Abstract—This paper studies 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, so each one gathers samples of it. The agents aim to cooperatively find, using their data, a solution to the optimization problem with guaranteed out-of-sample performance. The approach is to formulate a data-driven 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 function 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 of the dynamics converge asymptotically to a saddle point and hence an optimizer of the problem. Finally, we provide a class of functions that meet the min-max interchangeability criteria. Simulations illustrate our results.

I. INTRODUCTION

Stochastic optimization in the context of multiagent systems has numerous applications, such as target tracking, distributed estimation, and cooperative planning and learning. Solving these problems, in an exact sense, requires the knowledge of the probability distribution of the random variables. Often this information is unavailable and instead, agents gather samples and use the data to find a solution to the stochastic optimization in an approximate sense. If the available dataset is large, machine learning algorithms are able to find the optimizer of the original problem with arbitrary precision.

However, when the dataset is small, these algorithms fail to provide guarantees on the output obtained from the procedure. Scenarios with small datasets appear 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, leaving less room for gathering many samples. Distributionally robust optimization (DRO), instead, uses finite datasets to provide a solution with desirable out-of-sample performance guarantees. Motivated by this, we consider here the task for a group of agents to collaboratively find a data-driven solution for a stochastic optimization problem using the tools provided by the DRO framework. Instead of breaking the problem in two separate subproblems (model and estimate first the uncertainty using data-fusion algorithms and then use the estimates to solve the stochastic optimization problem), the DRO method tackles them jointly, providing approximations of the optimizers tailored to the quality and size of the gathered data.

Literature review: Stochastic optimization is a classical topic [2]. To the large set of methods available to solve this type of problems, a recent addition is data-driven distributionally robust optimization, see e.g. [3], [4], [5] and references therein. In this setup, the distribution of the random variable is unknown and so, a worst-case optimization is carried over a set of distributions (termed ambiguity set) that contains the true distribution with high probability. This worst-case optimization provides probabilistic performance bounds for the original stochastic optimization. One way of designing the ambiguity sets is to consider the set of distributions that are close (in some distance metric over the space of distributions) to some reference distribution constructed from the available data. Depending on the metric, one gets different ambiguity sets with different performance bounds. Some popular metrics are \(\phi\)-divergence [6], Prokhorov metric [7], and Wasserstein distance [3]. Here, we consider ambiguity sets defined using the Wasserstein metric. Tractable reformulations for the data-driven DRO methods have been well studied. However, designing coordination algorithms to find a data-driven solution when the data is gathered in a distributed way by a network of agents has not been investigated. This is the focus of this paper. Our work has connections with the growing body of literature on distribution optimization [8], [9], [10] problems and agreement-based algorithms to solve them, see e.g., [11], [12], [13], [14], [15], [16] and references therein. Our emphasis on guarantees with small datasets is in contrast with typical setups of distributed machine learning, see e.g. [17], which assume the availability of large datasets and provide asymptotic guarantees on the learning algorithms. Nonetheless, the coordination aspect in these works is similar in spirit to what we emphasize on here.

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 and instead, agents collect a finite set of samples of it. Given this data, each agent can individually find a data-driven solution of the stochastic optimization. However, agents wish to cooperate to leverage on the data collected by everyone in the group. Our approach consists of formulating a distributionally robust optimization problem over ambiguity sets defined as neigh-
borhoods of the empirical distribution under the Wasserstein metric. The solution of this problem has guaranteed out-of-sample performance for the stochastic optimization. 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 associated Lagrangian function, we identify a convex-concave function which under a min-max formulation.

A weighted graph is a triplet \( \mathcal{G} = (V, \mathcal{E}, \mathcal{A}) \), where \((V, \mathcal{E})\) is a digraph and \( \mathcal{A} \in \mathbb{R}_{\geq 0}^{m \times n}\) is the (symmetric) adjacency matrix of \( \mathcal{G} \), with the property that \(a_{ij} > 0\) if \((i, j) \in \mathcal{E}\) and \(a_{ij} = 0\), otherwise. The weighted degree of \(i \in [n]\) is \(w_i = \sum_{j=1}^{n} a_{ij}\). The weighted degree matrix \(D\) is the diagonal matrix defined by \(D_{ii} = w_i\) for all \(i \in [n]\). The Laplacian matrix is \(L = D - A\). Note that \(L = L^\top\) and \(L1_n = 0\). If \(G\) is connected, then zero is a simple eigenvalue of \(L\).

3) Convex analysis: Here we introduce elements from convex analysis following [19]. A set \(C \subseteq \mathbb{R}^n\) is convex if \((1 - \lambda)x + \lambda y \in C\) whenever \(x \in C\), \(y \in C\), and \(\lambda \in (0, 1)\). A vector \(\varphi \in \mathbb{R}^n\) is normal to a convex set \(C\) at a point \(x \in C\) if \((y - x)\top \varphi \leq 0\) for all \(y \in C\). The set of all vectors normal to \(C\) at \(x\), denoted \(N_C(x)\), is the normal cone to \(C\) at \(x\). The affine hull of \(S \subseteq \mathbb{R}^n\) is the smallest affine space containing \(S\),

\[
\text{aff}(S) := \{ \sum_{i=1}^{k} \lambda_i x_i | k \in \mathbb{Z}_{\geq 1}, x_i \in S, \lambda_i \in \mathbb{R}, \sum_{i=1}^{k} \lambda_i = 1 \}.
\]

The relative interior of a convex set \(C\) is the interior of \(C\) relative to the affine hull of \(C\). Formally,

\[
\text{ri}(C) := \{ x \in \text{aff}(C) | \exists \epsilon > 0, (x + \epsilon B_1(0)) \cap (\text{aff}(C)) \subset C \}.
\]

Given a convex set \(C\), a vector \(d\) is a direction of recession of \(C\) if \(x + \alpha d \in C\) for all \(x \in C\) and \(\alpha \geq 0\).

A convex function \(f : \mathbb{R}^n \rightarrow \mathbb{R}\) is proper if there exists \(x \in \mathbb{R}^n\) such that \(f(x) < +\infty\) and \(f\) does not take the value \(-\infty\) anywhere in \(\mathbb{R}^n\). The epigraph of \(f\) is the set

\[
\text{epi}f := \{(x, \lambda) \in (\mathbb{R}^n \times \mathbb{R}) | \lambda \geq f(x) \}.
\]

A function \(f\) is closed if \(\text{epi}f\) is a closed set. The function \(f\) is convex if and only if \(\text{epi}f\) is convex. 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 \(\text{epi}f\). Intuitively, it is the direction along which \(f\) is monotonically non-increasing. If \(f(x) \rightarrow +\infty\) whenever \(||x|| \rightarrow +\infty\), then \(f\) does not have a direction of recession.

A function \(F : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}\) is convex-concave (on \(\mathcal{X} \times \mathcal{Y}\)) if, given any point \((\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}\), \(x \mapsto F(x, \tilde{y})\) 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 this property as \(F\) being convex-concave in \((x, y)\). A point \((x_0, y_0) \in \mathcal{X} \times \mathcal{Y}\) is a saddle point of \(F\) over the set \(\mathcal{X} \times \mathcal{Y}\) if \(F(x_0, y_0) \leq F(x_0, y) \leq F(x_0, y_0)\) for all \(x \in \mathcal{X}\) and \(y \in \mathcal{Y}\). The set of saddle points of a convex-concave function \(F\) is convex. Each saddle point is a critical point of \(F\), i.e., if \(F\) is differentiable, then \(\nabla_x F(x_0, z_0) = 0\) and \(\nabla_z F(x_0, z_0) = 0\). Additionally, if \(F\) is twice differentiable, then \(\nabla_{xx} F(x_0, z_0) \preceq 0\) and \(\nabla_{zz} F(x_0, z_0) \succeq 0\). Given a convex-concave function \(F : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}\), define

\[
\mathcal{X} := \{ x \in \mathbb{R}^n | F(x, y) < +\infty \text{ for all } y \in \mathbb{R}^m \},\]

\[
\mathcal{Y} := \{ y \in \mathbb{R}^m | F(x, y) > -\infty \text{ for all } x \in \mathbb{R}^n \}.
\]

The product set \(\mathcal{X} \times \mathcal{Y}\) is called the effective domain of \(F\). The sets \(\mathcal{X}\) and \(\mathcal{Y}\) are convex. Note that \(F\) is finite
on $\mathcal{X} \times \mathcal{Y}$. If $\mathcal{X} \times \mathcal{Y}$ is nonempty, then $F$ is called proper. If the following equality holds
\[
\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} F(x, y) = \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} F(x, y),
\]
then this common value is called the saddle value of $F$. The function $F$ is closed if for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the functions $x \mapsto F(x, y)$ and $y \mapsto -F(x, y)$ are closed.

Theorem II.1. (Existence of finite saddle value and saddle point [19, Theorem 37.3 & 37.6]) Let $F: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a closed proper convex-concave function with effective domain $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^{n} \times \mathbb{R}^{m}$. If the following conditions hold,

(i) The convex functions $x \mapsto F(x, y)$ for $y \in \text{ri} (\mathcal{Y})$ have no common direction of recession;

(ii) The convex functions $y \mapsto -F(x, y)$ for $x \in \text{ri} (\mathcal{X})$ have no common direction of recession;

then the saddle value must be finite, there exists a saddle point of $F$ in the effective domain $\mathcal{X} \times \mathcal{Y}$ and the saddle value is attained at the saddle point.

4) Discontinuous dynamical systems: Here we present notions of discontinuous and projected dynamical systems from [20, 21, 22]. Let $f: \mathbb{R}^{n} \to \mathbb{R}^{n}$ be a Lebesgue measurable and locally bounded function, and consider
\[
\dot{x} = f(x).
\] (1)

A map $\gamma: [0, T] \to \mathbb{R}^{n}$ is a (Caratheodory) solution of (1) on the interval $[0, T]$ if it is absolutely continuous on $[0, T]$ and satisfies $\gamma(t) = f(\gamma(t))$ almost everywhere in $[0, T]$. We use the terms solution and trajectory interchangeably. A set $S \subset \mathbb{R}^{n}$ is invariant under (1) if every solution starting in $S$ remains in $S$. For a solution $\gamma$ of (1) defined on the time interval $[0, \infty)$, the omega-limit set $\Omega(\gamma)$ is defined by
\[
\Omega(\gamma) = \{ y \in \mathbb{R}^{n} \mid \text{there exists } \{t_{k}\}_{k=1}^{\infty} \subset [0, \infty) \text{ with } \lim_{k \to \infty} t_{k} = \infty \text{ and } \lim_{k \to \infty} \gamma(t_{k}) = y \}.
\]

If the solution $\gamma$ is bounded, then $\Omega(\gamma) \neq \emptyset$ by the Bolzano-Weierstrass theorem [23, p. 33]. Given a continuously differentiable function $V: \mathbb{R}^{n} \to \mathbb{R}$, the Lie derivative of $V$ along (1) at $x \in \mathbb{R}^{n}$ is $L_{f}V(x) = \nabla V(x)^{\top} f(x)$. The next result is a simplified version of [20, Proposition 3].

Proposition II.2. (Invariance principle for discontinuous Caratheodory systems): Let $S \subset \mathbb{R}^{n}$ be compact and invariant. Assume that, for each point $x_{0} \in S$, there exists a unique solution of (1) starting at $x_{0}$ and that its omega-limit set is invariant too. Let $V: \mathbb{R}^{n} \to \mathbb{R}$ be a continuously differentiable map such that $L_{f}V(x) \leq 0$ for all $x \in S$. Then, any solution of (1) starting at $S$ converges to the largest invariant set in $\text{cl}(\{x \in S \mid L_{f}V(x) = 0\})$.

Projected dynamical systems are a particular class of discontinuous dynamical systems. Let $K \subset \mathbb{R}^{n}$ be a closed convex set. Given a point $y \in \mathbb{R}^{n}$, the (point) projection of $y$ onto $K$ is $\text{proj}_{K}(y) = \arg\min_{z \in K} ||z - y||$. Note that $\text{proj}_{K}(y)$ is a singleton and the map $\text{proj}_{K}$ is Lipschitz on $\mathbb{R}^{n}$ with constant $L = 1$ [24, Proposition 2.4.1]. Given $x \in K$ and $v \in \mathbb{R}^{n}$, the (vector) projection of $v$ at $x$ with respect to $K$ is
\[
\Pi_{K}(x, v) = \lim_{\delta \to 0^{+}} \frac{\text{proj}_{K}(x + \delta v) - x}{\delta}.
\]

Given a vector field $f: \mathbb{R}^{n} \to \mathbb{R}^{n}$ and a closed convex polyhedron $K \subset \mathbb{R}^{n}$, the associated projected dynamical system is
\[
\dot{x} = \Pi_{K}(x, f(x)), \quad x(0) \in K.
\] (2)

One can verify easily that for any $x \in K$, there exists an element $\varphi_{x}$ belonging to the normal cone $N_{K}(x)$ such that $\Pi_{K}(x, f(x)) = f(x) - \varphi_{x}$. In particular, if $x$ is in the interior of $K$, then this element is the zero vector and we have $\Pi_{K}(x, f(x)) = f(x)$. At any boundary point of $K$, the projection operator restricts the flow of the vector field $f$ such that the solutions of (2) remain in $K$. Due to the projection, the dynamics (2) is in general discontinuous.

III. DATA-DRIVEN STOCHASTIC OPTIMIZATION

This section sets the stage for the formulation of our approach to deal with data-driven optimization in a distributed manner. The following material on data-driven stochastic optimization is taken from [3] and included here to provide a self-contained exposition. The reader familiar with these notions and tools can safely skip this section.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\xi$ be a random variable mapping this space to $(\mathbb{R}^{n}, B_{\mathbb{R}^{n}})$, where $B_{\mathbb{R}^{n}}$ is the Borel $\sigma$-algebra on $\mathbb{R}^{n}$. Let $\mathbb{P}$ and $\Xi \subset \mathbb{R}^{n}$ be the distribution and the support of the random variable $\xi$. Assume that $\Xi$ is closed and convex. Consider the stochastic optimization problem
\[
\inf_{x \in \mathcal{X}} \mathbb{E}_{\mathbb{P}}[f(x, \xi)],
\] (3)

where $\mathcal{X} \subset \mathbb{R}^{n}$ 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 the distribution $\mathbb{P}$. Assume that $\mathbb{P}$ is unknown and so, solving (3) is not possible. However, we are given $N$ independently drawn samples $\Xi := \{\xi(k)\}_{k=1}^{N} \subset \Xi$ of the random variable $\xi$. Note that, until it is revealed, $\Xi$ is a random object with probability distribution $\mathbb{P}^{N} := \prod_{k=1}^{N} \mathbb{P}$ supported on $\Xi^{N} := \prod_{k=1}^{N} \Xi$. The objective is to find a data-driven solution of (3), denoted $\hat{x}_{N} \in \mathcal{X}$, constructed using the dataset $\hat{\Xi}$, that has desirable properties for the expected cost $\mathbb{E}_{\mathbb{P}}[f(\hat{x}_{N}, \hat{\xi})]$ under a new sample. The property we are looking for is the finite-sample guarantee given by
\[
\mathbb{P}^{N} \left( \mathbb{E}_{\mathbb{P}}[f(\hat{x}_{N}, \hat{\xi})] \leq \hat{J}_{N} \right) \geq 1 - \beta,
\] (4)

where $\hat{J}_{N}$ might also depend on the training dataset and $\beta \in (0, 1)$ is the parameter which governs $\hat{x}_{N}$ and $\hat{J}_{N}$. The quantities $\hat{J}_{N}$ and $1 - \beta$ are referred to as the certificate and the reliability of the performance of $\hat{x}_{N}$. The goal is to find a data-driven solution with a low certificate and a high reliability. To do so, we use the available information $\hat{\Xi}$. The strategy is to determine a set $\hat{P}_{N}$ of probability distributions supported on $\Xi$ that contain the true distribution $\mathbb{P}$ with high confidence. The set $\hat{P}_{N}$ is referred to as the ambiguity set. Once such a set
is designed, the certificate \( \hat{J}_N \) is defined as the optimal value of the following distributionally robust optimization problem
\[
\hat{J}_N := \inf_{x \in X} \sup_{Q \in \tilde{P}_N} E_Q[f(x, \xi)].
\]  
(5)

Plugging this value in (9) yields \( \mathbb{P}^N(d_{W_2}(P, \tilde{P}_N) \geq \epsilon_N(\beta)) \leq \beta \). That is, if we let \( \hat{P}_N := B_{\epsilon_N(\beta)}(\tilde{P}_N) \), then
\[
\mathbb{P}^N(P \in \hat{P}_N) \geq 1 - \beta,
\]  
(11)
i.e., the true distribution belongs to the ambiguity set with probability at least \( 1 - \beta \). This leads us to the following result.

**Theorem III.2.** (Finite-sample guarantee of (5) with \( \hat{P}_N = B_{\epsilon_N(\beta)}(\tilde{P}_N) \)): For \( P \in \mathcal{M}_H(\Xi) \) and \( \beta \in (0, 1) \), let \( \hat{J}_N \) and \( \hat{J}_N \) be the optimal value and an optimizer of the distributionally robust optimization (5) with \( \hat{P}_N = B_{\epsilon_N(\beta)}(\tilde{P}_N) \). Then, the finite-sample guarantee (4) holds.

The proof follows by using (5) and (11) to yield (4). We end this section by discussing the tractability of solving (5) with \( \hat{P}_N = B_{\epsilon_N(\beta)}(\tilde{P}_N) \).

**Theorem III.3.** (Tractable reformulation of (5)): Assume that for all \( \xi \in \Xi \), the map \( x \mapsto f(x, \xi) \) is convex. Then, for any \( \beta \in (0, 1) \) and \( N \in \mathbb{Z}_{\geq 1} \), the optimal value of (5) with \( \hat{P}_N = B_{\epsilon_N(\beta)}(\tilde{P}_N) \) is equal to the optimum of the following convex optimization problem
\[
\inf_{\lambda \geq 0, x \in X} \left\{ \lambda \mathbb{E}_N^2(\beta) + \frac{1}{N} \sum_{k=1}^{N} \max_{\xi \in \Xi} \left( f(x, \xi) - \lambda \|\xi - \hat{\xi}_k\|^2 \right) \right\}.
\]

Theorem III.3 shows that under mild conditions on the objective function, one can reformulate the distributionally robust optimization problem as a convex optimization problem. This result plays a key role in our forthcoming discussion.

### IV. PROBLEM STATEMENT

Consider \( n \in \mathbb{Z}_{\geq 1} \) agents communicating over an undirected weighted graph \( G = (V, E, A) \). The set of vertices are enumerated as \( V := [n] \). Each agent \( i \in [n] \) can send and receive information from its neighbors \( N_i \) in \( G \). Let \( f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R} \) 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 \( 1_n \) and \( -1_n \) are not a direction of recession for the convex function \( x \mapsto f(x, \xi) \) is dense in \( \mathbb{R}^m \). As we progress, we stipulate additional conditions on \( f \) as necessary. Assume that all agents know the objective function \( f \). Given a random variable \( \xi \in \mathbb{R}^m \) with support \( \mathbb{R}^m \) and distribution \( P \), the original objective for the agents is to solve the following stochastic optimization problem
\[
\inf_{x \in \mathbb{R}^d} \mathbb{E}_P[f(x, \xi)].
\]  
(12)

However, agents cannot solve this problem as they do not know \( P \). Instead, each agent has a certain number (at least one) of independent and identically distributed realizations of the random variable \( \xi \). We denote the data available to agent \( i \) by \( \hat{\Xi}_i \) that is assumed to be nonempty. Assume that \( \hat{\Xi}_i \cap \hat{\Xi}_j = \emptyset \) for all \( i, j \in [n] \) and let \( \hat{\Xi} = \bigcup_{i=1}^{n} \hat{\Xi}_i \), containing \( N \) samples be the available data set.

The goal for the agents is then to collectively find, in a distributed manner, a data-driven solution \( \hat{x}_N \in \mathbb{R}^d \) to
approximate the optimizer of (12) with guaranteed performance bounds. To achieve this, we rely on the framework of distributionally robust optimization, cf. Section II-2. From Theorem III.3, a data-driven solution for (12) can be obtained by solving the following convex optimization problem

$$\inf_{\lambda, x} \left\{ \epsilon_N^2(\lambda) + \frac{1}{N} \sum_{k=1}^{N} \max_{\xi} \left( f(x, \xi) - \lambda \| \xi - \hat{\xi}^k \|^2 \right) \right\}$$

where $\beta \in (0, 1)$ and $\epsilon_N(\beta)$ is given in (10). We assume that there exists a finite optimizer $(x^*, \lambda^*)$ of (13), e.g., one of the conditions for existence of finite optimizers given in [26] is met. This optimizer has the following out-of-sample performance guarantee

$$\mathbb{P}^N \left( \mathbb{E}[f(x^*, \xi)] \leq \hat{J}_N \right) \geq 1 - \beta,$$

where $\hat{J}_N$ is the optimum value (13). Note that each agent can individually find a data-driven solution to (12) by using only the data available to it in the convex formulation (13). 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 (13) in a distributed manner, that is

(i) each agent $i$ has the information

$$I_i := \{ \mathbb{E}_i, f, a, c_i, e_i, A, \beta, n, N \},$$

where $a, c_i, e_i,$ and $A$ are parameters associated with the distribution $\mathbb{P}_i$, as defined in Section II-2 and $\beta \in (0, 1)$ is a parameter that agents agree upon beforehand.

(ii) each agent $i$ can only communicate with its neighbors $\mathcal{N}_i$ in the graph $G$.

(iii) each agent $i$ does not share with its neighbors any element of the dataset $\mathbb{E}_i$ available to it, and

(iv) there is no central coordinator or leader that can communicate with all agents.

The challenge in solving (13) in a distributed manner lies in the fact that the data is 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 (13) in a cooperative fashion is to let agents share their data with everyone in the network via some sort of flooding mechanism. This violates item (iii) of our definition of distributed algorithm given 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 execution of the algorithm.

Our strategy to tackle the problem is organized as follows: in Section V we reformulate the problem (13) to obtain a structure which allows us in Section VI to propose our distributed algorithm. Section VII discusses a class of objective functions $f$ for which the distributed algorithm provably converges.

V. DISTRIBUTED PROBLEM FORMULATION AND SADDLE POINTS

This section studies the structure of the optimization problem presented in Section IV with the ulterior goal of facilitating the design of a distributed algorithmic solution. Our first step is a reformulation of (13) that, by augmenting the decision variables of the agents, yields an optimization where the objective function is the aggregate of individual functions (that can be independently evaluated by the agents) and constraints which display a distributed structure. Our second step is the identification of a convex-concave function whose saddle points are the primal-dual optimizers of the reformulated problem under suitable conditions on the objective function $f$. This opens the way to consider the associated saddle-point dynamics as our candidate distributed algorithm. The structure of the original optimization problem makes this step particularly nontrivial.

A. Reformulation as distributed optimization problem

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

$$\min_{x_v, \lambda^v \geq 0} \ h(\lambda^v) + \frac{1}{N} \sum_{k=1}^{N} \max_{\xi} g_k(x^v_k, \lambda^v_k, \xi)$$

subject to $L \lambda^v = 0_n$, $L \otimes I_d) x_v = 0_{nd}$.

where $L \in \mathbb{R}^{n \times n}$ is the Laplacian of the graph $G$ and we have used the shorthand notation $h : \mathbb{R}^n \rightarrow \mathbb{R}$ for

$$h(\lambda^v) := \frac{\epsilon_N^2(\lambda)}{n}$$

and, for each $k \in [N]$, $g_k : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}$ for

$$g_k(x, \lambda, \xi) := f(x, \xi) - \lambda \| \xi - \hat{\xi}^k \|^2.$$

The following result establishes the correspondence between the optimizers of (13) and (15), respectively.

Lemma V.1. (One-to-one correspondence between optimizers of (13) and (15)): The following holds:

(i) If $(x^*, \lambda^*)$ is an optimizer of (13), then $(1_n \otimes x^*, \lambda^*1_n)$ is an optimizer of (15).

(ii) If $(x^*_v, \lambda^*_v)$ is an optimizer of (15), then there exists an optimizer $(x^*, \lambda^*)$ of (13) such that $x^*_v = 1_n \otimes x^*$ and $\lambda^*_v = \lambda^*1_n$.

Proof: The proof follows by noting that $G$ is connected and hence, (i) $L \lambda_v = 0_n$ if and only if $\lambda_v = \alpha 1_n$, $\alpha \in \mathbb{R}$; and (ii) $(L \otimes I_d) x_v = 0_{nd}$ if and only if $x_v = 1_n \otimes x, x \in \mathbb{R}^d$. Note that constraints (15b) and (15c) force agreement and that each of their components is computable by an agent of the
network using only local information. Moreover, the objective function (15a) can be written as \( \sum_{i=1}^{n} J_i(x^i, \lambda^i, \Xi^i) \), where 
\[
J_i(x^i, \lambda^i, \Xi^i) := \frac{c_i(x^i)^\lambda}{n} + \frac{1}{N} \sum_{k \in \Xi^i} \max_{x^k \in \Xi^i} g_k(x^i, \lambda^i, \xi),
\]
for all \( i \in [n] \). Therefore, the problem (15) 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 (15) with the following two properties: (i) it does not consist of an inner maximization, unlike the objective in (15a), and (ii) the primal-dual optimizers of (15) are saddle points of the newly introduced function. The availability of these two facts sets the stage for our ensuing algorithm design.

To proceed further, we first denote for convenience the objective function (15a) with \( F: \mathbb{R}^{nd} \times \mathbb{R}^n_\geq \rightarrow \mathbb{R} \),
\[
F(x_v, \lambda_v) := h(\lambda_v) + \frac{1}{N} \sum_{k=1}^{N} \max_{x^k \in \Xi^i} g_k(x^k_v, \lambda^k_v, \xi).
\]

Note that the Lagrangian of (15) is \( L: \mathbb{R}^{nd} \times \mathbb{R}^n_\geq \times \mathbb{R}^n \rightarrow \mathbb{R} \),
\[
L(x_v, \lambda_v, \nu, \eta) := F(x_v, \lambda_v) + \nu^T \lambda_v + \eta^T (L \otimes I_d)x_v,
\]
where \( \nu \in \mathbb{R}^n \) and \( \eta \in \mathbb{R}^{nd} \) are dual variables corresponding to the equality constraints (15b) and (15c), respectively. \( L \) is convex-concave in \( ((x_v, \lambda_v), (\nu, \eta)) \) on the domain \( \lambda_v \geq 0_n \).

The next result establishes important properties of the Lagrangian giving a correspondence between its saddle points and the optimizers of (15). We provide its proof, rather than establishing it as a consequence of results from the literature, because the objective (15a) takes values in the extended reals and in general might not be closed.

**Lemma V.2.** (Min-max equality for \( L \): The set of saddle points of \( L \) over the domain \( (\mathbb{R}^{nd} \times \mathbb{R}^n_\geq) \times (\mathbb{R}^n \times \mathbb{R}^{nd}) \) is nonempty and
\[
\inf_{x_v, \lambda_v \geq 0_n} \sup_{\nu, \eta} L(x_v, \lambda_v, \nu, \eta) = \sup_{\nu, \eta} \inf_{x_v, \lambda_v \geq 0_n} L(x_v, \lambda_v, \nu, \eta).
\]

Furthermore, the following holds:

(i) If \((\pi_v, \lambda_v)\) is a saddle point of \( L \) over \( (\mathbb{R}^{nd} \times \mathbb{R}^n_\geq) \times (\mathbb{R}^n \times \mathbb{R}^{nd}) \), then \((\pi_v, \lambda_v)\) is an optimizer of (15).

(ii) If \((\pi_v, \lambda_v)\) is an optimizer of (15), then there exists \((\bar{\nu}, \bar{\eta})\) such that \((\pi_v, \lambda_v, \bar{\nu}, \bar{\eta})\) is a saddle point of \( L \) over \( (\mathbb{R}^{nd} \times \mathbb{R}^n_\geq) \times (\mathbb{R}^n \times \mathbb{R}^{nd}) \).

**Proof:** Consider the left-hand side of (18). For a fixed \( x_v \) and \( \lambda_v \geq 0_n \), if \( L\lambda_v \neq 0_n \) or \( (L \otimes I_d)x_v \neq 0_{nd} \), then \( \sup_{\nu, \eta} L(x_v, \lambda_v, \nu, \eta) = +\infty \). Otherwise, \( \sup_{\nu, \eta} L(x_v, \lambda_v, \nu, \eta) = F(x_v, \lambda_v) \). That is,
\[
\sup_{\nu, \eta} L(x_v, \lambda_v, \nu, \eta) = \begin{cases} F(x_v, \lambda_v), & \text{if } L\lambda_v = 0_n \text{ and } (L \otimes I_d)x_v = 0_{nd}, \\ +\infty, & \text{otherwise}. \end{cases}
\]

By the definition of \( L \) and the assumption that there exists a finite optimizer of (13) and consequently that of (15), we get
\[
\inf_{x_v, \lambda_v \geq 0_n} \sup_{\nu, \eta} L(x_v, \lambda_v, \nu, \eta) = F(1_n \otimes x^*, \lambda^* 1_n),
\]
where \((x^*, \lambda^*)\) recall is an optimizer of (13). Now pick \((\bar{\nu}, \bar{\eta})\) such that \( L\bar{\nu} = 0_n \) and \( (L \otimes I_d)\bar{\eta} = 0_{nd} \). Then, one can show that \( \inf_{x_v, \lambda_v \geq 0_n} L(x_v, \lambda_v, \bar{\nu}, \bar{\eta}) = F(1_n \otimes x^*, \lambda^* 1_n) \). Thus, we deduce that
\[
\sup_{\nu, \eta} \inf_{x_v, \lambda_v \geq 0_n} L(x_v, \lambda_v, \nu, \eta) \geq F(1_n \otimes x^*, \lambda^* 1_n).
\]

Note that the following inequality holds (cf. [19, Lemma 36.11]),
\[
\inf_{x_v, \lambda_v \geq 0_n} \sup_{\nu, \eta} L(x_v, \lambda_v, \nu, \eta) \geq \sup_{\nu, \eta} \inf_{x_v, \lambda_v \geq 0_n} L(x_v, \lambda_v, \nu, \eta).
\]

Using the above inequality along with (19) and (20), we conclude that (18) holds. One can verify that the point \((1_n \otimes x^*, \lambda^* 1_n, \nu, \eta)\) with \( L\nu = 0_n \) and \((L \otimes I_d)\eta = 0_{nd}\) is a saddle point of \( L \) over the domain \( (\mathbb{R}^{nd} \times \mathbb{R}^n_\geq) \times (\mathbb{R}^n \times \mathbb{R}^{nd}) \). Hence, the set of saddle points is nonempty.

We proceed to establish (i) and (ii). Assume \((\pi_v, \lambda_v, \bar{\nu}, \bar{\eta})\) to be a saddle point of \( L \) over the considered domain. To arrive at a contradiction, assume that \((\pi_v, \lambda_v)\) is not an optimizer of (13). Recall that there exists a finite optimizer \((x^*, \lambda^*)\) of (13). This means that
\[
F(1_n \otimes x^*, \lambda^* 1_n) < F(\pi_v, \lambda_v).
\]

Note that \((L \otimes I_d)\pi_v = 0_{nd}\) and \(L\lambda_v = 0_n\) because otherwise \( \sup_{\nu} L(\pi_v, \lambda_v, \nu, \eta) = +\infty \), contradicting the fact that \((\pi_v, \lambda_v, \nu, \eta)\) is a saddle point. Therefore, using this property of \((\pi_v, \lambda_v)\) in (21), we obtain
\[
L(1_n \otimes x^*, \lambda^* 1_n, \bar{\nu}, \bar{\eta}) < L(\pi_v, \lambda_v, \bar{\nu}, \bar{\eta}).
\]

This is a contradiction as \((\pi_v, \lambda_v, \bar{\nu}, \bar{\eta})\) is a saddle point. Finally, assume \((\pi_v, \lambda_v)\) to be an optimizer of (13). Again, pick \((\bar{\nu}, \bar{\eta})\) such that \( L\bar{\nu} = 0_n \) and \((L \otimes I_d)\bar{\eta} = 0_{nd}\). One can show that \((\pi_v, \lambda_v, \bar{\nu}, \bar{\eta})\) is a saddle point of \( L \) over the specified domain, completing the proof.

Owing to the above result, one could potentially 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 in general not guaranteed to converge, see e.g., [27]. To overcome this hurdle, we augment the Lagrangian with quadratic terms in the primal variables. Let the augmented Lagrangian \( L_{aug} : \mathbb{R}^{nd} \times \mathbb{R}^n_\geq \times \mathbb{R}^n \times \mathbb{R}^{nd} \rightarrow \mathbb{R} \)
\[
L_{aug}(x_v, \lambda_v, \nu, \eta) := L(x_v, \lambda_v, \nu, \eta) + \frac{1}{2} x_v^T (L \otimes I_d)x_v + \frac{1}{2} \lambda_v^T L\lambda_v.
\]

Note that \( L_{aug} \) is also convex-concave in \((x_v, \lambda_v), (\nu, \eta))\) on the domain \( \lambda_v \geq 0_n \). The next result guarantees that this augmentation step does not change the saddle points.

**Lemma V.3.** (Saddle points of \( L \) and \( L_{aug} \) are the same): A point \((x^*_v, \lambda^*_v, \nu^*, \eta^*)\) is a saddle point of \( L \) over \((\mathbb{R}^{nd} \times \mathbb{R}^n_\geq) \times (\mathbb{R}^n \times \mathbb{R}^{nd}) \) 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 [28, Theorem 1.1]. The above result implies that finding the saddle points of \( L_{aug} \) would take us to the primal-dual optimizers of (15). However, a final roadblock remaining 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 \( \lambda_v = 0 \). Thus, our next task is to get rid of these internal optimization routines and identify a function for which the saddle-point dynamics is well defined over the feasible domain. Note that
\[
\begin{align*}
L_{aug}(x_v, \lambda_v, \nu, \eta) &= \max_{\{\xi\}} \tilde{L}_{aug}(x_v, \lambda_v, \nu, \eta, \{\xi\}) \\
&= \max_{\nu, \eta, \{\xi\}} \min_{(x_v, \lambda_v) \in C} \tilde{L}_{aug}(x_v, \lambda_v, \nu, \eta, \{\xi\}).
\end{align*}
\]
From the above equality and the fact that \( \tilde{L}_{aug} \) is convex-concave and finite-valued, we conclude from [19, Lemma 36.2] that the set of saddle points of \( \tilde{L}_{aug} \) over the domain \( C \times (R^n \times R^n \times mN) \) is nonempty. Further, this set is closed and convex again due to convexity-concavity of \( L_{aug} \). Finally, parts (ii) and (iii) follow from combining Lemmas V.2 and V.3 with the following two facts. First, from (25), if \((\tilde{\nu}, \tilde{\eta}, \tilde{\xi})\) is a saddle point of \( \tilde{L}_{aug} \), then \((\tilde{\nu}, \tilde{\eta}, \tilde{\xi})\) is a saddle point of \( L_{aug} \). Second, if \((\tilde{\nu}, \tilde{\eta}, \tilde{\xi})\) is a saddle point of \( L_{aug} \), then there exists \((\xi)\), which is the maximizer of \((\xi)\) → \( L_{aug}(\tilde{\nu}, \tilde{\eta}, \tilde{\xi}) \), such that \((\tilde{\nu}, \tilde{\eta}, \tilde{\xi})\) is a saddle point of \( L_{aug} \), completing the proof.

Section VII describes classes of objective functions for which the hypotheses of Proposition V.4 are met. We have introduced in Proposition V.4 the set \( C \) to increase the level of generality in preparation for the exposition of our algorithm that follows next. Specifically, since \( f \) is not necessarily convex-concave, the function \( L_{aug} \) might not be convex-concave over the entire domain \( (R^{nd} \times R^n) \times (R^n \times R^{nd}) \). For such cases, one can restrict the attention to the set \( C \times (R^n \times R^{nd} \times mN) \) provided the hypotheses of the above result are satisfied. As we show later, when the objective function \( f \) is convex-concave, one can employ the set \( C = R^{nd} \times R^n \).

VI. DISTRIBUTED ALGORITHM DESIGN AND CONVERGENCE ANALYSIS

Here we design and analyze our distributed algorithm to find the solutions of the optimization problem (13). Given the results of Section V, and specifically Proposition V.4, our algorithm seeks to find the saddle points of \( L_{aug} \) over the domain \( C \times (R^n \times R^{nd} \times mN) \). The dynamics consists of (projected) gradient-descent of \( L_{aug} \) in the convex variables and gradient-ascent in the concave ones. This is popularly termed as the saddle-point or the primal-dual dynamics [27, 29].

Given a closed, convex set \( C \subset R^{nd} \times R^n \), the saddle-point dynamics for \( L_{aug} \) is
\[
\frac{dx_v}{dt} = \Pi_C \left((x_v, \lambda_v), \left[ -\nabla_{x_v} L_{aug}(x_v, \lambda_v, \nu, \eta, \{\xi\}) \right] \right),
\]
\[
\frac{d\nu}{dt} = \nabla_{\nu} L_{aug}(x_v, \lambda_v, \nu, \eta, \{\xi\}),
\]
\[
\frac{d\eta}{dt} = \nabla_{\eta} L_{aug}(x_v, \lambda_v, \nu, \eta, \{\xi\}),
\]
\[
\frac{d\xi}{dt} = \nabla_{\xi} L_{aug}(x_v, \lambda_v, \nu, \eta, \{\xi\}), \forall k \in [N].
\]

For convenience, denote (26) by the vector field \( X_{dp} : R^{nd} \times R^n \times R^{nd+n+mN} \to R^{nd} \times R^{n} \times R^{nd+n+mN} \). In this notation, the first, second, and third components correspond to the dynamics of \( x_v \), \( \lambda_v \), and \((\nu, \eta, \{\xi\})\), respectively.

Remark VI.1. (Distributed implementation of \( X_{dp} \)): Here we discuss the distributed character of the dynamics (26). For
this, we rely on the set $C$ being decomposable into constraints on individual agent’s decision variables, i.e., $C := \Pi_{i=1}^n C_i$ with $C_i \subset \mathbb{R}_{\geq 0} \times \mathbb{R}^d$. This allows agents to perform the projection in (26a) in a distributed way (we show later that the set $C$ enjoys this structure for a broad class of objective functions $f$). Denote the components of the dual variables $\eta$ and $\nu$ 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 $\nu^i \in \mathbb{R}$. Further, let $K_i \subset [N]$ be the set of indices representing the samples held by $i$ ($k \in K_i$ if and only if $\tilde{\xi}^k \in \mathcal{X}_i$). For implementing $X_{sp}$, we assume that each agent $i$ maintains and updates the variables $(x^i, \nu^i, \eta^i, \xi^k)_{k \in K_i}$. The collection of these variables for all $i \in [n]$ forms $\{x_v, \lambda_v, \eta, (\xi^k)^i\}$. From (26), the dynamics of variables maintained by $i$ is

$$(\frac{dx^i}{dt}, \frac{d\lambda^i}{dt}) = \Pi_{C_i} \left( -\frac{1}{N} \sum_{k \in K_i} \nabla x g_k(x^i, \lambda^i, \xi^k) - \sum_{j \in N_i} a_{ij}(\eta^j - \eta^i) + (x^i - x^j) \right)$$

$$- \frac{c_i^\beta}{n} - \frac{1}{N} \sum_{k \in K_i} \nabla \lambda g_k(x^i, \lambda^i, \xi^k) - \sum_{j \in N_i} a_{ij}(\nu^j - \nu^i) + (\lambda^i - \lambda^j) ,$$

$d\nu^i = \sum_{j \in N_i} a_{ij}(\lambda^j - \lambda^i)$,

$d\lambda^i = \sum_{j \in N_i} a_{ij}(x^j - x^i)$,

$d\xi^k = \frac{1}{N} \nabla \xi g_k(x^i, \lambda^i, \xi^k)$, \quad $\forall k \in K_i$.

Observe that the right-hand side of the above dynamics is computable by agent $i$ using the variables that it maintains and information collected from its neighbors. Hence, $X_{sp}$ can be implemented in a distributed manner. Note that the number of variables in the set $\{\xi^k\}$, grows with the size of the data, whereas the size of all other variables is independent of the number of samples. Further, for any agent $i$, $\{\xi^k\}_{k \in K_i}$ can be interpreted as its internal state that is not communicated to its neighbors.

The following result establishes the convergence of the dynamics $X_{sp}$ to the saddle points of $\tilde{L}_{aug}$. In our previous work [27], we have extensively analyzed the convergence properties of saddle-point dynamics associated to convex-concave functions. However, those results do not apply directly to infer convergence for $X_{sp}$ because projection operators are involved in our algorithm design and the function $\tilde{L}_{aug}$ is not linear in the concave variable $\{\xi^k\}$ and not necessarily strictly convex in $(x_v, \lambda_v)$. Nonetheless, we borrow much insight from our previous analysis to prove the following result.

**Theorem VI.2.** (Convergence of trajectories of $X_{sp}$ to the optimizers of (15)): Suppose the hypotheses of Proposition VI.4 hold. Assume further that there exists a saddle point $(x_v^*, \lambda_v^*, \nu^*, \eta^*, \{\xi^k\}^*$) of $\tilde{L}_{aug}$ satisfying $(x_v^*, \lambda_v^*) \in \text{int}(C)$. Then, the trajectories of (26) starting in $C \times \mathbb{R}^{n} \times \mathbb{R}^{nd} \times \mathbb{R}^{mN}$ remain in this set and converge asymptotically to a saddle point of $\tilde{L}_{aug}$. As a consequence, the $(x_v, \lambda_v)$ component of the trajectory converges to an optimizer of (15).

**Proof:** We understand the trajectories of (26) in the Caratheodory sense, cf. Section II-4. Note that by definition of the projection operator, any solution $t \mapsto (x_v(t), \lambda_v(t), \nu(t), \eta(t), \{\xi^k(t)\})$ of (26) starting with $(x_v(0), \lambda_v(0)) \in C$ satisfies $(x_v(t), \lambda_v(t)) \in C$ for all $t \geq 0$.

**LaSalle function.** Let $(x_v^*, \lambda_v^*, \nu^*, \eta^*, \{\xi^k\}^*)$ be the equilibrium point of $\tilde{L}_{aug}$ satisfying $(x_v^*, \lambda_v^*) \in \text{int}(C)$. Using the definition of equilibrium point in (26b) and (26c), we get

$$(L \otimes I_d)x_v^* = 0_{nd} \text{ and } L\lambda_v^* = 0_n.$$  

(27)

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

$$V(x_v, \lambda_v, \zeta) := \frac{1}{2}(\|x_v - x_v^*\|^2 + \|\lambda_v - \lambda_v^*\|^2 + \|\zeta - \zeta^*\|^2),$$

where, for convenience, we use $\zeta := (\nu, \eta, \{\xi^k\})$ and, likewise, $\zeta^* := (\nu^*, \eta^*, \{\xi^k\}^*)$. The Lie derivative of $V$ along the dynamics (26) can be written as

$$L_{(26)}V(x_v, \lambda_v, \zeta) = M_1 + M_2,$$

where

$$M_1 = -(x_v - x_v^*)^T \nabla x \tilde{L}_{aug}(x_v, \lambda_v, \zeta),$$

$$- (\lambda_v - \lambda_v^*)^T \nabla \lambda \tilde{L}_{aug}(x_v, \lambda_v, \zeta) + (\zeta - \zeta^*)^T \nabla \zeta \tilde{L}_{aug}(x_v, \lambda_v, \zeta),$$

$$M_2 = -(x_v - x_v^*)^T \varphi_{x_v} - (\lambda_v - \lambda_v^*)^T \varphi_{\lambda_v},$$

and $(\varphi_{x_v}, \varphi_{\lambda_v})$ is an element of the normal cone $N_C(x_v, \lambda_v)$ (cf. Section II-3). This representation allows us to analyze the projection in the $(x_v, \lambda_v)$-dynamics separately. By definition of the normal cone, we get $M_2 \leq 0$. Recall that $\tilde{L}_{aug}$ is convex-concave in $(x_v, \lambda_v, \zeta)$. Thus, using the first-order convexity and concavity condition for these maps and following the computation as in [29, Proof of Lemma 4.1], we get

$$M_1 \leq \tilde{L}_{aug}(x_v^*, \lambda_v^*, \zeta) - \tilde{L}_{aug}(x_v, \lambda_v^*, \zeta^*) + \tilde{L}_{aug}(x_v^*, \lambda_v^*, \zeta^*) - \tilde{L}_{aug}(x_v, \lambda_v, \zeta^*).$$

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. Thus, $M_1 \leq 0$. Therefore, we conclude

$$L_{(26)}V(x_v, \lambda_v, \zeta) \leq 0.$$  

(28)

**Application of LaSalle invariance principle.** Using the property (28), we deduce two facts. First, given $\delta \geq 0$, any trajectory of (26) starting in $S_\delta := V^{-1}(\leq \delta) \cap (C \times \mathbb{R}^{nd+n+Nm})$ remains in $S_\delta$ at all times. In particular, every equilibrium point is stable under the dynamics. Second, the omega-limit set of each trajectory of (26) starting in $S_\delta$ is invariant under the dynamics. Thus, from the invariance principle for discontinuous dynamical systems, cf. Proposition II.2, any solution of (26) converges to the largest invariant set

$$\mathcal{M} \subset \{ (x_v, \lambda_v, \zeta) \mid L_{(26)}V(x_v, \lambda_v, \zeta) = 0, (x_v, \lambda_v) \in C \}. $$
Thus, plugging this equality in (32) and rearranging terms gives

\[
\sum_{k=1}^{N} g_k((x_v^*)^{v_k}, (\lambda_v^*)^{v_k}, (\xi^*)^k) = \sum_{k=1}^{N} g_k((x_v^*)^{v_k}, (\lambda_v^*)^{v_k}, (\xi^*)^k).
\]

Expanding the equality (a) and using (27), we obtain

\[
\sum_{k=1}^{N} g_k((x_v^*)^{v_k}, (\lambda_v^*)^{v_k}, (\xi^*)^k) = \sum_{k=1}^{N} g_k((x_v^*)^{v_k}, (\lambda_v^*)^{v_k}, (\xi^*)^k).
\]

Since \((x_v^*, \lambda_v^*) \in \text{int}(C)\) and \(C\) is a closed subset of \(\mathbb{R}^{nd} \times \mathbb{R}^{nu}\), we have \(\lambda_v^* > 0_n\). Further, from the saddle-point property, \(\{\xi^k\}\) maximizes the function \(\{\xi^k\} \mapsto \sum_{k=1}^{N} g_k((x_v^*)^{v_k}, (\lambda_v^*)^{v_k}, (\xi^*)^k)\). This map is strongly concave as \(\lambda_v^* > 0_n\) (which can be verified by checking that the Hessian with respect to \(\{\xi^k\}\) is negative definite). Therefore, (30) yields

\[
\xi^k = (\xi^*)^k, \quad \text{for all } k \in [N].
\]

Expanding the equality (b) in (29) and using (27), we get

\[
\text{h}(\lambda_v^*) + \frac{1}{N} \sum_{k=1}^{N} g_k((x_v^*)^{v_k}, (\lambda_v^*)^{v_k}, (\xi^*)^k) = h(\lambda_v^*)
\]

\[
+ \frac{1}{N} \sum_{k=1}^{N} g_k(x_v^{v_k}, \lambda_v^{v_k}, (\xi^*)^k) + (\nu^*)^T \lambda_v
\]

\[
+ (\eta^*)^T (L \otimes I_d)x_v + \frac{1}{2} x_v^T (L \otimes I_d)x_v + \frac{1}{2} \lambda_v^T L \lambda_v.
\]

For ease of notation, let \(y_v := (x_v; \lambda_v), y_v^* := (x_v^*; \lambda_v^*)\), and

\[
F(y_v) := h(\lambda_v) + \frac{1}{N} \sum_{k=1}^{N} g_k(x_v^{v_k}, \lambda_v^{v_k}, (\xi^*)^k).
\]

Then, the expression (31) can be written as

\[
F(y_v^*) = F(y_v) + (\nu^*)^T \lambda_v + (\eta^*)^T (L \otimes I_d)x_v
\]

\[
+ \frac{1}{2} y_v^T (L \otimes I_{d+1}) y_v.
\]

From the definition of saddle point, \((x_v^*, \lambda_v^*)\) minimizes the function \((x_v, \lambda_v) \mapsto \hat{L}_{\text{aug}}(x_v, \lambda_v, \xi^*)\) over the domain \(C\). Moreover, by assumption \((x_v^*, \lambda_v^*)\) lies in the interior of \(C\). Thus,

\[
\nabla_{x_v} \hat{L}_{\text{aug}}(x_v^*, \lambda_v^*, \xi^*) = 0_{nd},
\]

\[
\nabla_{\lambda_v} \hat{L}_{\text{aug}}(x_v^*, \lambda_v^*, \xi^*) = 0_n.
\]

The first of the above equalities yield

\[
(L \otimes I_d)\eta^* = -\nabla_{x_v} F(y_v^*).
\]

Plugging this equality in (32) and rearranging terms gives

\[
\frac{1}{2} y_v^T (L \otimes I_{d+1}) y_v = F(y_v^*) - F(y_v)
\]

\[
- (\nu^*)^T \lambda_v + x_v^T \nabla_{x_v} F(y_v^*).
\]

Note that \((x_v^*)^T \nabla_{x_v} F(y_v^*) = (x_v^*)^T (\nabla_{x_v} F(y_v^*) + (L \otimes I_d)\eta^* + (L \otimes I_d)x_v^*)\), where we have used (27). This in turn equals 0 because of (33a). Thus, we can rewrite (34) as

\[
\frac{1}{2} y_v^T (L \otimes I_{d+1}) y_v = F(y_v^*) - F(y_v)
\]

\[
- (\nu^*)^T \lambda_v + x_v^T \nabla_{x_v} F(y_v^*).
\]

Expanding (33b) gives

\[
\nabla_{\lambda_v} F(y_v^*) + L \nu^* + \frac{1}{2} \lambda_v^* = 0.
\]

Pre-multiplying the above equation with \((\lambda_v^*)^T\) and using (27), we get \((\lambda_v^*)^T \nabla_{\lambda_v} F(y_v^*) = 0\) and we can further rewrite (35) as

\[
\frac{1}{2} y_v^T (L \otimes I_{d+1}) y_v = F(y_v^*) - F(y_v) = - (\nu^*)^T \lambda_v
\]

\[
+ (x_v - x_v^*)^T \nabla_{x_v} F(y_v^*) - (\lambda_v^*)^T \nabla_{\lambda_v} F(y_v^*).
\]

Using (27) in (36) yields \(\nabla_{\lambda_v} F(y_v^*) = -L \nu^*\). That is, \(\lambda_v^T \nabla_{\lambda_v} F(y_v^*) = -\lambda_v^T L \nu^*\) which then replaced in (37) gives

\[
\frac{1}{2} y_v^T (L \otimes I_{d+1}) y_v = F(y_v^*) - F(y_v) + (y_v - y_v^*)^T \nabla_{y_v} F(y_v^*).
\]

The first-order convexity condition for \(F\) takes the form

\[
F(y_v) \geq F(y_v^*) + (y_v - y_v^*)^T \nabla_{y_v} F(y_v^*).
\]

Using the previous two expressions, we obtain

\[
y_v^* (L \otimes I_{d+1}) y_v \leq 0.
\]

This is only possible if this expression is zero because \((L \otimes I_{d+1}) y_v\) is positive semidefinite. Equating it to zero, we get \(x_v = 1_n \otimes x\) and \(\lambda_v = \lambda 1_n\) for some \((x, \lambda)\) and \((x, \lambda_v) \in C\). Collecting our derivations so far, we have that if \((x_v, \lambda_v, \xi) \in \mathcal{M}\), then

\[
\xi^k = (\xi^*)^k, \quad \forall k \in [N], \quad x_v = 1_n \otimes x,
\]

\[
\lambda_v = \lambda 1_n, \quad (x, \lambda_v) \in C.
\]

Identification of the largest invariant set. Consider a trajectory \(t \mapsto (x_v(t), \lambda_v(t), \xi(t))\) of (26) starting at \((x_v(0), \lambda_v(0), \xi(0)) \in \mathcal{M}\), and remaining in \(\mathcal{M}\) at all times (recall that \(\mathcal{M}\) is invariant). Then, the trajectory must satisfy (38) for all \(t \geq 0\), that is, there exists \(t \mapsto (x(t), \lambda(t))\) such that

\[
\xi^k(t) = (\xi^*)^k, \quad \forall k \in [N], \quad x_v(t) = 1_n \otimes x(t),
\]

\[
\lambda_v(t) = \lambda(t) 1_n, \quad (x_v(t), \lambda_v(t)) \in C.
\]

for all \(t \geq 0\). Plugging (39) in (26), we obtain that for all \(t \geq 0\), along the considered trajectory, we have \(\dot{v}(t) = 0_n, \dot{\eta}(t) = 0_n\), and \(\dot{\xi}(t) = 0_m\). This implies that the considered trajectory satisfies the following for all \(t \geq 0\),

\[
\begin{bmatrix}
\frac{dx_v(t)}{dt} \\
\frac{d\lambda_v(t)}{dt} \\
\end{bmatrix} = \Pi_C \begin{bmatrix}
x_v(t), \lambda_v(t), \xi(t) \\
\nabla_{x_v} \hat{L}_{\text{aug}}(x_v(t), \lambda_v(t), \xi(t)) \\
\nabla_{\lambda_v} \hat{L}_{\text{aug}}(x_v(t), \lambda_v(t), \xi(t)) \\
\end{bmatrix}
\]

which is a gradient descent dynamics of the convex function \((x_v, \lambda_v) \mapsto \hat{L}_{\text{aug}}(x_v, \lambda_v, \xi(0))\) projected over the set \(C\). Thus, either \(t \mapsto \hat{L}_{\text{aug}}(x_v(t), \lambda_v(t), \xi(0))\) decreases at some \(t\) or the
right-hand side of the above dynamics is zero at all times. Note that for all \( t \geq 0 \),
\[
\hat{L}_{\text{aug}}(x_v(t), \lambda_v(t), \zeta(0)) \overset{(e)}{=}
\hat{L}_{\text{aug}}(1_n \otimes x(t), \lambda(t)1_n, \zeta(0))
\]
\[
\overset{(b)}{=}
h(\lambda(t)1_n) + \frac{1}{N} \sum_{k=1}^{N} g_k(1_n \otimes x(t), \lambda(t)1_n, (\xi^k)^k)
\]
\[
\overset{(c)}{=}
\hat{L}_{\text{aug}}(1_n \otimes x(t), \lambda(t)1_n, \zeta^*) \overset{(d)}{=}
\hat{L}_{\text{aug}}(x_v^*, \lambda_v^*, \zeta^*).
\]
In the above set of expressions, equalities (a), (b), and (c) follow from conditions (39) and the definition of \( \hat{L}_{\text{aug}} \). Equality (d) follows from (29), which holds from every point in \( \mathcal{M} \). The above implies that \( t \mapsto \hat{L}_{\text{aug}}(x_v(t), \lambda_v(t), \zeta(0)) \) is a constant map. As a consequence, we conclude that \((x_v(0), \lambda_v(0), \zeta(0))\) is an equilibrium point of (26). Therefore, we have proved that the set \( \mathcal{M} \) is entirely composed of the equilibrium points of the dynamics (26). Convergence to an equilibrium point in the set of saddle points for each trajectory follows from this and the fact that each equilibrium point is stable, cf. [30].

\[ \]

**Remark VI.3.** (Convergence of algorithm for nonsmooth objective functions): Let \( f \) satisfy all assumptions outlined in Section IV except the differentiability and instead assume it is locally Lipschitz. This implies that the gradient of \( \hat{L}_{\text{aug}} \) with respect to variables \( x_v \) and \( \{\xi^k\} \) need not exist everywhere. However, the generalized gradients exist, see e.g., [21] for the definition. Therefore, one can replace gradients in (26a) and (26d) with the generalized counterparts and end up with a differential inclusion for the \( \{\xi^k\} \) dynamics and a projected differential inclusion for the \( x_v \) dynamics. Although we do not explore it here, we believe that, using analysis tools of nonsmooth dynamical systems, see [21] and references therein, one can show that the trajectories of the resulting nonsmooth dynamical system retain the convergence properties of Theorem VI.2. A promising route to establish this is to follow the exposition of [31], which studies saddle-point dynamics for a general class of functions.

**Remark VI.4.** (Constrained stochastic optimization): Certain constrained stochastic optimization problems can be cast in the form (12) and are therefore amenable to the distributed algorithmic solution techniques developed here. Given \( \delta \in (0, 1) \) and a measurable map \( F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \), consider the following constrained stochastic optimization problem
\[
\inf_{x \in \mathbb{R}^d} \{ F(x, \xi) \leq 0 \} \mathbb{P}_x \left[ f(x, \xi) \right]. \tag{40}
\]
The constraint is probabilistic in nature and so is commonly referred to as chance constraint [2]. One approach to solve this problem is to remove the constraint and add a convex function to the objective that penalizes its violation. Conditional value-at-risk (CVaR) is one such penalizing function. Formally, the CVaR of \( \xi \mapsto F(x, \xi) \) at level \( \delta \) is
\[
\text{CVaR}_\delta(F(x, \xi)) := \inf_{\tau \in \mathbb{R}} \mathbb{E}_\xi \left[ \tau + \frac{1}{\delta} \max \{ F(x, \xi) - \tau, 0 \} \right].
\]
Roughly speaking, this value represents the expectation of \( \xi \mapsto F(x, \xi) \) over the set of \( \xi \) that has measure \( \delta \) and that contain the highest values of this function. Note the fact [2, Chapter 6] that \( \text{CVaR}_\delta(F(x, \xi)) \leq 0 \) implies \( \mathbb{P}(F(x, \xi) \leq 0) \geq 1 - \delta \). Thus, using CVaR, problem (40) can be approximated by
\[
\inf_{x \in \mathbb{R}^d} \mathbb{E}_\xi \left[ f(x, \xi) \right] + \rho \text{CVaR}_\delta(F(x, \xi)),
\]
where \( \rho > 0 \) determines the trade-off between the two goals: minimizing the objective and satisfying the constraint. By invoking the definition of CVaR, the above problem can be written compactly as
\[
\inf_{x \in \mathbb{R}^d, \tau \in \mathbb{R}} \mathbb{E}_\xi \left[ f(x, \xi) + \rho(\tau + \frac{1}{\delta} \max \{ F(x, \xi) - \tau, 0 \} \right].
\]
This can be further recast as a stochastic optimization of the form (12). Therefore, under appropriate conditions on the function \( F \), one can solve a chance-constrained problem in a distributed way under the data-driven optimization paradigm using the algorithm design introduced here.

**VII. OBJECTIVE FUNCTIONS THAT MEET THE ALGORITHM CONVERGENCE CRITERIA**

In this section we report on two broad classes of objective functions \( f \) for which the hypotheses of Proposition V.4 hold. For both cases, we justify how the dynamics (26) serves as the distributed algorithm for solving (15).

A. Convex-concave functions

Here we focus on objective functions that are convex-concave in \((x, \xi)\). That is, 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 \( x \in \mathbb{R}^n \). We proceed to check the hypotheses of Theorem VI.2. To this end, let \( \mathcal{C} = \mathbb{R}^m \times \mathbb{R}^m \geq 0 \), which is closed, convex set with \( \text{int}(\mathcal{C}) \neq 0 \). Note that \( \hat{L}_{\text{aug}} \) is convex-concave on \( \mathcal{C} \times (\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m N) \) as \( f \) is convex-concave. The following result shows that (24) holds.

**Lemma VII.1.** (Min-max operators can be interchanged for \( \hat{L}_{\text{aug}} \): Let \( f \) be convex-concave in \((x, \xi)\). Then, for any \((\nu, \eta) \in \mathbb{R}^n \times \mathbb{R}^m \), the following holds
\[
\min_{x, \lambda_v \geq 0} \max_{\{\xi^k\}} \hat{L}_{\text{aug}}(x_v, \lambda_v, \nu, \eta, \{\xi^k\}) = \max_{\{\xi^k\}} \min_{x, \lambda_v \geq 0} \hat{L}_{\text{aug}}(x_v, \lambda_v, \nu, \eta, \{\xi^k\}). \tag{41}
\]

**Proof:** Given any \((\nu, \eta)\), denote the function \((x_v, \lambda_v, \{\xi^k\}) \mapsto \hat{L}_{\text{aug}}(x_v, \lambda_v, \nu, \eta, \{\xi^k\}) \) by \( \hat{L}_{\text{aug}}^{(\nu, \eta)} \). Since \( f \) is convex-concave, so is \( \hat{L}_{\text{aug}}^{(\nu, \eta)} \) in the variables \((x_v, \lambda_v, \{\xi^k\})\). We use Theorem VI.1 to prove the result. To do so, let us extend \( \hat{L}_{\text{aug}}^{(\nu, \eta)} \) over the entire domain \((\mathbb{R}^m \times \mathbb{R}^m) \times (\mathbb{R}^m N) \) as
\[
\hat{L}_{\text{aug}}^{(\nu, \eta)}(x_v, \lambda_v, \{\xi^k\}) = \begin{cases} 
\hat{L}_{\text{aug}}^{(\nu, \eta)}(x_v, \lambda_v, \{\xi^k\}), & \text{if } \lambda_v \geq 0, \\
+\infty, & \text{otherwise}.
\end{cases}
\]

One can see that \( \hat{L}_{\text{aug}}^{(\nu, \eta)} \) is closed, proper, and convex-concave (cf. Section II for definitions). Further, following [19, Theorem 36.3], the equality (41) holds if and only if the following holds
\[
\min_{x_v, \lambda_v \left\{ \xi^k \right\} \text{CVaR}_\delta(F(x, \xi))} \leq 0. \]

The rest of the proof establishes the above condition by checking the hypotheses of Theorem II.1 for $\mathcal{T}_{\text{aug}}$. For showing Theorem II.1(i), it is enough to identify $\{\xi_k\} \in \mathbb{R}^{m,N}$ for which the function $(x_v, \lambda_v) \mapsto \mathcal{T}_{\text{aug}}(x_v, \lambda_v, \{\xi_k\})$ does not have a direction of recession. By the assumptions on $f$, for each $k \in [N]$, there exists $\tilde{\xi}_k^* \in B_{\tilde{\xi}_k^*}(\beta)/\sqrt{\tilde{\xi}_k^*}$ such that $\mathcal{T}_{\text{aug}}(x_v, \lambda_v, \{\tilde{\xi}_k^*\})$ does not have a direction of recession. By the above expression, one has $\|\xi_k - \tilde{\xi}_k^*\|^2 \leq \mathcal{T}_{\text{aug}}(x_v, \lambda_v, \{\xi_k\})$. Picking these values, one has $\min_{\xi_k \in \mathbb{R}^m} \mathcal{T}_{\text{aug}}(x_v, \lambda_v, \{\xi_k\}) = \mathcal{T}_{\text{aug}}(x_v, \lambda_v, \{\xi_k\})$.

Thus, $\mathcal{T}_{\text{aug}}(x_v, \lambda_v, \{\xi_k\}) = e_N^2(\beta) + \frac{1}{n} \sum_{k=1}^N f(x_k, \tilde{\xi}_k^*) - \|\xi_k - \tilde{\xi}_k^*\|^2$.

Recall that for any $x \in \mathbb{R}^d$, $\xi \mapsto f(x, \xi)$ is concave. Hence, we deduce from the above expression that $\mathcal{T}_{\text{aug}}(x_v, \lambda_v, \{\xi_k\}) \to -\infty$ as $\|\{\xi_k\}\| \to \infty$. Therefore, $\{\xi_k\} \mapsto -\mathcal{T}_{\text{aug}}(x_v, \lambda_v, \{\xi_k\})$ does have a direction of recession, completing the proof.

As a consequence of the above discussion, we conclude that the hypotheses of Proposition V.4 hold true for the considered class of objective functions, and we can state, invoking Theorem VI.2, the following convergence result.

**Corollary VII.2.** (Convergence of trajectories of $\mathcal{X}_{\text{sp}}$ for convex-concave $f$): Let $f$ be convex-concave in $(x, \xi)$ and $\mathcal{C} = \mathbb{R}^d \times \mathbb{R}^{n_d}$. Assume further that there exists a saddle point $(x_{\text{sp}}, \lambda_{\text{sp}}, v^*, \nu^*, \{((\xi_k)^*)\})$ of $\mathcal{L}_{\text{aug}}$ satisfying $\lambda_{\text{sp}} > 0$. Then, the trajectories of (26) starting in $\mathcal{C} \times \mathbb{R}^d \times \mathbb{R}^{n_d} \times \mathbb{R}^{m_N}$ remain in this set and converge asymptotically to a saddle point of $\mathcal{L}_{\text{aug}}$. As a consequence, the $(x_v, \lambda_v)$ component of the trajectory converges to an optimizer of (15).

It is important to note that $\mathcal{C} = \Pi_{i=1}^n (\mathbb{R}^d \times \mathbb{R}_{\geq 0})$ and thus the projection in (26a) can be executed by individual agents. Following Remark VI.1, the dynamics (26) is implementable in a distributed way.

**B. Convex-concave 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 $\xi$. We further divide this classification into two.

1) **Quadratic function in $\xi$:** Assume additionally that the function $f$ is of the form

$$f(x, \xi) := \xi^T \mathcal{Q} \xi + x^T R \xi + \ell(x),$$

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

**Lemma VII.3.** (Characterizing where the objective function of (15) is finite-valued): Assume $f$ is of the form (42). Then, the function $F$ defined in (16) is finite-valued if and only if $\lambda^i \geq \lambda_{\text{max}}(Q)$ for all $i \in [n]$.

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

$$g_k(x_{\tilde{i}}, \lambda_{\tilde{i}}, \xi) = \xi^T (\mathcal{Q} - \mathcal{Q}_{\tilde{i}}) \xi + (x_{\tilde{i}})^T R \xi + 2\lambda^i (\mathcal{Q}_{\tilde{i}})^T \xi \xi + \ell(x_{\tilde{i}}) - \lambda^i \|\tilde{\xi}_k^*\|^2.$$ 

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

$$g_k(x_{\tilde{i}}, \lambda_{\tilde{i}}, \alpha w_{\text{max}}(Q)) = \alpha^2 (\lambda_{\text{max}}(Q) - \lambda^i) \|w_{\text{max}}(Q)\|^2 + \alpha \left((x_{\tilde{i}})^T R + 2\lambda^i (\mathcal{Q}_{\tilde{i}})^T \right) w_{\text{max}}(Q) \ell(x_{\tilde{i}}) - \lambda^i \|\tilde{\xi}_k^*\|^2.$$

Thus, we get $\max_k g_k(x_{\tilde{i}}, \lambda_{\tilde{i}}, \alpha w_{\text{max}}(Q)) = +\infty$ and so $\max_{\xi} g_k(x_{\tilde{i}}, \lambda_{\tilde{i}}, \xi) = +\infty$. Further note that for any $i$ and $k$ with $\tilde{\xi}_k^* \in \mathbb{R}^m$, $\max_{\lambda} g_k(x_{\tilde{i}}, \lambda_{\tilde{i}}, \xi) > -\infty$. This implies that $\sum_{k=1}^n \max_{\lambda} g_k(x_{\tilde{i}}, \lambda_{\tilde{i}}, \xi) = +\infty$ and so $F(x_{\tilde{i}}, \lambda_{\tilde{i}}) = +\infty$.

Now assume that $\lambda^i \geq \lambda_{\text{max}}(Q)$ for all $i \in [n]$. Then, for each $k$, evaluating the Hessian of $g_k$ with respect to $\xi$ leads to the conclusion that $g_k$ is a concave function of $\xi$.

Therefore, $\max_{\xi} g_k(x_{\tilde{i}}, \lambda_{\tilde{i}}, \xi) = \text{finite-valued}$ for each $k$. Hence, $F(x_{\tilde{i}}, \lambda_{\tilde{i}}) < +\infty$, completing the proof.

The above result implies that the optimizers of (15) for objective functions of the form (42) belong to the domain

$$\mathcal{C} := \mathbb{R}^{d \times \mathbb{R}_{\geq 0}} \times \mathbb{R}^{d \times \mathbb{R}_{\geq 0}} \times (\mathbb{R}^d \times \mathbb{R}^d).$$

Therefore, the saddle points of $\mathcal{L}_{\text{aug}}$ over the domain $(\mathbb{R}^{d \times \mathbb{R}_{\geq 0}} \times (\mathbb{R}^d \times \mathbb{R}^d))$ are contained in the set $\mathcal{C} \times (\mathbb{R}^d \times \mathbb{R}^d)$.

**Corollary VII.4.** (Convergence of trajectories of $\mathcal{X}_{\text{sp}}$ for quadratic $f$): Let $f$ be of the form (42) and $\mathcal{C}$ be given in (43). Assume further that there exists a saddle point $(x_{\text{sp}}, \lambda_{\text{sp}}, v^*, \nu^*, \{(\xi_k)^*)\})$ of $\mathcal{L}_{\text{aug}}$ satisfying $\lambda_{\text{sp}} > \lambda_{\text{max}}(Q)$. Then, the trajectories of (26) starting in $\mathcal{C} \times (\mathbb{R}^d \times \mathbb{R}^d) \times \mathbb{R}^{m_N}$ remain in this set and converge asymptotically to a saddle point of $\mathcal{L}_{\text{aug}}$. As a consequence, the $(x_v, \lambda_v)$ component of the trajectory converges to an optimizer of (15).
Note that $C$ given in (43) can be written as $C = \Pi_{i=1}^n (\mathbb{R}^d \times \{ \lambda \in \mathbb{R}_{\geq 0} \mid \lambda \geq \lambda_{\text{max}}(Q) \})$. Thus, following Remark VI.1, the dynamics (26) for this case 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

$$ f(x, \xi) := a((\xi_{m} - (\xi_{1:m-1}; 1)^T x))^2, \quad (44) $$

where $a > 0$ and $\xi_{1:m-1}$ denotes the vector $\xi$ without the last component $\xi_m$. Note that $f$ corresponds to the objective function for a least-squares problem. Further, note that it cannot be written in the form (42), as can be seen from its equivalent expression

$$ f(x, \xi) = a((\xi^T(-x_{1:m-1}; 1)(-x_{1:m-1}; 1)^T \xi - 2x_m(-x_{1:m-1}; 1)^T \xi + x_m^2)). $$

Our first step is to characterize, similarly to Lemma VII.3, the set over which the objective function (16) takes finite values.

**Lemma VII.5.** (Characterizing where the objective function of (15) is finite-valued): Assume $f$ is of the form (44). Then, the function $F$ defined in (16) is finite-valued if and only if $\lambda^i \geq a ||(x_{1:m-1}; 1)||^2$ for all $i \in [n]$.\hfill

**Proof:** The proof mimics the steps of the proof of Lemma VII.3. Assume there exists $i \in [n]$ such that $\lambda^i < a \|(x_{1:m-1}; 1)\|^2$. For any $k$ such that $\xi^k \in \Xi$, we have

$$ g_k(x^i, \lambda^i, a((x_{1:m-1}; 1)^T x^i))^2 - \lambda^i ||\xi - \hat{\xi}^k||^2. $$

Parametrizing $\xi = (x_{1:m-1}; 1)$, we obtain

$$ g_k(x^i, \lambda^i, a((x_{1:m-1}; 1)^T x^i)) = a(\lambda^i ||(x_{1:m-1}; 1)||^2 - x_m^2) - \lambda^2 ||(x_{1:m-1}; 1) - \hat{\xi}^k||^2. $$

This is a quadratic function in the parameter $\alpha$ and the coefficient of the second-order term is $||(-x_{1:m-1}; 1)||^2-a(((-x_{1:m-1}; 1)||^2 - \lambda^i)$. This coefficient is positive by the assumption stipulated above. Therefore, $\max_{\xi} g_k(x^i, \lambda^i, \xi) = +\infty$. Since $\max_{\xi} g_k(x^{w_k}, \lambda^{w_k}, \xi) > -\infty$ for all $k \in [N]$, we conclude that $F(x^i, \lambda^i) = +\infty$.

On the other hand, if $\lambda^i \geq a \|(x_{1:m-1}; 1)||^2$ for all $i \in [n]$, then $g_k(x^{i}, \lambda^{i}, \xi)$ is a concave function of $\xi$ for all $k \in [N]$. Thus, $\max_{\xi} g_k(x^{w_k}, \lambda^{w_k}, \xi)$ is finite-valued for each $k$ and so is $F(x^i, \lambda^i) < +\infty$.

Guided by the above result, let

$$ C := \mathbb{R}^{nd} \times \{ \lambda_n \in \mathbb{R}_{\geq 0} \mid \lambda^i \geq a \|(x_{1:m-1}; 1)||^2 \}. \quad (45) $$

As a consequence of Lemma VII.5, the optimizers of (15) belong to $C$ and so the saddle points of $L_{\text{aug}}$ over the domain $(\mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}) \times (\mathbb{R}^n \times \mathbb{R}^{nd})$ are contained in the set $C \times (\mathbb{R}^n \times \mathbb{R}^{nd})$. Further, $C$ is closed, convex with a nonempty interior and the function $L_{\text{aug}}$ is convex-concave on $C \times (\mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN})$. Finally, one can show that (24) holds in this case. Using these facts in Theorem VI.2 yields the following result.

**Corollary VII.6.** (Convergence of trajectories of $X_{\text{aug}}$ for least squares problem): Let $f$ be of the form (44) and $C$ be given in (45). Assume further that there exists a saddle point $(x^*_v, \lambda^*_v, v^*, \eta^*, \{(\xi^k)^v\}^*)$ of $L_{\text{aug}}$ satisfying $\{x^*_v, \lambda^*_v\} \in \text{int}(C)$. Then, the trajectories of (26) starting in $C \times \mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN}$ remain in this set and converge asymptotically to a saddle point of $L_{\text{aug}}$. As a consequence, the $(x_v, \lambda_v)$ component of the trajectory converges to an optimizer of (15).

In this case too, the saddle-point dynamics (26) is amenable to distributed implementation, cf. Remark VI.1, as one can write $C = \Pi_{i=1}^n \{ (x, \lambda) \in \mathbb{R}^d \times \mathbb{R}_{\geq 0} \mid \lambda \geq a \|(x_{1:m-1}; 1)||^2 \}$.

**VIII. SIMULATIONS**

Here we illustrate the application of the distributed algorithm (26) to find a data-driven solution for the regression problem with quadratic loss function and an affine predictor [32, Chapter 3], commonly termed as the least-squares problem. Assume $n = 10$ agents with communication topology defined by an undirected ring with additional edges $(1, 4), (2, 5), (3, 7), (6, 10)$. The weight of each edge is equal to one. We consider data points of the form $(\xi^k, (\bar{w}_k, g_k))$ consisting of the input $\bar{w}_k \in \mathbb{R}^2$ and the output $g_k \in \mathbb{R}^2$. The objective is to find an affine predictor $x \in \mathbb{R}^2$ using the dataset such that, ideally, for any new data point $(\xi, (w, y))$, the predictor $x^\top (w; 1)$ is equal to $y$. One way of finding such a predictor $x$ is to solve the following problem

$$ \inf_{x} \mathbb{E}_{P}[f(x, w, y)] \quad (46) $$

where $\mathbb{P}$ is the probability distribution of the data $(w, y)$ and $f : \mathbb{R}^5 \times \mathbb{R}^2 \rightarrow \mathbb{R}$ is the quadratic loss function, i.e., $f(x, w, y) = (x^\top (w; 1) - y)^2$, corresponding to the case considered in Section VII-B2.

To find the data-driven solution, we assume that each agent in the network has 30 i.i.d samples of $(w, y)$ and hence $N = 300$ is the total number of samples. The dataset is generated by assuming the input vector $w$ having a standard multivariate normal distribution, that is, zero mean and covariance as the identity matrix $I_4$. The output $y$ is assigned values $y = [1, 4, 3, 2] * w + v$ where $v$ is a random variable, uniformly distributed over the interval $[-1, 1]$. This defines completely the distribution $\mathbb{P}$ of $(w, y)$. Let $\beta \in (0, 1)$ such that $\epsilon_{\beta}(0) = 0.05$. This value is assumed to be computed by the agents beforehand. This defines completely the distributed optimization problem (15).

Figure 1 shows the execution of the distributed algorithm (26) that solves this problem. The trajectories converge to an equilibrium of the dynamics (26) whose $(x_v, \lambda_v)$ component corresponds to an optimizer of (15), consistent with Corollary VII.6.

To evaluate the quality of the obtained solution, we compute the average value of the loss function $f$ for a randomly generated validation dataset consisting of $N_{\text{val}} = 10^4$ data points $\{(\bar{w}_{\text{val}}, g_{\text{val}}))_{k=1}^N\}$. These points are i.i.d with the same distribution as that of the training dataset generated above. Given the obtained solution $(x_{1,10} \otimes x^*, \lambda^* 1_{10})$, see Figure 1.
we evaluate
\[ f_{\text{val}}^N(x^*) = \frac{1}{N_{\text{val}}} \sum_{k=1}^{N_{\text{val}}} f(x^*, \tilde{w}^{k}_{\text{val}}, \tilde{y}^{k}_{\text{val}}) \] (47)
and get \( f_{\text{val}}^N(x^*) = 0.3387 \). This is the average loss for the solution \( x^* \) obtained by the agents cooperating with each other, essentially fusing the information of the 300 data points. Note that each agent individually can also solve a data-driven solution with the samples gathered by it. However, the solution obtained in such a manner, in general, incurs a higher average loss. In the current setup, if agent 1 solves (13) only with the data available to it (and keeping other parameters equal), then it gets the optimizer as \( x_{\text{opt},1} = (0.8548; 3.8933; 2.8623; 2.1317; 0.2227) \). Using the validation dataset, we obtain \( f_{\text{val}}^N(x_{\text{opt},1}) = 0.4520 \), which is significantly greater than \( f_{\text{val}}^N(x^*) \). This shows the value of cooperation, that is, fusing the information contained in the data available to different agents leads to an optimizer with better out-of-sample performance. To highlight this fact further, Figure 2 shows the effect of the number of cooperating agents on the average loss incurred by the obtained solution to the data-driven optimization problem. As the plot shows, the improvement in performance due to coordination becomes more prominent as the size of the coordination agents grows.

IX. CONCLUSIONS

We have considered a cooperative stochastic optimization problem, where a group of agents rely on their individually collected data to collectively determine a data-driven solution with guaranteed out-of-sample performance. Our technical approach has proceeded by first developing a reformulation in the form of a distributed optimization problem, leading us to the identification of an augmented Lagrangian function whose saddle points have a one-to-one correspondence with the primal-dual optimizers. This characterization relies upon certain interchangeability properties between the min and max operators. Our discussion has identified several classes of objective functions for which these properties hold: convex-concave functions, convex-convex functions quadratic in the data, and convex-convex functions associated to least-squares problems. Building on the analytical results, we have designed a distributed saddle-point coordination algorithm where agents share their individual estimates about the solution, not the collected data. We have also formally established the asymptotic convergence of the algorithm to the solution of the cooperative stochastic optimization problem. Future work will explore the characterization of the algorithm convergence rate, the design of strategies capable of tracking the solution of the stochastic optimization problem when new data becomes available in an online fashion, and the analysis of scenarios with network chance constraints.

REFERENCES


