Distributed saddle-point subgradient algorithms with Laplacian averaging

David Mateos-Núñez Jorge Cortés

***Abstract*—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.**

I. Introduction

Saddle-point problems arise in constrained optimization via the Lagrangian formulation and, more generally, are equivalent to variational inequality problems. These formulations find applications in cooperative control of multi-agent systems, in machine learning and game theory, and in equilibrium problems in networked systems, motivating the study of distributed strategies that are guaranteed to converge, scale well with the number of agents, and are robust against a variety of failures and uncertainties. Our objective in this paper is to design and analyze distributed algorithms to solve general convex-concave saddle-point problems.

Literature review: This work builds on three related areas: iterative methods for saddle-point problems [2], [3], dual decompositions for constrained optimization [4, Ch. 5], [5], and consensus-based distributed optimization algorithms; see, e.g., [6], [7], [8], [9], [10], [11] 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 [12], [13], [14], leading to consensus approaches that allow different processors with local memories to update the same components of a vector by averaging their estimates. 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. [15], and are equivalent to the variational inequality framework [16], which includes as particular cases constrained optimization and many other equilibrium models relevant to networked systems, including traffic [17] and supply chain [18]. In a centralized scenario, the work [2] 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, [3] 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 [3], [19], 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 [20] 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.

Dual decomposition methods for constrained optimization are the melting pot where saddle-point approaches come together with methods for parallelizing computations, like the alternating direction method of multipliers [5]. These methods rely on a particular approach to split a sum of convex objectives by introducing agreement constraints on copies of the primal variable, leading to distributed strategies such as distributed primal-dual subgradient methods [8], [11] where the vector of Lagrange multipliers associated with the Laplacian's nullspace is updated by the agents using local communication. Ultimately, these methods allow to distribute global constraints that are sums of convex functions via *agreement on the multipliers* [21], [22], [23]. Regarding 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., [24], [9], [25], where all the agents know the constraint, or see, e.g., [26], [27], [9], where the constraint is given by the intersection of abstract closed convex sets. The second type *couples local* decision vectors across the network, and is addressed by [28] with linear equality constraints, by [21] with linear inequalities, by [22] with inequalities given by the sum of convex functions on local decision vectors, where each one is only known to the corresponding agent, and by [23] with semidefinite constraints. The work [28] considers a distinction, that we also adopt here, between *constraint graph* (where edges arise from participation in a constraint) and *communication graph*, generalizing other paradigms where each agent needs to communicate with all other agents involved in a particular constraint [29], [30]. When applied to distributed optimization, our work considers both kinds of constraints, and along with [28], [21], [22],

A preliminary version of this work has been accepted as [1] at the 2015 IEEE Conference on Decision and Control, Osaka, Japan

The authors are with the Department of Mechanical and Aerospace Engineering, University of California, San Diego, USA, davidmateosmateos@gmail.com, cortes@ucsd.edu.

[23], has the crucial feature that agents participating in the same constraint are able to coordinate their decisions without direct communication. This approach has been successfully applied to control of camera networks [31] and decomposable semidefinite programs [32]. 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. Our novel contribution tackles these constraints from a more general perspective, namely, we provide a multi-agent distributed approach for the *general saddle-point problems* 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 [3, Sec. 3] and the kind of linear proportional feedback on the disagreement typical of consensus-based approaches, see e.g., [6], [7], [9], in distributed convex optimization. The resulting family of algorithms solve more general saddle-point problems than existing algorithms in the literature in a decentralized way, and also 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 [22] uses primal-dual perturbed methods which enhance subgradient algorithms by evaluating the latter at precomputed arguments called *perturbation points*. These auxiliary computations require additional subgradient methods or proximal methods that add to the computation and the communication complexity. Similarly, the work [23] considers primal-dual methods, where each agent performs a minimization of the local component of the Lagrangian with respect to its primal variable (instead of computing a subgradient step). Notably, this work makes explicit the treatment of semidefinite constraints. The work [21] applies the *Cutting-Plane Consensus* algorithm to the dual optimization problem under linear constraints. The decentralization feature is the same but the computational complexity of the local problems grows with the number of agents. The generality of our approach stems from the fact that our saddle-point strategy is applicable beyond the case of Lagrangians in constrained optimization. In fact, we have recently considered in [33] distributed optimization problems with nuclear norm regularization via a min-max formulation of the nuclear norm where the convex-concave functions involved have, unlike Lagrangians, a quadratic concave part.

Statement of contributions: 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. Various simulations illustrate our results.

II. Preliminaries

Here we introduce basic notation and notions from graph theory and optimization used throughout the paper.

A. Notational conventions

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 $\mathbb{R}^n$, and by $\mathbb{1}_n \in \mathbb{R}^n$ the vector of all 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 $i = 1, \ldots, n$. Given a vector $v \in \mathbb{R}^n$, we denote its Euclidean norm, or two-norm, by $\|v\|_2 = \sqrt{\sum_{i=1}^n v_i^2}$ and the one-norm by $\|v\|_1 = \sum_{i=1}^n |v_i|$. Given a convex set $\mathcal{S} \subseteq \mathbb{R}^n$, a function $f : \mathcal{S} \to \mathbb{R}$ is convex if $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$ for all $\alpha \in [0,1]$ and $x, y \in \mathcal{S}$. A vector $\xi_x \in \mathbb{R}^n$ is a subgradient of f at $x \in \mathcal{S}$ if $f(y) - f(x) \geq \xi_x^\top (y - x)$, for all $y \in \mathcal{S}$. We denote by $\partial f(x)$ the set of all such subgradients. The function f is concave if $-f$ is convex. A vector $\xi_x \in \mathbb{R}^n$ is a subgradient of a concave function f at $x \in \mathcal{S}$ if $-\xi_x \in \partial(-f)(x)$. Given a *closed* convex set $\mathcal{S} \subseteq \mathbb{R}^n$, the orthogonal projection $\mathcal{P}_{\mathcal{S}}$ onto $\mathcal{S}$ is

$$\mathcal{P}_{\mathcal{S}}\big(x\big) \in \arg\min_{x' \in \mathcal{S}} \|x - x'\|_2. \tag{1}$$

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

$$\big(\mathcal{P}_{\mathcal{S}}(x') - x'\big)(x' - x) \leq 0. \tag{2}$$

For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, we denote by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ its minimum and maximum eigenvalues, and for any matrix A, we denote by $\sigma_{\max}(A)$ its maximum singular value. We use $\otimes$ to denote the Kronecker product of matrices.

B. Graph theory

We review basic notions from graph theory following [34]. A (weighted) digraph $\mathcal{G} := (\mathcal{I}, \mathcal{E}, \mathsf{A})$ is a triplet where $\mathcal{I} := \{1, \dots, N\}$ 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 $\big\{\mathcal{G}_t := (\mathcal{I}, \mathcal{E}_t, \mathsf{A}_t)\big\}_{t\geq 1}$ 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 Laplacian matrix $\mathsf{L} \in \mathbb{R}^{N\times N}$ of a digraph $\mathcal{G}$ is $\mathsf{L} := \operatorname{diag}(\mathsf{A}\mathbb{1}_N) - \mathsf{A}$. Note that $\mathsf{L}\mathbb{1}_N = 0_N$. The weighted out-degree and in-degree of $i \in \mathcal{I}$ are, respectively, $d_{\text{out}}(i) := \sum_{j=1}^N \mathsf{a}_{ij}$ and $d_{\text{in}}(i) := \sum_{j=1}^N \mathsf{a}_{ji}$. A digraph is weight-balanced if $d_{\text{out}}(i) = d_{\text{in}}(i)$ for all $i \in \mathcal{I}$, that is, $\mathbb{1}_N^\top \mathsf{L} = 0_N$. For convenience, we let $\mathsf{L}_{\mathcal{K}} := \mathrm{I}_N - \frac{1}{N}\mathbb{1}_N\mathbb{1}_N^\top$ denote the Laplacian of the complete graph with edge weights $1/N$. Note that $\mathsf{L}_{\mathcal{K}}$ is idempotent, i.e., $\mathsf{L}_{\mathcal{K}}^2 = \mathsf{L}_{\mathcal{K}}$. For the sake of the reader, Table I collects some shorthand notation.

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

TABLE I: Notation for graph matrices employed along the paper, where the dimension d depends on the context.

C. Optimization and saddle points

For any function $\mathcal{L} : \mathcal{W} \times \mathcal{M} \to \mathbb{R}$, the *max-min inequality* [35, Sec 5.4.1] states that

$$\inf_{w\in\mathcal{W}} \sup_{\mu\in\mathcal{M}} \mathcal{L}(w,\mu) \geq \sup_{\mu\in\mathcal{M}} \inf_{w\in\mathcal{W}} \mathcal{L}(w,\mu). \tag{3}$$

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

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

[15, 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

$$\min_{w\in\mathbb{R}^n} f(w) \quad \text{s.t.} \quad g(w) \leq 0,\ h(w) = 0, \tag{4}$$

is defined as

$$\mathcal{L}(w, \mu, \lambda) = f(w) + \mu^\top g(w) + \lambda^\top h(w) \tag{5}$$

for $(\mu, \lambda) \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p$. In this case, inequality (3) 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 (4) 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* [36, 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

$$f^\star(x) := \sup_{w\in\mathbb{R}^n} \{x^\top w - f(w)\}. \tag{6}$$

III. Distributed algorithms for saddle-point problems under agreement constraints

This section describes the problem of interest. Consider closed convex sets $\boldsymbol{\mathcal{W}} \subseteq \mathbb{R}^{d_{\boldsymbol{w}}}$, $\mathcal{D} \subseteq \mathbb{R}^{d_{\mathcal{D}}}$, $\boldsymbol{\mathcal{M}} \subseteq \mathbb{R}^{d_{\boldsymbol{\mu}}}$, $\mathcal{Z} \subseteq \mathbb{R}^{d_{\boldsymbol{z}}}$ and a function $\boldsymbol{\phi} : \boldsymbol{\mathcal{W}} \times \mathcal{D}^N \times \boldsymbol{\mathcal{M}} \times \mathcal{Z}^N \to \mathbb{R}$ 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:

$$\min_{\substack{\boldsymbol{w}\in\boldsymbol{\mathcal{W}},\, \boldsymbol{D}\in\mathcal{D}^N \\ D^i = D^j,\, \forall i,j}} \ \max_{\substack{\boldsymbol{\mu}\in\boldsymbol{\mathcal{M}},\, \boldsymbol{z}\in\mathcal{Z}^N \\ z^i = z^j,\, \forall i,j}} \boldsymbol{\phi}(\boldsymbol{w}, \boldsymbol{D}, \boldsymbol{\mu}, \boldsymbol{z}), \tag{7}$$

where $\boldsymbol{D} := (D^1, \dots, D^N)$ and $\boldsymbol{z} := (z^1, \dots, z^N)$. The motivation for distributed algorithms and the consideration of explicit agreement constraints in (7) 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.

A. 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 (7) 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 $f^i:\mathbb{R}^{n_i}\times\mathbb{R}^{d_D}\to\mathbb{R}$ and the components of $g^i:\mathbb{R}^{n_i}\times\mathbb{R}^{d_D}\to\mathbb{R}^m$ 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 $\mathcal{D}\subseteq\mathbb{R}^{d_D}$ convex. The optimization problem reads as

$$\begin{aligned}\min_{\substack{w^i\in\mathcal{W}_i,\,\forall i\\ D\in\mathcal{D}}}\ &\sum_{i=1}^N f^i(w^i,D)\\ &\text{s.t. } g^1(w^1,D)+\cdots+g^N(w^N,D)\le 0.\end{aligned}\tag{8}$$

This problem can be reformulated as a constrained saddle-point problem as follows. We first construct the corresponding Lagrangian function (5) and introduce copies $\{z^i\}_{i=1}^N$ of the Lagrange multiplier z associated to the global constraint in (8), 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 (8). Formally,

$$\min_{\substack{w^i\in\mathcal{W}_i\\ D\in\mathcal{D}}}\ \max_{z\in\mathbb{R}^m_{\ge 0}}\ \sum_{i=1}^N f^i(w^i,D)+z^\top\sum_{i=1}^N g^i(w^i,D)\tag{9a}$$

$$=\min_{\substack{w^i\in\mathcal{W}_i\\ D\in\mathcal{D}}}\ \max_{\substack{z^i\in\mathbb{R}^m_{\ge 0}\\ z^i=z^j,\,\forall i,j}}\ \sum_{i=1}^N\big(f^i(w^i,D)+{z^i}^\top g^i(w^i,D)\big)\tag{9b}$$

$$=\min_{\substack{w^i\in\mathcal{W}_i\\ D^i\in\mathcal{D}\\ D^i=D^j,\,\forall i,j}}\ \max_{\substack{z^i\in\mathbb{R}^m_{\ge 0}\\ z^i=z^j,\,\forall i,j}}\ \sum_{i=1}^N\big(f^i(w^i,D^i)+{z^i}^\top g^i(w^i,D^i)\big).\tag{9c}$$

This formulation has its roots in the classical dual decompositions surveyed in [5, Ch. 2], see also [37, Sec. 1.2.3] and [4, Sec. 5.4] for the particular case of resource allocation. While [5], [37] suggest to broadcast a centralized update of the multiplier, and the method in [4] has an implicit projection onto the probability simplex, the formulation (9) has the multiplier associated to the global constraint estimated in a decentralized way. The recent works [21], [22], [23] implicitly rest on the above formulation of *agreement on the multipliers* Section V particularizes our general saddle-point strategy to these distributed scenarios.

Remark III.1. (Distributed formulations via Fenchel conjugates): To illustrate the generality of the min-max problem (9c), 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

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

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 (6). 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

$$\max_{z\in\mathcal{Z}}\min_{w^i\in\mathbb{R}^{n_i},\,\forall i}\sum_{i=1}^N f^i(w^i)+\sum_{i=1}^N(z^\top A^i w^i - z^\top b^i)\tag{10a}$$

$$=\max_{z\in\mathcal{Z}}\sum_{i=1}^N\big(-{f^i}^\star(-{A^i}^\top z)-z^\top b^i\big)\tag{10b}$$

$$=\max_{\substack{z^i\in\mathcal{Z},\,\forall i\\ z^i=z^j,\,\forall i,j}}\sum_{i=1}^N\big(-{f^i}^\star(-{A^i}^\top z^i)-{z^i}^\top b^i\big),\tag{10c}$$

where $\mathcal{Z}$ is either $\mathbb{R}^m$ or $\mathbb{R}^m_{\ge 0}$ depending on whether we have equality or inequality ($\le$) constraints in (8). By [38, Prop. 11.3], the optimal primal values can be recovered locally as

$$w^{i^*}:=\partial {f^i}^\star(-{A^i}^\top z^{i*}),\qquad i\in\{1,\dots,N\}\tag{11}$$

without extra communication. Thus, our strategy generalizes the class of convex optimization problems with linear constraints studied in [28], 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. •

B. Saddle-point dynamics with Laplacian averaging

We propose a projected subgradient method to solve constrained saddle-point problems of the form (7). The agreement constraints are addressed via Laplacian averaging, allowing the design of distributed algorithms *when* the convex-concave functions are separable as in Sections III-A. The generality of this dynamics is inherited by the general structure of the convex-concave min-max problem (7). 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., [16] regarding the variational inequality framework, which is equivalent to the saddle-point framework. Formally, the dynamics is

$$\hat{\boldsymbol{w}}_{t+1}=\boldsymbol{w}_t-\eta_t g_{\boldsymbol{w}_t}\tag{12a}$$

$$\hat{\boldsymbol{D}}_{t+1}=\boldsymbol{D}_t-\sigma\mathbf{L}_t\boldsymbol{D}_t-\eta_t g_{\boldsymbol{D}_t}\tag{12b}$$

$$\hat{\boldsymbol{\mu}}_{t+1}=\boldsymbol{\mu}_t+\eta_t g_{\boldsymbol{\mu}_t}\tag{12c}$$

$$\hat{\boldsymbol{z}}_{t+1}=\boldsymbol{z}_t-\sigma\mathbf{L}_t\boldsymbol{z}_t+\eta_t g_{\boldsymbol{z}_t}\tag{12d}$$

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

where $\mathbf{L}_t=\mathsf{L}_t\otimes\mathrm{I}_{d_D}$ 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, $\{\eta_t\}_{t\ge 1}\subset\mathbb{R}_{>0}$ 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 $\boldsymbol{S} := \boldsymbol{\mathcal{W}} \times \mathcal{D}^N \times \boldsymbol{\mathcal{M}} \times \mathcal{Z}^N$ as defined in (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 III-A. In general, the dynamics (12) 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 (12) can be computed with *local* information *depends* on the structure of the function $\boldsymbol{\phi}$ in (7) 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 $\boldsymbol{\phi}$.

IV. Convergence analysis

Here we present our technical analysis on the convergence properties of the dynamics (12). Our starting point is the assumption that a solution to (7) exists, namely, a saddle point $(\boldsymbol{w}^*, \boldsymbol{D}^*, \boldsymbol{\mu}^*, \boldsymbol{z}^*)$ of $\boldsymbol{\phi}$ on $\boldsymbol{S} := \boldsymbol{\mathcal{W}} \times \mathcal{D}^N \times \boldsymbol{\mathcal{M}} \times \mathcal{Z}^N$ 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., [23])

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

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

$$t\boldsymbol{\phi}(\boldsymbol{w}_{t+1}^{\mathrm{av}}, \boldsymbol{D}_{t+1}^{\mathrm{av}}, \boldsymbol{\mu}_{t+1}^{\mathrm{av}}, \boldsymbol{z}_{t+1}^{\mathrm{av}}) - t\boldsymbol{\phi}(\boldsymbol{w}^*, \boldsymbol{D}^*, \boldsymbol{\mu}^*, \boldsymbol{z}^*). \quad (13)$$

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 IV-B we bound the cumulative disagreement in terms of the size of the subgradients and the learning rates. Finally, in Section IV-C we establish the saddle-point evaluation convergence result using the assumption that the estimates generated by the dynamics (12), 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 [39, Sec. 2.3.1] to find lower and upper bounds on (13) 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 $1/\sqrt{t}$.

A. 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 IV.1. (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 (12) 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*

$$\sigma \leq \big(\max\big\{\, d_{\mathrm{out,t}}(k) \;:\; k \in \mathcal{I},\; t \in \mathbb{Z}_{\geq 1} \,\big\}\big)^{-1}. \quad (14)$$

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

$$\begin{aligned} &2(\boldsymbol{\phi}(\boldsymbol{w}_t, \boldsymbol{D}_t, \boldsymbol{\mu}_t, \boldsymbol{z}_t) - \boldsymbol{\phi}(\boldsymbol{w}_p, \boldsymbol{D}_p, \boldsymbol{\mu}_t, \boldsymbol{z}_t)) \\ &\leq \tfrac{1}{\eta_t}\big(\|\boldsymbol{w}_t - \boldsymbol{w}_p\|_2^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{w}_p\|_2^2\big) \\ &\quad + \tfrac{1}{\eta_t}\big(\|\mathbf{M}\boldsymbol{D}_t - \boldsymbol{D}_p\|_2^2 - \|\mathbf{M}\boldsymbol{D}_{t+1} - \boldsymbol{D}_p\|_2^2\big) \\ &\quad + 6\eta_t \|g_{\boldsymbol{w}_t}\|_2^2 + 6\eta_t \|g_{\boldsymbol{D}_t}\|_2^2 \\ &\quad + 2\|g_{\boldsymbol{D}_t}\|_2(2 + \sigma\overline{\Lambda})\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_t\|_2 + 2\|g_{\boldsymbol{D}_t}\|_2\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_p\|_2. \end{aligned} \quad (15)$$

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

$$\begin{aligned} &2(\boldsymbol{\phi}(\boldsymbol{w}_t, \boldsymbol{D}_t, \boldsymbol{\mu}_t, \boldsymbol{z}_t) - \boldsymbol{\phi}(\boldsymbol{w}_t, \boldsymbol{D}_t, \boldsymbol{\mu}_p, \boldsymbol{z}_p)) \\ &\geq -\tfrac{1}{\eta_t}\big(\|\boldsymbol{\mu}_t - \boldsymbol{\mu}_p\|_2^2 - \|\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_p\|_2^2\big) \\ &\quad - \tfrac{1}{\eta_t}\big(\|\mathbf{M}\boldsymbol{z}_t - \boldsymbol{z}_p\|_2^2 - \|\mathbf{M}\boldsymbol{z}_{t+1} - \boldsymbol{z}_p\|_2^2\big) \\ &\quad - 6\eta_t \|g_{\boldsymbol{\mu}_t}\|_2^2 - 6\eta_t \|g_{\boldsymbol{z}_t}\|_2^2 \\ &\quad - 2\|g_{\boldsymbol{z}_t}\|_2(2 + \sigma\overline{\Lambda})\|\mathbf{L}_{\mathcal{K}}\boldsymbol{z}_t\|_2 - 2\|g_{\boldsymbol{z}_t}\|_2\|\mathbf{L}_{\mathcal{K}}\boldsymbol{z}_p\|_2. \end{aligned} \quad (16)$$

Building on Lemma IV.1, 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 IV.2. (Cumulative evaluation error of the states with respect to running time-averages in terms of disagreement): *Under the same assumptions of Lemma IV.1, for any $(\boldsymbol{w}_p, \boldsymbol{D}_p, \boldsymbol{\mu}_p, \boldsymbol{z}_p) \in \boldsymbol{\mathcal{W}} \times \mathcal{D}^N \times \boldsymbol{\mathcal{M}} \times \mathcal{Z}^N$, 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 $\frac{\mathrm{u}(t, \boldsymbol{w}_p, \boldsymbol{D}_p)}{2}$, 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*

$$\begin{aligned}
&\mathsf{u}(t,\boldsymbol{w}_p,\boldsymbol{D}_p) \equiv \mathsf{u}\big(t,\boldsymbol{w}_p,\boldsymbol{D}_p,\{\boldsymbol{w}_s\}_{s=1}^t,\{\boldsymbol{D}_s\}_{s=1}^t\big) \qquad (17)\\
&= \sum_{s=2}^{t}\big(\|\boldsymbol{w}_s-\boldsymbol{w}_p\|_2^2+\|\mathbf{M}\boldsymbol{D}_s-\boldsymbol{D}_p\|_2^2\big)\big(\tfrac{1}{\eta_s}-\tfrac{1}{\eta_{s-1}}\big)\\
&\quad+\tfrac{2}{\eta_1}\big(\|\boldsymbol{w}_1\|_2^2+\|\boldsymbol{w}_p\|_2^2+\|\boldsymbol{D}_1\|_2^2+\|\boldsymbol{D}_p\|_2^2\big)\\
&\quad+6\sum_{s=1}^{t}\eta_s(\|g_{\boldsymbol{w}_s}\|_2^2+\|g_{\boldsymbol{D}_s}\|_2^2)\\
&+2(2+\sigma\overline{\Lambda})\sum_{s=1}^{t}\|g_{\boldsymbol{D}_s}\|_2\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_s\|_2+2\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_p\|_2\sum_{s=1}^{t}\|g_{\boldsymbol{D}_s}\|_2, \qquad (18)
\end{aligned}$$

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 (15) over $s=1,\dots,t$, we obtain

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

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 sub-multiplicativity of the norm, and the identity $\|\mathbf{M}\|_2=1$. Finally, by the concavity of $\boldsymbol{\phi}$ in the last two arguments,

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

so the upper bound in the statement follows. Similarly, we obtain the lower bound by adding (16) over $s=1,\dots,t$ and using that $\boldsymbol{\phi}$ is jointly convex in the first two arguments,

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

which completes the proof. ∎

The combination of the pair of inequalities in Lemma IV.2 allows us to derive the saddle-point evaluation error of the running time-averages in the next result.

Proposition IV.3. (Saddle-point evaluation error of running time-averages): *Under the same hypotheses of Lemma IV.1, for any saddle point* $(\boldsymbol{w}^*,\boldsymbol{D}^*,\boldsymbol{\mu}^*,\boldsymbol{z}^*)$ *of* $\boldsymbol{\phi}$ *on* $\boldsymbol{\mathcal{W}}\times\mathcal{D}^N\times\boldsymbol{\mathcal{M}}\times\mathcal{Z}^N$ *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:*

$$\begin{aligned}
&-\mathsf{u}(t,\boldsymbol{\mu}^*,\boldsymbol{z}^*)-\mathsf{u}(t,\boldsymbol{w}_{t+1}^{av},\boldsymbol{D}_{t+1}^{av})\\
&\le 2t\boldsymbol{\phi}(\boldsymbol{w}_{t+1}^{av},\boldsymbol{D}_{t+1}^{av},\boldsymbol{\mu}_{t+1}^{av},\boldsymbol{z}_{t+1}^{av})-2t\boldsymbol{\phi}(\boldsymbol{w}^*,\boldsymbol{D}^*,\boldsymbol{\mu}^*,\boldsymbol{z}^*)\\
&\le \mathsf{u}(t,\boldsymbol{w}^*,\boldsymbol{D}^*)+\mathsf{u}(t,\boldsymbol{\mu}_{t+1}^{av},\boldsymbol{z}_{t+1}^{av})\,. \qquad (19)
\end{aligned}$$

Proof: We show the result in two steps, by evaluating the bounds from Lemma IV.2 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 IV.2; invoking the saddle-point relations

$$\begin{aligned}
\boldsymbol{\phi}(\boldsymbol{w}^*,\boldsymbol{D}^*,\boldsymbol{\mu}_{t+1}^{\mathrm{av}},\boldsymbol{z}_{t+1}^{\mathrm{av}}) &\le \boldsymbol{\phi}(\boldsymbol{w}^*,\boldsymbol{D}^*,\boldsymbol{\mu}^*,\boldsymbol{z}^*)\\
&\le \boldsymbol{\phi}(\boldsymbol{w}_{t+1}^{\mathrm{av}},\boldsymbol{D}_{t+1}^{\mathrm{av}},\boldsymbol{\mu}^*,\boldsymbol{z}^*)
\end{aligned}$$

where $(\boldsymbol{w}_t^{\mathrm{av}},\boldsymbol{D}_t^{\mathrm{av}},\boldsymbol{\mu}_t^{\mathrm{av}},\boldsymbol{z}_t^{\mathrm{av}})\in\boldsymbol{\mathcal{W}}\times\mathcal{D}^N\times\boldsymbol{\mathcal{M}}\times\mathcal{Z}^N$, for each $t\ge1$, by convexity; and combining the resulting inequalities, we obtain

$$\begin{aligned}
-\frac{\mathsf{u}(t,\boldsymbol{\mu}^*,\boldsymbol{z}^*)}{2} &\le \sum_{s=1}^{t}\boldsymbol{\phi}(\boldsymbol{w}_s,\boldsymbol{D}_s,\boldsymbol{\mu}_s,\boldsymbol{z}_s)-t\boldsymbol{\phi}(\boldsymbol{w}^*,\boldsymbol{D}^*,\boldsymbol{\mu}^*,\boldsymbol{z}^*)\\
&\le \frac{\mathsf{u}(t,\boldsymbol{w}^*,\boldsymbol{D}^*)}{2}. \qquad (20)
\end{aligned}$$

Choosing $(\boldsymbol{w}_p,\boldsymbol{D}_p,\boldsymbol{\mu}_p,\boldsymbol{z}_p)=(\boldsymbol{w}_{t+1}^{\mathrm{av}},\boldsymbol{D}_{t+1}^{\mathrm{av}},\boldsymbol{\mu}_{t+1}^{\mathrm{av}},\boldsymbol{z}_{t+1}^{\mathrm{av}})$ in the bounds of Lemma IV.2, multiplying each by -1 and combining them, we get

$$\begin{aligned}
-\frac{\mathsf{u}(t,\boldsymbol{w}_{t+1}^{\mathrm{av}},\boldsymbol{D}_{t+1}^{\mathrm{av}})}{2} &\le \Big(t\boldsymbol{\phi}(\boldsymbol{w}_{t+1}^{\mathrm{av}},\boldsymbol{D}_{t+1}^{\mathrm{av}},\boldsymbol{\mu}_{t+1}^{\mathrm{av}},\boldsymbol{z}_{t+1}^{\mathrm{av}})\\
&\quad-\sum_{s=1}^{t}\boldsymbol{\phi}(\boldsymbol{w}_s,\boldsymbol{D}_s,\boldsymbol{\mu}_s,\boldsymbol{z}_s)\Big)\le\frac{\mathsf{u}(t,\boldsymbol{\mu}_{t+1}^{\mathrm{av}},\boldsymbol{z}_{t+1}^{\mathrm{av}})}{2}. \qquad (21)
\end{aligned}$$

The result now follows by summing (20) and (21). ∎

B. Bounding the cumulative disagreement

Given the dependence of the saddle-point evaluation error obtained in Proposition IV.3 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 (12) 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 (12) 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,

$$\hat{\boldsymbol{D}}_{t+1} = \boldsymbol{D}_t-\sigma\mathbf{L}_t\boldsymbol{D}_t+\boldsymbol{u}_t^1 \qquad (22a)$$

$$\hat{\boldsymbol{z}}_{t+1} = \boldsymbol{z}_t-\sigma\mathbf{L}_t\boldsymbol{z}_t+\boldsymbol{u}_t^2 \qquad (22b)$$

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

where $\{\boldsymbol{u}_t^1\}_{t\ge1}\subset(\mathbb{R}^{d_{\mathcal{D}}})^N$, $\{\boldsymbol{u}_t^2\}_{t\ge1}\subset(\mathbb{R}^{d_{\mathcal{Z}}})^N$ 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 (1).

The next result characterizes the input-to-state stability properties of (22) with respect to the agreement space. The analysis builds on the proof strategy in our previous work [40, 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 IV.4. (Cumulative disagreement on (22) **over jointly-connected weight-balanced digraphs):** *Let* $\{\mathcal{G}_s\}_{s\ge1}$ *be a sequence of* B*-jointly connected,* δ*-nondegenerate, weight-balanced digraphs. For* $\tilde{\delta}'\in(0,1)$*, let*

$$\tilde{\delta} := \min\big\{\,\tilde{\delta}',\ (1-\tilde{\delta}')\frac{\delta}{d_{\max}}\,\big\}, \qquad (23)$$

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

$$\sigma \in \Big[\, \frac{\tilde{\delta}}{\delta},\; \frac{1-\tilde{\delta}}{d_{\max}} \,\Big], \tag{24}$$

the dynamics (22a) *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* $\{\boldsymbol{u}_s^1\}_{s=1}^{t-1} \subset (\mathbb{R}^{d_D})^N$*,*

$$\|\mathbf{L}_{\mathcal{K}} \boldsymbol{D}_t\|_2 \leq \tfrac{2^4 \|\boldsymbol{D}_1\|_2}{3^2} \Big(1 - \frac{\tilde{\delta}}{4N^2}\Big)^{\lceil \frac{t-1}{B} \rceil} + C_u \max_{1\leq s\leq t-1} \|\boldsymbol{u}_s^1\|_2\,, \tag{25}$$

where

$$C_u := \frac{2^5/3^2}{1 - \big(1 - \frac{\tilde{\delta}}{4N^2}\big)^{1/B}} \tag{26}$$

and the cumulative disagreement satisfies

$$\sum_{t=1}^{t'} \|\mathbf{L}_{\mathcal{K}} \boldsymbol{D}_t\|_2 \leq C_u \Big(\tfrac{\|\boldsymbol{D}_1\|_2}{2} + \sum_{t=1}^{t'-1} \|\boldsymbol{u}_t^1\|_2 \Big). \tag{27}$$

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

C. Convergence of saddle-point subgradient dynamics with Laplacian averaging

Here we characterize the convergence properties of the dynamics (12) 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 (12) decreases proportionally to $\frac{1}{\sqrt{t}}$. We select the learning rates according to the following scheme.

Assumption IV.5. (*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,*

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

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 IV.5 without communicating with each other. Figure 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 [14], [15], the Doubling Trick gives rise to a sequence of stepsizes that is not square summable.

Theorem IV.6. (Convergence of the *saddle-point dynamics with Laplacian averaging* (12)): *Let* $\{(\boldsymbol{w}_t, \boldsymbol{D}_t, \boldsymbol{\mu}_t, \boldsymbol{z}_t)\}_{t\geq 1}$ *be*

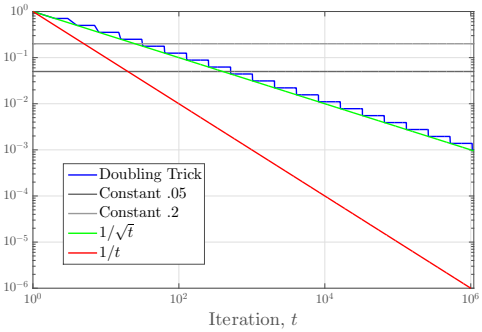


Fig. 1: Comparison of sequences of learning rates. We compare the sequence of learning rates resulting from the Doubling Trick in Assumption IV.5 against a constant stepsize, the sequence $\{1/\sqrt{t}\}_{t\geq 1}$, and the square-summable harmonic sequence $\{1/t\}_{t\geq 1}$.

generated by (12) *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* (24)*. 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 IV.5. Then, for any saddle point* $(\boldsymbol{w}^*, \boldsymbol{D}^*, \boldsymbol{\mu}^*, \boldsymbol{z}^*)$ *of* $\boldsymbol{\phi}$ *on* $\boldsymbol{\mathcal{W}} \times \mathcal{D}^N \times \boldsymbol{\mathcal{M}} \times \mathcal{Z}^N$ *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:*

$$-\frac{\alpha_{\boldsymbol{\mu},\boldsymbol{z}} + \alpha_{\boldsymbol{w},\boldsymbol{D}}}{2\sqrt{t-1}} \leq \boldsymbol{\phi}(\boldsymbol{w}_t^{av}, \boldsymbol{D}_t^{av}, \boldsymbol{z}_t^{av}, \boldsymbol{\mu}_t^{av}) - \boldsymbol{\phi}(\boldsymbol{w}^*, \boldsymbol{D}^*, \boldsymbol{z}^*, \boldsymbol{\mu}^*) \leq \frac{\alpha_{\boldsymbol{w},\boldsymbol{D}} + \alpha_{\boldsymbol{\mu},\boldsymbol{z}}}{2\sqrt{t-1}}, \tag{28}$$

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

$$\hat{\alpha}_{\boldsymbol{w},\boldsymbol{D}} := 4(B_{\boldsymbol{w}}^2 + B_{\boldsymbol{D}}^2) + 6(H_{\boldsymbol{w}}^2 + H_{\boldsymbol{D}}^2) + H_{\boldsymbol{D}}(3 + \sigma\overline{\Lambda}) C_u \big(B_{\boldsymbol{D}} + 2H_{\boldsymbol{D}}\big)\,, \tag{29}$$

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 IV.3 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 (19), we choose $\eta_t = \eta'$ for all $s \in \{1, \ldots, t'\}$ in both $\mathsf{u}(t', \boldsymbol{w}^*, \boldsymbol{D}^*)$ and $\mathsf{u}(t', \boldsymbol{w}_{t'+1}^{\text{av}}, \boldsymbol{D}_{t'+1}^{\text{av}})$. By doing this, we make zero the first two lines in (17), and then we upper-bound the remaining terms using the bounds on the estimates and the subgradients. The resulting inequality also holds replacing $(\boldsymbol{w}_{t'+1}^{\text{av}}, \boldsymbol{D}_{t'+1}^{\text{av}})$ by

$(\boldsymbol{w}^*, \boldsymbol{D}^*)$,

$$\mathsf{u}(t', \boldsymbol{w}^{\text{av}}_{t'+1}, \boldsymbol{D}^{\text{av}}_{t'+1}) \leq \tfrac{2}{\eta'}\big(\|\boldsymbol{w}_1\|_2^2 + B_{\boldsymbol{w}}^2 + \|\boldsymbol{D}_1\|_2^2 + B_{\boldsymbol{D}}^2\big) + 6(H_{\boldsymbol{w}}^2 + H_{\boldsymbol{D}}^2)\eta' t' + 2(2+\sigma\overline{\Lambda})H_{\boldsymbol{D}} \sum_{s=1}^{t'} \|\mathbf{L}_{\mathcal{K}} \boldsymbol{D}_s\|_2 + 2\|\mathbf{L}_{\mathcal{K}} \boldsymbol{D}^{\text{av}}_{t'+1}\|_2 H_{\boldsymbol{D}} t' \,. \quad (30)$$

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 $\mathsf{u}(t', \boldsymbol{w}^{\text{av}}_{t'+1}, \boldsymbol{D}^{\text{av}}_{t'+1})$, we note that, by the triangular inequality, we have

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

That is, we get

$$\begin{aligned} &\mathsf{u}(t', \boldsymbol{w}^*, \boldsymbol{D}^*) \leq \mathsf{u}(t', \boldsymbol{w}^{\text{av}}_{t'+1}, \boldsymbol{D}^{\text{av}}_{t'+1}) \\ &\leq \tfrac{2}{\eta'}\big(\|\boldsymbol{w}_1\|_2^2 + B_{\boldsymbol{w}}^2 + \|\boldsymbol{D}_1\|_2^2 + B_{\boldsymbol{D}}^2\big) + 6(H_{\boldsymbol{w}}^2 + H_{\boldsymbol{D}}^2)\eta' t' \\ &\quad + 2H_{\boldsymbol{D}}(3+\sigma\overline{\Lambda}) \sum_{s=1}^{t'} \|\mathbf{L}_{\mathcal{K}} \boldsymbol{D}_s\|_2. \end{aligned} \quad (31)$$

We now further bound $\sum_{s=1}^{t'} \|\mathbf{L}_{\mathcal{K}} \boldsymbol{D}_s\|_2$ in (27) noting that $\|\boldsymbol{u}_t^1\|_2 = \|\eta_t g_{D_t}\|_2 \leq \eta_t H_{\boldsymbol{D}} = \eta' H_{\boldsymbol{D}}$, 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 (31), taking $\eta' = \frac{1}{\sqrt{t'}}$ and noting that $1 \leq \sqrt{t'}$, we get

$$\mathsf{u}(t', \boldsymbol{w}^{\text{av}}_{t'+1}, \boldsymbol{D}^{\text{av}}_{t'+1}) \leq \alpha' \sqrt{t'}, \quad (32)$$

where

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

This bound is of the type $\mathsf{u}(t', \boldsymbol{w}^{\text{av}}_{t'+1}, \boldsymbol{D}^{\text{av}}_{t'+1}) \leq \alpha' \sqrt{t'}$, where α' depends on the initial conditions. This leads to step (ii). According to the Doubling Trick, for $m = 0, 1, \dots \lceil \log_2 t \rceil$, the dynamics is executed in each epoch of $t' = 2^m$ time steps $t = 2^m, \dots, 2^{m+1} - 1$, where at the beginning of each epoch the initial conditions are the final values in the previous epoch. The bound for $\mathsf{u}(t', \boldsymbol{w}^{\text{av}}_{t'+1}, \boldsymbol{D}^{\text{av}}_{t'+1})$ in each epoch is $\alpha' \sqrt{t'} = \alpha_m \sqrt{2^m}$, where α_m is the multiplicative constant in (32) 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,

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

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 (19) and dividing by $2t$. ■

In the statement of Theorem IV.6, the constant C_u appearing in (29) 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 IV.7. (Boundedness of estimates): The statement of Theorem IV.6 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 V-A. •

V. Applications to distributed constrained convex optimization

In this section we particularize our convergence result in Theorem IV.6 to the case of convex-concave functions arising from the Lagrangian of the constrained optimization (8) discussed in Section III-A. The Lagrangian formulation with explicit agreement constraints (9c) matches the general saddle-point problem (7) 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

$$\boldsymbol{\phi}(\boldsymbol{w}, \boldsymbol{D}, \boldsymbol{z}) = \sum_{i=1}^{N} \big(f^i(w^i, D^i) + {z^i}^\top g^i(w^i, D^i)\big). \quad (33)$$

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 (8) *only* under the agreement on the global decision vector and on the Lagrange multiplier associated to the global constraint, i.e.,

$$\mathcal{L}(\boldsymbol{w}, D, z) = \boldsymbol{\phi}(\boldsymbol{w}, D \otimes \mathbb{1}_N, z \otimes \mathbb{1}_N). \quad (34)$$

In this case, the *saddle-point dynamics with Laplacian averaging* (12) takes the following form: the updates of each agent

$i \in \{1,\dots,N\}$ are as follows,

$$\hat{w}^i_{t+1} = w^i_t - \eta_t(d_{f^i,w^i_t} + d^\top_{g^i,w^i_t} z^i), \tag{35a}$$

$$\hat{D}^i_{t+1} = D^i_t + \sigma \sum_{j=1}^N \mathsf{a}_{ij,t}(D^j_t - D^i_t) - \eta_t(d_{f^i,D^i_t} + d^\top_{g^i,D^i_t} z^i), \tag{35b}$$

$$\hat{z}^i_{t+1} = z^i_t + \sigma \sum_{j=1}^N \mathsf{a}_{ij,t}(z^j_t - z^i_t) + \eta_t g^i(w^i_t), \tag{35c}$$

$$\begin{bmatrix} w^i_{t+1} \\ D^i_{t+1} \\ z^i_{t+1} \end{bmatrix} = \begin{bmatrix} \mathcal{P}_{\mathcal{W}_i}(\hat{w}^i_{t+1}) \\ \mathcal{P}_{\mathcal{D}}(\hat{D}^i_{t+1}) \\ \mathcal{P}_{\mathbb{R}^m_{\geq 0}\cap\bar{\mathcal{B}}(0,r)}(\hat{z}^i_{t+1}) \end{bmatrix}, \tag{35d}$$

where the vectors $d_{f^i,w^i_t} \in \mathbb{R}^{n_i}$ and $d_{f^i,D^i_t} \in \mathbb{R}^{d_D}$ 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.,

$$d_{f^i,w^i_t} \in \partial_{w^i} f^i(w^i_t, D^i_t), \qquad d_{f^i,D^i_t} \in \partial_D f^i(w^i_t, D^i_t), \tag{36}$$

and the matrices $d_{g^i,w^i} \in \mathbb{R}^{m\times n_i}$ and $d_{g^i,D} \in \mathbb{R}^{m\times d_D}$ 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 $g^i : \mathbb{R}^{n_i} \times \mathbb{R}^{d_D} \to \mathbb{R}^m$.) 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 (35d) 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 V-A.

Algorithm 1: C-SP-SG algorithm

Data: Agents' data for Problem (8): $\{f^i, g^i, \mathcal{W}_i\}_{i=1}^N$, $\mathcal{D}$
Agents' adjacency values $\{\mathsf{A}_t\}_{t\geq 1}$
Consensus stepsize σ as in (24)
Learning rates $\{\eta_t\}_{t\geq 1}$ as in Assumption IV.5
Radius r s.t. $\bar{\mathcal{B}}(0,r)$ contains optimal dual set for (8)
Number of iterations T, indep. of rest of parameters
Result: Agent i outputs $(w^i_T)^{\text{av}}, (D^i_T)^{\text{av}}, (z^i_T)^{\text{av}}$
Initialization: Agent i sets $w^i_1 \in \mathbb{R}^{n_i}$, $D^i_1 \in \mathbb{R}^{d_D}$, $z^i_1 \in \mathbb{R}^m_{\geq 0}$, $(w^i_1)^{\text{av}} = w^i_1$, $(D^i_1)^{\text{av}} = D^i_1$, $(z^i_1)^{\text{av}} = z^i_1$
for $t \in \{2,\dots,T-1\}$ **do**
  for $i \in \{1,\dots,N\}$ **do**
    Agent i selects (sub-) gradients as in (36)
    Agent i updates (w^i_t, D^i_t, z^i_t) as in (35)
    Agent i updates $(w^i_{t+1})^{\text{av}} = \frac{t-1}{t}(w^i_t)^{\text{av}} + \frac{1}{t} w^i_t$, $(D^i_{t+1})^{\text{av}} = \frac{t-1}{t}(D^i_t)^{\text{av}} + \frac{1}{t} D^i_t$, $(z^i_{t+1})^{\text{av}} = \frac{t-1}{t}(z^i_t)^{\text{av}} + \frac{1}{t} z^i_t$
  end
end

The characterization of the saddle-point evaluation error under (35) is a direct consequence of Theorem IV.6.

Corollary V.1. (Convergence of the C-SP-SG algorithm): *For each $i \in \{1,\dots,N\}$, let the sequence $\{(w^i_t, D^i_t, z^i_t)\}_{t\geq 1}$ be generated by the coordination algorithm* (35)*, over a sequence of graphs $\{\mathcal{G}_t\}_{t\geq 1}$ satisfying the same hypotheses as Theorem IV.6. 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* (8)*. Assume also that the subgradient sets are bounded, in $\mathcal{W}_i \times \mathcal{D}$, as follows,*

$$\partial_{w^i} f^i \subseteq \bar{\mathcal{B}}(0, H_{f,w}),\ \partial_D f^i \subseteq \bar{\mathcal{B}}(0, H_{f,D}),$$
$$\partial_{w^i} g^i_l \subseteq \bar{\mathcal{B}}(0, H_{g,w}),\ \partial_D g^i_l \subseteq \bar{\mathcal{B}}(0, H_{g,D}),$$

for all $l \in \{1,\dots,m\}$. Let $(\boldsymbol{w}^*, D^*, z^*)$ be any saddle point of the Lagrangian $\mathcal{L}$ defined in (34) *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 IV.5 for the learning rates, the saddle-point evaluation error* (28) *holds for the running time-averages:*

$$-\frac{\alpha_{\boldsymbol{\mu},\boldsymbol{z}} + \alpha_{\boldsymbol{w},\boldsymbol{D}}}{2\sqrt{t-1}} \leq \boldsymbol{\phi}(\boldsymbol{w}^{av}_t, \boldsymbol{D}^{av}_t, \boldsymbol{z}^{av}_t) - \mathcal{L}(\boldsymbol{w}^*, D^*, z^*) \leq \frac{\alpha_{\boldsymbol{w},\boldsymbol{D}} + \alpha_{\boldsymbol{\mu},\boldsymbol{z}}}{2\sqrt{t-1}}, \tag{37}$$

for $\alpha_{\boldsymbol{w},\boldsymbol{D}}$ and $\alpha_{\boldsymbol{\mu},\boldsymbol{z}}$ as in (29)*, with*

$$B_{\boldsymbol{\mu}} = H_{\boldsymbol{\mu}} = 0, \qquad B_{\boldsymbol{z}} = \sqrt{N} r,$$
$$B_{\boldsymbol{w}} = (\sum_{i=1}^N \operatorname{diam}(\mathcal{W}_i)^2)^{1/2}, \qquad B_{\boldsymbol{D}} = \sqrt{N}\operatorname{diam}(\mathcal{D}),$$
$$H^2_{\boldsymbol{w}} = N(H_{f,w} + r\sqrt{m} H_{g,w})^2, \quad H^2_{\boldsymbol{z}} = \sum_{i=1}^N (\sup_{w^i \in \mathcal{W}_i} g_i(w^i))^2,$$
$$H^2_{\boldsymbol{D}} = N(H_{f,D} + r\sqrt{m} H_{g,D})^2,$$

where $\operatorname{diam}(\cdot)$ *refers to the diameter of the sets.*

The proof of this result follows by noting that the hypotheses of Theorem IV.6 are automatically satisfied. The only point to observe is that all the saddle points of the Lagrangian $\mathcal{L}$ defined in (34) 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 V.2. (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 V.1, the saddle-point evaluation error is smaller than ϵ if $\frac{\alpha_{\boldsymbol{w},\boldsymbol{D}} + \alpha_{\boldsymbol{\mu},\boldsymbol{z}}}{2\sqrt{t}} \leq \epsilon$. This provides a lower bound
$$t \geq \left(\frac{\alpha_{\boldsymbol{w},\boldsymbol{D}} + \alpha_{\boldsymbol{\mu},\boldsymbol{z}}}{2\epsilon}\right)^2,$$
on the number of required iterations.
- **Memory complexity**: Each agent i maintains the current updates $(w^i_t, D^i_t, z^i_t) \in \mathbb{R}^{n_i} \times \mathbb{R}^{d_D} \times \mathbb{R}^m$,

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 $\partial_{w^i} f^i \subseteq \mathbb{R}^{n_i}$, $\partial_D f^i \subseteq \mathbb{R}^{d_D}$, $\partial_{w^i} g_l^i \subseteq \mathbb{R}^{n_i}$, $\partial_D g_l^i \subseteq \mathbb{R}^{d_D}$, the latter for $l \in \{1, \dots, m\}$. Each agent also projects its estimates on the set $\mathcal{W}_i \times \mathcal{D} \times \mathbb{R}^m_{\geq 0} \cap \bar{\mathcal{B}}(0, r)$. 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 $\mathbb{R}^{d_D} \times \mathbb{R}^m$. With the information received, the agent updates the global decision variable D_t^i in (35b) and the Lagrange multiplier z_t^i in (35c). (Note that the variable D_t^i needs to be maintained and communicated only if the optimization problem (8) has a global decision variable.) •

A. 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 (35d) is the following. The subgradients of ϕ with respect to the primal variables are *linear* in the dual variables. To guarantee the boundedness of the subgradients of ϕ and of the dual variables, required by the application of Theorem IV.6, one can introduce a projection step onto a compact set that preserves the optimal dual set, a technique that has been used in [3], [19], [22]. These works select the bound for the projection *a priori*, whereas [9] 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 (8). 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., $f^i(w^i, D) = f^i(w^i)$ and $g^i(w^i, D) = g^i(w^i)$ 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,

$$\begin{aligned} \min_{w^i \in \mathcal{W}_i,\, \forall i} \; & \sum_{i=1}^N f^i(w^i) \\ \text{s.t. } & g^1(w^1) + \cdots + g^N(w^N) \leq 0 \end{aligned} \tag{38}$$

where each $\mathcal{W}_i$ is compact as in Corollary V.1. 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* [20, Sec. 7.2.3], called *Slater vector*, and define

$$\gamma := \min_{l \in \{1, \dots, m\}} -\sum_{i=1}^N g_l^i(\tilde{w}^i), \tag{39}$$

which is positive by construction. According to [19, 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,

$$\max_{z^* \in \mathcal{Z}^*} \|z^*\|_2 \leq \frac{1}{\gamma} \Big(\sum_{i=1}^N f^i(\tilde{w}^i) - q(\bar{z}) \Big), \tag{40}$$

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 (38),

$$\begin{aligned} q(z) &= \inf_{w^i \in \mathcal{W}_i,\, \forall i} \; \mathcal{L}(\boldsymbol{w}, z) \\ &= \inf_{w^i \in \mathcal{W}_i,\, \forall i} \; \sum_{i=1}^N \big(f^i(w^i) + z^\top g^i(w^i) \big) =: \sum_{i=1}^N q^i(z). \end{aligned} \tag{41}$$

Note that the right hand side in (40) 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 (40), we conclude that $\mathcal{Z}^* \subseteq \mathcal{Z}_c$, with

$$\mathcal{Z}_c := \mathbb{R}^m_{\geq 0} \cap \bar{\mathcal{B}} \Big(0, \frac{N}{\gamma} \Big(\max_{j \in \{1, \dots, N\}} f^j(\tilde{w}^j) - \min_{j \in \{1, \dots, N\}} q^j(\bar{z}) \Big) \Big). \tag{42}$$

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 V.1), agent i can compute

$$\tilde{w}^i \in \arg\min\nolimits_{w^i \in \mathcal{W}_i} g_l^i(w^i).$$

The resulting vector $(\tilde{w}^1, \dots, \tilde{w}^N)$ is a Slater vector, i.e., it belongs to the set

$$\begin{aligned} \{(w^1, \dots, w^N) \in \mathcal{W}_1 \times \cdots \times \mathcal{W}_N \; : \\ g^1(w^1) + \cdots + g^N(w^N) < 0\}, \end{aligned}$$

which is nonempty by the Strong Slater condition.

(i.b) Similarly, the agents compute the corresponding component $q^i(\bar{z})$ defined in (41). 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 (39) 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

$$y^i(k+1) = y^i(k) + \sigma \sum_{j=1}^{N} \mathsf{a}_{ij,t}(y^j(k) - y^i(k)), \tag{43a}$$

$$s_i(k+1) = s_i(k) + \sigma \sum_{j=1}^{N} \mathsf{a}_{ij,t}\big(\operatorname{sign}(y^j(k)) - \operatorname{sign}(y^i(k))\big), \tag{43b}$$

until an iteration k_i^* such that $N s_i(k_i^*) \leq -(N-1)$; see Lemma V.3 below for the justification of this termination condition. Then, agent i re-initializes $y^i(0) = y^i(k^*)$ and iterates

$$y^i(k+1) = \min\{y^j(k) \,:\, j \in \mathcal{N}^{\text{out}}(i) \cup \{i\}\} \tag{44}$$

(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

$$\hat{y} := N y^i(k^{**}) \leq \sum_{i=1}^{N} g^i(\tilde{w}^i),$$
$$\underline{\gamma} := \min_{l\in\{1,\dots,m\}} -\hat{y}_l \leq \gamma,$$

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 (44).

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 (42), bounds the optimal dual set for the constrained optimization (38),

$$\mathcal{Z}^* \subseteq \mathcal{Z}_c \subseteq \bar{\mathcal{B}}(0, r).$$

To conclude, we justify the termination condition of step (ii).

Lemma V.3. (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 IV.6, 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 (43a) is

$$\frac{1}{N}\sum_{i=1}^{N} y^i(0) = \frac{1}{N}\sum_{i=1}^{N} g^i(\tilde{w}^i) < 0. \tag{45}$$

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 (43a) to the point (45), it follows that there exists a finite time $k_i^* \in \mathbb{Z}_{\geq 1}$ such that $N s_i(k_i^*) \leq -(N-1)$. 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.

VI. 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 [41] with maximum degree $d_{\max} = 4$. Under this coordination strategy, the 50 agents solve collaboratively the following instance of problem (8) with nonlinear convex constraints:

$$\min_{w_i\in[0,1]} \sum_{i=1}^{50} c_i w_i$$
$$\text{s.t.} \sum_{i=1}^{50} -d_i \log(1+w_i) \leq -b. \tag{46}$$

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 derive from (24) the theoretically feasible consensus stepsize $\sigma = 0.2475$. For the projection step in (35d) of the C-SP-SG algorithm, the bound on the optimal dual set (42), using the Slater vector $\tilde{w} = \mathbb{1}_N$ and $\bar{z} = 0$, is

$$r = \frac{N \max_{j\in\{1,\dots,N\}} c_j}{\log(2)\sum_{i=1}^{N} d_i - N/10} = 3.313.$$

For comparison, we have also simulated the Consensus-Based Dual Decomposition (CoBa-DD) algorithm proposed in [23] 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

[1]The Matlab code is available at https://github.com/DavidMateosNunez/Consensus-based-Saddle-Point-Subgradient-Algorithm.git.

[2]The Matlab code is available at http://ens.ewi.tudelft.nl/~asimonetto/NumericalExample.zip.

same as above.) We should note that the analysis in [23] 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 IV.5 (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. 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 V.1, 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 (26), 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. 3 compares the network cost-error and the constraint satisfaction. We can observe that the C-SP-SG and the CoBa-DD [23] algorithms have some characteristics in common:

- They both benefit from using the Doubling Trick scheme.
- They approximate the solution, in all metrics of Fig. 2 and Fig. 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 [23] (wherein the theoretical bound concerns the practical convergence of the cost error using constant learning rates).

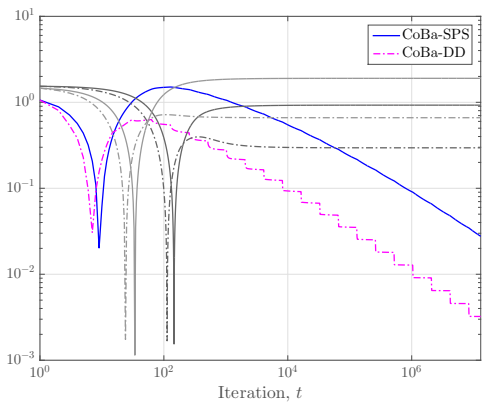
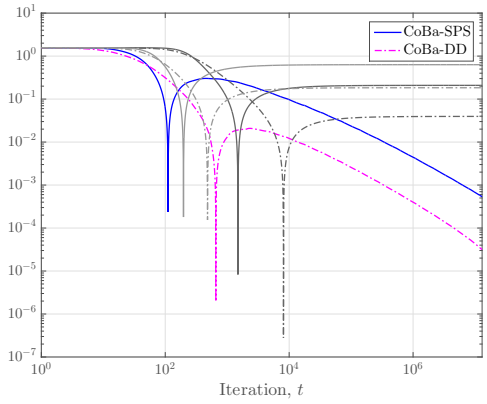
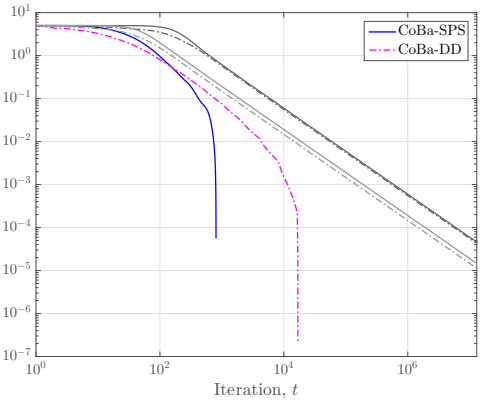


Fig. 2: Saddle-point evaluation error $|\phi(\boldsymbol{w}_t^{\text{av}}, \boldsymbol{z}_t^{\text{av}}) - \mathcal{L}(\boldsymbol{w}^*, z^*)|$. The lines in grey represent the same algorithms simulated with constant learning rates equal to 0.2 (lighter grey) and 0.05 (darker grey), respectively.



(a) Cost error



(b) Constraint satisfaction

Fig. 3: Cost error and constraint satisfaction. For the same instantiations as in Fig. 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 $-\sum_{i=1}^N d_i \log(1+(w_t^i)^{\text{av}}) + b$.

VII. Conclusions and Ideas for Future Work

We have studied 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 distributed both the objective function (via agreement on a subset of the primal variables) and the constraints (via agreement on the dual variables). This approach enables the use of existing consensus-based ideas to tackle the algorithmic solution to these problems in a distributed fashion. Future extensions will include, first, a refined analysis of convergence for constrained optimization in terms of the cost evaluation error instead of the saddle-point evaluation error. Second, more general distributed algorithms for computing bounds on Lagrange vectors and matrices, which are required in the design of truncated projections preserving the optimal dual sets. (An alternative route would explore the characterization of the intrinsic boundedness properties of the proposed distributed dynamics.) Third, the selection of other learning rates that improve the convergence rate of our proposed algorithms.

Finally, we envision applications to semidefinite programming, where chordal sparsity allows to tackle problems that have the dimension of the matrices grow with the network size, and also the treatment of low-rank conditions. Particular applications will include efficient optimization in wireless networks, control of camera networks, and estimation and control in smart grids.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their useful feedback that helped us improve the presentation of the paper. This research was partially supported by NSF Award CMMI-1300272 and Award FA9550-15-1-0108.

REFERENCES

[1] D. Mateos-Núñez and J. Cortés, "Distributed subgradient methods for saddle-point problems," in *IEEE Conf. on Decision and Control*, (Osaka, Japan), pp. 5462–5467, 2015.

[2] K. Arrow, L. Hurwitz, and H. Uzawa, *Studies in Linear and Non-Linear Programming*. Stanford, California: Stanford University Press, 1958.

[3] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *Journal of Optimization Theory & Applications*, vol. 142, no. 1, pp. 205–228, 2009.

[4] N. Parikh and S. Boyd, "Proximal algorithms," vol. 1, no. 3, pp. 123–231, 2013.

[5] 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*, vol. 3, no. 1, pp. 1–122, 2011.

[6] 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*, (Cancun, Mexico), pp. 4185–4190, 2008.

[7] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[8] J. Wang and N. Elia, "A control perspective for centralized and distributed convex optimization," in *IEEE Conf. on Decision and Control*, (Orlando, Florida), pp. 3800–3805, 2011.

[9] M. Zhu and S. Martínez, "On distributed convex optimization under inequality and equality constraints," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 151–164, 2012.

[10] R. Carli, G. Notarstefano, L. Schenato, and D. Varagnolo, "Distributed quadratic programming under asynchronous and lossy communications via Newton-Raphson consensus," in *European Control Conference*, (Lind, Austria), pp. 2514–2520, 2015.

[11] B. Gharesifard and J. Cortés, "Distributed continuous-time convex optimization on weight-balanced digraphs," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 781–786, 2014.

[12] J. N. Tsitsiklis, *Problems in Decentralized Decision Making and Computation*. PhD thesis, Massachusetts Institute of Technology, Nov. 1984. Available at http://web.mit.edu/jnt/www/Papers/PhD-84-jnt.pdf.

[13] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803–812, 1986.

[14] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific, 1997.

[15] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Belmont, MA: Athena Scientific, 1st ed., 2003.

[16] G. Scutari, D. P. Palomar, F. Facchinei, and J. S. Pang, "Convex optimization, game theory, and variational inequality theory," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 35–49, 2010.

[17] S. Dafermos, "Traffic equilibrium and variational inequalities," *Transportation Science*, vol. 14, no. 1, pp. 42–54, 1980.

[18] A. Nagurney, M. Yu, A. H. Masoumi, and L. S. Nagurney, *Networks Against Time: Supply Chain Analytics for Perishable Products*. SpringerBriefs in Optimization, Springer, 2010.

[19] A. Nedić and A. Ozdaglar, "Approximate primal solutions and rate analysis for dual subgradient methods," *SIAM J. Optimization*, vol. 19, no. 4, pp. 1757–1780, 2010.

[20] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms I*. Grundlehren Text Editions, New York: Springer, 1993.

[21] M. Bürger, G. Notarstefano, and F. Allgöwer, "A polyhedral approximation framework for convex and robust distributed optimization," *IEEE Transactions on Automatic Control*, vol. 59, no. 2, pp. 384–395, 2014.

[22] T.-H. Chang, A. Nedić, and A. Scaglione, "Distributed constrained optimization by consensus-based primal-dual perturbation method," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1524–1538, 2014.

[23] A. Simonetto and H. Jamali-Rad, "Primal recovery from consensus-based dual decomposition for distributed convex optimization," *Journal of Optimization Theory and Applications*, vol. 168, no. 1, pp. 172–197, 2016.

[24] 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*, vol. 41, no. 6, pp. 1715–1724, 2011.

[25] D. Yuan, D. W. C. Ho, and S. Xu, "Regularized primal-dual subgradient method for distributed constrained optimization," *IEEE Transactions on Cybernetics*, vol. 46, no. 9, pp. 2109–2118, 2016.

[26] A. Nedic, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, 2010.

[27] 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*, (Atlanta, GA), pp. 1366–1371, Dec. 2010.

[28] D. Mosk-Aoyama, T. Roughgarden, and D. Shah, "Fully distributed algorithms for convex optimization problems," *SIAM J. Optimization*, vol. 20, no. 6, pp. 3260–3279, 2010.

[29] G. Notarstefano and F. Bullo, "Distributed abstract optimization via constraints consensus: Theory and applications," *IEEE Transactions on Automatic Control*, vol. 56, no. 10, pp. 2247–2261, 2011.

[30] D. Richert and J. Cortés, "Robust distributed linear programming," *IEEE Transactions on Automatic Control*, vol. 60, no. 10, pp. 2567–2582, 2015.

[31] A. A. Morye, C. Ding, A. K. Roy-Chowdhury, and J. A. Farrell, "Distributed constrained optimization for Bayesian opportunistic visual sensing," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 6, pp. 2302–2318, 2014.

[32] A. Kalbat and J. Lavaei, "A fast distributed algorithm for decomposable semidefinite programs," *Mathematical Programming Computation*, 2015. Submitted.

[33] D. Mateos-Núñez and J. Cortés, "Distributed optimization for multi-task learning via nuclear-norm approximation," *IFAC-PapersOnLine*, vol. 48, no. 22, pp. 64–69, 2015. *IFAC Workshop on Distributed Estimation and Control in Networked Systems*, Philadelphia, PA.

[34] 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.

[35] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2009.

[36] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 2nd ed., 1999.

[37] A. Nedić and A. Ozdaglar, "Cooperative distributed multi-agent optimization," in *Convex Optimization in Signal Processing and Communications* (Y. Eldar and D. Palomar, eds.), Cambridge University Press, 2010.

[38] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, vol. 317 of *Comprehensive Studies in Mathematics*. New York: Springer, 1998.

[39] S. Shalev-Shwartz, *Online Learning and Online Convex Optimization*, vol. 12 of *Foundations and Trends in Machine Learning*. Now Publishers Inc, 2012.

[40] D. Mateos-Núñez and J. Cortés, "Distributed online convex optimization over jointly connected digraphs," *IEEE Transactions on Network Science and Engineering*, vol. 1, no. 1, pp. 23–37, 2014.

[41] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[42] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

APPENDIX

Here we present the proofs of the results Lemma IV.1 and Proposition IV.4 stated in Section IV.

Proof of Lemma IV.1: In this proof we extend the saddle-point analysis for the (centralized) subgradient methods

in [3, Lemma 3.1] by incorporating the treatment on the disagreement from our previous work in [40, Lemma V.2]. We first define

$$\begin{bmatrix}\mathbf{r}_{\boldsymbol{w},t+1}\\ \mathbf{r}_{\boldsymbol{D},t+1}\end{bmatrix} := \begin{bmatrix}\boldsymbol{w}_{t+1}-\hat{\boldsymbol{w}}_{t+1}\\ \boldsymbol{D}_{t+1}-\hat{\boldsymbol{D}}_{t+1}\end{bmatrix}. \tag{47}$$

Since $\mathrm{I}_d-\sigma\mathsf{L}_t$ is a stochastic matrix (because σ satisfies (14)), then its product by any vector is a convex combination of the entries of the vector. Hence, the fact that $\boldsymbol{D}_t\in\mathcal{D}^N$ implies that $\boldsymbol{D}_t-\sigma\mathbf{L}_t\boldsymbol{D}_t\in\mathcal{D}^N$. Using this together with the definition of orthogonal projection (1), we get

$$\begin{aligned}\|\mathbf{r}_{\boldsymbol{D},t+1}\|_2 &= \|\mathcal{P}_{\mathcal{D}^N}\big(\hat{\boldsymbol{D}}_{t+1}\big)-\hat{\boldsymbol{D}}_{t+1}\|_2\\ &\le \|(\boldsymbol{D}_t-\sigma\mathbf{L}_t\boldsymbol{D}_t)-\hat{\boldsymbol{D}}_{t+1}\|_2 = \eta_t\|g_{\boldsymbol{D}_t}\|_2.\end{aligned} \tag{48}$$

Similarly, since $\boldsymbol{w}_t\in\boldsymbol{\mathcal{W}}$, we also have

$$\begin{aligned}\|\mathbf{r}_{\boldsymbol{w},t+1}\|_2 &= \|\mathcal{P}_{\boldsymbol{\mathcal{W}}}\big(\hat{\boldsymbol{w}}_{t+1}\big)-\hat{\boldsymbol{w}}_{t+1}\|_2\\ &\le \|\boldsymbol{w}_t-\hat{\boldsymbol{w}}_{t+1}\|_2 = \eta_t\|g_{\boldsymbol{w}_t}\|_2.\end{aligned}$$

Left-multiplying the dynamics of $\boldsymbol{w}_t$ and $\boldsymbol{D}_t$ from (12a) and (12b) (in terms of the residual (47)) by the block-diagonal matrix $\mathrm{diag}(\mathrm{I}_{Nd},\mathbf{M})$, and using $\mathbf{M}\mathbf{L}_t=0$, we obtain

$$\begin{bmatrix}\boldsymbol{w}_{t+1}\\ \mathbf{M}\boldsymbol{D}_{t+1}\end{bmatrix} = \begin{bmatrix}\boldsymbol{w}_t\\ \mathbf{M}\boldsymbol{D}_t\end{bmatrix} + \begin{bmatrix}-\eta_t g_{\boldsymbol{w}_t}+\mathbf{r}_{\boldsymbol{w},t+1}\\ -\eta_t\mathbf{M}g_{\boldsymbol{D}_t}+\mathbf{M}\mathbf{r}_{\boldsymbol{D},t+1}\end{bmatrix}. \tag{49}$$

Subtracting $(\boldsymbol{w}_p,\boldsymbol{D}_p)\in\boldsymbol{\mathcal{W}}\times\mathcal{D}^N$ on each side, taking the norm, and noting that $\mathbf{M}^\top=\mathbf{M}$ and $\mathbf{M}^2=\mathbf{M}$, we get

$$\begin{aligned}&\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_p\|_2^2+\|\mathbf{M}\boldsymbol{D}_{t+1}-\boldsymbol{D}_p\|_2^2\\ &=\|\boldsymbol{w}_t-\boldsymbol{w}_p\|_2^2+\|\mathbf{M}\boldsymbol{D}_t-\boldsymbol{D}_p\|_2^2\\ &\quad+\|-\eta_t g_{\boldsymbol{w}_t}+\mathbf{r}_{\boldsymbol{w},t+1}\|_2^2+\|-\eta_t\mathbf{M}g_{\boldsymbol{D}_t}+\mathbf{M}\mathbf{r}_{\boldsymbol{D},t+1}\|_2^2\\ &\quad-2\eta_t g_{\boldsymbol{w}_t}^\top(\boldsymbol{w}_t-\boldsymbol{w}_p)-2\eta_t g_{\boldsymbol{D}_t}^\top(\mathbf{M}\boldsymbol{D}_t-\mathbf{M}\boldsymbol{D}_p)\\ &\quad+2\mathbf{r}_{\boldsymbol{w},t+1}^\top(\boldsymbol{w}_t-\boldsymbol{w}_p)+2\mathbf{r}_{\boldsymbol{D},t+1}^\top(\mathbf{M}\boldsymbol{D}_t-\mathbf{M}\boldsymbol{D}_p).\end{aligned} \tag{50}$$

We can bound the term $-g_{\boldsymbol{D}_t}^\top(\mathbf{M}\boldsymbol{D}_t-\mathbf{M}\boldsymbol{D}_p)$ by subtracting and adding $\boldsymbol{D}_t-\boldsymbol{D}_p$ inside the bracket and using convexity,

$$\begin{aligned}&-g_{\boldsymbol{w}_t}^\top(\boldsymbol{w}_t-\boldsymbol{w}_p)-g_{\boldsymbol{D}_t}^\top(\mathbf{M}\boldsymbol{D}_t-\mathbf{M}\boldsymbol{D}_p)\\ &=-g_{\boldsymbol{D}_t}^\top(\mathbf{M}\boldsymbol{D}_t-\boldsymbol{D}_t)-g_{\boldsymbol{D}_t}^\top(\boldsymbol{D}_p-\mathbf{M}\boldsymbol{D}_p)\\ &\quad-\begin{bmatrix}g_{\boldsymbol{w}_t}^\top & g_{\boldsymbol{D}_t}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{w}_t-\boldsymbol{w}_p\\ \boldsymbol{D}_t-\boldsymbol{D}_p\end{bmatrix}\\ &\le g_{\boldsymbol{D}_t}^\top\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_t-g_{\boldsymbol{D}_t}^\top\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_p\\ &\quad+\boldsymbol{\phi}(\boldsymbol{w}_p,\boldsymbol{D}_p,\boldsymbol{\mu}_t,\boldsymbol{z}_t)-\boldsymbol{\phi}(\boldsymbol{w}_t,\boldsymbol{D}_t,\boldsymbol{\mu}_t,\boldsymbol{z}_t),\end{aligned} \tag{51}$$

where we have used $\mathbf{L}_{\mathcal{K}}=\mathrm{I}_{Nd_{\mathcal{D}}}-\mathbf{M}$ and the fact that $g_{\boldsymbol{w}_t}\in\partial_{\boldsymbol{w}}\boldsymbol{\phi}(\boldsymbol{w}_t,\boldsymbol{D}_t,\boldsymbol{\mu}_t,\boldsymbol{z}_t)$ and $g_{\boldsymbol{D}_t}\in\partial_{\boldsymbol{D}}\boldsymbol{\phi}(\boldsymbol{w}_t,\boldsymbol{D}_t,\boldsymbol{\mu}_t,\boldsymbol{z}_t)$. Using this bound and (50), we get

$$\begin{aligned}&2(\boldsymbol{\phi}(\boldsymbol{w}_t,\boldsymbol{D}_t,\boldsymbol{\mu}_t,\boldsymbol{z}_t)-\boldsymbol{\phi}(\boldsymbol{w}_p,\boldsymbol{D}_p,\boldsymbol{\mu}_t,\boldsymbol{z}_t))\\ &\le\tfrac{1}{\eta_t}\big(\|\boldsymbol{w}_t-\boldsymbol{w}_p\|_2^2-\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_p\|_2^2\big)\\ &\quad+\tfrac{1}{\eta_t}\big(\|\mathbf{M}\boldsymbol{D}_t-\boldsymbol{D}_p\|_2^2-\|\mathbf{M}\boldsymbol{D}_{t+1}-\boldsymbol{D}_p\|_2^2\big)\\ &\quad+2g_{\boldsymbol{D}_t}^\top\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_t-2g_{\boldsymbol{D}_t}^\top\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_p\\ &\quad+\tfrac{1}{\eta_t}\|-\eta_t g_{\boldsymbol{w}_t}+\mathbf{r}_{\boldsymbol{w},t+1}\|_2^2+\tfrac{1}{\eta_t}\|-\eta_t\mathbf{M}g_{\boldsymbol{D}_t}+\mathbf{M}\mathbf{r}_{\boldsymbol{D},t+1}\|_2^2\\ &\quad+\tfrac{2}{\eta_t}\mathbf{r}_{\boldsymbol{w},t+1}^\top(\boldsymbol{w}_t-\boldsymbol{w}_p)+\tfrac{2}{\eta_t}\mathbf{r}_{\boldsymbol{D},t+1}^\top(\mathbf{M}\boldsymbol{D}_t-\boldsymbol{D}_p).\end{aligned} \tag{52}$$

We now bound each of the terms in the last three lines of (52). First, using the Cauchy-Schwarz inequality, we get

$$g_{\boldsymbol{D}_t}^\top\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_t-g_{\boldsymbol{D}_t}^\top\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_p\le\|g_{\boldsymbol{D}_t}\|_2(\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_t\|_2+\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_p\|_2). \tag{53}$$

For the terms in the second to last line, using the triangular inequality, the submultiplicativity of the norm, the fact that $\|\mathbf{M}\|_2\le 1$, and the bound (48), we have

$$\begin{aligned}&\|-\eta_t\mathbf{M}g_{\boldsymbol{D}_t}+\mathbf{M}\mathbf{r}_{\boldsymbol{D},t+1}\|_2\le\|-\eta_t\mathbf{M}g_{\boldsymbol{D}_t}\|_2+\|\mathbf{M}\mathbf{r}_{\boldsymbol{D},t+1}\|_2\\ &\le\eta_t\|\mathbf{M}\|_2\|g_{\boldsymbol{D}_t}\|_2+\|\mathbf{M}\|_2\|\mathbf{r}_{\boldsymbol{D},t+1}\|_2\le 2\eta_t\|g_{\boldsymbol{D}_t}\|_2,\end{aligned} \tag{54}$$

and, similarly,

$$\|-\eta_t g_{\boldsymbol{w}_t}+\mathbf{r}_{\boldsymbol{w},t+1}\|_2\le 2\eta_t\|g_{\boldsymbol{w}_t}\|_2.$$

Finally, regarding the term $\mathbf{r}_{\boldsymbol{D},t+1}^\top(\mathbf{M}\boldsymbol{D}_t-\boldsymbol{D}_p)$, we use the definition of $\mathbf{r}_{\boldsymbol{D},t+1}$ and also add and subtract $\hat{\boldsymbol{D}}_{t+1}$ inside the bracket. With the fact that $\mathbf{M}\boldsymbol{D}_p\in\mathcal{D}^N$ (because $\mathcal{D}$ is convex), we leverage the property (2) of the orthogonal projection to derive the first inequality. For the next two inequalities we use the Cauchy-Schwarz inequality, and then the bound in (48) for the residual, and also the definition of $\hat{\boldsymbol{D}}_{t+1}$, the fact that $\mathbf{M}\boldsymbol{D}_t-\boldsymbol{D}_t=-\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_t$, and the triangular inequality. Formally,

$$\begin{aligned}&\mathbf{r}_{\boldsymbol{D},t+1}^\top(\mathbf{M}\boldsymbol{D}_t-\mathbf{M}\boldsymbol{D}_p)=\mathbf{r}_{\boldsymbol{D},t+1}^\top(\mathbf{M}\boldsymbol{D}_t-\hat{\boldsymbol{D}}_{t+1})\\ &\quad+\big(\mathcal{P}_{\mathcal{D}^N}\big(\hat{\boldsymbol{D}}_{t+1}\big)-\hat{\boldsymbol{D}}_{t+1}\big)(\hat{\boldsymbol{D}}_{t+1}-\mathbf{M}\boldsymbol{D}_p)\\ &\le\mathbf{r}_{\boldsymbol{D},t+1}^\top(\mathbf{M}\boldsymbol{D}_t-\hat{\boldsymbol{D}}_{t+1})\le\|\mathbf{r}_{\boldsymbol{D},t+1}\|_2\|\mathbf{M}\boldsymbol{D}_t-\hat{\boldsymbol{D}}_{t+1}\|_2\\ &\le\eta_t\|g_{\boldsymbol{D}_t}\|_2\,\|-\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_t+\sigma\mathbf{L}_t\boldsymbol{D}_t+\eta_t g_{\boldsymbol{D}_t}\|_2\\ &\le\eta_t\|g_{\boldsymbol{D}_t}\|_2\Big((1+\sigma\overline{\Lambda})\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_t\|_2+\eta_t\|g_{\boldsymbol{D}_t}\|_2\Big),\end{aligned} \tag{55}$$

where in the last inequality we have also used a bound for the term $\|\mathbf{L}_t\boldsymbol{D}_t\|_2$ invoking $\overline{\Lambda}$ that we explain next. From the Courant-Fischer min-max Theorem [42] applied to the matrices $\mathsf{L}_t^\top\mathsf{L}_t$ and $\mathsf{L}_{\mathcal{K}}^2$ (which are symmetric with the same nullspace), we deduce that for any $x\in\mathbb{R}^N$,

$$\frac{x^\top\mathsf{L}_t^\top\mathsf{L}_t x}{\lambda_{\max}(\mathsf{L}_t^\top\mathsf{L}_t)}\le\frac{x^\top\mathsf{L}_{\mathcal{K}}^2 x}{\lambda_{n-1}(\mathsf{L}_{\mathcal{K}}^2)},$$

where $\lambda_{n-1}(\cdot)$ refers to the second smallest eigenvalue, which for the matrix $\mathsf{L}_{\mathcal{K}}^2=\mathsf{L}_{\mathcal{K}}$ is 1. (Note that all its eigenvalues are 1, except the smallest that is 0.) With the analogous inequality for Kronecker products with the identity I_d, the bound needed to conclude (55) is then

$$\begin{aligned}\|\mathbf{L}_t\boldsymbol{D}_t\|_2&=\sqrt{\boldsymbol{D}_t^\top\mathbf{L}_t^\top\mathbf{L}_t\boldsymbol{D}_t}\\ &\le\sqrt{\lambda_{\max}(\mathbf{L}_t^\top\mathbf{L}_t)\;\boldsymbol{D}_t^\top\mathbf{L}_{\mathcal{K}}^2\boldsymbol{D}_t}=\sigma_{\max}(\mathsf{L}_t)\;\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_t\|_2.\end{aligned}$$

Similarly to (55), now without the disagreement terms,

$$\begin{aligned}\mathbf{r}_{\boldsymbol{w},t+1}^\top(\boldsymbol{w}_t-\boldsymbol{w}_p)&=\mathbf{r}_{\boldsymbol{w},t+1}^\top(\boldsymbol{w}_t-\hat{\boldsymbol{w}}_{t+1})\\ &\quad+\big(\mathcal{P}_{\boldsymbol{\mathcal{W}}}\big(\hat{\boldsymbol{w}}_{t+1}\big)-\hat{\boldsymbol{w}}_{t+1}\big)(\hat{\boldsymbol{w}}_{t+1}-\boldsymbol{w}_p)\\ &\le\mathbf{r}_{\boldsymbol{w},t+1}^\top(\boldsymbol{w}_t-\hat{\boldsymbol{w}}_{t+1})\\ &\le\|\mathbf{r}_{\boldsymbol{w},t+1}\|_2\|\boldsymbol{w}_t-\hat{\boldsymbol{w}}_{t+1}\|_2\le\eta_t^2\|g_{\boldsymbol{w}_t}\|_2^2.\end{aligned}$$

Substituting the bounds (53), (54) and (55), and their counterparts for $\boldsymbol{w}_t$, in (52), we obtain

$$\begin{aligned} &2(\boldsymbol{\phi}(\boldsymbol{w}_t, \boldsymbol{D}_t, \boldsymbol{\mu}_t, \boldsymbol{z}_t) - \boldsymbol{\phi}(\boldsymbol{w}_p, \boldsymbol{D}_p, \boldsymbol{\mu}_t, \boldsymbol{z}_t)) \qquad (56)\\ &\leq \tfrac{1}{\eta_t}\big(\|\boldsymbol{w}_t - \boldsymbol{w}_p\|_2^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{w}_p\|_2^2\big)\\ &\quad + \tfrac{1}{\eta_t}\big(\|\mathbf{M}\boldsymbol{D}_t - \boldsymbol{D}_p\|_2^2 - \|\mathbf{M}\boldsymbol{D}_{t+1} - \boldsymbol{D}_p\|_2^2\big)\\ &\quad + 2\|g_{\boldsymbol{D}_t}\|_2(\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_t\|_2 + \|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_p\|_2) + 6\eta_t\|g_{\boldsymbol{w}_t}\|_2^2\\ &\quad + 4\eta_t\|g_{\boldsymbol{D}_t}\|_2^2 + 2\|g_{\boldsymbol{D}_t}\|_2\Big((1+\sigma\overline{\Lambda})\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_t\|_2 + \eta_t\|g_{\boldsymbol{D}_t}\|_2\Big) \end{aligned}$$

and (15) follows. The bound (16) can be derived similarly, requiring concavity of $\boldsymbol{\phi}$ in $(\boldsymbol{\mu}, \boldsymbol{z})$. ■

Proof of Proposition IV.4: Since both dynamics in (22) are structurally similar, we study the first one,

$$\boldsymbol{D}_{t+1} = \boldsymbol{D}_t - \sigma\mathbf{L}_t\boldsymbol{D}_t + \boldsymbol{u}_t^1 + \mathbf{r}_{\boldsymbol{D},t+1}, \qquad (57)$$

where $\mathbf{r}_{\boldsymbol{D},t+1}$ is as in (47) and satisfies (similarly to (48)) that

$$\begin{aligned} \|\mathbf{r}_{\boldsymbol{D},t+1}\|_2 &= \|\mathcal{P}_{\mathcal{D}^N}\big(\hat{\boldsymbol{D}}_{t+1}\big) - \hat{\boldsymbol{D}}_{t+1}\|_2\\ &\leq \|(\boldsymbol{D}_t - \sigma\mathbf{L}_t\boldsymbol{D}_t) - \hat{\boldsymbol{D}}_{t+1}\|_2 = \|\boldsymbol{u}_t^1\|_2. \end{aligned}$$

The dynamics (57) coincides with that of [40, eqn. (29)] where, in the notation of the reference, one sets $e_t := \boldsymbol{u}_t^1 + \mathbf{r}_{\boldsymbol{D},t+1}$. Therefore, we obtain a bound analogous to [40, eqn. (34)],

$$\|\mathbf{L}_{\mathcal{K}}\boldsymbol{D}_t\|_2 \leq \rho_{\tilde{\delta}}^{\lceil\frac{t-1}{B}\rceil-2}\|\boldsymbol{D}_1\|_2 + \sum_{s=1}^{t-1}\rho_{\tilde{\delta}}^{\lceil\frac{t-1-s}{B}\rceil-2}\|e_s\|_2\,, \quad (58)$$

where

$$\rho_{\tilde{\delta}} := 1 - \frac{\tilde{\delta}}{4N^2}.$$

To derive (25) we use three facts: first $\|e_t\|_2 \leq \|\boldsymbol{u}_t^1\|_2 + \|\mathbf{r}_{\boldsymbol{D},t+1}\|_2 \leq 2\|\boldsymbol{u}_t^1\|_2$; second, $\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$ for any $r \in (0,1)$ and in particular for $r = \rho_{\tilde{\delta}}^{1/B}$; and third,

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

The constant C_u in the statement is obtained recalling that

$$r = \rho_{\tilde{\delta}}^{1/B} = \Big(1 - \frac{\tilde{\delta}}{4N^2}\Big)^{1/B}.$$

To obtain (27), we sum (58) over the time horizon t' and bound the double sum as follows: using $r = \rho_{\tilde{\delta}}^{1/B}$ for brevity, we have

$$\begin{aligned} \sum_{t=2}^{t'}\sum_{s=1}^{t-1} r^{t-1-s}\|e_s\|_2 &= \sum_{s=1}^{t'-1}\sum_{t=s+1}^{t'} r^{t-1-s}\|e_s\|_2\\ \sum_{s=1}^{t'-1}\|e_s\|_2\sum_{t=s+1}^{t'} r^{t-1-s} &\leq \frac{1}{1-r}\sum_{s=1}^{t'-1}\|e_s\|_2. \end{aligned}$$

Finally, we use again the bound $\|e_t\|_2 \leq 2\|\boldsymbol{u}_t^1\|_2$. ■

David Mateos-Núñez received the Licenciatura degree in mathematics with first-class honors from the Universidad Autónoma de Madrid, Spain, in 2011, and his M. Sc. and Ph.D. degrees in Mechanical Engineering from the University of California, San Diego, in 2012 and 2015, respectively, under supervision from Professor Jorge Cortés. He is currently a postdoc at Fraunhofer, Germany. His research focuses on distributed optimization and cooperative control of networked systems.

Jorge Cortés received the Licenciatura degree in mathematics from Universidad de Zaragoza, Zaragoza, Spain, in 1997, and the Ph.D. degree in engineering mathematics from Universidad Carlos III de Madrid, Madrid, Spain, in 2001. He held postdoctoral positions with the University of Twente, Twente, The Netherlands, and the University of Illinois at Urbana-Champaign, Urbana, IL, USA. He was an Assistant Professor with the Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA, USA, from 2004 to 2007. He is currently a Professor in the Department of Mechanical and Aerospace Engineering, University of California, San Diego, CA, USA. He is the author of Geometric, Control and Numerical Aspects of Nonholonomic Systems (Springer-Verlag, 2002) and co-author (together with F. Bullo and S. Martínez) of Distributed Control of Robotic Networks (Princeton University Press, 2009). He is an IEEE Fellow and an IEEE Control Systems Society Distinguished Lecturer. His current research interests include distributed control, networked games, power networks, distributed optimization, spatial estimation, and geometric mechanics.