Data-driven distributed optimization using Wasserstein ambiguity sets

Ashish Cherukuri Jorge Cortés

Abstract— **This paper considers a general class of stochastic optimization problem for multiagent systems. We assume that the probability distribution of the uncertain parameters is unknown to the agents and instead, each agent gathers a certain number of samples of it. The objective for the agents is to cooperatively find, using the available data, a solution that has performance guarantees for the stochastic problem. To this end, we formulate a data-driven distributionally robust optimization (DRO) problem using Wasserstein ambiguity sets that has the desired performance guarantees. With the aim of solving this optimization in a distributed manner, we identify a convex-concave modified Lagrangian function whose saddle points are in correspondence with the optimizers of the DRO problem. We then design our distributed algorithm as the gradient descent in the convex variable and gradient ascent in the concave variable of this Lagrangian function. Our convergence analysis shows that the trajectories of this dynamics converge asymptotically to an optimizer of the DRO problem. Simulations illustrate our results.**

I. Introduction

Stochastic optimization for multiagent systems finds many applications, for example, distributed estimation and target tracking. For these settings, often, the probability distribution of the random variable is not known. Instead, agents gather instantiations of the variable and use it to find the solution of the stochastic optimization. When the dataset is large, machine learning algorithms are able to find the optimizer. 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 decisions must be taken in real time, leaving less room for gathering many samples. For these cases, distributionally robust optimization (DRO) uses the finite dataset to provide a solution that has desirable out-of-sample performance guarantees. Motivated by this, we consider the task for a group of agents to collaboratively find a data-driven solution for a stochastic optimization problem using the tools for DRO framework. Instead of breaking the problem in two separate steps (model first the uncertainty using data-fusion algorithms and then use it to solve the stochastic optimization problem), the DRO method jointly tackles them seeking to provide approximations to the optimizers tailored to the quality and size of the gathered data.

Literature review

Optimization under uncertainty is a classical topic [1]. A recent addition to the panoply of methods available to solve these problems is data-driven distributionally robust optimization (DRO), see e.g. [2], [3] 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 ϕ-divergence [4], Prohorov metric [5], and Wasserstein distance [2]. 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 our focus in the paper. In this context, our work has connections with the growing body of literature on distribution optimization, see e.g. [6] and references therein. Our work is in contrast with the setup of distributed machine learning, see e.g. [7]. Unlike our setup, these works 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 the definition of the stochastic optimization problem that a group of agents aim to solve. The probability distribution of the random variable involved in this problem is unknown and instead, agents collect a finite set of samples of the random variable. 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. With this perspective, we formulate a convex optimization problem that uses the group's collective dataset to seek a solution having guarantees on the out-of-sample performance for the stochastic optimization. By augmenting the decision variables, we reformulate the convex optimization problem to yield a structure amenable to the design of

A. Cherukuri is with the Automatic Control Laboratory, ETH Zürich, cashish@control.ee.ethz.ch and J. Cortés is with the Dept. of Mechanical and Aerospace Engineering, UC San Diego, cortes@ucsd.edu.

distributed algorithm. That is, the objective function is the aggregate of individual objectives and constraints involve consensus among decision variables. Next, we identify a convex-concave function that has two important properties: first, its saddle-points are in one-to-one correspondence with the optimizers of the reformulated problem; and second, the saddle-point dynamics written for this convex-concave function is implementable in a distributed manner. This forms our distributed algorithm. We establish that this saddle-point dynamics asymptotically converges to a saddle point of the convex-concave function and hence, to a solution of the convex optimization problem. Simulations illustrate our results. For reasons of space, the proofs are omitted and will appear elsewhere.

Organization

Section II introduces basic preliminary notions. Section III describes our problem statement. Section IV presents a reformulation geared towards our distributed algorithmic design, which we tackle in Section V, along with the establishment of its convergence guarantees. Section VI shows simulations results. We gather our conclusions in Section VII.

II. Preliminaries

This section introduces our notation and basic notions on graph theory, data-driven stochastic optimization, and saddle points.

A. Notation

We start with some notation and basic definitions. Let $\mathbb{R}$, $\mathbb{R}_{\geq 0}$, and $\mathbb{Z}_{\geq 1}$ denote the set of real, nonnegative real, and positive integer numbers. The extended reals are denoted as $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. We let $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ denote the 2-norm and the inner product on $\mathbb{R}^n$. Given $x, y \in \mathbb{R}^n$, x_i denotes the i-th component of x, and $x \leq y$ denotes $x_i \leq y_i$ for $i \in [n]$. For vectors $u \in \mathbb{R}^n$ and $w \in \mathbb{R}^m$, the vector $(u; w) \in \mathbb{R}^{n+m}$ denotes their concatenation. We use the shorthand notation $\mathbf{0}_n = (0, \dots, 0) \in \mathbb{R}^n$, $\mathbf{1}_n = (1, \dots, 1) \in \mathbb{R}^n$, and $\mathsf{I}_n \in \mathbb{R}^{n \times n}$ for the identity matrix. For $A \in \mathbb{R}^{n_1 \times n_2}$ and $B \in \mathbb{R}^{m_1 \times m_2}$, $A \otimes B \in \mathbb{R}^{n_1 m_1 \times n_2 m_2}$ is the Kronecker product. The Cartesian product of any set of objects $\{\mathcal{S}_i\}_{i=1}^n$ is denoted by $\prod_{i=1}^n \mathcal{S}_i := \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$. For a real-valued function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $(x, \xi) \to f(x, \xi)$, we denote the partial derivative of f with respect to the first argument by $\nabla_x f$ and with respect to the second argument by $\nabla_\xi f$.

A function $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ 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 $(x_*, y_*) \in \mathcal{X} \times \mathcal{Y}$ is a *saddle point* of F over the set $\mathcal{X} \times \mathcal{Y}$ if $F(x_*, y) \leq F(x_*, y_*) \leq F(x, y_*)$, 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 $\nabla_x F(x_*, z_*) = 0$ and $\nabla_z F(x_*, z_*) = 0$. Additionally, if F is twice differentiable, then $\nabla_{xx} F(x_*, z_*) \preceq 0$ and $\nabla_{zz} F(x_*, z_*) \succeq 0$.

B. Graph theory

Following [8], an *undirected graph*, or simply a *graph*, is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the vertex set and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set. For a graph $(v, u) \in \mathcal{E}$ if and only if $(u, v) \in \mathcal{E}$. A path is an ordered sequence such that any ordered pair of vertices appearing consecutively is an edge. A graph is *connected* if there is a path between any pair of distinct vertices. Let $\mathcal{N}_v \subseteq \mathcal{V}$ denote the set of neighbors of vertex $v \in \mathcal{V}$, i.e., $\mathcal{N}_v = \{u \in \mathcal{V} \mid (v, u) \in \mathcal{E}\}$. A *weighted graph* is a triplet $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathsf{A})$, where $(\mathcal{V}, \mathcal{E})$ is a digraph and $\mathsf{A} \in \mathbb{R}_{\geq 0}^{n \times n}$ is the *adjacency matrix* of $\mathcal{G}$, with the property that $a_{ij} > 0$ if $(v_i, v_j) \in \mathcal{E}$ and $a_{ij} = 0$, otherwise. Also, $a_{ij} = a_{ji}$ for all $(i, j) \in \mathcal{E}$. 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 $(\mathsf{D})_{ii} = w_i$, for all $i \in [n]$. The *Laplacian* matrix is $\mathsf{L} = \mathsf{D} - \mathsf{A}$. Note that $\mathsf{L} = \mathsf{L}^\top$ and $\mathsf{L}\mathbf{1}_n = 0$. If $\mathcal{G}$ is connected, then zero is a simple eigenvalue of L.

C. Data-driven stochastic optimization

Here we present notions on data-driven stochastic optimization following [2]. 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. Assume that Ξ is closed and convex. 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 continuously differentiable function, and $\mathbb{E}_{\mathbb{P}}[\,\cdot\,]$ is the expectation under the distribution $\mathbb{P}$. Assume that $\mathbb{P}$ is unknown and so, solving (1) is not possible. However, we are given N independently drawn samples $\widehat{\Xi} := \{\widehat{\xi}^k\}_{k=1}^N \subset \Xi$ of the random variable ξ. Note that, until it is revealed, $\widehat{\Xi}$ is a random object with probability distribution $\mathbb{P}^N := \prod_{i=1}^N \mathbb{P}$ supported on $\Xi^N := \prod_{i=1}^N \Xi$. 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 desirable properties for the expected cost $\mathbb{E}_{\mathbb{P}}[f(\widehat{x}_N, \xi)]$ under a new sample. The property we are looking for is the *finite-sample guarantee* given by

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

where $\widehat{J}_N$ might also depend on the training dataset and $\beta \in (0, 1)$ is the parameter which governs $\widehat{x}_N$ and $\widehat{J}_N$. The quantities $\widehat{J}_N$ and $1 - \beta$ are referred to as the *certificate* and the *reliability* of the performance of $\widehat{x}_N$. 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 $\widehat{\Xi}_N$. The strategy is to determine a set $\widehat{\mathcal{P}}_N$ of probability distributions supported on Ξ that contain the true distribution $\mathbb{P}$ with high confidence. The set $\widehat{\mathcal{P}}_N$ is referred to as the *ambiguity* set. Once such a set is designed, the certificate $\widehat{J}_N$ is defined as the optimal value of the following *distributionally robust optimization* problem

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

This is the worst-case optimal value considering all distributions in $\widehat{\mathcal{P}}_N$. A good candidate for $\widehat{\mathcal{P}}_N$ is the set of distributions that are close (under a certain metric) to the uniform distribution on $\widehat{\Xi}_N$, termed the *empirical distribution*. Formally, the empirical distribution is

$$\widehat{\mathbb{P}}_N := \frac{1}{N} \sum_{k=1}^{N} \delta_{\widehat{\xi}^k}, \tag{4}$$

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 first moment, i.e., $\mathbb{E}_{\mathbb{Q}}[\|\xi\|] = \int_{\Xi} \|\xi\| \mathbb{Q}(d\xi) < \infty$. The *2-Wasserstein metric* [1] $d_{W_2} : \mathcal{M}(\Xi) \times \mathcal{M}(\Xi) \to \mathbb{R}_{\geq 0}$ is

$$d_{W_2}(\mathbb{Q}_1, \mathbb{Q}_2) = \Big(\inf\Big\{ \int_{\Xi^2} \|\xi_1 - \xi_2\|^2 \Pi(d\xi_1, d\xi_2) \Big| \; \Pi \in \mathcal{H}(\mathbb{Q}_1, \mathbb{Q}_2) \Big\}\Big)^{\frac{1}{2}}, \tag{5}$$

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 $\epsilon \geq 0$, we use the notation

$$\mathcal{B}_\epsilon(\widehat{\mathbb{P}}_N) := \{\mathbb{Q} \in \mathcal{M}(\Xi) \mid d_{W_2}(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq \epsilon\} \tag{6}$$

to define the set of distributions that are ϵ-close to $\widehat{\mathbb{P}}_N$ under the defined metric. Let $\mathcal{M}_{\mathrm{lt}}(\Xi) \subset \mathcal{M}(\Xi)$ be the set of *light-tailed* distributions $\mathbb{P} \in \mathcal{M}_{\mathrm{lt}}(\Xi)$ for which there exists an exponent $a > 2$ such that

$$A := \mathbb{E}_{\mathbb{P}}[\exp(\|\xi\|^a)] = \int_{\Xi} \exp(\|\xi\|^a) \mathbb{P}(d\xi) < \infty.$$

The next result gives a lower bound on the probability with which the true distribution $\mathbb{P}$ is ϵ-close to $\widehat{\mathbb{P}}_N$.

Theorem II.1. *(Finite-sample guarantee of $\mathbb{P}$ belonging to the Wasserstein ambiguity set): Let $\mathbb{P} \in \mathcal{M}_{\mathrm{lt}}(\Xi)$. Then,*

$$\mathbb{P}^N(d_{W_2}(\mathbb{P}, \widehat{\mathbb{P}}_N) \geq \epsilon) \leq \begin{cases} c_1 e^{-c_2 N \epsilon^{\max\{4,m\}}}, & \text{if } \epsilon \leq 1, \\ c_1 e^{-c_2 N \epsilon^a}, & \text{if } \epsilon > 1, \end{cases} \tag{7}$$

for all $N \geq 1$, $m \neq 4$, and $\epsilon > 0$, where c_1, c_2 are positive constants that only depend on a, A, and m.

The proof is a direct application of [9, Theorem 2]. This result gives a method to construct the ambiguity set $\widehat{\mathcal{P}}_N$. Equating the right-hand side of (7) with the chosen $\beta \in (0, 1)$, we define for each $N \in \mathbb{Z}_{\geq 1}$,

$$\epsilon_N(\beta) = \begin{cases} \Big(\frac{\log(c_1 \beta^{-1})}{c_2 N}\Big)^{1/\max\{4,m\}}, & \text{if } N \geq \frac{\log(c_1 \beta^{-1})}{c_2}, \\ \Big(\frac{\log(c_1 \beta^{-1})}{c_2 N}\Big)^{1/a}, & \text{if } N < \frac{\log(c_1 \beta^{-1})}{c_2}. \end{cases} \tag{8}$$

Plugging this value in (7) yields $\mathbb{P}^N(d_{W_2}(\mathbb{P}, \widehat{\mathbb{P}}_N) \geq \epsilon_N(\beta)) \leq \beta$. That is, if we let $\widehat{\mathcal{P}}_N := \mathcal{B}_{\epsilon_N(\beta)}(\widehat{\mathbb{P}}_N)$, then

$$\mathbb{P}^N(\mathbb{P} \in \widehat{\mathcal{P}}_N) \geq 1 - \beta, \tag{9}$$

i.e., the true distribution belongs to the ambiguity set with probability at least $1 - \beta$. This leads us to the following result.

Theorem II.2. *(Finite-sample guarantee of (3) with $\widehat{\mathcal{P}}_N = \mathcal{B}_{\epsilon_N(\beta)}(\widehat{\mathbb{P}}_N)$): For $\mathbb{P} \in \mathcal{M}_{\mathrm{lt}}(\Xi)$ and $\beta \in (0, 1)$, let $\widehat{J}_N$ and $\widehat{x}_N$ be the optimal value and an optimizer of the distributionally robust optimization (3) with $\widehat{\mathcal{P}}_N = \mathcal{B}_{\epsilon_N(\beta)}(\widehat{\mathbb{P}}_N)$. Then, the finite-sample guarantee (2) holds.*

The proof follows by using (3) and (9) to yield (2). We end this section by discussing the tractability of solving (3) with $\widehat{\mathcal{P}}_N = \mathcal{B}_{\epsilon_N(\beta)}(\widehat{\mathbb{P}}_N)$. The next result shows that under mild conditions on the objective function, one can reformulate the problem as a convex optimization problem.

Theorem II.3. *(Tractable reformulation of (3)): Assume that for all $\tilde{x} \in \mathcal{X}$ and $\tilde{\xi} \in \Xi$, maps $\xi \mapsto -f(\tilde{x}, \xi)$ and $x \mapsto f(x, \tilde{\xi})$ are convex. Then, for any $\beta \in (0, 1)$ and $N \in \mathbb{Z}_{\geq 1}$, the optimal value of (3) with $\widehat{\mathcal{P}}_N = \mathcal{B}_{\epsilon_N(\beta)}(\widehat{\mathbb{P}}_N)$ is equal to the optimum of the following convex optimization problem*

$$\inf_{\lambda \geq 0, x \in \mathcal{X}} \Big\{ \lambda \epsilon_N^2(\beta) + \frac{1}{N} \sum_{k=1}^{N} \max_{\xi \in \Xi} \Big(f(x, \xi) - \lambda \|\xi - \widehat{\xi}^k\|^2 \Big) \Big\}.$$

III. Problem statement

Consider $n \in \mathbb{Z}_{\geq 1}$ agents communicating over an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathsf{A})$. The set of vertices are enumerated as $\mathcal{V} := [n]$. Each agent $i \in [n]$ can send and receive information from its neighbors $\mathcal{N}_i$ in $\mathcal{G}$. Let $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$, $(x, \xi) \to f(x, \xi)$, be a continuously differentiable, convex-concave function satisfying $f(x, \tilde{\xi}) \to \infty$ as $\|x\| \to \infty$ for all $\tilde{\xi} \in \mathbb{R}^m$. We assume all agents know this function. 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 following stochastic optimization problem

$$\inf_{x \in \mathbb{R}^d} \mathbb{E}_{\mathbb{P}} \Big[f(x, \xi) \Big], \tag{10}$$

which is not feasible given that $\mathbb{P}$ is unknown. Instead, each agent has a certain number of independent and identically distributed realizations of the random variable ξ. We denote the data available to agent i by $\widehat{\Xi}_i \subset \widehat{\Xi}$. 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 data set.

[1] We note that [2] employs the 1-Wasserstein metric instead of the 2-Wasserstein metric considered here.

The goal for the agents is then to collectively find, in a distributed manner, a data-driven solution $\widehat{x}_N \in \mathbb{R}^d$ to approximate the optimizer of (10) with guaranteed performance bounds. To achieve this, we rely on the framework of distributionally robust optimization, cf. Section II-2. From Theorem II.3, a data-driven solution for (10) can be obtained by solving the following convex optimization problem

$$\inf_{\lambda \geq 0, x} \Big\{ \lambda \epsilon_N^2(\beta) + \frac{1}{N} \sum_{k=1}^{N} \max_{\xi \in \mathbb{R}^m} \Big(f(x, \xi) - \lambda \|\xi - \widehat{\xi}^k\|^2 \Big) \Big\} \quad (11)$$

where $\beta \in (0,1)$ and $\epsilon_N(\beta)$ is given in (8). Assume that there exists a finite optimizer of (11), e.g., one of the conditions for existence of finite optimizers given in [10] is met. This optimizer, denoted $\widehat{x}_N$, has the performance guarantee

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

where $\widehat{J}_N$ is the optimum value (11). The agents in the network aim to solve (11) in a distributed manner, that is

(i) each agent i has the information
$$\mathcal{I}_i := \{\widehat{\Xi}_i, f, a, c_1, c_2, A, \beta, n, N\}, \quad (12)$$
where a, c_1, c_2, and A are parameters associated with the distribution $\mathbb{P}$, 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 $\mathcal{G}$,

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

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

The challenge in solving (11) in a distributed manner lies in the fact that the data is distributed over the network and the optimizer $\widehat{x}_N$ depends on it all. In Section IV, we overcome this hurdle by reformulating the problem allowing us, in Section V, to synthesize a distributed algorithm to solve it.

IV. Distributed problem formulation and saddle points

This section studies the structure of the optimization problem presented in Section III with the ulterior goal of facilitating the design of a distributed algorithmic solution. Our first step is a reformulation of (11) that, by augmenting the decision variables of the agents, leads us to 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 relevant step is the identification of a convex-concave function whose saddle points correspond to the optimizers of the reformulated problem, opening 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 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 $\lambda_{\mathrm{v}} := (\lambda^1; \dots; \lambda^n)$, and $x_{\mathrm{v}} := (x^1; \dots; x^n)$. Let $v_k \in [n]$ be the agent that holds the k-th sample $\widehat{\xi}^k$ of the dataset. Consider the following convex optimization problem

$$\min_{x_{\mathrm{v}}, \lambda_{\mathrm{v}} \geq \mathbf{0}_n} \quad h(\lambda_{\mathrm{v}}) + \frac{1}{N} \sum_{k=1}^{N} \max_{\xi \in \mathbb{R}^m} g_k(x^{v_k}, \lambda^{v_k}, \xi) \quad (13a)$$

$$\text{subject to} \quad \mathsf{L}\lambda_{\mathrm{v}} = \mathbf{0}_n, \quad (13b)$$

$$(\mathsf{L} \otimes \mathsf{I}_d) x_{\mathrm{v}} = \mathbf{0}_{nd}, \quad (13c)$$

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

$$h(\lambda_{\mathrm{v}}) := \frac{\epsilon_N^2(\beta)(\mathbf{1}_n^\top \lambda_{\mathrm{v}})}{n}$$

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 following result establishes the correspondence between the optimizers of (11) and (13), respectively.

Lemma IV.1. *(One-to-one correspondence between optimizers of* (11) *and* (13)*): The following holds:*

(i) *If (x^*, λ^*) is an optimizer of* (11)*, then $(\mathbf{1}_n \otimes x^*, \lambda^* \mathbf{1}_n)$ is an optimizer of* (13).

(ii) *If $(x_{\mathrm{v}}^*, \lambda_{\mathrm{v}}^*)$ is an optimizer of* (13)*, then there exists an optimizer (x^*, λ^*) of* (11) *such that $x_{\mathrm{v}}^* = \mathbf{1}_n \otimes x^*$ and $\lambda_{\mathrm{v}}^* = \lambda^* \mathbf{1}_n$.*

Note that constraints (13b) and (13c) force agreement and that each of their components is computable by an agent of the network using only local information. Moreover, the objective function (13a) can be written as $\sum_{i=1}^n J_i(x^i, \lambda^i, \widehat{\Xi}_i)$, where

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

for all $i \in [n]$. Therefore, the problem (13) has the adequate structure from a distributed optimization viewpoint: an aggregate objective function and locally computable constraints.

B. Optimizers as saddle points

Our next step is to map the optimizers of (13) to the saddle points of an (appropriate variant of the) Lagrangian function. Once this is established, the saddle-point dynamics associated to the Lagrangian function is the obvious candidate for a distributed algorithm to solve the original problem. Note that the Lagrangian of (13) is $L : \mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0} \times \mathbb{R}^n \times \mathbb{R}^{nd} \to \overline{\mathbb{R}}$,

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

$$+\nu^\top \mathsf{L}\lambda_{\mathrm{v}} + \eta^\top(\mathsf{L}\otimes \mathsf{I}_d)x_{\mathrm{v}}, \qquad (14)$$

where $\nu \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^{nd}$ are dual variables corresponding to the equality constraints (13b) and (13c), 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 establishes important properties of the Lagrangian giving a correspondence between its saddle points and the optimizers of (13).

Lemma IV.2. *(Min-max equality for L): The set of saddle points of L over the domain $(\mathbb{R}^{nd}\times\mathbb{R}^n_{\geq 0})\times(\mathbb{R}^n\times\mathbb{R}^{nd})$ 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:

(i) *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* (13).

(ii) *If $(\overline{x}_{\mathrm{v}},\overline{\lambda}_{\mathrm{v}})$ is an optimizer of* (13)*, 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 $(\mathbb{R}^{nd}\times\mathbb{R}^n_{\geq 0})\times(\mathbb{R}^n\times\mathbb{R}^{nd})$.*

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., [11]. In order to overcome this hurdle, we augment the Lagrangian with quadratic terms in the primal variables. 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 shows 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 $(x^*_{\mathrm{v}},\lambda^*_{\mathrm{v}},\nu^*,\eta^*)$ is a saddle point of L over $(\mathbb{R}^{nd}\times\mathbb{R}^n_{\geq 0})\times(\mathbb{R}^n\times\mathbb{R}^{nd})$ 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 [12, Theorem 1.1]. The above result implies that finding the saddle points of L_{aug} would take us to the optimizers of (13). However, there is one more roadblock remaining, and that is writing a gradient-based dynamics for L_{aug}. Notice that L_{aug} itself 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 routines and identify a function for which the gradient-based dynamics is well defined over the feasible domain. Note that

$$L_{\mathrm{aug}}(x_{\mathrm{v}},\lambda_{\mathrm{v}},\nu,\eta) = \max_{\{\xi^k\}}\tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}},\lambda_{\mathrm{v}},\nu,\eta,\{\xi^k\}), \qquad (15)$$

where

$$\tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}},\lambda_{\mathrm{v}},\nu,\eta,\{\xi^k\}) := h(\lambda_{\mathrm{v}}) + \frac{1}{N}\sum_{k=1}^N g_k(x^{v_k},\lambda^{v_k},\xi^k) + \nu^\top\mathsf{L}\lambda_{\mathrm{v}} + \eta^\top(\mathsf{L}\otimes\mathsf{I}_d)x_{\mathrm{v}} + \frac{1}{2}x_{\mathrm{v}}^\top(\mathsf{L}\otimes\mathsf{I}_d)x_{\mathrm{v}} + \frac{1}{2}\lambda_{\mathrm{v}}^\top\mathsf{L}\lambda_{\mathrm{v}}. \qquad (16)$$

Lemma IV.3 implies that saddle points of L_{aug} exist and so,

$$\min_{x_{\mathrm{v}},\lambda_{\mathrm{v}}\geq\mathbf{0}_n}\max_{\nu,\eta} L_{\mathrm{aug}}(x_{\mathrm{v}},\lambda_{\mathrm{v}},\nu,\eta) = \max_{\nu,\eta}\min_{x_{\mathrm{v}},\lambda_{\mathrm{v}}\geq\mathbf{0}_n} L_{\mathrm{aug}}(x_{\mathrm{v}},\lambda_{\mathrm{v}},\nu,\eta).$$

Using (15) in the above expression yields

$$\min_{x_{\mathrm{v}},\lambda_{\mathrm{v}}\geq\mathbf{0}_n}\max_{\nu,\eta}\max_{\{\xi^k\}}\tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}},\lambda_{\mathrm{v}},\nu,\eta,\{\xi^k\}) = \max_{\nu,\eta}\min_{x_{\mathrm{v}},\lambda_{\mathrm{v}}\geq\mathbf{0}_n}\max_{\{\xi^k\}}\tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}},\lambda_{\mathrm{v}},\nu,\eta,\{\xi^k\}). \qquad (17)$$

Note that $\tilde{L}_{\mathrm{aug}}$ is convex-concave in the variables $((x_{\mathrm{v}},\lambda_{\mathrm{v}}),(\nu,\eta,\{\xi^k\}))$ over the domain $\lambda_{\mathrm{v}}\geq\mathbf{0}_n$. The next result shows that, for a fixed (ν,η), the min and the max operator in the right-hand side of the above equality can be interchanged. This paves the way to show that saddle points of $\tilde{L}_{\mathrm{aug}}$ exist and are in correspondence with those of L_{aug}.

Lemma IV.4. *(Min-max operators can be interchanged for $\tilde{L}_{\mathrm{aug}}$): Given any $(\nu,\eta)\in\mathbb{R}^n\times\mathbb{R}^{nd}$, the following holds*

$$\min_{x_{\mathrm{v}},\lambda_{\mathrm{v}}\geq\mathbf{0}_n}\max_{\{\xi^k\}}\tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}},\lambda_{\mathrm{v}},\nu,\eta,\{\xi^k\}) = \max_{\{\xi^k\}}\min_{x_{\mathrm{v}},\lambda_{\mathrm{v}}\geq\mathbf{0}_n}\tilde{L}_{\mathrm{aug}}(x_{\mathrm{v}},\lambda_{\mathrm{v}},\nu,\eta,\{\xi^k\}). \qquad (18)$$

Using Lemma IV.4 in (17), we obtain

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

Since $\tilde{L}_{\mathrm{aug}}$ is convex-concave in $((x_{\mathrm{v}},\lambda_{\mathrm{v}}),(\nu,\eta,\{\xi^k\}))$, the above equality along with a straightforward computation establishes the following result.

Lemma IV.5. *(Correspondence between saddle points of L_{aug} and $\tilde{L}_{\mathrm{aug}}$): The set of saddle points of $\tilde{L}_{\mathrm{aug}}$ is nonempty, convex, and compact and the following holds:*

(i) *If $(x_{\mathrm{v}},\lambda_{\mathrm{v}},\nu,\eta)$ is a saddle point of L_{aug}, then there exists $\{\xi^k\}$ such that $(x_{\mathrm{v}},\lambda_{\mathrm{v}},\nu,\eta,\{\xi^k\})$ is a saddle point of $\tilde{L}_{\mathrm{aug}}$.*

(ii) *If $(x_{\mathrm{v}},\lambda_{\mathrm{v}},\nu,\eta,\{\xi^k\})$ is a saddle point of $\tilde{L}_{\mathrm{aug}}$, then $(x_{\mathrm{v}},\lambda_{\mathrm{v}},\nu,\eta)$ is a saddle point of L_{aug}.*

Finally, combining Lemmas **??**, IV.3 and IV.5, we arrive at the one-to-one correspondence between the saddle points of $\tilde{L}_{\mathrm{aug}}$ and the optimizers of (13).

Proposition IV.6. *(Correspondence between optimizers of* (13) *and the saddle points of $\tilde{L}_{\mathrm{aug}}$): The following holds:*

(i) *If* $((x_{\rm v}^*, \lambda_{\rm v}^*, \nu^*, \eta^*, \{(\xi^*)^k\})$ *is a saddle point of* $\tilde{L}_{\rm aug}$ *over* $(\mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}) \times (\mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN})$*, then* $(x_{\rm v}^*, \lambda_{\rm v}^*)$ *is an optimizer of* (13).

(ii) *If* $(x_{\rm v}^*, \lambda_{\rm v}^*)$ *is an optimizer of* (13)*, then there exists* $(\nu^*, \eta^*, \{(\xi^*)^k\})$ *such that* $((x_{\rm v}^*, \lambda_{\rm v}^*, \nu^*, \eta^*, \{(\xi^*)^k\})$ *is a saddle point of* $\tilde{L}_{\rm aug}$ *over* $(\mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}) \times (\mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN})$.

This result opens the way to our distributed algorithmic design based on the saddle-point dynamics associated with $\tilde{L}_{\rm aug}$, which we tackle in the next section.

V. Distributed algorithm design and convergence analysis

We introduce here our distributed algorithmic solution and present its convergence analysis. The purpose of the algorithm is to find the saddle points of $\tilde{L}_{\rm aug}$ (as these points correspond to the optimizers of (13), cf. Proposition IV.6). One way of reaching the saddle points of a convex-concave function is by performing gradient-descent in the convex variables and gradient-ascent in the concave ones. This is popularly termed as the saddle-point or the primal-dual dynamics [11], [13]. The saddle-point dynamics for $\tilde{L}_{\rm aug}$ is

$$\frac{dx_{\rm v}}{dt} = -\nabla_{x_{\rm v}} \tilde{L}_{\rm aug}(x_{\rm v}, \lambda_{\rm v}, \nu, \eta, \{\xi^k\}), \tag{19a}$$

$$\frac{d\lambda_{\rm v}}{dt} = [-\nabla_{\lambda_{\rm v}} \tilde{L}_{\rm aug}(x_{\rm v}, \lambda_{\rm v}, \nu, \eta, \{\xi^k\})]^+_{\lambda_{\rm v}}, \tag{19b}$$

$$\frac{d\nu}{dt} = \nabla_{\nu} \tilde{L}_{\rm aug}(x_{\rm v}, \lambda_{\rm v}, \nu, \eta, \{\xi^k\}), \tag{19c}$$

$$\frac{d\eta}{dt} = \nabla_{\eta} \tilde{L}_{\rm aug}(x_{\rm v}, \lambda_{\rm v}, \nu, \eta, \{\xi^k\}), \tag{19d}$$

$$\frac{d\xi^k}{dt} = \nabla_{\xi^k} \tilde{L}_{\rm aug}(x_{\rm v}, \lambda_{\rm v}, \nu, \eta, \{\xi^k\}), \forall k \in [N]. \tag{19e}$$

For convenience, denote (19) by the vector field $X_{\rm sp} : \mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0} \times \mathbb{R}^{nd+n+mN} \to \mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0} \times \mathbb{R}^{nd+n+mN}$. In this notation, the first, second, and third components correspond to the dynamics of $x_{\rm v}$, $\lambda_{\rm v}$, and $(\nu, \eta, \{\xi^k\})$, respectively.

Remark V.1. *(Distributed implementation of X_{sp}):* We denote the components of the dual variables η and ν by $\eta = (\eta^1; \eta^2; \dots; \eta^n)$ and $\nu = (\nu^1; \nu^2; \dots; \nu^n)$, so that agent $i \in [n]$ maintains $\eta^i \in \mathbb{R}^d$ and $\nu^i \in \mathbb{R}$. Further, 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 $X_{\rm sp}$, 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 $(x_{\rm v}, \lambda_{\rm v}, \nu, \eta, \{\xi^k\})$. From (19), one can write the dynamics of variables maintained by i and notice that the this dynamics is computable by agent i using the variables that it maintains and information collected from its neighbors. Hence, $X_{\rm sp}$ 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, $\{\xi^k\}_{k \in \mathcal{K}_i}$ can be interpreted as its internal state that is not communicated to its neighbors. •

The following result establishes the convergence of the dynamics $X_{\rm sp}$ to the saddle points of $\tilde{L}_{\rm aug}$ and hence to the desired optimizers.

Theorem V.2. *(Convergence of trajectories of X_{sp} to the optimizers of* (13)*): Assume there exists an optimizer* $(x_{\rm v}^*, \lambda_{\rm v}^*)$ *of* (13) *that satisfies* $\lambda_{\rm v}^* \neq 0$*. Then, starting from any initial condition* $(x_{\rm v}(0), \lambda_{\rm v}(0), \nu(0), \eta(0), \{\xi^k(0)\})$ *satisfying* $\lambda_{\rm v}(0) \geq \mathbf{0}_n$*, the trajectory of the saddle-point dynamics for* $\tilde{L}_{\rm aug}$ (19) *converges asymptotically to a saddle point of* $\tilde{L}_{\rm aug}$ *over* $(\mathbb{R}^{nd} \times \mathbb{R}^n_{\geq 0}) \times (\mathbb{R}^n \times \mathbb{R}^{nd} \times \mathbb{R}^{mN})$*. Consequently, the components* $(x_{\rm v}, \lambda_{\rm v})$ *of the trajectory converge to an optimizer of* (13).

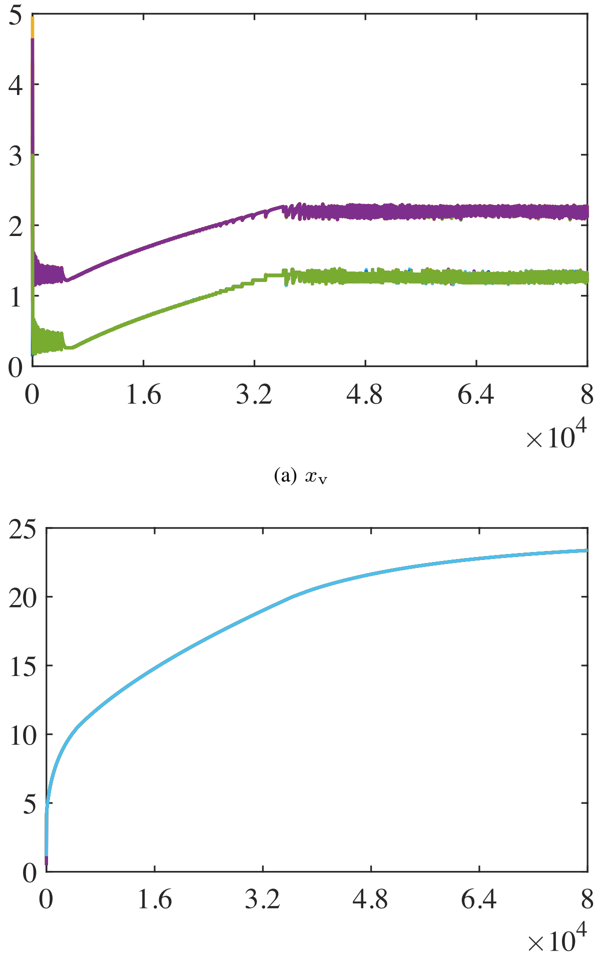
VI. Simulations

Here we illustrate the application of our distributed algorithm (19) to find, in a distributed manner, a data-driven solution for the least absolute deviations problem [14, Chapter 3]. Assume $n = 6$ agents with communication topology defined by an undirected ring with additional edges $(1, 4)$ and $(2, 6)$. The weight of each edge is equal to one. In this problem, each data point $\widehat{\xi}^k = (\widehat{w}^k, \widehat{y}^k)$ consists of a set of input $\widehat{w}^k \in \mathbb{R}$ and output $\widehat{y}^k \in \mathbb{R}$ pairs. The objective is to find a 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 popular way of finding such a predictor x is to solve the following problem

$$\inf_x \mathbb{E}_{\mathbb{P}} \Big[f(x, w, y) \Big] \tag{20}$$

where $\mathbb{P}$ is the probability distribution of the data (w, y) and $f : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ is the absolute value of the residual, i.e., $f(x, w, y) = \big| x^\top(w; 1) - y \big|$. Note that f is convex-concave.

For generating the dataset, we assume that the input vector w has a standard normal distribution and the output y is assigned values $y = 4w + v$ where v is a random variable, uniformly distributed over the interval $[-0.1, 0.1]$. This defines completely the distribution $\mathbb{P}$ of (w, y). For finding the data-driven solution, we assume that each agent in the network has 10 i.i.d samples of (w, y) and hence $N = 60$ is the total number of samples. Let β be some value belonging to the interval $(0, 1)$ such that $\epsilon_N(\beta) = 0.05$. This value is assumed to be computed by the agents beforehand. This defines completely the distributed optimization problem (13). Figure 1 shows the execution of the distributed algorithm (19) that solves this problem. Note that f is nondifferentiable and therefore, in (19) we replace the gradient operators with generalized gradients which makes (19) a differential inclusion. The trajectories converge to an equilibrium of the dynamics (19) establishing that the optimizers of (13) are obtained.



(a) x_{v}

(b) λ_{v}

Fig. 1. Illustration of the execution of the dynamics X_{sp} (19) to find a data-driven solution for the least absolute deviations problem (20). Plots (a) and (b) depict the evolution of the primal variables of the distributed optimization problem (13) defined for (20) with $\epsilon_N(\beta) = 0.05$ (for the sake of simplicity we have not shown the dual variables). The number of agents is 6 and each agent collects 10 i.i.d samples of the random variable as described in Section VI. The initial condition $(x_{\mathrm{v}}(0), \lambda_{\mathrm{v}}(0))$ is chosen randomly from the set $[0,5]^{12} \times [0,5]^6$ and $\nu(0) = \mathbf{0}_6$, $\eta(0) = \mathbf{0}_{12}$, and $\xi^k(0) = \mathbf{0}_2$ for all $k \in [N]$. The primal variables converge to the optimizer of (13), $\lambda_{\mathrm{v}}^* = \lambda^* \mathbf{1}_6$ and $x_{\mathrm{v}}^* = \mathbf{1}_6 \otimes x^*$ with $\lambda^* = 23.37$ and $x^* = [2.13; 1.27]$.

VII. Conclusions

We have considered a stochastic optimization problem where the probability distribution of the random variable is unknown and a group of agents collect samples of it. We have formulated a convex optimization problem that finds a data-driven solution to the stochastic optimization problem and designed a distributed algorithm that converges asymptotically to its solutions. The algorithm is a saddle-point dynamics for a convex-concave modified Lagrangian function whose saddle points correspond to the optimizers of the problem.

Future work will generalize the results for nonsmooth objective functions that are not necessarily concave in the random variable. We also wish to extend the algorithms to handle scenarios with streaming data and network chance-constrained optimization problems.

Acknowledgments

This work was supported by AFOSR Award FA9550-15-1-0108.

References

[1] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming*. Philadelphia, PA: SIAM, 2014.

[2] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations," 2015. Available at `https://arxiv.org/abs/1505.05116`.

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

[4] R. Jiang and Y. Guan, "Data-driven chance constrained stochastic program," *Mathematical Programming, Series A*, vol. 158, pp. 291–327, 2016.

[5] E. Erdoğan and G. Iyengar, "Ambiguous chance constrained problems and robust optimization," *Mathematical Programming, Series B*, vol. 107, pp. 37–61, 2006.

[6] A. Nedić, "Distributed optimization," in *Encyclopedia of Systems and Control* (J. Baillieul and T. Samad, eds.), New York: Springer, 2015.

[7] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: distributed machine learning for on-device intelligence," 2016. Available at `https://arxiv.org/abs/1610.02527`.

[8] F. Bullo, J. Cortés, and S. Martínez, *Distributed Control of Robotic Networks*. Applied Mathematics Series, Princeton University Press, 2009. Electronically available at `http://coordinationbook.info`.

[9] N. Fournier and A. Guillin, "On the rate of convergence in wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, vol. 162, no. 3, pp. 707–738, 2015.

[10] A. E. Ozdaglar and P. Tseng, "Existence of global minima for constrained optimization," *Journal of Optimization Theory & Applications*, vol. 128, no. 3, pp. 523–546, 2006.

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

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

[13] A. Cherukuri, E. Mallada, and J. Cortés, "Asymptotic convergence of primal-dual dynamics," *Systems & Control Letters*, vol. 87, pp. 10–15, 2016.

[14] B. Schölkopf and A. J. Smola, *Learning with kernels*. Cambridge, MA: MIT Press, 2002.