Distributed Stochastic Dual Subgradient for Constraint-Coupled Optimization

Distributed Stochastic Dual Subgradient for Constraint-Coupled Optimization

Andrea Camisa, *IEEE Member*, Ivano Notarnicola, *IEEE Member*, Giuseppe Notarstefano, *IEEE Member*

***Abstract*— In this paper we consider a distributed stochastic optimization framework in which agents in a network aim to cooperatively learn an optimal network-wide policy. The goal is to compute local functions to minimize the expected value of a given cost, subject to individual constraints and average coupling constraints. In order to handle the challenges of the distributed stochastic context, we resort to a Lagrangian duality approach that allows us to derive an associated stochastic dual problem with a separable structure. Thus, we propose a distributed algorithm, without a central coordinator, that exploits consensus iterations and stochastic approximation to find an optimal solution to the problem, with attractive scalability properties. We demonstrate convergence of the proposed scheme and validate its behavior through simulations.**

***Index Terms*— Optimization algorithms, Large-scale systems, Distributed control**

I. Introduction

We focus on a distributed optimization framework in which agents want to find local optimal policies that minimize the expected value of the sum of local cost functions subject to individual constraints and a global average coupling constraint. This *stochastic constraint-coupled* optimization set-up captures various applications in smart grid control, cooperative robotics and sensor networks in which multiple systems with their own dynamics have to negotiate how to share a common resource.

Literature. Constraint-coupled optimization is an emerging distributed optimization scenario that is getting more and more attention from the community. For the case of deterministic problems, authors have proposed dual approaches [1], [2], saddle-point methods [3]–[5] and others [6]–[10]. See also the tutorial [11] for a recent, comprehensive literature review. Although these algorithms can handle the complexities related to distributed computation, they have no guarantee for the stochastic setting. On the other hand, literature on distributed stochastic optimization typically consider scenarios with cost functions being all dependent on the *same* variable (with no coupling constraints). An examplary, non-exhaustive list of works tackling this set-up is [12]–[16]. We point out, once more, that these works address a set-up that is different from the constraint-coupled set-up considered in this paper. Centralized dual approaches for constrained stochastic optimization problems arising in the context of resource allocation have been introduced in [17]. These approaches are based on [18], which shows how Lagrangian duality can be used in nonconvex stochastic resource allocation problems to turn them into a tractable form. In [19] a similar resource allocation set-up is solved combining online optimization with two different accelerated stochastic algorithms. Although the setup considered in [19] is distributed, it requires that the sparsity of the problem matches the communication graph structure and does not allow for coupling constraints involving all the agents, hampering applicability in more general contexts.

Contributions. Inspired by the deterministic literature, we design a novel distributed algorithm to solve constraint-coupled *stochastic* optimization problems over networks. Our approach relies on a proper application of Lagrange duality by which we are able to reformulate the associated dual problem as a separable, stochastic optimization problem. Although the considered set-up is *nonconvex*, under continuity of the cumulative distribution function of the random variables representing the source of uncertainty, it is possible to guarantee that strong duality holds. The proposed distributed algorithm consists of an iterative procedure in which each agent performs the following steps: *(i)* observe a new realization of the random variable, *(ii)* solve a small (deterministic) local problem, *(iii)* update the estimate of the dual variable by using a consensus iteration and an ascent step. Notably, the proposed algorithm scales well with the network size, i.e., the amount of local computation stays constant as the number of agents increase. We formally prove that asymptotically agents reach consensus on a dual optimal solution. By virtue of strong duality, each agent is able to eventually retrieve a local policy that satisfies both local and average coupling constraints. A key feature of the proposed algorithm is that, along its evolution, it provides online solution estimates that satisfy the coupling constraints in an ergodic sense. Finally, a numerical example corroborates the theoretical findings.

Organization. In Section II, we describe the problem set-up with a motivating application. In Section III we present the proposed distributed algorithm, which is analyzed in Section IV. Simulation results are in Section V.

II. Stochastic Constraint-Coupled Optimization

In this section we formalize the stochastic constraint-coupled framework and present a motivating example.

A. Problem Set-up

Let w be a random vector with support set $\mathcal{W} \subseteq \mathbb{R}^q$ with w denoting a realization of w and let $\mathbb{E}[\cdot]$ denote the expectation operator with respect to w. We consider a network of N agents

This result is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 638992 – OPT4SMART). The authors are with the Department of Electrical, Electronic and Information Engineering, University of Bologna, Bologna, Italy, name.surname@unibo.it.

that must cooperatively solve the stochastic constraint-coupled problem

$$\begin{aligned} \min_{\chi_1(\cdot),\dots,\chi_N(\cdot)} \ & \sum_{i=1}^{N} \mathbb{E}\left[f_i(\chi_i(\mathrm{w});\mathrm{w})\right] \\ \text{subj. to } & \chi_i(w) \in X_i, \qquad \forall w \in \mathcal{W}, \forall i \\ & \sum_{i=1}^{N} \mathbb{E}\left[g_i(\chi_i(\mathrm{w});\mathrm{w})\right] \leq 0, \end{aligned} \tag{1}$$

where $\chi_i : \mathcal{W} \to \mathbb{R}^{n_i}$ is the local *policy* to be optimized, $f_i : \mathbb{R}^{n_i} \times \mathcal{W} \to \mathbb{R}$ is the i-th cost function, $g_i : \mathbb{R}^{n_i} \times \mathcal{W} \to \mathbb{R}^p$ is the i-th contribution to the inequality constraints, and $X_i \subseteq \mathbb{R}^{n_i}$ is the i-th set of individual constraints. We assume problem (1) is feasible with optimal cost $f^\star$. The random vector w represents the state of the environment and, as such, influences agents' decisions. That is, for all i, the policy $\chi_i(\cdot)$ yields a different outcome for each realization w of w.

In the considered distributed setting, agents aim to solve the problem without a centralized coordinator. They are assumed communicate according to a connected and undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \{1, \dots, N\}$ being the set of nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ being the set of edges. If $(i, j) \in \mathcal{E}$, then also also $(j, i) \in \mathcal{E}$ and agents i and j can exchange information. We denote by $\mathcal{N}_i$ the set of *neighbors* of agent i in $\mathcal{G}$, i.e., $\mathcal{N}_i = \{j \in \{1, \dots, N\} \mid (i, j) \in \mathcal{E}\}$. Agents are assumed to have a *partial knowledge* of problem (1). Specifically, each agent i is assumed to know only the functions f_i, g_i and the set X_i. Denoting by $(\chi_1^\star(\cdot), \dots, \chi_N^\star(\cdot))$ an optimal policy for problem (1), the goal is that each agent i computes its local policy $\chi_i^\star(\cdot)$.

We point out that, even under convexity of f_i, g_i and X_i, problem (1) is in general a nonconvex program due to the generality of the distribution of w. Moreover, the policies $\chi_i(\cdot)$ to be optimized are infinite dimensional variables. These issues make this set-up more challenging than other distributed set-ups investigated in the literature (such as [2], [8]), where typically the optimization variables are finite dimensional and the considered problems are convex. The approach that we will adopt allows us to turn problem (1) into a tractable form by relying on duality theory.

B. Motivating Application: Distributed Steady-state Control

Let us consider a distributed control scenario with N discrete-time linear dynamical systems that want to solve the optimization problem

$$\min_{\substack{z_1(\cdot),\dots,z_N(\cdot)\\ u_1(\cdot),\dots,u_N(\cdot)}} \sum_{i=1}^{N} \mathbb{E}\big[z_i(\mathrm{w})^\top Q_i z_i(\mathrm{w}) + u_i(\mathrm{w})^\top R_i u_i(\mathrm{w})\big] \tag{2a}$$

$$\text{subj. to } z_i(w) = A_i z_i(w) + B_i u_i(w), \qquad \forall w, \forall i \tag{2b}$$

$$\underline{z}_i \leq z_i(w) \leq \bar{z}_i,\ \underline{u}_i \leq u_i(w) \leq \bar{u}_i, \quad \forall w, \forall i \tag{2c}$$

$$\sum_{i=1}^{N} \mathbb{E}\left[C_i z_i(\mathrm{w}) + D_i u_i(\mathrm{w}) + H_i \mathrm{w}\right] \leq 0, \tag{2d}$$

where (2b) is the equilibrium dynamics of each system i, (2c) represent box constraints on the state and input of agent i, and the matrices Q_i, R_i, C_i, D_i, H_i have proper dimensions. The purpose of problem (2) is to determine for all i an optimal state-input map $(z_i^\star(\cdot), u_i^\star(\cdot))$, which, based on the influence of the environment (modeled by w), determines an equilibrium pair that each system i must stabilize. Such state-input map must minimize a given (quadratic) cost figure and satisfy local box constraints and average coupling constraints (2d). Problems in the form (2) typically arise, e.g., when controlling microgrids where the objective is to optimize the consumption schedule on a daily basis (see also [11]), in such a case the random vector w can be used to model the uncertainty due to unknown power production of renewable energy sources.

III. Distributed Stochastic Dual Subgradient

In this section we propose our novel Distributed Stochastic Dual Subgradient algorithm. We first introduce the dual approach and then provide the algorithm description.

A. Lagrange Duality Approach

The main idea to solve problem (1) is to formulate its dual problem and then reconstruct the primal solution. The procedure presented here is inspired to a duality-based stochastic approach proposed for resource allocation problems in [17]–[19]. Let us associate the multiplier $\mu \in \mathbb{R}^p$ to the stochastic inequality constraint $\sum_{i=1}^{N} \mathbb{E}\left[g_i(\chi_i(\mathrm{w});\mathrm{w})\right] \leq 0$. The Lagrangian of problem (1) is

$$\mathcal{L}(\{\chi_i(\cdot)\}_{\forall i}, \mu) \triangleq \mathbb{E}\left[\sum_{i=1}^{N} f_i(\chi_i(\mathrm{w});\mathrm{w}) + \mu^\top g_i(\chi_i(\mathrm{w});\mathrm{w})\right]$$

where the local constraints $\chi_i(w) \in X_i$ are kept implicit and thus not dualized. With the Lagrangian at reach, the dual function can be defined as

$$\begin{aligned} q(\mu) = \min_{\{\chi_i(\cdot)\}_{\forall i}} \ & \mathcal{L}(\{\chi_i(\cdot)\}_{\forall i}, \mu) \\ \text{subj. to } & \chi_i(w) \in X_i, \qquad \forall w \in \mathcal{W}, \forall i. \end{aligned} \tag{3}$$

Thus, the dual problem of (1) can be posed as

$$\max_{\mu \geq 0} \ q(\mu). \tag{4}$$

We denote by $q^\star$ the optimal cost of problem (4) and by $\mu^\star$ an optimal solution (if any), such that $q^\star = q(\mu^\star)$. Irrespective of the properties of the primal problem (1), the dual problem (4) enjoys a number of appealing features, [20]. Indeed, problem (4) is always a concave program; moreover, by the weak duality theorem, its cost provides a lower bound on the optimal cost of (1), i.e., $q^\star \leq f^\star$.

Note that computing the value of the dual function in (3) requires the minimization of an expected value over the probability distribution of w. However, in general, the distribution of w may not be known, and even if this was the case, this task is intractable within an iterative optimization process. Therefore, we pursue a stochastic approximation approach that involves reformulating problem (4) into a stochastic problem. To this end, we make the following assumption on problem (1).

Assumption 3.1: The functions f_i and g_i are continuous and the sets X_i are compact, for all $i \in \{1, \dots, N\}$. □

Note that problem (2) satisfies Assumption 3.1 since the cost is quadratic, the coupling constraints (2d) are linear and the functions $z_i(\cdot)$ and $u_i(\cdot)$ must satisfy box constraints (2c).

We restrict our attention to policies $\chi_i(\cdot)$ that are sufficiently regular, so as to guarantee that the expected values are well defined, i.e., we assume that the random variables $f_i(\chi_i(\mathrm{w});\mathrm{w})$ and $g_i(\chi_i(\mathrm{w});\mathrm{w})$ are integrable with respect to the distribution of w. All these assumptions guarantee that the minimization in (3) is well defined for all $\mu \in \mathbb{R}^p$. The following lemma is a key result that legitimates a stochastic dual approach and is important for the subsequent analysis.

Lemma 3.1: Let Assumption 3.1 hold and define for all $\mu \geq 0$ and $w \in \mathcal{W}$ the function

$$\tilde{q}(\mu; w) \triangleq \min_{x_1 \in X_1, \ldots, x_N \in X_N} \sum_{i=1}^{N} [f_i(x_i; w) + \mu^\top g_i(x_i; w)], \quad (5)$$

where $x_i \in \mathbb{R}^{n_i}$ for all i. Then, the stochastic problem

$$\max_{\mu \geq 0} \mathbb{E}\big[\tilde{q}(\mu; \mathrm{w})\big] \quad (6)$$

is equivalent to problem (4), i.e., the two problems have the same optimal solution set and the same optimal cost.

Proof: We follow the same arguments of [17, Proposition 1]. Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be the probability space associated to w and let us denote the function in the expectation in the Lagrangian as $\phi(x; y) = \sum_{i=1}^{N} f_i(x_i; y) + \mu^\top g_i(x_i; y)$, where $x = (x_1, \ldots, x_N)$ and y is a generic variable playing the role of w. Let us use $\mathcal{X}_i$ to denote the set of locally feasible policies, i.e., $\mathcal{X}_i \triangleq \{\chi_i(\cdot) : \chi_i(w) \in X_i, \forall w \in \mathcal{W}\}$. We obtain

$$\begin{aligned} q(\mu) &= \min_{\{\chi_i(\cdot) \in \mathcal{X}_i\}_{\forall i}} \mathbb{E}\left[\phi(\chi(\mathrm{w}); \mathrm{w})\right] \\ &= \min_{\{\chi_i(\cdot) \in \mathcal{X}_i\}_{\forall i}} \int_\Omega \phi(\chi(\mathrm{w}(\omega)); \mathrm{w}(\omega)) d\mathbb{P}(\omega). \\ &\overset{(a)}{=} \int_\Omega \min_{\{x_i \in X_i\}_{\forall i}} \phi(x; \mathrm{w}(\omega)) d\mathbb{P}(\omega) \\ &= \mathbb{E}\left[\min_{\{x_i \in X_i\}_{\forall i}} \phi(x; \mathrm{w})\right] = \mathbb{E}\big[\tilde{q}(\mu; \mathrm{w})\big], \end{aligned}$$

where in (a) we used linearity of the integral to carry out the minimization for each value in the sample space Ω separately, moreover we renamed to x the (finite dimensional) optimization variable $\chi(\mathrm{w}(\omega))$ for a fixed value of $\omega \in \Omega$ for notational convenience. The proof follows. ∎

To see the impact of Lemma 3.1, note that, contrarily to the previous formulation of the dual problem (4), with the new formulation (6) the cost can be evaluated at any given $\mu \geq 0$ by solving a deterministic, finite dimensional minimization (cf. (5)), which is parametric in w, and then compute the expected value with respect to w (either using its probability distribution or by simulation).

B. Distributed Algorithm Derivation and Description

Now we present our novel Distributed Stochastic Dual Subgradient algorithm. An important consequence of Lemma 3.1 is that the dual problem (4) can now be solved with a stochastic approximation approach. Further note that the function in (5) can be reformulated as the sum of N terms. Formally, let us define for each i the function

$$q_i(\mu; w) \triangleq \min_{x_i \in X_i} \left(f_i(x_i; w) + \mu^\top g_i(x_i; w)\right),$$

for all $\mu \geq 0$ and $w \in \mathcal{W}$. Then, $\tilde{q}(\mu; w) = \sum_{i=1}^{N} q_i(\mu; w)$, from which we see that problem (6) can be written as

$$\max_{\mu \geq 0} \mathbb{E}\left[\sum_{i=1}^{N} q_i(\mu; w)\right]. \quad (7)$$

The formulation (7) of the dual problem lends itself well to consensus-based distributed optimization schemes.

Denote with $t \in \mathbb{Z}$ the iteration index and introduce a stochastic process $\{\mathrm{w}^t\}_{t \geq 0}$, with realizations $\{w^t\}_{t \geq 0}$ having values w^t independent and identically distributed according to the distribution of w. We assume that, at each time instant t, agents have access to the functions $f_i(\cdot, w^t)$ and $g_i(\cdot, w^t)$. Each agent maintains (i) a primal variable $x_i^t \in \mathbb{R}^{n_i}$ and (ii) a dual solution estimate $\mu_i^t \in \mathbb{R}^p$, both of which are iteratively updated. Once a new realization w^t has been observed, neighboring agents first combine their dual solution estimates in a consensus fashion (cf. (8a)) using certain weights a_{ij}. Then, agent i performs a local minimization (cf. (8b)) and the result is used to update the local dual solution estimate (cf. (8c)), where $\alpha^t \geq 0$ denotes the step-size and $[\cdot]_+$ denotes the Euclidean projection onto the positive orthant. In Algorithm 1 we provide a pseudocode of the distributed algorithm from the perspective of node i. We stress once again that the algorithm is fully distributed and that all the agents perform the same steps.

Algorithm 1 Distributed Stochastic Dual Subgradient

Initialization: $\mu_i^0 \geq 0$

For $t = 0, 1, 2, \ldots$

Observe current realization w^t

Receive μ_j^t from $j \in \mathcal{N}_i$ and compute

$$v_i^t = \sum_{j \in \mathcal{N}_i} a_{ij} \mu_j^t \quad (8a)$$

Perform primal and dual updates

$$x_i^t \in \operatorname*{argmin}_{x_i \in X_i} \left\{f_i(x_i; w^t) + {v_i^t}^\top g_i(x_i; w^t)\right\} \quad (8b)$$

$$\mu_i^{t+1} = \left[v_i^t + \alpha^t g_i(x_i^t; w^t)\right]_+ \quad (8c)$$

Let us now outline some properties of the proposed algorithm. Recall that the original problem (1) is nonconvex. Remarkably, Algorithm 1 provides a computationally affordable method to compute the optimal policy in a distributed fashion, while also enjoying attractive scalability properties. Indeed, even though the local problem (8b) is nonconvex, it is much smaller than the original one and its size does not depend on the number of agents. If convexity of f_i and g_i is additionally assumed, a solution to problem (8b) can then be found in a reasonably short time interval even though problem (1) remains nonconvex and intractable in general.

IV. Algorithm Analysis

In this section, we provide a convergence analysis for Algorithm 1. We first characterize strong duality and then we provide results on both dual and primal quantities.

A. Strong Duality Characterization

As already mentioned in Section III-A, problem (4) (and thus also problem (7)) has an optimal cost $q^\star$ that, in general, is less than or equal to the optimal cost $f^\star$ of the nonconvex problem (1). Nevertheless, it can be proven that strong duality stems from existing results given the particular problem structure provided that the cumulative distribution function (cdf) of the random variable w is continuous, or, equivalently, its probability density function does not contain any Dirac delta function. Formally, denoting the cdf of w as $F_{\mathrm{w}}(w) = \mathbb{P}(\mathrm{w} \leq w)$, we make the following assumption.

Assumption 4.1: The cdf $F_{\mathrm{w}}(\cdot)$ is continuous. □

Moreover, as usually done in distributed approaches based on Lagrange duality, we assume some regularity conditions on problem (1), namely Slater's constraint qualification.

Assumption 4.2: There exist functions $\bar{\chi}_i(\cdot), \dots, \bar{\chi}_N(\cdot)$ that are feasible for problem (1) and satisfy the strict inequality $\sum_{i=1}^N \mathbb{E}\left[g_i(\bar{\chi}_i(\mathrm{w}); \mathrm{w})\right] < 0$. □

The next proposition formalizes the strong duality result for our stochastic constraint-coupled optimization set-up.

Proposition 4.1: Let Assumptions 3.1, 4.1 and 4.2 hold. Then, the duality gap is zero, i.e., $f^\star = q^\star$. □

The proof of Proposition 4.1 is based on the so-called Lyapunov's convexity theorem and can be obtained by following arguments that are similar to [18, Thm. 1], and hence is omitted.

B. Dual Convergence

Before stating the convergence properties of our Distributed Stochastic Dual Subgradient, we introduce assumptions on the network weights (cf. (8a)), on the step-size (cf. (8c)) and on the (sub)gradients of the dual function (3).

Assumption 4.3: The weights a_{ij}, $i, j \in \{1, \dots, N\}$ match the graph $\mathcal{G}$, i.e., $a_{ij} \neq 0$ for all $(i,j) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise. Moreover, they satisfy (i) $\sum_{j=1}^N a_{ij} = 1$ for all i, (ii) $\sum_{i=1}^N a_{ij} = 1$ for all j, (iii) $a_{ii} > 0$ for all i. □

We point out that Assumption 4.3 can be relaxed by using different consensus schemes that accommodate more general (e.g. directed and/or time-varying) networks, see e.g. [2].

Assumption 4.4: The step-size sequence $\{\alpha^t\}_{t\geq 0}$ is such that $0 \leq \alpha^{t+1} \leq \alpha^t$ for all $t \geq 0$, $\sum_{t=0}^\infty \alpha^t = \infty$, $\sum_{t=0}^\infty (\alpha^t)^2 < \infty$ and $\lim_{t\to\infty} t\alpha^t = \infty$. □

An example of step-size sequence satisfying Assumption 4.4 is $\alpha^t = K/t^\gamma$, with any $K > 0$ and $\frac{1}{2} \leq \gamma < 1$.

Assumption 4.5: The stochastic (sub)gradients are uniformly bounded, i.e., for all $i \in \{1, \dots, N\}$ it holds $\|g_i(x_i; w)\| \leq C_i$ for all $x_i \in X_i$ and $w \in \mathcal{W}$. □

We are now ready to provide the convergence result.

Theorem 4.2 (Dual convergence): Let Assumptions 3.1, 4.1, 4.2, 4.3, 4.4, and 4.5 hold and consider the sequences $\{x_1^t, \dots, x_N^t\}_{t\geq 0}$ and $\{\mu_1^t, \dots, \mu_N^t\}_{t\geq 0}$ generated by Algorithm 1. Then, almost surely, there hold

$$\lim_{t\to\infty} \sum_{i=1}^N \mathbb{E}[q_i(\mu_i^t; \mathrm{w}^t)] = q^\star, \tag{9a}$$

$$\lim_{t\to\infty} \mu_i^t = \mu^\star, \qquad \forall i \in \{1, \dots, N\}, \tag{9b}$$

for some $\mu^\star \in \mathbb{R}^p$ optimal solution of problem (4). □

Before proceeding with the proof, we point out that all the quantities generated by Algorithm 1 (i.e. x_i^t, μ_i^t, v_i^t) must be regarded as realizations of random processes, since they depend on the specific realization of the process w^t.

Proof: Let us define $\bar{q}_i(\mu) = \mathbb{E}[q_i(\mu; \mathrm{w})]$. First, we claim that the (sub)gradient estimators are unbiased, i.e.,

$$\mathbb{E}\left[g_i(x_i^t; \mathrm{w}^t)\right] = \widetilde{\nabla}\bar{q}_i\left(v_i^t\right). \tag{10}$$

To see this, let us follow the line of proof in [17, Prop. 1]. By (8b), the value of the function $\bar{q}_i$ at v_i^t is

$$\bar{q}_i(v_i^t) = \mathbb{E}\left[f_i(x_i^t; \mathrm{w}^t)\right] + {v_i^t}^\top \mathbb{E}\left[g_i(x_i^t; \mathrm{w}^t)\right].$$

Similarly, the value of the function $\bar{q}_i$ at a generic $\mu \in \mathbb{R}^p$ is

$$\begin{aligned}\bar{q}_i(\mu) &= \mathbb{E}\left[\min_{x_i \in X_i} \left(f_i(x_i; \mathrm{w}) + \mu^\top g_i(x_i; \mathrm{w})\right)\right] \\ &\leq \mathbb{E}\left[f_i(x_i^t; \mathrm{w})\right] + \mu^\top \mathbb{E}\left[g_i(x_i^t; \mathrm{w})\right],\end{aligned}$$

where the inequality follows since the expression $f_i(x_i; \mathrm{w}) + \mu^\top g_i(x_i; \mathrm{w})$ at any $x_i \in X_i$ is greater than or equal to its minimum over X_i. Subtracting the last two expressions we obtain that, by definition, $\mathbb{E}\left[g_i(x_i^t; w^t)\right]$ is a (sub)gradient of $-\tilde{q}_i$ at v_i^t.

Second, we show that the dual estimates μ_i^t are asymptotically consensual. To do so, rewrite the update of μ_i^t as $\mu_i^{t+1} = v_i^t + \epsilon_i^t$, with $\epsilon_i^t = \left[v_i^t + \alpha^t g_i(x_i^t; \mathrm{w}^t)\right]_+ - v_i^t$. Let us define the mean of the dual variables among the agents as $\bar{\mu}^t \triangleq \frac{1}{N}\sum_{i=1}^N \mu_i^t$ for all $t \geq 0$. The update of $\bar{\mu}^t$ reads

$$\bar{\mu}^{t+1} = \frac{1}{N}\sum_{i=1}^N \left(\sum_{j\in\mathcal{N}_i} a_{ij}\mu_j^t + \epsilon_i^t\right) = \bar{\mu}^t + \frac{1}{N}\sum_{i=1}^N \epsilon_i^t,$$

where the equality follows by Assumption 4.3. By denoting the column stack $\mu^t = (\mu_1^t, \dots, \mu_N^t)$, $\epsilon^t = (\epsilon_1^t, \dots, \epsilon_N^t)$ and the weight matrix $A \in \mathbb{R}^{Np\times Np}$ with block entry (i,j) equal to $a_{ij}I_p$ (I_p is the p-by-p identity), the consensus error is

$$\begin{aligned}\mu^{t+1} - \mathbf{1}\bar{\mu}^{t+1} &= A\mu^t + \epsilon^t - \mathbf{1}\left(\bar{\mu}^t + \frac{1}{N}\sum_{i=1}^N \epsilon_i^t\right) \\ &= (A - \mathbf{1}\mathbf{1}^\top/N)(\mu^t - \mathbf{1}\bar{\mu}^t) + (I - \mathbf{1}\mathbf{1}^\top/N)\epsilon^t\end{aligned}$$

in which $\mathbf{1} \in \mathbb{R}^{Np\times p}$ is the vertical stack of N times I_p. By taking the norm of both sides of the previous equality and by using the properties of the consensus weight matrix A (see also the proof of [11, Theorem 2.5]), it follows that

$$\left\|\mu^{t+1} - \mathbf{1}\bar{\mu}^{t+1}\right\| \leq \sigma\left\|\mu^t - \mathbf{1}\bar{\mu}^t\right\| + \left\|\epsilon^t\right\|,$$

where $\sigma \in (0,1)$ is the contraction factor associated to the consensus weight matrix A. Due to Assumptions 4.4 and 4.5, it can be proven that $\|\epsilon^t\|$ goes to 0 as $t \to \infty$ (cf. [11, Section

2.1]). It follows that $\lim_{t\to\infty}\sum_{i=1}^N \|\mu_i^t - \mathbf{1}\bar{\mu}^t\| = 0$, which concludes the proof of consensus of the dual estimates. Next, we show that the vectors μ_i^t converge to an optimal solution of problem (4). Let us analyze the aggregate evolution of μ_i^t for all $i \in \{1,\dots,N\}$

$$\sum_{i=1}^N \left\|\mu_i^{t+1} - \mu^\star\right\|^2 \overset{(a)}{\leq} \sum_{i=1}^N \left\|v_i^t + \alpha^t g_i(x_i^t;\mathrm{w}^t) - \mu^\star\right\|^2$$
$$\overset{(b)}{\leq} \sum_{i=1}^N \Big(\left\|\mu_i^t - \mu^\star\right\|^2 + (\alpha^t)^2 C_i^2 + 2\alpha^t \left(v_i^t - \mu^\star\right)^\top g_i(x_i^t;\mathrm{w}^t)\Big),$$

where in (a) we used (8c) and the non-expansiveness property of the projection [20] and in (b) we expanded the squared norm and we used Assumption 4.5.

After defining $C^2 = \sum_{i=1}^N C_i^2$ and after taking the conditional expected value with respect to all the history up to time t, represented by $\mathcal{F}^t = \{\{\mu_i^0\}_i, w^1, \{x_i^1\}_i, \{\mu_i^1\}_i, \dots, w^t, \{x_i^t\}_i, \{\mu_i^t\}_i\}$, we obtain

$$\mathbb{E}\Big[\sum_{i=1}^N \|\mu_i^{t+1} - \mu^\star\|^2 \mid \mathcal{F}^t\Big] \leq \sum_{i=1}^N \|\mu_i^t - \mu^\star\|^2 + (\alpha^t)^2 C^2$$
$$+ \mathbb{E}\Big[2\alpha^t \sum_{i=1}^N \left(v_i^t - \mu^\star\right)^\top g_i(x_i^t;\mathrm{w}^t) \mid \mathcal{F}^t\Big]$$
$$= \sum_{i=1}^N \|\mu_i^t - \mu^\star\|^2 + (\alpha^t)^2 C^2$$
$$+ 2\alpha^t \sum_{i=1}^N \left(v_i^t - \mu^\star\right)^\top \mathbb{E}\left[g_i(x_i^t;\mathrm{w}^t) \mid \{\mu_j^t\}_{j\in\mathcal{N}_i}\right],$$

where in the last equality we used the fact that, due to the conditional expectation, the value of v_i^t is fixed. Now, by using (10) with the definition of subgradient of $\sum_{i=1}^N \tilde{q}_i(v_i^t)$ we obtain $\sum_{i=1}^N \left(v_i^t - \mu^\star\right)^\top \mathbb{E}\left[g_i(x_i^t;\mathrm{w}^t) \mid \{\mu_j^t\}_{j\in\mathcal{N}_i}\right] \leq \left(\sum_{i=1}^N \tilde{q}_i\left(v_i^t\right) - q^\star\right)$, which we plug in the previous inequality to obtain

$$\mathbb{E}\left[\sum_{i=1}^N \|\mu_i^{t+1} - \mu^\star\|^2 \mid \mathcal{F}^t\right]$$
$$\leq \sum_{i=1}^N \|\mu_i^t - \mu^\star\|^2 + (\alpha^t)^2 C^2 + 2\alpha^t \left(\sum_{i=1}^N \tilde{q}_i\left(v_i^t\right) - q^\star\right)$$
$$= \sum_{i=1}^N \|\mu_i^t - \mu^\star\|^2 + (\alpha^t)^2 C^2 + 2\alpha^t \left(\sum_{i=1}^N \tilde{q}_i(\bar{\mu}^t) - q^\star\right)$$
$$- 2\alpha^t \sum_{i=1}^N \left(\tilde{q}_i(\bar{\mu}^t) - \tilde{q}_i\left(v_i^t\right)\right),$$

where in the last equality we added and subtracted $2\alpha^t \sum_{i=1}^N \tilde{q}_i(\bar{\mu}^t)$. Following similar arguments as in [11, Theorem 2.5], it can be seen that $\sum_{t=0}^\infty \{\alpha^t \sum_{i=1}^N \left(\tilde{q}_i(\bar{\mu}^t) - \tilde{q}_i\left(v_i^t\right)\right)\}_{t\geq 0} < \infty$ is summable. Thus, by using the Supermartingale convergence theorem [21, Proposition 8.2.10] and by following similar arguments as in [21, Proposition 8.2.13], it follows that, almost surely, $\lim_{t\to\infty} \bar{\mu}^t = \mu^\star$ and $\lim_{t\to\infty} q(\bar{\mu}^t) = q^\star$. Then, the proof follows by using the consensus result $\lim_{t\to\infty}\sum_{i=1}^N \|\mu_i^t - \mathbf{1}\bar{\mu}^t\| = 0$. ■

We are now able to clarify in what sense the iterations of Algorithm 1 lead to the computation of the optimal policy $(\chi_1^\star(\cdot),\dots,\chi_N^\star(\cdot))$ of problem (1) (cf. Section III). As we just shown, the sequences $\{\mu_i^t\}_{t\geq 0}$ converge to a dual optimal solution $\mu^\star$. Thus, step (8b) is "asymptotically performed" with $\mu^\star$ in place of v_i^t. Together with Proposition 4.1, this means that the minimization in (8b) eventually yields the optimal policy for given $w \in \mathcal{W}$, i.e.,

$$\chi_i^\star(w) \in \operatorname*{argmin}_{x_i\in X_i} \left\{f_i(x_i;w) + {\mu^\star}^\top g_i(x_i;w)\right\}.$$

C. Ergodic Primal Feasibility

Next, we provide a feasibility result of the primal iterates.

Theorem 4.3: In the same setting of Theorem 4.2, it holds $\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^N g_i(x_i^t;\mathrm{w}^t) \leq 0$ almost surely.

Proof: By using the update rule (8c), we have $\mu_i^{t+1} = [v_i^t + \alpha^t g_i(x_i^t;\mathrm{w}^t)]_+ \geq v_i^t + \alpha^t g_i(x_i^t;\mathrm{w}^t)$, for all $t \geq 0$. Reorganizing the previous inequality and summing over $i \in \{1,\dots,N\}$, it follows that, for all $t \geq 0$,

$$\sum_{i=1}^N g_i(x_i^t;\mathrm{w}^t) \leq \sum_{i=1}^N \frac{\mu_i^{t+1} - \sum_{j\in\mathcal{N}_i} a_{ij}\mu_j^t}{\alpha^t} = \sum_{i=1}^N \frac{\mu_i^{t+1} - \mu_i^t}{\alpha^t}.$$

By summing both sides of the previous inequality over t from 0 to $T-1$, we obtain

$$\sum_{t=0}^{T-1}\sum_{i=1}^N g_i(x_i^t;\mathrm{w}^t) \leq \sum_{t=0}^{T-1}\sum_{i=1}^N \frac{\mu_i^{t+1} - \mu_i^t}{\alpha^t} \leq \sum_{t=0}^{T-1}\sum_{i=1}^N \left(\frac{\mu_i^{t+1}}{\alpha^{t+1}} - \frac{\mu_i^t}{\alpha^t}\right)$$
$$= \frac{\sum_{i=1}^N \mu_i^T}{\alpha^T} - \frac{\sum_{i=1}^N \mu_i^0}{\alpha^0} \qquad \forall T \geq 1,$$

where the second inequality follows by the assumption $\alpha^{t+1} \leq \alpha^t$ and from non-negativity of μ_i^t for all $t \geq 0$ (by construction). Dividing by T yields, for all $T \geq 1$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^N g_i(x_i^t;\mathrm{w}^t) \leq \frac{1}{T}\left(\frac{\sum_{i=1}^N \mu_i^T}{\alpha^T} - \frac{\sum_{i=1}^N \mu_i^0}{\alpha^0}\right).$$

Since the sequences $\{\mu_i^t\}_{t\geq 0}$ converge almost surely for all $i \in \{1,\dots,N\}$ (Theorem 4.2), it follows that the sequences $\{\mu_i^t\}_{t\geq 0}$ are bounded. By taking the limit of both sides as T goes to infinity and using the assumption $\lim_{t\to\infty} t\alpha^t = \infty$, the right-hand side goes to zero, and the proof follows. ■

Note that the result provided by Theorem 4.3 is about the non-positivity of the limiting time average of the coupling constraint. Under ergodicity assumptions of the sequence $\{\sum_{i=1}^N g_i(x_i^t;\mathrm{w}^t)\}_{t\geq 0}$, this is equivalent to the constraint of problem (1) expressed in terms of expected value.

V. Numerical Example

In this section, we provide a numerical study that focuses on the motivating example presented in Section II-B. We consider $N = 30$ MIMO systems with state $z_i \in \mathbb{R}^5$ and input $u_i \in \mathbb{R}^3$. The random vector w is uniformly distributed in $\mathcal{W} = [0,1]^5$. The system matrices $A_i \in \mathbb{R}^{5\times 5}$, $B_i \in \mathbb{R}^{5\times 3}$, $C_i \in \mathbb{R}^{5\times 5}$, $D_i \in \mathbb{R}^{5\times 3}$ and $H_i \in \mathbb{R}^{5\times 5}$ are randomly generated with entries in $(0,1)$. The lower bounds $\underline{z}_i, \underline{u}_i$ have random values in $(-10,-5)$ while the upper bounds $\bar{z}_i, \bar{u}_i$ are random in

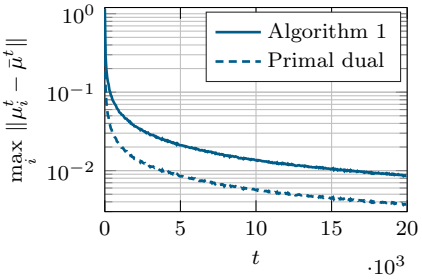

Fig. 1. Consensus error measured as the distance from the local dual estimates $\boldsymbol{\mu_i^t}$ to their mean $\bar{\boldsymbol{\mu}}^{\boldsymbol{t}}$.

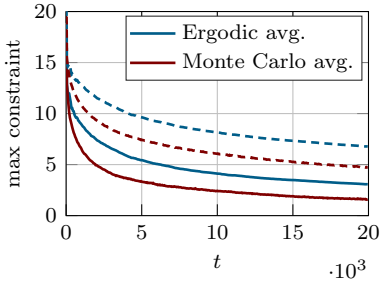

Fig. 2. Coupling constraint value. Ergodic avg is $\frac{1}{t}\sum_{\tau=1}^{t}\sum_{i=1}^{N} \boldsymbol{g_i(x_i^t, w^t)}$, Monte Carlo avg is $\mathbb{E}[\sum_{i=1}^{N} \boldsymbol{g_i(x_i^t, w)}]$ (estimated).

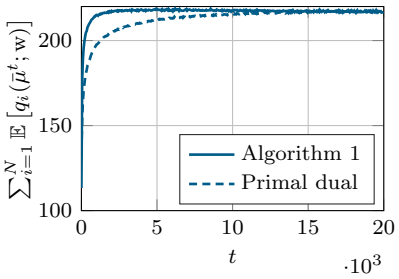

Fig. 3. Dual cost computed on the average dual variable $\bar{\boldsymbol{\mu}}^{\boldsymbol{t}}$. Expected values are estimated through Monte Carlo simulation.

$(5, 10)$. The matrices $Q_i \in \mathbb{R}^{5\times 5}$ and $R_i \in \mathbb{R}^{3\times 3}$ are identities scaled by a random factor in $(1, 5)$. Notice that Assumption 4.1 is satisfied therefore strong duality holds.

We run Algorithm 1 (solid lines in plots) and compare with a stochastic distributed primal-dual algorithm (dashed lines in plots) inspired to [4]. Agents communicate according to an Erdős-Rényi graph with edge probability 0.2. As for the weights a_{ij}, we use the Metropolis-Hastings rule, and initialize the local dual estimates μ_i^0 to random non-negative values. The step-size is $\alpha^t = 1/t^{0.7}$.

In Figure 1, the consensus error is shown as a function of the iteration t. The figure highlights that, as the algorithm evolves, the local dual estimates μ_i^t of the agents get closer to their mean, as expected from the convergence result of Theorem 4.2. Figure 2 shows horizontal and vertical averages of the maximum coupling constraint value (across its components) as a function of t. Both the Monte Carlo average (computed on 10^4 trials) and the time average asymptotically go to zero, as expected from Theorem 4.3. However, due to the diminishing step-size, this convergence can become slow. In Figure 3, we plot the dual cost computed at the mean dual solution estimate $\bar{\mu}^t = 1/N \sum_{i=1}^{N} \mu_i^t$. The expected values are approximated with Monte Carlo simulation over 10^4 trials. As expected from a dual method, the picture emphasizes that the cost increases toward optimality with the typical sublinear rate of subgradient methods.

VI. Conclusions

In this paper we proposed a distributed algorithm to solve constraint-coupled stochastic optimization problems over networks. The proposed algorithm relies on a stochastic duality-based approach, where the dual problem is solved with a stochastic distributed subgradient method. We proved dual convergence and average primal feasibility, confirmed by numerical computations.

References

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

[2] A. Falsone, K. Margellos, S. Garatti, and M. Prandini, "Dual decomposition for multi-agent distributed optimization with coupling constraints," *Automatica*, vol. 84, pp. 149–158, 2017.

[3] A. Simonetto, T. Keviczky, and M. Johansson, "A regularized saddle-point algorithm for networked optimization with resource allocation constraints," in *IEEE Conference on Decision and Control (CDC)*, 2012, pp. 7476–7481.

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

[5] A. Cherukuri and J. Cortés, "Distributed generator coordination for initialization and anytime optimization in economic dispatch," *IEEE Trans. on Control of Network Sys.*, vol. 2, no. 3, pp. 226–237, 2015.

[6] Y. Zhang and M. M. Zavlanos, "A consensus-based distributed augmented lagrangian method," in *IEEE Conference on Decision and Control*, 2018, pp. 1763–1768.

[7] A. Falsone, I. Notarnicola, G. Notarstefano, and M. Prandini, "Tracking-ADMM for distributed constraint-coupled optimization," *Automatica*, vol. 117, p. 108962, 2020.

[8] I. Notarnicola and G. Notarstefano, "Constraint-coupled distributed optimization: A relaxation and duality approach," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 1, pp. 483–492, 2020.

[9] A. Camisa, F. Farina, I. Notarnicola, and G. Notarstefano, "Distributed constraint-coupled optimization via primal decomposition over random time-varying graphs," *arXiv preprint arXiv:2010.14489*, 2020.

[10] I. Necoara, "Random coordinate descent algorithms for multi-agent convex optimization over networks," *IEEE Transactions on Automatic Control*, vol. 58, no. 8, pp. 2001–2012, 2013.

[11] G. Notarstefano, I. Notarnicola, and A. Camisa, "Distributed optimization for smart cyber-physical networks," *Foundations and Trends® in Systems and Control*, vol. 7, no. 3, pp. 253–383, 2019.

[12] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Journal of Optim. Theory and Applications*, vol. 147, no. 3, pp. 516–545, 2010.

[13] K. Srivastava and A. Nedic, "Distributed asynchronous constrained stochastic optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 772–790, 2011.

[14] R. Xin, A. K. Sahu, U. A. Khan, and S. Kar, "Distributed stochastic optimization with gradient tracking over strongly-connected networks," in *IEEE Conf. on Decision and Control (CDC)*, 2019, pp. 8353–8358.

[15] M. I. Qureshi, R. Xin, S. Kar, and U. A. Khan, "S-ADDOPT: Decentralized stochastic first-order optimization over directed graphs," *IEEE Control Systems Letters*, vol. 5, no. 3, pp. 953–958, 2020.

[16] F. Farina and G. Notarstefano, "Randomized block proximal methods for distributed stochastic big-data optimization," *IEEE Transactions on Automatic Control*, 2020.

[17] A. Ribeiro, "Ergodic stochastic optimization algorithms for wireless communication and networking," *IEEE Transactions on Signal Processing*, vol. 58, no. 12, pp. 6369–6386, 2010.

[18] A. Ribeiro and G. B. Giannakis, "Separation principles in wireless networking," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4488–4505, 2010.

[19] T. Chen, Q. Ling, and G. B. Giannakis, "Learn-and-adapt stochastic dual gradients for network resource allocation," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 4, pp. 1941–1951, 2018.

[20] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.

[21] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex analysis and optimization*. Athena Scientific, 2003.