Noise-to-state exponentially stable distributed convex optimization on weight-balanced digraphs

David Mateos-Núñez Jorge Cortés

Abstract— **This paper studies the robustness under additive persistent noise of a class of continuous-time distributed algorithms for convex optimization. A group of agents, each with its own private objective function and communicating over a weight-balanced digraph, seeks to determine the global decision vector that minimizes the sum of all the functions. Under mild conditions on the local objective functions, we establish that the distributed algorithm is noise-to-state exponentially stable in second moment with respect to the optimal solution. Our technical approach combines notions and tools from graph theory, stochastic differential equations, and Lyapunov stability analysis. Simulations illustrate our results.**

I. Introduction

Finding the minimizer of a sum of convex functions in a distributed way has been the focus of many works in the literature. Two main families of problems arise in this area: either the local objectives depend only on local variables that might be globally constrained, giving rise to network utility maximization problems; or the local objectives depend on a global decision vector. Here we focus on the second class of problems. Applications include large-scale optimization, machine learning,rendezvous and tracking in motion coordination of multi-agent systems, and source-localization in wireless sensor networks.

Literature review: Different networks (composed by either plain processors, sensors, or mobile robots) call for different modeling requirements and design techniques, so we offer here a partial list of references as a guide for the reader. Some algorithms evolve in discrete jumps with associated gradient stepsize that is vanishing [1], [2], [3], nonvanishing [4], [5], and/or might require to solve a local optimization at each time step [1], [6], [3], [7]; others evolve in continuous time [8], [9]; and some are hybrid [10]. Most algorithms converge asymptotically to the solution while others converge to an (arbitrarily good) approximation [4], [5]. Some examples of convergence rates, or size of the cost error as a function of the number of iterations, are $1/\sqrt{k}$ [1], [3] and $1/k$ [6]. The communication topologies might be undirected [8, Sec. IV], [4], [6], [9], [7], directed and weight-balanced (when the adjacency matrix is doubly stochastic) [1], [8, Sec. V], [2], [5], or just directed [3]; also, they can be fixed [8], [6], [9], [7], or change randomly over time (under a periodic connectivity assumption) [1], [4], [2], [3], [5]. On the other hand, the objective functions might range from being twice differentiable [9] or once differentiable [8, Sec. V], [7], to just Lipschitz [1], [8, Sec. IV], [4], [6], [2], [3], [5] (with bounded gradients [4], [5]); in addition, they might need to be strongly convex [9], strictly convex [6], [7] (uniformly), or just convex [1], [8, Sec. IV], [4], [2], [3], [5]. Some algorithms use the Hessian of the objective functions in addition to the gradients [9], [7]. Also, the agents might need to share their gradients [9] (in some instances) or even their objectives [7]. Some incorporate a global constraint known to all the agents using a projection method [1], [2], [3], [5] or a dual method [7], and in same cases each agent has a different constraint [2]. Remarkably, a family of algorithms impose a constraint on the initial condition [9], [7]. Some works consider additive noise affecting the dynamics through stochastically perturbed gradients with associated vanishing stepsize [1] or nonvanishing stepsize [5]. Finally, the algorithms can be synchronous [8], [6], [9] or allow gossip/randomized communication [1], [4], [2], [3], [5], [7], or might even use event-triggered communication [10].

There is lack of works in the literature merging continuous-time dynamics and additive persistent noise in the communication channels. These, however, might correspond to modeling requirements in applications like rendezvous and tracking in networks of mobile robots, where the motion coordination task is coupled with an optimization of local performance measures, and also in the calibration of networked analog systems, like satellites. Moreover, the convergence rate guarantees in works like [1], [3], [6] consider the error cost, rather than the estimates' distance to the global decision vector, which is relevant in the aforementioned applications. In this paper, we build on the approach presented in [11], [8] to address these novel features.

Statement of contributions: We study the robustness under additive persistent noise of a family of continuous-time distributed algorithms for convex optimization. This type of noise might be due to errors in the communication channels or in the computations performed by the agents. We show how the this family of algorithms allow the agents' estimates to converge exponentially fast, in a precise stochastic sense, to a neighborhood of the optimal solution, and the size of that neighborhood depends on the size of noise. Specifically, we establish that the evolution of the agents' estimates is noise-to-state exponentially stable in second moment with respect to the optimum global decision vector. As part of our technical approach, we study the interaction between local optimization and local consensus through the co-coercivity

The authors are with the Department of Mechanical and Aerospace Engineering, University of California, San Diego, CA 92093, USA, {dmateosn,cortes}@ucsd.edu.

of a family of vector fields that are the sum of a gradient of a convex function plus a nonsymmetric Laplacian. Specifically, we give sufficient conditions for this family of vector fields to be co-coercive under a class of linear transformations. These techniques allow us to overcome the challenges posed by directed communication topologies and channels affected by additive persistent noise. Most proofs are omitted for reasons of space and will appear elsewhere.

Organization: The paper is organized as follows. Section II introduces preliminary notions on graph theory and stochastic differential equations. Section III formulates the problem of interest. Section IV presents our main results as well as illustrative simulations. Finally, Section V discusses our conclusions and ideas for future work. The appendix gathers a relevant result for our technical approach.

II. Preliminaries

Here we introduce some notations and review basic notions on graph theory and stochastic differential equations.

Notational conventions: We let $\mathbb{R}$ and $\mathbb{R}_{\geq 0}$ be the sets of real and nonnegative real numbers, respectively. We also define the following vectors in $\mathbb{R}^n$: $\mathbb{1}_n \triangleq [1,\dots,1]^\top$, $0_n \triangleq [0,\dots,0]^\top$, and e_i has a 1 in the ith entry and the rest of its entries are zero; also, I_n is the identity matrix in $\mathbb{R}^{n\times n}$. We denote by $\|.\|_2$ the Euclidean norm for vectors. The Frobenius norm of the matrix $B \in \mathbb{R}^{n\times m}$ is $|B|_{\mathcal{F}} \triangleq \sqrt{\mathrm{trace}\,(B^\top B)} = \sqrt{\mathrm{trace}\,(BB^\top)}$. We also define $\|x\|_B \triangleq \|Bx\|_2$, which provides a seminorm on $\mathbb{R}^m$ whose nullset is the nullspace of B, defined as $\mathcal{N}(B) = \{x \in \mathbb{R}^m : Bx = 0\}$. (Some authors prefer to define $\|x\|_A \triangleq \sqrt{x^\top Ax}$, but this has the inconvenience that A has to be symmetric and positive semidefinite.) For a symmetric real matrix $A \in \mathbb{R}^{n\times n}$, we order its eigenvalues as $\lambda_{\max}(A) \triangleq \lambda_1(A) \geq \cdots \geq \lambda_n(A) \triangleq \lambda_{\min}(A)$. Similarly, we order the singular values of any matrix $B \in \mathbb{R}^{n\times m}$ as $\sigma_{\max}(B) \triangleq \sigma_1(B) \geq \cdots \geq \sigma_r(B) \triangleq \sigma_{\min}(A)$, where $r = \mathrm{rank}(B)$ is the rank of B. We recall that $\sigma_i(B) = \sqrt{\lambda_i(B^\top B)}$. Given a vector v whose entries are matrices, $\mathrm{diag}(v)$ is a block diagonal matrix with the blocks in v. The Kronecker product of any two matrices is denoted by $A \otimes B$.

A function $f : X \to \mathbb{R}$, where X is a convex set, is convex if $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for each $x, y \in X$ and any $\lambda \in [0,1]$; and f is concave if $-f$ is convex. A function $f : X_1 \to X_2$, for normed vector spaces X_1, X_2, is Lipschitz with constant κ if $\|f(x) - f(y)\|_{X_1} \leq \kappa\|x - y\|_{X_2}$ for each $x, y \in X_1$, where $\|.\|_X$ is the norm in the vector space X. If $f : \mathbb{R}^n \to \mathbb{R}$ is continuously twice differentiable, we denote its gradient and Hessian by ∇f and $\nabla^2 f$, respectively. Given a differentiable vector field $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$, we let $\mathbf{DF} : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ denote its Jacobian, where $\mathbf{DF}(x)_{ij} = \frac{\partial \mathbf{F}_i(x)}{\partial x_j}$ for all $x \in \mathbb{R}^n$.

Graph theory: The following notions in graph theory follow the exposition in [12]. A weighted digraph $\mathcal{G} = (\mathcal{I}, \mathcal{E}, \mathsf{A})$ is a triplet where $\mathcal{I} = \{1,\dots,N\}$ is the vertex set, $\mathcal{E} \subseteq \mathcal{I}\times\mathcal{I}$ is the edge set, and $\mathsf{A} \in \mathbb{R}_{\geq 0}^{N\times N}$ is the weighted adjacency matrix with the property that $\mathsf{a}_{ij} > 0$ if and only if $(i,j) \in \mathcal{E}$. The Laplacian matrix $\mathsf{L} \in \mathbb{R}^{N\times N}$ of $\mathcal{G}$ is $\mathsf{L} \triangleq \mathrm{diag}(\mathsf{A}\mathbb{1}_N) - \mathsf{A}$, which satisfies $\mathsf{L}\mathbb{1}_N = 0_N$. The complete graph is the digraph with edge set $\mathcal{E}_{\mathcal{K}} = \mathcal{I}\times\mathcal{I}$. For convenience, we let $\mathsf{L}_{\mathcal{K}}$ denote the Laplacian of the complete graph with edge weights $1/N$. Note that $\mathsf{L}_{\mathcal{K}} = \mathrm{I}_N - \mathrm{M}$, where $\mathrm{M} \triangleq \frac{1}{N}\mathbb{1}_N\mathbb{1}_N^\top$, and that $\mathsf{L}_{\mathcal{K}}$ is idempotent, i.e., $\mathsf{L}_{\mathcal{K}}^2 = \mathsf{L}_{\mathcal{K}}$. The weighted out-degree and in-degree of $i \in \mathcal{I}$ are, respectively, $d_{\text{out}}(i) = \sum_{j=1}^N \mathsf{a}_{ij}$ and $d_{\text{in}}(i) = \sum_{j=1}^N \mathsf{a}_{ji}$. A digraph is weight-balanced if $d_{\text{out}}(i) = d_{\text{in}}(i)$ for all $i \in \mathcal{I}$, that is, $\mathbb{1}_N^\top\mathsf{L} = 0_N$, which is also equivalent to the condition of $\mathsf{L} + \mathsf{L}^\top$ being positive semidefinite. A path is an ordered sequence of vertices such that any pair of vertices appearing consecutively is an edge. A digraph is strongly connected if there is a path between any pair of distinct vertices.

Stochastic differential equations: Informally speaking, a stochastic differential equation (SDE) [13], [14], [15] is an ordinary differential equation driven by a "random process" called Brownian motion, $\mathrm{B} : \Omega \times [t_0,\infty) \to \mathbb{R}^m$. Here, Ω is the outcome space and $\mathbb{P}$ is a probability measure defined on the sigma-algebra $\mathcal{F}$ of measurable events (subsets) of Ω. Together they form the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For each outcome $\omega \in \Omega$, the mapping $\mathrm{B}(\omega,.) : [t_0,\infty) \to \mathbb{R}^m$ is called a sample path of the Brownian motion and is continuous with probability 1 and with $\mathrm{B}(.,t_0) = 0$; and for each time $t \in [t_0,\infty)$, the function $\mathrm{B}(t) \triangleq \mathrm{B}(.,t) : \Omega \to \mathbb{R}^m$ is a random variable such that the increments $\mathrm{B}(t) - \mathrm{B}(s)$ have a multivariate Gaussian distribution of zero mean and covariance $(t-s)\mathrm{I}_m$ and are independent from $\mathrm{B}(s)$ for all $t_0 \leq s < t$. The SDE

$$\mathrm{d}x(\omega,t) = g(x(\omega,t),t)\mathrm{d}t + G(x(\omega,t),t)\Sigma(t)\mathrm{dB}(\omega,t) \quad (1)$$

can be regarded as the limiting case of the recurrence

$$\begin{aligned} x(\omega,t_{k+1}) = x(\omega,t_k) + g\big(x(\omega,t_k),t_k\big)(t_{k+1}-t_k) \\ +G\big(x(\omega,t_k),t_k\big)\Sigma(t_k)\Big(\mathrm{B}(\omega,t_{k+1}) - \mathrm{B}(\omega,t_k)\Big), \end{aligned}$$

defined in the partition $t_0 \leq \cdots \leq t_k \leq \ldots$, when $\max_{i\geq 0}\{t_{i+1} - t_i\} \to 0$. The vector field $g : \mathbb{R}^n \times [t_0,\infty) \to \mathbb{R}^n$ is the drift, the matrix valued function $G : \mathbb{R}^m \times [t_0,\infty) \to \mathbb{R}^{n\times q}$ is the diffusion term that models the way in which the noise enters the dynamics, and $\Sigma : [t_0,\infty) \to \mathbb{R}^{q\times m}$ determines the size of the noise. The matrix $\Sigma(t)\Sigma(t)^\top$ is called the infinitesimal covariance. Under some regularity conditions, the solution inherits some properties of the Brownian motion; for instance, $x : \Omega \times [t_0,\infty) \to \mathbb{R}^n$ has continuous sample paths with probability 1, and for each $t \geq t_0$, the function $x(.,t) : \Omega \to \mathbb{R}^n$ (written $x(t)$ for convenience) is a random variable with some distribution. That is, we are able to measure the probabilities of certain events concerning the random variables $\{x(t)\}_{t\geq t_0}$.

III. Problem statement

Consider a network of agents represented by a strongly connected and weight-balanced digraph $\mathcal{G}$. An edge $(i,j) \in \mathcal{E}$ corresponds to the ability of agent i to receive information sent from agent j. Now consider a function of the form

$$f(x) = \sum_{i=1}^{N} f_i(x), \tag{2}$$

where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex and only known to agent i. If at least one of the functions in the sum, say f_{i_0}, is strongly convex, then (2) has a unique minimizer $x_{\min} \in \mathbb{R}^d$. Our goal is to design a distributed algorithm that helps the network find such minimizer. More precisely, agents can communicate only to their neighbors in the directed network, and there is additive persistent noise affecting the communication channels as well as the computations.

a) Communication noise: The communication channels between agents are subject to Gaussian white noise. Specifically, when agent j sends the signal $x(t) \in \mathbb{R}^d$ to agent i at $t \geq t_0$, agent i receives the corrupted signal

$$x(t) + \hat{\mathbf{J}}_{ij}(t)\, W^{(i,j)}_{\text{comm}}(\omega, t), \tag{3}$$

where $\hat{\mathbf{J}}_{ij} : [t_0, \infty) \to \mathbb{R}^{d\times d}$ is measurable and essentially locally bounded such that $\hat{\mathbf{J}}_{ij}(t) = 0_{d\times d}$ if and only if $\mathsf{a}_{ij} = 0$. Informally, the random variable $W : \Omega \times [t_0, \infty) \to \mathbb{R}^d$ that represents the Gaussian white noise is the derivative of a Brownian motion $W = \text{“}\frac{\mathrm{dB}}{\mathrm{d}t}\text{”}$. Later, a rigorous implementation of this model will give rise to a stochastic differential equation because we can only quantify the effect of this noise as an integral over a time interval. A possible interpretation is that

$$\int_t^{t+\Delta t} \hat{\mathbf{J}}_{ij}(s)\hat{\mathbf{J}}_{ij}(s)^\top \mathrm{d}s$$

is the cumulative covariance matrix due to the Gaussian noise that affects the communication channel $(i,j) \in \mathcal{E}$ during the interval $[t, t+\Delta t]$.

The noise we are considering is additive because it is always present, no matter the value of the signal; and is persistent because it is not required to decay with time. Since the matrix $\hat{\mathbf{J}}_{ij}$ is a property of the communication network for each $(i,j) \in \mathcal{E}$, it might be unknown to agent i.

b) Computation noise: Similarly, any computation carried out by an agent is also corrupted by noise. Specifically, when agent i computes $\nabla f_i(x(t)) \in \mathbb{R}^d$ at time instant $t \geq t_0$, it actually obtains

$$\nabla f_i(x(t)) + \tilde{\mathbf{J}}_{ii}(t)\, W^{i}_{\text{comp}}(\omega, t), \tag{4}$$

where W^{i}_{comp} is an independent copy of $W^{(i,j)}_{\text{comm}}$, and $\tilde{\mathbf{J}}_{ii} : [t_0, \infty) \to \mathbb{R}^{d\times d}$ is measurable and essentially locally bounded. As before, agent i might not know the matrix $\tilde{\mathbf{J}}_{ii}$.

IV. Robust distributed optimization

Here we describe the coordination algorithm that the network of agents employ to solve the optimization problem in Section III. We also describe our hypotheses and present the main convergence results.

The intuition behind the algorithm is the following: each agent's dynamics is driven by the gradient descent on its own private objective function plus a second-order consensus protocol based on local communication of estimates with its neighbors in the directed graph. Let $x^i(t) \in \mathbb{R}^d$ be the current estimate of agent $i \in \{1, \dots, N\}$ about the minimizer of (2). The ith agent's dynamics is

$$\begin{aligned}
\mathrm{d}x^i(t) = {} & \tilde{\gamma} \sum_{j=1, j\neq i}^{N} \mathsf{a}_{ij} \Big(\big(x^j(t) - x^i(t)\big)\mathrm{d}t + \hat{\mathbf{J}}_{ij}(t)\mathrm{dB}^{1,(i,j)}(t) \Big) \\
& + \sum_{j=1, j\neq i}^{N} \mathsf{a}_{ij} \Big(\big(z^j(t) - z^i(t)\big)\mathrm{d}t + \hat{\mathbf{J}}_{ij}(t)\mathrm{dB}^{2,(i,j)}(t) \Big) \\
& - \nabla \tilde{f}_i(x^i(t))\mathrm{d}t - \tilde{\mathbf{J}}_{ii}(t)\mathrm{dB}^{3,i}(t)\,; \qquad (5) \\
\mathrm{d}z^i(t) = {} & -\sum_{j=1, j\neq i}^{N} \mathsf{a}_{ij} \Big(\big(x^j(t) - x^i(t)\big)\mathrm{d}t + \hat{\mathbf{J}}_{ij}(t)\mathrm{dB}^{1,(i,j)}(t) \Big),
\end{aligned}$$

where $\mathrm{B}^{1,(i,j)}$, $\mathrm{B}^{2,(i,j)}$ and $\mathrm{B}^{3,i}$ are independent d-dimensional Brownian motions for each directed channel $(i,j) \in \mathcal{E}$ and each agent $i \in \mathcal{I}$, respectively. Note that we have used the model for the noisy communication channels given by (3) and the model for the computation errors in (4). As described in Section II, the matrix A describes the topology of the directed network and hence determines the local interactions between agents. In the above dynamics, $z^i(t)$, for each agent, is an auxiliary (integrator) variable that helps agents reach consensus. The constant $\tilde{\gamma} > 0$ is a design parameter to be determined.

Next we write the above dynamics in compact form and further generalize the model for the noise. Let L be the Laplacian of the strongly connected and weight-balanced digraph $\mathcal{G}$ modeling inter-agent communication, and define $\mathbf{L} \triangleq \mathsf{L} \otimes \mathrm{I}_d$. Denote by $\boldsymbol{x} \triangleq [(x^1)^\top, \dots, (x^N)^\top]^\top \in (\mathbb{R}^d)^N$ the collection of estimates across the network. Finally, define the auxiliary function $\tilde{f} : (\mathbb{R}^d)^N \to \mathbb{R}$ by

$$\tilde{f}(\boldsymbol{x}) = \sum_{i=1}^{N} f_i(x^i), \tag{6}$$

which is the decentralized version of the function (2). We study the continuous time dynamics given by the following system of stochastic differential equations:

$$\begin{aligned}
\mathrm{d}\boldsymbol{x} &= -(\nabla \tilde{f}(\boldsymbol{x}) + \tilde{\gamma}\, \mathbf{L}\boldsymbol{x} + \mathbf{L}\boldsymbol{z})\mathrm{d}t + G^1(\boldsymbol{x}, \boldsymbol{z}, t)\Sigma^1(t)\mathrm{dB}(t), \\
\mathrm{d}\boldsymbol{z} &= \mathbf{L}\boldsymbol{x}\mathrm{d}t + G^2(\boldsymbol{x}, \boldsymbol{z}, t)\Sigma^2(t)\mathrm{dB}(t). \qquad (7)
\end{aligned}$$

Here, $\boldsymbol{z} \triangleq [(z^1)^\top, \dots, (z^N)^\top]^\top \in (\mathbb{R}^d)^N$ is the aggregate of the auxiliary states of the agents; the matrix-valued functions $G^1, G^2 : \mathbb{R}^{2Nd} \times [t_0, \infty) \to \mathbb{R}^{Nd\times q}$ are bounded and globally Lipschitz in the first two arguments and measurable

and essentially bounded in time; the matrix-valued functions $\Sigma^1, \Sigma^2 : [t_0, \infty) \to \mathbb{R}^{q\times m}$ are measurable and essentially locally bounded, where $m \geq 1$; and $\{\mathrm{B}(t)\}_{t\geq t_0}$ is an m-dimensional Brownian motion defined in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The compact form (7) offers an alternative motivation which is explained in [11] and [8, Sec. V]; summarizing, it can be regarded as a *modified* saddle-point dynamics to find the saddle point of the augmented Lagrangian $\tilde{f}(\boldsymbol{x}) + \frac{\tilde{\gamma}}{2}\boldsymbol{x}^\top \mathbf{L}\boldsymbol{x} + \boldsymbol{z}^\top \mathbf{L}\boldsymbol{x}$, where $\boldsymbol{z}$ is regarded as the Lagrange multiplier associated to the agreement constraint $\mathbf{L}\boldsymbol{x} = 0$. The modification comes from replacing $\mathbf{L}^\top$ in the saddle point dynamics, which is not distributed over the directed network, by $\mathbf{L}$. We emphasize that the dynamics (7) is distributed over the digraph $\mathcal{G}$ because the agent i can update $x^i(t)$ and $z^i(t)$ using only the information sent from its neighbors, $\{x^j(t),\, z^j(t)\}$, and its knowledge of the function f_i, which is private. This is the case because the gradient is distributed, $\nabla \tilde{f}(\boldsymbol{x}) = [\nabla \tilde{f}_1(x^1)^\top, \ldots, \nabla \tilde{f}_N(x^N)^\top]^\top$, and the agent i can compute the ith d-dimensional block $(\mathbf{L}\boldsymbol{x})^i \in \mathbb{R}^d$. Agents do not need to know the functions Σ^1, Σ^2, G^1, G^2.

The dynamics (7) contains a more general description of the agents' communication and computation capabilities than the one we discussed in Section III, as we show next.

Remark 4.1. (A particular model for the noise): Next we write Σ^1 and Σ^2 in the dynamics (7) that correspond to this model; here $G^1 = G^2 = \mathrm{I}_{Nd}$. Define the matrices $\hat{\Sigma}^1(t) \triangleq \begin{bmatrix}\tilde{\gamma}\mathbf{J}(t) & \mathbf{J}(t) & -\tilde{\mathbf{J}}(t)\end{bmatrix}$ and $\hat{\Sigma}^2(t) \triangleq \begin{bmatrix}-\mathbf{J}(t) & 0 & 0\end{bmatrix}$, both in $\mathbb{R}^{Nd\times 3Nd}$, where $\mathbf{J}(t) \in \mathbb{R}^{Nd\times Nd}$ is the matrix whose (i,j)th d-dimensional block is $\mathsf{a}_{ij}\hat{\mathbf{J}}_{ij}(t)$, and $\tilde{\mathbf{J}}(t) \in \mathbb{R}^{Nd\times Nd}$ is defined, following (4), as $\tilde{\mathbf{J}} \triangleq \mathrm{diag}\big(\tilde{\mathbf{J}}_{11}(t), \ldots, \tilde{\mathbf{J}}_{NN}(t)\big)$. Then, each matrix Σ^k, $k = 1, 2$, is formed by shifting $(i-1)Nd$ entries to the right each ith d-dimensional row-block of $\hat{\Sigma}^k$ and filling up with zeros:

$$\begin{bmatrix}\Sigma^1\\ \Sigma^2\end{bmatrix} \triangleq \begin{bmatrix}\big((e_1 e_1^\top)\otimes \mathrm{I}_d\big)\hat{\Sigma}^1 & \cdots & \big((e_N e_N^\top)\otimes \mathrm{I}_d\big)\hat{\Sigma}^1\\ \big((e_1 e_1^\top)\otimes \mathrm{I}_d\big)\hat{\Sigma}^2 & \cdots & \big((e_N e_N^\top)\otimes \mathrm{I}_d\big)\hat{\Sigma}^2\end{bmatrix},$$

which takes values in $\mathbb{R}^{2Nd\times 3N^2 d}$, so that each agent experiences an independent realization of the noise corresponding to the communication channel $(i,j) \in \mathcal{E}$. •

Now we prepare the hypotheses for our main result.

Assumption 4.2. (Objective functions): Assume that the functions $\{f_i : 1 \leq i \leq N\}$ are convex and twice continuously differentiable with uniformly upper-bounded Hessians, such that at least one of the functions is strongly convex; that is, there exist positive numbers R and r such that $0 \preccurlyeq \nabla^2 f_i \preccurlyeq R\,\mathrm{I}_d$ for all $i \in \{1, \ldots, N\}$, and $r\,\mathrm{I}_d \preccurlyeq \nabla^2 f_{i_0}$ for some $i_0 \in \{1, \ldots, N\}$. •

The bounded Hessians assumption is standard in the literature, whereas we have relaxed the strong convexity of all the objectives (standard to deal with robustness in general) to strong convexity of only one objective.

Part of our design requires selecting $\tilde{\gamma}$ in the dynamics (7). The provable interval for $\tilde{\gamma}$ for convergence and disturbance attenuation depends on the overall topology of the network, as we show next.

Assumption 4.3. (Choice of the design parameter): Before giving the provable interval for $\tilde{\gamma}$, we start defining some auxiliary quantities. Consider any fixed $\epsilon > 0$ and any $\delta \in \big(0,\, \tilde{k}\hat{K}^{-2}\big)$, where

$$\tilde{k} \triangleq \lambda_{\min}\left(r\, e_{i_0} e_{i_0}^\top + \epsilon\,(\mathsf{L} + \mathsf{L}^\top)\right), \quad \hat{K} \triangleq R + 2\epsilon\,\sigma_{\max}(\mathsf{L}).$$

Here, r and R come from the gradient bounds in Assumption 4.2. We select the design parameter $\tilde{\gamma}$ as

$$\tilde{\gamma}(\epsilon, \delta) \triangleq \tfrac{2+\beta^2}{\beta} + 2\epsilon, \qquad \beta \in \big(0,\, \min\{\beta_1^*(\delta, \epsilon), \beta_2^*(\delta)\}\big),$$

where $\beta_1^*(\delta, \epsilon)$ and $\beta_2^*(\delta)$ are chosen as follows: First,

$$\beta_1^*(\delta, \epsilon) \triangleq \sqrt{\tilde{k}^2\hat{K}^{-2} - \tilde{k}\delta}.$$

Second, we choose $\beta_2^*(\delta)$ such that

$$h(\beta, \delta) < 0 \qquad \forall \beta \in \big(0,\, \beta_2^*(\delta)\big),$$

where $h(., \delta) : (0, \infty) \to \mathbb{R}$ is given by

$$h(\beta, \delta) \triangleq \big(-\,y(\beta) + \sqrt{y(\beta)^2 - 1}\,\big)\lambda_2(\mathsf{L} + \mathsf{L}^\top) + \tfrac{\beta^2}{2\delta}, \quad (8)$$

with $y(\beta) \triangleq \frac{\beta^4 + 3\beta^2 + 2}{2\beta}$. •

Our main result shows that the dynamics of each agent's estimate is noise-to-state exponentially stable in second moment [16] with respect to $x_{\min}$.

Theorem 4.4. (Exponential Noise-to-State Stability of the dynamics): Under Assumption 4.2 and for the choice of the parameter $\tilde{\gamma}$ in Assumption 4.3, the dynamics (7) in a strongly connected and weight-balanced digraph has the following stability property: there exist constants $C_\mu, D_\mu, C_\theta > 0$ such that for all $t \geq t_0$ and for any initial values $\boldsymbol{x}_0 \triangleq \boldsymbol{x}(t_0)$, $\boldsymbol{z}_0 \triangleq \boldsymbol{z}(t_0)$ in $(\mathbb{R}^d)^N$, the following holds:

$$\begin{aligned}&\mathbb{E}\big[\|\boldsymbol{x}(t) - \mathbb{1}_N \otimes x_{\min}\|_2^2\big]\\ &\leq \mathbb{E}\big[\|\boldsymbol{x}(t) - \mathbb{1}_N \otimes x_{\min}\|_2^2 + \|\boldsymbol{z}(t) - \boldsymbol{z}^*\|_{\mathbf{L}_\mathcal{K}}^2\big]\\ &\leq C_\mu \mathrm{V}_0^2\, e^{-D_\mu\,(t-t_0)} + C_\theta\Big(\operatorname*{ess\,sup}_{t_0\leq\tau\leq t}\Big|\Big[\begin{smallmatrix}\Sigma^1(\tau)\\ \Sigma^2(\tau)\end{smallmatrix}\Big]\Big|_\mathcal{F}\Big)^2,\end{aligned}$$

where $x_{\min}$ is the unique minimizer of (2), $\boldsymbol{z}^*$ satisfies $\mathbf{L}\boldsymbol{z}^* = -\nabla\tilde{f}(\mathbb{1}_N \otimes x_{\min})$, $\mathbf{L}_\mathcal{K} \triangleq \mathsf{L}_\mathcal{K} \otimes \mathrm{I}_d$, and $\mathrm{V}_0^2 \triangleq \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 + \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|_{\mathbf{L}_\mathcal{K}}^2$.

This convergence result says that each agent's estimate converges exponentially fast, in second moment, to a neighborhood of the optimal solution, and the size of that neighborhood depends on the size of the noise. The explicit constants $C_\mu, D_\mu, C_\theta > 0$ are omitted here for lack of space. Figure 1 illustrates this convergence result.

The proof strategy to establish Theorem 4.4 relies on the identification of a NSS-Lyapunov function for (7) which in turn implies the Noise-to-State Stability property, cf. [16]. To verify the properties of the aforementioned Lyapunov function, we study the interaction between local optimization and local consensus through the co-coercivity of a family

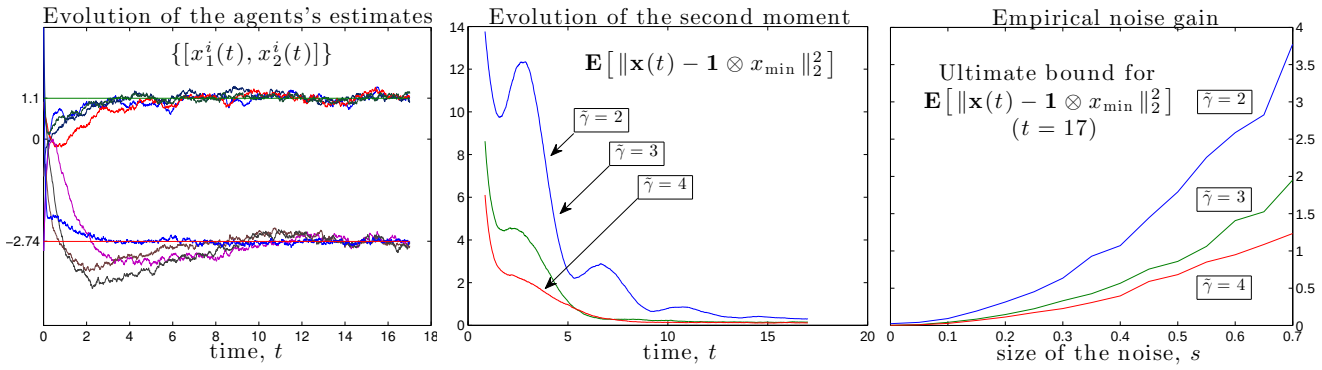

Fig. 1: Evolution of the stochastic distributed optimization algorithm (7) over a group of $N = 4$ agents communicating over a directed ring ($\mathcal{E} = \{(1,3),(3,2),(2,4),(4,1)\}$). The local objective functions are defined on $\mathbb{R}^2$ and given by $f_1(x_1,x_2) = \frac{1}{2}((x_1-4)^2 + (x_2-3)^2)$, $f_2(x_1,x_2) = x_1 + 3x_2 - 2$, $f_3(x_1,x_2) = \log(e^{x_1+3} + e^{x_2+1})$, and $f_4(x_1,x_2) = (x_1 + 2x_2 + 5)^2 + (x_1 - x_2 - 4)^2$. In all the cases, the initial conditions are $\boldsymbol{x} = [-3,-3,-1,-1,1,1,3,3]$, and $\boldsymbol{z} = \mathbb{1}_8$. Plot (a) shows the evolution of the first and second coordinates of the agents' estimates with $\tilde{\gamma} = 3$, $G^1 = G^2 = \mathrm{I}_8$, and $\Sigma^1 = \Sigma^2 = 0.2\,\mathrm{I}_8$. Despite the additive persistent noise, the estimates converge, in probability, to a neighborhood of the minimizer $x_{\min} = [1.10, -2.74]$. Plot (b) shows the asymptotic convergence in second moment to a neighborhood of the solution for three values of the design parameter. Plot (c) depicts the ultimate bound for the second moment when $\Sigma^1 = \Sigma^2 = s\,\mathrm{I}_8$ for increasing values of s from 0 to 0.7 and increments of 0.05; we observe that when the design parameter puts more emphasis on consensus, the noise gain is smaller. (The expectations have been computed averaging over 100 realizations of the noise.)

of vector fields that are the sum of a gradient of a convex function plus a nonsymmetric Laplacian. Specifically, Theorem A.2 gives sufficient conditions for this family of vector fields to be co-coercive. These techniques allow us to overcome the challenges posed by directed communication topologies and channels affected by additive persistent noise. In addition, they help us characterize the exponential rate of convergence, in second moment, to a neighborhood and the functional dependence of that neighborhood on the size of the disturbance.

As a particular case we obtain a refinement of the result in [8] showing exponential convergence of the algorithm.

Corollary 4.5. (Case without noise: global exponential stability): Under the same hypotheses of Theorem 4.4, if $\Sigma^1 = \Sigma^2 = 0$ in (7), then

$$\|\boldsymbol{x}(t) - \mathbb{1}_N \otimes x_{\min}\|_2^2 \leq \|\boldsymbol{x}(t) - \boldsymbol{x}^*\|_2^2 + \|\boldsymbol{z}(t) - \boldsymbol{z}^*\|_2^2 \leq C_\mu \mathrm{V}_0^2\, e^{-D_\mu\,(t-t_0)} + \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|_{\mathbf{M}}^2,$$

for all $t \geq t_0$, where $\mathbf{M} = \frac{1}{N}\mathbb{1}\mathbb{1}^\top \otimes \mathrm{I}_d$. In particular, choosing $\boldsymbol{z}^*$ such that $\mathbf{M}\boldsymbol{z}^* = \mathbf{M}\boldsymbol{z}_0$ (so that $\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|_{\mathbf{M}} = 0$), which is always possible because $\mathbf{M}\boldsymbol{z}(t) = \mathbf{M}\boldsymbol{z}(t_0)$, we obtain that the point $[\mathbb{1}_N^\top \otimes x_{\min}^\top, \boldsymbol{z}^{*\top}]^\top$ is globally asymptotically stable.

Proof. Since $\Sigma^1 = \Sigma^2 = 0$, for any value of $\boldsymbol{x}_0$, $\boldsymbol{z}_0$, the system of SDEs (7) becomes a system of ordinary differential equations. Let $\boldsymbol{z}_{\text{agree}}(t) \triangleq \mathbf{M}\boldsymbol{z}(t)$. By left-multiplying the dynamics of $\boldsymbol{z}(t)$ in (7) by $\mathbf{M}$, we obtain that $\dot{\boldsymbol{z}}_{\text{agree}} = 0$, where $\boldsymbol{z}_{\text{agree}}(t_0) = \mathbf{M}\boldsymbol{z}_0$. Therefore, $\boldsymbol{z}_{\text{agree}}(t) = \boldsymbol{z}_{\text{agree}}(t_0)$ for all $t \geq t_0$. Using that $\mathbf{M}$ is symmetric and $\mathbf{M} = \mathbf{M}^2$, if we define $\boldsymbol{z}^*_{\text{agree}} \triangleq \mathbf{M}\,\boldsymbol{z}^*$, then

$$\begin{aligned}
&(\boldsymbol{z}(t) - \boldsymbol{z}^*)^\top \mathbf{M} (\boldsymbol{z}(t) - \boldsymbol{z}^*) \\
&= (\boldsymbol{z}_{\text{agree}}(t) - \boldsymbol{z}^*_{\text{agree}})^\top \mathbf{M} (\boldsymbol{z}_{\text{agree}}(t) - \boldsymbol{z}^*_{\text{agree}}) \\
&= (\boldsymbol{z}_{\text{agree}}(t_0) - \boldsymbol{z}^*_{\text{agree}})^\top\, \mathbf{M}\, (\boldsymbol{z}_{\text{agree}}(t_0) - \boldsymbol{z}^*_{\text{agree}}) \\
&= (\boldsymbol{z}_0 - \boldsymbol{z}^*)^\top\, \mathbf{M}\, (\boldsymbol{z}_0 - \boldsymbol{z}^*) = \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|_{\mathbf{M}}^2.
\end{aligned}$$

Hence, using that $\mathbf{L}_{\mathcal{K}}^2 = \mathbf{L}_{\mathcal{K}}$,

$$\begin{aligned}
&\|\boldsymbol{z}(t) - \boldsymbol{z}^*\|_2^2 = (\boldsymbol{z}(t) - \boldsymbol{z}^*)^\top (\boldsymbol{z}(t) - \boldsymbol{z}^*) \\
&= (\boldsymbol{z}(t) - \boldsymbol{z}^*)^\top \big(\mathbf{L}_{\mathcal{K}} + \mathbf{M}\big)(\boldsymbol{z}(t) - \boldsymbol{z}^*) \\
&= \|\boldsymbol{z}(t) - \boldsymbol{z}^*\|_{\mathbf{L}_{\mathcal{K}}}^2 + \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|_{\mathbf{M}}^2,
\end{aligned}$$

so the result follows from Theorem 4.4. □

V. Conclusions

We have studied the robustness against additive persistent noise of a class of continuous-time algorithms for distributed convex optimization over weight-balanced digraphs. Specifically, we have shown that the agents' estimates converge exponentially fast, in second moment, to a neighborhood of the optimum global decision vector. The size of this neighborhood is proportional to the size of the noise. Future work will include distributed procedures to determine the values of the design parameter that guarantee convergence and disturbance attenuation, relaxing the weight-balanced property of the directed communication topology, and extensions to scenarios with delays and bandwidth limitations.

ACKNOWLEDGMENTS

This research was supported by NSF award CMMI-1300272.

REFERENCES

[1] J. Duchi, A. Agarwal, and M. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, 2012.

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

[3] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, "Push-sum distributed dual averaging for convex optimization," in *IEEE Conf. on Decision and Control*, (Maui, HI), pp. 5453–5458, 2012.

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

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

[6] E. Wei and A. Ozdaglar, "Distributed alternating direction method of multipliers," in *IEEE Conf. on Decision and Control*, (Maui, HI), pp. 5445–5450, 2012.

[7] S. Nikookhoy, J. Lu, and C. Y. Tang, "Distributed convex optimization with identical constraints," in *IEEE Conf. on Decision and Control*, (Orlando, FL), pp. 2926–2931, Dec. 2011.

[8] B. Gharesifard and J. Cortés, "Distributed continuous-time convex optimization on weight-balanced digraphs," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, 2014. To appear.

[9] J. Lu and C. Y. Tang, "Zero-gradient-sum algorithms for distributed convex optimization: the continuous-time case," in *American Control Conference*, (San Francisco, CA), pp. 5474 – 5479, June 2011.

[10] P. Wan and M. D. Lemmon, "Event-triggered distributed optimization in sensor networks," in *Symposium on Information Processing of Sensor Networks*, (San Francisco, CA), pp. 49–60, 2009.

[11] J. Wang and N. Elia, "Control approach to distributed optimization," in *Allerton Conf. on Communications, Control and Computing*, (Monticello, IL), pp. 557–561, Oct. 2010.

[12] 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.

[13] X. Mao, *Stochastic Differential Equations and Applications*. Woodhead Publishing, 2011.

[14] B. Öksendal, *Stochastic Differential Equations - An Introduction with Applications*. Universitext, Springer-Verlag, 2010.

[15] J. R. Movellan, "Tutorial on stochastic differential equations," tutorial, MPLab, UCSD, 2011.

[16] D. Mateos-Núñez and J. Cortés, "Stability of stochastic differential equations with additive persistent noise," in *American Control Conference*, (Washington, D.C.), pp. 5447–5452, June 2013.

[17] D. L. Zhu and P. Marcotte, "Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities," *SIAM Journal on Optimization*, vol. 6, no. 3, pp. 714–726, 1996.

APPENDIX

Our technical approach includes studying the interaction between local optimization and local consensus. First, we give the definition of co-coercivity of vector fields that are distorted by a linear transformation.

Definition A.1. (Co-coercivity of vector fields): A vector field $F : \mathbb{R}^m \to \mathbb{R}^m$ is $(S, \delta) - $ *co-coercive* with respect to $\bar{\boldsymbol{x}} \in \mathbb{R}^m$, for some matrix $S \in \mathbb{R}^{m\times m}$ and some number $\delta > 0$, if, for all $x \in \mathbb{R}^m$,

$$(\boldsymbol{x} - \bar{\boldsymbol{x}})^\top S(F(\boldsymbol{x}) - F(\bar{\boldsymbol{x}})) \geq \delta \, \|F(\boldsymbol{x}) - F(\bar{\boldsymbol{x}})\|_2^2. \quad (9)$$

(This is a restatement of $S^\top F(x)$ being co-coercive [17].)

Next, we study a family of vector fields that combine local gradient descent and local consensus via a nonsymmetric Laplacian. In particular, we provide sufficient conditions for these vector fields be co-coercive under a class of linear transformations. This plays a central role in our technical approach to deal with additive persistent noise both for undirected and directed topologies.

Theorem A.2. (A family of vector fields, including $(S, \delta)-$ co-coercivity): Let $\mathbf{G} : (\mathbb{R}^d)^N \to (\mathbb{R}^d)^N$ be a continuously differentiable vector field such that $\mathbf{DG}(\boldsymbol{x}) \in \mathbb{R}^{Nd\times Nd}$ is symmetric positive semidefinite for all $\boldsymbol{x} \in (\mathbb{R}^d)^N$, and let $\mathbf{T} : (\mathbb{R}^d)^N \to (\mathbb{R}^d)^N$ be the linear vector field given by $\mathbf{T}(\boldsymbol{x}) = 2(\mathsf{L} \otimes \mathrm{I}_d)\boldsymbol{x}$, where L is the Laplacian matrix of a strongly connected and weight-balanced digraph. Assume that there exist an integer $i_0 \in \{1, \dots, N\}$ and positive numbers r, R such that $r\,(e_{i_0} e_{i_0}^\top)\otimes \mathrm{I}_d \preccurlyeq \mathbf{DG}(\boldsymbol{x}) \preccurlyeq R\,\mathrm{I}_{Nd}$ for all $\boldsymbol{x} \in (\mathbb{R}^d)^N$. If we define $\tilde{k} \triangleq \lambda_{\min}\left(r\, e_{i_0} e_{i_0}^\top + \epsilon\,(\mathsf{L} + \mathsf{L}^\top)\right)$ and $\hat{K} \triangleq R + 2\epsilon\,\sigma_{\max}(\mathsf{L})$, then the following facts regarding the vector field $\mathbf{F} \triangleq \mathbf{G} + \epsilon\mathbf{T}$, for any $\epsilon > 0$, hold:

(i) $\tilde{k} > 0$ and $2\tilde{k}\,\mathrm{I}_{Nd} \preccurlyeq \mathbf{DF} + (\mathbf{DF})^\top$.

(ii) For any $\boldsymbol{x}, \bar{\boldsymbol{x}} \in (\mathbb{R}^d)^N$,

$$\tilde{k}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2 \leq \|\mathbf{F}(\boldsymbol{x}) - \mathbf{F}(\bar{\boldsymbol{x}})\|_2 \leq \hat{K}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2.$$

(iii) $\mathbf{F}$ is $(\mathrm{I} + \beta^2 S, \delta)-$co-coercive with respect to every $\bar{\boldsymbol{x}} \in (\mathbb{R}^d)^N$ for any nonzero matrix $\tilde{S} \in \mathbb{R}^{Nd\times Nd}$ if $\delta \in [0, \delta_1^*)$ and $\beta \in [0, \beta_1^*]$, where

$$\delta_1^* \triangleq \tilde{k}\hat{K}^{-2} \quad \text{and} \quad \beta_1^* \triangleq \sqrt{\left(\tilde{k}\hat{K}^{-2} - \delta\right)/(\|\tilde{S}\|_2\,\tilde{k}^{-1})}.$$

The following result is needed to establish that the parameter $\tilde{\gamma}$ is well defined.

Lemma A.3. Let the scalar function $h(., \delta) : (0, \infty) \to \mathbb{R}$ be defined as in (8), and assume L is the Laplacian matrix of a strongly connected and weight-balanced digraph. Then, for every $\delta > 0$, there exists $\hat{\beta} \triangleq \hat{\beta}(\delta) > 0$ such that $h(\beta, \delta) < 0$ for all $\beta \in (0, \hat{\beta})$.

The next remark notes that the proposed interval for the parameter $\tilde{\gamma}$ is guaranteed to be nonempty.

Remark 5.1. (The design parameter $\tilde{\gamma}$ is well defined): Since $\tilde{k} > 0$ by Theorem A.2 (i), it follows that in the construction of $\tilde{\gamma}$ we can take $\delta > 0$ and thereby take $\beta_1^*(\delta, \epsilon) > 0$. On the other hand, the quantity $\beta_2^*(\delta)$ can also be taken positive thanks to Lemma A.3. As a consequence, the proposed provable interval for $\tilde{\gamma}$ is nonempty.