is the case in the simulations) the stepsizes are selected as in (3.53) and (3.54), with $\|L\|$ replaced by $\eta\|L\|$ where $\eta = \theta^2 - 3\theta + 3$.

In the case of Algorithm 3.2, in order to satisfy Assumption 3.III we set the stepsizes as follows: if $\theta \neq 0$ then the stepsizes are selected as in (3.53) and

[1] https://github.com/kul-forbes/ProximalAlgorithms.jl

(3.54), with $\|L\|$ replaced by $\frac{\theta}{2}\|L\|$. If $\theta = 0$ then $\gamma = {}^{1.99}/_{\beta_f}$ and $\sigma = {}^{1.99}/_{\beta_l}$. Since in the example below $\beta_l = 0$, σ can be any positive number and is selected by trial for best performance.

Termination criteria: Another practical matter is to have a criterion for termination and performance comparisons of primal-dual algorithms. Consider $v = (v_1, v_2) \in Tz$, *i.e.*,

$$\begin{cases} v_1 \in \partial g(x) + \nabla f(x) + L^\top u, \\ v_2 \in \partial h^*(u) + \nabla l^*(u) - Lx. \end{cases}$$

This inclusion is satisfied with $v = 0$ at any primal-dual solution (x, u). Therefore, the norm of v is an appropriate measure of optimality. Consider the primal and dual updates in Algorithms 3.1 and 3.2 with $\Sigma = \sigma\mathrm{id}$ and $\Gamma = \gamma\mathrm{id}$. Using the definition of proximal mapping we have

$$\gamma^{-1}(x^k - \bar{x}^k) - L^\top(u^k - \bar{u}^k) + \nabla f(\bar{x}^k) - \nabla f(x^k) \in L^\top \bar{u}^k + \partial g(\bar{x}^k) + \nabla f(\bar{x}^k),$$

$$\sigma^{-1}(u^k - \bar{u}^k) + (1-\theta)L(x^k - \bar{x}^k) + \nabla l^*(\bar{u}^k) - \nabla l^*(u^k) \in \partial h^*(\bar{u}^k) + \nabla l^*(\bar{u}^k) - L\bar{x}^k.$$

Therefore, we use $\mathcal{O}(z^k, \bar{z}^k)$ given by

$$\begin{aligned} \mathcal{O}(z^k, \bar{z}^k) \coloneqq {} & \|\gamma^{-1}(x^k - \bar{x}^k) - L^\top(u^k - \bar{u}^k) + \nabla f(\bar{x}^k) - \nabla f(x^k)\|^2 \\ & + \|\sigma^{-1}(u^k - \bar{u}^k) + (1-\theta)L(x^k - \bar{x}^k) + \nabla l^*(\bar{u}^k) - \nabla l^*(u^k)\|^2 \end{aligned}$$

as a measure of optimality in the simulations. Note that computing $\mathcal{O}(z^k, \bar{z}^k)$ does not always require additional gradient evaluations. This is the case for several of the special cases discussed in Section 3.2.1 for which $x^{k+1} = \bar{x}^k$, such as SNCA with $\lambda = 1$, or SDCA, PDCA and dual of SPCA when $l = \delta_{\{0\}}$ and $\lambda = 1$. Although in other cases additional gradient evaluations may be unavoidable, $\mathcal{O}(z^k, \bar{z}^k)$ is used here for performance comparisons anyway. We also remark that in practice one may use the sum of residuals $\|\bar{x}^k - x^k\|/\gamma$ and $\|\bar{u}^k - u^k\|/\sigma$ as the termination criteria.

Simulations: We evaluate the performance of the algorithms with the above stepsizes on the *dual support vector machine* problem (1.3) with $C = {}^1/_{10}$. In our simulations we used two processed datasets, *iris* (first two classes) and *a1a* from the LIBSVM dataset [40].

We formulate the problem in the form of (3.1) by setting f equal to the quadratic cost, g the indicator of the box $[0, C]^N$, $h = \delta_{\{0\}}$ and $L = b^\top$ where $b = (b_1, \ldots, b_N)$. The results are presented in Figure 3.1 where we have only included some representative algorithms. Other variants have similar

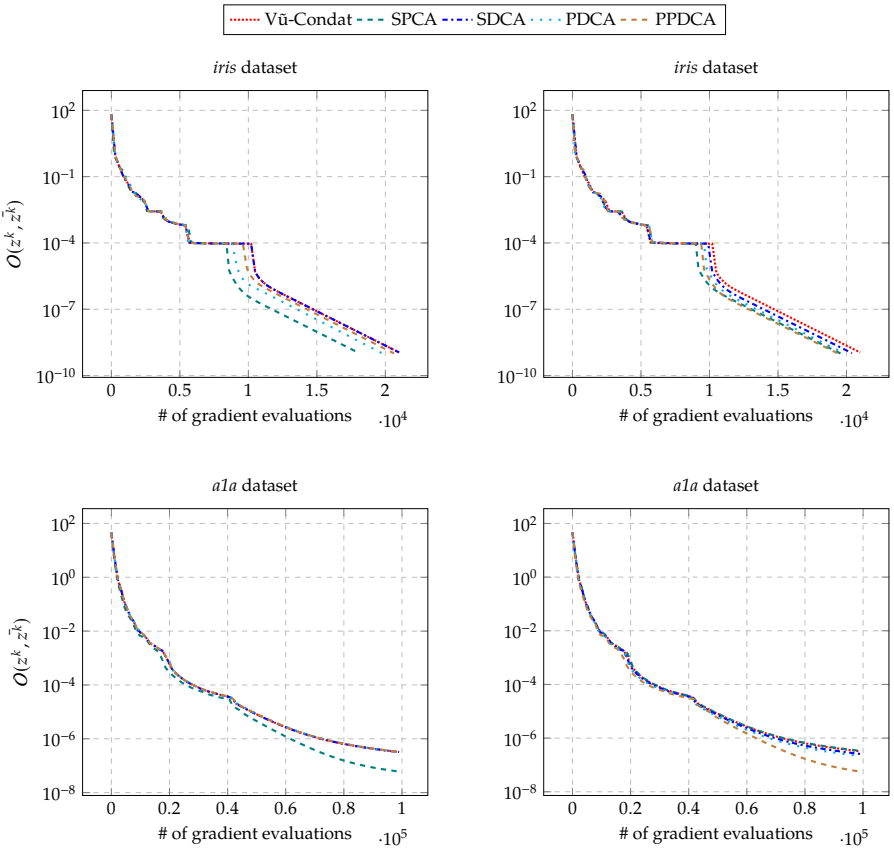

Figure 3.1: *Dual SVM on iris, a1a: (left column) variants of Algorithm 3.1 (right column) variants of Algorithm 3.2*

performances and are excluded for the sake of clarity of the plot. The stepsizes for SPCA can take larger values compared to other variants. Consequently, we often observe faster convergence for SPCA; however, it is observed that in general the algorithms have similar performance and the trajectories are very close to each other. Note that other than PPDCA the other five algorithms depicted require only two matrix-vector products excluding the gradient evaluation (PPDCA requires four). Moreover, for this problem the performance gain by using Algorithm 3.2 is not significant enough to justify the extra computation required for computing ν_k in (3.14).

Next, SPCA is compared with PD3O [180] discussed in Remark 3.1. We use the same formulation for both algorithms. The stepsize condition for PD3O is

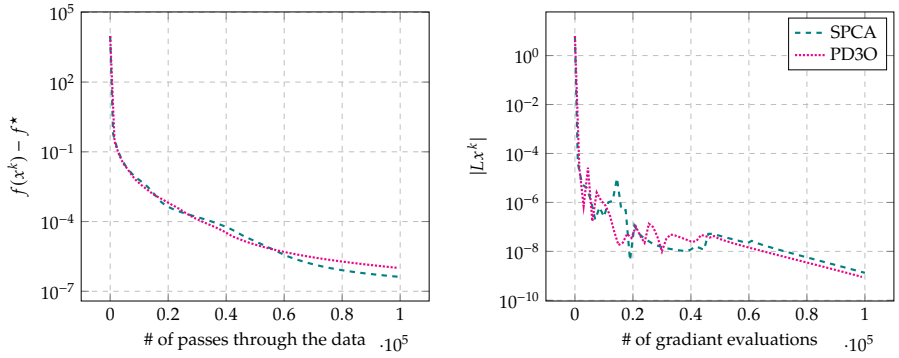

Figure 3.2: *Performance comparisons for SPCA and PD3O [180] for classifying a1a dataset*

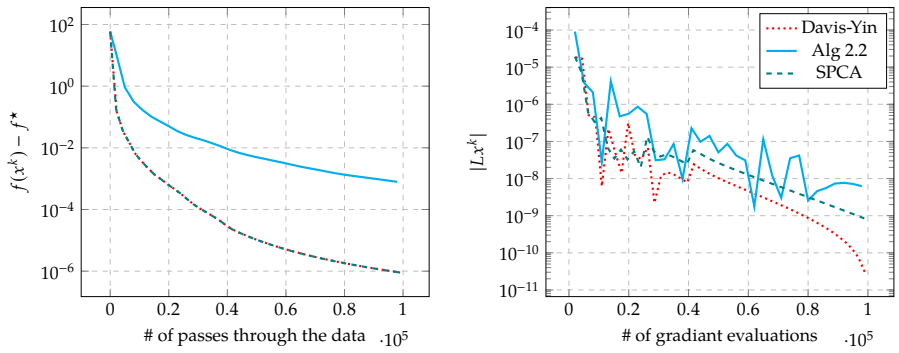

Figure 3.3: *Performance comparisons for SPCA, Alg. 2.2 and Davis-Yin algorithm for classifying a1a dataset*

identical to SPCA, and therefore we use the same rule of thumb described above for best performance. Note that the iterates are not necessarily feasible. In order to have a fair comparison, for both algorithms we use the primal variable that belongs to the domain of g, *i.e.*, x^{k+1} in (3.11) and $\bar{x}^{k+1}$ in (3.12). Figure 3.2 (left) shows the distance of the quadratic cost from the optimal cost and (right) plots $|Lx|$ (recall the constraint $Lx = 0$). It is observed that both algorithms have very similar performances.

In Chapter 2 a DRS-type algorithm was introduced, Algorithm 2.2, which in the optimization framework can be derived from Algorithm 3.1 with $\mu = 0$, by setting $l = \delta_{\{0\}}$, $L = \mathrm{id}$, and $\gamma = \sigma^{-1}$ after a change of variable. As noted in Remark 2.9, we compare its performance against the three term splitting of Davis and Yin [58]. For this set of simulations, problem (1.3) is

formulated as before with the difference that h is set to be the indicator of the set $\{x \in \mathbb{R}^N \mid \sum_{i=1}^N x_i b_i = 0\}$ (and $L = \mathrm{id}$). Note that the proximal mapping of h can be computed efficiently. The stepsizes of Algorithm 2.2 were selected according to Proposition 2.7 with $\theta = 1.5$, $\gamma = 0.5(4 - \theta^2)/\beta_f$ and ρ_k set to the maximum permitted value in (3.37). It is observed in Figure 3.3 that for this problem Algorithm 2.2 is slower, which is due to limiting the stepsizes to $\gamma = \sigma^{-1}$. We also plotted the result of SPCA with this formulation and observe that its trajectories are quite close to those of the Davis-Yin algorithm.

3.6 Conclusions

This section introduced a simple primal-dual framework for solving structured convex optimization problems. Depending on the value of two parameters, (extension of) known as well as many new primal-dual algorithms are obtained. Owing to this unified framework linear convergence rates were obtained for all the special cases of our framework provided that some mild regularity conditions for the cost functions are satisfied. Moreover, numerical simulations were performed on dual SVM where we discussed practical issues such as rules of thumb for stepsize selection and termination criteria.

Chapter 4

A randomized block-coordinate primal-dual proximal algorithm

This chapter is based on:

Latafat, P., Freris, N. M., and Patrinos, P. *A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization.* IEEE Transactions on Automatic Control 64, 10 (10 2019), 4050–4065.

4.1 Introduction

In this chapter we revisit the optimization problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x)+g(x)+h(Lx), \tag{4.1}$$

where L is a linear mapping, h and g are proper, closed, convex functions (possibly nonsmooth), and f is convex, continuously differentiable with Lipschitz continuous gradient. We further assume that the *proximal mappings* associated with h and g are efficiently computable [49]. This setup is quite general and captures a wide range of applications in signal processing, machine learning and control.

In problem (4.1), it is typically assumed that the gradient of the smooth term f is β_f-Lipschitz for some nonnegative constant β_f. We consider Lipschitz continuity of ∇f with respect to $\|\cdot\|_Q$ with $Q \succ 0$ in place of the canonical norm (*cf.* (4.2)). This is because in many applications of practical interest, a scalar Lipschitz constant fails to accurately capture the Lipschitz continuity of ∇f. A prominent example lies in distributed optimization, where f is separable,

i.e., $f(x) = \sum_{i=1}^{m} f_i(x_i)$. In this case, the metric Q is taken block-diagonal with blocks containing the Lipschitz constants of the ∇f_i's. Notice that in such settings considering a scalar Lipschitz constant results in using the largest of the Lipschitz constants, which leads to conservative stepsize selection and consequently slower convergence rates.

This chapter presents a primal-dual algorithm TriPD (Alg. 4.1), that consists of two proximal evaluations (corresponding to the two nonsmooth terms g and h), one gradient evaluation (for the smooth term f), and one correction step (*cf.* Alg. 4.1). We adopt the general Lipschitz continuity assumption (4.2) in our convergence analysis, which is essential for avoiding conservative stepsize conditions that depend on the *global* scalar Lipschitz constant.

In Section 4.2, it is shown that the sequence generated by TriPD (Alg. 4.1) is S-Fejér monotone (with respect to the set of primal-dual solutions), where S is a block diagonal positive definite matrix. This key property is exploited in Section 4.3 to develop a block-coordinate version of the algorithm with a general randomized activation scheme.

The connections of TriPD to other related primal-dual algorithms in the literature are discussed in Section 4.2.1. Most notably, TriPD transforms into the algorithm of Vũ and Condat [170, 53] by a change of order and elimination of the dual variable (refer to Section 4.2.1). Convergence of the Vũ-Condat scheme was established independently by Vũ [170] and Condat [53], by casting it in the form of the forward-backward splitting. In the original analysis a scalar constant is used to capture the Lipschitz continuity of the gradient of f, thus resulting in potentially smaller stepsizes (and slower convergence in practice). In [47], the authors assume the more general Lipschitz continuity property (4.2) by using a preconditioned variable metric forward-backward iteration. Nevertheless, the stepsize matrix is restricted to be proportional to Q^{-1}.

Block-coordinate (BC) minimization is a simple approach for tackling large-scale optimization problems. At each iteration, a subset of the coordinates is updated while others are held fixed. *Randomized* BC algorithms are of particular interest, and can be divided into two main categories:

Type a) comprises algorithms in which *only one* coordinate is randomly activated and updated at each iteration. The BC versions of gradient [125] and proximal gradient methods [139] belong in this category. A distinctive attribute of the aforementioned algorithms is the fact that the stepsizes are selected to be inversely proportional to the *coordinate-wise* Lipschitz constant of the smooth term rather than the global one. This results in applying larger stepsizes in directions with smaller Lipschitz constant, and therefore leads to faster convergence.

Type b) contains methods where *more than one* coordinate may be randomly activated and simultaneously updated [22, 51]. Note that this class may also capture the single active coordinate (type a) as a special case. The convergence condition for this class of BC algorithms is typically the same as in the full algorithm. In [22, 51] random BC is applied to α-averaged operators by establishing stochastic Fejér monotonicity, while [51] also considers quasi-nonexpansive operators. In [134, 22] the authors obtain randomized BC algorithms based on the primal-dual scheme of Vũ and Condat; the main drawback is that, just as in the full version of these algorithms, the use of conservative stepsize conditions leads to slower convergence in practice.

The BC version of TriPD (Alg. 4.1) falls into the second class, *i.e.*, it allows for a general randomized activation scheme (*cf.* Alg. 4.2). The proposed scheme converges under the same stepsize conditions as the full algorithm. As a consequence, in view of the characterization of Lipschitz continuity of ∇f in (4.2), when f is separable, *i.e.*, $f(x) = \sum_{i=1}^{m} f_i(x_i)$, our approach leads to algorithms that depend on the *local* Lipschitz constants (of ∇f_i's) rather than the global constant, thus assimilating the benefits of both categories. Notice that when f is separable, the coordinate-wise Lipschitz continuity assumption of [125, 139, 70] is equivalent to (4.2) with $\beta_f = 1$ and $Q = \mathrm{blkdiag}(\beta_1 I_{n_1}, \ldots, \beta_m I_{n_m})$, where m denotes the number of coordinate blocks, n_i denotes the dimension of the i-th coordinate block, and β_i denotes the Lipschitz constant of f_i. In the general setting, [125, Lem. 2] can be invoked to establish the connection between the metric Q and the coordinate-wise Lipschitz assumption. However, in many cases (most notably the separable case) this lemma is conservative.

In [70], the authors propose a randomized BC version of the Vũ-Condat scheme. Their analysis does not require the cost functions to be separable and utilizes a different Lyapunov function for establishing convergence. Notice that the block-coordinate scheme of [70] updates a single coordinate at every iteration (*i.e.*, it is a type a) algorithm) as opposed to the more general random sweeping of the coordinates. Additionally, in the case of f being separable, our proposed method (*cf.* Alg. 4.2) assigns a block stepsize that is inversely proportional to $\frac{\beta_i}{2}$ (where β_i denotes the Lipschitz constant for f_i), in place of β_i required by [70, Assumption 1(e)]: larger stepsizes are typically associated with faster convergence in primal-dual proximal algorithms.

In Section 3.4.1 linear convergence rates were obtain for a large class of primal-dual algorithms (special cases of Algorithms 3.1 and 3.2). In Section 4.4 we further explicate the required conditions for TriPD in terms of the objective functions, with two special cases of prevalent interest: a) when f, g and h satisfy a *quadratic growth* condition (*cf.* Lem. 4.8) (which is *much weaker than strong convexity*) or b) when f, g and h are *piecewise linear-quadratic* (*cf.* Lem. 4.9), a common scenario in many applications such as LPs, QPs, SVM and fitting

problems for a wide range of regularization functions; *e.g.* ℓ_1 norm, elastic nets, Huber loss and many more. Specifically, for the BC version of *TriPD*, linear conveergence rate is established under slightly stronger conditions (*cf.* Thm. 4.11).

Finally, as an important application, we consider a distributed structured optimization problem over a network of agents. In this context, each agent has its own private cost function of the form (4.1), while the communication among agents is captured by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$:

$$\begin{aligned} \underset{x_1,\dots,x_m}{\text{minimize}} \quad & \sum_{i=1}^{m} f_i(x_i) + g_i(x_i) + h_i(L_i x_i) \\ \text{subject to} \quad & A_{ij}x_i + A_{ji}x_j = b_{(i,j)} \qquad (i,j) \in \mathcal{E}. \end{aligned}$$

We use (i, j) to denote the unordered pair of agents i, j, and ij to denote the ordered pair. The goal is to solve the global optimization problem through local exchange of information. Notice that the linear constraints on the edges of the graph prescribe relations between neighboring agents' variables. This type of edge constraints was also considered in [182]. It is worthwhile noting that for the special case of two agents $i = 1, 2$, with $f_i, h_i \equiv 0$, one recovers the setup for the celebrated *alternating direction method of multipliers* (ADMM) algorithm. Another special case of particular interest is *consensus optimization*, when $A_{ij} = I$, $A_{ji} = -I$ and $b_{(i,j)} = 0$. A primal-dual algorithm for consensus optimization was introduced in [100] for the case of $f_i \equiv 0$, where a transformation was used to replace the edge variables with node variables.

This multi-agent optimization problem arises in many contexts such as sensor networks, power systems, transportation networks, robotics, water networks, distributed data-sharing, etc. [28, 89, 136]. In most of these applications, there are computation, communication and/or physical limitations on the system that render centralized management infeasible. This motivates the *fully* distributed synchronous and asynchronous algorithms developed in Section 4.5. Both versions are fully distributed in the sense that not only the iterations are performed locally, but also the stepsizes of each agent are selected based on local information without any prior global coordination (*cf.* Assumption 4.VI). The asynchronous variant of the algorithm is based on an instance of the randomized block-coordinate algorithm in Section 4.3. The protocol is as follows: at each iteration, a) agents are activated at random, and independently from one another, b) active agents perform local updates, c) they communicate the required updated values to their neighbors and d) return to an idle state.

4.2 Triangularly preconditioned primal-dual algorithm

In this section we present a primal-dual algorithm for problem (4.1). We adhere to the following assumptions throughout sections 4.2 to 4.4:

Assumption 4.I.

(i) $g : \mathbb{R}^n \to \overline{\mathbb{R}}$, $h : \mathbb{R}^r \to \overline{\mathbb{R}}$ are proper, closed, convex functions, and $L : \mathbb{R}^n \to \mathbb{R}^r$ is a linear mapping.

(ii) $f : \mathbb{R}^n \to \mathbb{R}$ is convex, continuously differentiable, and for some $\beta_f \in [0, \infty)$, ∇f is β_f-Lipschitz continuous with respect to the metric induced by $Q \succ 0$, *i.e.*,

$$\|\nabla f(x) - \nabla f(y)\|_{Q^{-1}} \leq \beta_f \|x - y\|_Q \quad \forall x, y \in \mathbb{R}^n. \tag{4.2}$$

(iii) The set of solutions to (4.1) is nonempty. Moreover, there exists $x \in \operatorname{ri} \operatorname{dom} g$ such that $Lx \in \operatorname{ri} \operatorname{dom} h$.

In Assumption 4.I*(ii)*, the constant $\beta_f \geq 0$ is not absorbed into the metric Q in order to also incorporate the case when ∇f is a constant (by setting $\beta_f = 0$).

The dual problem is to

$$\underset{u \in \mathbb{R}^r}{\text{minimize}} \; (g + f)^*(-L^\top u) + h^*(u). \tag{4.3}$$

With a slight abuse of terminology, we say that $(u^\star, x^\star)$ is a *primal-dual solution* (in place of dual-primal) if $u^\star$ solves the dual problem (4.3) and $x^\star$ solves the primal problem (4.1). We denote the set of primal-dual solutions by $\mathcal{S}$. Assumption 4.I*(iii)* guarantees that the set of solutions to the dual problem is nonempty and the duality gap is zero [143, Corollary 31.2.1]. Furthermore, the pair $(u^\star, x^\star)$ is a primal-dual solution if and only if it satisfies:

$$\begin{cases} 0 \in \partial h^*(u) - Lx, \\ 0 \in \partial g(x) + \nabla f(x) + L^\top u. \end{cases} \tag{4.4}$$

We proceed to present the new primal-dual scheme TriPD (Alg. 4.1). The motivation behind the name becomes apparent in the sequel after equation (4.12). The algorithm involves two proximal evaluations (respective to the nonsmooth terms g, h), and one gradient evaluation (for the Lipschitz differentiable term f). The stepsizes in TriPD (Alg. 4.1) are chosen so as to satisfy the following assumption:

Assumption 4.II (stepsize selection)**.** Both the dual stepsize matrix $\Sigma \in \mathbb{R}^{r\times r}$, and the primal stepsize matrix $\Gamma \in \mathbb{R}^{n\times n}$ are symmetric positive definite. In addition, they satisfy:

$$\Gamma^{-1} - \tfrac{\beta_f}{2} Q - L^\top \Sigma L \succ 0. \tag{4.5}$$

Selecting scalar primal and dual stepsizes, along with the standard definition of Lipschitz continuity, as is prevalent in the literature [170, 53], can plainly be treated by setting $\Sigma = \sigma I_r$, $\Gamma = \gamma I_n$, and $Q = I_n$, whence from (4.5) we require that

$$\gamma < \frac{1}{\frac{\beta_f}{2} + \sigma \|L\|^2}.$$

Algorithm 4.1 **Tri**angularly Preconditioned **P**rimal-**D**ual algorithm (TriPD)

Inputs: $x^0 \in \mathbb{R}^n$, $u^0 \in \mathbb{R}^r$

for $k = 0, 1, \ldots$ **do**

$\bar{u}^k = \operatorname{prox}_{h^*}^{\Sigma^{-1}}(u^k + \Sigma L x^k)$

$x^{k+1} = \operatorname{prox}_{g}^{\Gamma^{-1}}(x^k - \Gamma \nabla f(x^k) - \Gamma L^\top \bar{u}^k)$

$u^{k+1} = \bar{u}^k + \Sigma L (x^{k+1} - x^k)$

Remark 4.1. Each iteration of TriPD (Alg. 4.1) requires one application of L and one of $L^\top$ (even though it appears to require two applications of L). The reason is that, at iteration k, only $L^\top \bar{u}^k$, Lx^{k+1} need to be evaluated since $L(x^{k+1} - x^k) = Lx^{k+1} - Lx^k$ and Lx^k was computed during the previous iteration. □

TriPD (Alg. 4.1) can be compactly written as:

$$z^{k+1} = T z^k,$$

where $z^k := (u^k, x^k)$, and the operator T is given by:

$$\bar{u} = \operatorname{prox}_{h^*}^{\Sigma^{-1}}(u + \Sigma L x) \tag{4.6a}$$

$$\bar{x} = \operatorname{prox}_{g}^{\Gamma^{-1}}(x - \Gamma \nabla f(x) - \Gamma L^\top \bar{u}) \tag{4.6b}$$

$$Tz = (\bar{u} + \Sigma L(\bar{x} - x), \bar{x}). \tag{4.6c}$$

Remark 4.2 (relaxed iterations)**.** It is also possible to devise a *relaxed* version of TriPD (Alg. 4.1) as follows:

$$z^{k+1} = z^k + \Lambda(Tz^k - z^k),$$

where Λ is a positive definite matrix and $\Lambda \prec 2I_{n+r}$. For ease of exposition, we present the convergence analysis for the original version (*i.e.*, for $\Lambda = I_{n+r}$). Note that the analysis carries through with minor modifications for relaxed iterations. □

For compactness of exposition, we define the following operators:

$$A : (u, x) \mapsto (\partial h^*(u), \partial g(x)), \tag{4.7a}$$

$$M : (u, x) \mapsto (-Lx, L^\top u), \tag{4.7b}$$

$$C : (u, x) \mapsto (0, \nabla f(x)). \tag{4.7c}$$

The optimality condition (4.4) can then be written in the equivalent form of the *monotone inclusion*:

$$0 \in Az + Mz + Cz =: Fz, \tag{4.8}$$

where $z = (u, x)$. Observe that the linear operator M is monotone since it is skew-symmetric, *i.e.*, $M^\top = -M$. It is also easy to verify that the operator A is maximally monotone [13, Thm. 21.2 and Prop. 20.23], while operator C is cocoercive, being the gradient of $\tilde{f}(u, x) = f(x)$, and in light of Assumption 4.I*(ii)* and [13, Cor. 18.17].

We further define

$$P = \begin{pmatrix} \Sigma^{-1} & \frac{1}{2}L \\ \frac{1}{2}L^\top & \Gamma^{-1} \end{pmatrix}, \quad K = \begin{pmatrix} 0 & -\frac{1}{2}L \\ \frac{1}{2}L^\top & 0 \end{pmatrix}, \tag{4.9}$$

and set $H = P + K$. It is plain to check that condition (4.5) implies that the symmetric matrix P is positive definite (by a standard Schur complement argument). In addition, we set

$$S = \mathrm{blkdiag}(\Sigma^{-1}, \Gamma^{-1}). \tag{4.10}$$

Using these definitions, the operator T defined in (4.6) can be written as:

$$Tz := z + S^{-1}(H + M^\top)(\bar{z} - z), \tag{4.11}$$

where

$$\bar{z} = (H + A)^{-1}(H - M - C)z. \tag{4.12}$$

This compact representation simplifies the convergence analysis. A key consideration for choosing P and K as in (4.9) is to ensure that $H = P + K$ is lower block-triangular. Notice that when $M \equiv 0$, (4.11) can be viewed as a *triangularly preconditioned* forward-backward update, followed by a correction step. This motivates the name *TriPD*: **Tri**angularly Preconditioned **P**rimal-**D**ual algorithm. Due to the triangular structure of H, the backward step $(H + A)^{-1}$ in (4.12) can be carried out sequentially: an updated dual vector $\bar{u}$ is computed (through proximal mapping) using (u, x) and, subsequently, the primal vector $\bar{x}$ is computed using $\bar{u}$ and x, *cf.* (4.6). Furthermore, it follows from (4.11) that this choice makes $H + M^\top$ upper block-triangular which, alongside the diagonal structure of S, yields the efficiently computable update (4.6c) in view of:

$$S^{-1}(H + M^\top) = \begin{pmatrix} I & \Sigma L \\ 0 & I \end{pmatrix}. \tag{4.13}$$

Remark 4.3. Note that potentially larger stepsize parameter as in Section 3.2.2 can be considered here. Although this can potentially improve the rate of convergence in practice, we opt not to: the reason is that in the context of multi-agent optimization (that we especially target in this chapter) such design choice would require global coordination, that is contradictory to our objective of devising distributed algorithms. □

We proceed by showing that the set of primal-dual solutions coincides with the set of fixed points of T, $\operatorname{fix} T$:

$$\mathcal{S} = \{z \mid 0 \in Az + Mz + Cz\} = \operatorname{fix} T. \tag{4.14}$$

To see this note that from (4.11) and (4.12) we have:

$$\begin{aligned} z \in \operatorname{fix} T &\iff z = Tz \iff \bar{z} = z \\ &\iff (H + A)^{-1}(H - M - C)z = z \\ &\iff Hz - Mz - Cz \in Hz + Az \iff z \in \mathcal{S}, \end{aligned}$$

where in the second equivalence we used the fact that S is positive definite and $\langle (H + M^\top)z, z\rangle \geq \|z\|_P^2$ for all $z \in \mathbb{R}^{n+r}$ (since K is skew-adjoint and M is monotone).

Next, let us define

$$\tilde{P} \coloneqq \begin{pmatrix} \Sigma^{-1} & -\frac{1}{2}L \\ -\frac{1}{2}L^\top & \Gamma^{-1} - \frac{\beta_f}{4}Q \end{pmatrix}. \tag{4.15}$$

Observe that (from Schur complement) Assumption 4.II is necessary and sufficient for $2\tilde{P} - S$ to be symmetric positive definite (*cf.* the convergence result in Thm. 4.5). In particular, $\tilde{P}$ is positive definite since S is positive definite.

The next lemma establishes the key property of the operator T that is instrumental in our convergence analysis:

Lemma 4.4. *Let Assumptions 4.I and 4.II hold. Consider the operator T in (4.6) (equivalently (4.11)). Then, for any $z^\star \in \mathcal{S}$ and any $z \in \mathbb{R}^{n+r}$ we have*

$$\|Tz - z\|_{\tilde{P}}^2 \leq \langle z - z^\star, z - Tz \rangle_S. \tag{4.16}$$

Proof. Consider the operator T as in (4.11). By monotonicity of A at $z^\star$ and $\bar{z}$ along with (4.12) we have

$$0 \leq \langle -Mz^\star - Cz^\star + Mz + Cz - Hz + H\bar{z}, z^\star - \bar{z} \rangle. \tag{4.17}$$

For $\beta_f > 0$, assumption 4.I*(ii)* is equivalent to ∇f being *cocoercive* [13, Thm. 18.17], *i.e.*, for all $x, y \in \mathbb{R}^n$:

$$\tfrac{1}{\beta_f}\|\nabla f(x) - \nabla f(y)\|_{Q^{-1}}^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle. \tag{4.18}$$

On the other hand, for $\beta_f > 0$ we have

$$\begin{aligned}
\langle Cz - Cz^\star, z^\star - \bar{z} \rangle &= \langle \nabla f(x) - \nabla f(x^\star), x^\star - \bar{x} \rangle \\
&= \langle \nabla f(x) - \nabla f(x^\star), x - \bar{x} \rangle + \langle \nabla f(x) - \nabla f(x^\star), x^\star - x \rangle \\
&\leq \tfrac{1}{\beta_f}\|\nabla f(x) - \nabla f(x^\star)\|_{Q^{-1}}^2 + \tfrac{\beta_f}{4}\|x - \bar{x}\|_Q^2 + \langle \nabla f(x) - \nabla f(x^\star), x^\star - x \rangle \\
&\leq \langle \nabla f(x) - \nabla f(x^\star), x - x^\star \rangle + \tfrac{\beta_f}{4}\|x - \bar{x}\|_Q^2 + \langle \nabla f(x) - \nabla f(x^\star), x^\star - x \rangle, \\
&= \tfrac{\beta_f}{4}\|x - \bar{x}\|_Q^2,
\end{aligned} \tag{4.19}$$

where we have used (1.8) (with $V = \frac{2}{\beta_f}Q^{-1}$) in the first inequality, and (4.18) in the second inequality, respectively. Notice that if $\beta_f = 0$ then inequality (4.19) holds trivially with equality.

Using (4.19) in (4.17), along with skew-symmetry of K and M, we have

$$\begin{aligned}
0 \leq& \langle -Mz^\star - Cz^\star + Mz + Cz - Hz + H\bar{z}, z^\star - \bar{z} \rangle \\
\leq& \langle (M - K)(z - z^\star) + P(\bar{z} - z), z^\star - \bar{z} \rangle + \tfrac{\beta_f}{4}\|x - \bar{x}\|_Q^2
\end{aligned}$$

$$
\begin{aligned}
=&\langle (M-K)(z-z^\star)+P(\bar{z}-z), z^\star-z\rangle + \tfrac{\beta_f}{4}\|x-\bar{x}\|_Q^2 \\
&+\langle (M-K)(z-z^\star)+P(\bar{z}-z), z-\bar{z}\rangle \\
=&\langle P(\bar{z}-z), z^\star-z\rangle + \tfrac{\beta_f}{4}\|x-\bar{x}\|_Q^2 - \|\bar{z}-z\|_P^2 + \langle (M-K)(z-z^\star), z-\bar{z}\rangle \\
=&\langle z-z^\star, (H+M^\top)(z-\bar{z})\rangle + \tfrac{\beta_f}{4}\|x-\bar{x}\|_Q^2 - \|\bar{z}-z\|_P^2. \qquad (4.20)
\end{aligned}
$$

By definition, $S^{-1}(H+M^\top)(\bar{z}-z) = Tz - z$. Thus

$$
\langle z-z^\star, (H+M^\top)(z-\bar{z})\rangle = \langle z-z^\star, z-Tz\rangle_S. \qquad (4.21)
$$

On the other hand, we have $\bar{z}-z = (H+M^\top)^{-1}S(Tz-z)$. Using (4.9), (4.13) and (4.6c) we conclude

$$
\|\bar{z}-z\|_P^2 - \tfrac{\beta_f}{4}\|\bar{x}-x\|_Q^2 = \|Tz-z\|_{\tilde{P}}^2, \qquad (4.22)
$$

where $\tilde{P}$ is defined in (4.15). Combining (4.20), (4.21) and (4.22) completes the proof. □

The main convergence result for TriPD (Alg. 4.1) is established in the next theorem. In specific, it is shown that the generated sequence is S-Fejér monotone. We emphasize that the diagonal structure of S is the key property used in developing the block-coordinate version of the algorithm in Section 4.3.

Theorem 4.5. *Let Assumptions 4.I and 4.II hold. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by TriPD (Alg. 4.1). The following Fejér-type inequality holds for all $z^\star \in \mathcal{S}$:*

$$
\|z^{k+1}-z^\star\|_S^2 \le \|z^k-z^\star\|_S^2 - \|z^{k+1}-z^k\|_{2\tilde{P}-S}^2. \qquad (4.23)
$$

Consequently, $(z^k)_{k\in\mathbb{N}}$ converges to some $z^\star \in \mathcal{S}$.

Proof. We establish convergence by showing that the sequence $(z^k)_{k\in\mathbb{N}}$ is S-Fejér monotone with respect to $\mathcal{S} = \operatorname{fix} T$. We have

$$
\begin{aligned}
\|z^{k+1}-z^\star\|_S^2 =& \|Tz^k - z^k + z^k - z^\star\|_S^2 \\
=& \|z^k-z^\star\|_S^2 + \|Tz^k-z^k\|_S^2 + 2\langle z^k-z^\star, Tz^k-z^k\rangle_S \\
\le& \|z^k-z^\star\|_S^2 - \|Tz^k-z^k\|_{2\tilde{P}-S}^2, \qquad (4.24)
\end{aligned}
$$

where the inequality follows from Lemma 4.4. Note that $2\tilde{P} - S$ is symmetric positive-definite if and only if assumption 4.II holds. Therefore, by (4.24) the sequence $(z^k)_{k\in\mathbb{N}}$ is Fejér monotone in the space equipped with inner product $\langle\cdot,\cdot\rangle_S$; in particular, $(z^k)_{k\in\mathbb{N}}$ is bounded. Furthermore, it follows from (4.24) and the fact that $2\tilde{P} - S$ is positive-definite that

$$\|Tz^k - z^k\| \to 0. \tag{4.25}$$

The operator T is continuous (since it involves proximal and linear mappings that are continuous, and since ∇f is assumed continuous). Let z^c be a cluster point of $(z^k)_{k\in\mathbb{N}}$. It follows from the continuity of T and (4.25) that $Tz^c - z^c = 0$, *i.e.*, $z^c \in \operatorname{fix} T$. The result follows from Fejér monotonicity of $(z^k)_{k\in\mathbb{N}}$ with respect to $\mathcal{S} = \operatorname{fix} T$ and [13, Thm. 5.5]. □

4.2.1 Related primal-dual algorithms

In this section we elaborate on the relation between TriPD (Alg. 4.1) and other primal-dual algorithms. In Section 3.2 six main special cases of Algorithm 3.1 were discussed. The algorithm considered in this chapter, TriPD (Alg. 4.1), can be seen as SPCA (sequential primal corrector algorithm) applied to the dual problem when the smooth term is zero (see Section 3.2). Note that it is possible to derive and analyze a variant of TriPD (Alg. 4.1) for (3.1). However, we do not pursue this in this chapter and focus on problem (4.1) for clarity of exposition.

A closely related algorithm was proposed in [62] called PAPC (proximal alternating predictor-corrector) for solving saddle point problem (3.3) with $g \equiv 0$ and $l \equiv \delta_{\{0\}}$. Refer to the explanation for SPCA in Section 3.2 for more details. Another closely related primal-dual algorithms is the algorithm of Vũ and Condat [53, 170]. In fact TriPD (Alg. 4.1) transforms into this algorithm with a change of order (by starting from the primal update) and elimination of the dual variable u^k while keeping the axillary dual variable $\bar{u}^k$. Another connection with [53, 170] can be drawn by seeing it as a special case of Algorithm 3.1 as discussed in Section 3.2 under the name SNCA. One can verify that the operator defining the fixed-point iterations in the Vũ-Condat algorithm is given by (4.11) with $H = P + K$ and S defined as follows:

$$S = \begin{pmatrix} \Sigma^{-1} & L \\ L^\top & \Gamma^{-1} \end{pmatrix}, \tag{4.26}$$

$$P = \begin{pmatrix} \Sigma^{-1} & L \\ L^\top & \Gamma^{-1} \end{pmatrix}, \quad K = \begin{pmatrix} 0 & -L \\ L^\top & 0 \end{pmatrix}.$$

For such selection of S, P, K, it holds that $S^{-1}(H + M^\top) = I$, whence in proximal form, the operator defined in (4.11) becomes:

$$\bar{u} = \operatorname{prox}_{h^*}^{\Sigma^{-1}}(u - \Sigma\nabla l^*(u) + \Sigma L x)$$

$$\bar{x} = \operatorname{prox}_{g}^{\Gamma^{-1}}(x - \Gamma\nabla f(x) - \Gamma L^\top(2\bar{u} - u))$$

$$Tz = (\bar{u}, \bar{x}).$$

Observe the non-diagonal structure of S for the Vũ-Condat algorithm in (4.26), in contrast with the one for TriPD (Alg. 4.1) in (4.10). For the sake of comparison with [53, 170] we consider the relaxed iteration $z^{k+1} = z^k + \lambda(Tz^k - z^k)$ for some $\lambda \in (0, 2)$ and problem (3.1) with l satisfying Assumption 3.I*(iii)* in this subsection (which we opted to exclude from TriPD (Alg. 4.1) solely for the purpose of simplicity).

The analysis in Theorem 4.5 can be further used to establish convergence of the Vũ-Condat scheme for problem (3.1) under the sufficient conditions (3.8) (in place of Assumption 4.II). Notice that when $l = \delta_{\{0\}}$ (*i.e.*, for problem (4.1)), $l^* \equiv 0$ whence $\beta_l = 0$, and the condition simplifies to:

$$\Gamma^{-1} - \tfrac{\beta_f}{2(2-\lambda)} Q - L^\top \Sigma L \succ 0.$$

Given the stepsize condition (3.1) the following Fejér-type inequality holds.

$$\|z^{k+1} - z^\star\|_S^2 \leq \|z^k - z^\star\|_S^2 - \lambda\|z^{k+1} - z^k\|_{2\hat{P}-\lambda S}^2, \tag{4.27}$$

with S defined in (4.26) and $\hat{P}$ given by:

$$\hat{P} := \begin{pmatrix} \Sigma^{-1} - \frac{\beta_l}{4} R & L \\ L^\top & \Gamma^{-1} - \frac{\beta_f}{4} Q \end{pmatrix}.$$

This generalizes the result in [53, Thm. 3.1], [170, Cor. 4.2] and [96, Prop. 5.1] where $Q = I$ and the stepsizes are assumed to be scalar.

Our main goal here was to demonstrate the non-diagonal structure of S for the Vũ-Condat algorithm. In the sequel, we highlight that our analysis additionally leads to less conservative conditions as compared to [170, 53, 47]. Notice that the proofs in the aforementioned papers are based on casting the algorithm in the form of forward-backward iterations. Consequently, the stepsize condition obtained ensures that the underlying operator is *averaged*. In contradistinction, the sufficient condition in (3.1) only ensures that the Fejér-type inequality (4.27) holds, which is sufficient for convergence. Therefore, even in the case of scalar stepsizes (as in [170, 53]) condition (3.1) allows for larger stepsizes compared to

[170, 53].

In [47, 134] the authors propose a variable metric version of the algorithm with a preconditioning that accounts for the general Lipschitz metric. This is accomplished by fixing the stepsize matrix to be a constant times the inverse of the Lipschitz metric, and obtaining a condition on the constant. Our approach does not assume this restrictive form for the stepsize matrix; even when such a restriction is imposed it allows for *larger* stepsizes, thus achieving generally faster convergence. As an illustrative example, let us set $\Gamma = \mu Q^{-1}$ and $\Sigma = \nu R^{-1}$ for some $\mu, \nu > 0$. For simplicity and without loss of generality, let $\beta_l = 1$, $\beta_f = 1$. Then, (3.1) simplifies to:

$$(\mu^{-1} - \tfrac{1}{2(2-\lambda)})(\nu^{-1} - \tfrac{1}{2(2-\lambda)})Q - L^\top R^{-1} L \succ 0, \tag{4.28}$$

whereas the condition required in [47, 134] is $\lambda \in (0, 1]$ and

$$\frac{\delta}{1+\delta} > \frac{\max\{\mu, \nu\}}{2} \text{ with } \delta = \tfrac{1}{\sqrt{\nu\mu}} \|G^{-1/2} L Q^{-1/2}\|^{-1} - 1. \tag{4.29}$$

It is not difficult to check that condition, (4.28), is always less restrictive than (4.29). For instance, let $R^{-1/2} L Q^{-1/2} = I$ and set $\mu = 1.5$, then (4.28) requires that $\nu < \frac{1}{6.5}$ whereas (4.29) necessitates that $\nu < \frac{1}{24}$.

4.3 A randomized block-coordinate algorithm

In this section, we describe a randomized block-coordinate variant of TriPD (Alg. 4.1) and discuss important special cases pertaining to the randomized coordinate activation mechanism. The convergence analysis is based on establishing stochastic Fejér monotonicity [51] of the generated sequence. In addition, we establish linear convergence of the method under further assumptions in Section 4.4.

First, let us define a partitioning of the vector of primal-dual variables into m blocks of coordinates. Notice that each block might include a subset of primal or dual variables, or a combination of both. Respectively, let $U_i \in \mathbb{R}^{(n+r)\times(n+r)}$, for $i = 1, \ldots, m$, be a diagonal matrix with 0-1 diagonal entries that is used to select a subset of the coordinates (selected coordinates correspond to diagonal entries equal to 1). We call such matrix an *activation matrix*, as it is used to activate/select a subset of coordinates to update.

Let $\Phi = \{0, 1\}^m$ denote the set of binary strings of length m (with the elements considered as column vectors of dimension m). At the k-th iteration, the

algorithm draws a Φ-valued random activation vector ϵ^{k+1} which determines which *blocks of coordinates* will be updated. The i-th element of the vector ϵ^{k+1} is denoted as ϵ_i^{k+1}: the i-th block is updated at iteration k if $\epsilon_i^{k+1} = 1$. Notice that in general multiple blocks of coordinates may be concurrently updated. The conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}_k]$ is abbreviated by $\mathbb{E}_k[\cdot]$, where $\mathcal{F}_k$ is the filtration generated by $(\epsilon^1, \ldots, \epsilon^k)$. The following assumption summarizes the setup of the randomized coordinate selection.

Assumption 4.III.

(i) $\{U_i\}_{i=1}^m$ are 0-1 diagonal matrices and $\sum_{i=1}^m U_i = I$.

(ii) $(\epsilon^k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. Φ-valued random vectors with

$$p_i \coloneqq \mathbb{P}(\epsilon_i^1 = 1) > 0 \quad i = 1, \ldots, m. \tag{4.30}$$

(iii) The stepsize matrices Σ, Γ are diagonal.

The first condition implies that the activation matrices define a *partition* of the coordinates, while the second that each partition is activated with a positive probability.

We further define the (diagonal) coordinate activation probability matrix Π as follows:

$$\Pi \coloneqq \sum_{i=1}^m p_i U_i. \tag{4.31}$$

For $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$ we define the operator $\hat{T}^{(\epsilon)}$ by:

$$\hat{T}^{(\epsilon)} z \coloneqq z + \sum_{i=1}^m \epsilon_i U_i (Tz - z),$$

where T was defined in (4.6) (equivalently (4.11)). Observe that this is a compact notation for the update of only the selected blocks. The randomized scheme is then written as an iterative application of $\hat{T}^{(\epsilon^{k+1})}$ for $k = 0, 1, \ldots$ (this operator updates the active blocks of coordinates and leaves the others unchanged, *i.e.*, equal to their previous iterate values). The randomized block-coordinate scheme is summarized below.

Algorithm 4.2 Block-coordinate TriPD algorithm

Inputs: $x^0 \in \mathbb{R}^n$, $u^0 \in \mathbb{R}^r$

for $k = 0, 1, \ldots$ **do**

Select Φ-valued r.v. ϵ^{k+1}

$z^{k+1} = \hat{T}^{(\epsilon^{k+1})} z^k$

We emphasize that the randomized model that we adopt here is capable of capturing many stationary randomized activation mechanisms. To illustrate this, consider the following activation mechanisms (of specific interest in the realm of distributed multi-agent optimization *cf.* §4.5):

- *Multiple coordinate activation*: at each iteration, the j-th coordinate block is randomly activated with probability $p_j > 0$ independent of other coordinates blocks. This corresponds to the case that the sample space is equal to $\Phi = \{0, 1\}^m$. The general distributed algorithm of Section 4.5 assumes this mechanism.

- *Single coordinate activation*: at each iteration, one coordinate block is selected, *i.e.*, the sample space is

$$\{(1, 0, \ldots, 0), (0, 1, 0, \ldots, 0) \ldots, (0, \ldots, 0, 1)\}. \tag{4.32}$$

  We assign probability p_i to the event $\epsilon_i = 1$ (and $\epsilon_j = 0$ for $j \neq i$), whence the probabilities must satisfy $\sum_{i=1}^m p_i = 1$.

The next lemma establishes stochastic Fejér monotonicity for the generated sequence, by directly exploiting the diagonal structure of S. The proof technique is adapted from [22, Thm. 3] (see also [88, Thm. 2], [51, Thm. 2.5]), and is based on the Robbins-Siegmund lemma [142].

Theorem 4.6. *Let Assumptions 4.I–4.III hold. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by TriPD-BC (Alg. 4.2). The following Fejér-type inequality holds for all $z^\star \in \mathcal{S}$:*

$$\mathbb{E}_k\left[\|z^{k+1} - z^\star\|^2_{\Pi^{-1}S}\right] \leq \|z^k - z^\star\|^2_{\Pi^{-1}S} - \|Tz^k - z^k\|^2_{2\tilde{P}-S}. \tag{4.33}$$

Consequently, $(z^k)_{k\in\mathbb{N}}$ converges a.s. to some $z^\star \in \mathcal{S}$.

Proof. Let us define the operator $E^k := \sum_{i=1}^m \epsilon_i^k U_i$ that maps the elements of $(\mathbb{R}^{n+r}, \mathcal{F}_{k-1})$ to $(\mathbb{R}^{n+r}, \mathcal{F}_k)$. The iterations of TriPD-BC (Alg. 4.2) can be

written as $z^{k+1} = z^k + E^{k+1}(Tz^k - z^k)$. We have

$$\mathbb{E}_k \circ E^{k+1} = \sum_{\varepsilon\in\Psi} \mathbb{P}(\epsilon^{k+1} = \varepsilon) \sum_{j=1}^{m} \varepsilon_j U_j = \sum_{j=1}^{m} \sum_{\varepsilon\in\Psi} \mathbb{P}(\epsilon^{k+1} = \varepsilon)\varepsilon_j U_j$$

$$= \sum_{j=1}^{m} \sum_{\varepsilon\in\Psi,\varepsilon_j=1} \mathbb{P}(\epsilon^{k+1} = \varepsilon) U_j = \sum_{j=1}^{m} p_j U_j = \Pi, \tag{4.34}$$

where we used Assumptions 4.III*(i)* and 4.III*(ii)*. Therefore, we have

$$\begin{aligned}
\mathbb{E}_k\left[\|z^{k+1} - z^\star\|^2_{\Pi^{-1}S}\right] &= \mathbb{E}_k\left[\|z^k + E^{k+1}(Tz^k - z^k) - z^\star\|^2_{\Pi^{-1}S}\right] \\
&= \|z^k - z^\star\|^2_{\Pi^{-1}S} + 2\langle z^k - z^\star, \mathbb{E}_k\left[E^{k+1}(Tz^k - z^k)\right]\rangle_{\Pi^{-1}S} \\
&\quad + \mathbb{E}_k\left[\langle E^{k+1}(Tz^k - z^k), E^{k+1}(Tz^k - z^k)\rangle_{\Pi^{-1}S}\right] \\
&= \|z^k - z^\star\|^2_{\Pi^{-1}S} + \|Tz^k - z^k\|^2_S + 2\langle z^k - z^\star, Tz^k - z^k\rangle_S
\end{aligned}$$

where we used (4.34) and the fact E^k is self-adjoint and idempotent (since U_i are 0-1 matrices) in the last equality. Inequality (4.33) follows by using (4.16). The convergence of the sequence follows from (4.33) using the Robbins-Siegmund lemma [142] and arguing as in [22, Thm. 3] and [51, Prop. 2.3]. □

It is important to emphasize that a naive implementation of TriPD-BC (Alg. 4.2) (with regards to the partitioning of primal-dual variables) may involve wasteful computations. As an example, consider a BC algorithm in which, at every iteration, either all primal or all dual variables are updated. In such a case, if at iteration k the dual vector is to be updated, both x^{k+1}, u^{k+1} are computed (*cf.* Alg. 4.1), whereas only u^{k+1} is updated. This phenomenon is common to all primal-dual algorithms, and is due to the fact that the primal and dual updates need to be performed sequentially in the full version of the algorithm. As a consequence, the blocks of coordinates must be partitioned in such a way that computations are not discarded, so that the iteration cost of a BC algorithm is (substantially) smaller than computing the full operator T. This choice relies entirely on the structure of the optimization problem under consideration. A canonical example of prominent practical interest is the setting of multi-agent optimization in a network (*cf.* §4.5), where L is not diagonal, f and g are separable, and additional coupling between (primal) coordinates is present through h, see (4.51). In this example, the primal and dual coordinates are partitioned in such a way that no computation is discarded (*cf.* §4.5 for more details).

We proceed with another example where the coordinates may be grouped such that the BC algorithm does not incur any wasteful computations: consider problem (4.1) with $Lx = \operatorname{blkdiag}(L_1x_1, \dots, L_mx_m)$, and g, h separable functions *i.e.*,

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x) + \sum_{i=1}^{m} \big(g_i(x_i) + h_i(L_ix_i)\big).$$

In this problem, the coupling between the (primal) coordinates is carried via function f. For each $i = 1, \dots, m$, we can choose U_i such that it selects the i-th primal-dual coordinate block (u_i, x_i). Under such partitioning of coordinates, one may use TriPD-BC (Alg. 4.2) with any random activation pattern satisfying Assumption 4.III. For example, for the case of multiple independently activated coordinates, as discussed above, at iteration k the following is performed

$$\begin{cases} \bullet \text{ each block } (u_i, x_i) \text{ is activated with probability } p_i > 0 \\ \bullet \text{ for active block(s) } i \text{ compute:} \\ \bar{u}_i^k = \operatorname{prox}_{\sigma h_i^*}(u_i^k + \sigma L_i x_i^k) \\ x_i^{k+1} = \operatorname{prox}_{\gamma g_i}(x_i^k - \gamma \nabla_i f(x^k) - \gamma L_i^\top \bar{u}_i^k) \\ u_i^{k+1} = \bar{u}_i^k + \sigma L_i(x_i^{k+1} - x_i^k). \end{cases}$$

More generally, when g and h are separable in problem (4.1), and L is such that either each (block) row only has one nonzero element or each (block) column has one nonzero element, then the coordinates can be grouped together in such a way that no wasteful computations occur: in the first case the primal vector x_i and all dual vectors u_j that are required for its computation are selected by U_i (with the role of primal and dual reversed in the second case).

Remark 4.7. Note that in TriPD-BC (Alg. 4.2) the probabilities p_i are taken fixed, *i.e.*, the matrix Π is constant throughout the iterations. This is a non-restrictive assumption and can be relaxed by considering iteration-varying probabilities p_i^k in (4.30) and modifying TriPD-BC (Alg. 4.2) by setting:

$$z^{k+1} = z^k + \sum_{i=1}^{m} \frac{\epsilon_i^{k+1}}{m p_i^{k+1}} U_i(Tz^k - z^k).$$

Let Π^k denote the probability matrix defined as in (4.31) using p_i^k. Then, by arguing as in Theorem 4.6, it can be shown that the following stochastic Fejér monotonicity holds for the modified sequence:

$$\mathbb{E}_k\big[\|z^{k+1} - z^\star\|_S^2\big] \le \|z^k - z^\star\|_S^2 - \|Tz^k - z^k\|^2_{\frac{2}{m}\tilde{P} - \frac{1}{m^2}S(\Pi^{k+1})^{-1}}.$$

□

4.4 Linear convergence

In this section, we establish linear convergence of Algorithms 4.1 and 4.2 under additional conditions on the cost functions f, g and h. To this end, we show that linear convergence is attained if the monotone operator $F = A + M + C$ defining the primal-dual optimality conditions (*cf.* (4.8)) is *metrically subregular* (globally metrically subregular in the case of TriPD-BC (Alg. 4.2)). A notable consequence of our analysis is the fact that linear convergence is attained when the cost functions either a) belong in the class of *piecewise linear-quadratic* (PLQ) convex functions or b) when they satisfy a certain *quadratic growth* condition (which is much weaker than strong convexity). Moreover, notice that in the case of PLQ the solution need not be unique (*cf.* Thm.s 4.10 and 4.11).

Recall the notion of *metric subregularity* in Definition 1.2. Metric subregularity of the subdifferential operator has been studied thoroughly and is equivalent to the *quadratic growth condition* [6, 63] defined next. In particular, for a proper closed convex function f, the subdifferential ∂f is metrically subregular at $\bar{x}$ for $\bar{y}$ with $(\bar{x}, \bar{y}) \in \operatorname{gra} \partial f$ if and only if there exists a positive constant c and a neighborhood $\mathcal{U}$ of $\bar{x}$ such that the following growth condition holds [6, Thm. 3.3]:

$$f(x) \geq f(\bar{x}) + \langle \bar{y}, x - \bar{x} \rangle + c d^2(x, (\partial f)^{-1}(\bar{y})) \quad \forall x \in \mathcal{U}$$

Furthermore, ∂f is strongly subregular at $\bar{x}$ for $\bar{y}$ with $(\bar{x}, \bar{y}) \in \operatorname{gra} \partial f$, if and only if there exists a positive constant c and a neighborhood $\mathcal{U}$ of $\bar{x}$ such that [6, Thm. 3.5]:

$$f(x) \geq f(\bar{x}) + \langle \bar{y}, x - \bar{x} \rangle + c\|x - \bar{x}\|^2 \quad \forall x \in \mathcal{U} \tag{4.35}$$

Note that strongly convex functions satisfy (4.35), but (4.35) is *much weaker* than strong convexity, as it is a *local condition*: it only holds in a neighborhood of $\bar{x}$, and also only for $\bar{y}$.

The lemma below provides a sufficient condition for metric subregularity of the monotone operator $A + M + C$, in terms of strong subregularity of $\nabla f + \partial g$ and ∂h^* (equivalently the quadratic growth of $f + g$ and h^*, *cf.* (4.35)) as stated in the following assumption:

Assumption 4.IV (strong subregularity of $\nabla f + \partial g$ and ∂h^*)**.** There exists $z^\star = (u^\star, x^\star) \in \mathcal{S}$ satisfying:

(i) $\nabla f + \partial g$ is strongly subregular at $x^\star$ for $-L^\top u^\star$,

(ii) ∂h^* is strongly subregular at $u^\star$ for $Lx^\star$.

We say that f, g and h satisfy this assumption *globally* if the strong subregularity assumption of $\nabla f + \partial g$ and ∂h^* both hold globally (*cf.* Def. 1.2).

In particular, Assumption 4.IV holds globally if either f or g (or both) are strongly convex and h is continuously differentiable with Lipschitz continuous gradient, *i.e.*, h^* is strongly convex.

Lemma 4.8. *Let Assumptions 4.I and 4.IV hold. Then, $F = A + M + C$ (cf. (4.7)) is strongly subregular at $z^\star$ for 0. Moreover, if f, g and h satisfy Assumption 4.IV* globally*, then F is globally strongly subregular at $z^\star$ for 0. In both cases the set of primal-dual solutions is a singleton, $\mathcal{S} = \{z^\star\}$.*

Proof. From the equivalent characterization of strong subregularity in (4.35) we have that there exists a neighborhood $\mathcal{U}_{x^\star}$ of $x^\star$ such that for all $x \in \mathcal{U}_{x^\star}$

$$\begin{aligned}(f+g)(x) \geq &(f+g)(x^\star) + \langle -L^\top u^\star, x - x^\star\rangle \\ &+ c_1\|x - x^\star\|^2,\end{aligned} \tag{4.36}$$

and a neighborhood $\mathcal{U}_{u^\star}$ of $u^\star$ such that for all $u \in \mathcal{U}_{u^\star}$

$$h^*(u) \geq h^*(u^\star) + \langle Lx^\star, u - u^\star\rangle + c_2\|u - u^\star\|^2. \tag{4.37}$$

Fix $z = (u, x)$ with $u \in \mathcal{U}_{u^\star}$ and $x \in \mathcal{U}_{x^\star}$. Consider $v = (v_1, v_2) \in Fz \coloneqq Az + Mz + Cz$. By definition (cf. (4.7)) we have

$$\begin{cases} v_1 \in \partial h^*(u) - Lx, \\ v_2 \in \partial g(x) + \nabla f(x) + L^\top u. \end{cases}$$

Using this together with the definition of subdifferential yields:

$$\langle v_1 + Lx, u - u^\star\rangle \geq h^*(u) - h^*(u^\star), \tag{4.38}$$

$$\langle v_2 - L^\top u, x - x^\star\rangle \geq (f+g)(x) - (f+g)(x^\star). \tag{4.39}$$

Combining (4.38), (4.39) with (4.36), (4.37) and noting that

$$\langle L^\top(u^\star - u), x - x^\star\rangle + \langle L(x - x^\star), u - u^\star\rangle = 0,$$

yields:

$$\begin{aligned}\langle v, z - z^\star\rangle &= \langle v_1, u - u^\star\rangle + \langle v_2, x - x^\star\rangle \\ &\geq c_2\|u - u^\star\|^2 + c_1\|x - x^\star\|^2 \geq c\|z - z^\star\|^2,\end{aligned}$$

where $c = \min\{c_1, c_2\}$. Therefore, by the Cauchy-Schwarz inequality $\|v\| \geq c\|z - z^\star\|$. Since $\|z - z^\star\| \geq \operatorname{dist}(z, F^{-1}0)$, and $v \in Fz$ was selected arbitrarily, we have

$$\operatorname{dist}(z, F^{-1}0) \leq \tfrac{1}{c} \operatorname{dist}(0, Fz) \quad \forall z \in \mathcal{U}_{u^\star} \times \mathcal{U}_{x^\star}. \tag{4.40}$$

Thus F is metrically subregular at $z^\star$ for 0.

To establish uniqueness of the primal-dual solution consider:

$$\mathcal{L}(u, x) \coloneqq (f + g)(x) + \langle Lx, u\rangle - h^*(u).$$

Adding (4.36) and (4.37) yields

$$\mathcal{L}(u^\star, x) - \mathcal{L}(u, x^\star) \geq c\|z - z^\star\|^2 \quad \forall z \in \mathcal{U}_{u^\star} \times \mathcal{U}_{x^\star} \tag{4.41}$$

Let $\bar{z}^\star = (\bar{u}^\star, \bar{x}^\star) \in \mathcal{S}$ such that $\bar{z}^\star \in \mathcal{U}_{u^\star} \times \mathcal{U}_{x^\star}$. Since $\bar{z}^\star$ is also a primal-dual solution we have $\mathcal{L}(\bar{u}^\star, x^\star) - \mathcal{L}(u^\star, \bar{x}^\star) \geq 0$. Therefore, using (4.41) at $\bar{z}^\star$ yields $\bar{z}^\star = z^\star$. Since $\mathcal{S}$ is convex, we conclude that it is a singleton, *i.e.*, $\mathcal{S} = \{z^\star\}$. Consequently it follows from (4.40) that F is strongly subregular at $z^\star$ for 0.

The second part is a direct consequence of the first part and the fact that if Assumption 4.IV holds globally then also the quadratic growth conditions (4.36) and (4.37) hold globally, *i.e.*, $\mathcal{U}_{x^\star} = \mathbb{R}^n$, $\mathcal{U}_{u^\star} \in \mathbb{R}^r$. This can be shown by adapting the proof of [6, Thm. 3.3]. □

Lemma 4.9 establishes metric subregularity of the operator $A + M + C$ when the functions f, g and h are *piecewise linear-quadratic* (PLQ) (see Definition 1.4). The claim follows directly from Lemma 3.11. Note that this assumption does not imply that the set of solutions $\mathcal{S}$ is a singleton, nevertheless, linear convergence can still be established. The class of PLQ functions is closed under scalar multiplication, addition, conjugation and Moreau envelope [144]. A wide range of functions used in optimization applications belong to this class, for example: affine functions, quadratic forms, indicators of polyhedral sets, polyhedral norms (*e.g.*, the ℓ_1-norm), and regularizing functions such as elastic net, Huber loss, hinge loss, to name a few.

Lemma 4.9. *Let Assumption 4.I hold. In addition, assume that f, g and h are piecewise linear-quadratic. Then $F = A + M + C$ (cf. (4.7)) is metrically subregular with the same constant η at any z for any v with $(z, v) \in \operatorname{gra} F$.*

Our main convergence rate results are provided in Theorems 4.10 and 4.11. In this context, Lemmas 4.8 and 4.9 are used to establish sufficient conditions in terms of the cost functions. We omit the proof of Theorem 4.10; it is similar to that of Theorem 4.11, the main difference being that in Theorem 4.10 *local* (as opposed to global) metric subregularity is used: due to the Fejér-type inequality

(4.23), $\bar{z}^k$ will eventually be contained in a neighborhood of metric subregularity, where inequality (4.46) applies.

Theorem 4.10 (linear convergence of Alg. 4.1)**.** *Consider TriPD (Alg. 4.1) under the assumptions of Theorem 4.5. Suppose that $F = A+M+C$ is metrically subregular at all $z^\star \in \mathcal{S}$ for 0. Then, $(\operatorname{dist}_S(z^k, \mathcal{S}))_{k\in\mathbb{N}}$ converges Q-linearly to zero, and $(z^k)_{k\in\mathbb{N}}$ converges R-linearly to some $z^\star \in \mathcal{S}$.*

In particular, the metric subregularity assumption holds and the result follows if either one of the following holds:

(i) either f, g and h are PLQ,

(ii) or f, g and h satisfy Assumption 4.IV, in which case the solution is unique.

Theorem 4.11 (linear convergence of Alg. 4.2)**.** *Consider TriPD-BC (Alg. 4.2) under the assumptions of Theorem 4.6. Suppose that $F = A + M + C$ is globally metrically subregular for 0 (cf. Def. 1.2), i.e., there exists $\eta > 0$ such that*

$$\operatorname{dist}(z, F^{-1}0) \leq \eta \operatorname{dist}(0, Fz) \quad \forall z \in \mathbb{R}^{n+r}.$$

Then $(\mathbb{E}\left[d^2_{\Pi^{-1}S}(z^k, \mathcal{S})\right])_{k\in\mathbb{N}}$ converges Q-linearly to zero.

In particular, this result holds if

(i) either f, g, h are PLQ and there exists a compact set $\mathcal{C}$ such that $(z^k)_{k\in\mathbb{N}} \subseteq \mathcal{C}$ (as is the case if $\operatorname{dom} g$ and $\operatorname{dom} h^*$ are compact),

(ii) or f, g and h satisfy Assumption 4.IV globally, in which case the solution is unique.

Proof. For notational convenience let $\bar{S} = \Pi^{-1}S$ and note that $\mathcal{S} = \operatorname{zer} F$ (*cf.* (4.14)). By definition we have $\|z^k - \mathcal{P}^{\bar{S}}_{\mathcal{S}}(z^k)\|_{\bar{S}} = \operatorname{dist}_{\bar{S}}(z^k, \mathcal{S})$ (where the minimum is attained since $\mathcal{S}$ is a closed convex set). Consequently, it follows from (4.33) that

$$\begin{aligned}
\mathbb{E}_k\left[d^2_{\bar{S}}(z^{k+1}, \mathcal{S})\right] &\leq \mathbb{E}_k\left[\|z^{k+1} - \mathcal{P}^{\bar{S}}_{\mathcal{S}}(z^k)\|^2_{\bar{S}}\right] \\
&\leq \|z^k - \mathcal{P}^{\bar{S}}_{\mathcal{S}}(z^k)\|^2_{\bar{S}} - \|Tz^k - z^k\|^2_{2\bar{P}-S} \\
&= \operatorname{dist}^2_{\bar{S}}(z^k, \mathcal{S}) - \|Tz^k - z^k\|^2_{2\bar{P}-S}. \qquad (4.42)
\end{aligned}$$

By definition (4.11), we have

$$\begin{aligned}\|\bar{z}^k - z^k\|^2 &= \|(H + M^\top)^{-1} S(Tz^k - z^k)\|^2 \\ &\leq \|(H + M^\top)^{-1} S\|^2 \|(2\tilde{P} - S)^{-1}\| \|Tz^k - z^k\|^2_{2\tilde{P}-S}, \end{aligned} \tag{4.43}$$

where $\bar{z}^k$ is defined by (4.12) applied at $z = z^k$. Consider the projection of $\bar{z}^k$ onto $\mathcal{S}$, $\mathcal{P}_{\mathcal{S}}(\bar{z}^k)$. By definition $\|\bar{z}^k - \mathcal{P}_{\mathcal{S}}(\bar{z}^k)\| = \operatorname{dist}(\bar{z}^k, \mathcal{S})$, and we have

$$\begin{aligned}\operatorname{dist}^2_{\bar{S}}(z^k, \mathcal{S}) \leq \|z^k - \mathcal{P}_{\mathcal{S}}(\bar{z}^k)\|^2_{\bar{S}} &\leq \|\bar{S}\| \|z^k - \mathcal{P}_{\mathcal{S}}(\bar{z}^k)\|^2 \\ &\leq \|\bar{S}\| \big(\|\bar{z}^k - \mathcal{P}_{\mathcal{S}}(\bar{z}^k)\| + \|\bar{z}^k - z^k\|\big)^2 \\ &= \|\bar{S}\| \big(\operatorname{dist}(\bar{z}^k, \mathcal{S}) + \|\bar{z}^k - z^k\|\big)^2. \end{aligned} \tag{4.44}$$

In what follows we bound $\operatorname{dist}(\bar{z}^k, \mathcal{S})$ by $\|\bar{z}^k - z^k\|$. Define

$$v^k := -(H - M)(\bar{z}^k - z^k) + C\bar{z}^k - Cz^k. \tag{4.45}$$

It follows from (4.12) that $(H - M - C)z^k \in (H + D)\bar{z}^k$, which in turn implies

$$v^k \in F\bar{z}^k = (A + M + C)\bar{z}^k.$$

Consequently, using (global) metric subregularity of F yields

$$\operatorname{dist}(\bar{z}^k, \mathcal{S}) \leq \eta \|v^k\|. \tag{4.46}$$

By the triangle inequality and Lipschitz continuity of C,

$$\begin{aligned}\|v^k\| &= \|(H - M)(\bar{z}^k - z^k) - C\bar{z}^k + Cz^k\| \\ &\leq \|(H - M)(\bar{z}^k - z^k)\| + \|C\bar{z}^k - Cz^k\| \leq \xi \|\bar{z}^k - z^k\|, \end{aligned} \tag{4.47}$$

where $\xi = \|H - M\| + \beta_f \|Q\|$. By (4.46) and (4.47) we have

$$\operatorname{dist}(\bar{z}^k, \mathcal{S}) \leq \xi\eta \|\bar{z}^k - z^k\|.$$

Combine this with (4.43) and (4.44) to derive

$$\operatorname{dist}^2_{\bar{S}}(z^k, \mathcal{S}) \leq \phi \|Tz^k - z^k\|^2_{2\tilde{P}-S}, \tag{4.48}$$

where $\phi = (\xi\eta + 1)^2\|(H + M^\top)^{-1}S\|^2\|(2\tilde{P} - S)^{-1}\|\|\bar{S}\|$. Therefore, by (4.42) and (4.48) we have

$$\mathbb{E}_k\left[d^2_{\bar{S}}(z^{k+1}, \mathcal{S})\right] \leq \operatorname{dist}^2_{\bar{S}}(z^k, \mathcal{S}) - \tfrac{1}{\phi}\operatorname{dist}^2_{\bar{S}}(z^k, \mathcal{S}).$$

Taking expectation in both sides concludes the proof. For the case of PLQ functions, let $\mathcal{U}_{z_\star}$ denote an open subregularity neighborhood around $z_\star \in \mathcal{S}$, and set $\mathcal{U}_\star := \cup_{z^\star \in \mathcal{S}}\mathcal{U}_{z_\star}$. By Lemma 4.9 there exists a positive η such that $\operatorname{dist}(z, F^{-1}0) \leq \eta \operatorname{dist}(Fz, 0)$ for $z \in \mathcal{U}_\star$. Moreover, since $(z^k)_{k\in\mathbb{N}} \subseteq \mathcal{C}$ up to possibly enlarging $\mathcal{C}$ we have $(\bar{z}^k)_{k\in\mathbb{N}} \subseteq \mathcal{C}$. Note that since $(z^k)_{k\in\mathbb{N}} \subseteq \mathcal{C}$ and $\mathcal{C}$ is closed, $\mathcal{C} \cap \mathcal{S} \neq \emptyset$ and $\mathcal{C} \cap \mathcal{U}_\star \neq \emptyset$. It is sufficient to show that $\operatorname{dist}(z, F^{-1}0) \leq \eta \operatorname{dist}(0, Fz)$ for $z \in \mathcal{C}$. Let us define $D(z) := \operatorname{dist}(0, Fz)$. Since $\operatorname{gra} F$ is closed, $D(z)$ is lower semicontinuous [144, Thm. 5.7, Prop. 5.11(a)]. By [144, Cor. 1.10] $D(z)$ attains a minimum over the compact set $\mathcal{C} \setminus \mathcal{U}_\star$: $c_\mathcal{C} := \min_{z\in\mathcal{C}\setminus\mathcal{U}_\star} D(z) > 0$ where the strict inequality is due to the fact that the minmizer belongs to $\mathcal{C} \setminus \mathcal{U}_\star$. Moreover, $c_\mathcal{C} := \sup_{z\in\mathcal{C}} \operatorname{dist}(z, F^{-1}0) < \infty$ due to the fact that $\mathcal{C}$ is bounded. Hence $\operatorname{dist}(z, F^{-1}0) \leq c_\mathcal{C} \leq \frac{c_\mathcal{C}}{c_d}\operatorname{dist}(Fz, 0)$ for $z \in \mathcal{C} \setminus \mathcal{U}_\star$. Therefore, by combining the two cases we obtain $\operatorname{dist}(z, F^{-1}0) \leq \max\{\frac{c_\mathcal{C}}{c_d}, \eta\}\operatorname{dist}(Fz, 0)$ for $z \in \mathcal{C}$ as claimed. The second sufficient condition follows from Lemma 4.8. □

In the recent work [104] the authors establish linear convergence in the framework of non-expansive operators under the assumption that the residual mapping defined by $R = \mathrm{id} - T$ is metrically subregular. However, such a condition is not easily verifiable in terms of conditions on the cost functions. In the next lemma, we show that R is metrically subregular if and only if the monotone operator F is metrically subregular. This result connects the two assumptions and is interesting in its own right. More importantly, it enables the use of Lemmas 4.8 and 4.9 for establishing linear convergence for a wide array of problems.

Lemma 4.12. *Let Assumptions 4.I and 4.II hold. Consider the operator T defined in (4.11) and a point $z^\star \in \mathcal{S}$. Then, $F = A + M + C$ (cf. (4.7)) is metrically subregular at $z^\star$ for 0 if and only if the residual mapping $R := \mathrm{id} - T$ is metrically subregular at $z^\star$ for 0.*

4.5 Distributed optimization

In this section, we consider a general formulation for multi-agent optimization over a network, and leverage Algorithms 4.1 and 4.2 to devise both synchronous and randomized asynchronous distributed primal-dual algorithms. The setting is as follows. We consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ over a vertex set

$\mathcal{V} = \{1, \ldots, m\}$ with edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. Each vertex is associated with a corresponding *agent*, which is assumed to have a local memory and computational unit, and can only communicate with its neighbors. We define the *neighborhood* of agent i by $\mathcal{N}_i := \{j | (i,j) \in \mathcal{E}\}$. We use the terms vertex, agent, and node interchangeably. The goal is to solve the following global optimization problem in a distributed fashion:

$$\underset{x_1,\ldots,x_m}{\text{minimize}} \quad \sum_{i=1}^{m} f_i(x_i) + g_i(x_i) + h_i(L_i x_i) \tag{4.49a}$$

$$\text{subject to} \quad A_{ij}x_i + A_{ji}x_j = b_{(i,j)} \qquad (i,j) \in \mathcal{E}, \tag{4.49b}$$

where $x_i \in \mathbb{R}^{n_i}$. The cost functions f_i, g_i, $h_i \circ L_i$ are taken *private* to agent/node $i \in \mathcal{V}$, *i.e.*, our distributed methods operate solely by exchanging local variables among neighboring nodes that are unaware of each other's objectives. The coupling in the problem is represented through the edge constraints (4.49b).

Throughout this section the following assumptions hold:

Assumption 4.V. For each $i = 1, \ldots, m$:

(i) For $j \in \mathcal{N}_i$, $b_{(i,j)} \in \mathbb{R}^{l_{(i,j)}}$ and $A_{ij} \in \mathbb{R}^{n_i} \to \mathbb{R}^{l_{(i,j)}}$ is a linear mapping.

(ii) $g_i : \mathbb{R}^{n_i} \to \overline{\mathbb{R}}$, $h_i : \mathbb{R}^{r_i} \to \overline{\mathbb{R}}$ are proper closed convex functions, and $L_i : \mathbb{R}^{n_i} \to \mathbb{R}^{r_i}$ is a linear mapping.

(iii) $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ is convex, continuously differentiable, and for some $\beta_i \in [0, \infty)$, ∇f_i is β_i-Lipschitz continuous with respect to the metric $Q_i \succ 0$, *i.e.*,

$$\|\nabla f_i(x) - \nabla f_i(y)\|_{Q_i^{-1}} \leq \beta_i \|x - y\|_{Q_i} \quad x, y \in \mathbb{R}^{n_i}.$$

(iv) The graph $\mathcal{G}$ is connected.

(v) The set of solutions of (4.49) is nonempty. Moreover, there exists $x_i \in \operatorname{ri}\operatorname{dom} g_i$ such that $L_i x_i \in \operatorname{ri}\operatorname{dom} h_i$, for $i = 1, \ldots, m$, and $A_{ij}x_i + A_{ji}x_j = b_{(i,j)}$ for $(i,j) \in \mathcal{E}$.

Each agent $i \in \mathcal{V}$ maintains its own local primal variable $x_i \in \mathbb{R}^{n_i}$ and dual variables $y_i \in \mathbb{R}^{r_i}$, and $w_{(i,j),i} \in \mathbb{R}^{l_{(i,j)}}$ (for each $j \in \mathcal{N}_i$), where the former is related to the linear mapping L_i, and the latter is the local dual variable of agent i corresponding to the edge-constraint (4.49b). It is important to note that the updates in TriPD-Dist (Alg. 4.3) are performed locally through communication with neighbors: the only information that agent i shares with its neighbor $j \in \mathcal{N}_i$ is the quantity $A_{ij}x_i$, along with edge variable $w_{(i,j),i}$, while all other variables are kept *private*.

Algorithm 4.3 Synchronous & asynchronous versions of TriPD-Dist algorithm

Inputs: $x_i^0 \in \mathbb{R}^{n_i}$, $y_i^0 \in \mathbb{R}^{r_i}$, for $i = 1, \dots, m$, and $w_{(i,j),i} \in \mathbb{R}^{l_{(i,j)}}$ for $j \in \mathcal{N}_i$.

for $k = 0, 1, \dots$ **do**

I: Synchronous version	**II: Asynchronous version**
for all agents $i = 1, \dots, m$ **do**	Each agent $i = 1. \dots, m$ is activated independently with probability $p_i > 0$ **for** all active agents **do**

Local updates:

$$
\begin{aligned}
\bar{w}_{(i,j),i}^k &= \tfrac{1}{2}\big(w_{(i,j),i}^k + w_{(i,j),j}^k\big) + \tfrac{\kappa_{(i,j)}}{2}\big(A_{ij}x_i^k + A_{ji}x_j^k - b_{(i,j)}\big), \quad \forall j \in \mathcal{N}_i \\
\bar{y}_i^k &= \operatorname{prox}_{\sigma_i h_i^\star}\big(y_i^k + \sigma_i L_i x_i^k\big) \\
x_i^{k+1} &= \operatorname{prox}_{\tau_i g_i}\big(x_i^k - \tau_i L_i^\top \bar{y}_i^k - \tau_i \textstyle\sum_{j\in\mathcal{N}_i} A_{ij}^\top \bar{w}_{(i,j),i}^k - \tau_i \nabla f_i(x_i^k)\big) \\
y_i^{k+1} &= \bar{y}_i^k + \sigma_i L_i (x_i^{k+1} - x_i^k) \\
w_{(i,j),i}^{k+1} &= \bar{w}_{(i,j),i}^k + \kappa_{(i,j)} A_{ij}(x_i^{k+1} - x_i^k), \quad \forall j \in \mathcal{N}_i
\end{aligned}
$$

Transmission of information:

Send $A_{ij}x_i^{k+1}$, $w_{(i,j),i}^{k+1}$ to agent j, $\forall j \in \mathcal{N}_i$

The proposed distributed protocol features both a synchronous as well as an asynchronous implementation. In the synchronous version, at every iteration, all the agents update their variables. In the randomized asynchronous implementation, only a subset of randomly activated agents perform updates, at each iteration, and they do so using their local variables as well as information previously communicated to them by their neighbors. After an update is performed, in both cases, updated values are communicated to neighboring agents. Notice that the asynchronous scheme corresponds to the case of multiple coordinate blocks activation in TriPD-BC (Alg. 4.2). Other activation schemes can also be considered, and our convergence analysis plainly carries over; notably, the single agent activation which corresponds to the asynchronous model of [169, 74, 186] in which agents are assumed to 'wake-up' based on independent exponentially distributed tick-down timers.

Furthermore, in TriPD-Dist (Alg. 4.3) each agent i keeps positive local stepsizes σ_i, τ_i and $\big(\kappa_{(i,j)}\big)_{j\in\mathcal{N}_i}$. The edge weights/stepsizes $\kappa_{(i,j)}$ may alternatively be interpreted as inherent parameters of the communication graph. For example, they may be used to capture edge's 'fidelity,' *e.g.*, the channel quality in a communication link. The stepsizes are assumed to satisfy the following *local* assumption that is sufficient for the convergence of the algorithm (*cf.* Thm.s 4.13 and 4.14).

Assumption 4.VI (stepsizes of TriPD-Dist (Alg. 4.3))**.**

(i) (node stepsizes) Each agent i keeps two positive stepsizes σ_i, τ_i.

(ii) (edge stepsizes) A positive stepsize $\kappa_{(i,j)}$ is associated with edge $(i,j) \in \mathcal{E}$, and is shared between agents i, j.

(iii) (convergence condition) The stepsizes satisfy the following *local* condition

$$\tau_i < \frac{1}{\frac{\beta_i \|Q_i\|}{2} + \|\sigma_i L_i^\top L_i + \sum_{j \in \mathcal{N}_i} \kappa_{(i,j)} A_{ij}^\top A_{ij}\|}.$$

According to Assumption 4.VI*(iii)* the stepsizes τ_i, σ_i for each agent only depend on the local parameters β_i, $\|Q_i\|$, the edge weights, $\kappa_{(i,j)}$ and the linear mappings L_i, and A_{ij}, which are all known to agent i; therefore the stepsizes can be selected locally, in a decentralized fashion.

We proceed by casting the multi-agent optimization problem (4.49) in the form of the structured optimization problem (4.1). In doing so, we describe how TriPD-Dist (Alg. 4.3) is derived as an instance of Algorithms 4.1 and 4.2.

Define the linear operator

$$N_{(i,j)} : \mathsf{x} \mapsto (A_{ij} x_i, A_{ji} x_j),$$

and $\mathsf{N} \in \mathbb{R}^{2\sum_{(i,j)\in\mathcal{E}} l_{(i,j)} \times \sum_{i=1}^m n_i}$ by stacking $N_{(i,j)}$:

$$\mathsf{N} : \mathsf{x} \mapsto (N_{(i,j)}\mathsf{x})_{(i,j)\in\mathcal{E}}.$$

Its transpose is given by:

$$\mathsf{N}^\top : (w_{(i,j)})_{(i,j)\in\mathcal{E}} \mapsto \tilde{\mathsf{x}} = \sum_{(i,j)\in\mathcal{E}} N_{(i,j)}^\top w_{(i,j)},$$

with $\tilde{x}_i = \sum_{j\in\mathcal{N}_i} A_{ij}^\top w_{(i,j),i}$. We have set $w_{(i,j)} = (w_{(i,j),i}, w_{(i,j),j})$, *i.e.*, we consider two dual variables (of dimension $l_{(i,j)}$) for each edge constraint, where $w_{(i,j),i}$ is maintained by agent i and $w_{(i,j),j}$ by agent j.

Consider the set

$$C_{(i,j)} = \{(z_1, z_2) \in \mathbb{R}^{l_{(i,j)}} \times \mathbb{R}^{l_{(i,j)}} \mid z_1 + z_2 = b_{(i,j)}\}.$$

Then problem (4.49) can then be re-written as:

$$\text{minimize} \quad \sum_{i=1}^m f_i(x_i) + g_i(x_i) + h_i(L_i x_i) + \sum_{(i,j)\in\mathcal{E}} \delta_{C_{(i,j)}}(N_{(i,j)}\mathsf{x}) \tag{4.50}$$

Let $C = \bigtimes_{(i,j)\in\mathcal{E}} C_{(i,j)}$, $L = \text{blkdiag}(L_1, \dots, L_m)$, and $\mathsf{L}\mathsf{x} = (L\mathsf{x}, \mathsf{N}\mathsf{x}) =: (\tilde{\mathsf{y}}, \tilde{\mathsf{w}}) \in \mathbb{R}^{n_d}$ with $n_d = 2\sum_{(i,j)\in\mathcal{E}} l_{(i,j)} + \sum_{i=1}^m r_i$, and rewrite (4.50) in the following compact form:

$$\text{minimize} \quad \mathsf{f}(\mathsf{x}) + \mathsf{g}(\mathsf{x}) + \tilde{\mathsf{h}}(\mathsf{L}\mathsf{x}), \tag{4.51}$$

where $\mathsf{f}(\mathsf{x}) = \sum_{i=1}^m f_i(x_i)$, $\mathsf{g}(\mathsf{x}) = \sum_{i=1}^m g_i(x_i)$, $\tilde{\mathsf{h}}(\tilde{\mathsf{y}}, \tilde{\mathsf{w}}) = \mathsf{h}(\tilde{\mathsf{y}}) + \delta_C(\tilde{\mathsf{w}})$, $\mathsf{h}(\tilde{\mathsf{y}}) = \sum_{i=1}^m h_i(\tilde{y}_i)$.

In what follows, $\mathcal{S}$ refers to the set of primal-dual solutions of (4.51). As in Section 4.2, the primal-dual optimality conditions can be written in the form of monotone inclusion (4.8) with

$$A : (\mathsf{y}, \mathsf{w}, \mathsf{x}) \mapsto (\partial\mathsf{h}^*(\mathsf{y}), \partial\delta_C^*(\mathsf{w}), \partial\mathsf{g}(\mathsf{x})),$$

$$M : (\mathsf{y}, \mathsf{w}, \mathsf{x}) \mapsto (-L\mathsf{x}, -\mathsf{N}\mathsf{x}, L^\top\mathsf{y} + \mathsf{N}^\top\mathsf{w}),$$

$$C : (\mathsf{y}, \mathsf{w}, \mathsf{x}) \mapsto (0, 0, \nabla\mathsf{f}(\mathsf{x})),$$

where $\mathsf{u} = (\mathsf{y}, \mathsf{w})$ represents the dual vector.

We define the edge weight matrix as follows

$$W = \text{blkdiag}\left((\kappa_{(i,j)} I_{2l_{(i,j)}})_{(i,j)\in\mathcal{E}}\right),$$

where the weights $\kappa_{(i,j)}$ are repeated twice (for each of the two neighboring agents). Furthermore, we set

$$\Sigma = \text{blkdiag}(\sigma_1 I_{r_1}, \dots, \sigma_m I_{r_m}, W),$$

$$\Gamma = \text{blkdiag}(\tau_1 I_{n_1}, \dots, \tau_m I_{n_m}),$$

$$Q = \text{blkdiag}(\beta_1 Q_1, \dots, \beta_m Q_m).$$

Since $\text{prox}_{\tilde{\mathsf{h}}^\star}(\mathsf{y}, \mathsf{w}) = (\text{prox}_{\mathsf{h}^\star}(\mathsf{y}), \mathsf{w} - \mathcal{P}_C(\mathsf{w}))$ (using $\text{prox}_{\delta_C}(\cdot) = \mathcal{P}_C(\cdot)$ along with Moreau decomposition [13, Thm. 14.3]) the proximal updates of TriPD (Alg. 4.1), *cf.* (4.6), become:

$$\bar{y}_i = \text{prox}_{\sigma_i h_i^\star}(y_i + \sigma_i L_i x_i),$$

$$\bar{w}_{(i,j)} = w_{(i,j)} + \kappa_{(i,j)}(N_{(i,j)}\mathsf{x} - \mathcal{P}_{C_{(i,j)}}(\kappa_{(i,j)}^{-1} w_{(i,j)} + N_{(i,j)}\mathsf{x})),$$

$$\bar{x}_i = \text{prox}_{\tau_i g_i}(x_i - \tau_i L_i^\top \bar{y}_i - \tau_i(\mathsf{N}^\top\bar{\mathsf{w}})_i - \tau_i \nabla f(x_i)).$$

Note that for $w_1, w_2 \in \mathbb{R}^{l_{(i,j)}}$ the projection onto $C_{(i,j)}$ is

$$\mathcal{P}_{C_{(i,j)}}(w_1, w_2) = \frac{1}{2}\big(w_1 - w_2 + b_{(i,j)}, -w_1 + w_2 + b_{(i,j)}\big).$$

By assigning to agent i the primal coordinate x_i and dual coordinate y_i and $w_{(i,j),i}$ for all $j \in \mathcal{N}_i$, TriPD-Dist (Alg. 4.3) is obtained. Note that this assignment entails non-overlapping sets of coordinates, *i.e.*, Assumption 4.III*(i)* is satisfied.

The convergence results of TriPD-Dist (Alg. 4.3) are provided separately for the synchronous and asynchronous schemes in the next two theorems, along with a sufficient condition for linear convergence. The proofs follow directly from Theorems 4.10 and 4.11.

Theorem 4.13 (convergence of Algorithm 3-I)**.** *Let Assumptions 4.V and 4.VI hold. The sequence* $(\mathsf{z}^k)_{k\in\mathbb{N}} = (\mathsf{y}^k, \mathsf{w}^k, \mathsf{x}^k)_{k\in\mathbb{N}}$ *generated by Algorithm 3-I converges to some* $z^\star \in \mathcal{S}$*. Furthermore, if* f_i*,* g_i *and* h_i*,* $i = 1, \dots, m$ *are PLQ, then* $(\mathrm{dist}_S(\mathsf{z}^k, \mathcal{S}))_{k\in\mathbb{N}}$ *converges Q-linearly to zero, and* $(\mathsf{z}^k)_{k\in\mathbb{N}}$ *converges R-linearly to* $z^\star \in \mathcal{S}$*.*

Theorem 4.14 (convergence of Algorithm 3-II)**.** *Let Assumptions 4.V and 4.VI hold. The sequence* $(\mathsf{z}^k)_{k\in\mathbb{N}} = (\mathsf{y}^k, \mathsf{w}^k, \mathsf{x}^k)_{k\in\mathbb{N}}$ *generated by Algorithm 3-II converges almost surely to some* $z^\star \in \mathcal{S}$*. Furthermore, if* f_i*,* g_i *and* h_i*,* $i = 1, \dots, m$ *are PLQ and* $(\mathsf{z}^k)_{k\in\mathbb{N}} \subseteq \mathcal{C}$ *where* $\mathcal{C}$ *is a compact set, then* $(\mathbb{E}\left[\mathrm{dist}^2_{\Pi^{-1}S}(\mathsf{z}^k, \mathcal{S})\right])_{k\in\mathbb{N}}$ *converges Q-linearly to zero.*

4.6 Application: formation control

In this section we consider the problem of formation control of a group of robots [136, 148], where each robot/agent has its own local dynamics and cost function and the goal is to achieve a specific formation by communicating only with neighboring agents.

For simplicity of visualization we consider a 2D problem. Each subsystem (corresponding to a robot) has four states $x_i = (p_{x_i}, p_{y_i}, v_{x_i}, v_{y_i})$, where (p_{x_i}, p_{y_i}) and (v_{x_i}, v_{y_i}) denote the position and the velocity vectors, respectively. The input for each system is given by $u_i = (v^u_{x_i}, v^u_{y_i})$. The discrete-time LTI model of each system is given by

$$x_i(k+1) = \Phi_i x_i(k) + \Delta_i u_i(k), \quad k = 0, 1, \dots.$$

The state and input transition matrices are as follows

$$\Phi_i = \begin{pmatrix} I & 0 & X_1 & 0 \\ 0 & I & 0 & X_1 \\ 0 & 0 & X_2 & 0 \\ 0 & 0 & 0 & X_2 \end{pmatrix}, \quad \Delta_i = \begin{pmatrix} X_3 & 0 \\ 0 & X_3 \\ X_1 & 0 \\ 0 & X_1 \end{pmatrix},$$

where the parameters are given by $X_1 = -t_d(e^{-\frac{1}{t_d}} - 1)$, $X_2 = e^{-\frac{1}{t_d}}$ and $X_3 = t_d^2(e^{-\frac{1}{t_d}} - 1 + \frac{1}{t_d})$ with time constant $t_d = 5$ (s). This discrete-time model was derived from the continuous-time model of [148] using exact discretization with step length $\Delta T = 1$.

Let N denote the horizon length. Consider the stacked state and input vectors $\boldsymbol{x}_i \in \mathbb{R}^{4N}, \boldsymbol{u}_i \in \mathbb{R}^{2N}$:

$$\boldsymbol{x}_i \coloneqq (x_i(1), \ldots, x_i(N)), \ \boldsymbol{u}_i \coloneqq (u_i(0), \ldots, u_i(N-1)).$$

Then the dynamics of each agent can be represented as $\mathcal{A}_i \boldsymbol{x}_i + \mathcal{B}_i \boldsymbol{u}_i = b_i$ where $\mathcal{A}_i$, $\mathcal{B}_i$ are appropriate matrices and b_i depends on the initial state. The state and input constraints of each agent are represented by the sets $\mathcal{X}_i$, $\mathcal{U}_i$ and are assumed to be easy to project onto, *e.g.*, boxes, halfspaces, norm balls, etc. Moreover, we assume that each agent has its own private objective captured by input and state cost matrices $\mathcal{Q}_i$ and $\mathcal{R}_i$, and vectors q_i, t_i. The specific formation between agents is enforced using another quadratic term that penalizes deviation of two neighbors from the desired relative position. The optimization problem is described as follows:

$$\begin{aligned} \underset{\boldsymbol{x}_i, \boldsymbol{u}_i}{\text{minimize}} \ & \sum_{i=1}^{m} \tfrac{1}{2}\|\mathcal{Q}_i \boldsymbol{x}_i - q_i\|^2 + \tfrac{1}{2}\|\mathcal{R}_i \boldsymbol{u}_i - t_i\|^2 \\ & + \sum_{i=1}^{m} \sum_{j \in \mathcal{N}_i} \tfrac{\lambda_i}{2}\|\mathcal{C}(\boldsymbol{x}_i - \boldsymbol{x}_j) - d_{ij}\|^2 \\ \text{subject to} \ & \mathcal{A}_i \boldsymbol{x}_i + \mathcal{B}_i \boldsymbol{u}_i = b_i, \ \boldsymbol{x}_i \in \mathcal{X}_i, \ \boldsymbol{u}_i \in \mathcal{U}_i \\ & i = 1, \ldots, m \end{aligned} \tag{4.52}$$

The relative desired distance of agent i from its neighbor j is given by d_{ij}, $\mathcal{C}$ is an appropriate linear mapping that selects the position variables, and λ_i is a scalar weight to penalize deviation.

For each system that communicates with i, *i.e.*, $j \in \mathcal{N}_i$, we introduce a local variable $\boldsymbol{x}_{ij}$, that can be seen as the estimate of $\boldsymbol{x}_j$ kept locally by agent i. In order to be consistent hereafter the self variables $\boldsymbol{x}_i, \boldsymbol{u}_i$ are denoted by $\boldsymbol{x}_{ii}, \boldsymbol{u}_{ii}$.

For each agent $i = 1, \dots, m$ define the stacked vector

$$z_{\mathcal{N}_i} = \big((\boldsymbol{x}_{ij})_{j\in\mathcal{N}_i\cup\{i\}}, \boldsymbol{u}_{ii}\big) \in \mathbb{R}^{n_i},$$

where $n_i = 4N(|\mathcal{N}_i| + 1) + 2N$.

Let E_i be a linear mapping such that $E_i z_{\mathcal{N}_i} = \mathcal{A}_i \boldsymbol{x}_{ii} + \mathcal{B}_i \boldsymbol{u}_{ii}$. Hence, the set of points satisfying the dynamics are given by $\mathcal{D}_i = \{z \in \mathbb{R}^{n_i} | E_i z = b_i\}$. Consider the linear mapping L_i such that $L_i z_{\mathcal{N}_i} = (\boldsymbol{x}_{ii}, \boldsymbol{u}_{ii})$ and denote $\mathcal{Z}_i \coloneqq \mathcal{X}_i \times \mathcal{U}_i$. Moreover, let $h_i \coloneqq \delta_{\mathcal{Z}_i}$, $g_i \coloneqq \delta_{\mathcal{D}_i}$ and

$$f_i(z_{\mathcal{N}_i}) \coloneqq \tfrac{1}{2}\|\mathcal{Q}_i \boldsymbol{x}_{ii} - q_i\|^2 + \tfrac{1}{2}\|\mathcal{R}_i \boldsymbol{u}_{ii} - t_i\|^2 + \tfrac{\lambda_i}{2}\textstyle\sum_{j\in\mathcal{N}_i} \|\mathcal{C}(\boldsymbol{x}_{ii} - \boldsymbol{x}_{ij}) - d_{ij}\|^2.$$

With these definitions problem (4.52) is cast in the form of problem (4.49) (minimizing over $z_{\mathcal{N}_i}$, $i = 1, \dots, m$) where the linear mapping A_{ij}, for $j \in \mathcal{N}_i$, is such that $A_{ij} z_{\mathcal{N}_i} = (\boldsymbol{x}_{ii}, -\boldsymbol{x}_{ij})$ if $i < j$ and $A_{ij} z_{\mathcal{N}_i} = (-\boldsymbol{x}_{ij}, \boldsymbol{x}_{ii})$ otherwise. Therefore, we can readily apply TriPD-Dist (Alg. 4.3) to solve the problem in a fully distributed fashion yielding both synchronous and randomized asynchronous implementations.

In our simulations we used horizon length $N = 3$. For the input and state constraints of all agents we used box constraints: the positions p_{x_i} and p_{y_i} are assumed to be between 0 and 20 (m). The velocities v_{x_i} and v_{y_i} and inputs $v^u_{x_i}$ and $v^u_{y_i}$ are assumed to be between between 0 and 15 (m/s) (for all agents). The local state cost matrices are set $\mathcal{Q}_i = 0.1I$ for all i. The local input cost matrices are set $\mathcal{R}_i = I$ for half of the agents and $\mathcal{R}_i = 2I$ for the rest. Moreover, the vectors q_i, t_i are set equal to zero, and the penalty parameter $\lambda_i = 10$ is used for all the agents.

The stepsizes of TriPD-Dist (Alg. 4.3) were selected as follows: i) (edge stepsizes) $\kappa_{(i,j)} = 1$ for all $(i, j) \in \mathcal{E}$, ii) (node stepsizes) $\sigma_i = \beta_i/4$ and $\tau_i = 0.99/(\frac{\beta_i}{2} + \sigma_i + \sum_{j\in\mathcal{N}_i} \kappa_{(i,j)})$ for all i, where we used

$$\beta_i = \max\{\|\mathcal{Q}_i^\top \mathcal{Q}_i\| + \lambda_i(|\mathcal{N}_i| + 1), \|\mathcal{R}_i^\top \mathcal{R}_i\|\},$$

which is an upper bound for the Lipschitz constant of ∇f_i. It is plain to see that the above choice of stepsizes for the agents satisfy Assumption 4.VI*(iii)*. Note that the stepsize selection only requires local parameters $\mathcal{R}_i$, $\mathcal{Q}_i$, λ_i and the number of neighbors $|\mathcal{N}_i|$, *i.e.*, the algorithm can be implemented without *any* global coordination.

In our simulations, we considered m robots initially in a polygon configuration and enforced an arrow formation by appropriate selection of d_{ij} in (4.52). This scenario is depicted for $m = 5$ in Figure 4.2. The neighborhood relation in

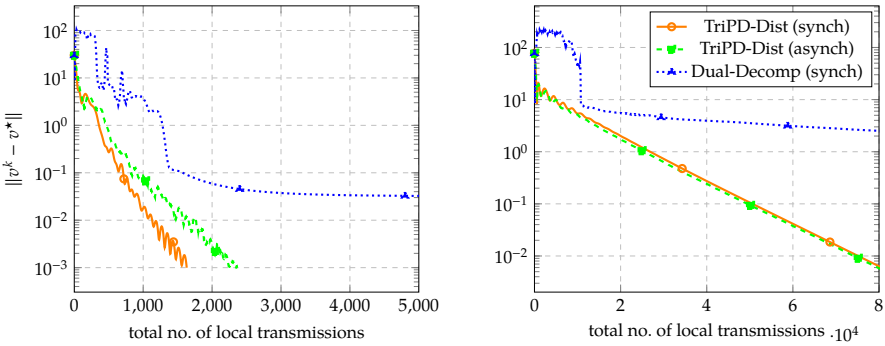

Figure 4.1: *Comparison for the convergence of the algorithms for $m = 5$ (left), and $m = 50$ (right).*

this case is taken to be the same arrow configuration, *i.e.*, all agents have two neighbors apart from two agents with only one neighbor.

For comparison we considered the dual decomposition approach of [136] (based on the subgradient method). Notice that dual decomposition with gradient or accelerated gradient methods can not be applied to this problem since f_i's are convex but not strongly convex. Recently, TriPD-Dist (Alg. 4.3) was compared against the dual accelerated proximal gradient method, in the context of distributed model predictive control (with strongly convex quadratic cost) [94].

In the simulations for Figure 4.1, we used the stepsize $10/k$ (as tuned for achieving better performance) for the dual decomposition method where k is the number of iterations. Notice that the dual decomposition approach for this problem can not achieve a full splitting of the operators involved: at every iteration agents need to solve an inner minimization (we used MATLAB's `quadprog` to perform this step), the result of which must be communicated to the neighbors for their computation, and is followed by another communication round. This extra need for synchronization would further slow down the algorithm in practice [73].

Figure 4.1 demonstrates the superior performance of both the synchronous and asynchronous versions of TriPD-Dist (Alg. 4.3) compared to the dual decomposition approach. The y-axis is the distance of $v^k := (\boldsymbol{x}_{11}^k, \boldsymbol{u}_{11}^k, \ldots, \boldsymbol{x}_{mm}^k, \boldsymbol{u}_{mm}^k)$ from the solution ($v^\star$ was computed by solving (4.52) in a centralized fashion). The x-axis denotes the total number of local transmissions between agents. In the asynchronous implementation we used independent activation probabilities $p_i = 0.5$ for all agents. It is observed that the total number of local iterations is similar to that of the synchronous implementation. Finally, as evident in Figure

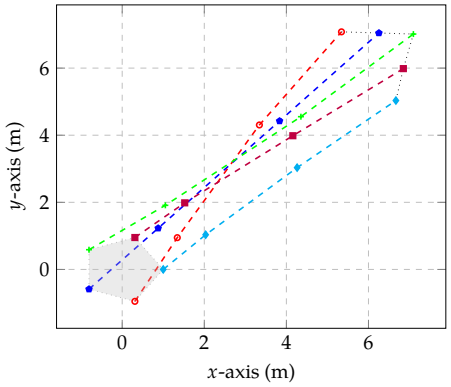

Figure 4.2: *Five agents reorganizing from a polygon to an arrow configuration.*

4.1 both versions of TriPD-Dist (Alg. 4.3) achieve linear convergence rate as predicted by Theorems 4.13 and 4.14 (the functions f_i, g_i and h_i are PLQ).

4.7 Conclusions

The primal-dual algorithm considered in this chapter enjoys several structural properties that distinguish it from other related methods in the literature. A key property, that has been instrumental in developing a block-coordinate version of the algorithm, is the fact that the generated sequence is S-Fejér monotone, where S is a block diagonal positive definite matrix. It is shown that the algorithm attains linear convergence under a metric subregularity assumption that holds for a wide range of cost functions that are not necessarily strongly convex. The block-coordinate version of the developed algorithm is exploited to devise a novel fully distributed asynchronous method for multi-agent optimization over graphs. Our future work includes designing a block-coordinate version of the *SuperMann* scheme of [157] that applies to quasi-nonexpansive operators. In light of the fact that this method enjoys superlinear convergence rates, such extension is especially attractive for multi-agent optimization yielding schemes with faster convergence and fewer communication rounds. Other research directions enlist investigating extensions to account for directed and time-varying topologies, communication delays, and designing efficient strategies for selecting activation probabilities and stepsizes.

Chapter 5

Plug and play distributed model predictive control

This chapter is based on:

Latafat, P., Bemporad, A., and Patrinos, P. *Plug and play distributed model predictive control with dynamic coupling: A randomized primal-dual proximal algorithm.* In European Control Conference (ECC) (June 2018), pp. 1160–1165.

5.1 Introduction

This chapter considers distributed model predictive control (DMPC) of a network of m dynamically coupled linear systems. For $i = 1, \dots, m$, the dynamics of system i is of the form

$$x_i(k+1) = \sum_{j=1}^{m} \Phi_{ij} x_j(k) + \Delta_{ij} u_j(k),$$

with $x_i(k) \in \mathbb{R}^{s_i}$, $u_i(k) \in \mathbb{R}^{t_i}$, subject to local state and input constraints. The structure of the network is defined by the coupling of the dynamics through matrices Φ_{ij} and Δ_{ij}. System j affects i if either one of Φ_{ij}, Δ_{ij} is nonzero. It is natural to assume that two systems can communicate if either one of them affects the dynamics of the other, in which case we say that they are neighbors. However, the systems need not be aware of the global structure of the network, or even existence of systems beyond their neighbors.

DMPC formulations considered in the literature vary depending on the nature of the coupling and can be grouped in two general categories. In applications

such as formation control where the systems are physically separate but share a common goal the DMPC problem involves coupling cost or constraints without dynamic coupling [171, 65, 173]. The second category involves DMPC problems with dynamic coupling with applications ranging from smart grids, sensor networks, water networks to transportation systems, and has been studied by many authors [79, 16, 77, 86, 69, 159]. This chapter is focused on the second category. We note that in our setting it is straightforward to extend the proposed setup to include coupling in cost and constraint between neighbors, however, this leads to complicated notation and has been avoided for the sake of clarity. Furthermore, this chapter is not concerned with the stability of the closed-loop system and looks at the DMPC problem from the optimization point of view.

A popular approach for solving the DMPC problem is to derive distributed algorithms using dual decomposition. Many authors have considered solving the dual problem using the proximal gradient method, the alternating direction method of multipliers (ADMM) or their variants [124, 77, 69, 86, 79]. These approaches are preferred to subgradient methods given that they allow constant stepsizes. Algorithms that are based on proximal gradient or its accelerated variants require the cost function to be strongly convex. Another common issue is the need for centralized computations for selecting the stepsizes. This is a major drawback that can hinder the implementation especially in applications where the network structure is subject to change. For example, applying proximal gradient requires the stepsize to be bounded by the inverse of the Lipschitz constant associated to the dual function [79]. In [76] a metric for Lipschitz continuity is used which requires solving a semidefinite program (SDP) globally. In [77] the authors provide a distributed method for selecting the metric that involves solving a series of local SDPs. Another recent work that involves distributed stepsize selection is [86] where the Lagrangian minimization step is modified with regularization terms. Each iteration in [77] and [86] involve a local inner minimization step the result of which is required by the neighbors, *i.e.*, each iteration involves two rounds of communication.

The main contributions are summarized below:

- The new algorithm is *fully*-distributed, involves simple computations for each subsystem without any inner loops, and requires one round of communication per update. At every iteration active subsystems perform local updates, communicate the necessary vectors to their neighbors, and go idle. The algorithm is presented in two forms: The synchronous case where all of the systems are active at every iteration, and the asynchronous case where subsystems are activated at random independently of one another.
- The stepsize of each subsystem is selected locally through a simple rule (cf. Assumption 5.II*(iii)*). Therefore, any modification to the network structure

would only affect the neighboring subsystems.

- The cost function must be convex but not necessarily strongly convex.
- The algorithm possesses linear convergence rate when the local input and state constraints are polyhedral sets, a common scenario.

5.2 Problem setup

We consider a distributed model predictive control problem with m dynamically coupled subsystems. We use an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to model the interaction between subsystems/agents. Each node $i \in \mathcal{V}$ is associated with a *subsystem*, maintains its own local variables, and can communicate with its neighbors. The goal is to solve the global model predictive control problem with only local exchange of information between neighbors.

Let Φ_{ij} and Δ_{ij} denote the state transition and input matrices from subsystem j to i. For all $i \in \mathcal{V}$, the *in-neighbor* and *out-neighbor* sets are defined by

$$\mathcal{N}_i^{\text{in}} = \{j \in \mathcal{V} \setminus \{i\} | \Phi_{ij} \neq 0 \text{ or } \Delta_{ij} \neq 0\},$$

$$\mathcal{N}_i^{\text{out}} = \{j \in \mathcal{V} \setminus \{i\} | \Phi_{ji} \neq 0 \text{ or } \Delta_{ji} \neq 0\},$$

and the neighborhood set is defined by $\mathcal{N}_i = \mathcal{N}_i^{\text{out}} \cup \mathcal{N}_i^{\text{in}}$, *i.e.*, the edge $(i, j) \in \mathcal{E}$ exists if $j \in \mathcal{N}_i$. The DMPC problem is written in the following standard form:

$$\begin{aligned}
\text{minimize} \quad & \frac{1}{2} \sum_{i=1}^{m} \Big(\sum_{k=1}^{N} x_i(k)^\top Q_i^k x_i(k) + \sum_{k=0}^{N-1} u_i(k)^\top R_i^k u_i(k) \Big) \\
\text{subject to} \quad & x_i(k+1) = \sum_{j \in \mathcal{N}_i^{\text{in}} \cup \{i\}} \Phi_{ij} x_j(k) + \Delta_{ij} u_j(k) \qquad (5.1a) \\
& u_i(k) \in \mathcal{U}_i, \quad \text{for } k = 0, \ldots, N-1, \\
& x_i(k) \in \mathcal{X}_i, \quad \text{for } k = 1, \ldots, N-1, \\
& x_i(N) \in \mathcal{X}_i^f \\
& \text{for all } i = 1, \ldots, m
\end{aligned}$$

where $x_i(0)$ is given, $x_i(k) \in \mathbb{R}^{s_i}$ and $u_i(k) \in \mathbb{R}^{t_i}$ denote the state and input variables of subsystem i at time k.

Note that the separable quadratic cost function is used for clarity of exposition. It may be replaced by any Lipschitz differentiable function without requiring the subsystems to solve inner minimizations (*cf.* Rem. 5.1). Furthermore, it is straightforward to modify our analysis to allow coupling between neighbors.

Throughout the chapter the following assumptions hold:

Assumption 5.I. For $i = 1, \dots, m$:

(i) Input and state constraint sets $\mathcal{X}_i, \mathcal{X}_i^f \subseteq \mathbb{R}^{s_i}$ and $\mathcal{U}_i \subseteq \mathbb{R}^{t_i}$ are nonempty, closed, and convex.

(ii) The cost matrices Q_i^k and R_i^k are positive semidefinite.

(iii) The graph $\mathcal{G}$ is connected.

(iv) The DMPC problem admits a solution. Moreover, for $i = 1, \dots, m$ there exists $x_i(k) \in \operatorname{ri} \mathcal{X}_i$ for $k = 1, \dots, N-1$, $x_i(N) \in \operatorname{ri} \mathcal{X}_i^f$, and $u_i(k) \in \operatorname{ri} \mathcal{U}_i$ for $k = 0, \dots, N-1$ such that the linear dynamics (5.1a) are satisfied.

The strict feasibility enforced in Assumption 5.I*(iv)* ensures that strong duality holds, and can be dropped whenever the constraint sets are polyhedral [143, Corollary 31.2.1].

For $i = 1, \dots, m$ define

$$z_i = \big(x_i(1), \cdots, x_i(N), u_i(0), \cdots, u_i(N-1)\big) \in \mathbb{R}^{r_i},$$

where $r_i = N(s_i + t_i)$. The quadratic cost function can be written as $\frac{1}{2}\sum_{i=1}^m z_i^\top G_i z_i$ where $G_i = \operatorname{blkdiag}(Q_i^1, \dots, Q_i^N, R_i^0, \dots, R_i^{N-1})$. The dynamics can be expressed as:

$$\sum_{j \in \mathcal{N}_i^{\text{in}} \cup \{i\}} L_{ij} z_j = \sum_{j \in \mathcal{N}_i^{\text{in}} \cup \{i\}} b_{ij} x_j(0), \quad \text{for } i = 1, \cdots, m,$$

where L_{ij} and b_{ij} are appropriate linear mappings [76]. With these definitions the distributed MPC problem becomes

$$\text{minimize} \quad \frac{1}{2}\sum_{i=1}^m z_i^\top G_i z_i \tag{5.2a}$$

$$\text{subject to} \quad \sum_{j \in \mathcal{N}_i^{\text{in}} \cup \{i\}} L_{ij} z_j = b_i, \quad i = 1, \cdots, m \tag{5.2b}$$

$$z_i \in \mathcal{Z}_i, \quad i = 1, \cdots, m \tag{5.2c}$$

where $b_i = \sum_{j \in \mathcal{N}_i^{\text{in}} \cup \{i\}} b_{ij} x_j(0)$, and the constraint sets $\mathcal{Z}_i$ denote the product of local input and state constraint sets:

$$\mathcal{Z}_i = \underbrace{\mathcal{X}_i \times \ldots \times \mathcal{X}_i}_{N-1} \times \mathcal{X}_i^f \times \underbrace{\mathcal{U}_i \times \ldots \times \mathcal{U}_i}_{N}.$$

5.3 A primal-dual algorithm for DMPC

Our goal is to solve (5.2) in a fully distributed fashion while keeping the number of communications to a minimum. For each subsystem that affects i, *i.e.*, $j \in \mathcal{N}_i^{\text{in}}$, we introduce a local variable z_{ij}, that can be seen as the estimate of z_j kept locally by agent i. For notation consistency, self-variables z_i are hereafter denoted by z_{ii}. We write the equivalent optimization problem:

$$\text{minimize} \quad \frac{1}{2} \sum_{i=1}^{m} z_{ii}^\top G_i z_{ii} \tag{5.3a}$$

$$\text{subject to} \quad \sum_{j \in \mathcal{N}_i^{\text{in}} \cup \{i\}} L_{ij} z_{ij} = b_i, \quad i = 1, \ldots, m \tag{5.3b}$$

$$z_{ii} \in \mathcal{Z}_i, \quad i = 1, \ldots, m \tag{5.3c}$$

$$z_{ij} = z_{jj}, \quad i = 1, \ldots, m \text{ and } j \in \mathcal{N}_i^{\text{in}} \tag{5.3d}$$

For $i \in \mathcal{V}$, let $n_i = \sum_{j \in \mathcal{N}_i^{\text{in}} \cup \{i\}} r_j$ and define:

$$z_{\mathcal{N}_i} = (z_{ij})_{j \in \mathcal{N}_i^{\text{in}} \cup \{i\}} \in \mathbb{R}^{n_i}, \; L_i = [L_{ij}]_{j \in \mathcal{N}_i^{\text{in}} \cup \{i\}} \in \mathbb{R}^{N s_i \times n_i}$$

The set of points satisfying the linear constraint is given by:

$$\mathcal{D}_i = \{z \in \mathbb{R}^{n_i} | L_i z = b_i\}.$$

We note that the variables are stacked in ascending order (index-wise). For example, consider the neighborhood relation described in Figure 5.1. Subsystem 1 is affected by subsystems 2 and 4, therefore, $z_{\mathcal{N}_1} = (z_{11}, z_{12}, z_{14})$. Our proposed algorithm is a primal-dual scheme. Therefore, in addition to primal variables each system holds dual variables. For each $i \in \mathcal{V}$ we introduce two sets of dual variables: the node variable $y_i \in \mathbb{R}^{r_i}$ and the edge variables $w_{ij,i} \in \mathbb{R}^{r_i}$ for $j \in \mathcal{N}_i^{\text{out}}$, and $w_{ji,i} \in \mathbb{R}^{r_j}$ for $j \in \mathcal{N}_i^{\text{in}}$. The first argument of the subscript denotes the edge relation and the second the ownership of the variable, *i.e.*, if system i affects j, then i and j will keep $w_{ij,i}$ and $w_{ij,j}$ respectively.

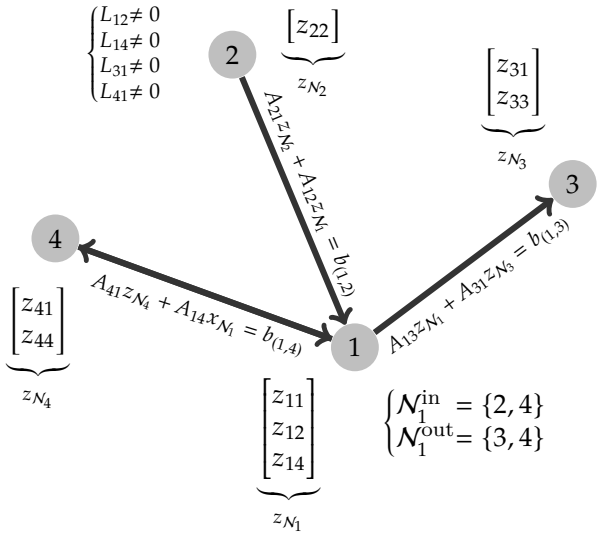

Figure 5.1: *Dynamic coupling in the DMPC problem*

Let $E_i \in \mathbb{R}^{r_i \times n_i}$ be a linear mapping such that $E_i z_{\mathcal{N}_i} = z_{ii}$. Define

$$g_i(z_{\mathcal{N}_i}) = \delta_{\mathcal{D}_i}(z_{\mathcal{N}_i}) + \tfrac{1}{2} z_{\mathcal{N}_i}^\top E_i^\top G_i E_i z_{\mathcal{N}_i}, \quad h_i = \delta_{\mathcal{Z}_i}, \tag{5.4}$$

where δ_X denotes the indicator function of a closed nonempty convex set, X. Problem (5.3) becomes

$$\text{minimize} \quad \sum_{i=1}^{m} g_i(z_{\mathcal{N}_i}) + h_i(E_i z_{\mathcal{N}_i}) \tag{5.5a}$$

$$\text{subject to} \quad A_{ij} z_{\mathcal{N}_i} + A_{ji} z_{\mathcal{N}_j} = 0 \qquad (i,j) \in \mathcal{E} \tag{5.5b}$$

where $A_{ij} \in \mathbb{R}^{l_{(i,j)} \times n_i}$ is defined based on the neighborhood relation as follows

$$A_{ij} z_{\mathcal{N}_i} = \begin{cases} z_{ii} & j \notin \mathcal{N}_i^{\text{in}},\ j \in \mathcal{N}_i^{\text{out}} \\ -z_{ij} & j \in \mathcal{N}_i^{\text{in}},\ j \notin \mathcal{N}_i^{\text{out}} \\ (z_{ii}, -z_{ij}) & j \in \mathcal{N}_i^{\text{in}},\ j \in \mathcal{N}_i^{\text{out}}. \end{cases} \tag{5.6}$$

Notice that depending on the neighborhood relation $l_{(i,j)}$ is either equal to r_i, r_j or $r_i + r_j$.

A primal-dual algorithm was introduced in [100] (see Chapter 6) for problems of the form (5.5) with consensus constraint. However, a consensus constraint can

not capture the coupling in the DMPC problem depicted in Figure 5.1. Another drawback of the aforementioned work is that the stepsize selection requires global coordination. Our analysis here is different from that work and is based on [95] (see Chapter 4). In Section 5.4 we describe how TriPD-Dist (Alg. 4.3) is applied to the DMPC problem to derive Algorithm 5.1.

Our proposed distributed scheme is summarized in Algorithm 5.1. It involves two versions. In the synchronous case at each iteration all systems perform their local updates and broadcast the result to the relevant neighbors. In the asynchronous case each system wakes up randomly independent of other systems, *i.e.*, there may be several active systems at each iteration. The stepsizes appearing in Algorithm 5.1 should satisfy the following:

Assumption 5.II (stepsizes in Algorithm 5.1)**.**

(i) (node stepsizes) Subsystem i keeps two positive stepsizes σ_i, τ_i associated to h_i and g_i, respectively.

(ii) (edge stepsizes) For each edge $(i,j) \in \mathcal{E}$ we associate a positive stepsize $\kappa_{(i,j)}$ that is shared between system i and j.

(iii) (convergence condition) The stepsizes satisfy the following local condition consensus

$$\tau_i < \frac{1}{\max\{\sum_{j\in\mathcal{N}_i^{\text{out}}} \kappa_{(i,j)} + \sigma_i, (\kappa_{(i,j)})_{j\in\mathcal{N}_i^{\text{in}}}\}}. \tag{5.7}$$

The dual updates for y_i in Algorithm 5.1 require projection onto the set $\mathcal{Z}_i$ which can often be performed efficiently, *e.g.* for boxes, halfspaces, norm balls. The primal updates are compactly written as $z_{\mathcal{N}_i} = \text{prox}_{\tau_i g_i}(c)$ where $c = (c_{ij})_{j\in\mathcal{N}_i^{\text{in}}\cup\{i\}}$ is given by

$$c_{ii} = z_{ii} - \tau_i\big(\bar{y}_i + \sum_{j\in\mathcal{N}_i^{\text{out}}} \bar{w}_{ij,i}\big), \tag{5.8a}$$

$$c_{ij} = z_{ij} + \tau_i \bar{w}_{ji,i}, \qquad \text{for all } j \in \mathcal{N}_i^{\text{in}}. \tag{5.8b}$$

The proximal mapping $\text{prox}_{\tau_i g_i}(c)$ involves the minimization of a strongly convex quadratic function over an affine subspace:

$$\underset{z}{\text{minimize}} \quad \tfrac{1}{2} z^\top (E_i^\top G_i E_i + \tfrac{1}{\tau_i} I_{n_i}) z - \tfrac{1}{\tau_i} c^\top z \tag{5.9a}$$

$$\text{subject to} \quad L_i z = b_i, \tag{5.9b}$$

Algorithm 5.1 Synchronous & asynchronous distributed primal-dual algorithm for DMPC

Inputs: $\sigma_i > 0$, $\tau_i > 0$ for $i = 1, \dots, m$, and $\kappa_{(i,j)} > 0$ for $(i,j) \in \mathcal{E}$
for $k = 0, 1, \dots$ **do**

I: Synchronous version	**II: Asynchronous version**
for all systems $i = 1, \dots, m$ **do**	draw r.v. ϵ_i^k according to $\mathbb{P}(\epsilon_i^0 = 1) = p_i > 0$ **for** all systems i with $\epsilon_i^k = 1$ **do**

Local updates:

$$
\begin{aligned}
\bar{w}_{ij,i}^k &= \tfrac{1}{2}\big(w_{ij,i}^k + w_{ij,j}^k\big) + \tfrac{\kappa_{(i,j)}}{2}\big(z_{ii}^k - z_{ji}^k\big), \forall j \in \mathcal{N}_i^{\text{out}} \\
\bar{w}_{ji,i}^k &= \tfrac{1}{2}\big(w_{ji,i}^k + w_{ji,j}^k\big) + \tfrac{\kappa_{(i,j)}}{2}\big(z_{jj}^k - z_{ij}^k\big), \forall j \in \mathcal{N}_i^{\text{in}} \\
\bar{y}_i^k &= y_i^k + \sigma_i z_{ii}^k - \sigma_i \mathcal{P}_{\mathcal{Z}_i}\big(\sigma_i^{-1} y_i^k + z_{ii}^k\big) \\
z_{\mathcal{N}_i}^{k+1} &= \text{prox}_{\tau_i g_i}(c) \text{ is given by (5.8) and (5.9).} \\
y_i^{k+1} &= \bar{y}_i^k + \sigma_i(z_{ii}^{k+1} - z_{ii}^k) \\
w_{ij,i}^{k+1} &= \bar{w}_{ij,i}^k + \kappa_{(i,j)}(z_{ii}^{k+1} - z_{ii}^k), \text{ for all } j \in \mathcal{N}_i^{\text{out}} \\
w_{ji,i}^{k+1} &= \bar{w}_{ji,i}^k - \kappa_{(i,j)}(z_{ij}^{k+1} - z_{ij}^k), \text{ for all } j \in \mathcal{N}_i^{\text{in}}
\end{aligned}
$$

Broadcast of information:
Send $z_{ii}^{k+1}, w_{ij,i}^{k+1}$ to $j \in \mathcal{N}_i^{\text{out}}$, and $z_{ij}^{k+1}, w_{ji,i}^{k+1}$ to $j \in \mathcal{N}_i^{\text{in}}$

and can be evaluated efficiently through solving the linear system defining its KKT optimality conditions. We stress that the matrix of the linear system is constant throughout iterations, and needs to be factored only once. Consequently, the evaluation of the primal step at every iteration amounts to forward and backward substitution steps [129, §III.C].

5.4 Deriving the algorithm and convergence results

In this section we detail the steps of applying Algorithms 4.1 and 4.2 to the DMPC problem.

Let $z = (z_{\mathcal{N}_1}, \dots, z_{\mathcal{N}_m})$ and define the linear operator

$$N_{(i,j)} : z \mapsto (A_{ij} z_{\mathcal{N}_i}, A_{ji} z_{\mathcal{N}_j}).$$

The edge constraints (5.5b) can be equivalently formulated in the cost as $\sum_{i=1}^m \delta_{C_{(i,j)}}\big(N_{(i,j)} z\big)$, where $C_{(i,j)} = \{(z_1, z_2) \in \mathbb{R}^{2l_{(i,j)}} \mid z_1 + z_2 = 0\}$. Conse-

quently, (5.5) can be formulated in the form of unconstrained optimization:

$$\text{minimize} \sum_{i=1}^{m} \big(g_i(z_{\mathcal{N}_i}) + h_i(E_i z_{\mathcal{N}_i})\big) + \sum_{(i,j)\in\mathcal{E}} \delta_{C_{(i,j)}}\big(N_{(i,j)} z\big)$$

In order to formulate the dual problem we introduce two sets of dual variables, $y_i \in \mathbb{R}^{r_i}$ and $w_{(i,j)} \in \mathbb{R}^{2l_{(i,j)}}$. The former corresponds to node and the latter to edge constraints. The edge variable $w_{(i,j)}$ consists of two blocks, $w_{(i,j)} = (w_{(i,j),i}, w_{(i,j),j})$, *i.e.*, we consider two dual variables for each constraint, where $w_{(i,j),i} \in \mathbb{R}^{l_{(i,j)}}$ is maintained by agent i and $w_{(j,i),j} \in \mathbb{R}^{l_{(i,j)}}$ by agent j. Notice that the edge variable $w_{(i,j),i}$ itself consists of either one or two blocks: $w_{(i,j),i} = (w_{ij,i}, w_{ji,i})$, where $w_{ij,i}$ and $w_{ji,i}$ are present when $j \in \mathcal{N}_i^{\text{out}}$ and $j \in \mathcal{N}_i^{\text{in}}$, respectively.

For clarity of exposition we rewrite the problem with compact notation. We use N without any subscript to denote the stacked linear mapping $\mathsf{N} = (N_{(i,j)})_{(i,j)\in\mathcal{E}}$, and $C = \bigtimes_{(i,j)\in\mathcal{E}} C_{(i,j)}$. The transpose of N is given by

$$\mathsf{N}^\top : (w_{(i,j)})_{(i,j)\in\mathcal{E}} \mapsto \tilde{\mathsf{z}} = \sum_{(i,j)\in\mathcal{E}} \mathsf{N}_{(i,j)}^\top w_{(i,j)},$$

with $\tilde{\mathsf{z}}_i = \sum_{j\in\mathcal{N}_i} A_{ij}^\top w_{(i,j),i}$. Furthermore, set $E = \operatorname{blkdiag}(E_1, \dots, E_m)$ and define $Lz = (Ez, \mathsf{N}z) =: (\tilde{y}, \tilde{w}) \in \mathbb{R}^{n_d}$, where $n_d = \sum_{(i,j)\in\mathcal{E}} 2l_{(i,j)} + \sum_{i=1}^m r_i$. Set $g(z) = \sum_{i=1}^m g_i(z_{\mathcal{N}_i})$, $\quad h(\tilde{y}, \tilde{w}) = \tilde{h}(\tilde{y}) + \delta_C(\tilde{w})$, where $\tilde{h}(\tilde{y}) = \sum_{i=1}^m h_i(\tilde{y}_i)$. Then, problem (5.5) can be casted as

$$\text{minimize} \quad g(z) + h(Lz). \tag{5.10}$$

Problem (5.10) may be solved by a range of primal-dual algorithms resulting in the full splitting of the nonsmooth functions and the linear mapping, see Chapter 3. Our goal is to derive algorithms in which: i) both the iterates and the stepsizes are computed locally, ii) involve one round of communication per iteration, iii) allow block coordinate updates. An ideal candidate for this purpose is the primal-dual algorithm introduced in [95, Alg. 1], see Chapter 4. In particular, the sequence generated by the algorithm is S-Fejér monotone where S is a block diagonal positive definite matrix.

Remark 5.1. In (5.4) the quadratic terms were captured by nonsmooth functions g_i. Our scheme requires calculating the proximal mapping of g_i which translates to solving the quadratic over affine minimization (5.9). Alternatively, one can model the quadratic cost functions using a third smooth term in (5.10) (see TriPD (Alg. 4.1)). This would result in a gradient step and a projection onto the set $\mathcal{D}_i$ in place of a quadratic over affine minimization. Hence, it is

possible to use general convex Lipschitz differentiable functions as cost in the DMPC problem. In that case the Lipschitz constant of the smooth term would affect the stepsizes. □

In order to represent the algorithm compactly we define the following set of diagonal matrices:

$$W = \operatorname{blkdiag}\left((\kappa_{(i,j)} I_{2l_{(i,j)}})_{(i,j)\in\mathcal{E}}\right),$$

$$\Sigma = \operatorname{blkdiag}(\sigma_1 I_{r_1}, \ldots, \sigma_m I_{r_m}),$$

$$\Gamma = \operatorname{blkdiag}(\tau_1 I_{n_1}, \ldots, \tau_m I_{n_m}).$$

Notice that $\kappa_{(i,j)}$ is repeated twice, *i.e.*, once for every node sharing the edge.

Let $v = (y, w, z)$, and define the operator T

$$Tv = (\bar{y} + \Sigma E(\bar{z} - z), \bar{w} + WN(\bar{z} - z), \bar{z}),$$

where

$$\bar{y} = \operatorname{prox}_{h^*}^{\Sigma^{-1}}(y + \Sigma E z) \tag{5.11a}$$

$$\bar{w} = \operatorname{prox}_{\delta_C^*}^{W^{-1}}(w + WNz) \tag{5.11b}$$

$$\bar{z} = \operatorname{prox}_{g}^{\Gamma^{-1}}(z - \Gamma E^\top \bar{y} - \Gamma N^\top \bar{w}). \tag{5.11c}$$

Then TriPD (Alg. 4.1) can be represented as the fixed-point iteration $v^{k+1} = Tv^k$. This iteration is amenable to *block coordinate* (BC) updates. A general BC scheme was proposed in TriPD-BC (Alg. 4.2). Our focus here is on the case where each coordinate has an independent probability to be active. Briefly put, the BC scheme is represented as

$$z^{k+1} = \sum_{i=1}^{m} \epsilon_i^k U_i (T z^k),$$

where U_i are diagonal matrices with zero and one diagonal elements, and are used to select the coordinates, while $\epsilon_i^k \in \{0, 1\}^m$ encodes if a coordinate i is updated at iteration k. The matrices U_i are assumed to be disjoint and $\sum_{i=1}^m U_i = \mathrm{I}$, where I is the identity matrix of appropriate dimensions. The partitioning described in this section satisfies these requirements, *i.e.*, for $i = 1, \ldots, m$, the matrix U_i selects $z_{\mathcal{N}_i}, y_i$ and $w_{(i,j),i}$ for $j \in \mathcal{N}_i$.

Since $\tilde{h}$ in (5.11a) is separable, using (5.4) and the Moreau identity [13], we have that for $i = 1, \ldots, m$

$$\bar{y}_i = y_i + \sigma_i z_{ii} - \mathcal{P}_{\mathcal{Z}_i}(y_i + \sigma_i z_{ii}),$$

and the projection onto $C_{(i,j)}$ is given by

$$\mathcal{P}_{C_{(i,j)}}(w_1, w_2) = \tfrac{1}{2}(w_1 - w_2, -w_1 + w_2).$$

Therefore, (5.11b) yields the updates for the edge variables in Algorithm 5.1. The $\bar{z}$ in (5.11c) can be evaluated as follows: For each $i \in \mathcal{V}$

$$\bar{z}_{\mathcal{N}_i} = \operatorname{prox}_{\tau_i g_i}\left(z_{\mathcal{N}_i} - \tau_i E_i^\top \bar{y}_i - \tau_i \sum_{j \in \mathcal{N}_i} A_{ij}^\top \bar{w}_{(i,j),i}\right).$$

Therefore, the primal update is carried out by solving (5.9) where c is given by (5.8). Finally, evaluation of the operator T requires matrix-vector products and straightforward substitution of the involved matrices yields Algorithm 5.1.

The next theorem summarizes the convergence results for Algorithm 5.1. The proof is omitted here and the interested reader is referred to Chapter 4.

Theorem 5.2. *Let Assumptions 5.I and 5.II hold. Consider the stacked vectors $z = (z_{\mathcal{N}_1}, \ldots, z_{\mathcal{N}_m})$, $y = (y_1, \ldots, y_m)$, $w = (w_{(i,j)})_{(i,j)\in\mathcal{E}}$. Then, in the case of synchronous updates, $(v^k)_{k\in\mathbb{N}} = (y^k, w^k, z^k)_{k\in\mathbb{N}}$ generated by Algorithm 5.1 converges to some $v^\star$, and in the case of asynchronous updates it converges almost surely to some $v^\star$-valued random variable, where $v^\star$ is a primal-dual solution to (5.5). In particular, $(z_{11}^k, \ldots, z_{mm}^k)_{k\in\mathbb{N}}$ converge to a solution of the DMPC problem (5.1). If in addition, $\mathcal{X}_i, \mathcal{X}^f$ and $\mathcal{U}_i$ are polyhedral sets then in the synchronous case the distance from the primal-dual solution set converges Q-linearly to zero.*

5.5 Numerical simulations

In this section, as a benchmark example we consider the problem of frequency control in power networks [141]. The network consists of power generation areas with the goal of maintaining nominal frequency levels despite changes in load and network configuration. The approach in [141] is based on modeling the dynamic coupling as disturbance. Clearly, this could lead to conservative control actions. In contrast our method solves the exact global optimization constrained by the dynamics through distributed computation and communication with the neighbors.

Each system consists of four states $x_i = (\Delta\theta_i, \Delta\omega_i, \Delta P_{\mathrm{m}_i}, \Delta P_{\mathrm{v}_i})$ and one control input $u_i = \Delta P_{\mathrm{ref}_i}$. The continuous-time LTI model of each system is given by

$$\dot{x}_i = \sum_{j \in \mathcal{N}_i^{\mathrm{in}} \cup \{i\}} A_{ij} x_j + B_i u_i.$$

Notice that the inputs are not coupled. The objective for each system is to track $x_i^r = (0, 0, \Delta P_{L_i}, \Delta P_{L_i})$ and $u_i^r = \Delta P_{L_i}$, where ΔP_{L_i} denotes the local power load. In our simulations we used five systems as described in Figure 5.2. The local constraints for each system are as follows: $\Delta\theta_i \in [-0.1, 0.1]$ for all i, and $\Delta P_{L_1}, \Delta P_{L_5} \in [-0.5, 0.5]$, $\Delta P_{L_2}, \Delta P_{L_3} \in [-0.65, 0.65]$, and $\Delta P_{L_4} \in [-0.55, 0.55]$. Furthermore, the quadratic costs $Q_i = 4I_{s_i}$ and $R_i = I_{t_i}$ are used for all systems along the horizon. We have omitted the details on the system dynamics here. The reader is referred to [141] and the references therein for details and parameter values. We used Euler's method for discretization of the dynamics with step length of 1 *sec*. This discretization has the advantage of maintaining the sparsity patterns of the transition matrices. In all our simulation we used horizon length $N = 20$.

In Algorithm 5.1 the stepsizes for each system must be selected in accordance to the simple condition of Assumption 5.II*(iii)*. Typically, in primal-dual proximal algorithms larger stepsizes yield faster convergence. However, there is a trade-off between edge parameters $\kappa_{(i,j)}$ and node parameters, σ_i, τ_i. We selected these values empirically as follows: i) $\kappa_{(i,j)} = 10$ for all $(i,j) \in \mathcal{E}$, ii) $\sigma_i = 1$ if $d_i^{\mathrm{out}} = 1$, and $10|d_i^{\mathrm{out}} - 1|$ otherwise, iii) $\tau_i = \frac{0.99}{\max\{10 d_i^{\mathrm{out}} + \sigma_i, 10\}}$, where d_i^{out} denotes the cardinality of $\mathcal{N}_i^{\mathrm{out}}$. Notice that due to this simple local rule, removal or addition of a node only affects the neighboring nodes through d_i^{out}.

Our simulations consist of two scenarios:

Scenario 1: In the first scenario we demonstrate the *plug and play* capability of our algorithm, *i.e.*, removal and addition of a new system only affects the network locally without the need for any global coordination. We consider systems $1, \ldots, 4$ with the dynamic coupling depicted in Figure 5.2. We assume that at time $t = 20$ system 5 is connected to systems 2 and 4. Furthermore, system 4 is disconnected from the network at time $t = 50$. Table 5.1 summarizes the load of power and network modification at given time steps. Figure 5.3 highlights the frequency deviation (the second state variable) for systems one and four. It is observed that the frequency control is achieved despite the load and configuration changes.

Scenario 2: In the second scenario, we considered a static network structure with 5 systems and load $\Delta P_{L_1} = 0.10$ with the same neighborhood structure and constraints as in the previous scenario. We compared our algorithm (referred

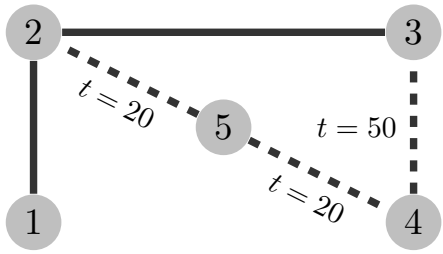

Figure 5.2: *Network structure in the DMPC problem for scenario* 1*: system* 5 *is added at* $t = 20$ *and system* 4 *is disconnected at* $t = 50$*.*

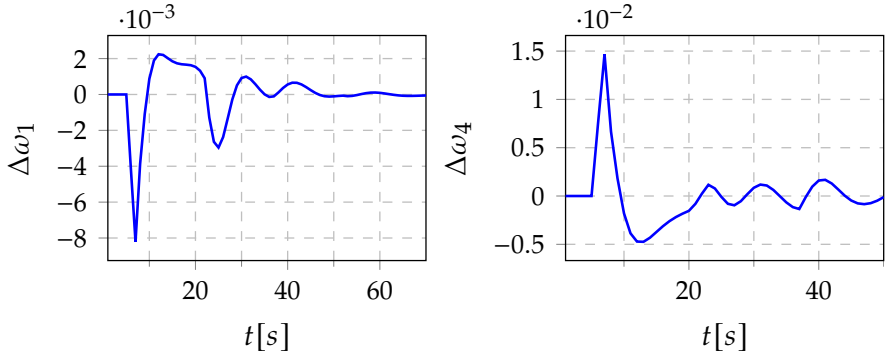

Figure 5.3: *Frequency deviation for systems one and four*

Table 5.1: *loads of power and network structure for Scenario* 1

time	5	5	20	20	35	35	50
system	1	4	2	5	5	3	4
ΔP_{L_i}	0.10	-0.12	0.08	added	0.05	-0.10	removed

here as PDDMPC) to [77, Alg. 3] (DGFG) that is based on applying the fast gradient method to the dual problem. The aforementioned paper proposes solving a series of convex semidefinite program (SDP) locally at the nodes in order to select the parameters of the algorithm in a distributed fashion. In order to have a fair comparison we solved the global optimization problem using MOSEK [5]. Figure 5.4 demonstrates the superior performance of our scheme. The y-axis is the error defined as the norm of the difference between current primal variables and the solution in both algorithms. The x-axis denotes the total number of local iterations. Notice that DGFG requires two rounds of communication at every iteration. Furthermore, we used the randomized version

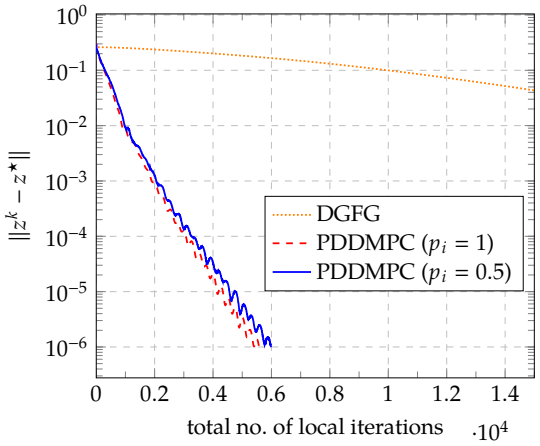

Figure 5.4: *Total number of local iterations: comparing synchronous PDDMPC, randomized PDDMPC and DGFG.*

of the algorithm where each system is activated independently with probability $p_i = 0.5$. It is observed that the random activation of nodes result in roughly the same number of total local iterations as the synchronous case.

5.6 Conclusions

This chapter introduced a fully distributed primal-dual proximal algorithm for the DMPC problem that includes both synchronous and randomized versions. In addition to simple local iterations, the stepsizes of the new algorithm are selected locally without any global coordination. Therefore, any changes to the network structure only affects the neighboring nodes. In addition, our algorithm enjoys a linear convergence rate under mild assumptions on the input and state constraints. Future works include devising efficient strategies for selecting the edge weights, and extending the algorithm for the case of lossy communications.

Chapter 6

A primal-dual proximal algorithm for distributed optimization over graphs

This chapter is based on:

Latafat, P., Stella, L., and Patrinos, P. *New primal-dual proximal algorithm for distributed optimization.* In 55th IEEE Conference on Decision and Control (CDC) (Dec 2016), pp. 1959–1964.

6.1 Introduction

In this chapter we deal with the distributed solution of the following optimization problem:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^{N} g_i(x) + h_i(C_i x) \tag{6.1}$$

where for $i = 1, \dots, N$, C_i is a linear operator, g_i and h_i are proper closed convex and possibly nonsmooth functions. We further assume that the *proximal mappings* associated with g_i and h_i are efficiently computable [49]. In a more general case we can include another continuously differentiable term with Lipschitz continuous gradient in (6.1) as in Chapter 3 but we opted not for clarity of exposition.

In machine learning and statistics the C_i are feature matrices and functions h_i measures the *fitting* of a predicted model with the observed data, while the g_i are *regularization* terms that enforce some prior knowledge in the solution (such as sparsity, or belonging to a certain constraint set). For example if h_i is the so-called hinge loss and $g_i = \frac{\lambda}{2}\|\cdot\|_2^2$, for some $\lambda > 0$, then one recovers

the standard SVM model. If instead $g_i = \lambda\|\cdot\|_1$ then one recovers the ℓ_1-norm SVM problem [185].

Problem (6.1) may be solved in a centralized fashion, when all the data of the problem (functions g_i, h_i and matrices C_i, for all $i \in \{1,\dots,N\}$) are available at one computing node. However, such a centralized approach is not realistic in many scenarios. For example, suppose that $h_i(C_i x)$ models least-squares terms and $C_1,\dots,C_N$ are very large features matrices. Then, collecting $C_1,\dots,C_N$ into a single computer may be infeasible due to communication costs, or even worse they may not fit into the computer's memory. Furthermore, the exchange of such information may not be possible at all due to privacy issues.

Our goal is therefore to solve problem (6.1) in a distributed fashion. Specifically, we consider a connected network of N computing agents, where the i-th agent is able to compute the proximal mappings of g_i, h_i, and matrix-vector products with C_i (and its adjoint operator). We want all the agents to iteratively converge to a *consensus* solution to (6.1), and to do so by only exchanging variables among neighbouring nodes, i.e, no centralized computations (*i.e.*, existence of a fusion center) are needed during the iterations.

To do so, we will propose a solution based on *Asymmetric Forward-Backward-Adjoint* (AFBA) splitting method, Algorithm 2.1. This splitting technique solves monotone inclusion problems involving three operators, however, in this chapter we will focus on a special case that involves two terms. Specifically, we develop a distributed algorithm which is based on a special case of AFBA applied to the monotone inclusion corresponding to the primal-dual optimality conditions of a suitable *graph splitting* of (6.1). Our algorithm involves a nonnegative parameter θ which serves as a tuning knob that allows to recover different algorithms. In particular, the algorithm of [38] is recovered in the special case when $\theta = 2$. We demonstrate how tuning this parameter affects the stepsizes and ultimately the convergence rate of the algorithm.

Other algorithms have been proposed for solving problems similar to (6.1) in a distributed way. As a reference framework, all algorithms aim at solving in a distributed way the problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^{N} F_i(x).$$

In [122] a distributed subgradient method is proposed, and in [64] this idea is extended to the projected subgradient method. More recently, several works focused on the use of ADMM for distributed optimization. In [29] the generic ADMM for consensus-type problems is illustrated. A drawback of this approach is that at every iteration the agents must solve a complicated subproblem that

might require an inner iterative procedure. In [131] another formulation is given for the case where $F_i = g_i + h_i$, and only proximal mappings with respect to g_i and h_i are separately computed in each node. Still, when either g_i or h_i is not separable (such as when they are composed with linear operators) these are not trivial to compute and may require inner iterative procedures, or factorization of the data matrices involved. Moreover, in both [29, 131] a central node is required for accumulating each agents variables at every iteration, therefore these formulations lead to *parallel* algorithms rather than distributed. In [153] the optimal parameter selection for ADMM is discussed in the case of distributed quadratic programming problems. In [174, 175, 115], fully distributed algorithms based on ADMM proposed, assuming that the proximal mapping of F_i is computable, which is impractical in many cases. In [21] the authors propose a variation of the Vũ-Condat algorithm [53, 170], having ADMM as a special case, and show its application to distributed optimization where $F_i = g_i + h_i$, but no composition with a linear operator is involved. Only proximal operations with respect to g_i and h_i and local exchange of variables (*i.e.*, among neighboring nodes) is required, and the method is analyzed in an asynchronous setting.

In this chapter we deal with the more general problem as in (6.1). The main features of our approach, that distinguish it from the related works mentioned above, are:

1) We deal with F_i that is the sum of two possibly nonsmooth functions one of which is composed with a linear operator.
2) Our algorithm only require local exchange of information, i.e., only neighboring nodes need to exchange local variables for the algorithms to proceed.
3) The iterations involve *direct* operations on the objective terms. Only evaluations of prox_{g_i}, $\mathrm{prox}_{h_i^\star}$ and matrix-vector products with C_i and C_i^T are involved. In particular, no inner subproblem needs to be solved iteratively by the computing agents, and no matrix inversions are required.

6.2 Problem formulation

Consider problem (6.1) under the following assumptions:

Assumption 6.I. For $i = 1, \ldots, N$:

(i) $C_i : \mathbb{R}^n \to \mathbb{R}^{r_i}$ are linear operators.

(ii) $g_i : \mathbb{R}^n \to \overline{\mathbb{R}}$, $h_i : \mathbb{R}^{r_i} \to \overline{\mathbb{R}}$ are proper closed convex functions.

(iii) The set of minimizers of (6.1), denoted by $S^\star$, is nonempty.

We are interested in solving problem (6.1) in a distributed fashion. Specifically, let $G = (V, E)$ be an undirected graph over the vertex set $V = \{1, \ldots, N\}$ with edge set $E \subset V \times V$. It is assumed that each node $i \in V$ is associated with a separate agent, and each agent maintains its own cost components g_i, h_i, C_i which are assumed to be private, and its own opinion of the solution $x_i \in \mathbb{R}^n$. The graph imposes communication constraints over agents. In particular, agent i can communicate directly only with its neighbors $j \in \mathcal{N}_i = \{j \in V \mid (i, j) \in E\}$. We make the following assumption.

Assumption 6.II. Graph G is connected.

With this assumption, we reformulate the problem as

$$\begin{aligned} \underset{\boldsymbol{x} \in \mathbb{R}^{Nn}}{\text{minimize}} \quad & \sum_{i=1}^{N} g_i(x_i) + h_i(C_i x_i) \\ \text{subject to} \quad & x_i = x_j \qquad (i, j) \in E \end{aligned}$$

where $\boldsymbol{x} = (x_1, \ldots, x_N)$. Associate any orientation to the unordered edge set E. Let $M = |E|$ and $B \in \mathbb{R}^{N \times M}$ be the *oriented node-arc incidence matrix*, where each column is associated with an edge $(i, j) \in E$ and has $+1$ and -1 in the i-th and j-th entry, respectively. Notice that the sum of each column of B is equal to 0. Let d_i denote the degree of a given vertex, that is, the number of vertices that are adjacent to it. We have $BB^\top = \mathcal{L} \in \mathbb{R}^{N \times N}$, where $\mathcal{L}$ is the graph *Laplacian* of G, i.e.,

$$\mathcal{L}_{ij} = \begin{cases} d_i & \text{if } i = j, \\ -1 & \text{if } i \neq j \text{ and node } i \text{ is adjacent to node } j, \\ 0 & \text{otherwise.} \end{cases}$$

Constraints $x_i = x_j$, $(i, j) \in E$ can be written in compact form as $A\boldsymbol{x} = 0$, where $A = B^\top \otimes I_n \in \mathbb{R}^{Mn \times Nn}$. Therefore, the problem is expressed as

$$\underset{\boldsymbol{x} \in \mathbb{R}^{Nn}}{\text{minimize}} \quad \sum_{i=1}^{N} g_i(x_i) + h_i(C_i x_i) + \delta_{\{0\}}(A\boldsymbol{x}). \tag{6.2}$$

The dual problem is:

$$\underset{\substack{y_i \in \mathbb{R}^{r_i} \\ \boldsymbol{w} \in \mathbb{R}^{Mn}}}{\text{minimize}} \quad \sum_{i=1}^{N} g_i^*(-A_i^\top \boldsymbol{w} - C_i^\top y_i) + h_i^*(y_i), \tag{6.3}$$

where $A_i \in \mathbb{R}^{Mn \times n}$ is the i-th (block) column of A. The primal-dual optimality conditions are

$$\begin{cases} 0 \in \partial g_i(x_i) + C_i^\top y_i + A_i^\top \boldsymbol{w}, & i = 1, \dots, N \\ C_i x_i \in \partial h_i^*(y_i) & i = 1, \dots, N \\ \sum_{i=1}^N A_i x_i = 0, \end{cases} \tag{6.4}$$

where $\boldsymbol{w} \in \mathbb{R}^{Mn}$, $y_i \in \mathbb{R}^{r_i}$, for $i = 1, \dots, N$. The following condition is assumed to hold throughout this chapter.

Assumption 6.III. There exist $x_i \in \operatorname{ri} \operatorname{dom} g_i$ such that $C_i x_i \in \operatorname{ri} \operatorname{dom} h_i$, $i = 1, \dots, N$ and $\sum_{i=1}^N A_i x_i = 0$.

This assumption implies that the set of solutions to (6.4) is nonempty (see [50, Prop. 4.3(iii)]). If $(\boldsymbol{x}^\star, \boldsymbol{y}^\star, \boldsymbol{w}^\star)$ is a solution to (6.4), then $\boldsymbol{x}^\star$ is a solution to the primal problem (6.2) and $(\boldsymbol{y}^\star, \boldsymbol{w}^\star)$ to its dual (6.3).

6.3 Distributed primal-dual algorithms

In this section we provide the main distributed algorithm that is based on *Asymmetric Forward-Backward-Adjoint* (AFBA) [96] (see Chapter 2). The developed algorithm belongs to the class of primal-dual algorithms. The convergence results include both primal and dual variables. However, the convergence analysis here focuses on the primal variables for clarity of exposition, with the understanding that similar claims holds for the dual variables.

Our distributed algorithm consists of two phases, a local phase and the phase in which each agent interacts with its neighbors according to the constraints imposed by the communication graph. Each iteration has the advantage of only requiring local matrix-vector products and proximal updates. Specifically, each agent performs 2 matrix-vector products per iteration and transmits a vector of dimension n to its neighbors.

Let $\boldsymbol{u} = (\boldsymbol{x}, \boldsymbol{v})$ where $\boldsymbol{v} = (\boldsymbol{y}, \boldsymbol{w})$ and $\boldsymbol{y} = (y_1, \dots, y_N)$. The optimality conditions in (6.4), can be written in the form of the following monotone inclusion:

$$0 \in D\boldsymbol{u} + M\boldsymbol{u} \tag{6.5}$$

with

$$M = \begin{pmatrix} 0 & \boldsymbol{C}^\top & A^\top \\ -\boldsymbol{C} & 0 & 0 \\ -A & 0 & 0 \end{pmatrix}, \quad D(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{w}) = (\partial \boldsymbol{g}(\boldsymbol{x}), \partial \boldsymbol{h}^*(\boldsymbol{y}), 0),$$

where $\boldsymbol{g}(\boldsymbol{x}) = \sum_{i=1}^{N} g_i(x_i)$, $\boldsymbol{h}^*(\boldsymbol{y}) = \sum_{i=1}^{N} h_i^*(y_i)$, $\boldsymbol{C} = \mathrm{blkdiag}(C_1, \dots, C_N)$. Notice that $A\boldsymbol{x} = \sum_{i=1}^{N} A_i x_i$, $A^\top \boldsymbol{w} = (A_1^\top w, \dots, A_N^\top w)$. The operator $D + M$ is maximally monotone [13, Prop. 20.23, Cor. 25.5].

Monotone inclusion (6.5) which is a restatement of the primal-dual optimality conditions (6.4) is solved by applying [96, Alg. 6]. This results in the following iteration:

$$\boldsymbol{x}^{k+1} = \mathrm{prox}_{\boldsymbol{g}}^{\Sigma^{-1}}(\boldsymbol{x}^k - \Sigma C^\top \boldsymbol{y}^k - \Sigma A^\top \boldsymbol{w}^k) \tag{6.6a}$$

$$\bar{\boldsymbol{y}}^k = \mathrm{prox}_{\boldsymbol{h}^*}^{\Gamma^{-1}}(\boldsymbol{y}^k + \Gamma C(\theta \boldsymbol{x}^{k+1} + (1-\theta)\boldsymbol{x}^k)) \tag{6.6b}$$

$$\bar{\boldsymbol{w}}^k = \boldsymbol{w}^k + \Pi A(\theta \boldsymbol{x}^{k+1} + (1-\theta)\boldsymbol{x}^k) \tag{6.6c}$$

$$\boldsymbol{y}^{k+1} = \bar{\boldsymbol{y}}^k + (2-\theta)\Gamma C(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) \tag{6.6d}$$

$$\boldsymbol{w}^{k+1} = \bar{\boldsymbol{w}}^k + (2-\theta)\Pi A(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) \tag{6.6e}$$

where matrices Σ, Γ, Π play the rule of stepsizes and are assumed to be positive definite. The iteration (6.6) can not be implemented in a distributed fashion because the dual vector w consists of M blocks corresponding to the edges. The key idea that allows distributed computations is to introduce the sequence

$$(\rho_i^k)_{k\in\mathbb{N}} = (A_i^\top \boldsymbol{w}^k)_{k\in\mathbb{N}}, \quad \text{for} \quad i = 1, \dots, N. \tag{6.7}$$

This transformation replaces the stacked edge vector $\boldsymbol{w}^k$ with corresponding node vectors ρ_i. More compactly, letting $\boldsymbol{\rho}^k = (\rho_1^k, \dots, \rho_N^k)$, it follows from (6.6c) and (6.6e) that

$$\boldsymbol{\rho}^{k+1} = \boldsymbol{\rho}^k + A^\top \Pi A(2\boldsymbol{x}^{k+1} - \boldsymbol{x}^k), \tag{6.8}$$

where $A^\top \Pi A$ is the *weighted graph Laplacian*. Since $\boldsymbol{w}^k$ in (6.6a) appear as $A^\top \boldsymbol{w}^k$ we can rewrite the iteration:

$$\boldsymbol{x}^{k+1} = \mathrm{prox}_{\boldsymbol{g}}^{\Sigma^{-1}}(\boldsymbol{x}^k - \Sigma C^\top \boldsymbol{y}^k - \Sigma \boldsymbol{\rho}^k)$$

$$\bar{\boldsymbol{y}}^k = \mathrm{prox}_{\boldsymbol{h}^*}^{\Gamma^{-1}}(\boldsymbol{y}^k + \Gamma C(\theta \boldsymbol{x}^{k+1} + (1-\theta)\boldsymbol{x}^k))$$

$$\boldsymbol{y}^{k+1} = \bar{\boldsymbol{y}}^k + (2-\theta)\Gamma C(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k)$$

$$\boldsymbol{\rho}^{k+1} = \boldsymbol{\rho}^k + A^\top \Pi A(2\boldsymbol{x}^{k+1} - \boldsymbol{x}^k)$$

Set

$$\Sigma = \mathrm{blkdiag}\,(\sigma_1 I_n, \dots, \sigma_N I_n),$$

$$\Gamma = \text{blkdiag}\,(\tau_1 I_{r_1}, \ldots, \tau_N I_{r_N}),$$

$$\Pi = \text{blkdiag}\,(\pi_1 I_n, \ldots, \pi_M I_n),$$

where $\sigma_i > 0, \tau_i > 0$ for $i = 1, \ldots, N$ and $\pi_l > 0$ for $l = 1, \ldots, M$. Consider a bijective mapping between $l = 1, \ldots, M$ and unordered pairs $(i, j) \in E$ such that $\kappa_{i,j} = \kappa_{j,i} = \pi_l$. Notice that π_l for $l = 1, \ldots, M$ are stepsizes to be selected by the algorithm and can be viewed as weights for the edges. Thus, iteration (6.6) gives rise to our distributed algorithm:

Algorithm 6.1

Inputs: $\sigma_i > 0$, $\tau_i > 0$, $\kappa_{i,j} > 0$ for $j \in \mathcal{N}_i$, $i = 1, \ldots, N$, $\theta \in [0, \infty)$, initial values $x_i^0 \in \mathbb{R}^n$, $y_i^0 \in \mathbb{R}^{r_i}$, $\rho_i^0 \in \mathbb{R}^n$.

for $k = 0, 1, \ldots$ **do**

for each agent $i = 1, \ldots, N$ **do**

Local steps:

$x_i^{k+1} = \text{prox}_{\sigma_i g_i}(x_i^k - \sigma_i \rho_i^k - \sigma_i C_i^\top y_i^k)$

$\bar{y}_i^k = \text{prox}_{\tau_i h_i^*}(y_i^k + \tau_i C_i(\theta x_i^{k+1} + (1-\theta) x_i^k))$

$y_i^{k+1} = \bar{y}_i^k + \tau_i (2-\theta) C_i (x_i^{k+1} - x_i^k)$

$u_i^k = 2x_i^{k+1} - x_i^k$

Exchange of information with neighbors:

$\rho_i^{k+1} = \rho_i^k + \sum_{j \in \mathcal{N}_i} \kappa_{i,j}(u_i^k - u_j^k)$

Notice that each agent i only requires $u_j^k \in \mathbb{R}^n$ for $j \in \mathcal{N}_i$ during the communication phase. Before proceeding with convergence results, we define the following for simplicity of notation:

$$\bar{\sigma} = \max\{\sigma_1, \ldots, \sigma_N\},$$

$$\bar{\tau} = \max\{\tau_1, \ldots, \tau_N, \pi_1, \ldots, \pi_M\},$$

$$L = \mathcal{L} \otimes I_n + \boldsymbol{C}^\top \boldsymbol{C}, \text{ where } \mathcal{L} \text{ is the graph Laplacian.}$$

It must be noted that the results in this section only provide choices of parameters that are sufficient for convergence. They can be selected much less conservatively by formulating and solving sufficient conditions that they must satisfy as *linear matrix inequalities* (LMIs).

Theorem 6.1. *Let Assumptions 6.II and 6.III hold true. Consider the sequence* $(\boldsymbol{x}^k)_{k \in \mathbb{N}} = (x_1^k, \ldots, x_N^k)_{k \in \mathbb{N}}$ *generated by Algorithm 6.1. Assume the maximum*

stepsizes, i.e., $\bar{\sigma}$ *and* $\bar{\tau}$ *defined above, are positive and satisfy*

$$\bar{\sigma}^{-1} - \bar{\tau}(\theta^2 - 3\theta + 3)\|L\| > 0, \tag{6.9}$$

for a fixed value of $\theta \in [0, \infty)$*. Then, the sequence* $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ *converges to* $(x^\star, \dots, x^\star)$ *for some* $x^\star \in S^\star$*. Furthermore, if* $\theta = 2$ *the strict inequality* (6.9) *is replaced with* $\bar{\sigma}^{-1} - \bar{\tau}\|L\| \geq 0$*.*

Proof. Algorithm 6.1 is an implementation of [96, Alg. 6]. Thus convergence of $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ to a solution of (6.2) is implied by [96, Prop. 5.4]. Combining this with Assumption 6.II yields the result. Notice that in that work the stepsizes are assumed to be scalars for simplicity. It is straightforward to adapt the result to the case of diagonal matrices. □

In Algorithm 6.1 when $\theta = 2$, we recover the algorithm of Chambolle and Pock [38]. One important observation is that the term $\theta^2 - 3\theta + 3$ in (6.9) is always positive and achieves its minimum at $\theta = 1.5$. This is a choice of interest for us since it results in larger stepsizes, $\sigma_i, \tau_i, \kappa_{i,j}$, and consequently better performance as we observe in numerical simulations.

Theorem 6.2 establishes linear convergence for the algorithm whenever g_i and h_i are *piecewise linear-quadratic* (PLQ), see Definition 1.4. The class of PLQ functions has been studied extensively and has many desirable properties (see [144, §10 and §11]). Many practical applications involve PLQ functions such as quadratic function, $\|\cdot\|_1$, indicator of polyhedral sets, hinge loss, *etc.* Thus, the R-linear convergence rate that we establish in Theorem 6.2 holds for a wide range of problems encountered in control, machine learning and signal processing.

Theorem 6.2. *Consider Algorithm 6.1 under the assumptions of Theorem 6.1. Assume* g_i *and* h_i *for* $i = 1, \dots, N$*, are piecewise linear-quadratic functions. Then the set valued mapping* $T = D + M$ *is* metrically subregular *at any* z *for any* z' *provided that* $(z, z') \in \operatorname{gra} T$*. Furthermore, the sequence* $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ *converges R-linearly to* $(x^\star, \dots, x^\star)$ *for some* $x^\star \in S^\star$*.*

Proof. The proof of the first claim is similar to Lemma 3.11 and is omitted. The second part of the proof follows directly by noting that [96, Alg. 6] used to derive Algorithm 6.1 is a special case of Algorithm 3.1 (with $\mu = 0$ and $\lambda = 1$). Therefore, linear convergence follows from Corollary 3.12*(i)*. The aforementioned theorem guarantees linear convergence for the stacked vector $\boldsymbol{u}$ in (6.5), however, here we consider the primal variables only. □

6.3.1 Special case

Consider the following problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^{N} g_i(x), \tag{6.10}$$

where $g_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ for $i = 1, \ldots, N$ are proper closed convex functions. This is a special case of (6.1) when $h_i \circ C_i \equiv 0$. Since functions h_i are absent, the dual variables y_i in Algorithm 6.1 vanish and for any choice of θ the algorithm reduces to:

$$\begin{aligned} x_i^{k+1} &= \operatorname{prox}_{\sigma_i g_i}(x_i^k - \sigma_i \rho_i^k) \\ u_i^k &= 2x_i^{k+1} - x_i^k \\ \rho_i^{k+1} &= \rho_i^k + \sum_{j\in\mathcal{N}_i} \kappa_{i,j}(u_i^k - u_j^k). \end{aligned}$$

Thus setting $\theta = 1.5$ in (6.9) to maximize the stepsizes yields $\bar{\sigma}^{-1} - \frac{3\bar{\tau}}{4}\|\mathcal{L}\| > 0$, where $\mathcal{L}$ is the graph Laplacian.

6.4 Numerical simulations

We now illustrate experimental results obtained by applying the proposed algorithm to the following problem:

$$\text{minimize} \quad \lambda\|x\|_1 + \sum_{i=1}^{N} \tfrac{1}{2}\|D_i x - d_i\|_2^2 \tag{6.11}$$

for a positive parameter λ. This is the ℓ_1 regularized least-squares problem. Problem (6.11) is of the form (6.1) if we set for $i = 1, \ldots, N$

$$g_i(x) = \tfrac{\lambda}{N}\|x\|_1, \quad h_i(z) = \tfrac{1}{2}\|z - d_i\|_2^2, \quad C_i = D_i, \tag{6.12}$$

where $D_i \in \mathbb{R}^{m_i \times n}$, $d_i \in \mathbb{R}^{m_i}$. For the experiments we used graphs of $N = 50$ computing agents, generated randomly according to the Erdős-Renyi model, with parameter $p = 0.05$. In the experiments we used $n = 500$ and generated D_i randomly with normally distributed entries, with $m_i = 50$ for all $i = 1, \ldots, N$.

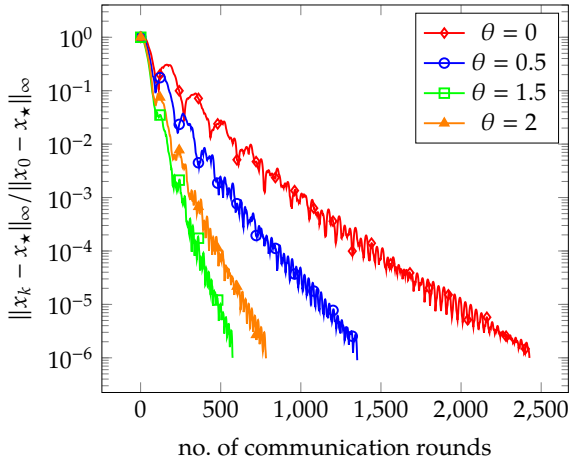

Figure 6.1: *Convergence of the relative error for the algorithms, in one of the considered instances.*

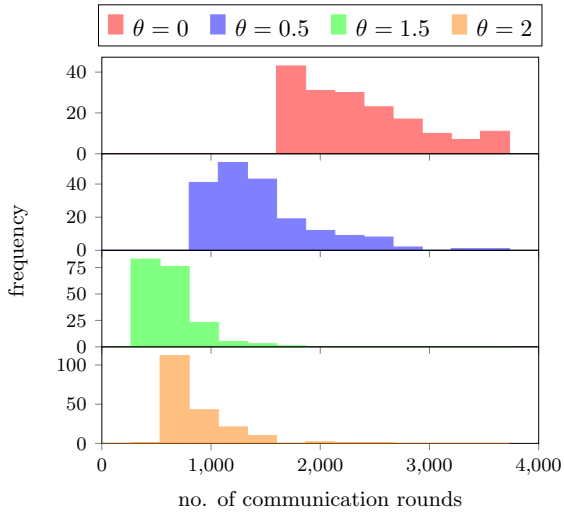

Figure 6.2: *Distribution of the number of communication rounds required by the algorithms to achieve a relative error of* 10^{-6}*, for fixed data and* 200 *randomly generated Erdős-Renyi graphs, with parameter* $p = 0.05$*.*

Then we generated vector d_i starting from a known solution for the problem and ensuring $\lambda < 0.1 \| \sum_i^N D_i^\top d_i \|_\infty$.

For the stepsize parameters we set $\sigma_i = \bar{\sigma}$, $\tau_i = \bar{\tau}$, for all $i = 1, \ldots, N$, and $\kappa_{i,j} = \kappa_{j,i} = \bar{\tau}$ for all edges $(i, j) \in E$, such that (6.9) is satisfied. In order to have a fair comparison we selected $\bar{\sigma} = \alpha/\|L\|$ and $\bar{\tau} = 0.99/(\alpha(\theta^2 - 3\theta + 3))$ with $\alpha = 20$ which was set empirically based on better performance of all the algorithms.

The results are illustrated in Figure 6.2, for several values of θ, where the distribution of the number of communication rounds required by the algorithms to reach a relative error of 10^{-6} is reported. In Figure 6.1 the convergence of algorithms is illustrated in one of the instances. It should be noted that the algorithm of Chambolle and Pock, that corresponds to $\theta = 2$, is generally slower than the case $\theta = 1.5$. This is mainly due to the larger stepsize parameters guaranteed by Theorem 6.1.

Chapter 7

Multi-agent structured optimization with bounded communication delays

This chapter is based on:

P. Latafat and P. Patrinos. *Primal-dual algorithms for multi-agent structured optimization over message-passing architectures with bounded communication delays* (submitted 2019).

7.1 Introduction

In this chapter we consider a class of structured optimization problems that can be represented as follows:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x)+\sum_{i=1}^{m}\big(g_i(x_i)+h_i(N_ix)\big), \tag{7.1}$$

where $x=(x_1,\dots,x_m)$, N_i is a linear mapping, h_i, g_i are proper closed convex (possibly) nonsmooth functions, g_i are in addition strongly convex, and f is convex, continuously differentiable with Lipschitz continuous gradient. The goal is to solve (7.1) over a network of agents through local communications. Each agent is assumed to maintain its own private cost functions g_i and h_i, while f and (possibly) the linear mappings N_i represent the coupling between the agents. In practice local communications between agents are subject to delays and/or dropouts which constitutes an important challenge addressed here.

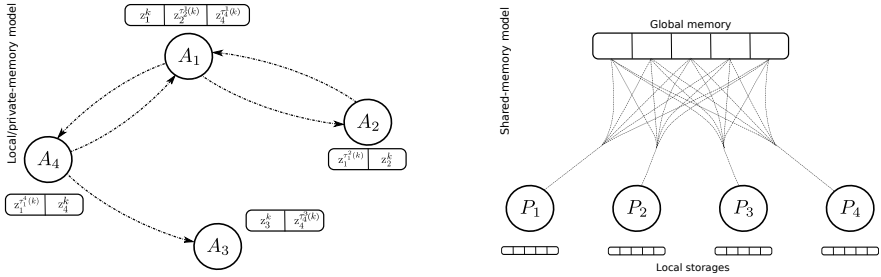

Figure 7.1: *The main two memory models; (left) agents cooperating to perform a task, (right) processors updating a global memory*

Most iterative algorithms for convex optimization can be written as

$$z^{k+1} = z^k - Tz^k, \tag{7.2}$$

where the mapping $\mathrm{id} - T$ (id is the identity operator) has some contractive property resulting in the convergence of the sequence to a zero of T. In distributed optimization the goal is to devise algorithms where a group of agents/processors distributively update certain coordinates of z while guaranteeing convergence to a zero of T.

There are two main computational models in distributed optimization (depicted in Fig. 7.1) with a range of hybrid models in between [20, §1]. These models are conceptually different and require different analysis. The model considered here is the local/private-memory model. Let us first describe the two models.

Shared-memory model: This model is characterized by the access of all agents/processors to a shared memory. A large body of literature exists for parallel coordinate descent algorithms for this problem. Typically, coordinate descent algorithms would require a memory lock to ensure consistent reading. Interesting recent works allow inconsistent reads [109, 133]. In this model, for the fixed point iteration (7.2), each processor reads the global memory and proceeds to choose a random coordinate $i \in \{1, \dots, m\}$ and to perform

$$z_i^{k+1} = z_i^k - T_i \hat{z}^k,$$

where $\hat{z}^k$ denotes the data loaded from the global memory to the local storage at the clock tick k, and T_i represents the operator that updates the i-th coordinate. This form of updates are asynchronous in the sense that the processors update the global memory simultaneously resulting in possibly inconsistent local copy

$\hat{z}^k$ due to other processors modifying the global memory during a read. The analysis of such algorithms would in general rely on either using the properties of the operator that updates the i-th coordinate when possible (coordinate-wise Lipschitz continuity in the case of the gradient [109]), or the properties of the global operator (see [133] for nonexpansive operators). A crucial point in the convergence analysis of such methods is the fact that for a given processor, the index of the coordinate to be updated is selected at random, but no matter which coordinate is selected the same local data $\hat{z}^k$ is used for the update. Let $\hat{T}_i z \coloneqq (0, \dots, 0, T_i z, 0 \dots, 0)$. Then, in a randomized scheme $\hat{T}_i$ can be summed over i:

$$\sum_{i=1}^{m} \hat{T}_i \hat{z}^k = T\hat{z}^k,$$

allowing one to use the properties known for the global operator (see the proof of [133, Lem. 7]). This type of argument is also used in [176] in the context of decentralized consensus optimization. See [34] for a detailed discussion on the assumptions that are often imposed in this model. As we discuss below, the difficulty in the local-memory model is precisely due to the fact that this summation no longer holds.

Local/private-memory model: In this model each agent/processor has its own private local memory. The agents can send and receive information to other agents as needed, and agent i can only update z_i. This model is also referred to as *message-passing model* [20].

In the absence of delay between agents, randomized block-coordinate updates may be used to develop distributed asynchronous algorithms. Such schemes would typically involve random independent activation of agents to perform their local updates, and are in this sense also referred to as asynchronous [88, 22, 95, 134]. Note that in these schemes while the agents may wake up to perform their updates at different times, the information used by each agent is assumed to be up to date, *i.e.*, synchronization is required.

In accordance with the notation of the seminal work [20, §7] we define the following local (outdated) version of the generic vector $z^k = (z_1^k, \dots, z_m^k)$ used by agent i:

$$z^k[i] \coloneqq \left(z_1^{\tau_1^i(k)}, \dots, z_m^{\tau_m^i(k)}\right), \tag{7.3}$$

where $\tau_j^i(k)$ is the latest time at which the value of z_j is transmitted to agent i by agent j. In our setting the delay is assumed to be bounded:

Assumption 7.I. There exists an integer B such that for all $k \geq 0$ the following holds

$$(\forall i, j) \quad 0 \leq k - \tau_j^i(k) \leq B, \quad \text{and} \quad \tau_i^i(k) = k.$$

The fact that each agent knows its own local variable without delay is projected in the assumption $\tau_i^i(k) = k$. This is a natural assumption and is satisfied in practice. Notice that for ease of notation we defined the complete outdated vector while in practice each agent would only keep a local copy of the coordinates that are required for its computation, see Fig. 7.1. The directions of the arrows in Fig. 7.1 signify the nature of the coupling between two agents. For example, the arrow from A_4 to A_3 indicates that agent A_3 requires z_4 for its computation. Such a relation between agents is dependent on the formulation and the nature of coupling between agents. For instance, in (7.1) the coupling is represented through f and possibly N_i. As we shall see in §7.2 the coupling through f may be one sided since agent i may require information from agent j for computing $\nabla_i f$ (the partial derivative of f with respect to i-th coordinate) without the reverse relation being true.

In summary, each agent controls only one block of coordinates and updates according to

$$z_i^{k+1} = z_i^k - T_i z^k[i],$$

the result of which will be sent (possibly with different delay) to the agents that require it in their computations. The difficulty in this model comes from the impossibility of summing $T_i z^k[i]$ over all i given that $z^k[i]$ is different for each i.

In addition to the above described delay, the *partially asynchronous* (PA) protocol considered in [20, §7] involves a second assumption: each agent must perform an update at least once during any time interval of length B. In [20, §7.5] a PA variant of the gradient method is studied. This analysis is further extended to the projected-gradient method in the convex case. In [161] a periodic linear convergence rate is established for the projected-gradient method. The recent work [184] extends this analysis to the proximal-gradient method.

The aforementioned primal methods are not well equipped for problems with more complex structures as in (7.1). An efficient way to tackle such problems is to employ a class of first-order methods, referred to as primal-dual algorithms. This approach leads to *fully split* algorithms eliminating the need for inverting matrices or solving inner loops. Developing PA schemes for primal-dual algorithms is not addressed here and remains a challenge. It it worth noting that [80] considers a primal-dual framework under a different asynchronous protocol where the primal variables follow a totally asynchronous model [20, §6]. However, the dual variables are required to be synchronized across agents.

It is worth noting that finite sum minimization over graphs is another popular problem that has been considered by many authors [122, 64, 150, 110, 90, 100]. Several asynchronous algorithms have been studied for this popular problem over master-worker architectures [1, 12, 71, 41, 183]. Moreover, asynchronous

subgradient type methods have been studied extensively for finite sum problems [121, 168, 154, 172, 105]. In contrast, in this work general optimization problem (7.1) is considered. This framework can be used to develop asynchronous distributed proximal algorithms for general finite sum minimization of the form

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^{m} \Psi_i(x) + \Theta_i(x) + \Phi_i(C_i x),$$

where Φ_i is smooth, Θ_i and Φ_i are (possibly) nonsmooth extended-real-valued and C_i is a linear mapping. This problem can be reformulated as (7.1) using a consensus reformulation according to the communication graph. However, this is deferred to future work.

7.1.1 Motivating examples

Consider the regularized logistic regression problem

$$\underset{w\in\mathbb{R}^n}{\text{minimize}} \quad \sum_{i=1}^{m} \sum_{j\in\mathcal{I}_i} \log\left(1+\exp\left(-y_{(j)}\langle x_{(j)}, w\rangle\right)\right) + \lambda\|w\|^2, \tag{7.4}$$

where $w = (w_1, \ldots, w_m) \in \mathbb{R}^n$ is the regression vector, λ is a positive constant, and the data is distributed between m machines; the pair $(x_{(j)}, y_{(j)})_{j\in\mathcal{I}_i}$ represents the data stored at the i-th machine. The goal is to solve the global minimization via local communications which may be subject to communication delays. Clearly (7.4) fits into the form of (7.1): let g represent the separable regularizer, $h_i(v) = \sum_{j\in\mathcal{I}_i} \log((1+\exp(-y_{(j)}v_j))$ the loss function, $f \equiv 0$ and the rows of N_i consisting of $x_{(j)}^\top$ for $j \in \mathcal{I}_i$. In this formulation the coupling is through the linear terms (*cf.* §7.5). Distributed elastic net problem is another such example with h_i representing the squared loss, N_i the locally stored data, g the elastic net regularizer and $f \equiv 0$. Note that in both examples g_i is strongly convex and h_i is continuously differentiable with Lipschitz continuous gradient, satisfying the requirements of §7.5 for the cost functions.

Another notable example is the problem of formation control [136], where each agent (vehicle) has its own private dynamics and cost function and the goal is to achieve a specific formation while communicating only with a selected number of agents. Let $w_i = (\xi_i, v_i)$ where ξ_i and v_i denote the local state and input sequences. The location of agent i is given by $y_i = C\xi_i$ and the set of its neighbors is denoted by $\mathcal{A}_i$. The linear dynamics of each agent over a control horizon is represented by the constraints $E_i w_i = b_i$. In order to enforce a formation between agents i and j the quadratic cost function $\|C(\xi_i - \xi_j) - d_{ij}\|^2$

is used where d_{ij} is the target relative distance between them (refer to [136] for details). Hence, the formation control problem is formulated as the following constrained minimization:

$$\text{minimize} \sum_{i=1}^{m} \tfrac{\lambda_i}{2} \sum_{j \in \mathcal{A}_i} \|C(\xi_i - \xi_j) - d_{ij}\|^2 + \tfrac{1}{2} \sum_{i=1}^{m} w_i^\top Q_i w_i \tag{7.5a}$$

$$\text{subject to } E_i w_i = b_i, \quad w_i \in \mathcal{W}_i, \quad i = 1, \ldots, m \tag{7.5b}$$

This problem can be easily cast in the form of (7.1) by setting f equal to the first term, g_i equal to the quadratic local cost, while $h_i \circ L_i$ captures the dynamics and input and state constraints (see Section 7.6 for more details). Therefore, the objective is to enforce a formation between agents by solving this optimization problem in presence of communication delays by allowing the agents to use outdated information. Notice that in this case the coupling between agents is enforced only through f. This special case of (7.1) is studied in §7.4.

7.1.2 Main contributions

• To the best of our knowledge this is the first work that considers the delay described in (7.3) in a message-passing model for primal-dual algorithms. Unlike primal methods (gradient or proximal-gradient), the proposed algorithms are applicable to problems with complex structures as in (7.1) without the need to solve inner loops or to invert matrices.

• The analysis of [20, 161, 184] rely on the use of the cost as the Lyapunov function. In contrast, we show that quasi-Fejér monotonicity is an effective tool in the analysis of bounded delays in our setting. While this paper focuses on two particular primal-dual algorithms, a similar analysis should be applicable to others such as those proposed in [50, 31, 62, 96, 95, 98].

• Two primal-dual algorithms are presented: (i) when the coupling between agents is enforced only through f, the algorithm of [53, 170] is considered (*cf.* §7.4), (ii) when the coupling is through f and the linear term, a new modified algorithm is developed (*cf.* §7.5). In addition, linear convergence rates are established with explicit convergence factors.

• In §7.5.2 an asynchronous protocol is considered; at every iteration agents are activated at random, and independently from one another, while performing their updates using outdated information. In practice, random activation can model the discrepancies in the speed of different agents.

7.2 Problem setup

Throughout this chapter the primal and dual vectors, denoted x and u, are assumed to be composed of m blocks as follows

$$x = (x_1, \dots, x_m) \in \mathbb{R}^n, \; u = (u_1, \dots, u_m) \in \mathbb{R}^r,$$

where $x_i \in \mathbb{R}^{n_i}$ and $u_i \in \mathbb{R}^{r_i}$. Moreover, we denote the stacked primal and dual variable as $z = (x, u)$.

Consider a linear mapping $L : \mathbb{R}^n \to \mathbb{R}^r$ that is partitioned as follows:

$$L = \begin{pmatrix} L_{11} & \cdots & L_{1m} \\ \vdots & \ddots & \vdots \\ L_{m1} & \cdots & L_{mm} \end{pmatrix}, \tag{7.6}$$

where $L_{ij} : \mathbb{R}^{n_i} \to \mathbb{R}^{r_j}$. Furthermore, the i-th (block) row of L is denoted by $L_{i\bullet} : \mathbb{R}^n \to \mathbb{R}^{r_i}$ and the i-th (block) column by $L_{\bullet i} : \mathbb{R}^{n_i} \to \mathbb{R}^r$, *i.e.*,

$$L = \begin{pmatrix} L_{1\bullet} \\ \vdots \\ L_{m\bullet} \end{pmatrix} = \begin{pmatrix} L_{\bullet 1} & \cdots & L_{\bullet m} \end{pmatrix}.$$

The following holds

$$\langle Lx, u \rangle = \sum_{i=1}^{m} \langle L_{i\bullet} x, u_i \rangle = \sum_{i=1}^{m} \langle x_i, L_{\bullet i}^\top u \rangle. \tag{7.7}$$

Consider the structured optimization problem (7.1) where the linear mapping N_i has been replaced by $L_{i\bullet}$ defined above in order to clarify the structure of the mapping:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) + \sum_{i=1}^{m} \big(g_i(x_i) + h_i(L_{i\bullet} x) \big). \tag{7.8}$$

The cost functions g_i and $h_i \circ L_{i\bullet}$ are private functions belonging to agent i. The coupling between agents is through the smooth term f and the linear term $L_{i\bullet} x$. An agent i is assumed to have access to the information required for its computation, be it outdated, *cf.* Algorithms 7.1 and 7.2.

Let the following assumptions hold

Assumption 7.II.

(i) For $i = 1, \dots, m$, $h_i : \mathbb{R}^{r_i} \to \overline{\mathbb{R}}$ is proper closed convex function, and $L_{i\bullet} : \mathbb{R}^n \to \mathbb{R}^{r_i}$ is a linear mapping.

(ii) (strong convexity) For $i = 1, \dots, m$, $g_i : \mathbb{R}^{n_i} \to \overline{\mathbb{R}}$ is proper closed μ_g^i-strongly convex for some $\mu_g^i > 0$.

(iii) $f : \mathbb{R}^n \to \mathbb{R}$ is convex, continuously differentiable, and for some $\beta \in [0, \infty)$, ∇f is β-Lipschitz continuous:

$$\|\nabla f(x) - \nabla f(x')\| \leq \beta \|x - x'\|, \quad \forall x, x' \in \mathbb{R}^n.$$

(iv) For every $i = 1, \dots, m$ there exists a nonnegative constant $\bar{\beta}_i$ such that for all $x, x' \in \mathbb{R}^n$ satisfying $x_i = x'_i$:

$$\|\nabla_i f(x) - \nabla_i f(x')\| \leq \bar{\beta}_i \|x - x'\|. \tag{7.9}$$

(v) The set of solutions to (7.8) is nonempty. Moreover, there exists $x_i \in \operatorname{ri} \operatorname{dom} g_i$, for $i = 1, \dots, m$ such that $L_{j\bullet} x \in \operatorname{ri} \operatorname{dom} h_j$, for $j = 1, \dots, m$.

Assumption 7.II*(iv)* quantifies the strength of the coupling (through f) between agents [20, §7.5]. In particular, if f is separable, *i.e.*, $f(x) = \sum_{i=1}^m f_i(x_i)$, then there is no coupling and $\bar{\beta}_i = 0$.

Problem (7.8) can be compactly represented as

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) + g(x) + h(Lx),$$

where $g(x) = \sum_{i=1}^m g_i(x_i)$, $h(u) = \sum_{i=1}^m h_i(u_i)$, and L is as in (7.6). The dual problem is given by

$$\underset{u \in \mathbb{R}^r}{\text{minimize}} \ (g + f)^*(-L^\top u) + h^*(u).$$

By Assumption 7.II*(ii)* the set of solutions to (7.8) is nonempty and unique. Under the constraint qualification of Assumption 7.II*(v)*, the set of solutions to the dual problem is nonempty (not necessarily a singleton) and the duality gap is zero [143, Cor. 31.2.1]. Furthermore, $x^\star$ is a primal solution and $u^\star$ is a dual solution if and only if the pair $(x^\star, u^\star)$ satisfies

$$\begin{cases} 0 \in \partial g(x^\star) + \nabla f(x^\star) + L^\top u^\star, \\ 0 \in \partial h^*(u^\star) - L x^\star. \end{cases} \tag{7.10}$$

Such a point is called a primal-dual solution and the set of all primal-dual solutions is denoted by $\mathcal{S}$.

For each agent $i \in \{1, \dots, m\}$ define positive stepsizes γ_i, σ_i associated with the primal and the dual variables, respectively. Let us also define the following parameters

$$\begin{aligned}
\bar{\beta} &\coloneqq (\bar{\beta}_1, \dots, \bar{\beta}_m), \\
\Gamma &\coloneqq \operatorname{blkdiag}(\gamma_1 \mathrm{I}_{n_1}, \dots, \gamma_m \mathrm{I}_{n_m}), \\
\Sigma &\coloneqq \operatorname{blkdiag}(\sigma_1 \mathrm{I}_{r_1}, \dots, \sigma_m \mathrm{I}_{r_m}), \\
D &= \operatorname{blkdiag}(\Gamma^{-1}, \Sigma^{-1}).
\end{aligned} \tag{7.11}$$

The algorithm of Vũ and Condat [170, 53] for solving (7.8) is given by the following updates for agent i at iteration k:

$$x_i^{k+1} = \operatorname{prox}_{\gamma_i g_i}\left(x_i^k - \gamma_i L_{\bullet i}^\top u^k - \gamma_i \nabla_i f(x^k)\right) \tag{7.12a}$$

$$u_i^{k+1} = \operatorname{prox}_{\sigma_i h_i^*}\left(u_i^k + \sigma_i L_{i\bullet}(2x^{k+1} - x^k)\right). \tag{7.12b}$$

In a synchronous implementation of the algorithm, each agent requires the latest variables x^k, x^{k+1} and u^k in the above updates, which may not be available due to communication delays. In the case when L is block-diagonal the coupling between agents is enforced only through the smooth function f (in (7.12) agent i requires the primal variables that are required for computing $\nabla_i f$). We refer to this type of coupling as *partial coupling*. In Section 7.4 the updates in (7.12) are considered for the case of partial coupling (*cf.* Alg. 7.2).

More generally when L is not block-diagonal, the coupling between agents is enacted through the linear mapping L (the linear operations $L_{\bullet i}^\top$ and $L_{i\bullet}$ in (7.12b) require additional communication between agents) and possibly the smooth function f. We refer to this type of coupling as *total coupling*. This case is considered in Section 7.5 where an *Arrow-Hurwicz-Uzawa* type [8] (referred hereafter as AHU-type) primal-dual algorithm is proposed in place of (7.12). The synchronous iterations of the AHU-type algorithm for agent i at iteration k is given by:

$$x_i^{k+1} = \operatorname{prox}_{\gamma_i g_i}\left(x_i^k - \gamma_i L_{\bullet i}^\top u^k - \gamma_i \nabla_i f(x^k)\right) \tag{7.13a}$$

$$u_i^{k+1} = \operatorname{prox}_{\sigma_i h_i^*}\left(u_i^k + \sigma_i L_{i\bullet} x^k\right). \tag{7.13b}$$

Differently from (7.12), in the dual update linear operator is applied to x^k in place of $2x^{k+1} - x^k$. When operating under the bounded delay assumption, the AHU-type primal-dual algorithm, (7.13), allows for larger stepsizes compared

to (7.12).

It is worth noting that (7.13) can be seen as a forward-backward iteration:

$$z^{k+1} = (D + T_2)^{-1}(D - T_1)z^k,$$

where $T_2 = (\partial g, \partial h^*)$, $T_1 : (x, u) \mapsto (\nabla f(x) + L^\top u, -Lx)$, and D as defined in (7.11). Even in the synchronous case this algorithm is not in general convergent. The convergence may be established when g and h^* are strongly convex [43, Assumption A]. Moreover, when g is the indicator of a set and h is the support of a set, the AHU-type algorithm resembles another primal-dual AHU-type algorithm considered in [123, 80] for solving saddle-point problems.

7.3 Notation and preliminary results

This section is devoted to establishing some preliminary results and notation.

Lemma 7.1. *Let $q : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper closed μ-convex function for some $\mu \geq 0$. For all $r \in \mathbb{R}^n$, $\omega \in \mathbb{R}^n$ and $\omega_\rho \coloneqq \operatorname{prox}_{\rho q}(\omega)$ the following holds*

$$q(r) - q(\omega_\rho) \geq \tfrac{1}{\rho}\langle \omega - \omega_\rho, r - \omega_\rho \rangle + \tfrac{\mu}{2}\|r - \omega_\rho\|^2. \tag{7.14}$$

Proof. The inequality follows immediately from the definition of strong convexity and the characterization of proximal mapping [13, Prop. 16.44]. □

For all $a, b, c \in \mathbb{R}^n$ and all positive definite matrices $V \in \mathbb{R}^{n\times n}$ the following elementary equality holds.

$$2\langle a - b, c - b\rangle_V = \|a - b\|_V^2 + \|c - b\|_V^2 - \|a - c\|_V^2. \tag{7.15}$$

We also make use (1.8) with $V = \epsilon \mathrm{I}_n$:

$$\langle x, y\rangle \leq \tfrac{\varepsilon}{2}\|x\|^2 + \tfrac{1}{2\varepsilon}\|y\|^2, \quad \forall x, y \in \mathbb{R}^n, \epsilon > 0. \tag{7.16}$$

Lemma 7.2 provides a basic inequality which is crucial in our analysis. Refer to [20, §7.5] and [184, Lem. 4] for the proof.

Lemma 7.2. *Let Assumption 7.I hold. Consider a vector $w^k = (w_1^k, \dots, w_m^k)$ and its outdated version $w^k[i]$, cf. (7.3). Then, the following inequality holds*

$$\|w^k - w^k[i]\| \leq \sum_{\tau=[k-B]_+}^{k-1} \|w^{\tau+1} - w^\tau\|. \tag{7.17}$$

Let l and d be two nonnegative scalars. For a given sequence $(w^t)_{t\in\mathbb{N}}$ we define the following for simplicity of notation:

$$S_l^d(w^t)_{t\le k} \coloneqq \sum_{\tau=[k-B+1-l]_+}^{k-d} \|w^{\tau+1}-w^\tau\|^2.$$

Summing $S_l^d(w^t)_{t\le k}$ over k from 0 to $p>0$ and noting that each term is repeated at most $B+l-d$ times we obtain:

$$\begin{aligned}\sum_{k=0}^{p} S_l^d(w^t)_{t\le k} &\le (B+l-d)\sum_{k=[1-B-l]_+}^{p-d}\|w^{k+1}-w^k\|^2\\ &\le (B+l-d)\sum_{k=0}^{p}\|w^{k+1}-w^k\|^2. \end{aligned} \tag{7.18}$$

This inequality will be used in the convergence analysis.

Lemma 7.3. *Suppose that $\mu_h^i, \mu_g^i > 0$, $i=1,\dots,m$, and in the case of Lem. 7.3(i) let Assumption 7.II(iv) hold. Then, the following hold for any positive constants ϵ_1, ϵ_2, ϵ_3, nonnegative integer q and generic vectors $v=(v_1,\dots,v_m)$, $y=(y_1,\dots,y_m)$ with $v_i\in\mathbb{R}^{n_i}$ and $y_i\in\mathbb{R}^{r_i}$:*

(i) $\sum_{i=1}^m \langle \nabla_i f(x^k[i]) - \nabla_i f(x^k), x_i^\star - v_i\rangle \le \frac{\epsilon_1}{2}\|v-x^\star\|_{M_g}^2 + \frac{B}{2\epsilon_1}\|\bar\beta\|_{M_g^{-1}}^2 S_1^1(x^t)_{t\le k}$

(ii) $\sum_{i=1}^m \langle L_{\bullet i}^\top\big(u^k[i]-u^{k+q}\big), x_i^\star - v_i\rangle \le \frac{\epsilon_2}{2}\|v-x^\star\|_{M_g}^2 + \frac{C_s(B+q)}{2\epsilon_2} S_1^{1-q}(u^t)_{t\le k}$

(iii) $\sum_{i=1}^m \langle L_{i\bullet}(x^k[i]-x^{k+q}), y_i - u_i^\star\rangle \le \frac{\epsilon_3}{2}\|y-u^\star\|_{M_h}^2 + \frac{R_s(B+q)}{2\epsilon_3} S_1^{1-q}(x^t)_{t\le k}$

where R_{s}, C_{s} are defined in (7.30), M_g, M_h in (7.20) and (7.31a).

Proof. We provide the proof for the first inequality and omit the rest noting that they are derived following a similar argument. Using the Cauchy–Schwarz inequality we have

$$\begin{aligned}\sum_{i=1}^m \langle x_i^\star - v_i, \nabla_i f(x^k[i]) - \nabla_i f(x^k)\rangle &\le \sum_{i=1}^m \|v_i - x_i^\star\|\|\nabla_i f(x^k) - \nabla_i f(x^k[i])\|\\ &\overset{(7.9)}{\le} \sum_{i=1}^m \bar\beta_i \|v_i - x_i^\star\|\|x^k - x^k[i]\|\\ &\overset{(7.17)}{\le} \sum_{i=1}^m \bar\beta_i\|v_i-x_i^\star\|\Big(\sum_{\tau=[k-B]_+}^{k-1}\|x^{\tau+1}-x^\tau\|\Big)\end{aligned}$$

$$= \sum_{i=1}^{m} \sum_{\tau=[k-B]_+}^{k-1} \bar{\beta}_i \|v_i - x_i^\star\| \|x^{\tau+1} - x^\tau\|$$

$$\left((7.16) \text{ with } \varepsilon = \tfrac{\mu_g^i \epsilon_1}{B}\right) \quad \leq \tfrac{1}{2} \sum_{i=1}^{m} \sum_{\tau=[k-B]_+}^{k-1} \left(\tfrac{\mu_g^i \epsilon_1}{B} \|v_i - x_i^\star\|^2 + \tfrac{\bar{\beta}_i^2 B}{\mu_g^i \epsilon_1} \|x^{\tau+1} - x^\tau\|^2 \right)$$

$$\leq \tfrac{\epsilon_1}{2} \|v - x^\star\|_{M_g}^2 + \tfrac{B}{2\epsilon_1} \|\bar{\beta}\|_{M_g^{-1}}^2 \sum_{\tau=[k-B]_+}^{k-1} \|x^{\tau+1} - x^\tau\|^2,$$

proving the claim. □

7.4 The case of partial coupling

Throughout this section we consider the optimization problem (7.8) with partial coupling (when L has a block-diagonal structure). In this case, the coupling between agents is enacted only through the smooth function f (and not through L). The example of formation control in Section 7.1.1 can be cast in this form.

Under this setting problem (7.8) becomes

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) + \sum_{i=1}^{m} \big(g_i(x_i) + h_i(L_{ii} x_i) \big),$$

where L_{ii} is the i-th diagonal block of L, see (7.6). In order to solve this problem with the iterates in (7.12), agent i must receive those x_j's that are required for the computation of $\nabla_i f$ and all other operations are local. Let us define two sets of indices: those that are required to send their variables to i:

$$\mathcal{N}_i^{\text{in}} := \{ j \mid \nabla_i f \text{ depends on } x_j \},$$

and those that i must send x_i to as $\mathcal{N}_i^{\text{out}} := \{ j \mid i \in \mathcal{N}_j^{\text{in}} \}$.

Algorithm 7.1 summarizes the proposed scheme. At every iteration each agent i performs the updates described in (7.12) using the last information it has received from agents $j \in \mathcal{N}_i^{\text{in}}$. It then transmits the updated x_i^{k+1} to the agents that require it (possibly with different delay). Note that $x^k[i]$ was defined as the outdated version of the full vector x^k for simplicity of notation, and in practical implementation it would only involve the coordinates that are required for the computation of $\nabla_i f$.

Algorithm 7.1 Vũ-Condat algorithm with bounded delays

Initialize: $x_i^0 \in \mathbb{R}^{n_i}$, $u_i^0 \in \mathbb{R}^{r_i}$ for $i \in \{1, \dots, m\}$.

For $k = 0, 1, \dots$ **do**

For each agent $i = 1, \dots, m$ **do**

% Local updates

perform the local updates using the last received information, i.e., the locally stored vector $x^k[i]$ as defined in (7.3):

1: $x_i^{k+1} = \operatorname{prox}_{\gamma_i g_i}\left(x_i^k - \gamma_i L_{ii}^\top u_i^k - \gamma_i \nabla_i f(x^k[i])\right)$

2: $u_i^{k+1} = \operatorname{prox}_{\sigma_i h_i^*}\left(u_i^k + \sigma_i L_{ii}(2x_i^{k+1} - x_i^k)\right)$

% Broadcasting to neighbors

3: send x_i^{k+1} to all $j \in \mathcal{N}_i^{\text{out}}$ (possibly with different delays)

As shown in Theorem 7.4, for small enough stepsizes the generated sequence converges to a primal-dual solution under the bounded delay assumption, and provided that functions g_i are strongly convex. Such needed requirements are summarized below:

Assumption 7.III. (stepsize condition) For $i = 1, \dots, m$, the stepsizes $\sigma_i, \gamma_i > 0$ satisfy the following assumption:

$$\gamma_i < \frac{1}{\sigma_i \|L_{ii}\|^2 + \beta + \frac{B^2}{2}\|\bar{\beta}\|^2_{M_g^{-1}}}, \tag{7.19}$$

where

$$M_g = \operatorname{blkdiag}\left(\mu_g^1 I_{n_1}, \dots, \mu_g^m I_{n_m}\right). \tag{7.20}$$

According to Assumption 7.III a one time global communication of $\|\bar{\beta}\|_{M_g^{-1}}$ and β is required when initiating the algorithm.

According to (7.19) as the upper bound on delay, B, and coupling constants $\bar{\beta}_i$ (as defined in (7.9)) increase, smaller stepsizes should be used. This is intuitive given that in either case the agents have a lower confidence in the currently stored vectors and thus should take smaller steps. Moreover, the higher the modulus of strong convexity, the larger steps agents are allowed to take, countering the effect of the delay.

In the case when the smooth term is separable, the problem is decoupled ($\bar{\beta}_i = 0$ for all i) and the stepsize condition for each agent does not depend on the delay. The same stepsize condition would be required in the case of synchronous

updates ($B = 0$). Note that, in this case (7.19) is still more conservative than the classical results which would require $\gamma_i < 1/(\sigma_i \|L_{ii}\|^2 + \beta/2)$ (the difference being a β appearing in place of $\beta/2$).

Before proceeding with the convergence results, let us define the following

$$P := \begin{pmatrix} \Gamma^{-1} & -L^\top \\ -L & \Sigma^{-1} \end{pmatrix}. \tag{7.21}$$

Noting that Σ, Γ are positive definite, and using Schur complement we have that P is positive definite if and only if $\Gamma^{-1} - L^\top \Sigma L$ is positive definite, a condition that holds if (7.19) is satisfied (since L has a block-diagonal structure).

Our analysis in Theorem 7.4 relies on showing that the generated sequence is quasi-Fejér monotone with respect to the set of primal-dual solutions in the space equipped with the inner product $\langle \cdot, \cdot \rangle_P$. Notice that without communication delays ($B \equiv 0$), this analysis leads to the usual Fejér monotonicity of the sequence. The use of outdated information introduces additional error terms that are shown to be tolerated by the algorithm if the stepsizes are small enough and the functions g_i are strongly convex.

Theorem 7.4. *Suppose that Assumptions 7.I–7.III are satisfied. Then, the sequence $(z^k)_{k\in\mathbb{N}} = (x^k, u^k)_{k\in\mathbb{N}}$ generated by Algorithm 7.1 is P-quasi-Fejér monotone with respect to $\mathcal{S}$. Furthermore, $(z^k)_{k\in\mathbb{N}}$ converges to some $z^\star \in \mathcal{S}$.*

Proof. In order to establish convergence we first derive the following intermediate result.

Lemma 7.5. *Suppose that Assumption 7.I, 7.II and 7.II(ii) are satisfied. Consider the sequence generated by Algorithm 7.1. Then, for any $(x^\star, u^\star) \in \mathcal{S}$ the following hold:*

$$\begin{aligned} &\|x^{k+1} - x^\star\|^2_{2M_g + \Gamma^{-1}} - \|x^k - x^\star\|^2_{\Gamma^{-1}} + \|x^k - x^{k+1}\|^2_{\Gamma^{-1} - \beta I} \\ &\leq 2 \sum_{i=1}^m \langle \nabla_i f(x^k[i]) - \nabla_i f(x^k), x_i^\star - x_i^{k+1} \rangle + 2 \langle L^\top (u^k - u^\star), x^\star - x^{k+1} \rangle, \end{aligned} \tag{7.22}$$

and

$$\|u^{k+1} - u^\star\|^2_{\Sigma^{-1}} - \|u^k - u^\star\|^2_{\Sigma^{-1}} + \|u^k - u^{k+1}\|^2_{\Sigma^{-1}} \leq 2 \langle L(2x^{k+1} - x^k - x^\star), u^{k+1} - u^\star \rangle. \tag{7.23}$$

Proof. To derive the first inequality use (7.14) with $q = g_i$, $r = x_i^\star$, $\omega_\rho = x_i^{k+1}$ and $\omega = x_i^k - \gamma_i L_{ii}^\top u^k[i] - \gamma_i \nabla_i f(x^k[i])$ (using the update for the primal variable x_i^{k+1} in the algorithm):

$$g_i(x_i^\star) - g_i(x_i^{k+1}) \geq \langle \nabla_i f(x^k[i]) + L_{ii}^\top u_i^k, x_i^{k+1} - x_i^\star \rangle$$
$$+ \tfrac{1}{\gamma_i}\langle x_i^k - x_i^{k+1}, x_i^\star - x_i^{k+1}\rangle + \tfrac{\mu_g^i}{2}\|x_i^\star - x_i^{k+1}\|^2. \tag{7.24}$$

Let $f^\star := f(x^\star)$, and $f^k := f(x^k)$. By convexity of f:

$$f^k - f^\star \leq \langle \nabla f(x^k), x^k - x^\star \rangle,$$

and by Lipschitz continuity of ∇f we have

$$f^{k+1} \leq f^k + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \tfrac{\beta}{2}\|x^{k+1} - x^k\|^2.$$

Summing the two yields

$$f^{k+1} - f^\star \leq \langle \nabla f(x^k), x^{k+1} - x^\star \rangle + \tfrac{\beta}{2}\|x^{k+1} - x^k\|^2. \tag{7.25}$$

For notational convenience we use $F := f + g$ and $F^\star := F(x^\star)$, $F^k := F(x^k)$. Noting that g is separable, sum (7.24) over i, add (7.25) and use (7.15) (with $V = \Gamma^{-1}$, $a = x^k$, $b = x^{k+1}$, $c = x^\star$) to obtain

$$\begin{aligned} F^\star - F^{k+1} \geq \quad & \tfrac{1}{2}\|x^k - x^{k+1}\|^2_{\Gamma^{-1} - \beta I} + \tfrac{1}{2}\|x^{k+1} - x^\star\|^2_{\Gamma^{-1} + M_g} \\ & - \tfrac{1}{2}\|x^k - x^\star\|^2_{\Gamma^{-1}} + \sum_{i=1}^m \langle -L_{ii}^\top u_i^k, x_i^\star - x_i^{k+1}\rangle \\ & + \sum_{i=1}^m \langle \nabla_i f(x^k) - \nabla_i f(x^k[i]), x_i^\star - x_i^{k+1}\rangle. \end{aligned}$$

On the other hand by convexity of f, strong convexity of g_i and (7.10) we have

$$F^{k+1} - F^\star \geq \langle -L^\top u^\star, x^{k+1} - x^\star \rangle + \tfrac{1}{2}\|x^{k+1} - x^\star\|^2_{M_g}.$$

Summing the last two inequalities, multiplying by 2 and a simple rearrangement yields (7.22).

For the second inequality, consider the update for u_i^{k+1} and use (7.14):

$$h_i^*(u_i^\star) - h_i^*(u_i^{k+1}) \geq \langle L_{ii}\big(2x_i^{k+1} - x_i^k\big), u_i^\star - u_i^{k+1}\rangle$$
$$+ \tfrac{1}{\sigma_i}\langle u_i^k - u_i^{k+1}, u_i^\star - u_i^{k+1}\rangle. \tag{7.26}$$

Furthermore, by convexity of h_i and using (7.10):

$$h^*(u^{k+1}) - h^*(u^\star) \geq \langle Lx^\star, u^{k+1} - u^\star \rangle.$$

Sum (7.26) over all i, add the last inequality, and use (7.15) to derive the inequality. □

Adding (7.22) and (7.23) we obtain

$$\begin{aligned}
&\|x^{k+1} - x^\star\|^2_{2M_g + \Gamma^{-1}} - \|x^k - x^\star\|^2_{\Gamma^{-1}} + \|x^k - x^{k+1}\|^2_{\Gamma^{-1}} \\
&+ \|u^{k+1} - u^\star\|^2_{\Sigma^{-1}} - \|u^k - u^\star\|^2_{\Sigma^{-1}} + \|u^k - u^{k+1}\|^2_{\Sigma^{-1}} \\
\leq \quad & 2\sum_{i=1}^m \langle x_i^\star - x_i^{k+1}, \nabla_i f(x^k[i]) - \nabla_i f(x^k) \rangle + \beta \|x^k - x^{k+1}\|^2 \\
&+ 2\langle L(2x^{k+1} - x^k - x^\star), u^{k+1} - u^\star \rangle + 2\langle L^\top (u^k - u^\star), x^\star - x^{k+1} \rangle. \quad (7.27)
\end{aligned}$$

The last two inner products can be rearranged as

$$2\langle L(x^{k+1} - x^k), u^{k+1} - u^k \rangle - 2\langle L(x^k - x^\star), u^k - u^\star \rangle + 2\langle L(x^{k+1} - x^\star), u^{k+1} - u^\star \rangle.$$

Replacing this term and using Lemma 7.3*(i)* (with $\epsilon_1 = 2$ and $v = x^{k+1}$) in (7.27) yields (with P defined in (7.21)):

$$\begin{aligned}
&\|z^{k+1} - z^\star\|^2_P - \|z^k - z^\star\|^2_P + \|z^{k+1} - z^k\|^2_P \\
&\leq \beta \|x^k - x^{k+1}\|^2 + \tfrac{B}{2} \|\bar{\beta}\|^2_{M_g^{-1}} S_1^1(x^t)_{t \leq k}. \quad (7.28)
\end{aligned}$$

Sum inequality (7.28) over k from 0 to $p > 0$ to obtain:

$$\begin{aligned}
&\|z^{p+1} - z^\star\|^2_P - \|z^0 - z^\star\|^2_P + \sum_{k=0}^p \|z^{k+1} - z^k\|^2_P \\
&\leq \beta \sum_{k=0}^p \|x^k - x^{k+1}\|^2 + \tfrac{B}{2} \|\bar{\beta}\|^2_{M_g^{-1}} \sum_{k=0}^p S_1^1(x^t)_{t \leq k}. \quad (7.29)
\end{aligned}$$

Let us define

$$\tilde{P} = \begin{pmatrix} \Gamma^{-1} - \frac{B^2}{2} \|\bar{\beta}\|^2_{M_g^{-1}} - \beta & -L^\top \\ -L & \Sigma^{-1} \end{pmatrix}.$$

Since Σ is positive definite ($\sigma_i > 0$), by Schur complement $\tilde{P}$ is positive definite provided that (7.19) holds (recall that L has a block-diagonal structure).

Use (7.18) (with $l = d = 1$) in (7.29) to derive

$$\|z^{p+1} - z^\star\|_P^2 + \sum_{k=0}^{p} \|z^{k+1} - z^k\|_{\tilde{P}}^2 \leq \|z^0 - z^\star\|_P^2.$$

Therefore, by letting p to infinity we obtain $\sum_{k=0}^{\infty} \|z^{k+1} - z^k\|_{\tilde{P}}^2 < \infty$. Hence, using (7.18) for the right-hand side in (7.28) we have

$$\sum_{k=0}^{\infty} \left(\tfrac{B}{2} \|\bar{\beta}\|_{M_g^{-1}}^2 S_1^1(x^t)_{t\leq k} + \beta \|x^k - x^{k+1}\|^2 \right) < \infty.$$

Therefore, in view of (7.28) we conclude that $(z^k)_{k\in\mathbb{N}}$ is P-quasi-Fejér monotone with respect to $\mathcal{S}$.

Consequently, the sequence $(x^k, u^k)_{k\in\mathbb{N}}$ is bounded [46, Lem. 3.1]. Let (x^c, u^c) be a cluster point of $(x^k, u^k)_{k\in\mathbb{N}}$, *i.e.*, $(x^{k_n}, u^{k_n}) \to (x^c, u^c)$. Using Lemma 7.2 also $x^{k_n}[i] \to x^c$. Noting that the proximal and linear maps as well as ∇f are continuous, for all $i = 1, \ldots, m$ we have

$$x_i^c = \operatorname{prox}_{\gamma_i g_i} \left(x_i^c - \gamma_i \nabla_i f(x^c) - \gamma_i L_{ii}^\top u_i^c \right)$$

$$u_i^c = \operatorname{prox}_{\sigma_i h_i^\star} \left(u_i^c + \sigma_i L_{ii} x_i^c \right),$$

which implies $(x^c, u^c) \in \mathcal{S}$. The convergence of the sequence follows [46, Thm. 3.8]. □

In the case of total coupling (when L is not block-diagonal), it is no longer possible to establish quasi-Fejér monotonicity of the Vũ-Condat generated sequence in the space equipped with $\langle \cdot, \cdot \rangle_P$. This is because the coupling linear mapping L, is operating on outdated vectors. In the next section we propose an AHU-type primal-dual algorithm that is better suited for problems with total coupling.

7.5 The case of total coupling

In this section we consider problem (7.8) with *total coupling*. That is, we assume that the coupling between agents is enforced through the linear maps (L is not block-diagonal), and possibly through the smooth term f.

7.5.1 An AHU-type primal-dual algorithm

We consider the primal-dual algorithm (7.13). Compared to (7.12), in the dual update the linear map $L_{i\bullet}$ operates on $x^k[i]$ in place of $2x^{k+1}[i] - x^k[i]$. This modification results in the possibility of using larger stepsizes since the terms $2x^{k+1}[i] - x^k[i]$ would introduce additional sources of error.

Let us define the following two sets of indices:

$$\mathcal{M}_i^{\mathrm{p}} := \{j \mid L_{ji} \neq 0\}, \quad \mathcal{M}_i^{\mathrm{d}} := \{j \mid L_{ij} \neq 0\},$$

where 0 denotes a zero matrix of appropriate dimensions. In Algorithm 7.2, due to the additional coupling through the linear maps, the primal vector of agent i must be transmitted to all $j \in \mathcal{M}_i^{\mathrm{p}} \cup \mathcal{N}_i^{\mathrm{out}}$ while the dual vector is to be transmitted to all $j \in \mathcal{M}_i^{\mathrm{d}}$. Notice that the outdated primal and dual vectors $x^k[i]$ and $u^k[i]$, need not have the same delay pattern and are arbitrary as long as Assumption 7.I is satisfied, *i.e.*, agent i may use the primal vector $x_j^{k_1}$ and the dual vector $u_j^{k_2}$ that were the variables of agent j at times k_1 and k_2.

Algorithm 7.2 An AHU-type primal-dual algorithm with bounded delays

Initialize: $x_i^0 \in \mathbb{R}^{n_i}$, $u_i^0 \in \mathbb{R}^{r_i}$ for $i \in \{1, \dots, m\}$.

For $k = 0, 1, \dots$ **do**

For each agent $i = 1, \dots, m$ **do**

% Local updates

perform the local updates using the last received information, i.e., the locally stored vectors $x^k[i]$ and $u^k[i]$ as defined in (7.3):

1: $x_i^{k+1} = \operatorname{prox}_{\gamma_i g_i}\left(x_i^k - \gamma_i L_{\bullet i}^\top u^k[i] - \gamma_i \nabla_i f(x^k[i])\right)$

2: $u_i^{k+1} = \operatorname{prox}_{\sigma_i h_i^\star}\left(u_i^k + \sigma_i L_{i\bullet} x^k[i]\right)$

% Broadcasting to neighbors

3: send x_i^{k+1} to all $j \in \mathcal{N}_i^{\mathrm{out}} \cup \mathcal{M}_i^{\mathrm{p}}$, and u_i^{k+1} to all $j \in \mathcal{M}_i^{\mathrm{d}}$ (possibly with different delays)

In Theorem 7.6 convergence is established for Algorithm 7.2 when the stepsizes are small enough, under the assumption that the functions g_i are strongly convex and h_i are continuously differentiable with Lipschitz continuous gradient. Note that under this extra assumption the set of primal-dual solutions is a singleton, $\mathcal{S} = \{z^\star\}$. We summarize these requirements below:

Assumption 7.IV. For all $i = 1, \dots, m$:

(i) (Lipschitz continuity) h_i is continuously differentiable, and ∇h_i is $\frac{1}{\mu_h^i}$-Lipschitz continuous for some $\mu_h^i > 0$. Equivalently, h_i^* is μ_h^i-strongly convex.

(ii) (stepsize condition) The stepsizes $\sigma_i, \gamma_i > 0$ satisfy the following inequalities

$$\sigma_i < \frac{1}{C_{\mathrm{s}}(B+1)^2}, \quad \gamma_i < \frac{1}{\beta + \frac{1}{2}R_{\mathrm{s}}(B+1)^2 + B^2\|\bar{\beta}\|^2_{M_g^{-1}}},$$

where

$$R_{\mathrm{s}} := \sum_{i=1}^m \tfrac{1}{\mu_h^i}\|L_{i\bullet}\|^2, \quad C_{\mathrm{s}} := \sum_{i=1}^m \tfrac{1}{\mu_g^i}\|L_{\bullet i}^\top\|^2. \tag{7.30}$$

Throughout this section we make use of the following positive definite matrices.

$$M_h := \mathrm{blkdiag}(\mu_h^1 I_{r_1}, \ldots, \mu_h^m I_{r_m}), \tag{7.31a}$$

$$M := \mathrm{blkdiag}(M_g, M_h). \tag{7.31b}$$

Note that according to Assumption 7.IV*(ii)* a one time global communication of R_{s}, C_{s}, β and $\|\bar{\beta}\|_{M_g^{-1}}$ is required.

The stepsize condition in Assumption 7.IV*(ii)* is more stringent than the condition derived in Section 7.4 for the case of partial coupling. This is due to the fact that $L_{i\bullet}$ and $L_{\bullet i}$ are operating on delayed vectors. It is shown in Theorem 7.6 that for the case of total coupling with AHU-type algorithm quasi-Fejér monotonicity holds in the space equipped with $\langle\cdot,\cdot\rangle_D$ (with D defined as in (7.11)).

We proceed with the convergence results for Algorithm 7.2.

Theorem 7.6. *Suppose that Assumptions 7.I, 7.II and 7.IV are satisfied. Then, the sequence $(z^k)_{k\in\mathbb{N}} = (x^k, u^k)_{k\in\mathbb{N}}$ generated by Algorithm 7.2 is D-quasi-Fejér monotone with respect to $\mathcal{S} = \{z^\star\}$, and converges to $z^\star$.*

Proof. The proof is similar to that of Theorem 7.4. In this case, the presence of coupling through the linear maps results in additional error terms. First we establish a key result for Algorithm 7.2.

Lemma 7.7. *Suppose that Assumption 7.I, 7.II and 7.IV(i) are satisfied. Consider the sequence generated by Algorithm 7.2. Then, for any $(x^\star, u^\star) \in \mathcal{S}$*

the following hold:

$$\|x^{k+1}-x^\star\|^2_{2M_g+\Gamma^{-1}} - \|x^k-x^\star\|^2_{\Gamma^{-1}} + \|x^k-x^{k+1}\|^2_{\Gamma^{-1}-\beta I}$$

$$\leq 2\sum_{i=1}^m \langle \nabla_i f(x^k[i]) - \nabla_i f(x^k) + L_{\bullet i}^\top (u^k[i]-u^\star), x_i^\star - x_i^{k+1}\rangle, \tag{7.32}$$

$$\|u^{k+1}-u^\star\|^2_{2M_h+\Sigma^{-1}} - \|u^k-u^\star\|^2_{\Sigma^{-1}} + \|u^k-u^{k+1}\|^2_{\Sigma^{-1}}$$

$$\leq 2\sum_{i=1}^m \langle L_{i\bullet} x^k[i], u_i^{k+1}-u_i^\star\rangle + 2\langle Lx^\star, u^\star - u^{k+1}\rangle. \tag{7.33}$$

Proof. The proof of the lemma is similar to that of Lemma 7.5 and is therefore omitted. □

Add (7.32) and (7.33), and rearrange the inner products using (7.7) to derive (with D defined in (7.11))

$$\begin{aligned}
&\|z^{k+1}-z^\star\|_D^2 - \|z^k-z^\star\|_D^2 + \|z^k-z^{k+1}\|_D^2 \\
&- \beta\|x^k-x^{k+1}\|^2 + \|x^{k+1}-x^\star\|^2_{2M_g} + \|u^{k+1}-u^\star\|^2_{2M_h} \\
\leq\ & 2\sum_{i=1}^m \langle \nabla_i f(x^k[i]) - \nabla_i f(x^k) + L_{\bullet i}^\top (u^k[i]-u^{k+1}), x_i^\star - x_i^{k+1}\rangle \\
&+ 2\sum_{i=1}^m \langle L_{i\bullet}(x^k[i]-x^{k+1}), u_i^{k+1}-u_i^\star\rangle
\end{aligned} \tag{7.34}$$

Using the inequalities in Lemma 7.3 with $q=1$, $\epsilon_1=\epsilon_2=1$, $\epsilon_3=2$, $v=x^{k+1}$ and $y=u^{k+1}$ yields:

$$\begin{aligned}
&\|z^{k+1}-z^\star\|_D^2 - \|z^k-z^\star\|_D^2 + \|z^k-z^{k+1}\|_D^2 \\
\leq\ & B\|\bar\beta\|^2_{M_g^{-1}} S_1^1(x^t)_{t\leq k} + \tfrac{1}{2}R_{\mathrm{s}}(B+1)S_1^0(x^t)_{t\leq k} \\
&+ C_{\mathrm{s}}(B+1)S_1^0(u^t)_{t\leq k} + \beta\|x^k-x^{k+1}\|^2.
\end{aligned} \tag{7.35}$$

Sum over k from 0 to $p > 0$, to derive

$$\begin{aligned}
&\|z^{p+1} - z^\star\|_D^2 - \|z^0 - z^\star\|_D^2 + \sum_{k=0}^{p} \|z^k - z^{k+1}\|_D^2 \\
\leq\ & B\|\bar{\beta}\|_{M_g^{-1}}^2 \sum_{k=0}^{p} S_1^1(x^t)_{t\leq k} + \tfrac{R_{\mathrm{s}}(B+1)}{2} \sum_{k=0}^{p} S_1^0(x^t)_{t\leq k} \\
&+ \beta \sum_{k=0}^{p} \|x^k - x^{k+1}\|^2 + C_{\mathrm{s}}(B+1) \sum_{k=0}^{p} S_1^0(u^t)_{t\leq k}. \qquad (7.36)
\end{aligned}$$

By repeated use of (7.18) in (7.36) we obtain

$$\begin{aligned}
&\|z^{p+1} - z^\star\|_D^2 - \|z^0 - z^\star\|_D^2 + \sum_{k=0}^{p} \|z^k - z^{k+1}\|_D^2 \\
&\leq \left(B^2\|\bar{\beta}\|_{M_g^{-1}}^2 + \tfrac{1}{2}R_{\mathrm{s}}(B+1)^2 + \beta\right) \sum_{k=0}^{p} \|x^k - x^{k+1}\|^2 \\
&\quad + C_{\mathrm{s}}(B+1)^2 \sum_{k=0}^{p} \|u^k - u^{k+1}\|^2.
\end{aligned}$$

If the stepsizes are small enough to satisfy Assumption 7.IV *(ii)*, letting p to infinity yields $\sum_{k=0}^{\infty} \|z^{k+1} - z^k\|^2 < \infty$. Therefore, it follows from (7.35) (using (7.18)) that $(z^k)_{k\in\mathbb{N}}$ is D-quasi-Fejér monotone with respect to $\mathcal{S}$. Arguing as in Theorem 7.4 completes the proof. □

The next theorem provides a sufficient condition for the stepsizes under which linear convergence is attained.

Theorem 7.8 (linear convergence)**.** *Suppose that Assumption 7.I, 7.II and 7.IV(i) are satisfied. Consider the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 7.2. Let c be a positive scalar and set $\gamma_i = \frac{c}{\mu_g^i}, \sigma_i = \frac{c}{\mu_h^i}$ for $i = 1, \ldots, m$. Let $\mu_g^{\min} = \min\{\mu_g^1, \ldots, \mu_g^m\}$, $\mu_h^{\min} = \min\{\mu_h^1, \ldots, \mu_h^m\}$. Then, the following linear convergence rate holds*

$$\|z^k - z^\star\|_D^2 \leq \left(\tfrac{1}{1+c}\right)^k \|z^0 - z^\star\|_D^2,$$

provided that $c \leq (1+c_2)^{\frac{1}{B+1}} - 1$ *where*

$$c_2 = \min\left\{ \frac{\mu_g^{\min}}{2B\|\bar{\beta}\|^2_{M_g^{-1}} + R_s(B+1) + \beta}, \frac{\mu_h^{\min}}{2C_s(B+1)} \right\}.$$

Proof. In Theorem 7.6 the strong convexity assumption was leveraged to counteract the error terms. In order to prove linear convergence we retain some of the strong convexity terms. Using the inequalities of Lemma 7.3 with $q = 1$, $\epsilon_1 = \epsilon_2 = 0.5$, $\epsilon_3 = 1$, $v = x^{k+1}$ and $y = u^{k+1}$ in (7.34) yields

$$\begin{aligned} &\|z^{k+1} - z^\star\|^2_D - \|z^k - z^\star\|^2_D + \|z^k - z^{k+1}\|^2_D \\ &\quad + \|x^{k+1} - x^\star\|^2_{M_g} + \|u^{k+1} - u^\star\|^2_{M_h} \\ &\leq \Big(2B\|\bar{\beta}\|^2_{M_g^{-1}} + R_s(B+1)\Big) S_1^0(x^t)_{t\leq k} \\ &\quad + 2C_s(B+1)S_1^0(u^t)_{t\leq k} + \beta\|x^k - x^{k+1}\|^2. \end{aligned} \tag{7.37}$$

Note that one may set these constants differently and obtain a different valid bound on the stepsizes.

Since we set $\gamma_i = \frac{c}{\mu_g^i}$, $\sigma_i = \frac{c}{\mu_h^i}$, we have $D = \operatorname{blkdiag}(\Gamma^{-1}, \Sigma^{-1}) = \frac{1}{c}\operatorname{blkdiag}(M_g, M_h)$, which together with (7.37) yields

$$\begin{aligned} (1+c)\|z^{k+1} - z^\star\|^2_D - \|z^k - z^\star\|^2_D \leq \Big(2B\|\bar{\beta}\|^2_{M_g^{-1}} + R_s(B+1) + \beta\Big) S_1^0(x^t)_{t\leq k} \\ + 2C_s(B+1)S_1^0(u^t)_{t\leq k} - \|z^{k+1} - z^k\|^2_D, \end{aligned} \tag{7.38}$$

where we used the conservative bound $\beta\|x^{k+1} - x^k\|^2 \leq \beta S_1^0(x^t)_{t\leq k}$ in order to avoid algebraic difficulties. The result follows by multiplying (7.38) by $(1+c)^k$ and summing over k from 0 to p, see [12, Lem. 1]. □

7.5.2 Randomized variant

In this subsection we propose a randomized variant of Algorithm 7.2 where agents are activated randomly according to independent probabilities, *i.e.*, at every iteration several agents may be active. Unlike the partially asynchronous protocol [20], in this scheme the agents are not required to perform at least one update in any interval of length B. In the randomized setting of Algorithm

Algorithm 7.3 A randomized variant of Algorithm 7.2

Initialize: $x_i^0 \in \mathbb{R}^{n_i}$, $u_i^0 \in \mathbb{R}^{r_i}$ for $i \in \{1, \dots, m\}$.

For $k = 0, 1, \dots$ **do**

each agent $i = 1, \dots, m$ is activated independantly with probability $p_i > 0$.

For active agents **do**

% Local updates

perform the local updates using the last received information, i.e., the locally stored vectors $x^k[i]$ and $u^k[i]$ as defined in (7.3):

1: $x_i^{k+1} = \operatorname{prox}_{\gamma_i g_i}\left(x_i^k - \gamma_i L_{\bullet i}^\top u^k[i] - \gamma_i \nabla_i f(x^k[i])\right)$

2: $u_i^{k+1} = \operatorname{prox}_{\sigma_i h_i^\star}\left(u_i^k + \sigma_i L_{i\bullet} x^k[i]\right)$

% Broadcasting to neighbors

3: send x_i^{k+1} to all $j \in \mathcal{N}_i^{\text{out}} \cup \mathcal{M}_i^{\text{p}}$, and u_i^{k+1} to all $j \in \mathcal{M}_i^{\text{d}}$ (possibly with different delays)

7.3, the stepsize condition in Assumption 7.IV *(ii)* is replaced by the following stepsize condition.

Assumption 7.V. (stepsize condition) For all $i = 1, \dots, m$, independent probabilities $p_i > 0$ and stepsizes $\sigma_i, \gamma_i > 0$ satisfy the following inequalities

$$\sigma_i < \frac{1}{2C_s(B^2 p_i + 1)}, \quad \gamma_i < \frac{1}{\beta + R_s(B^2 p_i + 1) + \|\bar{\beta}\|^2_{M_g^{-1}} B^2 p_i}.$$

Notice that compared to the non-randomized version, according to Assumption 7.V, an agent is allowed to take larger steps if its probability of activation is smaller.

Theorem 7.9. *Suppose that Assumption 7.I, 7.II, 7.IV(i) and 7.V are satisfied. Then the sequence $(z^k)_{k\in\mathbb{N}} = (x^k, u^k)_{k\in\mathbb{N}}$ generated by Algorithm 7.3 converges almost surely to $z^\star$.*

Proof. Let $\bar{z}_i^{k+1} = (\bar{x}_i^{k+1}, \bar{u}_i^{k+1})$ denote the updated vector belonging to agent i if that agent was to perform an update at iteration k. That is, in Algorithm 7.3, $z_i^{k+1} = \bar{z}_i^{k+1}$ if agent i is activated and $z_i^{k+1} = z_i^k$ if it remains idle. Let us define the global vector $\bar{z}^{k+1} = (\bar{z}_1^{k+1}, \dots, \bar{z}_m^{k+1})$ which corresponds to a deterministic update of all agents at iteration k. Using Lemma 7.7 as in (7.34) we have

$$\|\bar{z}^{k+1} - z^\star\|_D^2 - \|z^k - z^\star\|_D^2 + \|z^k - \bar{z}^{k+1}\|_D^2$$

$$
\begin{aligned}
&- \beta\|x^k - \bar{x}^{k+1}\|^2 + \|\bar{x}^{k+1} - x^\star\|^2_{2M_g} + \|\bar{u}^{k+1} - u^\star\|^2_{2M_h} \\
\leq\ & 2\sum_{i=1}^m \langle \nabla_i f(x^k[i]) - \nabla_i f(x^k) + L_{\bullet i}^\top (u^k[i] - \bar{u}^{k+1}), x_i^\star - \bar{x}_i^{k+1}\rangle \\
&+ 2\sum_{i=1}^m \langle L_{i\bullet}(x^k[i] - \bar{x}^{k+1}), \bar{u}_i^{k+1} - u_i^\star\rangle \qquad (7.39)
\end{aligned}
$$

Using Lemma 7.3*(ii)* with $q = 0$, $v = \bar{x}^{k+1}$, and (7.16) (with $\epsilon = \epsilon_2\mu_g^i$ for each term in the summation) we have

$$
\begin{aligned}
&\sum_{i=1}^m \langle L_{\bullet i}^\top (u^k[i] - u^k + u^k - \bar{u}^{k+1}), x_i^\star - \bar{x}_i^{k+1}\rangle \\
\leq\ & \tfrac{\epsilon_2}{2}\|x^\star - \bar{x}^{k+1}\|^2_{M_g} + \tfrac{BC_s}{2\epsilon_2} S_1^1(u^t)_{t\leq k} \\
&+ \sum_{i=1}^m \left(\|L_{\bullet i}^\top\|^2 \tfrac{1}{2\epsilon_2\mu_g^i}\|u^k - \bar{u}^{k+1}\|^2 + \tfrac{\epsilon_2\mu_g^i}{2}\|x_i^\star - \bar{x}_i^{k+1}\|^2 \right) \\
=\ & \epsilon_2\|x^\star - \bar{x}^{k+1}\|^2_{M_g} + \tfrac{C_s}{2\epsilon_2}\|\bar{u}^{k+1} - u^k\|^2 + \tfrac{BC_s}{2\epsilon_2} S_1^1(u^t)_{t\leq k}. \qquad (7.40)
\end{aligned}
$$

Similarly, using Lemma 7.3*(iii)* with $q = 0$, $y = \bar{u}^{k+1}$ and (7.16) (with $\epsilon = \epsilon_3\mu_h^i$ for each term in the summation) we obtain:

$$
\begin{aligned}
\sum_{i=1}^m \langle L_{i\bullet}(x^k[i] - \bar{x}^{k+1}), \bar{u}_i^{k+1} - u_i^\star\rangle &\leq \epsilon_3\|\bar{u}^{k+1} - u^\star\|^2_{M_h} + \tfrac{R_s}{2\epsilon_3}\|x^k - \bar{x}^{k+1}\|^2 \\
&+ \tfrac{BR_s}{2\epsilon_3} S_1^1(x^t)_{t\leq k}. \qquad (7.41)
\end{aligned}
$$

Using (7.40), (7.41) with $\epsilon_2 = 0.5$ and $\epsilon_3 = 1$ together with Lemma 7.3*(i)* with $\epsilon_1 = 1$, $v = \bar{x}^{k+1}$ in (7.39) yields

$$
\begin{aligned}
&\|\bar{z}^{k+1} - z^\star\|_D^2 - \|z^k - z^\star\|_D^2 + \|z^k - \bar{z}^{k+1}\|_D^2 \\
\leq\ & aS_1^1(x^t)_{t\leq k} + bS_1^1(u^t)_{t\leq k} + (\beta + R_s)\|x^k - \bar{x}^{k+1}\|^2 + 2C_s\|\bar{u}^{k+1} - u^k\|^2, \qquad (7.42)
\end{aligned}
$$

where $a := \left(BR_s + B\|\bar{\beta}\|^2_{M_g^{-1}}\right)$ and $b := 2BC_s$.

Let $\mathbb{E}_k[\cdot]$ denote the expectation conditioned on the knowledge until time k. Moreover, for notational convenience let us define the diagonal probability

matrix

$$\Pi = \mathrm{blkdiag}(p_1 \mathrm{I}_{n_1}, \dots, p_m \mathrm{I}_{n_m}, p_1 \mathrm{I}_{r_1}, \dots, p_m \mathrm{I}_{r_m}),$$

$$D_i = \mathrm{blkdiag}(\gamma_i^{-1} \mathrm{I}_{n_i} \sigma_i^{-1} \mathrm{I}_{r_i}),$$

$z_i^k = (x_i^k, u_i^k)$ and $z_i^\star = (x_i^\star, u_i^\star)$. Consequently, using the fact that D is diagonal we have

$$\begin{aligned}
\mathbb{E}_k\big\{\|z^{k+1} - z^\star\|^2_{\Pi^{-1}D}\big\} &= \mathbb{E}_k\left[\sum_{i=1}^m p_i^{-1}\|z_i^{k+1} - z_i^\star\|^2_{D_i}\right] \\
&= \sum_{i=1}^m p_i^{-1}\big(p_i\|\bar{z}_i^{k+1} - z_i^\star\|^2_{D_i} + (1-p_i)\|z_i^k - z_i^\star\|^2_{D_i}\big) \\
&= \sum_{i=1}^m \left(\|\bar{z}_i^{k+1} - z_i^\star\|^2_{D_i} + \tfrac{(1-p_i)}{p_i}\|z_i^k - z_i^\star\|^2_{D_i}\right) \\
&= \|\bar{z}^{k+1} - z^\star\|^2_D + \|z^k - z^\star\|^2_{\Pi^{-1}D} - \|z^k - z^\star\|^2_D.
\end{aligned} \tag{7.43}$$

Therefore, using (7.42) we obtain

$$\begin{aligned}
\mathbb{E}_k\big\{\|z^{k+1} - z^\star\|^2_{\Pi^{-1}D}\big\} \le \|z^k - z^\star\|^2_{\Pi^{-1}D} - \|z^k - \bar{z}^{k+1}\|^2_D + 2C_s\|\bar{u}^{k+1} - u^k\|^2 \\
+ aS_1^1(x^t)_{t\le k} + bS_1^1(u^t)_{t\le k} + (\beta + R_s)\|x^k - \bar{x}^{k+1}\|^2.
\end{aligned} \tag{7.44}$$

Let us define $X^k := \sum_{\tau=[k-B]_+}^{k-1}(\tau - (k-B) + 1)\|x^{\tau+1} - x^\tau\|^2$ and $U^k := \sum_{\tau=[k-B]_+}^{k-1}(\tau - (k-B) + 1)\|u^{\tau+1} - u^\tau\|^2$. It is easy to see that

$$X^{k+1} = X^k - \sum_{\tau=[k-B]_+}^{k-1}\|x^{\tau+1} - x^\tau\|^2 + B\|x^k - x^{k+1}\|^2.$$

Arguing as in (7.43) we have

$$\mathbb{E}_k\big\{\|x^{k+1} - x^k\|^2\big\} = \sum_{i=1}^m p_i\|\bar{x}_i^{k+1} - x_i^k\|^2.$$

Therefore

$$\mathbb{E}_k\{X^{k+1}\} \leq X^k - \sum_{\tau=[k-B]_+}^{k-1} \|x^{\tau+1} - x^\tau\|^2 + B\sum_{i=1}^m p_i\|\bar{x}_i^{k+1} - x_i^k\|^2, \quad (7.45)$$

Similarly for the dual variables we have

$$\mathbb{E}_k\{U^{k+1}\} \leq U^k - \sum_{\tau=[k-B]_+}^{k-1} \|u^{\tau+1} - u^\tau\|^2 + B\sum_{i=1}^m p_i\|\bar{u}_i^{k+1} - u_i^k\|^2. \quad (7.46)$$

Consider the following Lyapunov function:

$$v^k := \|z^k - z^\star\|^2_{\Pi^{-1}D} + aX^k + bU^k.$$

Then using (7.44), (7.45) and (7.46) we obtain

$$\begin{aligned}\mathbb{E}_k\{v^{k+1}\} \leq v^k &- \|z^k - \bar{z}^{k+1}\|_D^2 \\ &+ aB\sum_{i=1}^m p_i\|\bar{x}_i^{k+1} - x_i^k\|^2 + bB\sum_{i=1}^m p_i\|\bar{u}_i^{k+1} - u_i^k\|^2 \\ &+ (\beta + R_s)\|x^k - \bar{x}^{k+1}\|^2 + 2C_s\|\bar{u}^{k+1} - u^k\|^2.\end{aligned}$$

Therefore, if Assumption 7.V holds then there exists $\bar{c} > 0$ such that

$$\mathbb{E}_k\{v^{k+1}\} \leq v^k - \bar{c}\|z^k - \bar{z}^{k+1}\|^2.$$

Since $\|z^k - z^{k+1}\| \leq \|z^k - \bar{z}^{k+1}\|$, we conclude by the Robbins-Siegmund lemma [142] that almost surely $\|z^k - z^{k+1}\|$ converges to zero, and consequently by Lemma 7.2 so does $\|z^k - z^k[i]\|$. Moreover, as a second consequence of the Robbins-Siegmund lemma we have that $(v^k)_{k\in\mathbb{N}}$ and in particular $\|z^k - z^\star\|_{\Pi^{-1}D}$ converges to some $[0,\infty)$-valued variable. The convergence result follows by standard arguments as in [22, Thm. 3] and [51, Prop. 2.3] and using continuity of the proximal operator. □

In the next theorem we establish linear convergence for Algorithm 7.3 and provide an explicit convergence rate.

Theorem 7.10. *Suppose that Assumption 7.I, 7.II and 7.IV(i) are satisfied. Let* $c < \min\{p_1, \ldots, p_m\}$ *be a positive scalar and set* $\gamma_i = \frac{1}{(p_i/c-1)\mu_g^i}$ *and*

$\sigma_i = \frac{1}{(p_i/c-1)\mu_h^i}$ *for* $i = 1, \dots, m$. *Moreover, let*

$$\delta_1 = \frac{1}{2B\|\bar{\beta}\|^2_{M_g^{-1}} + 2BR_s + 2R_s + \beta} \min_i\{(p_i - c)\mu_g^i\},$$

$$\delta_2 = \frac{1}{4C_s(1+B)} \min_i\{(p_i - c)\mu_h^i\}.$$

Suppose that c *is such that the following inequality is satisfied (such* c *always exist close enough to zero)*

$$\frac{1}{(1-c)^B} + c \leq 1 + \min\{\delta_1, \delta_2\}.$$

Then, the following holds for the sequence $(z^k)_{k\in\mathbb{N}}$ *generated by Algorithm 7.3:*

$$\mathbb{E}\{\|z^k - z^\star\|_M^2\} \leq (1-c)^k \|z^0 - z^\star\|_M^2.$$

Proof. As in the deterministic case in Theorem 7.8, in order to show linear convergence we retain some of the strong convexity terms. Consider (7.39) and use Lemma 7.3*(i)* with $\epsilon_1 = 0.5$, $v = \bar{x}^{k+1}$, (7.40) and (7.41) with $\epsilon_2 = 0.25$, $\epsilon_3 = 0.5$ to derive:

$$\begin{aligned}
&\|\bar{z}^{k+1} - z^\star\|^2_{D+M} - \|z^k - z^\star\|^2_D + \|\bar{z}^{k+1} - z^k\|^2_D \\
&\leq 2B\left(\|\bar{\beta}\|^2_{M_g^{-1}} + R_s\right) S_1^1(x^t)_{t\leq k} + 4BC_s S_1^1(u^t)_{t\leq k} \\
&\quad + (2R_s + \beta)\|\bar{x}^{k+1} - x^k\|^2 + 4C_s\|\bar{u}^{k+1} - u^k\|^2, \qquad (7.47)
\end{aligned}$$

where $M := \mathrm{blkdiag}(M_g, M_h)$. Given the choice of stepsizes we have $D = \mathrm{blkdiag}(\Gamma^{-1}, \Sigma^{-1}) = \left(\frac{1}{c}\Pi - \mathrm{I}\right)M$. Using this and arguing as in (7.43) we have:

$$\begin{aligned}
&\mathbb{E}_k\left\{\|z^{k+1} - z^\star\|_M^2\right\} \\
&= \|\bar{z}^{k+1} - z^\star\|^2_{\Pi M} + \|z^k - z^\star\|^2_M - \|z^k - z^\star\|^2_{\Pi M} \\
&= c\|\bar{z}^{k+1} - z^\star\|^2_{D+M} + \|z^k - z^\star\|^2_M - c\|z^k - z^\star\|^2_{D+M}.
\end{aligned}$$

Combining this with (7.47) yields

$$\begin{aligned}
&\mathbb{E}_k\left\{\|z^{k+1} - z^\star\|_M^2\right\} \\
&\leq (1-c)\|z^k - z^\star\|_M^2 - c\|\bar{z}^{k+1} - z^k\|_D^2
\end{aligned}$$

$$+ 2cB\Big(\|\bar{\beta}\|^2_{M_g^{-1}} + R_{\mathrm{s}}\Big) S_1^1(x^t)_{t\le k} + 4cBC_s S_1^1(u^t)_{t\le k}$$
$$+ c(2R_s + \beta)\|\bar{x}^{k+1} - x^k\|^2 + 4cC_s\|\bar{u}^{k+1} - u^k\|^2. \tag{7.48}$$

Next, note that by definition of $\bar{z}^k = (\bar{x}^k, \bar{u}^k)$ we have

$$S_1^1(x^t)_{t\le k} \le \sum_{\tau=[k-B]_+}^{k-1} \|\bar{x}^{\tau+1} - x^\tau\|^2 \le \sum_{\tau=[k-B]_+}^{k} \|\bar{x}^{\tau+1} - x^\tau\|^2,$$

and similarly for the dual vector. Using this in (7.48) yields:

$$\begin{aligned}
&\mathbb{E}_k\big\{\|z^{k+1} - z^\star\|^2_M\big\} \\
&\le (1-c)\|z^k - z^\star\|^2_M - c\|\bar{z}^{k+1} - z^k\|^2_D \\
&\quad + c\Big(2B\|\bar{\beta}\|^2_{M_g^{-1}} + 2BR_{\mathrm{s}} + 2R_s + \beta\Big) \sum_{\tau=[k-B]_+}^{k} \|\bar{x}^{\tau+1} - x^\tau\|^2 \\
&\quad + 4cC_s(1+B) \sum_{\tau=[k-B]_+}^{k} \|\bar{u}^{\tau+1} - u^\tau\|^2.
\end{aligned}$$

The result follows by taking total expectation from both sides, dividing by $(1-c)^{k+1}$ and summing over k from 0 to p, see [12, Lem. 1]. □

Note that owing to the diagonal metric used in the proofs of Theorems 7.9 and 7.10, the independent activation pattern in Algorithm 7.3 can be replaced with the more general random sweeping strategy as in [22, 95].

7.6 Numerical simulations

In this section we revisit the formation control example defined in (7.5). For the dynamics of each agent/robot we used the model of [148] with exact discretization of steplength $\Delta T = 1$. The state and input cost matrices and the constraints sets $\mathcal{W}_i$ are as in Section 4.6.

Let $\hat{C}$ be a linear mapping such that $\hat{C}w_i = C\xi_i$ and $L_{i\bullet}$ be such that $L_{i\bullet}w = (E_i w_i, w_i)$. Minimization (7.5) can be formulated as an instance of (7.8) by setting $f(w) = \sum_{i=1}^m \frac{\lambda_i}{2} \sum_{j\in\mathcal{A}_i} \|\hat{C}(w_i - w_j) - d_{ij}\|^2$, $g_i(w_i) = \frac{1}{2} w_i^\top Q_i w_i$, $h_i(y_i, v_i) = \delta_{b_i}(y_i) + \delta_{\mathcal{W}_i}(v_i)$. Therefore, implementation of the algorithms pre-

sented in this paper would only require simple operations such as matrix-vector products, projections onto points and projections onto sets $\mathcal{W}_i$ (which are simple boxes).

In our simulations, horizon length 3 was used. The delays between agents are randomly generated integers in the interval $[0, B]$. We consider two numerical simulations. In the first one we set $m = 5$, $B = 1$ with initial polygon configuration and enforce an arrow formation by appropriate selection of d_{ij}. The interested reader may refer to Section 4.6 for additional details of the formation setup. As discussed in the introduction, minimization (7.5) is an example of partial coupling and Algorithm 7.1 is the suitable choice. In the first numerical experiment, depicted in Figure 7.2 (left), we use the theoretical stepsize bound in (7.19). For comparison, we also considered the dual decomposition approach of [136] that is based on the subgradient method (although this algorithm is not studied with communication delays). For comparison, Algorithms 7.2 and 7.3 are also plotted even though they are not designed for this type of problem. Algorithm 7.3 is used with probabilities of activation p_i set to 0.2 and 0.8. It is observed that the convergence rate of the proposed algorithms are linear.

In the second numerical experiment, depicted in Figure 7.2 (right), we considered a larger problem with $m = 50$ and the maximum delay $B = 10$. We simulated the algorithms with nominal stepsizes. It is observed that Algorithm 7.2 and the dual decomposition approach struggle to reach a high precision. Interestingly, the randomized algorithm Algorithm 7.3 is able to overcome this. Moreover, even with larger delays the algorithms are convergent with nominal stepsizes while the theoretical stepsize may become too small resulting in slow convergence in practice. Therefore, it would be interesting to study if the stepsize conditions presented in this paper can be relaxed for the special case when h_i are indicator functions and g_i are quadratic.

7.7 Conclusions

In this chapter we considered the application of primal-dual algorithms for solving structured optimization problems over message-passing architectures. The coupling between agents was classified as total and partial coupling. For each case a separate algorithm was studied and it was shown that the communication delay is tolerated provided that the stepsizes are small enough, and that some strong convexity assumption holds. In addition, in the case of total coupling a variant of the proposed algorithm was studied that allows random and independent activation of the agents. Future work consists of extending the

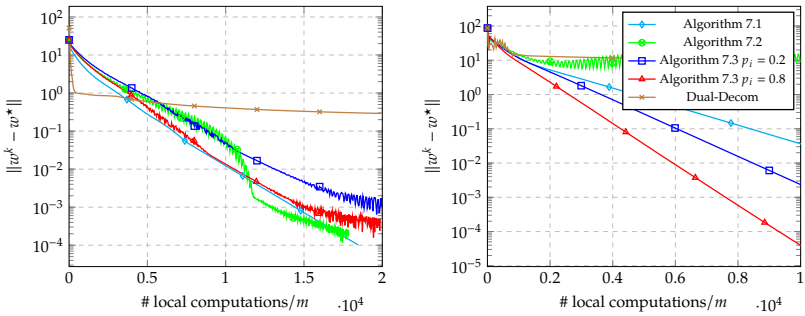

Figure 7.2: *Comparison for the convergence of the algorithms for $m = 5$, $B = 1$ (left) and $m = 50$, $B = 10$ (right).*

convergence analysis to the partially asynchronous framework and exploring different Lyapunov functions that allow for nonconvex cost functions.

Chapter 8

Block-coordinate and incremental aggregated proximal gradient methods

This chapter is based on:

P. Latafat, A. Themelis and P. Patrinos. *Block-coordinate and incremental aggregated proximal gradient methods for nonsmooth nonconvex problems.* arXiv:1906.10053 (submitted 2019).

8.1 Introduction

This chapter addresses block-coordinate (BC) proximal gradient methods for problems of the form

$$\underset{\boldsymbol{x}=(x_1,\dots,x_N)\in\mathbb{R}^{\sum_i n_i}}{\text{minimize}} \Phi(\boldsymbol{x}) \coloneqq F(\boldsymbol{x}) + G(\boldsymbol{x}), \quad \text{where} \quad F(\boldsymbol{x}) \coloneqq \tfrac{1}{N}\textstyle\sum_{i=1}^{N} f_i(x_i), \tag{8.1}$$

in the following setting.

Assumption 8.I (problem setting)**.** In problem (8.1) the following hold:

- A1 function f_i is L_{f_i}-smooth (Lipschitz differentiable with modulus L_{f_i}), $i \in [N]$;
- A2 function G is proper and lower semicontinuous (lsc);
- A3 a solution exists: $\arg\min \Phi \neq \emptyset$.

Unlike typical cases analyzed in the literature where G is separable [163, 166, 125, 15, 25, 139, 106, 44, 85, 179], we here consider the complementary case

where it is only the smooth term F that is assumed to be separable. The main challenge in analyzing convergence of BC schemes for (8.1) especially in the nonconvex setting is the fact that even in expectation the cost does not necessarily decrease along the trajectories. Instead, we demonstrate that the forward-backward envelope (FBE) [132, 158] is a suitable Lyapunov function for such problems.

Several BC-type algorithms that allow for a nonseparable nonsmooth term have been considered in the literature, however, all in convex settings. In [165, 167] a class of convex composite problems is studied that involves a linear constraint as the nonsmooth nonseparable term. A BC algorithm with a Gauss-Southwell-type rule is proposed and the convergence is established using the cost as Lyapunov function by exploiting linearity of the constraint to ensure feasibility. A refined analysis in [119, 120] extends this to a random coordinate selection strategy. Another approach in the convex case is to consider randomized BC updates applied to general averaged operators. Although this approach can allow for fully nonseparable problems, usually separable nonsmooth functions are considered in the literature. The convergence analysis of such methods relies on establishing quasi-Fejér monotonicity [88, 51, 134, 22, 133, 95]. In a primal-dual setting in [70] a combination of Bregman and Euclidean distance is employed as Lyapunov function. In [82] a BC algorithm is proposed for strongly convex algorithms that involves coordinate updates for the gradient followed by a full proximal step, and the distance from the (unique) solution is used as Lyapunov function. The analysis and the Lyapunov functions in all of the above mentioned works rely heavily on convexity and are not suitable for nonconvex settings.

Thanks to the nonconvexity and nonseparability of G, many machine learning problems can be formulated as in (8.1), a primary example being constrained and/or regularized finite sum problems [17, 149, 60, 59, 114, 138, 137, 147]

$$\text{minimize}_{\ x\in\mathbb{R}^n}\ \varphi(x) \coloneqq \tfrac{1}{N}\textstyle\sum_{i=1}^{N} f_i(x) + g(x), \tag{8.2}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ are smooth functions and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is possibly nonsmooth, and everything here can be nonconvex. In fact, one way to cast (8.2) into the form of problem (8.1) is by setting

$$G(\boldsymbol{x}) \coloneqq \tfrac{1}{N}\textstyle\sum_{i=1}^{N} g(x_i) + \delta_C(\boldsymbol{x}), \tag{8.3}$$

where $C \coloneqq \{\boldsymbol{x} \in \mathbb{R}^{nN} \mid x_1 = x_2 = \cdots = x_N\}$ is the consensus set, and δ_C is the indicator function of set C. Since the nonsmooth term g is allowed to be nonconvex, formulation (8.2) can account for nonconvex constraints such as rank constraints or zero norm balls, and nonconvex regularizers such as ℓ^p with $p \in [0, 1)$, [87].

Another prominent example in distributed applications is the *"sharing"* problem [29]:

$$\underset{\boldsymbol{x}\in\mathbb{R}^{nN}}{\text{minimize}}\ \Phi(\boldsymbol{x}) \coloneqq \tfrac{1}{N}\sum_{i=1}^{N} f_i(x_i) + g\Big(\sum_{i=1}^{N} x_i\Big). \tag{8.4}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ are smooth functions and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is nonsmooth, and all are possibly nonconvex. The sharing problem is cast as in (8.1) by setting $G \coloneqq g \circ A$, where $A \coloneqq [\mathrm{I}_n\ \ldots\ \mathrm{I}_n] \in \mathbb{R}^{n\times nN}$.

8.1.1 The main block-coordinate algorithm

While gradient evaluations are the building blocks of smooth minimization, a fundamental tool to deal with a nonsmooth lsc term $\psi : \mathbb{R}^r \to \overline{\mathbb{R}}$ is its *V-proximal mapping*

$$\operatorname{prox}_{\psi}^{V}(x) \coloneqq \operatorname*{arg\,min}_{w\in\mathbb{R}^r} \big\{\psi(w) + \tfrac{1}{2}\|w - x\|_V^2\big\}, \tag{8.5}$$

where V is a symmetric and positive definite matrix and $\|\cdot\|_V$ indicates the norm induced by the scalar product $(x, y) \mapsto \langle x, Vy\rangle$. It is common to take $V = t^{-1}\mathrm{I}_r$ as a multiple of the $r \times r$ identity matrix I_r, in which case the notation $\operatorname{prox}_{t\psi}$ is typically used and t is referred to as a stepsize. While this operator enjoys nice regularity properties when g is convex, such as (single valuedness and) Lipschitz continuity, for nonconvex g it may fail to be a well-defined function and rather has to be intended as a point-to-set mapping $\operatorname{prox}_{\psi}^{V} : \mathbb{R}^r \rightrightarrows \mathbb{R}^r$. Nevertheless, the value function associated to the minimization problem in the definition (8.5), namely the *Moreau envelope*

$$\psi^V(x) \coloneqq \min_{w\in\mathbb{R}^r} \big\{\psi(w) + \tfrac{1}{2}\|w - x\|_V^2\big\}, \tag{8.6}$$

is a well-defined real-valued function, in fact locally Lipschitz continuous, that lower bounds ψ and shares with ψ infima and minimizers. The proximal mapping is available in closed form for many useful functions, many of which are widely used regularizers in machine learning; for instance, the proximal mapping of the ℓ^0 and ℓ^1 regularizers amount to hard and soft thresholding operators.

In many applications the cost to be minimized is structured as the sum of a smooth term h and a proximable (*i.e.*, with easily computable proximal mapping) term ψ. In these cases, the *proximal gradient method* [75, 11] constitutes a cornerstone iterative method that interleaves gradient descent steps on the smooth function and proximal operations on the nonsmooth function, resulting in iterations of the form $x^+ \in \operatorname{prox}_{\gamma\psi}(x - \gamma\nabla h(x))$ for some suitable stepsize γ.

Our proposed scheme to address problem (8.1) is a BC variant of the proximal gradient method, in the sense that only some coordinates are updated according to the proximal gradient rule, while the others are left unchanged. This concept is synopsized in Algorithm 8.1, which constitutes the general algorithm addressed in this chapter.

Algorithm 8.1 General forward-backward block-coordinate scheme

REQUIRE $\boldsymbol{x}^0 \in \mathbb{R}^{\sum_i n_i}$, $\gamma_i \in (0, {}^{N}\!/_{L_{f_i}})$, $i \in [N]$

$\Gamma = \operatorname{blkdiag}(\gamma_1 \mathrm{I}_{n_1}, \dots, \gamma_N \mathrm{I}_{n_N})$, $k = 0$

REPEAT until convergence

1: $\boldsymbol{z}^k \in \operatorname{prox}_G^{\Gamma^{-1}}\big(\boldsymbol{x}^k - \Gamma \nabla F(\boldsymbol{x}^k)\big)$

2: select a set of indices $I^{k+1} \subseteq [N]$

3: update $x_i^{k+1} = z_i^k$ for $i \in I^{k+1}$ and $x_i^{k+1} = x_i^k$ for $i \notin I^{k+1}$, $k \leftarrow k+1$

RETURN $\boldsymbol{z}^k$

Although seemingly wasteful, in many cases one can efficiently compute individual blocks without the need of full operations. In fact BC Algorithm 8.1 bridges the gap between a BC framework and a class of incremental methods where a global computation typically involving the full gradient is carried out incrementally via performing computations only for a subset of coordinates. Two such broad applications, problems (8.2) and (8.4), are discussed in the dedicated Sections 8.3 and 8.4, where among other things we will show that Algorithm 8.1 leads to the well known Finito/MISO algorithm [60, 114].

8.1.2 Contributions

1) To the best of our knowledge this is the first analysis of BC schemes with a nonseparable nonsmooth term and in the fully nonconvex setting. While the original cost Φ cannot serve as a Lyapunov function, we show that the forward-backward envelope (FBE) [132, 158] decreases surely, not only in expectation (Lemma 8.5).

2) This allows for a quite general convergence analysis for different sampling criteria. This chapter in particular covers randomized strategies (Section 8.2.3) where at each iteration one or more coordinates are sampled with possibly time-varying probabilities, as well as essentially cyclic (and in particular cyclic and shuffled) strategies in case the nonsmooth term is convex (Section 8.2.4).

3) We exploit the Kurdyka-Łojasiewicz (KL) property to show global (as opposed to subsequential) and linear convergence when the sampling is essentially cyclic

and the nonsmooth function is convex, without imposing convexity requirements on the smooth functions (Theorem 8.11).

4) As immediate byproducts of our analysis we obtain **(a)** an incremental algorithm for the sharing problem [29] that to the best of our knowledge is novel (Section 8.4), and **(b)** the Finito/MISO algorithm [60, 114] leading to a much simpler and more general analysis than available in the literature with new convergence results both for randomized sampling strategies in the fully nonconvex setting and for essentially cyclic samplings when the nonsmooth term is convex (Section 8.3).

8.1.3 Organization

This chapter is organized as follows. The core of the chapter lies in the convergence analysis of Algorithm 8.1 detailed in Section 8.2: Section 8.2.1 introduces the FBE, fundamental tool of our methodology and lists some of its properties followed by other ancillary results documented in Section 8.7.1. The algorithmic analysis begins in Section 8.2.2 with a collection of facts that hold independently of the chosen sampling strategy, and later specializes to randomized and essentially cyclic samplings in the dedicated Sections 8.2.3 and 8.2.4. Sections 8.3 and 8.4 discuss two particular instances of the investigated algorithmic framework, namely (a generalization of) the Finito/MISO algorithm for finite sum minimization and an incremental scheme for the sharing problem, both for fully nonconvex and nonsmooth formulations. Convergence results are immediately inferred from those of the more general BC Algorithm 8.1. Section 8.6 concludes this chapter.

8.2 Convergence analysis

We begin by observing that Assumption 8.I is enough to guarantee the well definedness of the forward-backward operator in Algorithm 8.1, which for notational convenience will be henceforth denoted as $\mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x})$. Namely, $\mathrm{T}_\Gamma^{\mathrm{FB}} : \mathbb{R}^{\sum_i n_i} \rightrightarrows \mathbb{R}^{\sum_i n_i}$ is the point-to-set mapping

$$\begin{aligned}\mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) &\coloneqq \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{x} - \Gamma\nabla F(\boldsymbol{x})) \\ &= \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^{\sum_i n_i}} \left\{F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{w}-\boldsymbol{x}\rangle + G(\boldsymbol{w}) + \tfrac{1}{2}\|\boldsymbol{w}-\boldsymbol{x}\|_{\Gamma^{-1}}^2\right\}.\end{aligned} \tag{8.7}$$

Lemma 8.1. *Suppose that Assumption 8.I holds, and for* $\gamma_i \in (0, {}^{N}/_{L_{f_i}})$, $i \in [N]$ *let* $\Gamma \coloneqq \operatorname{blkdiag}(\gamma_1 \mathrm{I}_{n_1}, \dots, \gamma_N \mathrm{I}_{n_N})$. *Then,* $\operatorname{prox}_G^{\Gamma^{-1}}$ *and* $\mathrm{T}_\Gamma^{\mathrm{FB}}$ *are locally bounded, outer semicontinuous (osc), nonempty- and compact-valued mappings.*

Proof. For $\boldsymbol{x}^\star \in \operatorname{argmin} \Phi$ it follows from (8.41) that

$$\min \Phi \leq F(\boldsymbol{x}) + G(\boldsymbol{x}) \leq G(\boldsymbol{x}) + F(\boldsymbol{x}^\star) + \langle \nabla F(\boldsymbol{x}^\star), \boldsymbol{x} - \boldsymbol{x}^\star \rangle + \tfrac{1}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}\|^2_{\Lambda_F}.$$

Therefore, G is lower bounded by a quadratic function with quadratic term $-\frac{1}{2}\|\cdot\|^2_{\Lambda_F}$, and thus is prox-bounded in the sense of [144, Def. 1.23]. The claim then follows from [144, Thm. 1.25 and Ex. 5.23(b)] and the continuity of the forward mapping $\mathrm{id} - \Gamma \nabla F$. □

8.2.1 The forward-backward envelope

The fundamental challenge in the analysis of (8.1) is the fact that, without separability of G, descent on the cost function cannot be established even in expectation. Instead, we show that the *forward-backward envelope* (FBE) [132, 158] can be used as Lyapunov function. This subsection formally introduces the FBE, here generalized to account for a matrix-valued stepsize parameter Γ, and lists some of its basic properties needed for the convergence analysis of Algorithm 8.1. Although easy adaptations of the similar results in [132, 158, 156], for the sake of self-inclusiveness the proofs are included here.

Definition 8.2 (forward-backward envelope)**.** *In problem* (8.1)*, let* f_i *be differentiable functions,* $i \in [N]$*, and let* $\Gamma = \operatorname{blkdiag}(\gamma_1 \mathrm{I}_{n_1}, \dots, \gamma_N \mathrm{I}_{n_N})$ *for* $\gamma_1, \dots, \gamma_N > 0$*. The forward-backward envelope (FBE) associated to* (8.1) *with stepsize* Γ *is the function* $\Phi_\Gamma^{\mathrm{FB}} : \mathbb{R}^{\sum_i n_i} \to [-\infty, \infty)$ *defined as*

$$\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) \coloneqq \inf_{\boldsymbol{w} \in \mathbb{R}^{\sum_i n_i}} \left\{ F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{w} - \boldsymbol{x} \rangle + G(\boldsymbol{w}) + \tfrac{1}{2}\|\boldsymbol{w} - \boldsymbol{x}\|^2_{\Gamma^{-1}} \right\}. \tag{8.8a}$$

Definition 8.2 highlights an important symmetry between the Moreau envelope and the FBE: similarly to the relation between the Moreau envelope (8.6) and the proximal mapping (8.5), the FBE (8.8a) is the value function associated with the proximal gradient mapping (8.7). By replacing any minimizer $\boldsymbol{z} \in \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x})$ in the right-hand side of (8.8a) one obtains yet another interesting interpretation of the FBE in terms of the Γ^{-1}-augmented Lagrangian associated to (8.1)

$$\mathscr{L}_{\Gamma^{-1}}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}) \coloneqq F(\boldsymbol{x}) + G(\boldsymbol{z}) + \langle \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{z} \rangle + \tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2_{\Gamma^{-1}},$$

namely,

$$\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) = F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{x}\rangle + G(\boldsymbol{z}) + \tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2_{\Gamma^{-1}} \tag{8.8b}$$

$$= \mathscr{L}_{\Gamma^{-1}}(\boldsymbol{x}, \boldsymbol{z}, -\nabla F(\boldsymbol{x})). \tag{8.8c}$$

Lastly, by rearranging the terms it can easily be seen that

$$\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) = F(\boldsymbol{x}) - \tfrac{1}{2}\|\nabla F(\boldsymbol{x})\|^2_\Gamma + G^{\Gamma^{-1}}(\boldsymbol{x} - \Gamma\nabla F(\boldsymbol{x})), \tag{8.8d}$$

hence in particular that the FBE inherits regularity properties of $G^{\Gamma^{-1}}$ and ∇F, some of which are summarized in the next result.

Lemma 8.3 (FBE: fundamental inequalities)**.** *Suppose that Assumption 8.I is satisfied and let $\gamma_i \in (0, {}^{N}\!/_{L_{f_i}})$, $i \in [N]$. Then, the FBE $\Phi_\Gamma^{\mathrm{FB}}$ is a (real-valued and) locally Lipschitz continuous function. Moreover, the following hold for any $\boldsymbol{x} \in \mathbb{R}^{\sum_i n_i}$:*

(i) $\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) \le \Phi(\boldsymbol{x})$.

(ii) $\frac{1}{2}\|\boldsymbol{z}-\boldsymbol{x}\|^2_{\Gamma^{-1}-\Lambda_F} \le \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) - \Phi(\boldsymbol{z}) \le \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{x}\|^2_{\Gamma^{-1}+\Lambda_F}$ *for any* $\boldsymbol{z} \in \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x})$, *where* $\Lambda_F \coloneqq \frac{1}{N}\operatorname{blkdiag}(L_{f_1}\mathrm{I}_{n_1}, \dots, L_{f_n}\mathrm{I}_{n_N})$.

(iii) If in addition each f_i is μ_{f_i}-strongly convex and G is convex, then for every $\boldsymbol{x} \in \mathbb{R}^{\sum_i n_i}$

$$\tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}^\star\|^2_{\mu_F} \le \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) - \min\Phi$$

where $\boldsymbol{x}^\star \coloneqq \arg\min\Phi$, $\mu_F \coloneqq \frac{1}{N}\operatorname{blkdiag}(\mu_{f_1}\mathrm{I}_{n_1}, \dots, \mu_{f_N}\mathrm{I}_{n_N})$, *and* $\boldsymbol{z} = \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x})$.

Proof. Local Lipschitz continuity of the FBE follows from (8.8d) in light of Lemma 8.1 and [144, Ex. 10.32].

♠ 8.3*(i)* Follows by replacing $\boldsymbol{w} = \boldsymbol{x}$ in (8.8a).

♠ 8.3*(ii)* Directly follows from (8.42) and the identity $\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) = \mathcal{M}_\Gamma(\boldsymbol{z}, \boldsymbol{x})$ for $\boldsymbol{z} \in \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x})$.

♠ 8.3*(iii)* By strong convexity, denoting $\Phi_\star \coloneqq \min\Phi$, we have

$$\Phi_\star \le \Phi(\boldsymbol{z}) - \tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}^\star\|^2_{\mu_F} \le \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) - \tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}^\star\|^2_{\mu_F}$$

where the second inequality follows from Lemma 8.3*(ii)*. □

Another key property that the FBE shares with the Moreau envelope is that minimizing the extended-real-valued function Φ is equivalent to minimizing the

continuous function Φ_Γ^{FB}. Moreover, the former is level bounded iff so is the latter. This fact will be particularly useful for the analysis of Algorithm 8.1, as it will be shown in Lemma 8.5 that the FBE (surely) decreases along its iterates. As a consequence, despite the fact that the same does not hold for Φ (in fact, iterates may even be infeasible), coercivity of Φ is enough to guarantee boundedness of $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$.

Lemma 8.4 (FBE: minimization equivalence)**.** *Suppose that Assumption 8.I is satisfied and that $\gamma_i \in (0, {}^{N}\!/_{L_i})$, $i \in [N]$. Then, the following hold:*

(i) $\min \Phi_\Gamma^{\text{FB}} = \min \Phi$*;*

(ii) $\operatorname{arg\,min} \Phi_\Gamma^{\text{FB}} = \operatorname{arg\,min} \Phi$*;*

(iii) Φ_Γ^{FB} *is level bounded iff so is* Φ*.*

Proof.

♠ 8.4*(i)* and 8.4*(ii)* It follows from Lemma 8.3*(i)* that $\inf \Phi_\Gamma^{\text{FB}} \leq \min \Phi$. Conversely, let $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ be such that $\Phi_\Gamma^{\text{FB}}(\boldsymbol{x}^k) \to \inf \Phi_\Gamma^{\text{FB}}$ as $k \to \infty$, and for each k let $\boldsymbol{z}^k \in \mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{x}^k)$. It then follows from Lemmas 8.3*(i)* and 8.3*(ii)* that

$$\inf \Phi_\Gamma^{\text{FB}} \leq \min \Phi \leq \liminf_{k\to\infty} \Phi(\boldsymbol{z}^k) \leq \liminf_{k\to\infty} \Phi_\Gamma^{\text{FB}}(\boldsymbol{x}^k) = \inf \Phi_\Gamma^{\text{FB}},$$

hence $\min \Phi = \inf \Phi_\Gamma^{\text{FB}}$. Suppose now that $\boldsymbol{x} \in \operatorname{arg\,min} \Phi$ (which exists by Assumption 8.I); then it follows from Lemma 8.3*(ii)* that $\mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{x}) = \{\boldsymbol{x}\}$ (for otherwise another element would belong to a lower level set of Φ). Combining with Lemma 8.3*(i)* with $\boldsymbol{z} = \boldsymbol{x}$ we then have

$$\min \Phi = \Phi(\boldsymbol{z}) \leq \Phi_\Gamma^{\text{FB}}(\boldsymbol{x}) \leq \Phi(\boldsymbol{x}) = \min \Phi.$$

Since $\min \Phi = \inf \Phi_\Gamma^{\text{FB}}$, we conclude that $\boldsymbol{x} \in \operatorname{arg\,min} \Phi_\Gamma^{\text{FB}}$, and that in particular $\inf \Phi_\Gamma^{\text{FB}} = \min \Phi_\Gamma^{\text{FB}}$. Conversely, suppose $\boldsymbol{x} \in \operatorname{arg\,min} \Phi_\Gamma^{\text{FB}}$ and let $\boldsymbol{z} \in \mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{x})$. By combining Lemmas 8.3*(i)* and 8.3*(ii)* we have that $\boldsymbol{z} = \boldsymbol{x}$, that is, that $\mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{x}) = \{\boldsymbol{x}\}$. It then follows from Lemma 8.3*(ii)* and assertion 8.4*(i)* that

$$\Phi(\boldsymbol{x}) = \Phi(\boldsymbol{z}) \leq \Phi_\Gamma^{\text{FB}}(\boldsymbol{x}) = \min \Phi_\Gamma^{\text{FB}} = \min \Phi,$$

hence $\boldsymbol{x} \in \operatorname{arg\,min} \Phi$.

♠ 8.4*(iii)* Due to Lemma 8.3*(i)*, if Φ_Γ^{FB} is level bounded clearly so is Φ. Conversely, suppose that Φ_Γ^{FB} is not level bounded. Then, there exist $\alpha \in \mathbb{R}$ and $(\boldsymbol{x}^k)_{k\in\mathbb{N}} \subseteq \operatorname{lev}_{\leq\alpha} \Phi_\Gamma^{\text{FB}}$ such that $\|\boldsymbol{x}^k\| \to \infty$ as $k \to \infty$. Let $\lambda = \min_i \{\gamma_i^{-1} - L_{f_i} N^{-1}\} > 0$, and for each $k \in \mathbb{N}$ let $\boldsymbol{z}^k \in \mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{x}^k)$. It

then follows from Lemma 8.3*(ii)* that

$$\min \Phi \leq \Phi(\boldsymbol{z}^k) \leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \tfrac{\lambda}{2}\|\boldsymbol{x}^k - \boldsymbol{z}^k\|^2 \leq \alpha - \tfrac{\lambda}{2}\|\boldsymbol{x}^k - \boldsymbol{z}^k\|^2,$$

hence $(\boldsymbol{z}^k)_{k\in\mathbb{N}} \subseteq \operatorname{lev}_{\leq\alpha} \Phi$ and $\|\boldsymbol{x}^k - \boldsymbol{z}^k\|^2 \leq \frac{2}{\lambda}(\alpha - \min \Phi)$. Consequently, also the sequence $(\boldsymbol{z}^k)_{k\in\mathbb{N}} \subseteq \operatorname{lev}_{\leq\alpha} \Phi$ is unbounded, proving that Φ is not level bounded. □

We remark that the kinship of $\Phi_\Gamma^{\mathrm{FB}}$ and Φ extends also to local minimality; the interested reader is referred to [155, Thm. 3.6] for details.

8.2.2 A sure descent lemma

We now proceed to the theoretical analysis of Algorithm 8.1. Clearly, some assumptions on the index selection criterion are needed in order to establish reasonable convergence results, for little can be guaranteed if, for instance, one of the indices is never selected. Nevertheless, for the sake of a general analysis it is instrumental to first investigate which properties hold independently of such criteria. After listing some of these facts in Lemma 8.5, in Sections 8.2.3 and 8.2.4 we will specialize the results to randomized and (essentially) cyclic sampling strategies.

Lemma 8.5 (sure descent)**.** *Suppose that Assumption 8.1 is satisfied. Then, the following hold for the iterates generated by Algorithm 8.1:*

(i) $\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{k+1}) \leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \sum_{i\in I^{k+1}} \frac{\xi_i}{2\gamma_i}\|z_i^k - x_i^k\|^2$, *where* $\xi_i \coloneqq \frac{N-\gamma_i L_{f_i}}{N}$, $i \in [N]$, *are strictly positive;*

(ii) $(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ *monotonically decreases to a finite value* $\Phi_\star \geq \min \Phi$*;*

(iii) $\Phi_\Gamma^{\mathrm{FB}}$ *is constant (and equals* $\Phi_\star$ *as above) on the set of accumulation points of* $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$*;*

(iv) the sequence $(\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2)_{k\in\mathbb{N}}$ *has finite sum (and in particular vanishes);*

(v) if Φ *is coercive, then* $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ *and* $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ *are bounded.*

Proof.

♠ 8.5*(i)* To ease notation, let $\Lambda_F \coloneqq \frac{1}{N}\operatorname{blkdiag}(L_{f_1}\mathrm{I}_{n_1}, \ldots, L_{f_n}\mathrm{I}_{n_N})$ and for $\boldsymbol{w} \in \mathbb{R}^{\sum_i n_i}$ let $w_I \in \mathbb{R}^{\sum_i n_i}$ denote the slice $(w_i)_{i\in I}$, and let $\Lambda_{F_I}, \Gamma_I \in$

$\mathbb{R}^{\sum_{i\in I} n_i \times \sum_{i\in I} n_i}$ be defined accordingly. Start by observing that, since $\boldsymbol{z}^{k+1} \in \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{x}^{k+1} - \Gamma\nabla F(\boldsymbol{x}^{k+1}))$, from the proximal inequality on G it follows that

$$\begin{aligned} G(\boldsymbol{z}^{k+1}) - G(\boldsymbol{z}^k) &\leq \tfrac{1}{2}\|\boldsymbol{z}^k - \boldsymbol{x}^{k+1} + \Gamma\nabla F(\boldsymbol{x}^{k+1})\|^2_{\Gamma^{-1}} - \tfrac{1}{2}\|\boldsymbol{z}^{k+1} - \boldsymbol{x}^{k+1} + \Gamma\nabla F(\boldsymbol{x}^{k+1})\|^2_{\Gamma^{-1}} \\ &= \tfrac{1}{2}\|\boldsymbol{z}^k - \boldsymbol{x}^{k+1}\|^2_{\Gamma^{-1}} - \tfrac{1}{2}\|\boldsymbol{z}^{k+1} - \boldsymbol{x}^{k+1}\|^2_{\Gamma^{-1}} + \langle \nabla F(\boldsymbol{x}^{k+1}), \boldsymbol{z}^k - \boldsymbol{z}^{k+1}\rangle. \end{aligned} \tag{8.9}$$

We have

$$\begin{aligned} \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{k+1}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) &= F(\boldsymbol{x}^{k+1}) + \langle \nabla F(\boldsymbol{x}^{k+1}), \boldsymbol{z}^{k+1} - \boldsymbol{x}^{k+1}\rangle + G(\boldsymbol{z}^{k+1}) \\ &\quad + \tfrac{1}{2}\|\boldsymbol{z}^{k+1} - \boldsymbol{x}^{k+1}\|^2_{\Gamma^{-1}} \\ &\quad - \big(F(\boldsymbol{x}^k) + \langle \nabla F(\boldsymbol{x}^k), \boldsymbol{z}^k - \boldsymbol{x}^k\rangle + G(\boldsymbol{z}^k) + \tfrac{1}{2}\|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-1}}\big) \\ &\text{apply the upper bound in (8.41) with } \boldsymbol{w} = \boldsymbol{x}^{k+1} \text{ and the proximal inequality (8.9)} \\ &\leq \langle \nabla F(\boldsymbol{x}^k), \boldsymbol{x}^{k+1} - \boldsymbol{z}^k\rangle + \tfrac{1}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2_{\Lambda_F} + \langle \nabla F(\boldsymbol{x}^{k+1}), \boldsymbol{z}^k - \boldsymbol{x}^{k+1}\rangle \\ &\quad - \tfrac{1}{2}\|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-1}} + \tfrac{1}{2}\|\boldsymbol{z}^k - \boldsymbol{x}^{k+1}\|^2_{\Gamma^{-1}}. \end{aligned}$$

To conclude, notice that the ℓ-th block of $\nabla F(\boldsymbol{x}^k) - \nabla F(\boldsymbol{x}^{k+1})$ is zero for $\ell \notin I$, and that the ℓ-th block of $\boldsymbol{x}^{k+1} - \boldsymbol{z}^k$ is zero if $\ell \in I$. Hence, the scalar product vanishes. For similar reasons, one has $\|\boldsymbol{z}^k - \boldsymbol{x}^{k+1}\|^2_{\Gamma^{-1}} - \|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-1}} = -\|z_I^k - x_I^k\|^2_{\Gamma_I^{-1}}$ and $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2_{\Lambda_F} = \|z_I^k - x_I^k\|^2_{\Lambda_{F_I}}$, yielding the claimed expression.

♠ 8.5*(ii)* Monotonic decrease of $(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ is a direct consequence of assertion 8.5*(i)*. This ensures that the sequence converges to some value $\Phi_\star$, bounded below by $\min\Phi$ in light of Lemma 8.4*(i)*.

♠ 8.5*(iii)* Directly follows from assertion 8.5*(ii)* together with the continuity of $\Phi_\Gamma^{\mathrm{FB}}$, see Lemma 8.3.

♠ 8.5*(iv)* Denoting $\xi_{\min} \coloneqq \min_{i\in[N]}\{\xi_i\}$ which is a strictly positive constant, it follows from assertion 8.5*(i)* that for each $k \in \mathbb{N}$ it holds that

$$\begin{aligned} \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{k+1}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) &\leq -\sum_{i\in I^{k+1}} \tfrac{\xi_i}{2\gamma_i}\|z_i^k - x_i^k\|^2 \\ &\leq -\tfrac{\xi_{\min}}{2}\sum_{i\in I^{k+1}} \gamma_i^{-1}\|z_i^k - x_i^k\|^2 \\ &= -\tfrac{\xi_{\min}}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2_{\Gamma^{-1}}. \end{aligned} \tag{8.10}$$

By summing for $k \in \mathbb{N}$ and using the positive definiteness of Γ^{-1} together with the fact that $\min\Phi_\Gamma^{\mathrm{FB}} = \min\Phi > \infty$ as ensured by Lemma 8.4*(i)* and 8.IA3, we obtain that $\sum_{k\in\mathbb{N}}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 < \infty$.

♠ 8.5*(v)* It follows from assertion 8.5*(ii)* that the entire sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is contained in the sublevel set $\{\boldsymbol{w} \mid \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{w}) \leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^0)\}$, which is bounded provided that Φ is coercive as shown in Lemma 8.4*(iii)*. In turn, boundedness of $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ then follows from local boundedness of $\mathrm{T}_\Gamma^{\mathrm{FB}}$, *cf.* Lemma 8.1. □

8.2.3 Randomized sampling

In this section we provide convergence results for Algorithm 8.1 where the index selection criterion complies with the following requirement.

Assumption 8.II (randomized sampling requirements)**.** There exist constants $p_1, \dots, p_N > 0$ such that, at any iteration and independently of the past, each $i \in [N]$ is sampled with probability at least p_i.

Our notion of randomization is general enough to allow for time-varying probabilities and mini-batch selections. The role of parameters p_i in Assumption 8.II is to prevent that an index is sampled with arbitrarily small probability. In more rigorous terms, $\mathbb{P}_k[i \in I^{k+1}] \geq p_i$ shall hold for all $i \in [N]$, where $\mathbb{P}_k$ represents the probability conditional to the knowledge at iteration k. Notice that we do not require the p_i's to sum up to one, as multiple index selections are allowed, similar to the setting of [22, 95] in the convex case.

Due to the possible nonconvexity of problem (8.1), unless additional assumptions are made not much can be said about convergence of the iterates to a unique point. Nevertheless, the following result shows that any accumulation point $\boldsymbol{x}^\star$ of sequences $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ generated by Algorithm 8.1 is a stationary point, in the sense that it satisfies the necessary condition for minimality $0 \in \hat{\partial}\Phi(\boldsymbol{x}^\star)$, where $\hat{\partial}$ denotes the (regular) nonconvex subdifferential, see [144, Thm. 10.1].

Theorem 8.6 (randomized sampling: subsequential convergence)**.** *Suppose that Assumptions 8.I and 8.II are satisfied. Then, the following hold almost surely for the iterates generated by Algorithm 8.1:*

(i) the sequence $(\|\boldsymbol{x}^k - \boldsymbol{z}^k\|^2)_{k\in\mathbb{N}}$ *has finite sum (and in particular vanishes);*

(ii) the sequence $(\Phi(\boldsymbol{z}^k))_{k\in\mathbb{N}}$ *converges to* $\Phi_\star$ *as in Lemma 8.5(ii);*

(iii) $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ *and* $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ *have same cluster points, all stationary and on which* Φ *and* $\Phi_\Gamma^{\mathrm{FB}}$ *equal* $\Phi_\star$.

Proof. In what follows, $\mathbb{E}_k$ denotes the expectation conditional to the knowledge at iteration k.

♠ 8.6*(i)* Let $\xi_i := \frac{N-\gamma_i L_{f_i}}{N} > 0$, $i \in [N]$, be as in Lemma 8.5*(i)*. We have

$$\begin{aligned}
\mathbb{E}_k\big[\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{k+1})\big] &\overset{8.5(i)}{\leq} \mathbb{E}_k\left[\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \sum_{i\in I^{k+1}} \tfrac{\xi_i}{2\gamma_i}\|z_i^k - x_i^k\|^2\right] \\
&= \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \sum_{I\in\Omega} \mathbb{P}_k\big[\mathcal{I}^{k+1} = I\big] \sum_{i\in I} \tfrac{\xi_i}{2\gamma_i}\|z_i^k - x_i^k\|^2 \\
&= \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \sum_{i=1}^N \sum_{I\in\Omega, I\ni i} \mathbb{P}_k\big[\mathcal{I}^{k+1} = I\big] \tfrac{\xi_i}{2\gamma_i}\|z_i^k - x_i^k\|^2 \\
&\leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \sum_{i=1}^N \tfrac{p_i\xi_i}{2\gamma_i}\|z_i^k - x_i^k\|^2, \qquad (8.11)
\end{aligned}$$

where $\Omega \subseteq 2^{[N]}$ is the sample space ($2^{[N]}$ denotes the power set of $[N]$). Therefore,

$$\mathbb{E}_k\big[\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{k+1})\big] \leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \tfrac{\sigma}{2}\|\boldsymbol{x}^k - \boldsymbol{z}^k\|^2_{\Gamma^{-1}} \quad \text{where } \sigma := \min_{i=1\dots N} p_i\xi_i > 0. \qquad (8.12)$$

The claim follows from the Robbins-Siegmund supermartingale theorem, see *e.g.*, [142] or [17, Prop. 2].

♠ 8.6*(ii)* Observe that $\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-1}+\Lambda_F} \leq \Phi(\boldsymbol{z}^k) \leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-1}-\Lambda_F}$ holds (surely) for $k \in \mathbb{N}$ in light of Lemma 8.3*(ii)*. The claim then follows by invoking Lemma 8.5*(ii)* and assertion 8.6*(i)*.

♠ 8.6*(iii)* In the rest of the proof, for conciseness the "almost sure" nature of the results will be implied without mention. It follows from assertion 8.6*(i)* that a subsequence $(\boldsymbol{x}^k)_{k\in K}$ converges to some point $\boldsymbol{x}^\star$ iff so does the subsequence $(\boldsymbol{z}^k)_{k\in K}$. Since $\mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) \ni \boldsymbol{z}^k$ and both $\boldsymbol{x}^k$ and $\boldsymbol{z}^k$ converge to $\boldsymbol{x}^\star$ as $K \ni k \to \infty$, the inclusion $0 \in \hat{\partial}\Phi(\boldsymbol{x}^\star)$ follows from Lemma 8.21. Since the full sequences $(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ and $(\Phi(\boldsymbol{z}^k))_{k\in\mathbb{N}}$ converge to the same value $\Phi_\star$ (*cf.* Lem. 8.5*(ii)* and assertion 8.6*(ii)*), due to continuity of $\Phi_\Gamma^{\mathrm{FB}}$ (Lemma 8.3) it holds that $\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^\star) = \Phi_\star$, and in turn the bounds in Lemma 8.3*(ii)* together with assertion 8.6*(i)* ensure that $\Phi(\boldsymbol{x}^\star) = \Phi_\star$ too. □

When G is convex and F is strongly convex (that is, each of the functions f_i is strongly convex), the FBE decreases Q-linearly in expectation along the iterates generated by the randomized BC-Algorithm 8.1.

Theorem 8.7 (randomized sampling: linear convergence under strong convexity)**.** *Additionally to Assumptions 8.I and 8.II, suppose that G is convex and*

that each f_i is μ_{f_i}-strongly convex. Then, for all k the following hold for the iterates generated by Algorithm 8.1:

$$\mathbb{E}_k\big[\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{k+1}) - \min\Phi\big] \leq (1-c)\big(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \min\Phi\big) \tag{8.13a}$$

$$\mathbb{E}\big[\Phi(\boldsymbol{z}^k) - \min\Phi\big] \leq \big(\Phi(\boldsymbol{x}^0) - \min\Phi\big)(1-c)^k \tag{8.13b}$$

$$\tfrac{1}{2}\mathbb{E}\big[\|\boldsymbol{z}^k - \boldsymbol{x}^\star\|_{\mu_F}^2\big] \leq \big(\Phi(\boldsymbol{x}^0) - \min\Phi\big)(1-c)^k \tag{8.13c}$$

where $\boldsymbol{x}^\star \coloneqq \operatorname{arg\,min}\Phi$, $\mu_F \coloneqq \frac{1}{N}\operatorname{blkdiag}\big(\mu_{f_1}\mathrm{I}_{n_1}, \ldots \mu_{f_n}\mathrm{I}_{n_N}\big)$, and denoting $\xi_i = \frac{N-\gamma_i L_{f_i}}{N}$, $i \in [N]$,

$$c = \min_{i\in[N]}\left\{\tfrac{\xi_i p_i}{\gamma_i}\right\} \Big/ \max_{i\in[N]}\left\{\tfrac{N-\gamma_i\mu_{f_i}}{\gamma_i^2\mu_{f_i}}\right\}. \tag{8.14}$$

Moreover, by setting the stepsizes γ_i and minimum sampling probabilities p_i as

$$\gamma_i = \tfrac{N}{\mu_{f_i}}\Big(1 - \sqrt{1 - 1/\kappa_i}\Big) \quad \text{and} \quad p_i = \frac{\big(\sqrt{\kappa_i} + \sqrt{\kappa_i - 1}\big)^2}{\sum_{j=1}^N \big(\sqrt{\kappa_j} + \sqrt{\kappa_j - 1}\big)^2} \tag{8.15}$$

with $\kappa_i \coloneqq \frac{L_{f_i}}{\mu_{f_i}}$, $i \in [N]$, then the constant c in (8.13) can be tightened to

$$c = \frac{1}{\sum_{i=1}^N \big(\sqrt{\kappa_i} + \sqrt{\kappa_i - 1}\big)^2}. \tag{8.16}$$

Proof. Since $\boldsymbol{z}^k$ is a minimizer in (8.8a), the necessary stationarity condition reads $\Gamma^{-1}(\boldsymbol{x}^k - \boldsymbol{z}^k) - \nabla F(\boldsymbol{x}^k) \in \partial G(\boldsymbol{z}^k)$. Convexity of G then implies

$$G(\boldsymbol{x}^\star) \geq G(\boldsymbol{z}^k) + \langle \Gamma^{-1}(\boldsymbol{x}^k - \boldsymbol{z}^k) - \nabla F(\boldsymbol{x}^k), \boldsymbol{x}^\star - \boldsymbol{z}^k\rangle,$$

whereas from strong convexity of F we have

$$F(\boldsymbol{x}^\star) \geq F(\boldsymbol{x}^k) + \langle \nabla F(\boldsymbol{x}^k), \boldsymbol{x}^\star - \boldsymbol{x}^k\rangle + \tfrac{1}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|_{\mu_F}^2.$$

By combining these inequalities into (8.8b), and denoting $\Phi_\star \coloneqq \min\Phi = \min\Phi_\Gamma^{\mathrm{FB}}$ (*cf.* Lem. 8.4*(i)*), we have

$$\begin{aligned}\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \Phi_\star &\leq \tfrac{1}{2}\|\boldsymbol{z}^k - \boldsymbol{x}^k\|_{\Gamma^{-1}}^2 - \tfrac{1}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}^k\|_{\mu_F}^2 + \langle \Gamma^{-1}(\boldsymbol{z}^k - \boldsymbol{x}^k), \boldsymbol{x}^\star - \boldsymbol{z}^k\rangle \\ &= \tfrac{1}{2}\|\boldsymbol{z}^k - \boldsymbol{x}^k\|_{\Gamma^{-1} - \mu_F}^2 + \langle (\Gamma^{-1} - \mu_F)(\boldsymbol{z}^k - \boldsymbol{x}^k), \boldsymbol{x}^\star - \boldsymbol{z}^k\rangle - \tfrac{1}{2}\|\boldsymbol{x}^\star - \boldsymbol{z}^k\|_{\mu_F}^2.\end{aligned}$$

Next, by using the inequality $\langle \boldsymbol{a}, \boldsymbol{b}\rangle \leq \frac{1}{2}\|\boldsymbol{a}\|^2_{\mu_F} + \frac{1}{2}\|\boldsymbol{b}\|^2_{\mu_F^{-1}}$ to cancel out the last term, we obtain

$$\begin{aligned}\Phi^{\mathrm{FB}}_{\Gamma}(\boldsymbol{x}^k) - \Phi_\star &\leq \tfrac{1}{2}\|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-1}-\mu_F} + \tfrac{1}{2}\|(\Gamma^{-1}-\mu_F)(\boldsymbol{x}^k - \boldsymbol{z}^k)\|^2_{\mu_F^{-1}} \\ &= \tfrac{1}{2}\|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-2}\mu_F^{-1}(\mathrm{I}-\Gamma\mu_F)}, \end{aligned} \tag{8.17}$$

where the last identity uses the fact that the matrices are diagonal. Combined with (8.11) the claimed Q-linear convergence (8.13a) with factor c as in (8.14) is obtained. The R-linear rates in terms of the cost function and distance from the solution are obtained by repeated application of (8.13a) after taking (unconditional) expectation from both sides and using Lemma 8.3.

To obtain the tighter estimate (8.16), observe that (8.11) with the choice

$$p_i := \tfrac{1}{\gamma_i\mu_{f_i}}\tfrac{N-\gamma_i\mu_{f_i}}{N-\gamma_i L_{f_i}}\left(\textstyle\sum_j \tfrac{1}{\gamma_j\mu_{f_j}}\tfrac{N-\gamma_j\mu_{f_j}}{N-\gamma_j L_{f_j}}\right)^{-1},$$

which equals the one in (8.15) with γ_i as prescribed, yields

$$\begin{aligned}\mathbb{E}_k\left[\Phi^{\mathrm{FB}}_{\Gamma}(\boldsymbol{x}^{k+1})-\Phi_\star\right] &\leq \Phi^{\mathrm{FB}}_{\Gamma}(\boldsymbol{x}^k)-\Phi_\star - \left(2N\textstyle\sum_j \tfrac{1}{\gamma_j\mu_j}\tfrac{N-\gamma_j\mu_j}{N-\gamma_j L_j}\right)^{-1}\sum_{i=1}^N \tfrac{N-\gamma_i\mu_{f_i}}{\gamma_i^2\mu_{f_i}}\|z_i^k - x_i^k\|^2 \\ &= \Phi^{\mathrm{FB}}_{\Gamma}(\boldsymbol{x}^k)-\Phi_\star - \left(2N\textstyle\sum_j \tfrac{1}{\gamma_j\mu_j}\tfrac{N-\gamma_j\mu_j}{N-\gamma_j L_j}\right)^{-1}\|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-1}\mu_F^{-1}(\Gamma^{-1}-\mu_F)}.\end{aligned}$$

The result now follows by combining this with (8.17) and replacing the values of γ_i as proposed in (8.15). □

Notice that as κ_i's approach 1 the linear rate tends to $1 - 1/N$.

8.2.4 Cyclic, shuffled and essentially cyclic samplings

In this section we analyze the convergence of the BC-Algorithm 8.1 when a cyclic, shuffled cyclic or (more generally) an essentially cyclic sampling [164, 163, 85, 45, 179] is used. As formalized in the following standing assumption, an additional convexity requirement for the nonsmooth term G is needed.

Assumption 8.III (essentially cyclic sampling requirements)**.** In problem (8.1), function G is convex. Moreover, there exists $T \geq 1$ such that in Algorithm 8.1 each index is selected at least once within any interval of T iterations.

Note that having $T < N$ is possible because of our general sampling strategy where sets of indices can be sampled within the same iteration. For instance, $T = 1$ corresponds to $I^{k+1} = [N]$ for all k, in which case Algorithm 8.1 would reduce to a (full) proximal gradient scheme.

Two notable special cases of single index selection rules are the cyclic and shuffled cyclic sampling strategies.

Shuffled cyclic sampling: corresponds to setting

$$I^{k+1} = \left\{\pi_{\lfloor k/N \rfloor}\big(\mathrm{mod}(k, N) + 1\big)\right\} \quad \text{for all} \quad k \in \mathbb{N}, \tag{8.18}$$

where $\pi_0, \pi_1, \dots$ are permutations of the set of indices $[N]$ (chosen randomly or deterministically).

Cyclic sampling: corresponds to the case (8.18) with $\pi_{\lfloor k/N \rfloor} = \mathrm{id}$, *i.e.*,

$$I^{k+1} = \{\mathrm{mod}(k, N) + 1\} \quad \text{for all} \quad k \in \mathbb{N}. \tag{8.19}$$

Consistently with the deterministic nature of the essentially cyclic sampling, all results of the previous section hold surely, as opposed to almost surely.

Theorem 8.8 (essentially cyclic sampling: subsequential convergence)**.** *Suppose that Assumptions 8.I and 8.III are satisfied. Then, all the assertions of Theorem 8.6 hold surely.*

Proof. We first establish an important descent inequality for $\Phi_\Gamma^{\mathrm{FB}}$ after every T iterations, *cf.* (8.26). Convexity of G, entailing $\mathrm{prox}_G^{\Gamma^{-1}}$ being Lipschitz continuous (*cf.* Lem. 8.22*(i)*), allows the employment of techniques similar to those in [15, Lem. 3.3]. Since all indices are updated at least once every T iterations, one has that

$$t_\nu(i) \coloneqq \min\{t \in [T] \mid i \text{ is sampled at iteration } T\nu + t - 1\} \tag{8.20}$$

is well defined for each index $i \in [N]$ and $\nu \in \mathbb{N}$. Since i is sampled at iteration $T\nu + t_\nu(i) - 1$ and $x_i^{T\nu} = x_i^{T\nu+1} = \cdots = x_i^{T\nu+t_\nu(i)-1}$ by definition of $t_\nu(i)$, it holds that

$$\begin{aligned} x_i^{T\nu+t_\nu(i)} &= x_i^{T\nu+t_\nu(i)-1} + U_i^\top\Big(\mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+t_\nu(i)-1}) - \boldsymbol{x}^{T\nu+t_\nu(i)-1}\Big) \\ &= x_i^{T\nu} + U_i^\top\Big(\mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+t_\nu(i)-1}) - \boldsymbol{x}^{T\nu+t_\nu(i)-1}\Big), \end{aligned} \tag{8.21}$$

where $U_i \in \mathbb{R}^{\sum_j n_j \times n_i}$ denotes the i-th block column of the identity matrix so that for a vector $v \in \mathbb{R}^{n_i}$

$$U_i v = (0, \dots, 0, \overset{i\text{-th}}{\boxed{v}}, 0, \dots, 0)^\top. \tag{8.22}$$

For all $t \in [T]$ the following holds

$$\begin{aligned}
\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T(\nu+1)}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu}) &= \sum_{\tau=1}^{T} \left(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+\tau}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+\tau-1})\right) \\
&\leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+t}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+t-1}) \\
&\leq -\tfrac{\xi_{\min}}{2} \|\boldsymbol{x}^{T\nu+t} - \boldsymbol{x}^{T\nu+t-1}\|^2_{\Gamma^{-1}},
\end{aligned} \tag{8.23}$$

where $\xi_i := \frac{N - \gamma_i L_{f_i}}{N}$ as in Lemma 8.5*(i)*, $\xi_{\min} := \min_{i\in[N]} \{\xi_i\}$, and the two inequalities follow from Lemma 8.5*(i)*. Moreover, using triangular inequality for $i \in [N]$ yields

$$\begin{aligned}
\|\boldsymbol{x}^{T\nu+t_\nu(i)-1} - \boldsymbol{x}^{T\nu}\|_{\Gamma^{-1}} &\leq \sum_{\tau=1}^{t_\nu(i)-1} \|\boldsymbol{x}^{T\nu+\tau} - \boldsymbol{x}^{T\nu+\tau-1}\|_{\Gamma^{-1}} \\
&\leq \tfrac{T}{\sqrt{\xi_{\min}/2}} \Big(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T(\nu+1)})\Big)^{1/2},
\end{aligned} \tag{8.24}$$

where the second inequality follows from (8.23) together with the fact that $t_\nu(i) \leq T$. For all $i \in [N]$, from the triangular inequality and the $L_{\mathbf{T}}$-Lipschitz continuity of $\mathrm{T}_\Gamma^{\mathrm{FB}}$ (Lemma 8.22*(iv)*) we have

$$\begin{aligned}
\gamma_i^{-1/2} \|U_i^\top(\boldsymbol{x}^{T\nu} - \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu}))\| &\leq \gamma_i^{-1/2} \|U_i^\top\left(\boldsymbol{x}^{T\nu} - \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+t_\nu(i)-1})\right)\| \\
&\quad + \gamma_i^{-1/2} \|U_i^\top\left(\mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+t_\nu(i)-1}) - \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu})\right)\| \\
&\leq \gamma_i^{-1/2} \|x_i^{T\nu+t_\nu(i)-1} - x_i^{T\nu+t_\nu(i)}\| \\
&\quad + \|\mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+t_\nu(i)-1}) - \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu})\|_{\Gamma^{-1}} \\
&\leq \|\boldsymbol{x}^{T\nu+t_\nu(i)-1} - \boldsymbol{x}^{T\nu+t_\nu(i)}\|_{\Gamma^{-1}} \\
&\quad + L_{\mathbf{T}} \|\boldsymbol{x}^{T\nu+t_\nu(i)-1} - \boldsymbol{x}^{T\nu}\|_{\Gamma^{-1}} \\
&\overset{(8.23),\,(8.24)}{\leq} \tfrac{1+TL_{\mathbf{T}}}{\sqrt{\xi_{\min}/2}} \Big(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T(\nu+1)})\Big)^{1/2}.
\end{aligned} \tag{8.25}$$

By squaring and summing over $i \in [N]$, we obtain

$$\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T(\nu+1)}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu}) \leq -\tfrac{\xi_{\min}}{2N(1+TL_{\boldsymbol{T}})^2}\|\boldsymbol{z}^{T\nu} - \boldsymbol{x}^{T\nu}\|^2_{\Gamma^{-1}}. \tag{8.26}$$

By telescoping the inequality and using the fact that $\min \Phi_\Gamma^{\mathrm{FB}} = \min \Phi$ shown in Lemma 8.4*(i)*, we obtain that $(\|\boldsymbol{z}^{T\nu} - \boldsymbol{x}^{T\nu}\|^2_{\Gamma^{-1}})_{\nu\in\mathbb{N}}$ has finite sum, and in particular vanishes. Clearly, by suitably shifting, for every $t \in [T]$ the same can be said for the sequence $(\|\boldsymbol{z}^{T\nu+t} - \boldsymbol{x}^{T\nu+t}\|^2_{\Gamma^{-1}})_{\nu\in\mathbb{N}}$. The whole sequence $(\|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2)_{k\in\mathbb{N}}$ is thus summable, and we may now infer the claim as done in the proof of Theorem 8.6. □

In the next theorem explicit linear convergence rates are derived under the additional strong convexity assumption for the smooth functions. The cyclic and shuffled cyclic cases are treated separately, as tighter bounds can be obtained by leveraging the fact that within cycles of N iterations every index is updated exactly once.

Theorem 8.9 (essentially cyclic sampling: linear convergence under strong convexity)**.** *Additionally to Assumptions 8.I and 8.III, suppose that each function f_i is μ_{f_i}-strongly convex. Then, denoting $\delta := \min_{i\in[N]}\left\{\frac{\gamma_i\mu_{f_i}}{N}\right\}$ and $\Delta := \max_{i\in[N]}\left\{\frac{\gamma_i L_{f_i}}{N}\right\}$, for all $\nu \in \mathbb{N}$ the following hold for the iterates generated by Algorithm 8.1:*

$$\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T(\nu+1)}) - \min\Phi \leq (1-c)\big(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu}) - \min\Phi\big) \tag{8.27a}$$

$$\Phi(\boldsymbol{z}^{T\nu}) - \min\Phi \leq \big(\Phi(\boldsymbol{x}^0) - \min\Phi\big)(1-c)^\nu \tag{8.27b}$$

$$\tfrac{1}{2}\|\boldsymbol{z}^{T\nu} - \boldsymbol{x}^\star\|^2_{\mu_F} \leq \big(\Phi(\boldsymbol{x}^0) - \min\Phi\big)(1-c)^\nu \tag{8.27c}$$

where $\boldsymbol{x}^\star := \arg\min\Phi$, $\mu_F := \frac{1}{N}\operatorname{blkdiag}\big(\mu_{f_1}\mathrm{I}_{n_1}, \ldots \mu_{f_n}\mathrm{I}_{n_N}\big)$, and

$$c = \frac{\delta(1-\Delta)}{N\big(1+T(1-\delta)\big)^2(1-\delta)}. \tag{8.28}$$

In the case of shuffled cyclic (8.18) or cyclic (8.19) sampling, the inequalities can be tightened by replacing T with N and with

$$c = \frac{\delta(1-\Delta)}{N(2-\delta)^2(1-\delta)}. \tag{8.29}$$

Proof.

♠ *The general essentially cyclic case.* Since $\mathrm{T}_\Gamma^{\mathrm{FB}}$ is $L_{\mathbf{T}}$-Lipschitz continuous with $L_{\mathbf{T}} = 1-\delta$ as shown in Lemma 8.22*(iv)*, inequality (8.26) becomes

$$\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T(\nu+1)}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu}) \leq -\tfrac{1-\Delta}{2N(1+T(1-\delta))^2}\|\boldsymbol{z}^{T\nu} - \boldsymbol{x}^{T\nu}\|_{\Gamma^{-1}}^2.$$

Moreover, it follows from (8.17) that

$$\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu}) - \Phi_\star \leq \tfrac{1}{2}(\delta^{-1}-1)\|\boldsymbol{z}^{T\nu} - \boldsymbol{x}^{T\nu}\|_{\Gamma^{-1}}^2. \tag{8.30}$$

By combining the two inequalities the claimed Q-linear convergence (8.27a) with factor c as in (8.28) is obtained. In turn, the R-linear rates (8.27b) and (8.27c) follow from Lemma 8.3.

♠ *The shuffled cyclic case.* Let us now suppose that the sampling strategy follows a shuffled rule as in (8.18) with permutations $\pi_0, \pi_1, \dots$ (hence in the cyclic case $\pi_\nu = \mathrm{id}$ for all $\nu \in \mathbb{N}$). Let U_i be as in (8.22) and $\xi_{\min}$ as in the proof of Theorem 8.8. Observe that $t_\nu(i) = \pi_\nu^{-1}(i) \leq N$ for $t_\nu(i)$ as defined in (8.20). For all $t \in [N]$

$$\begin{aligned}
\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{N(\nu+1)}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{N\nu}) &\leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{N\nu+t-1}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{N\nu}) \\
&\leq -\tfrac{\xi_{\min}}{2}\sum_{\tau=1}^{t-1}\|\boldsymbol{x}^{N\nu+\tau} - \boldsymbol{x}^{N\nu+\tau-1}\|_{\Gamma^{-1}}^2 \\
&= -\tfrac{\xi_{\min}}{2}\|\boldsymbol{x}^{N\nu+t-1} - \boldsymbol{x}^{N\nu}\|_{\Gamma^{-1}}^2,
\end{aligned} \tag{8.31}$$

where the equality follows from the fact that at every iteration a different coordinate is updated (and that Γ is diagonal), and the inequalities from Lemma 8.5*(i)*. Similarly, (8.23) holds with T replaced by N (despite the fact that T is not necessarily N, but is rather bounded as $T \leq 2N-1$). By using (8.31) in place of (8.24), inequality (8.25) is tightened as follows

$$\gamma_i^{-1/2}\|U_i^\top(\boldsymbol{x}^{N\nu} - \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{N\nu}))\| \leq \tfrac{1+L_{\mathbf{T}}}{\sqrt{\xi_{\min}/2}}\Big(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{N\nu}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{N(\nu+1)})\Big)^{1/2}.$$

By squaring and summing for $i \in [N]$ we obtain

$$\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{N(\nu+1)}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{N\nu}) \leq -\tfrac{\xi_{\min}}{2N(1+L_{\mathbf{T}})^2}\|\boldsymbol{z}^{N\nu} - \boldsymbol{x}^{N\nu}\|_{\Gamma^{-1}}^2 = \tfrac{-(1-\Delta)}{2N(1+L_{\mathbf{T}})^2}\|\boldsymbol{z}^{N\nu} - \boldsymbol{x}^{N\nu}\|_{\Gamma^{-1}}^2, \tag{8.32}$$

where $L_{\mathbf{T}} = 1-\delta$ as discussed above. By combining this and (8.30) (with T replaced by N) the improved coefficient (8.29) is obtained. □

Note that if one sets $\gamma_i = \alpha N/L_{f_i}$ for some $\alpha \in (0,1)$, then $\delta =$

$\alpha \min_{i\in[N]} \{\mu_{f_i}/L_{f_i}\}$ and $\Delta = \alpha$. With this selection, as the condition number approaches 1 the rate in (8.29) tends to $1 - \frac{\alpha}{N(2-\alpha)^2}$.

8.2.5 Global and linear convergence with KL inequality

The convergence analyses of the randomized and essentially cyclic cases both rely on a descent property on the FBE that quantifies the progress in the minization of Φ_Γ^{FB} in terms of the squared forward-backward residual $\|\boldsymbol{x} - \boldsymbol{z}\|^2$. A subtle but important difference, however, is that the inequality (8.12) in the former case involves a conditional expectation, whereas (8.26) in the latter does not. The *sure* descent property occurring for essentially cyclic sampling strategies is the key for establishing global (as opposed to subsequential) convergence based on the Kurdyka-Łojasiewicz (KL) property [111, 112, 93]. A similar result is achieved in [179], which however considers the complementary case to problem (8.1) where the nonsmooth function G is assumed to be separable, and thus the cost function itself can serve as Lyapunov function.

Definition 8.10 (KL property with exponent θ)**.** *A proper lsc function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to have the* Kurdyka-Łojasiewicz *(KL) property with exponent $\theta \in (0,1)$ at $\bar{w} \in \operatorname{dom} h$ if there exist $\varepsilon, \eta, \varrho > 0$ such that*

$$\psi'(h(w) - h(\bar{w}))\operatorname{dist}(0, \partial h(w)) \geq 1$$

holds for all w such that $\|w - \bar{w}\| < \varepsilon$ and $h(\bar{w}) < h(w) < h(\bar{w}) + \eta$, where $\psi(s) \coloneqq \varrho s^{1-\theta}$. We say that h satisfies the KL property with exponent θ (without mention of $\bar{w}$) if it satisfies the KL property with exponent θ at any $\bar{w} \in \operatorname{dom} \partial h$.

Semialgebraic functions comprise a wide class of functions that enjoy this property [24, 23], which has been extensively exploited to provide convergence rates of optimization algorithms [9, 10, 11, 25, 72, 128, 102, 178]. Based on this, in the next result we provide sufficient conditions ensuring global and R-linear convergence of Algorithm 8.1 with essentially cyclic sampling.

Theorem 8.11 (essentially cyclic sampling: global and linear convergence)**.** *Additionally to Assumptions 8.I and 8.III, suppose that Φ has the KL property with exponent $\theta \in (0,1)$ (as is the case when f_i and G are semialgebraic), and is coercive. Then, any sequences $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ generated by Algorithm 8.1 converge to (the same) stationary point $\boldsymbol{x}^\star$. Moreover, if $\theta \leq 1/2$ then $(\|\boldsymbol{z}^k - \boldsymbol{x}^k\|)_{k\in\mathbb{N}}$, $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ converge at R-linear rate.*

Proof. Let $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ be sequences generated by Algorithm 8.1 with essentially cyclic sampling, and let $\Phi_\star$ be the limit of the sequence $(\Phi_\Gamma^{\text{FB}}(\boldsymbol{x}^k))_{k\in\mathbb{N}}$

as in Lemma 8.5*(ii)*. To avoid trivialities, we may assume that $\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) \gneqq \Phi_\star$ for all k, for otherwise the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is asymptotically constant, and thus so is $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$. Let Ω be the set of accumulation points of $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$, which is compact and such that $\Phi_\Gamma^{\mathrm{FB}} \equiv \Phi_\star$ on Ω, as ensured by Theorem 8.8. It follows from Lemma 8.23 and [9, Lem. 1(ii)] that $\Phi_\Gamma^{\mathrm{FB}}$ enjoys a *uniform* KL property on Ω; in particular, $\psi'(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \Phi_\star)\operatorname{dist}(0, \partial\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k)) \geq 1$ holds for all k large enough such that $\boldsymbol{x}^k$ is sufficiently close to Ω and $\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k)$ is sufficiently close to $\Phi_\star$, where $\psi(s) = \varrho s^{1-\theta'}$ for some $\varrho > 0$ and $\theta' = \max\{\theta, 1/2\}$. Combined with Lemma 8.22*(iii)*, for all k large enough we thus have

$$\psi'(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \Phi_\star) \geq \frac{c}{\|\boldsymbol{x}^k - \boldsymbol{z}^k\|_{\Gamma^{-1}}}, \tag{8.33}$$

where $c \coloneqq \frac{N \min_i\{\sqrt{\gamma_i}\}}{N+\max_i\{\gamma_i L_{f_i}\}} > 0$. Let $\Delta_k \coloneqq \psi(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \Phi_\star)$. By combining (8.33) and (8.26) we have that there exists a constant $c' > 0$ such that

$$\Delta_{(\nu+1)T} - \Delta_{\nu T} \leq \psi'(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{\nu T}) - \Phi_\star)\big(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{(\nu+1)T}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{\nu T})\big) \leq -c'\|\boldsymbol{x}^{\nu T} - \boldsymbol{z}^{\nu T}\|_{\Gamma^{-1}} \tag{8.34}$$

holds for all $\nu \in \mathbb{N}$ large enough (the first inequality uses concavity of ψ). By summing over ν (sure) summability of the sequence $(\|\boldsymbol{x}^{\nu T} - \boldsymbol{z}^{\nu T}\|)_{\nu\in\mathbb{N}}$ is obtained. By suitably shifting, for every $t \in [T]$ the same can be said for the sequence $(\|\boldsymbol{z}^{T\nu+t} - \boldsymbol{x}^{T\nu+t}\|)_{\nu\in\mathbb{N}}$, and since T is finite we conclude that the whole sequence $(\|\boldsymbol{z}^k - \boldsymbol{x}^k\|)_{k\in\mathbb{N}}$ is summable. Since $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| \leq \|\boldsymbol{z}^k - \boldsymbol{x}^k\|$ we conclude that $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ has finite length and is thus convergent (to a single point), and consequently so is $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$.

Suppose now that $\theta \leq 1/2$, so that $\psi(s) = \varrho\sqrt{s}$. Then,

$$\|\boldsymbol{x}^{\nu T} - \boldsymbol{z}^{\nu T}\|_{\Gamma^{-1}} \overset{(8.33)}{\geq} \tfrac{2c}{\varrho}\sqrt{\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{\nu T}) - \Phi_\star} = \tfrac{2c}{\varrho^2}\psi(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{\nu T}) - \Phi_\star) = \tfrac{2c}{\varrho^2}\Delta_{\nu T}.$$

Combined with (8.34) it follows that $(\Delta_{\nu T})_{\nu\in\mathbb{N}}$ conveges Q-linearly. By rearranging (8.34) as

$$c'\|\boldsymbol{x}^{\nu T} - \boldsymbol{z}^{\nu T}\|_{\Gamma^{-1}} \leq \Delta_{\nu T} - \Delta_{(\nu+1)T} \leq \Delta_{\nu T},$$

R-linear convergence of $(\|\boldsymbol{x}^{\nu T} - \boldsymbol{z}^{\nu T}\|)_{\nu\in\mathbb{N}}$ follows. By suitably shifting, for every $t \in [T]$ the same can be said for the sequence $(\|\boldsymbol{z}^{T\nu+t} - \boldsymbol{x}^{T\nu+t}\|)_{\nu\in\mathbb{N}}$, and since T is finite we conclude that the whole sequence $(\|\boldsymbol{z}^k - \boldsymbol{x}^k\|)_{k\in\mathbb{N}}$ converges R-linearly. On the other hand, since $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| \leq \|\boldsymbol{z}^k - \boldsymbol{x}^k\|$, also $(\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|)_{k\in\mathbb{N}}$ converges R-linearly, hence so does $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$. By combining the two, we conclude that also $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ converges R-linearly. □

8.3 Nonconvex finite sum problems: the Finito/MISO algorithm

As mentioned in Section 8.1, if G is of the form (8.3) then problem (8.1) reduces to the finite sum minimization presented in (8.2). Most importantly, the proximal mapping of the original nonsmooth function G can be easily expressed in terms of that of the small function g in the reduced finite sum reformulation, as shown in the next lemma.

Lemma 8.12. *Given $\gamma_i > 0$, $i \in [N]$, let $\Gamma := \mathrm{blkdiag}(\gamma_1 I_n, \ldots, \gamma_N I_n)$ and $\hat{\gamma} := \left(\sum_{i=1}^N \gamma_i^{-1}\right)^{-1}$. Then, for G as in (8.3) and any $\boldsymbol{u} \in \mathbb{R}^{Nn}$*

$$\mathrm{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}) = \left\{(\hat{v}, \ldots, \hat{v}) \mid \hat{v} \in \mathrm{prox}_{\hat{\gamma} g}(\hat{u})\right\} \quad \text{where} \quad \hat{u} := \hat{\gamma} \textstyle\sum_{i=1}^N \gamma_i^{-1} u_i.$$

Proof. Observe first that for every $w \in \mathbb{R}^n$ one has

$$\begin{aligned}\textstyle\sum_i \gamma_i^{-1}\|w-u_i\|^2 &= \textstyle\sum_i \gamma_i^{-1}\|\hat{u}-u_i\|^2 + \sum_i \gamma_i^{-1}\|w-\hat{u}\|^2 + \overbrace{2\textstyle\sum_i \gamma_i^{-1}\langle \hat{u}-u_i, w-\hat{u}\rangle}^{=0} \\ &= \textstyle\sum_i \gamma_i^{-1}\|\hat{u}-u_i\|^2 + \hat{\gamma}^{-1}\|w-\hat{u}\|^2. \end{aligned} \tag{8.35}$$

Next, observe that since $\mathrm{dom}\, G \subseteq C$ (the consensus set),

$$\begin{aligned}\mathrm{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}) &= \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^{Nn}} \left\{G(\boldsymbol{w}) + \textstyle\sum_{i=1}^N \frac{1}{2\gamma_i}\|w_i - u_i\|^2\right\} \\ &= \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^{Nn}} \left\{G(\boldsymbol{w}) + \textstyle\sum_{i=1}^N \frac{1}{2\gamma_i}\|w_i - u_i\|^2 \mid w_1 = \cdots = w_N\right\} \\ &= \operatorname*{arg\,min}_{(w,\ldots,w)} \left\{g(w) + \textstyle\sum_{i=1}^N \frac{1}{2\gamma_i}\|w - u_i\|^2\right\} \\ &\overset{(8.35)}{=} \operatorname*{arg\,min}_{(w,\ldots,w)} \left\{g(w) + \tfrac{1}{2\hat{\gamma}}\|w - \hat{u}\|^2\right\} = \left\{(\hat{v}, \ldots, \hat{v}) \mid \hat{v} \in \mathrm{prox}_{\hat{\gamma} g}(\hat{u})\right\}\end{aligned}$$

as claimed. □

If all stepsizes are set to a same value γ, so that $\Gamma = \gamma \mathrm{I}_{Nn}$, then the forward-backward step reduces to

$$\begin{aligned}\boldsymbol{z} \in \mathrm{prox}_G^{\Gamma^{-1}}(\boldsymbol{x} - \Gamma\nabla F(\boldsymbol{x})) \quad &\Leftrightarrow \quad \boldsymbol{z} = (\bar{z}, \ldots, \bar{z}), \\ &\bar{z} \in \mathrm{prox}_{\gamma g/N}\left(\tfrac{1}{N}\textstyle\sum_{j=1}^N \left(x_j - \frac{\gamma}{N}\nabla f_j(x_j)\right)\right).\end{aligned} \tag{8.36}$$

The argument of $\operatorname{prox}_{\gamma g/N}$ is the (unweighted) average of the forward operator. By applying Algorithm 8.1 with (8.36), Finito/MISO [60, 114] is recovered. Differently from the existing convergence analyses, ours covers fully nonconvex and nonsmooth problems, more general sampling strategies and the possibility to select different stepsizes γ_i for each block, which can have a significant impact on the performance compared to the case where all stepsizes are equal. Moreover, to the best of our knowledge this is the first work that shows global convergence and linear rates even when the smooth functions are nonconvex. The resulting scheme is presented in Algorithm 8.2. We remark that the consensus formulation to recover Finito/MISO (although from a different umbrella algorithm) was also observed in [56] in the convex case. Moreover, the Finito/MISO algorithm with cyclic sampling is also studied in [117] when $g \equiv 0$ and f_i are strongly convex functions; consistently with Assumption 8.III, our analysis covers the more general essentially cyclic sampling even in the presence of a nonsmooth convex term g and allowing the smooth functions f_i to be nonconvex. Randomized Finito/MISO with $g \equiv 0$ is also studied in the recent work [135]; although their analysis is limited to a single stepsize, in the convex case it is allowed to be larger than our worst-case stepsize $\min_i \gamma_i$.

Algorithm 8.2 Nonconvex proximal Finito/MISO for problem (8.2)

REQUIRE $x^{\text{init}} \in \mathbb{R}^n$; $\gamma_i \in (0, {}^N\!/_{L_{f_i}})$, $i \in [N]$

INITIALIZE $\hat{\gamma} := \left(\sum_{i=1}^N \gamma_i^{-1}\right)^{-1}$; $s_i^0 = x^{\text{init}} - \frac{\gamma_i}{N}\nabla f_i(x^{\text{init}})$, $i \in [N]$; $\hat{s}^0 = \sum_{i=1}^N \frac{\hat{\gamma}}{\gamma_i} s_i^0$

REPEAT FOR $k = 0, 1, \ldots$ UNTIL CONVERGENCE

1: $z^k \in \operatorname{prox}_{\hat{\gamma} g}(\hat{s}^k)$

2: select a set of indices $I^k \subseteq [N]$, set $\hat{s}^{k+1} = \hat{s}^k$ and $s_i^{k+1} = s_i^k, i \notin I^k$

3: **for** $i \in I^k$ **do**

4: $\quad s_i^{k+1} \leftarrow z^k - \frac{\gamma_i}{N}\nabla f_i(z^k)$ ▷ buffer update

5: $\quad \hat{s}^{k+1} \leftarrow \hat{s}^{k+1} + \frac{\hat{\gamma}}{\gamma_i}(s_i^{k+1} - s_i^k)$ ▷ average update

RETURN z^k

The convergence results from Section 8.2 are immediately translated to this setting by noting that the bold variable $\boldsymbol{z}^k$ corresponds to $(z^k, \ldots, z^k)$. Therefore, $\Phi(\boldsymbol{z}^k) = \varphi(z^k)$ where φ is the cost function for the finite sum problem.

Corollary 8.13 (subsequential convergence of Algorithm 8.2)**.** *In the finite sum problem (8.2) suppose that* $\operatorname{argmin}\varphi$ *is nonempty, g is proper and lsc, and each f_i is L_{f_i}-Lipschitz differentiable, $i \in [N]$. Then, the following hold almost surely (resp. surely) for the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 8.2 with*

randomized sampling strategy as in Assumption 8.II (resp. with any essentially cyclic sampling strategy and g convex as required in Assumption 8.III):

(i) *the sequence $(\varphi(z^k))_{k\in\mathbb{N}}$ converges to a finite value $\varphi_\star \leq \varphi(x^{\text{init}})$;*

(ii) *all cluster points of the sequence $(z^k)_{k\in\mathbb{N}}$ are stationary and on which φ equals $\varphi_\star$.*

If, additionally, φ is coercive, then the following also hold:

(iii) *$(z^k)_{k\in\mathbb{N}}$ is bounded (in fact, this holds surely for arbitrary sampling criteria).*

Corollary 8.14 (linear convergence of Algorithm 8.2 under strong convexity)**.** *Additionally to the assumptions of Corollary 8.13, suppose that g is convex and that each f_i is μ_{f_i}-strongly convex. The following hold for the iterates generated by Algorithm 8.2:*

RANDOMIZED SAMPLING: *under Assumption 8.II,*

$$\mathbb{E}\big[\varphi(z^k) - \min\varphi\big] \leq (\varphi(x^{\text{init}}) - \min\varphi)(1-c)^k$$

$$\tfrac{1}{2}\mathbb{E}\big[\|z^k - x^\star\|^2\big] \leq \frac{N(\varphi(x^{\text{init}}) - \min\varphi)}{\sum_i \mu_{f_i}}(1-c)^k$$

holds for all $k \in \mathbb{N}$, where c is as in (8.14) and $x^\star \coloneqq \operatorname{argmin}\varphi$. If the stepsizes γ_i and the sampling probabilities p_i are set as in Theorem 8.7, then the tighter constant c as in (8.16) is obtained.

SHUFFLED CYCLIC OR CYCLIC SAMPLING: *under either sampling strategy (8.18) or (8.19),*

$$\varphi(z^{\nu N}) - \min\varphi \leq (\varphi(x^{\text{init}}) - \min\varphi)(1-c)^\nu$$

$$\tfrac{1}{2}\mathbb{E}\big[\|z^{\nu N} - x^\star\|^2\big] \leq \frac{N(\varphi(x^{\text{init}}) - \min\varphi)}{\sum_i \mu_{f_i}}(1-c)^\nu$$

holds surely for all $\nu \in \mathbb{N}$, where c is as in (8.29).

The next result follows from Theorem 8.11 once the needed properties of Φ as in the umbrella formulation (8.1) are shown to hold.

Corollary 8.15 (global convergence of Algorithm 8.2)**.** *In the finite sum problem (8.2), suppose that φ has the KL property with exponent $\theta \in (0,1)$ (as is the case when f_i and g are semialgebraic) and coercive, g is proper convex*

and lsc, and each f_i is L_{f_i}-Lipschitz differentiable, $i \in [N]$. Then, the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 8.2 with any essentially cyclic sampling strategy as in Assumption 8.III converges surely to a stationary point for φ. Moreover, if $\theta \leq 1/2$ then it converges at R-linear rate.

Proof. Function $\Phi = F + G$ be as in (8.3) clearly is coercive and satisfies Assumption 8.I. In order to invoke Theorem 8.11 is suffices to show that there exists a constant $c > 0$ such that

$$\operatorname{dist}(0, \partial\Phi(\boldsymbol{x})) \geq c \operatorname{dist}(0, \partial\varphi(x)) \quad \text{for all } x \in \mathbb{R}^n \text{ and } \boldsymbol{x} = (x, \dots, x), \tag{8.37}$$

as this will ensure that Φ enjoys the KL property at $\boldsymbol{x}^\star = (x^\star, \dots, x^\star)$ with same desingularizing function (up to a positive scaling). Notice that for $x \in \mathbb{R}^n$ and $\boldsymbol{x} = (x, \dots, x)$, one has $\boldsymbol{v} \in \partial G(\boldsymbol{x})$ iff $\frac{1}{N}\sum_{i=1}^N v_i \in \partial g(x)$. Since $\partial\Phi(\boldsymbol{x}) = \frac{1}{N} \times_{i=1}^N \nabla f_i(x_i) + \partial G(\boldsymbol{x})$ and $\partial\varphi(x) = \frac{1}{N}\sum_{i=1}^N \nabla f_i(x) + \partial g(x)$, see [144, Ex. 8.8(c) and Prop. 10.5], for $x \in \mathbb{R}^n$ and denoting $\boldsymbol{x} = (x, \dots, x)$ we have

$$\begin{aligned}\operatorname{dist}(0, \partial\varphi(x)) &\leq \inf_{\boldsymbol{v}\in\partial G(\boldsymbol{x})} \left\| \tfrac{1}{N}\textstyle\sum_{i=1}^N \nabla f_i(x) + \tfrac{1}{N}\sum_{i=1}^N v_i \right\| \\ &\leq \tfrac{1}{N} \inf_{\boldsymbol{v}\in\partial G(\boldsymbol{x})} \textstyle\sum_{i=1}^N \|\nabla f_i(x) + v_i\| = \tfrac{1}{N} \inf_{\boldsymbol{u}\in\partial\Phi(\boldsymbol{x})} |||\boldsymbol{u}|||,\end{aligned}$$

where $|||\cdot|||$ is the norm in $\mathbb{R}^{Nn}$ given by $|||\boldsymbol{w}||| = \sum_{i=1}^N \|w_i\|$. Inequality (8.37) then follows by observing that $\inf_{\boldsymbol{u}\in\partial\Phi(\boldsymbol{x})} |||\boldsymbol{u}|||$ is the distance of 0 from $\partial\Phi(\boldsymbol{x})$ in the norm $|||\cdot|||$, hence that $|||\cdot||| \leq c'\|\cdot\|$ for some $c' > 0$. □

8.3.1 A low memory variant

The main drawback of the Finito/MISO algorithm is the high memory requirement. Implementing Algorithm 8.2 requires keeping a table of s_i's in the memory, which can be prohibitively large in some applications. In this section we present Algorithm 8.3 (L-Finito), a low memory variant of Algorithm 8.2, that is inspired by stochastic variance reduced gradient method (SVRG) [91, 177]. SVRG is designed for strongly convex problems. In [3] a variant of SVRG is studied that allows f_i to be nonconvex while requiring their sum as well as g to be convex. In [137] SVRG is studied under a different stepsize condition for nonconvex f_i, and with $g = 0$. This analysis is extended for proxSVRG in [138] with convex g. In contrast, while Algorithm 8.3 (L-Finito) enjoys the low memory requirements of SVRG, it covers fully nonconvex regularized finite sum problems with simple-to-choose constant stepsizes.

The proposed algorithm involves a full proximal gradient update followed by a (shuffled) cyclic sweeping of indices. Note that in the inner loop any sweeping strategy may be used as long as no index is selected twice. This is unlike SVRG which allows for an index to be selected (at random) more than once in the inner loop. The idea here is that after the full update step, the common point z^k is used for the update of the average $\hat{s}$, thus eliminating the need for storing s_i's. The reduction in memory requirements comes at the cost of repeated gradient evaluations $\nabla f_i(z^k)$ within the inner loop. Just as is the case for SVRG [91] this can be avoided by storing a table of gradients which can be maintained cheaply in applications such as logistic regression and least squares.

Algorithm 8.3 Nonconvex low-memory Finito/MISO (L-Finito) for problem (8.2)

REQUIRE $x^{\text{init}} \in \mathbb{R}^n$; $\gamma_i \in (0, {}^N\!/_{L_{f_i}})$, $i \in [N]$

INITIALIZE $\hat{\gamma} := \left(\sum_{i=1}^N \gamma_i^{-1}\right)^{-1}$; average $\hat{s}^0 = x^{\text{init}} - \frac{\hat{\gamma}}{N}\sum_{i=1}^N \nabla f_i(x^{\text{init}})$

REPEAT FOR $k = 0, 1, \ldots$ UNTIL CONVERGENCE

1: $z^k \in \operatorname{prox}_{\hat{\gamma} g}(\hat{s}^k)$

2: $\hat{s}^{k+1} = z^k - \frac{\hat{\gamma}}{N}\sum_{i=1}^N \nabla f_i(z^k)$ ▷ full update

3: Select ℓ^k sets of disjoint indices $I_1^k, \ldots, I_{\ell^k}^k \subseteq [N]$

4: **for** $j = 1, \ldots, \ell^k$ **do** ▷ inner loop

5: $\quad z^{\text{temp}} \in \operatorname{prox}_{\hat{\gamma} g}(\hat{s}^{k+1})$

6: $\quad$ **for** $i \in I_j^k$ **do**

7: $\quad\quad \hat{s}^{k+1} \leftarrow \hat{s}^{k+1} + \frac{\hat{\gamma}}{\gamma_i}(z^{\text{temp}} - z^k) - \frac{\hat{\gamma}}{N}(\nabla f_i(z^{\text{temp}}) - \nabla f_i(z^k))$

RETURN z^k

The convergence of Alg. 8.3 is a consequence of the sure descent property and is summarized below. We omit convergence rate results for the strongly convex case.

Corollary 8.16 (convergence of Algorithm 8.3)**.** *In the finite sum problem (8.2) suppose that* $\operatorname{argmin}\varphi$ *is nonempty, g is proper and lsc, and each f_i is L_{f_i}-Lipschitz differentiable, $i \in [N]$. Then, the assertions in Cor.s 8.13(i) and 8.13(ii) hold surely for the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 8.3. If in addition φ has the KL property with exponent $\theta \in (0, 1)$ (as is the case when f_i and g are semialgebraic) and is coercive, then $(z^k)_{k\in\mathbb{N}}$ converges surely to a stationary point for φ. Moreover, if $\theta \leq 1/2$ then it converges at R-linear rate.*

8.4 Nonconvex sharing problem

In this section we consider the sharing problem (8.4). As discussed in Section 8.1, (8.4) fits into the problem framework (8.1) by simply letting $G := g \circ A$, where $A := [\mathbf{I}_n \ \dots \ \mathbf{I}_n] \in \mathbb{R}^{n \times nN}$. By arguing as in [14, Thm. 6.15] it can be shown that, when A has full row rank, the proximal mapping of $G = g \circ A$ is given by

$$\operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}) = \boldsymbol{u} + \Gamma A^\top (A\Gamma A^\top)^{-1} \Big(\operatorname{prox}_g^{(A\Gamma A^\top)^{-1}}(A\boldsymbol{u}) - A\boldsymbol{u} \Big). \tag{8.38}$$

Since $A\Gamma A^\top = \sum_{i=1}^N \gamma_i$ for the sharing problem (8.4),

$$\boldsymbol{v} \in \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}) \quad \Leftrightarrow \quad \boldsymbol{v} = (u_1 + \gamma_1 w, \dots, u_N + \gamma_N w)$$

$$w \in \tilde{\gamma}^{-1} \big(\operatorname{prox}_{\tilde{\gamma} g}(\tilde{u}) - \tilde{u} \big), \quad \tilde{\gamma} := \textstyle\sum_{i=1}^N \gamma_i, \quad \tilde{u} := \sum_{i=1}^N u_i.$$

Consequently, the general BC Algorithm 8.1 when applied to the sharing problem (8.4) reduces to Algorithm 8.4.

Algorithm 8.4 Block-coordinate method for nonconvex sharing problem

REQUIRE $x_i^{\text{init}} \in \mathbb{R}^n$, $\gamma_i \in (0, {}^N\!/_{L_{f_i}})$, $i \in [N]$

INITIALIZE $\tilde{\gamma} := \sum_{i=1}^N \gamma_i$, $s_i^0 = x_i^{\text{init}} - \frac{\gamma_i}{N} \nabla f_i(x_i^{\text{init}})$ $i \in [N]$, $\tilde{s}^0 = \sum_{i=1}^N s_i^0$

REPEAT FOR $k = 0, 1, \dots$ UNTIL CONVERGENCE

1: $w^k \leftarrow \tilde{\gamma}^{-1} (\operatorname{prox}_{\tilde{\gamma} g}(\tilde{s}^k) - \tilde{s}^k)$

2: select a set of indices $I^k \subseteq [N]$, set $\hat{s}^{k+1} = \hat{s}^k$ and $s_i^{k+1} = s_i^k, i \notin I^k$

3: **for** $i \in I^k$ **do**

4: $\quad s_i^{k+1} \leftarrow s_i^k + \gamma_i w^k - \frac{\gamma_i}{N} \nabla f_i(s_i^k + \gamma_i w^k)$ ▷ update buffer

5: $\quad \tilde{s}^{k+1} \leftarrow \tilde{s}^{k+1} + (s_i^{k+1} - s_i^k)$ ▷ update sum

RETURN $\boldsymbol{z}^k = (s_1^k + \gamma_1 w^k, \dots, s_N^k + \gamma_N w^k)$

Remark 8.17 (generalized sharing constraint)**.** Another notable instance of $G = g \circ A$ well suited for the BC framework of Algorithm 8.1 is when $g = \delta_{\{0\}}$ and $A = [A_1 \ \dots \ A_N]$, $A_i \in \mathbb{R}^{n \times n_i}$ such that A is full row rank. This models the generalized sharing problem

$$\underset{\boldsymbol{x} \in \mathbb{R}^{\sum_i n_i}}{\text{minimize}} \ \tfrac{1}{N} \textstyle\sum_{i=1}^N f_i(x_i) \quad \text{subject to} \ \ \sum_{i=1}^N A_i x_i = 0.$$

In this case (8.38) simplifies to

$$\left(\operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u})\right)_i = u_i - \gamma_i A_i^\top \mathcal{A}^{-1} \sum_{i=1}^N A_i u_i,$$

where $\mathcal{A} \coloneqq A\Gamma A^\top$ can be factored offline and $\sum_{i=1}^N A_i x_i$ can be updated in an incremental fashion in the same spirit of Algorithm 8.4. □

The convergence results for Algorithm 8.4 summarized below fall as special cases of those in Section 8.2.

Corollary 8.18 (convergence of Algorithm 8.4)**.** *In the sharing problem* (8.4)*, suppose that* $\operatorname{arg\,min}\Phi$ *is nonempty,* g *is proper and lsc, and each* f_i *is* L_{f_i}*-Lipschitz differentiable,* $i \in [N]$*. Consider the sequences* $(w^k)_{k\in\mathbb{N}}$ *and* $(\boldsymbol{s}^k)_{k\in\mathbb{N}}$ *generated by Algorithm 8.4 and let* $(\boldsymbol{z}^k)_{k\in\mathbb{N}} = (s_1^k + \gamma_1 w^k, \dots, s_N^k + \gamma_N w^k)_{k\in\mathbb{N}}$*. Then, the following hold almost surely (resp. surely) with randomized sampling strategy as in Assumption 8.II (resp. with any essentially cyclic sampling strategy and* g *convex as required in Assumption 8.III):*

(i) the sequence $(\Phi(\boldsymbol{z}^k))_{k\in\mathbb{N}}$ *converges to a finite value* $\Phi_\star \le \Phi(\boldsymbol{x}^{\text{init}})$*;*

(ii) all cluster points of the sequence $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ *are stationary and on which* Φ *equals* $\Phi_\star$*.*

If, additionally, Φ *is coercive, then the following also hold:*

(iii) $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ *is bounded (in fact, this holds surely for arbitrary sampling criteria).*

Corollary 8.19 (linear convergence of Algorithm 8.4 under strong convexity)**.** *Additionally to the assumptions of Corollary 8.18, suppose that* g *is convex and that each* f_i *is* μ_{f_i}*-strongly convex. The following hold:*

Randomized sampling: *under Assumption 8.II,*

$$\mathbb{E}\big[\Phi(\boldsymbol{z}^k) - \min\Phi\big] \le \big(\Phi(\boldsymbol{x}^{\text{init}}) - \min\Phi\big)(1-c)^k$$

$$\tfrac{1}{2}\mathbb{E}\big[\|\boldsymbol{z}^k - \boldsymbol{x}^\star\|_{\mu_F}^2\big] \le \big(\Phi(\boldsymbol{x}^{\text{init}}) - \min\Phi\big)(1-c)^k$$

holds for all $k \in \mathbb{N}$*, where* $\boldsymbol{x}^\star \coloneqq \operatorname{arg\,min}\Phi$*,* c *is as in* (8.14) *and* $\mu_F \coloneqq \frac{1}{N}\operatorname{blkdiag}\big(\mu_{f_1}\mathrm{I}_{n_1}, \dots \mu_{f_n}\mathrm{I}_{n_N}\big)$*. If the stepsizes* γ_i *and the sampling probabilities* p_i *are set as in Theorem 8.7, then the tighter constant* c *as in* (8.16) *is obtained.*

Shuffled cyclic or cyclic sampling: *under either* (8.18) *or* (8.19),

$$\Phi(\boldsymbol{z}^{N\nu}) - \min \Phi \leq \big(\Phi(\boldsymbol{x}^{\text{init}}) - \min \Phi\big)(1-c)^{\nu}$$

$$\tfrac{1}{2}\|\boldsymbol{z}^{N\nu} - \boldsymbol{x}^{\star}\|_{\mu_F}^2 \leq \big(\Phi(\boldsymbol{x}^{\text{init}}) - \min \Phi\big)(1-c)^{\nu}$$

holds surely for all $\nu \in \mathbb{N}$*, where* c *is as in* (8.29)*.*

We conclude with an immediate consequence of Theorem 8.11 that shows that (strong) convexity is in fact not necessary for global or linear convergence to hold.

Corollary 8.20 (global and linear convergence of Algorithm 8.4)**.** *In problem* (8.4)*, suppose that* Φ *has the KL property with exponent* $\theta \in (0,1)$ *(as is the case when* g *and* f_i *are semialgebraic) and is coercive,* g *is proper convex lsc, and each* f_i *is* L_{f_i}*-Lipschitz differentiable,* $i \in [N]$*. Then, the sequence* $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ *as defined in Corollary 8.18 with any essentially cyclic sampling strategy as in Assumption 8.III converges surely to a stationary point for* Φ*. Moreover, if* $\theta \leq 1/2$ *it converges with R-linear rate.*

8.5 Numerical simulations for the regularized finite sum problem

In this section we evaluate the performance of Algorithms 8.2 and 8.3 with two sets of experiments on ℓ_1-regularized and ℓ_0-norm ball constrained least squares problems. The simulations are performed using the open-source julia package *CIAOAlgorithms*[1]. In all of our simulations, stepsizes $\gamma_i = {}^{0.99N}/_{L_{f_i}}$ are used (independent of minibatch size). Whenever a single stepsize is used, *i.e.*, $\gamma = \min_i \gamma_i$, this is indicated in the legend by adding -γ to the algorithm's name. For Algorithm 8.3 (L-Finito), we used $I_i^k = \{i\}$ and $\ell^k = N$, *i.e.*, a cyclical inner loop. In the figures, each iteration of Algorithm 8.3 is counted as 3 passes through the data due to the repeated $\nabla f_i(z^k)$ update in the inner loop.

Lasso: Consider the regularized least squares problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \frac{1}{N}\sum_{i=1}^{N} \underbrace{\tfrac{1}{2}(\langle a_i, x\rangle - b_i)^2}_{f_i(x)} + \|x\|_1, \tag{8.39}$$

[1] https://github.com/kul-forbes/CIAOAlgorithms

where the pair (a_i, b_i) represents the ith data. In our simulations we used $N = 1000$, $n = 5000$ and the sparsity parameter for the solution was set to $p = 0.01$. The data for our simulations were randomly generated according to the procedure described in [126, §6]. In order to evaluate the performance of the algorithms we generated the matrix B in [126, §6] with two different condition numbers $\operatorname{cond}(B) = 20$ and $\operatorname{cond}(B) = 2.5$ and normalized its columns. This leads to $\max_i L_{f_i} / \min_i L_{f_i}$ equal to 144 and 3.3. The simulations are provided in Figure 8.1 with two different initializations. See our discussion at the end of the section.

Figure 8.2 depicts the behavior of the algorithms for different batch sizes. We only plot the result for cyclical sampling with different batch sizes b, since it outperformed randomized and shuffled variants. It is observed that lower batch sizes result in faster convergence. This effect was less pronounced in the low memory variant which is due to the presence of a full proximal gradient update within each iteration.

Sparsity constrained least squares: Consider the ℓ_0-norm ball constrained least squares problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} (\langle a_i, x \rangle - b_i)^2, \quad \text{subject to } \|x\|_0 \leq l. \tag{8.40}$$

For this example we used the same data as described in the previous section. In the simulation $l = 50$ was used. The results are depicted in Figure 8.3. Note that all algorithms converged to the same cost $\varphi^\star$. Hence, the distance of the cost from $\varphi^\star$ is used to measure their performance.

We make the following observations:

- The cyclical variant consistently outperforms the shuffled and randomized variants.
- The low memory variant, Algorithm 8.3 (L-Finito), has comparable performance to randomized variant of Algorithm 8.2.
- When using a single stepsize the speed of the algorithms is dictated by the largest L_{f_i}, and as expected, it can be much slower when the data is not uniform in the sense that $\max_i L_{f_i} / \min_i L_{f_i}$ is large.
- The choice of initial point affects the convergence rate on nonstrongly convex problems.

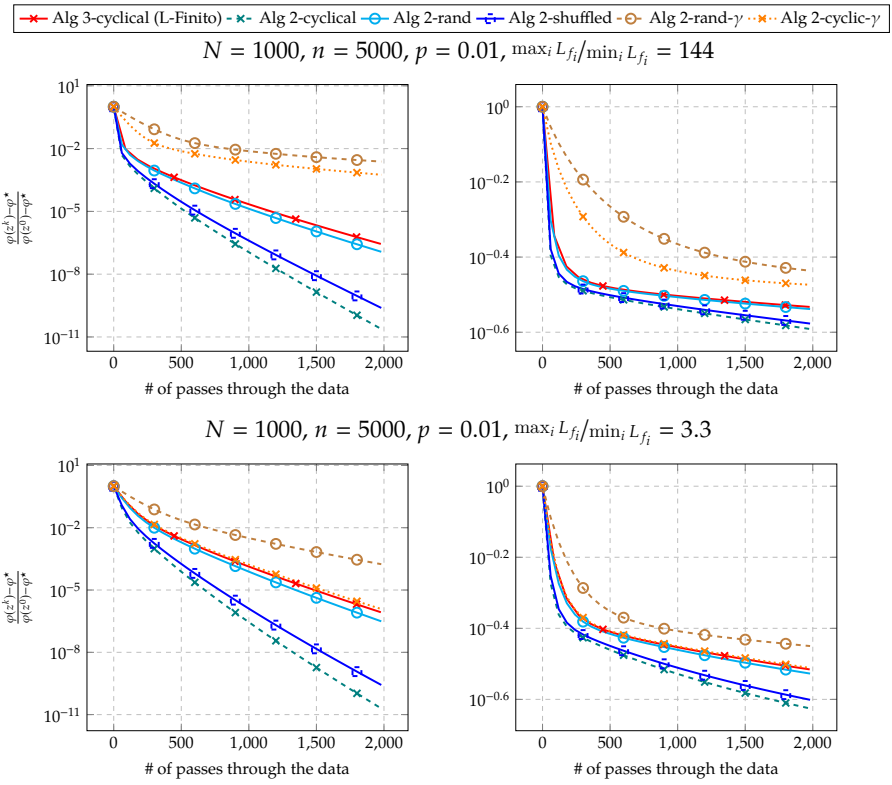

Figure 8.1: *Performance comparisons for Algorithms 8.2 and 8.3 for problem (8.39) with different initial points: (left)* $x^{\text{init}} = 0_n$*, (right)* $x^{\text{init}} = 1_n$*. Algorithm names ending with -*γ *indicate the use of a single stepsize* $\gamma = \min_i \gamma_i$*.*

8.6 Conclusions

We presented a general block-coordinate forward-backward algorithm for minimizing the sum of a separable smooth and a nonseparable nonsmooth function, both allowed to be nonconvex. The framework is general enough to encompass regularized finite sum minimization and sharing problems, and leads to (a generalization of) the Finito/MISO algorithm [60, 114] with new convergence results, and to another novel incremental algorithm (for the sharing problem). The forward-backward envelope is shown to be a particularly suitable Lyapunov function for establishing convergence: additionally to enjoying favorable continuity properties, *sure* descent (as opposed to in expectation) occurs along the iterates. Possible future developments include extending the framework to ac-

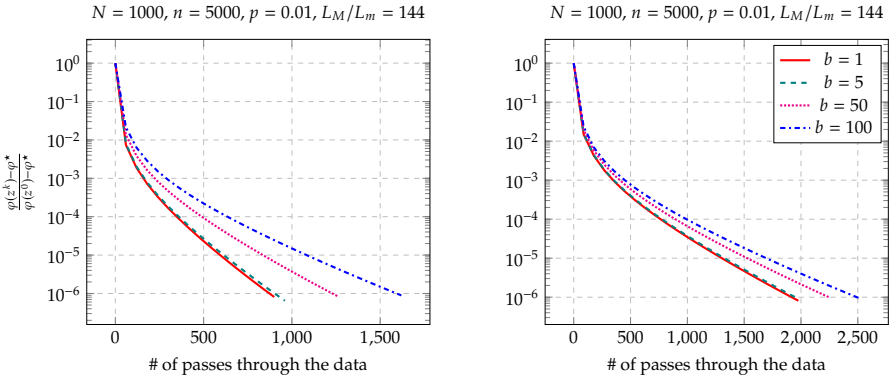

Figure 8.2: *(left) Algorithm 8.2 and (right) Algorithm 8.3 (L-Finito) with cyclical sampling and different batch sizes for solving problem (8.39).*

count for a nonseparable smooth term, for instance by "quantifying the strength of coupling" between blocks of variables as in [20, §7.5].

8.7 The key tool: the forward-backward envelope

This appendix contains some proofs and auxiliary results omitted in the main body. We begin by observing that, since F and $-F$ are 1-smooth in the metric induced by $\Lambda_F \coloneqq \frac{1}{N}\operatorname{blkdiag}(L_{f_1}\mathrm{I}_{n_1}, \dots, L_{f_N}\mathrm{I}_{n_N})$, one has

$$F(\boldsymbol{x})+\langle \nabla F(\boldsymbol{x}), \boldsymbol{w}-\boldsymbol{x}\rangle - \tfrac{1}{2}\|\boldsymbol{w}-\boldsymbol{x}\|^2_{\Lambda_F} \le F(\boldsymbol{w}) \le F(\boldsymbol{x})+\langle \nabla F(\boldsymbol{x}), \boldsymbol{w}-\boldsymbol{x}\rangle + \tfrac{1}{2}\|\boldsymbol{w}-\boldsymbol{x}\|^2_{\Lambda_F} \tag{8.41}$$

for all $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^{\sum_i n_i}$, see [19, Prop. A.24]. Let us denote

$$\mathcal{M}_\Gamma(\boldsymbol{w}, \boldsymbol{x}) \coloneqq F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{w}-\boldsymbol{x}\rangle + G(\boldsymbol{w}) + \tfrac{1}{2}\|\boldsymbol{w}-\boldsymbol{x}\|^2_{\Gamma^{-1}}$$

the quantity being minimized (with respect to $\boldsymbol{w}$) in the definition (8.8a) of the FBE. It follows from (8.41) that

$$\Phi(\boldsymbol{w}) + \tfrac{1}{2}\|\boldsymbol{w}-\boldsymbol{x}\|^2_{\Gamma^{-1}-\Lambda_F} \le \mathcal{M}_\Gamma(\boldsymbol{w}, \boldsymbol{x}) \le \Phi(\boldsymbol{w}) + \tfrac{1}{2}\|\boldsymbol{w}-\boldsymbol{x}\|^2_{\Gamma^{-1}+\Lambda_F} \tag{8.42}$$

holds for all $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^{\sum_i n_i}$. In particular, $\mathcal{M}_\Gamma$ is a *majorizing model* for Φ, in the sense that $\mathcal{M}_\Gamma(\boldsymbol{x}, \boldsymbol{x}) = \Phi(\boldsymbol{x})$ and $\mathcal{M}_\Gamma(\boldsymbol{w}, \boldsymbol{x}) \ge \Phi(\boldsymbol{w})$ for all $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^{\sum_i n_i}$. In fact, as explained in Section 8.2.1, while a Γ-forward-backward step $\boldsymbol{z} \in \mathrm{T}^{\mathrm{FB}}_\Gamma(\boldsymbol{x})$ amounts to evaluating a minimizer of $\mathcal{M}_\Gamma(\,\cdot\,, \boldsymbol{x})$, the FBE is defined instead as

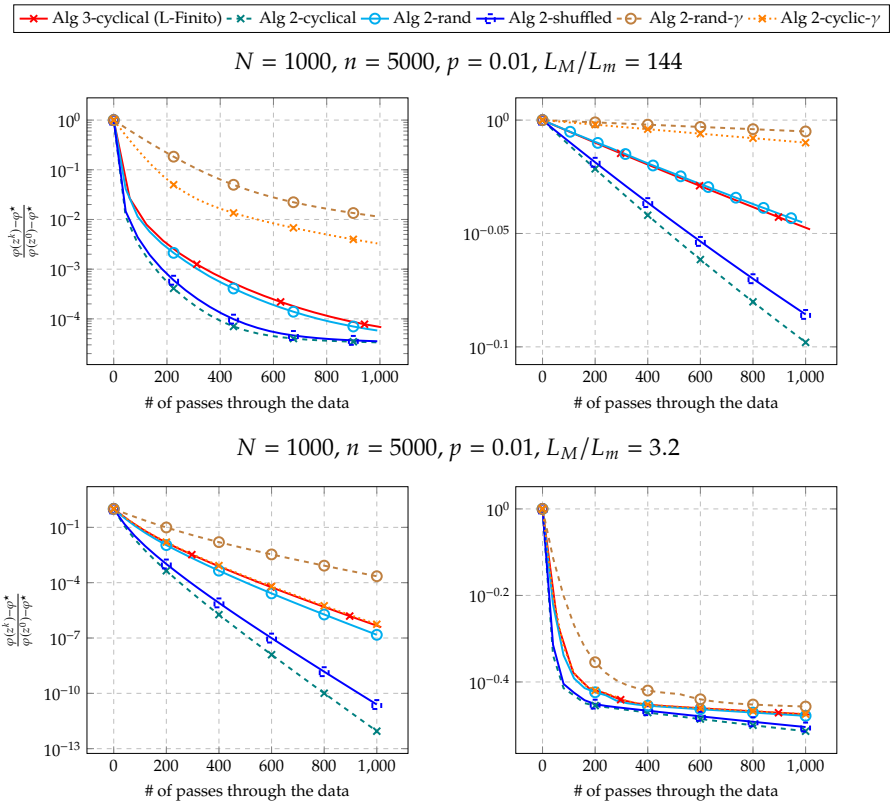

Figure 8.3: *Performance comparisons for Algorithms 8.2 and 8.3 for problem (8.40) with different initial points: (left)* $x^{\text{init}} = 0_n$*, (right)* $x^{\text{init}} = 1_n$*. Algorithm names ending with -*γ *indicate the use of a single stepsize* $\gamma = \min_i \gamma_i$

the minimization value, namely $\Phi_\Gamma^{\text{FB}}(\boldsymbol{x}) = \mathcal{M}_\Gamma(\boldsymbol{z}, \boldsymbol{x})$ where $\boldsymbol{z}$ is any element of $\mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{x})$.

8.7.1 Further results

This section contains a list of auxiliary results invoked in the main proofs of Section 8.2.

Lemma 8.21. *Suppose that Assumption 8.1 holds, and let two sequences* $(\boldsymbol{u}^k)_{k\in\mathbb{N}}$ *and* $(\boldsymbol{v}^k)_{k\in\mathbb{N}}$ *satisfy* $\boldsymbol{v}^k \in \mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{u}^k)$ *for all* k *and be such that both*

converge to a point $\boldsymbol{u}^\star$ *as* $k \to \infty$. *Then,* $\boldsymbol{u}^\star \in \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{u}^\star)$, *and in particular* $0 \in \hat{\partial}\Phi(\boldsymbol{u}^\star)$.

Proof. Since ∇F is continuous, it holds that $\boldsymbol{u}^k - \Gamma\nabla F(\boldsymbol{u}^k) \to \boldsymbol{u}^\star - \Gamma\nabla F(\boldsymbol{u}^\star)$ as $k \to \infty$. From outer semicontinuity of $\operatorname{prox}_G^{\Gamma^{-1}}$ [144, Ex. 5.23(b)] it then follows that

$$\begin{aligned}\boldsymbol{u}^\star = \lim_{k\to\infty} \boldsymbol{v}^k &\in \limsup_{k\to\infty} \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}^k - \Gamma\nabla F(\boldsymbol{u}^k)) \\ &\subseteq \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}^\star - \Gamma\nabla F(\boldsymbol{u}^\star)) = \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{u}^\star),\end{aligned}$$

where the limit superior is meant in the Painlevé-Kuratowski sense, *cf.* [144, Def. 4.1]. The optimality conditions defining $\operatorname{prox}_G^{\Gamma^{-1}}$ [144, Thm. 10.1] then read

$$\begin{aligned}0 &\in \hat{\partial}\big(G + \tfrac{1}{2}\|\cdot - (\boldsymbol{u}^\star - \Gamma\nabla F(\boldsymbol{u}^\star))\|^2_{\Gamma^{-1}}\big)(\boldsymbol{u}^\star) \\ &= \hat{\partial}G(\boldsymbol{u}^\star) + \Gamma^{-1}(\boldsymbol{u}^\star - (\boldsymbol{u}^\star - \Gamma\nabla F(\boldsymbol{u}^\star))) \\ &= \hat{\partial}G(\boldsymbol{u}^\star) + \nabla F(\boldsymbol{u}^\star) = \hat{\partial}\Phi(\boldsymbol{u}^\star),\end{aligned}$$

where the first and last equalities follow from [144, Ex. 8.8(c)]. □

Lemma 8.22. *Suppose that Assumption 8.I holds and that function* G *is convex. Then, the following hold:*

(i) $\operatorname{prox}_G^{\Gamma^{-1}}$ *is (single-valued and) firmly nonexpansive (FNE) in the metric* $\|\cdot\|_{\Gamma^{-1}}$*; namely, for all* $\boldsymbol{u}$, $\boldsymbol{v}$

$$\begin{aligned}\|\operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}) - \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{v})\|^2_{\Gamma^{-1}} &\le \langle \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}) - \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{v}), \Gamma^{-1}(\boldsymbol{u}-\boldsymbol{v})\rangle \\ &\le \|\boldsymbol{u}-\boldsymbol{v}\|^2_{\Gamma^{-1}}; \qquad (8.43)\end{aligned}$$

(ii) the Moreau envelope $G^{\Gamma^{-1}}$ *is differentiable:* $\nabla G^{\Gamma^{-1}} = \Gamma^{-1}(\mathrm{id} - \operatorname{prox}_G^{\Gamma^{-1}})$*;*

(iii) for every $\boldsymbol{x} \in \mathbb{R}^{\sum_i n_i}$ *it holds that*

$$\operatorname{dist}(0, \partial\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x})) \le \tfrac{N + \max_i\{\gamma_i L_{f_i}\}}{N \min_i\{\sqrt{\gamma_i}\}} \|\boldsymbol{x} - \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x})\|_{\Gamma^{-1}};$$

(iv) $\mathrm{T}_\Gamma^{\mathrm{FB}}$ *is* $L_{\mathbf{T}}$*-Lipschitz continuous in the metric* $\|\cdot\|_{\Gamma^{-1}}$ *for some* $L_{\mathbf{T}} \ge 0$*; if in addition* f_i *is* μ_{f_i}*-strongly convex,* $i \in [N]$*, then* $L_{\mathbf{T}} \le 1 - \delta$ *for* $\delta = \frac{1}{N}\min_{i\in[N]}\{\gamma_i \mu_{f_i}\}$.

Proof.

♠ 8.22*(i)* and 8.22*(ii)* See [13, Prop.s 12.28 and 12.30].

♠ 8.22*(iii)* Let $D \subseteq \mathbb{R}^{\sum_i n_i}$ be the set of points at which ∇F is differentiable. From the chain rule of differentiation applied to the expression (8.8d) and using assertion 8.22*(ii)*, we have that Φ_Γ^{FB} is differentiable on D with gradient

$$\nabla\Phi_\Gamma^{\text{FB}}(\boldsymbol{x}) = \big[\mathrm{I} - \Gamma\nabla^2 F(\boldsymbol{x})\big]\Gamma^{-1}\big[\boldsymbol{x} - \mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{x})\big] \quad \forall \boldsymbol{x} \in D.$$

Since D is dense in $\mathbb{R}^{\sum_i n_i}$ owing to Lipschitz continuity of ∇F, we may invoke [144, Thm. 9.61] to infer that $\partial\Phi_\Gamma^{\text{FB}}(\boldsymbol{x})$ is nonempty for every $\boldsymbol{x} \in \mathbb{R}^{\sum_i n_i}$ and

$$\begin{aligned}\partial\Phi_\Gamma^{\text{FB}}(\boldsymbol{x}) \supseteq \partial_B\Phi_\Gamma^{\text{FB}}(\boldsymbol{x}) &= \big[\mathrm{I} - \Gamma\partial_B\nabla F(\boldsymbol{x})\big]\Gamma^{-1}\big[\boldsymbol{x} - \mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{x})\big] \\ &= \big[\Gamma^{-1} - \partial_B\nabla F(\boldsymbol{x})\big]\big[\boldsymbol{x} - \mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{x})\big],\end{aligned}$$

where ∂_B denotes the (set-valued) Bouligand differential [68, §7.1]. The claim now follows by observing that each element of $\partial_B\nabla f_i(x_i)$ has norm bounded by L_{f_i} and $\partial_B\nabla F(\boldsymbol{x}) = \frac{1}{N}\operatorname{blkdiag}(\partial_B\nabla f_1(x_1), \dots, \partial_B\nabla f_N(x_N))$.

♠ 8.22*(iv)* Lipschitz continuity follows from assertion 8.22*(i)* together with the fact that Lipschitz continuity is preserved by composition. Suppose now that f_i is μ_{f_i}-strongly convex, $i \in [N]$. By [127, Thm 2.1.12] for all $x_i, y_i \in \mathbb{R}^{n_i}$

$$\langle \nabla f_i(x_i) - \nabla f_i(y_i), x_i - y_i\rangle \geq \tfrac{\mu_{f_i} L_{f_i}}{\mu_{f_i}+L_{f_i}}\|x_i - y_i\|^2 + \tfrac{1}{\mu_{f_i}+L_{f_i}}\|\nabla f_i(x_i) - \nabla f_i(y_i)\|^2. \tag{8.44}$$

For the forward operator we have

$$\begin{aligned}&\|(\mathrm{id} - \tfrac{\gamma_i}{N}\nabla f_i)(x_i) - (\mathrm{id} - \tfrac{\gamma_i}{N}\nabla f_i)(y_i)\|^2 \\ &= \|x_i - y_i\|^2 + \tfrac{\gamma_i^2}{N^2}\|\nabla f_i(x_i) - \nabla f_i(y_i)\|^2 - \tfrac{2\gamma_i}{N}\langle x_i - y_i, \nabla f_i(x_i) - \nabla f_i(y_i)\rangle \\ &\overset{(8.44)}{\leq} \Big(1 - \tfrac{\gamma_i^2\mu_{f_i}L_{f_i}}{N^2}\Big)\|x_i - y_i\|^2 - \tfrac{\gamma_i}{N}\Big(2 - \tfrac{\gamma_i}{N}(\mu_{f_i} + L_{f_i})\Big)\langle \nabla f_i(x_i) - \nabla f_i(y_i), x_i - y_i\rangle \\ &\leq \Big(1 - \tfrac{\gamma_i^2\mu_{f_i}L_{f_i}}{N^2}\Big)\|x_i - y_i\|^2 - \tfrac{\gamma_i\mu_{f_i}}{N}\big(2 - \tfrac{\gamma_i}{N}(\mu_{f_i} + L_{f_i})\big)\|x_i - y_i\|^2 \\ &= \big(1 - \tfrac{\gamma_i\mu_{f_i}}{N}\big)^2\|x_i - y_i\|^2,\end{aligned}$$

where strong convexity and the fact that $\gamma_i < {}^N\!/_{L_{f_i}} \leq {}^{2N}\!/_{(\mu_{f_i}+L_{f_i})}$ was used in the second inequality. Multiplying by γ_i^{-1} and summing over i shows that $\mathrm{id} - \Gamma\nabla F$ is $(1-\delta)$-contractive in the metric $\|\cdot\|_{\Gamma^{-1}}$, and so is $\mathrm{T}_\Gamma^{\text{FB}} = \operatorname{prox}_G^{\Gamma^{-1}} \circ (\mathrm{id} - \Gamma\nabla F)$ as it follows from assertion 8.22*(i)*. □

The next result recaps an important property that the FBE inherits from the cost function Φ that is instrumental for establishing global convergence and asymptotic linear rates for the BC-Algorithm 8.1. The result falls as special case of [181, Thm. 5.2] after observing that

$$\Phi_{\Gamma}^{\mathrm{FB}}(\boldsymbol{x}) = \inf_{\boldsymbol{w}} \{\Phi(\boldsymbol{w}) + D_H(\boldsymbol{w}, \boldsymbol{x})\},$$

where $D_H(\boldsymbol{w}, \boldsymbol{x}) = H(\boldsymbol{w}) - H(\boldsymbol{x}) - \langle \nabla H(\boldsymbol{x}), \boldsymbol{w} - \boldsymbol{x} \rangle$ is the Bregman distance with kernel $H = \frac{1}{2}\|\cdot\|_{\Gamma^{-1}}^2 - F$.

Lemma 8.23 ([181, Thm. 5.2])**.** *Suppose that Assumption 8.I holds and for* $\gamma_i \in (0, {}^{N}/_{L_{f_i}})$*,* $i \in [N]$*, let* $\Gamma = \mathrm{blkdiag}(\gamma_1 \mathrm{I}_{n_1}, \ldots, \gamma_N \mathrm{I}_{n_N})$*. If* Φ *has the KL property with exponent* $\theta \in (0, 1)$ *(as is the case when* f_i *and* G *are semialgebraic), then so does* $\Phi_{\Gamma}^{\mathrm{FB}}$ *with exponent* $\max\{{}^{1}/_{2}, \theta\}$*.*

Conclusions

In this thesis a new three term splitting was proposed that solves monotone inclusions involving the sum of maximally monotone, cocoercive and monotone linear operators. Classical methods such as forward-backward and Douglas-Rachford splittings are special cases of the new splitting. This new splitting is leveraged for developing a unifying framework for primal-dual algorithms. It is shown that, by selecting different parameters, (extensions to) known algorithms as well as novel primal-dual methods are obtained. Moreover, owing to this unified analysis linear convergence is deduced for many primal-dual algorithms based on mild regularity conditions for the cost functions. Most notably, said algorithms achieve linear convergence when the cost functions are piecewise linear-quadratic.

A randomized block-coordinate primal-dual algorithm was developed that is particularly suitable for distributed applications. It is shown that it leads to asynchronous (randomized) distributed algorithms where both the updates and stepsizes only depend on local information. This is of great practical importance, not only because it allows *plug-and-play* implementations but also because it leads to larger stepsizes that depend on local Lipschitz constants rather than on the global one. Our computational tests have shown its superior performance in distributed model predictive control with dynamic coupling.

In the context of structured finite sum minimization over graphs, we have seen that solving the corresponding monotone inclusion with AFBA leads to novel distributed algorithms. In the context of general structured optimization over message-passing architectures, we have seen that bounded communication delays are tolerated provided that some strong convexity assumptions hold.

This thesis has also studied a block-coordinate (BC) variant of proximal gradient method for nonconvex problems. In the nonconvex setting neither the distance from the solution nor the cost function are viable Lyapunov function candidates. We have shown that *forward-backward envelope* is a particularly suitable candidate in this setting. A sure descent is established for the block-coordinate

scheme and convergence is analyzed under very general sampling strategies. When applied to regularized finite sum minimization, the proposed BC scheme leads to a generalization of the Finitio/MISO algorithm. As another prominent example, when applied to the regularized sharing problem it results in a novel incremental aggregated method.

Future directions

Several future research directions were discussed in each of the individual chapters. Here, we outline some other general ideas for future research.

The algorithms considered in this thesis comprise of simple fixed point iterations, employ constant stepsizes, scale well, and are suitable for distributed and parallel computations. However, like other first-order methods they struggle in reaching high precision, and are sensitive to ill conditioning, which can result in deterioration of performance. Therefore, an important future research direction is to develop synchronous and asynchronous distributed Newton-type algorithms.

Another important future research direction relates to incremental aggregated algorithms. While several such algorithms have been proposed over the years, their convergence analyses are typically rather complicated and/or are limited to the strongly convex case. In Chapter 8 a connection was established between (a class of) such algorithms and a block-coordinate proximal gradient method in the fully nonconvex case. However, there are several other incremental algorithms that are not covered by this analysis. In addition to this, in Chapter 8 it was shown that the forward-backward envelope is a particularly suitable Lyapunov function for the nonconvex BC proximal gradient algorithm. Based on this, ongoing research is investigating a linesearch strategy that extends that of [158] to the block-coordinate case. This in turn opens up the possibility of developing Newton-type incremental aggregated algorithms. Another important extension is to study the Bregman variants by exploiting Bregman forward-backward envelope [2] which would greatly expand the scope of the algorithms.

Bibliography

[1] Agarwal, A., and Duchi, J. C. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems 24*. 2011, pp. 873–881.

[2] Ahookhosh, M., Themelis, A., and Patrinos, P. A Bregman forward-backward linesearch algorithm for nonconvex composite optimization: superlinear convergence to nonisolated local minima. *arXiv:1905.11904* (2019).

[3] Allen-Zhu, Z., and Yuan, Y. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning* (2016), p. 1080–1089.

[4] Alotaibi, A., Combettes, P. L., and Shahzad, N. Solving coupled composite monotone inclusions by successive Fejér approximations of their Kuhn-Tucker set. *SIAM Journal on Optimization 24*, 4 (2014), 2076–2095.

[5] ApS, M. *The MOSEK optimization toolbox for MATLAB manual. Version 7.1 (Revision 63).*, 2015.

[6] Aragón Artacho, F. J., and Geoffroy, M. H. Characterization of metric regularity of subdifferentials. *Journal of Convex Analysis 15*, 2 (2008), 365–380.

[7] Aravkin, A. Y., Burke, J. V., and Pillonetto, G. Sparse/robust estimation and Kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory. *Journal of Machine Learning Research 14* (2013), 2689–2728.

[8] Arrow, K. J., Hurwicz, L., and Uzawa, H. *Studies in linear and non-linear programming.* Stanford University Press: Stanford, 1958.

[9] Attouch, H., and Bolte, J. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming 116*, 1-2 (2009), 5–16.

[10] Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research 35*, 2 (2010), 438–457.

[11] Attouch, H., Bolte, J., and Svaiter, B. F. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming 137*, 1 (2 2013), 91–129.

[12] Aytekin, A., Feyzmahdavian, H. R., and Johansson, M. Analysis and implementation of an asynchronous optimization algorithm for the parameter server. *arXiv:1610.05507* (2016).

[13] Bauschke, H. H., and Combettes, P. L. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics. Springer, 2017.

[14] Beck, A. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

[15] Beck, A., and Tetruashvili, L. On the convergence of block coordinate descent type methods. *SIAM journal on Optimization 23*, 4 (2013), 2037–2060.

[16] Bemporad, A., and Barcelli, D. Decentralized model predictive control. In *Networked control systems*. Springer, 2010, pp. 149–178.

[17] Bertsekas, D. P. Incremental proximal methods for large scale convex optimization. *Mathematical programming 129*, 2 (2011), 163–195.

[18] Bertsekas, D. P. *Convex Optimization Theory*. Athena Scientific, 2015.

[19] Bertsekas, D. P. *Nonlinear Programming*. Athena Scientific, 2016.

[20] Bertsekas, D. P., and Tsitsiklis, J. N. *Parallel and distributed computation: numerical methods*, vol. 23. Prentice-Hall, 1989.

[21] Bianchi, P., and Hachem, W. A primal-dual algorithm for distributed optimization. In *IEEE 53rd Annual Conference on Decision and Control (CDC)* (Dec 2014), pp. 4240–4245.

[22] Bianchi, P., Hachem, W., and Iutzeler, F. A coordinate descent primal-dual algorithm and application to distributed asynchronous optimization. *IEEE Transactions on Automatic Control 61*, 10 (2016), 2947–2957.

[23] Bolte, J., Daniilidis, A., and Lewis, A. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization 17*, 4 (2007), 1205–1223.

[24] Bolte, J., Daniilidis, A., Lewis, A., and Shiota, M. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization 18*, 2 (2007), 556–572.

[25] Bolte, J., Sabach, S., and Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming 146*, 1-2 (2014), 459–494.

[26] Boţ, R. I., Csetnek, E. R., and Hendrich, C. Recent developments on primal-dual splitting methods with applications to convex minimization. In *Mathematics Without Boundaries*. Springer, 2014, pp. 57–99.

[27] Boţ, R. I., and Hendrich, C. A Douglas-Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators. *SIAM Journal on Optimization 23*, 4 (2013), 2541–2565.

[28] Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. Randomized gossip algorithms. *IEEE Transactions on Information Theory 52*, 6 (2006), 2508–2530.

[29] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning 3*, 1 (2011), 1–122.

[30] Briceño-Arias, L. M. Forward-Douglas-Rachford splitting and forward-partial inverse method for solving monotone inclusions. *Optimization 64*, 5 (2015), 1239–1261.

[31] Briceño-Arias, L. M., and Combettes, P. L. A monotone + skew splitting model for composite monotone inclusions in duality. *SIAM Journal on Optimization 21*, 4 (2011), 1230–1250.

[32] Briceño-Arias, L. M., and Davis, D. Forward-backward-half forward algorithm for solving monotone inclusions. *SIAM Journal on Optimization 28*, 4 (2018), 2839–2871.

[33] Cai, X., Han, D., and Yuan, X. On the convergence of the direct extension of ADMM for three-block separable convex minimization models with one strongly convex function. *Computational Optimization and Applications 66*, 1 (2017), 39–73.

[34] Cannelli, L., Facchinei, F., Kungurtsev, V., and Scutari, G. Asynchronous parallel algorithms for nonconvex optimization. *Mathematical Programming* (Jun 2019).

[35] Cegielski, A. *Iterative methods for fixed point problems in Hilbert spaces*, vol. 2057. Springer, 2012.

[36] Cevher, V., and Vũ, B. C. A reflected forward-backward splitting method for monotone inclusions involving Lipschitzian operators. *Set-valued and Variational analysis* (Mar 2020).

[37] Chambolle, A. An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision 20*, 1-2 (2004), 89–97.

[38] Chambolle, A., and Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision 40*, 1 (2011), 120–145.

[39] Chambolle, A., and Pock, T. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming 159*, 1 (Sep 2016), 253–287.

[40] Chang, C.-C., and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST) 2* (2011), 1–27.

[41] Chang, T., Hong, M., Liao, W., and Wang, X. Asynchronous distributed ADMM for large-scale optimization—part i: Algorithm and convergence analysis. *IEEE Transactions on Signal Processing 64*, 12 (June 2016), 3118–3130.

[42] Chen, C., He, B., Ye, Y., and Yuan, X. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming 155*, 1-2 (2016), 57–79.

[43] Chen, G., and Rockafellar, R. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization 7*, 2 (1997), 421–444.

[44] Chouzenoux, E., Pesquet, J.-C., and Repetti, A. A block coordinate variable metric forward–backward algorithm. *Journal of Global Optimization 66*, 3 (11 2016), 457–485.

[45] Chow, Y. T., Wu, T., and Yin, W. Cyclic coordinate-update algorithms for fixed-point problems: Analysis and applications. *SIAM Journal on Scientific Computing 39*, 4 (2017), A1280–A1300.

[46] Combettes, P. L. Quasi-Fejérian analysis of some optimization algorithms. *Studies in Computational Mathematics 8* (2001), 115–152.

[47] Combettes, P. L., Condat, L., Pesquet, J.-C., and Vũ, B. C. A forward-backward view of some primal-dual optimization methods in image recovery. In *Image Processing (ICIP), 2014 IEEE International Conference on* (2014), IEEE, pp. 4141–4145.

[48] Combettes, P. L., and Eckstein, J. Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions. *Mathematical Programming* (Jul 2016).

[49] Combettes, P. L., and Pesquet, J.-C. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*. Springer New York, 2011, pp. 185–212.

[50] Combettes, P. L., and Pesquet, J.-C. Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. *Set-Valued and variational analysis 20*, 2 (2012), 307–330.

[51] Combettes, P. L., and Pesquet, J.-C. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization 25*, 2 (2015), 1221–1248.

[52] Combettes, P. L., and Yamada, I. Compositions and convex combinations of averaged nonexpansive operators. *Journal of Mathematical Analysis and Applications 425*, 1 (2015), 55–70.

[53] Condat, L. A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications 158*, 2 (2013), 460–479.

[54] Condat, L. A generic proximal algorithm for convex optimization-application to total variation minimization. *Signal Processing Letters, IEEE 21*, 8 (2014), 985–989.

[55] Davis, D. Convergence rate analysis of primal-dual splitting schemes. *SIAM Journal on Optimization 25*, 3 (2015), 1912–1943.

[56] Davis, D. Smart: The stochastic monotone aggregated root-finding algorithm. *arXiv:1601.00698* (2016).

[57] Davis, D., and Yin, W. *Convergence Rate Analysis of Several Splitting Schemes*. Springer International Publishing, Cham, 2016, pp. 115–163.

[58] Davis, D., and Yin, W. A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis 25*, 4 (Dec 2017), 829–858.

[59] Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems* (2014), pp. 1646–1654.

[60] Defazio, A., and Domke, J. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning* (2014), pp. 1125–1133.

[61] Dontchev, A. L., and Rockafellar, R. T. Implicit functions and solution mappings. *Springer Monographs in Mathematics. Springer 208* (2009).

[62] Drori, Y., Sabach, S., and Teboulle, M. A simple algorithm for a class of nonsmooth convex-concave saddle-point problems. *Operations Research Letters 43*, 2 (2015), 209–214.

[63] Drusvyatskiy, D., and Lewis, A. S. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research 43*, 3 (2018), 919–948.

[64] Duchi, J. C., Agarwal, A., and Wainwright, M. J. Dual averaging for distributed optimization: convergence analysis and network scaling. *IEEE Transactions on Automatic control 57*, 3 (2012), 592–606.

[65] Dunbar, W. B., and Caveney, D. S. Distributed receding horizon control of vehicle platoons: Stability and string stability. *IEEE Transactions on Automatic Control 57*, 3 (2012), 620–633.

[66] Eckstein, J., and Bertsekas, D. P. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming 55*, 1-3 (1992), 293–318.

[67] Esser, E., Zhang, X., and Chan, T. F. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences 3*, 4 (2010), 1015–1046.

[68] Facchinei, F., and Pang, J.-S. *Finite-dimensional variational inequalities and complementarity problems*, vol. II. Springer, 2003.

[69] Farokhi, F., Shames, I., and Johansson, K. H. Distributed MPC via dual decomposition and alternating direction method of multipliers. In *Distributed model predictive control made easy*. Springer, 2014, pp. 115–131.

[70] Fercoq, O., and Bianchi, P. A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions. *SIAM Journal on Optimization 29*, 1 (2019), 100–134.

[71] Feyzmahdavian, H. R., Aytekin, A., and Johansson, M. An asynchronous mini-batch algorithm for regularized stochastic optimization. *IEEE Transactions on Automatic Control 61*, 12 (2016), 3740–3754.

[72] Frankel, P., Garrigos, G., and Peypouquet, J. Splitting methods with variable metric for Kurdyka-łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications 165*, 3 (2015), 874–900.

[73] Freris, N., Graham, S., and Kumar, P. Fundamental limits on synchronizing clocks over networks. *IEEE Transactions on Automatic Control 56*, 6 (2011), 1352–1364.

[74] Freris, N., and Zouzias, A. Fast distributed smoothing of relative measurements. In *51st IEEE Conference on Decision and Control (CDC)* (2012), pp. 1411–1416.

[75] Fukushima, M., and Mine, H. A generalized proximal point algorithm for certain non-convex minimization problems. *International Journal of Systems Science 12*, 8 (1981), 989–1000.

[76] Giselsson, P. A generalized distributed accelerated gradient method for distributed model predictive control with iteration complexity bounds. In *American Control Conference (ACC)* (2013), pp. 327–333.

[77] Giselsson, P. Improved dual decomposition for distributed model predictive control. *IFAC Proceedings Volumes 47*, 3 (2014), 1203–1209.

[78] Giselsson, P. Nonlinear forward-backward splitting with projection correction. *arXiv:1908.07449* (2019).

[79] Giselsson, P., Doan, M. D., Keviczky, T., Schutter, B. D., and Rantzer, A. Accelerated gradient methods and dual decomposition in distributed model predictive control. *Automatica 49*, 3 (2013), 829–833.

[80] Hale, M. T., Nedić, A., and Egerstedt, M. Asynchronous multiagent primal-dual optimization. *IEEE Transactions on Automatic Control 62*, 9 (Sept 2017), 4421–4435.

[81] Han, D., and Yuan, X. A note on the alternating direction method of multipliers. *Journal of Optimization Theory and Applications 155*, 1 (2012), 227–238.

[82] Hanzely, F., Mishchenko, K., and Richtarik, P. SEGA: Variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems 31*. 2018, pp. 2082–2093.

[83] Hastie, T., Friedman, J., and Tibshirani, R. *The Elements of Statistical Learning*. Springer New York, 2001.

[84] He, B., and Yuan, X. Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective. *SIAM Journal on Imaging Sciences 5*, 1 (2012), 119–149.

[85] Hong, M., Wang, X., Razaviyayn, M., and Luo, Z.-Q. Iteration complexity analysis of block coordinate descent methods. *Mathematical Programming 163*, 1-2 (2017), 85–114.

[86] Hou, X., Xiao, Y., Cai, J., Hu, J., and Braun, J. E. Distributed model predictive control via proximal Jacobian ADMM for building control applications. In *American Control Conference (ACC)* (May 2017), pp. 37–43.

[87] Hou, Y., Song, I., Min, H.-K., and Park, C. H. Complexity-reduced scheme for feature extraction with linear discriminant analysis. *IEEE transactions on neural networks and learning systems 23*, 6 (2012), 1003–1009.

[88] Iutzeler, F., Bianchi, P., Ciblat, P., and Hachem, W. Asynchronous distributed optimization using a randomized alternating direction method of multipliers. In *52nd IEEE Conference on Decision and Control (CDC)* (2013), pp. 3671–3676.

[89] Jadbabaie, A., Lin, J., and Morse, A. S. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control 48*, 6 (2003), 988–1001.

[90] Johansson, B., Rabi, M., and Johansson, M. A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization 20*, 3 (2010), 1157–1170.

[91] Johnson, R., and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26* (2013), pp. 315–323.

[92] Komodakis, N., and Pesquet, J. C. Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems. *IEEE Signal Processing Magazine 32*, 6 (Nov 2015), 31–54.

[93] Kurdyka, K. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier 48*, 3 (1998), 769–783.

[94] Latafat, P., Bemporad, A., and Patrinos, P. Plug and play distributed model predictive control with dynamic coupling: A randomized primal-dual proximal algorithm. In *European Control Conference (ECC)* (June 2018), pp. 1160–1165.

[95] Latafat, P., Freris, N. M., and Patrinos, P. A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. *IEEE Transactions on Automatic Control 64*, 10 (10 2019), 4050–4065.

[96] Latafat, P., and Patrinos, P. Asymmetric forward–backward–adjoint splitting for solving monotone inclusions involving three operators. *Computational Optimization and Applications 68*, 1 (Sep 2017), 57–93.

[97] Latafat, P., and Patrinos, P. Multi-agent structured optimization over message-passing architectures with bounded communication delays. In *2018 IEEE Conference on Decision and Control (CDC)* (Dec 2018), pp. 1688–1693.

[98] Latafat, P., and Patrinos, P. Primal-dual proximal algorithms for structured convex optimization: A unifying framework. In *Large-Scale and Distributed Optimization*, P. Giselsson and A. Rantzer, Eds. Springer International Publishing, 2018, pp. 97–120.

[99] Latafat, P., and Patrinos, P. Primal-dual algorithms for multi-agent structured optimization over message-passing architectures with bounded communication delays. *arXiv:1809.07199* (2019).

[100] Latafat, P., Stella, L., and Patrinos, P. New primal-dual proximal algorithm for distributed optimization. In *55th IEEE Conference on Decision and Control (CDC)* (Dec 2016), pp. 1959–1964.

[101] Latafat, P., Themelis, A., and Patrinos, P. Block-coordinate and incremental aggregated proximal gradient methods for nonsmooth nonconvex problems. *arXiv:1906.10053* (2019).

[102] Li, G., and Pong, T. K. Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. *Mathematical Programming 159*, 1 (9 2016), 371–401.

[103] Li, M., Sun, D., and Toh, K.-C. A convergent 3-block semi-proximal ADMM for convex minimization problems with one strongly convex block. *Asia-Pacific Journal of Operational Research* (2015), 1550024.

[104] Liang, J., Fadili, J., and Peyré, G. Convergence rates with inexact non-expansive operators. *Mathematical Programming 159*, 1 (Sep 2016), 403–434.

[105] Lin, P., Ren, W., and Song, Y. Distributed multi-agent optimization subject to nonidentical constraints and communication delays. *Automatica 65* (2016), 120 – 131.

[106] Lin, Q., Lu, Z., and Xiao, L. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization 25*, 4 (2015), 2244–2273.

[107] Lin, T., Ma, S., and Zhang, S. Iteration complexity analysis of multi-block ADMM for a family of convex minimization without strong convexity. *Journal of Scientific Computing 69*, 1 (2016), 52–81.

[108] Lions, P.-L., and Mercier, B. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis 16*, 6 (1979), 964–979.

[109] Liu, J., and Wright, S. J. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization 25*, 1 (2015), 351–376.

[110] Lobel, I., Ozdaglar, A., and Feijer, D. Distributed multi-agent optimization with state-dependent communication. *Mathematical programming 129*, 2 (2011), 255–284.

[111] Łojasiewicz, S. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles* (1963), 87–89.

[112] Łojasiewicz, S. Sur la géométrie semi- et sous- analytique. *Annales de l'institut Fourier 43*, 5 (1993), 1575–1595.

[113] Luke, D. R., and Shefi, R. A globally linearly convergent method for pointwise quadratically supportable convex–concave saddle point problems. *Journal of Mathematical Analysis and Applications 457*, 2 (2018), 1568 – 1590. Special Issue on Convex Analysis and Optimization: New Trends in Theory and Applications.

[114] Mairal, J. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization 25*, 2 (2015), 829–855.

[115] Makhdoumi, A., and Ozdaglar, A. Broadcast-based distributed alternating direction method of multipliers. In *52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (2014), IEEE, pp. 270–277.

[116] Malitsky, Y., and Tam, M. K. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization 30*, 2 (2020), 1451–1472.

[117] Mokhtari, A., Gürbüzbalaban, M., and Ribeiro, A. Surpassing gradient descent provably: A cyclic incremental method with linear convergence rate. *SIAM Journal on Optimization 28*, 2 (2018), 1420–1447.

[118] Moreau, J.-J. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France 93* (1965), 273–299.

[119] Necoara, I. Random coordinate descent algorithms for multi-agent convex optimization over networks. *IEEE Transactions on Automatic Control 58*, 8 (8 2013), 2001–2012.

[120] Necoara, I., and Patrascu, A. A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints. *Computational Optimization and Applications 57*, 2 (3 2014), 307–337.

[121] Nedić, A., Bertsekas, D., and Borkar, V. Distributed asynchronous incremental subgradient methods. In *Inherently Parallel Algorithms in Feasibility and Optimization and their Applications*, D. Butnariu, Y. Censor, and S. Reich, Eds., vol. 8 of *Studies in Computational Mathematics*. Elsevier, 2001, pp. 381 – 407.

[122] Nedić, A., and Ozdaglar, A. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control 54*, 1 (2009), 48–61.

[123] Nedić, A., and Ozdaglar, A. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications 142*, 1 (Jul 2009), 205–228.

[124] Negenborn, R., Schutter, B. D., and Hellendoorn, J. Multi-agent model predictive control for transportation networks: Serial versus parallel schemes. *Engineering Applications of Artificial Intelligence 21*, 3 (apr 2008), 353–366.

[125] Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization 22*, 2 (2012), 341–362.

[126] Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical Programming 140*, 1 (Aug 2013), 125–161.

[127] Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer Science & Business Media, 2013.

[128] Ochs, P., Chen, Y., Brox, T., and Pock, T. iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences 7*, 2 (2014), 1388–1419.

[129] O'Donoghue, B., Stathopoulos, G., and Boyd, S. A splitting method for optimal control. *IEEE Transactions on Control Systems Technology 21*, 6 (Nov 2013), 2432–2442.

[130] Parikh, N., and Boyd, S. Proximal algorithms. *Foundations and Trends in optimization 1* (2013), 123–231.

[131] Parikh, N., and Boyd, S. Block splitting for distributed optimization. *Mathematical Programming Computation 6*, 1 (2014), 77–102.

[132] Patrinos, P., and Bemporad, A. Proximal Newton methods for convex composite optimization. In *52nd IEEE Conference on Decision and Control* (12 2013), pp. 2358–2363.

[133] Peng, Z., Xu, Y., Yan, M., and Yin, W. ARock: An algorithmic framework for asynchronous parallel coordinate updates. *SIAM Journal on Scientific Computing 38*, 5 (2016), A2851–A2879.

[134] Pesquet, J.-C., and Repetti, A. A class of randomized primal-dual algorithms for distributed optimization. *Journal of Nonlinear and Convex Analysis 16*, 12 (2015), 2453–2490.

[135] Qian, X., Sailanbayev, A., Mishchenko, K., and Richtárik, P. MISO is making a comeback with better proofs and rates. *arXiv:1906.01474* (2019).

[136] Raffard, R. L., Tomlin, C. J., and Boyd, S. P. Distributed optimization for cooperative agents: application to formation flight. In *2004 43rd IEEE Conference on Decision and Control (CDC)* (Dec 2004), vol. 3, pp. 2453–2459.

[137] Reddi, S. J., Hefny, A., Sra, S., Poczos, B., and Smola, A. J. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning* (2016), pp. 314–323.

[138] Reddi, S. J., Sra, S., Poczos, B., and Smola, A. J. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems* (2016), pp. 1145–1153.

[139] Richtárik, P., and Takáč, M. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming 144*, 1-2 (2014), 1–38.

[140] Rieger, J., and Tam, M. K. Backward-forward-reflected-backward splitting for three operator monotone inclusions. *Applied Mathematics and Computation 381* (2020), 125248.

[141] Riverso, S., Farina, M., and Ferrari-Trecate, G. Plug-and-play decentralized model predictive control for linear systems. *IEEE Transactions on Automatic Control 58*, 10 (2013), 2608–2614.

[142] Robbins, H., and Siegmund, D. A convergence theorem for non negative almost supermartingales and some applications. In *Herbert Robbins Selected Papers*. Springer, 1985, pp. 111–135.

[143] Rockafellar, R. T. *Convex analysis*. Princeton University Press, 1970.

[144] Rockafellar, R. T., and Wets, R. J.-B. *Variational analysis*, vol. 317. Springer Science & Business Media, 2009.

[145] Rudin, L. I., Osher, S., and Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena 60*, 1 (1992), 259 – 268.

[146] Ryu, E. K., and Vũ, B. C. Finding the forward-Douglas–Rachford-forward method. *Journal of Optimization Theory and Applications 184*, 3 (Mar 2020), 858–876.

[147] Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming 162*, 1 (3 2017), 83–112.

[148] Schouwenaars, T., How, J., and Feron, E. Decentralized cooperative trajectory planning of multiple aircraft with hard safety guarantees. In *AIAA Guidance, Navigation, and Control Conference and Exhibit* (2004), pp. 1–14.

[149] Shalev-Shwartz, S., and Zhang, T. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research 14*, Feb (2013), 567–599.

[150] Shi, W., Ling, Q., Wu, G., and Yin, W. A proximal gradient algorithm for decentralized composite optimization. *IEEE Transactions on Signal Processing 63*, 22 (Nov 2015), 6013–6023.

[151] Solodov, M. V., and Tseng, P. Modified projection-type methods for monotone variational inequalities. *SIAM Journal on Control and Optimization 34*, 5 (1996), 1814–1830.

[152] Sra, S., Nowozin, S., and Wright, S. J. *Optimization for machine learning*. Mit Press, 2012.

[153] Teixeira, A., Ghadimi, E., Shames, I., Sandberg, H., and Johansson, M. Optimal scaling of the ADMM algorithm for distributed quadratic programming. In *IEEE 52nd Annual Conference on Decision and Control (CDC)* (2013), pp. 6868–6873.

[154] Terelius, H., Topcu, U., and Murray, R. M. Decentralized multi-agent optimization via dual decomposition. *IFAC Proceedings Volumes 44*, 1 (2011), 11245 – 11251. 18th IFAC World Congress.

[155] Themelis, A. *Proximal Algorithms for Structured Nonconvex Optimization*. PhD thesis, KU Leuven, 12 2018.

[156] Themelis, A., Ahookhosh, M., and Patrinos, P. On the acceleration of forward-backward splitting via an inexact Newton method. In *Splitting Algorithms, Modern Operator Theory, and Applications*, H. H. Bauschke, R. S. Burachik, and D. R. Luke, Eds. Springer International Publishing, Cham, 2019, pp. 363–412.

[157] Themelis, A., and Patrinos, P. Supermann: A superlinearly convergent algorithm for finding fixed points of nonexpansive operators. *IEEE Transactions on Automatic Control 64*, 12 (Dec 2019), 4875–4890.

[158] Themelis, A., Stella, L., and Patrinos, P. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization 28*, 3 (2018), 2274–2303.

[159] Trnka, P., Pekař, J., and Havlena, V. Application of distributed MPC to Barcelona water distribution network. *IFAC Proceedings Volumes 44*, 1 (2011), 9139 – 9144.

[160] Tseng, P. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM Journal on Control and Optimization 29*, 1 (1991), 119–138.

[161] Tseng, P. On the rate of convergence of a partially asynchronous gradient projection algorithm. *SIAM Journal on Optimization 1*, 4 (1991), 603–619.

[162] Tseng, P. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization 38*, 2 (2000), 431–446.

[163] Tseng, P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications 109*, 3 (2001), 475–494.

[164] Tseng, P., and Bertsekas, D. P. Relaxation methods for problems with strictly convex separable costs and linear constraints. *Mathematical Programming 38*, 3 (10 1987), 303–321.

[165] Tseng, P., and Yun, S. Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *Journal of Optimization Theory and Applications 140*, 3 (7 2008), 513–535.

[166] Tseng, P., and Yun, S. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming 117*, 1 (3 2009), 387–423.

[167] Tseng, P., and Yun, S. A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. *Computational Optimization and Applications 47*, 2 (10 2010), 179–206.

[168] Tsianos, K. I., and Rabbat, M. G. Distributed dual averaging for convex optimization under communication delays. In *2012 American Control Conference (ACC)* (June 2012), pp. 1067–1072.

[169] Tsitsiklis, J., Bertsekas, D., and Athans, M. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control 31*, 9 (1986), 803–812.

[170] Vũ, B. C. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics 38*, 3 (2013), 667–681.

[171] Wakasa, Y., Arakawa, M., Tanaka, K., and Akashi, T. Decentralized model predictive control via dual decomposition. In *47th IEEE Conference on Decision and Control (CDC)* (Dec 2008), pp. 381–386.

[172] Wang, H., Liao, X., Huang, T., and Li, C. Cooperative distributed optimization in multiagent networks with delays. *IEEE Transactions on Systems, Man, and Cybernetics: Systems 45*, 2 (Feb 2015), 363–369.

[173] Wang, Z., and Ong, C. J. Distributed model predictive control of linear discrete-time systems with local and global constraints. *Automatica 81* (2017), 184 – 195.

[174] Wei, E., and Ozdaglar, A. Distributed alternating direction method of multipliers. In *IEEE 51st Annual Conference on Decision and Control (CDC)* (2012), pp. 5445–5450.

[175] Wei, E., and Ozdaglar, A. On the $O(1/k)$ convergence of asynchronous distributed alternating direction method of multipliers. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE* (2013), pp. 551–554.

[176] Wu, T., Yuan, K., Ling, Q., Yin, W., and Sayed, A. H. Decentralized consensus optimization with asynchrony and delays. *IEEE Transactions on Signal and Information Processing over Networks 4*, 2 (June 2018), 293–307.

[177] Xiao, L., and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization 24*, 4 (2014), 2057–2075.

[178] Xu, Y., and Yin, W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences 6*, 3 (2013), 1758–1789.

[179] Xu, Y., and Yin, W. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *Journal of Scientific Computing 72*, 2 (8 2017), 700–734.

[180] Yan, M. A new primal–dual algorithm for minimizing the sum of three functions with a linear operator. *Journal of Scientific Computing 76*, 3 (2018), 1698–1717.

[181] Yu, P., Li, G., and Pong, T. K. Deducing Kurdyka-Łojasiewicz exponent via inf-projection. *arXiv:1902.03635* (2019).

[182] Zhang, G., and Heusdens, R. Bi-alternating direction method of multipliers over graphs. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015), pp. 3571–3575.

[183] Zhang, R., and Kwok, J. T. Asynchronous distributed ADMM for consensus optimization. In *Proceedings of the 31st International Conference on International Conference on Machine Learning* (2014), pp. 1701–1709.

[184] Zhou, Y., Liang, Y., Yu, Y., Dai, W., and Xing, E. P. Distributed proximal gradient algorithm for partially asynchronous computer clusters. *Journal of Machine Learning Research 19*, 19 (2018), 1–32.

[185] Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. 1-norm support vector machines. *Advances in neural information processing systems 16*, 1 (2004), 49–56.

[186] Zouzias, A., and Freris, N. Randomized gossip algorithms for solving Laplacian systems. In *European Control Conference (ECC)* (July 2015), pp. 1920–1925.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
STADIUS
Kasteelpark Arenberg 10, bus 2446
B-3001 Leuven
puya.latafat@kuleuven.be
https://www.esat.kuleuven.be/stadius/