Cooperative Data-Driven Distributionally Robust Optimization

Cherukuri, Ashish; Cortes, Jorge

Published in:
IEEE Transactions on Automatic Control

DOI:
10.1109/TAC.2019.2955031

Cooperative Data-Driven Distributionally Robust Optimization

Ashish Cherukuri and Jorge Cortés

***Abstract*—We study a class of multiagent stochastic optimization problems where the objective is to minimize the expected value of a function which depends on a random variable. The probability distribution of the random variable is unknown to the agents. The agents aim to cooperatively find, using their collected data, a solution with guaranteed out-of-sample performance. The approach is to formulate a data-driven distributionally robust optimization problem using Wasserstein ambiguity sets, which turns out to be equivalent to a convex program. We reformulate the latter as a distributed optimization problem and identify a convex–concave augmented Lagrangian, whose saddle points are in correspondence with the optimizers, provided a min–max interchangeability criteria is met. Our distributed algorithm design, then consists of the saddle-point dynamics associated to the augmented Lagrangian. We formally establish that the trajectories converge asymptotically to a saddle point and, hence, an optimizer of the problem. Finally, we identify classes of functions that meet the min–max interchangeability criteria.**

***Index Terms*—Data-driven methods, distributed optimization, distributionally robust optimization, multiagent systems.**

I. Introduction

Stochastic optimization in the context of multiagent systems has numerous applications, such as target tracking, distributed estimation, and cooperative planning and learning. Due to the expectation operator, solving these problems is computationally burdensome even when the probability distribution of the random variable is known. To address this intractability, researchers have studied numerous sample-based methods. Such methods might be subject to overfitting, and hence a major concern is obtaining out-of-sample performance guarantees. This is particularly relevant when only a few samples are available, typically in applications where acquiring samples is expensive due to the size and complexity of the system or when decisions must be taken in real time. Distributionally robust optimization (DRO) provides a regularization framework that guarantees good out-of-sample performance even when the data are disturbed and not sampled from the true distribution. We consider here the task for a group of agents to collaboratively find a data-driven solution for a stochastic optimization problem using the DRO framework.

A. Literature Review

To the large set of methods available to solve stochastic optimization problems [2], a recent addition is data-driven DRO (see, e.g., [3]–[6] and references therein). In this setup, the distribution of the random variable is unknown and a worst-case optimization is carried over a set of distributions, termed ambiguity set. This optimization provides probabilistic performance bounds for the original problem [3], [7] and overcomes the problem of overfitting. One way of designing the ambiguity sets is to consider the set of distributions that are close (in some metric) to some reference distribution constructed from the data. Popular metrics are ϕ-divergence [8], Prohorov metric [9], and Wasserstein distance [3] (adopted here). In [4], the ambiguity set is constructed with distributions that pass a goodness-of-fit test. In addition to data-driven methods, other works on DRO consider ambiguity sets defined using moment constraints [10], [11] and the Kullback–Leibler (KL)-divergence distance [12]. Tractable reformulations for the data-driven DRO have been well studied [3], [5], [13]. However, designing coordination algorithms to solve them when the data are gathered in a distributed way by a group of agents has not been investigated. This is the focus of this article. Besides data-driven DRO, one can solve the stochastic optimization problem considered here via other sampling-based methods, see [14]. Among these, sample average approximation (SAA) and stochastic approximation (SA) yield simple implementations and finite-sample guarantees independent of the dimension of the uncertainty (see, e.g., [2, Ch. 5] and [15]). However, such guarantees may not hold when the samples are corrupted and may require stricter assumptions on the cost function and the feasibility set. In contrast, the sample guarantees of the data-driven DRO method hold for more general settings (see, e.g., [3] and [7]), but are more conservative and do not scale well with the size of the uncertainty parameter. Additionally, the complexity of solving a data-driven DRO is often worse than that of the SAA and SA methods. Finally, our work also has connections with the growing body of literature on distributed optimization problems [16] and agreement-based algorithms to solve them (see, e.g., [17] and references therein).

B. Statement of Contributions

Our starting point is a multiagent stochastic optimization problem involving the minimization of the expected value of an objective function with a decision variable and a random variable as arguments. The probability distribution of the random variable is unknown. Agents collect a finite set of samples and wish to cooperatively solve a DRO problem over ambiguity sets defined as neighborhoods of the empirical distribution under the Wasserstein metric. Our first contribution is the reformulation of the DRO problem to display a structure amenable to distributed algorithm design. We achieve this by augmenting the decision variables to yield a convex optimization whose objective function is the aggregate of individual objectives and whose constraints involve consensus among neighboring agents. Building on an augmented version of the Lagrangian, we identify a convex–concave function, under a min–max interchangeability condition, whose saddle points are in one-to-one correspondence with the optimizers of the reformulated problem. Our second contribution is the design of the saddle-point

Manuscript received December 16, 2018; revised August 7, 2019; accepted November 15, 2019. Date of publication November 22, 2019; date of current version September 25, 2020. This work was presented at the 2017 Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA [1]. This work was supported by the AFOSR under Award FA9550-15-1-0108. Recommended by Associate Editor Prof. G. C. Calafiore. *(Corresponding author: Ashish Cherukuri.)*

A. Cherukuri is with the ENTEG, University of Groningen, 9712 CP Groningen, The Netherlands (e-mail: a.k.cherukuri@rug.nl).

J. Cortés is with the Department of Mechanical and Aerospace Engineering, University of California San Diego, San Diego, CA 92093 USA (e-mail: cortes@ucsd.edu).

Digital Object Identifier 10.1109/TAC.2019.2955031

dynamics for the identified convex–concave Lagrangian function. We show that the proposed dynamics is distributed and provably correct (its trajectories asymptotically converge to a solution of the original problem). Our third contribution is the identification of two broad classes of objective functions for which the min–max interchangeability holds. The first class is the set of functions that are convex–concave in the decision and the random variable, respectively. The second class is where functions are convex–convex and have some additional structure: they are either quadratic in the random variable or they correspond to the loss function of the least-squares problem. For space reasons, some proofs and additional material are available at [18].

II. Data-Driven Stochastic Optimization

This section[1] sets the stage for the formulation of our approach to deal with data-driven optimization in a distributed manner. The material is taken from [3] and included here for a self-contained exposition. The reader familiar with it can safely skip it. Let $(\Omega, \mathcal{F}, P)$ be a probability space and ξ be a random variable mapping this space to $(\mathbb{R}^m, B_\sigma(\mathbb{R}^m))$, where $B_\sigma(\mathbb{R}^m)$ is the Borel σ-algebra on $\mathbb{R}^m$. Let $\mathbb{P}$ and $\Xi \subseteq \mathbb{R}^m$ be the distribution and the support of the random variable ξ. Consider the stochastic optimization problem

$$\inf_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}[f(x,\xi)] \tag{1}$$

where $\mathcal{X} \subseteq \mathbb{R}^n$ is a closed convex set, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a continuous function, and $\mathbb{E}_{\mathbb{P}}[\cdot]$ is the expectation under $\mathbb{P}$. Assume that $\mathbb{P}$ is unknown and we are given N independently drawn samples $\widehat{\Xi} := \{\widehat{\xi}^k\}_{k=1}^N \subset \Xi$ of ξ. Note that, until revealed, $\widehat{\Xi}$ is a random object with distribution $\mathbb{P}^N := \prod_{i=1}^N \mathbb{P}$ and support $\Xi^N := \prod_{i=1}^N \Xi$. The objective is to find a *data-driven* solution of (1), denoted $\widehat{x}_N \in \mathcal{X}$, constructed using the dataset $\widehat{\Xi}$, that has a *finite-sample guarantee* given by

$$\mathbb{P}^N\Big(\mathbb{E}_{\mathbb{P}}[f(\widehat{x}_N,\xi)] \le \widehat{J}_N\Big) \ge 1-\beta \tag{2}$$

where $\widehat{J}_N$ might depend on $\widehat{\Xi}$ and $\beta \in (0,1)$ is the parameter governing $\widehat{x}_N$ and $\widehat{J}_N$. The goal is to find $\widehat{x}_N$ with low $\widehat{J}_N$ and β. To do so, the strategy is to determine a set $\widehat{\mathcal{P}}_N$ of probability distributions supported on Ξ and minimize the worst-case cost over $\widehat{\mathcal{P}}_N$. The set $\widehat{\mathcal{P}}_N$ is referred to as the *ambiguity* set. Once such a set is designed, $\widehat{J}_N$ and $\widehat{x}_N$ are defined as the optimal value and an optimizer, respectively, of the *DRO* problem

$$\widehat{J}_N := \inf_{x\in\mathcal{X}} \sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_N} \mathbb{E}_{\mathbb{Q}}[f(x,\xi)]. \tag{3}$$

We consider ambiguity sets $\widehat{\mathcal{P}}_N$ constructed using data. Formally, the *empirical distribution* is $\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{k=1}^N \delta_{\widehat{\xi}^k}$, where $\delta_{\widehat{\xi}^k}$ is the unit point mass at $\widehat{\xi}^k$. Let $\mathcal{M}(\Xi)$ be the space of probability distributions $\mathbb{Q}$ supported on Ξ with finite second moment, i.e., $\mathbb{E}_{\mathbb{Q}}[\|\xi\|^2] = \int_\Xi \|\xi\|^2\mathbb{Q}(d\xi) < +\infty$. The *2-Wasserstein metric* $d_{W_2} : \mathcal{M}(\Xi)\times\mathcal{M}(\Xi) \to \mathbb{R}_{\ge 0}$ is

$$d_{W_2}(\mathbb{Q}_1,\mathbb{Q}_2) = \left(\inf\left\{\int_{\Xi^2}\|\xi_1-\xi_2\|^2\Pi(d\xi_1,d\xi_2)\,\middle|\, \Pi\in\mathcal{H}(\mathbb{Q}_1,\mathbb{Q}_2)\right\}\right)^{(1/2)} \tag{4}$$

where $\mathcal{H}(\mathbb{Q}_1,\mathbb{Q}_2)$ is the set of all distributions on $\Xi\times\Xi$ with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$. Given $\epsilon \ge 0$, denote

$$\mathcal{B}_\epsilon(\widehat{\mathbb{P}}_N) := \{\mathbb{Q}\in\mathcal{M}(\Xi) \mid d_{W_2}(\widehat{\mathbb{P}}_N,\mathbb{Q}) \le \epsilon\}. \tag{5}$$

For an appropriately chosen radius ϵ, the ambiguity set $\widehat{\mathcal{P}}_N = \mathcal{B}_\epsilon(\widehat{\mathbb{P}}_N)$, plugged in problem (3), results into a finite-sample guarantee (2). There might be different ways of establishing this fact. For example, Esfahani and Kuhn [3] provided a bound for ϵ under $\mathbb{P}$ being light-tailed. The work [7] considers more general distributions and gives a different, potentially tighter, finite-sample guarantee for f being either quadratic or log-exponential loss function. The focus here is on the design of distributed algorithms to solve (3) with $\mathcal{B}_\epsilon(\widehat{\mathbb{P}}_N)$ as the ambiguity set. To this end, the next reformulation is key.

Theorem II.1. (Reformulation of (3)): For $N\in\mathbb{Z}_{\ge 1}$, the optimal value of (3) with the choice $\widehat{\mathcal{P}}_N = \mathcal{B}_\epsilon(\widehat{\mathbb{P}}_N)$ is equal to the optimum of the problem

$$\inf\nolimits_{\lambda\ge 0, x\in\mathcal{X}} \left\{\lambda\epsilon^2 + \tfrac{1}{N}\sum_{k=1}^N \max_{\xi\in\Xi}\Big(f(x,\xi) - \lambda\|\xi - \widehat{\xi}^k\|^2\Big)\right\}.$$

This problem is convex if $x \mapsto f(x,\tilde{\xi})$ is convex for all $\tilde{\xi}\in\Xi$.

This result and its proof are similar to [3, Th. 4.2] and its corresponding proof, respectively. While our metric is 2-Wasserstein, the referred result's is 1-Wasserstein. Theorem II.1 holds under weaker set of conditions on f (see, e.g., [5] and [13]). We however avoid this generality as it complicates the design and analysis of the distributed algorithm.

III. Problem Statement

Consider $n\in\mathbb{Z}_{\ge 1}$ agents communicating over an undirected weighted connected graph [19] $\mathcal{G} = (\mathcal{V},\mathcal{E},\mathsf{A})$. The set of vertices are enumerated as $\mathcal{V} := [n]$. Each agent $i\in[n]$ can send and receive information from its neighbors $\mathcal{N}_i = \{j\in\mathcal{V} \mid (i,j)\in\mathcal{E}\}$. Let $f : \mathbb{R}^d\times\mathbb{R}^m \to \mathbb{R}$, $(x,\xi)\mapsto f(x,\xi)$ be a continuously differentiable objective function. Assume that for any $\xi\in\mathbb{R}^m$, the map $x\mapsto f(x,\xi)$ is convex and that for any $x\in\mathbb{R}^d$, the map $\xi\mapsto f(x,\xi)$ is either convex or concave. Suppose that the set of $\xi\in\mathbb{R}^m$ for which $\mathbf{1}_n$ and $-\mathbf{1}_n$ are not a direction of recession for the convex function $x\mapsto f(x,\xi)$ is dense in $\mathbb{R}^m$. Assume that all agents know f. Given a random variable $\xi\in\mathbb{R}^m$ with support $\mathbb{R}^m$ and distribution $\mathbb{P}$, the original objective for the agents is to solve the stochastic optimization problem (1) over $\mathcal{X} = \mathbb{R}^d$ (the

[1] We use the following notation. Let $\mathbb{R}$, $\mathbb{R}_{\ge 0}$, and $\mathbb{Z}_{\ge 1}$ denote the set of real, non-negative real, and positive integer numbers. The extended reals are denoted as $\overline{\mathbb{R}} = \mathbb{R}\cup\{-\infty,\infty\}$. For $n\in\mathbb{Z}_{\ge 1}$, we let $[n] := \{1,\dots,n\}$. We let $\|\cdot\|$ denote the 2-norm on $\mathbb{R}^n$. Given $x,y\in\mathbb{R}^n$, $x\le y$ means $x_i\le y_i$ for $i\in[n]$. For $u\in\mathbb{R}^n$ and $w\in\mathbb{R}^m$, $(u;w)\in\mathbb{R}^{n+m}$ is its concatenation. We let $\mathbf{0}_n = (0,\dots,0)\in\mathbb{R}^n$, $\mathbf{1}_n = (1,\dots,1)\in\mathbb{R}^n$, and $\mathsf{I}_n\in\mathbb{R}^{n\times n}$ be the identity matrix. For $A\in\mathbb{R}^{n_1\times n_2}$ and $B\in\mathbb{R}^{m_1\times m_2}$, $A\otimes B\in\mathbb{R}^{n_1m_1\times n_2m_2}$ is the Kronecker product. The Cartesian product of $\{\mathcal{S}_i\}_{i=1}^n$ is $\prod_{i=1}^n \mathcal{S}_i := \mathcal{S}_1\times\cdots\times\mathcal{S}_n$. The interior of $\mathcal{S}\subset\mathbb{R}^n$ is $\operatorname{int}(\mathcal{S})$. For $f:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$, $(x,\xi)\mapsto f(x,\xi)$, we denote by $\nabla_x f$ and $\nabla_\xi f$ the partial derivatives of f with respect to its first and second arguments, respectively. Given $V:\mathcal{X}\to\mathbb{R}_{\ge 0}$, we let $V^{-1}(\le\delta) := \{x\in\mathcal{X}\mid V(x)\le\delta\}$. The projection of $y\in\mathbb{R}^n$ onto a closed convex set $\mathcal{K}\subset\mathbb{R}^n$ is $\operatorname{proj}_{\mathcal{K}}(y) = \operatorname{argmin}_{z\in\mathcal{K}}\|z-y\|$. The projection of $v\in\mathbb{R}^n$ at $x\in\mathcal{K}$ with respect to $\mathcal{K}$ is $\Pi_{\mathcal{K}}(x,v) = \lim_{\delta\to 0+}\big(\operatorname{proj}_{\mathcal{K}}(x+\delta v)-x\big)/\delta$. A vector $\varphi\in\mathbb{R}^n$ is *normal* to a convex set C at $x\in C$ if $(y-x)^\top\varphi\le 0$ for all $y\in C$. The set of all such vectors is the *normal cone* $N_C(x)$ to C at x. A vector d is a *direction of recession* of C if $x+\alpha d\in C$ for all $x\in C$ and $\alpha\ge 0$. A convex function $f:\mathbb{R}^n\to\overline{\mathbb{R}}$ is *proper* if there is $x\in\mathbb{R}^n$ such that $f(x)<+\infty$ and f does not take the value $-\infty$ anywhere in $\mathbb{R}^n$. The *epigraph* of f is $\operatorname{epi} f := \{(x,\lambda)\in(\mathbb{R}^n\times\overline{\mathbb{R}})\mid\lambda\ge f(x)\}$. A function f is closed if $\operatorname{epi} f$ is closed. For a closed proper convex function f, a vector d is a *direction of recession* of f if $(d,0)$ is a direction of recession of the set $\operatorname{epi} f$. If $f(x)\to+\infty$ whenever $\|x\|\to+\infty$, then f does not have a direction of recession. A function $F:\mathcal{X}\times\mathcal{Y}\to\overline{\mathbb{R}}$ is *convex–concave* if, for any $(\tilde{x},\tilde{y})\in\mathcal{X}\times\mathcal{Y}$, $x\mapsto F(x,\tilde{y})$ is convex and $y\mapsto F(\tilde{x},y)$ is concave. When the space $\mathcal{X}\times\mathcal{Y}$ is clear from the context, we refer to it as F being convex–concave in (x,y). A point $(x_*,y_*)\in\mathcal{X}\times\mathcal{Y}$ is a *saddle point* of F over $\mathcal{X}\times\mathcal{Y}$ if $F(x_*,y)\le F(x_*,y_*)\le F(x,y_*)$ for all $x\in\mathcal{X}$ and $y\in\mathcal{Y}$.

proposed method can handle generalizations to a generic closed convex set $\mathcal{X}$ by assuming that each agent knows a subset of $\mathbb{R}^d$ such that their intersection is $\mathcal{X}$). We assume that each agent has a certain number (at least one) of i.i.d realizations of the random variable ξ. We denote the data available to agent i by $\widehat{\Xi}_i$. Assume that $\widehat{\Xi}_i \cap \widehat{\Xi}_j = \emptyset$ for all $i, j \in [n]$ and let $\widehat{\Xi} = \cup_{i=1}\widehat{\Xi}_i$ containing N samples be the available dataset. To obtain a data-driven solution $\widehat{x}_N \in \mathbb{R}^d$ that has guaranteed performance bounds for the stochastic problem, using the framework presented in Section II, the agents aim to solve, in a distributed manner, the problem

$$\inf_{\lambda \geq 0, x} \left\{ \lambda\epsilon^2 + \frac{1}{N}\sum_{k=1}^{N} \max_{\xi \in \mathbb{R}^m} \left(f(x,\xi) - \lambda \|\xi - \widehat{\xi}^k\|^2 \right) \right\}. \tag{6}$$

The following is assumed to hold throughout the paper.

Assumption III.1. (Nontrivial Feasibility and Existence of Finite Optimizers of (6)): We assume that there exists a finite optimizer of (6) and the subset of $\mathbb{R}_{\geq 0} \times \mathbb{R}^d$ where the objective function in (6) takes finite values has a nonempty interior. •

The existence of finite optimizers is ensured if one of the sets of conditions for such existence given in [20] are met. Each agent could individually find a data-driven solution to (1) by using only its own data in the convex formulation (6). However, such a solution, in general, will have an inferior out-of-sample guarantee as compared to the one obtained collectively. In the cooperative setting, agents aim to solve (6) in a distributed manner, i.e., 1) each agent i has the information

$$\mathcal{I}_i := \{\widehat{\Xi}_i, f, \epsilon, n, N\} \tag{7}$$

where ϵ is the radius of the ambiguity set that agents agree upon beforehand, 2) each agent i can only communicate with its neighbors $\mathcal{N}_i$, 3) each agent i does not share with its neighbors any element of its own dataset $\widehat{\Xi}_i$, and 4) there is no central coordinator that can communicate with all agents.

Solving (6) in a distributed manner is challenging because the data are distributed over the network and the optimizer x^* depends on it all. Moreover, the inner maximization can be a nonconvex problem, in general. One way of solving (6) in a cooperative fashion is to let agents share their data with everyone via some sort of flooding mechanism. This violates 3) above. We specifically keep such methods out of scope due to two reasons. First, the data would not be private anymore, creating a possibility of adversarial action. Second, the communication burden of such a strategy is higher than our proposed distributed strategy when the size of the network and the dataset grows along the algorithm execution.

Remark III.2. (Alternative Distributed Algorithmic Solutions): The problem (6) can possibly be solved using other distributed methods. For instance, (6) can be written as a semi-infinite program, and then a distributed cutting-surface method can be designed following the centralized algorithm in [6]. If f is piecewise affine in ξ, (6) takes the form of a conic program (without the $\max$ operator in the objective), which can potentially be solved via primal-dual distributed solvers. Following [7] and [21], for certain f (linear form or objective of LASSO or logistic regression), (6) is equivalent to minimizing the empirical cost plus a regularizer. For such cases, primal-dual distributed solvers may be a valid solution strategy. The advantage of our methodology is its generality, not requiring to write different algorithms depending on the form of f. •

IV. Distributed Problem Formulation

We study the structure of the optimization (6) with the ulterior goal of facilitating the distributed algorithm design. Our first step is a reformulation that, by augmenting the agents' decision variables, yields an optimization where the objective is the aggregate of individual agent functions and constraints, which have a distributed structure. Our second step identifies a convex–concave function whose saddle points are the primal-dual optimizers of the reformulated problem under suitable conditions on the objective function. The structure of the original optimization makes this step particularly nontrivial.

A. Reformulation as Distributed Optimization Problem

We have each agent $i \in [n]$ maintain a copy of λ and x, denoted by $\lambda^i \in \mathbb{R}$ and $x^i \in \mathbb{R}^d$, respectively. Thus, the decision variables for i are (x^i, λ^i). For notational ease, let the concatenated vectors be $\lambda_{\mathrm{v}} := (\lambda^1; \ldots; \lambda^n)$, and $x_{\mathrm{v}} := (x^1; \ldots; x^n)$. Let $v_k \in [n]$ be the agent that holds the kth sample $\widehat{\xi}^k$ of the dataset. Consider the convex optimization

$$\min_{x_{\mathrm{v}}, \lambda_{\mathrm{v}} \geq \mathbf{0}_n} \quad h(\lambda_{\mathrm{v}}) + \frac{1}{N}\sum_{k=1}^{N} \max_{\xi \in \mathbb{R}^m} g_k(x^{v_k}, \lambda^{v_k}, \xi) \tag{8a}$$

$$\text{subject to} \quad \mathsf{L}\lambda_{\mathrm{v}} = \mathbf{0}_n, \tag{8b}$$

$$(\mathsf{L} \otimes \mathsf{I}_d) x_{\mathrm{v}} = \mathbf{0}_{nd}, \tag{8c}$$

where $\mathsf{L} \in \mathbb{R}^{n \times n}$ is the Laplacian of $\mathcal{G}^2$, we have used the shorthand notation $h : \mathbb{R}^n \to \mathbb{R}$ for $h(\lambda_{\mathrm{v}}) := \frac{\epsilon^2(\mathbf{1}_n^\top \lambda_{\mathrm{v}})}{n}$, and, for each $k \in [N]$, $g_k : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}$ for $g_k(x, \lambda, \xi) := f(x, \xi) - \lambda\|\xi - \widehat{\xi}^k\|^2$.

The next result establishes the correspondence between problems (6) and (8). The proof uses connectivity of the graph and is available in [18].

Lemma IV.1. (One-to-One Correspondence Between Optimizers of (6) and (8)): The following holds.

1) If (x^*, λ^*) is an optimizer of (6), then $(\mathbf{1}_n \otimes x^*, \lambda^* \mathbf{1}_n)$ is an optimizer of (8).
2) If $(x_{\mathrm{v}}^*, \lambda_{\mathrm{v}}^*)$ is an optimizer of (8), then an optimizer (x^*, λ^*) of (6) exists with $x_{\mathrm{v}}^* = \mathbf{1}_n \otimes x^*$ and $\lambda_{\mathrm{v}}^* = \lambda^* \mathbf{1}_n$.

Note that constraints (8b) and (8c) force agreement and that each of their components is computable by an agent of the network using only local information. Moreover, the objective function (8a) can be written as $\sum_{i=1}^n J_i(x^i, \lambda^i, \widehat{\Xi}_i)$, where

$$J_i(x^i, \lambda^i, \widehat{\Xi}_i) := \frac{\epsilon^2 \lambda^i}{n} + \frac{1}{N} \sum_{k : \widehat{\xi}^k \in \widehat{\Xi}_i} \max_{\xi \in \mathbb{R}^m} g_k(x^i, \lambda^i, \xi)$$

for all $i \in [n]$. Therefore, the problem (8) has the adequate structure from a distributed optimization viewpoint: an aggregate objective function and locally computable constraints.

B. Augmented Lagrangian and Saddle Points

Our next step is to identify an appropriate variant of the Lagrangian function of (8) such that 1) it does not consist of an inner maximization, unlike the objective in (8a), and 2) the primal-dual optimizers of (8) are saddle points of the newly introduced function. To proceed, we first denote for convenience the objective function (8a) with $F : \mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0} \to \mathbb{R}$

$$F(x_{\mathrm{v}}, \lambda_{\mathrm{v}}) := h(\lambda_{\mathrm{v}}) + \frac{1}{N}\sum_{k=1}^{N} \max_{\xi \in \mathbb{R}^m} g_k(x^{v_k}, \lambda^{v_k}, \xi). \tag{9}$$

The Lagrangian of (8) is $L : \mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0} \times \mathbb{R}^n \times \mathbb{R}^{nd} \to \overline{\mathbb{R}}$

$$L(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta) := F(x_{\mathrm{v}}, \lambda_{\mathrm{v}}) + \nu^\top \mathsf{L}\lambda_{\mathrm{v}} + \eta^\top (\mathsf{L} \otimes \mathsf{I}_d) x_{\mathrm{v}}, \tag{10}$$

[2] The *degree* matrix D is diagonal with $(\mathsf{D})_{ii} = \sum_{j=1}^n a_{ij}$, for $i \in [n]$. The *Laplacian* matrix is $\mathsf{L} = \mathsf{D} - \mathsf{A}$, where A is a weighted adjacency matrix of $\mathcal{G}$. Note $\mathsf{L} = \mathsf{L}^\top$. For connected $\mathcal{G}$, zero is a simple eigenvalue of L.

where $\nu \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^{nd}$ are dual variables corresponding to the equality constraints (8b) and (8c), respectively. L is convex–concave in $((x_{\mathrm{v}}, \lambda_{\mathrm{v}}), (\nu, \eta))$ on the domain $\lambda_{\mathrm{v}} \geq \mathbf{0}_n$. The next result states that (8) has zero duality gap, and follows from [22, Cor. 28.22 and Th. 28.3] using Assumption III.1.

Lemma IV.2. (Min–Max Equality for L): The set of saddle points of L over $(\mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}) \times (\mathbb{R}^n \times \mathbb{R}^{nd})$ is nonempty and

$$\inf_{x_{\mathrm{v}}, \lambda_{\mathrm{v}} \geq \mathbf{0}_n} \sup_{\nu, \eta} L(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta) = \sup_{\nu, \eta} \inf_{x_{\mathrm{v}}, \lambda_{\mathrm{v}} \geq \mathbf{0}_n} L(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta).$$

Furthermore, the following holds.

1) If $(\overline{x}_{\mathrm{v}}, \overline{\lambda}_{\mathrm{v}}, \bar{\nu}, \bar{\eta})$ is a saddle point of L over $(\mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}) \times (\mathbb{R}^n \times \mathbb{R}^{nd})$, then $(\overline{x}_{\mathrm{v}}, \overline{\lambda}_{\mathrm{v}})$ is an optimizer of (8).
2) If $(\overline{x}_{\mathrm{v}}, \overline{\lambda}_{\mathrm{v}})$ is an optimizer of (8), then there exists $(\bar{\nu}, \bar{\eta})$ such that $(\overline{x}_{\mathrm{v}}, \overline{\lambda}_{\mathrm{v}}, \bar{\nu}, \bar{\eta})$ is a saddle point of L over $(\mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}) \times (\mathbb{R}^n \times \mathbb{R}^{nd})$.

Based on this, one could write a saddle-point dynamics for the Lagrangian L as a distributed algorithm to find the optimizers. However, without strict or strong convexity assumptions on the objective function, the resulting dynamics is not guaranteed to converge (see, e.g., [23]). To overcome this hurdle, we augment the Lagrangian with quadratic terms. Let the augmented Lagrangian $L_{\mathrm{aug}} : \mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0} \times \mathbb{R}^n \times \mathbb{R}^{nd} \to \overline{\mathbb{R}}$ be

$$L_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta) = L(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta) + \frac{1}{2} x_{\mathrm{v}}^\top (\mathsf{L} \otimes \mathsf{I}_d) x_{\mathrm{v}} + \frac{1}{2} \lambda_{\mathrm{v}}^\top \mathsf{L} \lambda_{\mathrm{v}}.$$

Note that L_{aug} is also convex–concave in $((x_{\mathrm{v}}, \lambda_{\mathrm{v}}), (\nu, \eta))$ on the domain $\lambda_{\mathrm{v}} \geq \mathbf{0}_n$. The next result guarantees that this augmentation step does not change the saddle points.

Lemma IV.3. (Saddle points of L and L_{aug} are the same): A point is a saddle point of L over $(\mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}) \times (\mathbb{R}^n \times \mathbb{R}^{nd})$ if and only if it is a saddle point of L_{aug} over the same domain.

The proof follows by using the convexity property of the objective function in [24, Th. 1.1]. The above result implies that finding the saddle points of L_{aug} would take us to the primal-dual optimizers of (8). A final roadblock is writing a gradient-based dynamics for L_{aug}, given that this function involves a set of maximizations in its definition and so the gradient of L_{aug} with respect to x_{v} is undefined for $\lambda_{\mathrm{v}} = 0$. Thus, our next task is to get rid of these internal optimization and identify a function for which the saddle-point dynamics is well defined over the feasible domain. Note

$$L_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta) = \max_{\{\xi^k\}} \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta, \{\xi^k\}), \tag{11a}$$

$$\tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta, \{\xi^k\}) := h(\lambda_{\mathrm{v}}) + \frac{1}{N} \sum_{k=1}^{N} g_k(x^{v_k}, \lambda^{v_k}, \xi^k)$$
$$+ \nu^\top \mathsf{L} \lambda_{\mathrm{v}} + \eta^\top (\mathsf{L} \otimes \mathsf{I}_d) x_{\mathrm{v}} + \frac{1}{2} x_{\mathrm{v}}^\top (\mathsf{L} \otimes \mathsf{I}_d) x_{\mathrm{v}} + \frac{1}{2} \lambda_{\mathrm{v}}^\top \mathsf{L} \lambda_{\mathrm{v}}. \tag{11b}$$

The next result shows that, under appropriate conditions, $\tilde{L}_{\mathrm{aug}}$ is the function we need. The proof is available in [18].

Proposition IV.4. (Saddle Points of $\tilde{L}_{\mathrm{aug}}$ and Correspondence With Optimizers of (8)): Let $\mathcal{C} \subset \mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}$ with $\mathrm{int}(\mathcal{C}) \neq \emptyset$ be a closed, convex set such that

1) the saddle points of L_{aug} over $(\mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}) \times (\mathbb{R}^n \times \mathbb{R}^{nd})$ are contained in the set $\mathcal{C} \times (\mathbb{R}^n \times \mathbb{R}^{nd})$;
2) $\tilde{L}_{\mathrm{aug}}$ is convex–concave on $\mathcal{C} \times (\mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN})$;
3) for any (ν, η)

$$\min_{(x_{\mathrm{v}}, \lambda_{\mathrm{v}}) \in \mathcal{C}} \max_{\{\xi^k\}} \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta, \{\xi^k\})$$
$$= \max_{\{\xi^k\}} \min_{(x_{\mathrm{v}}, \lambda_{\mathrm{v}}) \in \mathcal{C}} \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta, \{\xi^k\}). \tag{12}$$

Then, the following holds.

1) The set of saddle points of $\tilde{L}_{\mathrm{aug}}$ over $\mathcal{C} \times (\mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN})$ is nonempty, convex, and closed.
2) If $(\overline{x}_{\mathrm{v}}, \overline{\lambda}_{\mathrm{v}}, \bar{\nu}, \bar{\eta}, \{(\bar{\xi})^k\})$ is a saddle point of $\tilde{L}_{\mathrm{aug}}$ over $\mathcal{C} \times (\mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN})$, $(\overline{x}_{\mathrm{v}}, \overline{\lambda}_{\mathrm{v}})$ is an optimizer of (8).
3) If $(\overline{x}_{\mathrm{v}}, \overline{\lambda}_{\mathrm{v}}) \in \mathcal{C}$ is an optimizer of (8), then there exists $(\bar{\nu}, \bar{\eta}, \{(\bar{\xi})^k\})$ such that $(\overline{x}_{\mathrm{v}}, \overline{\lambda}_{\mathrm{v}}, \bar{\nu}, \bar{\eta}, \{(\bar{\xi})^k\})$ is a saddle point of $\tilde{L}_{\mathrm{aug}}$ over $\mathcal{C} \times (\mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN})$.

Section VI identifies objective functions for which the hypotheses of Proposition IV.4 are met. We have introduced the set $\mathcal{C}$ to increase the level of generality in preparation for the exposition of our algorithm. Specifically, since f is not necessarily convex–concave, $\tilde{L}_{\mathrm{aug}}$ might not be convex–concave over the entire $(\mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}) \times (\mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN})$. For such cases, one can restrict the attention to $\mathcal{C} \times (\mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN})$ provided the hypotheses of the result are satisfied. We show later that, if f is convex–concave, one can set $\mathcal{C} = \mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}$.

V. Distributed Algorithm Design and Analysis

Here, we design and analyze our distributed algorithm to find the solutions of (6). Given the results of Section IV, and specifically Proposition IV.4, our algorithm seeks to find the saddle points of $\tilde{L}_{\mathrm{aug}}$ over $\mathcal{C} \times (\mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN})$. The dynamics consists of (projected) gradient-descent of $\tilde{L}_{\mathrm{aug}}$ in the convex variables and gradient-ascent in the concave ones. This is popularly termed as the saddle-point or the primal-dual dynamics [23], [25]. Given a closed, convex set $\mathcal{C} \subset \mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}$, the saddle-point dynamics for $\tilde{L}_{\mathrm{aug}}$ is

$$\begin{bmatrix} \frac{dx_{\mathrm{v}}}{dt} \\ \frac{d\lambda_{\mathrm{v}}}{dt} \end{bmatrix} = \Pi_{\mathcal{C}}\left((x_{\mathrm{v}}, \lambda_{\mathrm{v}}), \begin{bmatrix} -\nabla_{x_{\mathrm{v}}} \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta, \{\xi^k\}) \\ -\nabla_{\lambda_{\mathrm{v}}} \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta, \{\xi^k\}) \end{bmatrix} \right), \tag{13a}$$

$$\frac{d\nu}{dt} = \nabla_{\nu} \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta, \{\xi^k\}), \tag{13b}$$

$$\frac{d\eta}{dt} = \nabla_{\eta} \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta, \{\xi^k\}), \tag{13c}$$

$$\frac{d\xi^k}{dt} = \nabla_{\xi^k} \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta, \{\xi^k\}) \quad \forall k \in [N], \tag{13d}$$

where Π is the projection operator. For convenience, denote (13) by $X_{\mathrm{sp}} : \mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0} \times \mathbb{R}^{nd+n+mN} \to \mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0} \times \mathbb{R}^{nd+n+mN}$, where the first, second, and third components correspond to the dynamics of x_{v}, λ_{v}, and $(\nu, \eta, \{\xi^k\})$, respectively.

Remark V.1. (Distributed Implementation of (13)): To discuss the distributed character of the dynamics (13), we rely on $\mathcal{C}$ being decomposable into constraints on individual agent's decision variables, i.e., $\mathcal{C} := \Pi_{i=1}^n \mathcal{C}_i$ with $\mathcal{C}_i \subset \mathbb{R}^d \times \mathbb{R}_{\geq 0}$. This allows agents to perform the projection in (13a) in a distributed way. Denote the components of the dual variables η and ν by $\eta = (\eta^1; \eta^2; \ldots; \eta^n)$ and $\nu = (\nu^1; \nu^2; \ldots; \nu^n)$, so that agent $i \in [n]$ maintains $\eta^i \in \mathbb{R}^d$ and $\nu^i \in \mathbb{R}$. Furthermore, let $\mathcal{K}_i \subset [N]$ be the set of indices representing the samples held by i ($k \in \mathcal{K}_i$ if and only if $\widehat{\xi}^k \in \widehat{\Xi}_i$). For implementing X_{sp}, we assume that each agent i maintains and updates the variables $(x^i, \lambda^i, \nu^i, \eta^i, \{\xi^k\}_{k \in \mathcal{K}_i})$. The collection of these variables for all $i \in [n]$ forms $(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta, \{\xi^k\})$. From (13), the dynamics of variables maintained by i is computable by i using its variables and information collected from its neighbors. Hence, X_{sp} can be implemented in a distributed manner. Note that the number of variables in $\{\xi^k\}$ grows with the size of the data, whereas the size of all other variables is independent of the number of samples. Furthermore, for any agent i, $\{\xi^k\}_{k \in \mathcal{K}_i}$ is an internal state that is not communicated to its neighbors. •

Remark V.2. (Discretization and Implementation of (13)): The practical implementation of the dynamics (13) requires a proper discretization. A first-order discretization with standard conditions on stepsizes, as illustrated in [26], provides convergence guarantees for the running averages of the iterates. Alternatively, since our analysis rests on Lyapunov arguments, one can use the decay of the certificate to design a triggering mechanism, leading to discretizations with adaptive stepsizes and guaranteed convergence rates (see, e.g., [27] and [28]). Such discretization scheme can also be made robust against practical challenges such as asynchronicity in updates, noisy communication, and packet dropouts. This reasoning is the motivation to carry out the analysis in continuous time. •

The next result establishes the convergence of the dynamics $X_{\rm sp}$ to the saddle points of $\tilde{L}_{\rm aug}$. In previous work [23], [25], [28], we have extensively analyzed the convergence properties of saddle-point dynamics for convex–concave functions. However, those results do not apply directly to infer convergence for $X_{\rm sp}$ because projection operators are involved in the algorithm definition, $\tilde{L}_{\rm aug}$ is linear in some convex ($\lambda_{\rm v}$), and concave (ν, η) variables (thus, it is neither strictly/strongly convex, nor strictly/strongly concave, ruling out the possibility of using results that rely on either of these hypotheses) but is not linear in the convex variable $x_{\rm v}$ or in the concave one $\{\xi^k\}$.

Theorem V.3. (Convergence of X_{sp} to the Optimizers of (8)): Suppose the hypotheses of Proposition IV.4 hold. Assume further that there exists a saddle point $(x_{\rm v}^*, \lambda_{\rm v}^*, \nu^*, \eta^*, \{(\xi^k)^*\})$ of $\tilde{L}_{\rm aug}$ with $(x_{\rm v}^*, \lambda_{\rm v}^*) \in \operatorname{int}(\mathcal{C})$ such that $\xi \mapsto g_k((x_{\rm v}^*)^{v_k}, (\lambda_{\rm v}^*)^{v_k}, \xi)$ is strongly concave for all $k \in [N]$. Then, the trajectories of (13) starting in $\mathcal{C} \times \mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN}$ remain in this set and converge asymptotically to a saddle point of $\tilde{L}_{\rm aug}$. As a consequence, the $(x_{\rm v}, \lambda_{\rm v})$ component of the trajectory converges to an optimizer of (8).

Proof: The trajectories of (13) are understood in the Caratheodory sense [29]. By definition of the projection, any solution $t \mapsto (x_{\rm v}(t), \lambda_{\rm v}(t), \nu(t), \eta(t), \{\xi^k(t)\})$ starting with $(x_{\rm v}(0), \lambda_{\rm v}(0)) \in \mathcal{C}$ satisfies $(x_{\rm v}(t), \lambda_{\rm v}(t)) \in \mathcal{C}$ for all $t \geq 0$. *LaSalle function.* Let $(x_{\rm v}^*, \lambda_{\rm v}^*, \nu^*, \eta^*, \{(\xi^*)^k\})$ be the equilibrium point of $\tilde{L}_{\rm aug}$ satisfying $(x_{\rm v}^*, \lambda_{\rm v}^*) \in \operatorname{int}(\mathcal{C})$. Using the definition of equilibrium point in (13b) and (13c), we get

$$(\mathsf{L} \otimes \mathsf{I}_d) x_{\rm v}^* = \mathbf{0}_{nd} \text{ and } \mathsf{L}\lambda_{\rm v}^* = \mathbf{0}_n. \tag{14}$$

Consider the function $V : \mathcal{C} \times \mathbb{R}^{nd+n+Nm} \to \mathbb{R}_{\geq 0}$

$$V(x_{\rm v}, \lambda_{\rm v}, \zeta) := \frac{1}{2}(\|x_{\rm v} - x_{\rm v}^*\|^2 + \|\lambda_{\rm v} - \lambda_{\rm v}^*\|^2 + \|\zeta - \zeta^*\|^2),$$

where $\zeta := (\nu, \eta, \{\xi^k\})$ and, likewise, $\zeta^* := (\nu^*, \eta^*, \{(\xi^*)^k\})$. Writing the dynamics (13) as $(-\nabla_{x_{\rm v}} \tilde{L}_{\rm aug}; -\nabla_{\lambda_{\rm v}} \tilde{L}_{\rm aug}; \nabla_\zeta \tilde{L}_{\rm aug}) - (\varphi_{x_{\rm v}}; \varphi_{\lambda_{\rm v}}; \mathbf{0}_{nd+n+Nm})$, where $(\varphi_{x_{\rm v}}, \varphi_{\lambda_{\rm v}})$ is an element of the normal cone $N_{\mathcal{C}}(x_{\rm v}, \lambda_{\rm v})$ and following the steps of [25, Proof of Lemma 4.1], we obtain that the Lie derivative of V

$$\begin{aligned}\mathcal{L}_{X_{\rm sp}} V(x_{\rm v}, \lambda_{\rm v}, \zeta) \leq{}& \tilde{L}_{\rm aug}(x_{\rm v}^*, \lambda_{\rm v}^*, \zeta) - \tilde{L}_{\rm aug}(x_{\rm v}^*, \lambda_{\rm v}^*, \zeta^*) \\ &+ \tilde{L}_{\rm aug}(x_{\rm v}^*, \lambda_{\rm v}^*, \zeta^*) - \tilde{L}_{\rm aug}(x_{\rm v}, \lambda_{\rm v}, \zeta^*).\end{aligned} \tag{15}$$

From the definition of saddle point, the sum of the first two terms of the right-hand side are nonpositive and so is the sum of the last two. Therefore, we conclude $\mathcal{L}_{X_{\rm sp}} V(x_{\rm v}, \lambda_{\rm v}, \zeta) \leq 0$.

Application of LaSalle Invariance Principle. Using the monotonic evolution of V, we deduce two facts. First, given $\delta \geq 0$, any trajectory of (13) starting in $\mathcal{S}_\delta := V^{-1}(\leq \delta) \cap (\mathcal{C} \times \mathbb{R}^{n+nd+mN})$ remains in $\mathcal{S}_\delta$. In particular, every equilibrium point is stable. Second, the omega-limit set of each trajectory of (13) starting in $\mathcal{S}_\delta$ is invariant under the dynamics (see, e.g., [29] for relevant definitions). Thus, from the invariance principle for discontinuous dynamical systems [30, Prop. 3], any solution of (13) converges to the largest invariant set

$$\mathcal{M} \subset \{(x_{\rm v}, \lambda_{\rm v}, \zeta) \mid \mathcal{L}_{X_{\rm sp}} V(x_{\rm v}, \lambda_{\rm v}, \zeta) = 0, (x_{\rm v}, \lambda_{\rm v}) \in \mathcal{C}\}.$$

Properties of the Largest Invariant Set. Let $(x_{\rm v}, \lambda_{\rm v}, \zeta) \in \mathcal{M}$. Then, from $\mathcal{L}_{X_{\rm sp}} V(x_{\rm v}, \lambda_{\rm v}, \zeta) = 0$ and (15), we get

$$\tilde{L}_{\rm aug}(x_{\rm v}^*, \lambda_{\rm v}^*, \zeta) \overset{(a)}{=} \tilde{L}_{\rm aug}(x_{\rm v}^*, \lambda_{\rm v}^*, \zeta^*) \overset{(b)}{=} \tilde{L}_{\rm aug}(x_{\rm v}, \lambda_{\rm v}, \zeta^*). \tag{16}$$

Expanding the equality (a) and using (14), we obtain

$$\begin{aligned}&\sum_{k=1}^N g_k((x_{\rm v}^*)^{v_k}, (\lambda_{\rm v}^*)^{v_k}, \xi^k) \\ &\quad = \sum_{k=1}^N g_k((x_{\rm v}^*)^{v_k}, (\lambda_{\rm v}^*)^{v_k}, (\xi^*)^k).\end{aligned} \tag{17}$$

From the saddle-point property, $\{(\xi^*)^k\}$ maximizes $\{\xi^k\} \mapsto \sum_{k=1}^N g_k((x^*)^{v_k}, (\lambda^*)^{v_k}, \xi^k)$. This map is strongly concave by hypothesis. Thus, (17) yields $\xi^k = (\xi^*)^k$, for all $k \in [N]$. Expanding the equality (b) in (16) and using (14), we get

$$\begin{aligned}h(\lambda_{\rm v}^*) &+ \frac{1}{N}\sum_{k=1}^N g_k(x_{\rm v}^*)^{v_k}, (\lambda_{\rm v}^*)^{v_k}, (\xi^*)^k) = h(\lambda_{\rm v}) \\ &+ \frac{1}{N}\sum_{k=1}^N g_k(x_{\rm v}^{v_k}, \lambda_{\rm v}^{v_k}, (\xi^*)^k) + (\nu^*)^\top \mathsf{L}\lambda_{\rm v} \\ &+ (\eta^*)^\top (\mathsf{L} \otimes \mathsf{I}_d) x_{\rm v} + \frac{1}{2} x_{\rm v}^\top (\mathsf{L} \otimes \mathsf{I}_d) x_{\rm v} + \frac{1}{2}\lambda_{\rm v}^\top \mathsf{L}\lambda_{\rm v}.\end{aligned} \tag{18}$$

For ease of notation, let $y_{\rm v} := (x_{\rm v}; \lambda_{\rm v})$, $y_{\rm v}^* := (x_{\rm v}^*; \lambda_{\rm v}^*)$, and

$$G(y_{\rm v}) := h(\lambda_{\rm v}) + \frac{1}{N}\sum_{k=1}^N g_k(x_{\rm v}^{v_k}, \lambda_{\rm v}^{v_k}, (\xi^*)^k).$$

Then, the expression (18) can be written as

$$\begin{aligned}G(y_{\rm v}^*) = G(y_{\rm v}) &+ (\nu^*)^\top \mathsf{L}\lambda_{\rm v} + (\eta^*)^\top (\mathsf{L} \otimes \mathsf{I}_d) x_{\rm v} \\ &+ \frac{1}{2} y_{\rm v}^\top (\mathsf{L} \otimes \mathsf{I}_{d+1}) y_{\rm v}.\end{aligned} \tag{19}$$

From the definition of saddle point, $(x_{\rm v}^*, \lambda_{\rm v}^*)$ minimizes $(x_{\rm v}, \lambda_{\rm v}) \mapsto \tilde{L}_{\rm aug}(x_{\rm v}, \lambda_{\rm v}, \zeta^*)$ over $\mathcal{C}$. Moreover, by assumption $(x_{\rm v}^*, \lambda_{\rm v}^*)$ lies in the interior of $\mathcal{C}$. Thus

$$\nabla_{x_{\rm v}} \tilde{L}_{\rm aug}(x_{\rm v}^*, \lambda_{\rm v}^*, \zeta^*) = \mathbf{0}_{nd}, \tag{20a}$$

$$\nabla_{\lambda_{\rm v}} \tilde{L}_{\rm aug}(x_{\rm v}^*, \lambda_{\rm v}^*, \zeta^*) = \mathbf{0}_n. \tag{20b}$$

Here, (20a) yields $(\mathsf{L} \otimes \mathsf{I}_d)\eta^* = -\nabla_{x_{\rm v}} G(y_{\rm v}^*)$. Plugging this equality in (19) and rearranging terms gives

$$\begin{aligned}\frac{1}{2} y_{\rm v}^\top (\mathsf{L} \otimes \mathsf{I}_{d+1}) y_{\rm v} &= G(y_{\rm v}^*) - G(y_{\rm v}) \\ &\quad - (\nu^*)^\top \mathsf{L}\lambda_{\rm v} + x_{\rm v}^\top \nabla_{x_{\rm v}} G(y_{\rm v}^*).\end{aligned}$$

Note that $(x_{\rm v}^*)^\top \nabla_{x_{\rm v}} G(y_{\rm v}^*) = (x_{\rm v}^*)^\top (\nabla_{x_{\rm v}} G(y_{\rm v}^*) + (\mathsf{L} \otimes \mathsf{I}_d)\eta^* + (\mathsf{L} \otimes \mathsf{I}_d) x_{\rm v}^*) = (x_{\rm v}^*)^\top \nabla_{x_{\rm v}} \tilde{L}_{\rm aug}(x_{\rm v}^*, \lambda_{\rm v}^*, \zeta^*)$, where we have used (14). This, in turn, equals 0 because of (20a). Thus

$$\begin{aligned}\frac{1}{2} y_{\rm v}^\top (\mathsf{L} \otimes \mathsf{I}_{d+1}) y_{\rm v} &= G(y_{\rm v}^*) - G(y_{\rm v}) \\ &\quad - (\nu^*)^\top \mathsf{L}\lambda_{\rm v} + (x_{\rm v} - x_{\rm v}^*)^\top \nabla_{x_{\rm v}} G(y_{\rm v}^*).\end{aligned} \tag{21}$$

Expanding (20b) gives

$$\nabla_{\lambda_{\rm v}} G(y_{\rm v}^*) + \mathsf{L}\nu^* + \frac{1}{2}\mathsf{L}\lambda_{\rm v}^* = 0. \tag{22}$$

Premultiplying the above equation with $(\lambda_{\mathrm{v}}^*)^\top$ and using (14), we get $(\lambda_{\mathrm{v}}^*)^\top \nabla_{\lambda_{\mathrm{v}}} G(y_{\mathrm{v}}^*) = 0$ and we can rewrite (21) as

$$\frac{1}{2} y_{\mathrm{v}}^\top (\mathsf{L} \otimes \mathsf{I}_{d+1}) y_{\mathrm{v}} = G(y_{\mathrm{v}}^*) - G(y_{\mathrm{v}}) - (\nu^*)^\top \mathsf{L} \lambda_{\mathrm{v}} \\ + (x_{\mathrm{v}} - x_{\mathrm{v}}^*)^\top \nabla_{x_{\mathrm{v}}} G(y_{\mathrm{v}}^*) - (\lambda_{\mathrm{v}}^*)^\top \nabla_{\lambda_{\mathrm{v}}} G(y_{\mathrm{v}}^*). \tag{23}$$

Using (14) in (22) yields $\nabla_{\lambda_{\mathrm{v}}} G(y_{\mathrm{v}}^*) = -\mathsf{L}\nu^*$. That is, $\lambda_{\mathrm{v}}^\top \nabla_{\lambda_{\mathrm{v}}} G(y_{\mathrm{v}}^*) = -\lambda_{\mathrm{v}}^\top \mathsf{L} \nu^*$ which when replaced in (23) gives

$$\frac{1}{2} y_{\mathrm{v}}^\top (\mathsf{L} \otimes \mathsf{I}_{d+1}) y_{\mathrm{v}} = G(y_{\mathrm{v}}^*) - G(y_{\mathrm{v}}) + (y_{\mathrm{v}} - y_{\mathrm{v}}^*)^\top \nabla_{y_{\mathrm{v}}} G(y_{\mathrm{v}}^*).$$

The first-order convexity condition for F takes the form

$$G(y_{\mathrm{v}}) \geq G(y_{\mathrm{v}}^*) + (y_{\mathrm{v}} - y_{\mathrm{v}}^*)^\top \nabla_{y_{\mathrm{v}}} G(y_{\mathrm{v}}^*).$$

Using the previous two expressions, we get $y_{\mathrm{v}}^\top (\mathsf{L} \otimes \mathsf{I}_{d+1}) y_{\mathrm{v}} \leq 0$. This is only possible if this expression is zero because $\mathsf{L} \otimes \mathsf{I}_{d+1}$ is positive semidefinite. Equating it to zero, we get $x_{\mathrm{v}} = \mathbf{1}_n \otimes x$ and $\lambda_{\mathrm{v}} = \lambda \mathbf{1}_n$ for some (x, λ) and $(x_{\mathrm{v}}, \lambda_{\mathrm{v}}) \in \mathcal{C}$. So far, we have proved that if $(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \zeta) \in \mathcal{M}$, then

$$\xi^k = (\xi^*)^k \quad \forall k \in [N], \qquad x_{\mathrm{v}} = \mathbf{1}_n \otimes x, \tag{24a}$$

$$\lambda_{\mathrm{v}} = \lambda \mathbf{1}_n, \;\; (x_{\mathrm{v}}, \lambda_{\mathrm{v}}) \in \mathcal{C}. \tag{24b}$$

Identification of the Largest Invariant Set. Consider a trajectory $t \mapsto (x_{\mathrm{v}}(t), \lambda_{\mathrm{v}}(t), \zeta(t))$ of (13) starting and remaining in $\mathcal{M}$. Then, it must satisfy (24) for all $t \geq 0$, i.e., there exists $t \mapsto (x(t), \lambda(t))$ such that

$$\xi^k(t) = (\xi^*)^k \quad \forall k \in [N], \qquad x_{\mathrm{v}}(t) = \mathbf{1}_n \otimes x(t), \tag{25a}$$

$$\lambda_{\mathrm{v}}(t) = \lambda(t) \mathbf{1}_n, \;\; (x_{\mathrm{v}}(t), \lambda_{\mathrm{v}}(t)) \in \mathcal{C} \tag{25b}$$

for all $t \geq 0$. Plugging (25) in (13), we obtain that for all $t \geq 0$, along the considered trajectory, we have $\dot{\nu}(t) = \mathbf{0}_n$, $\dot{\eta}(t) = \mathbf{0}_{nd}$, and $\dot{\xi}(t) = \mathbf{0}_{mN}$. This implies that, for all $t \geq 0$

$$\begin{bmatrix} \frac{dx_{\mathrm{v}}(t)}{dt} \\ \frac{d\lambda_{\mathrm{v}}(t)}{dt} \end{bmatrix} = \Pi_{\mathcal{C}} \left((x_{\mathrm{v}}(t), \lambda_{\mathrm{v}}(t)), \begin{bmatrix} -\nabla_{x_{\mathrm{v}}} \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}(t), \lambda_{\mathrm{v}}(t), \zeta(0)) \\ -\nabla_{\lambda_{\mathrm{v}}} \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}(t), \lambda_{\mathrm{v}}(t), \zeta(0)) \end{bmatrix} \right)$$

which is a gradient descent dynamics of the convex function $(x_{\mathrm{v}}, \lambda_{\mathrm{v}}) \mapsto \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \zeta(0))$ projected over $\mathcal{C}$. Thus, either $t \mapsto \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}(t), \lambda_{\mathrm{v}}(t), \zeta(0))$ decreases at some t or the right-hand side of the above dynamics is zero at all times. Note

$$\tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}(t), \lambda_{\mathrm{v}}(t), \zeta(0)) \overset{(a)}{=} \tilde{L}_{\mathrm{aug}}(\mathbf{1}_n \otimes x(t), \lambda(t)\mathbf{1}_n, \zeta(0)) \\ \overset{(b)}{=} h(\lambda(t)\mathbf{1}_n) + \tfrac{1}{N} \sum_{k=1}^{N} g_k(\mathbf{1}_n \otimes x(t), \lambda(t)\mathbf{1}_n, (\xi^*)^k) \\ \overset{(c)}{=} \tilde{L}_{\mathrm{aug}}(\mathbf{1}_n \otimes x(t), \lambda(t)\mathbf{1}_n, \zeta^*) \overset{(d)}{=} \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}^*, \lambda_{\mathrm{v}}^*, \zeta^*)$$

for all $t \geq 0$. Equalities (a), (b), and (c) follow from (25) and the definition of $\tilde{L}_{\mathrm{aug}}$. Equality (d) follows from (16), which holds from every point in $\mathcal{M}$. The above implies that $t \mapsto \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}(t), \lambda_{\mathrm{v}}(t), \zeta(0))$ is a constant map. Hence, $(x_{\mathrm{v}}(0), \lambda_{\mathrm{v}}(0), \zeta(0))$ is an equilibrium of (13). Therefore, the set $\mathcal{M}$ is entirely composed of the equilibria of (13). Convergence to an equilibrium in the set of saddle points follows from this and the fact that each equilibrium point is stable. ■

VI. Objective Functions That Meet the Algorithm Convergence Criteria

We identify two broad classes of objective functions f for which the hypotheses of Proposition IV.4 hold. In both cases, we justify how (13) serves as a distributed solver of (8).

A. Convex–Concave Functions

We focus on objective functions that are convex–concave in (x, ξ): in addition to $x \mapsto f(x, \xi)$ being convex for each $\xi \in \mathbb{R}^m$, the function $\xi \mapsto f(x, \xi)$ is concave for each $x \in \mathbb{R}^d$. We proceed to check the hypotheses of Theorem V.3. To this end, let $\mathcal{C} = \mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}$, which is closed and convex with $\mathrm{int}(\mathcal{C}) \neq \emptyset$. Note that $\tilde{L}_{\mathrm{aug}}$ is convex–concave on $\mathcal{C} \times (\mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN})$ as f is convex–concave.

Lemma VI.1. (Interchange of Min–Max Operators): Let f be convex–concave in (x, ξ). Then, for any $(\nu, \eta) \in \mathbb{R}^n \times \mathbb{R}^{nd}$

$$\min_{x_{\mathrm{v}}, \lambda_{\mathrm{v}} \geq \mathbf{0}_n} \max_{\{\xi^k\}} \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta, \{\xi^k\}) \\ = \max_{\{\xi^k\}} \min_{x_{\mathrm{v}}, \lambda_{\mathrm{v}} \geq \mathbf{0}_n} \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta, \{\xi^k\}). \tag{26}$$

Proof: Given any (ν, η), denote $(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \{\xi^k\}) \mapsto \tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \nu, \eta, \{\xi^k\})$ by $\tilde{L}_{\mathrm{aug}}^{(\nu,\eta)}$. Since f is convex–concave, so is $\tilde{L}_{\mathrm{aug}}^{(\nu,\eta)}$ in the variables $((x_{\mathrm{v}}, \lambda_{\mathrm{v}}), \{\xi^k\})$. Consider an extension of $\tilde{L}_{\mathrm{aug}}^{(\nu,\eta)}$ over the entire $(\mathbb{R}^{nd} \times \mathbb{R}^n) \times (\mathbb{R}^{mN})$

$$\overline{L}_{\mathrm{aug}}^{(\nu,\eta)}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \{\xi^k\}) = \begin{cases} \tilde{L}_{\mathrm{aug}}^{(\nu,\eta)}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \{\xi^k\}), & \text{if } \lambda_{\mathrm{v}} \geq \mathbf{0}_n \\ +\infty, & \text{otherwise.} \end{cases}$$

One can see that $\overline{L}_{\mathrm{aug}}^{(\nu,\eta)}$ is closed, proper, and convex–concave. Furthermore, following [22, Th. 36.3], (26) holds iff

$$\min_{x_{\mathrm{v}}, \lambda_{\mathrm{v}}} \max_{\{\xi^k\}} \overline{L}_{\mathrm{aug}}^{(\nu,\eta)}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \{\xi^k\}) = \max_{\{\xi^k\}} \min_{x_{\mathrm{v}}, \lambda_{\mathrm{v}}} \overline{L}_{\mathrm{aug}}^{(\nu,\eta)}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \{\xi^k\}).$$

We establish this condition by checking the hypotheses of [22, Th. 37.3] for $\overline{L}_{\mathrm{aug}}^{(\nu,\eta)}$. For this, we show that: 1) there exists $\{\overline{\xi}^k\} \in \mathbb{R}^{mN}$ for which $(x_{\mathrm{v}}, \lambda_{\mathrm{v}}) \mapsto \overline{L}_{\mathrm{aug}}^{(\nu,\eta)}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \{\overline{\xi}^k\})$ does not have a direction of recession, and 2) there exists $(\overline{x}_{\mathrm{v}}, \overline{\lambda}_{\mathrm{v}}) \in \mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}$ with $\overline{\lambda}_{\mathrm{v}} > 0$ such that $\{\xi^k\} \mapsto -\overline{L}_{\mathrm{aug}}^{(\nu,\eta)}(\overline{x}_{\mathrm{v}}, \overline{\lambda}_{\mathrm{v}}, \{\xi^k\})$ does not have a direction of recession. For 1), by the assumptions on f, for each $k \in [N]$, there exists $\overline{\xi}^k \in B_{\epsilon_N(\beta)\sqrt{N}/\sqrt{2n}}(\widehat{\xi}^k)$ such that $\mathbf{1}_n$ and $-\mathbf{1}_n$ are not directions of recession for $x \mapsto f(x, \overline{\xi}^k)$. Picking these values, $\|\overline{\xi}^k - \widehat{\xi}^k\|^2 \leq \epsilon^2 N/2n$ for all $k \in [N]$. Thus

$$\overline{L}_{\mathrm{aug}}^{(\nu,\eta)}(x_{\mathrm{v}}, \lambda_{\mathrm{v}}, \{\overline{\xi}^k\}) = \frac{\epsilon^2 (z^t \lambda_{\mathrm{v}})}{n} + \frac{1}{N} \sum_{k=1}^{N} f(x^{v_k}, \overline{\xi}^k) \\ + \nu^\top \mathsf{L} \lambda_{\mathrm{v}} + \eta^\top (\mathsf{L} \otimes \mathsf{I}_d) x_{\mathrm{v}} + \frac{1}{2} x_{\mathrm{v}}^\top (\mathsf{L} \otimes \mathsf{I}_d) x_{\mathrm{v}} + \frac{1}{2} \lambda_{\mathrm{v}}^\top \mathsf{L} \lambda_{\mathrm{v}},$$

where $z \in \mathbb{R}^n$ with $z_i > 0$ for all $i \in [n]$. The right-hand side of the above expression as a function of $(x_{\mathrm{v}}, \lambda_{\mathrm{v}})$ does not have a direction of recession, that is, 1) holds. Next, we check 2). To this end, pick $\overline{x}_{\mathrm{v}} = \mathbf{1}_{nd}$ and $\overline{\lambda}_{\mathrm{v}} = \mathbf{1}_n$. Then

$$\overline{L}_{\mathrm{aug}}^{(\nu,\eta)}(\overline{x}_{\mathrm{v}}, \overline{\lambda}_{\mathrm{v}}, \{\xi^k\}) = \epsilon^2 + \frac{1}{N} \sum_{k=1}^{N} f(\mathbf{1}_d, \xi^k) - \|\xi^k - \widehat{\xi}^k\|^2.$$

Since $\xi \mapsto f(x, \xi)$ is concave for any $x \in \mathbb{R}^d$, we deduce $\overline{L}_{\mathrm{aug}}^{(\nu,\eta)}(\overline{x}_{\mathrm{v}}, \overline{\lambda}_{\mathrm{v}}, \{\xi^k\}) \to -\infty$ as $\|\{\xi^k\}\| \to \infty$, and 2) holds. ■

Hence, we conclude that the hypotheses of Proposition IV.4 hold for the considered class of objective functions, and we can state, invoking Theorem V.3, the next convergence result.

Corollary VI.2. (Convergence of Trajectories of X_{sp} for Convex–Concave f): Let f be convex–concave in (x, ξ) and $\mathcal{C} = \mathbb{R}^{nd} \times \mathbb{R}_{\geq 0}$. Assume there exists a saddle point $(x_{\mathrm{v}}^*, \lambda_{\mathrm{v}}^*, \nu^*, \eta^*, \{(\xi^k)^*\})$ of $\tilde{L}_{\mathrm{aug}}$ satisfying $\lambda_{\mathrm{v}}^* > \mathbf{0}_n$. Then, the trajectories of (13) starting in $\mathcal{C} \times \mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN}$ remain in this set and converge asymptotically to a saddle point of $\tilde{L}_{\text{aug}}$. As a consequence, the $(x_{\text{v}}, \lambda_{\text{v}})$ component of the trajectory converges to an optimizer of (8).

Note that $\mathcal{C} = \Pi_{i=1}(\mathbb{R}^d \times \mathbb{R}_{\geq 0})$ and, thus, (13) is implementable in a distributed way, cf. Remark V.1.

B. Convex–Convex Function

Here, we focus on objective functions for which both $x \mapsto f(x, \xi)$ and $\xi \mapsto f(x, \xi)$ are convex maps for all $x \in \mathbb{R}^d$ and $\xi \in \mathbb{R}^m$. Note that f need not be jointly convex in x and ξ. We further divide this classification into two.

1) Quadratic Function in ξ: Assume f is of the form

$$f(x, \xi) := \xi^\top Q \xi + x^\top R \xi + \ell(x), \tag{27}$$

where $Q \in \mathbb{R}^{m \times m}$ is positive definite, $R \in \mathbb{R}^{d \times m}$, and ℓ is a continuously differentiable convex function. Our next result is useful in identifying a domain that contains the saddle points of L_{aug} over $(\mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}) \times (\mathbb{R}^n \times \mathbb{R}^{nd})$.

Lemma VI.3. (Characterizing Where F is Finite): Assume f is of the form (27). Then, the function F defined in (9) is finite-valued only if $\lambda^i \geq \lambda_{\max}(Q)$ for all $i \in [n]$.

Proof: Assume there exists $\tilde{i} \in [n]$ such that $\lambda^{\tilde{i}} < \lambda_{\max}(Q)$. We wish to show that $F(x_{\text{v}}, \lambda_{\text{v}}) = +\infty$ in this case. For any k such that $\widehat{\xi}^k \in \widehat{\Xi}_{\tilde{i}}$, we have

$$\begin{aligned} g_k(x^{\tilde{i}}, \lambda^{\tilde{i}}, \xi) = \xi^\top (Q - \lambda^{\tilde{i}} \mathsf{I}_m) \xi + (x^{\tilde{i}})^\top R \xi + 2\lambda^{\tilde{i}} (\widehat{\xi}^k)^\top \xi \\ + \ell(x^{\tilde{i}}) - \lambda^{\tilde{i}} \|\widehat{\xi}^k\|^2. \end{aligned}$$

Let $w_{\max}(Q) \in \mathbb{R}^m$ be an eigenvector of Q corresponding to the eigenvalue $\lambda_{\max}(Q)$. Parameterizing $\xi = \alpha w_{\max}(Q)$

$$\begin{aligned} g_k(x^{\tilde{i}}, \lambda^{\tilde{i}}, \alpha w_{\max}(Q)) = \alpha^2 (\lambda_{max}(Q) - \lambda^{\tilde{i}}) \|w_{\max}(Q)\|^2 \\ + \alpha \Big((x^{\tilde{i}})^\top R + 2\lambda^{\tilde{i}} (\widehat{\xi}^k)^\top \Big) w_{\max}(Q) + \ell(x^{\tilde{i}}) - \lambda^{\tilde{i}} \|\widehat{\xi}^k\|^2. \end{aligned}$$

Thus, we get $\max_\alpha g_k(x^{\tilde{i}}, \lambda^{\tilde{i}}, \alpha w_{\max}(Q)) = +\infty$ and so $\max_\xi$ $g_k(x^{\tilde{i}}, \lambda^{\tilde{i}}, \xi) = +\infty$. Also, note that for any i and k with $\widehat{\xi}^k \in \widehat{\Xi}_i$, $\max_\xi g_k(x^i, \lambda^i, \xi) > -\infty$. This implies that $\sum_{k=1}^N \max_\xi g_k(x^{v_k}, \lambda^{v_k}, \xi) = +\infty$ and $F(x_{\text{v}}, \lambda_{\text{v}}) = +\infty$. ■

The above result implies that the optimizers of (8) for objective functions of the form (27) belong to the domain

$$\mathcal{C} := \mathbb{R}^{nd} \times \{\lambda_{\text{v}} \in \mathbb{R}^n_{\geq 0} \mid \lambda_{\text{v}} \geq \lambda_{\max}(Q) \mathbf{1}_n\}. \tag{28}$$

Therefore, the saddle points of L_{aug} over $(\mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}) \times (\mathbb{R}^n \times \mathbb{R}^{nd})$ are contained in $\mathcal{C} \times (\mathbb{R}^n \times \mathbb{R}^{nd})$. Note that $\mathcal{C}$ is closed, convex with a nonempty interior. Furthermore, following the proof of Lemma VI.3, $\tilde{L}_{\text{aug}}$ is convex–concave on $\mathcal{C} \times (\mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN})$ (an easy way to validate this fact is by noting that the Hessian of $\tilde{L}_{\text{aug}}$ with respect to the convex (concave) variables is positive (negative) semidefinite). Finally, repeating the proof of Lemma VI.1, we arrive at the equality (12). Using these facts in Theorem V.3 yields the next result.

Corollary VI.4. (Convergence of Trajectories of X_{sp} for Quadratic f): Let f be of the form (27) and $\mathcal{C}$ be given in (28). Assume further that there exists a saddle point $(x^*_{\text{v}}, \lambda^*_{\text{v}}, \nu^*, \eta^*, \{(\xi^k)^*\})$ of $\tilde{L}_{\text{aug}}$ satisfying $\lambda^*_{\text{v}} > \lambda_{\max}(Q) \mathbf{1}_n$. Then, the trajectories of (13) starting in $\mathcal{C} \times \mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN}$ remain in this set and converge asymptotically to a saddle point of $\tilde{L}_{\text{aug}}$. As a consequence, the $(x_{\text{v}}, \lambda_{\text{v}})$ component of the trajectory converges to an optimizer of (8).

Note that $\mathcal{C}$ given in (28) can be written as $\mathcal{C} = \Pi_{i=1}^n (\mathbb{R}^d \times \{\lambda \in \mathbb{R}_{\geq 0} \mid \lambda \geq \lambda_{\max}(Q)\})$. Thus, following Remark V.1, the dynamics (13) can be implemented in a distributed manner.

2) Least-Squares Problem: Let $d = m$ and assume additionally that the function f is of the form

$$f(x, \xi) := a(\xi_m - (\xi_{1:m-1}; 1)^\top x)^2, \tag{29}$$

where $a > 0$ and $\xi_{1:m-1}$ denotes the vector ξ without the last component ξ_m. Note that f corresponds to the objective function for a least-squares problem and it cannot be written in the form (27). We first characterize the set over which the objective function (9) takes finite values. The proof [18] mimics the steps of the proof of Lemma VI.3.

Lemma VI.5. (Characterizing Where F is Finite): Assume f is of the form (29). Then, the function F defined in (9) is finite-valued only if $\lambda^i \geq a\|(x^i_{1:m-1}; 1)\|^2$ for all $i \in [n]$.

Guided by the above result, let

$$\mathcal{C} := \mathbb{R}^{nd} \times \{\lambda_{\text{v}} \in \mathbb{R}^n_{\geq 0} \,|\, \lambda^i \geq a\|(x^i_{1:m-1}; 1)\|^2 \quad \forall i \in [n]\}. \tag{30}$$

Owing to Lemma VI.5, the optimizers of (8) belong to $\mathcal{C}$ and so, the saddle points of L_{aug} over $(\mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}) \times (\mathbb{R}^n \times \mathbb{R}^{nd})$ are contained in $\mathcal{C} \times (\mathbb{R}^n \times \mathbb{R}^{nd})$. Furthermore, $\mathcal{C}$ is closed, convex with a nonempty interior and $\tilde{L}_{\text{aug}}$ is convex–concave on $\mathcal{C} \times (\mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN})$. Finally, one can show that (12) holds here. Using these facts in Theorem V.3 yields the next result.

Corollary VI.6. (Convergence of Trajectories of X_{sp} for Least-Squares Problem): Let f be of the form (29) and $\mathcal{C}$ be given in (30). Assume there exists a saddle point $(x^*_{\text{v}}, \lambda^*_{\text{v}}, \nu^*, \eta^*, \{(\xi^k)^*\})$ of $\tilde{L}_{\text{aug}}$ satisfying $(x^*_{\text{v}}, \lambda^*_{\text{v}}) \in \text{int}(\mathcal{C})$. Then, the trajectories of (13) starting in $\mathcal{C} \times \mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN}$ remain in this set and converge asymptotically to a saddle point of $\tilde{L}_{\text{aug}}$. As a consequence, the $(x_{\text{v}}, \lambda_{\text{v}})$ component of the trajectory converges to an optimizer of (8).

The saddle-point dynamics (13) is amenable to distributed implementation too, cf. Remark V.1, as one can write $\mathcal{C} = \Pi_{i=1}^n \{(x, \lambda) \in \mathbb{R}^d \times \mathbb{R}_{\geq 0} \mid \lambda \geq a\|(x_{1:m-1}; 1)\|^2\}$. We present in [18] an example where this dynamics is employed to find a data-driven solution for a regression problem with quadratic loss function and an affine predictor.

VII. Conclusion

We have studied a stochastic optimization problem, where a group of agents rely on their individually collected data to jointly determine a data-driven solution with guaranteed out-of-sample performance. Our approach identifies an augmented Lagrangian whose saddle points are in one-to-one correspondence with the primal-dual optimizers. This characterization relies upon certain interchangeability properties, which are satisfied by several classes of objective functions (convex–concave, convex–convex quadratic in the data, and convex–convex associated to least-squares problems). We have designed a provably correct distributed saddle-point algorithm where agents share individual solution estimates, not the collected data. Future work will explore the characterization of the convergence rate, the design of strategies capable of tracking the optimal solution with streaming data, and the analysis of scenarios with network chance constraints.

References

[1] A. Cherukuri and J. Cortés, "Data-driven distributed optimization using Wasserstein ambiguity sets," in *Proc. Allerton Conf. Commun., Control Comput.*, 2017, Monticello, IL, USA, pp. 38–44.

[2] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming*. Philadelphia, PA, USA: SIAM, 2014.

[3] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Math. Program.*, vol. 171, no. 1, pp. 115–166, 2018.

[4] D. Bertsimas, V. Gupta, and N. Kallus, "Robust sample average approximation," *Math. Program.*, vol. 171, no. 1, pp. 217–282, 2018.

[5] R. Gao and A. J. Kleywegt, "Distributionally robust stochastic optimization with Wasserstein distance," 2016. [Online]. Available: https://arxiv.org/abs/1604.02199

[6] F. Luo and S. Mehrotra, "Decomposition algorithm for distributionally robust optimization using Wasserstein metric with an application to a class of regression models," *Eur. J. Oper. Res.*, vol. 278, no. 1, pp. 20–35, 2019.

[7] J. Blanchet, Y. Kang, and K. Murthy, "Robust Wasserstein profile inference and applications to machine learning," *J. Appl. Probabil.*, vol. 56, no. 3, pp. 830–857, 2019.

[8] R. Jiang and Y. Guan, "Data-driven chance constrained stochastic program," *Math. Program., Ser. A*, vol. 158, pp. 291–327, 2016.

[9] E. Erdoğan and G. Iyengar, "Ambiguous chance constrained problems and robust optimization," *Math. Program., Ser. B*, vol. 107, pp. 37–61, 2006.

[10] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Oper. Res.*, vol. 58, pp. 595–612, 2010.

[11] W. Wieseman, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Oper. Res.*, vol. 62, no. 6, pp. 1358–1376, 2014.

[12] Z. Hu and L. J. Hong, "Kullback–Leibler divergence constrained distributionally robust optimization," 2013. [Online]. Available: http://www.optimization-online.org/DB_FILE/2012/11/3677.pdf

[13] J. Blanchet and K. Murthy, "Quantifying distributional model risk via optimal transport," *Math. Oper. Res.*, vol. 44, no. 2, pp. 565–600, 2019.

[14] T. H. de Mello and G. Bayraksan, "Monte Carlo sampling-based methods for stochastic optimization," *Surveys Oper. Res. Manage. Sci.*, vol. 19, no. 1, pp. 56–85, 2014.

[15] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1574–1609, 2009.

[16] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Belmont, MA, USA: Athena Sci., 1997.

[17] A. Nedić, "Distributed optimization," in *Encyclopedia of Systems and Control*, J. Baillieul and T. Samad, Eds. New York, NY, USA: Springer, 2015.

[18] A. Cherukuri and J. Cortés, "Cooperative data-driven distributionally robust optimization," 2017. [Online]. Available: arxiv.org/abs/1711.04839

[19] F. Bullo, J. Cortés, and S. Martinez, *Distributed Control of Robotic Networks* Applied Mathematics Series. Princeton, NJ, USA: Princeton Univ. Press, 2009.

[20] A. E. Ozdaglar and P. Tseng, "Existence of global minima for constrained optimization," *J. Optim. Theory Appl.*, vol. 128, no. 3, pp. 523–546, 2006.

[21] R. Gao, X. Chen, and A. J. Kleywegt, "Distributional robustness and regularization in statistical learning," 2017. [Online]. Available: https://arxiv.org/abs/1712.06050

[22] R. T. Rockafellar, *Convex Analysis* Princeton Landmarks in Mathematics and Physics., Princeton, NJ, USA: Princeton Univ. Press, 1997.

[23] A. Cherukuri, B. Gharesifard, and J. Cortés, "Saddle-point dynamics: conditions for asymptotic stability of saddle points," *SIAM J. Control Optim.*, vol. 55, no. 1, pp. 486–511, 2017.

[24] X. L. Sun, D. Li, and K. I. M. Mckinnon, "On saddle points of augmented Lagrangians for constrained nonconvex optimization," *SIAM J. Optim.*, vol. 15, no. 4, pp. 1128–1146, 2005.

[25] A. Cherukuri, E. Mallada, and J. Cortés, "Asymptotic convergence of constrained primal-dual dynamics," *Syst. Control Lett.*, vol. 87, pp. 10–15, 2016.

[26] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *J. Optim. Theory Appl.*, vol. 142, no. 1, pp. 205–228, 2009.

[27] S. S. Kia, J. Cortés, and S. Martinez, "Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication," *Automatica*, vol. 55, pp. 254–264, 2015.

[28] A. Cherukuri, E. Mallada, S. H. Low, and J. Cortés, "The role of convexity in saddle-point dynamics: Lyapunov function and robustness," *IEEE Trans. Autom. Control*, vol. 63, no. 8, pp. 2449–2464, Aug. 2018.

[29] J. Cortés, "Discontinuous dynamical systems—A tutorial on solutions, nonsmooth analysis, and stability," *IEEE Control Syst.*, vol. 28, no. 3, pp. 36–73, Jun. 2008.

[30] A. Bacciotti and F. Ceragioli, "Nonpathological Lyapunov functions and discontinuous Caratheodory systems," *Automatica*, vol. 42, no. 3, pp. 453–458, 2006.