UNIVERSITY OF CALIFORNIA, SAN DIEGO

Distributed Algorithms for Convex Optimization: noisy channels, online computation, nuclear norm regularization, and separable constraints

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Engineering Sciences (Mechanical Engineering)

by

David Mateos-Núñez

Committee in charge:

Professor Jorge Cortés, Chair
Professor Philip E. Gill
Professor Tara Javidi
Professor Miroslav Krstić
Professor Maurício de Oliveira

2016

The dissertation of David Mateos-Núñez is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2016

DEDICATION

To my parents,

María Luz Núñez Díaz and Carlos Mateos Mateos

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Preface

Algorithms are detailed sequences of instructions, and are useful to describe and automate processes. Hence, an expression of our Language and Culture. With the help of Science and Engineering, they can infuse behavior in inanimate matter, our computers. Companies, indeed, are harnessing these advances to profit from products and services very efficiently, and still, with the unprecedented access to knowledge and technology by individuals, a fundamental question arises of how we construct organizations, companies, and institutions that preserve and create opportunity, as members of a complex system that is far from the socio-economic equilibrium both geographically and in time. John M. Culkin wrote, rephrasing Marshall McLuhan, that "we shape our tools and then our tools shape us" [Cul67], including the conditions that surround us: our buildings, our media, our education... It is within this feedback loop where we *hope* that the Arts will inform our pursue, Politics will push the collective wisdom, and Companies will try to realize their vision. It is a complex way of engaging with the future, each other, and the environment, and is an experiment never done before. Poised to follow the technological race to some "inevitable" conclusion, we nonetheless hope to load *our* values in the infrastructures of our society, but the challenge remains that in all the negotiations the currency will be the collective perception of both our potential and also the *way* we can realize that potential from the current state. Although the unequal share of the machines and the means to capture resources influence our participation in defining currencies and meaning, there must be hope to engage the best of ourselves, protect the things we love, and use our resourcefulness. Our actions will echo in the collective narrative.

ACKNOWLEDGEMENTS

For those traveling with me
and for you, dear reader,
I say first the "why"
and then the "who".

(You are about to read
the one thing I ask you
to take from me.)

For four years,
society backed me up,
letting me and my peers
be the only judge.

People seldom are trusted,
supported,
and respected like this.

The outcome of this gift,
I need to justify.
The aim for sure I'll miss,

but I will give a try.

How do you take a gift
and make it grow?
How do we echo
the efforts of all?

Dear reader,
think of this;
the answer to society
is constantly at risk.

Wanting to say more,
I should stop,
‘cause the important ritual
can be ignored.

But for weakness I go on
and cheers I share;
since our happiness we owe,
be idle... don’t you dare!

To the ones who work hard
for necessity or convenience;
because you play your card,
you build our resilience.

So I am glad to say thanks
to you pupils of Reason
for working in peace
and honoring people.

March 4, 2015.

To the departments at UCSD

Special thanks I want to give to Jorge, for helping me manage the rigors of this journey and for respecting so much this profession. After nearly 200 weeks of constructive attitude on his side, I have learned many lessons that time will put into perspective, specially the organization and way of working that favors the cumulative effort. He also listened to more concerns that is generally accustomed and made sure I could continue my search through all the obstacles.

Seemingly a coincidence, I have known the people in this committee, Jorge, Philip, Tara, Maurício and Miroslav, for the quality of their teaching, which I celebrate and love for very particular reasons. I also enjoyed the courses taught by Bob Bitmead, to whom we owe the personality of Kalman Filters, and by Jiawang Nie, who has the tenderest inclination to student participation. I was lucky to be a teaching assistant with Maurício, twice! I will always treasure that. Bahman Gharesifard, our post-doc at the time I arrived, gave me a starting point for my research and a warm welcome to the profession. Mike Ouimet encouraged my scientific distractions across the board, and one day, after some insisting, finally partnered with Beth to take us to sail in Catamaran. Dean Richert, and later Ashish Cherukuri, gave me a constant example of discipline and vocation in research. Thanks to the company of Dean in the lab, day after day, I slowly got convinced that *every obstacle tells you the way around it after enough effort.* Thanks to this I was able to pull off the first two papers during that difficult initiation period. A similar merit goes to Letty Malouf, my landlady for two years, who became a family member and a friend to me and listened untiringly to my stories from Spain. Hamed Foroush was a brother at that time and María Barrantes came to complete our unlikely gang with her amazing humor. With them, I saw another life beyond the horizons of the campus. I also wish to thank Daniele Cavaglieri and Minyi Ji for creating the other home for the rest of us international students to celebrate Thanksgiving and Easter. My regards to César, Andrés, Aman, Minu, Lucas, Eduardo, Johanna, Robert, Azad, Farinaz, Solmaz, and Shuxia, for sharing a bit of their lives and interests along the way. The Southern California Control Workshops gave me the adrenaline of the one minute talk in Caltech, the discovery of Santa Barbara, and the beginning of a friendship with Justin Pearson and David A. Copp. Another appreciated ritual was our Friday paper presentation promoted by Jorge and Sonia Martínez and made possible through the participation of the people in our group. The University of California San Diego afforded all the conference trips to present the results of this thesis, and J.V. Agnew, Marina Robenko, Linda McKamey, and others in the humorously decorated offices did a great job behind the scenes.

To the journals and research grants

This thesis is the fruit of many patient iterations of improvement and refinement to add a solid brick on top of a formidable body of previous work in optimization, dynamical systems and networks. The following is an account of the journals where our results have been published.

Chapter 3 is taken from the journal paper “Noise-to-state exponentially stable distributed convex optimization on weight-balanced digraphs”, by D. Mateos-Núñez and J. Cortés, submitted for publication in *SIAM Journal on Control and Optimization.*

Chapter 4 is taken from the journal paper “Distributed online convex optimization over jointly connected digraphs,” by D. Mateos-Núñez and J. Cortés, published in *IEEE Transactions on Network Science and Engineering 1 (1) (2014), 23-37.*

Chapter 5 is taken from the journal paper “Distributed saddle-point subgradient algorithms with Laplacian averaging,” by D. Mateos-Núñez, J. Cortés, submitted for publication in *IEEE Transactions on Automatic Control.*

Chapter 6 is an extended version of the conference paper “Distributed optimization for multi-task learning via nuclear-norm approximation,” by D. Mateos-Núñez, J. Cortés, presented in *IFAC Workshop on Distributed Estimation and Control in Networked Systems, Philadelphia, Pennsylvania, USA, 2015.*

Chapter 7 is taken from the journal paper “pth moment noise-to-state stability of stochastic differential equations with persistent noise,” by D. Mateos-Núñez, J. Cortés, published in *SIAM Journal on Control and Optimization 52 (4) (2014), 2399-2421.* A revised version can be found in *arXiv:1501.05008v1.*

The authors would like to thank the anonymous reviewers from these journals and conference proceedings because their comments helped improve our manuscripts and thus the presentation of this thesis. This research was supported by NSF awards CMMI-0908508 and CMMI-1300272.

To the welcome encounters

Research is by definition always an incomplete journey, and although I yearned for the mystic mountain, the places beyond the map defy any personal romantic vision. We belong to a collective, iterative search, wisely constrained by our communities to impose the necessary alliance of necessity and joint courage. At my own scale, I sought knowledge everywhere, and found people who encouraged and mentored me. Borja Ibarz accompanied me through the landmarks of life in San Diego with wit and wisdom, being loyal to his friends across the years and bringing more company to the rest of us, like Romi Thakur and Dhanya Gopal, who became the fellow explorers of the natural beauty of California. Jean Baptiste Passot shared with us his garden, his cooking, and the magical spaces that he creates. The same goes for Eileen Motta, whose resourcefulness made us enjoy the most pleasurable gatherings. Juan Nogales filled the silent breaks of Summer afternoons with entertaining explanations about metabolic networks. Alex Breslow and Min Zhang brought me back the lost habit of jogging while talking about books and other curiosities. Israel Lopez Coto became my bouldering partner, at which time we also met Melisa, Andreas, Helena, Jacob, Martin, Tim, and Will, who got us fully immersed in the climbing experience. To Jacob Huffman I owe a lasting enthusiasm in the work led by Jeff Hawkins on Hierarchical Temporal Memories by sharing his book “On Intelligence”. Along these lines, Bruno Pedroni was always eager to explain concisely his research on Neuromorphic computing. Yicong Ma and Yanjin Kuang invited me to paddle in their team for the Dragon

Boat Race, and thanks to Poy Supanee, I was able to make it a second year. Bruno Buchberger, from the Research Institute for Symbolic Computation in Austria, answered my questions about his Theorema Project during a period where I was deeply interested in symbolic search engines and collaborative knowledge bases assisted by machine learning algorithms. TEDxUCSD brought me the pleasure of hugging my admired Science-Fiction writer David Brin, some of his books I had read before I ever dreamed to come to the city where he studied and lives. Fengzhong Li contacted me from Shandong University, in China, to inform me that my proof about noise dissipative Lyapunov functions appeared to be incomplete. His careful reading and kindness produced a chain of emails where he criticized my attempts until the proof looked satisfactory. I cherish that encounter for its many implications ranging from the amazement of his reaching out, and my journey from nightmare to confrontation, to the confirmation that *a good reasoning is one that let others pinpoint where you are wrong.* Helen Henninger transformed the conference trip to Groningen into a tale trip of discovery. Karan Sikka became my mentor in kitchen rituals in my second home and shared many Indian teas and discussions on machine learning and optimization. Federico Pistono helped me transcend the *online* medium last Christmas by gifting me the personal satisfaction of translating into Spanish his book “A Tale of Two Futures”. During this task, I explored another mind with such engagement that I can confirm the insight of Marshall McLuhan that *the medium is the message.* Saket Navlakha, from the Salk Institute, engaged me in discussions about learning and the brain, inspiring me much during the last months. For the people whose company I enjoyed without ever meeting them, I would like to mention John Green for his History online crash courses, the sublime show “Cosmos” by Neil deGrasse Tyson, and also my late discovery of the science fiction writers Robert Heinlein, Stephen Baxter and Karl

Schroeder.

To the dancing communities

Dancing is a language for reconciliation that flows through the depths of the body. It also welcomes great teachers and makes lasting friends. For the ones in Alcalá de Henares, I wish to thank Alberto and Rodrigo Villarrubia, Esperanza Alarcón, Talía Gamarra, and Noemí Chao Palacios. For the people in San Diego, I am grateful to the voluntary organizers and instructors of the UCSD Argentine Tango club across the years for sharing their passion: Grigor Aslanyan, Ten Mendoza, Ivana Guarrasi, Erika Cheng, Reza Farsian, Eduardo Santana, Effie Quattrociocchi, Avi Jain, Rie Avinash, Trista Brophy, Viraj Deshpande, and Min Zhang. The same goes to the adorable people in the broader Tango scene, including the instructors and old friends Florentino Guizar, Isabelle Simonne, Linda Garwood, Amanda, Sandy, Richard, Serena, Keiko, Jittra, Irina, Mariana, and so many others whose close embrace I hope to keep warm for many years. I thank Vera Alverdi and Vincent for their unscripted brilliance teaching us the choreography and rekindling the long forgotten joys of team work when I needed them most. My love for the team members Alexia, Mario, Autumn, Gamo, Cinzia, Alex, Julie, Marcus, César, Yesenia, and Riz. Mariza Suárez gave me the sweetest year of the Ph.D as my cheerful and supportive partner, introducing me to West Coast Swing and planning the best trips. My love also goes to Scampy, my favorite dog in the World, for being so happy to be with us.

To the people in Spain

The farther in time, the more the experiences are entangled with the process of growing up. These people I can only thank enough during the most lucid of moments, when all the memories make their pass and years are felt in seconds. The friendship and hobbies shared with Javier Pérez Álvaro convinced me that I could continue my learning journey in Mathematics during a period where that goal appeared to be unreachable. Luckily enough, some time down the road I got to spend many hours doing so, and met Aurelio Labanda Alonso, María González García, and Coralina Hernández Trujillo, who contributed to create the atmosphere to better ourselves and overcome difficulties that makes everything memorable. The gatherings at the cafeteria with María Aguirre, Guillermo Rey, and other dear people, completed the blessing of being a student at a University, which all in all is a experience that we should strive to protect despite all its faults. For the privilege of learning Maths, my tenderest love to all the students and professors in the Department of Mathematics at University Autónoma de Madrid. From my experiences in High School and before, where sports and games, and books and new friends reach the prehistory of my journey, my special thanks to Enrique Valencia Gómez and Víctor González Martín for caring and sharing over the years. Wanting to acknowledge the teachers in Middle School and High School, my thoughts fly to Teresa, Ramón, Jesús, Carmen Hermida, Eva Aranzana, Ángel Barahona, and Concha Albertos, and others, the people who I am truly most identified with. I thank my neighbors in Alcalá de Henares and Robleda, the places where I grew up and still inspire me in ways that I am beginning to comprehend. My family was specially responsible to make me feel what it means to arrive home and to leave home. Aunts, uncles and cousins made life sweet with hugs and familiarity. My aunt and my grandmother where always there for me, keeping a house full of pretty and dear things, where we were comfortable and could make friends and play and have adventures. My admiration goes to my sister Eva, for fulfilling every task with dignity and beauty and for discovering still another passion in the sign language. To my brother Jesús, my respect for the many hours that he surprised us studying in the last years. My cousin Raquel used to be my attention rapture time and again with the things she did and said, and used to listen untiringly to everything that inspired me. My parents worked hard to make our entire life comfortable and helped us grow up with many good memories.

To the people who I cannot mention

The nicest memories are the less deserved thanks to our inheritance and the generosity of life awakening. I conclude saying thanks for a lasting legacy of pro-active passion in a poem that I think of as “Loyal to no end” started in March 19, 2015, that honors a poem of Borges brought up one day by Concha Albertos in High School.

(Ongoing) poem of the gifts.

For the enlightenment experiment
that a postman spread,
by the writer of San Diego,
with his big, big, head.

For the movie of Brave and Merida’s mind,
heir of Ulysses and Wallace combined,
subduing a castle with measured speech,

winning her allies in a warehouse of meat.

For the visible faces,
Diamandis and Musk,
rocketing Heinlein
above Earth's crust.

For the playground of *Consensus*,
where iterating the greeks,
with Lagrangians and Laplacians
we agree with greed.

For dogs
who crave outdoors,
and own a mind
that can be kind.

For Tartaglia, Cardano and Ferrari,
for playing well in their time,
because their work and passion
made History sublime.

For "how I met your mother"
and self-conscious living,
and the laughable things
like slaps in Thanksgiving.

For crying,
as someone said,
when there's more beauty
than we expect.

For Home and the sisters
intellectually hungry,
Asu and Angelia,
and this dear country.

For Wikipedia,
wanting to be born.
Never saw a child
that we needed more.

For the people in our team,
and the work and the lead
of Jorge and Sonia,
through these many weeks.

For the shoulders of giants,
like boulders of History
that we climb defiant
and sweating for mystery.

For the eloquent truth
of working machines;
if I give you my youth,
redeem then my genes.

For dwelling clear thoughts
between completion and loss.
Before surrender to love,
we struggle, why, my gosh?

For the patience of Minsky
and resourceful thinking.
Let us look at the living
with the eyes of building.

For the sharp symbols in the screen
and for my friends dancing serene;
for the hours devoted to think,
and the strangeness of this dream.

For the explosions postponed
towards the gems of Space
in these skies that were loaned.
Soon the asteroids we'll chase!

For my friends from Spain

and their lasting example,
your presence in my brain
often gives me the angle.

For the things that will come...

VITA

2011	B. S. in Mathematics with distinction, Universidad Autónoma de Madrid, Spain
2012	M. S. in Engineering Sciences (Mechanical Engineering), University of California, San Diego
2014	Graduate Teaching Assistant, University of California, San Diego
2015	Ph. D. in Engineering Sciences (Mechanical Engineering), University of California, San Diego

PUBLICATIONS

Journals

D. Mateos-Núñez, J. Cortés, "Distributed saddle-point subgradient algorithms with Laplacian averaging," *IEEE Transactions on Automatic Control, submitted.*

D. Mateos-Núñez, J. Cortés, "Noise-to-state exponentially stable distributed convex optimization on weight-balanced digraphs,"*SIAM Journal on Control and Optimization, submitted.*

D. Mateos-Núñez, J. Cortés, "Distributed online convex optimization over jointly connected digraphs," *IEEE Transactions on Network Science and Engineering 1 (1) (2014), 23-37.*

D. Mateos-Núñez, J. Cortés, "pth moment noise-to-state stability of stochastic differential equations with persistent noise," *SIAM Journal on Control and Optimization 52 (4) (2014), 2399-2421.*

Conferences

D. Mateos-Núñez, J. Cortés, "Distributed optimization for multi-task learning via nuclear-norm approximation," *IFAC Workshop on Distributed Estimation and Control in Networked Systems, Philadelphia, Pennsylvania, USA, 2015, to appear.*

D. Mateos-Núñez, J. Cortés, "Distributed subgradient methods for saddle-point problems," *Proceedings of the IEEE Conference on Decision and Control, Osaka, Japan, 2015, to appear.*

D. Mateos-Núñez, J. Cortés, "Distributed online second-order dynamics for convex optimization over switching connected graphs," *International Symposium on Mathematical Theory of Networks and Systems, Groningen, The Netherlands, 2014, pp. 15-22.*

D. Mateos-Núñez, J. Cortés, "Noise-to-state stable distributed convex optimization on weight-balanced digraphs," *Proceedings of the IEEE Conference on Decision and Control, Florence, Italy, 2013, pp. 2781-2786.*

D. Mateos-Núñez, J. Cortés, "Stability of stochastic differential equations with additive persistent noise," *Proceedings of the American Control Conference, Washington, D.C., USA, 2013, pp. 5447-5452.*

ABSTRACT OF THE DISSERTATION

Distributed Algorithms for Convex Optimization:
noisy channels, online computation, nuclear norm regularization,
and separable constraints

by

David Mateos-Núñez

Doctor of Philosophy in Engineering Sciences (Mechanical Engineering)

University of California, San Diego, 2016

Professor Jorge Cortés, Chair

This thesis contributes to the body of research in the design and analysis of distributed algorithms for the optimization of a sum of convex functions, that finds applications in networked and multi-agent systems. In this framework, a group of agents cooperate with each other to optimize the sum of their local objectives in a decentralized way by means of local interactions. We consider four aspects. In the first scenario, the agents need to agree on a global decision vector that minimizes the unconstrained sum. In this case, we study a family of distributed,

continuous-time algorithms that have each agent update its estimate of the global optimizer doing gradient descent on its local cost function while, at the same time, seeking to agree with its neighbors' estimates via proportional-integral feedback on their disagreement. Our aim is to characterize the algorithm robustness properties against the additive persistent noise resulting from errors in communication and computation. We model this algorithm as a stochastic differential equation and develop a novel Lyapunov technique to establish the noise-to-state stability property in 2nd moment.

In the second scenario, we consider the online case, whereby each agent in the network commits to a decision and incurs a local cost given by functions that are revealed over time and whose unknown evolution might be adversarially adaptive to the agent's behavior. The goal of each agent is to incur a cumulative cost over time with respect to the sum of local functions across the network that is competitive with the best centralized decision in hindsight. The proposed algorithms evolve in discrete time using first-order information of the objectives in the form of subgradients, and the communication topology is modeled as a sequence of time-varying weight-balanced digraphs such that the consecutive unions over time periods of some length are strongly connected. We illustrate our results in an application to medical diagnosis, where networked hospitals use patient data to improve their decision models cooperatively in an online fashion.

In the third scenario, we depart from the cooperative search of a global decision vector. Instead, the agents now wish to estimate local decision vectors that minimize the sum of their objectives and are coupled through a constraint that is a sum of convex functions. Motivated by dual-decompositions of constrained optimization problems through the Lagrangian formulation, we consider subgradient algorithms to find a saddle-point of general convex-concave functions under agreement constraints. These distributed strategies are suitable for monotone variational inequality problems, which are equivalent to convex-concave saddle-point problems.

In the fourth scenario, we show a distributed treatment of nuclear-norm regularization, a widely used convex surrogate of the rank function on the spectral ball. To this end, we exploit our previous strategies for saddle-point problems using two variational characterizations of the nuclear norm that are separable under an agreement condition on auxiliary matrices that are independent of the size of the network. As a result, we derive two novel distributed algorithms to address standard optimization models for multi-task learning and low-rank matrix completion.

Chapter 1

Introduction

Algorithms are often employed to *optimize*, to produce answers with properties that are the best with respect to some criteria. Hence the fourth part of the title *distributed algorithms for convex optimization.* The *distributed* aspect is the main single topic of this thesis, and among other connotations that we shortly discuss, it means that the optimization goal is achieved through the *participation* of entities, or *agents*, that use resources that are *spread* as opposed to *centralized*, pointing to the idea of *local resources.* One can think of these resources as information, sensing, computational, and storage capabilities, for instance. When we use the word *network*, we may refer to the collective of agents or, most commonly, to the *communication* network, that codifies which agents can communicate information with whom in the sequence of instructions in our algorithm. This points to the idea of *local communication*, or communication between *neighboring* agents in the network. Summarizing, distributed optimization refers to the participation of agents to optimize some criteria using local resources and local communication. Finally, the word *convex* is a mathematical property of the criterion that we are optimizing. It refers to real-valued functions in arbitrary dimensions that admit a linear estimator from below at every point, and define many useful performance models in applications. In this thesis, we, as designers of the algorithms, prescribe the *response* of the agents with respect to the set of criteria that needs to be optimized. These are the *objective* functions associated to each agent, whose sum determines the collective performance. The *design of the objectives* is still another *design stage* independent from the broader *design of the algorithm* that allows to frame the specific problem to be solved. The distributed paradigm that we employ for cooperative optimization is synonym of *decentralized*, or *peer-to-peer*, which is a theme of active research nowadays, including *decentralized trust* and *reputation systems*.

To mention some examples, the tools that we present can be applied to *resource allocation* and *cooperative control* in networked systems using *ad hoc* infrastructures and *peer-to-peer* interactions, and they can also be applied to large-scale machine learning problems. To make a connection with the latter, let us recall that machine learning is about the *estimation* of unknown parameters, which is done via optimization. However, the creation of models of how the available information relates to the unknown parameters is not the focus of this thesis, although we do illustrate the suitability of our tools in this context. Our focus is on the design and analysis of algorithms to solve *general* families of convex optimization problems in a distributed way. These formulations can then be *specified* by expert modelers to fit specific applications. Our algorithms automatically particularize to those cases, and the convergence properties that we establish hold if fairly general hypotheses like convexity are satisfied by the models. While the optimization problems that are solved in machine learning are not always convex (like in *deep neural networks*, to mention an example that is currently the focus of intense research), there are widely used models for *regression* and *classification* and *multi-task feature learning* that fit the assumptions of convexity. There is great modeling power in the intersection of optimization, learning, decision making, and networks, so it is better to search first for the problem and then for the tool. Next we present briefly the problems that we address.

1.1 Problems considered

In a nutshell, we consider the optimization of a sum of convex functions, with aspects such as agreement on the optimizer, nuclear norm regularization, noisy communication channels, time-varying objectives, and constraints coupling the agents' variables. We now present briefly each of them.

In the first class of problems the agents wish to *agree* on an optimal parameter vector, which we call *global* decision vector,

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^{N} f^i(x), \tag{1.1}$$

where $f^i : \mathbb{R}^d \to \mathbb{R}$ is the cost function available to agent i. In regression and classification problems this function represents the *fitness* of a model with parameter vector x with respect to the *data set* of agent i. This motivates calling the above problem *cooperative data fusion*. Other applications demand the global decision vector to be replaced by a set of parameter vectors, which we call *local* decision vectors, constrained in a more flexible way than agreement to capture *patterns* in the decentralized data. In particular, the nuclear norm of the matrix composed of the local parameter vectors across the network promotes *low-rank* solutions, and as such is less rigid than the agreement constraint. This general principle motivates our second class of problems,

$$\min_{w^i \in \mathbb{R}^d, \forall i} \sum_{i=1}^{N} f^i(w^i) + \gamma \|W\|_*, \tag{1.2}$$

where $\|\cdot\|_*$ is the nuclear norm[1] of the matrix $W = [w^1|\ldots|w^N] \in \mathbb{R}^{d\times N}$ formed by aggregating the vectors $\{w^i\}_{i=1}^N$ as columns. The nuclear norm is weighted by the design parameter $\gamma \in \mathbb{R}_{>0}$ and the hope from a modeler perspective is that, by tuning this parameter appropriately, one induces the set of local decision vectors to belong approximately to a low-dimensional vector space.

What we mean by solving problems (1.1) and (1.2) in a distributed way is the following, which we can call the **distributed *imperative***: agent i updates iteratively an estimate of the optimal values by using information from f^i and by sharing its estimate with their neighbors in the communication network. The agents are allowed to share additional *auxiliary* variables as long as the communication and computational cost is non prohibitive. In addition, each agent can *project* their iterates into *simple* convex sets. We anticipate that our algorithms employ *first-order* information from the objective functions in the form of *subgradients*, and the agents' interactions occur through *Laplacian* averaging, which is essentially linear averaging. The auxiliary variables employed by the agents are usually motivated by *Lagrange multipliers* or, in the case of nuclear norm, by a characterization in terms of a *min-max* problem employing auxiliary local matrices. The dimension of these matrices is $d(d+1)/2$, ideally independent of N, giving an idea of what it means to get close to prohibitive communication costs. The **appeal of the distributed framework** is many-fold:

- **Privacy concerns** are respected because the private data sets are codified

[1] The sum of the singular values.

"wholesale" in the agents' objective functions, which are not shared.

- **Communication bandwidth** is efficiently used because communications are *sparse* and because the local estimates are a "compressed" version, in a sense, of the entire data sets.

- **Commodity hardware** is aggregated across the network of agents, including computation, storage and data-collection/sensing capabilities.

- **Fault tolerance** is assured because the solution to the problem is achieved cooperatively through the *non-privileged* interaction of many agents.

- **Scaling** only requires localized additional infrastructure and adjustment of algorithm parameters.

In addition to these features, in problem (1.1) we also include the treatment of **noise** in the communication channels and scenarios with **time-varying** objective functions, and in problem (1.2) we also consider, instead of the nuclear norm regularization, **constraints** coupling the local variables through a sum of convex functions.

The case of noise concerns the modeling of the algorithm as a stochastic dynamical system, and indeed the language of dynamical systems and their asymptotic behavior, quantified in detail, is in the core of our contributions. Precisely, our model in this case is a *stochastic differential equation*, which is an ordinary differential equation whose integration is perturbed by Brownian motion. The performance of our distributed algorithm with noisy communication channels is then characterized using the notion of ***noise-to-state stability*** in second moment, which describes a specific type of stochastic convergence to a neighborhood whose size depends on the magnitude of the noise.

The above problems are *off-line* in the sense that the data set defining the problem is available from the start of the execution of the algorithm. In contrast, we consider an alternative scenario wherein the objective functions, or the data, are revealed sequentially in an *unpredictable* manner. This refers to the **online**, or real-time, scenario, which for an arbitrary agent j consists of showing the ***agent regret*** of a sequence of estimates $\{x_t^j\}_{t=1}^T$ with respect to the best *centralized* choice in *hindsight* over some time horizon T,

$$\mathcal{R}_T^j := \sum_{t=1}^T \sum_{i=1}^N f_t^i(x_t^j) - \min_{y \in \mathbb{R}^d} \sum_{t=1}^T \sum_{i=1}^N f_t^i(y).$$

The function f_t^i is the cost incurred by agent i at time t. The **online *imperative*** is that each agent i in the network *observes* f_t^i only upon calculating its estimate x_t^i, and for this it can only use *historical* knowledge from previous objectives $\{f_s^i\}_{s=1}^{t-1}$ (usually just the last one) and also the estimates from its neighbors. The **intuition about the regret** comes from the fact that *if* it can be bounded by a sub-linear function of T, *then* we can guarantee that in temporal average, an arbitrary agent j is doing *nearly as well*, asymptotically, as the best choice had all the information been centrally available. The **appeal of the online framework** complements the distributed framework:

- **Real time** processing of data streams provides adaptability.
- **Data rate** can be faster than the diffusion of information across the network.
- **Sub-linear regret** (*only*) says that trends that can be captured by a single decision in hindsight, can also be approximated "on the fly".

The addition of **constraints coupling the local decision vectors** through a sum of convex functions has applications in multi-agent systems outside machine learning, including

- **Traffic** and **routing** problems where the constraints are given by conservation equations.
- **Resource allocation** problems where the constraints include budgets and/or demand satisfaction.
- **Optimal control** problems in discrete time where the constraints refer to the system evolution.
- **Network formation** where the constraints refer to relative distances and angles.
- **Metric learning** where the optimization constraints are given, for instance, by Large Margin Nearest Neighbor relations.

Many of these problems can be written in the following form,

$$\begin{aligned} \min_{w^i \in \mathcal{W}_i, \forall i, \, D \in \mathcal{D}} \quad & \sum_{i=1}^{N} f^i(w^i, D) \\ \text{s.t.} \quad & \sum_{i=1}^{N} g^i(w^i, D) \leq 0, \end{aligned}$$

where f^i and g^i are functions available to agent i that might depend on both a local decision vector and a global decision vector in which the agents need to agree.

A crucial aspect is that we consider constraints that couple the local variables of agents *even* if they are **not** neighbors in the communication network, motivating the distinction between ***constraint*** **graph** and ***communication*** **graph**. The *nodes* of these graphs represent the agents, but the *edges* codify different things. In the constraint graph there is an edge between two nodes whenever there is a constraint coupling the decisions of the corresponding agents. In the communication graph, there is a directed edge, pointing from one agent to another, if the first agent can receive information from the second. As an example, consider the agreement constraint in problem (1.1). It turns out that the constraint graph associated to agreement can be represented in terms of **any** *connected graph*, meaning that all the agents agree if they agree *pair-wise* over a set of relations that connect all the agents. In fact, the agreement condition has a special character in distributed optimization, partly due to the extensive research in *consensus* algorithms. Thanks to the current understanding of these algorithms, we know that consensus can be achieved under very general assumptions on the connectivity of the communication graphs. For instance, the communication graph can *vary* with time and only the consecutive *unions* over bounded periods of time need to contain a directed *path* between any two nodes. This makes consensus-based distributed strategies very valuable as we explain next.

To address the dichotomy between constraint and communication graphs, we use the following insight. In the Lagrangian formulation of the above constrained optimization problems, the *linearity* with respect to the component functions in the constraints allows to introduce *copies* of the Lagrange multipliers subject to agreement. The constraints can then be split in their component functions among the agents by assigning them the corresponding copy of the multipliers. This is a good strategy, because the agents can deal with the agreement condition on the copies of the multipliers by relying just on the communication graph. For the sake of generality, we consider formulations where the Lagrangian is replaced by a general convex-concave function and study the corresponding saddle-point problems with explicit agreement constraints on a subset of the arguments of both the convex and concave parts. This holds the key for the treatment of the nuclear norm introduced in problem (1.2) thanks to a characterization of the nuclear norm as a min-max problem in additional variables. In this case, a preliminary formulation as a minimization problem reveals an additive structure in the objective function under an agreement condition, while a further transformation into a min-max problem through explicit *Fenchel conjugacy* avoids the computation of the inverse of local matrices by candidate subgradient algorithms. Crossing this conceptual bridge in the opposite direction, in the case of minimization problems with linear constraints, one can also eliminate the *primal* variables in the Lagrange formulation in favor of the maximization of the sum of Fenchel conjugates under agreement on the multipliers, which also favors the distributed strategies studied in this thesis. With the unifying role of agreement, we complete our overview.

1.2 Literature review

The following presentation is divided in four categories: the broad field of distributed optimization, including the constrained and the unconstrained cases; the regret perspective for online optimization; the significance and treatment of nuclear norm regularization; and finally the stability analysis of stochastic differential equations that places in context the development of our tools for noise-to-state stability.

1.2.1 Distributed optimization

Our work on distributed optimization builds on three related areas: iterative methods for saddle-point problems [AHU58, NO09b], dual decompositions for constrained optimization [PB13, Ch. 5], [BPC$^+$11], and consensus-based distributed optimization algorithms; see, e.g., [NO09a, JKJJ08, WO12, ZM12, GC14, WE11] and references therein. Historically, these fields have been driven by the need of solving constrained optimization problems and by an effort of parallelizing the computations [Tsi84, BT97], leading to consensus approaches that allow different processors with local memories to update the same components of a vector by averaging their estimates (see the pioneer work [TBA86]).

Saddle-point or min-max problems arise in optimization contexts such as worst-case design, exact penalty functions, duality theory, and zero-sum games, see e.g. [BNO03]. In a centralized scenario, the work [AHU58] studies iterative subgradient methods to find saddle points of a Lagrangian function and establishes convergence to an arbitrarily small neighborhood depending on the gradient stepsize. Along these lines, [NO09b] presents an analysis for general convex-concave functions and studies the evaluation error of the running time-averages, showing convergence to an arbitrarily small neighborhood assuming boundedness of the estimates. In [NO09b, NO10a], the boundedness of the estimates in the case of Lagrangians is achieved using a truncated projection onto a closed set that preserves the optimal dual set, which [HUL93] shows to be bounded when the strong Slater condition holds. This bound on the Lagrange multipliers depends on global information and hence must be known beforehand for its use in distributed implementations.

Dual decomposition methods for constrained optimization are the melting pot where saddle-point approaches come together with methods for parallelizing the computations, like the alternating direction method of multipliers (ADMM) and primal-dual subgradient methods. These methods constitute a particular approach to split a sum of convex objectives by introducing agreement constraints on the primal variable, leading to distributed strategies such as distributed ADMM [WO12] and distributed primal-dual subgradient methods [GC14, WE11]. Ultimately, these methods allow to distribute constraints that are also sums of convex functions via agreement on the multipliers [CNS14].

In distributed constrained optimization, we highlight two categories of constraints that determine the technical analysis and the applications: the first type concerns a *global* decision vector in which agents need to agree. See, e.g., [YXZ11, ZM12, YHX15], where all the agents know the constraint, or see, e.g., [Ozd07, NOP10, NDS10, ZM12], where the constraint is given by the intersection of abstract closed convex sets. The second type *couples* the *local* decision vectors across the network. Examples of the latter include [CNS14], where the inequality constraint is a sum of convex functions and each one is only known to the corresponding agent. Another example is [MARS10], where in the case of linear equality constraints there is a distinction between constraint graph and communication graph. In this case, the algorithm is proved to be distributed with respect to the communication graph, deepening on previous paradigms where each agent needs to communicate with all other agents involved in a particular constraint [RC15]. Employing dual decomposition methods previously discussed, this thesis addresses a combination of the two types of constraints, including the least studied second type. This is possible using a strategy that allows an agreement condition to play an independent role on a subset of both primal and dual variables. We in fact tackle these constraints from a more general perspective, namely, we provide a multi-agent distributed approach for the class of saddle-point problems in [NO09b] under an additional agreement condition on a subset of the variables of both the convex and concave parts. We do this by combining the saddle-point subgradient methods in [NO09b, Sec. 3] and the kind of linear proportional feedback on the disagreement employed by [NO09a] for the minimization of a sum of convex functions. The resulting family of algorithms particularize to a novel class of primal-dual consensus-based subgradient methods when the convex-concave function is the Lagrangian of the minimization of a sum of convex functions under a constraint of the same form.

In this particular case, the recent work [CNS14] uses primal-dual perturbed methods, which require the extra updates of the perturbation points to guarantee asymptotic convergence of the running time-averages to a saddle point. These computations require subgradient methods or proximal methods that add to the computation and the communication complexity.

We can also provide a taxonomy of distributed algorithms for convex optimization depending on how the particular work deals with a multiplicity of aspects that include the network topology, the type of implementation, and the assumptions on the objective functions and the constraints, and the obtained convergence guarantees. Some algorithms evolve in discrete time with associated gradient stepsize that is vanishing [DAW12, SN11, TLR12, ZM12], nonvanishing [NO09a, RNV10, SN11], or might require the solution of a local optimization at each iteration [DAW12, WO12, TLR12, NLT11]; others evolve in continuous time [WE10, GC14, LT12] and even use separation of time scales [ZVC$^+$11]; and some are hybrid [WL09]. Most algorithms converge asymptotically to the solution, while others converge to an arbitrarily good approximation [NO09a, RNV10]. Some examples of convergence rates, or size of the cost error as a function of the number of iterations, are $1/\sqrt{k}$ [DAW12, TLR12] and $1/k$ [WO12]. The communication topologies might be undirected [NO09a, WO12, LT12, NLT11, WE10, ZVC$^+$11], directed and weight-balanced or with a doubly stochastic adjacency matrix [DAW12, GC14, ZM12, RNV10, SN11], or just directed under some knowledge about the number of in-neighbors and out-neighbors [TLR12]; also, they can be fixed [GC14, WO12, LT12, NLT11, ZVC$^+$11], or change over time under joint connectivity [DAW12, NO09a, ZM12, TLR12, RNV10, NLT11, SN11]. On the other hand, the objective functions might be required to be twice continuously differentiable [LT12, ZVC$^+$11] or once differentiable [GC14, Sec. V], [NLT11], or just Lipschitz [DAW12], [GC14, Sec. IV], [NO09a, WO12, ZM12, TLR12, RNV10, SN11]; in addition, they might need to be strongly convex [LT12], strictly convex [WO12, NLT11, ZVC$^+$11], or just convex [DAW12, GC14, NO09a, ZM12, TLR12, RNV10, SN11]. Some algorithms use the Hessian of the objective functions in addition to the gradients [LT12, NLT11, ZVC$^+$11]. Also, the agents might need to share their gradients or second derivatives [LT12, ZVC$^+$11] or even their objectives [NLT11]. Some incorporate a global constraint known to all the agents using a projection method [DAW12, ZM12, TLR12, RNV10] or a dual method [NLT11], and in same cases each agent has a different constraint [ZM12, SN11]. Some algorithms impose a constraint on the initial condition [LT12, NLT11] in order to guarantee convergence. The algorithm execution can be synchronous [GC14, WO12, LT12], allow gossip/randomized communication [LTRB11, SN11], or use event-triggered communication [WL09, KCM15]. Of particular interest to one of our chapters are the works that consider noise affecting the dynamics through stochastically perturbed gradients with associated vanishing stepsize [DAW12] or nonvanishing stepsize [RNV10], while [SN11] considers both noisy communication links and subgradient errors. The characterization of the (discrete-time) algorithm performance under noise provided in these works builds on the fact that the projection onto a compact constraint set at every iteration effectively provides a uniform bound on the subgradients of the component functions.

Our work on distributed unconstrained optimization under noise generalizes the class of continuous-time algorithms studied in [WE10] for undirected graphs and in [GC14] for weight-balanced digraphs. Specifically, in the case of weight-balanced communication digraphs, we also account for the presence of noise in the communication channels and in the agent computations. Under this strategy, each agent updates its estimate of the global solution using the gradient of its local objective function while, at the same time, performing proportional-integral distributed feedback on the disagreement among neighboring agents. As a result, the set of equilibria is given by the solution of the optimization problem together with an affine subspace of the integrator variables. The introduction of noise makes the resulting dynamical system a stochastic differential equation [Mao11, Ö10, Kha12], with the particular feature that the stochastic perturbations do not decay with time and are present even at the equilibria of the underlying deterministic dynamics. The persistent nature of the noise rules out many classical stochastic notions of stability [Thy97, Mao99, Mao11]. Instead, the concept of noise-to-state stability (NSS) [DK00] with respect to an equilibrium of the underlying ordinary differential equation is a notion of stochastic convergence to a neighborhood of that point. More precisely, it provides an ultimate bound for the state of the stochastic system, in probability, that depends on the magnitude of the covariance of the noise. Asymptotic convergence to the equilibrium follows in the absence of noise. In this regard, we build on our extension [MNC14b] of this concept to NSS in pth moment with respect to subspaces to establish NSS in second moment with respect to the subspace of equilibria of the underlying ordinary differential equation.

1.2.2 Distributed online optimization

Online learning is about sequential decision making given historical observations on the loss incurred by previous decisions, even when the loss functions are adversarially adaptive to the behavior of the decision maker. Interestingly, in online convex optimization [Zin03, CBL06, SS12, Haz11], it is doable to be competitive with the best single decision in hindsight. These works show how the regret, i.e., the difference between the cumulative cost over time and the cost of the best single decision in hindsight, is sublinear in the time horizon. Online convex optimization has applications to information theory [CBL06], game theory [SSS07], supervised online machine learning [SS12], online advertisement placement, and portfolio selection [Haz11]. Algorithmic approaches include online gradient descent [Zin03], online Newton step [HAK07], follow-the-approximate-leader [HAK07], and online alternating directions [WB12]. A few recent works have explored the combination of distributed and online convex optimization. The work [DGBSX12] proposes distributed online strategies that rely on the computation and maintenance of spanning trees for global vector-sum operations and work under suitable statistical assumptions on the sequence of objectives. The work [RKW11] studies decentralized online convex programming for groups of agents whose interaction topology is a chain. The works [YSVQ13, HCM13] study agent regret without any statistical assumptions on the sequence of objectives. On the one hand [YSVQ13] introduces distributed online projected subgradient descent and shows square-root regret (for convex cost functions) and logarithmic regret (for strongly convex cost functions). The analysis critically relies on a projection step onto a compact set at each time step (which automatically guarantees the uniform boundedness of the estimates), and therefore excludes the unconstrained case (given the non-compactness of the whole state space). In contrast, [HCM13] introduces distributed online dual averaging and shows square-root regret (for convex cost functions) using a general regularized projection that admits both unconstrained and constrained optimization, but the logarithmic bound is not established. Both works only consider static and strongly-connected interaction digraphs. Our approach to online optimization generalizes a family of distributed saddle-point subgradient algorithms [WE11, GC14] that enjoy asymptotic (exponenial) convergence with constant stepsizes and robust asymptotic behavior in the presence of noise [MNC13].

1.2.3 Nuclear norm regularization

Mathematical models that use a low-rank matrix estimate are key in applications such as recommender systems through matrix completion [CR09], dimension reduction in multivariate regression [YL07], multi-task feature learning [AZ05, AEP06, AEP08], and convex relaxations of optimal power flow [MFSL14], where it is necessary to recover a low-rank solution of a semidefinite program. The basic underlying structure is the same: an estimate of a matrix that is *assumed or postulated* to be of low rank. While the rank function is nonconvex, it turns out that the nuclear norm, defined as the one norm of the vector of singular values, is the convex surrogate of the rank function [Faz02]. When used as a regularization in optimization problems, the nuclear norm promotes a low-rank solution and in some cases even allows to recover the exact low-rank solution [CT10, RFP10]. The applications of nuclear norm regularization described above have inspired research in parallel computation following the model of stochastic gradient descent [RR13], but these developments emphasize the parallel aspect alone, rather than other aspects such as geographically distributed data, communication bandwidth, and privacy. Other strategies to address the problem do not consider the parallel or the distributed aspects, but instead try to overcome the nonsmooth nature of the nuclear norm using techniques such as approximate singular value decompositions [WLRT08, WC15]; coordinate descent and subspace selection [DHM12, HO14]; and successive over-relaxation [WYZ12]. Our approach builds on our recent work [MNC15], presented in Chapter 5, which develops a general analytical framework combining distributed optimization and subgradient methods for saddle-point problems.

1.2.4 Stability of stochastic differential equations

Stochastic differential equations (SDEs) go beyond ordinary differential equations (ODEs) to deal with systems subject to stochastic perturbations of a particular type, known as white noise. Applications are numerous and include option pricing in the stock market, networked systems with noisy communication channels, and, in general, scenarios whose complexity cannot be captured by deterministic models. In this thesis we study SDEs subject to *persistent noise* (including the case of additive noise), i.e., systems for which the noise is present even at the equilibria of the underlying ODE and does not decay with time. Such scenarios arise, for instance, in control-affine systems when the input is corrupted by persistent noise. For these systems, the presence of persistent noise makes it impossible to establish in general a stochastic notion of asymptotic stability for the (possibly unbounded) set of equilibria of the underlying ODE. Our focus is therefore to develop notions and tools to study the stability properties of these systems and provide probabilistic guarantees on the size of the state of the system.

In general, it is not possible to obtain explicit descriptions of the solutions of SDEs. Fortunately, the Lyapunov techniques used to study the qualitative behavior of ODEs [Kha02, LMS91] can be adapted to study the stability properties of SDEs as well [Kha12, Thy97, Mao99]. Depending on the notion of stochastic convergence used, there are several types of stability results in SDEs. These include stochastic stability (or stability in probability), stochastic asymptotic stability, almost sure exponential stability, and pth moment asymptotic stability, see e.g., [Thy97, Mao99, Mao11, Tan03]. However, these notions are not appropriate in the presence of persistent noise because they require the effect of the noise on the set of equilibria to either vanish or decay with time. To deal with persistent noise, as well as other system properties like delays, a concept of ultimate boundedness is required that generalizes the notion of convergence. As an example, for stochastic delay differential equations, [WK13] considers a notion of ultimate bound in pth moment [Sch01] and employs Lyapunov techniques to establish it. More generally, for mean-square random dynamical systems, the concept of forward attractor [KL12] describes a notion of convergence to a dynamic neighborhood and employs contraction analysis to establish it. Similar notions of ultimate boundedness for the state of a system, now in terms of the size of the disturbance, are also used for differential equations, and many of these notions are inspired by dissipativity properties of the system that are captured via *dissipation inequalities* of a suitable Lyapunov function: such inequalities encode the fact that the Lyapunov function decreases along the trajectories of the system as long as the state is "big enough" with regards to the disturbance. As an example, the concept of input-to-state stability (ISS) goes hand in hand with the concept of ISS-Lyapunov function, since the existence of the second implies the former (and, in many cases, a converse result is also true [SW95]). Along these lines, the notion of practical stochastic input-to-state stability (SISS) in [LZJ08, WXZ07] generalizes the concept of ISS to SDEs where the disturbance or input affects both the deterministic term of the dynamics and the diffusion term modeling the role of the noise. Under this notion, the state bound is guaranteed in probability, and also depends, as in the case of ISS, on a decaying effect of the initial condition plus an increasing function of the sum of the size of the input and a positive constant related to the persistent noise. For systems where the input modulates the covariance of the noise, SISS corresponds to noise-to-state-stability (NSS) [DK00], which guarantees, in probability, an ultimate bound for the state that depends on the magnitude of the noise covariance. That is, the noise in this case plays the main role, since the covariance can be modulated arbitrarily and can be unknown. This is the appropriate notion of stability for the class of SDEs with persistent noise considered in this thesis, which are nonlinear systems affine in the input, where the input corresponds to white noise with locally bounded covariance. Such systems cannot be studied under the ISS umbrella, because the stochastic integral against Brownian motion has infinite variation, whereas the integral of a legitimate input for ISS must have finite variation.

1.3 Contributions

The contributions are organized according to chapter of the same title, which can be read independently.

1.3.1 Noise-to-state exponentially stable distributed convex optimization on weight-balanced digraphs

In the distributed approach to the optimization of an unconstrained sum of convex functions, we assume that both inter-agent communications and agent computations are corrupted by Gaussian white noise of locally bounded covariance, and the communication topology is represented by a strongly connected weight-balanced digraph. We study a family of distributed continuous-time coordination algorithms where each agent keeps track and interchanges with its neighbors two variables: one corresponding to its current estimate of the global optimizer and the other one being an auxiliary variable to guide agents towards agreement. According to this coordination strategy, each agent updates its estimate using gradient information of its local cost function while, at the same time, seeking to agree with its neighbors' estimates via proportional-integral feedback on their disagreement. The presence of noise both in the communication channels and the agent computations introduces errors in the algorithm execution that do not decay with time and are present even at the equilibria.

Our main contribution establishes that the resulting dynamics, modeled as a stochastic differential equation, is noise-to-state exponentially stable in second moment and, therefore, robust against additive persistent noise. Precisely, we characterize the exponential rate of convergence in second moment to a neighborhood that depends on the size of the disturbance. Our technical approach relies on the construction of a suitable candidate noise-to-state (NSS) Lyapunov function whose nullspace is the affine subspace corresponding to the solution of the convex optimization problem plus a direction of the auxiliary variables that absorbs the variance of the noise. To verify that the candidate function is in fact an NSS Lyapunov function, we analyze the interaction between local optimization and local consensus through the co-coercivity of a family of vector fields comprising a gradient of a convex function plus a linear transformation with a nonsymmetric Laplacian. Specifically, we give sufficient conditions for this family of vector fields to be co-coercive under small perturbations. When noise is present, the expected rate of change of the NSS Lyapunov function is proportional to the difference between the square Frobenius norm of the covariance of the noise and the distance to its nullspace. In the absence of noise, our NSS-Lyapunov function renders the system exponentially stable with respect to the solution.

1.3.2 Distributed online convex optimization over jointly connected digraphs

We consider the online unconstrained convex optimization scenario where no model is assumed about the evolution of the local objectives available to the agents. In this scenario, we propose a class of distributed coordination algorithms and study the associated agent regret in the optimization of the sum of the local cost functions across the network. Our algorithm design combines subgradient descent on the local objectives revealed in the previous round and proportional-integral (and/or higher-order) distributed feedback on the disagreement among neighboring agents. Assuming bounded subgradients of the local cost functions, we establish logarithmic agent regret bounds under local strong convexity and square-root agent regret under convexity plus a mild geometric condition. We also characterize the dependence of the regret bounds on the network parameters. Our technical approach uses the concept of network regret, which captures the performance of the sequence of collective estimates across the group of agents. The derivation of the sublinear regret bounds results from three main steps: the study of the difference between network and agent regret; the analysis of the cumulative disagreement of the online estimates via the input-to-state stability property of a generalized Laplacian consensus dynamics; and the uniform boundedness of the online estimates (and auxiliary variables) when the set of local optimizers is uniformly bounded. With respect to previous work, the contributions advance the current state of the art because of the consideration of unconstrained formulations of the online optimization problem, which makes the discussion valid for regression and classification and raises major technical challenges to ensure the uniform boundedness of estimates; the synthesis of a novel family of coordination algorithms that generalize distributed online subgradient descent and saddle-point dynamics; and the development of regret guarantees under jointly connected interaction digraphs. Our novel analysis framework modularizes the main technical ingredients (the disagreement evolution via linear decoupling and input-to-state stability; the boundedness of estimates and auxiliary states through marginalizing the role of disagreement and learning rates; and the role played by network topology and the convexity properties) and extends and integrate techniques from distributed optimization (e.g., Lyapunov techniques for consensus under joint connectivity) and online optimization (e.g., Doubling Trick bounding techniques for square-root regret). We illustrate our results in a medical diagnosis example.

1.3.3 Distributed saddle-point subgradient algorithms with Laplacian averaging

We consider general saddle-point problems with explicit agreement constraints on a subset of the arguments of both the convex and concave parts. These problems appear in dual decompositions of constrained optimization problems, and in other saddle-point problems where the convex-concave functions, unlike Lagrangians, are not necessarily linear in the arguments of the concave part. This is a substantial improvement over prior work that only focuses on dual decompositions of constrained optimization. When considering constrained optimization problems, the agreement constraints are introduced as an artifact to distribute both primal and dual variables independently. For instance, separable constraints can be decomposed using agreement on dual variables, while a subset of the primal variables can still be subject to agreement or eliminated through Fenchel conjugation; local constraints can be handled through projections; and part of the objective can be expressed as a maximization problem in extra variables. Driven by these important classes of problems, our main contribution is the design and analysis of distributed coordination algorithms to solve general convex-concave saddle-point problems with agreement constraints, and to do so with subgradient methods, which have less computationally complexity. The coordination algorithms that we study can be described as projected saddle-point subgradient methods with Laplacian averaging, which naturally lend themselves to distributed implementation. For these algorithms we characterize the asymptotic convergence properties in terms of the network topology and the problem data, and provide the convergence rate. The technical analysis entails computing bounds on the saddle-point evaluation error in terms of the disagreement, the size of the subgradients, the size of the states of the dynamics, and the subgradient stepsizes. Finally, under assumptions on the boundedness of the estimates and the subgradients, we further bound the cumulative disagreement under joint connectivity of the communication graphs, regardless of the interleaved projections, and make a choice of decreasing stepsizes that guarantees convergence of the evaluation error as $1/\sqrt{t}$, where t is the iteration step. We particularize our results to the case of distributed constrained optimization with objectives and constraints that are a sum of convex functions coupling local decision vectors across a network. For this class of problems, we also present a distributed strategy that lets the agents compute a bound on the optimal dual set. This bound enables agents to project the estimates of the multipliers onto a compact set (thus guaranteeing the boundedness of the states and subgradients of the resulting primal-dual projected subgradient dynamics) in a way that preserves the optimal dual set. We illustrate our results in simulation for an optimization scenario with nonlinear constraints coupling the decisions of agents that cannot communicate directly.

1.3.4 Distributed optimization for multi-task learning via nuclear-norm approximation

We motivate the nuclear norm regularization in two problems that can benefit from distributed strategies: multi-task feature learning and matrix completion. Then we introduce two distributed formulations of the resulting optimization problems: a separable convex minimization, and a separable saddle-point problem, and we make the presentation systematic as to the automatic derivation of distributed coordination algorithms. After introducing each formulation, we establish the existence of critical points that solve the original problem and also present the corresponding distributed subgradient dynamics. To the best of our knowledge, our subgradient saddle-point method is a novel coordination algorithm even in its centralized version and we argue its advantages and general application to each of the motivational problems. For both families of distributed strategies, we establish the convergence guarantees. The subgradient saddle-point method relies on the input-to-state stability properties of auxiliary states, necessary for the boundedness of the estimates. The convergence results are illustrated in a simulation example of low-rank matrix completion.

1.3.5 pth moment noise-to-state stability of stochastic differential equations with persistent noise

The contributions in this topic are twofold. Our first contribution concerns the noise-to-state stability of systems described by SDEs with persistent noise. We generalize the notion of noise-dissipative Lyapunov function, which is a positive semidefinite function that satisfies a dissipation inequality that can be nonexponential (by this we mean that the inequality admits a convex $\mathcal{K}_\infty$ gain instead of the linear gain characteristic of exponential dissipativity). We also introduce the concept of $p$thNSS-Lyapunov function with respect to a closed set, which is a noise-dissipative Lyapunov function that in addition is proper with respect to the set with a convex lower-bound gain function. Using this framework, we show that noise-dissipative Lyapunov functions have NSS dynamics and we characterize the overshoot gain. More importantly, we show that the existence of a $p$thNSS-Lyapunov function with respect to a closed set implies that the system is NSS in pth moment with respect to the set. Our second contribution is driven by the aim of providing alternative, structured ways to check the hypotheses of the above results. We introduce the notion of two functions being proper with respect to each other as a generalization of the notion of properness with respect to a set. We then develop a methodology to verify whether two functions are proper with respect to each other by analyzing the associated pair of inequalities with increasingly strong refinements that involve the classes $\mathcal{K}$, $\mathcal{K}_\infty$, and $\mathcal{K}_\infty$ plus a convexity property. We show that these refinements define equivalence relations between pairs of functions, thereby producing nested partitions on the space of functions. This provides a useful way to deal with these inequalities because the construction of the gains is explicit when the transitivity property is exploited. This formalism motivates our characterization of positive semidefinite functions that are proper, in the various refinements, with respect to the Euclidean distance to their nullset. This characterization is technically challenging because we allow the set to be noncompact, and thus the pre-comparison functions can be discontinuous. We devote special attention to the case when the set is a subspace and examine the connection with seminorms. Finally, we show how this framework allows us to develop an alternative formulation of our stability results.

1.4 Organization

The technical chapters can be read independently. Chapter 2 presents some notational conventions and preliminary notions in Optimization, Graph Theory and Stochastic Differential Equations. In Chapters 3 and 4, we present our distributed coordination algorithms for the unconstrained minimization of a sum of convex functions in two scenarios: first having the agents communicate under noisy communication channels, a scenario in which we study the noise-to state stability property in second moment; and, second, having the agents make decisions “on the fly” using information that is incrementally revealed over time, a scenario in which we study the agent regret. In Chapter 5, we develop a distributed strategy for saddle-point problems with convex-concave functions with explicit agreement constraints in a subset of the arguments. These algorithms particularize to primal-dual subgradient algorithms for distributed constrained optimization. This framework also encodes the distributed treatment of nuclear norm regularization that we present in Chapter 6. The development of a novel Lyapunov technique to asses the stability in second moment of stochastic differential equations is given in Chapter 7. Finally, Chapter 8 gathers our conclusions and ideas for future research directions.

Saddle-point problems are the most general *formulations* that we address, because these include the variational formulation of the nuclear norm and also the treatment of constrained optimization problems through the Lagrangian, which naturally includes the unconstrained case. However, this hierarchy does not reflect the classification of the mathematical tools and models that we develop to address aspects such as noise in the communication channels or different performance metrics such as the agent regret. Therefore, we stick to the order in which we have developed the results, with the exception of the Lyapunov techniques to assess the stability of stochastic differential equations that we relegate to the end.

Chapter 2

Preliminaries

In this chapter we introduce some notational conventions and review basic notions about convex analysis, characterizations of the nuclear norm, graph theory, and stochastic differential equations.

2.1 Notational conventions

We let $\mathbb{R}$ and $\mathbb{R}_{\geq 0}$ denote the sets of real and nonnegative real numbers, respectively. We denote by $\mathbb{R}^n$ the n-dimensional Euclidean space, by $\mathrm{I}_n \in \mathbb{R}^{n \times n}$ the identity matrix in $\mathbb{R}^n$, by $e_i \in \mathbb{R}^n$ the ith column of I_n, and by $\mathbb{1}$ the vector of ones. Given two vectors, u, $v \in \mathbb{R}^n$, we denote by $u \geq v$ the entry-wise set of inequalities $u_i \geq v_i$ for each $i = 1, \ldots, n$. The linear subspace generated by a set $\{u_1, ..., u_m\} \subseteq \mathbb{R}^n$ of vectors is denoted by $\mathrm{span}\{u_1, ..., u_n\}$. For simplicity, we often use $(v_1, \ldots v_n)$ to represent the column vector $[v_1^\top, \ldots v_n^\top]^\top$. Given an array v whose entries are matrices, we denote by $\mathrm{diag}(v)$ the block-diagonal matrix whose blocks are the entries of v.

Given a vector $v \in \mathbb{R}^n$, we denote its one-norm by $\|v\|_1 = \sum_{i=1}^n |v_i|$, the Euclidean norm, or two-norm, by $\|v\|_2 = \sqrt{\sum_{i=1}^n v_i^2}$, and the Euclidean distance from x to a set $\mathcal{U} \subseteq \mathbb{R}^n$ by $|x|_{\mathcal{U}} := \inf\{\|x-u\|_2 : u \in \mathcal{U}\}$. The function $|.|_{\mathcal{U}}$ is continuous when $\mathcal{U}$ is closed. The Euclidean open ball of radius ϵ centered at x is represented by $\mathcal{B}(x,\epsilon) := \{y \in \mathbb{R}^n : \|y-x\|_2 < \epsilon\}$, while $\bar{\mathcal{B}}(x,\epsilon)$ is the closed counterpart. Given $\mathcal{D} \subseteq \mathbb{R}^n$, we denote by $\mathcal{C}(\mathcal{D};\mathbb{R}_{\geq 0})$ and $\mathcal{C}^2(\mathcal{D};\mathbb{R}_{\geq 0})$ the set of positive semidefinite functions defined on $\mathcal{D}$ that are continuous and continuously twice differentiable (if $\mathcal{D}$ is open), respectively. Given normed vector spaces X_1, X_2, a function $f : X_1 \to X_2$ is Lipschitz with constant κ if $\|f(x)-f(y)\|_{X_1} \leq \kappa\|x-y\|_{X_2}$ for each $x,y \in X_1$, where $\|.\|_X$ denotes the norm in X. Given $f,g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, we say that $f(s)$ is in $\mathcal{O}(g(s))$ as $s \to \infty$ if there exist constants $\kappa, s_0 > 0$ such that $f(s) < \kappa g(s)$ for all $s > s_0$.

Given a *closed* convex set $\mathcal{C} \subseteq \mathbb{R}^n$, the orthogonal projection $\mathcal{P}_{\mathcal{C}}(\cdot)$ onto $\mathcal{C}$ is

$$\mathcal{P}_{\mathcal{C}}(x) \in \arg\min_{x' \in \mathcal{C}} \|x - x'\|_2. \tag{2.1}$$

This value exists and is unique. (Note that *compactness* could be assumed without loss of generality taking the intersection of $\mathcal{C}$ with balls centered at x.) We use the following basic property of the orthogonal projection: for every $x \in \mathcal{C}$ and $x' \in \mathbb{R}^n$,

$$\Big(\mathcal{P}_{\mathcal{C}}(x') - x'\Big)(x' - x) \leq 0. \tag{2.2}$$

If $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, we denote its gradient and Hessian by ∇f and $\nabla^2 f$, respectively. Given a differentiable vector field $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$, we let $\mathbf{DF} : \mathbb{R}^n \to \mathbb{R}^{m \times n}$ denote its Jacobian, where $\mathbf{DF}(x)_{ij} = \frac{\partial \mathbf{F}_i(x)}{\partial x_j}$ for all $x \in \mathbb{R}^n$.

2.1.1 Seminorms

A seminorm is a function $S:\mathbb{R}^n \to \mathbb{R}$ that is positively homogeneous, i.e., $S(\lambda x) = |\lambda| S(x)$ for any $\lambda \in \mathbb{R}$, and satisfies the triangular inequality, i.e., $S(x+y) \leq S(x) + S(y)$ for any $x, y \in \mathbb{R}^n$. From these properties it can be deduced that $S \in \mathcal{C}(\mathbb{R}^n; \mathbb{R}_{\geq 0})$ and its nullset is always a subspace. If, moreover, the function S is positive definite, i.e., $S(x) = 0$ implies $x = 0$, then S is a norm. For any matrix $A \in \mathbb{R}^{m \times n}$, the function $\|x\|_A \triangleq \|Ax\|_2$ is a seminorm and can be viewed as a distance to $\mathcal{N}(A)$. (Note that we depart here from the usual convention of defining $\|x\|_A := \sqrt{x^\top A x}$, which has the inconvenience of requiring A to be symmetric and positive semidefinite). The nullset of the seminorm corresponds to the nullspace of A, $\mathcal{N}(A) = \{x \in \mathbb{R}^n : Ax = 0\}$.

2.1.2 Matrix analysis

For a square matrix $A \in \mathbb{R}^{n \times n}$, the set of eigenvalues is denoted by $\operatorname{spec}(A)$. If the eigenvalues of A are real (for instance if A is real and symmetric), we label them in increasing order from the minimum to the maximum as $\lambda_{\min}(A) = \lambda_1(A), \dots, \lambda_n(A) = \lambda_{\max}(A)$, except in Chapter 3 and Chapter 7 where the order is the opposite, i.e., $\lambda_{\max}(A) = \lambda_1(A)$ and $\lambda_{\min}(A) = \lambda_n(A)$. For convenience, we also use the notation $\lambda_{\max}^{\varnothing}(A)$ to denote the maximum nonzero eigenvalue of A. Given a subspace $\mathcal{U} \subsetneqq \mathbb{R}^n$, and a symmetric matrix $A \in \mathbb{R}^{n \times n}$, we let $\lambda_{\max}^{\mathcal{U}^\perp}(A) := \max_{\{x^\top u = 0 \,:\, u \in \mathcal{U},\, \|x\|_2 = 1\}} x^\top A x$. The singular values of $A \in \mathbb{R}^{n \times m}$ are the square roots of the eigenvalues of $A^\top A$. We order them according to $\sigma_{\max}(A) := \sigma_1(A) \geq \cdots \geq \sigma_r(A) := \sigma_{\min}(A)$, where $r = \operatorname{rank}(A)$ is the rank of A. We denote by $A^\dagger$ the Moore-Penrose pseudoinverse of A, and by $\mathcal{C}(A)$ the column space of A, i.e., the vector space generated by the columns of A. The Kronecker product of

$A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{p\times q}$ is denoted by $A \otimes B \in \mathbb{R}^{np\times mq}$. Recall that $\mathrm{spec}(A\otimes B) = \mathrm{spec}(A) \times \mathrm{spec}(B)$. A matrix $A \in \mathbb{R}^{n\times n}$ is diagonalizable if it can be written as $A = S_A D_A S_A^{-1}$, where $D_A \in \mathbb{R}^{n\times n}$ is a diagonal matrix whose entries are the eigenvalues of A, and $S_A \in \mathbb{R}^{n\times n}$ is an invertible matrix whose columns are the corresponding eigenvectors. The sets $\mathbb{S}^d$, $\mathbb{S}^d_{\succeq 0}$, $\mathbb{O}^d \subseteq \mathbb{R}^{d\times d}$ represent, respectively the symmetric, positive semidefinite, and orthogonal matrices in $\mathbb{R}^{d\times d}$. When a matrix $A \in \mathbb{R}^{d\times d}$ is symmetric positive semidefinite, we often write $A \succeq 0$, while $A \succeq B$ is an equivalent notation for $A - B \succeq 0$.

2.1.3 Matrix norms

The spectral norm, or two-norm, of a rectangular matrix $A \in \mathbb{R}^{n\times m}$ is defined by $\|A\|_2 := \sigma_{\max}(A)$, and its condition number is given by $\kappa(A) := \|A\|_2\|A^{-1}\|_2 = \sigma_{\max}(A)/\sigma_{\min}(A)$. The nuclear norm, or trace norm, is $\|A\|_* = \mathrm{trace}(\sqrt{A^\top A})$. This coincides with the sum of the singular values of A, $\|A\|_* = \sum_{i=1}^{\min\{n,m\}} \sigma_i$. The Frobenius norm is given by $\|A\|_{\mathcal{F}} = \sqrt{\mathrm{trace}(A^\top A)} = \sqrt{\mathrm{trace}(AA^\top)} = \sqrt{\sum_{i=1}^{\min\{n,m\}} \sigma_i^2}$. Note that for any $A \in \mathbb{R}^{n\times m}$ with rank r, the nuclear norm and the Frobenius norm are related by

$$\|A\|_* \leq \sqrt{r}\|A\|_{\mathcal{F}} \leq \sqrt{\min\{n,m\}}\|A\|_{\mathcal{F}}. \tag{2.3}$$

We also denote the $L_{2,1}$-norm of A by $\|A\|_{2,1} := \|(\|a_1\|_2, \ldots, \|a_m\|_2)\|_1$, which is the one-norm of the vector of two-norms of the columns of A.

2.2 Convex functions

Given a convex set $\mathcal{C} \subseteq \mathbb{R}^n$, a function $f : \mathcal{C} \to \mathbb{R}$ is convex if $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$ for all $\alpha \in [0,1]$ and $x, y \in \mathcal{C}$. A vector $\xi_x \in \mathbb{R}^n$ is a subgradient of f at $x \in \mathcal{C}$ if $f(y) - f(x) \geq \xi_x^\top (y-x)$, for all $y \in \mathcal{C}$. We denote by $\partial f(x)$ the set of all such subgradients. Alternatively, the function f is concave if $-f$ is convex, and a subgradient of a concave function is defined through a subgradient of $-f$. The characterization in [Nes04, Lemma 3.1.6] asserts that a function $f : \mathcal{C} \to \mathbb{R}$ is convex if and only if $\partial f(x)$ is nonempty for each $x \in \mathcal{C}$. Equivalently, f is convex if $\partial f(x)$ is nonempty and for each $x \in \mathcal{C}$ and $\xi_x \in \partial f(x)$,

$$f(y) - f(x) \geq \xi_x^\top (y-x) + \tfrac{p(x,y)}{2} \|y-x\|_2^2,$$

for all $y \in \mathcal{C}$, where the nonnegative-valued function $p : \mathcal{C} \times \mathcal{C} \to \mathbb{R}_{\geq 0}$ is the modulus of strong convexity (whose value may be 0). For $p > 0$, a function f is p-strongly convex on $\mathcal{C}$ if $p(x,y) = p$ for all $x, y \in \mathcal{C}$. Equivalently, f is p-strongly convex on $\mathcal{C}$ if

$$(\xi_y - \xi_x)^\top (y-x) \geq p\|y-x\|_2^2,$$

for each $\xi_x \in \partial f(x)$, $\xi_y \in \partial f(y)$, for all $x, y \in \mathcal{C}$. For convenience, we denote by $\operatorname{argmin}(f)$ the set of minimizers of a convex function f in its domain. The following definition comes in handy when we introduce the next class of functions. Given $w \in \mathbb{R}^n \setminus \{0\}$ and $c \in [0,1]$, we let

$$\mathcal{F}_c(w) := \Big\{ v \in \mathbb{R}^n \ : \ v^\top w \geq c\|v\|_2 \|w\|_2 \Big\}$$

denote the convex cone of vectors in $\mathbb{R}^n$ whose angle with w has a cosine lower bounded by c. Using this notation, for $\beta \in [0,1]$, a convex function $f : \mathbb{R}^n \to \mathbb{R}$ with $\operatorname{argmin}(f) \neq \emptyset$ is β-central on $\mathcal{Z} \subseteq \mathbb{R}^n \setminus \operatorname{argmin}(f)$ if for each $x \in \mathcal{Z}$, there exists $y \in \operatorname{argmin}(f)$ such that $-\partial f(x) \subset \mathcal{F}_\beta(y-x)$, i.e.,

$$-\xi_x^\top (y-x) \geq \beta \, \|\xi_x\|_2 \|y-x\|_2,$$

for all $\xi_x \in \partial f(x)$. Note that any convex function $f : \mathbb{R}^n \to \mathbb{R}$ with a nonempty set of minimizers is at least 0-central on $\mathbb{R}^n \setminus \operatorname{argmin}(f)$. Finally, a convex function f has H-bounded subgradient sets if there exists $H \in \mathbb{R}_{>0}$ such that $\|\xi_x\|_2 \leq H$ for all $\xi_x \in \partial f(x)$ and $x \in \mathbb{R}^n$.

2.2.1 Comparison functions

The following classes of comparison functions [Kha02] are useful in our technical treatment in Chapter 3 and Chapter 7. A continuous function $\alpha : [0,b) \to \mathbb{R}_{\geq 0}$, for $b > 0$ or $b = \infty$, is class $\mathcal{K}$ if it is strictly increasing and $\alpha(0) = 0$, and it belongs to class $\mathcal{K}_\infty$ if $\alpha \in \mathcal{K}$ and is unbounded. A continuous function $\mu : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is class $\mathcal{KL}$ if, for each fixed $s \geq 0$, the function $r \mapsto \mu(r,s)$ is class $\mathcal{K}$, and, for each fixed $r \geq 0$, the function $s \mapsto \mu(r,s)$ is decreasing and $\lim_{s \to \infty} \mu(r,s) = 0$. If α_1, α_2 are class $\mathcal{K}$ and the domain of α_1 contains the range of α_2, then their composition $\alpha_1 \circ \alpha_2$ is class $\mathcal{K}$ too. If α_3, α_4 are class $\mathcal{K}_\infty$, then both the inverse function α_3^{-1} and their composition $\alpha_3 \circ \alpha_4$ are class $\mathcal{K}_\infty$. In our technical treatment, it is sometimes convenient to require comparison functions to satisfy additional convexity properties. By [BV09, Ex. 3.3], if $f : [a,b] \to [f(a), f(b)]$ is a strictly increasing convex (respectively, concave) function, then the inverse function $f^{-1} : [f(a), f(b)] \to [a,b]$ is strictly increasing and concave (respectively, convex).

Also, following [BV09, Section 3], if $f, g : \mathbb{R} \to \mathbb{R}$ are convex (respectively, concave) and f is nondecreasing, then the composition $f \circ g$ is also convex (respectively, concave).

2.3 Optimization

For any function $\mathcal{L} : \mathcal{W} \times \mathcal{M} \to \mathbb{R}$, the *max-min inequality* [BV09, Sec 5.4.1] states that

$$\inf_{w \in \mathcal{W}} \sup_{\mu \in \mathcal{M}} \mathcal{L}(w, \mu) \geq \sup_{\mu \in \mathcal{M}} \inf_{w \in \mathcal{W}} \mathcal{L}(w, \mu). \tag{2.4}$$

When equality holds, we say that $\mathcal{L}$ satisfies the *strong max-min property* (also called the *saddle-point* property). A point $(w^*, \mu^*) \in \mathcal{W} \times \mathcal{M}$ is called a *saddle point* if

$$w^* = \inf_{w \in \mathcal{W}} \mathcal{L}(w, \mu^*) \text{ and } \mu^* = \sup_{\mu \in \mathcal{M}} \mathcal{L}(w^*, \mu).$$

[BNO03, Sec. 2.6] discusses sufficient conditions to guarantee the existence of saddle points. Note that the existence of saddle points implies the strong max-min property. Given functions $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^m \to \mathbb{R}$ and $h : \mathbb{R}^p \to \mathbb{R}$, the *Lagrangian* for the problem

$$\min_{w \in \mathbb{R}^n} f(w) \quad \text{s.t.} \quad g(w) \leq 0,\ h(w) = 0, \tag{2.5}$$

is defined as

$$\mathcal{L}(w, \mu, \lambda) = f(w) + \mu^\top g(w) + \lambda^\top h(w) \tag{2.6}$$

for $(\mu, \lambda) \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p$. In this case, inequality (2.4) is called *weak-duality*, and if equality holds, then we say that *strong-duality* (or Lagrangian duality) holds. If a point (w^*, μ^*, λ^*) is a saddle point for the Lagrangian, then w^* solves the constrained minimization problem (2.5) and (μ^*, λ^*) solves the *dual problem*, which is maximizing the *dual function* $q(\mu, \lambda) := \inf_{w \in \mathbb{R}^n} \mathcal{L}(w, \mu, \lambda)$ over $\mathbb{R}^m_{\geq 0} \times \mathbb{R}^p$. This implication is part of the *Saddle Point Theorem.* (The reverse implication establishes the existence of a saddle-point –and thus strong duality– adding a *constraint qualification* condition.) Under the saddle-point condition, the *optimal dual vectors* (μ^*, λ^*) coincide with the *Lagrange multipliers* [Ber99, Prop. 5.1.4]. In the case of affine linear constraints, the dual function can be written using the *Fenchel conjugate* of f, defined in $\mathbb{R}^n$ as

$$f^\star(x) := \sup_{w \in \mathbb{R}^n} \{x^\top w - f(w)\}. \tag{2.7}$$

2.4 Variational characterizations of the nuclear norm

The following characterizations of the nuclear norm play a key role in the distributed formulations that we study in Chapter 6,

$$2\|W\|_* = \min_{\substack{D \in \mathbb{S}^d_{\succeq 0} \\ \mathcal{C}(W) \subseteq \mathcal{C}(D)}} \operatorname{trace}\left(D^\dagger W W^\top\right) + \operatorname{trace}(D), \tag{2.8a}$$

$$\|W\|_*^2 = \min_{\substack{D \in \mathbb{S}^d_{\succeq 0}, \operatorname{trace}(D) \leq 1 \\ \mathcal{C}(W) \subseteq \mathcal{C}(D)}} \operatorname{trace}\left(D^\dagger W W^\top\right). \tag{2.8b}$$

Defining $C := WW^\top$, the minimizers are, respectively,

$$D_1^* := \sqrt{C} \quad \text{and} \quad D_2^* := \frac{\sqrt{C}}{\operatorname{trace}(\sqrt{C})}. \tag{2.9}$$

A proof sketch of the latter can be found in [AEP06, Thm 4.1]. A different proof, valid when C is positive definite, can also be found in [AEP08, Appendix A]. Adding the penalty $\epsilon \operatorname{trace}(D^\dagger)$ in either minimization, and factoring out $D^\dagger$, gives $C_\epsilon = WW^\top + \epsilon \mathrm{I}_d$ in the formula for the optimizers (2.9). The optimal values then change according to

$$\begin{aligned}\operatorname{trace}\left(\sqrt{WW^\top + \epsilon \mathrm{I}_d}\right) &= \operatorname{trace}\left(\sqrt{[W|\sqrt{\epsilon}\mathrm{I}_d][W|\sqrt{\epsilon}\mathrm{I}_d]^\top}\right) \\ &= \|[W|\sqrt{\epsilon}\mathrm{I}_d]\|_*,\end{aligned}$$

which is the nuclear norm of the block matrix comprised of W and $\sqrt{\epsilon}\mathrm{I}_d$. Also, for any $W \in \mathbb{R}^{d\times N}$, one has

$$\|W\|_* = \min_{U \in \mathbb{O}^d} \|W^\top U\|_{2,1}\,. \tag{2.10}$$

This result can be found in the proof by [AEP06, Thm 4.1], where we clarify that the cited reference is not consistent in the use of the notation $\|\cdot\|_{2,1}$ (mixing columns and rows).

For convenience, we also define the following sets that appear in the optimization problems of Chapter 6. For any $c, r \in \mathbb{R}_{>0}$, let

$$\mathfrak{D}(c,r) := \{D \in \mathbb{S}^d_{\succeq 0} : D \succeq c\mathrm{I},\, \|D\|_{\mathcal{F}} \le r\}\,, \tag{2.11a}$$

$$\mathit{\Delta}(c) := \{D \in \mathbb{S}^d_{\succeq 0} : D \succeq c\mathrm{I},\, \operatorname{trace}(D) \le 1\}\,. \tag{2.11b}$$

We refer to these sets as *reduced ice-cream* and *reduced spectraplex*, respectively, based on the fact that they correspond to the intersection of the *reduced* cone $\{D \in \mathbb{S}^d : D \succeq c\mathrm{I}_d\} \subseteq \mathbb{S}^d_{\succeq 0}$ with the ball given by the Frobenius norm and with the trace constraint, respectively.

2.5 Graph theory

The following notions in graph theory follow the exposition in [BCM09]. A (weighted) digraph $\mathcal{G} := (\mathcal{I}, \mathcal{E}, \mathsf{A})$ is a triplet where $\mathcal{I} := \{1, \ldots, N\}$ is the vertex set, $\mathcal{E} \subseteq \mathcal{I} \times \mathcal{I}$ is the edge set, and $\mathsf{A} \in \mathbb{R}^{N\times N}_{\geq 0}$ is the weighted adjacency matrix with the property that $\mathsf{a}_{ij} := A_{ij} > 0$ if and only if $(i, j) \in \mathcal{E}$. The complete graph is the digraph with edge set $\mathcal{I} \times \mathcal{I}$. Given $\mathcal{G}_1 = (\mathcal{I}, \mathcal{E}_1, \mathsf{A}_1)$ and $\mathcal{G}_2 = (\mathcal{I}, \mathcal{E}_2, \mathsf{A}_2)$, their union is the digraph $\mathcal{G}_1 \cup \mathcal{G}_2 = (\mathcal{I}, \mathcal{E}_1 \cup \mathcal{E}_2, \mathsf{A}_1 + \mathsf{A}_2)$. A path is an ordered sequence of vertices such that any pair of vertices appearing consecutively is an edge. A digraph is strongly connected if there is a path between any pair of distinct vertices. A sequence of digraphs $\left\{\mathcal{G}_t := (\mathcal{I}, \mathcal{E}_t, \mathsf{A}_t)\right\}_{t\geq 1}$ is δ-nondegenerate, for $\delta \in \mathbb{R}_{>0}$, if the weights are uniformly bounded away from zero by δ whenever positive, i.e., for each $t \in \mathbb{Z}_{\geq 1}$, $\mathsf{a}_{ij,t} := (\mathsf{A}_t)_{ij} > \delta$ whenever $\mathsf{a}_{ij,t} > 0$. A sequence $\{\mathcal{G}_t\}_{t\geq 1}$ is B-jointly connected, for $B \in \mathbb{Z}_{\geq 1}$, if for each $k \in \mathbb{Z}_{\geq 1}$, the digraph $\mathcal{G}_{kB} \cup \cdots \cup \mathcal{G}_{(k+1)B-1}$ is strongly connected. The (out-)Laplacian matrix $\mathsf{L} \in \mathbb{R}^{N\times N}$ of a digraph $\mathcal{G}$ is $\mathsf{L} := \mathrm{diag}(\mathsf{A}\mathbb{1}_N) - \mathsf{A}$. Note that $\mathsf{L}\mathbb{1}_N = 0$. The weighted out-degree and in-degree of $i \in \mathcal{I}$ are, respectively, $d_{\mathrm{out}}(i) := \sum_{j=1}^{N} \mathsf{a}_{ij}$ and $d_{\mathrm{in}}(i) := \sum_{j=1}^{N} \mathsf{a}_{ji}$. A digraph is weight-balanced if $d_{\mathrm{out}}(i) = d_{\mathrm{in}}(i)$ for all $i \in \mathcal{I}$, that is, $\mathbb{1}_N^\top \mathsf{L} = 0$, which is also equivalent to the condition of $\mathsf{L} + \mathsf{L}^\top$ being positive semidefinite. If $\mathcal{G}$ is weight-balanced and strongly connected, then $\mathsf{L} + \mathsf{L}^\top$ is positive semidefinite and $\mathcal{N}(\mathsf{L} + \mathsf{L}^\top) = \mathrm{span}\{\mathbb{1}_N\}$. For comparison purposes, we let $\mathsf{L}_{\mathcal{K}}$ denote the

Laplacian of the complete graph with edge weights $1/N$, i.e., $\mathsf{L}_{\mathcal{K}} := \mathrm{I}_N - \mathrm{M}$, where $\mathrm{M} := \frac{1}{N}\mathbb{1}_N\mathbb{1}_N^\top$. Note that $\mathsf{L}_{\mathcal{K}}$ is idempotent, i.e., $\mathsf{L}_{\mathcal{K}}^2 = \mathsf{L}_{\mathcal{K}}$. For the reader's convenience, Table 2.1 collects the shorthand notation combining Laplacian matrices and Kronecker products used in some chapters of the thesis.

Table 2.1: Shorthand notation for graph matrices employed throughout the thesis. Here, $\{\mathcal{G}_t\}_{t\geq 1}$, $K \in \mathbb{Z}_{\geq 1}$, and $E \in \mathbb{R}^{K\times K}$.

$\mathrm{M} = \frac{1}{N}\mathbb{1}_N\mathbb{1}_N^\top$	$\mathbf{M} = \mathrm{M} \otimes \mathrm{I}_d$	
$\mathsf{L}_{\mathcal{K}} = \mathrm{I}_N - \mathrm{M}$	$\mathbf{L}_{\mathcal{K}} = \mathsf{L}_{\mathcal{K}} \otimes \mathrm{I}_d$	$\hat{\mathbf{L}}_{\mathcal{K}} = \mathrm{I}_K \otimes \mathbf{L}_{\mathcal{K}}$
$\mathsf{L}_t = \mathrm{diag}(\mathsf{A}_t\mathbb{1}_N) - \mathsf{A}_t$	$\mathbf{L}_t = \mathsf{L}_t \otimes \mathrm{I}_d$	$\mathbb{L}_t = E \otimes \mathbf{L}_t$

2.6 Stochastic differential equations

This section is intended to provide the basic notation and results used in Chapter 3, including a distillation of the main result of Chapter 7. We relegate a more thorough introduction to the subject in Chapter 7. A stochastic differential equation (SDE) [Mao11, Ö10] is, roughly speaking, an ordinary differential equation driven by a "random process" called Brownian motion, $\mathrm{B} : \Omega \times [t_0, \infty) \to \mathbb{R}^m$. Here, Ω is the outcome space and $\mathbb{P}$ is a probability measure defined on the sigma-algebra $\mathcal{F}$ of measurable events (subsets) of Ω. These elements together form the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For each outcome $\omega \in \Omega$, the mapping $\mathrm{B}(\omega, .) : [t_0, \infty) \to \mathbb{R}^m$ is a sample path of the Brownian motion and is continuous with probability 1 and with $\mathrm{B}(., t_0) = 0$; and for each time $t \in [t_0, \infty)$, the function $\mathrm{B}(t) := \mathrm{B}(., t) : \Omega \to \mathbb{R}^m$ is a random variable such that the increments $\mathrm{B}(t) - \mathrm{B}(s)$ have a multivariate Gaussian distribution of zero mean and covariance $(t-s)\mathrm{I}_m$ and are independent from $\mathrm{B}(s)$ for all $t_0 \leq s < t$. Formally, we consider the SDE

$$\mathrm{d}x(\omega,t) = g(x(\omega,t),t)\mathrm{d}t + G(x(\omega,t),t)\Sigma(t)\mathrm{d}\mathrm{B}(\omega,t)\,, \tag{2.12}$$

where $x(\omega, t_0) = x_0$ with probability 1 for some $x_0 \in \mathbb{R}^n$. The vector field $g : \mathbb{R}^n \times [t_0, \infty) \to \mathbb{R}^n$ is sometimes called the drift, the matrix valued function $G : \mathbb{R}^n \times [t_0, \infty) \to \mathbb{R}^{n \times q}$ is the diffusion term and models the way in which the noise enters the dynamics, and the matrix $\Sigma : [t_0, \infty) \to \mathbb{R}^{q \times m}$ modulates the covariance of the noise. The matrix $\Sigma(t)\Sigma(t)^\top$ is called the infinitesimal covariance. The following result, from [Mao11, Th. 3.6, p. 58], guarantees the existence and uniqueness of solutions.

Lemma 6.1. (Existence and uniqueness). *Let g and G be measurable, and let Σ be measurable and essentially locally bounded. For any $T > t_0$ and $n \geq 1$, let $K_{T,n} \in \mathbb{R}_{>0}$ be such that, for almost every $t \in [t_0, T]$ and all $x, y \in \mathbb{R}^n$ with $\max\Big\{\|x\|_2, \|y\|_2\Big\} \leq n$, it holds that*

$$\max\Big\{\, \|g(x,t) - g(y,t)\|_2^2 \,,\, \|G(x,t) - G(y,t)\|_{\mathcal{F}}^2 \,\Big\} \leq K_{T,n} \|x - y\|_2^2.$$

Furthermore, assume that for any $T > t_0$, there exists $K_T > 0$ such that, for almost every $t \in [t_0, T]$ and all $x \in \mathbb{R}^n$,

$$x^\top g(x,t) + \tfrac{1}{2} \|G(x,t)\|_{\mathcal{F}}^2 \leq K_T (1 + \|x\|_2^2).$$

Then, the SDE (2.12) *enjoys global existence and uniqueness of solutions for each initial condition $x_0 \in \mathbb{R}^n$.*

In particular, under the hypotheses of Lemma 6.1, the solution inherits some properties of the Brownian motion. For instance, $x : \Omega \times [t_0, \infty) \to \mathbb{R}^n$ has continuous sample paths $x(\omega, .) : [t_0, \infty) \to \mathbb{R}^n$ with probability 1, and for each $t \geq t_0$, $x(t) := x(., t) : \Omega \to \mathbb{R}^n$ is a random variable with certain distribution (so that we are able to measure the probabilities of certain events that involve them).

Looking at (2.12), during a small time interval δ, the random outcome $x(\omega,t)$ changes approximately its value by an amount that is normally distributed with expectation $g(x(\omega,t))\delta$ and covariance $G(x(\omega,t),t)\Sigma(t)\Sigma(t)^\top G(x(\omega,t),t)^\top\delta$, and this change is independent of the previous history of the solution $\{x(s)\}_{s\leq t}$.

Next we introduce an important operator in the stability analysis of stochastic differential equations. For any twice continuously differentiable function $\mathrm{V}:\mathbb{R}^n \to \mathbb{R}$, we denote the generator of the SDE (2.12) acting on the function V as the mapping $\mathcal{L}[\mathrm{V}]:\mathbb{R}^n \times [t_0,\infty) \to \mathbb{R}$ given by

$$\mathcal{L}[\mathrm{V}](x,t) := \nabla\mathrm{V}(x)^\top g(x) + \tfrac{1}{2}\,\mathrm{trace}\Big(\Sigma(t)^\top G(x,t)^\top \nabla^2\mathrm{V}(x) G(x,t)\Sigma(t)\Big). \tag{2.13}$$

The above quantity is the expected rate of change of the function V along the solutions of the SDE (2.12) that take the value x at time t. Because of this, it can be considered a generalization of Lie derivative to SDEs. In fact, the SDE that the function $\mathrm{V}(x(\omega,t))$ itself satisfies, called Itô formula [Mao11, Th. 6.4, p. 36] contains the evaluation of (2.13) as a term. The following result provides a useful tool to study the stability properties of SDEs, and we use it in Chapter 3. It is a distilled version of one of our main results in Chapter 7.

Theorem 6.2. (Exponential pth moment noise-to-state stability). *Under the hypotheses of Lemma 6.1, further assume that Σ is continuous, and let $\mathrm{V}\in\mathcal{C}^2(\mathbb{R}^n;\mathbb{R}_{\geq 0})$ satisfy the following properties with respect to a closed set $\mathcal{U}\subseteq\mathbb{R}^n$: there exist $p>0$ and class $\mathcal{K}_\infty$ functions α_1 and α_2, where α_1 is convex, such that*

$$\alpha_1(|x|_{\mathcal{U}}^p) \leq \mathrm{V}(x) \leq \alpha_2(|x|_{\mathcal{U}}^p),$$

for all $x\in\mathbb{R}^n$, and there exist $\mathrm{W}\in\mathcal{C}(\mathbb{R}^n;\mathbb{R}_{\geq 0})$, $\sigma\in\mathcal{K}$, and concave $\eta\in\mathcal{K}_\infty$ such *that*

$$\mathcal{L}[\mathrm{V}](x,t) \leq -\mathrm{W}(x) + \sigma\Big(\|\Sigma(t)\|_{\mathcal{F}}\Big),$$

for all $(x,t) \in \mathbb{R}^n \times [t_0, \infty)$*, where, in addition,* $\mathrm{V}(x) \leq \eta(\mathrm{W}(x))$*, for all* $x \in \mathbb{R}^n$*. Then the system* (2.12) *is* pth moment noise-to-state stable ($p$thNSS) with respect to $\mathcal{U}$*, i.e., there exist* $\mu \in \mathcal{KL}$ *and* $\theta \in \mathcal{K}$ *such that*

$$\mathbb{E}\Big[|x(t)|_{\mathcal{U}}^p\Big] \leq \mu\Big(|x_0|_{\mathcal{U}}, t - t_0\Big) + \theta\Big(\max_{t_0 \leq s \leq t} \|\Sigma(s)\|_{\mathcal{F}}\Big),$$

for all $t \geq t_0$ *and any* $x_0 \in \mathbb{R}^n$*. Specifically,* $\mu(r,s) := \alpha_1^{-1}\Big(2\tilde{\mu}(\alpha_2(r^p), s)\Big)$ *and* $\theta(r) := \alpha_1^{-1}\Big(2\eta(2\sigma(r))\Big)$*, where the class* $\mathcal{KL}$ *function* $(r,s) \mapsto \tilde{\mu}(r,s)$ *is well defined as the solution* $y(s)$ *to the initial value problem*

$$\dot{y}(s) = -\tfrac{1}{2}\eta^{-1}(y(s)), \quad y(0) = r.$$

We refer to the function V satisfying the hypotheses of this result as a *pth moment NSS-Lyapunov function with respect to* $\mathcal{U}$ for the system (2.12). If the functions α_1 and η are linear, then we refer to the above property as *pth moment noise-to-state exponential stability.*

Chapter 3

Noise-to-state exponentially stable distributed convex optimization

Our first technical chapter considers the scenario where the agents need to agree on a global decision vector that minimizes an unconstrained sum of convex functions. We study a family of distributed, continuous-time algorithms that have each agent update its estimate of the global optimizer doing gradient descent on its local cost function while, at the same time, seeking to agree with its neighbors' estimates via proportional-integral feedback on their disagreement. Our aim is to characterize the algorithm robustness properties against the additive persistent noise resulting from the errors in communication and computation. We model this algorithm with a stochastic differential equation and apply a novel Lyapunov technique to establish the noise-to-state stability property in 2nd moment.

3.1 Network model and problem statement

This section describes the model for the network of agents and the optimization problem we set out to solve in a distributed way. Consider a group of N agents with identities $\{1,\dots,N\}$ whose communication topology is modeled by a strongly connected and weight-balanced digraph $\mathcal{G}$. An edge $(i,j)\in\mathcal{E}$ represents the ability of agent i to receive information sent from agent j. We consider scenarios where the inter-agent communication is corrupted by Gaussian white noise signals. Our description here is only meant to motivate the rigorous formalization of the dynamics presented in the forthcoming section using stochastic differential equations. For now, if agent j sends the signal $x(t)\in\mathbb{R}^d$ to agent i at time $t\geq t_0$, agent i receives the corrupted signal

$$x(t)+\mathbf{J}^{ij}(t)\,W_{\mathrm{cmm}}^{(i,j)}(\omega,t), \tag{3.1}$$

where the vector $W_{\mathrm{cmm}}^{(i,j)}(\omega,t)\in\mathbb{R}^d$ contains d independent Gaussian white noise signals, and $\mathbf{J}^{ij}(t)\in\mathbb{R}^{d\times d}$ is a weighting matrix. The noise we consider is additive, might be always present no matter what the value of the transmitted signal is, and we call it persistent because it is not assumed to decay with time. Our forthcoming algorithm design does not require that agent $i\in\{1,\dots,N\}$ knows the weighting matrices $\mathbf{J}^{ij}$ for any $(i,j)\in\mathcal{E}$. We also consider the possibility of the information available to any given agent being corrupted by noise when incorporated into its computations. Specifically, if agent i attempts to incorporate the quantity $q_i(t)\in\mathbb{R}^d$ into its computations, what the agent instead uses is

$$q_i(t)+\tilde{\mathbf{J}}^{ii}(t)\,W_{\mathrm{cmp}}^{i}(\omega,t), \tag{3.2}$$

where the entries of $W^i_{\mathrm{cmp}}(\omega,t) \in \mathbb{R}^d$ are independent Gaussian white noise signals, and $\tilde{\mathbf{J}}^{ii}(t) \in \mathbb{R}^{d\times d}$ is a weighting matrix. As before, our algorithm design does not require that agent $i \in \{1,\ldots,N\}$ knows the weighting matrix $\tilde{\mathbf{J}}^{ii}$.

With the model for the network in place, we next define the network objective. Consider a function $f : \mathbb{R}^d \to \mathbb{R}$ of the form

$$f(x) = \sum_{i=1}^{N} f_i(x), \tag{3.3}$$

where the local function $f_i : \mathbb{R}^d \to \mathbb{R}$ is only known to agent $i \in \{1,\ldots,N\}$. We assume each f_i is convex and that at least one of them is strongly convex, so that the function f has a unique minimizer, which we denote by $x_{\min} \in \mathbb{R}^d$. Our goal is to design a distributed continuous-time coordination algorithm that helps the network collectively find the minimizer $x_{\min}$ in the presence of noise both in the communication channels and in the agent computations.

3.2 Robust distributed optimization

This section introduces a distributed coordination algorithm that allows the network of agents to solve the optimization problem as described in Section 7.2. Our study here generalizes the work in [GC14] to scenarios where the communication channels and the computations performed by the agents are subject to noise. In order to synthesize a strategy that allows the network to agree on the solution of the optimization problem, we have each agent $i \in \{1,\ldots,N\}$ keep an estimate $x^i \in \mathbb{R}^d$ about the minimizer of the function f in (3.3). For convenience, we denote by $\boldsymbol{x} := [(x^1)^\top, \ldots, (x^N)^\top]^\top \in (\mathbb{R}^d)^N$ the collection of estimates across the network and consider the function $\tilde{f} : (\mathbb{R}^d)^N \to \mathbb{R}$ defined by

$$\tilde{f}(\boldsymbol{x}) := \sum_{i=1}^{N} f_i(x^i). \tag{3.4}$$

In this computation, each agent can evaluate f_i at its own estimate x^i and the network objective function in (3.3) can be evaluated when agreement holds, $\tilde{f}(\mathbb{1} \otimes x) = f(x)$.

The continuous-time algorithm we consider is then given by the following system of stochastic differential equations,

$$\mathrm{d}\boldsymbol{x} = -(\nabla \tilde{f}(\boldsymbol{x}) + \tilde{\gamma}\mathbf{L}\boldsymbol{x} + \mathbf{L}\boldsymbol{z})\mathrm{d}t + G^1(\boldsymbol{x}, \boldsymbol{z}, t)\Sigma^1(t)\mathrm{dB}(t), \tag{3.5a}$$

$$\mathrm{d}\boldsymbol{z} = \mathbf{L}\boldsymbol{x}\mathrm{d}t + G^2(\boldsymbol{x}, \boldsymbol{z}, t)\Sigma^2(t)\mathrm{dB}(t), \tag{3.5b}$$

where we use the shorthand notation $\mathbf{L} := \mathsf{L} \otimes \mathrm{I}_d$ and L is the Laplacian of the digraph $\mathcal{G}$ modeling inter-agent communication. We assume that the matrix-valued functions $G^1, G^2 : \mathbb{R}^{2Nd} \times [t_0, \infty) \to \mathbb{R}^{Nd \times q}$ are uniformly bounded and uniformly globally Lipschitz in the first two arguments, and measurable and essentially bounded in time. Also, we assume that the matrix-valued functions $\Sigma^1, \Sigma^2 : [t_0, \infty) \to \mathbb{R}^{q \times m}$, with $m \geq 1$, are continuous and locally bounded and that $\{\mathrm{B}(t)\}_{t \geq t_0}$ is an m-dimensional Brownian motion defined in the probability space.

We next provide some intuition behind the algorithm design in (3.5) and properly justify its distributed character. The deterministic part of the dynamics prescribes that each agent updates its estimate by following the gradient of its local cost function while, at the same time, seeking to agree with its neighbors' estimates. The latter is implemented through a second-order system of differential equations that involves the auxiliary variables $\boldsymbol{z} := [(z^1)^\top, \ldots, (z^N)^\top]^\top \in (\mathbb{R}^d)^N$ and employs proportional-integral feedback on the disagreement. When the graph $\mathcal{G}$

is undirected, one can in fact see [GC14] that the deterministic part corresponds exactly to the saddle-point dynamics associated with the augmented Lagrangian $\mathcal{L}(\boldsymbol{x},\boldsymbol{z}) = \tilde{f}(\boldsymbol{x}) + \tilde{\gamma}\,\boldsymbol{x}^\top \mathbf{L}\boldsymbol{x} + \boldsymbol{z}^\top \mathbf{L}\boldsymbol{x}$, corresponding to the minimization of $\tilde{f}$ under the constraints $\mathbf{L}\boldsymbol{x} = 0$. The stochastic part of the dynamics (3.5) is motivated by the desire to capture the presence of noise affecting the execution of the coordination algorithm. In particular, Remark 2.3 below discusses how the noise model described in Section 7.2 affecting the communication channels and the agent computations is captured by the stochastic differential equation (3.5). Finally, the dynamics is distributed over the digraph $\mathcal{G}$ because each agent $i \in \{1,\dots,N\}$ can update its variables x^i and z^i using only the information sent from its neighbors and its knowledge of its local function f_i. This is not difficult to see from the observation that the gradient of $\tilde{f}$ takes the form $\nabla \tilde{f}(\boldsymbol{x}) = [\nabla f_1(x^1)^\top, \dots, \nabla f_N(x^N)^\top]^\top$ and that the agent i can compute the ith d-dimensional block $(\mathbf{L}\boldsymbol{x})^i \in \mathbb{R}^d$.

Remark 2.3. (Noise model for communication and computation is captured by the dynamics (3.5)). Although the dynamics described by (3.5) cannot be exactly implemented in practice, it is a reasonable model of evolution in continuous time with network communications also in continuous time. We justify this statement here as follows. When communication along an edge $(i,j) \in \mathcal{E}$ occurs continuously over time, the model (3.1) gives rise to functions $\mathbf{J}^{ij} : [t_0,\infty) \to \mathbb{R}^{d\times d}$, which we assume measurable and essentially locally bounded, and $W^{(i,j)}_{\text{cmm}} : \Omega \times [t_0,\infty) \to \mathbb{R}^d$. Similarly, when considering continuous-time dynamics, the computation model (3.2) gives rise to functions $\tilde{\mathbf{J}}^{ii} : [t_0,\infty) \to \mathbb{R}^{d\times d}$, which we also assume measurable and essentially locally bounded, and $W^i_{\text{cmp}} : \Omega \times [t_0,\infty) \to \mathbb{R}^d$. Under this noise model, the implementation of the dynamics $\dot{\boldsymbol{x}} = -(\nabla \tilde{f}(\boldsymbol{x}) + \tilde{\gamma}\mathbf{L}\boldsymbol{x} + \mathbf{L}\boldsymbol{z})$ and $\dot{\boldsymbol{z}} = \mathbf{L}\boldsymbol{x}$ by the agent i actually results in the dynamics,

$$\begin{aligned}
\mathrm{d}x^i(t) = \tilde{\gamma} \sum_{j=1}^{N} \mathsf{a}_{ij} \Big(\big(x^j(t) - x^i(t) \big) \mathrm{d}t + \mathbf{J}^{ij}(t) \mathrm{dB}^{1,(i,j)}(t) \Big) \\
+ \sum_{j=1}^{N} \mathsf{a}_{ij} \Big(\big(z^j(t) - z^i(t) \big) \mathrm{d}t + \mathbf{J}^{ij}(t) \mathrm{dB}^{2,(i,j)}(t) \Big) \\
- \nabla f_i(x^i(t)) \mathrm{d}t - \tilde{\mathbf{J}}^{ii}(t) \mathrm{dB}^{3,i}(t), \qquad (3.6a) \\
\mathrm{d}z^i(t) = - \sum_{j=1}^{N} \mathsf{a}_{ij} \Big(\big(x^j(t) - x^i(t) \big) \mathrm{d}t + \mathbf{J}^{ij}(t) \mathrm{dB}^{1,(i,j)}(t) \Big), \qquad (3.6b)
\end{aligned}$$

where $\mathrm{B}^{1,(i,j)}$, $\mathrm{B}^{2,(i,j)}$ and $\mathrm{B}^{3,i}$ are independent d-dimensional Brownian motions for each edge $(i,j) \in \mathcal{E}$ and each agent $i \in \{1, \ldots, N\}$, respectively. We next show how this dynamics is captured by (3.5). First, we set $G^1(\boldsymbol{x}, \boldsymbol{z}, t) = G^2(\boldsymbol{x}, \boldsymbol{z}, t) = \mathrm{I}_{Nd}$ for all $\boldsymbol{x}$, $\boldsymbol{z}$, t. Second, let $\mathbf{J}(t) \in \mathbb{R}^{Nd \times Nd}$ be the matrix whose (i,j)th d-dimensional block is $\mathsf{a}_{ij}\mathbf{J}^{ij}(t)$ and $\tilde{\mathbf{J}}(t) = \operatorname{diag}\Big(\tilde{\mathbf{J}}_{11}(t), \ldots, \tilde{\mathbf{J}}_{NN}(t)\Big) \in \mathbb{R}^{Nd \times Nd}$. Define $\hat{\Sigma}^1(t) := \begin{bmatrix} \tilde{\gamma}\mathbf{J}(t) & \mathbf{J}(t) & -\tilde{\mathbf{J}}(t) \end{bmatrix} \in \mathbb{R}^{Nd \times 3Nd}$ and $\hat{\Sigma}^2(t) := \begin{bmatrix} -\mathbf{J}(t) & 0 & 0 \end{bmatrix} \in \mathbb{R}^{Nd \times 3Nd}$, and set

$$\begin{bmatrix} \Sigma^1(t) \\ \Sigma^2(t) \end{bmatrix} := \begin{bmatrix} \Big((e_1 e_1^\top) \otimes \mathrm{I}_d\Big)\hat{\Sigma}^1(t) & \cdots & \Big((e_N e_N^\top) \otimes \mathrm{I}_d\Big)\hat{\Sigma}^1(t) \\ \Big((e_1 e_1^\top) \otimes \mathrm{I}_d\Big)\hat{\Sigma}^2(t) & \cdots & \Big((e_N e_N^\top) \otimes \mathrm{I}_d\Big)\hat{\Sigma}^2(t) \end{bmatrix} \in \mathbb{R}^{2Nd \times 3N^2 d}.$$

Then, the dynamics (3.5) with this selection of functions G^1, G^2, Σ^1, and Σ^2 corresponds to (3.6). •

The main result of the paper is the characterization of the asymptotic stability properties of the stochastic differential equation (3.5) with respect to the solution of the optimization problem. To achieve this, we rely on selecting appropriately the design parameter $\tilde{\gamma}$. This is described precisely in the following assumption.

Assumption 2.4. (Selection of the parameter $\tilde{\gamma}$). *Given any $\epsilon > 0$, let $K_1 :=$*

$\lambda_{min}\Big(r\, e_{i_0}e_{i_0}^\top + \epsilon\,(\mathsf{L}+\mathsf{L}^\top)\Big)$ and $K_2 := R + 2\epsilon\,\sigma_{max}(\mathsf{L})$, and, for any $\delta \in (0, K_1K_2^{-2})$, let $\beta_1^* \equiv \beta_1^*(\delta,\epsilon) := \sqrt{K_1^2K_2^{-2} - K_1\delta}$ and $\beta_2^* \equiv \beta_2^*(\delta)$ be such that $h(\beta,\delta) < 0$ for $\beta \in (0, \beta_2^*(\delta))$, where

$$h(\beta,\delta) := \Big(-\tfrac{\beta^4+3\beta^2+2}{2\beta} + \sqrt{\Big(\tfrac{\beta^4+3\beta^2+2}{2\beta}\Big)^2 - 1}\;\Big)\lambda_2(\mathsf{L}+\mathsf{L}^\top) + \tfrac{\beta^2}{2\delta}. \qquad (3.7)$$

Under the above selections, select the design parameter $\tilde{\gamma}$ as,

$$\tilde{\gamma}(\epsilon,\delta) := \tfrac{2+\beta^2}{\beta} + 2\epsilon, \qquad \beta \in (0, \min\{\beta_1^*(\delta,\epsilon), \beta_2^*(\delta)\}).$$

Note that the selection of $\tilde{\gamma}$ is determined by the bounds on the Hessians of the objective functions and the network topology. The reasons behind the specific form of the functions employed in Assumption 2.4 will become fully clear later in our technical derivations, but we provide the basic insight in Remark 2.7 below.

Our main result states that, under Assumption 2.4, the dynamics of $\boldsymbol{x}$ is noise-to-state exponentially stable in second moment with respect to $\mathbb{1} \otimes x_{\min}$.

Theorem 2.5. (Exponential noise-to-state stability of the dynamics (3.5)). *Assume the functions $\{f_i\}_{i=1}^N$ are convex and twice continuously differentiable with uniformly upper-bounded Hessians, i.e., there exists $R > 0$ such that $0 \preccurlyeq \nabla^2 f_i \preccurlyeq R\,\mathrm{I}_d$, for $i \in \{1,\dots,N\}$. Further assume that at least one of the functions is strongly convex, i.e., there exists $r > 0$ such that $r\,\mathrm{I}_d \preccurlyeq \nabla^2 f_{i_0}$ for some $i_0 \in \{1,\dots,N\}$. If, in addition, $\tilde{\gamma}$ is selected according to Assumption 2.4, then the dynamics* (3.5) *executed over a strongly connected and weight-balanced digraph has the following stability property: there exist constants C_μ, D_μ, $C_\theta > 0$ such that, for any initial condition*

$(\boldsymbol{x}_0, \boldsymbol{z}_0) \in (\mathbb{R}^d)^N \times (\mathbb{R}^d)^N$ *and all* $t \geq t_0$*, it holds that*

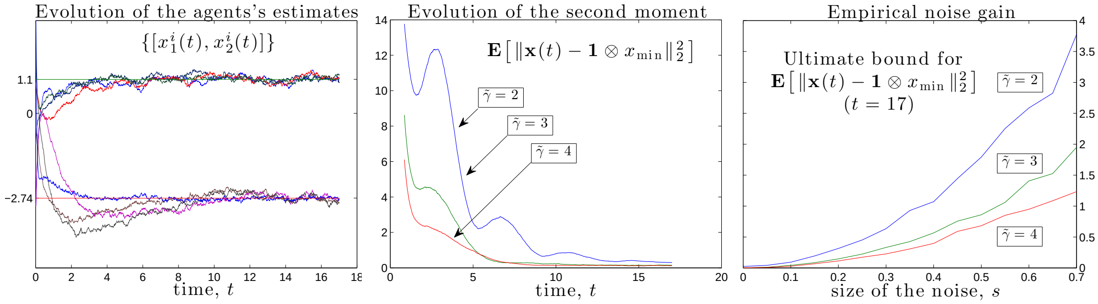
$$\begin{aligned}\mathbb{E}\Big[\|\boldsymbol{x}(t) - \mathbb{1} \otimes x_{\min}\|_2^2\Big] &\leq \mathbb{E}\Big[\|\boldsymbol{x}(t) - \mathbb{1} \otimes x_{\min}\|_2^2 + \|\boldsymbol{z}(t) - \boldsymbol{z}^*\|_{\mathbf{L}_{\mathcal{K}}}^2\Big] \\ &\leq C_\mu (\|\boldsymbol{x}_0 - \mathbb{1} \otimes x_{\min}\|_2^2 + \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|_{\mathbf{L}_{\mathcal{K}}}^2)\, e^{-D_\mu (t-t_0)} + C_\theta \Big(\max_{t_0 \leq \tau \leq t} \|\Sigma(\tau)\|_{\mathcal{F}}\Big)^2, \quad (3.8)\end{aligned}$$

where $\mathbf{L}_{\mathcal{K}} := \mathsf{L}_{\mathcal{K}} \otimes \mathrm{I}_d$*,* $\Sigma(t) := [\Sigma^1(t)^\top, \Sigma^2(t)^\top]^\top$*,* $x_{\min} \in \mathbb{R}^d$ *is the unique minimizer of* (3.3)*, and* $\boldsymbol{z}^* \in \mathbb{R}^d$ *is any point satisfying* $\mathbf{L}\boldsymbol{z}^* = -\nabla \tilde{f}(\mathbb{1} \otimes x_{\min})$*.*

The expression (3.8) states that the dynamics (3.5) is noise-to-state stable in second moment with respect to the affine subspace of equilibria. In other words, the agreement direction of the agents' auxiliary states in $\boldsymbol{z}$ absorbs the cumulative variance of the noise while the estimates in $\boldsymbol{x}$ converge asymptotically, in second moment, to a neighborhood of the minimizer of (3.3). The size of this neighborhood depends on the size of the noise, quantified by $\|\Sigma(t)\|_{\mathcal{F}} := \sqrt{\mathrm{trace}\,(\Sigma(t)\Sigma(t)^\top)}$, which is related to the infinitesimal covariance $\Sigma(t)\Sigma(t)^\top$.

Example 2.6 (4-agent network over directed ring)**.** Here we briefly illustrate the results of Theorem 2.5. Consider the evolution of the distributed algorithm (3.5) with noise over a group of $N = 4$ agents communicating over a directed ring with edge set $\mathcal{E} = \{(1,3), (3,2), (2,4), (4,1)\}$. This digraph is indeed strongly connected and weight-balanced, with Laplacian matrix

$$\mathsf{L} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}$$



(a) (b) (c)

Figure 3.1: Simulation example of our distributed continuous-time algorithm (3.5) under persistent noise. The three plots correspond to the 4-agent network in Example 2.50. Plot (a) shows the evolution of the first and second coordinates of the agents' estimates with $\tilde{\gamma} = 3$, $G^1 = G^2 = \mathrm{I}_8$, and $\Sigma^1 = \Sigma^2 = 0.2\,\mathrm{I}_8$. Despite the additive persistent noise, the estimates converge, in probability, to a neighborhood of the minimizer $x_{\min} = (1.10, -2.74)$. For three different values of the design parameter $\tilde{\gamma}$, plot (b) shows the asymptotic convergence in second moment to a neighborhood of the solution. Plot (c) depicts the ultimate bound for the second moment of the error when the size of the noise varies as $\Sigma^1 = \Sigma^2 = s\,\mathrm{I}_8$, with s ranging from 0 to 0.7 with increments of 0.05. It is worth observing that, as the design parameter gets larger (putting more emphasis on consensus among the agents) the effective error gets smaller. In all plots, the initial conditions are $\boldsymbol{x}_0 = (-3, -3, -1, -1, 1, 1, 3, 3)$, and $\boldsymbol{z}_0 = \mathbb{1}_8$. The dynamics is simulated using the Euler discretization with stepsize 0.01, and the expectations are computed averaging over 100 realizations of the noise.

The local objective functions, defined on $\mathbb{R}^2$, are given by

$$f_1(x_1,x_2) = \tfrac{1}{2}((x_1-4)^2+(x_2-3)^2), \quad f_2(x_1,x_2) = x_1+3x_2-2,$$

$$f_3(x_1,x_2) = \log(e^{x_1+3}+e^{x_2+1}), \quad f_4(x_1,x_2) = (x_1+2x_2+5)^2+(x_1-x_2-4)^2.$$

The first set of hypotheses of Theorem 2.5 concerns the Hessians

$$\nabla^2 f_1(x_1,x_2) = \mathrm{I}_2, \qquad \nabla^2 f_2(x_1,x_2) = 0_{2\times 2}, \qquad \nabla^2 f_4(x_1,x_2) = \begin{bmatrix} 4 & 2 \\ 2 & 10 \end{bmatrix},$$

$$\nabla^2 f_3(x_1,x_2) = \begin{bmatrix} \frac{u^2}{(u+v)^2} & \frac{uv}{(u+v)^2} \\ \frac{uv}{(u+v)^2} & \frac{v^2}{(u+v)^2} \end{bmatrix} + \begin{bmatrix} \frac{u}{u+v} & 0 \\ 0 & \frac{v}{u+v} \end{bmatrix},$$

where $u := e^{x_1+3} > 0$ and $v := e^{x_2+1} > 0$. Upper bounds for the Hessians can be found as $\nabla^2 f_3(x_1,x_2) \preceq 2\mathrm{I}_2$ and $\nabla^2 f_4(x_1,x_2) \preceq 11\mathrm{I}_2$ for all $(x_1,x_2) \in \mathbb{R}^2$. Hence, $\nabla^2 f_i \preceq 11\mathrm{I}_2$ for all $i \in \{1,\dots,4\}$, i.e., $R = 11$. Also, at least one Hessian is lower bounded; for instance, $\nabla^2 f_1 \succeq 1\mathrm{I}_2$, i.e., $r = 1$. Using this information, we compute now an interval for $\tilde{\gamma}$ following Assumption 2.4. Taking $\epsilon = 0.1$, we obtain $K_1 \approx 0.05$ and $K_2 = 11.4$. Choosing then $\delta = 0.98 K_1 K_2^{-2}$ so that $\beta_1^*(\delta,\epsilon) \approx 6.1 \times 10^{-4}$ is approximately the biggest attainable value smaller than $\beta_2^*(\delta)$, one gets $\tilde{\gamma} \in [3.25 \times 10^3, \infty)$. Our experiments suggest that this range is conservative because we observe a correct behavior for values as comparatively low as $\tilde{\gamma} = 3$. Figure 7.1 illustrates the evolution of our algorithm using several realizations of the noise and also shows the bounds on the error as a function of the size of the noise. •

Remark 2.7. (Dependencies of the constant C_θ). It is worth observing that the constant C_θ in (3.8) is *independent* of the infinitesimal covariance of the noise. Hence, the size of the noise has a quadratic influence on the ultimate error bound. The explicit expression depends on the bounds on the Hessians of the objective functions and the network topology as follows,

$$C_\theta = \frac{4\kappa_2^2}{\min\{1, K_1(1+K_2^2)^{-1}\}} \frac{\max\{1, \lambda_{\max}(\mathsf{L}+\mathsf{L}^\top)\}}{\min\{1, \lambda_{N-1}(\mathsf{L}+\mathsf{L}^\top)\}} \frac{\lambda_{\max}(\mathrm{P}_\beta)}{\lambda_{(2N-1)d}(\mathrm{P}_\beta)} \frac{\mathrm{trace}(\mathrm{P}_\beta)}{\lambda_{(3N-2)d}(\mathrm{Q}_\beta)},$$

where κ_2 is a global bound on the functions G^1 and G^2 (in the sense of essential supremum with respect to the last argument) and

$$\mathrm{P}_\beta := \begin{bmatrix} \mathrm{I}+\beta^2\mathbf{L}_{\mathcal{K}} & \beta\mathbf{L}_{\mathcal{K}} \\ \beta\mathbf{L}_{\mathcal{K}} & \mathbf{L}_{\mathcal{K}} \end{bmatrix} \in \mathbb{R}^{2Nd\times 2Nd},$$

$$\mathrm{Q}_\beta := \begin{bmatrix} \begin{bmatrix} \beta^3+2\beta+\frac{2}{\beta} & (1+\beta^2) \\ (1+\beta^2) & \beta \end{bmatrix} \otimes (\mathbf{L}+\mathbf{L}^\top) & \begin{matrix} 0 \\ \beta\mathbf{L}_{\mathcal{K}} \end{matrix} \\ \begin{matrix} 0 & \beta\mathbf{L}_{\mathcal{K}} \end{matrix} & 2\delta\mathrm{I} \end{bmatrix} \in \mathbb{R}^{3Nd\times 3Nd}.$$

Interestingly, C_θ is *also independent* of the parameter $\tilde{\gamma}$. The above matrices play a crucial role in our technical approach. In a nutshell, our candidate Lyapunov function V is a quadratic function defined by P_β and the generator of the SDE (3.5) acting on this function, $\mathcal{L}[\mathrm{V}]$, is bounded by a quadratic function W defined by the matrix Q_β in an embedding of $\mathbb{R}^{2Nd}$ in $\mathbb{R}^{3Nd}$. Thus, the matrices P_β and Q_β are key in characterizing the pth moment NSS-Lyapunov function in the hypotheses of Theorem 6.2. In particular, the value of the design parameter $\tilde{\gamma}$ is chosen to establish the negative semidefiniteness of Q_β. •

We devote Section 3.3 to prove Theorem 2.5, where we provide explicit characterizations of the class $\mathcal{KL}$ function $\mu(r,s) := C_\mu r^2 e^{-D_\mu s}$ and also derive the class $\mathcal{K}_\infty$ function $\theta(r) := C_\theta r^2$. We end this section by noting that, in the noiseless case, a byproduct of Theorem 2.5 is a refinement of the result in [GC14], showing exponential convergence to the solution.

Corollary 2.8. (Global exponential stability in the noiseless case). *In the noiseless case (i.e., $\Sigma^1 = \Sigma^2 = 0$), and under the hypotheses of Theorem 2.5, the trajectory of the dynamics* (3.5) *starting from an arbitrary initial condition* $(\boldsymbol{x}_0, \boldsymbol{z}_0) \in (\mathbb{R}^d)^N \times$

$(\mathbb{R}^d)^N$ *satisfies, for all* $t \geq t_0$,

$$\|\boldsymbol{x}(t) - \mathbb{1} \otimes x_{\min}\|_2^2 \leq \|\boldsymbol{x}(t) - \mathbb{1} \otimes x_{\min}\|_2^2 + \|\boldsymbol{z}(t) - \boldsymbol{z}^*\|_2^2$$
$$\leq C_\mu(\|\boldsymbol{x}_0 - \mathbb{1} \otimes x_{\min}\|_2^2 + \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|_{\mathbf{L}_\mathcal{K}}^2)e^{-D_\mu(t-t_0)} + \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|_{\mathbf{M}}^2, \tag{3.9}$$

where $\mathbf{M} = \frac{1}{N}\mathbb{1}\mathbb{1}^\top \otimes \mathrm{I}_d$, *and* $\boldsymbol{z}^* \in \mathbb{R}^d$ *is any point satisfying* $\mathbf{L}\boldsymbol{z}^* = -\nabla\tilde{f}(\mathbb{1} \otimes x_{\min})$. *In particular, choosing* $\boldsymbol{z}^* \in \mathbb{R}^d$ *such that* $\mathbf{M}\boldsymbol{z}^* = \mathbf{M}\boldsymbol{z}_0$ *shows that the convergence of the trajectory starting from* $(\boldsymbol{x}_0, \boldsymbol{z}_0)$ *to the point* $(\mathbb{1} \otimes x_{\min}, \boldsymbol{z}^*)$ *is exponential.*

Proof. Since $\Sigma^1 = \Sigma^2 = 0$, the system of SDEs (3.5) becomes a system of ordinary differential equations. Let $\boldsymbol{z}_{\text{agree}}(t) := \mathbf{M}\boldsymbol{z}(t)$. By left-multiplying the dynamics of $\boldsymbol{z}(t)$ in (3.5) by $\mathbf{M}$, we obtain that $\dot{\boldsymbol{z}}_{\text{agree}} = 0$ and therefore $\boldsymbol{z}_{\text{agree}}(t) = \boldsymbol{z}_{\text{agree}}(t_0)$ for all $t \geq t_0$. Using that $\mathbf{M}$ is symmetric and $\mathbf{M} = \mathbf{M}^2$, if we define $\boldsymbol{z}^*_{\text{agree}} := \mathbf{M}\,\boldsymbol{z}^*$, then

$$\begin{aligned}(\boldsymbol{z}(t) - \boldsymbol{z}^*)^\top \mathbf{M}(\boldsymbol{z}(t) - \boldsymbol{z}^*) &= (\boldsymbol{z}_{\text{agree}}(t) - \boldsymbol{z}^*_{\text{agree}})^\top \mathbf{M}(\boldsymbol{z}_{\text{agree}}(t) - \boldsymbol{z}^*_{\text{agree}}) \\ &= (\boldsymbol{z}_{\text{agree}}(t_0) - \boldsymbol{z}^*_{\text{agree}})^\top \mathbf{M}\,(\boldsymbol{z}_{\text{agree}}(t_0) - \boldsymbol{z}^*_{\text{agree}}) \\ &= (\boldsymbol{z}_0 - \boldsymbol{z}^*)^\top \mathbf{M}\,(\boldsymbol{z}_0 - \boldsymbol{z}^*) = \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|_{\mathbf{M}}^2.\end{aligned}$$

On the other hand, using that $\mathrm{I}_{Nd} = \mathbf{L}_\mathcal{K} + \mathbf{M}$ and $\mathbf{L}_\mathcal{K}^2 = \mathbf{L}_\mathcal{K}$, we obtain

$$\|\boldsymbol{z}(t) - \boldsymbol{z}^*\|_2^2 = (\boldsymbol{z}(t) - \boldsymbol{z}^*)^\top \Big(\mathbf{L}_\mathcal{K} + \mathbf{M}\Big)(\boldsymbol{z}(t) - \boldsymbol{z}^*) = \|\boldsymbol{z}(t) - \boldsymbol{z}^*\|_{\mathbf{L}_\mathcal{K}}^2 + \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|_{\mathbf{M}}^2.$$

Equation (3.9) follows from this fact together with (3.8). Finally, noting that $\mathbf{M}\boldsymbol{z} = \mathbb{1} \otimes (\frac{1}{N}\sum_{i=1}^N z^i)$ and $\mathbf{L}(\mathbb{1} \otimes a) = 0$ for any $a \in \mathbb{R}^d$, it is clear that, given an initial condition $\boldsymbol{z}_0 \in (\mathbb{R}^d)^N$, one can choose $\boldsymbol{z}^*$ that satisfies at the same time $\mathbf{L}\boldsymbol{z}^* = -\nabla\tilde{f}(\mathbb{1} \otimes x_{\min})$ and $\mathbf{M}\boldsymbol{z}^* = \mathbf{M}\boldsymbol{z}_0$. If this is the case, $\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|_{\mathbf{M}} = 0$, and (3.9) shows exponential convergence of the trajectory starting from $(\boldsymbol{x}_0, \boldsymbol{z}_0)$ to $(\mathbb{1} \otimes x_{\min}, \boldsymbol{z}^*)$. □

3.3 Algorithm properties and stability analysis

In this section, we establish a series of properties of the distributed coordination algorithm (3.5) leading up to the characterization of its asymptotic correctness stated in Theorem 2.5. We begin by expressing the dynamics in compact form. Let $v := (\boldsymbol{x}^\top, \boldsymbol{z}^\top)^\top \in \mathbb{R}^{2Nd}$ and consider

$$\begin{aligned} \mathrm{d}v &= (\mathbf{A}v + \mathbf{N}(\boldsymbol{x}))\mathrm{d}t + G(v,t)\Sigma(t)\mathrm{dB} \\ &:= \left(\begin{bmatrix} -\gamma\mathbf{L} & -\mathbf{L} \\ \mathbf{L} & 0 \end{bmatrix} v + \begin{bmatrix} -\nabla \tilde{f}(\boldsymbol{x}) - 2\epsilon \mathbf{L}x \\ 0 \end{bmatrix} \right) \mathrm{d}t + \begin{bmatrix} G^1(\boldsymbol{x},\boldsymbol{z},t) & 0 \\ 0 & G^2(\boldsymbol{x},\boldsymbol{z},t) \end{bmatrix} \Sigma(t)\mathrm{dB}, \end{aligned} \tag{3.10}$$

where, for convenience, we have split the parameter $\tilde{\gamma}$ as $\tilde{\gamma} = \gamma + 2\epsilon$. This dynamics fits the model (2.12) with $g(v) := \mathbf{A}v + \mathbf{N}(\boldsymbol{x})$. As mentioned earlier, $\Sigma : [t_0, \infty) \to \mathbb{R}^{q \times m}$ is continuous and locally bounded, and $G : \mathbb{R}^{2Nd} \times [t_0, \infty) \to \mathbb{R}^{2Nd \times q}$ is measurable in time, uniformly globally Lipschitz in the first argument, say with Lipschitz constant $\kappa_1 \in \mathbb{R}_{>0}$, and bounded in its domain (essentially in time) by $\kappa_2 \in \mathbb{R}_{>0}$. Formally,

$$\|G(v,t) - G(v',t)\|_{\mathcal{F}} \leq \kappa_1 \|v - v'\|_2, \qquad \sup_{v \in \mathbb{R}^{2Nd}} \operatorname*{ess\,sup}_{t \geq t_0} \|G(v,t)\|_{\mathcal{F}} \leq \kappa_2, \tag{3.11}$$

for all $v, v' \in \mathbb{R}^{2Nd}$.

With this notation in place, we proceed with our technical analysis leading up to the proof of our main result. We structure the discussion as follows: In

Section 3.3.1 we characterize the equilibrium points of the deterministic part of the dynamics in terms of the solution that we are seeking for our optimization problem. In Section 3.3.2, we define and prove several properties of the vector field that governs the flow; this vector field combines the gradients of the local objective functions and the non-symmetric Laplacian. As a side result, in Section 3.3.3 we establish the existence and uniqueness of solutions for the stochastic differential equation modeling the dynamics with noise. Finally, in Section 3.3.4 we present the core of our analysis, which is the identification of a 2nd moment noise-to-state stability Lyapunov function satisfying the hypotheses of Theorem 6.2.

3.3.1 Equilibrium points

In this section we show, for completeness, the correspondence between the equilibrium points o (3.10) in the absence of noise and the solutions of the optimization problem stated in Section 7.2.

Lemma 3.9. (Equilibrium points and Karush-Kuhn-Tucker conditions). *Let $\mathcal{G}$ be weight-balanced and strongly connected. Then, there exists $\boldsymbol{x}^*$ such that $[\boldsymbol{x}^{*\top}, \boldsymbol{z}^{*\top}]^\top$ satisfies the equilibrium conditions for the dynamics* (3.5) *without noise,*

$$\nabla \tilde{f}(\boldsymbol{x}^*) + \mathbf{L}\boldsymbol{z}^* = 0_{Nd} \quad \textit{and} \quad \mathbf{L}\boldsymbol{x}^* = 0_{Nd}, \tag{3.12}$$

for some $\boldsymbol{z}^* \in (\mathbb{R}^d)^N$, if and only if there exists $\boldsymbol{x}_{KKT}$ such that $[\boldsymbol{x}_{KKT}^\top, \boldsymbol{z}_{KKT}^\top]^\top$ satisfies the Karush-Kuhn-Tucker conditions for the minimization of $\tilde{f}$ in (3.4) *subject to $\mathbf{L}\boldsymbol{x} = 0$,*

$$\nabla \tilde{f}(\boldsymbol{x}_{KKT}) + \mathbf{L}^\top \boldsymbol{z}_{KKT} = 0_{Nd} \quad \textit{and} \quad \mathbf{L}\boldsymbol{x}_{KKT} = 0_{Nd}, \tag{3.13}$$

for some $\boldsymbol{z}_{KKT} \in (\mathbb{R}^d)^N$. *Moreover, both* (3.12) *and* (3.13) *are equivalent to*

$$(\mathbb{1}^\top \otimes \mathrm{I}_d)\nabla \tilde{f}(\boldsymbol{x}) = 0_{Nd} \quad \textit{and} \quad \mathbf{L}\boldsymbol{x} = 0_{Nd}, \tag{3.14}$$

and, if either $\boldsymbol{x}^*$ *or* $\boldsymbol{x}_{KKT}$ *exists and is unique, then so is the other one and* $\boldsymbol{x}^* = \boldsymbol{x}_{KKT}$.

Proof. Since $\mathcal{G}$ is weight-balanced and strongly connected, then $\mathcal{N}(\mathsf{L} + \mathsf{L}^\top) = \operatorname{span}\{\mathbb{1}\}$. The first equation in (3.14) follows by left-multiplying the first equation in (3.12) and (3.13) by $(\mathbb{1}^\top \otimes \mathrm{I}_d)$ and using that $\mathbb{1}^\top \mathsf{L} = 0$ because $\mathcal{G}$ is weight-balanced. The reason why (3.14) is equivalent to both (3.12) and (3.13) is the following: if there exists any $\boldsymbol{x}$ such that $(\mathbb{1}^\top \otimes \mathrm{I}_d)\nabla \tilde{f}(\boldsymbol{x}) = 0_d$, then $\nabla \tilde{f}(\boldsymbol{x})$ is in the column space of both $\mathbf{L}$ and $\mathbf{L}^\top$, which means that there exist $\boldsymbol{z}^*$ and $\boldsymbol{z}_{\mathrm{KKT}}$, respectively, that satisfy (3.12) and (3.13). This is because $\mathbf{L}(\mathbb{1} \otimes \mathrm{I}_d) = \mathbf{L}^\top(\mathbb{1} \otimes \mathrm{I}_d) = 0_{Nd \times d}$, and $\operatorname{rank}(\mathbf{L}) + \operatorname{rank}(\mathbb{1} \otimes \mathrm{I}_d) = \operatorname{rank}(\mathbf{L}^\top) + \operatorname{rank}(\mathbb{1} \otimes \mathrm{I}_d) = (N-1)d + d = Nd$. The result now follows by observing that $\boldsymbol{x}^*$ and $\boldsymbol{x}_{\mathrm{KKT}}$ are both defined by (3.14). □

As a consequence of this result and since there exists a unique minimizer $x_{\min}$ of (3.3), we deduce that the equilibrium points of the dynamics (3.5) in the absence of noise are $\boldsymbol{x}^* = \mathbb{1} \otimes x_{\min} \in (\mathbb{R}^d)^N$ and any $\boldsymbol{z}^* \in (\mathbb{R}^d)^n$ with $\mathbf{L}\boldsymbol{z}^* = -\nabla \tilde{f}(\mathbb{1} \otimes x_{\min})$.

3.3.2 Co-coercivity properties of the dynamics

In this section, we study the co-coercivity properties of the vector field $\mathbf{N}$ in the dynamics (3.10). Our results here play a key role later in establishing the global existence and uniqueness of the solutions and the noise-to-state stability properties of the dynamics. We first provide a general discussion on co-coercivity and then focus our attention on the properties of the dynamics (3.10). Given $S \in \mathbb{R}^{m \times m}$ and $\delta > 0$, we refer to a vector field $F : \mathbb{R}^m \to \mathbb{R}^m$ as $(S, \delta)-$ *co-coercive* with respect to

$\bar{\boldsymbol{x}} \in \mathbb{R}^m$ if,

$$(\boldsymbol{x}-\bar{\boldsymbol{x}})^\top S(F(\boldsymbol{x})-F(\bar{\boldsymbol{x}})) \geq \delta \, \|F(\boldsymbol{x})-F(\bar{\boldsymbol{x}})\|_2^2, \tag{3.15}$$

for all $\boldsymbol{x} \in \mathbb{R}^m$. This corresponds to the notion of co-coercivity of $S^\top F$ as defined in [ZM96] but here we define it for a vector field that is not necessarily the gradient of a scalar function. The following result provides sufficient conditions for a family of vector fields to be co-coercive under transformations that are small perturbations of the identity.

Theorem 3.10. (Sufficient conditions for $(\mathrm{I}+\beta^2\tilde{S},\delta)-$co-coercivity). *Let* $\mathbf{G} : (\mathbb{R}^d)^N \to (\mathbb{R}^d)^N$ *be a continuously differentiable vector field such that* $\mathbf{DG}(\boldsymbol{x}) \in \mathbb{R}^{Nd\times Nd}$ *is symmetric positive semidefinite for all* $\boldsymbol{x} \in (\mathbb{R}^d)^N$*. Also, let* $\mathbf{T} : (\mathbb{R}^d)^N \to (\mathbb{R}^d)^N$ *be the linear vector field* $\mathbf{T}(\boldsymbol{x}) = 2(\mathsf{L}\otimes \mathrm{I}_d)\boldsymbol{x}$*, where* L *is the Laplacian matrix of a strongly connected and weight-balanced digraph. Assume that there exist* $i_0 \in \{1,\dots,N\}$ *and* $r, R > 0$ *such that* $r\,(e_{i_0}e_{i_0}{}^\top)\otimes \mathrm{I}_d \preccurlyeq \mathbf{DG}(\boldsymbol{x}) \preccurlyeq R\,\mathrm{I}_{Nd}$ *for all* $\boldsymbol{x} \in (\mathbb{R}^d)^N$*. Given* $\epsilon > 0$*, let* $K_1 := \lambda_{min}\Big(r\,e_{i_0}e_{i_0}^\top + \epsilon(\mathsf{L}+\mathsf{L}^\top)\Big)$*,* $K_2 := R + 2\epsilon\,\sigma_{max}(\mathsf{L})$*, and* $\mathbf{F} := \mathbf{G}+\epsilon\mathbf{T}$*. Then,*

(i) $K_1 > 0$ *and* $2K_1\,\mathrm{I}_{Nd} \preccurlyeq \mathbf{DF}(\boldsymbol{x}) + (\mathbf{DF}(\boldsymbol{x}))^\top$ *for any* $\boldsymbol{x} \in (\mathbb{R}^d)^N$*.*

(ii) $K_1\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|_2 \leq \|\mathbf{F}(\boldsymbol{x})-\mathbf{F}(\bar{\boldsymbol{x}})\|_2 \leq K_2\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|_2$ *for any* $\boldsymbol{x},\bar{\boldsymbol{x}} \in (\mathbb{R}^d)^N$*.*

(iii) $\mathbf{F}$ *is* $(\mathrm{I}+\beta^2\tilde{S},\delta)-$*co-coercive with respect to every* $\bar{\boldsymbol{x}} \in (\mathbb{R}^d)^N$ *for any nonzero matrix* $\tilde{S} \in \mathbb{R}^{Nd\times Nd}$ *if* $\delta \in [0, K_1K_2^{-2})$ *and*

$$\beta \in \Big[0,\ \sqrt{\Big(K_1K_2^{-2}-\delta\Big)/(\|\tilde{S}\|_2\,K_1^{-1})}\ \Big].$$

Proof. Regarding *(i)*, we first show that $\lambda_{\min}\Big(r\,e_{i_0}e_{i_0}^\top + \epsilon(\mathsf{L}+\mathsf{L}^\top)\Big) > 0$. For this, note that the matrices $r\,e_{i_0}e_{i_0}{}^\top$ and $\epsilon(\mathsf{L}+\mathsf{L}^\top)$ are positive semidefinite. In addition, their sum has rank N as we show next. Arguing by contradiction, assume that $y \in \mathbb{R}^N \setminus \{0\}$ is in its nullspace, i.e., $\left(r\, e_{i_0} {e_{i_0}}^\top + \epsilon(\mathsf{L}+\mathsf{L}^\top)\right) y = 0$. Pre-multiplying by $y^\top$, it follows then that $0 \leq \epsilon y^\top(\mathsf{L}+\mathsf{L}^\top) y^\top = -r\,(y_{i_0})^2 \leq 0$, which implies that $y_{i_0} = 0$ and $y^\top(\mathsf{L}+\mathsf{L}^\top)y^\top = 0$. As $\mathsf{L}+\mathsf{L}^\top$ is symmetric positive semidefinite (because the graph is weight-balanced), we have $y \in \mathcal{N}(\mathsf{L}+\mathsf{L}^\top)$. Since $\mathcal{N}(\mathsf{L}+\mathsf{L}^\top) = \operatorname{span}\{\mathbb{1}_N\}$, because the graph is strongly connected, and $y_{i_0} = 0$, we obtain that $y = 0_N$, which is a contradiction. Therefore, $r\,(e_{i_0} {e_{i_0}}^\top) \otimes \mathrm{I}_d + \epsilon(\mathsf{L}+\mathsf{L}^\top) \otimes \mathrm{I}_d$ is positive definite, and hence $K_1 > 0$. On the other hand,

$$\begin{aligned} 2K_1\,\mathrm{I}_{Nd} &\preccurlyeq 2\left(r\, e_{i_0} {e_{i_0}}^\top + \epsilon(\mathsf{L}+\mathsf{L}^\top)\right) \otimes \mathrm{I}_d \\ &\preccurlyeq 2\mathbf{DG}(\boldsymbol{x}) + \mathbf{DT}(\boldsymbol{x}) + (\mathbf{DT}(\boldsymbol{x}))^\top \preccurlyeq \mathbf{DF}(\boldsymbol{x}) + (\mathbf{DF}(\boldsymbol{x}))^\top, \end{aligned}$$

for any $\boldsymbol{x} \in (\mathbb{R}^d)^N$, as required. Before proving *(ii)* and *(iii)*, we derive some useful expressions. We start by defining $j : [0,1] \to (\mathbb{R}^d)^N$ as $j(t) := \mathbf{F}\left(\bar{\boldsymbol{x}} + t(\boldsymbol{x} - \bar{\boldsymbol{x}})\right) - \mathbf{F}(\bar{\boldsymbol{x}})$. By the Fundamental Theorem of Calculus, we have that

$$j(1) = j(1) - j(0) = \int_0^1 j'(t)\mathrm{d}t = \mathrm{E}(\boldsymbol{x})(\boldsymbol{x} - \bar{\boldsymbol{x}}), \tag{3.16}$$

where the integral is taken component-wise and the matrix-valued function $\mathrm{E} : (\mathbb{R}^d)^N \to \mathbb{R}^{Nd \times Nd}$ is defined by

$$\begin{aligned} \mathrm{E}(\boldsymbol{x}) &:= \int_0^1 \mathbf{DF}\left(\bar{\boldsymbol{x}} + t(\boldsymbol{x} - \bar{\boldsymbol{x}})\right)\mathrm{d}t = \int_0^1 \mathbf{DG}\left(\bar{\boldsymbol{x}} + t(\boldsymbol{x} - \bar{\boldsymbol{x}})\right)\mathrm{d}t + 2\epsilon(\mathsf{L} \otimes \mathrm{I}_d) \\ &:= \mathrm{D}(\boldsymbol{x}) + 2\epsilon(\mathsf{L} \otimes \mathrm{I}_d), \end{aligned}$$

for $\boldsymbol{x} \in (\mathbb{R}^d)^N$. We derive next some useful facts about E.

(a) Since $\mathrm{D}(\boldsymbol{x})$ is symmetric positive semidefinite and $\mathrm{D}(\boldsymbol{x}) \preccurlyeq R\,\mathrm{I}$ for all

$\boldsymbol{x} \in (\mathbb{R}^d)^N$, using [Ber05, Fact 5.11.2], we deduce

$$\begin{aligned}\sigma_{\max}(\mathrm{E}(\boldsymbol{x})) &\leq \sigma_{\max}(\mathrm{D}(\boldsymbol{x})) + \sigma_{\max}(2\epsilon(\mathsf{L} \otimes \mathrm{I}_d)) \\ &= \lambda_{\max}(\mathrm{D}(\boldsymbol{x})) + 2\epsilon\,\sigma_{\max}(\mathsf{L} \otimes \mathrm{I}_d) \leq R + 2\epsilon\,\sigma_{\max}(\mathsf{L}) = K_2, \end{aligned} \tag{3.17}$$

where in the last inequality we have used $\sigma_{\max}(\mathsf{L} \otimes \mathrm{I}_d) = \sqrt{\lambda_{\max}((\mathsf{L}^\top \mathsf{L}) \otimes \mathrm{I}_d)} = \sqrt{\lambda_{\max}(\mathsf{L}^\top \mathsf{L})} = \sigma_{\max}(\mathsf{L})$.

(b) Using *(i)*, we deduce

$$2K_1\,\mathrm{I} \preccurlyeq \mathrm{E}(\boldsymbol{x}) + \mathrm{E}(\boldsymbol{x})^\top. \tag{3.18}$$

(c) Using [Ber05, Fact 8.14.4] and (3.18), we get

$$\sigma_{\min}(\mathrm{E}(\boldsymbol{x})) \geq \tfrac{1}{2}\,\lambda_{\min}(\mathrm{E}(\boldsymbol{x}) + \mathrm{E}(\boldsymbol{x})^\top) \geq K_1 > 0. \tag{3.19}$$

(d) Since $\mathrm{E}(\boldsymbol{x})$ is a square matrix, we have $\lambda_i(\mathrm{E}(\boldsymbol{x})\mathrm{E}(\boldsymbol{x})^\top) = \lambda_i(\mathrm{E}(\boldsymbol{x})^\top \mathrm{E}(\boldsymbol{x})) = \Big(\sigma_i(\mathrm{E}(\boldsymbol{x}))\Big)^2$ for $i = 1, \dots, Nd$, and, therefore, both $\mathrm{E}(\boldsymbol{x})\mathrm{E}(\boldsymbol{x})^\top$ and $\mathrm{E}(\boldsymbol{x})^\top \mathrm{E}(\boldsymbol{x})$ are lower and upper bounded by $(\sigma_{\min}(\mathrm{E}(\boldsymbol{x})))^2\,\mathrm{I}$ and $(\sigma_{\max}(\mathrm{E}(\boldsymbol{x})))^2\,\mathrm{I}$, respectively.

(e) Taking the invertible congruence given by the matrix $\mathrm{E}(\boldsymbol{x})^{-1} \in \mathbb{R}^{Nd \times Nd}$ (which is invertible by (c)) on both sides of (3.18), that is, multiplying on the left by $(\mathrm{E}(\boldsymbol{x})^\top)^{-1} = (\mathrm{E}(\boldsymbol{x})^{-1})^\top := \mathrm{E}(\boldsymbol{x})^{-\top}$ and on the right by $\mathrm{E}(\boldsymbol{x})^{-1}$, we get

$$2K_1\,\mathrm{E}(\boldsymbol{x})^{-\top}\mathrm{E}(\boldsymbol{x})^{-1} \preccurlyeq \mathrm{E}(\boldsymbol{x})^{-\top} + \mathrm{E}(\boldsymbol{x})^{-1}. \tag{3.20}$$

Now, since $\mathrm{E}(\boldsymbol{x})^{-\top}\mathrm{E}(\boldsymbol{x})^{-1} = \Big(\mathrm{E}(\boldsymbol{x})\mathrm{E}(\boldsymbol{x})^{\top}\Big)^{-1}$ we obtain from (3.20) that

$$\mathrm{E}(\boldsymbol{x})^{-\top} + \mathrm{E}(\boldsymbol{x})^{-1} \succcurlyeq \frac{2K_1}{\lambda_{\max}\Big(\mathrm{E}(\boldsymbol{x})\mathrm{E}(\boldsymbol{x})^{\top}\Big)}\mathrm{I} = 2K_1\Big(\sigma_{\max}(\mathrm{E}(\boldsymbol{x}))\Big)^{-2}\mathrm{I} \succcurlyeq 2K_1K_2^{-2}\mathrm{I}, \tag{3.21}$$

for all $\boldsymbol{x} \in (\mathbb{R}^d)^N$, where we used (d) in the identity and (a) in the last inequality. Equipped with these facts, we are ready to establish items *(ii)* and *(iii)*.

Regarding *(ii)*, notice that $\|\mathbf{F}(\boldsymbol{x}) - \mathbf{F}(\bar{\boldsymbol{x}})\|_2^2 = \|j(1)\|_2^2 = (\boldsymbol{x}-\bar{\boldsymbol{x}})^{\top}\mathrm{E}(\boldsymbol{x})^{\top}\mathrm{E}(\boldsymbol{x})(\boldsymbol{x}-\bar{\boldsymbol{x}})$, and therefore the result follows from (d) using the bound for $\sigma_{\min}(\mathrm{E}(\boldsymbol{x}))$ in (3.19) and for $\sigma_{\max}(\mathrm{E}(\boldsymbol{x}))$ in (3.17).

Regarding *(iii)*, we rewrite the inequality (3.15), which we need to establish for the vector field $\mathbf{F}$ and the matrix transformation $S := \mathrm{I} + \beta^2\tilde{S}$, as $(\boldsymbol{x}-\bar{\boldsymbol{x}})^{\top}Sj(1) \geq \delta\, j(1)^{\top}j(1)$, for all $\boldsymbol{x} \in (\mathbb{R}^d)^N$. Using (3.16), this becomes

$$(\boldsymbol{x}-\bar{\boldsymbol{x}})^{\top}S\mathrm{E}(\boldsymbol{x})(\boldsymbol{x}-\bar{\boldsymbol{x}}) \geq \delta\,(\boldsymbol{x}-\bar{\boldsymbol{x}})^{\top}\mathrm{E}(\boldsymbol{x})^{\top}\mathrm{E}(\boldsymbol{x})(\boldsymbol{x}-\bar{\boldsymbol{x}}), \quad \forall \boldsymbol{x} \in (\mathbb{R}^d)^N,$$

which follows from the stronger condition given by

$$\tfrac{1}{2}\Big(\mathrm{E}(\boldsymbol{x})^{\top}S^{\top} + S\mathrm{E}(\boldsymbol{x})\Big) \succcurlyeq \delta\,\mathrm{E}(\boldsymbol{x})^{\top}\mathrm{E}(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in (\mathbb{R}^d)^N. \tag{3.22}$$

We now proceed verifying an equivalent linear matrix inequality. Taking now on both sides of (3.22) the same congruence as in (e) and substituting $S = \mathrm{I} + \beta^2\tilde{S}$, we get

$$(\mathrm{I} + \beta^2\tilde{S}^{\top})\mathrm{E}(\boldsymbol{x})^{-1} + \mathrm{E}(\boldsymbol{x})^{-\top}(\mathrm{I} + \beta^2\tilde{S}) \succcurlyeq 2\delta\mathrm{I}, \quad \forall \boldsymbol{x} \in (\mathbb{R}^d)^N,$$

which, after reordering terms and defining $\tilde{\mathrm{E}}(\boldsymbol{x}) := \mathrm{E}(\boldsymbol{x})^{-1} - \delta \mathrm{I}$, becomes

$$\tilde{\mathrm{E}}(\boldsymbol{x}) + \tilde{\mathrm{E}}(\boldsymbol{x})^\top \succcurlyeq -\beta^2 \Big(\tilde{S}^\top \mathrm{E}(\boldsymbol{x})^{-1} + \mathrm{E}(\boldsymbol{x})^{-\top} \tilde{S}\Big), \quad \forall \boldsymbol{x} \in (\mathbb{R}^d)^N. \tag{3.23}$$

To guarantee that (3.23) holds, we seek bounds on both sides that are uniform. Regarding the left-hand side of (3.23), we get from (3.21) that

$$\tilde{\mathrm{E}}(\boldsymbol{x}) + \tilde{\mathrm{E}}(\boldsymbol{x})^\top = \mathrm{E}(\boldsymbol{x})^{-1} + \mathrm{E}(\boldsymbol{x})^{-\top} - 2\delta \mathrm{I} \succcurlyeq 2\Big(K_1 K_2^{-2} - \delta\Big)\mathrm{I}, \quad \forall \boldsymbol{x} \in (\mathbb{R}^d)^N. \tag{3.24}$$

Regarding the right-hand side of (3.23), using (3.19) we first observe that

$$\|\mathrm{E}(\boldsymbol{x})^{-1}\|_2 = \sigma_{\max}(\mathrm{E}(\boldsymbol{x})^{-1}) = \Big(\sigma_{\min}(\mathrm{E}(\boldsymbol{x}))\Big)^{-1} \leq K_1^{-1}, \quad \forall \boldsymbol{x} \in (\mathbb{R}^d)^N.$$

Thus, using the triangular inequality, the fact that $\|A\|_2 = \|A^\top\|_2$, and the submultiplicativity of the norm, we get that for all $\boldsymbol{x} \in (\mathbb{R}^d)^N$,

$$\|\tilde{S}^\top \mathrm{E}(\boldsymbol{x})^{-1} + \mathrm{E}(\boldsymbol{x})^{-\top}\tilde{S}\|_2 \leq 2\|\tilde{S}^\top \mathrm{E}(\boldsymbol{x})^{-1}\|_2 \leq 2\|\tilde{S}\|_2 \, \|\mathrm{E}(\boldsymbol{x})^{-1}\|_2 \leq 2\|\tilde{S}\|_2 \, K_1^{-1}.$$

Since $\pm A \preccurlyeq \|A\|_2 \, \mathrm{I}$, we deduce

$$-\Big(\tilde{S}^\top \mathrm{E}(\boldsymbol{x})^{-1} + \mathrm{E}(\boldsymbol{x})^{-1}\tilde{S}\Big) \preccurlyeq 2\,\|\tilde{S}\|_2 \, K_1^{-1}\,\mathrm{I}, \quad \forall \boldsymbol{x} \in (\mathbb{R}^d)^N. \tag{3.25}$$

Therefore, relating the uniform bounds (3.24) and (3.25), we conclude that if $\beta \leq \beta_1^*$, then (3.23) holds for every $\boldsymbol{x} \in (\mathbb{R}^d)^N$ because

$$\tilde{\mathrm{E}}(\boldsymbol{x}) + \tilde{\mathrm{E}}(\boldsymbol{x})^\top \succcurlyeq 2\Big(K_1 K_2^{-2} - \delta\Big)\,\mathrm{I} \succcurlyeq 2\,\|\tilde{S}\|_2 \, K_1^{-1} \, \beta^2 \,\mathrm{I} \succcurlyeq -\beta^2 \Big(\tilde{S}^\top \mathrm{E}(\boldsymbol{x})^{-1} + \mathrm{E}(\boldsymbol{x})^{-1}\tilde{S}\Big),$$

which concludes the proof. □

Note that, under the hypotheses of Theorem 2.5, the above result is applicable to $\mathbf{G} = \nabla \tilde{f}$ (so that $\mathbf{DG} = \nabla^2 \tilde{f}$ is symmetric and conveniently lower and upper bounded by the hypotheses on the local functions), $\mathbf{F}(\boldsymbol{x}) = \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) := \nabla \tilde{f}(\boldsymbol{x}) + 2\epsilon \mathbf{L}\boldsymbol{x}$, and $\tilde{S} = \mathbf{L}_\mathcal{K}$ (which has $\|\tilde{S}\|_2 = \|\mathbf{L}_\mathcal{K}\|_2 = 1$). In particular, we have

$$K_1 \|\boldsymbol{x}' - \boldsymbol{x}\|_2 \leq \|\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}') - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x})\|_2 \leq K_2 \|\boldsymbol{x}' - \boldsymbol{x}\|_2, \tag{3.26}$$

for all $\boldsymbol{x}', \boldsymbol{x} \in (\mathbb{R}^d)^N$, and

$$(\boldsymbol{x} - \boldsymbol{x}^*)^\top (\mathrm{I} + \beta^2 \mathbf{L}_\mathcal{K})(\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*)) \geq \delta \, \|\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*)\|_2^2, \tag{3.27}$$

for all $\boldsymbol{x} \in (\mathbb{R}^d)^N$, $\delta \in (0, \, K_1 K_2^{-2})$ and $\beta \in \left[0, \, \sqrt{K_1^2 K_2^{-2} - K_1 \delta}\right]$.

3.3.3 Global existence and uniqueness of solutions

Here we establish the global existence and uniqueness of solutions of the dynamics (3.5) by verifying the hypotheses in Lemma 6.1. We obtain the following bound for almost every $t \geq t_0$:

$$\begin{aligned}
&\max \Big\{ \|g(v) - g(v')\|_2 \, , \, \|G(v,t) - G(v',t)\|_\mathcal{F} \Big\} \\
&\leq \|\mathbf{A}(v - v')\|_2 + \|\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}')\|_2 + \|G(v,t) - G(v',t)\|_\mathcal{F} \\
&\leq \|\mathbf{A}\|_2 \|v - v'\|_2 + K_2 \|\boldsymbol{x} - \boldsymbol{x}'\|_2 + \kappa_1 \|v - v'\|_2 \leq \Big(\|\mathbf{A}\|_2 + K_2 + \kappa_1 \Big) \|v - v'\|_2,
\end{aligned}$$

where in the second inequality we have used (3.26) and the Lipschitz condition for G in (3.11). In addition, for almost every $t \geq t_0$,

$$\begin{aligned}
&v^\top g(v) + \tfrac{1}{2}\|G(v,t)\|_{\mathcal{F}}^2 = v^\top \mathbf{A} v + v^\top \mathbf{N}(\boldsymbol{x}) + \tfrac{1}{2}\|G(v,t)\|_{\mathcal{F}}^2 \\
&\leq \|\mathbf{A}\|_2 \|v\|_2^2 + \|v\|_2 \|\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x})\|_2 + \tfrac{1}{2}\kappa_2^2 \\
&\leq \|\mathbf{A}\|_2 \|v\|_2^2 + \|v\|_2 \Big(K_2 \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 + \|\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*)\|_2\Big) + \tfrac{1}{2}\kappa_2^2 \\
&\leq \|\mathbf{A}\|_2 \|v\|_2^2 + \|v\|_2 \Big(K_2 \|\boldsymbol{x}\|_2 + K_2 \|\boldsymbol{x}^*\|_2 + \|\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*)\|_2\Big) + \tfrac{1}{2}\kappa_2^2 \\
&\leq \Big(\|\mathbf{A}\|_2 + K_2\Big) \|v\|_2^2 + \|v\|_2 \Big(K_2 \|\boldsymbol{x}^*\|_2 + \|\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*)\|_2\Big) + \tfrac{1}{2}\kappa_2^2 \\
&\leq \Big(1 + \|v\|_2^2\Big) \Big(\|\mathbf{A}\|_2 + K_2 + K_2 \|\boldsymbol{x}^*\|_2 + \|\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*)\|_2 + \tfrac{1}{2}\kappa_2^2\Big).
\end{aligned}$$

The global existence and uniqueness of the solutions of the dynamics (3.5) now follows from Lemma 6.1 as a consequence of these two facts.

3.3.4 NSS Lyapunov function

Our strategy to establish the noise-to-state stability properties of the distributed coordination algorithm (3.5) is based on identifying a suitable NSS Lyapunov function for the dynamics. Our first result of this section identifies a candidate Lyapunov function whose derivative in the sense of Itô can be conveniently upper bounded. To obtain this bound, we build on the co-coercivity properties stated in Theorem 3.10 of the vector fields that combine local gradient descent and local consensus.

Proposition 3.11. (Candidate second moment NSS-Lyapunov function). *Under* *the hypotheses of Theorem 2.5, let*

$$\mathrm{P}_\beta := \begin{bmatrix} \mathrm{I}+\beta^2\mathbf{L}_{\mathcal{K}} & \beta\mathbf{L}_{\mathcal{K}} \\ \beta\mathbf{L}_{\mathcal{K}} & \mathbf{L}_{\mathcal{K}} \end{bmatrix} \in \mathbb{R}^{2Nd\times 2Nd}, \tag{3.28a}$$

$$\mathrm{Q}_\beta := \begin{bmatrix} \begin{bmatrix} \beta^3+2\beta+\frac{2}{\beta} & (1+\beta^2) \\ (1+\beta^2) & \beta \end{bmatrix} \otimes (\mathbf{L}+\mathbf{L}^\top) & \begin{matrix} 0 \\ \beta\mathbf{L}_{\mathcal{K}} \end{matrix} \\ \begin{matrix} 0 & \beta\mathbf{L}_{\mathcal{K}} \end{matrix} & 2\delta\mathrm{I} \end{bmatrix} \in \mathbb{R}^{3Nd\times 3Nd}, \tag{3.28b}$$

and define the functions $\mathrm{V},\mathrm{W}:\mathbb{R}^{2Nd}\to\mathbb{R}$ *by*

$$\mathrm{V}(v) := \tfrac{1}{2}[(\boldsymbol{x}-\boldsymbol{x}^*)^\top,(\boldsymbol{z}-\boldsymbol{z}^*)^\top]\,\mathrm{P}_\beta \begin{bmatrix} \boldsymbol{x}-\boldsymbol{x}^* \\ \boldsymbol{z}-\boldsymbol{z}^* \end{bmatrix},$$

$$\mathrm{W}(v) := \tfrac{1}{2}\begin{bmatrix} (\boldsymbol{x}-\boldsymbol{x}^*)^\top & (\boldsymbol{z}-\boldsymbol{z}^*)^\top & (\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x})-\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*))^\top \end{bmatrix}\,\mathrm{Q}_\beta \begin{bmatrix} \boldsymbol{x}-\boldsymbol{x}^* \\ \boldsymbol{z}-\boldsymbol{z}^* \\ \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x})-\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*) \end{bmatrix},$$

where $\boldsymbol{x}^* = \mathbb{1}\otimes x_{\min} \in (\mathbb{R}^d)^N$ *and* $\boldsymbol{z}^* \in (\mathbb{R}^d)^n$ *is such that* $\mathbf{L}\boldsymbol{z}^* = -\nabla\tilde{f}(\mathbb{1}\otimes x_{\min})$. *Then the following holds:*

(i) The matrix P_β *is positive semidefinite for any* $\beta\in\mathbb{R}$ *with nullspace*

$$\mathcal{N}(\mathrm{P}_\beta) = \mathrm{span}\Big\{\begin{bmatrix} 0 & (\mathbb{1}\otimes b)^\top \end{bmatrix}^\top : b\in\mathbb{R}^d\Big\}.$$

(ii) The matrix Q_β *is positive semidefinite for the range of values of* β *specified in Theorem 2.5, and has nullspace*

$$\mathcal{N}(\mathrm{Q}_\beta) = \mathrm{span}\Big\{\begin{bmatrix} (\mathbb{1}\otimes b_1)^\top & 0 & 0 \end{bmatrix}^\top, \begin{bmatrix} 0 & (\mathbb{1}\otimes b_2)^\top & 0 \end{bmatrix}^\top : b_1,b_2\in\mathbb{R}^d\Big\}.$$

(iii) The function V *is twice continuously differentiable and bounded by*

$$\alpha_1\Big(\|v-v^*\|_{\hat{\mathrm{I}}}^2\Big) \le \mathrm{V}(v) \le \alpha_2\Big(\|v-v^*\|_{\hat{\mathrm{I}}}^2\Big), \tag{3.29}$$

where $v^* := (\boldsymbol{x}^*, \boldsymbol{z}^*)$*,* $\alpha_1(r) := \lambda_{(2N-1)d}(\mathrm{P}_\beta)r$*,* $\alpha_2(r) := \lambda_{max}(\mathrm{P}_\beta)r$*, and the matrix* $\hat{\mathrm{I}} \in \mathbb{R}^{2Nd}$ *is defined as*

$$\hat{\mathrm{I}} := \mathrm{diag}\Big(\mathrm{I}_{Nd}, \mathbf{L}_{\mathcal{K}}\Big).$$

(iv) The function W *is continuous, and the following dissipation inequality holds,*

$$\mathcal{L}[\mathrm{V}](v,t) \le -\mathrm{W}(v) + \sigma\Big(\|\Sigma(t)\|_{\mathcal{F}}\Big), \tag{3.30}$$

for all $(v,t) \in \mathbb{R}^{2Nd} \times [t_0, \infty)$*, where* $\sigma(r) := \mathrm{trace}(\mathrm{P}_\beta)\,\kappa_2^2 r^2$*.*

Proof. To show *(i)*, we note that P_β is a congruence by an invertible matrix of the positive semidefinite matrix $\hat{\mathrm{I}}$,

$$\mathrm{P}_\beta = \begin{bmatrix} \mathrm{I} & 0 \\ \beta\mathrm{I} & \mathrm{I} \end{bmatrix}^\top \begin{bmatrix} \mathrm{I} & 0 \\ 0 & \mathbf{L}_{\mathcal{K}} \end{bmatrix} \begin{bmatrix} \mathrm{I} & 0 \\ \beta\mathrm{I} & \mathrm{I} \end{bmatrix}.$$

Therefore, $\mathrm{rank}\,(\mathrm{P}_\beta) = \mathrm{rank}\,(\hat{\mathrm{I}}) = \mathrm{rank}\,(\mathrm{I}) + \mathrm{rank}\,(\mathbf{L}_{\mathcal{K}}) = Nd + (N-1)d = (2N-1)d$. The statement follows now by noting that the subspace $\mathrm{span}\Big\{\begin{bmatrix} 0 & (\mathbb{1}\otimes b)^\top \end{bmatrix}^\top : b \in \mathbb{R}^d\Big\}$ has dimension d and lies in the nullspace of P_β.

To establish *(ii)*, we show that $-\mathrm{Q}_\beta$ is negative semidefinite for the range of values of β in the statement. For convenience, define the matrices

$$\begin{aligned} \mathrm{B} &:= \begin{bmatrix} \beta^3+2\beta+\frac{2}{\beta} & (1+\beta^2) \\ (1+\beta^2) & \beta \end{bmatrix}, \\ \mathrm{Q}_1 &:= -\mathrm{B}\otimes(\mathbf{L}+\mathbf{L}^\top), \end{aligned}$$

and note that Q_1 corresponds to the first block of $-\mathrm{Q}_\beta$. Since B is symmetric, $\det(\mathrm{B})=1$, and $\mathrm{trace}(\mathrm{B})=\beta^3+3\beta+\frac{2}{\beta}>0$ for $\beta>0$, we deduce that $-\mathrm{B}\prec 0$ for any $\beta>0$. Therefore, Q_1 is symmetric negative semidefinite with nullspace

$$\mathcal{N}(\mathrm{Q}_1)=\mathrm{span}\left\{\begin{bmatrix}(\mathbb{1}\otimes b_1)^\top & 0\end{bmatrix}^\top, \begin{bmatrix}0 & (\mathbb{1}\otimes b_2)^\top\end{bmatrix}^\top : b_1,b_2\in\mathbb{R}^d\right\}.$$

Next, defining

$$\mathrm{Q}_2 := \begin{pmatrix} 0 & 0 \\ 0 & \frac{\beta^2}{2\delta} \end{pmatrix}\otimes\mathbf{L}_\mathcal{K}$$

and using $\mathbf{L}_{\mathcal{K}}^2 = \mathbf{L}_{\mathcal{K}}$, we simplify the following invertible congruence,

$$
\begin{aligned}
&-\begin{bmatrix} \mathrm{I} & 0 & 0 \\ 0 & \mathrm{I} & 0 \\ 0 & -\frac{\beta}{2\delta}\mathbf{L}_{\mathcal{K}} & \mathrm{I} \end{bmatrix}^{\top} \mathrm{Q}_{\beta} \begin{bmatrix} \mathrm{I} & 0 & 0 \\ 0 & \mathrm{I} & 0 \\ 0 & -\frac{\beta}{2\delta}\mathbf{L}_{\mathcal{K}} & \mathrm{I} \end{bmatrix} \\
&= \begin{bmatrix} \mathrm{I} & 0 & 0 \\ 0 & \mathrm{I} & -\frac{\beta}{2\delta}\mathbf{L}_{\mathcal{K}} \\ 0 & 0 & \mathrm{I} \end{bmatrix} \left(\left[\begin{array}{cc|c} \multicolumn{2}{c|}{\multirow{2}{*}{Q_1}} & 0 \\ & & 0 \\ \hline 0 & 0 & 0 \end{array}\right] + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\beta\mathbf{L}_{\mathcal{K}} \\ 0 & -\beta\mathbf{L}_{\mathcal{K}} & -2\delta\mathrm{I} \end{bmatrix} \right) \begin{bmatrix} \mathrm{I} & 0 & 0 \\ 0 & \mathrm{I} & 0 \\ 0 & -\frac{\beta}{2\delta}\mathbf{L}_{\mathcal{K}} & \mathrm{I} \end{bmatrix} \\
&= \begin{bmatrix} \mathrm{Q}_1 + \begin{pmatrix} 0 & 0 \\ 0 & \frac{\beta^2}{2\delta} \end{pmatrix} \otimes \mathbf{L}_{\mathcal{K}} & 0 \\ 0 & -2\delta\mathrm{I} \end{bmatrix} = \begin{bmatrix} \mathrm{Q}_1 + \mathrm{Q}_2 & 0 \\ 0 & -2\delta\mathrm{I} \end{bmatrix}.
\end{aligned}
$$

Since this is a block-diagonal matrix whose lower block, $-2\delta\mathrm{I}$, is negative definite, to establish the result is sufficient to show that for the specified values of β, the sum $\mathrm{Q}_1 + \mathrm{Q}_2$ is negative semidefinite. Note the maximum nonzero eigenvalue of Q_1, denoted $\lambda_{\max}^{\varnothing}(\mathrm{Q}_1)$, is

$$
\Big(-\tfrac{\beta^4+3\beta^2+2}{2\beta} + \sqrt{\Big(\tfrac{\beta^4+3\beta^2+2}{2\beta}\Big)^2 - 1} \Big) \lambda_2(\mathsf{L} + \mathsf{L}^{\top}).
$$

On the other hand, Q_2 is symmetric positive semidefinite with $\mathrm{rank}(\mathrm{Q}_2) = \mathrm{rank}(\mathbf{L}_{\mathcal{K}}) = (N-1)d$ and $\mathrm{spec}(\mathrm{Q}_2) = \{0, \frac{\beta^2}{2\delta}\}$, so the maximum eigenvalue of Q_2 is $\lambda_{\max}(R) = \frac{\beta^2}{2\delta}$. Now, since $\mathcal{N}(\mathrm{Q}_1) \subseteq \mathcal{N}(\mathrm{Q}_2)$, it follows that $\mathcal{N}(\mathrm{Q}_1) \subseteq \mathcal{N}(\mathrm{Q}_1 + \mathrm{Q}_2)$. Thus, in order to check the semidefiniteness of $\mathrm{Q}_1 + \mathrm{Q}_2$, we can restrict our attention to the subspace

$\mathcal{U}^{\perp} := \mathcal{N}(\mathrm{Q}_1)^{\perp}$. By Weyl's Theorem [HJ85, Theorem 4.3.7],

$$\begin{aligned}\lambda_{\max}^{\varnothing}(\mathrm{Q}_1+\mathrm{Q}_2) = \lambda_{\max}^{\mathcal{U}^{\perp}}(\mathrm{Q}_1+\mathrm{Q}_2) &\le \lambda_{\max}^{\mathcal{U}^{\perp}}(\mathrm{Q}_1)+\lambda_{\max}^{\mathcal{U}^{\perp}}(\mathrm{Q}_2)\\ &= \lambda_{\max}^{\varnothing}(\mathrm{Q}_1)+\lambda_{\max}(\mathrm{Q}_2) = h(\beta,\delta)\,.\end{aligned}$$

Since, by Lemma 0.65, $h(\beta,\delta)<0$ for $\delta\in(0,\,K_1K_2^{-2})$ and $\beta\in(0,\min\{\beta_1^*(\delta,\epsilon),\beta_2^*(\delta)\})$, we deduce that $\mathrm{Q}_1+\mathrm{Q}_2$ is negative definite in the subspace $\mathcal{N}(\mathrm{Q}_1)^{\perp}$. Therefore, $\mathcal{N}(\mathrm{Q}_1+\mathrm{Q}_2)=\mathcal{N}(\mathrm{Q}_1)$, which in turn implies that $\mathcal{N}(\mathrm{Q}_\beta)=\operatorname{span}\left\{[u^\top,0]^\top \,:\, u\in\mathcal{N}(\mathrm{Q}_1)\right\}$.

Regarding *(iii)*, it is clear from its definition that V is twice (in fact, infinitely) continuously differentiable. Furthermore, notice that $\hat{\mathrm{I}}$ and P_β are symmetric positive semidefinite with the same nullspace, so that

$$\frac{\lambda_{(2N-1)d}(\mathrm{P}_\beta)}{\lambda_{\max}(\hat{\mathrm{I}})}\,y^\top\hat{\mathrm{I}}y \le y^\top\mathrm{P}_\beta y \le \frac{\lambda_{\max}(\mathrm{P}_\beta)}{\lambda_{(2N-1)d}(\hat{\mathrm{I}})}\,y^\top\hat{\mathrm{I}}y,$$

for all $y\in\mathbb{R}^{2Nd}$. Since $\hat{\mathrm{I}}$ is idempotent, $\hat{\mathrm{I}}=\hat{\mathrm{I}}^2$, we have $y^\top\hat{\mathrm{I}}y=\|y\|_{\hat{\mathrm{I}}}^2$. The result now follows by observing that all nonzero eigenvalues of $\hat{\mathrm{I}}$ are 1.

Finally, we turn our attention to *(iv)*. We first compute the elements of $\mathcal{L}[\mathrm{V}]$ in (2.13). With the notation of (3.10), using that $\mathrm{P}_\beta=\mathrm{P}_\beta^\top$ and the sub-multiplicativity of the Frobenius norm, the diffusion term yields

$$\begin{aligned}&\tfrac{1}{2}\operatorname{trace}\Big(\Sigma(t)^\top G(v,t)^\top\nabla_v^2\mathrm{V}(v)G(v,t)\Sigma(t)\Big)\\ &\quad=\tfrac{1}{2}\operatorname{trace}\Big(\Sigma(t)^\top G(v,t)^\top\mathrm{P}_\beta G(v,t)\Sigma(t)\Big)\\ &\quad=\|\mathrm{P}_\beta^{1/2}G(v,t)\Sigma(t)\|_{\mathcal{F}}^2 \le \|\mathrm{P}_\beta^{1/2}\|_{\mathcal{F}}^2\,\|G(v,t)\|_{\mathcal{F}}^2\,\|\Sigma(t)\|_{\mathcal{F}}^2\\ &\quad\le\operatorname{trace}(\mathrm{P}_\beta)\,\kappa_2^2\,\|\Sigma(t)\|_{\mathcal{F}}^2=\sigma(\|\Sigma(t)\|_{\mathcal{F}}).\end{aligned}$$

On the other hand, defining $\tilde{\mathrm{Q}}_1 := 2\operatorname{sym}\Big(\mathrm{P}_\beta\mathbf{A}\Big) := \mathrm{P}_\beta\mathbf{A}+\mathbf{A}^\top\mathrm{P}_\beta$ and $\tilde{v}:=v-v^*$, and subtracting the quantity $\mathbf{A}v^* + \mathbf{N}(\boldsymbol{x}^*) = 0$, the drift term yields

$$\begin{aligned}\nabla_v \mathrm{V}(v)^\top \Big(\mathbf{A}v + \mathbf{N}(\boldsymbol{x})\Big) &= \nabla_v \mathrm{V}(v)^\top \Big(\mathbf{A}\tilde{v} - \mathbf{N}(\boldsymbol{x}^*) + \mathbf{N}(\boldsymbol{x})\Big) \\ &= \tfrac{1}{2}\tilde{v}^\top \tilde{\mathrm{Q}}_1 \tilde{v} + \tilde{v}^\top \mathrm{P}_\beta(-\mathbf{N}(\boldsymbol{x}^*) + \mathbf{N}(\boldsymbol{x})).\end{aligned}$$

Summarizing, we have

$$\mathcal{L}[\mathrm{V}](v,t) \le \tfrac{1}{2}\tilde{v}^\top \tilde{\mathrm{Q}}_1 \tilde{v} + \tilde{v}^\top \mathrm{P}_\beta(-\mathbf{N}(\boldsymbol{x}^*) + \mathbf{N}(\boldsymbol{x})) + \sigma\Big(\|\Sigma(t)\|_{\mathcal{F}}\Big) \tag{3.31}$$

for all $(v,t) \in \mathbb{R}^{2Nd} \times [t_0, \infty)$. We look first at the quadratic term in (3.31) arising from the linear part of the dynamics. Since $\mathbf{L}_{\mathcal{K}}\mathbf{L} = \mathrm{I}\,\mathbf{L} = \mathbf{L}$, splitting the matrix P_β, we obtain the factorization

$$\begin{aligned}\tilde{\mathrm{Q}}_1 &= 2\,\mathrm{sym}\Bigg(\Bigg(\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} \otimes \mathrm{I}_{Nd} + \begin{bmatrix}\beta^2 & \beta\\ \beta & 1\end{bmatrix} \otimes \mathbf{L}_{\mathcal{K}}\Bigg)\Bigg(\begin{bmatrix}-\gamma & -1\\ 1 & 0\end{bmatrix} \otimes \mathbf{L}\Bigg)\Bigg) \\ &= 2\,\mathrm{sym}\Bigg(\Bigg(\begin{bmatrix}1+\beta^2 & \beta\\ \beta & 1\end{bmatrix}\begin{bmatrix}-\gamma & -1\\ 1 & 0\end{bmatrix}\Bigg) \otimes \Big(\mathbf{L}_{\mathcal{K}}\mathbf{L}\Big)\Bigg) \\ &= 2\,\mathrm{sym}\Bigg(\begin{bmatrix}-\gamma(1+\beta^2)+\beta & -(1+\beta^2)\\ -\gamma\beta + 1 & -\beta\end{bmatrix} \otimes \mathbf{L}\Bigg).\end{aligned}$$

Now, recalling that $(2+\beta^2)/\beta + 2\epsilon = \tilde{\gamma} = \gamma + 2\epsilon$, we have $\gamma = (2+\beta^2)/\beta$, so the first matrix is indeed symmetric and we can factor out $2\,\mathrm{sym}(\mathbf{L}) := \mathbf{L} + \mathbf{L}^\top$ using the Kronecker product. In fact, $-\gamma(1+\beta^2) + \beta = -\beta^3 - 2\beta - \frac{2}{\beta}$, and we deduce

$$\tilde{\mathrm{Q}}_1 = -\begin{bmatrix}\beta^3 + 2\beta + \frac{2}{\beta} & (1+\beta^2)\\ (1+\beta^2) & \beta\end{bmatrix} \otimes (\mathbf{L} + \mathbf{L}^\top) = \mathrm{Q}_1. \tag{3.32}$$

Next, we turn our attention to the nonlinear term in (3.31). Note that

$$\begin{aligned}
&\tilde{v}^\top \mathrm{P}_\beta \Big(-\mathbf{N}(\boldsymbol{x}^*) + \mathbf{N}(\boldsymbol{x}) \Big) \\
&\quad = \tilde{v}^\top \begin{bmatrix} \mathrm{I} + \beta^2 \mathbf{L}_\mathcal{K} & \beta \mathbf{L}_\mathcal{K} \\ \beta \mathbf{L}_\mathcal{K} & \mathbf{L}_\mathcal{K} \end{bmatrix} \begin{bmatrix} \mathrm{I} \\ 0 \end{bmatrix} \Big(-\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*) + \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) \Big) \\
&\quad = -(\boldsymbol{x} - \boldsymbol{x}^*)^\top (\mathrm{I} + \beta^2 \mathbf{L}_\mathcal{K}) \Big(\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*) \Big) - (\boldsymbol{z} - \boldsymbol{z}^*)^\top \beta \mathbf{L}_\mathcal{K} \Big(\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*) \Big) \\
&\quad \leq -\delta \| \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*) \|_2^2 - (\boldsymbol{z} - \boldsymbol{z}^*)^\top \beta \mathbf{L}_\mathcal{K} \Big(\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*) \Big) .
\end{aligned}$$

Here, the last inequality follows from (3.27). Therefore, the nonlinear term can be expressed as

$$\begin{aligned}
&\tilde{v}^\top \mathrm{P}_\beta \Big(-\mathbf{N}(\boldsymbol{x}^*) + \mathbf{N}(\boldsymbol{x}) \Big) \\
&\quad = \frac{1}{2} \Big[\tilde{v}^\top, \Big(\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*) \Big)^\top \Big] \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\beta \mathbf{L}_\mathcal{K} \\ 0 & -\beta \mathbf{L}_\mathcal{K} & -2\delta \mathrm{I} \end{bmatrix} \begin{bmatrix} \tilde{v} \\ \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*) \end{bmatrix} . \qquad (3.33)
\end{aligned}$$

The result now follows from substituting (3.32) and (3.33) into (3.31). □

Given the result in Proposition 3.11, the missing piece to establish that V is a second moment NSS-Lyapunov function with respect to span$\Big\{ \Big[0 \quad (\mathbb{1} \otimes b)^\top \Big]^\top : b \in \mathbb{R}^d \Big\}$ is to relate its value to that of W. To this end, we define the constraint set

$$\mathcal{D}_{\boldsymbol{x}^*} := \Big\{ y \in \mathbb{R}^{3Nd} : y^3 = \tilde{\mathbf{F}}_\epsilon(y^1 + \boldsymbol{x}^*) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*) \Big\} \subset \mathbb{R}^{3Nd},$$

and the quadratic functions $\mathrm{V}_{\tilde{\mathrm{P}}_\beta}, \mathrm{W}_{\mathrm{Q}_\beta} : \mathcal{D}_{\boldsymbol{x}^*} \to \mathbb{R}_{\geq 0}$,

$$\mathrm{V}_{\tilde{\mathrm{P}}_\beta}(y) := \tfrac{1}{2} y^\top \tilde{\mathrm{P}}_\beta y, \qquad \tilde{\mathrm{P}}_\beta := \begin{bmatrix} \mathrm{P}_\beta & 0 \\ 0 & 0 \end{bmatrix},$$
$$\mathrm{W}_{\mathrm{Q}_\beta}(y) := \tfrac{1}{2} y^\top \mathrm{Q}_\beta y.$$

Note that $\mathrm{V}(v) = \mathrm{V}_{\tilde{\mathrm{P}}_\beta}(v - v^*, \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*))$ and $\mathrm{W}(v) = \mathrm{W}_{\mathrm{Q}_\beta}(v - v^*, \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*))$ for all $v \in \mathbb{R}^{2Nd}$. The following result relates the value of these quadratic functions.

Proposition 3.12. (Bound on candidate second moment NSS-Lyapunov function). *Under the hypotheses of Theorem 2.5, the next bound holds,*

$$\mathrm{V}_{\tilde{\mathrm{P}}_\beta}(y) \leq \eta(\mathrm{W}_{\mathrm{Q}_\beta}(y)), \quad \forall y \in \mathcal{D}_{\boldsymbol{x}^*}, \tag{3.34}$$

with linear gain $\eta(r) := C_\eta r$, for $r \geq 0$, where

$$C_\eta := \frac{\max\{1, \lambda_{max}(\mathsf{L} + \mathsf{L}^\top)\}\, \lambda_{max}(\mathrm{P}_\beta)}{\min\{1, \frac{K_1}{(1+K_2^2)}\}\, \lambda_{(3N-2)d}(\mathrm{Q}_\beta)\, \min\{1, \lambda_{N-1}(\mathsf{L} + \mathsf{L}^\top)\}}.$$

and P_β and Q_β are defined in (3.28).

Proof. For $A := \operatorname{diag}\left(\mathrm{I}, \sqrt{\mathbf{L} + \mathbf{L}^\top}, \mathrm{I}\right) \in \mathbb{R}^{3Nd \times 3Nd}$, whose nullspace is $\mathcal{N}(A) = \operatorname{span}\Big\{ \begin{bmatrix} 0 & (\mathbb{1} \otimes b)^\top & 0 \end{bmatrix}$ we define the functions $\phi_{2,A}, \psi_{2,A} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$,

$$\phi_{2,A}(s) := \sup_{\{y \in \mathcal{D}_{\boldsymbol{x}^*} : \|y\|_A^2 \leq s\}} \mathrm{V}_{\tilde{\mathrm{P}}_\beta}(y),$$
$$\psi_{2,A}(s) := \inf_{\{y \in \mathcal{D}_{\boldsymbol{x}^*} : \|y\|_A^2 \geq s\}} \mathrm{W}_{\mathrm{Q}_\beta}(y).$$

Before preceding with our proof strategy we show that the infimum and supremum are taken over nonempty sets. Consider the bijective map $\ell : \mathbb{R}^{2Nd} \to \mathbb{R}^{3Nd}$ given by

$$\ell(\boldsymbol{x},\boldsymbol{z}) := (\boldsymbol{x}-\boldsymbol{x}^*, \boldsymbol{z}-\boldsymbol{z}^*, \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*)),$$

which is continuous with image $\ell(\mathbb{R}^{2Nd}) = \mathcal{D}_{\boldsymbol{x}^*}$. We deduce that, as $(\boldsymbol{x},\boldsymbol{z})$ ranges over $\mathbb{R}^{2Nd}$, the norm $\|\ell(\boldsymbol{x},\boldsymbol{z})\|_A^2 = \|\boldsymbol{x}-\boldsymbol{x}^*\|_2^2 + q(\boldsymbol{z}-\boldsymbol{z}^*) + \|\tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*)\|_2^2$ takes all the values in $\mathbb{R}_{\geq 0}$ (because the composition is continuous), where

$$q(y) := y^\top(\mathbf{L}+\mathbf{L}^\top)y$$

for $y \in (\mathbb{R}^d)^N$. Therefore, the sets $\{y \in \mathcal{D}_{\boldsymbol{x}^*} : \|y\|_A \geq s\}$ and $\{y \in \mathcal{D}_{\boldsymbol{x}^*} : \|y\|_A \leq s\}$ are nonempty for each $s \geq 0$.

Our proof strategy consists of showing that for all $y \in \mathcal{D}_{\boldsymbol{x}^*}$ it holds that

$$\mathrm{V}_{\hat{\mathrm{P}}_\beta}(y) \leq \phi_{2,A}\Big(\|y\|_A^2\Big) \leq \bar{\alpha}_2\Big(\|y\|_A^2\Big), \tag{3.35a}$$

$$\bar{\alpha}_3\Big(\|y\|_A^2\Big) \leq \psi_{2,A}\Big(\|y\|_A^2\Big) \leq \mathrm{W}_{\mathrm{Q}_\beta}(y). \tag{3.35b}$$

If this were the case, then the result would follow by defining $\eta(r) = \bar{\alpha}_2(\bar{\alpha}_3^{-1}(r))$. For convenience, we use the shorthand notation $\tilde{\boldsymbol{x}} := \boldsymbol{x}-\boldsymbol{x}^*$, $\tilde{\boldsymbol{z}} := \boldsymbol{z}-\boldsymbol{z}^*$, and $\Delta\tilde{\mathbf{F}}_\epsilon(\tilde{\boldsymbol{x}}) := \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*)$. Also, we define the matrices

$$\hat{\mathrm{Q}} := \operatorname{diag}\Big(\mathbf{L}+\mathbf{L}^\top, \mathbf{L}+\mathbf{L}^\top, \mathrm{I}\Big), \quad \hat{\mathrm{P}} := \operatorname{diag}\Big(\mathrm{I}, \mathbf{L}+\mathbf{L}^\top, 0\Big).$$

Regarding (3.35b), note that $\hat{\mathrm{Q}}$ and Q_β are positive semidefinite with $\mathcal{N}(\hat{\mathrm{Q}}) = \mathcal{N}(\mathrm{Q}_\beta)$ by Proposition 3.11*(ii)*, and hence $c_1 \tilde{w}^\top \hat{\mathrm{Q}} \tilde{w} \leq \tilde{w}^\top \mathrm{Q}_\beta \tilde{w}$ for all $\tilde{w} \in \mathcal{D}_{\boldsymbol{x}^*}$, with

$c_1 := \lambda_{(3N-2)d}(\mathrm{Q}_\beta)/\lambda_{\max}(\hat{\mathrm{Q}})$. For each $s > 0$, we then have

$$\begin{aligned}
\psi_{2,A}(s) &= \inf_{\{\|\tilde{\boldsymbol{x}}\|_2^2 + q(\tilde{\boldsymbol{z}}) + \|\Delta\tilde{\mathbf{F}}_\epsilon(\tilde{\boldsymbol{x}})\|_2^2 \geq s\}} \mathrm{W}_{\mathrm{Q}_\beta}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{z}}, \Delta\tilde{\mathbf{F}}_\epsilon(\tilde{\boldsymbol{x}})) \\
&\geq \inf_{\{\|\tilde{\boldsymbol{x}}\|_2^2 + q(\tilde{\boldsymbol{z}}) + \|\Delta\tilde{\mathbf{F}}_\epsilon(\tilde{\boldsymbol{x}})\|_2^2 \geq s\}} c_1\Big(q(\tilde{\boldsymbol{x}}) + q(\tilde{\boldsymbol{z}}) + \|\Delta\tilde{\mathbf{F}}_\epsilon(\tilde{\boldsymbol{x}})\|_2^2\Big) \\
&\geq \inf_{\{q(\tilde{\boldsymbol{z}}) + (1+K_2^2)\|\tilde{\boldsymbol{x}}\|_2^2 \geq s\}} c_1\Big(q(\tilde{\boldsymbol{x}}) + q(\tilde{\boldsymbol{z}}) + K_1\|\tilde{\boldsymbol{x}}\|_2^2\Big) \\
&\geq \inf_{\{q(\tilde{\boldsymbol{z}}) + (1+K_2^2)\|\tilde{\boldsymbol{x}}\|_2^2 \geq s\}} \min\{c_1\hat{c}s\,,\, c_1 s\} = c_1 \min\{\hat{c}, 1\}\, s,
\end{aligned}$$

where $\hat{c} := K_1/(1+K_2^2)$, in the second inequality we have used that for each $s > 0$,

$$\Big\{\|\tilde{\boldsymbol{x}}\|_2^2 + q(\tilde{\boldsymbol{z}}) + \|\Delta\tilde{\mathbf{F}}_\epsilon(\tilde{\boldsymbol{x}})\|_2^2 \geq s\Big\} \subseteq \Big\{q(\tilde{\boldsymbol{z}}) + (1+K_2^2)\|\tilde{\boldsymbol{x}}\|_2^2 \geq s\Big\},$$

(which follows from (3.26)), and in the last inequality we have used

$$q(\tilde{\boldsymbol{x}}) + q(\tilde{\boldsymbol{z}}) + K_1\|\tilde{\boldsymbol{x}}\|_2^2 \geq \min\{\hat{c}, 1\}\Big(q(\tilde{\boldsymbol{z}}) + (1+K_2^2)\|\tilde{\boldsymbol{x}}\|_2^2\Big).$$

Thus, the linear gain in (3.35b) is $\bar{\alpha}_3(r) := c_1 \min\{\hat{c}, 1\} r$. Regarding (3.35a), we proceed similarly: $\hat{\mathrm{P}}$ and $\tilde{\mathrm{P}}_\beta$ are positive semidefinite with $\mathcal{N}(\hat{\mathrm{P}}) = \mathcal{N}(\tilde{\mathrm{P}}_\beta)$ by Proposition 3.11*(i)*, and hence $\tilde{w}^\top \tilde{\mathrm{P}}_\beta \tilde{w} \leq \bar{c}_2 \tilde{w}^\top \hat{\mathrm{P}} \tilde{w}$ for all $\tilde{w} \in \mathcal{D}_{\boldsymbol{x}^*}$, with $\bar{c}_2 := \lambda_{\max}(\tilde{\mathrm{P}}_\beta)/\lambda_{(2N-1)d}(\hat{\mathrm{P}})$. We then have

$$\begin{aligned}
\phi_{2,A}(s) &= \sup_{\{\|\tilde{\boldsymbol{x}}\|_2^2 + q(\tilde{\boldsymbol{z}}) + \|\Delta\tilde{\mathbf{F}}_\epsilon(\tilde{\boldsymbol{x}})\|_2^2 \leq s\}} \mathrm{V}_{\tilde{\mathrm{P}}_\beta}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{z}}, \Delta\tilde{\mathbf{F}}_\epsilon(\tilde{\boldsymbol{x}})) \\
&\leq \sup_{\{\|\tilde{\boldsymbol{x}}\|_2^2 + q(\tilde{\boldsymbol{z}}) + \|\Delta\tilde{\mathbf{F}}_\epsilon(\tilde{\boldsymbol{x}})\|_2^2 \leq s\}} \bar{c}_2\Big(\|\tilde{\boldsymbol{x}}\|_2^2 + q(\tilde{\boldsymbol{z}})\Big) \\
&\leq \sup_{\{\|\tilde{\boldsymbol{x}}\|_2^2 + q(\tilde{\boldsymbol{z}}) + \|\Delta\tilde{\mathbf{F}}_\epsilon(\tilde{\boldsymbol{x}})\|_2^2 \leq s\}} \bar{c}_2\Big(\|\tilde{\boldsymbol{x}}\|_2^2 + q(\tilde{\boldsymbol{z}}) + \|\Delta\tilde{\mathbf{F}}_\epsilon(\tilde{\boldsymbol{x}})\|_2^2\Big) = \bar{c}_2 s.
\end{aligned}$$

Thus, the linear gain in (3.35a) is $\bar{\alpha}_2(r) := \bar{c}_2 r$. Tracking now the composition of functions $\eta(r) = \bar{\alpha}_2(\bar{\alpha}_3^{-1}(r)) := C_\eta r$, we have

$$C_\eta = \frac{\lambda_{\max}(\hat{\mathrm{Q}})\,\lambda_{\max}(\tilde{\mathrm{P}}_\beta)}{\min\{1, \frac{K_1}{(1+K_2^2)}\}\,\lambda_{(3N-2)d}(\mathrm{Q}_\beta)\,\lambda_{(2N-1)d}(\hat{\mathrm{P}})},$$

which yields the expression in the statement. □

3.3.5 Completing the proof of the main result

The combination of the above developments leads us here to the proof of Theorem 2.5.

Proof of Theorem 2.5. By Proposition 3.11, the function V also satisfies (3.29) and (3.30). Additionally, from Proposition 3.12, for all $v \in \mathbb{R}^{2Nd}$ we have

$$\mathrm{V}(v) = \mathrm{V}_{\tilde{\mathrm{P}}_\beta}(v - v^*, \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*)) \leq \eta(\mathrm{W}_{\mathrm{Q}_\beta}(v - v^*, \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}) - \tilde{\mathbf{F}}_\epsilon(\boldsymbol{x}^*))) = \eta(\mathrm{W}(v)).$$

Therefore, as defined in the hypotheses of Theorem 6.2, V is a second moment NSS-Lyapunov function for the dynamics (3.5) with respect to the affine subspace

$$[\mathbb{1}^\top \otimes x_{\min}^\top, \boldsymbol{z}^{*\top}]^\top + \mathcal{N}(\hat{\mathrm{I}}) = [\mathbb{1}^\top \otimes x_{\min}^\top, \boldsymbol{z}^{*\top}]^\top + \operatorname{span}\Big\{\Big[0 \quad (\mathbb{1} \otimes b)^\top\Big]^\top : b \in \mathbb{R}^d\Big\}.$$

Applying Theorem 6.2, we conclude that the dynamics (3.5) is second moment NSS stable with respect to the same affine subspace with

$$\mu(r,s) := \alpha_1^{-1}\Big(2\tilde{\mu}(\alpha_2(r^p), s)\Big) = \frac{2\,\lambda_{\max}(\mathrm{P}_\beta) r^2}{\lambda_{(2N-1)d}(\mathrm{P}_\beta)} \exp\Big(-\tfrac{1}{2C_\eta}\, s\Big),$$

$$\theta(r) := \alpha_1^{-1}\Big(2\eta(2\sigma(r))\Big) = \frac{4C_\eta \operatorname{trace}(\mathrm{P}_\beta)\,\kappa_2^2}{\lambda_{(2N-1)d}(\mathrm{P}_\beta)}\, r^2,$$

where κ_2 is such that (3.11) holds, and C_η is defined in Proposition 3.12. Substituting the value of C_η, the function $\theta(r)$ can be written as $C_\theta r^2$ for the expression of C_θ in Remark 2.7. □

3.4 Discussion

We have considered a multi-agent network communicating over a weight-balanced, strongly connected digraph that seeks to collectively solve a convex optimization problem defined by a sum of local functions, one per agent, in the presence of noise both in the communication channels and in the agent computations. We have studied the robustness properties against additive persistent noise of a family of distributed continuous-time algorithms that have each agent update its estimate by following the gradient of its local cost function while, at the same time, seeking to agree with its neighbors' estimates via proportional-integral feedback on their disagreement. Specifically, we have established that the proposed class of algorithms is noise-to-state exponentially stable in second moment. Our strategy to establish this result has relied on constructing a function whose nullset is the solution to the optimization problem plus a direction of variance accumulation in some auxiliary variables, and then showing that in fact this is a NSS-Lyapunov function using the co-coercivity properties of the vector fields that define the dynamics.

Acknowledgments

This chapter is taken from a work submitted by D. Mateos-Núñez and J. Cortés to *SIAM Journal on Control and Optimization,* with the title "Noise-to-state exponentially stable distributed convex optimization on weight-balanced digraphs." A preliminary version appeared in the *Proceedings of the IEEE Conference on Decision and Control, Florence, Italy, 2013, pp. 2781-2786.* This research was supported by the US National Science Foundation (NSF) Award CMMI-1300272.

Chapter 4

Distributed online convex optimization over jointly connected digraphs

In this chapter we consider networked online convex optimization scenarios from a regret analysis perspective. At each round, each agent in the network commits to a decision and incurs in a local cost given by functions that are revealed over time and whose evolution might be adversarially adaptive to the agent's behavior. The goal of each agent is to incur a cumulative cost over time with respect to the sum of local functions across the network that is competitive with the best single centralized decision in hindsight. To achieve this, agents cooperate with each other using local averaging over time-varying weight-balanced digraphs as well as subgradient descent on the local cost functions revealed in the previous round. We propose a class of coordination algorithms that generalize distributed online subgradient descent and saddle-point dynamics, allowing proportional-integral (and higher-order) feedback on the disagreement among neighboring agents. We show that our algorithm design achieves logarithmic agent regret (when local objectives are strongly convex), or square-root agent regret (when local objectives are convex) in scenarios where the communication graphs are jointly connected. We illustrate the application of our results for medical diagnosis

4.1 Problem statement

We begin by describing the online convex optimization problem for one player and then present the networked version, which is the focus of the paper. In online convex optimization, given a time horizon $T \in \mathbb{Z}_{\geq 1}$, in each round $t \in \{1,\dots,T\}$ a player chooses a point $x_t \in \mathbb{R}^d$. After committing to this choice, a convex cost function $f_t : \mathbb{R}^d \to \mathbb{R}$ is revealed. Consequently, the 'cost' incurred by the player is $f_t(x_t)$. Given the temporal sequence of objectives $\{f_t\}_{t=1}^T$, the regret of the player using $\{x_t\}_{t=1}^T$ with respect to a single choice $u \in \mathbb{R}^d$ in hindsight over a time horizon T is defined by

$$\mathcal{R}(u, \{f_t\}_{t=1}^T) := \sum_{t=1}^T f_t(x_t) - \sum_{t=1}^T f_t(u), \tag{4.1}$$

i.e., the difference between the total cost incurred by the online estimates $\{x_t\}_{t=1}^T$ and the cost of a single hindsight decision u. A logical choice, if it exists, is the best decision over a time horizon T had all the information been available a priori, i.e.,

$$u = \hat{x}_T \in \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \sum_{t=1}^T f_t(x).$$

In the case when no information is available about the evolution of the functions $\{f_t\}_{t=1}^T$, one is interested in designing algorithms whose worst-case regret is upper bounded sublinearly in the time horizon T with respect to any decision in hindsight.

This ensures that, on average, the algorithm performs nearly as well as the best single decision in hindsight.

We now explain the distributed version of the online convex optimization problem where the online player is replaced by a network of N agents, each with access to partial information. In the round $t \in \{1,\dots,T\}$, agent $i \in \{1,\dots,N\}$ chooses a point x_t^i corresponding to what it thinks the network as a whole should have chosen. After committing to this choice, the agent has access to a convex cost function $f_t^i : \mathbb{R}^d \to \mathbb{R}$ and the network cost is then given by the evaluation of

$$f_t(x) := \sum_{i=1}^{N} f_t^i(x). \tag{4.2}$$

Note that this function is not known to any of the agents and is not available at any single location. In this scenario, the regret of agent $j \in \{1,\dots,N\}$ using $\{x_t^j\}_{t=1}^T$ with respect to a single choice u in hindsight over a time horizon T is

$$\mathcal{R}^j(u, \{f_t\}_{t=1}^T) := \sum_{t=1}^{T}\sum_{i=1}^{N} f_t^i(x_t^j) - \sum_{t=1}^{T}\sum_{i=1}^{N} f_t^i(u).$$

The goal then is to design coordination algorithms among the agents that guarantee that the worst-case agent regret is upper bounded sublinearly in the time horizon T with respect to any decision in hindsight. This would guarantee that each agent incurs an average cost over time with respect to the sum of local cost functions across the network that is nearly as low as the cost of the best single choice had all the information been centrally available a priori. Since information is now distributed across the network, agents must collaborate with each other to determine their decisions for the next round. We assume that the network communication topology is time-dependent and described by a sequence of weight-balanced digraphs $\{\mathcal{G}_t\}_{t=1}^T = \{(\{1,\dots,N\}, \mathcal{E}_t, \mathsf{A}_t)\}_{t=1}^T$. At each round, agents can use historical observations of locally revealed cost functions and become aware through local communication of the choices made by their neighbors in the previous round.

4.2 Dynamics for distributed online optimization

In this section we propose a distributed coordination algorithm to solve the networked online convex optimization problem described in Section 4.1. In each round $t \in \{1,\dots,T\}$, agent $i \in \{1,\dots,N\}$ performs

$$\begin{aligned} x^i_{t+1} &= x^i_t + \sigma\Big(a\sum_{j=1}^N \mathsf{a}_{ij,t}(x^j_t - x^i_t) + \sum_{j=1}^N \mathsf{a}_{ij,t}(z^j_t - z^i_t)\Big) - \eta_t g_{x^i_t}, \\ z^i_{t+1} &= z^i_t - \sigma\sum_{j=1}^N \mathsf{a}_{ij,t}(x^j_t - x^i_t), \end{aligned} \tag{4.3}$$

where $g_{x^i_t} \in \partial f^i_t(x^i_t)$, the scalars σ, $a \in \mathbb{R}_{>0}$ are design parameters, and $\eta_t \in \mathbb{R}_{>0}$ is the learning rate at time t. Agent i is responsible for the variables x^i, z^i, and shares their values with its neighbors according to the time-dependent digraph $\mathcal{G}_t$. Note that (4.3) is both consistent with the notion of incremental access to information by individual agents and is distributed over $\mathcal{G}_t$: each agent updates its estimate by following a subgradient of the cost function revealed to it in the previous round while, at the same time, seeking to agree with its neighbors' estimates. The latter is implemented through a second-order process that employs proportional-integral feedback on the disagreement. Our design is inspired by and extends the distributed algorithms for distributed optimization of a sum of convex functions studied in [WE11, GC14]. We use the term *online subgradient descent algorithm with proportional-integral disagreement feedback* to refer to (4.3).

We next rewrite the dynamics in compact form. To do so, we introduce the notation $\boldsymbol{x} := (x^1,\dots,x^N) \in (\mathbb{R}^d)^N$ and $\boldsymbol{z} := (z^1,\dots,z^N) \in (\mathbb{R}^d)^N$ to denote the aggregate of the agents' online decisions and the aggregate of the agents' auxiliary variables, respectively. For $t \in \{1,\dots,T\}$, we also define the convex function $\tilde{f}_t : (\mathbb{R}^d)^N \to \mathbb{R}$ by

$$\tilde{f}_t(\boldsymbol{x}) := \sum_{i=1}^{N} f_t^i(x^i). \tag{4.4}$$

When all agents agree on the same choice, one recovers the value of the network cost function (4.2), $\tilde{f}_t(\mathbb{1}_N \otimes x) = f_t(x)$. With this notation in place, the algorithm (4.3) takes the form

$$\begin{bmatrix} \boldsymbol{x}_{t+1} \\ \boldsymbol{z}_{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{z}_t \end{bmatrix} - \sigma \begin{bmatrix} a\mathbf{L}_t & \mathbf{L}_t \\ -\mathbf{L}_t & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{z}_t \end{bmatrix} - \eta_t \begin{bmatrix} \tilde{g}_{\boldsymbol{x}_t} \\ 0 \end{bmatrix}, \tag{4.5}$$

where $\mathbf{L}_t := \mathsf{L}_t \otimes \mathrm{I}_d$ and $\tilde{g}_{\boldsymbol{x}_t} = (g_{x_t^1}, \dots, g_{x_t^N}) \in \partial \tilde{f}_t(\boldsymbol{x}_t)$.

This compact-form representation suggests a more general class of distributed dynamics that includes (4.3) as a particular case. For $K \in \mathbb{Z}_{\geq 1}$, let $E \in \mathbb{R}^{K \times K}$ be diagonalizable with real positive eigenvalues, and define $\mathbb{L}_t := E \otimes \mathbf{L}_t$. Consider the dynamics on $((\mathbb{R}^d)^N)^K$ defined by

$$\boldsymbol{v}_{t+1} = (\mathrm{I}_{KNd} - \sigma \mathbb{L}_t)\boldsymbol{v}_t - \eta_t \mathbf{g}_t, \tag{4.6}$$

where $\mathbf{g}_t \in ((\mathbb{R}^d)^N)^K$ takes the form

$$\mathbf{g}_t := (\tilde{g}_{\boldsymbol{x}_t}, 0, \dots, 0), \tag{4.7}$$

and we use the decomposition $\boldsymbol{v} = (\boldsymbol{x}, \boldsymbol{v}^2, \dots, \boldsymbol{v}^K)$. Throughout the paper, our convergence results are formulated for this dynamics because of its generality, which we discuss in the following remark.

Remark 2.13. (Online subgradient descent algorithms with proportional and proportional-integral disagreement feedback). *The online subgradient descent algorithm with proportional-integral disagreement feedback* (4.3) *corresponds to the dynamics* (4.6) *with the choices* $K = 2$ *and*

$$E = \begin{bmatrix} a & 1 \\ -1 & 0 \end{bmatrix}.$$

For $a \in (2, \infty)$, E *has positive eigenvalues* $\lambda_{min}(E) = \frac{a}{2} - \sqrt{(\frac{a}{2})^2 - 1}$ *and* $\lambda_{max}(E) = \frac{a}{2} + \sqrt{(\frac{a}{2})^2 - 1}$. *Interestingly, the online subgradient descent algorithm with proportional disagreement feedback proposed in [YSVQ13] (without the projection component onto a bounded convex set) also corresponds to the dynamics* (4.6) *with the choices* $K = 1$ *and* $E = [1]$. ●

Our forthcoming exposition presents the technical approach to establish the properties of the distributed dynamics (4.6) with respect to the agent regret defined in Section 4.1. An informal description of our main results is as follows. Under mild conditions on the connectivity of the communication network, a suitable choice of σ, and the assumption that the time-dependent local cost functions have bounded subgradient sets and uniformly bounded optimizers, the following bounds hold:

Logarithmic agent regret: if each local cost function is locally p-strongly convex and $\eta_t = \frac{1}{pt}$, then any sequence generated by the dynamics (4.6) satisfies, for each $j \in \{1, \dots, N\}$,

$$\mathcal{R}^j\Big(u, \{f_t\}_{t=1}^T\Big) \in \mathcal{O}(\|u\|_2^2 + \log T).$$

Square-root agent regret: if each local cost function is convex (plus a mild geometric assumption) and, for $m = 0,1,2,\dots,\lceil \log_2 T \rceil$, we take $\eta_t = \frac{1}{\sqrt{2^m}}$ in each period of 2^m rounds $t = 2^m,\dots,2^{m+1}-1$, then any sequence generated by the dynamics (4.6) satisfies, for each $j \in \{1,\dots,N\}$,

$$\mathcal{R}^j\Big(u,\{f_t\}_{t=1}^T\Big) \in \mathcal{O}(\|u\|_2^2\sqrt{T}).$$

In our technical approach to establish these sublinear agent regret bounds, we find it useful to consider the notion of network regret [DGBSX12, TR12] with respect to a single hindsight choice $u \in \mathbb{R}^d$ over the time horizon T,

$$\mathcal{R}_{\mathcal{N}}(u,\{\tilde{f}_t\}_{t=1}^T) := \sum_{t=1}^T \tilde{f}_t(\boldsymbol{x}_t) - \sum_{t=1}^T \tilde{f}_t(\mathbb{1}_N \otimes u),$$

to capture the performance of the sequence of collective estimates $\{\boldsymbol{x}_t\}_{t=1}^T \subseteq (\mathbb{R}^d)^N$. Our proof strategy builds on this concept and relies on bounding the following terms:

(i) both the network regret and the difference between the agent and network regrets;

(ii) the cumulative disagreement of the collective estimates;

(iii) the sequence of collective estimates uniformly in the time horizon.

Section 4.3 presents the formal discussion for these results. The combination of these steps allows us in Section 4.4 to formally establish the sublinear agent regret bounds outlined above.

4.3 Regret analysis

This section presents the results outlined above on bounding the agent and network regrets, the cumulative disagreement of the collective estimates, and the sequence of collective estimates for executions of the distributed dynamics (4.6). These results are instrumental later in the derivation of the sublinear agent regret bounds, but are also of independent interest.

4.3.1 Bounds on network and agent regret

Our first result relates the agent and network regrets for any sequence of collective estimates (regardless of the algorithm that generates them) in terms of their cumulative disagreement.

Lemma 3.14. (Bound on the difference between agent and network regret). *For $T \in \mathbb{Z}_{\geq 1}$, let $\{f_t^1, \dots, f_t^N\}_{t=1}^T$ be convex functions on $\mathbb{R}^d$ with H-bounded subgradient sets. Then, any sequence $\{\boldsymbol{x}_t\}_{t=1}^T \subset (\mathbb{R}^d)^N$ satisfies, for any $j \in \{1, \dots, N\}$ and $u \in \mathbb{R}^d$,*

$$\mathcal{R}^j(u, \{f_t\}_{t=1}^T) \leq \mathcal{R}_{\mathcal{N}}(u, \{\tilde{f}_t\}_{t=1}^T) + NH \sum_{t=1}^T \|\mathbf{L}_{\mathcal{K}} \boldsymbol{x}_t\|_2,$$

where $\mathbf{L}_{\mathcal{K}} := \mathsf{L}_{\mathcal{K}} \otimes \mathrm{I}_d$.

Proof. Since $\tilde{f}_t(\mathbb{1}_N \otimes x) = f_t(x)$ for all $x \in \mathbb{R}^d$, we have

$$\mathcal{R}^j(u, \{f_t\}_{t=1}^T) - \mathcal{R}_{\mathcal{N}}(u, \{\tilde{f}_t\}_{t=1}^T) = \sum_{t=1}^T \Big(\tilde{f}_t(\mathbb{1}_N \otimes x_t^j) - \tilde{f}_t(\boldsymbol{x}_t) \Big). \qquad (4.8)$$

The convexity of $\tilde{f}_t$ implies that, for any $\xi \in \partial \tilde{f}_t(\mathbb{1}_N \otimes x_t^j)$,

$$\begin{aligned}\tilde{f}_t(\mathbb{1}_N \otimes x_t^j) - \tilde{f}_t(\boldsymbol{x}_t) \leq & \xi^\top(\mathbb{1}_N \otimes x_t^j - \boldsymbol{x}_t) \\ & \leq \|\xi\|_2 \|\mathbb{1}_N \otimes x_t^j - \boldsymbol{x}_t\|_2 \leq \sqrt{N} H \|\mathbb{1}_N \otimes x_t^j - \boldsymbol{x}_t\|_2, \end{aligned} \tag{4.9}$$

where we have used the Cauchy-Schwarz inequality and the fact that the subgradient sets are H-bounded. In addition,

$$\begin{aligned}\|\mathbb{1}_N \otimes x_t^j - \boldsymbol{x}_t\|_2^2 = & \sum_{i=1}^N \|x_t^j - x_t^i\|_2^2 \\ & \leq \frac{1}{2} \sum_{j=1}^N \sum_{i=1}^N \|x_t^j - x_t^i\|_2^2 = N \boldsymbol{x}_t^\top \mathbf{L}_\mathcal{K} \boldsymbol{x}_t. \end{aligned} \tag{4.10}$$

The fact that $\mathbf{L}_\mathcal{K}^2 = \mathbf{L}_\mathcal{K} = \mathbf{L}_\mathcal{K}^\top$ allows us to write $\boldsymbol{x}_t^\top \mathbf{L}_\mathcal{K} \boldsymbol{x}_t = \|\mathbf{L}_\mathcal{K} \boldsymbol{x}_t\|_2^2$. The result now follows using (4.9) and (4.10) in conjunction with (4.8). □

Next, we bound the network regret for executions of the coordination algorithm (4.6) in terms of the learning rates and the cumulative disagreement. The bound holds regardless of the connectivity of the communication network as long as the digraph remains weight-balanced.

Lemma 3.15. (Bound on network regret). *For $T \in \mathbb{Z}_{\geq 1}$, let $\{f_t^1, \dots, f_t^N\}_{t=1}^T$ be convex functions on $\mathbb{R}^d$ with H-bounded subgradient sets. Let the sequence $\{\boldsymbol{x}_t\}_{t=1}^T$ be generated by the coordination algorithm* (4.6) *over a sequence of arbitrary weight-balanced digraphs $\{\mathcal{G}_t\}_{t=1}^T$. Then, for any $u \in \mathbb{R}^d$, and any sequence of* *learning rates* $\{\eta_t\}_{t=1}^T \subset \mathbb{R}_{>0}$,

$$2\mathcal{R}_{\mathcal{N}}\Big(u, \{\tilde{f}_t\}_{t=1}^T\Big) \leq \sum_{t=2}^T \|\mathbf{M}\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 \Big(\tfrac{1}{\eta_t} - \tfrac{1}{\eta_{t-1}} - p_t(\boldsymbol{u}, \boldsymbol{x}_t)\Big) \\ + 2\sqrt{N}H \sum_{t=1}^T \|\mathbf{L}_{\mathcal{K}}\boldsymbol{x}_t\|_2 + NH^2 \sum_{t=1}^T \eta_t + \tfrac{1}{\eta_1}\|\mathbf{M}\boldsymbol{x}_1 - \boldsymbol{u}\|_2^2,$$

where $\mathbf{M} := \mathrm{M} \otimes \mathrm{I}_d$, $\boldsymbol{u} := \mathbb{1}_N \otimes u$ *and* $p_t : (\mathbb{R}^d)^N \times (\mathbb{R}^d)^N \to \mathbb{R}_{\geq 0}$ *is the modulus of strong convexity of* $\tilde{f}_t$.

Proof. Left-multiplying the dynamics (4.6) by the block-diagonal matrix $\mathrm{diag}(1, 0, \ldots, 0) \otimes \mathbf{M} \in \mathbb{R}^{(Nd)K \times (Nd)K}$, and using $\mathbf{M}\mathbf{L}_t = 0$, we obtain the following projected dynamics

$$\mathbf{M}\boldsymbol{x}_{t+1} = \mathbf{M}\boldsymbol{x}_t - \eta_t \mathbf{M}\tilde{g}_{\boldsymbol{x}_t}. \tag{4.11}$$

Note that this dynamics is decoupled from the dynamics of $\boldsymbol{v}_t^2, \ldots \boldsymbol{v}_t^K$. Subtracting $\boldsymbol{u}$ and taking the norm on both sides, we get $\|\mathbf{M}\boldsymbol{x}_{t+1} - \boldsymbol{u}\|_2^2 = \|\mathbf{M}\boldsymbol{x}_t - \boldsymbol{u} - \eta_t \mathbf{M}\tilde{g}_{\boldsymbol{x}_t}\|_2^2$, so that

$$\begin{aligned} \|\mathbf{M}\boldsymbol{x}_{t+1} - \boldsymbol{u}\|_2^2 =& \|\mathbf{M}\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 + \eta_t^2 \|\mathbf{M}\tilde{g}_{\boldsymbol{x}_t}\|_2^2 - 2\eta_t (\mathbf{M}\tilde{g}_{\boldsymbol{x}_t})^\top (\mathbf{M}\boldsymbol{x}_t - \boldsymbol{u}) \\ =& \|\mathbf{M}\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 + \eta_t^2 \|\mathbf{M}\tilde{g}_{\boldsymbol{x}_t}\|_2^2 - 2\eta_t \tilde{g}_{\boldsymbol{x}_t}^\top (\mathbf{M}\boldsymbol{x}_t - \boldsymbol{u}), \end{aligned} \tag{4.12}$$

where we have used $\mathbf{M}^2 = \mathbf{M}$ and $\mathbf{M}\boldsymbol{u} = \boldsymbol{u}$. Regarding the last term, note that

$$\begin{aligned} -\tilde{g}_{\boldsymbol{x}_t}^\top (\mathbf{M}\boldsymbol{x}_t - \boldsymbol{u}) =& -\tilde{g}_{\boldsymbol{x}_t}^\top (\mathbf{M}\boldsymbol{x}_t - \boldsymbol{x}_t) - \tilde{g}_{\boldsymbol{x}_t}^\top (\boldsymbol{x}_t - \boldsymbol{u}) \\ &\leq \tilde{g}_{\boldsymbol{x}_t}^\top \mathbf{L}_{\mathcal{K}}\boldsymbol{x}_t + \tilde{f}_t(\boldsymbol{u}) - \tilde{f}_t(\boldsymbol{x}_t) - \tfrac{p_t(\boldsymbol{u}, \boldsymbol{x}_t)}{2}\|\boldsymbol{u} - \boldsymbol{x}_t\|_2^2, \end{aligned}$$

where we have used $\mathbf{L}_{\mathcal{K}} = \mathrm{I}_{Nd} - \mathbf{M}$. Substituting into (4.12), we obtain

$$\begin{aligned}\|\mathbf{M}\boldsymbol{x}_{t+1} - \boldsymbol{u}\|_2^2 \leq \|\mathbf{M}\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 + \eta_t^2 \|\mathbf{M}\tilde{g}_{\boldsymbol{x}_t}\|_2^2 \\ + 2\eta_t \Big(\tilde{g}_{\boldsymbol{x}_t}^\top \mathbf{L}_{\mathcal{K}} \boldsymbol{x}_t + \tilde{f}_t(\boldsymbol{u}) - \tilde{f}_t(\boldsymbol{x}_t) - \tfrac{p_t(\boldsymbol{u},\boldsymbol{x}_t)}{2}\|\boldsymbol{u} - \boldsymbol{x}_t\|_2^2\Big),\end{aligned}$$

so that, reordering terms,

$$\begin{aligned}2(\tilde{f}_t(\boldsymbol{x}_t) - \tilde{f}_t(\boldsymbol{u})) \leq \tfrac{1}{\eta_t}\Big(\|\mathbf{M}\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 - \|\mathbf{M}\boldsymbol{x}_{t+1} - \boldsymbol{u}\|_2^2\Big) \qquad (4.13)\\ - p_t(\boldsymbol{u},\boldsymbol{x}_t)\|\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 + 2\tilde{g}_{\boldsymbol{x}_t}^\top \mathbf{L}_{\mathcal{K}} \boldsymbol{x}_t + \eta_t \|\mathbf{M}\tilde{g}_{\boldsymbol{x}_t}\|_2^2. \qquad (4.14)\end{aligned}$$

Next, we bound each of the terms appearing in the last line of (4.13). For the term $p_t(\boldsymbol{u},\boldsymbol{x}_t)\|\boldsymbol{x}_t - \boldsymbol{u}\|_2^2$, note that

$$\begin{aligned}\|\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 = \|(\mathbf{M} + \mathbf{L}_{\mathcal{K}})(\boldsymbol{x}_t - \boldsymbol{u})\|_2^2 = \|\mathbf{M}(\boldsymbol{x}_t - \boldsymbol{u})\|_2^2 \\ + \|\mathbf{L}_{\mathcal{K}}(\boldsymbol{x}_t - \boldsymbol{u})\|_2^2 + 2(\boldsymbol{x}_t - \boldsymbol{u})^\top \mathbf{M}\mathbf{L}_{\mathcal{K}}(\boldsymbol{x}_t - \boldsymbol{u}) \\ = \|\mathbf{M}\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 + \|\mathbf{L}_{\mathcal{K}}\boldsymbol{x}_t\|_2^2, \qquad (4.15)\end{aligned}$$

where we have used $\mathbf{M}\mathbf{L}_{\mathcal{K}} = 0$ and $\mathbf{M}\boldsymbol{u} = \boldsymbol{u}$. Regarding the term $2\tilde{g}_{\boldsymbol{x}_t}^\top \mathbf{L}_{\mathcal{K}} \boldsymbol{x}_t$, note that $\|\tilde{g}_{\boldsymbol{x}_t}\|_2^2 = \sum_{i=1}^N \|g_{x_t^i}\|_2^2 \leq NH^2$ because the subgradient sets are bounded by H. Hence, using the Cauchy-Schwarz inequality,

$$\tilde{g}_{\boldsymbol{x}_t}^\top \mathbf{L}_{\mathcal{K}} \boldsymbol{x}_t \leq \|\tilde{g}_{\boldsymbol{x}_t}\|_2 \|\mathbf{L}_{\mathcal{K}}\boldsymbol{x}_t\|_2 \leq \sqrt{N}H\|\mathbf{L}_{\mathcal{K}}\boldsymbol{x}_t\|_2. \qquad (4.16)$$

Finally, regarding the term $\eta_t \|\mathbf{M}\tilde{g}_{\boldsymbol{x}_t}\|_2^2$ in (4.13), note that

$$\begin{aligned}
\|\mathbf{M}\tilde{g}_{\boldsymbol{x}_t}\|_2^2 &= \|\mathbb{1}_N \otimes \tfrac{1}{N}\sum_{i=1}^{N} g_{x_t^i}\|_2^2 = N\Big\|\tfrac{1}{N}\sum_{i=1}^{N} g_{x_t^i}\Big\|_2^2 \\
&= \tfrac{1}{N}\sum_{l=1}^{d}\Big(\sum_{i=1}^{N} g_{x_t^i}\Big)_l^2 \leq \tfrac{1}{N}\sum_{l=1}^{d}\Big(N\sum_{i=1}^{N}\big(g_{x_t^i}\big)_l^2\Big) \\
&= \sum_{i=1}^{N}\sum_{l=1}^{d}\big(g_{x_t^i}\big)_l^2 = \sum_{i=1}^{N}\|g_{x_t^i}\|_2^2 \leq NH^2,
\end{aligned} \tag{4.17}$$

where in the first inequality we have used the inequality of quadratic and arithmetic means [Bul03]. The result now follows from summing the expression in (4.13) over the time horizon T, discarding the negative terms, and using the upper bounds in (4.15)-(4.17). □

The combination of Lemmas 3.14 and 3.15 provides a bound on the agent regret in terms of the learning rates and the cumulative disagreement of the collective estimates. This motivates our next section.

4.3.2 Bound on cumulative disagreement

In this section we study the evolution of the disagreement among the agents' estimates under (4.6). Our analysis builds on the input-to-state stability (ISS) properties of the linear part of the dynamics with respect to the agreement subspace, where we treat the subgradient term as a perturbation. Consequently, here we study the dynamics

$$\boldsymbol{v}_{t+1} = (\mathrm{I}_{KNd} - \sigma\mathbb{L}_t)\boldsymbol{v}_t + \boldsymbol{d}_t, \tag{4.18}$$

where $\{\boldsymbol{d}_t\}_{t\geq 1} \subset ((\mathbb{R}^d)^N)^K$ is an arbitrary sequence of disturbances. Our first result shows that, for the purpose of studying the ISS properties of (4.18), the dynamics can be decoupled into K first-order linear consensus dynamics.

Lemma 3.16. (Decoupling into a collection of first-order consensus dynamics). *Given a diagonalizable matrix* $E \in \mathbb{R}^{K\times K}$ *with real eigenvalues, let* S_E *be the matrix of eigenvectors in the decomposition* $E = S_E D_E S_E^{-1}$*, with* $D_E = \operatorname{diag}(\lambda_1(E), \ldots, \lambda_K(E))$*. Then, under the change of variables*

$$\boldsymbol{w}_t := (S_E^{-1} \otimes \mathrm{I}_{Nd})\boldsymbol{v}_t, \tag{4.19}$$

the dynamics (4.18) *is equivalently represented by the collection of first-order dynamics on* $(\mathbb{R}^d)^N$ *defined by*

$$\boldsymbol{w}_{t+1}^l = (\mathrm{I}_{Nd} - \sigma\,\lambda_l(E)\mathbf{L}_t)\boldsymbol{w}_t^l + \boldsymbol{e}_t^l, \tag{4.20}$$

where $l \in \{1, \ldots, K\}$*,* $\boldsymbol{w}_t = (\boldsymbol{w}_t^1, \ldots, \boldsymbol{w}_t^K) \in ((\mathbb{R}^d)^N)^K$ *and*

$$\boldsymbol{e}_t^l := \Big((S_E^{-1} \otimes \mathrm{I}_{Nd})\boldsymbol{d}_t\Big)^l \in (\mathbb{R}^d)^N. \tag{4.21}$$

Moreover, for each $t \in \mathbb{Z}_{\geq 1}$*,*

$$\|\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{v}_t\|_2 \leq \|S_E\|_2\sqrt{K}\max_{1\leq l\leq K}\|\mathbf{L}_{\mathcal{K}}\boldsymbol{w}_t^l\|_2, \tag{4.22}$$

where $\hat{\mathbf{L}}_{\mathcal{K}} := \mathrm{I}_K \otimes \mathbf{L}_{\mathcal{K}}$*.*

Proof. We start by noting that

$$\begin{aligned}\mathbb{L}_t &= S_E\, D_E\, S_E^{-1} \otimes \mathrm{I}_{Nd}\, \mathbf{L}_t\, \mathrm{I}_{Nd} \\ &= (S_E \otimes \mathrm{I}_{Nd})\,(D_E \otimes \mathbf{L}_t)\,(S_E \otimes \mathrm{I}_{Nd})^{-1},\end{aligned}$$

and therefore we obtain the factorization

$$\mathrm{I}_{KNd} - \sigma \mathbb{L}_t = (S_E \otimes \mathrm{I}_{Nd})(\mathrm{I}_{KNd} - \sigma D_E \otimes \mathbf{L}_t)(S_E \otimes \mathrm{I}_{Nd})^{-1}.$$

Now, under the change of variables (4.19), the dynamics (4.18) takes the form

$$\boldsymbol{w}_{t+1} = (\mathrm{I}_{KNd} - \sigma\, D_E \otimes \mathbf{L}_t)\boldsymbol{w}_t + (S_E^{-1} \otimes \mathrm{I}_{Nd})\boldsymbol{d}_t, \tag{4.23}$$

which corresponds to the set of dynamics (4.20). Moreover,

$$\begin{aligned}\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{v}_t &= (\mathrm{I}_K \otimes \mathbf{L}_{\mathcal{K}})(S_E \otimes \mathrm{I}_{Nd})\boldsymbol{w}_t \\ &= (S_E \otimes \mathrm{I}_{Nd})(\mathrm{I}_K \otimes \mathbf{L}_{\mathcal{K}})\boldsymbol{w}_t = (S_E \otimes \mathrm{I}_{Nd})\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{w}_t.\end{aligned}$$

Hence, the sub-multiplicativity of the norm together with [Ber05, Fact 9.12.22] for the norms of Kronecker products, yields

$$\begin{aligned}\|\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{v}_t\|_2 &\le \|S_E \otimes \mathrm{I}_{Nd}\|_2 \|\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{w}_t\|_2 = \|S_E\|_2 \|\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{w}_t\|_2 \\ &= \|S_E\|_2 \Big(\sum_{l=1}^{K} \|\mathbf{L}_{\mathcal{K}}\boldsymbol{w}_t^l\|_2^2\Big)^{1/2} \le \|S_E\|_2 \sqrt{K} \max_{1\le l\le K} \|\mathbf{L}_{\mathcal{K}}\boldsymbol{w}_t^l\|_2,\end{aligned}$$

as claimed. □

In the next result, we use Lemma 3.16 to bound the cumulative disagreement of the collective estimates over time.

Proposition 3.17. (Input-to-state stability and cumulative disagreement of (4.18) over jointly connected weight-balanced digraphs). *Let $E \in \mathbb{R}^{K\times K}$ be a diagonalizable matrix with real positive eigenvalues and $\{\mathcal{G}_s\}_{s\ge 1}$ a sequence of B-jointly connected,*

δ-nondegenerate, weight-balanced digraphs. For $\tilde{\delta}' \in (0,1)$, let

$$\tilde{\delta} := \min\Big\{\tilde{\delta}',\,(1-\tilde{\delta}')\frac{\lambda_{min}(E)\,\delta}{\lambda_{max}(E)\,d_{\max}}\Big\}, \tag{4.24}$$

where

$$d_{\max} := \max\Big\{d_{\text{out,t}}(k)\;:\;k \in \mathcal{I}, 1 \le t \le T\Big\}. \tag{4.25}$$

Then, for any choice

$$\sigma \in \Big[\frac{\tilde{\delta}}{\lambda_{min}(E)\delta}, \frac{1-\tilde{\delta}}{\lambda_{max}(E) d_{\max}}\Big], \tag{4.26}$$

the dynamics (4.18) *over $\{\mathcal{G}_s\}_{s\ge1}$ is input-to-state stable with respect to the nullspace of the matrix $\hat{\mathbf{L}}_{\mathcal{K}}$. Specifically, for any $t \in \mathbb{Z}_{\ge1}$ and any $\{\boldsymbol{d}_s\}_{s=1}^{t-1} \subset ((\mathbb{R}^d)^N)^K$,*

$$\|\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{v}_t\|_2 \le C_{\mathcal{I}}\|\boldsymbol{v}_1\|_2\Big(1-\frac{\tilde{\delta}}{4N^2}\Big)^{\lceil\frac{t-1}{B}\rceil} + C_{\mathcal{U}} \max_{1\le s\le t-1}\|\boldsymbol{d}_s\|_2, \tag{4.27}$$

where

$$C_{\mathcal{I}} := \kappa(S_E)\sqrt{K}\,(\tfrac{4}{3})^2, \quad C_{\mathcal{U}} := \frac{C_{\mathcal{I}}}{1-\Big(1-\frac{\tilde{\delta}}{4N^2}\Big)^{1/B}}. \tag{4.28}$$

And the cumulative disagreement satisfies, for $T \in \mathbb{Z}_{\ge1}$,

$$\sum_{t=1}^{T}\|\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{v}_t\|_2 \le C_{\mathcal{U}}\Big(\|\boldsymbol{v}_1\|_2 + \sum_{t=1}^{T-1}\|\boldsymbol{d}_t\|_2\Big). \tag{4.29}$$

Proof. The strategy of the proof is the following. We use Lemma 3.16 to decouple (4.18) into K copies (for each eigenvalue of E) of the same first-order linear consensus dynamics. We then analyze the convergence properties of the latter using [NO10b, Th. 1.2]. Finally, we bound the disagreement in the original network variables using again Lemma 3.16.

We start by noting that the selection of $\tilde{\delta}$ makes the set in (5.18) nonempty and consequently the selection of σ feasible. We write the dynamics (4.20), omitting the dependence on $l \in \{1,\dots,K\}$ for the sake of clarity, as

$$\boldsymbol{y}_{t+1} = (\mathrm{I}_{Nd} - \hat{\sigma}\,\mathbf{L}_t)\,\boldsymbol{y}_t + e_t, \tag{4.30}$$

where $\hat{\sigma} := \sigma\,\lambda_l(E) > 0$ and $e_t := \boldsymbol{e}_t^l$. From (4.21), we have

$$\|e_t\|_2 \le \|(S_E^{-1} \otimes \mathrm{I}_{Nd})\boldsymbol{d}_t\|_2 \le \|S_E^{-1}\|_2 \|\boldsymbol{d}_t\|_2, \tag{4.31}$$

for each $t \in \mathbb{Z}_{\ge 1}$. Next, let

$$\mathsf{P}_t := \mathrm{I}_N - \hat{\sigma}\,\mathsf{L}_t = \hat{\sigma}\mathsf{A}_t + \mathrm{I}_N - \hat{\sigma}\,\mathsf{D}_{\mathrm{out}t}, \tag{4.32}$$

and define $\boldsymbol{\Phi}(k,s) := \Big(\mathsf{P}_k \mathsf{P}_{k-1} \cdots \mathsf{P}_{s+1} \mathsf{P}_s\Big) \otimes \mathrm{I}_d$, for each $k \ge s \ge 1$. The trajectory of (4.30) can then be expressed as

$$\boldsymbol{y}_{t+1} = \boldsymbol{\Phi}(t,1)\boldsymbol{y}_1 + \sum_{s=1}^{t-1} \boldsymbol{\Phi}(t,s+1)e_s + e_t,$$

for $t \ge 2$. If we now multiply this equation by $\mathbf{L}_{\mathcal{K}}$, take norms on each side, and use the triangular inequality, we obtain

$$\begin{aligned}\|\mathbf{L}_{\mathcal{K}}\boldsymbol{y}_{t+1}\|_2 \leq &\|\mathbf{L}_{\mathcal{K}}\mathbf{\Phi}(t,1)\boldsymbol{y}_1\|_2 \\ &+\sum_{s=1}^{t-1}\|\mathbf{L}_{\mathcal{K}}\mathbf{\Phi}(t,s+1)e_s\|_2+\|\mathbf{L}_{\mathcal{K}}e_t\|_2 \\ =&\sqrt{V(\mathbf{\Phi}(t,1)\boldsymbol{y}_1)}+\sum_{s=1}^{t-1}\sqrt{V(\mathbf{\Phi}(t,s+1)e_s)}+\sqrt{V(e_t)},\end{aligned} \tag{4.33}$$

where $V:(\mathbb{R}^d)^N \to \mathbb{R}$ is defined by

$$V(\boldsymbol{y}) := \|\mathbf{L}_{\mathcal{K}}\boldsymbol{y}\|_2^2 = \sum_{i=1}^{N}\|\boldsymbol{y}^i-(\mathbf{M}\boldsymbol{y})^i\|_2^2.$$

Our next step is to verify the hypotheses of [NO10b, Th. 1.2] to conclude from [NO10b, (1.23)] that, for every $\boldsymbol{y}\in(\mathbb{R}^d)^N$ and every $k\geq s\geq 1$, the following holds,

$$V(\mathbf{\Phi}(k,s)\boldsymbol{y}) \leq \Big(1-\frac{\tilde{\delta}}{2N^2}\Big)^{\lceil\frac{k-s+1}{B}\rceil-2}V(\boldsymbol{y}). \tag{4.34}$$

Consider the matrices $\{\mathsf{P}_t\}_{t\geq 1}$ defined in (4.32). Since the digraphs are weight-balanced, i.e., $\mathbb{1}_N^\top\mathsf{L}_t=0$, we have $\mathbb{1}_N^\top\mathsf{P}_t=\mathbb{1}_N^\top$, and since $\mathsf{L}_t\mathbb{1}_N=0$, it follows that $\mathsf{P}_t\mathbb{1}_N=\mathbb{1}_N$. Moreover, according to (4.25) and (5.18), for each $t\in\mathbb{Z}_{\geq 1}$,

$$\begin{aligned}(\mathsf{P}_t)_{ii} &\geq 1-\hat{\sigma}\,d_{\text{out,t}}(i)=1-\sigma\lambda_l(E)\,d_{\text{out,t}}(i)\\ &\geq 1-\sigma\,\lambda_{\max}(E)d_{\max}\geq\tilde{\delta},\end{aligned}$$

for every $i\in\mathcal{I}$. On the other hand, for $i\neq j$, $(\mathsf{P}_t)_{ij}=\hat{\sigma}\mathsf{a}_{ij,t}\geq 0$ and therefore, if $\mathsf{a}_{ij,t}>0$, then the nondegeneracy of the adjacency matrices together with (5.18) implies that

$$(\mathsf{P}_t)_{ij} = \sigma\lambda_l(E)\mathsf{a}_{ij,t} \geq \sigma\,\lambda_{\min}(E)\delta \geq \tilde{\delta}.$$

Summarizing, the matrices in the sequence $\{\mathsf{P}_t\}_{t\geq 1}$ are doubly stochastic with entries uniformly bounded away from 0 by $\tilde{\delta}$ whenever positive. These are the sufficient conditions in [NO10b, Th. 1.2], along with B-joint connectivity, to guarantee that (4.34) holds. Plugging (4.34) into (4.33), and noting that

$$\rho_{\tilde{\delta}} := 1 - \frac{\tilde{\delta}}{4N^2} \geq \sqrt{1 - \tfrac{\tilde{\delta}}{2N^2}},$$

because $(1-x/2)^2 \geq 1-x$ for any $x \in [0,1]$, we get

$$\|\mathbf{L}_{\mathcal{K}}\boldsymbol{y}_{t+1}\|_2 \leq \rho_{\tilde{\delta}}^{\lceil\frac{t}{B}\rceil-2}\|\boldsymbol{y}_1\|_2 + \sum_{s=1}^{t}\rho_{\tilde{\delta}}^{\lceil\frac{t-s}{B}\rceil-2}\|e_s\|_2. \tag{4.35}$$

Here we have used that $\sqrt{V(\boldsymbol{y})} \leq \|\mathbf{L}_{\mathcal{K}}\|_2\|\boldsymbol{y}\|_2 \leq \|\boldsymbol{y}\|_2$ because $\|\mathbf{L}_{\mathcal{K}}\|_2 = 1$ (as $\hat{\mathbf{L}}_{\mathcal{K}}$ is symmetric and all its nonzero eigenvalues are equal to 1). We now proceed to bound $\|\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{v}_t\|_2$ in terms of $\boldsymbol{v}_1$ and the inputs $\{\boldsymbol{d}_t\}_{t\geq 1}$ of the original dynamics (4.18). To do this, we rely on Lemma 3.16. In fact, from (4.22), and using (4.35) for each of the K first-order consensus algorithms, we obtain

$$\|\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{v}_t\|_2 \leq \|S_E\|_2\sqrt{K}\max_{1\leq l\leq K}\Big\{\rho_{\tilde{\delta}}^{\lceil\frac{t-1}{B}\rceil-2}\|\boldsymbol{w}_1^l\|_2 + \sum_{s=1}^{t-1}\rho_{\tilde{\delta}}^{\lceil\frac{t-1-s}{B}\rceil-2}\|e_s\|_2\Big\}.$$

Recalling now (4.19), so that $\|\boldsymbol{w}_1^l\|_2 \leq \|\boldsymbol{w}_1\|_2 \leq \|S_E^{-1}\|_2\|\boldsymbol{v}_1\|_2$ for each $l \in \{1,\dots,K\}$, and using also (4.31), we obtain

$$\begin{aligned}\|\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{v}_t\|_2 \leq & \|S_E\|_2\sqrt{K}\rho_{\tilde{\delta}}^{-2}\Big(\rho_{\tilde{\delta}}^{\lceil\frac{t-1}{B}\rceil}\|S_E^{-1}\|_2\|\boldsymbol{v}_1\|_2 \\ & + \sum_{s=1}^{t-1}\rho_{\tilde{\delta}}^{\lceil\frac{t-1-s}{B}\rceil}\|S_E^{-1}\|_2\|\boldsymbol{d}_s\|_2\Big),\end{aligned} \tag{4.36}$$

for all $t \geq 2$ (and for $t = 1$ the inequality holds trivially). Equation (4.27) follows from (4.36) noting two facts. First, $\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$ for any $r \in (0,1)$ and in particular for $r = \rho_{\tilde{\delta}}^{1/B}$. Second, since $\tilde{\delta} \in (0,1)$, we have

$$\rho_{\tilde{\delta}}^{-1} = \tfrac{1}{1-\tilde{\delta}/(4N^2)} \leq \tfrac{1}{1-1/(4N^2)} = \tfrac{4N^2}{4N^2-1} \leq \tfrac{4}{3}.$$

To obtain (4.29), we sum (4.36) over the time horizon T to get

$$\sum_{t=1}^{T}\|\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{v}_t\|_2 \leq \kappa(S_E)\sqrt{K}\rho_{\tilde{\delta}}^{-2}\Big(\frac{1}{1-\rho_{\tilde{\delta}}^{1/B}}\|\boldsymbol{v}_1\|_2 + \sum_{t=2}^{T}\sum_{s=1}^{t-1}\rho_{\tilde{\delta}}^{\lceil\frac{t-1-s}{B}\rceil}\|\boldsymbol{d}_s\|_2\Big),$$

and using $r = \rho_{\tilde{\delta}}^{1/B}$ for brevity, the last sum is bounded as

$$\begin{aligned}\sum_{t=2}^{T}\sum_{s=1}^{t-1} r^{t-1-s}\|\boldsymbol{d}_s\|_2 &= \sum_{s=1}^{T-1}\sum_{t=s+1}^{T} r^{t-1-s}\|\boldsymbol{d}_s\|_2 \\ &= \sum_{s=1}^{T-1}\|\boldsymbol{d}_s\|_2\sum_{t=s+1}^{T} r^{t-1-s} \leq \frac{1}{1-r}\sum_{s=1}^{T-1}\|\boldsymbol{d}_s\|_2.\end{aligned}$$

This yields (4.29) and the proof is complete. □

The combination of the bound on the cumulative disagreement stated in Proposition 3.17 with the bound on agent regret that follows from Lemmas 3.14 and 3.15 leads us to the next result.

Corollary 3.18. (Bound on agent regret under the dynamics (4.6) for arbitrary learning rates). *For $T \in \mathbb{Z}_{\geq 1}$, let $\{f_t^1, \dots, f_t^N\}_{t=1}^T$ be convex functions on $\mathbb{R}^d$ with*

H-bounded subgradient sets. Let $E \in \mathbb{R}^{K\times K}$ be a diagonalizable matrix with real positive eigenvalues and $\{\mathcal{G}_t\}_{t\geq 1}$ a sequence of B-jointly connected, δ-nondegenerate, weight-balanced digraphs. If σ is chosen according to (5.18)*, then the agent regret associated to a sequence $\{\boldsymbol{x}_t = (x_t^1,\dots,x_t^N)\}_{t=1}^T$ generated by the coordination algorithm* (4.6) *satisfies, for any $j \in \{1,\dots,N\}$, $u \in \mathbb{R}^d$, and $\{\eta_t\}_{t=1}^T \subset \mathbb{R}_{>0}$, the bound*

$$\begin{aligned}2\mathcal{R}^j(u,\{f_t\}_{t=1}^T) \leq N\sum_{t=2}^T \|\tfrac{1}{N}\sum_{i=1}^N x_t^i - u\|_2^2\Big(\tfrac{1}{\eta_t} - \tfrac{1}{\eta_{t-1}} - p_t(\boldsymbol{u},\boldsymbol{x}_t)\Big)\\ + 4NHC_{\mathcal{U}}\|\boldsymbol{v}_1\|_2 + NH^2(4\sqrt{N}C_{\mathcal{U}}+1)\sum_{t=1}^T \eta_t\\ + \frac{N}{\eta_1}\|\tfrac{1}{N}\sum_{i=1}^N x_1^i - u\|_2^2,\end{aligned} \tag{4.37}$$

where $C_{\mathcal{U}}$ is given in (4.28) *and $p_t : (\mathbb{R}^d)^N \times (\mathbb{R}^d)^N \to \mathbb{R}_{\geq 0}$ is the modulus of strong convexity of $\tilde{f}_t$.*

Proof. From Lemmas 3.14 and 3.15, we can write

$$\begin{aligned}2\mathcal{R}^j(u,\{f_t\}_{t=1}^T) \leq \sum_{t=2}^T \|\mathbf{M}\boldsymbol{x}_t - \boldsymbol{u}\|_2^2\Big(\tfrac{1}{\eta_t} - \tfrac{1}{\eta_{t-1}} - p_t(\boldsymbol{u},\boldsymbol{x}_t)\Big)\\ + \Big(2\sqrt{N}H + 2NH\Big)\sum_{t=1}^T \|\mathbf{L}_{\mathcal{K}}\boldsymbol{x}_t\|_2 + NH^2\sum_{t=1}^T \eta_t\\ + \tfrac{1}{\eta_1}\|\mathbf{M}\boldsymbol{x}_1 - \boldsymbol{u}\|_2^2.\end{aligned} \tag{4.38}$$

On the one hand, note that

$$\|\mathbf{M}\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 = N\|\tfrac{1}{N}\sum_{i=1}^N x_t^i - u\|_2^2. \tag{4.39}$$

On the other hand, taking $\boldsymbol{d}_s = -\eta_s(\tilde{g}_{\boldsymbol{x}_s},0,\dots,0) \in ((\mathbb{R}^d)^N)^K$ in (4.29) of Proposition 3.17 and noting that $\|\boldsymbol{d}_s\|_2 = \eta_s \|\tilde{g}_{\boldsymbol{x}_s}\|_2 \leq \eta_s \sqrt{N}H$, we get

$$\sum_{t=1}^{T} \|\mathbf{L}_{\mathcal{K}} \boldsymbol{x}_t\|_2 \leq \sum_{t=1}^{T} \|\hat{\mathbf{L}}_{\mathcal{K}} \boldsymbol{v}_t\|_2 \leq C_{\mathcal{U}} \Big(\|\boldsymbol{v}_1\|_2 + \sum_{t=1}^{T-1} \eta_t \sqrt{N}H \Big).$$

We obtain the result by substituting this and (4.39) into (4.38), and using the bound $2\sqrt{N}H + 2NH \leq 4NH$. □

To establish the desired logarithmic and square-root regret bounds we need a suitable selection of learning rates in the bound obtained in Corollary 3.18. This step is enabled by the final ingredient in our analysis: bounding the evolution of the online estimates and all the auxiliary states uniformly in the time horizon T. We tackle this next.

4.3.3 Uniform bound on the trajectories

Here we show that the trajectories of (4.6) are bounded uniformly in the time horizon. For this, we first bound the mean of the online estimates and then use the ISS property of the disagreement evolution studied in the previous section.

Our first result establishes a useful bound on how far from the origin one should be so that a certain important inclusion among convex cones is satisfied. This plays a key role in the technical developments of this section.

Lemma 3.19. (Convex cone inclusion). *Given $\beta \in (0,1]$, $\epsilon \in (0,\beta)$, and any scalars $C_{\mathcal{X}}, C_{\mathcal{IU}} \in \mathbb{R}_{>0}$, let*

$$\hat{r}_\beta := \frac{C_{\mathcal{X}} + C_{\mathcal{IU}}}{\beta\sqrt{1-\epsilon^2} - \epsilon\sqrt{1-\beta^2}}. \tag{4.40}$$

Then, $\hat{r}_\beta \in (C_\mathcal{X} + C_{\mathcal{IU}}, \infty)$ *and, for any* $x \in \mathbb{R}^d \setminus \mathcal{B}(0, \hat{r}_\beta)$*,*

$$\bigcup_{w \in \bar{\mathcal{B}}(-x, C_\mathcal{X} + C_{\mathcal{IU}})} \mathcal{F}_\beta(w) \subseteq \mathcal{F}_\epsilon(-x), \tag{4.41}$$

where the set on the left is convex.

Proof. Throughout the proof, we consider the functions arccos and arcsin in the domain $[0,1]$. Since $\epsilon \in (0, \beta)$ and $\beta \in (0,1]$, it follows that $\arccos(\epsilon) - \arccos(\beta) \in (0, \pi/2)$. Now, using the angle-difference formula and noting that $\sin(\arccos(\alpha)) = \sqrt{1-\alpha^2}$ for any $\alpha \in [0,1]$, we have

$$\sin\Big(\arccos(\epsilon) - \arccos(\beta)\Big) = (\sqrt{1-\epsilon^2})\beta - \epsilon\sqrt{1-\beta^2},$$

which belongs to the set $(0,1)$ by the observation above. Therefore, $\hat{r}_\beta \in (C_\mathcal{X} + C_{\mathcal{IU}}, \infty)$. Let $x \in \mathbb{R}^d \setminus \mathcal{B}(0, \hat{r}_\beta)$. Since $\|x\|_2 \geq \hat{r}_\beta > C_\mathcal{X} + C_{\mathcal{IU}}$, then $\bar{\mathcal{B}}(-x, C_\mathcal{X} + C_{\mathcal{IU}}) \subseteq \mathbb{R}^d \setminus \{0\}$, and the intersection of $\bigcup_{w \in \bar{\mathcal{B}}(-x, C_\mathcal{X} + C_{\mathcal{IU}})} \mathcal{F}_\beta(w)$ with any plane passing through the origin and $-x$ forms a two-dimensional cone (cf. Figure 4.1) with angle

$$2 \arcsin\Big(\tfrac{C_\mathcal{X} + C_{\mathcal{IU}}}{\|x\|_2}\Big) + 2\arccos(\beta). \tag{4.42}$$

In the case of the intersection of $\mathcal{F}_\epsilon(-x)$ with any plane passing through the origin and $-x$, the angle is $2\arccos(\epsilon)$ (which is less than π because $\epsilon < \beta \leq 1$). Now, given the axial symmetry of both cones with respect to the line passing through the origin and $-x$, (4.41) is satisfied if and only if

$$\arcsin\Big(\tfrac{C_\mathcal{X} + C_{\mathcal{IU}}}{\|x\|_2}\Big) + \arccos(\beta) \leq \arccos(\epsilon), \tag{4.43}$$

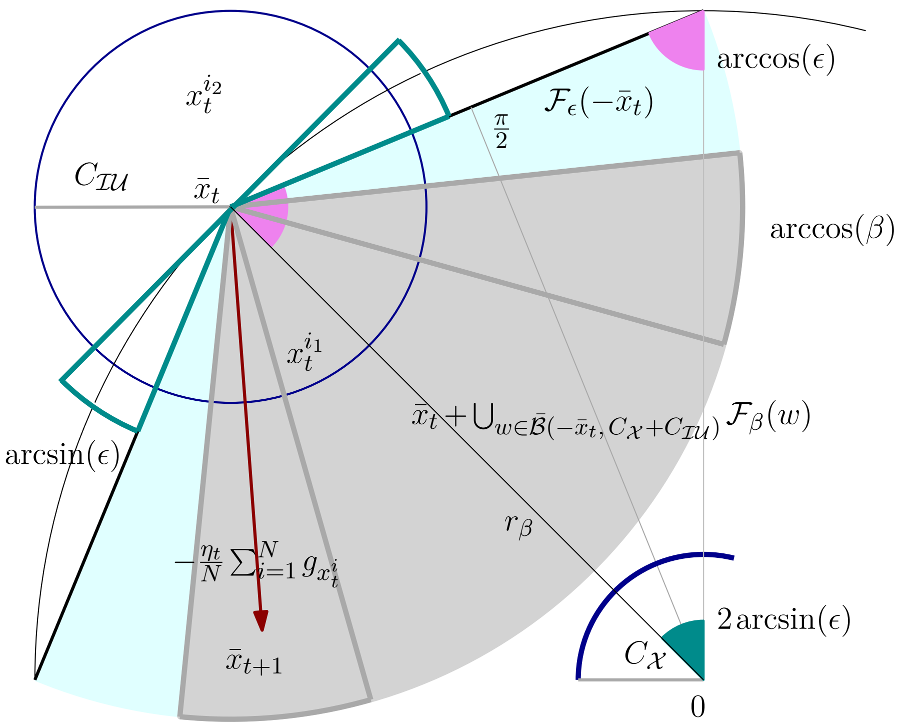

$\arccos(\epsilon)$
$\mathcal{F}_\epsilon(-\bar{x}_t)$
$\frac{\pi}{2}$
$x_t^{i_2}$
$C_{\mathcal{IU}}$
$\bar{x}_t$
$\arccos(\beta)$
$x_t^{i_1}$
$\bar{x}_t + \bigcup_{w \in \bar{\mathcal{B}}(-\bar{x}_t, C_{\mathcal{X}}+C_{\mathcal{IU}})} \mathcal{F}_\beta(w)$
$\arcsin(\epsilon)$
r_β
$-\frac{\eta_t}{N}\sum_{i=1}^N g_{x_t^i}$
$2\arcsin(\epsilon)$
$C_{\mathcal{X}}$
$\bar{x}_{t+1}$
0

Figure 4.1: Visual aid in two dimensions for the proof of Lemmas 3.19 and 3.20 (where the shaded cones are actually infinite).

as implied by $\|x\|_2 \geq \hat{r}_\beta$ because sin is increasing in $(0, \pi/2)$. On the other hand, the inclusion (4.41) also guarantees that $\bigcup_{w \in \bar{\mathcal{B}}(-x, C_{\mathcal{X}}+C_{\mathcal{IU}})} \mathcal{F}_\beta(w)$ is a convex cone because each $\mathcal{F}_\beta(w)$ is convex, the union is taken over elements in a convex set, and (4.43) implies that the angle in (4.42) is less than π. □

The following result bounds the mean of the online estimates for arbitrary learning rates uniformly in the time horizon.

Lemma 3.20. (Uniform bound on the mean of the online estimates). *For $T \in \mathbb{Z}_{\geq 1}$, let $\{f_t^1, \ldots, f_t^N\}_{t=1}^T$ be convex functions on $\mathbb{R}^d$ with H-bounded subgradient sets and nonempty sets of minimizers. Let $\cup_{t=1}^T \cup_{i=1}^N \operatorname{argmin}(f_t^i) \subseteq \bar{\mathcal{B}}(0, C_{\mathcal{X}})$ for some $C_{\mathcal{X}} \in \mathbb{R}_{>0}$ independent of T, and assume $\{f_t^1, \ldots, f_t^N\}_{t=1}^T$ are β-central on $\mathbb{R}^d \backslash \bar{\mathcal{B}}(0, C_{\mathcal{X}})$ for some $\beta \in (0, 1]$. Let $E \in \mathbb{R}^{K \times K}$ be a diagonalizable matrix with real* *positive eigenvalues and* $\{\mathcal{G}_s\}_{s\geq 1}$ *a sequence of* B*-jointly connected,* δ*-nondegenerate, weight-balanced digraphs. Let* σ *be chosen according to* (5.18) *and denote by* $\{\boldsymbol{x}_t = (x_t^1, \dots, x_t^N)\}_{t=1}^T$ *the sequence generated by the coordination algorithm* (4.6)*. For* $t \in \{1, \dots, T\}$*, let* $\bar{x}_t := \frac{1}{N}\sum_{i=1}^N x_t^i$ *denote the mean of the online estimates. Then, for any sequence of learning rates* $\{\eta_t\}_{t=1}^T \subset \mathbb{R}_{>0}$*,*

$$\|\bar{x}_t\|_2 \leq r_\beta + H \max_{s\geq 1} \eta_s, \tag{4.44}$$

where, for some $\epsilon \in (0, \beta)$*,*

$$r_\beta := \max\Big\{\frac{C_{\mathcal{X}} + C_{\mathcal{IU}}}{\beta\sqrt{1-\epsilon^2} - \epsilon\sqrt{1-\beta^2}}, \frac{H}{2\epsilon}\max_{s\geq 1}\eta_s\Big\} \tag{4.45}$$

(which is well defined as shown in Lemma 3.19), and

$$C_{\mathcal{IU}} := C_{\mathcal{I}}\|\boldsymbol{v}_1\|_2 + C_{\mathcal{U}}\sqrt{N}H\max_{s\geq 1}\eta_s, \tag{4.46}$$

where $C_{\mathcal{U}}$ *and* $C_{\mathcal{I}}$ *are given in* (4.28)*.*

Proof. To guide the reasoning, Figure 4.1 depicts some of the elements of the proof and intends to be a visual aid. The dynamics of the mean of the agents' estimates is described by (4.11), which in fact corresponds to N copies of

$$\bar{x}_{t+1} = \bar{x}_t - \eta_t \tfrac{1}{N}\sum_{i=1}^N g_{x_t^i}, \tag{4.47}$$

where $g_{x_t^i} \in \partial f_t^i(x_t^i)$. Our proof strategy is based on showing that, for any $t \in \{1, \dots, T\}$, if $\bar{x}_t$ belongs to the set

$$\mathbb{R}^d \setminus \mathcal{B}(0, r_\beta), \tag{4.48}$$

then $\|\bar{x}_{t+1}\|_2 \leq \|\bar{x}_t\|_2$. To establish this fact, we study both the direction and the magnitude of the increment $-\frac{\eta_t}{N}\sum_{i=1}^N g_{x_t^i}$ in (4.47). Since the subgradients in the latter expression are not evaluated at the mean, but at the agents' estimates, we first show that the agents' estimates are sufficiently close to the mean. According to the input-to-state stability property (4.27) from Proposition 3.17 with the choice $\boldsymbol{d}_s = -\eta_s(\tilde{g}_{\boldsymbol{x}_s}, 0, \ldots, 0) \in ((\mathbb{R}^d)^N)^K$, so that $\|\boldsymbol{d}_s\|_2 = \eta_s\|\tilde{g}_{\boldsymbol{x}_s}\|_2 \leq \eta_s\sqrt{N}H$, we get

$$\begin{aligned}\|\mathbf{L}_{\mathcal{K}}\boldsymbol{x}_t\|_2 \leq \|\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{v}_t\|_2 \leq & C_{\mathcal{I}}\|\boldsymbol{v}_1\|_2\Big(1-\frac{\tilde{\delta}}{4N^2}\Big)^{\lceil\frac{t-1}{B}\rceil} \\ & + C_{\mathcal{U}}\sqrt{N}H\max_{1\leq s\leq t-1}\eta_s \leq C_{\mathcal{IU}}, \qquad (4.49)\end{aligned}$$

where $C_{\mathcal{IU}}$ is defined in (4.46). Hence,

$$\begin{aligned}\max_i \|x_t^i - \bar{x}_t\|_2 &\leq \Big(\sum_{i=1}^n \|x_t^i - \bar{x}_t\|_2^2\Big)^{1/2} \\ &= \|\mathbf{L}_{\mathcal{K}}\boldsymbol{x}_t\|_2 \leq C_{\mathcal{IU}}. \qquad (4.50)\end{aligned}$$

This allows to exploit the starting assumption that $\bar{x}_t$ belongs to the set (4.48) when we study the increment $-\frac{\eta_t}{N}\sum_{i=1}^N g_{x_t^i}$.

Regarding the direction of the increment $-\frac{\eta_t}{N}\sum_{i=1}^N g_{x_t^i}$, the β-centrality of the function f_t^i for each $i \in \{1,\ldots,N\}$ and $t \in \{1,\ldots,T\}$ on $\mathbb{R}^d \setminus \bar{\mathcal{B}}(0, C_{\mathcal{X}})$ implies that, for any $z \in \mathbb{R}^d \setminus \bar{\mathcal{B}}(0, C_{\mathcal{X}})$, we have

$$-\partial f_t^i(z) \subseteq \bigcup_{y\in\operatorname{argmin}(f_t^i)} \mathcal{F}_\beta(y-z) \subseteq \bigcup_{y\in\bar{\mathcal{B}}(0,C_{\mathcal{X}})} \mathcal{F}_\beta(y-z), \qquad (4.51)$$

where the last inclusion follows from the hypothesis that $\cup_{t=1}^T \cup_{i=1}^N \operatorname{argmin}(f_t^i) \subseteq$

$\bar{\mathcal{B}}(0, C_{\mathcal{X}})$. Now, using the change of variables $w := y - z$, we have

$$\bigcup_{\substack{y \in \bar{\mathcal{B}}(0, C_{\mathcal{X}}) \\ z \in \bar{\mathcal{B}}(x, C_{\mathcal{IU}})}} \mathcal{F}_\beta(y - z) = \bigcup_{w \in \bar{\mathcal{B}}(-x, C_{\mathcal{X}} + C_{\mathcal{IU}})} \mathcal{F}_\beta(w). \tag{4.52}$$

The representation on the right shows that the set is convex whenever x belongs to the set in (4.48) thanks to Lemma 3.19 (essentially because $\mathcal{F}_\beta(w)$ is convex, the union is taken over elements in a convex set, and the intersection with any plane passing through $-x$ and the origin is a two-dimensional cone with angle less than π). Hence, taking the union when $z \in \bar{\mathcal{B}}(x, C_{\mathcal{IU}})$ on both sides of (4.51) and using (4.52), we obtain

$$\begin{aligned} \operatorname{conv}\Big(\bigcup_{z \in \bar{\mathcal{B}}(x, C_{\mathcal{IU}})} -\partial f_t^i(z) \Big) &\subseteq \bigcup_{w \in \bar{\mathcal{B}}(-x, C_{\mathcal{X}} + C_{\mathcal{IU}})} \mathcal{F}_\beta(w) \\ &\subseteq \mathcal{F}_\epsilon(-x), \end{aligned}$$

where the last inclusion holds for any x in the set (4.48) by Lemma 3.19 (noting from (4.40) that $r_\beta \geq \hat{r}_\beta$). Taking now $x = \bar{x}_t$ and noting that $x_t^i \in \bar{\mathcal{B}}(\bar{x}_t, C_{\mathcal{IU}})$ by (4.50), we deduce

$$-\frac{1}{N} \sum_{i=1}^N g_{x_t^i} \in \operatorname{conv}\Big(\bigcup_{z \in \bar{\mathcal{B}}(\bar{x}_t, C_{\mathcal{IU}})} -\partial f_t^i(z) \Big) \subseteq \mathcal{F}_\epsilon(-\bar{x}_t).$$

This guarantees that $\bar{x}_{t+1} = \bar{x}_t - \frac{\eta_t}{N} \sum_{i=1}^N g_{x_t^i}$ is contained in a convex cone with vertex at $\bar{x}_t$ and strictly contained in the semi-space tangent to the ball $\bar{\mathcal{B}}(0, \|\bar{x}_t\|_2)$ at $\bar{x}_t$ (with a tolerance-angle between them of $\arcsin(\epsilon)$).

Regarding the magnitude $\| -\frac{\eta_t}{N} \sum_{i=1}^N g_{x_t^i} \|_2 \leq H\eta_t$, we need to show, based on the starting assumption that $\bar{x}_t$ belongs to the set (4.48), that $H \max_{s \geq 1} \eta_s$ is no larger than the chords of angle $\arccos(\epsilon)$ with respect to the radii of $\bar{\mathcal{B}}(0, r_\beta)$.

Now, any such chord defines an isosceles triangle in the plane containing the chord and the segment joining the origin and $\bar{x}_t$. Since the angle subtended by the chord at the origin is $2\arcsin(\epsilon)$, then the length of the chord is $2r_\beta\epsilon$. Therefore, since $r_\beta \geq \frac{H}{2\epsilon}\max_{s\geq 1}\eta_s$ by the hypothesis (4.45), we conclude that the length of the chord is larger or equal than $H\max_{s\geq 1}\eta_s$. This guarantees that $\bar{x}_{t+1} = \bar{x}_t - \frac{\eta_t}{N}\sum_{i=1}^N g_{x_t^i}$ is in the ball $\bar{\mathcal{B}}(0, \|\bar{x}_t\|_2)$.

The above argument guarantees that, if the starting assumption that $\bar{x}_t$ belongs to the set (4.48) holds, then $\|\bar{x}_{t+1}\|_2 \leq \|\bar{x}_t\|_2$. However, if the starting assumption is not true, then the previous inequality might not hold. Since the magnitude of the increment in an arbitrary direction is upper bounded by $H\max_{s\geq 1}\eta_s$, adding this value to the threshold r_β in the definition of (4.48) yields the desired bound (4.44) for $\{\bar{x}_t\}_{t=1}^T$, uniformly in T. □

The next result bounds the online estimates for arbitrary learning rates in terms of the initial conditions and the uniform bound on the sets of local minimizers. The fact that the bound includes the auxiliary states follows from the ISS property and the invariance of the mean of the auxiliary states.

Proposition 3.21. (Boundedness of the online estimates and the auxiliary states). *Under the hypotheses of Lemma 3.20, the trajectories of the coordination algorithm* (4.6) *are uniformly bounded in the time horizon T, for any $\{\eta_t\}_{t=1}^T \subset \mathbb{R}_{>0}$, as*

$$\|\boldsymbol{v}_t\|_2 \leq C(\beta),$$

for $t \in \{1, \dots, T\}$, where

$$C(\beta) := \sqrt{N}\Big(r_\beta + H\max_{s\geq 1}\eta_s\Big) + \sqrt{K}\|\boldsymbol{v}_1\|_2 + C_{\mathcal{IU}}, \tag{4.53}$$

and where r_β *is given in* (4.45) *and* $C_{\mathcal{IU}}$ *in* (4.46)*.*

Proof. We start by noting the useful decomposition $\boldsymbol{v}_t = (\mathrm{I}_K \otimes \mathbf{M})\boldsymbol{v}_t + \hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{v}_t$. Using the triangular inequality, we obtain

$$\begin{aligned}\|\boldsymbol{v}_t\|_2 &\leq \|(\mathrm{I}_K \otimes \mathbf{M})\boldsymbol{v}_t\|_2 + \|\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{v}_t\|_2 \\ &\leq \|\mathbf{M}\boldsymbol{x}_t\|_2 + \sum_{l=2}^{K}\|\mathbf{M}\boldsymbol{v}_t^l\|_2 + \|\hat{\mathbf{L}}_{\mathcal{K}}\boldsymbol{v}_t\|_2.\end{aligned}$$

The first term can be upper bounded by noting that $\|\mathbf{M}\boldsymbol{x}_t\|_2 = \sqrt{N}\|\frac{1}{N}\sum_{i=1}^n x_t^i\|_2$ and invoking (4.44) in Lemma 3.20. The second term does not actually depend on t. To see this, we use the fact that $(\mathrm{I}_K \otimes \mathbf{M})(\mathrm{I}_{KNd} - \sigma\mathbb{L}_t) = \mathrm{I}_K \otimes \mathbf{M}$ in the dynamics (4.6) with the choice (4.7) to obtain the following invariance property of the mean of the auxiliary states,

$$\mathbf{M}\boldsymbol{v}_{t+1}^l = \mathbf{M}\boldsymbol{v}_t^l = \mathbf{M}\boldsymbol{v}_1^l$$

for $l \in \{2,\ldots,K\}$. Then, using the sub-multiplicativity of the norm and [Ber05, Fact 9.12.22] for the norms of Kronecker products in $\|\mathbf{M}\|_2 = \|\mathrm{M}\otimes\mathrm{I}_d\|_2 = \|\mathrm{M}\|_2\|\mathrm{I}_d\|_2 = 1$, we get

$$\sum_{l=2}^{K}\|\mathbf{M}\boldsymbol{v}_1^l\|_2 \leq \|\mathbf{M}\|_2\sum_{l=2}^{K}\|\boldsymbol{v}_1^l\|_2 \leq \sqrt{K}\|\boldsymbol{v}_1\|_2,$$

where the last inequality follows from the inequality of arithmetic and quadratic means [Bul03]. Finally, the third term is upper bounded in (4.49), and the result follows. □

The previous statements about uniform boundedness of the trajectories can also be concluded when the objectives are strongly convex. The following result says that local strong convexity and bounded subgradient sets imply β-centrality.

Lemma 3.22. (Local strong convexity and bounded subgradients implies centrality away from the minimizer). *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a convex function on $\mathbb{R}^d$ that is also γ-strongly convex on $\bar{\mathcal{B}}(y,\zeta)$, for some $\gamma,\ \zeta \in \mathbb{R}_{>0}$ and $y \in \mathbb{R}^d$. Then, for any $x \in \mathbb{R}^d \setminus \bar{\mathcal{B}}(y,\zeta)$ and $g_x \in \partial h(x)$, $g_y \in \partial h(y)$,*

$$\left(g_x - g_y\right)^\top (x-y) \geq \gamma\zeta \, \|x-y\|_2. \tag{4.54}$$

If in addition h has H-bounded subgradient sets and $0 \in \partial h(y)$, then h is $\frac{\gamma\zeta}{H}$-central in $\mathbb{R}^d \setminus \bar{\mathcal{B}}(y,\zeta)$. (Note that if $0 \in \partial h(y)$, then $\arg\min_{x\in\mathbb{R}^d} h(x) = \{y\}$ is a singleton by strong convexity in the ball $\bar{\mathcal{B}}(y,\zeta)$.)

Proof. Given any $y \in \mathbb{R}^d$ and $x \in \mathbb{R}^d \setminus \bar{\mathcal{B}}(y,\zeta)$, let $\tilde{x} \in \bar{\mathcal{B}}(y,\zeta)$ be any point in the line segment between x and y. Consequently, for some $\nu \in (0,1)$, we can write

$$\tilde{x} - y = \nu(x-y) = \frac{\nu}{1-\nu}(x-\tilde{x}). \tag{4.55}$$

Then, for any $g_x \in \partial h(x)$, $g_y \in \partial h(y)$, and $g_{\tilde{x}} \in \partial h(\tilde{x})$,

$$\begin{aligned}\left(g_x - g_y\right)^\top (x-y) &= \left(g_x - g_{\tilde{x}} + g_{\tilde{x}} - g_y\right)^\top (x-y)\\ &= \frac{1}{1-\nu}(g_x - g_{\tilde{x}})^\top (x-\tilde{x}) + \frac{1}{\nu}(g_{\tilde{x}} - g_y)^\top (\tilde{x}-y)\\ &\geq 0 + \frac{\gamma}{\nu}\|\tilde{x}-y\|_2^2 = \gamma\|\tilde{x}-y\|_2\|x-y\|_2,\end{aligned}$$

where in the inequality we have used convexity for the first term and strong convexity for the second term. To derive (4.54) we choose $\tilde{x}$ satisfying $\|\tilde{x}-y\|_2 = \zeta$, while the second part follows taking $g_y = 0$ in (4.54) and multiplying the right-hand side by $\frac{\|g_x\|_2}{H}$ because the latter quantity is less than 1. □

4.4 Logarithmic and square-root agent regret

In this section, we build on our technical results of the previous section: the general agent regret bound for arbitrary learning rates (cf. Corollary 3.18), and the uniform boundedness of the trajectories of the general dynamics (4.6) (cf. Proposition 3.21). Equipped with these results, we are ready to select the learning rates to deduce the agent regret bounds outlined in Section 4.2.

Our first main result establishes the logarithmic agent regret for the general dynamics (4.6) under harmonic learning rates.

Theorem 4.23. (Logarithmic agent regret for the dynamics (4.6)). *For $T \in \mathbb{Z}_{\geq 1}$, let $\{f_t^1, \dots, f_t^N\}_{t=1}^T$ be convex functions on $\mathbb{R}^d$ with H-bounded subgradient sets and nonempty sets of minimizers. Let $\cup_{t=1}^T \cup_{i=1}^N \operatorname{argmin}(f_t^i) \subseteq \bar{\mathcal{B}}(0, C_{\mathcal{X}}/2)$ for some $C_{\mathcal{X}} \in \mathbb{R}_{>0}$ independent of T, and assume $\{f_t^1, \dots, f_t^N\}_{t=1}^T$ are p-strongly convex on $\bar{\mathcal{B}}(0, C(\frac{pC_{\mathcal{X}}}{2H}))$, for some $p \in \mathbb{R}_{>0}$, where $C(\cdot)$ is defined in* (4.53). *Let $E \in \mathbb{R}^{K \times K}$ be a diagonalizable matrix with real positive eigenvalues and $\{\mathcal{G}_t\}_{t\geq 1}$ a sequence of B-jointly connected, δ-nondegenerate, weight-balanced digraphs. Let σ be chosen according to* (5.18) *and denote by $\{\boldsymbol{x}_t = (x_t^1, \dots, x_t^N)\}_{t=1}^T$ the sequence generated by the coordination algorithm* (4.6). *Then, taking $\eta_t = \frac{1}{\tilde{p}t}$, for any $\tilde{p} \in (0, p]$, the following regret bound holds for any $j \in \{1, \dots, N\}$ and $u \in \mathbb{R}^d$:*

$$
\begin{aligned}
2\mathcal{R}^j(u, \{f_t\}_{t=1}^T) \leq & \frac{NH^2\left(4\sqrt{N}C_{\mathcal{U}} + 1\right)}{\tilde{p}}(1 + \log T) \\
& + 4NHC_{\mathcal{U}}\|\boldsymbol{v}_1\|_2 + N\tilde{p}\,\|\tfrac{1}{N}\sum_{i=1}^N x_1^i - u\|_2^2, \qquad (4.56)
\end{aligned}
$$

where $C_{\mathcal{U}}$ is given by (4.28).

Proof. First we note that $C_{\mathcal{X}} < C(\frac{pC_{\mathcal{X}}}{2H})$ because r_β in (4.45) is a lower bound for the function $C(\cdot)$ in (4.53) and $C_{\mathcal{X}} < r_\beta$ as a consequence of Lemma 3.19.

Thus, the fact that each f_t^i is p-strongly convex on $\bar{\mathcal{B}}(0, C(\frac{pC_{\mathcal{X}}}{2H}))$ implies that it is also p-strongly convex on $\bar{\mathcal{B}}(0, C_{\mathcal{X}})$. Let x_t^{*i} denote the unique minimizer of f_t^i. Then, $\operatorname{argmin}(f_t^i) \subseteq \bar{\mathcal{B}}(0, C_{\mathcal{X}}/2)$ implies that $\bar{\mathcal{B}}(x_t^{*i}, C_{\mathcal{X}}/2) \subseteq \bar{\mathcal{B}}(0, C_{\mathcal{X}})$. The application of Lemma 3.22 with $\gamma = p$, $\zeta = C_{\mathcal{X}}/2$ and $y = x_t^{*i}$ implies then that each f_t^i is β'-central on $\mathbb{R}^d \setminus \bar{\mathcal{B}}(0, C_{\mathcal{X}})$ for any $\beta' \leq \frac{pC_{\mathcal{X}}/2}{H}$. Hence, the hypotheses of Proposition 3.21 are satisfied with $\beta = \frac{pC_{\mathcal{X}}}{2H}$ and therefore the estimates satisfy the bound $\|\boldsymbol{x}_t\|_2 \leq \|\boldsymbol{v}_t\|_2 \leq C(\frac{pC_{\mathcal{X}}}{2H})$ for $t \geq 1$, independent of T, which means they are confined to the region where the modulus of strong convexity of each f_t^i is p. Now, the modulus of strong convexity of $\tilde{f}_t$ is the same as for the functions $\{f_t^i\}_{i=1}^N$. That is, for each $\tilde{\xi}_{\boldsymbol{y}} = (\xi_{y^1}, \ldots, \xi_{y^N}) \in \partial \tilde{f}_t(\boldsymbol{y})$ and $\tilde{\xi}_{\boldsymbol{x}} = (\xi_{x^1}, \ldots, \xi_{x^N}) \in \partial \tilde{f}_t(\boldsymbol{x})$, for all $\boldsymbol{y}, \boldsymbol{x} \in \bar{\mathcal{B}}(0, C(\frac{pC_{\mathcal{X}}}{2H})) \subset (\mathbb{R}^d)^N$, one has

$$\begin{aligned}(\tilde{\xi}_{\boldsymbol{y}} - \tilde{\xi}_{\boldsymbol{x}})^\top (\boldsymbol{y} - \boldsymbol{x}) &= \sum_{i=1}^N (\xi_{y^i} - \xi_{x^i})^\top (y^i - x^i) \\ &\geq p \sum_{i=1}^N \|y^i - x^i\|_2^2 = p\, \|\boldsymbol{y} - \boldsymbol{x}\|_2^2.\end{aligned}$$

Thus, for all $\boldsymbol{y}$, $\boldsymbol{x} \in \bar{\mathcal{B}}(0, C(\frac{pC_{\mathcal{X}}}{2H}))$, we can take $p_t(\boldsymbol{y}, \boldsymbol{x}) = p$ in (4.37) and hence Corollary 3.18 implies the result by noting

$$\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - p_t(\boldsymbol{u}, \boldsymbol{x}_t) = \tilde{p}t - \tilde{p}(t-1) - p = \tilde{p} - p \leq 0,$$

so the first sum in (4.37) can be bounded by 0. Finally, $\sum_{t=1}^T \eta_t = \frac{1}{\tilde{p}} \sum_{t=1}^T \frac{1}{t} < \frac{1}{\tilde{p}}(1 + \log T)$. □

Our second main result establishes the square-root agent regret for the general dynamics (4.6). Its proof follows from Corollary 3.18, this time by using a bounding technique called the Doubling Trick [SS12, Sec. 2.3.1] in the learning rates selection.

Theorem 4.24. (Square-root agent regret). *For $T \in \mathbb{Z}_{\geq 1}$, let $\{f_t^1, \ldots, f_t^N\}_{t=1}^T$ be convex functions on $\mathbb{R}^d$ with H-bounded subgradient sets and nonempty sets of minimizers. Let $\cup_{t=1}^T \cup_{i=1}^N \operatorname{argmin}(f_t^i) \subseteq \bar{\mathcal{B}}(0, C_{\mathcal{X}})$ for some $C_{\mathcal{X}} \in \mathbb{R}_{>0}$ independent of T, and assume $\{f_t^1, \ldots, f_t^N\}_{t=1}^T$ are also β-central on $\mathbb{R}^d \setminus \bar{\mathcal{B}}(0, C_{\mathcal{X}})$ for some $\beta \in (0,1]$. Let $E \in \mathbb{R}^{K \times K}$ be a diagonalizable matrix with real positive eigenvalues and $\{\mathcal{G}_t\}_{t \geq 1}$ a sequence of B-jointly connected, δ-nondegenerate, weight-balanced digraphs. Let σ be chosen according to (5.18) and denote by $\{\boldsymbol{x}_t = (x_t^1, \ldots, x_t^N)\}_{t=1}^T$ the sequence generated by the coordination algorithm (4.6). Consider the following choice of learning rates called Doubling Trick scheme: for $m = 0, 1, 2, \ldots, \lceil \log_2 T \rceil$, we take $\eta_t = \frac{1}{\sqrt{2^m}}$ in each period of 2^m rounds $t = 2^m, \ldots, 2^{m+1} - 1$. Then, the following regret bound holds for any $j \in \{1, \ldots, N\}$ and $u \in \mathbb{R}^d$:*

$$2\mathcal{R}^j(u, \{f_t\}_{t=1}^T) \leq \frac{\sqrt{2}}{\sqrt{2}-1} \alpha \sqrt{T}, \tag{4.57}$$

where

$$\alpha := N^{3/2} H^2 C_{\mathcal{U}} C(\beta) \left(\frac{4}{\sqrt{N} H} + \frac{4}{C(\beta)} + \frac{1}{\sqrt{N} C_{\mathcal{U}} C(\beta)} \right) + N \Big(r_\beta + H + \|u\|_2 \Big)^2,$$

where $C_{\mathcal{U}}$ is given in (4.28) and $C(\cdot)$ is defined in (4.53).

Proof. We divide the proof in two steps. In step (i), we use the general agent regret bound of Corollary 3.18 making a choice of constant learning rates over a fixed known time horizon T'. In step (ii), we use multiple times this bound together with the Doubling Trick [SS12, Sec. 2.3.1] to produce an implementation procedure in which no knowledge of the time horizon is required. Regarding (i), the choice

$\eta_t = \eta'$ for all $t \in \{1,\dots,T'\}$ in (4.37) yields

$$\begin{aligned} 2\mathcal{R}^j(u,\{f_t\}_{t=1}^{T'}) \leq & 4NHC_{\mathcal{U}}\|\boldsymbol{v}_1\|_2 \\ & + NH^2\Big(4\sqrt{N}C_{\mathcal{U}}+1\Big)T'\eta' + \frac{N}{\eta'}\|\tfrac{1}{N}\sum_{i=1}^{N} x_1^i - u\|_2^2, \end{aligned} \tag{4.58}$$

where the first sum in (4.37) is upper-bounded by 0 because $\frac{1}{\eta'} - \frac{1}{\eta'} - p_t(\boldsymbol{u},\boldsymbol{x}_t) \leq 0$. Taking now $\eta' = 1/\sqrt{T'}$ in (4.58), factoring out $\sqrt{T'}$ and using $1 \leq \sqrt{T'}$, we obtain

$$\begin{aligned} 2\mathcal{R}^j(u,\{f_t\}_{t=1}^{T'}) \leq & \Big(4NHC_{\mathcal{U}}\|\boldsymbol{v}_1\|_2 \\ & + NH^2\Big(4\sqrt{N}C_{\mathcal{U}}+1\Big) + N\|\tfrac{1}{N}\sum_{i=1}^{N} x_1^i - u\|_2^2\Big)\sqrt{T'}. \end{aligned} \tag{4.59}$$

This bound is of the type $2\mathcal{R}^j(u,\{f_t\}_{t=1}^{T'}) \leq \alpha'\sqrt{T'}$, where α' depends on the initial conditions. This leads to step (ii). According to the Doubling Trick [SS12, Sec. 2.3.1], for $m = 0,1,\dots\lceil\log_2 T\rceil$, the dynamics is executed in each period of $T' = 2^m$ rounds $t = 2^m,\dots,2^{m+1}-1$, where at the beginning of each period the initial conditions are the final values in the previous period. The regret bound for each period is $\alpha'\sqrt{T'} = \alpha_m\sqrt{2^m}$, where α_m is the multiplicative constant in (4.59) that depends on the initial conditions in the corresponding period. To eliminate the dependence on the latter, by Proposition 3.21, we have that $\|\boldsymbol{v}_t\|_2 \leq C(\beta)$, for $C(\cdot)$ in (4.53) with $\max_{s\geq 1}\eta_s = 1$. Also, using (4.44), we have

$$\|\tfrac{1}{N}\sum_{i=1}^{N} x_t^i - u\|_2 \leq \|\bar{x}_t\|_2 + \|u\|_2 \leq r_\beta + H + \|u\|_2.$$

Since $C(\beta)$ only depends on the initial conditions at the beginning of the implementation procedure, the regret on each period is now bounded as $\alpha_m\sqrt{2^m} \leq \alpha\sqrt{2^m}$, for α in the statement. Consequently, the total regret can be bounded by

$$\sum_{m=0}^{\lceil \log_2 T \rceil} \alpha\sqrt{2^m} = \alpha \tfrac{1-\sqrt{2}^{\lceil \log_2 T \rceil + 1}}{1-\sqrt{2}} \leq \alpha \tfrac{1-\sqrt{2T}}{1-\sqrt{2}} \leq \tfrac{\sqrt{2}}{\sqrt{2}-1}\alpha\sqrt{T},$$

which yields the desired bound. □

Remark 4.25. (Asymptotic dependence of logarithmic agent regret bound on network properties). *Here we analyze the asymptotic dependence of the logarithmic regret bound in Theorem 4.23 on the number of agents. It is not difficult to see that, when $N \to \infty$, then*

$$\frac{C_{\mathcal{U}}}{C_{\mathcal{I}}} = \frac{1}{1-(1-\frac{\tilde{\delta}}{4N^2})^{1/B}} \sim \frac{4N^2 B}{\tilde{\delta}}.$$

Hence, for any B that guarantees B-joint connectivity, the asymptotic behavior as $N \to \infty$ of the bound (4.56) *scales as*

$$\frac{N^{3+1/2}B}{\tilde{\delta}} o(T), \tag{4.60}$$

where $\lim_{T\to\infty} \frac{o(T)}{T} = 0$. In contrast to (4.60)*, the asymptotic dependence on the number of agents in [YSVQ13, HCM13], which assume strong connectivity every time step and a doubly stochastic adjacency matrix A, is*

$$\frac{N^{1+1/2}}{1-\sigma_2(\mathsf{A})} o(T), \tag{4.61}$$

where $\sigma_2(\mathsf{A})$ is the second smallest singular value of A. (Here we are taking into account the fact that [HCM13] divides the regret by the number of agents.) The bounds (4.60) *and* (4.61) *are comparable in the case of sparse connected graphs that fail to be good expanders (i.e., for sparse graphs with low algebraic connectivity given* *by the second smallest eigenvalue of the Laplacian). This is the most reasonable comparison given our joint-connectivity assumption. For simplicity, we examine the case of undirected graphs because then* $1-\sigma_2(\mathsf{A}) = 1-\lambda_2(\mathsf{A}) = \lambda_2(\mathsf{L})$*, where* $\mathsf{L} = \mathrm{diag}(\mathsf{A}\mathbb{1}_N) - \mathsf{A} = \mathrm{I} - \mathsf{A}$ *is the Laplacian corresponding to* A*. The sparsity of the graph implies that* δ*,* $d_{\max} \approx 1$*, so that one can compute the maximum feasible* $\tilde{\delta}$ *from* (5.17) *to be* $\tilde{\delta}^* := (1+\frac{\lambda_{max}(E) d_{\max}}{\lambda_{min}(E)\delta})^{-1} \approx (1+\lambda_{max}(E)/\lambda_{min}(E))^{-1}$*. The algebraic connectivity* $\lambda_2(\mathsf{L})$ *can vary even for sparse graphs. Paths and cycles are two examples of graphs that fail to be expander graphs and their algebraic connectivity [Fie73] is* $2(1-\cos(\pi/N))$ *and* $2(1-\cos(2\pi/N))$*, respectively (for edge-weights equal to* 1*), and thus proportional to* $1-\cos(1/N) \sim \frac{1}{N^2}$ *when* $N \to \infty$*. With these values of* $\tilde{\delta}^*$ *and* $\lambda_2(\mathsf{L})$ *(up to a constant independent of* N*),* (4.60) *and* (4.61) *become*

$$N^{3+1/2} B\, o(T) \quad \textit{and} \quad N^{3+1/2} o(T),$$

respectively. Expression (4.60) *highlights the trade-offs between the degree of parallelization and the regret behavior for a given time horizon. Such trade-offs must be considered in the light of factors like the serial processor speed and the rate of data-collection as well as the cost and bandwidth limitations of transmitting spatially distributed data.* •

4.5 Simulation: application to medical diagnosis

In this section we illustrate the performance of the coordination algorithm (4.6) in a binary classification problem from medical diagnosis. We specifically consider the online gradient descent with proportional and with proportional-integral disagreement feedback. Inspired by [SWV$^+$01], we consider a clinical decision problem involving the use of Computerized Tomography (CT) for patients with minor head injury. We consider a network of hospitals that works cooperatively to develop a set of rules to determine whether a patient requires immediately a CT for possible neurological intervention, or if an alternative follow-up protocol should be applied to further inform the decision. The hospitals estimate local prediction models using the data collected from their patients while coordinating their efforts according to (4.6) to benefit from the model parameters updated by other hospitals.

We start by describing the data collected by the hospitals. Suppose that at round t, hospital i collects a vector $w_t^i \in \mathbb{R}^c$ encoding a set of features corresponding to patient data. In our case, $c = 10$ and the components of w_t^i correspond to factors or symptoms like "age", "amnesia before impact", "open skull fracture", "loss of consciousness", "vomiting", etc. The ultimate goal of each hospital is to decide if any acute brain finding would be revealed by the CT, and the true answer is denoted by $y_t^i \in \{-1, 1\}$, where $-1 =$ "no" and $1 =$ "yes" are the two possible classes. The true assessment is only found once the CT or the follow-up protocol have been used.

To cast this scenario in the networked online optimization framework described in Section 4.1, it is enough to specify the cost function $f_t^i : \mathbb{R}^d \to \mathbb{R}$ for each hospital $i \in \{1, \ldots, N\}$ and each round $t \in \{1, \ldots, T\}$. In this scenario, the cost function measures the fitness of the model parameters estimated by the hospital with respect to the data collected from its patients, as we explain next. Each hospital i seeks to estimate a vector of model parameters $x_t^i \in \mathbb{R}^d$, $d = c + 1$, that weigh the correspondence between the symptoms and the actual brain damage (up to an additional affine term). More precisely, hospital i employs a model h to assign the quantity $h(x_t^i, w_t^i)$, called decision or prediction, to the data point w_t^i using the estimated model parameters x_t^i. For instance, a linear predictor is based on the model $h(x_t^i, w) = x_t^{i\top}(w_t^i, 1)$, with the corresponding class predictor being $\mathrm{sign}(h(x_t^i, w_t^i))$. The loss incurred by hospital i is then $f_t^i(x_t^i) = l(x_t^i, w_t^i, y_t^i)$, where the loss function l is decreasing in the so-called margin $y_t^i h(x_t^i, w_t^i)$. This is because correct predictions (when the margin is positive) should be penalized less than incorrect predictions (when the margin is negative). Common loss functions are the logistic (smooth) function, $l(x, w, y) = \log\left(1 + e^{-2y\, h(x,w)}\right)$ or the hinge (nonsmooth) function, $l(x, w, y) = \max\{0, 1 - y\, h(x, w)\}$.

In the scenario just described, each hospital $i \in \{1, \ldots, N\}$ updates to x_{t+1}^i its estimated model parameters x_t^i according to the dynamics (4.6) as the data (w_t^i, y_t^i) becomes available. We simulate here two cases, the online gradient descent with proportional disagreement feedback, corresponding to $K = 1$ and $E = [1]$, and the online gradient descent with proportional-integral disagreement feedback, corresponding to

$$K = 2 \quad \text{and} \quad E = \begin{bmatrix} a & 1 \\ -1 & 0 \end{bmatrix};$$

cf. Remark 2.13. Both the online and distributed aspects of our approach are relevant for this kind of large-scale supervised learning. On one hand, data streams can be analyzed rapidly and with low storage to produce a real-time service using first-order information of the corresponding cost functions (for single data points or for mini-batches). On the other hand, hospitals can benefit from the prediction models updated by other hospitals. Under (4.6), each hospital i only shares the provisional vector of model parameters x_t^i with neighboring hospitals and maintains its patient data (w_t^i, y_t^i) private. In addition, the joint connectivity assumption is a flexible condition on how frequently hospitals communicate with each other. With regards to communication latency, note that the potential delays in communication among hospitals are small compared to the rate at which data is collected from patients. Also, the fitness of provisional local models can always be computed with respect to mini-batches of variable size when one hospital collects a different amount of data than others in the given time scales of coordination.

In our simulation, a network of 5 hospitals uses the time-varying communication topology shown in Figure 4.2. This results in the executions displayed in Figures 4.3 and 4.4, where provisional local models are shown to asymptotically agree and achieve sublinear regret with respect to the best model obtained in hindsight with all the data centrally available. For completeness, the plots also compare their performance against a centralized online gradient descent algorithm [Zin03, HAK07].

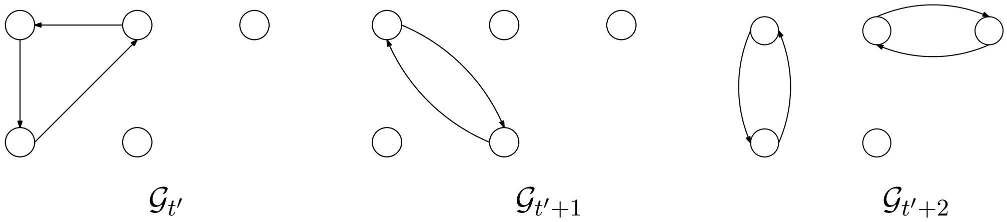


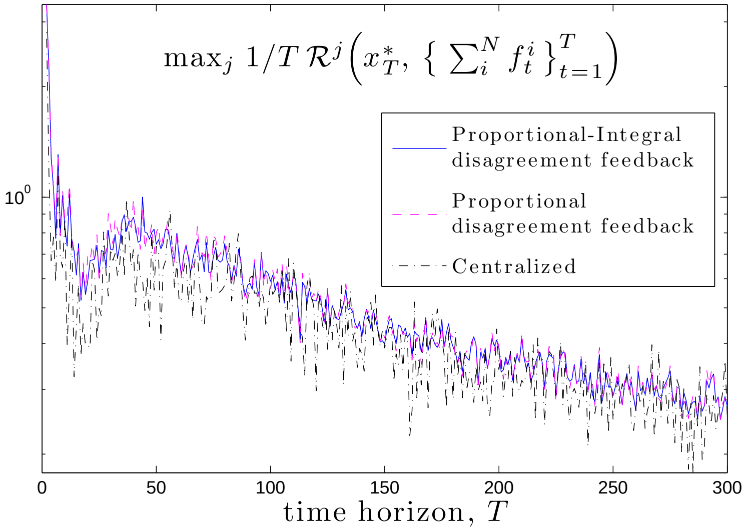
Figure 4.2: The communication topology in our simulation example corresponds to the periodic repetition of the displayed sequence of weight-balanced digraphs (where all nonzero edge weights are 1). The resulting sequence is 3-jointly connected, 1-nondegenerate, and the maximum out-degree is 1, i.e., $B = 3$, $\delta = 1$, and $d_{\max} = 1$.

4.6 Discussion

We have studied a networked online convex optimization scenario where each agent has access to partial information that is increasingly revealed over time in the form of a local cost function. The goal of the agents is to generate a sequence of decisions that achieves sublinear regret with respect to the best single decision in hindsight had all the information been centrally available. We have proposed a class of distributed coordination algorithms that allow agents to fuse their local decision parameters and incorporate the information of the local objectives as it becomes available. Our algorithm design uses first-order local information about the cost functions revealed in the previous round, in the form of subgradients, and only requires local communication of decision parameters among neighboring agents over a sequence of weight-balanced, jointly connected digraphs. We have shown that our distributed strategies achieve the same logarithmic and square-root agent regret bounds that centralized implementations enjoy. We have also characterized the dependence of the agent regret bounds on the network parameters. Our technical approach has built on an innovative combination of network and agent regret bounds, the cumulative disagreement of the collective estimates, and the boundedness of the sequence of collective estimates uniformly in the time horizon.

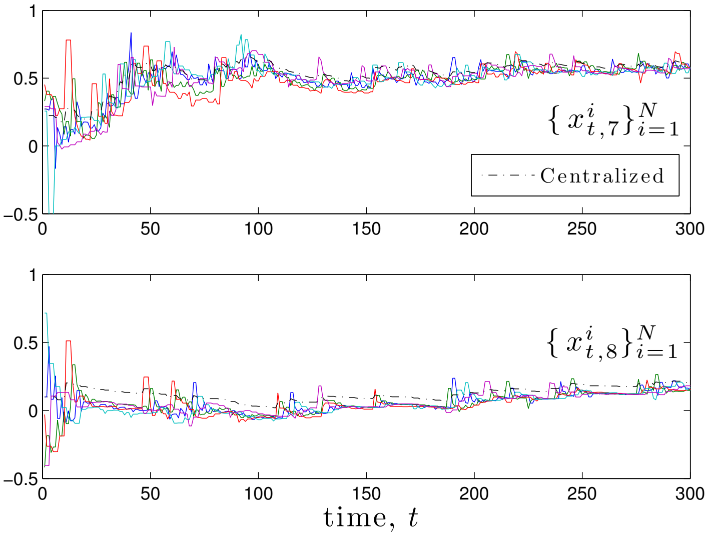
Acknowledgments

This chapter is taken from a journal published by D. Mateos-Núñez and J. Cortés in *IEEE Transactions on Network Science and Engineering 1 (1) (2014), 23-37*, with the title "Distributed online convex optimization over jointly connected digraphs." The authors thank the anonymous reviewers of this journal for their useful feedback that helped us improve the presentation of the manuscript. A preliminary version appeared at the *2014 International Symposium on Mathematical Theory of Networks and Systems.* This research was supported by the US National Science Foundation (NSF) Award CMMI-1300272.



Average regret

Figure 4.3: Simulated temporal average regret of the online gradient descent algorithms with proportional and proportional-integral disagreement feedback (the latter with $a = 4$) versus the centralized online gradient descent algorithm. The dynamics involves $N = 5$ agents communicating over the periodic sequence of digraphs displayed in Figure 4.2. Each local objective $f_t^i : \mathbb{R}^d \to \mathbb{R}$, with $d = 11$, is given by $f_t^i(x) = l(x, w_t^i, y_t^i)$ with loss function $l(x,w,y) = \log\big(1 + e^{-2yx^\top(w,1)}\big)$, for the data set from http://www.stats4stem.org/r-headinjury-data.html. Since $w \in \{0,1\}^{d-1}$ and $y \in \{-1,1\}$, we have $\|\partial_x l(x,w,y)\| \le \|-2y(w,1)\| \le 2d$, so the local cost functions are globally Lipschitz with $H = 2d$. The learning rates are $\eta_t = 1/\sqrt{t}$ (with the same asymptotic behavior as for the Doubling Trick scheme employed in Theorem 4.24). With $\tilde{\delta}' = 0.01$ in (5.17), so that $\tilde{\delta} = \tilde{\delta}'$, equation (5.18) yields $\sigma \in (0.01/\lambda_{\min}(E), 0.99/\lambda_{\max}(E))$, so we take $\sigma = 0.1$ for both dynamics. The initial condition $\boldsymbol{x}_1$ is randomly generated and, in the second-order case, we take $\boldsymbol{z}_1 = \mathbb{1}_5 \otimes \mathbb{1}_{11}$. The scale is logarithmic and the evolutions are bounded by a line of negative slope, as should correspond to a regret bound proportional to $\log(\sqrt{T}/T) = -\frac{1}{2}\log T$. The centralized estimate is computed by the centralized online gradient descent $c_{t+1} = c_t - \eta_t \frac{1}{N}\sum_{i=1}^N \nabla f_t^i(c_t)$ with the same learning rates. The global optimal solution in hindsight, $\hat{x}_T$, for each time horizon T, is computed offline using centralized gradient descent. (As a side note, the agent regret of an algorithm can be sublinear regardless of the design of the cost functions, which ultimately determines the pertinence of the centralized model in hindsight and hence the pertinence of the online distributed performance.)



Agents' estimates using proportional-integral disagreement feedback

Figure 4.4: Agents' estimates for our online gradient descent algorithm with proportional-integral disagreement feedback, where $a = 4$, versus the centralized online gradient descent algorithm. The problem data and algorithm parameters have been chosen as for Figure 4.3. The plot on the top shows the evolution of the 7th coordinate of each agent's estimate, which is the gain associated to the feature "assessed by clinician as high risk for neurological intervention," versus the evolution of the centralized estimate. The centralized estimate is computed by the centralized online gradient descent $c_{t+1} = c_t - \eta_t \frac{1}{N} \sum_{i=1}^{N} \nabla f^i_t(c_t)$ with the same learning rates. The plot on the bottom shows a similar comparison for the gain associated to the feature "loss of consciousness."

Chapter 5

Distributed saddle-point subgradient algorithms with Laplacian averaging

In this chapter we present distributed subgradient methods for min-max problems with agreement constraints on a subset of the arguments of both the convex and concave parts. Applications include constrained minimization problems where each constraint is a sum of convex functions in the local variables of the agents. In the latter case, the proposed algorithm reduces to primal-dual updates using local subgradients and Laplacian averaging on local copies of the multipliers associated to the global constraints. For the case of general convex-concave saddle-point problems, our analysis establishes the convergence of the running time-averages of the local estimates to a saddle point under periodic connectivity of the communication digraphs. Specifically, choosing the gradient step-sizes in a suitable way, we show that the evaluation error is proportional to $1/\sqrt{t}$, where t is the iteration step. We illustrate our results in simulation for an optimization scenario with nonlinear constraints coupling the decisions of agents that cannot communicate directly.

5.1 Distributed algorithms for saddle-point problems under agreement constraints

This section describes the problem of interest. Consider closed convex sets $\boldsymbol{W} \subseteq \mathbb{R}^{d_{\boldsymbol{w}}}$, $\mathcal{D} \subseteq \mathbb{R}^{d_{D}}$, $\boldsymbol{M} \subseteq \mathbb{R}^{d_{\boldsymbol{\mu}}}$, $\mathcal{Z} \subseteq \mathbb{R}^{d_{z}}$ and a function $\boldsymbol{\phi} : \boldsymbol{W} \times \mathcal{D}^N \times \boldsymbol{M} \times \mathcal{Z}^N \to \mathbb{R}$ which is jointly convex on the first two arguments and jointly concave on the last two arguments. We seek to solve the constrained saddle-point problem:

$$\min_{\substack{\boldsymbol{w} \in \boldsymbol{W},\, \boldsymbol{D} \in \mathcal{D}^N \\ D^i = D^j, \forall i,j}} \max_{\substack{\boldsymbol{\mu} \in \boldsymbol{M},\, \boldsymbol{z} \in \mathcal{Z}^N \\ z^i = z^j, \forall i,j}} \boldsymbol{\phi}(\boldsymbol{w}, \boldsymbol{D}, \boldsymbol{\mu}, \boldsymbol{z}), \tag{5.1}$$

where $\boldsymbol{D} := (D^1, \ldots, D^N)$ and $\boldsymbol{z} := (z^1, \ldots, z^N)$. The motivation for distributed algorithms and the consideration of explicit agreement constraints in (5.1) comes from decentralized or parallel computation approaches in network optimization and machine learning. In such scenarios, global decision variables, which need to be determined from the aggregation of local data, can be duplicated into distinct ones so that each agent has its own local version to operate with. Agreement constraints are then imposed across the network to ensure the equivalence to the original optimization problem. We explain this procedure next, specifically through the dual decomposition of optimization problems where objectives and constraints are a sum of convex functions.

5.1.1 Optimization problems with separable constraints

We illustrate here how optimization problems with constraints given by a sum of convex functions can be reformulated in the form (5.1) to make them amenable to distributed algorithmic solutions. Our focus are constraints coupling the local decision vectors of agents that cannot communicate directly.

Consider a group of agents $\{1,\dots,N\}$, and let $f^i : \mathbb{R}^{n_i} \times \mathbb{R}^{d_{\boldsymbol{D}}} \to \mathbb{R}$ and the components of $g^i : \mathbb{R}^{n_i} \times \mathbb{R}^{d_{\boldsymbol{D}}} \to \mathbb{R}^m$ be convex functions associated to agent $i \in \{1,\dots,N\}$. These functions depend on both a local decision vector $w^i \in \mathcal{W}_i$, with $\mathcal{W}_i \subseteq \mathbb{R}^{n_i}$ convex, and on a global decision vector $D \in \mathcal{D}$, with $\mathcal{D} \subseteq \mathbb{R}^{d_{\boldsymbol{D}}}$ convex. The optimization problem reads as

$$\begin{aligned} \min_{\substack{w^i \in \mathcal{W}_i, \forall i \\ D \in \mathcal{D}}} & \sum_{i=1}^{N} f^i(w^i, D) \\ \text{s.t.}\, & g^1(w^1, D) + \dots + g^N(w^N, D) \le 0. \end{aligned} \tag{5.2}$$

This problem can be reformulated as a constrained saddle-point problem as follows. We first construct the corresponding Lagrangian function (2.6) and introduce copies $\{z^i\}_{i=1}^N$ of the Lagrange multiplier z associated to the global constraint in (5.2), then associate each z^i to g^i, and impose the agreement constraint $z^i = z^j$ for all i, j. Similarly, we also introduce copies $\{D^i\}_{i=1}^N$ of the global decision vector D subject to agreement, $D^i = D^j$ for all i, j. The existence of a saddle point implies that strong duality is attained and there exists a solution of the optimization (5.2).

Formally,

$$\min_{\substack{w^i\in\mathcal{W}_i\\ D\in\mathcal{D}}} \max_{z\in\mathbb{R}^m_{\geq 0}} \sum_{i=1}^N f^i(w^i,D)+z^\top\sum_{i=1}^N g^i(w^i,D) \tag{5.3a}$$

$$= \min_{\substack{w^i\in\mathcal{W}_i\\ D\in\mathcal{D}}} \max_{\substack{z^i\in\mathbb{R}^m_{\geq 0}\\ z^i=z^j,\forall i,j}} \sum_{i=1}^N \Big(f^i(w^i,D)+{z^i}^\top g^i(w^i,D)\Big) \tag{5.3b}$$

$$= \min_{\substack{w^i\in\mathcal{W}_i\\ D^i\in\mathcal{D}\\ D^i=D^j,\forall i,j}} \max_{\substack{z^i\in\mathbb{R}^m_{\geq 0}\\ z^i=z^j,\forall i,j}} \sum_{i=1}^N \Big(f^i(w^i,D^i)+{z^i}^\top g^i(w^i,D^i)\Big). \tag{5.3c}$$

This formulation has its roots in the classical dual decompositions surveyed in [BPC$^+$11, Ch. 2], see also [NO10b, Sec. 1.2.3] and [PB13, Sec. 5.4] for the particular case of resource allocation. While [BPC$^+$11, NO10b] suggest to broadcast a centralized update of the multiplier, and the method in [PB13] has an implicit projection onto the probability simplex, the formulation (5.3) has the multiplier associated to the global constraint estimated in a decentralized way. The recent works [BNA14, CNS14, SJR16] implicitly rest on the above formulation of *agreement on the multipliers* Section 5.3 particularizes our general saddle-point strategy to these distributed scenarios.

Remark 1.26. (Distributed formulations via Fenchel conjugates). To illustrate the generality of the min-max problem (5.3c), we show here how *only* the particular case of *linear* constraints can be reduced to a maximization problem under agreement. Consider the particular case of $\min_{w^i\in\mathbb{R}^{n_i}}\sum_{i=1}^N f^i(w^i)$, subject to a linear constraint

$$\sum_{i=1}^N A^i w^i - b \leq 0,$$

with $A^i\in\mathbb{R}^{m\times n_i}$ and $b\in\mathbb{R}^m$. The above formulation suggests a distributed strategy that *eliminates* the primal variables using Fenchel conjugates (2.7). Taking $\{b^i\}_{i=1}^N$

such that $\sum_{i=1}^N b^i = b$, this problem can be transformed, if a saddle-point exists (so that strong duality is attained), into

$$\max_{z \in \mathcal{Z}} \min_{w^i \in \mathbb{R}^{n_i}, \forall i} \sum_{i=1}^N f^i(w^i) + \sum_{i=1}^N (z^\top A^i w^i - z^\top b^i) \tag{5.4a}$$

$$= \max_{z \in \mathcal{Z}} \sum_{i=1}^N \Big(- f^{i\star}(-A^{i^\top} z) - z^\top b^i \Big) \tag{5.4b}$$

$$= \max_{\substack{z^i \in \mathcal{Z}, \forall i \\ z^i = z^j, \forall i,j}} \sum_{i=1}^N \Big(- f^{i\star}(-A^{i^\top} z^i) - z^{i^\top} b^i \Big), \tag{5.4c}$$

where $\mathcal{Z}$ is either $\mathbb{R}^m$ or $\mathbb{R}^m_{\geq 0}$ depending on whether we have equality or inequality ($\leq$) constraints in (5.2). By [RW98, Prop. 11.3], the optimal primal values can be recovered locally as

$$w^{i^*} := \partial f^{i\star}(-A^{i^\top} z^{i^*}), \qquad i \in \{1, \dots, N\} \tag{5.5}$$

without extra communication. Thus, our strategy generalizes the class of convex optimization problems with linear constraints studied in [MARS10], which distinguishes between the *constraint graph* (where edges arise from participation in a constraint) and the *network graph*, and defines *distributed* with respect to the latter.

●

5.1.2 Saddle-point dynamics with Laplacian averaging

We propose a projected subgradient method to solve constrained saddle-point problems of the form (5.1). The agreement constraints are addressed via Laplacian averaging, allowing the design of distributed algorithms *when* the convex-concave functions are separable as in Sections 5.1.1. The generality of this dynamics is inherited by the general structure of the convex-concave min-max problem (5.1).

We have chosen this structure both for convenience of analysis, from the perspective of the saddle-point evaluation error, and, more importantly, because it allows to model problems beyond constrained optimization; see, e.g., [SPFP10] regarding the variational inequality framework, which is equivalent to the saddle-point framework. Formally, the dynamics is

$$\hat{\boldsymbol{w}}_{t+1} = \boldsymbol{w}_t - \eta_t g_{\boldsymbol{w}_t} \tag{5.6a}$$

$$\hat{\boldsymbol{D}}_{t+1} = \boldsymbol{D}_t - \sigma \mathbf{L}_t \boldsymbol{D}_t - \eta_t g_{\boldsymbol{D}_t} \tag{5.6b}$$

$$\hat{\boldsymbol{\mu}}_{t+1} = \boldsymbol{\mu}_t + \eta_t g_{\boldsymbol{\mu}_t} \tag{5.6c}$$

$$\hat{\boldsymbol{z}}_{t+1} = \boldsymbol{z}_t - \sigma \mathbf{L}_t \boldsymbol{z}_t + \eta_t g_{\boldsymbol{z}_t} \tag{5.6d}$$

$$(\boldsymbol{w}_{t+1}, \boldsymbol{D}_{t+1}, \boldsymbol{\mu}_{t+1}, \boldsymbol{z}_{t+1}) = \mathcal{P}_{\boldsymbol{S}}\Big(\hat{\boldsymbol{w}}_{t+1}, \hat{\boldsymbol{D}}_{t+1}, \hat{\boldsymbol{\mu}}_{t+1}, \hat{\boldsymbol{z}}_{t+1}\Big),$$

where $\mathbf{L}_t = \mathsf{L}_t \otimes \mathrm{I}_{d_{\boldsymbol{D}}}$ or $\mathbf{L}_t = \mathsf{L}_t \otimes \mathrm{I}_{d_{\boldsymbol{z}}}$, depending on the context, with L_t the Laplacian matrix of $\mathcal{G}_t$; $\sigma \in \mathbb{R}_{>0}$ is the consensus stepsize, $\{\eta_t\}_{t \geq 1} \subset \mathbb{R}_{>0}$ are the learning rates;

$$\begin{aligned} g_{\boldsymbol{w}_t} &\in \partial_{\boldsymbol{w}} \boldsymbol{\phi}(\boldsymbol{w}_t, \boldsymbol{D}_t, \boldsymbol{\mu}_t, \boldsymbol{z}_t), \\ g_{\boldsymbol{D}_t} &\in \partial_{\boldsymbol{D}} \boldsymbol{\phi}(\boldsymbol{w}_t, \boldsymbol{D}_t, \boldsymbol{\mu}_t, \boldsymbol{z}_t), \\ g_{\boldsymbol{\mu}_t} &\in \partial_{\boldsymbol{\mu}} \boldsymbol{\phi}(\boldsymbol{w}_t, \boldsymbol{D}_t, \boldsymbol{\mu}_t, \boldsymbol{z}_t), \\ g_{\boldsymbol{z}_t} &\in \partial_{\boldsymbol{z}} \boldsymbol{\phi}(\boldsymbol{w}_t, \boldsymbol{D}_t, \boldsymbol{\mu}_t, \boldsymbol{z}_t), \end{aligned}$$

and $\mathcal{P}_{\boldsymbol{S}}$ represents the orthogonal projection onto the closed convex set $\boldsymbol{S} := \boldsymbol{W} \times \mathcal{D}^N \times \boldsymbol{M} \times \mathcal{Z}^N$ as defined in (2.1). This family of algorithms particularize to a novel class of primal-dual consensus-based subgradient methods *when* the convex-concave function takes the Lagrangian form discussed in Section 5.1.1. In general, the dynamics (5.6) goes beyond any specific multi-agent model. However, when interpreted in this context, the Laplacian component corresponds to the model for the interaction among the agents.

In the upcoming analysis, we make network considerations that affect the evolution of $\mathbf{L}_t\boldsymbol{D}_t$ and $\mathbf{L}_t\boldsymbol{z}_t$, which measure the disagreement among the corresponding components of $\boldsymbol{D}_t$ and $\boldsymbol{z}_t$ via the Laplacian of the time-dependent adjacency matrices. These quantities are amenable for distributed computation, i.e., the computation of the ith block requires the blocks D_t^j and z_t^j of the network variables corresponding to indexes j with $\mathsf{a}_{ij,t} := (\mathsf{A}_t)_{ij} > 0$. On the other hand, whether the subgradients in (5.6) can be computed with *local* information *depends* on the structure of the function $\boldsymbol{\phi}$ in (5.1) in the context of a given networked problem. Since this issue is anecdotal for our analysis, for the sake of generality we consider a general convex-concave function $\boldsymbol{\phi}$.

5.2 Convergence analysis

Here we present our technical analysis on the convergence properties of the dynamics (5.6). Our starting point is the assumption that a solution to (5.1) exists, namely, a saddle point $(\boldsymbol{w}^*, \boldsymbol{D}^*, \boldsymbol{\mu}^*, \boldsymbol{z}^*)$ of $\boldsymbol{\phi}$ on $\boldsymbol{S} := \boldsymbol{W} \times \mathcal{D}^N \times \boldsymbol{M} \times \mathcal{Z}^N$ under the agreement condition on $\mathcal{D}^N$ and $\mathcal{Z}^N$. That is, with $\boldsymbol{D}^* = D^* \otimes \mathbb{1}_N$ and $\boldsymbol{z}^* = z^* \otimes \mathbb{1}_N$ for some $(D^*, z^*) \in \mathcal{D} \times \mathcal{Z}$. (We cannot actually conclude the feasibility property of the original problem *from* the evolution of the estimates.) We then study the evolution of the *running time-averages* (sometimes called *ergodic sums*;

see, e.g., [SJR16])

$$\begin{aligned} \boldsymbol{w}_{t+1}^{\mathrm{av}} &= \frac{1}{t}\sum_{s=1}^{t} \boldsymbol{w}_s, \quad & \boldsymbol{D}_{t+1}^{\mathrm{av}} &= \frac{1}{t}\sum_{s=1}^{t} \boldsymbol{D}_s, \\ \boldsymbol{\mu}_{t+1}^{\mathrm{av}} &= \frac{1}{t}\sum_{s=1}^{t} \boldsymbol{\mu}_s, \quad & \boldsymbol{z}_{t+1}^{\mathrm{av}} &= \frac{1}{t}\sum_{s=1}^{t} \boldsymbol{z}_s. \end{aligned}$$

We summarize next our overall strategy to provide the reader with a *roadmap* of the forthcoming analysis. In Section 5.2.1, we bound the saddle-point evaluation error

$$t\boldsymbol{\phi}(\boldsymbol{w}_{t+1}^{\mathrm{av}}, \boldsymbol{D}_{t+1}^{\mathrm{av}}, \boldsymbol{\mu}_{t+1}^{\mathrm{av}}, \boldsymbol{z}_{t+1}^{\mathrm{av}}) - t\boldsymbol{\phi}(\boldsymbol{w}^*, \boldsymbol{D}^*, \boldsymbol{\mu}^*, \boldsymbol{z}^*). \tag{5.7}$$

in terms of the following quantities: the initial conditions, the size of the states of the dynamics, the size of the subgradients, and the cumulative disagreement of the running time-averages. Then, in Section 5.2.2 we bound the cumulative disagreement in terms of the size of the subgradients and the learning rates. Finally, in Section 5.2.3 we establish the saddle-point evaluation convergence result using the assumption that the estimates generated by the dynamics (5.6), as well as the subgradient sets, are uniformly bounded. (This assumption can be met in applications by designing projections that preserve the saddle points, particularly in the case of distributed constrained optimization that we discuss later.) In our analysis, we conveniently choose the learning rates $\{\eta_t\}_{t\geq 1}$ using the Doubling Trick scheme [SS12, Sec. 2.3.1] to find lower and upper bounds on (5.7) proportional to $\sqrt{t}$. Dividing by t finally allows us to conclude that the saddle-point evaluation error of the running time-averages is bounded by $1/\sqrt{t}$.

5.2.1 Saddle-point evaluation error in terms of the disagreement

Here, we establish the saddle-point evaluation error of the running time-averages in terms of the disagreement. Our first result, whose proof is presented in the Appendix, establishes a pair of inequalities regarding the evaluation error of the states of the dynamics with respect to a generic point in the variables of the convex and concave parts, respectively.

Lemma 2.27. (Evaluation error of the states in terms of the disagreement). *Let the sequence* $\{(\boldsymbol{w}_t, \boldsymbol{D}_t, \boldsymbol{\mu}_t, \boldsymbol{z}_t)\}_{t\geq 1}$ *be generated by the coordination algorithm* (5.6) *over a sequence of arbitrary weight-balanced digraphs* $\{\mathcal{G}_t\}_{t\geq 1}$ *such that* $\sup_{t\geq 1}\sigma_{max}(\mathsf{L}_t) \leq \overline{\Lambda}$, *and with*

$$\sigma \leq \Big(\max\Big\{d_{\mathrm{out,t}}(k) \,:\, k\in\mathcal{I},\, t\in\mathbb{Z}_{\geq 1}\Big\}\Big)^{-1}. \tag{5.8}$$

Then, for any sequence of learning rates $\{\eta_t\}_{t\geq 1}\subset\mathbb{R}_{>0}$ *and any* $(\boldsymbol{w}_p, \boldsymbol{D}_p)\in \boldsymbol{W}\times\mathcal{D}^N$, *the following holds:*

$$\begin{aligned}
&2(\boldsymbol{\phi}(\boldsymbol{w}_t, \boldsymbol{D}_t, \boldsymbol{\mu}_t, \boldsymbol{z}_t) - \boldsymbol{\phi}(\boldsymbol{w}_p, \boldsymbol{D}_p, \boldsymbol{\mu}_t, \boldsymbol{z}_t)) \\
\leq\, & \tfrac{1}{\eta_t}\Big(\|\boldsymbol{w}_t - \boldsymbol{w}_p\|_2^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{w}_p\|_2^2\Big) \\
&+ \tfrac{1}{\eta_t}\Big(\|\mathbf{M}\boldsymbol{D}_t - \boldsymbol{D}_p\|_2^2 - \|\mathbf{M}\boldsymbol{D}_{t+1} - \boldsymbol{D}_p\|_2^2\Big) \\
&+ 6\eta_t\|g_{\boldsymbol{w}_t}\|_2^2 + 6\eta_t\|g_{\boldsymbol{D}_t}\|_2^2 \\
&+ 2\|g_{\boldsymbol{D}_t}\|_2(2+\sigma\overline{\Lambda})\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_t\|_2 + 2\|g_{\boldsymbol{D}_t}\|_2\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_p\|_2.
\end{aligned} \tag{5.9}$$

Also, for any $(\boldsymbol{\mu}_p, \boldsymbol{z}_p) \in \boldsymbol{M} \times \mathcal{Z}^N$*, the analogous holds,*

$$\begin{aligned}
&2(\boldsymbol{\phi}(\boldsymbol{w}_t, \boldsymbol{D}_t, \boldsymbol{\mu}_t, \boldsymbol{z}_t) - \boldsymbol{\phi}(\boldsymbol{w}_t, \boldsymbol{D}_t, \boldsymbol{\mu}_p, \boldsymbol{z}_p)) \qquad (5.10)\\
\geq &-\tfrac{1}{\eta_t}\Big(\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_p\|_2^2 - \|\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_p\|_2^2\Big)\\
&-\tfrac{1}{\eta_t}\Big(\|\mathbf{M}\boldsymbol{z}_t - \boldsymbol{z}_p\|_2^2 - \|\mathbf{M}\boldsymbol{z}_{t+1} - \boldsymbol{z}_p\|_2^2\Big)\\
&-6\eta_t\|g_{\boldsymbol{\mu}_t}\|_2^2 - 6\eta_t\|g_{\boldsymbol{z}_t}\|_2^2\\
&-2\|g_{\boldsymbol{z}_t}\|_2(2+\sigma\overline{\Lambda})\|\mathbf{L}_{\mathcal{K}}\boldsymbol{z}_t\|_2 - 2\|g_{\boldsymbol{z}_t}\|_2\|\mathbf{L}_{\mathcal{K}}\boldsymbol{z}_p\|_2.
\end{aligned}$$

Building on Lemma 2.27, we next obtain bounds for the sum over time of the evaluation errors with respect to a generic point and the running time-averages.

Lemma 2.28. (Cumulative evaluation error of the states with respect to running time-averages in terms of disagreement). *Under the same assumptions of Lemma 2.27, for any* $(\boldsymbol{w}_p, \boldsymbol{D}_p, \boldsymbol{\mu}_p, \boldsymbol{z}_p) \in \boldsymbol{W} \times \mathcal{D}^N \times \boldsymbol{M} \times \mathcal{Z}^N$*, the difference*

$$\sum_{s=1}^{t} \boldsymbol{\phi}(\boldsymbol{w}_s, \boldsymbol{D}_s, \boldsymbol{\mu}_s, \boldsymbol{z}_s) - t\boldsymbol{\phi}(\boldsymbol{w}_p, \boldsymbol{D}_p, \boldsymbol{\mu}_{t+1}^{av}, \boldsymbol{z}_{t+1}^{av})$$

is upper-bounded by $\frac{\mathsf{u}(t,\boldsymbol{w}_p,\boldsymbol{D}_p)}{2}$*, while the difference*

$$\sum_{s=1}^{t} \boldsymbol{\phi}(\boldsymbol{w}_s, \boldsymbol{D}_s, \boldsymbol{\mu}_s, \boldsymbol{z}_s) - t\boldsymbol{\phi}(\boldsymbol{w}_{t+1}^{av}, \boldsymbol{D}_{t+1}^{av}, \boldsymbol{\mu}_p, \boldsymbol{z}_p)$$

is lower-bounded by $-\frac{\mathsf{u}(t,\boldsymbol{\mu}_p,\boldsymbol{z}_p)}{2}$, *where*

$$\begin{aligned}
&\mathsf{u}(t,\boldsymbol{w}_p,\boldsymbol{D}_p) \equiv \mathsf{u}\Big(t,\boldsymbol{w}_p,\boldsymbol{D}_p,\{\boldsymbol{w}_s\}_{s=1}^t,\{\boldsymbol{D}_s\}_{s=1}^t\Big) &(5.11)\\
&= \sum_{s=2}^t \Big(\|\boldsymbol{w}_s-\boldsymbol{w}_p\|_2^2+\|\mathbf{M}\boldsymbol{D}_s-\boldsymbol{D}_p\|_2^2\Big)\Big(\tfrac{1}{\eta_s}-\tfrac{1}{\eta_{s-1}}\Big)\\
&\quad+\tfrac{2}{\eta_1}\Big(\|\boldsymbol{w}_1\|_2^2+\|\boldsymbol{w}_p\|_2^2+\|\boldsymbol{D}_1\|_2^2+\|\boldsymbol{D}_p\|_2^2\Big)\\
&\quad+6\sum_{s=1}^t \eta_s(\|g_{\boldsymbol{w}_s}\|_2^2+\|g_{\boldsymbol{D}_s}\|_2^2)\\
&\quad+2(2+\sigma\overline{\Lambda})\sum_{s=1}^t \|g_{\boldsymbol{D}_s}\|_2\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_s\|_2+2\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_p\|_2\sum_{s=1}^t\|g_{\boldsymbol{D}_s}\|_2, &(5.12)
\end{aligned}$$

and $\mathsf{u}(t,\boldsymbol{\mu}_p,\boldsymbol{z}_p) \equiv \mathsf{u}\Big(t,\boldsymbol{\mu}_p,\boldsymbol{z}_p,\{\boldsymbol{\mu}_s\}_{s=1}^t,\{\boldsymbol{z}_s\}_{s=1}^t\Big)$.

Proof. By adding (5.9) over $s=1,\ldots,t$, we obtain

$$\begin{aligned}
&2\sum_{s=1}^t(\boldsymbol{\phi}(\boldsymbol{w}_s,\boldsymbol{D}_s,\boldsymbol{\mu}_s,\boldsymbol{z}_s)-\boldsymbol{\phi}(\boldsymbol{w}_p,\boldsymbol{D}_p,\boldsymbol{\mu}_s,\boldsymbol{z}_s))\\
&\le \sum_{s=2}^t \Big(\|\boldsymbol{w}_s-\boldsymbol{w}_p\|_2^2+\|\mathbf{M}\boldsymbol{D}_s-\boldsymbol{D}_p\|_2^2\Big)\Big(\tfrac{1}{\eta_s}-\tfrac{1}{\eta_{s-1}}\Big)\\
&\quad+\tfrac{1}{\eta_1}\Big(\|\boldsymbol{w}_1-\boldsymbol{w}_p\|_2^2+\|\mathbf{M}\boldsymbol{D}_1-\boldsymbol{D}_p\|_2^2\Big)\\
&\quad+6\sum_{s=1}^t \eta_s(\|g_{\boldsymbol{w}_s}\|_2^2+\|g_{\boldsymbol{D}_s}\|_2^2)\\
&\quad+2(2+\sigma\overline{\Lambda})\sum_{s=1}^t \|g_{\boldsymbol{D}_s}\|_2\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_s\|_2+2\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_p\|_2\sum_{s=1}^t\|g_{\boldsymbol{D}_s}\|_2.
\end{aligned}$$

This is bounded from above by $\mathsf{u}(t,\boldsymbol{w}_p,\boldsymbol{D}_p)$ because $\|\mathbf{M}\boldsymbol{D}_1-\boldsymbol{D}_p\|_2^2 \le 2\|\boldsymbol{D}_1\|_2^2+2\|\boldsymbol{D}_p\|_2^2$, which follows from the triangular inequality, Young's inequality, the submultiplicativity of the norm, and the identity $\|\mathbf{M}\|_2=1$. Finally, by the concavity of $\boldsymbol{\phi}$ in the last two arguments,

$$\boldsymbol{\phi}(\boldsymbol{w}_p,\boldsymbol{D}_p,\boldsymbol{\mu}_{t+1}^{\text{av}},\boldsymbol{z}_{t+1}^{\text{av}}) \ge \frac{1}{t}\sum_{s=1}^t \boldsymbol{\phi}(\boldsymbol{w}_p,\boldsymbol{D}_p,\boldsymbol{\mu}_s,\boldsymbol{z}_s),$$

so the upper bound in the statement follows. Similarly, we obtain the lower bound by adding (5.10) over $s=1,\dots,t$ and using that $\boldsymbol{\phi}$ is jointly convex in the first two arguments,

$$\boldsymbol{\phi}(\boldsymbol{w}_{t+1}^{\mathrm{av}},\boldsymbol{D}_{t+1}^{\mathrm{av}},\boldsymbol{\mu}_s,\boldsymbol{z}_s)\leq\frac{1}{t}\sum_{s=1}^{t}\boldsymbol{\phi}(\boldsymbol{w}_s,\boldsymbol{D}_s,\boldsymbol{\mu}_s,\boldsymbol{z}_s),$$

which completes the proof. □

The combination of the pair of inequalities in Lemma 2.28 allows us to derive the saddle-point evaluation error of the running time-averages in the next result.

Proposition 2.29. (Saddle-point evaluation error of running time-averages). *Under the same hypotheses of Lemma 2.27, for any saddle point* $(\boldsymbol{w}^*,\boldsymbol{D}^*,\boldsymbol{\mu}^*,\boldsymbol{z}^*)$ *of* $\boldsymbol{\phi}$ *on* $\boldsymbol{W}\times\mathcal{D}^N\times\boldsymbol{M}\times\mathcal{Z}^N$ *with* $\boldsymbol{D}^*=D^*\otimes\mathbb{1}_N$ *and* $\boldsymbol{z}^*=z^*\otimes\mathbb{1}_N$ *for some* $(D^*,z^*)\in\mathcal{D}\times\mathcal{Z}$*, the following holds:*

$$\begin{aligned}&-\mathsf{u}(t,\boldsymbol{\mu}^*,\boldsymbol{z}^*)-\mathsf{u}(t,\boldsymbol{w}_{t+1}^{av},\boldsymbol{D}_{t+1}^{av})\\ \leq&\,2t\boldsymbol{\phi}(\boldsymbol{w}_{t+1}^{av},\boldsymbol{D}_{t+1}^{av},\boldsymbol{\mu}_{t+1}^{av},\boldsymbol{z}_{t+1}^{av})-2t\boldsymbol{\phi}(\boldsymbol{w}^*,\boldsymbol{D}^*,\boldsymbol{\mu}^*,\boldsymbol{z}^*)\\ \leq&\,\mathsf{u}(t,\boldsymbol{w}^*,\boldsymbol{D}^*)+\mathsf{u}(t,\boldsymbol{\mu}_{t+1}^{av},\boldsymbol{z}_{t+1}^{av}).\end{aligned}\tag{5.13}$$

Proof. We show the result in two steps, by evaluating the bounds from Lemma 2.28 in two sets of points and combining them. First, choosing $(\boldsymbol{w}_p,\boldsymbol{D}_p,\boldsymbol{\mu}_p,\boldsymbol{z}_p)=(\boldsymbol{w}^*,\boldsymbol{D}^*,\boldsymbol{\mu}^*,\boldsymbol{z}^*)$ in the bounds of Lemma 2.28; invoking the saddle-point relations

$$\boldsymbol{\phi}(\boldsymbol{w}^*,\boldsymbol{D}^*,\boldsymbol{\mu}_{t+1}^{\mathrm{av}},\boldsymbol{z}_{t+1}^{\mathrm{av}})\leq\boldsymbol{\phi}(\boldsymbol{w}^*,\boldsymbol{D}^*,\boldsymbol{\mu}^*,\boldsymbol{z}^*)\leq\boldsymbol{\phi}(\boldsymbol{w}_{t+1}^{\mathrm{av}},\boldsymbol{D}_{t+1}^{\mathrm{av}},\boldsymbol{\mu}^*,\boldsymbol{z}^*)$$

where $(\boldsymbol{w}_t^{\mathrm{av}},\boldsymbol{D}_t^{\mathrm{av}},\boldsymbol{\mu}_t^{\mathrm{av}},\boldsymbol{z}_t^{\mathrm{av}})\in\boldsymbol{W}\times\mathcal{D}^N\times\boldsymbol{M}\times\mathcal{Z}^N$, for each $t\geq1$, by convexity; and combining the resulting inequalities, we obtain

$$-\frac{\mathsf{u}(t,\boldsymbol{\mu}^*,\boldsymbol{z}^*)}{2} \leq \sum_{s=1}^{t} \boldsymbol{\phi}(\boldsymbol{w}_s,\boldsymbol{D}_s,\boldsymbol{\mu}_s,\boldsymbol{z}_s) - t\boldsymbol{\phi}(\boldsymbol{w}^*,\boldsymbol{D}^*,\boldsymbol{\mu}^*,\boldsymbol{z}^*) \leq \frac{\mathsf{u}(t,\boldsymbol{w}^*,\boldsymbol{D}^*)}{2}. \tag{5.14}$$

Choosing $(\boldsymbol{w}_p,\boldsymbol{D}_p,\boldsymbol{\mu}_p,\boldsymbol{z}_p) = (\boldsymbol{w}_{t+1}^{\text{av}},\boldsymbol{D}_{t+1}^{\text{av}},\boldsymbol{\mu}_{t+1}^{\text{av}},\boldsymbol{z}_{t+1}^{\text{av}})$ in the bounds of Lemma 2.28, multiplying each by -1 and combining them, we get

$$\begin{aligned}-\frac{\mathsf{u}(t,\boldsymbol{w}_{t+1}^{\text{av}},\boldsymbol{D}_{t+1}^{\text{av}})}{2} \leq \Big(t\boldsymbol{\phi}(\boldsymbol{w}_{t+1}^{\text{av}},\boldsymbol{D}_{t+1}^{\text{av}},\boldsymbol{\mu}_{t+1}^{\text{av}},\boldsymbol{z}_{t+1}^{\text{av}}) \\ -\sum_{s=1}^{t}\boldsymbol{\phi}(\boldsymbol{w}_s,\boldsymbol{D}_s,\boldsymbol{\mu}_s,\boldsymbol{z}_s)\Big) \leq \frac{\mathsf{u}(t,\boldsymbol{\mu}_{t+1}^{\text{av}},\boldsymbol{z}_{t+1}^{\text{av}})}{2}.\end{aligned} \tag{5.15}$$

The result now follows by summing (5.14) and (5.15). □

5.2.2 Bounding the cumulative disagreement

Given the dependence of the saddle-point evaluation error obtained in Proposition 2.29 on the cumulative disagreement of the estimates $\boldsymbol{D}_t$ and $\boldsymbol{z}_t$, here we bound their disagreement over time. We treat the subgradient terms as perturbations in the dynamics (5.6) and study the input-to-state stability properties of the latter. This approach is well suited for scenarios where the size of the subgradients can be uniformly bounded. Since the coupling in (5.6) with $\boldsymbol{w}_t$ and $\boldsymbol{\mu}_t$, as well as among the estimates $\boldsymbol{D}_t$ and $\boldsymbol{z}_t$ themselves, takes place only through the subgradients, we focus on the following pair of decoupled dynamics,

$$\hat{\boldsymbol{D}}_{t+1} = \boldsymbol{D}_t - \sigma\mathbf{L}_t\boldsymbol{D}_t + \boldsymbol{u}_t^1 \tag{5.16a}$$

$$\hat{\boldsymbol{z}}_{t+1} = \boldsymbol{z}_t - \sigma\mathbf{L}_t\boldsymbol{z}_t + \boldsymbol{u}_t^2 \tag{5.16b}$$

$$(\boldsymbol{D}_{t+1},\boldsymbol{z}_{t+1}) = \mathcal{P}_{\mathcal{D}^N\times\mathcal{Z}^N}\Big(\hat{\boldsymbol{D}}_{t+1},\hat{\boldsymbol{z}}_{t+1}\Big),$$

where $\{\boldsymbol{u}_t^1\}_{t\geq 1} \subset (\mathbb{R}^{d_{\boldsymbol{D}}})^N$, $\{\boldsymbol{u}_t^2\}_{t\geq 1} \subset (\mathbb{R}^{d_{\boldsymbol{z}}})^N$ are arbitrary sequences of disturbances, and $\mathcal{P}_{\mathcal{D}^N \times \mathcal{Z}^N}$ is the orthogonal projection onto $\mathcal{D}^N \times \mathcal{Z}^N$ as defined in (2.1).

The next result characterizes the input-to-state stability properties of (5.16) with respect to the agreement space. The analysis builds on the proof strategy in our previous work [MNC14a, Prop. V.4]. The main trick here is to bound the projection residuals in terms of the disturbance. The proof is presented in the Appendix.

Proposition 2.30. (Cumulative disagreement on (5.16) over jointly-connected weight-balanced digraphs). *Let $\{\mathcal{G}_s\}_{s\geq 1}$ be a sequence of B-jointly connected, δ-nondegenerate, weight-balanced digraphs. For $\tilde{\delta}' \in (0,1)$, let*

$$\tilde{\delta} := \min\Big\{\, \tilde{\delta}',\, (1-\tilde{\delta}')\frac{\delta}{d_{\max}}\,\Big\}, \tag{5.17}$$

where

$$d_{\max} := \max\Big\{\, d_{\text{out,t}}(k) \,:\, k \in \mathcal{I},\, t \in \mathbb{Z}_{\geq 1}\,\Big\}.$$

Then, for any choice of consensus stepsize such that

$$\sigma \in \left[\frac{\tilde{\delta}}{\delta}, \frac{1-\tilde{\delta}}{d_{\max}}\right], \tag{5.18}$$

the dynamics (5.16a) *over $\{\mathcal{G}_t\}_{t\geq 1}$ is input-to-state stable with respect to the nullspace of the matrix $\hat{\mathbf{L}}_{\mathcal{K}}$. Specifically, for any $t \in \mathbb{Z}_{\geq 1}$ and any $\{\boldsymbol{u}_s^1\}_{s=1}^{t-1} \subset (\mathbb{R}^{d_{\boldsymbol{D}}})^N$,*

$$\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_t\|_2 \leq \tfrac{2^4\|\boldsymbol{D}_1\|_2}{3^2}\Big(1 - \frac{\tilde{\delta}}{4N^2}\Big)^{\lceil \frac{t-1}{B} \rceil} + C_u \max_{1\leq s\leq t-1} \|\boldsymbol{u}_s^1\|_2\,, \tag{5.19}$$

where

$$C_u := \frac{2^5/3^2}{1-\left(1-\frac{\tilde{\delta}}{4N^2}\right)^{1/B}} \tag{5.20}$$

and the cumulative disagreement satisfies

$$\sum_{t=1}^{t'} \|\mathbf{L}_{\mathcal{K}} \boldsymbol{D}_t\|_2 \le C_u \Big(\tfrac{\|\boldsymbol{D}_1\|_2}{2} + \sum_{t=1}^{t'-1} \|\boldsymbol{u}_t^1\|_2 \Big). \tag{5.21}$$

Analogous bounds hold interchanging $\boldsymbol{D}_t$ with $\boldsymbol{z}_t$.

5.2.3 Convergence of saddle-point subgradient dynamics with Laplacian averaging

Here we characterize the convergence properties of the dynamics (5.6) using the developments above. In informal terms, our main result states that, under a mild connectivity assumption on the communication digraphs, a suitable choice of decreasing stepsizes, and assuming that the agents' estimates and the subgradient sets are uniformly bounded, the saddle-point evaluation error under (5.6) decreases proportionally to $\frac{1}{\sqrt{t}}$. We select the learning rates according to the following scheme.

Assumption 2.31. (*Doubling Trick scheme* for the learning rates). *The agents define a sequence of epochs numbered by $m = 0, 1, 2, \ldots$, and then use the constant value $\eta_s = \frac{1}{\sqrt{2^m}}$ in each epoch m, which has 2^m time steps $s = 2^m, \ldots, 2^{m+1}-1$. Namely,*

$$\begin{aligned} \eta_1 &= 1, & \eta_2 &= \eta_3 = 1/\sqrt{2}, \\ \eta_4 &= \cdots = \eta_7 = 1/2, & \eta_8 &= \cdots = \eta_{15} = 1/\sqrt{8}, \end{aligned}$$

and so on. In general,

$$\eta_{2^m} = \cdots = \eta_{2^{m+1}-1} = 1/\sqrt{2^m}. \qquad \bullet$$

Note that the agents can compute the values in Assumption 2.31 without communicating with each other. Figure 5.1 provides an illustration of this learning rate selection and compares it against constant and other sequences of stepsizes. Note that, unlike other choices commonly used in optimization [BT97, BNO03], the Doubling Trick gives rise to a sequence of stepsizes that is not square summable.

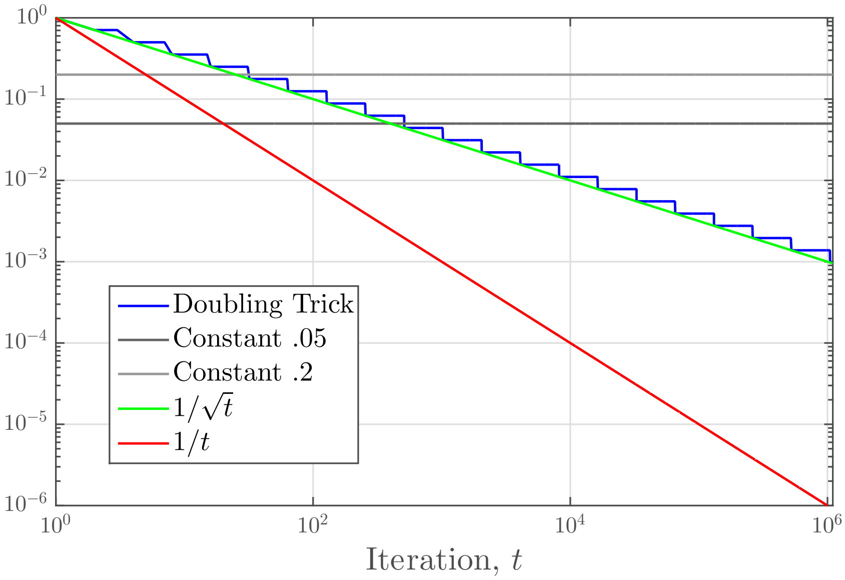


Figure 5.1: Comparison of sequences of learning rates. We compare the sequence of learning rates resulting from the Doubling Trick in Assumption 2.31 against a constant stepsize, the sequence $\{1/\sqrt{t}\}_{t\geq 1}$, and the square-summable harmonic sequence $\{1/t\}_{t\geq 1}$.

Theorem 2.32. (Convergence of the *saddle-point dynamics with Laplacian averaging* (5.6)). *Let* $\{(\boldsymbol{w}_t, \boldsymbol{D}_t, \boldsymbol{\mu}_t, \boldsymbol{z}_t)\}_{t\geq 1}$ *be generated by* (5.6) *over a sequence* $\{\mathcal{G}_t\}_{t\geq 1}$ *of* B*-jointly connected,* δ*-nondegenerate, weight-balanced digraphs satisfying*

$\sup_{t\geq 1}\sigma_{max}(\mathsf{L}_t) \leq \overline{\Lambda}$ *with* σ *selected as in* (5.18). *Assume*

$$\|\boldsymbol{w}_t\|_2 \leq B_{\boldsymbol{w}},\ \ \|\boldsymbol{D}_t\|_2 \leq B_{\boldsymbol{D}},\ \ \|\boldsymbol{\mu}_t\|_2 \leq B_{\boldsymbol{\mu}},\ \ \|\boldsymbol{z}_t\|_2 \leq B_{\boldsymbol{z}},$$

for all $t \in \mathbb{Z}_{\geq 1}$ *whenever the sequence of learning rates* $\{\eta_t\}_{t\geq 1} \subset \mathbb{R}_{>0}$ *is uniformly bounded. Similarly, assume*

$$\|g_{\boldsymbol{w}_t}\|_2 \leq H_{\boldsymbol{w}},\ \|g_{\boldsymbol{D}_t}\|_2 \leq H_{\boldsymbol{D}},\ \|g_{\boldsymbol{\mu}_t}\|_2 \leq H_{\boldsymbol{\mu}},\ \|g_{\boldsymbol{z}_t}\|_2 \leq H_{\boldsymbol{z}}$$

for all $t \in \mathbb{Z}_{\geq 1}$. *Let the learning rates be chosen according to the Doubling Trick in Assumption 2.31. Then, for any saddle point* $(\boldsymbol{w}^*, \boldsymbol{D}^*, \boldsymbol{\mu}^*, \boldsymbol{z}^*)$ *of* $\boldsymbol{\phi}$ *on* $\boldsymbol{W} \times \mathcal{D}^N \times \boldsymbol{M} \times \mathcal{Z}^N$ *with* $\boldsymbol{D}^* = D^* \otimes \mathbb{1}_N$ *and* $\boldsymbol{z}^* = z^* \otimes \mathbb{1}_N$ *for some* $(D^*, z^*) \in \mathcal{D} \times \mathcal{Z}$, *which is assumed to exist, the following holds for the running time-averages:*

$$\begin{aligned} -\frac{\alpha_{\boldsymbol{\mu},\boldsymbol{z}} + \alpha_{\boldsymbol{w},\boldsymbol{D}}}{2\sqrt{t-1}} \leq & \boldsymbol{\phi}(\boldsymbol{w}_t^{av}, \boldsymbol{D}_t^{av}, \boldsymbol{z}_t^{av}, \boldsymbol{\mu}_t^{av}) - \boldsymbol{\phi}(\boldsymbol{w}^*, \boldsymbol{D}^*, \boldsymbol{z}^*, \boldsymbol{\mu}^*) \\ \leq & \frac{\alpha_{\boldsymbol{w},\boldsymbol{D}} + \alpha_{\boldsymbol{\mu},\boldsymbol{z}}}{2\sqrt{t-1}}, \end{aligned} \tag{5.22}$$

where $\alpha_{\boldsymbol{w},\boldsymbol{D}} := \frac{\sqrt{2}}{\sqrt{2}-1}\hat{\alpha}_{\boldsymbol{w},\boldsymbol{D}}$ *with*

$$\begin{aligned} \hat{\alpha}_{\boldsymbol{w},\boldsymbol{D}} := & 4(B_{\boldsymbol{w}}^2 + B_{\boldsymbol{D}}^2) + 6(H_{\boldsymbol{w}}^2 + H_{\boldsymbol{D}}^2) \\ & + H_{\boldsymbol{D}}(3 + \sigma\overline{\Lambda})C_u\Big(B_{\boldsymbol{D}} + 2H_{\boldsymbol{D}}\Big), \end{aligned} \tag{5.23}$$

and $\alpha_{\boldsymbol{z},\boldsymbol{\mu}}$ *is analogously defined.*

Proof. We divide the proof in two steps. In step (i), we use the general bound of Proposition 2.29 making a choice of constant learning rates over a fixed time horizon t'. In step (ii), we use multiple times this bound together with the Doubling Trick to produce the implementation procedure in the statement. In (i), to further bound (5.13), we choose $\eta_t = \eta'$ for all $s \in \{1,\dots,t'\}$ in both $\mathsf{u}(t',\boldsymbol{w}^*,\boldsymbol{D}^*)$ and $\mathsf{u}(t',\boldsymbol{w}^{\mathrm{av}}_{t'+1},\boldsymbol{D}^{\mathrm{av}}_{t'+1})$. By doing this, we make zero the first two lines in (5.11), and then we upper-bound the remaining terms using the bounds on the estimates and the subgradients. The resulting inequality also holds replacing $(\boldsymbol{w}^{\mathrm{av}}_{t'+1},\boldsymbol{D}^{\mathrm{av}}_{t'+1})$ by $(\boldsymbol{w}^*,\boldsymbol{D}^*)$,

$$\begin{aligned}\mathsf{u}(t',\boldsymbol{w}^{\mathrm{av}}_{t'+1},\boldsymbol{D}^{\mathrm{av}}_{t'+1}) \le{}& \tfrac{2}{\eta'}\Big(\|\boldsymbol{w}_1\|_2^2 + B_{\boldsymbol{w}}^2 + \|\boldsymbol{D}_1\|_2^2 + B_{\boldsymbol{D}}^2\Big)\\ &+6(H_{\boldsymbol{w}}^2+H_{\boldsymbol{D}}^2)\eta' t'\\ +2(2+\sigma\overline{\Lambda})H_{\boldsymbol{D}}\sum_{s=1}^{t'}&\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_s\|_2 + 2\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}^{\mathrm{av}}_{t'+1}\|_2 H_{\boldsymbol{D}} t'\,. \end{aligned} \tag{5.24}$$

Regarding the bound for $\mathsf{u}(t',\boldsymbol{w}^*,\boldsymbol{D}^*)$, we just note that $\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}^*\|_2 = 0$, whereas for $\mathsf{u}(t',\boldsymbol{w}^{\mathrm{av}}_{t'+1},\boldsymbol{D}^{\mathrm{av}}_{t'+1})$, we note that, by the triangular inequality, we have

$$\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}^{\mathrm{av}}_{t'+1}\|_2 = \frac{1}{t'}\|\mathbf{L}_{\mathcal{K}}\Big(\sum_{s=1}^{t'}\boldsymbol{D}_s\Big)\|_2 \le \frac{1}{t'}\sum_{s=1}^{t'}\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_s\|_2.$$

That is, we get

$$\begin{aligned}\mathsf{u}(t',\boldsymbol{w}^*,\boldsymbol{D}^*) &\le \mathsf{u}(t',\boldsymbol{w}^{\mathrm{av}}_{t'+1},\boldsymbol{D}^{\mathrm{av}}_{t'+1})\\ &\le \tfrac{2}{\eta'}\Big(\|\boldsymbol{w}_1\|_2^2 + B_{\boldsymbol{w}}^2 + \|\boldsymbol{D}_1\|_2^2 + B_{\boldsymbol{D}}^2\Big) + 6(H_{\boldsymbol{w}}^2+H_{\boldsymbol{D}}^2)\eta' t'\\ &\quad + 2H_{\boldsymbol{D}}(3+\sigma\overline{\Lambda})\sum_{s=1}^{t'}\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_s\|_2. \end{aligned} \tag{5.25}$$

We now further bound $\sum_{s=1}^{t'}\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_s\|_2$ in (5.21) noting that $\|\boldsymbol{u}_t^1\|_2 = \|\eta_t g_{\boldsymbol{D}_t}\|_2 \le$

$\eta_t H_{\boldsymbol{D}} = \eta' H_{\boldsymbol{D}}$, to obtain

$$\begin{aligned}\sum_{s=1}^{t'} \|\mathbf{L}_{\mathcal{K}} \boldsymbol{D}_s\|_2 &\leq C_u\Big(\tfrac{\|\boldsymbol{D}_1\|_2}{2} + \sum_{s=1}^{t'-1} \eta' H_{\boldsymbol{D}}\Big) \\ &\leq C_u\Big(\tfrac{\|\boldsymbol{D}_1\|_2}{2} + t'\eta' H_{\boldsymbol{D}}\Big).\end{aligned}$$

Substituting this bound in (5.25), taking $\eta' = \frac{1}{\sqrt{t'}}$ and noting that $1 \leq \sqrt{t'}$, we get

$$\mathsf{u}(t', \boldsymbol{w}_{t'+1}^{\text{av}}, \boldsymbol{D}_{t'+1}^{\text{av}}) \leq \alpha' \sqrt{t'}, \tag{5.26}$$

where

$$\begin{aligned}\alpha' :=& 2(\|\boldsymbol{w}_1\|_2^2 + \|\boldsymbol{D}_1\|_2^2 + B_{\boldsymbol{w}}^2 + B_{\boldsymbol{D}}^2) + 6(H_{\boldsymbol{w}}^2 + H_{\boldsymbol{D}}^2) \\ &+ 2H_{\boldsymbol{D}}(3 + \sigma\overline{\Lambda}) C_u\Big(\tfrac{\|\boldsymbol{D}_1\|_2}{2} + H_{\boldsymbol{D}}\Big).\end{aligned}$$

This bound is of the type $\mathsf{u}(t', \boldsymbol{w}_{t'+1}^{\text{av}}, \boldsymbol{D}_{t'+1}^{\text{av}}) \leq \alpha' \sqrt{t'}$, where α' depends on the initial conditions. This leads to step (ii). According to the Doubling Trick, for $m = 0, 1, \ldots \lceil \log_2 t \rceil$, the dynamics is executed in each epoch of $t' = 2^m$ time steps $t = 2^m, \ldots, 2^{m+1} - 1$, where at the beginning of each epoch the initial conditions are the final values in the previous epoch. The bound for $\mathsf{u}(t', \boldsymbol{w}_{t'+1}^{\text{av}}, \boldsymbol{D}_{t'+1}^{\text{av}})$ in each epoch is $\alpha' \sqrt{t'} = \alpha_m \sqrt{2^m}$, where α_m is the multiplicative constant in (5.26) that depends on the initial conditions in the corresponding epoch. Using the assumption that the estimates are bounded, i.e., $\alpha_m \leq \hat{\alpha}_{\boldsymbol{w},\boldsymbol{D}}$, we deduce that the bound in each epoch is $\hat{\alpha}_{\boldsymbol{w},\boldsymbol{D}} \sqrt{2^m}$. By the Doubling Trick,

$$\sum_{m=0}^{\lceil \log_2 t \rceil} \sqrt{2^m} = \frac{1 - \sqrt{2}^{\lceil \log_2 t \rceil + 1}}{1 - \sqrt{2}} \leq \frac{1 - \sqrt{2t}}{1 - \sqrt{2}} \leq \frac{\sqrt{2}}{\sqrt{2} - 1} \sqrt{t},$$

we conclude that

$$\mathsf{u}(t, \boldsymbol{w}^*, \boldsymbol{D}^*) \leq \mathsf{u}(t, \boldsymbol{w}^{\text{av}}_{t+1}, \boldsymbol{D}^{\text{av}}_{t+1}) \leq \tfrac{\sqrt{2}}{\sqrt{2}-1} \hat{\alpha}_{\boldsymbol{w},\boldsymbol{D}} \sqrt{t}.$$

Similarly,

$$-\mathsf{u}(t, \boldsymbol{\mu}^*, \boldsymbol{z}^*) \geq -\mathsf{u}(t, \boldsymbol{\mu}^{\text{av}}_{t+1}, \boldsymbol{z}^{\text{av}}_{t+1}) \geq -\tfrac{\sqrt{2}}{\sqrt{2}-1} \hat{\alpha}_{\boldsymbol{\mu},\boldsymbol{z}} \sqrt{t}.$$

The desired pair of inequalities follows substituting these bounds in (5.13) and dividing by $2t$. □

In the statement of Theorem 2.32, the constant C_u appearing in (5.23) encodes the dependence on the network properties. The running time-averages can be updated sequentially as $\boldsymbol{w}^{\text{av}}_{t+1} := \frac{t-1}{t}\boldsymbol{w}^{\text{av}}_t + \frac{1}{t}\boldsymbol{w}_t$ without extra memory. Note also that we assume feasibility of the problem because this property does not follow from the behavior of the algorithm.

Remark 2.33. (Boundedness of estimates). The statement of Theorem 2.32 requires the subgradients and the estimates produced by the dynamics to be bounded. In the literature of distributed (sub-) gradient methods, it is fairly common to assume the boundedness of the subgradient sets relying on their continuous dependence on the arguments, which in turn are assumed to belong to a compact domain. Our assumption on the boundedness of the estimates, however, concerns a saddle-point subgradient dynamics for general convex-concave functions, and its consequences vary depending on the application. We come back to this point and discuss the treatment of dual variables for distributed constrained optimization in Section 5.3.1. •

5.3 Applications to distributed constrained convex optimization

In this section we particularize our convergence result in Theorem 2.32 to the case of convex-concave functions arising from the Lagrangian of the constrained optimization (5.2) discussed in Section 5.1.1. The Lagrangian formulation with explicit agreement constraints (5.3c) matches the general saddle-point problem (5.1) for the convex-concave function $\boldsymbol{\phi} : (\mathcal{W}_1 \times \cdots \times \mathcal{W}_N) \times \mathcal{D}^N \times (\mathbb{R}^m_{\geq 0})^N \to \mathbb{R}$ defined by

$$\boldsymbol{\phi}(\boldsymbol{w}, \boldsymbol{D}, \boldsymbol{z}) = \sum_{i=1}^{N} \Big(f^i(w^i, D^i) + {z^i}^\top g^i(w^i, D^i) \Big). \tag{5.27}$$

Here the arguments of the convex part are, on the one hand, the local primal variables across the network, $\boldsymbol{w} = (w^1, \dots, w^N)$ (not subject to agreement), and, on the other hand, the copies across the network of the global decision vector, $\boldsymbol{D} = (D^1, \dots, D^N)$ (subject to agreement). The arguments of the concave part are the network estimates of the Lagrange multiplier, $\boldsymbol{z} = (z^1, \dots, z^N)$ (subject to agreement). Note that this convex-concave function is the associated Lagrangian for (5.2) *only* under the agreement on the global decision vector and on the Lagrange multiplier associated to the global constraint, i.e.,

$$\mathcal{L}(\boldsymbol{w}, D, z) = \boldsymbol{\phi}(\boldsymbol{w}, D \otimes \mathbb{1}_N, z \otimes \mathbb{1}_N). \tag{5.28}$$

In this case, the *saddle-point dynamics with Laplacian averaging* (5.6) takes the following form: the updates of each agent $i \in \{1, \dots, N\}$ are as follows,

$$\hat{w}^i_{t+1} = w^i_t - \eta_t (d_{f^i, w^i_t} + d^\top_{g^i, w^i_t} z^i), \tag{5.29a}$$

$$\begin{aligned}\hat{D}^i_{t+1} = {} & D^i_t + \sigma \sum_{j=1}^N \mathsf{a}_{ij,t}(D^j_t - D^i_t) \\ & - \eta_t (d_{f^i, D^i_t} + d^\top_{g^i, D^i_t} z^i),\end{aligned} \tag{5.29b}$$

$$\hat{z}^i_{t+1} = z^i_t + \sigma \sum_{j=1}^N \mathsf{a}_{ij,t}(z^j_t - z^i_t) + \eta_t g^i(w^i_t), \tag{5.29c}$$

$$\begin{bmatrix} w^i_{t+1} \\ D^i_{t+1} \\ z^i_{t+1} \end{bmatrix} = \begin{bmatrix} \mathcal{P}_{\mathcal{W}_i}\big(\hat{w}^i_{t+1}\big) \\ \mathcal{P}_{\mathcal{D}}\big(\hat{D}^i_{t+1}\big) \\ \mathcal{P}_{\mathbb{R}^m_{\geq 0} \cap \bar{\mathcal{B}}(0,r)}\big(\hat{z}^i_{t+1}\big) \end{bmatrix}, \tag{5.29d}$$

where the vectors $d_{f^i, w^i_t} \in \mathbb{R}^{n_i}$ and $d_{f^i, D^i_t} \in \mathbb{R}^{d_{\boldsymbol{D}}}$ are subgradients of f^i with respect to the first and second arguments, respectively, at the point (w^i_t, D^i_t), i.e.,

$$d_{f^i, w^i_t} \in \partial_{w^i} f^i(w^i_t, D^i_t), \qquad d_{f^i, D^i_t} \in \partial_D f^i(w^i_t, D^i_t), \tag{5.30}$$

and the matrices $d_{g^i, w^i} \in \mathbb{R}^{m \times n_i}$ and $d_{g^i, D} \in \mathbb{R}^{m \times d_{\boldsymbol{D}}}$ contain in the lth row an element of the subgradient sets $\partial_{w^i} g^i_l(w^i_t, D^i_t)$ and $\partial_D g^i_l(w^i_t, D^i_t)$, respectively. (Note that these matrices correspond, in the differentiable case, to the Jacobian block-matrices of the vector function $g^i : \mathbb{R}^{n_i} \times \mathbb{R}^{d_{\boldsymbol{D}}} \to \mathbb{R}^m$.) We refer to this strategy as the *Consensus-based Saddle-Point (Sub-) Gradient (C-SP-SG) algorithm* and present it in pseudo-code format in Algorithm 1.

Note that the orthogonal projection of the estimates of the multipliers in (5.29d) is unique. The radius r employed in its definition is a design parameter that is either set *a priori* or determined by the agents. We discuss this point in detail below in Section 5.3.1.

Algorithm 1: C-SP-SG algorithm

```
Data: Agents' data for Problem (5.2): $\{f^i, g^i, \mathcal{W}_i\}_{i=1}^N$, $\mathcal{D}$
  Agents' adjacency values $\{\mathsf{A}_t\}_{t\geq 1}$
  Consensus stepsize $\sigma$ as in (5.18)
  Learning rates $\{\eta_t\}_{t\geq 1}$ as in Assumption 2.31
  Radius $r$ s.t. $\bar{\mathcal{B}}(0,r)$ contains optimal dual set for (5.2)
  Number of iterations $T$, indep. of rest of parameters
Result: Agent $i$ outputs $(w_T^i)^{\text{av}}, (D_T^i)^{\text{av}}, (z_T^i)^{\text{av}}$
Initialization: Agent $i$ sets $w_1^i \in \mathbb{R}^{n_i}$, $D_1^i \in \mathbb{R}^{d_D}$, $z_1^i \in \mathbb{R}_{\geq 0}^m$,
                $(w_1^i)^{\text{av}} = w_1^i$, $(D_1^i)^{\text{av}} = D_1^i$, $(z_1^i)^{\text{av}} = z_1^i$
for $t \in \{2, \dots, T-1\}$ do
    for $i \in \{1, \dots, N\}$ do
        Agent $i$ selects (sub-) gradients as in (5.30)
        Agent $i$ updates $(w_t^i, D_t^i, z_t^i)$ as in (5.29)
        Agent $i$ updates $(w_{t+1}^i)^{\text{av}} = \frac{t-1}{t}(w_t^i)^{\text{av}} + \frac{1}{t} w_t^i$,
                  $(D_{t+1}^i)^{\text{av}} = \frac{t-1}{t}(D_t^i)^{\text{av}} + \frac{1}{t} D_t^i$ ,
                  $(z_{t+1}^i)^{\text{av}} = \frac{t-1}{t}(z_t^i)^{\text{av}} + \frac{1}{t} z_t^i$
    end
end
```

The characterization of the saddle-point evaluation error under (5.29) is a direct consequence of Theorem 2.32.

Corollary 3.34. (Convergence of the C-SP-SG algorithm). *For each $i \in \{1, \dots, N\}$, let the sequence $\{(w_t^i, D_t^i, z_t^i)\}_{t\geq 1}$ be generated by the coordination algorithm* (5.29), *over a sequence of graphs $\{\mathcal{G}_t\}_{t\geq 1}$ satisfying the same hypotheses as Theorem 2.32. Assume that the sets $\mathcal{D}$ and $\mathcal{W}_i$ are compact (besides being convex), and the radius r is such that $\bar{\mathcal{B}}(0,r)$ contains the optimal dual set of the constrained optimization* (5.2). *Assume also that the subgradient sets are bounded, in $\mathcal{W}_i \times \mathcal{D}$, as follows,*

$$\partial_{w^i} f^i \subseteq \bar{\mathcal{B}}(0, H_{f,w}),\ \partial_D f^i \subseteq \bar{\mathcal{B}}(0, H_{f,D}),$$

$$\partial_{w^i} g_l^i \subseteq \bar{\mathcal{B}}(0, H_{g,w}),\ \partial_D g_l^i \subseteq \bar{\mathcal{B}}(0, H_{g,D}),$$

for all $l \in \{1,\dots,m\}$*. Let* $(\boldsymbol{w}^*, D^*, z^*)$ *be any saddle point of the Lagrangian* $\mathcal{L}$ *defined in* (5.28) *on the set* $(\mathcal{W}_1 \times \cdots \times \mathcal{W}_N) \times \mathcal{D} \times \mathbb{R}^m$*. (The existence of such saddle-point implies that strong duality is attained.) Then, under Assumption 2.31 for the learning rates, the saddle-point evaluation error* (5.22) *holds for the running time-averages:*

$$\begin{aligned} -\frac{\alpha_{\boldsymbol{\mu},\boldsymbol{z}}+\alpha_{\boldsymbol{w},\boldsymbol{D}}}{2\sqrt{t-1}} &\leq \boldsymbol{\phi}(\boldsymbol{w}_t^{av},\boldsymbol{D}_t^{av},\boldsymbol{z}_t^{av}) - \mathcal{L}(\boldsymbol{w}^*, D^*, z^*) \\ &\leq \frac{\alpha_{\boldsymbol{w},\boldsymbol{D}}+\alpha_{\boldsymbol{\mu},\boldsymbol{z}}}{2\sqrt{t-1}}, \end{aligned} \tag{5.31}$$

for $\alpha_{\boldsymbol{w},\boldsymbol{D}}$ *and* $\alpha_{\boldsymbol{\mu},\boldsymbol{z}}$ *as in* (5.23)*, with*

$$\begin{aligned} B_{\boldsymbol{\mu}} &= H_{\boldsymbol{\mu}} = 0, \qquad B_{\boldsymbol{z}} = \sqrt{N}r, \\ B_{\boldsymbol{w}} &= \Big(\sum_{i=1}^{N} \operatorname{diam}(\mathcal{W}_i)^2\Big)^{1/2}, \qquad B_{\boldsymbol{D}} = \sqrt{N}\operatorname{diam}(\mathcal{D}), \\ H_{\boldsymbol{w}}^2 &= N(H_{f,w} + r\sqrt{m}H_{g,w})^2, \quad H_{\boldsymbol{z}}^2 = \sum_{i=1}^{N}\Big(\sup_{w^i \in \mathcal{W}_i} g_i(w^i)\Big)^2, \\ H_{\boldsymbol{D}}^2 &= N(H_{f,D} + r\sqrt{m}H_{g,D})^2, \end{aligned}$$

where $\operatorname{diam}(\cdot)$ *refers to the diameter of the sets.*

The proof of this result follows by noting that the hypotheses of Theorem 2.32 are automatically satisfied. The only point to observe is that all the saddle points of the Lagrangian $\mathcal{L}$ defined in (5.28) on the set $(\mathcal{W}_1 \times \cdots \times \mathcal{W}_N) \times \mathcal{D} \times \mathbb{R}^m_{\geq 0}$, are also contained in $(\mathcal{W}_1 \times \cdots \times \mathcal{W}_N) \times \mathcal{D} \times \bar{\mathcal{B}}(0,r)$. Note also that we assume feasibility of the problem because this property does not follow from the behavior of the algorithm.

Remark 3.35. (Time, memory, computation, and communication complexity of the C-SP-SG algorithm). We discuss here the complexities associated with the execution of the C-SP-SG algorithm:

- **Time complexity**: According to Corollary 3.34, the saddle-point evaluation error is smaller than ϵ if $\frac{\alpha_{\boldsymbol{w},\boldsymbol{D}}+\alpha_{\boldsymbol{\mu},\boldsymbol{z}}}{2\sqrt{t}} \le \epsilon$. This provides a lower bound

$$t \ge \Big(\frac{\alpha_{\boldsymbol{w},\boldsymbol{D}}+\alpha_{\boldsymbol{\mu},\boldsymbol{z}}}{2\epsilon}\Big)^2,$$

  on the number of required iterations.

- **Memory complexity**: Each agent i maintains the current updates $(w_t^i, D_t^i, z_t^i) \in \mathbb{R}^{n_i} \times \mathbb{R}^{d_{\boldsymbol{D}}} \times \mathbb{R}^m$, and the corresponding current running time-averages $((w_t^i)^{\text{av}}$, $(D_t^i)^{\text{av}}$, $(z_t^i)^{\text{av}})$ with the same dimensions.

- **Computation complexity**: Each agent i makes a choice/evaluation of subgradients, at each iteration, from the subdifferentials $\partial_{w^i} f^i \subseteq \mathbb{R}^{n_i}$, $\partial_D f^i \subseteq \mathbb{R}^{d_{\boldsymbol{D}}}$, $\partial_{w^i} g_l^i \subseteq \mathbb{R}^{n_i}$, $\partial_D g_l^i \subseteq \mathbb{R}^{d_{\boldsymbol{D}}}$, the latter for $l \in \{1,\dots,m\}$. Each agent also projects its estimates on the set $\mathcal{W}_i \times \mathcal{D} \times \mathbb{R}^m_{\ge 0} \cap \bar{\mathcal{B}}(0,r)$. The complexity of this computation depends on the sets $\mathcal{W}_i$ and $\mathcal{D}$.

- **Communication complexity:** Each agent i shares with its neighbors at each iteration a vector in $\mathbb{R}^{d_{\boldsymbol{D}}} \times \mathbb{R}^m$. With the information received, the agent updates the global decision variable D_t^i in (5.29b) and the Lagrange multiplier z_t^i in (5.29c). (Note that the variable D_t^i needs to be maintained and communicated only if the optimization problem (5.2) has a global decision variable.) •

5.3.1 Distributed strategy to bound the optimal dual set

The motivation for the design choice of *truncating* the projection of the dual variables onto a bounded set in (5.29d) is the following. The subgradients of $\boldsymbol{\phi}$ with respect to the primal variables are *linear* in the dual variables. To guarantee the boundedness of the subgradients of $\boldsymbol{\phi}$ and of the dual variables, required by the application of Theorem 2.32, one can introduce a projection step onto a compact set that preserves the optimal dual set, a technique that has been used in [NO09b, NO10a, CNS14]. These works select the bound for the projection *a priori*, whereas [ZM12] proposes a distributed algorithm to compute a bound preserving the optimal dual set, *for* the case of a global inequality constraint *known to all the agents*. Here, we deal with a complementary case, where the constraint is a sum of functions, each known to the corresponding agents, that couple the local decision vectors across the network. For this case, we next describe how the agents can compute, in a distributed way, a radius $r \in \mathbb{R}_{>0}$ such that the ball $\bar{\mathcal{B}}(0,r)$ contains the optimal dual set for the constrained optimization (5.2). A radius with such property is not unique, and estimates with varying degree of conservativeness are possible.

In our model, each agent i has only access to the set $\mathcal{W}_i$ and the functions f^i and g^i. In turn, we make the important assumption that there are no variables subject to agreement, i.e., $f^i(w^i, {\scriptstyle D}) = f^i(w^i)$ and $g^i(w^i, {\scriptstyle D}) = g^i(w^i)$ for all $i \in \{1, \dots, N\}$, and we leave for future work the generalization to the case where agreement variables are present. Consider then the following problem,

$$\begin{aligned} \min_{w^i \in \mathcal{W}_i, \forall i} & \sum_{i=1}^{N} f^i(w^i) \\ \text{s.t.}\, & g^1(w^1) + \cdots + g^N(w^N) \leq 0 \end{aligned} \tag{5.32}$$

where each $\mathcal{W}_i$ is compact as in Corollary 3.34. We first propose a bound on the optimal dual set and then describe a distributed strategy that allows the agents to compute it. Let $(\tilde{w}^1, \dots, \tilde{w}^N) \in \mathcal{W}_1 \times \cdots \times \mathcal{W}_N$ be a vector satisfying the *Strong*

Slater condition [HUL93, Sec. 7.2.3], called *Slater vector*, and define

$$\gamma := \min_{l\in\{1,\dots,m\}} -\sum_{i=1}^{N} g_l^i(\tilde{w}^i), \tag{5.33}$$

which is positive by construction. According to [NO10a, Lemma 1] (which we amend imposing that the Slater vector belongs to the abstract constraint set $(\mathcal{W}_1 \times \cdots \times \mathcal{W}_N)$), we get that the optimal dual set $\mathcal{Z}^* \subseteq \mathbb{R}^m_{\geq 0}$ associated to the constraint $g^1(w^1) + \cdots + g^N(w^N) \leq 0$ is bounded as follows,

$$\max_{z^*\in\mathcal{Z}^*} \|z^*\|_2 \leq \frac{1}{\gamma}\Big(\sum_{i=1}^{N} f^i(\tilde{w}^i) - q(\bar{z})\Big), \tag{5.34}$$

for any $\bar{z} \in \mathbb{R}^m_{\geq 0}$, where $q : \mathbb{R}^m_{\geq 0} \to \mathbb{R}$ is the dual function associated to the optimization (5.32),

$$\begin{aligned} q(z) &= \inf_{w^i\in\mathcal{W}_i,\,\forall i} \quad \mathcal{L}(\boldsymbol{w}, z) \\ &= \inf_{w^i\in\mathcal{W}_i,\,\forall i} \quad \sum_{i=1}^{N} \Big(f^i(w^i) + z^\top g^i(w^i)\Big) =: \sum_{i=1}^{N} q^i(z). \end{aligned} \tag{5.35}$$

Note that the right hand side in (5.34) is nonnegative by weak duality, and that $q(\bar{z})$ does not coincide with $-\infty$ for any $\bar{z} \in \mathbb{R}^m_{\geq 0}$ because each set $\mathcal{W}_i$ is compact. With this notation,

$$\sum_{i=1}^{N} f^i(\tilde{w}^i) - q(\bar{z}) \leq N\Big(\max_{j\in\{1,\dots,N\}} f^j(\tilde{w}^j) - \min_{j\in\{1,\dots,N\}} q^j(\bar{z})\Big).$$

Using this bound in (5.34), we conclude that $\mathcal{Z}^* \subseteq \mathcal{Z}_c$, with

$$\mathcal{Z}_c := \mathbb{R}^m_{\geq 0} \cap \bar{\mathcal{B}}\Big(0, \frac{N}{\gamma}\Big(\max_{j\in\{1,\dots,N\}} f^j(\tilde{w}^j) - \min_{j\in\{1,\dots,N\}} q^j(\bar{z})\Big)\Big). \tag{5.36}$$

Now we briefly describe the distributed strategy that the agents can use to bound the set $\mathcal{Z}_c$. The algorithm can be divided in three stages:

(i.a) Each agent finds the corresponding component $\tilde{w}^i$ of a Slater vector.

For instance, if $\mathcal{W}_i$ is *compact* (as is the case in Corollary 3.34), agent i can compute

$$\tilde{w}^i \in \operatorname*{arg\,min}_{w^i \in \mathcal{W}_i} g^i_l(w^i).$$

The resulting vector $(\tilde{w}^1, \dots, \tilde{w}^N)$ is a Slater vector, i.e., it belongs to the set

$$\{(w^1, \dots, w^N) \in \mathcal{W}_1 \times \dots \times \mathcal{W}_N : \\ g^1(w^1) + \dots + g^N(w^N) < 0\},$$

which is nonempty by the Strong Slater condition.

(i.b) Similarly, the agents compute the corresponding component $q^i(\bar{z})$ defined in (5.35). The common value $\bar{z} \in \mathbb{R}^m_{\geq 0}$ does not depend on the problem data and can be 0 or any other value agreed upon by the agents beforehand.

(ii) The agents find a lower bound for γ in (5.33) in two stages: first they use a distributed consensus algorithm and at the same time they estimate the fraction of agents that have a positive estimate. Second, when each agent is convinced that every other agent has a positive approximation, given by a precise termination condition that is satisfied in finite time, they broadcast their estimates to their neighbors to agree on the minimum value across the network.

Formally, each agent sets $y^i(0) := g^i(\tilde{w}^i) \in \mathbb{R}^m$ and $s_i(0) := \operatorname{sign}(y^i(0))$, and executes the following iterations

$$y^i(k+1) = y^i(k) + \sigma \sum_{j=1}^{N} \mathsf{a}_{ij,t}(y^j(k) - y^i(k)), \tag{5.37a}$$

$$s_i(k+1) = s_i(k) + \sigma \sum_{j=1}^{N} \mathsf{a}_{ij,t}\Big(\operatorname{sign}(y^j(k)) - \operatorname{sign}(y^i(k))\Big), \tag{5.37b}$$

until an iteration k_i^* such that $Ns_i(k_i^*) \leq -(N-1)$; see Lemma 3.36 below for the justification of this termination condition. Then, agent i re-initializes $y^i(0) = y^i(k^*)$ and iterates

$$y^i(k+1) = \min\{y^j(k) : j \in \mathcal{N}^{\text{out}}(i) \cup \{i\}\} \tag{5.38}$$

(where agent i does not need to know if a neighbor has re-initialized). The agents reach agreement about $\min_{i\in\{1,\dots,n\}} y^i(0) = \min_{i\in\{1,\dots,n\}} y^i(k^*)$ in a number of iterations no greater than $(N-1)B$ counted after $k^{**} := \max_{j\in\{1,\dots,N\}} k_j^*$ (which can be computed if each agent broadcasts once k_i^*). Therefore, the agents obtain the same lower bounds

$$\hat{y} := Ny^i(k^{**}) \leq \sum_{i=1}^{N} g^i(\tilde{w}^i),$$

$$\underline{\gamma} := \min_{l\in\{1,\dots,m\}} -\hat{y}_l \leq \gamma,$$

where the first lower bound is coordinate-wise.

(iii) The agents exactly agree on $\max_{j\in\{1,\dots,N\}} f^j(\tilde{w}^j)$ and $\min_{j\in\{1,\dots,N\}} q^j(\bar{z})$ using the finite-time algorithm analogous to (5.38).

In summary, the agents obtain the same upper bound

$$r := \frac{N}{\underline{\gamma}}\Big(\max_{j\in\{1,\dots,N\}} f^j(\tilde{w}^j) - \min_{j\in\{1,\dots,N\}} q^j(\bar{z})\Big),$$

which, according to (5.36), bounds the optimal dual set for the constrained optimization (5.32),

$$\mathcal{Z}^* \subseteq \mathcal{Z}_c \subseteq \bar{\mathcal{B}}(0,r).$$

To conclude, we justify the termination condition of step (ii).

Lemma 3.36. (Termination condition of step (ii)). *If each agent knows the size of the network N, then under the same assumptions on the communication graphs and the parameter σ as in Theorem 2.32, the termination time k_i^* is finite.*

Proof. Note that $y^i(0)$ is not guaranteed to be negative but, by construction of each $\{g^i(\tilde{w}^i)\}_{i=1}^N$ in step (i), it holds that the convergence point for (5.37a) is

$$\frac{1}{N}\sum_{i=1}^{N} y^i(0) = \frac{1}{N}\sum_{i=1}^{N} g^i(\tilde{w}^i) < 0. \tag{5.39}$$

This, together with the fact that Laplacian averaging preserves the convex hull of the initial conditions, it follows (inductively) that s_i decreases monotonically to -1. Thanks to the exponential convergence of (5.37a) to the point (5.39), it follows that there exists a finite time $k_i^* \in \mathbb{Z}_{\geq 1}$ such that $Ns_i(k_i^*) \leq -(N-1)$. This termination time is determined by the constant B of joint connectivity and the constant δ of nondegeneracy of the adjacency matrices. □

The complexity of the entire procedure corresponds to

- each agent computing the minimum of two convex functions;

- executing Laplacian average consensus until the agents' estimates fall within a centered interval around the average of the initial conditions; and
- running two agreement protocols on the minimum of quantities computed by the agents.

5.4 Simulation example

Here we simulate[1] the performance of the Consensus-based Saddle-Point (Sub-) Gradient algorithm (cf. Algorithm 1) in a network of $N = 50$ agents whose communication topology is given by a fixed connected small world graph [WS98] with maximum degree $d_{\max} = 4$. Under this coordination strategy, the 50 agents solve collaboratively the following instance of problem (5.2) with nonlinear convex constraints:

$$\begin{aligned} \min_{w_i \in [0,1]} & \sum_{i=1}^{50} c_i w_i \\ \text{s.t.} & \sum_{i=1}^{50} -d_i \log(1+w_i) \leq -b. \end{aligned} \tag{5.40}$$

Problems with constraints of this form arise, for instance, in wireless networks to ensure quality-of-service. For each $i \in \{1,\dots,50\}$, the constants c_i, d_i are taken randomly from a uniform distribution in $[0,1]$, and $b = 5$. We compute the solution to this problem, to use it as a benchmark, with the Optimization Toolbox using the solver *fmincon* with an *interior point* algorithm. Since the graph is connected, it follows that $B = 1$ in the definition of joint connectivity. Also, the constant of nondegeneracy is $\delta = 0.25$ and $\sigma_{\max}(\mathsf{L}) \approx 1.34$. With these values, we

[1] The Matlab code is available at `https://github.com/DavidMateosNunez/Consensus-based-Saddle-Point-Subgradient-Algorithm.git`.

derive from (5.18) the theoretically feasible consensus stepsize $\sigma = 0.2475$. For the projection step in (5.29d) of the C-SP-SG algorithm, the bound on the optimal dual set (5.36), using the Slater vector $\tilde{w} = \mathbb{1}_N$ and $\bar{z} = 0$, is

$$r = \frac{N \max_{j \in \{1,\ldots,N\}} c_j}{\log(2) \sum_{i=1}^{N} d_i - N/10} = 3.313.$$

For comparison, we have also simulated the Consensus-Based Dual Decomposition (CoBa-DD) algorithm proposed in [SJR16] using (and adapting to this problem) the code made available online by the authors[2]. (The bound for the optimal dual set used in the projection of the estimates of the multipliers is the same as above.) We should note that the analysis in [SJR16] only considers constant learning rates, which necessarily results in steady-state error in the algorithm convergence.

We have simulated the C-SP-SG and the CoBa-DD algorithms in two scenarios: under the Doubling Trick scheme of Assumption 2.31 (solid blue and magenta dash-dot lines, respectively), and under constant learning rates equal to 0.05 (darker grey) and 0.2 (lighter grey). Fig. 5.2 shows the saddle-point evaluation error for both algorithms. The saddle-point evaluation error of our algorithm is well within the theoretical bound established in Corollary 3.34, which for this optimization problem is approx. $1.18 \times 10^9/\sqrt{t}$. (This theoretical bound is overly conservative for connected digraphs because the ultimate bound for the disagreement C_u in (5.20), here $C_u \approx 3.6 \times 10^6$, is tailored for sequences of digraphs that are B-jointly connected instead of relying on the second smallest eigenvalue of the Laplacian of connected graphs.) Fig. 5.3 compares the network cost-error and the constraint satisfaction. We can observe that the C-SP-SG and the CoBa-DD [SJR16] algorithms have some characteristics in common:

[2] The Matlab code is available at `http://ens.ewi.tudelft.nl/~asimonetto/NumericalExample.zip`.

- They both benefit from using the Doubling Trick scheme.

- They approximate the solution, in all metrics of Fig. 5.2 and Fig. 5.3 at a similar rate. Although the factor in logarithmic scale of the C-SP-SG algorithm is larger, we note that this algorithm does not require the agents to solve a local optimization problem at each iteration for the updates of the primal variables, while both algorithms share the same communication complexity.

- The empirical convergence rate for the saddle-point evaluation error under the Doubling Trick scheme is of order $1/\sqrt{t}$ (logarithmic slope $-1/2$), while the empirical convergence rate for the cost error under constant learning rates is of order $1/t$ (logarithmic slope -1). This is consistent with the theoretical results here and in [SJR16] (wherein the theoretical bound concerns the practical convergence of the cost error using constant learning rates).

5.5 Discussion

We have proposed provably correct projected subgradient methods for saddle-point problems under explicit agreement constraints. We have shown that separable constrained optimization problems can be written in this form, where agreement plays a role in making the objectives (via agreement on a subset of the primal variables) as well as the constraints (via agreement on the dual variables) distributed. This approach enables the use of existing consensus-based ideas to tackle the algorithmic solution to these problems in a distributed fashion. We have illustrated how the general saddle-point formulation adopted in this paper encompasses optimization problems with separable and semidefinite constraints.

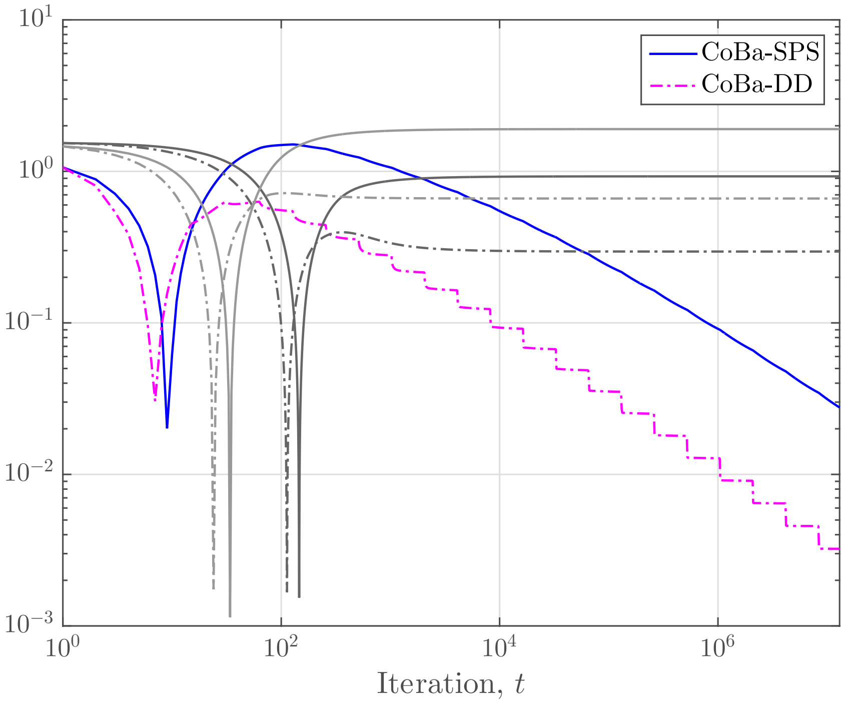
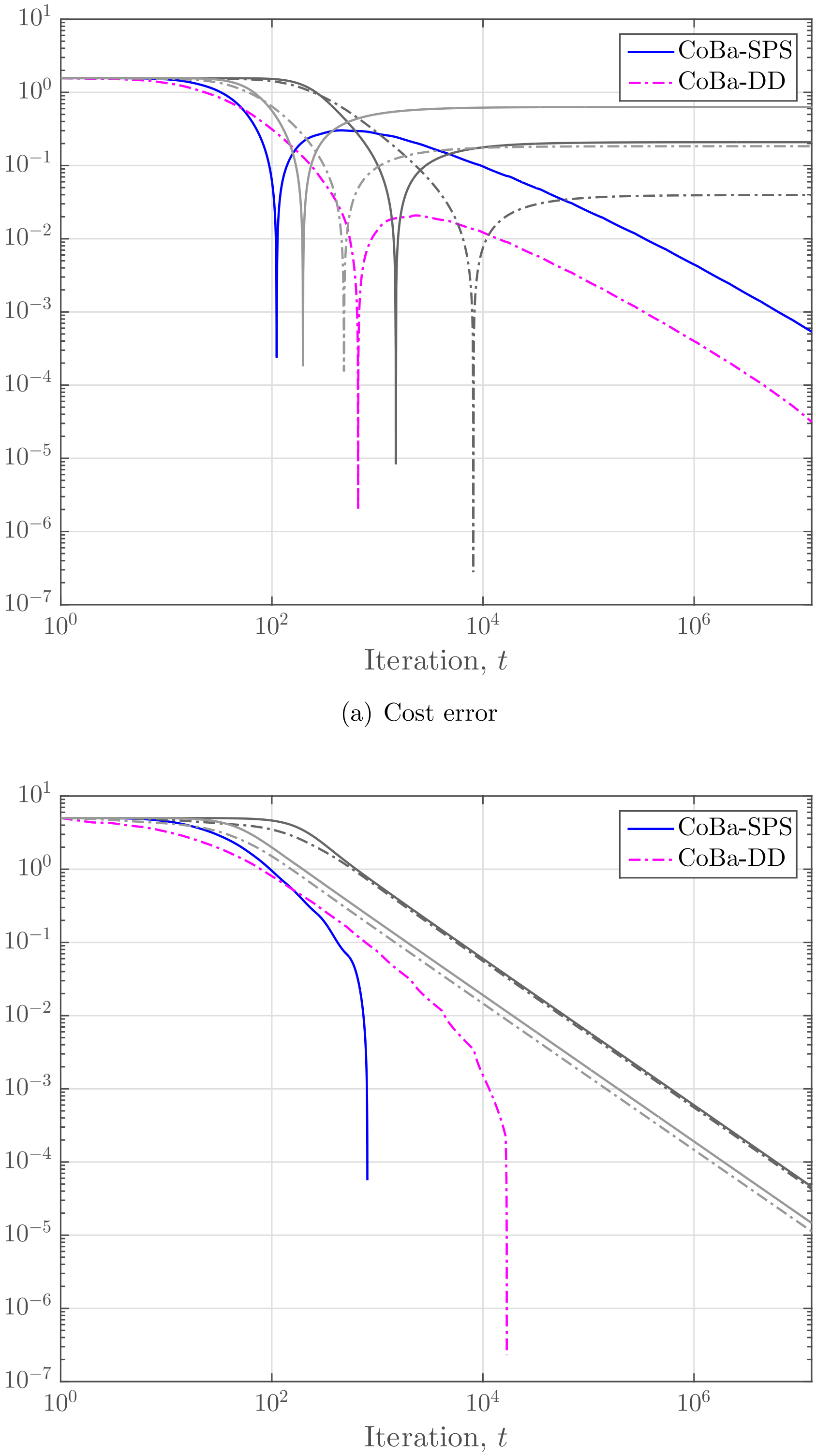


Figure 5.2: Saddle-point evaluation error $|\boldsymbol{\phi}(\boldsymbol{w}_t^{\text{av}}, \boldsymbol{z}_t^{\text{av}}) - \mathcal{L}(\boldsymbol{w}^*, z^*)|$. The lines in grey represent the same algorithms simulated with constant learning rates equal to 0.2 (lighter grey) and 0.05 (darker grey), respectively.

Acknowledgments

This chapter is taken from a work submitted by D. Mateos-Núñez and J. Cortés to *IEEE Transactions on Automatic Control,* with the title “Distributed saddle-point subgradient algorithms with Laplacian averaging.” A preliminary version will appear in the *Proceedings of the IEEE Conference on Decision and Control, Osaka, Japan, to be held in December 2015.* This research was supported by the US National Science Foundation (NSF) Award CMMI-1300272.



(a) Cost error

(b) Constraint satisfaction

Figure 5.3: Cost error and constraint satisfaction. For the same instantiations as in Fig. 5.2, (a) represents the evolution of the network cost error $|\sum_{i=1}^{N} c_i (w_t^i)^{\text{av}} - \sum_{i=1}^{N} c_i w_i^*|$, and (b) the evolution of the network constraint satisfaction $-\sum_{i=1}^{N} d_i \log(1+(w_t^i)^{\text{av}})+b$.

Chapter 6

Distributed optimization for multi-task learning via nuclear-norm approximation

In this chapter we finish our incursions in distributed optimization. Here we exploiting a variational characterization of the nuclear norm to extend the framework of distributed convex optimization to machine learning problems that focus on the sparsity of the aggregate solution. We propose two distributed dynamics that can be used for multi-task feature learning and recommender systems in scenarios with more tasks or users than features. Our first dynamics tackles a convex minimization on local decision variables subject to agreement on a set of local auxiliary matrices. Our second dynamics employs a saddle-point reformulation through Fenchel conjugation of quadratic forms, avoiding the computation of the inverse of the local matrices. We establish the correctness of both coordination algorithms using the general analytical framework of the previous chapter. Finally, we illustrate these results in a simulation example of low-rank matrix completion.

6.1 Optimization with nuclear norm regularization

We are interested in developing distributed coordination algorithms to solve the optimization problem

$$\min_{\substack{w_i \in \mathcal{W}, \\ i \in \{1,\ldots,N\}}} \sum_{i=1}^{N} f_i(w_i) + \gamma \Omega(W), \tag{6.1}$$

where $\mathcal{W} \subseteq \mathbb{R}^d$ is a closed convex set; the matrix $W \in \mathbb{R}^{d \times N}$ aggregates the vectors $\{w_i\}_{i=1}^N$ as columns, i.e., $W := [w_1|\ldots|w_N]$; each function $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex; $\gamma \in \mathbb{R}_{>0}$ is a design parameter; and $\Omega : \mathbb{R}^{d \times N} \to$ is a joint regularizer to promote solutions with low rank or other sparsity patterns. We next motivate the distributed optimization problem with nuclear-norm regularization in two scenarios.

6.1.1 Multi-task feature learning

In data-driven optimization problems each function f_i often codifies the loss incurred by the vector of weighting parameters w_i with respect to a set of n_i data points $\{p_j, y_j\}_{j=1}^{n_i}$. As such, this loss can be called *residual* or *margin*, depending on whether we are considering regression or classification problems. The work [AEP08] exploits the relation (2.10) as follows. For a given $W \in \mathbb{R}^{d \times N}$, the following regularizer is used,

$$\begin{aligned} \Omega(W) &= \min_{\substack{U \in \mathbb{O}^d,\, A \in \mathbb{R}^{d \times N} \\ W = UA}} \|A^\top\|_{2,1} \\ &= \min_{U \in \mathbb{O}^d} \|W^\top U\|_{2,1} = \|W\|_*. \end{aligned}$$

This minimization promotes a *dictionary* matrix U of orthonormal columns such that the columns of W are sparse linear combinations of them. The latter is achieved through $\|A^\top\|_{2,1}$, which 'favors' rows of small size because the one-norm is the convex surrogate of the zero-norm, or number of nonzero elements. This offers an interesting perspective on minimization problems that are convex on the *product* UA, with $U \in \mathbb{O}^d$, and have a penalty term $\|A^\top\|_{2,1}$. As pointed in [AEP08], the above characterization enables a convex reformulation on the matrix variable $W = UA$.

6.1.2 Matrix completion for recommender systems

The estimation of a low-rank matrix from a set of entries, or matrix completion, see, e.g., [MHT10], also fits naturally in the framework of (6.1) with nuclear-norm regularization. This is because the nuclear norm is the convex surrogate of the rank function [Faz02]. Let $Z \in \mathbb{R}^{d\times N}$ be a low-rank matrix of unknown rank for which only a few entries per column are known. The goal is then to determine a matrix W that minimizes the Frobenius norm across the revealed entries while keeping small the nuclear norm,

$$\min_{\substack{w_i \in \mathcal{W},\\ i\in\{1,\ldots,N\}}} \sum_{i=1}^{N} \sum_{j\in\Upsilon_i} (W_{ji} - Z_{ji})^2 + \gamma \|W\|_* \tag{6.2}$$

where, for each $i \in \{1,\ldots,N\}$,

$$\Upsilon_i := \{j \in \{1,\ldots,d\} : Z_{ji} \text{ is a revealed entry of } Z\}.$$

6.1.3 A case for distributed optimization

The optimization problem (6.1) can be formulated as a convex and separable minimization when the joint regularizer is $\|\cdot\|_*$ or $\|\cdot\|_*^2$ using the characterizations (2.8a) or (2.8b). Assuming that a minimum exists, we can write

$$\begin{aligned}
&\min_{W\in\mathbb{R}^{d\times N}} \sum_{i=1}^{N} f_i(w_i) + \gamma\|W\|_*^2 \\
=& \min_{\substack{W\in\mathbb{R}^{d\times N} \\ D\in\mathbb{S}^d_{\succeq 0},\, \mathrm{trace}(D)\leq 1 \\ w_i\in\mathcal{C}(D), \forall i}} \sum_{i=1}^{N} f_i(w_i) + \gamma\sum_{i=1}^{N} w_i^\top D^\dagger w_i\,. \\
=& \min_{\substack{w_i\in\mathcal{W}, \forall i \\ D_i\in\mathbb{S}^d_{\succeq 0},\, \mathrm{trace}(D_i)\leq 1,\, \forall i \\ w_i\in\mathcal{C}(D_i),\, \forall i \\ D_i=D_j,\, \forall i,j}} \sum_{i=1}^{N} f_i(w_i) + \gamma\sum_{i=1}^{N} w_i^\top D_i^\dagger w_i\,,
\end{aligned} \tag{6.3}$$

and similarly for $\Omega(W) = 2\|W\|_*$ replacing the constraint $\mathrm{trace}(D) \leq 1$ by the penalty functions $\gamma\sum_{i=1}^{N}\frac{1}{N}\mathrm{trace}(D_i)$. When $d \ll N$, it is reasonable to design distributed strategies that use local gradient descent and consensus to solve this problem because the objective can be split across a network of agents, and the only coupling constraint is the agreement on the matrix arguments, $D_i = D_j$ for each i, j, whose dimensions do not grow with the network size. The condition $d \ll N$ in multi-task feature learning implies that there are far less features than tasks or users (for instance, there are less diseases or symptoms than people). The same observation applies to matrix completion in collaborative filtering where the rows represent features and the columns represent users.

However, the design of distributed strategies to solve (6.3) raises the following challenges,

(i) The constraint set $\{w \in \mathbb{R}^d, D \in \mathbb{S}^d_{\succeq 0} : w \in \mathcal{C}(D)\}$ is convex but not closed, which is a difficulty when designing a projection among the local variables.

Note that for any fixed matrix D_i, one could project w_i onto $\mathcal{C}(D_i)$ by computing $D_i D_i^\dagger w$, but this projection is state-dependent.

(ii) The computation of $D_i^\dagger$ is a concern because D_i might be rank deficient and the pseudoinverse might be *discontinuous* when the rank of D_i changes.

We avoid these difficulties by enforcing the solution to be within a margin of the boundary of the positive semidefinite cone. This is achieved by considering an approximate regularization that we introduce in Section 6.2.1. Our first dynamics solves the *nuclear-norm regularization as a separable minimization with agreement constraint.* Even with (ii) addressed, an additional challenge involves the efficient computation of the inverse:

- Iterative algorithms involving the computation of D^{-1} are computationally expensive and potentially lead to numerical instabilities.

We eliminate the necessity of computing D^{-1} altogether in Section 6.2.2 by transforming the convex minimization into a saddle-point problem. This transformation is general and does not require the approximate treatment of the nuclear norm regularization in Section 6.2.1. Our second dynamics solves the *nuclear-norm regularization as a separable min-max problem with agreement constraint.*

6.2 Distributed coordination algorithms

Here we address the three challenges outlined in Section 6.1 to solve the optimization problem (6.3). In the forthcoming discussion, we present two reformulations of this problem and two distributed coordination algorithms to solve them.

6.2.1 Nuclear norm approximate regularization

In relation to the first two challenges outlined above, note that the optimal values D_1^* and D_2^* in (2.9) for the variational characterizations of $\|\cdot\|_*$ and $\|\cdot\|_*^2$ are in general positive semidefinite. To enforce these optimal values to be in the interior of the positive semidefinite cone, following the technique in [AEP08, Sec. 4], we consider an approximate problem by introducing in (6.3) the barrier function $\epsilon\,\mathrm{trace}(D^\dagger)$ for some $\epsilon \in \mathbb{R}_{>0}$. We next justify how the optimizer of the approximate problem, which depends on ϵ, is farther than some *margin* from the boundary of $\mathbb{S}^d_{\succ 0}$ (in turn, this fact allows to insert in our optimization problem a dummy constraint of the form $D \succeq c\mathrm{I}$, where c is what we refer to as the margin). For $\Omega_\epsilon(W) = 2\|[W|\sqrt{\epsilon}\mathrm{I}_d]\|_*$, this is easy to see because, in view of (2.9),

$$D_{1,\epsilon}^* := \sqrt{WW^\top + \epsilon \mathrm{I}_d} \succeq \sqrt{\epsilon}\mathrm{I}_d.$$

For $\Omega_\epsilon(W) = \|[W|\sqrt{\epsilon}\mathrm{I}_d]\|_*^2$, we need more care and we offer next a result using the notation for the reduced spectraplex defined in Section 2.4.

Lemma 2.37. (Dummy constraint for ϵ-approximate regularization under $\Omega(W) = \|W\|_*^2$). *Let $W \in \mathbb{R}^{d\times N}$ be any matrix whose columns have two-norm bounded by r_w. Then, the optimizer of*

$$\min_{\substack{D\in\mathbb{S}^d_{\succeq 0},\,\mathrm{trace}(D)\leq 1,\\ \mathcal{C}(W)\subseteq\mathcal{C}(D)}} \mathrm{trace}\left(D^\dagger(WW^\top + \epsilon\mathrm{I})\right), \tag{6.4}$$

(whose optimal value is $\|[W\,|\,\sqrt{\epsilon}\mathrm{I}_d]\|_*^2$), also minimizes

$$\min_{D\in\Delta(c_\epsilon)} \mathrm{trace}\left(D^\dagger(WW^\top + \epsilon\mathrm{I})\right),$$

where the margin c_ϵ of the reduced spectraplex $\mathit{\Delta}(c_\epsilon)$ is

$$c_\epsilon := \frac{\sqrt{\epsilon}}{\sqrt{d}\sqrt{N r_w^2 + \epsilon\, d}}\,. \tag{6.5}$$

Furthermore, c_ϵ in (6.5) *satisfies $c_\epsilon \leq 1/d$ for any $\epsilon, r_w \in \mathbb{R}_{>0}$. Hence, $\mathit{\Delta}(c_\epsilon)$ is nonempty for any $\epsilon, r_w \in \mathbb{R}_{>0}$.*

Proof. In view of (2.8b) and (2.9), the optimizer of (6.4) is

$$D_{2,\epsilon}^* := \frac{\sqrt{WW^\top + \epsilon \mathrm{I}_d}}{\mathrm{trace}(\sqrt{WW^\top + \epsilon \mathrm{I}_d})}\,. \tag{6.6}$$

Using the inequality for the nuclear and the Frobenius norm (2.3), we have

$$\begin{aligned} D_{2,\epsilon}^* &= \frac{\sqrt{WW^\top + \epsilon \mathrm{I}_d}}{\|[W \,|\, \sqrt{\epsilon}\mathrm{I}_d]\|_*} \geq \frac{\sqrt{\epsilon}}{\sqrt{d}\,\|[W \,|\, \sqrt{\epsilon}\mathrm{I}_d]\|_{\mathcal{F}}}\mathrm{I}_d \\ &= \frac{\sqrt{\epsilon}}{\sqrt{d}\sqrt{\|W\|_{\mathcal{F}}^2 + \|\sqrt{\epsilon}\mathrm{I}_d\|_{\mathcal{F}}^2}}\mathrm{I}_d \geq \frac{\sqrt{\epsilon}}{\sqrt{d}\sqrt{N r_w^2 + \epsilon\, d}}\mathrm{I}_d\,, \end{aligned}$$

where in the last two equations we have used the fact that the square Frobenius norm is the sum of square two-norms of the columns. The second fact is trivial. □ □

As a result, when we add the barrier terms $\sum_{i=1}^N \frac{\epsilon}{N}\,\mathrm{trace}(D_i^\dagger)$ to the optimization in (6.3), the constraints $D_i \in \mathbb{S}_{\succeq 0}^d$ and $w_i \in \mathcal{C}(D_i)$ can be replaced by $D_i \succeq c_\epsilon \mathrm{I}_d$. Hence, the variational characterization of $\|[W \,|\, \sqrt{\epsilon}\mathrm{I}_d]\|_*^2$ can be written over the compact domain $\mathit{\Delta}(c_\epsilon)$. Alternatively, in the case of $2\|[W \,|\, \sqrt{\epsilon}\mathrm{I}_d]\|_*$, we saw above that we can use the constraint $D_i \succeq \sqrt{\epsilon}\mathrm{I}_d$ to achieve the same effect. However, because the trace constraint is now absent, we construct a compact domain containing the optimal value $D_{1,\epsilon}^*$ by introducing one more dummy constraint $\|D_i\|_{\mathcal{F}} \leq r_\epsilon$, with

$$r_\epsilon := \sqrt{N} r_w + \sqrt{\epsilon d}. \tag{6.7}$$

This, together with the constraint $D_i \succeq \sqrt{\epsilon} \mathrm{I}_d$, yields the compact domain given by the reduced ice-cream $\mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$. The derivation is similar to the proof of Lemma 2.37; here we compute an upper bound as opposed to a lower bound. In both cases, we use the fact that the columns of W are contained in the ball $\bar{\mathcal{B}}(0, r_w) \subseteq \mathbb{R}^d$.

The following results summarizes our discussion above.

Corollary 2.38. (Separable minimization with agreement constraint). *Let $\mathcal{W} \in \bar{\mathcal{B}}(0, r_w)$ and define c_ϵ as in* (6.5). *Then*

$$\min_{W \in \mathbb{R}^{d \times N}} \sum_{i=1}^{N} f_i(w_i) + \gamma \Omega_\epsilon(W), \tag{6.8}$$

with $\Omega_\epsilon(W) = \| [W \,|\, \sqrt{\epsilon} \mathrm{I}_d] \|_*^2$ is equal to

$$\min_{\substack{w_i \in \mathcal{W}, \forall i, \\ D_i \in \Delta(c_\epsilon),\, \forall i, \\ D_i = D_j,\, \forall i,j}} \sum_{i=1}^{N} f_i(w_i) + \gamma \sum_{i=1}^{N} \Big(w_i^\top D_i^{-1} w_i + \tfrac{\epsilon}{N} \operatorname{trace}(D_i^{-1}) \Big). \tag{6.9}$$

The analogous result is valid for $\Omega_\epsilon(W) = 2 \| [W \,|\, \sqrt{\epsilon} \mathrm{I}_d] \|_*$ replacing $\Delta(c_\epsilon)$ by $\mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$ and including the penalty functions $\gamma \sum_{i=1}^{N} \frac{1}{N} \operatorname{trace}(D_i)$.

In both cases of Corollary 2.38, Weierstrass' Theorem guarantees that the minimum is reached since we are minimizing a continuous function over a compact set. This leads to our first candidate dynamics.

Distributed subgradient dynamics for nuclear optimization.

Our first coordination algorithm for the distributed optimization with nuclear norm (6.8) is a subgradient algorithm with proportional feedback on the disagreement on the matrix variables:

$$\begin{aligned}
\hat{w}_i(k+1) &= w_i(k) - \eta_k\Big(g_i(k) + 2\gamma D_i(k)^{-1} w_i(k)\Big),\\
\hat{D}_i(k+1) &= D_i(k) - \eta_k \gamma\Big(- D_i^{-1}(k) w_i(k) w_i(k)^\top D_i^{-1}(k)\\
&\qquad\qquad + \tfrac{\alpha}{N}\mathrm{I}_d - \tfrac{\epsilon}{N} D_i^{-2}(k)\Big) + \sigma \sum_{j=1}^{N} \mathsf{a}_{ij,t}(D_j(k) - D_i(k)),\\
w_i(k+1) &= \mathcal{P}_{\mathcal{W}}(\hat{w}_i(k+1)),\\
D_i(k+1) &= \mathcal{P}_{\mathcal{D}}(\hat{D}_i(k+1)),
\end{aligned} \tag{6.10}$$

where $g_i(k) \in \partial f_i(w_i(k))$, for each $i \in \{1,\dots,N\}$, and $\mathcal{P}_{\mathcal{W}}(\cdot)$ and $\mathcal{P}_{\mathcal{D}}(\cdot)$ denote the projections onto the compact convex sets $\mathcal{W}$ and $\mathcal{D}$. This notation allows us to consider both approximate regularizers: for the case $2\|[W \,|\, \sqrt{\epsilon}\mathrm{I}_d]\|_*$, the trace acts as a penalty, i.e., $\alpha = 1$, and the domain is $\mathcal{D} = \mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$; for the case $\|[W \,|\, \sqrt{\epsilon}\mathrm{I}_d]\|_*^2$, the trace acts as a constraint, i.e., $\alpha = 0$, and $\mathcal{D} = \mathit{\Delta}(c_\epsilon)$.

Remark 2.39. (Implementation of the orthogonal projections). *The fact that projections need to be orthogonal raises additional conceptual challenges. We make here the following observations:*

- *If we represent the set $\mathfrak{D}(c,r)$ as the intersection $\{D \in \mathbb{S}^d : D \succeq c\mathrm{I}_d\} \cap \{D \in \mathbb{R}^{d\times d} : \|D\|_{\mathcal{F}} \le r\}$, then a candidate projection onto $\mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$ can be performed by the composition of a matrix square root,*

$$\mathcal{P}_{\{D \in \mathbb{S}^d : D \succeq c\mathrm{I}_d\}}(D) = \sqrt{(D - c\mathrm{I}_d)^2} + c\mathrm{I}_d,$$

and, if $\|D\|_{\mathcal{F}} \geq r_\epsilon$, the normalization step

$$\mathcal{P}_{\bar{\mathcal{B}}_{\mathcal{F}}(0,r_\epsilon)}(D) := \frac{r}{\|D\|_{\mathcal{F}}} D.$$

(The matrix square root can be computed efficiently using Newton method [Hig86].) Each of these functions is an orthogonal projection onto the corresponding set. However, each of them is guaranteed to preserve the set of the other only if $c = 0$. This implies that, in general, they need to be applied iteratively.

- *Similar considerations apply to the projection onto $\Delta(c)$, but in contrast with $\mathfrak{D}(c,r)$, the orthogonal projections onto $\{D \in \mathbb{S}^d : D \succeq c\mathrm{I}_d\}$, and $\{D \in \mathbb{R}^{d\times d} : \mathrm{trace}(D) \leq 1\}$ do not preserve each other's corresponding sets even when $c = 0$.* •

6.2.2 Separable saddle-point formulation

In the previous section we have written the optimization (6.8) with approximate nuclear norm regularization as a separable convex optimization with an agreement constraint on auxiliary local matrices. Here we derive an equivalent min-max problem that is also separable and has the advantage of enabling iterative distributed strategies that avoid the computation of the inverse of the local matrices. To achieve this aim, the next result expresses the quadratic forms $w^\top D^\dagger w$ and $\mathrm{trace}(D^\dagger) = \sum_{j=1}^d e_j^\top D^\dagger e_j$ as the maximum of concave functions in additional auxiliary variables. We write these expressions using Fenchel conjugacy of quadratic forms, and in doing this, we avoid the need to compute the pseudoinverse of D.

Proposition 2.40. (Min-max formulation via Fenchel conjugacy). *For $i \in \{1, \ldots, N\}$*

and $\alpha \in \mathbb{R}_{\geq 0}$, let $F_i : \mathcal{W} \times \mathbb{R}^{d\times d} \times \mathbb{R}^d \times \mathbb{R}^{d\times d} \to \mathbb{R}$ be defined by

$$\begin{aligned} F_i(w,D,x,Y) := {} & f_i(w) + \gamma \operatorname{trace}\Big(D(-xx^\top - \tfrac{\epsilon}{N} YY^\top)\Big) \\ & - 2\gamma w^\top x - 2\gamma \frac{\epsilon}{N} \operatorname{trace}(Y) + \frac{\alpha}{N} \operatorname{trace}(D). \end{aligned} \tag{6.11}$$

Then, the following two optimizations are equivalent

$$\begin{aligned} & \min_{\substack{D \in \mathbb{S}^d_{\succeq 0}, \\ w \in \mathcal{W} \cap \mathcal{C}(D)}} \Big\{ f_i(w) + \gamma\Big(w^\top D^\dagger w + \tfrac{\epsilon}{N} \operatorname{trace}(D^\dagger) + \tfrac{\alpha}{N} \operatorname{trace}(D)\Big)\Big\} \\ = {} & \min_{w \in \mathcal{W},\, D \in \mathbb{R}^{d\times d}} \ \sup_{x \in \mathbb{R}^d, Y \in \mathbb{R}^{d\times d}} F_i(w,D,x,Y). \end{aligned} \tag{6.12}$$

Moreover, the minimization on the right does not change with the addition of the constraints $D \in \mathbb{S}^d_{\succeq 0}$ and $w \in \mathcal{C}(D)$ (which allows to replace the operator sup *by* max*).*

Proof. For any $w \in \mathcal{W}$ and $D \in \mathbb{R}^{d\times d}$, it holds that

$$\sup_{x \in \mathbb{R}^d} -(x^\top D x + 2 w^\top x) = \begin{cases} w^\top D^\dagger w & \text{if } D \in \mathbb{S}^d_{\succeq 0}, w \in \mathcal{C}(D), \\ \infty & \text{otherwise.} \end{cases}$$

This transformation is the same as Fenchel conjugacy up to a factor, see, e.g., [BV09, P. 649]. When $D \in \mathbb{S}^d_{\succeq 0}$ and $w \in \mathcal{C}(D)$, the maximizer is $x^* = -D^\dagger w$. Since

$$w^\top D^\dagger w + \frac{\epsilon}{N} \operatorname{trace}(D^\dagger) = w^\top D^\dagger w + \frac{\epsilon}{N} \sum_{j=1}^d e_j^\top D^\dagger e_j\,,$$

this term can be substituted in the minimization (6.12) by

$$\begin{aligned}
&\sup_{x\in\mathbb{R}^d} -(x^\top Dx + 2w^\top x) + \frac{\epsilon}{N}\sum_{j=1}^{d}\sup_{y^j\in\mathbb{R}^d} -({y^j}^\top Dy^j + 2e_j^\top y^j)\\
=&\sup_{x\in\mathbb{R}^d, Y\in\mathbb{R}^{d\times d}} \Big(-x^\top Dx - 2w^\top x\\
&\qquad\qquad - \frac{\epsilon}{N}\operatorname{trace}(DYY^\top) - \frac{2\epsilon}{N}\operatorname{trace}(Y)\Big),
\end{aligned}$$

where $Y = [y_1|\cdots|y_d] \in \mathbb{R}^{d\times d}$, completing the proof. □ □

The function $w^\top D^\dagger w$ is jointly convex in the convex domain $\{w \in \mathcal{W}, D \in \mathbb{S}^d_{\succeq 0} : w \in \mathcal{C}(D)\}$ because it is a point-wise maximum of linear functions indexed by x. (The function is also proper but not closed because the domain is not closed). The same considerations apply adding the constraint $\operatorname{trace}(D) \leq 1$. We are now ready to establish the main equivalence between optimization problems.

Corollary 2.41. (Separable min-max problem with agreement constraint).. *The optimization* (6.8) *with* $\Omega_\epsilon(W) = \|[W\,|\,\sqrt{\epsilon}\mathrm{I}_d]\|_*^2$ *is equivalent to*

$$\min_{\substack{w_i\in\mathcal{W},\, D_i\in\mathbb{R}^{d\times d},\\ \operatorname{trace}(D_i)\leq 1,\,\forall i,\\ D_i=D_j\;\forall i,j}} \;\sup_{\substack{x_i\in\mathbb{R}^d,\forall i\\ Y_i\in\mathbb{R}^{d\times d},\forall i}} \;\sum_{i=1}^{N} F_i(w_i, D_i, x_i, Y_i)\,, \tag{6.13}$$

without the penalty on the trace in F_i *(i.e.,* $\alpha = 0$*) for each* $i \in \{1,\ldots,N\}$*. As long as* c_ϵ *is given by* (6.5) *and* $\mathcal{W} \subseteq \bar{\mathcal{B}}(0, r_w)$*, the constraints* $D_i \in \mathit{\Delta}(c_\epsilon)$ *are not necessary, but including them allows to replace the operator* sup *by* max*. An analogous result holds for* $\Omega_\epsilon(W) = 2\|[W\,|\,\sqrt{\epsilon}\mathrm{I}_d]\|_*$ *when, instead of the trace constraints, one has the penalty terms* $\sum_{i=1}^N \frac{1}{N}\operatorname{trace}(D_i)$ *(i.e.,* $\alpha = 1$*). In this case, as long as* r_ϵ *is given by* (6.7) *and* $\mathcal{W} \subseteq \bar{\mathcal{B}}(0, r_w)$*, the constraints* $D_i \in \mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$ *are not necessary.*

Proof. The proof for both cases is analogous so, for brevity, we only present it for $\Omega_\epsilon(W) = \|[\,W\,|\,\sqrt{\epsilon}\mathrm{I}_d\,]\|_*^2$. Introducing the constraint $\mathrm{trace}(D) \leq 1$ on both sides of (6.12), and adding over $i = 1, \ldots, N$ under the agreement constraint $D_i = D_j$ for each $1 \leq i,\, j \leq N$, we get

$$\begin{aligned}
&\min_{\substack{w_i\in\mathcal{W},\forall i,\\ D_i\in\mathbb{S}^d_{\succeq 0},\,\mathrm{trace}(D_i)\leq 1,\,\forall i,\\ w_i\in\mathcal{C}(D_i),\,\forall i,\\ D_i=D_j,\,\forall i,j}} \Bigg(\sum_{i=1}^N f_i(w_i) \\
&\qquad\qquad + \gamma \sum_{i=1}^N \Big(w_i^\top D_i^\dagger w_i + \tfrac{\epsilon}{N}\,\mathrm{trace}(D_i^\dagger)\Big)\Bigg) \\
&= \min_{\substack{w_i\in\mathcal{W},\,\forall i,\\ D_i\in\mathbb{R}^{d\times d},\,\mathrm{trace}(D_i)\leq 1,\,\forall i,\\ D_i=D_j\ \forall i,j}} \ \max_{\substack{x_i\in\mathbb{R}^d,\forall i\\ Y_i\in\mathbb{R}^{d\times d},\forall i}} \ \sum_{i=1}^N F_i(w_i, D_i, x_i, Y_i).
\end{aligned}$$

On the right hand side the solution does not change if we introduce the constraints $D_i \in \mathbb{S}^d_{\succeq 0}$, $w_i \in \mathcal{C}(D_i)$ thanks to Proposition 2.40. Moreover, the solution remains also the same by substituting both constraints on either side by $D \in \mathit{\Delta}(c_\epsilon)$ thanks to Lemma 2.37 as long as c_ϵ is given by (6.5) and $\mathcal{W} \in \bar{\mathcal{B}}(0, r_w)$. □ □

The next result establishes the existence of a saddle-point for the convex-concave formulation of the ϵ-approximate minimization. For convenience, define $F : \mathcal{W}^N \times \mathit{\Delta}(c_\epsilon) \times (\mathbb{R}^d)^N \times (\mathbb{R}^{d\times d})^N \to \mathbb{R}$ as

$$F(\boldsymbol{w}, D, \boldsymbol{x}, \boldsymbol{Y}) := \sum_{i=1}^N F_i(w_i, D, x_i, Y_i), \tag{6.14}$$

where $\boldsymbol{w} := (w_1, \ldots w_N)$, $\boldsymbol{x} := (x_1, \ldots x_N)$, $\boldsymbol{Y} := (Y_1, \ldots Y_N)$.

Proposition 2.42. (Existence of saddle points). *For $\mathcal{W} \subseteq \bar{\mathcal{B}}(0, r_w)$ and $\mathcal{D}$ equal to either $\mathit{\Delta}(c_\epsilon)$ or $\mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$, the set of saddle points of F on $\mathcal{W}^N \times \mathcal{D} \times (\mathbb{R}^d)^N \times$*

$(\mathbb{R}^{d\times d})^N$ is nonempty and compact, and, as a consequence,

$$\max_{x_i\in\mathbb{R}^d,\, Y_i\in\mathbb{R}^{d\times d},\,\forall i}\ \min_{w_i\in\mathcal{W},\,\forall i,\, D\in\varDelta(c_\epsilon)} \sum_{i=1}^{N} F_i(w_i, D, x_i, Y_i)$$
$$= \min_{w_i\in\mathcal{W},\,\forall i,\, D\in\varDelta(c_\epsilon)}\ \max_{x_i\in\mathbb{R}^d,\, Y_i\in\mathbb{R}^{d\times d},\,\forall i} \sum_{i=1}^{N} F_i(w_i, D, x_i, Y_i).$$

(The agreement constraints $D_i = D_j$ for all $i, j \in \{1,\dots,N\}$ are written implicitly because the existence of saddle-points is established within those agreement constraints.)

Proof. We use the Saddle-Point Theorem [BNO03, Thm. 2.6.9, p. 150]. First we need to verify the hypotheses of [BNO03, Assumption 2.6.1, p. 144]. For each $(\boldsymbol{x}, Y) \in (\mathbb{R}^d)^N \times (\mathbb{R}^{d\times d})^N$, we introduce the function $t_{\boldsymbol{x},\boldsymbol{Y}} : \mathcal{W}^N \times \mathcal{D} \to \mathbb{R}\cup\{\infty\}$ defined by

$$t_{\boldsymbol{x},\boldsymbol{Y}}(\boldsymbol{w}, D) := \begin{cases} F(\boldsymbol{w}, D, \boldsymbol{x}, \boldsymbol{Y}) & \text{if } (\boldsymbol{w}, D) \in \mathcal{W}^N \times \mathcal{D}, \\ \infty & \text{if } (\boldsymbol{w}, D) \notin \mathcal{W}^N \times \mathcal{D}, \end{cases}$$

and for each $(\boldsymbol{w}, D) \in \mathcal{W}^N \times \mathcal{D}$, we introduce the function $r_{\boldsymbol{w},D} : (\mathbb{R}^d)^N \times (\mathbb{R}^{d\times d})^N \to \mathbb{R}\cup\infty$ given by

$$r_{\boldsymbol{w},D}(\boldsymbol{x}, \boldsymbol{Y}) := -F(\boldsymbol{w}, D, \boldsymbol{x}, \boldsymbol{Y}).$$

We observe that for each $(\boldsymbol{x}, \boldsymbol{Y}) \in (\mathbb{R}^d)^N \times (\mathbb{R}^{d\times d})^N$, the function $t_{\boldsymbol{x},\boldsymbol{Y}}$ is closed and convex, and similarly, for each $(\boldsymbol{w}, D) \in \mathcal{W}^N \times \mathcal{D}$, the function $r_{\boldsymbol{w},D}$ is also closed and convex. Hence the aforementioned assumption holds. Going back to [BNO03, Thm. 2.6.9, p. 150], we verify that the set $\mathcal{W}^N \times \mathcal{D}$ is compact, and the last step is to show that there exists $(\bar{\boldsymbol{w}}, \bar{D}) \in \mathcal{W}^N \times \mathcal{D}$ and $\alpha \in \mathbb{R}$ such that the level set

$$\{(\boldsymbol{x}, \boldsymbol{Y}) \in (\mathbb{R}^d)^N \times (\mathbb{R}^{d\times d})^N \,:\, F(\bar{\boldsymbol{w}}, \bar{D}, \boldsymbol{x}, \boldsymbol{Y}) \geq \alpha\} \tag{6.15}$$

is nonempty and compact. We first prove that is closed, then that is bounded, and finally that is nonempty for some $\alpha \in \mathbb{R}$. The level set is closed because F is continuous. To show that is bounded, we will prove that it is contained in a ball. For this, we use [Ber05, Prop 8.4.13] on the trace of a product of a symmetric matrix and a positive semidefinite matrix, which can be bounded as $\operatorname{trace}(\bar{D}YY^\top) \leq \lambda_{\max}(\bar{D}) \operatorname{trace}(YY^\top) \leq \operatorname{trace}(YY^\top) = \|Y\|_{\mathcal{F}}^2$ because $\operatorname{trace}(\bar{D}) \leq 1$ and $\bar{D} \succeq 0$, and similarly we bound $x^\top \bar{D} x \leq \|x\|_2^2$. Therefore,

$$\begin{aligned}
&F(\bar{\boldsymbol{w}}, \bar{D}, \boldsymbol{x}, \boldsymbol{Y}) - \sum_{i=1}^N f_i(\bar{w}_i) \\
=&\gamma \sum_{i=1}^N \Big(\operatorname{trace}\big(\bar{D}(-x_i x_i^\top - \tfrac{\epsilon}{N} Y_i Y_i^\top + \tfrac{\alpha}{N})\big) \\
&- 2\bar{w}_i^\top x_i - 2\tfrac{\epsilon}{N} \operatorname{trace}(Y_i)\Big) \\
\geq& -\gamma \sum_{i=1}^N \Big(\|x_i\|_2^2 + \tfrac{\epsilon}{N}\|Y_i\|_{\mathcal{F}}^2 + \tfrac{\alpha d}{N} + 2\bar{w}_i^\top x_i + 2\tfrac{\epsilon}{N}\operatorname{trace}(Y_i)\Big) \\
\geq& -\gamma \sum_{i=1}^N \Big(2\|x_i\|_2^2 + r_w^2 + 2\tfrac{\epsilon}{N}\|Y_i\|_{\mathcal{F}}^2 + (\alpha+\epsilon)\tfrac{d}{N}\Big),
\end{aligned}$$

where in the last inequality we have used that $2\operatorname{trace}(Y) \leq d + \operatorname{trace}(YY^\top)$, which follows because

$$\begin{aligned}
0 \leq \|\mathrm{I}_d - Y\|_{\mathcal{F}}^2 &= \operatorname{trace}(\mathrm{I} - Y)(\mathrm{I} - Y^\top) \\
&= \operatorname{trace}(\mathrm{I}_d + YY^\top - Y - Y^\top) = d + \|Y\|_{\mathcal{F}}^2 - 2\operatorname{trace}(Y),
\end{aligned}$$

and, similarly, $2\bar{w}_i^\top x \leq \|\bar{w}_i\|_2^2 + \|x\|_2^2 \leq r_w^2 + \|x\|_2^2$. Therefore, the level set in (6.15) is contained in the set

$$\Big\{(\boldsymbol{x},\boldsymbol{Y})\in(\mathbb{R}^d)^N\times(\mathbb{R}^{d\times d})^N \ : \\ c^*-\gamma\sum_{i=1}^N\Big(2\|x_i\|_2^2+r_w^2+2\tfrac{\epsilon}{N}\|Y_i\|_{\mathcal{F}}^2+\tfrac{\epsilon}{N}d\Big)\geq\alpha\Big\}, \tag{6.16}$$

where $c^*:=\min_{\boldsymbol{w}\in\mathcal{W}^N}\sum_{i=1}^N f_i(w_i)$. Note that c^* is well defined thanks to Weierstrass' Theorem. The boundedness of the set (6.15) then follows because the super set (6.16) is bounded for any $\alpha\in\mathbb{R}$ (although it may be empty). Finally, we need to find some α for which the level set in (6.15) is nonempty. For this, note that the point $(\boldsymbol{x},\boldsymbol{Y})=(0_{Nd},0_{d\times Nd})$ belongs to the level set (6.15) for $\alpha\leq\sum_{i=1}^N f_i(\bar{w}_i)=F(\bar{\boldsymbol{w}},\bar{D},0,0)$. (We could have shown that in fact for any $(\bar{\boldsymbol{x}},\bar{\boldsymbol{Y}})\in(\mathbb{R}^d)^N\times(\mathbb{R}^{d\times d})^N$ there exists $\bar{\alpha}$ such that $(\bar{\boldsymbol{x}},\bar{\boldsymbol{Y}})$ is in the level set of level $\bar{\alpha}$, but this is not required.) □ □

The above discussion leads us to introduce our second candidate dynamics.

Distributed saddle-point dynamics for nuclear optimization.

Our second coordination algorithm for the distributed optimization with nuclear norm (6.8) is a saddle-point subgradient dynamics with proportional feedback on the disagreement of a subset of the variables:

$$\begin{aligned} w_i(k+1)&=\mathcal{P}_{\mathcal{W}}\Big(w_i(k)-\eta_k\Big(g_i(k)-2\gamma x_i(k)\Big)\Big),\\ D_i(k+1)&=\mathcal{P}_{\mathcal{D}}\Big(D_i(k)-\eta_k\gamma\Big(-x_ix_i^\top-\tfrac{\epsilon}{N}Y_iY_i^\top+\tfrac{\alpha}{N}\mathrm{I}_d\Big)\\ &\qquad+\sigma\sum_{j=1}^N\mathsf{a}_{ij,t}(D_j(k)-D_i(k))\Big),\\ x_i(k+1)&=x_i(k)+\eta_k\gamma\Big(-2D_ix_i(k)-2w_i(k)\Big),\\ Y_i(k+1)&=Y_i(k)+\eta_k\gamma\Big(-\frac{2\epsilon}{N}D_i(k)Y_i(k)-\frac{2\epsilon}{N}\mathrm{I}_d\Big), \end{aligned} \tag{6.17}$$

where $g_i(k) \in \partial f_i(w_i(k))$, for each $i \in \{1,\ldots,N\}$, and $\mathcal{P}_{\mathcal{W}}(\cdot)$ and $\mathcal{P}_{\mathcal{D}}(\cdot)$ denote the projections onto the compact convex sets $\mathcal{W}$ and $\mathcal{D}$. For the case of the regularizer $2\|[W \,|\, \sqrt{\epsilon}\mathrm{I}_d]\|_*$ we set $\alpha = 1$ and $\mathcal{D} = \mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$, and for the regularizer $\|[W \,|\, \sqrt{\epsilon}\mathrm{I}_d]\|_*^2$, we set $\alpha = 0$ and $\mathcal{D} = \Delta(c_\epsilon)$.

6.3 Convergence analysis

The convergence result of the distributed strategies (6.10) and (6.17) follows from Theorem 2.32, as we outline next.

Theorem 3.43. (Convergence of the coordination algorithms (6.10) and (6.17)). *Let the convex compact set $\mathcal{W} \subseteq \mathbb{R}^d$ be contained in $\bar{\mathcal{B}}(0, r_w)$ and let the bounds c_ϵ and r_ϵ be defined as in* (6.5) *and* (6.7). *Assume that each dynamics evolves over a sequence $\{\mathcal{G}_t\}_{t\geq 1}$ of B-jointly connected, δ-nondegenerate, weight-balanced digraphs with uniformly bounded Laplacian eigenvalues. Let σ be as follows: for any $\tilde{\delta}' \in (0,1)$, let $\tilde{\delta} := \min\left\{\tilde{\delta}',\, (1-\tilde{\delta}')\frac{\delta}{d_{\max}}\right\}$, where $d_{\max} := \max\left\{d_{\mathrm{out,t}}(j) \;:\; j \in \mathcal{I}, t \in \mathbb{Z}_{\geq 1}\right\}$, and choose*

$$\sigma \in \left[\frac{\tilde{\delta}}{\delta}, \frac{1-\tilde{\delta}}{d_{\max}}\right].$$

Assume also that the learning rates be chosen according to the Doubling Trick in Assumption 2.31. Then both the dynamics (6.10) *and* (6.17) *converge to an optimizer of* (6.8). *The evaluation error with respect to any minimum of* (6.9), *or with respect to any saddle point of the convex-concave function* (6.13), *is proportional to* $1/\sqrt{t}$.

Proof. The convergence result for both dynamics follows from the results in Theorem 2.32. Regarding the dynamics (6.17), the hypotheses of Theorem 2.32 are satisfied as we summarize next: the existence of a saddle point is proved in Proposition 2.42; the evolution of the estimates $\{w_i(k)\}_{i=1}^N$ and $\{D_i(k)\}_{i=1}^N$ in (6.17) is uniformly bounded thanks to the projections onto the compact sets $\mathcal{W}$ and $\mathit{\Delta}(c_\epsilon)$. On the other hand, the auxiliary states $\{x_i(k)\}_{i=1}^N$ and $\{Y_i(k)\}_{i=1}^N$ are uniformly bounded by virtue of Proposition 3.44 below. As a consequence of the bounded evolution of the dynamics (6.17), one can also construct bounds for the subgradients of each convex-concave function F_i. The other assumptions regarding the communication graphs and the choice of the design parameters σ and the learning rates $\{\eta_t\}_{t\geq 1}$ complete the hypotheses in Theorem 2.32. The convergence follows similarly from this result for the dynamics (6.10), which is also a particular case of the general dynamics (5.6). □ □

The proof of correctness above requires the evolution of the estimates of the dynamics to be bounded. In particular, the current analysis establishes the boundedness of the auxiliary states $\{x_i(k)\}_{i=1}^N$ and $\{Y_i(k)\}_{i=1}^N$ in the dynamics (6.17) by heavily relying on the fact that $w_i(k) \in \bar{\mathcal{B}}(0, r_w)$ and $D_i(k) \succeq c\mathrm{I}_d$ for all $k \geq 1$. Precisely, we use the latter to show the input-to-state stability (ISS) property (see [JSW99]) of the system

$$x_{k+1} = x_k + \eta_k \gamma\Big(-2D_k x_k - 2w_k\Big), \tag{6.18a}$$

$$Y_{k+1} = Y_k + \eta_k \gamma\Big(-\frac{2\epsilon}{N} D_k Y_k - \frac{2\epsilon}{N} F_k\Big), \tag{6.18b}$$

for arbitrary sequences of disturbances $\{w_k\}_{k\geq 1} \subset \mathbb{R}^d$ and $\{F_k\}_{k\geq 1} \subset \mathbb{R}^{d\times d}$, where $D_k \succeq c\mathrm{I}_d$.

Proposition 3.44. (ISS Lyapunov function for the auxiliary states in (6.17)). *Assume $D_k \succeq c\mathrm{I}_d$ and $\|D_k\|_2 \leq r$ for some $c \in \mathbb{R}_{>0}$ and $r \in [1, \infty)$, respectively, and* *let* $\{\eta_k\} \subset \mathbb{R}_{>0}$ *be any sequence of learning rates such that*

$$\eta_k \le \min\Big\{\frac{c-\kappa}{4\gamma r^2}, \frac{N(c-\kappa)}{4\gamma\epsilon r^2}\Big\}, \tag{6.19}$$

for some $\kappa \in (0,c)$ *for all* $k \ge 1$. *Then, the function* $\|\cdot\|_2^2$ *is an ISS-Lyapunov function for the evolution of* x_k *in* (6.18a) *when* $\{w_k\}_{k\ge1} \subset \mathbb{R}^d$ *is considered as a sequence of disturbances,*

$$\|x_{k+1}\|_2^2 \le \Big(1-\eta_k\gamma\kappa\Big)\|x_k\|_2^2 + \Big(8\eta_k^2\gamma^2 + \frac{\eta_k\gamma}{\kappa}\Big)\|w_k\|_2^2,$$

for all $k \ge 1$. *Similarly, the function* $\|\cdot\|_{\mathcal{F}}^2$ *is an ISS Lyapunov function for the evolution of* Y_k *in* (6.18b) *under the sequence of disturbances* $\{F_k\}_{k\ge1} \subset \mathbb{R}^{d\times d}$,

$$\|Y_{k+1}\|_{\mathcal{F}}^2 \le \Big(1-\frac{\eta_k\gamma\kappa\epsilon}{N}\Big)\|Y_k\|_{\mathcal{F}}^2 + \Big(\frac{8\eta_k^2\gamma^2\epsilon^2}{N^2} + \frac{\eta_k\gamma\epsilon}{\kappa N}\Big)\|F_k\|_{\mathcal{F}}^2,$$

for all $k \ge 1$.

Proof. Taking the square norm on both sides of the dynamics (6.18a), we get

$$\begin{aligned}\|x_{k+1}\|_2^2 &= \|x_k + \eta_k\gamma\Big(-2D_kx_k - 2w_k\Big)\|_2^2 \\ &= \|x_k\|_2^2 + 4\eta_k^2\gamma^2\|D_kx_k + w_k\|_2^2 - 2\eta_k\gamma x_k^\top(D_kx_k + w_k).\end{aligned}$$

Next we use Young's inequality, $2x^\top w \le \kappa\|x\|_2^2 + \frac{1}{\kappa}\|w\|_2^2$, for any $\kappa \in \mathbb{R}_{>0}$ and all $x, w \in \mathbb{R}^d$,

$$\begin{aligned}\|x_{k+1}\|_2^2 - \|x_k\|_2^2 \le{}& \eta_k^2\gamma^2\|D_kx_k + w_k\|_2^2 - 2\eta_k\gamma x_k^\top D_kx_k \\ &+ \eta_k\gamma\kappa\|x_k\|_2^2 + \frac{\eta_k\gamma}{\kappa}\|w_k\|_2^2 \\ \le{}& -\eta_k\gamma\Big(2c-\kappa-8\eta_k\gamma r^2\Big)\|x_k\|_2^2 + \Big(8\eta_k^2\gamma^2 + \frac{\eta_k\gamma}{\kappa}\Big)\|w_k\|_2^2, \end{aligned} \tag{6.20}$$

where in the last inequality we have used that $D_k \succeq c\mathrm{I}_d$ and also

$$
\begin{aligned}
\|D_k x_k + w_k\|_2^2 \leq & 2\|D_k x_k\|_2^2 + 2\|w_k\|_2^2 \\
& \leq 2\|D_k\|_2^2 \|x_k\|_2^2 + 2\|w_k\|_2^2 \leq 2\max\{1, r^2\}\|x_k\|_2^2 + 2\|w_k\|_2^2,
\end{aligned}
$$

which follows by Young's inequality. The first term in the minimum of (6.19) comes then from imposing the condition $2c - \kappa - 8\eta_k \gamma r^2 \geq \kappa$ in (6.20).

We derive the analogous result for the dynamics (6.18a) using the Frobenius norm $\|\cdot\|_{\mathcal{F}}$ in place of the Euclidean norm $\|\cdot\|_2$ and using the inequality $\operatorname{trace}(D^\top W) \leq \kappa \|D\|_{\mathcal{F}}^2 + \frac{1}{\kappa}\|W\|_2^2$. (This inequality is the same as for vectors because the Frobenius norm of a matrix is the Euclidean norm of the vectorization of the matrix and the trace of the product of two matrices is the same as the scalar product of the vectorization of the matrices.)

$$
\begin{aligned}
\|Y_{k+1}\|_{\mathcal{F}}^2 = & \|Y_k\|_{\mathcal{F}}^2 + \frac{4\eta_k^2 \gamma^2 \epsilon^2}{N^2} \|D_k Y_k + F_k\|_{\mathcal{F}}^2 \\
& - \frac{2\eta_k \gamma \epsilon}{N} \operatorname{trace}\Big(Y_k^\top (D_k Y_k + F_k)\Big)
\end{aligned}
$$

so, using that $-\operatorname{trace}(Y_k^\top D_k Y_k) \leq -c\|Y_k\|_{\mathcal{F}}$,

$$
\begin{aligned}
\|Y_{k+1}\|_{\mathcal{F}}^2 - \|Y_k\|_{\mathcal{F}}^2 \leq & \frac{4\eta_k^2 \gamma^2 \epsilon^2}{N^2} \|D_k Y_k + F_k\|_{\mathcal{F}}^2 \\
& - \frac{2\eta_k \gamma \epsilon}{N} \operatorname{trace}\Big(Y_k^\top D_k Y_k\Big) + \frac{\eta_k \gamma \kappa \epsilon}{N} \|Y_k\|_{\mathcal{F}}^2 + \frac{\eta_k \gamma \epsilon}{\kappa N} \|F_k\|_{\mathcal{F}}^2 \\
\leq & - \frac{\eta_k \gamma \epsilon}{N} \Big(2c - \kappa - 8\frac{\eta_k \gamma \epsilon r^2}{N}\Big) \|Y_k\|_{\mathcal{F}}^2 + \Big(\frac{8\eta_k^2 \gamma^2 \epsilon^2}{N^2} + \frac{\eta_k \gamma \epsilon}{\kappa N}\Big) \|F_k\|_{\mathcal{F}}^2,
\end{aligned}
\tag{6.21}
$$

where in the last inequality we have used again that $D_k \succeq c\mathrm{I}_d$ and also

$$\begin{aligned}\|D_kY_k+F_k\|_{\mathcal{F}}^2 &\leq 2\|D_kY_k\|_{\mathcal{F}}^2+2\|F_k\|_{\mathcal{F}}^2\\ &\leq 2\|D_k\|_{\mathcal{F}}^2\|Y_k\|_{\mathcal{F}}^2+2\|F_k\|_2^2 \leq 2\max\{1,r^2\}\|Y_k\|_{\mathcal{F}}^2+2\|F_k\|_{\mathcal{F}}^2,\end{aligned}$$

which follows by Young's inequality as above. The second term in the minimum of (6.19) comes from imposing $2c-\kappa-8\frac{\eta_k\gamma\epsilon r^2}{N}\geq\kappa$ in (6.21). □

As a consequence of the above result, the dynamics of x_k and Y_k in (6.18) are ISS [JSW99, Lemma 3.5 and Remark 3.6]. The ISS property in turn implies the uniform boundedness of the auxiliary states because in the dynamics (6.17), we have $w_k \in \bar{\mathcal{B}}(0,r_w)$ and $F_k = \mathrm{I}_d$ for all $k \geq 1$.

6.4 Simulation example

Here we illustrate the performance of the distributed saddle-point algorithm (6.17) on a matrix completion problem, cf. Section 6.1.2. The matrix $Z \in \mathbb{R}^{8\times 20}$ has rank 2 and each agent is assigned a column. From each column, only 5 entries have been revealed, and with this partial information, and without knowledge about the rank of Z, the agents execute the coordination algorithm (6.17) to solve the optimization (6.2). In this application each local function $f_i(w_i)=\sum_{j\in\Upsilon_i}(W_{ji}-Z_{ji})^2$ is not strongly convex, but just convex, in line with the hypotheses of Theorem 3.43. Figure 6.1 illustrates the matrix fitting error, the evolution of the network cost function, and the disagreement of the local auxiliary matrices.

6.5 Discussion

We have considered a class of optimization problems that involve the joint minimization over a set of local variables of a sum of convex functions together with a regularizing term that favors sparsity patterns in the resulting aggregate solution. Particular instances of these optimization problems include multi-task feature learning and matrix completion. We have exploited the separability property of a variational characterization of the nuclear norm to design two types of provably-correct distributed coordination algorithms. Our analysis relies on the body of work on distributed convex optimization and saddle-point dynamics. To the best of our knowledge, the proposed coordination algorithms are the first distributed dynamics for convex optimization with nuclear-norm regularization.

Acknowledgments

This chapter is an extended version of a work by D. Mateos-Núñez and J. Cortés presented in the *IFAC Workshop on Distributed Estimation and Control in Networked Systems, Philadelphia, Pennsylvania, USA, 2015,* with the title "Distributed optimization for multi-task learning via nuclear-norm approximation." The author of this thesis wishes to thank Ashish Cherukuri and Karan Sikka for stimulating discussions during the creation of the manuscript, and Mariza Suárez for her encouragement since the inception of these ideas. This research was supported by NSF Award CMMI-1300272.

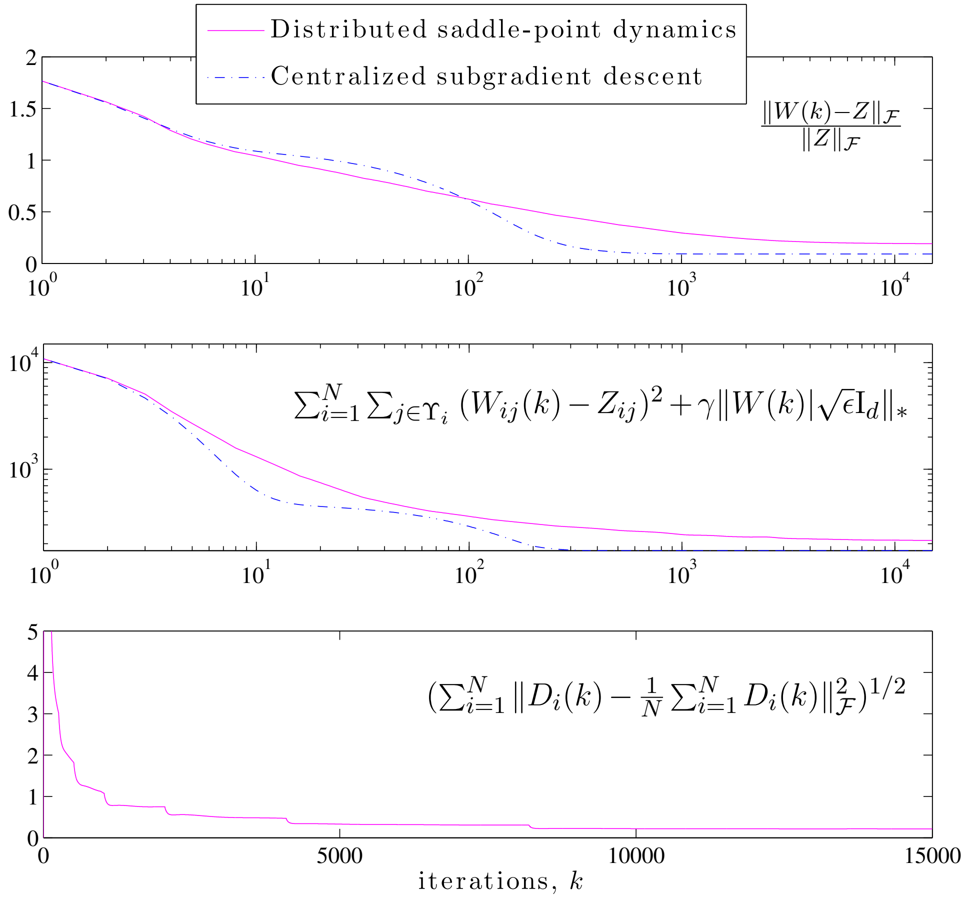


Figure 6.1: Simulation example of nuclear norm regularization for distributed low-rank matrix completion. Here we represent the evolution of algorithm (6.17) (magenta solid line). In the top we represent the matrix fitting error; in the center, the evolution of network cost function, and, in the bottom, the disagreement of the local matrices. The comparison is made with respect to a standard subgradient descent algorithm (blue dashed line) with constant gradient stepsize equal to 0.1. (The subgradient of the nuclear norm employed therein takes the form $U_r V_r^\top \in \partial\|W(k)\|_*$, where $U_r \Sigma_r V_r^\top$ is the reduced singular value decomposition of $W(k)$.) The optimization parameter weighting the nuclear norm is $\gamma = 2$, and the parameter of the approximate regularization is $\epsilon = 10^{-3}$. We use as constraint set $\mathcal{W} = \bar{\mathcal{B}}(0, r_w)$ with $r_w = 800$. In the distributed algorithm, the constraint set for the auxiliary matrices is $\mathcal{D} = \mathfrak{D}(\sqrt{\epsilon}, r_\epsilon)$, the consensus stepsize is $\sigma = 0.5$, and the communication topology is a ring connecting the 20 agents. Our algorithm is slower because it halves the learning rates (subgradient stepsizes) according to the doubling trick. This is necessary for asymptotic convergence in Theorem 3.43, in sharp contrast with standard (centralized) gradient descent that uses constant subgradient stepsize. The third plot shows the disagreement among the auxiliary matrices for our distributed algorithm. For decreasing learning rates, which is our case, the disagreement is guaranteed to converge to zero.

Chapter 7

pth moment noise-to-state stability of stochastic differential equations with persistent noise

We devote our last chapter to the stability properties of stochastic differential equations subject to persistent noise (including the case of additive noise), which is noise that is present even at the equilibria of the underlying differential equation and does not decay with time. A condensed version of the main result of this chapter was presented earlier in Section 2.6 because of its application to our continuous-time distributed algorithm with noisy communication channels. The class of systems we consider exhibit disturbance attenuation outside a closed, not necessarily bounded, set. We identify conditions, based on the existence of Lyapunov functions, to establish the noise-to-state stability in probability and in pth moment of the system with respect to a closed set. As part of our analysis, we study the concept of two functions being proper with respect to each other formalized via pair of inequalities with comparison functions. We show that such inequalities define several equivalence relations for increasingly strong refinements on the comparison functions. We also provide a complete characterization of the properties that a pair of functions must satisfy to belong to the same equivalence class. This characterization allows us to provide checkable conditions to determine whether a function satisfies the requirements to be a strong NSS-Lyapunov function in probability or a pth moment NSS-Lyapunov function.

7.1 Basic notions

We start with the basic definitions following [Mao11] with slightly more detail than in Chapter 2.

7.1.1 Brownian motion

Throughout the chapter, we assume that $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ is a complete probability space, where $\mathbb{P}$ is a probability measure defined on the σ-algebra $\mathcal{F}$, which contains all the subsets of Ω of probability 0. The filtration $\{\mathcal{F}_t\}_{t\geq 0}$ is a family of sub-σ-algebras of $\mathcal{F}$ satisfying $\mathcal{F}_t \subseteq \mathcal{F}_s \subseteq \mathcal{F}$ for any $0 \leq t < s < \infty$; we assume it is right continuous, i.e., $\mathcal{F}_t = \cap_{s>t}\mathcal{F}_s$ for any $t \geq 0$, and $\mathcal{F}_0$ contains all the subsets of Ω of probability 0. The Borel σ-algebra in $\mathbb{R}^n$, denoted by $\mathcal{B}^n$, or in $[t_0, \infty)$, denoted by $\mathcal{B}([t_0,\infty))$, are the smallest σ-algebras that contain all the open sets in $\mathbb{R}^n$ or $[t_0, \infty)$, respectively. A function $X : \Omega \to \mathbb{R}^n$ is $\mathcal{F}$-measurable if the set $\{\omega \in \Omega : X(\omega) \in A\}$ belongs to $\mathcal{F}$ for any $A \in \mathcal{B}^n$. We call such function a ($\mathcal{F}$-measurable) $\mathbb{R}^n$-valued random variable. If X is a real-valued random variable that is integrable with respect to $\mathbb{P}$, its expectation is $\mathbb{E}[X] = \int_\Omega X(\omega)\mathrm{d}\mathbb{P}(\omega)$. A function $f : \Omega \times [t_0, \infty) \to \mathbb{R}^n$ is $\mathcal{F} \times \mathcal{B}$-measurable (or just measurable) if the set $\{(\omega, t) \in \Omega \times [t_0, \infty) : f(\omega, t) \in A\}$ belongs to $\mathcal{F} \times \mathcal{B}([t_0, \infty))$ for any $A \in \mathcal{B}^n$. We call such function an $\{\mathcal{F}_t\}$-adapted process if $f(.,t):\Omega\to\mathbb{R}^n$ is $\mathcal{F}_t$-measurable for every $t\ge t_0$. At times, we omit the dependence on "ω", in the sense that we refer to the indexed family of random variables, and refer to the random process $f=\{f(t)\}_{t\ge t_0}$. We define $\mathcal{L}^1([t_0,\infty);\mathbb{R}^n)$ as the set of all $\mathbb{R}^n$-valued measurable $\{\mathcal{F}_t\}$-adapted processes f such that $\mathbb{P}(\{\omega\in\Omega: \int_{t_0}^T \|f(\omega,s)\|_2\,\mathrm{d}s<\infty\})=1$ for every $T>t_0$. Similarly, $\mathcal{L}^2([t_0,\infty);\mathbb{R}^{n\times m})$ denotes the set of all $\mathbb{R}^{n\times m}$-matrix-valued measurable $\{\mathcal{F}_t\}$-adapted processes G such that $\mathbb{P}(\{\omega\in\Omega: \int_{t_0}^T \|G(\omega,s)\|_{\mathcal{F}}^2\,\mathrm{d}s<\infty\})=1$ for every $T>t_0$.

A one-dimensional Brownian motion $\mathrm{B}:\Omega\times[t_0,\infty)\to\mathbb{R}$ defined in the probability space $(\Omega,\mathcal{F},\{\mathcal{F}_t\}_{t\ge0},\mathbb{P})$ is an $\{\mathcal{F}_t\}$-adapted process such that

- $\mathbb{P}(\{\omega\in\Omega: \mathrm{B}(\omega,t_0)=0\})=1$;
- the mapping $\mathrm{B}(\omega,.):[t_0,\infty)\to\mathbb{R}$, called sample path, is continuous also with probability 1;
- the increment $\mathrm{B}(.,t)-\mathrm{B}(.,s):\Omega\to\mathbb{R}$ is independent of $\mathcal{F}_s$ for $t_0\le s<t<\infty$ (i.e., if $S_b\triangleq\{\omega\in\Omega:\mathrm{B}(\omega,t)-\mathrm{B}(\omega,s)\in(-\infty,b)\}$, for $b\in\mathbb{R}$, then $\mathbb{P}(A\cap S_b)=\mathbb{P}(A)\mathbb{P}(S_b)$ for all $A\in\mathcal{F}_s$ and all $b\in\mathbb{R}$). In addition, this increment is normally distributed with zero mean and variance $t-s$.

An m-dimensional Brownian motion $\mathrm{B}:\Omega\times[t_0,\infty)\to\mathbb{R}^m$ is given by $\mathrm{B}(\omega,t)=[\mathrm{B}_1(\omega,t),\ldots,\mathrm{B}_m(\omega,t)]^\top$, where each B_i is a one-dimensional Brownian motion and, for each $t\ge t_0$, the random variables $\mathrm{B}_1(t),...,\mathrm{B}_m(t)$ are independent.

7.1.2 Stochastic differential equations

Here we review some basic notions on stochastic differential equations (SDEs) following [Mao11]; other useful references are [Kha12, Ö10, Mov11]. Consider the

n-dimensional SDE

$$\mathrm{d}x(\omega,t) = f\Big(x(\omega,t),t\Big)\mathrm{d}t + G\Big(x(\omega,t),t\Big)\Sigma(t)\,\mathrm{dB}(\omega,t), \tag{7.1}$$

where $x(\omega,t) \in \mathbb{R}^n$ is a realization at time t of the random variable $x(.,t) : \Omega \to \mathbb{R}^n$, for $t \in [t_0,\infty)$. The initial condition is given by $x(\omega,t_0) = x_0$ with probability 1 for some $x_0 \in \mathbb{R}^n$. The functions $f : \mathbb{R}^n \times [t_0,\infty) \to \mathbb{R}^n$, $G : \mathbb{R}^n \times [t_0,\infty) \to \mathbb{R}^{n\times q}$, and $\Sigma : [t_0,\infty) \to \mathbb{R}^{q\times m}$ are measurable. The functions f and G are regarded as a model for the architecture of the system and, instead, Σ is part of the model for the stochastic disturbance; at any given time Σ determines a linear transformation of the m-dimensional Brownian motion $\{\mathrm{B}(t)\}_{t\geq t_0}$, so that at time $t \geq t_0$ the input to the system is the process $\{\Sigma(t)\mathrm{B}(t)\}_{t\geq t_0}$, with covariance $\int_{t_0}^{t} \Sigma(t)\Sigma(t)^\top \mathrm{d}s$. The distinction between the roles of G and Σ is irrelevant for the SDE; both together determine the effect of the Brownian motion. The integral form of (7.1) is given by

$$x(\omega,t) = x_0 + \int_{t_0}^{t} f\Big(x(\omega,s),s\Big)\mathrm{d}s + \int_{t_0}^{t} G\Big(x(\omega,s),s\Big)\Sigma(s)\,\mathrm{dB}(\omega,s),$$

where the second integral is an stochastic integral [Mao11, p. 18]. A $\mathbb{R}^n$-valued random process $\{x(t)\}_{t\geq t_0}$ is a solution of (7.1) with initial value x_0 if

(i) is continuous with probability 1, $\{\mathcal{F}_t\}$-adapted, and satisfies $x(\omega,t_0) = x_0$ with probability 1,

(ii) the processes $\{f(x(t),t)\}_{t\geq t_0}$ and $\{G(x(t),t)\}_{t\geq t_0}$ belong to $\mathcal{L}^1([t_0,\infty);\mathbb{R}^n)$ and $\mathcal{L}^2([t_0,\infty);\mathbb{R}^{n\times m})$ respectively, and

(iii) equation (7.1) holds for every $t \geq t_0$ with probability 1.

A solution $\{x(t)\}_{t\geq t_0}$ of (7.1) is unique if any other solution $\{\bar{x}(t)\}_{t\geq t_0}$ with $\bar{x}(t_0) = x_0$ differs from it only in a set of probability 0, that is, $\mathbb{P}(\Big\{x(t) = \bar{x}(t) \;\; \forall t \geq t_0\Big\}) = 1$.

We make the following assumptions on the objects defining (7.1) to guarantee existence and uniqueness of solutions.

Assumption 1.45. *We assume Σ is essentially locally bounded. Furthermore, for any $T > t_0$ and $n \geq 1$, we assume there exists $K_{T,n} > 0$ such that, for almost every $t \in [t_0, T]$ and all $x, y \in \mathbb{R}^n$ with $\max\Big\{\|x\|_2, \|y\|_2\Big\} \leq n$,*

$$\max\Big\{\, \|f(x,t) - f(y,t)\|_2^2,\, \|G(x,t) - G(y,t)\|_{\mathcal{F}}^2 \,\Big\} \leq K_{T,n}\|x - y\|_2^2.$$

Finally, we assume that for any $T > t_0$, there exists $K_T > 0$ such that, for almost every $t \in [t_0, T]$ and all $x \in \mathbb{R}^n$, $x^\top f(x,t) + \frac{1}{2}\|G(x,t)\|_{\mathcal{F}}^2 \leq K_T(1 + \|x\|_2^2)$.

According to [Mao11, Th. 3.6, p. 58], Assumption 1.45 is sufficient to guarantee global existence and uniqueness of solutions of (7.1) for each initial condition $x_0 \in \mathbb{R}^n$.

We conclude this section by presenting a useful operator in the stability analysis of SDEs. Given a function $\mathrm{V} \in \mathcal{C}^2(\mathbb{R}^n; \mathbb{R}_{\geq 0})$, we define the generator of (7.1) acting on the function V as the mapping $\mathcal{L}[\mathrm{V}] : \mathbb{R}^n \times [t_0, \infty) \to \mathbb{R}$ given by

$$\mathcal{L}[\mathrm{V}](x,t) \triangleq \nabla \mathrm{V}(x)^\top f(x,t) + \tfrac{1}{2}\operatorname{trace}\Big(\Sigma(t)^\top G(x,t)^\top \nabla^2 \mathrm{V}(x) G(x,t)\Sigma(t)\Big). \tag{7.2}$$

It can be shown that $\mathcal{L}[\mathrm{V}](x,t)$ gives the expected rate of change of V along a solution of (7.1) that passes through the point x at time t, so it is a generalization of the Lie derivative. According to [Mao11, Th. 6.4, p. 36], if we evaluate V along the solution $\{x(t)\}_{t \geq t_0}$ of (7.1), then the process $\{\mathrm{V}(x(t))\}_{t \geq t_0}$ satisfies the new SDE

$$\mathrm{V}(x(t)) = \mathrm{V}(x_0) + \int_{t_0}^{t} \mathcal{L}[\mathrm{V}](x(s),s)\mathrm{d}s + \int_{t_0}^{t} \nabla \mathrm{V}(x(s))^\top G(x(s),s)\Sigma(s)\mathrm{dB}(s). \tag{7.3}$$

Equation (7.3) is known as Itô's formula and corresponds to the stochastic version of the chain rule.

7.2 Noise-to-state stability via noise-dissipative Lyapunov functions

In this section, we study the stability of stochastic differential equations subject to persistent noise. Our first step is the introduction of a novel notion of stability. This captures the behavior of the pth moment of the distance (of the state) to a given closed set, as a function of two objects: the initial condition and the maximum size of the covariance. After this, our next step is to derive several Lyapunov-type stability results that help determine whether a stochastic differential equation enjoys these stability properties. The following definition generalizes the concept of noise-to-state stability given in [DK00].

Definition 2.46. (Noise-to-state stability with respect to a set). *The system* (7.1) *is* noise-to-state stable (NSS) in probability with respect to the set $\mathcal{U} \subseteq \mathbb{R}^n$ *if for any $\epsilon > 0$ there exist $\mu \in \mathcal{KL}$ and $\theta \in \mathcal{K}$ (that might depend on ϵ), such that*

$$\mathbb{P}\Big\{ |x(t)|_{\mathcal{U}}^p > \mu\Big(|x_0|_{\mathcal{U}}, t-t_0\Big) + \theta\Big(\operatorname*{ess\,sup}_{t_0 \le s \le t} \|\Sigma(s)\|_{\mathcal{F}}\Big) \Big\} \le \epsilon, \tag{7.4}$$

for all $t \geq t_0$ and any $x_0 \in \mathbb{R}^n$. And the system (7.1) *is* pth moment noise-to-state stable ($p$thNSS) with respect to $\mathcal{U}$ *if there exist $\mu \in \mathcal{KL}$ and $\theta \in \mathcal{K}$, such that*

$$\mathbb{E}\Big[|x(t)|_{\mathcal{U}}^p\Big] \le \mu\Big(|x_0|_{\mathcal{U}}, t-t_0\Big) + \theta\Big(\operatorname*{ess\,sup}_{t_0 \le s \le t} \|\Sigma(s)\|_{\mathcal{F}}\Big), \tag{7.5}$$

for all $t \geq t_0$ and any $x_0 \in \mathbb{R}^n$. The gain functions μ and θ are the overshoot gain *and the* noise gain*, respectively.*

The quantity $\|\Sigma(t)\|_{\mathcal{F}} = \sqrt{\operatorname{trace}\left(\Sigma(t)\Sigma(t)^\top\right)}$ is a measure of the size of the noise because it is related to the infinitesimal covariance $\Sigma(t)\Sigma(t)^\top$. The choice of the pth power is irrelevant in the statement in probability since one could take any $\mathcal{K}_\infty$ function evaluated at $|x(t)|_{\mathcal{U}}$. However, this would make a difference in the statement in expectation. (Also, we use the same power for convenience.) When the set $\mathcal{U}$ is a subspace, we can substitute $|.|_{\mathcal{U}}$ by $\|.\|_A$, for some matrix $A \in \mathbb{R}^{m\times n}$ with $\mathcal{N}(A) = \mathcal{U}$. In such a case, the definition above does not depend on the choice of the matrix A.

Remark 2.47. (NSS is not a particular case of ISS). The concept of NSS is not a particular case of input-to-state stability (ISS) [Son08] for systems that are affine in the input, namely,

$$\dot{y} = f(y,t) + G(y,t)u(t) \;\Leftrightarrow\; y(t) = y(t_0) + \int_{t_0}^{t} f(y(s),s)\,\mathrm{d}s + \int_{t_0}^{t} G(y(s),s)u(s)\,\mathrm{d}s,$$

where $u : [t_0,\infty) \to \mathbb{R}^q$ is measurable and essentially locally bounded [Son98, Sec. C.2]. The reason is the following: the components of the vector-valued function $\int_{t_0}^{t} G(y(s),s)u(s)\,\mathrm{d}s$ are differentiable almost everywhere by the Lebesgue fundamental theorem of calculus [MW99, p. 289], and thus absolutely continuous [MW99, p. 292] and with bounded variation [MW99, Prop. 8.5]. On the other hand, at any time previous to $t_k(t) \triangleq \min\{t, \inf\{s \geq t_0 : \|x(s)\|_2 \geq k\}\}$, the driving disturbance of (7.1) is the vector-valued function $\int_{t_0}^{t_k(t)} G(x(s),s)\Sigma(s)\mathrm{dB}(s)$, whose ith component has quadratic variation [Mao11, Th. 5.14, p. 25] equal to

$$\int_{t_0}^{t_k(t)} \sum_{j=1}^{m} |\sum_{l=1}^{q} G(x(s),s)_{il}\Sigma(s)_{lj}|^2 \mathrm{d}s > 0.$$

Since a continuous process that has positive quadratic variation must have infinite variation [Kle05, Th. 1.10], we conclude that the driving disturbance in this case is not allowed in the ISS framework. •

Our first goal now is to provide tools to establish whether a stochastic differential equation enjoys the noise-to-state stability properties given in Definition 2.46. To achieve this, we look at the dissipativity properties of a special kind of energy functions along the solutions of (7.1).

Definition 2.48. (Noise-dissipative Lyapunov function). *A function* $\mathrm{V} \in \mathcal{C}^2(\mathbb{R}^n;\mathbb{R}_{\geq 0})$ *is a* noise-dissipative Lyapunov function *for* (7.1) *if there exist* $\mathrm{W} \in \mathcal{C}(\mathbb{R}^n;\mathbb{R}_{\geq 0})$*,* $\sigma \in \mathcal{K}$*, and concave* $\eta \in \mathcal{K}_\infty$ *such that*

$$\mathrm{V}(x) \leq \eta(\mathrm{W}(x)), \tag{7.6}$$

for all $x \in \mathbb{R}^n$*, and the following dissipation inequality holds:*

$$\mathcal{L}[\mathrm{V}](x,t) \leq -\mathrm{W}(x) + \sigma\Big(\|\Sigma(t)\|_{\mathcal{F}}\Big), \tag{7.7}$$

for all $(x,t) \in \mathbb{R}^n \times [t_0,\infty)$.

Remark 2.49. (Itô formula and exponential dissipativity). Interestingly, the conditions (7.6) and (7.7) are equivalent to

$$\mathcal{L}[\mathrm{V}](x,t) \leq -\eta^{-1}(\mathrm{V}(x)) + \sigma\Big(\|\Sigma(t)\|_{\mathcal{F}}\Big), \tag{7.8}$$

for all $x \in \mathbb{R}^n$, where $\eta^{-1} \in \mathcal{K}_\infty$ is convex. Note that, since $\mathcal{L}[\mathrm{V}]$ is not the Lie derivative of V (as it contains the Hessian of V), one cannot directly deduce from (7.8) the existence of a continuously twice differentiable function $\tilde{\mathrm{V}}$ such that

$$\mathcal{L}[\tilde{\mathrm{V}}](x,t) \leq -c\tilde{\mathrm{V}}(x) + \tilde{\sigma}\Big(\|\Sigma(t)\|_{\mathcal{F}}\Big), \tag{7.9}$$

as instead can be done in the context of ISS, see e.g. [PW96]. •

Example 2.50. (A noise-dissipative Lyapunov function). Assume that $h : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and verifies

$$\gamma(\|x - x'\|_2^2) \leq (x - x')^\top (\nabla h(x) - \nabla h(x')) \tag{7.10}$$

for some convex function $\gamma \in \mathcal{K}_\infty$ for all $x, x' \in \mathbb{R}^n$. In particular, this implies that h is strictly convex. (Incidentally, any strongly convex function verifies (7.10) for some choice of γ linear and strictly increasing.) Consider now the dynamics

$$\mathrm{d}x(\omega,t) = -\Big(\delta \mathsf{L} x(\omega,t) + \nabla h(x(\omega,t))\Big)\mathrm{d}t + \Sigma(t)\,\mathrm{dB}(\omega,t), \tag{7.11}$$

for all $t \in [t_0, \infty)$, where $x(\omega, t_0) = x_0$ with probability 1 for some $x_0 \in \mathbb{R}^n$, and $\delta > 0$. Here, the matrix $\mathsf{L} \in \mathbb{R}^{n\times n}$ is symmetric and positive semidefinite, and the matrix-valued function $\Sigma : [t_0, \infty) \to \mathbb{R}^{n\times m}$ is continuous. This dynamics corresponds to the SDE (7.1) with $f(x,t) \triangleq -\delta \mathsf{L} x - \nabla h(x)$ and $G(x,t) \triangleq \mathrm{I}_n$ for all $(x,t) \in \mathbb{R}^n \times [t_0, \infty)$.

Let $x^* \in \mathbb{R}^n$ be the unique solution of the Karush-Kuhn-Tucker [BV09] condition $\delta \mathsf{L} x^* = -\nabla h(x^*)$, corresponding to the unconstrained minimization of $F(x) \triangleq \frac{\delta}{2} x^\top \mathsf{L} x + h(x)$. Consider then the candidate Lyapunov function $\mathrm{V} \in \mathcal{C}^2(\mathbb{R}^n; \mathbb{R}_{\geq 0})$

given by $\mathrm{V}(x) \triangleq \frac{1}{2}(x-x^*)^\top(x-x^*)$. Using (7.2), we obtain that, for all $x \in \mathbb{R}^n$,

$$\begin{aligned}\mathcal{L}[\mathrm{V}](x,t) &= -(x-x^*)^\top\Big(\delta\mathsf{L}x + \nabla h(x)\Big) + \tfrac{1}{2}\operatorname{trace}\Big(\Sigma(t)^\top\Sigma(t)\Big) \\ &= -\delta(x-x^*)^\top\mathsf{L}(x-x^*) - (x-x^*)^\top\Big(\nabla h(x) - \nabla h(x^*)\Big) + \tfrac{1}{2}\|\Sigma(t)\|_{\mathcal{F}}^2 \\ &\leq -\gamma(\|x-x^*\|_2^2) + \tfrac{1}{2}\|\Sigma(t)\|_{\mathcal{F}}^2.\end{aligned}$$

We note that $\mathrm{W} \in \mathcal{C}(\mathbb{R}^n;\mathbb{R}_{\geq 0})$ defined by $\mathrm{W}(x) \triangleq \gamma(\|x-x^*\|_2^2)$ verifies

$$\mathrm{V}(x) = \tfrac{1}{2}\gamma^{-1}\Big(\mathrm{W}(x)\Big) \qquad \forall x \in \mathbb{R}^n,$$

where γ^{-1} is concave and belongs to the class $\mathcal{K}_\infty$ as explained in Section 2.2.1. Therefore, V is a noise-dissipative Lyapunov function for (7.11), with concave $\eta \in \mathcal{K}_\infty$ given by $\eta(r) = 1/2\gamma^{-1}(r)$ and $\sigma \in \mathcal{K}$ given by $\sigma(r) \triangleq 1/2\, r^2$. •

The next result generalizes [DKW01, Th. 4.1] to positive semidefinite Lyapunov functions that satisfy weaker dissipativity properties (cf. (7.8)) than the typical exponential-like inequality (7.9), and characterizes the overshoot gain.

Theorem 2.51. (Noise-dissipative Lyapunov functions have an NSS dynamics). *Under Assumption 1.45, and further assuming that* Σ *is continuous, suppose that* V *is a noise-dissipative Lyapunov function for* (7.1). *Then,*

$$\mathbb{E}\Big[\mathrm{V}(x(t))\Big] \leq \tilde{\mu}\Big(\mathrm{V}(x_0), t-t_0\Big) + \eta\Big(2\sigma\Big(\max_{t_0 \leq s \leq t}\|\Sigma(s)\|_{\mathcal{F}}\Big)\Big), \tag{7.12}$$

for all $t \geq t_0$, *where the class* $\mathcal{KL}$ *function* $(r,s) \mapsto \tilde{\mu}(r,s)$ *is well defined as the solution* $y(s)$ *to the initial value problem*

$$\dot{y}(s) = -\tfrac{1}{2}\eta^{-1}(y(s)), \quad y(0) = r. \tag{7.13}$$

Proof. Recall that Assumption 1.45 guarantees the global existence and uniqueness of solutions of (7.1). Given the process $\{\mathrm{V}(x(t))\}_{t\geq t_0}$, the proof strategy is to obtain a differential inequality for $\mathbb{E}\Big[\mathrm{V}(x(t))\Big]$ using Itô formula (7.3), and then use a comparison principle to translate the problem into one of standard input-to-state stability for an appropriate choice of the input.

To carry out this strategy, we consider Itô formula (7.3) with respect to an arbitrary reference time instant $t' \geq t_0$,

$$\mathrm{V}(x(t)) = \mathrm{V}(x(t')) + \int_{t'}^{t} \mathcal{L}[\mathrm{V}](x(s),s)\mathrm{d}s + \int_{t'}^{t} \nabla \mathrm{V}(x(s))^\top G(x(s),s)\Sigma(s)\mathrm{dB}(s), \tag{7.14}$$

and we first ensure that the expectation of the integral against Brownian motion is 0. Let $S_k = \{x \in \mathbb{R}^n : \|x\|_2 \leq k\}$ be the ball of radius k centered at the origin. Fix $x_0 \in \mathbb{R}^n$ and denote by τ_k the first exit time of $x(t)$ from S_k for integer values of k greater than $\|x(t_0)\|_2$, namely, $\tau_k \triangleq \inf\{s \geq t_0 : \|x(s)\|_2 \geq k\}$, for $k > \lceil \|x(t_0)\|_2 \rceil$. Since the event $\{\omega \in \Omega : \tau_k \leq t\}$ belongs to $\mathcal{F}_t$ for each $t \geq t_0$, it follows that τ_k is an $\{\mathcal{F}_t\}$-stopping time for each $t \geq t_0$. Now, for each k fixed, if we consider the random variable $t_k(t) \triangleq \min\{t, \tau_k\}$ and define $I(t',t)$ as the stochastic integral in (7.14) for any fixed $t' \in [t_0, t_k(t)]$, then the process $I(t', t_k(t))$ has zero expectation as we show next. The function $X : S_k \times [t',t] \to \mathbb{R}$ given by $X(x,s) \triangleq \nabla \mathrm{V}(x)^\top G(x,s)\Sigma(s)$ is essentially bounded (in its domain), and thus $\mathbb{E}\Big[\int_{t'}^{t} \mathbf{1}_{[t',t_k(t)]}(s)\, X(x(s),t)^2\, \mathrm{d}s\Big] < \infty$, where $\mathbf{1}_{[t',t_k(t)]}(s)$ is the indicator function of the set $[t', t_k(t)]$. Therefore, $\mathbb{E}\Big[I(t',t_k(t))\Big] = 0$ by [Mao11, Th. 5.16, p. 26]. Define now $\bar{\mathrm{V}}(t) \triangleq \mathbb{E}\Big[\mathrm{V}(x(t))\Big]$ and $\bar{\mathrm{W}}(t) \triangleq \mathbb{E}\Big[\mathrm{W}(x(t))\Big]$ in $\Gamma(t_0) \triangleq \{t \geq t_0 : \bar{\mathrm{V}}(t) < \infty\}$. By the above, taking expectations in (7.14) and using (7.7), we obtain that

$$\begin{aligned}\bar{\mathrm{V}}(t_k(t)) &= \bar{\mathrm{V}}(t') + \mathbb{E}\Big[\int_{t'}^{t_k(t)} \mathcal{L}[\mathrm{V}](x(s),s)\mathrm{d}s\Big] \\ &\leq \bar{\mathrm{V}}(t') - \mathbb{E}\Big[\int_{t'}^{t_k(t)} \mathrm{W}(x(s))\mathrm{d}s\Big] + \mathbb{E}\Big[\int_{t'}^{t_k(t)} \sigma(\|\Sigma(s)\|_{\mathcal{F}})\,\mathrm{d}s\Big] \qquad (7.15)\end{aligned}$$

for all $t \in \Gamma(t_0)$ and any $t' \in [t_0, t_k(t)]$. Next we use the fact that V is continuous and $\{x(t)\}_{t\geq t_0}$ is also continuous with probability 1. In addition, according to Fatou's lemma [MW99, p. 123] for convergence in the probability measure, we get that

$$\begin{aligned}\bar{\mathrm{V}}(t) &= \mathbb{E}\Big[\mathrm{V}(x(\liminf_{k\to\infty} t_k(t)))\Big] = \mathbb{E}\Big[\liminf_{k\to\infty} \mathrm{V}(x(t_k(t)))\Big] \qquad (7.16) \\ &\leq \liminf_{k\to\infty} \mathbb{E}\Big[\mathrm{V}(x(t_k(t)))\Big] = \liminf_{k\to\infty} \bar{\mathrm{V}}(t_k(t))\end{aligned}$$

for all $t \in \Gamma(t_0)$. Moreover, using the monotone convergence [MW99, p. 176] when $k \to \infty$ in both Lebesgue integrals in (7.15) (because both integrands are nonnegative and $\mathbf{1}_{[t',t_k(t)]}$ converges monotonically to $\mathbf{1}_{[t',t]}$ as $k \to \infty$ for any $t' \in [t_0, t_k(t)]$), we obtain from (7.16) that

$$\bar{\mathrm{V}}(t) \leq \bar{\mathrm{V}}(t') - \mathbb{E}\Big[\int_{t'}^{t} \mathrm{W}(x(s))\mathrm{d}s\Big] + \int_{t'}^{t} \sigma(\|\Sigma(s)\|_{\mathcal{F}})\,\mathrm{d}s \qquad (7.17)$$

for all $t \in \Gamma(t_0)$ and any $t' \in [t_0, t]$. Before resuming the argument we make two observations. First, applying Tonelli's theorem [MW99, p. 212] to the nonnegative process $\{\mathrm{W}(x(s))\}_{s\geq t'}$, it follows that

$$\mathbb{E}\Big[\int_{t'}^{t} \mathrm{W}(x(s))\mathrm{d}s\Big] = \int_{t'}^{t} \bar{\mathrm{W}}(x(s))\mathrm{d}s. \qquad (7.18)$$

Second, using (7.6) and Jensen's inequality [Bor95, Ch. 3], we get that

$$\bar{\mathrm{V}}(t) = \mathbb{E}\Big[\mathrm{V}(x(t))\Big] \leq \mathbb{E}\Big[\eta(\mathrm{W}(x(t)))\Big] \leq \eta\Big(\mathbb{E}\Big[\mathrm{W}(x(t))\Big]\Big) = \eta\Big(\bar{\mathrm{W}}(t)\Big), \tag{7.19}$$

because η is concave, so $\bar{\mathrm{W}}(t) \geq \eta^{-1}(\bar{\mathrm{V}}(t))$. Hence, (7.17) and (7.18) yield

$$\begin{aligned} \bar{\mathrm{V}}(t) &\leq \bar{\mathrm{V}}(t') - \int_{t'}^{t} \bar{\mathrm{W}}(s)\,\mathrm{d}s + \int_{t'}^{t} \sigma(\|\Sigma(s)\|_{\mathcal{F}})\,\mathrm{d}s \\ &\leq \bar{\mathrm{V}}(t') + \int_{t'}^{t} \Big(-\eta^{-1}(\bar{\mathrm{V}}(s)) + \sigma(\|\Sigma(s)\|_{\mathcal{F}})\Big)\,\mathrm{d}s \end{aligned} \tag{7.20}$$

for all $t \in \Gamma(t_0)$ and any $t' \in [t_0, t]$, which in particular shows that $\Gamma(t_0)$ can be taken equal to $[t_0, \infty)$.

Now the strategy is to compare $\bar{\mathrm{V}}$ with the unique solution of an ordinary differential equation that represents an input-to-state stable (ISS) system. First we leverage the integral inequality (7.20) to show that $\bar{\mathrm{V}}$ is continuous in $[t_0, \infty)$, which allows us then to rewrite (7.20) as a differential inequality at t'. To to show that $\bar{\mathrm{V}}$ is continuous, we use the dominated convergence theorem [Mao11, Thm. 2.3, P. 6] applied to $V_k(\hat{t}) \triangleq \mathrm{V}(x(\hat{t})) - \mathrm{V}(x(\hat{t}+1/k))$, for $\hat{t} \in [t_0, t]$, and similarly taking $\hat{t} - 1/k$ (excluding, respectively, the cases when $\hat{t} = t$ or $\hat{t} = t_0$). The hypotheses are satisfied because V_k can be majorized using (7.20) as

$$|V_k(\hat{t})| \;\leq\; \mathrm{V}(x(\hat{t})) + \mathrm{V}(x(\hat{t}+1/k)) \;\leq\; 2\Big(\mathrm{V}(x_0) + \int_{t_0}^{t} \sigma(\|\Sigma(s)\|_{\mathcal{F}})\,\mathrm{d}s\Big), \tag{7.21}$$

where the term on the right is not a random variable and thus coincides with its expectation. Therefore, for every $\hat{t} \in [t_0, t]$,

$$\lim_{s \to \hat{t}} \mathbb{E}\Big[\mathrm{V}(x(s))\Big] = \mathbb{E}\Big[\lim_{s \to \hat{t}} \mathrm{V}(x(s))\Big] = \mathbb{E}\Big[\mathrm{V}(x(\hat{t}))\Big],$$

so $\bar{\mathrm{V}}$ is continuous on $[t_0,t]$, for any $t \geq t_0$. Now, using again (7.20) and the continuity of the integrand, we can bound the upper right-hand derivative [Kha02, Appendix C.2] (also called upper Dini derivative), as

$$\begin{aligned} D^{+}\bar{\mathrm{V}}(t') &\triangleq \limsup_{t\to t',\, t>t'} \frac{\bar{\mathrm{V}}(t)-\bar{\mathrm{V}}(t')}{t-t'} \\ &\leq \limsup_{t\to t',\, t>t'} \frac{1}{t-t'}\int_{t'}^{t}\Big(-\eta^{-1}(\bar{\mathrm{V}}(s))+\sigma(\|\Sigma(s)\|_{\mathcal{F}})\Big)\,\mathrm{d}s = h(\bar{\mathrm{V}}(t'),b(t')), \end{aligned}$$

for any $t' \in [t_0,\infty)$, where the function $h : \mathbb{R}_{\geq 0}\times\mathbb{R}_{\geq 0}\to\mathbb{R}$ is given by

$$h(y,b) \triangleq -\eta^{-1}(y)+b,$$

and $b(t) \triangleq \sigma(\|\Sigma(t)\|_{\mathcal{F}})$, which is continuous in $[t_0,\infty)$. Therefore, according to the comparison principle [Kha02, Lemma 3.4, P. 102], using that $\bar{\mathrm{V}}$ is continuous in $[t_0,\infty)$ and $D^{+}\bar{\mathrm{V}}(t') \leq h(\bar{\mathrm{V}}(t'),b(t'))$, for any $t' \in [t_0,\infty)$, the solutions [Son98, Sec. C.2] of the initial value problem

$$\dot{\mathrm{U}}(t) = h(\mathrm{U}(t),b(t)), \qquad \mathrm{U}_0 \triangleq \mathrm{U}(t_0) = \bar{\mathrm{V}}(t_0) \tag{7.22}$$

(where h is locally Lipschitz in the first argument as we show next), satisfy that $\mathrm{U}(t) \geq \bar{\mathrm{V}}(t)\,(\geq 0)$ in the common interval of existence. We argue the global existence and uniqueness of solutions of (7.22) as follows. Since $\alpha \triangleq \eta^{-1}$ is convex and class $\mathcal{K}_\infty$ (see Section 2.2.1), it holds that

$$\alpha(s') \leq \alpha(s) \leq \alpha(s') + \tfrac{\alpha(s'')-\alpha(s')}{s''-s'}(s-s')$$

for all $s \in [s',s'']$, for any $s'' > s' \geq 0$. Thus, $|\alpha(s)-\alpha(s')| = \alpha(s)-\alpha(s') \leq L(s-s')$, for any $s'' \geq s \geq s' \geq 0$, where $L \triangleq (\alpha(s'')-\alpha(s'))/(s''-s')$, so η^{-1} is locally

Lipschitz. Hence, h is locally Lipschitz in $\mathbb{R}_{\geq 0}\times\mathbb{R}_{\geq 0}$. Therefore, given the input function b and any $\mathrm{U}_0\geq 0$, there is a unique maximal solution of (7.22), denoted by $\mathrm{U}(\mathrm{U}_0,t_0;t)$, defined in a maximal interval $[t_0,t_{\max}(\mathrm{U}_0,t_0))$. (As a by-product, the initial value problem (7.13), which can be written as $\dot{y}(s)=\frac{1}{2}h(y(s),0)$, $y(0)=r$, has a unique and strictly decreasing solution in $[0,\infty)$, so $\tilde{\mu}$ in the statement is well defined and in class $\mathcal{KL}$.) To show that (7.22) is ISS we follow a similar argument as in the proof of [Son08, Th. 5] (and note that, as a consequence, we obtain that $t_{\max}(\mathrm{U}_0,t_0)=\infty$). Firstly, if $\eta^{-1}(\mathrm{U})\geq 2b$, then $\dot{\mathrm{U}}(t)=-\frac{1}{2}\eta^{-1}(\mathrm{U}(t))$, which implies that U is nonincreasing outside the set $S\triangleq\{t\geq t_0:\mathrm{U}(t)\leq\eta(2b(t))\}$. Thus, if some $t^*\geq t_0$ belongs to S, then so does every $t\in[t^*,t_{\max}(\mathrm{U}_0,t_0))$ implying that U is locally bounded because b is locally bounded (in fact, continuous). (Note that $\mathrm{U}(t)\geq 0$ because $\dot{\mathrm{U}}(t)\geq 0$ whenever $\mathrm{U}(t)=0$.) Therefore, for all $t\geq t_0$, and for $\tilde{\mu}$ as in the statement (which we have shown is well defined), we have that

$$\bar{\mathrm{V}}(t)\leq\mathrm{U}(t)\leq\max\Big\{\tilde{\mu}\Big(\bar{\mathrm{V}}(t_0),t-t_0\Big),\eta\Big(2\max_{t_0\leq s\leq t}b(s)\Big)\Big\}.$$

Since the maximum of two quantities is upper bounded by the sum, and using the definition of b together with the monotonicity of σ, it follows that

$$\bar{\mathrm{V}}(t)\leq\mathrm{U}(t)\leq\tilde{\mu}\Big(\mathrm{V}(x_0),t-t_0\Big)+\eta\Big(2\sigma\Big(\max_{t_0\leq s\leq t}\|\Sigma(s)\|_{\mathcal{F}}\Big)\Big),\tag{7.23}$$

for all $t\geq t_0$, where we also used that $\bar{\mathrm{V}}(t_0)=\mathrm{V}(x_0)$, and the proof is complete. □

Of particular interest to us is the case when the function V is lower and upper bounded by class $\mathcal{K}_\infty$ functions of the distance to a closed, not necessarily bounded, set.

Definition 2.52. (NSS-Lyapunov functions). *A function* $\mathrm{V}\in\mathcal{C}^2(\mathbb{R}^n;\mathbb{R}_{\geq 0})$ *is a* strong NSS-Lyapunov function in probability with respect to $\mathcal{U} \subseteq \mathbb{R}^n$ *for* (7.1) *if* V *is a noise-dissipative Lyapunov function and, in addition, there exist* $p > 0$ *and class* $\mathcal{K}_\infty$ *functions* α_1 *and* α_2 *such that*

$$\alpha_1(|x|_{\mathcal{U}}^p) \leq \mathrm{V}(x) \leq \alpha_2(|x|_{\mathcal{U}}^p), \quad \forall x \in \mathbb{R}^n. \tag{7.24}$$

If, moreover, α_1 *is convex, then* V *is a* pth moment NSS-Lyapunov function with respect to $\mathcal{U}$.

Note that a strong NSS-Lyapunov function in probability with respect to a set satisfies an inequality of the type (7.24) for any $p > 0$, whereas the choice of p is relevant when α_1 is required to be convex. The reason for the 'strong' terminology is that we require (7.8) to be satisfied with convex $\eta^{-1} \in \mathcal{K}_\infty$. Instead, a standard NSS-Lyapunov function in probability satisfies the same inequality with a class $\mathcal{K}_\infty$ function which is not necessarily convex. We also note that (7.24) implies that $\mathcal{U} = \{x \in \mathbb{R}^n : \mathrm{V}(x) = 0\}$, which is closed because V is continuous.

Example 2.53. (Example 2.50–revisited: an NSS-Lyapunov function). Consider the function V introduced in Example 2.50. For each $p \in (0, 2]$, note that

$$\alpha_{1p}(\|x - x^*\|_2^p) \leq \mathrm{V}(x) \leq \alpha_{2p}(\|x - x^*\|_2^p) \qquad \forall x \in \mathbb{R}^n,$$

for the convex functions $\alpha_{1p}(r) = \alpha_{2p}(r) \triangleq r^{2/p}$, which are in the class $\mathcal{K}_\infty$. (Recall that α_2 in Definition 2.52 is only required to be $\mathcal{K}_\infty$.) Thus, the function V is a pth moment NSS-Lyapunov function for (7.11) with respect to x^* for $p \in (0, 2]$. •

The notion of NSS-Lyapunov function plays a key role in establishing our main result on the stability of SDEs with persistent noise.

Corollary 2.54. (The existence of an NSS-Lyapunov function implies the corresponding NSS property). *Under Assumption 1.45, and further assuming that* Σ *is continuous, given a closed set* $\mathcal{U} \subset \mathbb{R}^n$*,*

(i) if $\mathrm{V} \in \mathcal{C}^2(\mathbb{R}^n;\mathbb{R}_{\geq 0})$ *is a strong NSS-Lyapunov function in probability with respect to* $\mathcal{U}$ *for* (7.1)*, then the system is NSS in probability with respect to* $\mathcal{U}$ *with gain functions*

$$\mu(r,s) \triangleq \alpha_1^{-1}\Big(\tfrac{2}{\epsilon}\tilde{\mu}(\alpha_2(r^p),s)\Big), \quad \theta(r) \triangleq \alpha_1^{-1}\Big(\tfrac{2}{\epsilon}\eta(2\sigma(r))\Big); \tag{7.25}$$

(ii) if $\mathrm{V} \in \mathcal{C}^2(\mathbb{R}^n;\mathbb{R}_{\geq 0})$ *is a pthNSS-Lyapunov function with respect to* $\mathcal{U}$ *for* (7.1)*, then the system is pth moment NSS with respect to* $\mathcal{U}$ *with gain functions* μ *and* θ *as in* (7.25) *setting* $\epsilon = 1$*.*

Proof. To show *(i)*, note that, since $\alpha_1(|x|^p_{\mathcal{U}}) \leq \mathrm{V}(x)$ for all $x \in \mathbb{R}^n$, with $\alpha_1 \in \mathcal{K}_\infty$, it follows that for any $\hat{\rho} > 0$ and $t \geq t_0$,

$$\begin{aligned}\mathbb{P}\Big\{|x(t)|^p_{\mathcal{U}} > \hat{\rho}\Big\} = \mathbb{P}\Big\{\alpha_1(|x(t)|^p_{\mathcal{U}}) > \alpha_1(\hat{\rho})\Big\} \leq \mathbb{P}\Big\{\mathrm{V}(x(t)) > \alpha_1(\hat{\rho})\Big\} \leq \frac{\mathbb{E}\Big[\mathrm{V}(x(t))\Big]}{\alpha_1(\hat{\rho})} \\ \leq \frac{1}{\alpha_1(\hat{\rho})}\bigg(\tilde{\mu}\Big(\alpha_2(|x_0|^p_{\mathcal{U}}), t - t_0\Big) + \eta\Big(2\sigma\Big(\max_{t_0 \leq s \leq t}\|\Sigma(s)\|_{\mathcal{F}}\Big)\Big)\bigg),\end{aligned} \tag{7.26}$$

where we have used the strict monotonicity of α_1 in the first equation, Chebyshev's inequality [Bor95, Ch. 3] in the second inequality, and the upper bound for $\mathbb{E}\Big[\mathrm{V}(x(t))\Big]$ obtained in Theorem 2.51, cf. (7.12), in the last inequality (leveraging the monotonicity of $\tilde{\mu}$ in the first argument and the fact that $\mathrm{V}(x) \leq \alpha_2(|x|^p_{\mathcal{U}})$ for all $x \in \mathbb{R}^n$). Also, for any function $\alpha \in \mathcal{K}$, we have that $\alpha(2r) + \alpha(2s) \geq \alpha(r+s)$

for all $r, s \geq 0$. Thus,

$$\begin{aligned}\rho(\epsilon, x_0, t) &\triangleq \mu\Big(|x_0|_{\mathcal{U}}, t-t_0\Big) + \theta\Big(\max_{t_0 \leq s \leq t} \|\Sigma(s)\|_{\mathcal{F}}\Big) \qquad (7.27)\\ &\geq \alpha_1^{-1}\Bigg(\frac{1}{\epsilon}\tilde{\mu}\Big(\alpha_2(|x_0|_{\mathcal{U}}^p), t-t_0\Big) + \frac{1}{\epsilon}\eta\Big(2\sigma\Big(\max_{t_0 \leq s \leq t}\|\Sigma(s)\|_{\mathcal{F}}\Big)\Big)\Bigg) \triangleq \hat{\rho}(\epsilon).\end{aligned}$$

Substituting now $\hat{\rho} \triangleq \hat{\rho}(\epsilon)$ in (7.26), and using that $\rho(\epsilon, x_0, t) \geq \hat{\rho}(\epsilon)$, we get that $\mathbb{P}\Big\{|x(t)|_{\mathcal{U}}^p > \rho(\epsilon, x_0, t)\Big\} \leq \mathbb{P}\Big\{|x(t)|_{\mathcal{U}}^p > \hat{\rho}(\epsilon)\Big\} \leq \epsilon$.

To show *(ii)*, since α_1^{-1} is concave, applying Jensen's inequality [Bor95, Ch. 3], we get

$$\mathbb{E}\Big[|x(t)|_{\mathcal{U}}^p\Big] \leq \mathbb{E}\Big[\alpha_1^{-1}\Big(\mathrm{V}(x(t))\Big)\Big] \leq \alpha_1^{-1}\Big(\mathbb{E}\Big[\mathrm{V}(x(t))\Big]\Big) \leq \hat{\rho}(1) \leq \rho(1, x_0, t),$$

where in the last two inequalities we have used the bound for $\mathbb{E}\Big[\mathrm{V}(x(t))\Big]$ in (7.26) and the definition of $\hat{\rho}(\epsilon)$ in (7.27). □

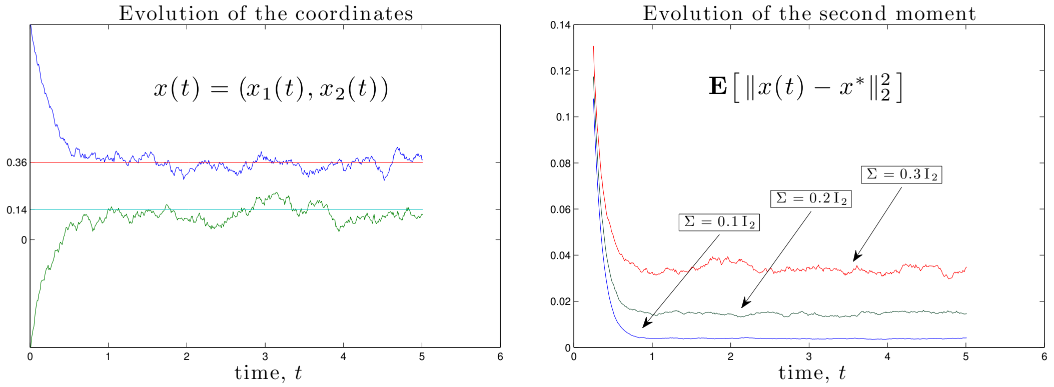
Example 2.55. (Example 2.50–revisited: illustration of Corollary 2.54). Consider again Example 2.50. Since V is a pth moment NSS-Lyapunov function for (7.11) with respect to the point x^* for $p \in (0, 2]$, as shown in Example 2.53, Corollary 2.54 implies that

$$\mathbb{E}\Big[\|x - x^*\|_2^p\Big] \leq \mu\Big(\|x_0 - x^*\|_2, t - t_0\Big) + \theta\Big(\max_{t_0 \leq s \leq t} \|\Sigma(s)\|_{\mathcal{F}}\Big), \qquad (7.28)$$

for all $t \geq t_0$, $x_0 \in \mathbb{R}^n$, and $p \in (0, 2]$, where

$$\mu(r, s) = \Big(2\tilde{\mu}(r^2, s)\Big)^{p/2}, \quad \theta(r) = \Big(\gamma^{-1}(r^2)\Big)^{p/2},$$

and the class $\mathcal{KL}$ function $\tilde{\mu}$ is defined as the solution to the initial value prob-



(a) (b)

Figure 7.1: Simulation example of the notion of noise to state stability in second moment. Evolution of the dynamics (7.11) with $\mathsf{L} = 0$, $h(x_1, x_2) = \log\big(e^{(x_1-2)} + e^{(x_2+1)}\big) + 0.5(x_1 + x_2 - 1)^2 + (x_1 - x_2)^2$, and initial condition $[x_1(0), x_2(0)] = (1, -0.5)$. Since h is a sum of convex functions, and the Hessian of the quadratic part of h has eigenvalues $\{2, 4\}$, we can take γ given by $\gamma(r) = 2r$, for $r \geq 0$. Plot (a) shows the evolution of the first and second coordinates with $\Sigma = 0.1\,\mathrm{I}_2$. Plot (b) illustrates the noise-to-state stability property in second moment with respect to $x^* = (0.36, 0.14)$, where the matrix $\Sigma(t)$ is a constant multiple of the identity. (The expectation is computed averaging over 500 realizations of the noise.)

lem (7.13) with $\eta(r) = \frac{1}{2}\gamma^{-1}(r)$. Figure 7.1 illustrates this noise-to-state stability property. We note that if the function h is strongly convex, i.e., if $\gamma(r) = c_\gamma r$ for some constant $c_\gamma > 0$, then $\tilde{\mu} : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ becomes $\tilde{\mu}(r, s) = re^{-c_\gamma s}$, and $\mu(r, s) = 2^{p/2}\, r^p e^{-c_\gamma p/2\, s}$, so the bound for $\mathbb{E}\Big[\|x - x^*\|_2^{p}\Big]$ in (7.28) decays exponentially with time to $\theta\Big(\max_{t_0 \leq s \leq t} \|\Sigma(s)\|_{\mathcal{F}}\Big)$. •

7.3 Refinements of the notion of proper functions

In this section, we analyze in detail the inequalities between functions that appear in the definition of noise-dissipative Lyapunov function, strong NSS-Lyapunov function in probability, and pth moment NSS-Lyapunov function. In

Section 7.3.1, we establish that these inequalities can be regarded as equivalence relations. In Section 7.3.2, we make a complete characterization of the properties of two functions related by these equivalence relations. Finally, in Section 7.3.3, these results lead us to obtain an alternative formulation of Corollary 2.54.

7.3.1 Proper functions and equivalence relations

Here, we provide a refinement of the notion of proper functions with respect to each other. Proper functions play an important role in stability analysis, see e.g., [Kha02, Son08].

Definition 3.56. (Refinements of the notion of proper functions with respect to each other). *Let $\mathcal{D} \subseteq \mathbb{R}^n$ and the functions* $\mathrm{V}, \mathrm{W} : \mathcal{D} \to \mathbb{R}_{\geq 0}$ *be such that*

$$\alpha_1(\mathrm{W}(x)) \leq \mathrm{V}(x) \leq \alpha_2(\mathrm{W}(x)), \quad \forall x \in \mathcal{D},$$

for some functions $\alpha_1, \alpha_2 : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. Then,

(i) if $\alpha_1, \alpha_2 \in \mathcal{K}$, we say that V *is* $\mathcal{K}$ - dominated by W *in $\mathcal{D}$, and write* $\mathrm{V} \lhd^{\mathcal{K}} \mathrm{W}$ *in $\mathcal{D}$;*

(ii) if $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, we say that V *and* W *are* $\mathcal{K}_\infty$ - proper with respect to each other in $\mathcal{D}$, *and write* $\mathrm{V} \sim^{\mathcal{K}_\infty} \mathrm{W}$ *in $\mathcal{D}$;*

(iii) if $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ are convex and concave, respectively, we say that V *and* W *are* $\mathcal{K}_\infty^{cc}$ - (convex-concave) proper with respect to each other in $\mathcal{D}$, *and write* $\mathrm{V} \sim^{\mathcal{K}_\infty^{cc}} \mathrm{W}$ *in $\mathcal{D}$;*

(iv) if $\alpha_1(r) \triangleq c_{\alpha_1} r$ and $\alpha_2(r) \triangleq c_{\alpha_2} r$, for some constants $c_{\alpha_1}, c_{\alpha_2} > 0$, we say that V *and* W *are* equivalent in $\mathcal{D}$, *and write* $\mathrm{V} \sim \mathrm{W}$ *in $\mathcal{D}$.*

Note that the relations in Definition 3.56 are nested, i.e., given $\mathrm{V},\mathrm{W} : \mathcal{D} \to \mathbb{R}_{\geq 0}$, the following chain of implications hold in $\mathcal{D}$:

$$\mathrm{V} \sim \mathrm{W} \Rightarrow \mathrm{V} \sim^{\mathcal{K}_\infty^{cc}} \mathrm{W} \Rightarrow \mathrm{V} \sim^{\mathcal{K}_\infty} \mathrm{W} \Rightarrow \mathrm{V} \lhd^{\mathcal{K}} \mathrm{W}. \tag{7.29}$$

Also, note that if $\mathrm{W}(x) = \|x\|_2$, $\mathcal{D}$ is a neighborhood of 0, and α_1, α_2 are class $\mathcal{K}$, then we recover the notion of V being a *proper* function [Kha02]. If $\mathcal{D} = \mathbb{R}^n$, and V and W are seminorms, then the relation $\sim$ corresponds to the concept of equivalent seminorms.

The relation $\sim^{\mathcal{K}_\infty}$ is relevant for ISS and NSS in probability, whereas the relation $\sim^{\mathcal{K}_\infty^{cc}}$ is important for pth moment NSS. The latter is because the inequalities in $\sim^{\mathcal{K}_\infty^{cc}}$ are still valid, thanks to Jensen inequality, if we substitute V and W by their expectations along a stochastic process. Another fact about the relation $\sim^{\mathcal{K}_\infty^{cc}}$ is that $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, convex and concave, respectively, must be asymptotically linear if $\mathrm{V}(\mathcal{D}) \supseteq [s_0, \infty)$, for some $s_0 \geq 0$, so that $\alpha_1(s) \leq \alpha_2(s)$ for all $s \geq s_0$. This follows from Lemma 0.66.

Remark 3.57. (Quadratic forms in a constrained domain). It is sometimes convenient to view the functions $\mathrm{V},\mathrm{W} : \mathcal{D} \to \mathbb{R}_{\geq 0}$ as defined in a domain where their functional expression becomes simpler. To make this idea precise, assume there exist $i : \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}^m$, with $m \geq n$, and $\hat{\mathrm{V}}, \hat{\mathrm{W}} : \hat{\mathcal{D}} \to \mathbb{R}_{\geq 0}$, where $\hat{\mathcal{D}} = i(\mathcal{D})$, such that $\mathrm{V} = \hat{\mathrm{V}} \circ i$ and $\mathrm{W} = \hat{\mathrm{W}} \circ i$. If this is the case, then the existence of $\alpha_1, \alpha_2 : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that $\alpha_1\Big(\hat{\mathrm{W}}(\hat{x})\Big) \leq \hat{\mathrm{V}}(\hat{x}) \leq \alpha_2\Big(\hat{\mathrm{W}}(\hat{x})\Big)$, for all $\hat{x} \in \hat{\mathcal{D}}$, implies that $\alpha_1\Big(\mathrm{W}(x)\Big) \leq \mathrm{V}(x) \leq \alpha_2\Big(\mathrm{W}(x)\Big)$, for all $x \in \mathcal{D}$. The reason is that for any $x \in \mathcal{D}$ there exists $\hat{x} \in \hat{\mathcal{D}}$, given by $\hat{x} = i(x)$, such that $\mathrm{V}(x) = \hat{\mathrm{V}}(\hat{x})$ and $\mathrm{W}(x) = \hat{\mathrm{W}}(\hat{x})$, so

$$\alpha_1\Big(\mathrm{W}(x)\Big) = \alpha_1\Big(\hat{\mathrm{W}}(\hat{x})\Big) \leq \mathrm{V}(x) = \hat{\mathrm{V}}(\hat{x}) \leq \alpha_2\Big(\hat{\mathrm{W}}(\hat{x})\Big) = \alpha_2\Big(\mathrm{W}(x)\Big).$$

Consequently, if any of the relations given in Definition 3.56 is satisfied by $\hat{\mathrm{V}}$, $\hat{\mathrm{W}}$ in $\hat{\mathcal{D}}$, then the corresponding relation is satisfied by V, W in $\mathcal{D}$. For instance, in some scenarios this procedure can allow us to rewrite the original functions V, W as quadratic forms $\hat{\mathrm{V}}$, $\hat{\mathrm{W}}$ in a constrained set of an extended Euclidean space, where it is easier to establish the appropriate relation between the functions. We make use of this observation in Section 7.3.3 below. •

Lemma 3.58. (Powers of seminorms with the same nullspace). *Let A and B in $\mathbb{R}^{m\times n}$ be nonzero matrices with the same nullspace, $\mathcal{N}(A)=\mathcal{N}(B)$. Then, for any $p,q>0$, the inequalities $\alpha_1\big(\|x\|_A^p\big) \leq \|x\|_B^q \leq \alpha_2\big(\|x\|_A^p\big)$ are verified with*

$$\alpha_1(r) \triangleq \left(\frac{\lambda_{n-k}(B^\top B)}{\lambda_{max}(A^\top A)}\right)^{\frac{q}{2}} r^{q/p}; \quad \alpha_2(r) \triangleq \left(\frac{\lambda_{max}(B^\top B)}{\lambda_{n-k}(A^\top A)}\right)^{\frac{q}{2}} r^{q/p},$$

where $k \triangleq \dim(\mathcal{N}(A))$. In particular, $\|.\|_A^p \sim \|.\|_B^p$ and $\|.\|_A^p \sim^{\mathcal{K}_\infty} \|.\|_B^q$ in $\mathbb{R}^n$ for any real numbers $p,q>0$.

Proof. For $\mathcal{U} \triangleq \mathcal{N}(A)$, write any $x\in\mathbb{R}^n$ as $x = x_{\mathcal{U}} + x_{\mathcal{U}^\perp}$, where $x_{\mathcal{U}} \in \mathcal{U}$ and $x_{\mathcal{U}^\perp} \in \{x\in\mathbb{R}^n : x^\top u = 0 ,\forall u \in \mathcal{U}\}$, so that $Ax = A(x_{\mathcal{U}} + x_{\mathcal{U}^\perp}) = Ax_{\mathcal{U}^\perp}$ and $Bx = Bx_{\mathcal{U}^\perp}$ because $\mathcal{N}(A) = \mathcal{N}(B) = \mathcal{U}$. Using the formulas for the eigenvalues in [HJ85, p. 178], we see that the next chain of inequalities hold:

$$\begin{aligned}
\alpha_1\Big(\|x\|_A^p\Big) &= \alpha_1\Big(\big(x_{\mathcal{U}^\perp}^\top A^\top A x_{\mathcal{U}^\perp}\big)^{\frac{p}{2}}\Big) \leq \alpha_1\Big(\big(\lambda_{\max}(A^\top A) x_{\mathcal{U}^\perp}^\top x_{\mathcal{U}^\perp}\big)^{\frac{p}{2}}\Big)\\
&\leq \big(\lambda_{n-k}(B^\top B) x_{\mathcal{U}^\perp}^\top x_{\mathcal{U}^\perp}\big)^{\frac{q}{2}} \leq \big(x_{\mathcal{U}^\perp}^\top B^\top B x_{\mathcal{U}^\perp}\big)^{\frac{q}{2}} \leq \big(\lambda_{\max}(B^\top B) x_{\mathcal{U}^\perp}^\top x_{\mathcal{U}^\perp}\big)^{\frac{q}{2}}\\
&\leq \alpha_2\Big(\big(\lambda_{n-k}(A^\top A) x_{\mathcal{U}^\perp}^\top x_{\mathcal{U}^\perp}\big)^{\frac{p}{2}}\Big) \leq \alpha_2\Big(\big(x_{\mathcal{U}^\perp}^\top A^\top A x_{\mathcal{U}^\perp}\big)^{\frac{p}{2}}\Big) = \alpha_2\Big(\|x\|_A^p\Big),
\end{aligned}$$

where $\|x\|_B^q = \big(x_{\mathcal{U}^\perp}^\top B^\top B x_{\mathcal{U}^\perp}\big)^{\frac{q}{2}}$. From this we conclude that $\|.\|_A^p \sim^{\mathcal{K}_\infty} \|.\|_B^q$ in $\mathbb{R}^n$. Finally, when $p=q$, the class $\mathcal{K}_\infty$ functions α_1, α_2 in the statement are linear, so we obtain that $\|.\|_A^p \sim \|.\|_B^p$ in $\mathbb{R}^n$. □

Next we show that $\sim^{\mathcal{K}_\infty}$ and $\sim^{\mathcal{K}_\infty^{cc}}$ are reflexive, symmetric, and transitive, and hence define equivalence relations.

Lemma 3.59. (The $\mathcal{K}_\infty$- and $\mathcal{K}_\infty^{cc}$-proper relations are equivalence relations). *The relations $\sim^{\mathcal{K}_\infty}$ and $\sim^{\mathcal{K}_\infty^{cc}}$ in any set $\mathcal{D} \subseteq \mathbb{R}^n$ are both equivalence relations.*

Proof. For convenience, we represent both relations by $\sim^*$. Both are reflexive, i.e., $\mathrm{V} \sim^* \mathrm{V}$, because one can take $\alpha_1(r) = \alpha_2(r) = r$ noting that a linear function is both convex and concave. Both are symmetric, i.e., $\mathrm{V} \sim^* \mathrm{W}$ if and only if $\mathrm{W} \sim^* \mathrm{V}$, because if $\alpha_1 \circ \mathrm{W} \leq \mathrm{V} \leq \alpha_2 \circ \mathrm{W}$ in $\mathcal{D}$, then $\alpha_2^{-1} \circ \mathrm{V} \leq \mathrm{W} \leq \alpha_1^{-1} \circ \mathrm{V}$ in $\mathcal{D}$. In the case of $\sim^{\mathcal{K}_\infty}$, the inverse of a class $\mathcal{K}_\infty$ function is class $\mathcal{K}_\infty$. Additionally, in the case of $\sim^{\mathcal{K}_\infty^{cc}}$, if $\alpha \in \mathcal{K}_\infty$ is convex (respectively, concave), then $\alpha^{-1} \in \mathcal{K}_\infty$ is concave (respectively, convex). Finally, both are transitive, i.e., $\mathrm{U} \sim^* \mathrm{V}$ and $\mathrm{V} \sim^* \mathrm{W}$ imply $\mathrm{U} \sim^* \mathrm{W}$, because if $\alpha_1 \circ \mathrm{V} \leq \mathrm{U} \leq \alpha_2 \circ \mathrm{V}$ and $\tilde{\alpha}_1 \circ \mathrm{W} \leq \mathrm{V} \leq \tilde{\alpha}_2 \circ \mathrm{W}$ in $\mathcal{D}$, then $\alpha_1 \circ \tilde{\alpha}_1 \circ \mathrm{W} \leq \mathrm{U} \leq \alpha_2 \circ \tilde{\alpha}_2 \circ \mathrm{W}$ in $\mathcal{D}$. In the case of $\sim^{\mathcal{K}_\infty}$, the composition of two class $\mathcal{K}_\infty$ functions is class $\mathcal{K}_\infty$. Additionally, in the case of $\sim^{\mathcal{K}_\infty^{cc}}$, if $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ are both convex (respectively, concave), then the compositions $\alpha_1 \circ \alpha_2$ and $\alpha_2 \circ \alpha_1$ belong to $\mathcal{K}_\infty$ and are convex (respectively, concave), as explained in Section 2.2.1. □

Remark 3.60. (The relation $\lhd^{\mathcal{K}}$ is not an equivalence relation). The proof above also shows that the relation $\lhd^{\mathcal{K}}$ is reflexive and transitive. However, it is not symmetric: consider $\mathrm{V}, \mathrm{W} \in \mathcal{C}(\mathbb{R}^n; \mathbb{R}_{\geq 0})$ given by $\mathrm{V}(x) = 1 - e^{-\|x\|_2}$ and $\mathrm{W}(x) = \|x\|_2$. Clearly, $\mathrm{V} \lhd^{\mathcal{K}} \mathrm{W}$ in $\mathbb{R}^n$ by taking $\alpha_1 = \alpha_2 = \alpha \in \mathcal{K}$, with $\alpha(s) = 1 - e^{-s}$. On the other hand, if there exist $\tilde{\alpha}_1, \tilde{\alpha}_2 \in \mathcal{K}$ such that $\tilde{\alpha}_1(\mathrm{V}(x)) \leq \mathrm{W}(x) \leq \tilde{\alpha}_2(\mathrm{V}(x))$ for all $x \in \mathbb{R}^n$, then we reach the contradiction, by continuity of $\tilde{\alpha}_2$, that $\lim_{\|x\|_2 \to \infty} \|x\|_2 \leq \tilde{\alpha}_2\Big(\lim_{\|x\|_2 \to \infty} \Big(1 - e^{-\|x\|_2}\Big)\Big) = \tilde{\alpha}_2(1) < \infty$. •

7.3.2 Characterization of proper functions with respect to each other

In this section, we provide a complete characterization of the properties that two functions must satisfy to be related by the equivalence relations defined in Section 7.3.1. For $\mathcal{D} \subseteq \mathbb{R}^n$, consider $\mathrm{V}_1, \mathrm{V}_2 : \mathcal{D} \to \mathbb{R}_{\geq 0}$. Given a real number $p > 0$, define

$$\begin{aligned}\phi_p(s) &\triangleq \sup_{\{y \in \mathcal{D} : \mathrm{V}_2(y) \leq \sqrt[p]{s}\}} \mathrm{V}_1(y),\\ \psi_p(s) &\triangleq \inf_{\{y \in \mathcal{D} : \mathrm{V}_2(y) \geq \sqrt[p]{s}\}} \mathrm{V}_1(y),\end{aligned}$$

for $s \geq 0$. The value $\phi_p(s)$ gives the supremum of the function V_1 in the $\sqrt[p]{s}$ - sublevel set of V_2, and $\psi_p(s)$ is the infimum of V_1 in the $\sqrt[p]{s}$ - superlevel set of V_2. Thus, the functions ϕ_p and ψ_p satisfy

$$\psi_p\Big(\mathrm{V}_2(x)^p\Big) = \inf_{\substack{\{y \in \mathcal{D} :\\ \mathrm{V}_2(y) \geq \mathrm{V}_2(x)\}}} \mathrm{V}_1(y) \leq \mathrm{V}_1(x) \leq \sup_{\substack{\{y \in \mathcal{D} :\\ \mathrm{V}_2(y) \leq \mathrm{V}_2(x)\}}} \mathrm{V}_1(y) = \phi_p\Big(\mathrm{V}_2(x)^p\Big), \tag{7.30}$$

for all $x \in \mathcal{D}$, which suggests ϕ_p and ψ_p as pre-comparison functions to construct α_1 and α_2 in Definition 3.56. To this end, we find it useful to formulate the following properties of the function V_1 with respect to V_2:

P0: The set $\{x \in \mathcal{D} : \mathrm{V}_2(x) = s\}$ is nonempty for all $s \geq 0$.

P1: The nullsets of V_1 and V_2 are the same, i.e., $\{x \in \mathcal{D} : \mathrm{V}_1(x) = 0\} = \{x \in \mathcal{D} : \mathrm{V}_2(x) = 0\}$.

P2: The function ϕ_1 is locally bounded in $\mathbb{R}_{\geq 0}$ and right continuous at 0, and ψ_1 is positive definite.

P3: The next limit holds: $\lim_{s\to\infty}\psi_1(s)=\infty$.

P4 (as a function of $p>0$): The asymptotic behavior of ϕ_p and ψ_p is such that $\phi_p(s)$ and $s^2/\psi_p(s)$ are both in $\mathcal{O}(s)$ as $s\to\infty$.

The next result shows that these properties completely characterize whether the functions V_1 and V_2 are related through the equivalence relations defined in Section 7.3.1. This result generalizes [Kha02, Lemma 4.3] in several ways: the notions of proper functions considered here are more general and are not necessarily restricted to a relationship between an arbitrary function and the distance to a compact set.

Theorem 3.61. (Characterizations of proper functions with respect to each other). *Let* $\mathrm{V}_1,\mathrm{V}_2:\mathcal{D}\to\mathbb{R}_{\geq 0}$*, and assume* V_2 *satisfies* P0*. Then*

(i) V_1 *satisfies* $\{\mathrm{P}i\}_{i=1}^2$ *with respect to* V_2 $\Leftrightarrow$ $\mathrm{V}_1 \lhd^{\mathcal{K}} \mathrm{V}_2$ *in* $\mathcal{D}$*;*

(ii) V_1 *satisfies* $\{\mathrm{P}i\}_{i=1}^3$ *with respect to* V_2 $\Leftrightarrow$ $\mathrm{V}_1 \sim^{\mathcal{K}_\infty} \mathrm{V}_2$ *in* $\mathcal{D}$*;*

(iii) V_1 *satisfies* $\{\mathrm{P}i\}_{i=1}^4$ *with respect to* V_2 *for* $p>0$ $\Leftrightarrow$ $\mathrm{V}_1 \sim^{\mathcal{K}^{cc}_\infty} \mathrm{V}_2^p$ *in* $\mathcal{D}$*.*

Proof. We begin by establishing a few basic facts about the pre-comparison functions ψ_p and ϕ_p. By definition and by P0, it follows that $0\leq\psi_1(s)\leq\phi_1(s)$ for all $s\geq 0$. Since ϕ_1 is locally bounded by P2, then so is ψ_1. In particular, ϕ_1 and ψ_1 are well defined in $\mathbb{R}_{\geq 0}$. Moreover, both ϕ_1 and ψ_1 are nondecreasing because if $s_2\geq s_1$, then the supremum is taken in a larger set, $\{x\in\mathcal{D}:\mathrm{V}_2(x)\leq s_2\}\supseteq\{x\in\mathcal{D}:\mathrm{V}_2(x)\leq s_1\}$, and the infimum is taken in a smaller set, $\{x\in\mathcal{D}:\mathrm{V}_2(x)\geq s_2\}\subseteq\{x\in\mathcal{D}:\mathrm{V}_2(x)\geq s_1\}$. Furthermore, for any $q>0$, the functions ϕ_q and ψ_q are also monotonic and positive definite because $\phi_q(s)=\phi_1(\sqrt[q]{s})$ and $\psi_q(s)=\psi_1(\sqrt[q]{s})$ for all $s\geq 0$. We now use these properties of the pre-comparison functions to construct α_1, α_2 in Definition 3.56 required by the implications from left to right in each statement.

Proof of *(i)* ($\Rightarrow$). To show the existence of $\alpha_2 \in \mathcal{K}$ such that $\alpha_2(s) \geq \phi_1(s)$ for all $s \in \mathbb{R}_{\geq 0}$, we proceed as follows. Since ϕ_1 is locally bounded and nondecreasing, given a strictly increasing sequence $\{b_k\}_{k\geq 1} \subseteq \mathbb{R}_{\geq 0}$ with $\lim_{k\to\infty} b_k = \infty$, we choose the sequence $\{M_k\}_{k\geq 1} \subseteq \mathbb{R}_{\geq 0}$, setting $M_0 = 0$, in the following way:

$$M_k \triangleq \max\Big\{ \sup_{s\in[0,b_k]} \phi_1(s)\,,\, M_{k-1} + 1/k^2 \Big\} = \max\Big\{ \phi_1(b_k)\,,\, M_{k-1} + 1/k^2 \Big\}. \tag{7.31}$$

This choice guarantees that $\{M_k\}_{k\geq 1}$ is strictly increasing and, for each $k \geq 1$,

$$0 \leq M_k - \phi_1(b_k) \leq \sum_{i=1}^{k} \frac{1}{i^2} \leq \pi^2/6. \tag{7.32}$$

Also, since ϕ_1 is right continuous at 0, we can choose $b_1 > 0$ such that there exists $\alpha_2 : [0, b_1] \to \mathbb{R}_{\geq 0}$ continuous, positive definite and strictly increasing, satisfying that $\alpha_2(s) \geq \phi_1(s)$ for all $s \in [0, b_1]$ and with $\alpha_2(b_1) = M_2$. (This is possible because the only function that cannot be upper bounded by an arbitrary continuous function in some arbitrarily small interval $[0, b_1]$ is the function that has a jump at 0.) The rest of the construction is explicit. We define α_2 as a piecewise linear function in (b_1, ∞) in the following way: for each $k \geq 2$, we define

$$\alpha_2(s) \triangleq \alpha_2(b_{k-1}) + \frac{M_{k+1} - \alpha_2(b_{k-1})}{b_k - b_{k-1}}\,(s - b_{k-1}), \qquad \forall s \in (b_{k-1}, b_k].$$

The resulting α_2 is continuous by construction. Also, $\alpha_2(b_1) = M_2$, so that, inductively, $\alpha_2(b_{k-1}) = M_k$ for $k \geq 2$. Two facts now follow: first, $M_{k+1} - \alpha_2(b_{k-1}) = M_{k+1} - M_k \geq 1/(k+1)^2$ for $k \geq 2$, so α_2 has positive slope in each interval $(b_{k-1}, b_k]$ and thus is strictly increasing in (b_1, ∞); second, $\alpha_2(s) > \alpha_2(b_{k-1}) = M_k \geq \phi_1(b_k) \geq \phi_1(s)$ for all $s \in (b_{k-1}, b_k]$, for each $k \geq 2$, so $\alpha_2(s) \geq \phi_1(s)$ for all $s \in (b_1, \infty)$.

We have left to show the existence of $\alpha_1 \in \mathcal{K}$ such that $\alpha_1(s) \leq \psi_1(s)$ for all

$s \in \mathbb{R}_{\geq 0}$. First, since $0 \leq \psi_1(s) \leq \phi_1(s)$ for all $s \geq 0$ by definition and by P0, using the sandwich theorem [LL88, p. 107], we derive that ψ_1 is right continuous at 0 the same as ϕ_1. In addition, since ψ_1 is nondecreasing, it can only have a countable number of jump discontinuities (none of them at 0). Therefore, we can pick $c_1 > 0$ such that a continuous and nondecreasing function $\hat{\psi}_1$ can be constructed in $[0, c_1)$ by removing the jumps of ψ_1, so that $\hat{\psi}_1(s) \leq \psi_1(s)$. Moreover, since ψ_1 is positive definite and right continuous at 0, then $\hat{\psi}_1$ is also positive definite. Thus, there exists α_1 in $[0, c_1)$ continuous, positive definite, and strictly increasing, such that, for some $r < 1$,

$$\alpha_1(s) \leq r\hat{\psi}_1(s) \leq r\psi_1(s) \tag{7.33}$$

for all $s \in [0, c_1)$. To extend α_1 to a function in class $\mathcal{K}$ in $\mathbb{R}_{\geq 0}$, we follow a similar strategy as for α_2. Given a strictly increasing sequence $\{c_k\}_{k \geq 2} \subseteq \mathbb{R}_{\geq 0}$ with $\lim_{k \to \infty} c_k = \infty$, we define a sequence $\{m_k\}_{k \geq 1} \subseteq \mathbb{R}_{\geq 0}$ in the following way:

$$m_k \triangleq \inf_{s \in [c_k, c_{k+1})} \psi_1(s) - \tfrac{\psi_1(c_1) - \alpha_1(c_1)}{1+k^2} = \psi_1(c_k) - \tfrac{\psi_1(c_1) - \alpha_1(c_1)}{1+k^2}. \tag{7.34}$$

Next we define α_1 in $[c_1, \infty)$ as the piecewise linear function

$$\alpha_1(s) \triangleq \alpha_1(c_k) + \frac{m_k - \alpha_1(c_k)}{c_{k+1} - c_k}(s - c_k), \qquad \forall s \in [c_k, c_{k+1}),$$

for all $k \geq 1$, so α_1 is continuous by construction. It is also strictly increasing because $\alpha_1(c_2) = m_1 = (\psi_1(c_1) + \alpha_1(c_1))/2 > \alpha_1(c_1)$ by (7.33), and also, for each $k \geq 2$, the slopes are positive because $m_k - \alpha_1(c_k) = m_k - m_{k-1} > 0$ (due to the fact that $\{m_k\}_{k \geq 1}$ in (7.34) is strictly increasing because ψ_1 is nondecreasing). Finally, $\alpha_1(s) < \alpha_1(c_{k+1}) = m_k < \psi_1(c_k) \leq \psi_1(s)$ for all $s \in [c_k, c_{k+1})$, for all $k \geq 1$ by (7.34).

Equipped with α_1, α_2 as defined above, and as a consequence of (7.30), we have that

$$\alpha_1(\mathrm{V}_2(x)) \leq \psi_1(\mathrm{V}_2(x)) \leq \mathrm{V}_1(x) \leq \phi_1(\mathrm{V}_2(x)) \leq \alpha_2(\mathrm{V}_2(x)), \quad \forall x \in \mathcal{D}. \tag{7.35}$$

This concludes the proof of *(i)* ($\Rightarrow$).

As a preparation for *(ii)-(iii)* ($\Rightarrow$), and assuming P3, we derive two facts regarding the functions α_1 and α_2 constructed above. First, we establish that

$$\alpha_2(s) \in \mathcal{O}(\phi_1(s)) \text{ as } s \to \infty. \tag{7.36}$$

To show this, we argue that

$$\lim_{k\to\infty} \sup_{s\in(b_{k-1},b_k]} \Big(\alpha_2(s) - \phi_1(s)\Big) \leq \lim_{k\to\infty} \Big(\phi_1(b_{k+1}) - \phi_1(b_{k-1})\Big) + \pi^2/6, \tag{7.37}$$

so that there exist $C, s_1 > 0$ such that $\alpha_2(s) \leq 3\phi_1(s) + C$, for all $s \geq s_1$. Thus, noting that $\lim_{s\to\infty} \phi_1(s) = \infty$ as a consequence of P3, the expression (7.36) follows. To establish (7.37), we use the monotonicity of α_2 and ϕ_1, (7.31) and (7.32). For $k \geq 2$,

$$\begin{aligned}
&\sup_{s\in(b_{k-1},b_k]} \Big(\alpha_2(s) - \phi_1(s)\Big) \leq \alpha_2(b_k) - \phi_1(b_{k-1}) = M_{k+1} - \phi_1(b_{k-1}) \\
&= \max\Big\{\phi_1(b_{k+1}) - \phi_1(b_{k-1}),\, M_k + 1/(k+1)^2 - \phi_1(b_{k-1})\Big\} \\
&\leq \max\Big\{\phi_1(b_{k+1}) - \phi_1(b_{k-1}),\, \phi_1(b_k) + \pi^2/6 + 1/(k+1)^2 - \phi_1(b_{k-1})\Big\}.
\end{aligned}$$

Second, the construction of α_1 guarantees that

$$\psi_1(s) \in \mathcal{O}(\alpha_1(s)) \text{ as } s \to \infty, \tag{7.38}$$

because, as we show next,

$$\lim_{k\to\infty} \sup_{s\in[c_k,c_{k+1})} \Big(\psi_1(s)-\alpha_1(s)\Big) \leq \lim_{k\to\infty} \Big(\alpha_1(c_{k+2})-\alpha_1(c_k)\Big), \tag{7.39}$$

so there exists $s_2>0$ such that $\psi_1(s)\leq 3\alpha_1(s)$ for all $s\geq s_2$. To obtain (7.39), we leverage the monotonicity of ψ_1 and α_1, and (7.34); namely, for $k\geq 2$,

$$\begin{aligned}&\sup_{s\in[c_k,c_{k+1})} \Big(\psi_1(s)-\alpha_1(s)\Big) \leq \psi_1(c_{k+1})-\alpha_1(c_k)\\ &= m_{k+1}+\tfrac{\psi_1(c_1)-\alpha_1(c_1)}{1+(k+1)^2}-\alpha_1(c_k)=\alpha_1(c_{k+2})+\tfrac{\psi_1(c_1)-\alpha_1(c_1)}{1+(k+1)^2}-\alpha_1(c_k).\end{aligned}$$

Equipped with (7.36) and (7.38), we prove next *(ii)-(iii)* (⇒).

Proof of *(ii)* (⇒): If, in addition, P3 holds, then $\lim_{s\to\infty}\phi_1(s)\geq\lim_{s\to\infty}\psi_1(s)=\infty$. This guarantees that $\alpha_2\in\mathcal{K}_\infty$. Also, according to (7.38), P3 implies that α_1 is unbounded, and thus in $\mathcal{K}_\infty$ as well. The result now follows by (7.35).

Proof of *(iii)* (⇒): Finally, assume that P4 also holds for some $p>0$. We show next the existence of the required convex and concave functions involved in the relation $\sim^{\mathcal{K}_\infty^{cc}}$. Let $\alpha_{1,p}(s)\triangleq\alpha_1(\sqrt[p]{s})$ and $\alpha_{2,p}(s)\triangleq\alpha_2(\sqrt[p]{s})$ for $s\geq 0$, so that

$$\alpha_{1,p}(s)=\alpha_1(\sqrt[p]{s})\leq\psi_1(\sqrt[p]{s})=\psi_p(s) \quad\text{and}\quad \phi_p(s)=\phi_1(\sqrt[p]{s})\leq\alpha_2(\sqrt[p]{s})=\alpha_{2,p}(s).$$

From (7.36) and P4, it follows that there exist s', c_1, $c_2>0$ such that $\alpha_2(s)\leq c_1\phi_1(s)$ and $\phi_p(s)\leq c_2 s$ for all $s\geq s'$. Thus,

$$\alpha_{2,p}(s)=\alpha_2(\sqrt[p]{s})\leq c_1\phi_1(\sqrt[p]{s})=c_1\phi_p(s)\leq c_1c_2 s,$$

for all $s\geq s'$, so $\alpha_{2,p}(s)$ is in $\mathcal{O}(s)$ as $s\to\infty$. Similarly, according to (7.38) and P4, there are constants s'', c_3, $c_4>0$ such that $\psi_1(s)\leq c_3\,\alpha_1(s)$ and $s^2\leq c_4 s\,\psi_p(s)$ for all $s \geq s''$. Thus,

$$s\,\alpha_{1,p}(s) = s\,\alpha_1(\sqrt[p]{s}) \geq s\,\tfrac{1}{c_3}\psi_1(\sqrt[p]{s}) = s\,\tfrac{1}{c_3}\psi_p(s) \geq \tfrac{1}{c_3 c_4}s^2,$$

for all $s \geq s''$, so $s^2/\alpha_{1,p}(s)$ is in $\mathcal{O}(s)$ as $s \to \infty$. Summarizing, the construction of α_1, α_2 guarantees, under P4, that $\alpha_{1,p}$, $\alpha_{2,p}$ satisfy that $s^2/\alpha_{1,p}(s)$ and $\alpha_{2,p}(s)$ are in $\mathcal{O}(s)$ as $s \to \infty$ (and, as a consequence, so are $s^2/\alpha_{2,p}(s)$ and $\alpha_{1,p}(s)$). Therefore, according to Lemma 0.66, we can leverage (7.35) by taking $\tilde{\alpha}_1$, $\tilde{\alpha}_2 \in \mathcal{K}_\infty$, convex and concave, respectively, such that, for all $x \in \mathcal{D}$,

$$\begin{aligned}\tilde{\alpha}_1\Big(\mathrm{V}_2(x)^p\Big) \leq& \,\alpha_{1,p}(\mathrm{V}_2(x)^p) = \alpha_1(\mathrm{V}_2(x)) \leq \psi_1\Big(\mathrm{V}_2(x)\Big) \leq \mathrm{V}_1(x) \\ \leq& \,\phi_1\Big(\mathrm{V}_2(x)\Big) \leq \alpha_2(\mathrm{V}_2(x)) = \alpha_{2,p}(\mathrm{V}_2(x)^p) \leq \tilde{\alpha}_2\Big(\mathrm{V}_2(x)^p\Big).\end{aligned}$$

Proof of *(i)* ($\Leftarrow$): If there exist class $\mathcal{K}$ functions α_1, α_2 such that $\alpha_1(\mathrm{V}_2(x)) \leq \mathrm{V}_1(x) \leq \alpha_2(\mathrm{V}_2(x))$ for all $x \in \mathcal{D}$, then the nullsets of V_1 and V_2 are the same, which is the property P1. In addition, $0 \leq \phi_1(s) \leq \alpha_2(s)$ for all $s \geq 0$, so ϕ_1 is locally bounded and, moreover, the sandwich theorem guarantees that ϕ_1 is right continuous at 0. Also, since $\alpha_1(s) \leq \psi_1(s)$, for all $s \geq 0$, and $\psi_1(0) = 0$, it follows that ψ_1 is positive definite. Therefore, P2 also holds.

Proof of *(ii)* ($\Leftarrow$): Since $\psi_1(s) \geq \alpha_1(s)$ for all $s \geq 0$, the property P3 follows because

$$\lim_{s\to\infty} \psi_1(s) \geq \lim_{s\to\infty} \alpha_1(s) = \infty.$$

Proof of *(iii)* ($\Leftarrow$): If $\mathrm{V}_1 \sim^{\mathcal{K}_\infty^{cc}} \mathrm{V}_2^p$, then $\mathrm{V}_1 \sim^{\mathcal{K}_\infty} \mathrm{V}_2^p$ by (7.29). Also, we have trivially that $\mathrm{V}_2^p \sim^{\mathcal{K}_\infty} \mathrm{V}_2$. Since $\sim^{\mathcal{K}_\infty}$ is an equivalence relation by Lemma 3.59, it follows that $\mathrm{V}_1 \sim^{\mathcal{K}_\infty} \mathrm{V}_2$, so the properties $\{\mathrm{Pi}\}_{i=1}^3$ hold as in *(ii)* ($\Leftarrow$). We have left to derive P4. If $\mathrm{V}_1 \sim^{\mathcal{K}_\infty^{cc}} \mathrm{V}_2^p$, then there exist $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ convex and concave, respectively, such that $\alpha_1\Big(\mathrm{V}_2(x)^p\Big) \leq \mathrm{V}_1(x) \leq \alpha_2\Big(\mathrm{V}_2(x)^p\Big)$ for all $x \in \mathcal{D}$. Hence, by the definition of ψ_p and ϕ_p, and P0, and by the monotonicity of α_1 and α_2, we have that, for all $s \geq 0$,

$$\begin{aligned}\alpha_1(s) &\leq \inf_{\{x \in \mathcal{D} : \mathrm{V}_2(x)^p \geq s\}} \alpha_1\Big(\mathrm{V}_2(x)^p\Big) \leq \inf_{\{x \in \mathcal{D} : \mathrm{V}_2(x)^p \geq s\}} \mathrm{V}_1(x) = \psi_p(s) \\ &\leq \phi_p(s) = \sup_{\{x \in \mathcal{D} : \mathrm{V}_2(x)^p \leq s\}} \mathrm{V}_1(x) \leq \sup_{\{x \in \mathcal{D} : \mathrm{V}_2(x)^p \leq s\}} \alpha_2\Big(\mathrm{V}_2(x)^p\Big) \leq \alpha_2(s).\end{aligned} \tag{7.40}$$

Now, since $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ are convex and concave, respectively, it follows by Lemma 0.66 that $s^2/\alpha_1(s)$ and $\alpha_2(s)$ are in $\mathcal{O}(s)$ as $s \to \infty$. Knowing from (7.40) that $\alpha_1(s) \leq \psi_p(s) \leq \phi_p(s) \leq \alpha_2(s)$ for all $s \geq 0$, we conclude that the functions $s^2/\psi_p(s)$ and $\phi_p(s)$ are also in $\mathcal{O}(s)$ as $s \to \infty$, which is the property P4. □

The following example shows ways in which the conditions of Theorem 3.61 might fail.

Example 3.62. (Illustration of Theorem 3.61). Let $\mathrm{V}_2 : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ be the distance to the set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 = 0\}$, i.e., $\mathrm{V}_2(x_1, x_2) = |x_1|$. Consider the following cases:

P2 fails (ψ_1 is not positive definite): Let $\mathrm{V}_1(x_1, x_2) = |x_1| e^{-|x_2|}$ for $(x_1, x_2) \in \mathbb{R}^2$. Note that V_1 is not $\mathcal{K}$-dominated by V_2 because, given any $\alpha_1 \in \mathcal{K}$, for every $x_1 \in \mathbb{R}$ with $|x_1| > 0$ there exists $x_2 \in \mathbb{R}$ such that the inequality $\alpha_1(|x_1|) \leq |x_1| e^{-|x_2|}$ does not hold (just choose x_2 satisfying $|x_2| > \log\Big(\frac{|x_1|}{\alpha_1(|x_1|)}\Big)$). Thus, there must be some of the hypotheses on Theorem 3.61 that fail to be true. In this case, we observe that

$$\psi_1(s) = \inf_{\{(x_1,x_2)\in\mathbb{R}^2\,:\,|x_1|\ge s\}} |x_1|e^{-|x_2|}$$

is identically 0 for all $s \geq 0$, so it is not positive definite as required in P2.

P2 fails (ϕ_1 is not locally bounded): Let $\mathrm{V}_1(x_1,x_2) = |x_1|e^{|x_2|}$ for $(x_1,x_2)\in\mathbb{R}^2$. As above, one can show that α_2 does not exist in the required class; in this case, the hypothesis P2 is not satisfied because ϕ_1 is not locally bounded in $(0,\infty)$:

$$\phi_1(s) = \sup_{\{(x_1,x_2)\in\mathbb{R}^2\,:\,|x_1|\le s\}} |x_1|e^{|x_2|} = \infty, \quad \forall\, s>0.$$

P2 fails (ϕ_1 is not right continuous): Let $\mathrm{V}_1(x_1,x_2) = |x_1|^4 + |\sin(x_1x_2)|$ for $(x_1,x_2)\in\mathbb{R}^2$. For every $p>0$, we have that

$$\phi_p(s) = \sup_{\{(x_1,x_2)\in\mathbb{R}^2\,:\,|x_1|^p\le s\}} |x_1|^4 + |\sin(x_1x_2)| \le s^{4/p} + 1,$$

so ϕ_p is locally bounded in $\mathbb{R}_{\ge 0}$, and, again for every $p>0$,

$$\psi_p(s) = \inf_{\{(x_1,x_2)\in\mathbb{R}^2\,:\,|x_1|^p\ge s\}} |x_1|^4 + |\sin(x_1x_2)| \ge s^{4/p},$$

so ψ_p is positive definite. However, ϕ_p is not right continuous at 0 because $\sin(x_1x_2)=0$ when $x_1=0$, but $\sup_{\{(x_1,x_2)\in\mathbb{R}^2\,:\,|x_1|^p\le s_0\}} \sin(x_1x_2) = 1$ for any $s_0>0$, so by Theorem 3.61 *(i)*, it follows that V_1 is not $\mathcal{K}$-dominated by V_2.

P4 fails (non-compliant asymptotic behavior): Let $\mathrm{V}_1(x_1,x_2) = |x_1|^4$ for $(x_1,x_2)\in\mathbb{R}^2$. Then P2 is satisfied and P3 also holds because $\lim_{s\to\infty}\psi_1(s) = \lim_{s\to\infty} s^4 = \infty$, so Theorem 3.61 *(ii)* implies that V_1 and V_2 are $\mathcal{K}_\infty$-proper with respect to each other. However, in this case $\phi_p(s) = \psi_p(s) = s^{4/p}$, which implies that

ϕ_p is not in $\mathcal{O}(s)$ as $s \to \infty$ when $p \in (0,4)$, and $s^2/\psi_p(s)$ is not in $\mathcal{O}(s)$ as $s \to \infty$ when $p > 4$. Thus P4 is satisfied only for $p = 4$, so Theorem 3.61 *(iii)* implies that only in this case V_1 and V_2^p are $\mathcal{K}_\infty^{cc}$- proper with respect to each other. Namely, for $p > 4$, one cannot choose a convex $\alpha_1 \in \mathcal{K}_\infty$ such that $\alpha_1(|x_1|^p) \leq |x_1|^4$ for all $x_1 \in \mathbb{R}$ and, if $p < 4$, one cannot choose a concave $\alpha_2 \in \mathcal{K}_\infty$ such that $|x_1|^4 \leq \alpha_2(|x_1|^p)$ for all $x_1 \in \mathbb{R}$. •

7.3.3 Application to noise-to-state stability

In this section we use the results of Sections 7.3.1 and 7.3.2 to study the noise-to-state stability properties of stochastic differential equations of the form (7.1). Our first result provides a way to check whether a candidate function that satisfies a dissipation inequality of the type (7.6) is in fact a noise-dissipative Lyapunov function, a strong NSS-Lyapunov function in probability, or a pth moment NSS-Lyapunov function.

Corollary 3.63. (Establishing proper relations between pairs of functions through seminorms). *Consider* $\mathrm{V}_1, \mathrm{V}_2 : \mathcal{D} \to \mathbb{R}_{\geq 0}$ *such that their nullset is a subspace* $\mathcal{U}$*. Let* $A, \tilde{A} \in \mathbb{R}^{m \times n}$ *be such that* $\mathcal{N}(A) = \mathcal{U} = \mathcal{N}(\tilde{A})$*. Assume that* V_1 *and* V_2 *satisfy* $\{\mathrm{P}i\}_{i=0}^3$ *with respect to* $\|.\|_A$ *and* $\|.\|_{\tilde{A}}$*, respectively. Then, for any* $q > 0$*,*

$$\mathrm{V}_1 \sim^{\mathcal{K}_\infty} \mathrm{V}_2, \quad \mathrm{V}_1 \sim^{\mathcal{K}_\infty} \|.\|_A^q, \quad \mathrm{V}_2 \sim^{\mathcal{K}_\infty} \|.\|_{\tilde{A}}^q \quad in \quad \mathcal{D}.$$

If, in addition, V_1 *and* V_2 *satisfy* P4 *with respect to* $\|.\|_A$ *and* $\|.\|_{\tilde{A}}$*, respectively, for some* $p > 0$*, then*

$$\mathrm{V}_1 \sim^{\mathcal{K}_\infty^{cc}} \mathrm{V}_2, \quad \mathrm{V}_1 \sim^{\mathcal{K}_\infty^{cc}} \|.\|_A^p, \quad \mathrm{V}_2 \sim^{\mathcal{K}_\infty^{cc}} \|.\|_{\tilde{A}}^p \quad in \quad \mathcal{D}.$$

Proof. The statements follow from the characterizations in Theorem 3.61 *(ii)* and *(iii)*, and from the fact that the relations $\sim^{\mathcal{K}_\infty}$ and $\sim^{\mathcal{K}_\infty^{cc}}$ are equivalence relations as shown in Lemma 3.59. That is, under the hypothesis P0,

$$\left.\begin{aligned} &\mathrm{V}_1 \text{ satisfies } \{\mathrm{Pi}\}_{i=1}^{3} \text{ w/ respect to } \|.\|_A \ (\Leftrightarrow\ \mathrm{V}_1 \sim^{\mathcal{K}_\infty} \|.\|_A \text{ in } \mathcal{D})\\ &\mathrm{V}_2 \text{ satisfies } \{\mathrm{Pi}\}_{i=1}^{3} \text{ w/ respect to } \|.\|_{\tilde{A}} \ (\Leftrightarrow\ \mathrm{V}_2 \sim^{\mathcal{K}_\infty} \|.\|_{\tilde{A}} \text{ in } \mathcal{D}) \end{aligned}\right\} \Rightarrow \mathrm{V}_1 \sim^{\mathcal{K}_\infty} \mathrm{V}_2 \text{ in } \mathcal{D},$$

$$\left.\begin{aligned} &\mathrm{V}_1 \text{ satisfies } \{\mathrm{Pi}\}_{i=1}^{4} \text{ w/ respect to } \|.\|_A \ (\Leftrightarrow\ \mathrm{V}_1 \sim^{\mathcal{K}_\infty^{cc}} \|.\|_A^p \text{ in } \mathcal{D})\\ &\mathrm{V}_2 \text{ satisfies } \{\mathrm{Pi}\}_{i=1}^{4} \text{ w/ respect to } \|.\|_{\tilde{A}} \ (\Leftrightarrow\ \mathrm{V}_2 \sim^{\mathcal{K}_\infty^{cc}} \|.\|_{\tilde{A}}^p \text{ in } \mathcal{D}) \end{aligned}\right\} \Rightarrow \mathrm{V}_1 \sim^{\mathcal{K}_\infty^{cc}} \mathrm{V}_2 \text{ in } \mathcal{D}.$$

Note that, by Lemma 3.58 and (7.29), the equivalences

$$\|.\|_A \sim^{\mathcal{K}_\infty} \|.\|_{\tilde{A}}^q \ \text{ in } \mathcal{D}, \qquad \|.\|_A^p \sim^{\mathcal{K}_\infty^{cc}} \|.\|_{\tilde{A}}^p \ \text{ in } \mathcal{D}$$

hold for any $p, q > 0$ and any matrices $A, \tilde{A} \in \mathbb{R}^{m \times n}$ with $\mathcal{N}(A) = \mathcal{N}(\tilde{A})$. □

We next build on this result to provide an alternative formulation of Corollary 2.54. To do so, we employ the observation made in Remark 3.57 about the possibility of interpreting the candidate functions as defined on a constrained domain of an extended Euclidean space.

Corollary 3.64. (The existence of a $p$thNSS-Lyapunov function implies pth moment NSS –revisited). *Under Assumption 1.45, let* $\mathrm{V} \in \mathcal{C}^2(\mathbb{R}^n;\mathbb{R}_{\geq 0})$, $\mathrm{W} \in \mathcal{C}(\mathbb{R}^n;\mathbb{R}_{\geq 0})$ *and* $\sigma \in \mathcal{K}$ *be such that the dissipation inequality* (7.7) *holds. Let* $R : \mathbb{R}^n \to \mathbb{R}^{(m-n)}$*, with* $m \geq n$*,* $\mathcal{D} \subset \mathbb{R}^m$*,* $\hat{\mathrm{V}} \in \mathcal{C}^2(\mathcal{D};\mathbb{R}_{\geq 0})$ *and* $\hat{\mathrm{W}} \in \mathcal{C}(\mathcal{D};\mathbb{R}_{\geq 0})$ *be such that, for* $i(x) = [x^\top, R(x)^\top]^\top$*, one has*

$$\mathcal{D} = i(\mathbb{R}^n), \quad \mathrm{V} = \hat{\mathrm{V}} \circ i, \quad \textit{and} \quad \mathrm{W} = \hat{\mathrm{W}} \circ i.$$

Let $A = \operatorname{diag}(A_1, A_2)$ *and* $\tilde{A} = \operatorname{diag}(\tilde{A}_1, \tilde{A}_2)$ *be block-diagonal matrices, with* $A_1, \tilde{A}_1 \in$

$\mathbb{R}^{n\times n}$ and $A_2, \tilde{A}_2 \in \mathbb{R}^{(m-n)\times(m-n)}$, such that $\mathcal{N}(A) = \mathcal{N}(\tilde{A})$ and

$$\|R(x)\|_{A_2}^2 \leq \kappa \|x\|_{A_1}^2 \tag{7.41}$$

for some $\kappa > 0$, for all $x \in \mathbb{R}^n$. Assume that $\hat{\mathrm{V}}$ and $\hat{\mathrm{W}}$ satisfy the properties $\{\mathrm{P}i\}_{i=0}^4$ with respect to $\|.\|_A$ and $\|.\|_{\tilde{A}}$, respectively, for some $p > 0$. Then the system (7.1) is NSS in probability and in pth moment with respect to $\mathcal{N}(A_1)$.

Proof. By Corollary 3.63, we have that

$$\hat{\mathrm{V}} \sim^{\mathcal{K}_\infty^{cc}} \hat{\mathrm{W}}, \quad \text{and} \quad \hat{\mathrm{V}} \sim^{\mathcal{K}_\infty^{cc}} \|.\|_{\mathrm{diag}(A_1,A_2)}^p \quad \text{in} \quad \mathcal{D}. \tag{7.42}$$

As explained in Remark 3.57, the first relation implies that $\mathrm{V} \sim^{\mathcal{K}_\infty^{cc}} \mathrm{W}$ in $\mathbb{R}^n$. This, together with the fact that (7.7) holds, implies that V is a noise-dissipative Lyapunov function for (7.1). Also, setting $\hat{x} = i(x)$ and using (7.41), we obtain that

$$\|x\|_{A_1}^2 \leq \|\hat{x}\|_{\mathrm{diag}(A_1,A_2)}^2 = \|x\|_{A_1}^2 + \|R(x)\|_{A_2}^2 \leq (1+\kappa)\|x\|_{A_1}^2,$$

so, in particular, $\|[.,R(.)]\|_{\mathrm{diag}(A_1,A_2)}^p \sim \|.\|_{A_1}^p$ in $\mathbb{R}^n$. Now, from the second relation in (7.42), by Remark 3.57, it follows that $\hat{\mathrm{V}} \circ i \sim^{\mathcal{K}_\infty^{cc}} \|[.,R(.)]\|_{\mathrm{diag}(A_1,A_2)}^p$ in $\mathbb{R}^n$. Thus, using (7.29) and Lemma 3.59, we conclude that $\mathrm{V} \sim^{\mathcal{K}_\infty^{cc}} \|.\|_{A_1}^p$ in $\mathbb{R}^n$. In addition, the Euclidean distance to the set $\mathcal{N}(A_1)$ is equivalent to $\|.\|_{A_1}$, i.e., $|.|_{\mathcal{N}(A_1)} \sim \|.\|_{A_1}$. This can be justified as follows: choose $B \in \mathbb{R}^{n\times k}$, with $k = \dim(\mathcal{N}(A_1))$, such that the columns of B form an orthonormal basis of $\mathcal{N}(A_1)$. Then,

$$|x|_{\mathcal{N}(A_1)} = \|(\mathrm{I} - BB^\top)x\|_2 = \|x\|_{\mathrm{I}-BB^\top} \sim \|.\|_{A_1}, \tag{7.43}$$

where the last relation follows from Lemma 3.58 because $\mathcal{N}(\mathrm{I} - BB^\top) = \mathcal{N}(A_1)$.

Summarizing, $\mathrm{V} \sim^{\mathcal{K}_\infty^{cc}} \|.\|_{A_1}^p$ and $\|.\|_{A_1}^p \sim |x|_{\mathcal{N}(A_1)}^p$ in $\mathbb{R}^n$ (because the pth power is irrelevant for the relation $\sim$). As a consequence,

$$\mathrm{V} \sim^{\mathcal{K}_\infty^{cc}} |.|_{\mathcal{N}(A_1)}^p \quad \text{in} \quad \mathbb{R}^n, \tag{7.44}$$

which implies condition (7.24) with convex $\alpha_1 \in \mathcal{K}_\infty$, concave $\alpha_2 \in \mathcal{K}_\infty$, and $\mathcal{U} = \mathcal{N}(A_1)$. Therefore, V is a pth moment NSS-Lyapunov function with respect to the set $\mathcal{N}(A_1)$, and the result follows from Corollary 2.54. □

7.4 Discussion

We have studied the stability properties of SDEs subject to persistent noise (including the case of additive noise). We have generalized the concept of noise-dissipative Lyapunov function and introduced the concepts of strong NSS-Lyapunov function in probability and pth moment NSS-Lyapunov function, both with respect to a closed set. We have shown that noise-dissipative Lyapunov functions have NSS dynamics and established that the existence of an NSS-Lyapunov function, of either type, with respect to a closed set, implies the corresponding NSS property of the system with respect to the set. In particular, pth moment NSS with respect to a set provides a bound, at each time, for the pth power of the distance from the state to the set, and this bound is the sum of an increasing function of the size of the noise covariance and a decaying effect of the initial conditions. This bound can be achieved regardless of the possibility that inside the set some combination of the states accumulates the variance of the noise. This is a meaningful stability property for the aforementioned class of systems because the presence of persistent noise makes it impossible to establish in general a stochastic notion of asymptotic stability for the set of equilibria of the underlying differential equation. We have also studied in depth the inequalities between pairs of functions that appear in the various notions of Lyapunov functions mentioned above. We have shown that these inequalities define equivalence relations and have developed a complete characterization of the properties that two functions must satisfy to be related by them. Finally, building on this characterization, we have provided an alternative statement of our stochastic stability results.

Acknowledgments

This chapter is taken from a revised version of a journal published by D. Mateos-Núñez and J. Cortés in *SIAM Journal on Control and Optimization 52 (4) (2014), 2399-2421,* with the title “pth moment noise-to-state stability of stochastic differential equations with persistent noise.” The revised version can be found in *arXiv:1501.05008v1.* The author of this thesis would like to thank Dean Richert for useful discussions during the preparation of the manuscript. In addition, both authors would like to thank Fengzhong Li, from Shandong University, for his kind observations that made possible an important correction of the proof of Theorem 2.51. A preliminary version appeared at the *2014 International Symposium on Mathematical Theory of Networks and Systems.* This research was supported by the US National Science Foundation (NSF) Award CMMI-1300272.

Chapter 8

Conclusions

We have developed distributed multi-agent strategies to solve several families of convex optimization problems. The network objective is a sum of convex functions under a variety of scenarios: time varying objectives, nuclear norm regularization, and constraints that are also a sum of convex functions. We have placed mild assumptions on the communication network, just requiring local, time varying, and asynchronous communication over weight-balanced digraphs whose consecutive unions are strongly connected over bounded time horizons. In the case of our continuous-time algorithm, we strengthen this hypothesis to time-invariant strongly connected digraphs, but we develop a novel Lyapunov technique for stochastic differential equations to establish the noise-to-state stability in second moment, letting us model persistent white noise in communications and computations. All of our strategies require the agents to use only local information about their objective functions and their component constraints in the form of subgradients. In the scenario with time-varying objective functions, only historic information about previously revealed functions is used. In this case, the agent regret compares each agent's sequence of decisions with the centralized solution in hindsight. Our extension of the classical sublinear regret bounds to the distributed case means that trends that can be captured in hindsight by a single decision computed with all the information centrally available, can also be approximated “on the fly” by the agents with historic local information.

From the perspective of the applications, the optimization models considered can be specified and tuned by machine learning experts for scenarios such as regression, classification, multi-task and online learning, all of which can benefit from our distributed strategies as we explained in the introduction. Our proofs show that our distributed coordination algorithms have analogous correctness guarantees as the centralized counterparts. The capacity to add constraints given by a sum of convex functions offers additional modeling flexibility for problems such as formation control or network resource allocation, where the constraints are motivated by physical objectives, such as relative positions and angles, or limitations, such as budgets. The key insight here is the agreement on the Lagrange multipliers that allow agents’ decisions to be constrained even when the agents involved cannot communicate with each other directly.

8.1 Future directions

The following are suggestions for future research on the various aspects of distributed optimization studied in this thesis.

8.1.0.1 Noisy communication channels

One avenue of research is the study of models merging continuous-time gradient updates and discrete-time communications under noise, using the framework of hybrid stochastic systems. Other aspects of general interest are the relaxations of the weight-balanced property for directed communication graphs as well as the joint connectivity assumption in the context of continuous-time stochastic evolution, as well as the effect of delays and bandwidth limitations.

8.1.0.2 Online optimization

Directions of interest are the refinement of the regret bounds when partial knowledge about the evolution of the cost functions is available, the study of the impact of practical implementation considerations such as disturbances, noise, communication delays, and asynchronism in the algorithm performance, and the specific application to large-scale learning scenarios involving the distributed interaction of many users and devices.

8.1.0.3 Nuclear norm regularization

The characterizations that we have proposed admit a modification using Fenchel duality in place of Fenchel conjugacy. Other directions include the construction of convex domains that favor the implementation of orthogonal projections, the treatment of other barrier functions like the logarithm of the determinant, and the extension to applications where chordal sparsity plays a role.

8.1.0.4 Saddle-point problems and constrained optimization

Future extensions in this topic include, first, more general distributed algorithms for computing bounds on Lagrange multiplier vectors and matrices, which are required in the design of projections preserving the optimal dual sets. An alternative route would explore the characterization of the intrinsic boundedness properties of the distributed saddle-point dynamics studied in Chapter 5. Second, a refined analysis of the convergence bound for constrained optimization in terms of the cost error as opposed to the saddle-point evaluation error. Third, we envision applications to semidefinite programming where chordal sparsity allows to tackle problems where the dimension of the matrices grows with the size of the network.

8.1.0.5 Noise to state stability

In the context of stochastic stability notions like noise-to-state stability in probability or in pth moment for SDEs, future work can consider the effect of delays and impulsive right-hand sides in the class of SDEs employed in this thesis.

Chapter 9

Epilogue

Tierra,

en los hogares alumbrados de noches asombradas,
nace otro pulso, con otro ritmo,
y tus carnes tiemblan electrificadas,
y la luz obedece al algoritmo.

En las terminales nerviosas de tus nuevos ríos,
que suben y bajan con binarios suspiros,
las memorias, sin nombre todavía,
reflejan compulsiones sin guía.

Mientras tu cuerpo quema las grasas prehistóricas,
llegar hasta el Sol quiere con hazañas metafóricas,
resoplando al legado de los imperios verdes
erigiendo templos que recuerden.

Grande es el esfuerzo de tu piel,
fina, fresca y vaporosa,

herida tu orilla ansiosa
de La Pupila el iris miel.

Quiera la música de tu ser antiguo,
acompañar al hormigueo de tus nuevos hijos.
Quiera la fertilidad de tu sueño ambiguo,
que florezcan sus palabras y sus ritos.

Pues ellos siembran y construyen
historias de amor y de valor,
y en lo más hondo te aman,
mudos a veces de estupor.

Ellos te han dado nombre de diosa,
y estudian asombrados tus ciclos
y cuentan las estrellas de tus noches,
y pueblan tu aliento de mitos.

May 5, 2014

Appendix A

Auxiliary results

Appendix of Chapter 3

The following result concerning the function h defined by (3.7) is employed in the proof of Proposition 3.11.

Lemma 0.65. *For $\delta > 0$, let $h(.,\delta) : (0,\infty) \to \mathbb{R}$ be defined by* (3.7) *and L be the Laplacian matrix of a strongly connected and weight-balanced digraph. Then, there exists $\hat{\beta} \equiv \hat{\beta}(\delta) > 0$ such that $h(\beta,\delta) < 0$ for all $\beta \in (0,\hat{\beta})$.*

Proof. Since the function $h(.,\delta)$ is continuous in the first argument, it is enough to show that the next two limits hold,

$$\lim_{\beta\to+\infty} h(\beta,\delta) = \infty, \quad \text{and} \quad \lim_{\beta\to 0^+} h(\beta,\delta) = 0^-,$$

to deduce the result from the by Bolzano Intermediate Value Theorem. Note that

$$-r+\sqrt{r^2-1} = \frac{(-r+\sqrt{r^2-1})(-r-\sqrt{r^2-1})}{-r-\sqrt{r^2-1}} = \frac{|r^2-1|-r^2}{r+\sqrt{r^2-1}},$$

which behaves asymptotically as $-\frac{1}{2r}$ when $r \to \infty$. Since $r := \frac{\beta^4+3\beta^2+2}{2\beta}$ goes to ∞ for both cases in which $\beta \to \infty$ or $\beta \to 0$, it follows that $-\frac{\beta^4+3\beta^2+2}{2\beta} + \sqrt{\left(\frac{\beta^4+3\beta^2+2}{2\beta}\right)^2 - 1}$ behaves as $-\frac{\beta}{\beta^4+3\beta^2+2}$ in both cases. Therefore,

$$\begin{aligned}\lim_{\beta\to+\infty} h(\beta,\delta) &= \lim_{\beta\to+\infty}\left(-\tfrac{\beta}{\beta^4+3\beta^2+2}\lambda_2(\mathsf{L}+\mathsf{L}^\top) + \tfrac{\beta^2}{2\delta}\right)\\ &= \lim_{\beta\to+\infty}\left(-\tfrac{1}{\beta^3}\lambda_2(\mathsf{L}+\mathsf{L}^\top) + \tfrac{\beta^2}{2\delta}\right) = \infty,\end{aligned}$$

and

$$\lim_{\beta\to 0^+} h(\beta,\delta) = \lim_{\beta\to 0^+}\left(-\tfrac{\beta}{\beta^4+3\beta^2+2}\lambda_2(\mathsf{L}+\mathsf{L}^\top) + \tfrac{\beta^2}{2\delta}\right)$$
$$= \lim_{\beta\to 0^+}\left(-\tfrac{\beta}{2}\lambda_2(\mathsf{L}+\mathsf{L}^\top) + \tfrac{\beta^2}{2\delta}\right) = \lim_{\beta\to 0^+}\beta\Big(-\tfrac{\lambda_2(\mathsf{L}+\mathsf{L}^\top)}{2} + \beta\Big) = 0^-,$$

and the result follows. □

Appendix of Chapter 7

The next result is used in the proof of Theorem 3.61.

Lemma 0.66. (Existence of bounding convex and concave functions in $\mathcal{K}_\infty$). *Let α be a class $\mathcal{K}_\infty$ function. Then the following are equivalent:*

(i) There exist $s_0 \geq 0$ and $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, convex and concave, respectively, such that $\alpha_1(s) \leq \alpha(s) \leq \alpha_2(s)$ for all $s \geq s_0$, and

(ii) $\alpha(s), s^2/\alpha(s)$ are in $\mathcal{O}(s)$ as $s \to \infty$.

Proof. The implication *(i)* ⇒ *(ii)* follows because, for any $s \geq s_0 > 0$,

$$\tfrac{\alpha_1(s_0)}{s_0}s \leq \alpha_1(s) \leq \alpha(s) \leq \alpha_2(s) \leq \tfrac{\alpha_2(s_0)}{s_0}s,$$

by convexity and concavity, respectively, where $\alpha_1(s_0), \alpha_2(s_0) > 0$.

To show *(ii)* $\Rightarrow$ *(i)*, we proceed to construct α_1, α_2 as in the statement using the correspondence between functions, graphs and epigraphs (or hypographs). Let $\alpha_1 : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be the function whose epigraph is the convex hull of the epigraph of α, i.e., $\operatorname{epi}\alpha_1 \triangleq \operatorname{conv}(\operatorname{epi}\alpha)$. Thus, α_1 is convex, nondecreasing, and $0 \leq \alpha_1(s) \leq \alpha(s)$ for all $s \geq 0$ because $\mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \supseteq \operatorname{epi}\alpha_1 = \operatorname{conv}(\operatorname{epi}\alpha) \supseteq \operatorname{epi}\alpha$. Moreover, α_1 is continuous in $(0, \infty)$ by convexity [Roc70, Th. 10.4], and is also continuous at 0 by the sandwich theorem [LL88, p. 107] because $\alpha \in \mathcal{K}_\infty$. To show that $\alpha_1 \in \mathcal{K}_\infty$, we have to check that it is unbounded, positive definite in $\mathbb{R}_{\geq 0}$, and strictly increasing. First, since $s^2/\alpha(s) \in \mathcal{O}(s)$ as $s \to \infty$, there exist constants $c_1, s_0 > 0$ such that $\alpha(s) \geq c_1 s$ for all $s > s_0$. Now, define $g_1(s) \triangleq \alpha(s)$ if $s \leq s_0$ and $g_1(s) \triangleq c_1 s$ if $s > s_0$, and $g_2(s) \triangleq -c_1 s_0 + c_1 s$ for all $s \geq 0$, so that $g_2 \leq g_1 \leq \alpha$. Then, $\operatorname{epi}\alpha_1 = \operatorname{conv}(\operatorname{epi}\alpha) \subseteq \operatorname{conv}(\operatorname{epi} g_1) \subseteq \operatorname{epi} g_2$, because $\operatorname{epi} g_2$ is convex, and thus α_1 is unbounded. Also, since $\operatorname{conv}(\operatorname{epi} g_1) \cap \mathbb{R}_{\geq 0} \times \{0\} = \{(0,0)\}$, it follows that α_1 is positive definite. To show that α_1 is strictly increasing, we use two facts: since α_1 is convex, we know that the set in which α_1 is allowed to be constant must be of the form $[0, b]$ for some $b > 0$; on the other hand, since α_1 is positive definite, it is nonconstant in any neighborhood of 0. As a result, α_1 is nonconstant in any subset of its domain, so it is strictly increasing.

Next, let $\alpha_2 : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be the function whose hypograph is the convex hull of the hypograph of α, i.e., $\operatorname{hyp}\alpha_2 \triangleq \operatorname{conv}(\operatorname{hyp}\alpha)$. The function α_2 is well-defined because $\alpha(s) \in \mathcal{O}(s)$ as $s \to \infty$, i.e., there exist constants $c_2, s_0 > 0$ such that $\alpha(s) \leq c_2 s$ for all $s > s_0$, so if we define $g(s) \triangleq c_2 s_0 + c_2 s$ for all $s \geq 0$, then $\operatorname{hyp}\alpha_2 = \operatorname{conv}(\operatorname{hyp}\alpha) \subseteq \operatorname{hyp} g$, because $\operatorname{hyp} g$ is convex, and thus $\alpha_2(s) \leq g(s)$. Also, by construction, α_2 is concave, nondecreasing, and $\alpha_2 \geq \alpha$ because $\operatorname{hyp}\alpha_2 \supseteq \operatorname{hyp}\alpha$, which also implies that α_2 is unbounded. Moreover, α_2 is continuous in $(0, \infty)$ by concavity [Roc70, Th. 10.4], and is also continuous at 0 because the possibility of an infinite jump is excluded by the fact that $\alpha_2 \leq g$. To show that $\alpha_2 \in \mathcal{K}_\infty$, we have to check that it is positive definite in $\mathbb{R}_{\geq 0}$ and strictly increasing. Note that α_2 is positive definite because $\alpha_2(0) = 0$ and $\alpha_2 \geq \alpha$. To show that α_2 is strictly increasing, we reason by contradiction. Assume that α_2 is constant in some closed interval of the form $[s_1, s_2]$, for some $s_2 > s_1 \geq 0$. Then, as α_2 is concave, we conclude that it is nonincreasing in (s_2, ∞). Now, since α_2 is continuous, we reach the contradiction that $\lim_{s\to\infty} \alpha(s) \leq \lim_{s\to\infty} \alpha_2(s) \leq \alpha_2(s_1) < \infty$. Hence, α_2 is strictly increasing. $\square$

Bibliography

[AEP06] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, volume 19, pages 41–48, 2006.

[AEP08] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[AHU58] K. Arrow, L Hurwitz, and H. Uzawa. *Studies in Linear and Non-Linear Programming.* Stanford University Press, Stanford, California, 1958.

[AZ05] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(11):1817–1853, 2005.

[BCM09] F. Bullo, J. Cortés, and S. Martínez. *Distributed Control of Robotic Networks.* Applied Mathematics Series. Princeton University Press, 2009. Electronically available at `http://coordinationbook.info`.

[Ber99] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 2nd edition, 1999.

[Ber05] D. S. Bernstein. *Matrix Mathematics.* Princeton University Press, Princeton, NJ, 2005.

[BNA14] M. Bürger, G. Notarstefano, and F. Allgöwer. A polyhedral approximation framework for convex and robust distributed optimization. *IEEE Transactions on Automatic Control*, 59(2):384–395, 2014.

[BNO03] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar. *Convex Analysis and Optimization.* Athena Scientific, Belmont, MA, 1st edition, 2003.

[Bor95] V. S. Borkar. *Probability Theory: An Advanced Course.* Springer, New York, 1995.

[BPC$^+$11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[BT97] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific, 1997.

[Bul03] P. S. Bullen. *Handbook of Means and Their Inequalities*, volume 560 of *Mathematics and Its Applications*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.

[BV09] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2009.

[CBL06] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[CNS14] T-H. Chang, A. Nedić, and A. Scaglione. Distributed constrained optimization by consensus-based primal-dual perturbation method. *IEEE Transactions on Automatic Control*, 59(6):1524–1538, 2014.

[CR09] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[CT10] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[Cul67] J. M. Culkin. Each culture develops its own sense ratio to meet the demands of its environment. In G. Stearn, editor, *McLuhan: Hot & Cool*, pages 49–57. 1967.

[DAW12] J. C. Duchi, A. Agarwal, and M. J. Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606, 2012.

[DGBSX12] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(Jan):165–202, 2012.

[DHM12] M. Dudík, Z. Harchaoui, and J. Malick. Lifted coordinate descent for learning with trace-norm regularization. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22, pages 327–336. JMLR Workshop and Conference Proceedings, 2012.

[DK00] H. Deng and M. Krstić. Output-feedback stabilization of stochastic nonlinear systems driven by noise of unknown covariance. *Systems & Control Letters*, 39:173–182, 2000.

[DKW01] H. Deng, M. Krstić, and R. J. Williams. Stabilization of stochastic nonlinear systems driven by noise of unknown covariance. *IEEE Transactions on Automatic Control*, 46(8):1237–1253, 2001.

[Faz02] M. Fazel. *Matrix rank minimization with applications.* PhD thesis, Stanford University, 2002.

[Fie73] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305, 1973.

[GC14] B. Gharesifard and J. Cortés. Distributed continuous-time convex optimization on weight-balanced digraphs. *IEEE Transactions on Automatic Control*, 59(3):781–786, 2014.

[HAK07] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[Haz11] E. Hazan. The convex optimization approach to regret minimization. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*, pages 287–304. MIT Press, Cambridge, MA, 2011.

[HCM13] S. Hosseini, A. Chapman, and M. Mesbahi. Online distributed optimization via dual averaging. In *IEEE Conf. on Decision and Control*, Florence, Italy, 2013.

[Hig86] N. J. Higham. Newton's method for the matrix square root. *Mathematics of Computation*, 46(174):537–549, 1986.

[HJ85] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.

[HO14] C.-J. Hsieh and P. A. Olsen. Nuclear norm minimization via active subspace selection. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32. JMLR Workshop and Conference Proceedings, 2014.

[HUL93] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I.* Grundlehren Text Editions. Springer, New York, 1993.

[JKJJ08] B. Johansson, T. Keviczky, M. Johansson, and K. H. Johansson. Subgradient methods and consensus algorithms for solving convex optimization problems. In *IEEE Conf. on Decision and Control*, pages 4185–4190, Cancun, Mexico, 2008.

[JSW99] Z.-P. Jiang, E.D. Sontag, and Y. Wang. Input-to-state stability for discrete-time nonlinear systems. In *14th IFAC World Congress*, pages 277–282, 1999.

[KCM15] S. S. Kia, J. Cortés, and S. Martínez. Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication. *Automatica*, 55:254–264, 2015.

[Kha02] H. K. Khalil. *Nonlinear Systems*. Prentice Hall, 3 edition, 2002.

[Kha12] R. Khasminskii. *Stochastic Stability of Differential Equations*, volume 66 of *Stochastic Modelling and Applied Probability*. Springer, 2012.

[KL12] P. E. Kloeden and T. Lorenz. Mean-square random dynamical systems. *Journal of Differential Equations*, 253(5):1422–1438, 2012.

[Kle05] F. C. Klebaner. *Introduction to Stochastic Calculus With Applications*. Imperial College Press, 2005.

[LL88] J. Lewin and M. Lewin. *An Introduction to Mathematical Analysis*. BH mathematics series. The Random House, 1988.

[LMS91] V. Lakshmikantham, V. M. Matrosov, and S. Sivasundaram. *Vector Lyapunov Functions and Stability Analysis of Nonlinear Systems*, volume 63 of *Mathematics and its Applications*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.

[LT12] J. Lu and C. Y. Tang. Zero-gradient-sum algorithms for distributed convex optimization: the continuous-time case. *IEEE Transactions on Automatic Control*, 57(9):2348–2354, 2012.

[LTRB11] J. Lu, C. Y. Tang, P. R. Regier, and T. D. Bow. Gossip algorithms for convex consensus optimization over networks. *IEEE Transactions on Automatic Control*, 56(12):2917–2923, 2011.

[LZJ08] S.-J. Liu, J.-F. Zhang, and Z.-P. Jiang. A notion of stochastic input-to-state stability and its application to stability of cascaded stochastic nonlinear systems. *Acta Mathematicae Applicatae Sinica, English Series*, 24(1):141–156, 2008.

[Mao99] X. Mao. Stochastic versions of the LaSalle theorem. *Journal of Differential Equations*, 153(1):175–195, 1999.

[Mao11] X. Mao. *Stochastic Differential Equations and Applications*. Woodhead Publishing, 2nd edition, 2011.

[MARS10] D. Mosk-Aoyama, T. Roughgarden, and D. Shah. Fully distributed algorithms for convex optimization problems. *SIAM J. Optimization*, 20(6):3260–3279, 2010.

[MFSL14] R. Madani, G. Fazelnia, S. Sojoudi, and J. Lavaei. Low-rank solutions of matrix inequalities with applications to polynomial optimization and matrix completion problems. In *IEEE Conf. on Decision and Control*, Los Angeles, CA, 2014.

[MHT10] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

[MNC13] D. Mateos-Núñez and J. Cortés. Noise-to-state stable distributed convex optimization on weight-balanced digraphs. In *IEEE Conf. on Decision and Control*, pages 2781–2786, Florence, Italy, 2013.

[MNC14a] D. Mateos-Núñez and J. Cortés. Distributed online convex optimization over jointly connected digraphs. *IEEE Transactions on Network Science and Engineering*, 1(1):23–37, 2014.

[MNC14b] D. Mateos-Núñez and J. Cortés. pth moment noise-to-state stability of stochastic differential equations with persistent noise. *SIAM Journal on Control and Optimization*, 52(4):2399–2421, 2014.

[MNC15] D. Mateos-Núñez and J. Cortés. Distributed subgradient methods for saddle-point problems. In *IEEE Conf. on Decision and Control*, pages 5462–5467, Osaka, Japan, 2015.

[Mov11] J. R. Movellan. Tutorial on stochastic differential equations. Tutorial, MPLab, UCSD, 2011.

[MW99] J. N. McDonald and N. A. Weiss. *A Course in Real Analysis.* Elsevier, Oxford, UK, 1999.

[NDS10] I. Necoara, I. Dumitrache, and J. A. K. Suykens. Fast primal-dual projected linear iterations for distributed consensus in constrained convex optimization. In *IEEE Conf. on Decision and Control*, pages 1366–1371, Atlanta, GA, December 2010.

[Nes04] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization.* Springer, 2004.

[NLT11] S. Nikookhoy, J. Lu, and C. Y. Tang. Distributed convex optimization with identical constraints. In *IEEE Conf. on Decision and Control*, pages 2926–2931, Orlando, FL, December 2011.

[NO09a] A. Nedic and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

[NO09b] A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory & Applications*, 142(1):205–228, 2009.

[NO10a] A. Nedić and A. Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM J. Optimization*, 19(4):1757–1780, 2010.

[NO10b] A. Nedić and A. Ozdaglar. Cooperative distributed multi-agent optimization. In Y. Eldar and D. Palomar, editors, *Convex Optimization in Signal Processing and Communications*. Cambridge University Press, 2010.

[NOP10] A. Nedic, A. Ozdaglar, and P. A. Parrilo. Constrained consensus and optimization in multi-agent networks. *IEEE Transactions on Automatic Control*, 55(4):922–938, 2010.

[Ö10] B. Öksendal. *Stochastic Differential Equations - An Introduction with Applications*. Universitext. Springer-Verlag, 2010.

[Ozd07] A. Ozdaglar. Constrained consensus and alternating projections. In *Allerton Conf. on Communications, Control and Computing*, Monticello, IL, September 2007.

[PB13] N. Parikh and S. Boyd. Proximal algorithms. 1(3):123–231, 2013.

[PW96] L. Praly and Y. Wang. Stabilization in spite of matched unmodeled dynamics and an equivalent definition of input-to-state stability. *Mathematics of Control, Signals and Systems*, 9(1):1–33, 1996.

[RC15] D. Richert and J. Cortés. Robust distributed linear programming. *IEEE Transactions on Automatic Control*, 60(10):2567–2582, 2015.

[RFP10] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[RKW11] M. Raginsky, N. Kiarashi, and R. Willett. Decentralized online convex programming with local information. In *American Control Conference*, pages 5363–5369, San Francisco, CA, 2011.

[RNV10] S. S. Ram, A. Nedić, and V. V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory & Applications*, 147(3):516–545, 2010.

[Roc70] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[RR13] B. Recht and C. Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

[RW98] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*, volume 317 of *Comprehensive Studies in Mathematics*. Springer, New York, 1998.

[Sch01] H. Schurz. On moment-dissipative stochastic dynamical systems. *Dynamic systems and applications*, 10:11–44, 2001.

[SJR16] A. Simonetto and H. Jamali-Rad. Primal recovery from consensus-based dual decomposition for distributed convex optimization. *Journal of Optimization Theory and Applications*, 168(1):172–197, 2016.

[SN11] K. Srivastava and A. Nedić. Distributed asynchronous constrained stochastic optimization. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):772–790, 2011.

[Son98] E. D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, volume 6 of *TAM*. Springer, 2 edition, 1998.

[Son08] E. D. Sontag. Input to state stability: Basic concepts and results. *Nonlinear and Optimal Control Theory*, 1932:163–220, 2008.

[SPFP10] G. Scutari, D. P. Palomar, F. Facchinei, and J. S. Pang. Convex optimization, game theory, and variational inequality theory. *IEEE Signal Processing Magazine*, 27(3):35–49, 2010.

[SS12] S. Shalev-Shwartz. *Online Learning and Online Convex Optimization*, volume 12 of *Foundations and Trends in Machine Learning*. Now Publishers Inc, 2012.

[SSS07] S. Shalev-Shwartz and Y. Singer. Convex repeated games and fenchel duality. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19, Cambridge, MA, 2007. MIT Press.

[SW95] E. D. Sontag and Y. Wang. On characterizations of the input-to-state stability property. *Systems & Control Letters*, 24(5):351–359, 1995.

[SWV$^+$01] I. G. Stiell, G. A. Wells, K. Vandemheen, C. Clement, H. Lesiuk, A. Laupacis, R. D. McKnight, R. Verbeek, R. Brison, D. Cass, M. A. Eisenhauer, G. H. Greenberg, and J. Worthington. The Canadian CT Head Rule for patients with minor head injury. *The Lancet*, 357(9266):1391–1396, 2001.

[Tan03] T. Taniguchi. The asymptotic behaviour of solutions of stochastic functional differential equations with finite delays by the Lyapunov-Razumikhin method. In A. A. Martynyuk, editor, *Advances in Stability Theory at the End of the 20th Century*, volume 13 of *Stability and Control: Theory, Methods and Applications*, pages 175–188. Taylor and Francis, 2003.

[TBA86] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, 1986.

[Thy97] U. Høgsbro Thygesen. A survey of Lyapunov techniques for stochastic differential equations. Technical Report 18-1997, Department of Mathematical Modeling, Technical University of Denmark, 1997.

[TLR12] K. I. Tsianos, S. Lawlor, and M. G. Rabbat. Push-sum distributed dual averaging for convex optimization. In *IEEE Conf. on Decision and Control*, pages 5453–5458, Maui, HI, 2012.

[TR12] K. I. Tsianos and M. G. Rabbat. Distributed strongly convex optimization. In *Allerton Conf. on Communications, Control and Computing*, pages 593–600, Monticello, IL, October 2012.

[Tsi84] J. N. Tsitsiklis. *Problems in Decentralized Decision Making and Computation.* PhD thesis, Massachusetts Institute of Technology, November 1984. Available at http://web.mit.edu/jnt/www/Papers/PhD-84-jnt.pdf.

[WB12] H. Wang and A. Banerjee. Online alternating direction method. In *International Conference on Machine Learning*, pages 1119–1126, Edinburgh, Scotland, July 2012.

[WC15] R. Witten and E. Candès. Randomized algorithms for low-rank matrix factorizations: Sharp performance bounds. *Algorithmica*, 72(1):264–281, 2015.

[WE10] J. Wang and N. Elia. Control approach to distributed optimization. In *Allerton Conf. on Communications, Control and Computing*, pages 557–561, Monticello, IL, October 2010.

[WE11] J. Wang and N. Elia. A control perspective for centralized and distributed convex optimization. In *IEEE Conf. on Decision and Control*, pages 3800–3805, Orlando, Florida, 2011.

[WK13] F. Wu and P. E. Kloeden. Mean-square random attractors of stochastic delay differential equations with random delay. *Discrete and Continuous Dynamical Systems - Series B (DCDS-B)*, 18:1715–1734, 2013.

[WL09] P. Wan and M. D. Lemmon. Event-triggered distributed optimization in sensor networks. In *Symposium on Information Processing of Sensor Networks*, pages 49–60, San Francisco, CA, 2009.

[WLRT08] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25(3):335–366, 2008.

[WO12] E. Wei and A. Ozdaglar. Distributed alternating direction method of multipliers. In *IEEE Conf. on Decision and Control*, pages 5445–5450, Maui, HI, 2012.

[WS98] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[WXZ07] Z.-J. Wu, X.-J. Xie, and S.-Y. Zhang. Adaptive backstepping controller design using stochastic small-gain theorem. *Automatica*, 43(4):608–620, 2007.

[WYZ12] Z. Wen, W. Yin, and Y. Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361, 2012.

[YHX15] D. Yuan, D. W. C. Ho, and S. Xu. Regularized primal-dual subgradient method for distributed constrained optimization. *IEEE Transactions on Cybernetics*, 2015. To appear.

[YL07] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B*, 68(1):49–57, 2007.

[YSVQ13] F. Yan, S. Sundaram, S. V. N. Vishwanathan, and Y. Qi. Distributed autonomous online learning: regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2483–2493, 2013.

[YXZ11] D. Yuan, S. Xu, and H. Zhao. Distributed primal-dual subgradient method for multiagent optimization via consensus algorithms. *IEEE Trans. Systems, Man, and Cybernetics- Part B*, 41(6):1715–1724, 2011.

[Zin03] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 928–936, Washighton, D.C., 2003.

[ZM96] D. L. Zhu and P. Marcotte. Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM Journal on Optimization*, 6(3):714–726, 1996.

[ZM12] M. Zhu and S. Martínez. On distributed convex optimization under inequality and equality constraints. *IEEE Transactions on Automatic Control*, 57(1):151–164, 2012.

[ZVC$^+$11] F. Zanella, D. Varagnolo, A. Cenedese, G. Pillonetto, and L. Schenato. Newton-Raphson consensus for distributed convex optimization. In *IEEE Conf. on Decision and Control*, pages 5917–5922, Orlando, Florida, December 2011.