On the Learning Behavior of Adaptive Networks — Part I: Transient Analysis

Jianshu Chen, *Member, IEEE,* and Ali H. Sayed, *Fellow, IEEE*

***Abstract*—This work carries out a detailed transient analysis of the learning behavior of multi-agent networks, and reveals interesting results about the learning abilities of distributed strategies. Among other results, the analysis reveals how combination policies influence the learning process of networked agents, and how these policies can steer the convergence point towards any of many possible Pareto optimal solutions. The results also establish that the learning process of an adaptive network undergoes three (rather than two) well-defined stages of evolution with distinctive convergence rates during the first two stages, while attaining a finite mean-square-error (MSE) level in the last stage. The analysis reveals what aspects of the network topology influence performance directly and suggests design procedures that can optimize performance by adjusting the relevant topology parameters. Interestingly, it is further shown that, in the adaptation regime, each agent in a sparsely connected network is able to achieve the same performance level as that of a centralized stochastic-gradient strategy even for left-stochastic combination strategies. These results lead to a deeper understanding and useful insights on the convergence behavior of coupled distributed learners. The results also lead to effective design mechanisms to help diffuse information more thoroughly over networks.**

***Index Terms*—Multi-agent learning, multi-agent adaptation, distributed strategies, diffusion of information, Pareto solutions.**

I. INTRODUCTION

In multi-agent systems, agents interact with each other to solve a problem of common interest, such as an optimization problem in a distributed manner. Such networks of interacting agents are useful in solving distributed estimation, learning and decision making problems [2]–[39]. They are also useful in modeling biological networks [40]–[42], collective rational behavior [12], [13], and in developing biologically-inspired designs [2], [43]. Two useful strategies that can be used to guide the interactions of agents over a network are consensus strategies [5]–[11] and diffusion strategies [16]–[27]. Both classes of algorithms involve self-learning and social-learning steps. During self-learning, each agent updates its state using its local data. During social learning, each agent aggregates information from its neighbors. A useful feature that results from these localized interactions is that the network ends up exhibiting global patterns of behavior. For example, in distributed estimation and learning, each agent is able to attain the performance of centralized solutions by relying solely on local cooperation [6], [19], [22], [23].

In this article, and the accompanying Part II [44], we consider a general class of distributed strategies, which includes diffusion and consensus updates as special cases, and study the resulting global learning behavior by addressing four important questions: (i) where does the distributed algorithm converge to? (ii) when does it converge? (iii) how fast does it converge? and (iv) how close does it converge to the intended point? We answer questions (i)–(iii) in Part I and question (iv) in Part II [44]. We study these four questions by characterizing the learning dynamics of the network in some great detail. An interesting conclusion that follows from our analysis is that the learning curve of a multi-agent system will be shown to exhibit *three* different phases. In the first phase (Transient Phase I), the convergence rate of the network is determined by the second largest eigenvalue of the combination matrix in magnitude, which is related to the degree of network connectivity. In the second phase (Transient Phase II), the convergence rate is determined by the entries of the right-eigenvector of the combination matrix corresponding to the eigenvalue at one. And, in the third phase (the steady-state phase) the mean-square performance of the algorithm turns out to depend on this same right-eigenvector in a revealing way. Even more surprisingly, we shall discover that the agents have the same learning behavior starting at Transient Phase II, and are able to achieve a performance level that matches that of a fully connected network or a centralized stochastic-gradient strategy. Actually, we shall show that the consensus and diffusion strategies can be represented as perturbed versions of a centralized *reference* recursion in a certain transform domain. We quantify the effect of the perturbations and establish the aforementioned properties for the various phases of the learning behavior of the networks. The results will reveal the manner by which the network topology influences performance in some unexpected ways.

There have been of course many insightful works in the literature on distributed strategies and their convergence behavior. In Sections II-B and IV-A further ahead, we explain in what ways the current manuscript extends these earlier investigations and what novel contributions this work leads

Manuscript received December 28, 2013; revised November 21, 2014; accepted April 08, 2015. This work was supported in part by NSF grants CCF-1011918 and ECCS-1407712. A short and limited version of this work appears in the conference publication [1] without proofs and under more restrictive conditions than considered in this broader and expanded work.

J. Chen was with Department of Electrical Engineering, University of California, Los Angeles, and is currently with Microsoft Research, Redmond, WA 98052. This work was performed while he was a PhD student at UCLA. Email: cjs09@ucla.edu.

A. H. Sayed is with Department of Electrical Engineering, University of California, Los Angeles, CA 90095. Email: sayed@ee.ucla.edu.

Communicated by Prof. Nicolò Cesa-Bianchi, Associate Editor for Pattern Recognition, Statistical Learning, and Inference.

to. In particular, it will be seen that several new insights are discovered that clarify how distributed networks learn. For the time being, in these introductory remarks, we limit ourselves to mentioning one important aspect of our development. Most prior studies on distributed optimization and estimation tend to focus on the performance and convergence of the algorithms under *diminishing* step-size conditions [5]–[10], [28]–[30], [45], or on convergence under deterministic conditions on the data [10]. This is perfectly fine for applications involving *static* optimization problems where the objective is to locate the fixed optimizer of some aggregate cost function of interest. In this paper, however, we examine the learning behavior of distributed strategies under *constant* step-size conditions. This is because constant step-sizes are necessary to enable continuous adaptation, learning, and tracking in the presence of streaming data and drifting conditions. These features would enable the algorithms to perform well even when the location of the optimizer drifts with time. Nevertheless, the use of constant step-sizes enriches the dynamics of (stochastic-gradient) distributed algorithms in that the gradient update term does not die out with time anymore, in clear contrast to the diminishing step-size case where the influence of the gradient term is annihilated over time due to the decaying value of the step-size parameter. For this reason, more care is needed to examine the learning behavior of distributed strategies in the constant step-size regime since their updates remain continually active and the effect of gradient noise is always present. This work also generalizes and extends in non-trivial ways the studies in [18], [20]. For example, while reference [18] assumed that the individual costs of all agents have the *same* minimizer, and reference [20] assumed that each of these individual costs is strongly convex, these requirements are not needed in the current study: individual costs can have distinct minimizers and they do not even need to be convex (see the discussion after expression (32)). This fact widens significantly the class of distributed learning problems that are covered by our framework. Moreover, the network behavior is studied under less restrictive assumptions and for broader scenarios, including a close study of the various phases of evolution during the transient phase of the learning process. We also study a larger class of distributed strategies that includes diffusion and consensus strategies as special cases.

To examine the learning behavior of adaptive networks under broader and more relaxed conditions than usual, we pursue a new analysis route by introducing a *reference* centralized recursion and by studying the perturbation of the diffusion and consensus strategies relative to this centralized solution over time. Insightful new results are obtained through this perturbation analysis. For example, we are now able to examine closely *both* the transient phase behavior and the steady-state phase behavior of the learning process and to explain how behavior in these two stages relate to the behavior of the centralized solution (see Fig. 2 further ahead). Among several other results, the mean-square-error expression (52) derived later in Part II [44] following some careful analysis, which builds on the results of this Part I, is one of the new (compact and powerful) insights; it reveals how the performance of each agent is closely related to that of the centralized stochastic approximation strategy — see the discussion right after (52). As the reader will ascertain from the derivations in the appendices, arriving at these conclusions for a broad class of distributed strategies and under weaker conditions than usual is demanding and necessitates a careful study of the evolution of the error dynamics over the network and its stability. When all is said and done, Parts I and II [44] lead to several novel insights into the learning behavior of adaptive networks.

Notation. All vectors are column vectors. We use boldface letters to denote random quantities (such as $\boldsymbol{u}_{k,i}$) and regular font to denote their realizations or deterministic variables (such as $u_{k,i}$). We use $\mathrm{diag}\{x_1,\ldots,x_N\}$ to denote a (block) diagonal matrix consisting of diagonal entries (blocks) $x_1,\ldots,x_N$, and use $\mathrm{col}\{x_1,\ldots,x_N\}$ to denote a column vector formed by stacking $x_1,\ldots,x_N$ on top of each other. The notation $x \preceq y$ means each entry of the vector x is less than or equal to the corresponding entry of the vector y, and the notation $X \preceq Y$ means each entry of the matrix X is less than or equal to the corresponding entry of the matrix Y. The notation $x = \mathrm{vec}(X)$ denotes the vectorization operation that stacks the columns of a matrix X on top of each other to form a vector x, and $X = \mathrm{vec}^{-1}(x)$ is the inverse operation. The operators ∇_w and ∇_{w^T} denote the column and row gradient vectors with respect to w. When ∇_{w^T} is applied to a column vector s, it generates a matrix. The notation $a(\mu) = O(b(\mu))$ means that there exists a constant $C > 0$ such that for all μ, $a(\mu) \leq C \cdot b(\mu)$.

II. PROBLEM FORMULATION

A. Distributed Strategies: Consensus and Diffusion

We consider a connected network of N agents that are linked together through a topology — see Fig. 1. Each agent k implements a distributed algorithm of the following form to update its state vector from $\boldsymbol{w}_{k,i-1}$ to $\boldsymbol{w}_{k,i}$:

$$\boldsymbol{\phi}_{k,i-1} = \sum_{l=1}^{N} a_{1,lk}\boldsymbol{w}_{l,i-1} \tag{1}$$

$$\boldsymbol{\psi}_{k,i} = \sum_{l=1}^{N} a_{0,lk}\boldsymbol{\phi}_{l,i-1} - \mu_k\hat{\boldsymbol{s}}_{k,i}(\boldsymbol{\phi}_{k,i-1}) \tag{2}$$

$$\boldsymbol{w}_{k,i} = \sum_{l=1}^{N} a_{2,lk}\boldsymbol{\psi}_{l,i} \tag{3}$$

where $\boldsymbol{w}_{k,i} \in \mathbb{R}^M$ is the state of agent k at time i, usually an estimate for the solution of some optimization problem, $\boldsymbol{\phi}_{k,i-1} \in \mathbb{R}^M$ and $\boldsymbol{\psi}_{k,i} \in \mathbb{R}^M$ are intermediate variables generated at node k before updating to $\boldsymbol{w}_{k,i}$, μ_k is a non-negative constant step-size parameter used by node k, and $\hat{\boldsymbol{s}}_{k,i}(\cdot)$ is an $M \times 1$ update vector function at node k. In deterministic optimization problems, the update vectors $\hat{\boldsymbol{s}}_{k,i}(\cdot)$ can be the gradient or Newton steps associated with the cost functions [10]. On the other hand, in stocastic approximation problems, such as adaptation, learning and estimation problems [5]–[9], [14]–[23], [25]–[29], the update vectors are usually computed from realizations of data samples that arrive sequentially at the nodes. In the stochastic setting, the quantities appearing

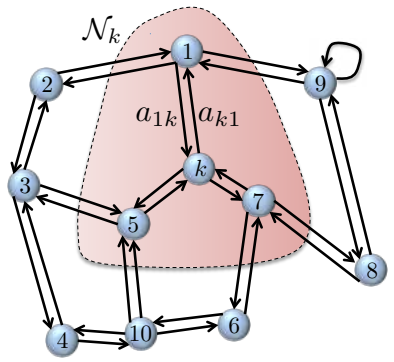


Fig. 1. A network representing a multi-agent system. The set of all agents that can communicate with node k is denoted by $\mathcal{N}_k$.The edge linking any two agents is represented by two directed arrows to emphasize that information can flow in both directions.

in (1)–(3) become random and we use boldface letters to highlight their stochastic nature. In Example 1 below, we illustrate choices for $\hat{\boldsymbol{s}}_{k,i}(w)$ in different contexts.

The combination coefficients $a_{1,lk}$, $a_{0,lk}$ and $a_{2,lk}$ in (1)–(3) are nonnegative weights that each node k assigns to the information arriving from node l; these coefficients are required to satisfy:

$$\sum_{l=1}^{N} a_{1,lk} = 1, \quad \sum_{l=1}^{N} a_{0,lk} = 1, \quad \sum_{l=1}^{N} a_{2,lk} = 1 \tag{4}$$

$$a_{1,lk} \geq 0, \qquad a_{0,lk} \geq 0, \qquad a_{2,lk} \geq 0 \tag{5}$$

$$a_{1,lk} = a_{2,lk} = a_{0,lk} = 0, \qquad \text{if } l \notin \mathcal{N}_k \tag{6}$$

Observe from (6) that the combination coefficients are zero if $l \notin \mathcal{N}_k$, where $\mathcal{N}_k$ denotes the set of neighbors of node k. Therefore, each summation in (1)–(3) is actually confined to the neighborhood of node k. In algorithm (1)–(3), each node k first combines the states $\{\boldsymbol{w}_{l,i-1}\}$ from its neighbors and updates $\boldsymbol{w}_{k,i-1}$ to the intermediate variable $\boldsymbol{\phi}_{k,i-1}$. Then, the $\{\boldsymbol{\phi}_{l,i-1}\}$ from the neighbors are aggregated and updated to $\boldsymbol{\psi}_{k,i}$ along the opposite direction of $\hat{\boldsymbol{s}}_{k,i}(\boldsymbol{\phi}_{k,i-1})$. Finally, the intermediate estimators $\{\boldsymbol{\psi}_{l,i}\}$ from the neighbors are combined to generate the new state $\boldsymbol{w}_{k,i}$ at node k.

Example 1: The distributed algorithm (1)–(3) can be applied to optimize aggregate costs of the following form:

$$J^{\text{glob}}(w) = \sum_{k=1}^{N} J_k(w) \tag{7}$$

or to find Pareto-optimal solutions to multi-objective optimization problems, such as:

$$\min_{w} \{J_1(w), \ldots, J_N(w)\} \tag{8}$$

where $J_k(w)$ is an individual convex cost associated with each agent k. Optimization problems of the form (7)–(8) arise in various applications — see [3]–[31]. Depending on the context, the update vector $\hat{\boldsymbol{s}}_{k,i}(\cdot)$ may be chosen in different ways:

- In deterministic optimization problems, the expressions for $\{J_k(w)\}$ are known and the update vector $\hat{\boldsymbol{s}}_{k,i}(\cdot)$ at node k is chosen as the deterministic gradient (column) vector $\nabla_w J_k(\cdot)$.
- In distributed estimation and learning, the individual cost function at each node k is usually selected as the expected value of some loss function $Q_k(\cdot,\cdot)$, i.e., $J_k(w) = \mathbb{E}\{Q_k(w, \boldsymbol{x}_{k,i})\}$ [18], where the expectation is with respect to the randomness in the data samples $\{\boldsymbol{x}_{k,i}\}$ collected at node k at time i. The exact expression for $\nabla_w J_k(w)$ is usually unknown since the probability distribution of the data is not known beforehand. In these situations, the update vector $\hat{\boldsymbol{s}}_{k,i}(\cdot)$ is chosen as an instantaneous approximation for the true gradient vector, such as, $\hat{\boldsymbol{s}}_{k,i}(\cdot) = \widehat{\nabla_w J_k}(\cdot) = \nabla_w Q_k(\cdot, \boldsymbol{x}_{k,i})$, which is known as *stochastic gradient*. Note that the update vector $\hat{\boldsymbol{s}}_{k,i}(w)$ is now evaluated from the random data sample $\boldsymbol{x}_{k,i}$. Therefore, it is also random and time dependent.

The update vectors $\{\hat{\boldsymbol{s}}_{k,i}(\cdot)\}$ may not necessarily be the gradients of cost functions or their stochastic approximations. They may take other forms for different reasons. For example, in [6], a certain gain matrix K is multiplied to the left of the stochastic gradient vector $\widehat{\nabla_w J_k}(\cdot)$ to make the estimator asymptotically efficient for a linear observation model. ■

Returning to the general distributed strategy (1)–(3), we note that it can be specialized into various useful algorithms. We let A_1, A_0 and A_2 denote the $N \times N$ matrices that collect the coefficients $\{a_{1,lk}\}$, $\{a_{0,lk}\}$ and $\{a_{2,lk}\}$. Then, condition (4) is equivalent to

$$A_1^T \mathbb{1} = \mathbb{1}, \quad A_0^T \mathbb{1} = \mathbb{1}, \quad A_2^T \mathbb{1} = \mathbb{1} \tag{9}$$

where $\mathbb{1}$ is the $N \times 1$ vector with all its entries equal to one. Condition (9) means that the matrices $\{A_0, A_1, A_2\}$ are left-stochastic (i.e., the entries on each of their columns add up to one). Different choices for A_1, A_0 and A_2 correspond to different distributed strategies, as summarized in Table I. Specifically, the traditional consensus [5]–[11] and diffusion (ATC and CTA) [16]–[23] algorithms with *constant* step-sizes are given by the following iterations:

$$\text{Consensus}: \begin{cases} \boldsymbol{\phi}_{k,i-1} = \displaystyle\sum_{l \in \mathcal{N}_k} a_{0,lk} \boldsymbol{w}_{l,i-1} \\ \boldsymbol{w}_{k,i} = \boldsymbol{\phi}_{k,i-1} - \mu_k \hat{\boldsymbol{s}}_{k,i}(\boldsymbol{w}_{k,i-1}) \end{cases} \tag{10}$$

$$\text{CTA diffusion}: \begin{cases} \boldsymbol{\phi}_{k,i-1} = \displaystyle\sum_{l \in \mathcal{N}_k} a_{1,lk} \boldsymbol{w}_{l,i-1} \\ \boldsymbol{w}_{k,i} = \boldsymbol{\phi}_{k,i-1} - \mu_k \hat{\boldsymbol{s}}_{k,i}(\boldsymbol{\phi}_{k,i-1}) \end{cases} \tag{11}$$

$$\text{ATC diffusion}: \begin{cases} \boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu_k \hat{\boldsymbol{s}}_{k,i}(\boldsymbol{w}_{k,i-1}) \\ \boldsymbol{w}_{k,i} = \displaystyle\sum_{l \in \mathcal{N}_k} a_{2,lk} \boldsymbol{\psi}_{l,i} \end{cases} \tag{12}$$

Therefore, the convex combination steps appear in different locations in the consensus and diffusion implementations. For instance, observe that the consensus strategy (10) evaluates the update direction $\hat{\boldsymbol{s}}_{k,i}(\cdot)$ at $\boldsymbol{w}_{k,i-1}$, which is the estimator *prior* to the aggregation, while the diffusion strategy (11) evaluates the update direction at $\boldsymbol{\phi}_{k,i-1}$, which is the estimator *after* the

TABLE I
DIFFERENT CHOICES FOR A_1, A_0 AND A_2 CORRESPOND TO DIFFERENT DISTRIBUTED STRATEGIES.

Distributed Strategeis	A_1	A_0	A_2	$A_1A_0A_2$
Consensus	I	A	I	A
ATC diffusion	I	I	A	A
CTA diffusion	A	I	I	A

aggregation. In our analysis, we will proceed with the general form (1)–(3) to study all three schemes, and other possibilities, within a unifying framework.

We observe that there are two types of learning processes involved in the dynamics of each agent k: (i) self-learning in (2) from locally sensed data and (ii) social learning in (1) and (3) from neighbors. All nodes implement the same self- and social learning structure. As a result, the learning dynamics of all nodes in the network are coupled; knowledge exploited from local data at node k will be propagated to its neighbors and from there to their neighbors in a diffusive learning process. It is expected that some global performance pattern will emerge from these localized interactions in the multi-agent system. In this work and the accompanying Part II [44], we address the following questions:

- Limit point: where does each state $\boldsymbol{w}_{k,i}$ converge to?
- Stability: under which condition does convergence occur?
- Learning rate: how fast does convergence occur?
- Performance: how close is $\boldsymbol{w}_{k,i}$ to the limit point?
- Generalization: can $\boldsymbol{w}_{k,i}$ match the performance of a centralized solution?

We address the first three questions in this part, and examine the last two questions pertaining to performance in Part II [44]. We address the five questions by characterizing analytically the learning dynamics of the network to reveal the global behavior that emerges in the small step-size regime. The answers to these questions will provide useful and novel insights about how to tune the algorithm parameters in order to reach desired performance levels — see Sec. VI in Part II [44].

B. Relation to Prior Work

In comparison with the existing literature [5]–[10], [28]–[30], [45]–[48], it is worth noting that most prior studies on distributed optimization algorithms focus on studying their performance and convergence under *diminishing* step-size conditions and for *doubly-stochastic* combination policies (i.e., matrices for which the entries on each of their columns *and* on each of their rows add up to one). These are of course useful conditions, especially when the emphasis is on solving *static* optimization problems. We focus instead on the case of *constant* step-sizes because, as explained earlier, they enable continuous adaptation and learning under drifting conditions; in contrast, diminishing step-sizes turn off learning once they approach zero. By using constant step-sizes, the resulting algorithms are able to track *dynamic* solutions that may slowly drift as the underlying problem conditions change.

Moreover, constant step-size implementations have merits even for stationary environments where the solutions remain *static*. This is because, as we are going to show later in this work and its accompanying Part II [44], constant step-size learning converges at a geometric rate, in the order of $O(\gamma^i)$ for some $0 < \gamma < 1$, towards a small mean-square error in the order of the step-size parameter. This means that these solutions can attain satisfactory performance even after short intervals of time. In comparison, implementations that rely on a diminishing step-size of the form $\mu(i) = \mu_o/i$, for some constant μ_o, converge almost surely to the solution albeit at the slower rate of $O(1/i)$. Furthermore, the choice of the parameter μ_o is critical to guarantee the $O(1/i)$ rate [49, p.54]; if μ_o is not large enough, the resulting convergence rate can be considerably slower than $O(1/i)$. To avoid this slowdown in convergence, a large initial value μ_o is usually chosen in practice, which ends up leading to an overshoot in the learning curve; the curve grows up initially before starting its decay at the asymptotic rate, $O(1/i)$.

We remark that we also do not limit the choice of combination policies to being doubly-stochastic; we only require condition (9). It turns out that left-stochastic matrices lead to superior mean-square error performance (see, e.g., expression (63) in Part II [44] and also [17]). The use of both constant step-sizes and left-stochastic combination policies enrich the learning dynamics of the network in interesting ways, as we are going to discover. In particular, under these conditions, we will derive an interesting result that reveals how the topology of the network determines the limit point of the distributed strategies. We will show that the combination weights steer the convergence point away from the expected solution and towards any of many possible Pareto optimal solutions. This is in contrast to commonly-used doubly-stochastic combination policies where the limit point of the network is fixed and cannot be changed regardless of the topology. We will show that the limit point is determined by the right eigenvector that is associated with the eigenvalue at one for the matrix product $A_1A_0A_2$. We will also be able to characterize in Part II [44] how close each agent in the network gets to this limit point and to explain how the limit point plays the role of a Pareto optimal solution for a suitably defined aggregate cost function.

We note that the concept of a limit point in this work is different from earlier studies on the limit point of consensus implementations that deal exclusively with the problem of evaluating the weighted average of initial state values at the agents (e.g., [50]). In these implementations, there are no adaptation steps and no streaming data; the step-size parameters $\{\mu_k\}$ are set to zero in (2), (10), (11) and (12). In contrast, the general distributed strategy (1)–(3) is meant to solve continuous adaptation and learning problems from streaming data arriving at the agents. In this case, the adaptation term $\hat{\boldsymbol{s}}_{k,i}(\cdot)$ (self-learning) is necessary, in addition to the combination step (social-learning). There is a non-trivial coupling between both steps and across the agents. For this reason, identifying the actual limit point of the distributed strategy is rather challenging and requires a close examination of the evolution of the network dynamics, as demonstrated by the technical tools used in this work. In comparison, while the evolution of traditional average-consensus implementations can be described by linear first-order recursions, the same is not true for adaptive networks where the dynamics evolves

according to nonlinear stochastic difference recursions.

III. MODELING ASSUMPTIONS

In this section, we collect the assumptions and definitions that are used in the analysis and explain why they are justified and how they relate to similar assumptions used in several prior studies in the literature. As the discussion will reveal, in most cases, the assumptions that we adopt here are relaxed (i.e., weaker) versions than conditions used before in the literature such as in [5]–[7], [10], [12], [18]–[20], [28]–[30], [49], [51]. We do so in order to analyze the learning behavior of networks under conditions that are similar to what is normally assumed in the prior art, albeit ones that are generally less restrictive.

Assumption 1 (Strongly-connected network): The $N \times N$ matrix product $A \triangleq A_1 A_0 A_2$ is assumed to be a primitive left-stochastic matrix, i.e., $A^T \mathbb{1} = \mathbb{1}$ and there exists a finite integer j_o such that all entries of A^{j_o} are strictly positive. ■

This condition is satisfied for most networks and is not restrictive. Let $A = [a_{lk}]$ denote the entries of A. Assumption 1 is automatically satisfied if the product A corresponds to a connected network and there exists at least one $a_{kk} > 0$ for some node k (i.e., at least one node with a nontrivial self-loop) [21], [23]. It then follows from the Perron-Frobenius Theorem [52] that the matrix $A_1 A_0 A_2$ has a single eigenvalue at one of multiplicity one and all other eigenvalues are strictly less than one in magnitude, i.e.,

$$1 = \lambda_1(A) > |\lambda_2(A)| \geq \cdots \geq |\lambda_N(A)| \tag{13}$$

Obviously, $\mathbb{1}^T$ is a left eigenvector for $A_1 A_0 A_2$ corresponding to the eigenvalue at one. Let θ denote the right eigenvector corresponding to the eigenvalue at one (the Perron vector) and whose entries are normalized to add up to one, i.e.,

$$A\theta = \theta, \qquad \mathbb{1}^T \theta = 1 \tag{14}$$

Then, the Perron-Frobenius Theorem further ensures that all entries of θ satisfy $0 < \theta_k < 1$. Note that, unlike [5]–[11], [28], [29], we do not require the matrix $A_1 A_0 A_2$ to be doubly-stochastic (in which case θ would be $\mathbb{1}/N$ and, therefore, all its entries will be identical to each other). Instead, we will study the performance of the algorithms in the context of general left-stochastic matrices $\{A_1, A_0, A_2\}$ and we will examine the influence of (the generally non-equal entries of) θ on both the limit point and performance of the network.

Definition 1 (Step-sizes): Without loss of generality, we express the step-size at each node k as $\mu_k = \mu_{\max}\beta_k$, where $\mu_{\max} \triangleq \max\{\mu_k\}$ is the largest step-size, and $0 \leq \beta_k \leq 1$. We assume $\beta_k > 0$ for at least one k. Thus, observe that we are allowing the possibility of zero step-sizes by some of the agents.

■

Definition 2 (Useful vectors): Let π and p be the following $N \times 1$ vectors:

$$\pi \triangleq A_2 \theta \tag{15}$$

$$p \triangleq \mathrm{col}\{\pi_1\beta_1, \ldots, \pi_N\beta_N\} \tag{16}$$

where π_k is the kth entry of the vector π. ■

The vector p will play a critical role in the performance of the distributed strategy (1)–(3). Furthermore, we introduce the following assumptions on the update vectors $\hat{\boldsymbol{s}}_{k,i}(\cdot)$ in (1)–(3).

Assumption 2 (Update vector: Randomness): There exists an $M \times 1$ deterministic vector function $s_k(w)$ such that, for all $M \times 1$ vectors $\boldsymbol{w}$ in the filtration $\mathcal{F}_{i-1}$ generated by the past history of iterates $\{\boldsymbol{w}_{k,j}\}$ for $j \leq i-1$ and all k, it holds that

$$\mathbb{E}\,\{\hat{\boldsymbol{s}}_{k,i}(\boldsymbol{w})|\mathcal{F}_{i-1}\} = s_k(\boldsymbol{w}) \tag{17}$$

for all i, k. Furthermore, there exist $\alpha \geq 0$ and $\sigma_v^2 \geq 0$ such that for all i, k and $\boldsymbol{w} \in \mathcal{F}_{i-1}$:

$$\mathbb{E}\left\{\|\hat{\boldsymbol{s}}_{k,i}(\boldsymbol{w}) - s_k(\boldsymbol{w})\|^2 \,\middle|\, \mathcal{F}_{i-1}\right\} \leq \alpha \cdot \|\boldsymbol{w}\|^2 + \sigma_v^2 \tag{18}$$

■

Condition (18) requires the conditional variance of the random update direction $\hat{\boldsymbol{s}}_{k,i}(\boldsymbol{w})$ to be bounded by the square-norm of $\boldsymbol{w}$. Condition (18) is a generalized version of Assumption 2 from [18], [20]; it is also a generalization of the assumptions from [28], [49], [51], where $\hat{\boldsymbol{s}}_{k,i}(\boldsymbol{w})$ was instead modeled as the following perturbed version of the true gradient vector:

$$\hat{\boldsymbol{s}}_{k,i}(\boldsymbol{w}) = \widehat{\nabla_w J_k}(\boldsymbol{w}) = \nabla_w J_k(\boldsymbol{w}) + \boldsymbol{v}_{k,i}(\boldsymbol{w}) \tag{19}$$

with $s_k(\boldsymbol{w}) = \nabla_w J_k(\boldsymbol{w})$, in which case conditions (17)–(18) translate into the following requirements on the gradient noise $\boldsymbol{v}_{k,i}(\boldsymbol{w})$:

$$\mathbb{E}\,\{\boldsymbol{v}_{k,i}(\boldsymbol{w})|\mathcal{F}_{i-1}\} = 0 \qquad (\text{zero mean}) \tag{20}$$

$$\mathbb{E}\left\{\|\boldsymbol{v}_{k,i}(\boldsymbol{w})\|^2 \,\middle|\, \mathcal{F}_{i-1}\right\} \leq \alpha \cdot \|\boldsymbol{w}\|^2 + \sigma_v^2 \tag{21}$$

In Example 2 of [18], we explained how these conditions are satisfied automatically in the context of mean-square-error adaptation over networks. Assumption 2 given by (17)–(18) is more general than (20)–(21) because we are allowing the update vector $\hat{\boldsymbol{s}}_{k,i}(\cdot)$ to be constructed in forms other than (19). Furthermore, Assumption (21) is also more relaxed than the following variant used in [49], [51]:

$$\mathbb{E}\left\{\|\boldsymbol{v}_{k,i}(\boldsymbol{w})\|^2 \,\middle|\, \mathcal{F}_{i-1}\right\} \leq \alpha \cdot \|\nabla_w J_k(\boldsymbol{w})\|^2 + \sigma_v^2 \tag{22}$$

This is because (22) implies a condition of the form (21). Indeed, note that

$$\begin{aligned}
&\mathbb{E}\left\{\|\boldsymbol{v}_{k,i}(\boldsymbol{w})\|^2 \,\middle|\, \mathcal{F}_{i-1}\right\} \\
&\quad = \alpha \cdot \|\nabla_w J_k(\boldsymbol{w}) - \nabla_w J_k(0) + \nabla_w J_k(0)\|^2 + \sigma_v^2 \\
&\quad \overset{(a)}{\leq} 2\alpha \cdot \|\nabla_w J_k(\boldsymbol{w}) - \nabla_w J_k(0)\|^2 + 2\alpha \|\nabla_w J_k(0)\|^2 + \sigma_v^2 \\
&\quad \overset{(b)}{\leq} 2\alpha\lambda_U^2 \cdot \|\boldsymbol{w}\|^2 + 2\alpha \|\nabla_w J_k(0)\|^2 + \sigma_v^2 \\
&\quad \triangleq \alpha' \cdot \|\boldsymbol{w}\|^2 + \sigma_{v'}^2
\end{aligned} \tag{23}$$

where step (a) uses the relation $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$, and step (b) used (24) to be assumed next.

Assumption 3 (Update vector: Lipschitz): There exists a nonnegative λ_U such that for all $x, y \in \mathbb{R}^M$ and all k:

$$\|s_k(x) - s_k(y)\| \leq \lambda_U \cdot \|x - y\| \tag{24}$$

where the subscript "U" in λ_U means "upper bound". ■

A similar assumption to (24) was used before in the literature for the model (19) by requiring the gradient vector of the individual cost functions $J_k(w)$ to be Lipschitz [5], [12], [30], [49], [51]. Again, condition (24) is more general because we are not limiting the construction of the update direction to (19).

Assumption 4 (Update vector: Strong monotonicity): Let p_k denote the kth entry of the vector p defined in (16). There exists $\lambda_L > 0$ such that for all $x, y \in \mathbb{R}^M$:

$$(x-y)^T \cdot \sum_{k=1}^{N} p_k \Big[s_k(x) - s_k(y) \Big] \geq \lambda_L \cdot \|x - y\|^2 \tag{25}$$

where the subscript "L" in λ_L means "lower bound". ■

Remark 1: Applying the Cauchy-Schwartz inequality [52, p.15] to the left-hand side of (25) and using (24), we deduce the following relation between λ_L and λ_U:

$$\lambda_U \cdot \|p\|_1 \geq \lambda_L \tag{26}$$

where $\|\cdot\|_1$ denotes the $1-$norm of the vector argument. ■

The following lemma gives the equivalent forms of Assumptions 3–4 when the $\{s_k(w)\}$ happen to be differentiable.

Lemma 1 (Equivalent conditions on update vectors): Suppose $\{s_k(w)\}$ are differentiable in an open set $\mathcal{S} \subseteq \mathbb{R}^M$. Then, having conditions (24) and (25) hold on $\mathcal{S}$ is equivalent to the following conditions, respectively,

$$\|\nabla_{w^T} s_k(w)\| \leq \lambda_U \tag{27}$$

$$\frac{1}{2}[H_c(w) + H_c^T(w)] \geq \lambda_L \cdot I_M \tag{28}$$

for any $w \in \mathcal{S}$, where $\|\cdot\|$ denotes the 2-induced norm (largest singular value) of its matrix argument and

$$H_c(w) \triangleq \sum_{k=1}^{n} p_k \nabla_{w^T} s_k(w) \tag{29}$$

Proof: See Appendix B. ■

Since in Assumptions 3–4 we require conditions (24) and (25) to hold over the entire $\mathbb{R}^M$, then the equivalent conditions (27)–(28) will need to hold over the entire $\mathbb{R}^M$ when the $\{s_k(w)\}$ are differentiable. In the context of distributed optimization problems of the form (7)–(8) with twice-differentiable $J_k(w)$, where the stochastic gradient vectors are constructed as in (19), Lemma 1 implies that the above Assumptions 3–4 are equivalent to the following conditions on the Hessian matrix of each $J_k(w)$ [49, p.10]:

$$\big\|\nabla_w^2 J_k(w)\big\| \leq \lambda_U \tag{30}$$

$$\sum_{k=1}^{N} p_k \nabla_w^2 J_k(w) \geq \lambda_L I_M > 0 \tag{31}$$

Condition (31) is in turn equivalent to requiring the following weighted sum of the individual cost functions $\{J_k(w)\}$ to be strongly convex:

$$J^{\mathrm{glob},\star}(w) \triangleq \sum_{k=1}^{N} p_k J_k(w) \tag{32}$$

We note that strong convexity conditions are prevalent in many studies on optimization techniques in the literature. For example, each of the individual costs $J_k(w)$ is assumed to be stronlgy convex in [29] in order to derive upper bounds on the limit superior ("$\limsup$") of the mean-square-error of the estimates $\boldsymbol{w}_{k,i}$ or the expected value of the cost function at $\boldsymbol{w}_{k,i}$. In comparison, the framework in this work does not require the individual costs to be strongly convex or even convex. Actually, some of the costs $\{J_k(w)\}$ can be non-convex as long as the aggregate cost (32) remains strongly convex. Such relaxed assumptions on the individual costs introduce challenges into the analysis, and we need to develop a systematic approach to characterize the limiting behavior of adaptive networks under such less restrictive conditions.

Example 2: The strong-convexity condition (31) on the aggregate cost (32) can be related to a global observability condition similar to [6]–[8]. To illustrate this point, we consider an example dealing with quadratic costs. Thus, consider a network of N agents that are connected according to a certain topology. The data samples received at each agent k at time i consist of the observation signal $\boldsymbol{d}_k(i) \in \mathbb{R}$ and the regressor vector $\boldsymbol{u}_{k,i} \in \mathbb{R}^{1\times M}$, which are assumed to be related according to the following linear model:

$$\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i} w^o + \boldsymbol{v}_k(i) \tag{33}$$

where $\boldsymbol{v}_k(i) \in \mathbb{R}$ is a zero-mean additive white noise that is uncorrelated with the regressor vector $\boldsymbol{u}_{\ell,j}$ for all k, ℓ, i, j. Each agent in the network would like to estimate $w^o \in \mathbb{R}^M$ by learning from the local data stream $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$ and by collaborating with its intermediate neighbors. The problem can be formulated as minimizing the aggregate cost (7) with $J_k(w)$ chosen to be

$$J_k(w) = \frac{1}{2}\mathbb{E}|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i} w|^2 \tag{34}$$

i.e.,

$$J^{\mathrm{glob}}(w) = \sum_{k=1}^{N} \frac{1}{2}\mathbb{E}|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i} w|^2 \tag{35}$$

This is a distributed least-mean-squares (LMS) estimation problem studied in [16], [17], [19]. We would like to explain that condition (31) amounts to a global observability condition. First, note that the Hessian matrix of $J_k(w)$ in this case is the covariance matrix of the regressor $\boldsymbol{u}_{k,i}$:

$$R_{u,k} \triangleq \mathbb{E}\{\boldsymbol{u}_{k,i}^T \boldsymbol{u}_{k,i}\} \tag{36}$$

Therefore, condition (31) becomes that there exists a $\lambda_L > 0$ such that

$$\sum_{k=1}^{N} p_k R_{u,k} \geq \lambda_L I_M > 0 \tag{37}$$

Furthermore, it can be verified that the above inequality holds for any positive $\{p_k\}$ as long as the following *global observability* condition holds:

$$\sum_{k=1}^{N} R_{u,k} > 0 \tag{38}$$

To see this, let $p_{\min} = \min_k p_k$, and write the left-hand side of (37) as

$$\begin{aligned}\sum_{k=1}^{N} p_k R_{u,k} &= p_{\min} \sum_{k=1}^{N} R_{u,k} + \sum_{k=1}^{N} (p_k - p_{\min}) R_{u,k} \\ &\geq p_{\min} \sum_{k=1}^{N} R_{u,k} > \underbrace{p_{\min} \lambda_{u,\min}}_{\triangleq \lambda_L} \cdot I_M \end{aligned} \tag{39}$$

where $\lambda_{u,\min}$ denotes the minimum eigenvalue of $\sum_{k=1}^{N} R_{u,k}$. Note that the left-hand side of (38) is the Hessian of $J^{\rm glob}(w)$ in (35). Therefore, condition (38) means that the aggregate cost function (35) is strongly convex so that the information provided by the linear observation model (33) over the entire network is sufficient to uniquely identify the minimizer of (35). Similar global observability conditions were used in [6]–[8] to study the performance of distributed parameter estimation problems. Such conditions are useful because it implies that even if w^o is not locally observable to any agent in the network but is globally observable, i.e., $R_{u,k} > 0$ does not hold for any $k = 1, \ldots, N$ but (38) holds, the distributed strategies (10)–(12) will still enable each agent to estimate the correct w^o through local cooperation. In Part II [44], we provide more insights into how cooperation benefits the learning at each agent. ■

Assumption 5 (Jacobian matrix: Lipschitz): Let w^o denote the limit point of the distributed strategy (1)–(3), which is defined further ahead as the unique solution to (42). Then, in a small neighborhood around w^o, we assume that $s_k(w)$ is differentiable with respect to w and satisfies

$$\|\nabla_{w^T} s_k(w^o + \delta w) - \nabla_{w^T} s_k(w^o)\| \leq \lambda_H \cdot \|\delta w\| \tag{40}$$

for all $\|\delta w\| \leq r_H$ for some small r_H, and where λ_H is a nonnegative number independent of δw.

■

In the context of distributed optimization problems of the form (7)–(8) with twice-differentiable $J_k(w)$, where the stochastic gradient vectors are constructed as in (19), the above Assumption translates into the following Lipschitz Hessian condition:

$$\|\nabla_w^2 J_k(w^o + \delta w) - \nabla_w^2 J_k(w^o)\| \leq \lambda_H \cdot \|\delta w\| \tag{41}$$

Condition (40) is useful when we examine the convergence rate of the algorithm later in this article. It is also useful in deriving the steady-state mean-square-error expression (52) in Part II [44].

IV. Learning Behavior

A. Overview of Main Results

Before we proceed to the formal analysis, we first give a brief overview of the main results that we are going to establish in this part on the learning behavior of the distributed strategies (1)–(3) for sufficiently small step-sizes. The first major conclusion is that for general *left-stochastic* matrices $\{A_1, A_0, A_2\}$, the agents in the network will have their estimators $\boldsymbol{w}_{k,i}$ converge, in the mean-square-error sense, to the *same* vector w^o that corresponds to the unique solution of the following algebraic equation:

$$\sum_{k=1}^{N} p_k s_k(w) = 0 \tag{42}$$

For example, in the context of distributed optimization problems of the form (7), this result implies that for left-stochastic matrices $\{A_1, A_0, A_2\}$, the distributed strategies represented by (1)–(3) will *not* converge to the global minimizer of the original aggregate cost (7), which is the unique solution to the alternative algebraic equation

$$\sum_{k=1}^{N} \nabla_w J_k(w) = 0 \tag{43}$$

Instead, these distributed solutions will converge to the global minimizer of the *weighted* aggregate cost $J^{\rm glob,\star}(w)$ defined by (32) in terms of the entries p_k, i.e., to the unique solution of

$$\sum_{k=1}^{N} p_k \nabla_w J_k(w) = 0 \tag{44}$$

Result (42) also means that the distributed strategies (1)–(3) converge to a Pareto optimal solution of the multi-objective problem (8); one Pareto solution for each selection of the topology parameters $\{p_k\}$. The distinction between the aggregate costs $J^{\rm glob}(w)$ and $J^{\rm glob,\star}(w)$ does not appear in earlier studies on distributed optimization [5]–[11], [28]–[30] mainly because these studies focus on *doubly-stochastic* combination matrices, for which the entries $\{p_k\}$ will all become equal to each other for uniform step-sizes $\mu_k \equiv \mu$ or $\mu_k(i) \equiv \mu(i)$. In that case, the minimizations of (7) and (32) become equivalent and the solution of (43) and (44) would then coincide. In other words, regardless of the choice of the doubly stochastic combination weights, when the $\{p_k\}$ are identical, the limit point will be unique and correspond to the solution of

$$\sum_{k=1}^{N} s_k(w) = 0 \tag{45}$$

In contrast, result (42) shows that left-stochastic combination policies add more flexibility into the behavior of the network. By selecting different combination weights, or even different topologies, the entries $\{p_k\}$ can be made to change and the limit point can be steered towards other desired Pareto optimal solutions. Even in the traditional case of consensus-type implementations for computing averages, as opposed to learning from streaming data, it also holds that it is beneficial to relax the requirement of a doubly-stochastic combination policy in order to enable broadcast algorithms without feedback [53].

The second major conclusion of the paper is that we will show in (129) further ahead that there always exist sufficiently small step-sizes such that the learning process over the network

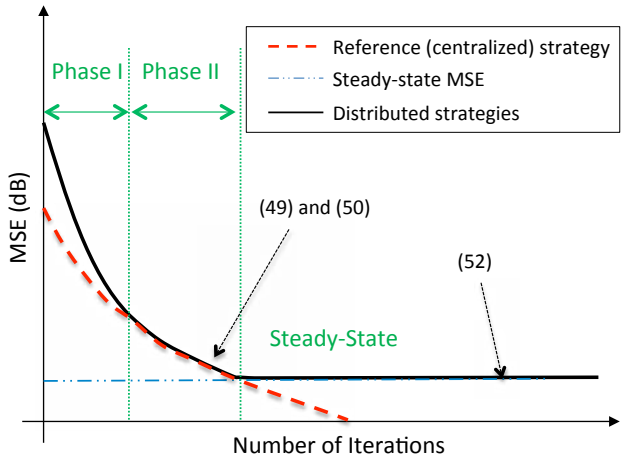


Fig. 2. A typical mean-square-error (MSE) learning curve includes a transient stage that consists of two phases and a steady-state phase. The plot shows how the learning curve of a network of agents compares to the learning curve of a centralized reference solution. The analysis in this work, and in the accompanying Part II [44] characterizes in detail the parameters that determine the behavior of the network (rate, stability, and performance) during each phase of the learning process.

is mean-square stable. This means that the weight error vectors relative to w^o will satisfy

$$\limsup_{i\to\infty} \mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 \leq O(\mu_{\max}) \tag{46}$$

so that the steady-state mean-square-error at each agent will be of the order of $O(\mu_{\max})$.

The third major conclusion of our analysis is that we will show that, during the convergence process towards the limit point w^o, the learning curve at each agent exhibits *three* distinct phases: Transient Phase I, Transient Phase II, and Steady-State Phase. These phases are illustrated in Fig. 2 and they are interpreted as follows. Let us first introduce a *reference* (centralized) procedure that is described by the following centralized-type recursion:

$$\bar{w}_{c,i} = \bar{w}_{c,i-1} - \mu_{\max} \sum_{k=1}^{N} p_k s_k(\bar{w}_{c,i-1}) \tag{47}$$

which is initialized at

$$\bar{w}_{c,0} = \sum_{k=1}^{N} \theta_k w_{k,0} \tag{48}$$

where θ_k is the kth entry of the eigenvector θ, $\mu_{\max}$, and $\{p_k\}$ are defined in Definitions 1–2, $w_{k,0}$ is the initial value of the distributed strategy at agent k, and $\bar{w}_{c,i}$ is an $M \times 1$ vector generated by the reference recursion (47). The three phases of the learning curve will be shown to have the following features:

- **Transient Phase I:**
  If agents are initialized at different values, then the estimates of the various agents will initially evolve in such a way to make each $\boldsymbol{w}_{k,i}$ get closer to the reference recursion $\bar{w}_{c,i}$. The rate at which the agents approach $\bar{w}_{c,i}$ will be determined by $|\lambda_2(A)|$, the second largest eigenvalue of A in magnitude. If the agents are initialized at the same value, say, e.g., $\boldsymbol{w}_{k,0} = 0$, then the learning curves start at Transient Phase II directly.
- **Transient Phase II:**
  In this phase, the trajectories of all agents are uniformly close to the trajectory of the reference recursion; they converge in a coordinated manner to steady-state. The learning curves at this phase are well modeled by the same reference recursion (47) since we will show in (145) that:

$$\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 = \|\tilde{w}_{c,i}\|^2 + O(\mu_{\max}^{1/2}) \cdot \gamma_c^i + O(\mu_{\max}) \tag{49}$$

  Furthermore, for small step-sizes and during the later stages of this phase, $\bar{w}_{c,i}$ will be close enough to w^o and the convergence rate r will be shown to satisfy:

$$r = \left[\rho(I_M - \mu_{\max} H_c)\right]^2 + O\big((\mu_{\max}\epsilon)^{\frac{1}{2(M-1)}}\big) \tag{50}$$

  where $\rho(\cdot)$ denotes the spectral radius of its matrix argument, ϵ is an arbitrarily small positive number, and H_c is the same matrix that results from evaluating (29) at $w = w^o$, i.e.,

$$H_c \triangleq \sum_{k=1}^{N} p_k H_k = H_c(w^o) \tag{51}$$

  where $H_k \triangleq \nabla_{w^T} s_k(w^o)$.
- **Steady-State Phase:**
  The reference recursion (47) continues converging towards w^o so that $\|\tilde{w}_{c,i}\|^2 = \|w^o - \bar{w}_{c,i}\|^2$ will converge to zero ($-\infty$ dB in Fig. 2). However, for the distributed strategy (1)–(3), the mean-square-error $\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 = \mathbb{E}\|w^o - \boldsymbol{w}_{k,i}\|^2$ at each agent k will converge to a *finite* steady-state value. We will be able to characterize this value in terms of the vector p in Part II [44] as follows:[1]

$$\lim_{i\to\infty} \mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 = \mu_{\max} \cdot \mathrm{Tr}\left\{X(p^T \otimes I_M)\mathcal{R}_v(p \otimes I_M)\right\} + o(\mu_{\max}) \tag{52}$$

  where X is the solution to the Lyapunov equation described later in (41) of Part II [44] (when $\Sigma = I$), and $o(\mu_{\max})$ denotes a strictly higher order term of $\mu_{\max}$. Expression (52) is a revealing result. It is a non-trivial extension of a classical result pertaining to the mean-square-error performance of stand-alone adaptive filters [54]–[57] to the more demanding context when a multitude of adaptive agents are coupled together in a cooperative manner through a network topology. This result has an important ramification, which we pursue in Part II [44]. We will show there that no matter how the agents are connected to each other, there is always a way to select the combination weights such that the performance of the network is invariant to the topology. This will also imply that, for any connected topology, there is always a way to select the combination weights such that the performance of the network matches that of the centralized solution.

[1]The interpretation of the limit in (52) is explained in more detail in Sec. IV of Part II [44].

Note that the above results are obtained for the general distributed strategy (1)–(3). Therefore, the results can be specialized to the consensus, CTA diffusion, and ATC diffusion strategies in (10)–(12) by choosing the matrices A_1, A_2, and A_0 according to Tab. I. The results in this paper and its accompanying Part II [44] not only generalize the analysis from earlier works [16]–[20] but, more importantly, they also provide deeper insights into the learning behavior of these adaptation and learning strategies.

V. Study of Error Dynamics

A. Error Quantities

We shall examine the learning behavior of the distributed strategy (1)–(3) by examining how the perturbation between the distributed solution (1)–(3) and the reference solution (47) evolves over time — see Fig. 3. Specifically, let $\check{\boldsymbol{w}}_{k,i}$ denote the discrepancy between $\boldsymbol{w}_{k,i}$ and $\bar{w}_{c,i}$, i.e.,

$$\check{\boldsymbol{w}}_{k,i} \triangleq \boldsymbol{w}_{k,i} - \bar{w}_{c,i} \tag{53}$$

and let $\boldsymbol{w}_i$ and $\check{\boldsymbol{w}}_i$ denote the global vectors that collect the $\boldsymbol{w}_{k,i}$ and $\check{\boldsymbol{w}}_{k,i}$ from across the network, respectively:

$$\boldsymbol{w}_i \triangleq \mathrm{col}\{\boldsymbol{w}_{1,i}, \ldots, \boldsymbol{w}_{N,i}\} \tag{54}$$

$$\check{\boldsymbol{w}}_i \triangleq \mathrm{col}\{\check{\boldsymbol{w}}_{1,i}, \ldots, \check{\boldsymbol{w}}_{N,i}\} = \boldsymbol{w}_i - \mathbb{1} \otimes \bar{w}_{c,i} \tag{55}$$

It turns out that it is insightful to study the evolution of $\check{\boldsymbol{w}}_i$ in a *transformed* domain where it is possible to express the distributed recursion (1)–(3) as a perturbed version of the reference recursion (47).

Definition 3 (Network basis transformation): We define the transformation by introducing the Jordan canonical decomposition of the matrix $A = A_1 A_0 A_2$. Let

$$A^T = UDU^{-1} \tag{56}$$

where U is an invertible matrix whose columns correspond to the right-eigenvectors of A^T, and D is a block Jordan matrix with a single eigenvalue at one with multiplicity one while all other eigenvalues are strictly less than one. The Kronecker form of A then admits the decomposition:

$$\mathcal{A}^T \triangleq A^T \otimes I_M = \mathcal{U}\mathcal{D}\mathcal{U}^{-1} \tag{57}$$

where

$$\mathcal{U} \triangleq U \otimes I_M, \qquad \mathcal{D} \triangleq D \otimes I_M \tag{58}$$

We use $\mathcal{U}$ to define the following basis transformation:

$$\boldsymbol{w}'_i \triangleq \mathcal{U}^{-1}\boldsymbol{w}_i = (U^{-1} \otimes I_M)\boldsymbol{w}_i \tag{59}$$

$$\check{\boldsymbol{w}}'_i \triangleq \mathcal{U}^{-1}\check{\boldsymbol{w}}_i = (U^{-1} \otimes I_M)\check{\boldsymbol{w}}_i \tag{60}$$

The relations between the quantities in transformations (59)–(60) are illustrated in Fig. 3(a). ■

We can gain useful insight into the nature of this transformation by exploiting more directly the structure of the matrices $\mathcal{U}$, $\mathcal{D}$, and $\mathcal{U}^{-1}$. By Assumption 1, the matrix A^T has an eigenvalue one of multiplicity one, with the corresponding left- and right-eigenvectors being θ^T and $\mathbb{1}$, respectively. All other eigenvalues of D are strictly less than one in magnitude. Therefore, the matrices D, U, and U^{-1} can be partitioned as

$$D = \left[\begin{array}{c|c} 1 & \\ \hline & D_{N-1} \end{array}\right] \quad U = \left[\begin{array}{c|c} \mathbb{1} & U_L \end{array}\right] \quad U^{-1} = \left[\begin{array}{c} \theta^T \\ \hline U_R \end{array}\right] \tag{61}$$

where D_{N-1} is an $(N-1)\times(N-1)$ Jordan matrix with all diagonal entries strictly less than one in magnitude, U_L is an $N \times (N-1)$ matrix, and U_R is an $(N-1)\times N$ matrix. Then, the Kronecker forms $\mathcal{D}$, $\mathcal{U}$, and $\mathcal{U}^{-1}$ can be expressed as

$$\mathcal{D} = \left[\begin{array}{c|c} I_M & \\ \hline & \mathcal{D}_{N-1} \end{array}\right], \; \mathcal{U} = \left[\begin{array}{c|c} \mathbb{1}\otimes I_M & \mathcal{U}_L \end{array}\right], \; \mathcal{U}^{-1} = \left[\begin{array}{c} \theta^T \otimes I_M \\ \hline \mathcal{U}_R \end{array}\right] \tag{62}$$

where

$$\mathcal{U}_L \triangleq U_L \otimes I_M \tag{63}$$

$$\mathcal{U}_R \triangleq U_R \otimes I_M \tag{64}$$

$$\mathcal{D}_{N-1} \triangleq D_{N-1} \otimes I_M \tag{65}$$

It is important to note that $U^{-1}U = I_N$ and that

$$\theta^T\mathbb{1} = 1, \quad \theta^T U_L = 0, \quad U_R\mathbb{1} = 0, \quad U_R U_L = I_{N-1} \tag{66}$$

We first study the structure of $\boldsymbol{w}'_i$ defined in (59) using (61):

$$\boldsymbol{w}'_i = \mathrm{col}\{\underbrace{(\theta^T \otimes I_M)\boldsymbol{w}_i}_{\triangleq \boldsymbol{w}_{c,i}}, \; \underbrace{(U_R \otimes I_M)\boldsymbol{w}_i}_{\triangleq \boldsymbol{w}_{e,i}}\} \tag{67}$$

The two components $\boldsymbol{w}_{c,i}$ and $\boldsymbol{w}_{e,i}$ have useful interpretations. Recalling that θ_k denotes the kth entry of the vector θ, then $\boldsymbol{w}_{c,i}$ can be expressed as

$$\boldsymbol{w}_{c,i} = \sum_{k=1}^{N} \theta_k \boldsymbol{w}_{k,i} \tag{68}$$

As we indicated after Assumption 1, the entries $\{\theta_k\}$ are positive and add up to one. Therefore, $\boldsymbol{w}_{c,i}$ is a weighted average (i.e., the centroid) of the estimates $\{\boldsymbol{w}_{k,i}\}$ across all agents. To interpret $\boldsymbol{w}_{e,i}$, we examine the inverse mapping of (59) from $\boldsymbol{w}'_i$ to $\boldsymbol{w}_i$ using the block structure of $\mathcal{U}$ in (61):

$$\begin{aligned} \boldsymbol{w}_i &= (U \otimes I_M)\boldsymbol{w}'_i \\ &= (\mathbb{1} \otimes I_M)\boldsymbol{w}_{c,i} + (U_L \otimes I_M)\boldsymbol{w}_{e,i} \\ &= \mathbb{1} \otimes \boldsymbol{w}_{c,i} + (U_L \otimes I_M)\boldsymbol{w}_{e,i} \end{aligned} \tag{69}$$

which implies that the individual estimates at the various agents satisfy:

$$\boldsymbol{w}_{k,i} = \boldsymbol{w}_{c,i} + (u_{L,k} \otimes I_M)\boldsymbol{w}_{e,i} \tag{70}$$

where $u_{L,k}$ denotes the kth row of the matrix U_L. The network basis transformation defined by (59) represents the cluster of iterates $\{\boldsymbol{w}_{k,i}\}$ by its centroid $\boldsymbol{w}_{c,i}$ and their positions $\{u_{L,k}\otimes I_M)\boldsymbol{w}_{e,i}\}$ relative to the centroid as shown in Fig. 3. The two parts, $\boldsymbol{w}_{c,i}$ and $\boldsymbol{w}_{e,i}$, of $\boldsymbol{w}'_i$ in (67) are the coordinates in this new transformed representation. Then, the actual error quantity $\tilde{\boldsymbol{w}}_{k,i}$ relative to w^o can be represented as

$$\begin{aligned} \tilde{\boldsymbol{w}}_{k,i} &= w^o - \bar{w}_{c,i} - (\boldsymbol{w}_{k,i} - \bar{w}_{c,i}) \\ &= w^o - \bar{w}_{c,i} - (\boldsymbol{w}_{c,i} + (u_{L,k} \otimes I_M)\boldsymbol{w}_{e,i} - \bar{w}_{c,i}) \end{aligned} \tag{71}$$

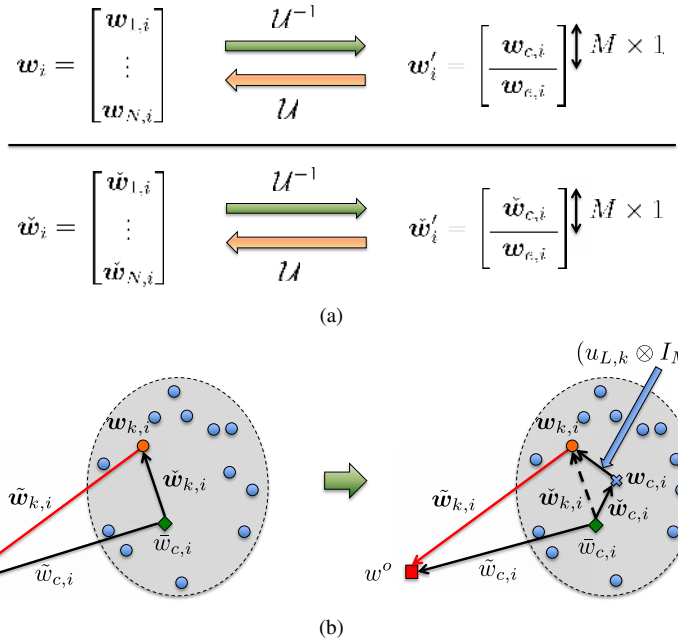


Fig. 3. (a) Network basis transformation. (b) The diagrams show how the iterate $\boldsymbol{w}_{k,i}$ is decomposed relative to the reference $\bar{w}_{c,i}$ and relative to the centroid, $\boldsymbol{w}_{c,i}$, of the N iterates across the network.

Introduce

$$\tilde{w}_{c,i} \triangleq w^o - \bar{w}_{c,i} \tag{72}$$

$$\check{\boldsymbol{w}}_{c,i} \triangleq \boldsymbol{w}_{c,i} - \bar{w}_{c,i} \tag{73}$$

Then, from (71) we arrive at the following critical relation for our analysis in the sequel:

$$\tilde{\boldsymbol{w}}_{k,i} = \tilde{w}_{c,i} - \check{\boldsymbol{w}}_{c,i} - (u_{L,k} \otimes I_M)\boldsymbol{w}_{e,i} \tag{74}$$

This relation is also illustrated in Fig. 3. Then, the behavior of the error quantities $\{\tilde{\boldsymbol{w}}_{k,i}\}$ can be studied by examining $\tilde{w}_{c,i}$, $\check{\boldsymbol{w}}_{c,i}$ and $\boldsymbol{w}_{e,i}$, respectively, which is pursued in Sec. VI further ahead. The first term is the error between the reference recursion and w^o, which is studied in Theorems 1–3. The second quantity is the difference between the weighted centroid $\boldsymbol{w}_{c,i}$ of the cluster and the reference vector $\bar{w}_{c,i}$, and the third quantity characterizes the positions of the individual iterates $\{\boldsymbol{w}_{k,i}\}$ relative to the centroid $\boldsymbol{w}_{c,i}$. As long as the second and the third terms in (74), or equivalently, $\check{\boldsymbol{w}}_{c,i}$ and $\boldsymbol{w}_{e,i}$, are small (which will be shown in Theorem 4), the behavior of each $\boldsymbol{w}_{k,i}$ can be well approximated by the behavior of the reference vector $\bar{w}_{c,i}$. Indeed, $\check{\boldsymbol{w}}_{c,i}$ and $\boldsymbol{w}_{e,i}$ are the coordinates of the transformed vector $\check{\boldsymbol{w}}'_i$ defined by (60). To see this, we substitute (61) and (55) into (60) to get

$$\begin{aligned} \check{\boldsymbol{w}}'_i &= (U^{-1} \otimes I_M)(\boldsymbol{w}_i - \mathbb{1} \otimes \bar{w}_{c,i}) \\ &= \boldsymbol{w}'_i - (U^{-1}\mathbb{1}) \otimes \bar{w}_{c,i} \end{aligned} \tag{75}$$

Recalling (66) and the expression for U^{-1} in (61), we obtain

$$\begin{aligned} U^{-1}\mathbb{1} &= \mathrm{col}\{\theta^T\mathbb{1},\ U_R\mathbb{1}\} \\ &= \mathrm{col}\{1,\ 0_{N-1}\} \end{aligned} \tag{76}$$

where 0_{N-1} denotes an $(N-1)\times 1$ vector with all zero entries. Substituting (76) and (67) into (75), we get

$$\check{\boldsymbol{w}}'_i = \mathrm{col}\{\boldsymbol{w}_{c,i} - \bar{w}_{c,i},\ \boldsymbol{w}_{e,i}\} = \mathrm{col}\{\check{\boldsymbol{w}}_{c,i},\ \boldsymbol{w}_{e,i}\} \tag{77}$$

Therefore, it suffices to study the dynamics of $\check{\boldsymbol{w}}'_i$ and its mean-square performance. We will establish joint recursions for $\boldsymbol{w}_{c,i}$ and $\boldsymbol{w}_{e,i}$ in Sec. V-B, and joint recursions for $\check{\boldsymbol{w}}_{c,i}$ and $\boldsymbol{w}_{e,i}$ in Sec. V-C. Table II summarizes the definitions of the various quantities, the recursions that they follow, and their relations.

B. Signal Recursions

We now derive the joint recursion that describes the evolution of the quantities $\check{\boldsymbol{w}}_{c,i} = \boldsymbol{w}_{c,i} - \bar{w}_{c,i}$ and $\boldsymbol{w}_{e,i}$. Since $\bar{w}_{c,i}$ follows the reference recursion (47), it suffices to derive the joint recursion for $\boldsymbol{w}_{c,i}$ and $\boldsymbol{w}_{e,i}$. To begin with, we introduce the following global quantities:

$$\mathcal{A} = A \otimes I_M \tag{78}$$

$$\mathcal{A}_0 = A_0 \otimes I_M \tag{79}$$

$$\mathcal{A}_1 = A_1 \otimes I_M \tag{80}$$

$$\mathcal{A}_2 = A_2 \otimes I_M \tag{81}$$

$$\mathcal{M} = \Omega \otimes I_M \tag{82}$$

$$\Omega = \mathrm{diag}\{\mu_1, \ldots, \mu_N\} \tag{83}$$

We also let the notation $x = \mathrm{col}\{x_1, \ldots, x_N\}$ denote an arbitrary $N\times 1$ block column vector that is formed by stacking $M\times 1$ sub-vectors $x_1, \ldots, x_N$ on top of each other. We further define the following global update vectors:

$$\hat{\boldsymbol{s}}_i(x) \triangleq \mathrm{col}\{\hat{\boldsymbol{s}}_{1,i}(x_1), \ldots, \hat{\boldsymbol{s}}_{N,i}(x_N)\} \tag{84}$$

$$s(x) \triangleq \mathrm{col}\{s_1(x_1), \ldots, s_N(x_N)\} \tag{85}$$

Then, the general recursion for the distributed strategy (1)–(3) can be rewritten in terms of these extended quantities as follows:

$$\boldsymbol{w}_i = \mathcal{A}^T\boldsymbol{w}_{i-1} - \mathcal{A}_2^T\mathcal{M}\hat{\boldsymbol{s}}_i(\boldsymbol{\phi}_{i-1}) \tag{86}$$

TABLE II
SUMMARY OF VARIOUS ITERATES, ERROR QUANTITIES, AND THEIR RELATIONS.

	Original system			Transformed system[a]			Reference system	
Quantity	$\boldsymbol{w}_{k,i}$	$\tilde{\boldsymbol{w}}_{k,i}$	$\check{\boldsymbol{w}}_{k,i}$	$\boldsymbol{w}_{c,i}$	$\check{\boldsymbol{w}}_{c,i}$	$\boldsymbol{w}_{e,i}$	$\bar{w}_{c,i}$	$\tilde{w}_{c,i}$
Definition	Iterate at agent k	$w^o-\boldsymbol{w}_{k,i}$	$\boldsymbol{w}_{k,i}-\bar{w}_{c,i}$	$\sum_{k=1}^{N}\theta_k\boldsymbol{w}_{k,i}$	$\boldsymbol{w}_{c,i}-\bar{w}_{c,i}$	$\mathcal{U}_R\boldsymbol{w}_i$	Ref. Iterate	$w^o-\bar{w}_{c,i}$
Recursion	Eqs. (1)–(3)			Eq. (96)	Eq. (103)	Eq. (104)	Eq. (47)	

[a] The transformation is defined by (59)–(60).

where

$$\boldsymbol{\phi}_i \triangleq \mathrm{col}\{\boldsymbol{\phi}_{1,i},\ldots,\boldsymbol{\phi}_{N,i}\} \tag{87}$$

and is related to $\boldsymbol{w}_i$ and $\boldsymbol{w}'_i$ via the following relation

$$\boldsymbol{\phi}_i = \mathcal{A}_1^T\boldsymbol{w}_i = \mathcal{A}_1^T\mathcal{U}\boldsymbol{w}'_i \tag{88}$$

Applying the transformation (59) to both sides of (86), we obtain the transformed global recursion:

$$\boldsymbol{w}'_i = \mathcal{D}\boldsymbol{w}'_{i-1} - \mathcal{U}^{-1}\mathcal{A}_2^T\mathcal{M}\hat{\boldsymbol{s}}_i\left(\boldsymbol{\phi}_{i-1}\right) \tag{89}$$

We can now use the block structures in (62) and (67) to derive recursions for $\boldsymbol{w}_{c,i}$ and $\boldsymbol{w}_{e,i}$ from (89). Substituting (62) and (67) into (89), and using properties of Kronecker products [58, p.147], we obtain

$$\begin{aligned}\boldsymbol{w}_{c,i} &= \boldsymbol{w}_{c,i-1} - (\theta^T\otimes I_M)\mathcal{A}_2^T\mathcal{M}\hat{\boldsymbol{s}}_i\left(\boldsymbol{\phi}_{i-1}\right)\\ &= \boldsymbol{w}_{c,i-1} - (\theta^T A_2^T\Omega\otimes I_M)\hat{\boldsymbol{s}}_i\left(\boldsymbol{\phi}_{i-1}\right)\\ &= \boldsymbol{w}_{c,i-1} - \mu_{\max}\cdot(p^T\otimes I_M)\hat{\boldsymbol{s}}_i\left(\boldsymbol{\phi}_{i-1}\right)\end{aligned} \tag{90}$$

and

$$\boldsymbol{w}_{e,i} = \mathcal{D}_{N-1}\boldsymbol{w}_{e,i-1} - \mathcal{U}_R\mathcal{A}_2^T\mathcal{M}\hat{\boldsymbol{s}}_i\left(\boldsymbol{\phi}_{i-1}\right) \tag{91}$$

where in the last step of (90) we used the relation

$$\mu_{\max}\cdot p = \Omega A_2\theta \tag{92}$$

which follows from Definitions 1 and 2. Furthermore, by adding and subtracting identical factors, the term $\hat{\boldsymbol{s}}_i\left(\boldsymbol{\phi}_{i-1}\right)$ that appears in (90) and (91) can be expressed as

$$\begin{aligned}\hat{\boldsymbol{s}}_i\left(\boldsymbol{\phi}_{i-1}\right) = s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1}) &+ \underbrace{\hat{\boldsymbol{s}}_i\left(\boldsymbol{\phi}_{i-1}\right)-s\left(\boldsymbol{\phi}_{i-1}\right)}_{\triangleq\boldsymbol{v}_i(\boldsymbol{\phi}_{i-1})}\\ &+ \underbrace{s\left(\boldsymbol{\phi}_{i-1}\right)-s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})}_{\triangleq\boldsymbol{z}_{i-1}}\end{aligned} \tag{93}$$

where the first perturbation term $\boldsymbol{v}_i(\boldsymbol{\phi}_{i-1})$ consists of the difference between the true update vectors $\{s_k(\boldsymbol{\phi}_{k,i-1})\}$ and their stochastic approximations $\{\hat{\boldsymbol{s}}_{k,i}(\boldsymbol{\phi}_{k,i-1})\}$, while the second perturbation term $\boldsymbol{z}_{i-1}$ represents the difference between the same $\{s_k(\boldsymbol{\phi}_{k,i-1})\}$ and $\{s_k(\boldsymbol{w}_{c,i-1})\}$. The subscript $i-1$ in $\boldsymbol{z}_{i-1}$ implies that this variable depends on data up to time $i-1$ and the subscript i in $\boldsymbol{v}_i(\boldsymbol{\phi}_{i-1})$ implies that its value depends on data up to time i (since, in general, $\hat{\boldsymbol{s}}_i(\cdot)$ can depend on data from time i — see Eq. (33) in Part II for an example). Then, $\hat{\boldsymbol{s}}_i\left(\boldsymbol{\phi}_{i-1}\right)$ can be expressed as

$$\hat{\boldsymbol{s}}_i\left(\boldsymbol{\phi}_{i-1}\right) = s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1}) + \boldsymbol{v}_i + \boldsymbol{z}_{i-1} \tag{94}$$

Lemma 2 (Signal dynamics): In summary, the previous derivation shows that the weight iterates at each agent evolve according to the following dynamics:

$$\boldsymbol{w}_{k,i} = \boldsymbol{w}_{c,i} + (u_{L,k}\otimes I_M)\boldsymbol{w}_{e,i} \tag{95}$$

$$\boldsymbol{w}_{c,i} = \boldsymbol{w}_{c,i-1} - \mu_{\max}\cdot(p^T\otimes I_M)\hat{\boldsymbol{s}}_i\left(\boldsymbol{\phi}_{i-1}\right) \tag{96}$$

$$\boldsymbol{w}_{e,i} = \mathcal{D}_{N-1}\boldsymbol{w}_{e,i-1} - \mathcal{U}_R\mathcal{A}_2^T\mathcal{M}\hat{\boldsymbol{s}}_i\left(\boldsymbol{\phi}_{i-1}\right) \tag{97}$$

$$\hat{\boldsymbol{s}}_i\left(\boldsymbol{\phi}_{i-1}\right) = s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1}) + \boldsymbol{v}_i + \boldsymbol{z}_{i-1} \tag{98}$$

■

C. Error Dynamics

To simplify the notation, we introduce the centralized operator $T_c:\mathbb{R}^M\to\mathbb{R}^M$ as the following mapping for any $x\in\mathbb{R}^M$:

$$\begin{aligned}T_c(x) &\triangleq x - \mu_{\max}\cdot(p^T\otimes I_M)\,s(\mathbb{1}\otimes x)\\ &= x - \mu_{\max}\sum_{k=1}^{N}p_k s_k(x)\end{aligned} \tag{99}$$

Substituting (94) into (96)–(97) and using (99), we find that we can rewrite (96) and (97) in the alternative form:

$$\boldsymbol{w}_{c,i} = T_c(\boldsymbol{w}_{c,i-1}) - \mu_{\max}\cdot(p^T\otimes I_M)\left[\boldsymbol{z}_{i-1}+\boldsymbol{v}_i\right] \tag{100}$$

$$\boldsymbol{w}_{e,i} = \mathcal{D}_{N-1}\boldsymbol{w}_{e,i-1} - \mathcal{U}_R\mathcal{A}_2^T\mathcal{M}\left[s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})+\boldsymbol{z}_{i-1}+\boldsymbol{v}_i\right] \tag{101}$$

Likewise, we can write the reference recursion (47) in the following compact form:

$$\bar{w}_{c,i} = T_c(\bar{w}_{c,i-1}) \tag{102}$$

Comparing (100) with (102), we notice that the recursion for the centroid vector, $\boldsymbol{w}_{c,i}$, follows the same update rule as the reference recursion except for the two driving perturbation terms $\boldsymbol{z}_{i-1}$ and $\boldsymbol{v}_i$. Therefore, we would expect the trajectory of $\boldsymbol{w}_{c,i}$ to be a perturbed version of that of $\bar{w}_{c,i}$. Recall from (73) that

$$\check{\boldsymbol{w}}_{c,i} \triangleq \boldsymbol{w}_{c,i} - \bar{w}_{c,i}$$

To obtain the dynamics of $\check{\boldsymbol{w}}_{c,i}$, we subtract (102) from (100).

Lemma 3 (Error dynamics): The error quantities that appear on the right-hand side of (77) evolve according to the following dynamics:

$$\begin{aligned}\check{\boldsymbol{w}}_{c,i} &= T_c(\boldsymbol{w}_{c,i-1}) - T_c(\bar{w}_{c,i-1})\\ &\quad - \mu_{\max}\cdot(p^T\otimes I_M)\left[\boldsymbol{z}_{i-1}+\boldsymbol{v}_i\right]\end{aligned} \tag{103}$$

$$\begin{aligned}\boldsymbol{w}_{e,i} &= \mathcal{D}_{N-1}\boldsymbol{w}_{e,i-1}\\ &\quad - \mathcal{U}_R\mathcal{A}_2^T\mathcal{M}\left[s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})+\boldsymbol{z}_{i-1}+\boldsymbol{v}_i\right]\end{aligned} \tag{104}$$

■

The analysis in sequel will study the dynamics of the variances of the error quantities $\check{\boldsymbol{w}}_{c,i}$ and $\boldsymbol{w}_{e,i}$ based on (103)–(104). The main challenge is that these two recursions are coupled with each other through $\boldsymbol{z}_{i-1}$ and $\boldsymbol{v}_i$. To address the difficulty, we will extend the energy operator approach developed in [20] to the general scenario under consideration.

D. Energy Operators

To carry out the analysis, we need to introduce the following operators.

Definition 4 (Energy vector operator): Suppose $x = \mathrm{col}\{x_1, \ldots, x_N\}$ is an arbitrary $N \times 1$ block column vector that is formed by stacking $M_0 \times 1$ vectors $x_1, \ldots, x_N$ on top of each other. The energy vector operator $P_{M_0} : \mathbb{C}^{M_0 N} \to \mathbb{R}^N$ is defined as the mapping:

$$P_{M_0}[x] \triangleq \mathrm{col}\{\|x_1\|^2, \ldots, \|x_N\|^2\} \tag{105}$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector. ■

Definition 5 (Norm matrix operator): Suppose X is an arbitrary $K \times N$ block matrix consisting of blocks $\{X_{kn}\}$ of size $M_0 \times M_0$:

$$X = \begin{bmatrix} X_{11} & \cdots & X_{1N} \\ \vdots & & \vdots \\ X_{K1} & \cdots & X_{KN} \end{bmatrix} \tag{106}$$

The norm matrix operator $\bar{P}_{M_0} : \mathbb{C}^{M_0 K \times M_0 N} \to \mathbb{R}^{K \times N}$ is defined as the mapping:

$$\bar{P}_{M_0}[x] \triangleq \begin{bmatrix} \|X_{11}\| & \cdots & \|X_{1N}\| \\ \vdots & & \vdots \\ \|X_{K1}\| & \cdots & \|X_{KN}\| \end{bmatrix} \tag{107}$$

where $\|\cdot\|$ denotes the $2-$induced norm of a matrix. ■

By default, we choose M_0 to be M, the size of the vector $\boldsymbol{w}_{k,i}$. In this case, we will drop the subscript in$P_{M_0}[\cdot]$ and use $P[\cdot]$ for convenience. However, in other cases, we will keep the subscript to avoid confusion. Likewise, $\bar{P}_{M_0}[\cdot]$ characterizes the norms of different parts of a matrix it operates on. We will also drop the subscript if $M_0 = M$. In Appendix A, we collect several properties of the above energy operators, and which will be used in the sequel to characterize how the energy of the error quantities propagates through the dynamics (103)–(104).

VI. Transient Analysis

Using the energy operators and the various properties, we can now examine the transient behavior of the learning curve more closely. Recall from (74) that $\tilde{\boldsymbol{w}}_{k,i}$ consists of three parts: the error of the reference recursion, $\tilde{w}_{c,i}$, the difference between the centroid and the reference, $\check{\boldsymbol{w}}_{c,i}$, and the position of individual iterates relative to the centroid, $(u_{L,k} \otimes I_M)\boldsymbol{w}_{e,i}$. The main objective in the sequel is to study the convergence of the reference error, $\tilde{w}_{c,i}$, and establish non-asymptotic bounds for the mean-square values of $\check{\boldsymbol{w}}_{c,i}$ and $\boldsymbol{w}_{e,i}$, which will allow us to understand how fast and how close the iterates at the individual agents, $\{\boldsymbol{w}_{k,i}\}$, get to the reference recursion. Recalling from (77) that $\check{\boldsymbol{w}}_{c,i}$ and $\boldsymbol{w}_{e,i}$ are the two blocks of the transformed vector $\check{\boldsymbol{w}}'_i$ defined by (60), we can examine instead the evolution of

$$\begin{aligned} \check{\mathcal{W}}'_i &\triangleq \mathbb{E}P[\check{\boldsymbol{w}}'_i] = \mathrm{col}\left\{\mathbb{E}P[\check{\boldsymbol{w}}_{c,i}], \mathbb{E}P[\boldsymbol{w}_{e,i}]\right\} \\ &= \mathrm{col}\left\{\mathbb{E}\|\check{\boldsymbol{w}}_{c,i}\|^2, \mathbb{E}P[\boldsymbol{w}_{e,i}]\right\} \end{aligned} \tag{108}$$

Specifically, we will study the convergence of $\tilde{w}_{c,i}$ in Sec. VI-A, the stability of $\check{\mathcal{W}}'_i$ in Sec. VI-B, and the two transient phases of $\tilde{\boldsymbol{w}}_{k,i}$ in Sec. VI-C.

A. Limit Point

Before we proceed to study $\check{\mathcal{W}}'_i$, we state the following theorems on the existence of a limit point and on the convergence of the reference recursion (102).

Theorem 1 (Limit point): Given Assumptions 3–4, there exists a unique $M \times 1$ vector w^o that solves

$$\sum_{k=1}^{N} p_k s_k(w^o) = 0 \tag{109}$$

where p_k is the kth entry of the vector p defined in (16).

Proof: See Appendix E. ■

Theorem 2 (Convergence of the reference recursion): Let $\tilde{w}_{c,i} \triangleq w^o - \bar{w}_{c,i}$ denote the error vector of the reference recursion (102). Then, the following non-asymptotic bound on the squared error holds for all $i \geq 0$:

$$(1-2\mu_{\max}\|p\|_1\lambda_U)^i \cdot \|\tilde{w}_{c,0}\|^2 \leq \|\tilde{w}_{c,i}\|^2 \leq \gamma_c^{2i} \cdot \|\tilde{w}_{c,0}\|^2 \tag{110}$$

where

$$\gamma_c \triangleq 1 - \mu_{\max}\lambda_L + \frac{1}{2}\mu_{\max}^2\|p\|_1^2\lambda_U^2 \tag{111}$$

Furthermore, if the following condition on the step-size holds

$$0 < \mu_{\max} < \frac{2\lambda_L}{\|p\|_1^2\lambda_U^2} \tag{112}$$

then, the iterate $\tilde{w}_{c,i}$ converges to zero.

Proof: See Appendix F. ■

Note from (110) that, when the step-size is sufficiently small, the reference recursion (47) converges at a geometric rate between $1 - 2\mu_{\max}\|p\|_1\lambda_U$ and $\gamma_c^2 = 1 - 2\mu_{\max}\lambda_L + o(\mu_{\max})$. Note that this is a *non-asymptotic* result. That is, the convergence rate r_{Ref} of the reference recursion (102) is always lower and upper bounded by these two rates:

$$1 - 2\mu_{\max}\|p\|_1\lambda_U \leq r_{\mathrm{Ref}} \leq \gamma_c^2, \quad \forall i \geq 0 \tag{113}$$

We can obtain a more precise characterization of the convergence rate of the reference recursion in the asymptotic regime (for large enough i), as follows.

Theorem 3 (Convergence rate of the reference recursion): Specifically, for any small $\epsilon > 0$, there exists a time instant i_0 such that, for $i \geq i_0$, the error vector $\tilde{w}_{c,i}$ converges to zero at the following rate:

$$r_{\mathrm{Ref}} = \left[\rho(I_M - \mu_{\max}H_c)\right]^2 + O\big((\mu_{\max}\epsilon)^{\frac{1}{2(M-1)}}\big) \tag{114}$$

Proof: See Appendix G. ■

Note that since (114) holds for arbitrary $\epsilon > 0$, we can choose ϵ to be an arbitrarily small positive number. Therefore, the convergence rate of the reference recursion is arbitrarily close to $[\rho(I_M - \mu_{\max} H_c)]^2$.

B. Mean-Square Stability

Now we apply the properties from Lemmas 5–6 to derive an inequality recursion for the transformed energy vector $\check{\mathcal{W}}_i' = \mathbb{E}P[\check{\boldsymbol{w}}_i']$. The results are summarized in the following lemma.

Lemma 4 (Inequality recursion for $\check{\mathcal{W}}_i'$): The $N \times 1$ vector $\check{\mathcal{W}}_i'$ defined by (108) satisfies the following relation for all time instants:

$$\check{\mathcal{W}}_i' \preceq \Gamma \check{\mathcal{W}}_{i-1}' + \mu_{\max}^2 b_v \tag{115}$$

where

$$\Gamma \triangleq \Gamma_0 + \mu_{\max}^2 \psi_0 \cdot \mathbb{1}\mathbb{1}^T \in \mathbb{R}^{N\times N} \tag{116}$$

$$\Gamma_0 \triangleq \begin{bmatrix} \gamma_c & \mu_{\max} h_c(\mu_{\max}) \cdot \mathbb{1}^T \\ 0 & \Gamma_e \end{bmatrix} \in \mathbb{R}^{N\times N} \tag{117}$$

$$b_v \triangleq \mathrm{col}\{b_{v,c},\ b_{v,e}\cdot \mathbb{1}\} \in \mathbb{R}^N \tag{118}$$

$$\Gamma_e \triangleq \begin{bmatrix} |\lambda_2(A)| & \frac{2}{1-|\lambda_2(A)|} & & \\ & \ddots & \ddots & \\ & & \ddots & \frac{2}{1-|\lambda_2(A)|} \\ & & & |\lambda_2(A)| \end{bmatrix} \in \mathbb{R}^{(N-1)\times(N-1)} \tag{119}$$

The scalars ψ_0, $h_c(\mu)$, $b_{v,c}$ and $b_{v,e}$ are defined as

$$\begin{aligned} \psi_0 \triangleq \max\Bigg\{ & 4\alpha\|p\|_1^2,\ \ 4\alpha\|p\|_1^2 \cdot \big\|\bar{P}[\mathcal{A}_1^T \mathcal{U}_L]\big\|_\infty^2, \\ & 4N \cdot \big\|\bar{P}[\mathcal{U}_R \mathcal{A}_2^T]\big\|_\infty^2 \lambda_U^2 \left(\frac{3}{1-|\lambda_2(A)|} + \frac{\alpha}{\lambda_U^2}\right), \\ & 4N \cdot \big\|\bar{P}[\mathcal{U}_R \mathcal{A}_2^T]\big\|_\infty^2 \cdot \big\|\bar{P}[\mathcal{A}_1^T \mathcal{U}_L]\big\|_\infty^2 \lambda_U^2 \\ & \cdot \left(\frac{1}{1-|\lambda_2(A)|} + \frac{\alpha}{\lambda_U^2}\right) \Bigg\} \end{aligned} \tag{120}$$

$$h_c(\mu_{\max}) \triangleq \|p\|_1^2 \cdot \big\|\bar{P}[\mathcal{A}_1^T \mathcal{U}_L]\big\|_\infty^2 \lambda_U^2 \cdot \Big[\frac{1}{\lambda_L - \frac{1}{2}\mu_{\max}\|p\|_1^2 \lambda_U^2}\Big] \tag{121}$$

$$b_{v,c} \triangleq \|p\|_1^2 \cdot \big[4\alpha(\|\tilde{w}_{c,0}\|^2 + \|w^o\|^2) + \sigma_v^2\big] \tag{122}$$

$$\begin{aligned} b_{v,e} \triangleq N \big\|\bar{P}[\mathcal{U}_R \mathcal{A}_2^T]\big\|_\infty^2 \Bigg(& 12 \frac{\lambda_U^2 \|\tilde{w}_{c,0}\|^2 + \|g^o\|_\infty}{1-|\lambda_2(A)|} \\ & + 4\alpha(\|\tilde{w}_{c,0}\|^2 + \|w^o\|^2) + \sigma_v^2 \Bigg) \end{aligned} \tag{123}$$

where $g^o \triangleq P[s(\mathbb{1} \otimes w^o)]$.

Proof: See Appendix H. ∎

From (116)–(117), we see that as the step-size $\mu_{\max}$ becomes small, we have $\Gamma \approx \Gamma_0$, since the second term in the expression for Γ depends on the square of the step-size. Moreover, note that Γ_0 is an upper triangular matrix. Therefore, $\check{\boldsymbol{w}}_{c,i}$ and $\boldsymbol{w}_{e,i}$ are weakly coupled for small step-sizes; $\mathbb{E}P[\boldsymbol{w}_{e,i}]$ evolves on its own, but it will seep into the evolution of $\mathbb{E}P[\check{\boldsymbol{w}}_{c,i}]$ via the off-diagonal term in Γ_0, which is $O(\mu_{\max})$. This insight is exploited to establish a non-asymptotic bound on $\check{\mathcal{W}}_i' = \mathrm{col}\{\mathbb{E}\|\check{\boldsymbol{w}}_{c,i}\|^2, \mathbb{E}P[\boldsymbol{w}_{e,i}]\}$ in the following theorem.

Theorem 4 (Non-asymptotic bound for $\check{\mathcal{W}}_i'$): Suppose the matrix Γ defined in (116) is stable, i.e., $\rho(\Gamma) < 1$. Then, the following non-asymptotic bound holds for all $i \geq 0$:

$$\begin{aligned} \mathbb{E}P[\check{\boldsymbol{w}}_{c,i}] \preceq\ & \mu_{\max} h_c(\mu_{\max}) \cdot \mathbb{1}^T (\gamma_c I - \Gamma_e)^{-1} \left(\gamma_c^i I - \Gamma_e^i\right) \mathcal{W}_{e,0} \\ & + \check{\mathcal{W}}_{c,\infty}^{\mathrm{ub}'} \end{aligned} \tag{124}$$

$$\mathbb{E}P[\boldsymbol{w}_{e,i}] \preceq \Gamma_e^i \mathcal{W}_{e,0} + \check{\mathcal{W}}_{e,\infty}^{\mathrm{ub}'} \tag{125}$$

where $\mathcal{W}_{e,0} \triangleq \mathbb{E}P[\boldsymbol{w}_{e,0}]$, $\check{\mathcal{W}}_{c,\infty}^{\mathrm{ub}'}$ and $\check{\mathcal{W}}_{e,\infty}^{\mathrm{ub}'}$ are the $\limsup$ bounds of $\mathbb{E}P[\check{\boldsymbol{w}}_{c,i}]$ and $\mathbb{E}P[\boldsymbol{w}_{e,i}]$, respectively:

$$\begin{aligned} \check{\mathcal{W}}_{c,\infty}^{\mathrm{ub}'} = & \mu_{\max} \cdot \frac{\psi_0\big(\lambda_L + h_c(0)\big)\mathbb{1}^T (I - \Gamma_e)^{-1} \mathcal{W}_{e,0} + b_{v,c}\lambda_L}{\lambda_L^2} \\ & + o(\mu_{\max}) \end{aligned} \tag{126}$$

$$\begin{aligned} \check{\mathcal{W}}_{e,\infty}^{\mathrm{ub}'} = & \mu_{\max}^2 \cdot \frac{\psi_0\big(\lambda_L + h_c(0)\big)\mathbb{1}^T (I - \Gamma_e)^{-1} \mathcal{W}_{e,0} + b_{v,e}\lambda_L}{\lambda_L} \\ & \times (I - \Gamma_e)^{-1}\mathbb{1} + o(\mu_{\max}^2) \end{aligned} \tag{127}$$

where $o(\cdot)$ denotes strictly higher order terms, and $h_c(0)$ is the value of $h_c(\mu_{\max})$ (see (121)) evaluated at $\mu_{\max} = 0$. An important implication of (124) and (126) is that

$$\mathbb{E}P[\check{\boldsymbol{w}}_{c,i}] \leq O(\mu_{\max}), \qquad \forall i \geq 0 \tag{128}$$

Furthermore, a sufficient condition that guarantees the stability of the matrix Γ is that

$$\begin{aligned} 0 < \mu_{\max} < \min\Bigg\{ & \frac{\lambda_L}{\frac{1}{2}\|p\|_1^2 \lambda_U^2 + \frac{1}{3}\psi_0 \left(\frac{1-|\lambda_2(A)|}{2}\right)^{-2N}}, \\ & \sqrt{\frac{3(1-|\lambda_2(A)|)^{2N+1}}{2^{2N+2}\psi_0}}, \\ & \frac{\lambda_L}{\|p\|_1^2 \lambda_U^2 \big(\|\bar{P}_1[A^T U_L]\|_\infty^2 + \frac{1}{2}\big)} \Bigg\} \end{aligned} \tag{129}$$

Proof: See Appendix J. ∎

Corollary 1 (Asymptotic bounds): It holds that

$$\limsup_{i\to\infty} \mathbb{E}\|\check{\boldsymbol{w}}_{c,i}\|^2 \leq O(\mu_{\max}) \tag{130}$$

$$\limsup_{i\to\infty} \mathbb{E}\|\boldsymbol{w}_{e,i}\|^2 \leq O(\mu_{\max}^2) \tag{131}$$

Proof: The bound (130) holds since $\mathbb{E}\|\check{\boldsymbol{w}}_{c,i}\|^2 = \mathbb{E}P[\check{\boldsymbol{w}}_{c,i}] \leq O(\mu_{\max})$ for all $i \geq 0$ according to (128). Furthermore, inequality (131) holds because

$$\begin{aligned} \limsup_{i\to\infty} \mathbb{E}\|\boldsymbol{w}_{e,i}\|^2 & \overset{(a)}{=} \limsup_{i\to\infty} \mathbb{1}^T \mathbb{E}P[\boldsymbol{w}_{e,i}] \\ & \overset{(b)}{\preceq} \mathbb{1}^T \check{\mathcal{W}}_{e,\infty}^{\mathrm{ub}'} \\ & \overset{(c)}{=} O(\mu_{\max}^2) \end{aligned} \tag{132}$$

where step (a) uses property (157) of the energy operator $P[\cdot]$, step (b) uses (125), and step (c) uses (127). ∎

Finally, we present following main theorem that characterizes the difference between the learning curve of $\tilde{\boldsymbol{w}}_{k,i}$ at each agent k and that of $\tilde{w}_{c,i}$ generated by the reference recursion (102).

Theorem 5 (Learning behavior of $\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2$): Suppose the stability condition (129) holds. Then, the difference between the learning curve of the mean-square-error $\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2$ at each agent k and the learning curve of $\|\tilde{w}_{c,i}\|^2$ is bounded *non-asymptotically* as

$$\begin{aligned}\big|\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 - \|\tilde{w}_{c,i}\|^2\big| &\le 2\|u_{L,k}\otimes I_M\|^2\cdot \mathbb{1}^T\Gamma_e^i\mathcal{W}_{e,0} \\ &\quad + 2\|\tilde{w}_{c,0}\|\cdot\|u_{L,k}\otimes I_M\|\cdot\sqrt{\mathbb{1}^T\Gamma_e^i\mathcal{W}_{e,0}} \\ &\quad + \gamma_c^i\cdot O(\mu_{\max}^{\frac{1}{2}}) + O(\mu_{\max}) \qquad \text{for all } i\ge 0 \end{aligned} \tag{133}$$

where γ_c was defined earlier in (111).

Proof: See Appendix L. ∎

C. Interpretation of Results

The result established in Theorem 5 is significant because it allows us to examine the learning behavior of $\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2$. First, note that the bound (133) established in Theorem 5 is non-asymptotic; that is, it holds for any $i\ge 0$. Let $e_1(i)$, $e_2(i)$, $e_3(i)$ and e_4 denote the four terms in (133):

$$e_1(i) \triangleq 2\|u_{L,k}\otimes I_M\|^2\cdot\mathbb{1}^T\Gamma_e^i\mathcal{W}_{e,0} \tag{134}$$

$$e_2(i) \triangleq 2\|\tilde{w}_{c,0}\|\cdot\|u_{L,k}\otimes I_M\|\cdot\sqrt{\mathbb{1}^T\Gamma_e^i\mathcal{W}_{e,0}} \tag{135}$$

$$e_3(i) \triangleq \gamma_c^i\cdot O(\mu_{\max}^{\frac{1}{2}}) \tag{136}$$

$$e_4 \triangleq O(\mu_{\max}) \tag{137}$$

Then, inequality (133) can be rewritten as

$$\begin{aligned}&\|\tilde{w}_{c,i}\|^2 - [e_1(i)+e_2(i)+e_3(i)+e_4] \\ &\quad\le \mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 \le \|\tilde{w}_{c,i}\|^2 + [e_1(i)+e_2(i)+e_3(i)+e_4]\end{aligned} \tag{138}$$

for all $i\ge 0$. That is, the learning curve of $\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2$ is a perturbed version of the learning curve of reference recursion $\|\tilde{w}_{c,i}\|^2$, where the perturbation consists of four parts: $e_1(i)$, $e_2(i)$, $e_3(i)$ and e_4. We now examine the non-asymptotic convergence rates of these four perturbation terms relative to that of the reference term $\|\tilde{w}_{c,i}\|^2$ to show how the learning behavior of $\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2$ can be approximated by $\|\tilde{w}_{c,i}\|^2$. From their definitions in (134)–(135), we note that $e_1(i)$ and $e_2(i)$ converge to zero at the *non-asymptotic* rates of $\rho(\Gamma_e)$ and $[\rho(\Gamma_e)]^{\frac{1}{2}}$ over $i\ge 0$. According to (119), we have $\rho(\Gamma_e)=|\lambda_2(A)|$, the magnitude of the second largest eigenvalue of A. Let r_{e_1} and r_{e_2} denote the convergence rates of $e_1(i)$ and $e_2(i)$. We have

$$r_{e_1} = |\lambda_2(A)|,\quad r_{e_2} = |\lambda_2(A)|^{\frac{1}{2}} \tag{139}$$

By Assumption 1 and (13), the value of $|\lambda_2(A)|$ is strictly less than one, and is determined by the network connectivity and the design of the combination matrix; it is also independent of the step-size parameter $\mu_{\max}$. For this reason, the terms $e_1(i)$ and $e_2(i)$ converge to zero at rates that are determined by the network and are independent of $\mu_{\max}$. Furthermore, the term $e_3(i)$ is always small, i.e., $O(\mu_{\max}^{\frac{1}{2}})$, for all $i\ge 0$, and converges to zero as $i\to\infty$. The last term e_4 is also small, namely, $O(\mu_{\max})$. On the other hand, as revealed by Theorem 2 and (113), the non-asymptotic convergence rate of the term $\|\tilde{w}_{c,i}\|^2$ in (138) is bounded by

$$1-2\mu_{\max}\|p\|_1\lambda_U \le r_{\rm Ref} \le \gamma_c^2 = 1-2\mu_{\max}\lambda_L + o(\mu_{\max}) \tag{140}$$

Therefore, as long as $\mu_{\max}$ is small enough so that[2]

$$|\lambda_2(A)| < |\lambda_2(A)|^{\frac{1}{2}} < 1-2\mu_{\max}\|p\|_1\lambda_U \tag{141}$$

which is equivalent to requiring

$$\mu_{\max} < \frac{1-|\lambda_2(A)|^{\frac{1}{2}}}{2\|p\|_1\lambda_U} \tag{142}$$

we have the following relation regarding the non-asymptotic rates of $e_1(i)$, $e_2(i)$ and $\|\tilde{w}_{c,i}\|^2$:

$$r_{e_1} < r_{e_2} < r_{\rm Ref} \tag{143}$$

This means that, for sufficiently small $\mu_{\max}$ satisfying (142), $e_1(i)$ and $e_2(i)$ converge faster than $\|\tilde{w}_{c,i}\|^2$. For this reason, the perturbation terms $e_1(i)$ and $e_2(i)$ in (138) die out earlier than $\|\tilde{w}_{c,i}\|^2$. When they are negligible compared to $\|\tilde{w}_{c,i}\|^2$, we reach the end of Transient Phase I. Then, in Transient Phase II, we have

$$\|\tilde{w}_{c,i}\|^2 - [e_3(i)+e_4] \le \mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 \le \|\tilde{w}_{c,i}\|^2 + [e_3(i)+e_4] \tag{144}$$

By (136) and (137), the above inequality (144) is equivalent to the following relation:

$$\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 = \|\tilde{w}_{c,i}\|^2 + O(\mu_{\max}^{\frac{1}{2}})\cdot\gamma_c^i + O(\mu_{\max}) \tag{145}$$

This means that $\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2$ is close to $\|\tilde{w}_{c,i}\|^2$ in Transient Phase II, and the convergence rate of $\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2$ is the same as that of $\|\tilde{w}_{c,i}\|^2$. Furthermore, in the later stage of Transient Phase II, the convergence rate of $\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2$ would be close to the asymptotic rate of $\|\tilde{w}_{c,i}\|^2$ given by (114). Afterwards, as $i\to\infty$, both $\|\tilde{w}_{c,i}\|^2$ and $O(\mu_{\max}^{\frac{1}{2}})\cdot\gamma_c^i$ converge to zero and the only term remaining will be $e_4 = O(\mu_{\max})$, which contributes to the steady-state MSE. More specifically, taking the $\limsup$ of both sides of (133) leads to

$$\begin{aligned}\limsup_{i\to\infty}\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 &= \limsup_{i\to\infty}\big|\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 - \|\tilde{w}_{c,i}\|^2\big| \\ &\le O(\mu_{\max})\end{aligned} \tag{146}$$

We will go a step further and evaluate this steady-state MSE in closed-form for small step-sizes in Part II [44]. Therefore, $\boldsymbol{w}_{k,i}$ converges to w^o with a small steady-state MSE that is on the order of $O(\mu_{\max})$. And the steady-state MSE can be made arbitrarily small for small step-sizes.

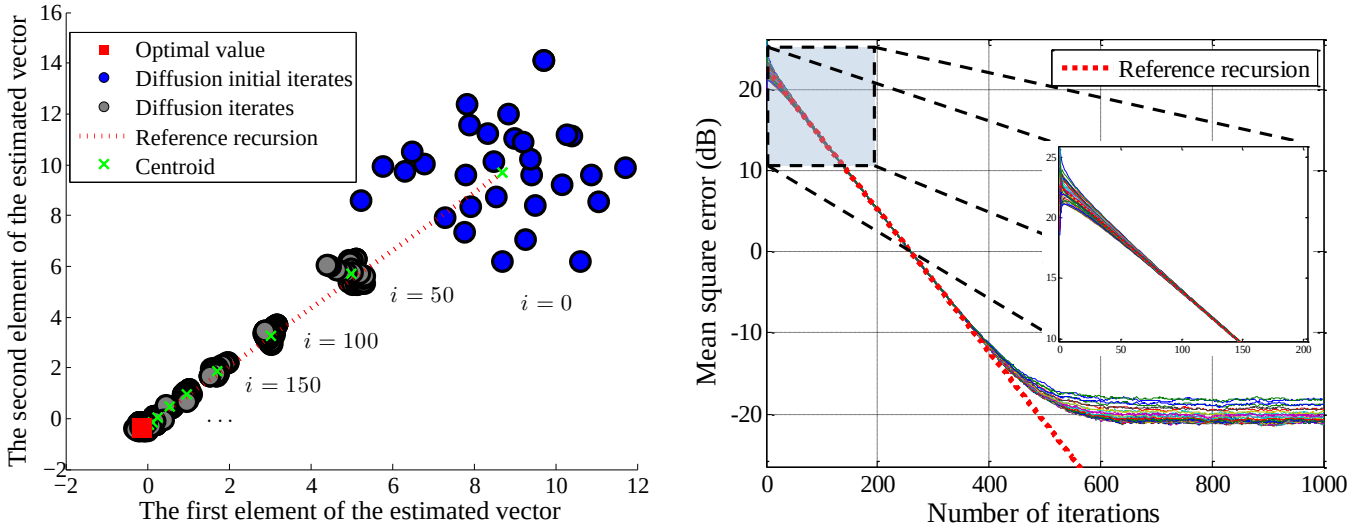
In view of this discussion, we now recognize that the results established in Theorems 1–4 reveal the evolution of the three components, $\tilde{w}_{c,i}$, $\tilde{\boldsymbol{w}}_{c,i}$ and $\boldsymbol{w}_{e,i}$ in (74) during three distinct phases of the learning process. From (124), the centroid $\boldsymbol{w}_{c,i}$ of the distributed algorithm (1)–(3) stays close to $\bar{w}_{c,i}$ over the entire time for sufficiently small step-sizes since the mean-square error $\mathbb{E}\|\boldsymbol{w}_{c,i}-\bar{w}_{c,i}\|^2 = \mathbb{E}P[\tilde{\boldsymbol{w}}_{c,i}]$ is always of the order of $O(\mu_{\max})$. However, $\mathcal{W}_{e,0} = \mathbb{E}P[\boldsymbol{w}_{e,i}]$ in (125) is

[2] $|\lambda_2(A)| < |\lambda_2(A)|^{\frac{1}{2}}$ always holds since $|\lambda_2(A)|$ is strictly less than one.

TABLE III
BEHAVIOR OF ERROR QUANTITIES IN DIFFERENT PHASES.

Error quantity	Transient Phase I		Transient Phase II		Steady-State [c]
	Convergence rate r [a]	Value	Convergence rate r [b]	Value	Value
$\|\tilde{w}_{c,i}\|^2$	$1-2\mu_{\max}\|p\|_1\lambda_U \le r \le \gamma_c^2$	$\gg O(\mu_{\max})$	$r=[\rho(I_M-\mu_{\max}H_c)]^2$	$\gg O(\mu_{\max})$	0
$\mathbb{E}\|\check{\boldsymbol{w}}_{c,i}\|^2$	converged	$O(\mu_{\max})$	converged	$O(\mu_{\max})$	$O(\mu_{\max})$
$\mathbb{E}P[\check{\boldsymbol{w}}_{e,i}]$	$r \le \lvert\lambda_2(A)\rvert$	$\gg O(\mu_{\max})$	converged	$O(\mu_{\max}^2)$	$O(\mu_{\max}^2)$
$\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2$	Multiple modes	$\gg O(\mu_{\max})$	$r=[\rho(I_M-\mu_{\max}H_c)]^2$	$\gg O(\mu_{\max})$	$O(\mu_{\max})$

[a] γ_c is defined in (111), and $\gamma_c^2 = 1-2\mu_{\max}\lambda_L + o(\mu_{\max})$.
[b] We only show the leading term of the convergence rate for r. More precise expression can be found in (114).
[c] Closer studies of the steady-state performance can be found in Part II [44].



(a) (b)

Fig. 4. The evolution and learning curves of various quantities in a diffusion LMS adaptive network (best viewed in color), where $M=2$, and the regressors are spatially and temporally white, and isotropic across agents. (a) The evolution of the iterates $\{\boldsymbol{w}_{k,i}\}$ at all agents, the centroid $\boldsymbol{w}_{c,i}$, and the reference recursion $\bar{w}_{c,i}$ on the two-dimensional solution space; the horizontal axis and vertical axis are the first and second elements of $\boldsymbol{w}_{k,i}$, respectively. The clusters of $\{\boldsymbol{w}_{k,i}\}$ are plotted every 50 iterations. (b) The MSE learning curves, averaged over 1000 trials, for the iterates $\{\boldsymbol{w}_{k,i}\}$ at all agents, and the reference recursion $\bar{w}_{c,i}$. The zoom-in region shows the learning curves for different agents, which quick shrink together in Phase I.

not necessarily small at the beginning. This is because, as we pointed out in (70) and Fig. 3, $\boldsymbol{w}_{e,i}$ characterizes the deviation of the agents from their centroid. If the agents are initialized at different values, then $\mathbb{E}P[\boldsymbol{w}_{e,0}] \neq 0$, and it takes some time for $\mathbb{E}P[\boldsymbol{w}_{e,i}]$ to decay to a small value of $O(\mu_{\max}^2)$. By (125), the rate at which $\mathbb{E}P[\boldsymbol{w}_{e,i}]$ decays is $\rho(\Gamma_e) = |\lambda_2(A)|$. On the other hand, recall from Theorems 2–3 that the error of the reference recursion, $\tilde{w}_{c,i}$ converges at a rate between $1-2\mu_{\max}\|p\|_1\lambda_U$ and $r_c^2 = 1-2\mu_{\max}\lambda_L + o(\lambda_{\max})$ at beginning and then $[\rho(I_M-\mu_{\max}H_c)]^2$ later on, which is slower than the convergence rate of $\mathbb{E}P[\boldsymbol{w}_{e,i}]$ for small step-size $\mu_{\max}$. Now, returning to relation (74):

$$\tilde{\boldsymbol{w}}_{k,i} = \tilde{w}_{c,i} - \check{\boldsymbol{w}}_{c,i} - (u_{L,k}\otimes I_M)\boldsymbol{w}_{e,i} \qquad (147)$$

this means that during the initial stage of adaptation, the third term in (147) decays to $O(\mu_{\max}^2)$ at a faster rate than the first term, although $\tilde{w}_{c,i}$ will eventually converge to zero. Recalling from (70) and Fig. 3 that $\boldsymbol{w}_{e,i}$ characterizes the deviation of the agents from their centroid, the decay of $\boldsymbol{w}_{e,i}$ implies that the agents are coordinating with each other so that their estimates $\boldsymbol{w}_{k,i}$ are close to the same $\boldsymbol{w}_{c,i}$ — we call this stage Transient Phase I. Moreover, as we just pointed out, the term $\mathbb{E}P[\check{\boldsymbol{w}}_{c,i}]$ is $O(\mu_{\max})$ over the entire time domain so that the second term in (147) is always small. This also means that the centroid of the cluster in Fig. 3, i.e., $\boldsymbol{w}_{c,i}$, is always close to the reference recursion $\bar{w}_{c,i}$ since $\check{\boldsymbol{w}}_{c,i} = \boldsymbol{w}_{c,i} - \bar{w}_{c,i}$ is always small. Now that $\mathbb{E}\|\check{\boldsymbol{w}}_{c,i}\|^2$ is $O(\mu_{\max})$ and $\mathbb{E}P[\boldsymbol{w}_{e,i}]$ is $O(\mu_{\max}^2)$, the error $\tilde{\boldsymbol{w}}_{k,i}$ at each agent k is mainly dominated by the first term, $\tilde{w}_{c,i}$, in (147), and the estimates $\{\boldsymbol{w}_{k,i}\}$ at different agents converge together at the same rate as the reference recursion, given by (114), to steady-state — we call this stage Transient Phase II. Furthermore, if $\mathcal{W}_{e,0}=0$, i.e., the iterates $\boldsymbol{w}_{k,i}$ are initialized at the same value (e.g., zero vector), then (125) shows that $\mathbb{E}P[\boldsymbol{w}_{e,i}]$ is $O(\mu_{\max}^2)$ over the entire time domain so that the learning dynamics start at Transient Phase II directly. Finally, all agents reach the third phase, steady-state, where $\tilde{w}_{c,i} \to 0$ and $\tilde{\boldsymbol{w}}_{k,i}$ is dominated by the second and third terms in (147) so that $\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2$ becomes $O(\mu_{\max})$. We summarize the above results in Table III and illustrate the evolution of the quantities in the simulated example in Fig. 4. We observe from Fig. 4 that the radius of the cluster shrinks quickly at the early stage of the transient phase, and then converges towards the optimal solution.

VII. CONCLUSION

In this work, we studied the learning behavior of adaptive networks under fairly general conditions. We showed that, in the constant and small step-size regime, a typical learning curve of each agent exhibits three phases: Transient Phase I, Transient Phase II, and Steady-state Phase. A key observation

is that, the second and third phases approach the performance of a centralized strategy. Furthermore, we showed that the right eigenvector of the combination matrix corresponding to the eigenvalue at one influences the limit point, the convergence rate, and the steady-state mean-square-error (MSE) performance of the distributed optimization and learning strategies in a critical way. Analytical expressions that illustrate these effects were derived. Various implications were discussed and illustrative examples were also considered.

APPENDIX A
PROPERTIES OF THE ENERGY OPERATORS

In this appendix, we state lemmas on properties of the operators $P_{M_0}[\cdot]$ and $\bar{P}_{M_0}[\cdot]$. We begin with some basic properties.

Lemma 5 (Basic properties): Consider $N\times 1$ block vectors $x=\mathrm{col}\{x_1,\ldots,x_N\}$ and $y=\mathrm{col}\{y_1,\ldots,y_N\}$ with $M\times 1$ entries $\{x_k,y_k\}$. Consider also the $K\times N$ block matrix X with blocks of size $M\times M$. Then, the operators $P[\cdot]$ and $\bar{P}[\cdot]$ satisfy the following properties:

1) (***Nonnegativity***)**:** $P[x]\succeq 0$, $\bar{P}[X]\succeq 0$.
2) (***Scaling***)**:** For any scalar $a\in\mathbb{C}$, we have $P[ax]=|a|^2P[x]$ and $\bar{P}[aX]=|a|\cdot\bar{P}[X]$.
3) (***Convexity***)**:** suppose $x^{(1)},\ldots,x^{(K)}$ are $N\times 1$ block vectors formed in the same manner as x, $X^{(1)},\ldots,X^{(K)}$ are $K\times N$ block matrices formed in the same manner as X, and let $a_1,\ldots,a_K$ be non-negative real scalars that add up to one. Then,

$$\begin{aligned}&P[a_1x^{(1)}+\cdots+a_Kx^{(K)}]\\&\qquad\preceq a_1P[x^{(1)}]+\cdots+a_KP[x^{(K)}]\end{aligned}\tag{148}$$

$$\begin{aligned}&\bar{P}[a_1X^{(1)}+\cdots+a_KX^{(K)}]\\&\qquad\preceq a_1\bar{P}[X^{(1)}]+\cdots+a_K\bar{P}[X^{(K)}]\end{aligned}\tag{149}$$

4) (***Additivity***)**:** Suppose $\boldsymbol{x}=\mathrm{col}\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$ and $\boldsymbol{y}=\mathrm{col}\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N\}$ are $N\times 1$ block random vectors that satisfy $\mathbb{E}\boldsymbol{x}_k^*\boldsymbol{y}_k=0$ for $k=1,\ldots,N$, where $*$ denotes complex conjugate transposition. Then,

$$\mathbb{E}P[\boldsymbol{x}+\boldsymbol{y}]=\mathbb{E}P[\boldsymbol{x}]+\mathbb{E}P[\boldsymbol{y}]\tag{150}$$

5) (***Triangular inequality***)**:** Suppose X and Y are two $K\times N$ block matrices of same block size M. Then,

$$\bar{P}[X+Y]\preceq\bar{P}[X]+\bar{P}[Y]\tag{151}$$

6) (***Submultiplicity***)**:** Suppose X and Z are $K\times N$ and $N\times L$ block matrices of the same block size M, respectively. Then,

$$\bar{P}[XZ]\preceq\bar{P}[X]\bar{P}[Z]\tag{152}$$

7) (***Kronecker structure***)**:** Suppose $X\in\mathbb{C}^{K\times N}$, $a\in\mathbb{C}^N$ and $b\in\mathbb{C}^M$. Then,

$$\bar{P}[X\otimes I_M]=\bar{P}_1[X]\tag{153}$$

$$P[a\otimes b]=\|b\|^2\cdot P_1[a]\tag{154}$$

where by definition, $\bar{P}_1[\cdot]$ and $P_1[\cdot]$ denote the operators that work on the scalar entries of their arguments. When X consists of nonnegative entries, relation (153) becomes

$$\bar{P}[X\otimes I_M]=X\tag{155}$$

8) (***Relation to norms***)**:** The $\infty-$norm of $P[x]$ is the squared block maximum norm of x:

$$\|P[x]\|_\infty=\|x\|_{b,\infty}^2\triangleq\big(\max_{1\le k\le N}\|x_k\|\big)^2\tag{156}$$

Moreover, the sum of the entries in $P[x]$ is the squared Euclidean norm of x:

$$\mathbb{1}^TP[x]=\|x\|^2=\sum_{k=1}^N\|x_k\|^2\tag{157}$$

9) (***Inequality preservation***)**:** Suppose vectors x, y and matrices F, G have nonnegative entries, then $x\preceq y$ implies $Fx\preceq Fy$, and $F\preceq G$ implies $Fx\preceq Gx$.
10) (***Upper bounds***)**:** It holds that

$$\bar{P}[X]\preceq\|\bar{P}[X]\|_1\cdot\mathbb{1}\mathbb{1}^T\tag{158}$$

$$\bar{P}[X]\preceq\|\bar{P}[X]\|_\infty\cdot\mathbb{1}\mathbb{1}^T\tag{159}$$

where $\|\cdot\|_\infty$ denotes the $\infty-$induced norm of a matrix (maximum absolute row sum).

Proof: See Appendix C. ∎

More importantly, the following variance relations hold for the energy and norm operators. These relations show how error variances propagate after a certain operator is applied to a random vector.

Lemma 6 (Variance relations): Consider $N\times 1$ block vectors $x=\mathrm{col}\{x_1,\ldots,x_N\}$ and $y=\mathrm{col}\{y_1,\ldots,y_N\}$ with $M\times 1$ entries $\{x_k,y_k\}$. The following variance relations are satisfied by the energy vector operator $P[\cdot]$:

1) (***Linear transformation***)**:** Given a $K\times N$ block matrix Q with the size of each block being $M\times M$, Qx defines a linear operator on x and its energy satisfies

$$P[Qx]\preceq\|\bar{P}[Q]\|_\infty\cdot\bar{P}[Q]\;P[x]\tag{160}$$

$$\preceq\|\bar{P}[Q]\|_\infty^2\cdot\mathbb{1}\mathbb{1}^T\cdot P[x]\tag{161}$$

As a special case, for a left-stochastic $N\times N$ matrix A, we have

$$P[(A^T\otimes I_M)x]\preceq A^TP[x]\tag{162}$$

2) (***Update operation***)**:** The global update vector defined by (85) satisfies the following variance relation:

$$P[s(x)-s(y)]\preceq\lambda_U^2P[x-y]\tag{163}$$

3) (***Centralized operation***)**:** The centralized operator $T_c(x)$ defined by (99) satisfies the following variance relations:

$$P[T_c(x)-T_c(y)]\preceq\gamma_c^2\cdot P[x-y]\tag{164}$$

$$P[T_c(x)-T_c(y)]\succeq(1-2\mu_{\max}\|p\|_1\lambda_U)\cdot P[x-y]\tag{165}$$

where

$$\gamma_c\triangleq 1-\mu_{\max}\lambda_L+\frac{1}{2}\mu_{\max}^2\|p\|_1^2\lambda_U^2\tag{166}$$

Moreover, it follows from (26) that

$$\begin{aligned}\gamma_c &\geq 1-\mu_{\max}\lambda_L+\frac{1}{2}\mu_{\max}^2\lambda_L^2\\ &=(1-\frac{1}{2}\mu_{\max}\lambda_L)^2+\frac{1}{4}\mu_{\max}^2\lambda_L^2>0\end{aligned}\tag{167}$$

4) (***Stable Jordan operation***)**:** Suppose D_L is an $L\times L$ Jordan matrix of the following block form:

$$D_L\triangleq \mathrm{diag}\{D_{L,2},\ldots,D_{L,n_0}\}\tag{168}$$

where the nth $L_n\times L_n$ Jordan block is defined as (note that $L=L_2+\cdots+L_{n_0}$)

$$D_{L,n}\triangleq\begin{bmatrix}d_n & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1\\ & & & d_n\end{bmatrix}\tag{169}$$

We further assume D_L to be stable with $0\leq|d_{n_0}|\leq\cdots\leq|d_2|<1$. Then, for any $L\times 1$ vectors x' and y', we have

$$P_1[D_Lx'+y']\preceq\Gamma_e\cdot P_1[x']+\frac{2}{1-|d_2|}\cdot P_1[y']\tag{170}$$

where Γ_e is the $L\times L$ matrix defined as

$$\Gamma_e\triangleq\begin{bmatrix}|d_2| & \frac{2}{1-|d_2|} & & \\ & \ddots & \ddots & \\ & & \ddots & \frac{2}{1-|d_2|}\\ & & & |d_2|\end{bmatrix}\tag{171}$$

5) (***Stable Kronecker Jordan operator***)**:** Suppose $\mathcal{D}_L=D_L\otimes I_M$, where D_L is the $L\times L$ Jordan matrix defined in (168)–(169). Then, for any $LM\times 1$ vectors x_e and y_e, we have

$$P[\mathcal{D}_Lx_e+y_e]\preceq\Gamma_e\cdot P[x_e]+\frac{2}{1-|d_2|}\cdot P[y_e]\tag{172}$$

Proof: See Appendix D. ∎

APPENDIX B
PROOF OF LEMMA 1

First, we establish that conditions (27) and (28) imply (24) and (25), respectively. Using the mean-value theorem [49, p.6], we have for any $x,y\in\mathcal{S}$:

$$\begin{aligned}\|s_k(x)-s_k(y)\|&=\left\|\int_0^1\nabla_{w^T}s_k(y+t(x-y))dt\cdot(x-y)\right\|\\ &\leq\int_0^1\|\nabla_{w^T}s_k(y+t(x-y))\|\,dt\cdot\|x-y\|\\ &\leq\lambda_U\cdot\|x-y\|\end{aligned}\tag{173}$$

where we used the fact that $y+t(x-y)=tx+(1-t)y\in\mathcal{S}$ given $x,y\in\mathcal{S}$ and $0\leq t\leq 1$. Likewise, we have

$$\begin{aligned}&(x-y)^T\cdot\sum_{k=1}^N p_k[s_k(x)-s_k(y)]\\ &=(x-y)^T\cdot\sum_{k=1}^N p_k\int_0^1\nabla_{w^T}s_k(y+t(x-y))dt\cdot(x-y)\\ &=(x-y)^T\cdot\int_0^1\sum_{k=1}^N p_k\nabla_{w^T}s_k(y+t(x-y))dt\cdot(x-y)\\ &\overset{(29)}{=}(x-y)^T\cdot H_c(y+t(x-y))\cdot(x-y)\\ &=(x-y)^T\cdot\frac{H_c(y+t(x-y))+H_c^T(y+t(x-y))}{2}\cdot(x-y)\\ &\geq\lambda_L\cdot\|x-y\|^2\end{aligned}\tag{174}$$

Next, we establish the reverse direction that conditions (24) and (25) imply (27) and (28). Choosing $x=w+t\cdot\delta w$ and $y=w$ in (24) for any $\delta w\neq 0$ and any small positive t, we get

$$\begin{aligned}&\|s_k(w+t\cdot\delta w)-s_k(w)\|\leq t\cdot\lambda_U\cdot\|\delta w\|\\ \Rightarrow\quad&\lim_{t\to 0^+}\left\|\frac{s_k(w+t\cdot\delta w)-s_k(w)}{t}\right\|\leq\lambda_U\cdot\|\delta w\|\\ \Rightarrow\quad&\left\|\lim_{t\to 0^+}\frac{s_k(w+t\cdot\delta w)-s_k(w)}{t}\right\|\leq\lambda_U\cdot\|\delta w\|\\ \Rightarrow\quad&\|\nabla_{w^T}s_k(w)\delta w\|\leq\lambda_U\cdot\|\delta w\|\\ \Rightarrow\quad&\|\nabla_{w^T}s_k(w)\|\triangleq\sup_{\delta w\neq 0}\frac{\|\nabla_{w^T}s_k(w)\delta w\|}{\|\delta w\|}\leq\lambda_U\end{aligned}\tag{175}$$

Likewise, choosing $x=w+t\cdot\delta w$ and $y=w$ in (25) for any $\delta w\neq 0$ and any small positive t, we obtain

$$\begin{aligned}&t\cdot\delta w^T\cdot\sum_{k=1}^N p_k[s_k(w+t\cdot\delta w)-s_k(w)]\geq t^2\cdot\lambda_L\cdot\|\delta w\|^2\\ \Rightarrow\quad&\delta w^T\cdot\sum_{k=1}^N p_k\left(\lim_{t\to 0^+}\frac{s_k(w+t\cdot\delta w)-s_k(w)}{t}\right)\\ &\qquad\geq\lambda_L\cdot\|\delta w\|^2\\ \Rightarrow\quad&\delta w^T\cdot\sum_{k=1}^N p_k\nabla_{w^T}s_k(w)\cdot\delta w\geq\lambda_L\cdot\|\delta w\|^2\\ \Rightarrow\quad&\delta w^T H_c(w)\delta w\geq\lambda_L\cdot\|\delta w\|^2\\ \Rightarrow\quad&\delta w^T\frac{H_c(w)+H_c^T(w)}{2}\delta w\geq\lambda_L\cdot\|\delta w\|^2\\ \Rightarrow\quad&\frac{H_c(w)+H_c^T(w)}{2}\geq\lambda_L\cdot I_M\end{aligned}\tag{176}$$

APPENDIX C
PROOF OF LEMMA 5

Properties 1-2 are straightforward from the definitions of $P[\cdot]$ and $\bar{P}[\cdot]$. Property 4 was proved in [20]. We establish the remaining properties.

(***Property 3: Convexity***) The convexity of $P[\cdot]$ has already been proven in [20]. We now establish the convexity of the operator $\bar{P}[\cdot]$. Let $X_{qn}^{(k)}$ denote the $(q,n)-$th $M\times M$ block of the matrix $X^{(k)}$, where $q=1,\ldots,K$ and $n=1,\ldots,N$. Then,

$$\bar{P}\left[\sum_{k=1}^K a_kX^{(k)}\right]$$

$$\preceq \begin{bmatrix} \sum_{k=1}^K a_k \left\| X_{11}^{(k)} \right\| & \cdots & \sum_{k=1}^K a_k \left\| X_{1N}^{(k)} \right\| \\ \vdots & & \vdots \\ \sum_{k=1}^K a_k \left\| X_{K1}^{(k)} \right\| & \cdots & \sum_{k=1}^K a_k \left\| X_{KN}^{(k)} \right\| \end{bmatrix}$$
$$= \sum_{k=1}^K a_k \bar{P}[X^{(k)}] \tag{177}$$

(***Property 5: Triangular inequality***) Let X_{qn} and Y_{qn} denote the $(q,n)-$th $M \times M$ blocks of the matrices X and Y, respectively, where $q = 1, \ldots, K$ and $n = 1, \ldots, N$. Then, by the triangular inequality of the matrix norm $\|\cdot\|$, we have

$$\bar{P}\left[X+Y\right] \preceq \begin{bmatrix} \|X_{11}\| + \|Y_{11}\| & \cdots & \|X_{1N}\| + \|Y_{1N}\| \\ \vdots & & \vdots \\ \|X_{K1}\| + \|Y_{K1}\| & \cdots & \|X_{KN}\| + \|Y_{KN}\| \end{bmatrix}$$
$$= \bar{P}[X] + \bar{P}[Y] \tag{178}$$

(***Property 6: Submultiplicity***) Let X_{kn} and Z_{nl} be the $(k,n)-$th and $(n,l)-$th $M \times M$ blocks of X and Z, respectively. Then, the $(k,l)-$th $M \times M$ block of the matrix product XZ, denoted by $[XZ]_{k,l}$, is

$$[XZ]_{k,l} = \sum_{n=1}^N X_{kn} Z_{nl} \tag{179}$$

Therefore, the $(k,l)-$th entry of the matrix $\bar{P}[XZ]$ can be bounded as

$$\left[\bar{P}[XZ]\right]_{k,l} = \left\| \sum_{n=1}^N X_{kn} Z_{nl} \right\| \leq \sum_{n=1}^N \|X_{kn}\| \cdot \|Z_{nl}\| \tag{180}$$

Note that $\|X_{kn}\|$ and $\|Z_{nl}\|$ are the $(k,n)-$th and $(n,l)-$th entries of the matrices $\bar{P}[X]$ and $\bar{P}[Z]$, respectively. The right-hand side of the above inequality is therefore the $(k,l)-$th entry of the matrix product $\bar{P}[X]\bar{P}[Z]$. Therefore, we obtain

$$\bar{P}[XZ] \preceq \bar{P}[X]\bar{P}[Z] \tag{181}$$

(***Property 7: Kronecker structure***) For (153), we note that the $(k,n)-$th $M \times M$ block of $X \otimes I_M$ is $x_{kn} I_M$. Therefore, by the definition of $\bar{P}[\cdot]$, we have

$$\bar{P}[X \otimes I_M] = \begin{bmatrix} |x_{11}| & \cdots & |x_{1N}| \\ \vdots & & \\ |x_{K1}| & \cdots & |x_{KN}| \end{bmatrix} = \bar{P}_1[X] \tag{182}$$

In the special case when X consists of nonnegative entries, $\bar{P}_1[X] = X$, and we recover (155). To prove (154), we let $a = \mathrm{col}\{a_1, \ldots, a_N\}$ and $b = \mathrm{col}\{b_1, \ldots, b_M\}$. Then, by the definitin of $P[\cdot]$, we have

$$P[a \otimes b] = \mathrm{col}\{|a_1|^2 \cdot \|b\|^2, \ldots, |a_N|^2 \cdot \|b\|^2\} = \|b\|^2 \cdot P_1[a] \tag{183}$$

(***Property 8: Relation to norms***) Relations (156) and (157) are straightforward and follow from the definition.

(***Property 9: Inequality preservation***) The proof that $x \preceq y$ implies $Fx \preceq Fy$ can be found in [20]. We now prove that $F \preceq G$ implies $Fx \preceq Gx$. This can be proved by showig that $(G-F)x \succeq 0$, which is true because all entries of $G-F$ and x are nonnegative due to $F \preceq G$ and $x \succeq 0$.

(***Property 10: Upper bounds***) By the definition of $\bar{P}[X]$ in (107), we get

$$\bar{P}[X] \preceq \left(\max_{l,k} \|X_{lk}\| \right) \cdot \mathbb{1}\mathbb{1}^T$$
$$\preceq \max_l \left(\sum_{k=1}^N \|X_{lk}\| \right) \cdot \mathbb{1}\mathbb{1}^T$$
$$= \|\bar{P}[X]\|_\infty \cdot \mathbb{1}\mathbb{1}^T \tag{184}$$

Likewise, we can establish that $\bar{P}[X] \preceq \|\bar{P}[X]\|_1 \cdot \mathbb{1}\mathbb{1}^T$.

APPENDIX D
PROOF OF LEMMA 6

(***Property 1: Linear transformation***) Let Q_{kn} be the (k,n)-th $M \times M$ block of Q. Then

$$P[Qx] = \mathrm{col}\left\{ \left\| \sum_{n=1}^N Q_{1n} x_n \right\|^2, \ldots, \left\| \sum_{n=1}^N Q_{Kn} x_n \right\|^2 \right\} \tag{185}$$

Using the convexity of $\|\cdot\|^2$, we have the following bound on each n-th entry:

$$\left\| \sum_{n=1}^N Q_{kn} x_n \right\|^2$$
$$\overset{(a)}{=} \left[\sum_{n=1}^N \|Q_{kn}\| \right]^2 \cdot \left\| \sum_{n=1}^N \frac{\|Q_{kn}\|}{\sum_{l=1}^N \|Q_{kl}\|} \cdot \frac{Q_{kn}}{\|Q_{kn}\|} x_n \right\|^2$$
$$\overset{(b)}{\leq} \left[\sum_{n=1}^N \|Q_{kn}\| \right]^2 \cdot \sum_{n=1}^N \frac{\|Q_{kn}\|}{\sum_{l=1}^N \|Q_{kl}\|} \cdot \frac{\|Q_{kn}\|^2}{\|Q_{kn}\|^2} \|x_n\|^2$$
$$= \left[\sum_{n=1}^N \|Q_{kn}\| \right] \cdot \sum_{n=1}^N \|Q_{kn}\| \cdot \|x_n\|^2$$
$$\leq \max_k \left[\sum_{n=1}^N \|Q_{kn}\| \right] \cdot \sum_{n=1}^N \|Q_{kn}\| \cdot \|x_n\|^2$$
$$= \|\bar{P}[Q]\|_\infty \cdot \sum_{n=1}^N \|Q_{kn}\| \cdot \|x_n\|^2 \tag{186}$$

where in step (b) we applied Jensen's inequlity to $\|\cdot\|^2$. Note that if some $\|Q_{kn}\|$ in step (a) is zero, we eliminate the corresponding term from the sum and it can be verified that the final result still holds. Substituting into (185), we establish (160). The special case (162) can be obtained by using $\bar{P}[A^T \otimes I_M] = A^T$ and that $\|A^T\|_\infty = 1$ (left-stochastic) in (160). Finally, the upper bound (161) can be proved by applying (159) to $\bar{P}[Q]$.

(***Property 2: Update operation***) By the definition of $P[\cdot]$ and the Lipschitz Assumption 3, we have

$$P[s(x) - s(y)]$$
$$= \mathrm{col}\{\|s_1(x_1) - s_1(y_1)\|^2, \ldots \|s_N(x_N) - s_N(y_N)\|^2\}$$
$$\preceq \mathrm{col}\{\lambda_U^2 \cdot \|x_1 - y_1\|^2, \ldots \lambda_U^2 \cdot \|x_N - y_N\|^2\}$$
$$= \lambda_U^2 \cdot P[x-y] \tag{187}$$

(***Property 3: Centralized opertion***) Since $T_c : \mathbb{R}^M \to \mathbb{R}^M$, the output of $P[T_c(x) - T_c(y)]$ becomes a scalar. From the definition, we get

$$\begin{aligned}
&P\left[T_c(x) - T_c(y)\right] \\
&\quad = \left\|x - y - \mu_{\max} \cdot (p^T \otimes I_M)\left[s(\mathbb{1} \otimes x) - s(\mathbb{1} \otimes y)\right]\right\|^2 \\
&\quad = \left\|x - y - \mu_{\max} \cdot \sum_{k=1}^{N} p_k \left[s_k(x) - s_k(y)\right]\right\|^2 \\
&\quad = \|x - y\|^2 - 2\mu_{\max} \cdot (x - y)^T \sum_{k=1}^{N} p_k \left[s_k(x) - s_k(y)\right] \\
&\qquad + \mu_{\max}^2 \left\|(p^T \otimes I_M)\left[s(\mathbb{1} \otimes x) - s(\mathbb{1} \otimes y)\right]\right\|^2 \\
&\quad = \|x - y\|^2 - 2\mu_{\max} \cdot (x - y)^T \sum_{k=1}^{N} p_k \left[s_k(x) - s_k(y)\right] \\
&\qquad + \mu_{\max}^2 P\left[(p^T \otimes I_M)\left[s(\mathbb{1} \otimes x) - s(\mathbb{1} \otimes y)\right]\right] \qquad (188)
\end{aligned}$$

We first prove the upper bound (164) as follows:

$$\begin{aligned}
&P\left[T_c(x) - T_c(y)\right] \\
&\quad = \left\|x - y - \mu_{\max} \cdot \sum_{k=1}^{N} p_k [s_k(x) - s_k(y)]\right\|^2 \\
&\quad = \|x - y\|^2 - 2\mu_{\max} \cdot (x - y)^T \sum_{k=1}^{N} p_k [s_k(x) - s_k(y)] \\
&\qquad + \mu_{\max}^2 \cdot \Big\| \sum_{k=1}^{N} p_k [s_k(x) - s_k(y)]\Big\|^2 \\
&\quad \overset{(25)}{\leq} \|x - y\|^2 - 2\mu_{\max} \cdot \lambda_L \cdot \|x - y\|^2 \\
&\qquad + \mu_{\max}^2 \cdot \Big\| \sum_{k=1}^{N} p_k [s_k(x) - s_k(y)]\Big\|^2 \\
&\quad \leq \|x - y\|^2 - 2\mu_{\max} \cdot \lambda_L \cdot \|x - y\|^2 \\
&\qquad + \mu_{\max}^2 \cdot \Big[\sum_{k=1}^{N} p_k \|s_k(x) - s_k(y)\|\Big]^2 \\
&\quad \overset{(24)}{\leq} \|x - y\|^2 - 2\mu_{\max} \cdot \lambda_L \cdot \|x - y\|^2 \\
&\qquad + \mu_{\max}^2 \cdot \left[\sum_{k=1}^{N} p_k \cdot \lambda_U \cdot \|x - y\|\right]^2 \\
&\quad = \|x - y\|^2 - 2\mu_{\max} \cdot \lambda_L \cdot \|x - y\|^2 \\
&\qquad + \mu_{\max}^2 \cdot \|p\|_1^2 \lambda_U^2 \cdot \|x - y\|^2 \\
&\quad = (1 - 2\mu_{\max}\lambda_L + \mu_{\max}^2 \lambda_U^2 \|p\|_1^2) \cdot \|x - y\|^2 \\
&\quad \leq \left(1 - \mu_{\max}\lambda_L + \frac{1}{2}\mu_{\max}^2 \lambda_U^2 \|p\|_1^2\right)^2 \cdot \|x - y\|^2 \qquad (189)
\end{aligned}$$

where in the last step we used the relation $(1-x) \leq (1-\frac{1}{2}x)^2$.

Next, we prove the lower bound (165). From (188), we notice that the last term in (188) is always nonnegative so that

$$\begin{aligned}
&P\left[T_c(x) - T_c(y)\right] \\
&\quad \succeq \|x - y\|^2 - 2\mu_{\max} \cdot (x - y)^T \sum_{k=1}^{N} p_k \left[s_k(x) - s_k(y)\right] \\
&\quad \overset{(a)}{\succeq} \|x - y\|^2 - 2\mu_{\max} \cdot \|x - y\| \cdot \Big\| \sum_{k=1}^{N} p_k \left[s_k(x) - s_k(y)\right] \Big\| \\
&\quad \succeq \|x - y\|^2 - 2\mu_{\max} \cdot \|x - y\| \cdot \sum_{k=1}^{N} p_k \left\|s_k(x) - s_k(y)\right\| \\
&\quad \overset{(b)}{\succeq} \|x - y\|^2 - 2\mu_{\max} \cdot \|x - y\| \cdot \sum_{k=1}^{N} p_k \lambda_U \|x - y\| \\
&\quad = (1 - 2\mu_{\max}\lambda_U \|p\|_1) \cdot \|x - y\|^2 \qquad (190)
\end{aligned}$$

where in step (a), we used the Cauchy-Schwartz inequality $x^T y \leq |x^T y| \leq \|x\| \cdot \|y\|$, and in step (b) we used (24).

(***Property 4: Stable Jordan operator***) First, we notice that matrix $D_{L,n}$ can be written as

$$D_{L,n} = d_n \cdot I_{L_n} + \Theta_{L_n} \qquad (191)$$

where Θ_{L_n} is an $L_n \times L_n$ strictly upper triangular matrix of the following form:

$$\Theta_{L_n} \triangleq \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} \qquad (192)$$

Define the following matrices:

$$\Lambda_L \triangleq \mathrm{diag}\{d_2 I_{L_2}, \ldots, d_{n_0} I_{L_{n_0}}\} \qquad (193)$$

$$\Theta'_L \triangleq \mathrm{diag}\{\Theta_{L_2}, \ldots, \Theta_{L_{n_0}}\} \qquad (194)$$

Then, the original Jordan matrix D_L can be expressed as

$$D_L = \Lambda_L + \Theta'_L \qquad (195)$$

so that

$$\begin{aligned}
&P_1[D_L x' + y'] \\
&\quad = P_1\left[\Lambda_L x' + \Theta'_L x' + y'\right] \\
&\quad = P_1\left[|d_2| \cdot \frac{1}{|d_2|}\Lambda_L x' + \frac{1 - |d_2|}{2} \cdot \frac{2}{1 - |d_2|}\Theta'_L x' \right. \\
&\qquad \left. + \frac{1 - |d_2|}{2} \cdot \frac{2}{1 - |d_2|} y'\right] \\
&\quad \overset{(a)}{\preceq} |d_2| \cdot P_1\left[\frac{1}{|d_2|}\Lambda_L x'\right] + \frac{1 - |d_2|}{2} \cdot P_1\left[\frac{2}{1 - |d_2|}\Theta'_L x'\right] \\
&\qquad + \frac{1 - |d_2|}{2} \cdot P_1\left[\frac{2}{1 - |d_2|} y'\right] \\
&\quad \overset{(b)}{=} \frac{1}{|d_2|} \cdot P_1\left[\Lambda_L x'\right] + \frac{2}{1 - |d_2|} \cdot P_1\left[\Theta'_L x'\right] + \frac{2}{1 - |d_2|} \cdot P_1\left[y'\right] \\
&\quad \overset{(c)}{\preceq} \frac{\|\bar{P}_1[\Lambda_L]\|_\infty}{|d_2|} \cdot \bar{P}_1[\Lambda_L] \cdot P_1\left[x'\right] \\
&\qquad + \frac{2\|\bar{P}_1[\Theta'_L]\|_\infty}{1 - |d_2|} \cdot \bar{P}_1[\Theta'_L] \cdot P_1\left[x'\right] + \frac{2}{1 - |d_2|} \cdot P_1\left[y'\right] \\
&\quad \overset{(d)}{\preceq} \bar{P}_1[\Lambda_L] \cdot P_1\left[x'\right] + \frac{2}{1 - |d_2|} \cdot \Theta'_L \cdot P_1\left[x'\right] + \frac{2}{1 - |d_2|} \cdot P_1\left[y'\right] \\
&\quad \overset{(e)}{\preceq} |d_2| \cdot I_L \cdot P_1\left[x'\right] + \frac{2}{1 - |d_2|} \cdot \Theta_L \cdot P_1\left[x'\right] + \frac{2}{1 - |d_2|} \cdot P_1\left[y'\right] \\
&\quad \overset{(f)}{=} \Gamma_e \cdot P_1\left[x'\right] + \frac{2}{1 - |d_2|} \cdot P_1\left[y'\right] \qquad (196)
\end{aligned}$$

where step (a) uses the convexity property (148), step (b) uses the scaling property, step (c) uses variance relation (160), step (d) uses $\|\bar{P}_1[\Lambda_L]\|_\infty = |d_2|$, $\bar{P}_1[\Theta'_L] = \Theta'_L$ and $\|\bar{P}_1[\Theta'_L]\|_\infty = \|\Theta'_L\|_\infty = 1$, step (e) uses $\bar{P}_1[\Lambda_L] \preceq |d_2| \cdot I_L$ and $\Theta'_L \preceq \Theta_L$, where Θ_L denotes a matrix of the same form as (192) but of size $L \times L$, step (f) uses the definition of the matrix Γ_e in (171). The above derivation assumes $|d_2| \neq 0$. When $|d_2| = 0$, we can verify that the above inequality still holds. To see this, we first notice that when $|d_2| = 0$, the relation $0 \leq |d_{n_0}| \leq \cdots \leq |d_2|$ implies that $d_{n_0} = \cdots = d_2 = 0$ so that $\Lambda_L = 0$ and $D_L = \Theta'_L$ — see (193) and (195). Therefore, similar to the steps (a)–(f) in (196), we get

$$\begin{aligned}
P_1[D_L x' + y'] &= P_1[\Theta'_L x' + y'] \\
&= P_1\big[\frac{1}{2} \cdot 2\Theta'_L x' + \frac{1}{2} \cdot 2y'\big] \\
&\preceq \frac{1}{2} \cdot P_1[2\Theta'_L x'] + \frac{1}{2} \cdot P_1[2y'] \\
&= \frac{1}{2} \cdot 2^2 \cdot P_1[\Theta'_L x'] + \frac{1}{2} \cdot 2^2 \cdot P_1[y'] \\
&= 2P_1[\Theta'_L x'] + 2P_1[y'] \\
&\preceq 2\|\bar{P}_1[\Theta'_L]\|_\infty \cdot \bar{P}_1[\Theta'_L] P_1[x'] + 2P_1[y'] \\
&= 2\Theta'_L P_1[x'] + 2P_1[y'] \\
&\preceq 2\Theta_L P_1[x'] + 2P_1[y'] \qquad (197)
\end{aligned}$$

By (171), we have $\Gamma_e = 2\Theta_L$ when $|d_2| = 0$. Therefore, the above expression is the same as the one on the right-hand side of (196).

(***Property 5: Stable Kronecker Jordan operator***) Using (195) we have

$$\begin{aligned}
&P\left[\mathcal{D}_L x_e + y_e\right] \\
&= P\left[(\Lambda_L \otimes I_M)x_e + (\Theta'_L \otimes I_M)x_e + y_e\right] \\
&= P\Big[|d_2| \cdot \frac{1}{|d_2|} \cdot (\Lambda_L \otimes I_M)x_e + \frac{1-|d_2|}{2} \cdot \frac{2}{1-|d_2|} \cdot (\Theta'_L \otimes I_M)x_e \\
&\qquad + \frac{1-|d_2|}{2} \cdot \frac{2}{1-|d_2|} \cdot y_e\Big] \\
&\overset{(a)}{\preceq} |d_2| \cdot P\Big[\frac{1}{|d_2|} \cdot (\Lambda_L \otimes I_M)x_e\Big] \\
&\quad + \frac{1-|d_2|}{2} \cdot P\Big[\frac{2}{1-|d_2|} \cdot (\Theta'_L \otimes I_M)x_e\Big] \\
&\quad + \frac{1-|d_2|}{2} \cdot P\Big[\frac{2}{1-|d_2|} \cdot y_e\Big] \\
&\overset{(b)}{=} \frac{1}{|d_2|} \cdot P[(\Lambda_L \otimes I_M)x_e] + \frac{2}{1-|d_2|} \cdot P[(\Theta'_L \otimes I_M)x_e] \\
&\quad + \frac{2}{1-|d_2|} \cdot P[y_e] \\
&\overset{(c)}{\preceq} \frac{\|\bar{P}[(\Lambda_L \otimes I_M)]\|_\infty}{|d_2|} \cdot \bar{P}[(\Lambda_L \otimes I_M)] \cdot P[x_e] \\
&\quad + \frac{2\|\bar{P}[\Theta'_L \otimes I_M]\|_\infty}{1-|d_2|} \cdot \bar{P}[\Theta'_L \otimes I_M] \cdot P[x_e] + \frac{2}{1-|d_2|} \cdot P[y_e] \\
&\overset{(d)}{\preceq} \bar{P}[(\Lambda_L \otimes I_M)] \cdot P[x_e] + \frac{2}{1-|d_2|} \cdot \Theta'_L \cdot P[x_e] \\
&\quad + \frac{2}{1-|d_2|} \cdot P[y_e] \\
&\overset{(e)}{\preceq} |d_2| \cdot I_L \cdot P[x_e] + \frac{2}{1-|d_2|} \cdot \Theta_L \cdot P[x_e] + \frac{2}{1-|d_2|} \cdot P[y_e] \\
&\overset{(f)}{=} \Gamma_e \cdot P[x_e] + \frac{2}{1-|d_2|} \cdot P[y_e] \qquad (198)
\end{aligned}$$

where step (a) uses the convexity property (148), step (b) uses the scaling property, step (c) uses variance relation (160), step (d) uses $\|\bar{P}[\Lambda_L \otimes I_M]\|_\infty = |d_2|$ and $\bar{P}[\Theta'_L \otimes I_M] = \Theta'_L$, step (e) uses $\bar{P}[\Lambda_L \otimes I_M] \preceq |d_2| \cdot I_L$ and $\Theta'_L \preceq \Theta_L$, and step (f) uses the definition of the matrix Γ_e in (171). Likewise, we can also verify that the above inequalty holds for the case $|d_2| = 0$.

APPENDIX E
PROOF OF THEOREM 1

Consider the following operator:

$$T_0(w) \triangleq w - \frac{\lambda_L}{\|p\|_1^2 \lambda_U^2} \sum_{k=1}^{N} p_k s_k(w) \qquad (199)$$

As long as we are able to show that $T_0(w)$ is a strict contraction mapping, i.e., $\forall\, x, y$, $\|T_0(x) - T_0(y)\| \leq \gamma_0 \|x - y\|$ with $\gamma_o < 1$, then we can invoke the Banach fixed point theorem [59, pp.299-300] to conclude that there exists a unique w^o such that $w^o = T_0(w^o)$, i.e.,

$$w^o = w^o - \frac{\lambda_L}{\|p\|_1^2 \lambda_U^2} \sum_{k=1}^{N} p_k s_k(w^o) \;\Leftrightarrow\; \sum_{k=1}^{N} p_k s_k(w^o) = 0 \qquad (200)$$

as desired. Now, to show that $T_0(\cdot)$ defined in (199) is indeed a contraction, we compare $T_0(\cdot)$ with $T_c(\cdot)$ in (99) and observe that $T_0(w)$ has the same form as $T_c(\cdot)$ if we set $\mu_{\max} = \frac{\lambda_L}{\|p\|_1^2 \lambda_U^2}$ in (99). Therefore, calling upon property (164) and using $\mu_{\max} = \frac{\lambda_L}{\|p\|_1^2 \lambda_U^2}$ in the expression for γ_c in (166), we obtain

$$\begin{aligned}
&P[T_0(x) - T_0(y)] \\
&\preceq \left(1 - \frac{\lambda_L}{\|p\|_1^2 \lambda_U^2} \lambda_L + \frac{1}{2}\left(\frac{\lambda_L}{\|p\|_1^2 \lambda_U^2}\right)^2 \|p\|_1^2 \lambda_U^2\right)^2 \cdot P[x - y] \\
&= \left(1 - \frac{1}{2}\frac{\lambda_L^2}{\|p\|_1^2 \lambda_U^2}\right)^2 \cdot P[x - y] \qquad (201)
\end{aligned}$$

By the definition of $P[\cdot]$ in (105), the above inequality is equivalent to

$$\|T_0(x) - T_0(y)\|^2 \leq \left(1 - \frac{1}{2}\frac{\lambda_L^2}{\|p\|_1^2 \lambda_U^2}\right)^2 \cdot \|x - y\|^2 \qquad (202)$$

It remains to show that $|1 - \lambda_L^2/(2\|p\|_1^2 \lambda_U^2)| < 1$. By (26) and the fact that λ_L, $\|p\|_1^2$ and λ_U^2 are positive, we have

$$\frac{1}{2} < 1 - \frac{1}{2}\frac{\lambda_L^2}{\|p\|_1^2 \lambda_U^2} < 1 \qquad (203)$$

Therefore, $T_0(w)$ is a strict contraction mapping.

APPENDIX F
PROOF OF THEOREM 2

By Theorem 1, w^o is the unique solution to equation (109). Subtracting both sides of (109) from w^o, we recognize that w^o is also the unique solution to the following equation:

$$w^o = w^o - \mu_{\max}\sum_{k=1}^{N} p_k s_k(w^o) \tag{204}$$

so that $w^o = T_c(w^o)$. Applying property (164), we obtain

$$\begin{aligned}\|\tilde{w}_{c,i}\|^2 &= P[w^o - \bar{w}_{c,i}]\\ &= P[T_c(w^o) - T_c(\bar{w}_{c,i-1})]\\ &\preceq \gamma_c^2 \cdot P[w^o - \bar{w}_{c,i-1}]\\ &\preceq \gamma_c^{2i} \cdot P[w^o - \bar{w}_{c,0}]\\ &= \gamma_c^{2i} \cdot \|\tilde{w}_{c,0}\|^2\end{aligned} \tag{205}$$

Since $\gamma_c > 0$, the upper bound on the right-hand side will converge to zero if $\gamma_c < 1$. From its definition (166), this condition is equivalent to requiring

$$1 - \mu_{\max}\lambda_L + \frac{1}{2}\mu_{\max}^2\|p\|_1^2\lambda_U^2 < 1 \tag{206}$$

Solving the above quadratic inequality in $\mu_{\max}$, we obtain (112). On the other hand, to prove the lower bound in (112), we apply (165) and obtain

$$\begin{aligned}\|\tilde{w}_{c,i}\|^2 &= P[w^o - \bar{w}_{c,i}]\\ &= P[T_c(w^o) - T_c(\bar{w}_{c,i-1})]\\ &\succeq (1 - 2\mu_{\max}\|p\|_1\lambda_U) \cdot P[w^o - \bar{w}_{c,i-1}]\\ &\succeq (1 - 2\mu_{\max}\|p\|_1\lambda_U)^i \cdot P[w^o - \bar{w}_{c,0}]\\ &= (1 - 2\mu_{\max}\|p\|_1\lambda_U)^i \cdot \|\tilde{w}_{c,0}\|^2\end{aligned} \tag{207}$$

APPENDIX G
PROOF OF THEOREM 3

Since (110) already establishes that $\bar{w}_{c,i}$ approaches w^o asymptotically (so that $\tilde{w}_{c,i} \to 0$), and since from Assumption 5 we know that $s_k(w)$ is differentiable when $\|\tilde{w}_{c,i}\| \le r_H$ for large enough i, we are justified to use the mean-value theorem [49, p.24] to obtain the following useful relation:

$$\begin{aligned}&s_k(\bar{w}_{c,i-1}) - s_k(w^o)\\ &= -\left[\int_0^1 \nabla_{w^T} s_k(w^o - t\tilde{w}_{c,i-1})dt\right]\tilde{w}_{c,i-1}\\ &= -\nabla_{w^T} s_k(w^o)\cdot\tilde{w}_{c,i-1}\\ &\quad - \int_0^1 \big[\nabla_{w^T} s_k(w^o - t\tilde{w}_{c,i-1}) - \nabla_{w^T} s_k(w^o)\big]dt\cdot\tilde{w}_{c,i-1}\end{aligned} \tag{208}$$

Therefore, subtracting w^o from both sides of (47) and using (109) we get,

$$\begin{aligned}\tilde{w}_{c,i} &= \tilde{w}_{c,i-1} + \mu_{\max}\sum_{k=1}^{N} p_k(s_k(\bar{w}_{c,i-1}) - s_k(w^o))\\ &= [I - \mu_{\max}H_c]\tilde{w}_{c,i-1} - \mu_{\max}\cdot e_{i-1}\end{aligned} \tag{209}$$

where

$$H_c \triangleq \sum_{k=1}^{N} p_k \nabla_{w^T} s_k(w^o) \tag{210}$$

$$e_{i-1} \triangleq \sum_{k=1}^{N} p_k \int_0^1 \big[\nabla_{w^T} s_k(w^o - t\tilde{w}_{c,i-1}) - \nabla_{w^T} s_k(w^o)\big]dt\cdot\tilde{w}_{c,i-1} \tag{211}$$

Furthermore, the perturbation term e_{i-1} satisfies the following bound:

$$\begin{aligned}\|e_{i-1}\| &\le \sum_{k=1}^{N} p_k \int_0^1 \big\|\nabla_{w^T} s_k(w^o - t\tilde{w}_{c,i-1}) - \nabla_{w^T} s_k(w^o)\big\|dt\\ &\quad\cdot\|\tilde{w}_{c,i-1}\|\\ &\le \sum_{k=1}^{N} p_k \int_0^1 \lambda_H\cdot t\cdot\|\tilde{w}_{c,i-1}\|dt\cdot\|\tilde{w}_{c,i-1}\|\\ &= \frac{1}{2}\|p\|_1\lambda_H\cdot\|\tilde{w}_{c,i-1}\|^2\end{aligned} \tag{212}$$

Evaluating the weighted Euclidean norm of both sides of (209), we get

$$\begin{aligned}\|\tilde{w}_{c,i}\|_\Sigma^2 &= \|\tilde{w}_{c,i-1}\|_{B_c^T\Sigma B_c}^2 - 2\mu_{\max}\cdot\tilde{w}_{c,i-1}^T B_c^T\Sigma e_{i-1}\\ &\quad + \mu_{\max}^2\cdot\|e_{i-1}\|_\Sigma^2\end{aligned} \tag{213}$$

where

$$B_c = I - \mu_{\max}H_c \tag{214}$$

Moreover, $\|x\|_\Sigma^2 = x^T\Sigma x$, and Σ is an arbitrary positive semi-definite weighting matrix. The second and third terms on the right-hand side of (213) satisfy the following bounds:

$$\begin{aligned}&\big|\tilde{w}_{c,i-1}^T B_c^T\Sigma e_{i-1}\big|\\ &\le \|\tilde{w}_{c,i-1}\|\cdot\|B_c^T\|\cdot\|\Sigma\|\cdot\|e_{i-1}\|\\ &\overset{(a)}{\le} \|\tilde{w}_{c,i-1}\|\cdot\|B_c^T\|\cdot\mathrm{Tr}(\Sigma)\cdot\|e_{i-1}\|\\ &\le \|\tilde{w}_{c,i-1}\|\cdot\|B_c^T\|\cdot\mathrm{Tr}(\Sigma)\cdot\frac{\lambda_H\|p\|_1}{2}\cdot\|\tilde{w}_{c,i-1}\|^2\end{aligned} \tag{215}$$

and

$$\begin{aligned}\|e_{i-1}\|_\Sigma^2 &\le \|\Sigma\|\cdot\|e_{i-1}\|^2\\ &\overset{(b)}{\le} \mathrm{Tr}(\Sigma)\cdot\|e_{i-1}\|^2\\ &\le \mathrm{Tr}(\Sigma)\cdot\frac{\lambda_H^2\|p\|_1^2}{4}\cdot\|\tilde{w}_{c,i-1}\|^4\end{aligned} \tag{216}$$

where for steps (a) and (b) of the above inequalities we used the property $\|\Sigma\| \le \mathrm{Tr}(\Sigma)$. This is because we consider here the spectral norm, $\|\Sigma\| = \sigma_{\max}(\Sigma)$, where $\sigma_{\max}(\cdot)$ denotes the maximum singular value. Since Σ is symmetric and positive semidefinite, its singular values are the same as its eigenvalues so that $\|\Sigma\| = \lambda_{\max}(\Sigma) \le \sum_m \lambda_m(\Sigma) = \mathrm{Tr}(\Sigma)$. Now, for any given small $\epsilon > 0$, there exists i_0 such that, for $i \ge i_0$, we have $\|\tilde{w}_{c,i-1}\| \le \epsilon$ so that

$$\big|\tilde{w}_{c,i-1}^T B_c^T\Sigma e_{i-1}\big| \le \epsilon\cdot\|B_c^T\|\cdot\mathrm{Tr}(\Sigma)\cdot\frac{\lambda_H\|p\|_1}{2}\cdot\|\tilde{w}_{c,i-1}\|^2 \tag{217}$$

$$\|e_{i-1}\|_\Sigma^2 \le \epsilon^2\cdot\mathrm{Tr}(\Sigma)\cdot\frac{\lambda_H^2\|p\|_1^2}{4}\cdot\|\tilde{w}_{c,i-1}\|^2 \tag{218}$$

Substituting (217)–(218) into (213), we obtain

$$\|\tilde{w}_{c,i-1}\|^2_{B_c^T\Sigma B_c-\Delta} \le \|\tilde{w}_{c,i}\|^2_{\Sigma} \le \|\tilde{w}_{c,i-1}\|^2_{B_c^T\Sigma B_c+\Delta} \quad (219)$$

where

$$\Delta \triangleq \mu_{\max}\epsilon\cdot\lambda_H\|p\|_1\cdot\big[\|B_c^T\|+\mu_{\max}\epsilon\frac{\lambda_H\|p\|_1}{4}\big]\cdot\mathrm{Tr}(\Sigma)\cdot I_M$$
$$= O(\mu_{\max}\epsilon)\cdot\mathrm{Tr}(\Sigma)\cdot I_M \quad (220)$$

Let $\sigma=\mathrm{vec}(\Sigma)$ denote the vectorization operation that stacks the columns of a matrix Σ on top of each other. We shall use the notation $\|x\|^2_\sigma$ and $\|x\|^2_\Sigma$ interchangeably to denote the weighted squared Euclidean norm of a vector. Using the Kronecker product property [58, p.147]: $\mathrm{vec}(U\Sigma V)=(V^T\otimes U)\mathrm{vec}(\Sigma)$, we can vectorize the matrices $B_c^T\Sigma B_c+\Delta$ and $B_c^T\Sigma B_c-\Delta$ in (219) as $\mathcal{F}_+\sigma$ and $\mathcal{F}_-\sigma$, respectively, where

$$\mathcal{F}_+ \triangleq B_c^T\otimes B_c^T+\mu_{\max}\epsilon\cdot\lambda_H\|p\|_1\cdot\big[\|B_c^T\|+\mu_{\max}\epsilon\frac{\lambda_H\|p\|_1}{4}\big]qq^T$$
$$= B_c^T\otimes B_c^T+O(\mu_{\max}\epsilon) \quad (221)$$

$$\mathcal{F}_- \triangleq B_c^T\otimes B_c^T-\mu_{\max}\epsilon\cdot\lambda_H\|p\|_1\cdot\big[\|B_c^T\|+\mu_{\max}\epsilon\frac{\lambda_H\|p\|_1}{4}\big]qq^T$$
$$= B_c^T\otimes B_c^T-O(\mu_{\max}\epsilon) \quad (222)$$

where $q\triangleq\mathrm{vec}(I_M)$, and we have used the fact that $\mathrm{Tr}(\Sigma)=\mathrm{Tr}(\Sigma I_M)=\mathrm{vec}(I_M)^T\mathrm{vec}(\Sigma)=q^T\sigma$. In this way, we can write relation (219) as

$$\|\tilde{w}_{c,i-1}\|^2_{\mathcal{F}_-\sigma} \le \|\tilde{w}_{c,i}\|^2_{\sigma} \le \|\tilde{w}_{c,i-1}\|^2_{\mathcal{F}_+\sigma} \quad (223)$$

Using a state-space technique from [60, pp.344-346], we conclude that $\|\tilde{w}_{c,i}\|^2_\Sigma$ converges at a rate that is between $\rho(\mathcal{F}_-)$ and $\rho(\mathcal{F}_+)$. Recalling from (221)–(222) that $\mathcal{F}_+$ and $\mathcal{F}_-$ are perturbed matrices of $B_c^T\otimes B_c^T$, and since the perturbation term is $O(\epsilon\mu_{\max})$ which is small for small ϵ, we would expect the spectral radii of $\mathcal{F}_+$ and $\mathcal{F}_-$ to be small perturbations of $\rho(B_c^T\otimes B_c^T)$. This claim is justified below.

Lemma 7 (Perturbation of spectral radius): Let $\epsilon\ll 1$ be a sufficiently small positive number. For any $M\times M$ matrix X, the spectral radius of the perturbed matrix $X+E$ for $E=O(\epsilon)$ is

$$\rho(X+E)=\rho(X)+O\big(\epsilon^{\frac{1}{2(M-1)}}\big) \quad (224)$$

Proof: Let $X=TJT^{-1}$ be the Jordan canonical form of the matrix X. Without loss of generality, we consider the case where there are two Jordan blocks:

$$J=\mathrm{diag}\{J_1,J_2\} \quad (225)$$

where $J_1\in\mathbb{R}^{L\times L}$ and $J_2\in\mathbb{R}^{(M-L)\times(M-L)}$ are Jordan blocks of the form

$$J_k=\begin{bmatrix}\lambda_k & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1\\ & & & \lambda_k\end{bmatrix} \quad (226)$$

with $|\lambda_1|>|\lambda_2|$. Since $X+E$ is similar to $T^{-1}(X+E)T$, the matrix $X+E$ has the same set of eigenvalues as $J+E_0$ where

$$E_0\triangleq T^{-1}ET=O(\epsilon) \quad (227)$$

Let

$$\epsilon_0\triangleq\epsilon^{\frac{1}{2(M-1)}} \quad (228)$$

$$D_{\epsilon_0}\triangleq\mathrm{diag}\big\{1,\epsilon_0,\ldots,\epsilon_0^{M-1}\big\} \quad (229)$$

Then, by similarity again, the matrix $J+E_0$ has the same set of eigenvalues as

$$D_{\epsilon_0}^{-1}(J+E_0)D_{\epsilon_0}=D_{\epsilon_0}^{-1}JD_{\epsilon_0}+E_1 \quad (230)$$

where $E_1\triangleq D_{\epsilon_0}^{-1}E_0D_{\epsilon_0}$. Note that the ∞-induced norm (the maximum absolute row sum) of E_1 is bounded by

$$\|E_1\|_\infty \le \|D_{\epsilon_0}^{-1}\|_\infty\cdot\|E_0\|_\infty\cdot\|D_{\epsilon_0}\|_\infty$$
$$=\frac{1}{\epsilon_0^{M-1}}\cdot O(\epsilon)\cdot 1=\frac{1}{\epsilon^{\frac{1}{2}}}\cdot O(\epsilon)=O(\epsilon^{\frac{1}{2}}) \quad (231)$$

and that

$$D_{\epsilon_0}^{-1}JD_{\epsilon_0}=\mathrm{diag}\{J_1',J_2'\} \quad (232)$$

where

$$J_k'=\begin{bmatrix}\lambda_k & \epsilon_0 & & \\ & \ddots & \ddots & \\ & & \ddots & \epsilon_0\\ & & & \lambda_k\end{bmatrix} \quad (233)$$

Then, by appealing to Geršgorin Theorem [52, p.344], we conclude that the eigenvalues of the matrix $D_{\epsilon_0}^{-1}JD_{\epsilon_0}+E_1$, which are also the eigenvalues of the matrices $J+E_0$ and $X+E$, lie inside the union of the Geršgorin discs, namely,

$$\bigcup_{m=1}^{M}\mathcal{G}_m \quad (234)$$

where $\mathcal{G}_m$ is the mth Geršgorin disc defined as

$$\mathcal{G}_m \triangleq \begin{cases}\Big\{\lambda:\ |\lambda-\lambda_1|\le\epsilon_0+\sum_{\ell=1}^{M}|E_{1,m\ell}|\Big\}, & 1\le m\le L\\ \Big\{\lambda:\ |\lambda-\lambda_2|\le\epsilon_0+\sum_{\ell=1}^{M}|E_{1,m\ell}|\Big\}, & L<m\le M\end{cases}$$
$$=\begin{cases}\big\{\lambda:\ |\lambda-\lambda_1|\le O\big(\epsilon^{\frac{1}{2(M-1)}}\big)\big\}, & 1\le m\le L\\ \big\{\lambda:\ |\lambda-\lambda_2|\le O\big(\epsilon^{\frac{1}{2(M-1)}}\big)\big\}, & L<m\le M\end{cases} \quad (235)$$

and where $E_{1,m\ell}$ denotes the (m,ℓ)-th entry of the matrix E_1. In the last step we used (228) and (231). Observe from (235) that there are two clusters of Geršgorin discs that are centered around λ_1 and λ_2, respectively, and have radii on the order of $O(\epsilon^{\frac{1}{2(M-1)}})$. A further statement from Geršgorin theorem shows that if the these two clusters of discs happen to be disjoint, which is true in our case since $|\lambda_1|>|\lambda_2|$ and we can select ϵ to be sufficiently small to ensure this property. Then there are exactly L eigenvalues of $X+E$ in $\cup_{m=1}^{L}\mathcal{G}_m$ while the remaining $M-L$ eigenvalues are in $\cup_{m=M-L}^{M}\mathcal{G}_m$. From $|\lambda_1|>|\lambda_2|$, we conclude that the largest eigenvalue of $D_{\epsilon_0}^{-1}JD_{\epsilon_0}+E_1$ is $\lambda_1+O(\epsilon^{\frac{1}{2(M-1)}})$, which establishes (224). ■

Using (224) for $\mathcal{F}_+$ and $\mathcal{F}_-$ in (221)–(222), we conclude that

$$\rho(\mathcal{F}_+) = \big[\rho(I_M - \mu_{\max}H_c)\big]^2 + O\big((\mu_{\max}\epsilon)^{\frac{1}{2(M-1)}}\big) \quad (236)$$

$$\rho(\mathcal{F}_-) = \big[\rho(I_M - \mu_{\max}H_c)\big]^2 + O\big((\mu_{\max}\epsilon)^{\frac{1}{2(M-1)}}\big) \quad (237)$$

which holds for arbitrarily small ϵ. Since the convergence rate of $\|\tilde{w}_{c,i}\|^2$ is between $\rho(\mathcal{F}_+)$ and $\rho(\mathcal{F}_-)$, we arrive at (114).

APPENDIX H
PROOF OF LEMMA 4

From the definition in (108), it suffices to establish a joint inequality recursion for both $\mathbb{E}P[\check{\boldsymbol{w}}_{c,i}]$ and $\mathbb{E}P[\boldsymbol{w}_{e,i}]$. To begin with, we state the following bounds on the perturbation terms in (93).

Lemma 8 (Bounds on the perturbation terms): The following bounds hold for any $i \geq 0$.

$$P[\boldsymbol{z}_{i-1}] \preceq \lambda_U^2 \cdot \big\|\bar{P}_1[A_1^T U_L]\big\|_\infty^2 \cdot \mathbb{1}\mathbb{1}^T \cdot P[\boldsymbol{w}_{e,i-1}] \quad (238)$$

$$P[s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})] \preceq 3\lambda_U^2 \cdot P[\check{\boldsymbol{w}}_{c,i-1}]\cdot\mathbb{1} + 3\lambda_U^2\|\tilde{w}_{c,0}\|^2\cdot\mathbb{1} + 3g^o \quad (239)$$

$$\begin{aligned}\mathbb{E}\{P[\boldsymbol{v}_i]|\mathcal{F}_{i-1}\} &\preceq 4\alpha\cdot\mathbb{1}\cdot P[\check{\boldsymbol{w}}_{c,i-1}] \\ &\quad + 4\alpha\cdot\|\bar{P}[\mathcal{A}_1^T\mathcal{U}_L]\|_\infty^2\cdot\mathbb{1}\mathbb{1}^T P[\boldsymbol{w}_{e,i-1}] \\ &\quad + \big[4\alpha\cdot(\|\tilde{w}_{c,0}\|^2 + \|w^o\|^2) + \sigma_v^2\big]\cdot\mathbb{1} \quad (240)\\ \mathbb{E}P[\boldsymbol{v}_i] &\preceq 4\alpha\cdot\mathbb{1}\cdot\mathbb{E}P[\check{\boldsymbol{w}}_{c,i-1}] \\ &\quad + 4\alpha\cdot\|\bar{P}[\mathcal{A}_1^T\mathcal{U}_L]\|_\infty^2\cdot\mathbb{1}\mathbb{1}^T\mathbb{E}P[\boldsymbol{w}_{e,i-1}] \\ &\quad + \big[4\alpha\cdot(\|\tilde{w}_{c,0}\|^2 + \|w^o\|^2) + \sigma_v^2\big]\cdot\mathbb{1} \quad (241)\end{aligned}$$

where $P[\check{\boldsymbol{w}}_{c,i-1}] = \|\check{\boldsymbol{w}}_{c,i-1}\|^2$, and $g^o \triangleq P[s(\mathbb{1}\otimes w^o)]$.

Proof: See Appendix I. ∎

Now, we derive an inequality recursion for $\mathbb{E}P[\check{\boldsymbol{w}}_{c,i}]$ from (103). Note that

$$\begin{aligned}&\mathbb{E}P[\check{\boldsymbol{w}}_{c,i}] = \mathbb{E}\|\check{\boldsymbol{w}}_{c,i}\|^2 \\ &= \mathbb{E}P\big[T_c(\boldsymbol{w}_{c,i-1}) - T_c(\bar{w}_{c,i-1}) - \mu_{\max}\cdot(p^T\otimes I_M)\boldsymbol{z}_{i-1} \\ &\qquad - \mu_{\max}\cdot(p^T\otimes I_M)\boldsymbol{v}_i)\big] \\ &\overset{(a)}{=} \mathbb{E}P\left[T_c(\boldsymbol{w}_{c,i-1}) - T_c(\bar{w}_{c,i-1}) - \mu_{\max}\cdot(p^T\otimes I_M)\boldsymbol{z}_{i-1}\right] \\ &\quad + \mu_{\max}^2\cdot\mathbb{E}P\left[(p^T\otimes I_M)\boldsymbol{v}_i)\right] \\ &= \mathbb{E}P\Bigg[\gamma_c\cdot\frac{1}{\gamma_c}\left(T_c(\boldsymbol{w}_{c,i-1}) - T_c(\bar{w}_{c,i-1})\right) \\ &\qquad + (1-\gamma_c)\cdot\frac{-\mu_{\max}}{1-\gamma_c}(p^T\otimes I_M)\boldsymbol{z}_{i-1}\Bigg] \\ &\quad + \mu_{\max}^2\cdot\mathbb{E}P\left[(p^T\otimes I_M)\boldsymbol{v}_i)\right] \\ &\overset{(b)}{\preceq} \gamma_c\cdot\frac{1}{\gamma_c^2}\mathbb{E}P\left[T_c(\boldsymbol{w}_{c,i-1}) - T_c(\bar{w}_{c,i-1})\right] \\ &\quad + (1-\gamma_c)\cdot\frac{\mu_{\max}^2}{(1-\gamma_c)^2}\mathbb{E}P\left[(p^T\otimes I_M)\boldsymbol{z}_{i-1}\right] \\ &\quad + \mu_{\max}^2\mathbb{E}P\left[(p^T\otimes I_M)\boldsymbol{v}_i\right] \\ &\overset{(c)}{\preceq} \gamma_c\cdot\mathbb{E}P\left[\check{\boldsymbol{w}}_{c,i-1}\right] + \frac{\mu_{\max}^2}{1-\gamma_c}\mathbb{E}P\left[(p^T\otimes I_M)\boldsymbol{z}_{i-1}\right] \\ &\quad + \mu_{\max}^2\mathbb{E}P\left[(p^T\otimes I_M)\boldsymbol{v}_i\right] \\ &= \gamma_c\cdot\mathbb{E}P\left[\check{\boldsymbol{w}}_{c,i-1}\right] + \frac{\mu_{\max}^2}{1-\gamma_c}\mathbb{E}\left\|(p^T\otimes I_M)\boldsymbol{z}_{i-1}\right\|^2 \\ &\quad + \mu_{\max}^2\mathbb{E}\left\|(p^T\otimes I_M)\boldsymbol{v}_i\right\|^2 \\ &\overset{(d)}{=} \gamma_c\cdot\mathbb{E}P\left[\check{\boldsymbol{w}}_{c,i-1}\right] + \frac{\mu_{\max}^2}{1-\gamma_c}\mathbb{E}\left\|\sum_{k=1}^N p_k\boldsymbol{z}_{k,i-1}\right\|^2 \\ &\quad + \mu_{\max}^2\mathbb{E}\left\|\sum_{k=1}^N p_k\boldsymbol{v}_{k,i}\right\|^2 \\ &= \gamma_c\cdot\mathbb{E}P\left[\check{\boldsymbol{w}}_{c,i-1}\right] \\ &\quad + \frac{\mu_{\max}^2}{1-\gamma_c}\cdot\left(\sum_{l=1}^N p_l\right)^2\cdot\mathbb{E}\left\|\sum_{k=1}^N\frac{p_k}{\sum_{l=1}^N p_l}\boldsymbol{z}_{k,i-1}\right\|^2 \\ &\quad + \mu_{\max}^2\cdot\left(\sum_{l=1}^N p_l\right)^2\cdot\mathbb{E}\left\|\sum_{k=1}^N\frac{p_k}{\sum_{l=1}^N p_l}\boldsymbol{v}_{k,i}\right\|^2 \\ &\overset{(e)}{\leq} \gamma_c\cdot\mathbb{E}P\left[\check{\boldsymbol{w}}_{c,i-1}\right] \\ &\quad + \frac{\mu_{\max}^2}{1-\gamma_c}\cdot\left(\sum_{l=1}^N p_l\right)^2\cdot\sum_{k=1}^N\frac{p_k}{\sum_{l=1}^N p_l}\mathbb{E}\left\|\boldsymbol{z}_{k,i-1}\right\|^2 \\ &\quad + \mu_{\max}^2\cdot\left(\sum_{l=1}^N p_l\right)^2\cdot\sum_{k=1}^N\frac{p_k}{\sum_{l=1}^N p_l}\mathbb{E}\left\|\boldsymbol{v}_{k,i}\right\|^2 \\ &= \gamma_c\cdot\mathbb{E}P\left[\check{\boldsymbol{w}}_{c,i-1}\right] + \frac{\mu_{\max}^2}{1-\gamma_c}\cdot\left(\sum_{l=1}^N p_l\right)\cdot\sum_{k=1}^N p_k\mathbb{E}\left\|\boldsymbol{z}_{k,i-1}\right\|^2 \\ &\quad + \mu_{\max}^2\cdot\left(\sum_{l=1}^N p_l\right)\cdot\sum_{k=1}^N p_k\mathbb{E}\left\|\boldsymbol{v}_{k,i}\right\|^2 \\ &= \gamma_c\cdot\mathbb{E}P\left[\check{\boldsymbol{w}}_{c,i-1}\right] + \frac{\mu_{\max}^2}{1-\gamma_c}\cdot\|p\|_1\cdot p^T\mathbb{E}P[\boldsymbol{z}_{i-1}] \\ &\quad + \mu_{\max}^2\cdot\|p\|_1\cdot p^T\mathbb{E}P[\boldsymbol{v}_i] \\ &\overset{(f)}{=} \gamma_c\cdot\mathbb{E}P\left[\check{\boldsymbol{w}}_{c,i-1}\right] + \frac{\mu_{\max}\cdot\|p\|_1}{\lambda_L - \frac{1}{2}\mu_{\max}\|p\|_1^2\lambda_U^2}\cdot p^T\mathbb{E}P[\boldsymbol{z}_{i-1}] \\ &\quad + \mu_{\max}^2\cdot\|p\|_1\cdot p^T\mathbb{E}P[\boldsymbol{v}_i] \\ &\overset{(g)}{\preceq} \gamma_c\cdot\mathbb{E}P\left[\check{\boldsymbol{w}}_{c,i-1}\right] \\ &\quad + \frac{\mu_{\max}\|p\|_1}{\lambda_L - \mu_{\max}\frac{1}{2}\|p\|_1^2\lambda_U^2} \\ &\quad \cdot p^T\Big\{\lambda_U^2\cdot\big\|\bar{P}[\mathcal{A}_1^T\mathcal{U}_L]\big\|_\infty^2\cdot\mathbb{1}\mathbb{1}^T\cdot\mathbb{E}P[\boldsymbol{w}_{e,i-1}]\Big\} \\ &\quad + \mu_{\max}^2\cdot\|p\|_1\cdot p^T\Big\{4\alpha\cdot\mathbb{1}\cdot\mathbb{E}P[\check{\boldsymbol{w}}_{c,i-1}] \\ &\qquad + 4\alpha\cdot\|\bar{P}[\mathcal{A}_1^T\mathcal{U}_L]\|_\infty^2\cdot\mathbb{1}\mathbb{1}^T\mathbb{E}P[\boldsymbol{w}_{e,i-1}] \\ &\qquad + \big[4\alpha\cdot(\|\tilde{w}_{c,0}\|^2 + \|w^o\|^2) + \sigma_v^2\big]\cdot\mathbb{1}\Big\} \\ &\overset{(h)}{=} \big[\gamma_c + \mu_{\max}^2\cdot 4\alpha\|p\|_1^2\big]\cdot\mathbb{E}P[\check{\boldsymbol{w}}_{c,i-1}] \\ &\quad + \|p\|_1^2\cdot\big\|\bar{P}[\mathcal{A}_1^T\mathcal{U}_L]\big\|_\infty^2\cdot\lambda_U^2 \\ &\quad \cdot\Big[\frac{\mu_{\max}}{\lambda_L - \frac{1}{2}\mu_{\max}\|p\|_1^2\lambda_U^2} + 4\mu_{\max}^2\frac{\alpha}{\lambda_U^2}\Big]\cdot\mathbb{1}^T\mathbb{E}P[\boldsymbol{w}_{e,i-1}] \\ &\quad + \mu_{\max}^2\cdot\|p\|_1^2\cdot\big[4\alpha\left(\|\tilde{w}_{c,0}\|^2 + \|w^o\|^2\right) + \sigma_v^2\big] \quad (242)\end{aligned}$$

where step (a) uses the additivity property in Lemma 5 since the definition of $\boldsymbol{z}_{i-1}$ and $\boldsymbol{v}_i$ in (93) and the definition of $\boldsymbol{w}_{c,i-1}$ in (67) imply that $\boldsymbol{z}_{i-1}$ and $\boldsymbol{w}_{c,i-1}$ depend on all $\{\boldsymbol{w}_j\}$

for $j \leq i-1$, meaning that the cross terms are zero:

$$\mathbb{E}[\boldsymbol{v}_i \boldsymbol{z}_{i-1}^T] = \mathbb{E}\left\{\mathbb{E}\left[\boldsymbol{v}_i|\mathcal{F}_{i-1}\right]\boldsymbol{z}_{i-1}^T\right\} = 0$$
$$\mathbb{E}\big\{\boldsymbol{v}_i[T_c(\boldsymbol{w}_{c,i-1}) - T_c(\bar{w}_{c,i-1})]^T\big\}$$
$$= \mathbb{E}\left\{\mathbb{E}\left[\boldsymbol{v}_i|\mathcal{F}_{i-1}\right][T_c(\boldsymbol{w}_{c,i-1}) - T_c(\bar{w}_{c,i-1})]^T\right\} = 0$$

Step (b) uses the convexity property in Lemma 5, step (c) uses the variance property (164), step (d) uses the notation $\boldsymbol{z}_{k,i-1}$ and $\boldsymbol{v}_{k,i}$ to denote the kth $M\times 1$ block of the $NM\times 1$ vector $\boldsymbol{z}_{i-1}$ and $\boldsymbol{v}_i$, respectively, step (e) applies Jensen's inequality to the convex function $\|\cdot\|^2$, step (f) substitutes expression (166) for γ_c, step (g) substitutes the bounds for the perturbation terms from (238), (239), and (241), step (h) uses the fact that $p^T\mathbb{1} = \|p\|_1$.

Next, we derive the bound for $\mathbb{E}P[\boldsymbol{w}_{e,i}]$ from the recursion for $\boldsymbol{w}_{e,i}$ in (101):

$$\begin{aligned}
&\mathbb{E}P[\boldsymbol{w}_{e,i}]\\
&= \mathbb{E}P\big[\mathcal{D}_{N-1}\boldsymbol{w}_{e,i-1} - \mathcal{U}_R\mathcal{A}_2^T\mathcal{M}\, s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})\\
&\quad - \mathcal{U}_R\mathcal{A}_2^T\mathcal{M}\boldsymbol{z}_{i-1} - \mathcal{U}_R\mathcal{A}_2^T\mathcal{M}\boldsymbol{v}_i\big]\\
&\overset{(a)}{=} \mathbb{E}P\left[\mathcal{D}_{N-1}\boldsymbol{w}_{e,i-1} - \mathcal{U}_R\mathcal{A}_2^T\mathcal{M}\ \left(s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1}) + \boldsymbol{z}_{i-1}\right)\right]\\
&\quad + \mathbb{E}P\left[\mathcal{U}_R\mathcal{A}_2^T\mathcal{M}\boldsymbol{v}_i\right]\\
&\overset{(b)}{\preceq} \Gamma_e\cdot\mathbb{E}P\left[\boldsymbol{w}_{e,i-1}\right]\\
&\quad + \frac{2}{1-|\lambda_2(A)|}\cdot\mathbb{E}P\left[\mathcal{U}_R\mathcal{A}_2^T\mathcal{M}\big(s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1}) + \boldsymbol{z}_{i-1}\big)\right]\\
&\quad + \mathbb{E}P\left[\mathcal{U}_R\mathcal{A}_2^T\mathcal{M}\boldsymbol{v}_i\right]\\
&\overset{(c)}{\preceq} \Gamma_e\cdot\mathbb{E}P\left[\boldsymbol{w}_{e,i-1}\right]\\
&\quad + \frac{2}{1-|\lambda_2(A)|}\cdot\left\|\bar{P}[\mathcal{U}_R\mathcal{A}_2^T\mathcal{M}]\right\|_\infty^2\\
&\quad\quad\cdot\mathbb{1}\mathbb{1}^T\cdot\mathbb{E}P\left[s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1}) + \boldsymbol{z}_{i-1}\right]\\
&\quad + \left\|\bar{P}[\mathcal{U}_R\mathcal{A}_2^T\mathcal{M}]\right\|_\infty^2\cdot\mathbb{1}\mathbb{1}^T\cdot\mathbb{E}P\left[\boldsymbol{v}_i\right]\\
&\overset{(d)}{\preceq} \Gamma_e\cdot\mathbb{E}P\left[\boldsymbol{w}_{e,i-1}\right]\\
&\quad + \mu_{\max}^2\cdot\frac{4\left\|\bar{P}[\mathcal{U}_R\mathcal{A}_2^T]\right\|_\infty^2}{1-|\lambda_2(A)|}\cdot\mathbb{1}\mathbb{1}^T\\
&\quad\quad\cdot\big\{\mathbb{E}P\left[s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})\right] + \mathbb{E}P\left[\boldsymbol{z}_{i-1}\right]\big\}\\
&\quad + \mu_{\max}^2\cdot\left\|\bar{P}[\mathcal{U}_R\mathcal{A}_2^T]\right\|_\infty^2\cdot\mathbb{1}\mathbb{1}^T\cdot\mathbb{E}P\left[\boldsymbol{v}_i\right]\\
&\overset{(e)}{\preceq} \Bigg[\Gamma_e + 4\mu_{\max}^2\cdot\left\|\bar{P}[\mathcal{U}_R\mathcal{A}_2^T]\right\|_\infty^2\cdot\left\|\bar{P}[\mathcal{A}_1^T\mathcal{U}_L]\right\|_\infty^2\lambda_U^2 N\\
&\quad\quad\times\left(\frac{1}{1-|\lambda_2(A)|} + \frac{\alpha}{\lambda_U^2}\right)\mathbb{1}\mathbb{1}^T\Bigg]\cdot\mathbb{E}P[\boldsymbol{w}_{e,i-1}]\\
&\quad + 4\mu_{\max}^2\cdot\left\|\bar{P}[\mathcal{U}_R\mathcal{A}_2^T]\right\|_\infty^2\lambda_U^2 N\left(\frac{3}{1-|\lambda_2(A)|} + \frac{\alpha}{\lambda_U^2}\right)\\
&\quad\quad\cdot\mathbb{1}\cdot\mathbb{E}\|\check{\boldsymbol{w}}_{c,i-1}\|^2\\
&\quad + \mu_{\max}^2\cdot\left\|\bar{P}[\mathcal{U}_R\mathcal{A}_2^T]\right\|_\infty^2\cdot\Bigg[12\frac{\lambda_U^2\|\tilde{w}_{c,0}\|^2 N + \mathbb{1}^T g^o}{1-|\lambda_2(A)|}\\
&\quad\quad + N[4\alpha(\|\tilde{w}_{c,0}\|^2 + \|w^o\|^2) + \sigma_v^2]\Bigg]\cdot\mathbb{1}\\
&\overset{(f)}{\preceq} \Bigg[\Gamma_e + 4\mu_{\max}^2\cdot\left\|\bar{P}[\mathcal{U}_R\mathcal{A}_2^T]\right\|_\infty^2\cdot\left\|\bar{P}[\mathcal{A}_1^T\mathcal{U}_L]\right\|_\infty^2\lambda_U^2 N\\
&\quad\quad\times\left(\frac{1}{1-|\lambda_2(A)|} + \frac{\alpha}{\lambda_U^2}\right)\mathbb{1}\mathbb{1}^T\Bigg]\cdot\mathbb{E}P[\boldsymbol{w}_{e,i-1}]\\
&\quad + 4\mu_{\max}^2\cdot\left\|\bar{P}[\mathcal{U}_R\mathcal{A}_2^T]\right\|_\infty^2\lambda_U^2 N\left(\frac{3}{1-|\lambda_2(A)|} + \frac{\alpha}{\lambda_U^2}\right)\\
&\quad\quad\cdot\mathbb{1}\cdot\mathbb{E}\|\check{\boldsymbol{w}}_{c,i-1}\|^2\\
&\quad + \mu_{\max}^2\cdot N\left\|\bar{P}[\mathcal{U}_R\mathcal{A}_2^T]\right\|_\infty^2\cdot\Bigg[12\frac{\lambda_U^2\|\tilde{w}_{c,0}\|^2 + \|g^o\|_\infty}{1-|\lambda_2(A)|}\\
&\quad\quad + 4\alpha(\|\tilde{w}_{c,0}\|^2 + \|w^o\|^2) + \sigma_v^2\Bigg]\cdot\mathbb{1}
\end{aligned}\tag{243}$$

where step (a) uses the additivity property in Lemma 5 since the definition of $\boldsymbol{z}_{i-1}$ and $\boldsymbol{v}_i$ in (93) and the definitions of $\boldsymbol{w}_{c,i-1}$ and $\boldsymbol{w}_{e,i-1}$ in (67) imply that $\boldsymbol{z}_{i-1}$, $\boldsymbol{w}_{c,i-1}$ and $\boldsymbol{w}_{e,i-1}$ depend on all $\{\boldsymbol{w}_j\}$ for $j\leq i-1$, meaning that the cross terms between $\boldsymbol{v}_i$ and all other terms are zero, just as in step (a) of (242), step (b) uses the variance relation of stable Kronecker Jordan operators from (172) with $d_2 = \lambda_2(A)$, step (c) uses the variance relation of linear operator (161), step (d) uses the submultiplictive property (152) and $P[x+y]\preceq 2P[x]+2P[y]$ derived from the convexity property (148) and the scaling property in (238), (239), and (241), step (e) substitutes the bounds on the perturbation terms from (238)–(241), and step (f) uses the inequality $|\mathbb{1}^T g^o| \leq N\|g^o\|_\infty$.

Finally, using the quantities defined in (120)–(123), we can rewrite recursions (242) and (243) as

$$\begin{aligned}
\mathbb{E}P[\check{\boldsymbol{w}}_{c,i}] &\preceq (\gamma_c + \mu_{\max}^2\psi_0)\cdot\mathbb{E}P[\check{\boldsymbol{w}}_{c,i-1}]\\
&\quad + (\mu_{\max}h_c(\mu_{\max}) + \mu_{\max}^2\psi_0)\cdot\mathbb{1}^T\mathbb{E}P[\boldsymbol{w}_{e,i-1}]\\
&\quad + \mu_{\max}^2 b_{v,c}
\end{aligned}\tag{244}$$

$$\begin{aligned}
\mathbb{E}P[\boldsymbol{w}_{e,i}] &\preceq \mu_{\max}^2\psi_0\mathbb{1}\cdot\mathbb{E}P[\check{\boldsymbol{w}}_{c,i-1}]\\
&\quad + (\Gamma_e + \mu_{\max}^2\psi_0\mathbb{1}\mathbb{1}^T)\cdot\mathbb{E}P[\boldsymbol{w}_{e,i-1}]\\
&\quad + \mu_{\max}^2 b_{v,e}\cdot\mathbb{1}
\end{aligned}\tag{245}$$

where $\mathbb{E}P[\check{\boldsymbol{w}}_{c,i}] = \mathbb{E}\|\check{\boldsymbol{w}}_{c,i}\|^2$. Using the matrices and vectors defined in (116)–(118), we can write the above two recursions in a joint form as in (115).

Appendix I
Proof of Lemma 8

First, we establish the bound for $P[\boldsymbol{z}_{i-1}]$ in (238). Substituting (69) and (88) into the definition of $\boldsymbol{z}_{i-1}$ in (93) we get:

$$\begin{aligned}
&P[\boldsymbol{z}_{i-1}]\\
&\preceq P\big[s\left(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1} + (A_1^T U_L\otimes I_M)\boldsymbol{w}_{e,i-1}\right) - s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})\big]\\
&\overset{(a)}{\preceq} \lambda_U^2\cdot P\left[(A_1^T U_L\otimes I_M)\boldsymbol{w}_{e,i-1}\right]\\
&\overset{(b)}{\preceq} \lambda_U^2\cdot\left\|\bar{P}[\mathcal{A}_1^T\mathcal{U}_L]\right\|_\infty^2\cdot\mathbb{1}\mathbb{1}^T\cdot P[\boldsymbol{w}_{e,i-1}]
\end{aligned}\tag{246}$$

where step (a) uses the variance relation (163), and step (b) uses property (161).

Next, we prove the bound on $P[s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})]$. It holds that

$$\begin{aligned}
&P[s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1})]\\
&= P\Big[\frac{1}{3}\cdot 3\big(s(\mathbb{1}\otimes\boldsymbol{w}_{c,i-1}) - s(\mathbb{1}\otimes\bar{w}_{c,i-1})\big)
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{3}\cdot 3\big(s(\mathbb{1}\otimes\bar{w}_{c,i-1})-s(\mathbb{1}\otimes w^o)\big)\\
&+\frac{1}{3}\cdot 3\cdot s(\mathbb{1}\otimes w^o)\Big]\\
&\overset{(a)}{\preceq}\frac{1}{3}\cdot P\big[3\big(s(\mathbb{1}\otimes \boldsymbol{w}_{c,i-1})-s(\mathbb{1}\otimes\bar{w}_{c,i-1})\big)\big]\\
&+\frac{1}{3}\cdot P\big[3\big(s(\mathbb{1}\otimes\bar{w}_{c,i-1})-s(\mathbb{1}\otimes w^o)\big)\big]\\
&+\frac{1}{3}\cdot P\big[3\cdot s(\mathbb{1}\otimes w^o)\big]\\
&\overset{(b)}{=} 3P\big[s(\mathbb{1}\otimes \boldsymbol{w}_{c,i-1})-s(\mathbb{1}\otimes\bar{w}_{c,i-1})\big]\\
&+3P\big[s(\mathbb{1}\otimes\bar{w}_{c,i-1})-s(\mathbb{1}\otimes w^o)\big]+3P\big[s(\mathbb{1}\otimes w^o)\big]\\
&\overset{(c)}{\preceq} 3\lambda_U^2\cdot P\big[\mathbb{1}\otimes(\boldsymbol{w}_{c,i-1}-\bar{w}_{c,i-1})\big]\\
&+3\lambda_U^2\cdot P[\mathbb{1}\otimes(\bar{w}_{c,i-1}-w^o)]+3P[s(\mathbb{1}\otimes w^o)]\\
&\overset{(d)}{=} 3\lambda_U^2\cdot\|\check{\boldsymbol{w}}_{c,i-1}\|^2\cdot\mathbb{1}+3\lambda_U^2\cdot\|\bar{w}_{c,i-1}-w^o\|^2\cdot\mathbb{1}\\
&+3P[s(\mathbb{1}\otimes w^o)]\\
&\overset{(e)}{\preceq} 3\lambda_U^2\cdot\|\check{\boldsymbol{w}}_{c,i-1}\|^2\cdot\mathbb{1}+3\lambda_U^2\|\tilde{w}_{c,0}\|^2\cdot\mathbb{1}+3P[s(\mathbb{1}\otimes w^o)]
\end{aligned}
\tag{247}
$$

where step (a) uses the convexity property (148), step (b) uses the scaling property in Lemma 5, step (c) uses the variance relation (163), step (d) uses property (154), and step (e) uses the bound (110) and the fact that $\gamma_c<1$.

Finally, we establish the bounds on $P[\boldsymbol{v}_i]$ in (240)–(241). Introduce the $MN\times 1$ vector $\boldsymbol{x}$:

$$
\boldsymbol{x}\triangleq\mathbb{1}\otimes\boldsymbol{w}_{c,i-1}+\mathcal{A}_1^T\mathcal{U}_L\boldsymbol{w}_{e,i-1}\equiv\boldsymbol{\phi}_{i-1}
\tag{248}
$$

We partition $\boldsymbol{x}$ in block form as $\boldsymbol{x}=\mathrm{col}\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$, where each $\boldsymbol{x}_k$ is $M\times 1$. Then, by the definition of $\boldsymbol{v}_i$ from (93), we have

$$
\begin{aligned}
\mathbb{E}\{P[\boldsymbol{v}_i]|\mathcal{F}_{i-1}\}&=\mathbb{E}\{P[\hat{\boldsymbol{s}}_i(\boldsymbol{x})-s(\boldsymbol{x})]|\mathcal{F}_{i-1}\}\\
&=\mathrm{col}\big\{\mathbb{E}\left[\|\hat{\boldsymbol{s}}_{1,i}(\boldsymbol{x}_1)-s_1(\boldsymbol{x}_1)\|^2\big|\mathcal{F}_{i-1}\right],\\
&\qquad\ldots,\mathbb{E}\left[\|\hat{\boldsymbol{s}}_{N,i}(\boldsymbol{x}_N)-s_N(\boldsymbol{x}_N)\|^2\big|\mathcal{F}_{i-1}\right]\big\}\\
&\overset{(a)}{\preceq}\mathrm{col}\big\{\alpha\cdot\|\boldsymbol{x}_1\|^2+\sigma_v^2,\ \ldots,\ \alpha\cdot\|\boldsymbol{x}_N\|^2+\sigma_v^2\big\}\\
&=\alpha\cdot P[\boldsymbol{x}]+\sigma_v^2\mathbb{1}
\end{aligned}
\tag{249}
$$

where step (a) uses Assumption (18). Now we bound $P[\boldsymbol{x}]$:

$$
\begin{aligned}
P[\boldsymbol{x}]&=P\left[\mathbb{1}\otimes\boldsymbol{w}_{c,i-1}+\mathcal{A}_1^T\mathcal{U}_L\boldsymbol{w}_{e,i-1}\right]\\
&=P\Big[\frac{1}{4}\cdot 4\cdot\mathbb{1}\otimes(\boldsymbol{w}_{c,i-1}-\bar{w}_{c,i-1})+\frac{1}{4}\cdot 4\cdot\mathbb{1}\otimes(\bar{w}_{c,i-1}-w^o)\\
&\quad+\frac{1}{4}\cdot 4\cdot\mathcal{A}_1^T\mathcal{U}_L\boldsymbol{w}_{e,i-1}+\frac{1}{4}\cdot 4\cdot\mathbb{1}\otimes w^o\Big]\\
&=P\Big[\frac{1}{4}\cdot 4\cdot\mathbb{1}\otimes\check{\boldsymbol{w}}_{c,i-1}+\frac{1}{4}\cdot 4\cdot\mathbb{1}\otimes\tilde{w}_{c,i-1}\\
&\quad+\frac{1}{4}\cdot 4\cdot\mathcal{A}_1^T\mathcal{U}_L\boldsymbol{w}_{e,i-1}+\frac{1}{4}\cdot 4\cdot\mathbb{1}\otimes w^o\Big]\\
&\overset{(a)}{\preceq}\frac{1}{4}\cdot 4^2\cdot P[\mathbb{1}\otimes\check{\boldsymbol{w}}_{c,i-1}]+\frac{1}{4}\cdot 4^2\cdot P[\mathbb{1}\otimes\tilde{w}_{c,i-1}]\\
&\quad+\frac{1}{4}\cdot 4^2\cdot P[\mathcal{A}_1^T\mathcal{U}_L\boldsymbol{w}_{e,i-1}]+\frac{1}{4}\cdot 4^2\cdot P[\mathbb{1}\otimes w^o]\\
&\overset{(b)}{=}4\cdot\|\check{\boldsymbol{w}}_{c,i-1}\|^2\cdot\mathbb{1}+4\cdot\|\tilde{w}_{c,i-1}\|^2\cdot\mathbb{1}\\
&\quad+4\cdot P[\mathcal{A}_1^T\mathcal{U}_L\boldsymbol{w}_{e,i-1}]+4\cdot\|w^o\|^2\cdot\mathbb{1}\\
&\overset{(c)}{\preceq}4\cdot\|\check{\boldsymbol{w}}_{c,i-1}\|^2\cdot\mathbb{1}+4\cdot\|\bar{P}[\mathcal{A}_1^T\mathcal{U}_L]\|_\infty^2\cdot\mathbb{1}\mathbb{1}^T\cdot P[\boldsymbol{w}_{e,i-1}]\\
&\quad+4\cdot\|\tilde{w}_{c,0}\|^2\cdot\mathbb{1}+4\cdot\|w^o\|^2\cdot\mathbb{1}
\end{aligned}
\tag{250}
$$

where step (a) uses the convexity property (148) and the scaling property in Lemma 5, step (b) uses the Kronecker property (154), step (c) uses the variance relation (160) and the bound (110). Substituting (250) into (249), we obtain (240), and taking expectation of (240) with respect to $\mathcal{F}_{i-1}$ leads to (241).

APPENDIX J
PROOF OF THEOREM 4

Assume initially that the matrix Γ is stable (we show further ahead how the step-size parameter $\mu_{\max}$ can be selected to ensure this property). Then, we can iterate the inequality recursion (115) and obtain

$$
\begin{aligned}
\check{\mathcal{W}}_i'&\preceq\Gamma^i\check{\mathcal{W}}_0'+\mu_{\max}^2\sum_{j=0}^{i-1}\Gamma^j b_v\\
&\preceq\sum_{j=0}^{\infty}\Gamma^j\check{\mathcal{W}}_0'+\mu_{\max}^2\sum_{j=0}^{\infty}\Gamma^j b_v\\
&\preceq(I-\Gamma)^{-1}(\check{\mathcal{W}}_0'+\mu_{\max}^2 b_v)
\end{aligned}
\tag{251}
$$

where the first two inequalities use the fact that all entries of Γ are nonnegative. Moreover, substituting (116) into (115), we get

$$
\check{\mathcal{W}}_i'\preceq\Gamma_0\check{\mathcal{W}}_{i-1}'+\mu_{\max}^2\psi_0\mathbb{1}\mathbb{1}^T\check{\mathcal{W}}_{i-1}'+\mu_{\max}^2 b_v
\tag{252}
$$

Substituting (251) into the second term on the right-hand side of (252) leads to

$$
\check{\mathcal{W}}_i'\preceq\Gamma_0\check{\mathcal{W}}_{i-1}'+\mu_{\max}^2\cdot c_v(\mu_{\max})
\tag{253}
$$

where

$$
c_v(\mu_{\max})\triangleq\psi_0\cdot\mathbb{1}^T(I-\Gamma)^{-1}(\check{\mathcal{W}}_0'+\mu_{\max}^2 b_v)\cdot\mathbb{1}+b_v
\tag{254}
$$

Now iterating (253) leads to the following non-asymptotic bound:

$$
\check{\mathcal{W}}_i{}'\preceq\Gamma_0^i\check{\mathcal{W}}_0'+\sum_{j=0}^{i-1}\mu_{\max}^2\Gamma_0^j\cdot c_v(\mu_{\max})\preceq\Gamma_0^i\check{\mathcal{W}}_0'+\check{\mathcal{W}}_\infty^{\mathrm{ub}'}
\tag{255}
$$

where

$$
\check{\mathcal{W}}_\infty^{\mathrm{ub}'}\triangleq\mu_{\max}^2(I-\Gamma_0)^{-1}\cdot c_v(\mu_{\max})
\tag{256}
$$

We now derive the non-asymptotic bounds (124)–(125) from (255). To this end, we need to study the structure of the term $\Gamma_0^i\check{\mathcal{W}}_0'$. Our approach relies on applying the unilateral $z-$transform to the causal matrix sequence $\{\Gamma_0^i,i\geq 0\}$ to get

$$
\Gamma_0(z)\triangleq\mathcal{Z}\left\{\Gamma_0^i\right\}=z(zI-\Gamma_0)^{-1}
\tag{257}
$$

since Γ_0 is a stable matrix. Note from (117) that Γ_0 is a 2×2 block upper triangular matrix. Substituting (117) into the above expression and using the formula for inverting 2×2

block upper triangular matrices (see formula (4) in [58, p.48]), we obtain

$$\Gamma_0(z)=\begin{bmatrix}\frac{z}{z-\gamma_c} & \mu_{\max}h_c(\mu_{\max})\cdot\frac{z}{z-\gamma_c}\cdot\mathbb{1}^T(zI-\Gamma_e)^{-1}\\ 0 & z(zI-\Gamma_e)^{-1}\end{bmatrix} \tag{258}$$

Next we compute the inverse $z-$transform to obtain Γ_0^i. Thus, observe that the inverse $z-$transform of the $(1,1)$ entry, the $(2,1)$ block, and the $(2,2)$ block are the causal sequences γ_c^i, 0, and Γ_e^i, respectively. For the $(1,2)$ block, it can be expressed in partial fractions as

$$\begin{aligned}&\mu_{\max}h_c(\mu_{\max})\cdot\frac{z}{z-\gamma_c}\cdot\mathbb{1}^T(zI-\Gamma_e)^{-1}\\&=\mu_{\max}h_c(\mu_{\max})\cdot\mathbb{1}^T(\gamma_cI-\Gamma_e)^{-1}\left(\frac{z}{z-\gamma_c}I-z(zI-\Gamma_e)^{-1}\right)\end{aligned}$$

from which we conclude that the inverse $z-$transform of the $(1,2)$ block is

$$\mu_{\max}h_c(\mu_{\max})\cdot\mathbb{1}^T(\gamma_cI-\Gamma_e)^{-1}\left(\gamma_c^iI-\Gamma_e^i\right),\quad i\geq 0 \tag{259}$$

It follows that

$$\Gamma_0^i=\begin{bmatrix}\gamma_c^i & \mu_{\max}h_c(\mu_{\max})\cdot\mathbb{1}^T(\gamma_cI-\Gamma_e)^{-1}\left(\gamma_c^iI-\Gamma_e^i\right)\\ 0 & \Gamma_e^i\end{bmatrix} \tag{260}$$

Furthermore, as indicated by (48) in Sec. IV-A, the reference recursion (47) is initialized at the centroid of the network, i.e., $\bar{w}_{c,0}=\sum_{k=1}^N\theta_kw_{k,0}$. This fact, together with (68) leads to $\bar{w}_{c,0}=w_{c,0}$, which means that $\breve{\boldsymbol{w}}_{c,0}=0$. As a result, we get the following form for $\mathcal{W}_0'$:

$$\check{\mathcal{W}}_0'=\mathrm{col}\left\{0,\ \mathbb{E}P[\boldsymbol{w}_{e,0}]\right\} \tag{261}$$

Multiplying (260) to the left of (261) gives

$$\Gamma_0^i\check{\mathcal{W}}_0'=\begin{bmatrix}\mu_{\max}h_c(\mu_{\max})\cdot\mathbb{1}^T(\gamma_cI-\Gamma_e)^{-1}\left(\gamma_c^iI-\Gamma_e^i\right)\mathcal{W}_{e,0}\\ \Gamma_e^i\mathcal{W}_{e,0}\end{bmatrix} \tag{262}$$

where $\mathcal{W}_{e,0}=\mathbb{E}P[\boldsymbol{w}_{e,0}]$. Substituting (262) into (255), we obtain

$$\begin{aligned}\check{\mathcal{W}}_i'\preceq&\begin{bmatrix}\mu_{\max}h_c(\mu_{\max})\cdot\mathbb{1}^T(\gamma_cI-\Gamma_e)^{-1}\left(\gamma_c^iI-\Gamma_e^i\right)\mathbb{E}P[\boldsymbol{w}_{e,0}]\\ \Gamma_e^i\ \mathbb{E}P[\boldsymbol{w}_{e,0}]\end{bmatrix}\\&+\check{\mathcal{W}}_\infty^{\mathrm{ub}'}\end{aligned} \tag{263}$$

Finally, we study the behavior of the asymptotic bound $\check{\mathcal{W}}_\infty^{\mathrm{ub}'}$ by calling upon the following lemma.

Lemma 9 (Useful matrix expressions): It holds that

$$\begin{aligned}&\mathbb{1}^T(I-\Gamma)^{-1}\\&=\zeta(\mu_{\max})\\&\cdot\begin{bmatrix}\frac{\mu_{\max}^{-1}}{\lambda_L-\frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2} & \left(1+\frac{h_c(\mu_{\max})}{\lambda_L-\frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2}\right)\mathbb{1}^T(I-\Gamma_e)^{-1}\end{bmatrix}\end{aligned} \tag{264}$$

$$\begin{aligned}&(I-\Gamma_0)^{-1}\\&=\begin{bmatrix}\frac{\mu_{\max}^{-1}}{\lambda_L-\frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2} & \frac{h_c(\mu_{\max})}{\lambda_L-\frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2}\mathbb{1}^T(I-\Gamma_e)^{-1}\\ 0 & (I-\Gamma_e)^{-1}\end{bmatrix}\end{aligned} \tag{265}$$

where

$$\begin{aligned}\zeta(\mu_{\max})=&\Bigg\{1-\psi_0\cdot\Big[\frac{\mu_{\max}}{\lambda_L-\frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2}\\&+\mu_{\max}^2\left(1+\frac{h_c(\mu_{\max})}{\lambda_L-\frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2}\right)\mathbb{1}^T(I-\Gamma_e)^{-1}\mathbb{1}\Big]\Bigg\}^{-1}\end{aligned} \tag{266}$$

Proof: See Appendix K. ∎

Substituting (254), (261), (264) and (265) into (256) and after some algebra, we obtain

$$\begin{aligned}\check{\mathcal{W}}_\infty^{\mathrm{ub}'}=&\psi_0\cdot\zeta(\mu_{\max})f(\mu_{\max})\\&\cdot\begin{bmatrix}\mu_{\max}\frac{1+\mu_{\max}h_c(\mu_{\max})\mathbb{1}^T(I-\Gamma_e)^{-1}\mathbb{1}}{\lambda_L-\frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2}\\ \mu_{\max}^2\cdot(I-\Gamma_e)^{-1}\mathbb{1}\end{bmatrix}\\&+\begin{bmatrix}\mu_{\max}\frac{b_{v,c}+\mu_{\max}h_c(\mu_{\max})\mathbb{1}^T(I-\Gamma_e)^{-1}\mathbb{1}b_{v,e}}{\lambda_L-\frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2}\\ \mu_{\max}^2b_{v,e}\cdot(I-\Gamma_e)^{-1}\mathbb{1}\end{bmatrix}\end{aligned} \tag{267}$$

where

$$\begin{aligned}f(\mu_{\max})\triangleq&\frac{\mu_{\max}b_{v,c}}{\lambda_L-\frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2}\\&+\left(1+\frac{h_c(\mu_{\max})}{\lambda_L-\frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2}\right)\cdot\mathbb{1}^T(I-\Gamma_e)^{-1}\\&\times\left(\mathbb{E}P[\boldsymbol{w}_{e,0}]+\mu_{\max}^2\mathbb{1}b_{v,e}\right)\end{aligned} \tag{268}$$

Introduce

$$\begin{aligned}\check{\mathcal{W}}_{c,\infty}^{\mathrm{ub}'}\triangleq&\psi_0\cdot\zeta(\mu_{\max})f(\mu_{\max})\\&\cdot\mu_{\max}\frac{1+\mu_{\max}h_c(\mu_{\max})\mathbb{1}^T(I-\Gamma_e)^{-1}\mathbb{1}}{\lambda_L-\frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2}\\&+\mu_{\max}\frac{b_{v,c}+\mu_{\max}h_c(\mu_{\max})\mathbb{1}^T(I-\Gamma_e)^{-1}\mathbb{1}b_{v,e}}{\lambda_L-\frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2}\end{aligned} \tag{269}$$

$$\begin{aligned}\check{\mathcal{W}}_{e,\infty}^{\mathrm{ub}'}\triangleq&\psi_0\cdot\zeta(\mu_{\max})f(\mu_{\max})\cdot\mu_{\max}^2\cdot(I-\Gamma_e)^{-1}\mathbb{1}\\&+\mu_{\max}^2b_{v,e}\cdot(I-\Gamma_e)^{-1}\mathbb{1}\end{aligned} \tag{270}$$

Then, we have

$$\check{\mathcal{W}}_\infty^{\mathrm{ub}'}=\mathrm{col}\{\check{\mathcal{W}}_{c,\infty}^{\mathrm{ub}'},\ \check{\mathcal{W}}_{e,\infty}^{\mathrm{ub}'}\} \tag{271}$$

Substituting (271) into (263), we conclude (124)–(125). Now, to prove (126)–(127), it suffices to prove

$$\lim_{\mu_{\max}\to 0}\frac{\check{\mathcal{W}}_{c,\infty}^{\mathrm{ub}'}}{\mu_{\max}}=\frac{\psi_0\big(\lambda_L+h_c(0)\big)\mathbb{1}^T(I-\Gamma_e)^{-1}\mathcal{W}_{e,0}+b_{v,c}\lambda_L}{\lambda_L^2} \tag{272}$$

$$\begin{aligned}\lim_{\mu_{\max}\to 0}\frac{\check{\mathcal{W}}_{e,\infty}^{\mathrm{ub}'}}{\mu_{\max}^2}=&\frac{\psi_0\big(\lambda_L+h_c(0)\big)\mathbb{1}^T(I-\Gamma_e)^{-1}\mathcal{W}_{e,0}+b_{v,e}\lambda_L}{\lambda_L}\\&\cdot(I-\Gamma_e)^{-1}\mathbb{1}\end{aligned} \tag{273}$$

Substituting (269) and (270) into the left-hand side of (272) and (273), respectively, we get

$$\begin{aligned}&\lim_{\mu_{\max}\to 0}\frac{\check{\mathcal{W}}_{c,\infty}^{\mathrm{ub}'}}{\mu_{\max}}\\&=\lim_{\mu_{\max}\to 0}\Bigg\{\\&\qquad\psi_0\cdot\zeta(\mu_{\max})f(\mu_{\max})\cdot\frac{1+\mu_{\max}h_c(\mu_{\max})\mathbb{1}^T(I-\Gamma_e)^{-1}\mathbb{1}}{\lambda_L-\frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2}\end{aligned}$$

$$
\begin{aligned}
&+\frac{b_{v,c}+\mu_{\max}h_c(\mu_{\max})\mathbb{1}^T(I-\Gamma_e)^{-1}\mathbb{1}b_{v,e}}{\lambda_L-\frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2}\Big\}\\
&=\psi_0\cdot\zeta(0)f(0)\cdot\frac{1}{\lambda_L}+\frac{b_{v,c}}{\lambda_L}\\
&\overset{(a)}{=}\psi_0\cdot 1\cdot\left[\left(1+\frac{h_c(0)}{\lambda_L}\right)\cdot\mathbb{1}^T(I-\Gamma_e)^{-1}\mathbb{E}P[\boldsymbol{w}_{e,0}]\right]\cdot\frac{1}{\lambda_L}+\frac{b_{v,c}}{\lambda_L}\\
&=\frac{\psi_0\big(\lambda_L+h_c(0)\big)\mathbb{1}^T(I-\Gamma_e)^{-1}\mathcal{W}_{e,0}+b_{v,c}\lambda_L}{\lambda_L^2} \qquad (274)
\end{aligned}
$$

$$
\begin{aligned}
&\lim_{\mu_{\max}\to 0}\frac{\check{\mathcal{W}}_{e,\infty}^{\mathrm{ub}'}}{\mu_{\max}^2}\\
&=\lim_{\mu_{\max}\to 0}\big\{\psi_0\cdot\zeta(\mu_{\max})f(\mu_{\max})\cdot(I-\Gamma_e)^{-1}\mathbb{1}\\
&\qquad+b_{v,e}\cdot(I-\Gamma_e)^{-1}\mathbb{1}\big\}\\
&=\psi_0\cdot\zeta(0)f(0)\cdot(I-\Gamma_e)^{-1}\mathbb{1}+b_{v,e}\cdot(I-\Gamma_e)^{-1}\mathbb{1}\\
&\overset{(b)}{=}\psi_0\cdot\left[\left(1+\frac{h_c(0)}{\lambda_L}\right)\cdot\mathbb{1}^T(I-\Gamma_e)^{-1}\mathbb{E}P[\boldsymbol{w}_{e,0}]\right]\cdot(I-\Gamma_e)^{-1}\mathbb{1}\\
&\quad+b_{v,e}\cdot(I-\Gamma_e)^{-1}\mathbb{1}\\
&=\frac{\psi_0\big(\lambda_L+h_c(0)\big)\mathbb{1}^T(I-\Gamma_e)^{-1}\mathcal{W}_{e,0}+b_{v,e}\lambda_L}{\lambda_L}\cdot(I-\Gamma_e)^{-1}\mathbb{1} \qquad (275)
\end{aligned}
$$

where steps (a) and (b) use the expressions for $\zeta(\mu_{\max})$ and $f(\mu_{\max})$ from (266) and (268).

Now we proceed to prove (128). We already know that the second term on the right-hand side of (124), $\check{\mathcal{W}}_{c,\infty}^{\mathrm{ub}'}$, is $O(\mu_{\max})$ because of (126). Therefore, we only need to show that the first term on the right-hand side of (124) is $O(\mu_{\max})$ for all $i\geq 0$. To this end, it suffices to prove that

$$
\begin{aligned}
&\lim_{\mu_{\max}\to 0}\frac{\left\|\mu_{\max}h_c(\mu_{\max})\cdot\mathbb{1}^T(\gamma_c I-\Gamma_e)^{-1}\left(\gamma_c^i I-\Gamma_e^i\right)\mathcal{W}_{e,0}\right\|}{\mu_{\max}}\\
&\qquad\leq\text{constant} \qquad (276)
\end{aligned}
$$

where the constant on the right-hand side should be independent of i. This can be proved as below:

$$
\begin{aligned}
&\lim_{\mu_{\max}\to 0}\frac{\left\|\mu_{\max}h_c(\mu_{\max})\cdot\mathbb{1}^T(\gamma_c I-\Gamma_e)^{-1}\left(\gamma_c^i I-\Gamma_e^i\right)\mathcal{W}_{e,0}\right\|}{\mu_{\max}}\\
&=\lim_{\mu_{\max}\to 0}\left\|h_c(\mu_{\max})\cdot\mathbb{1}^T(\gamma_c I-\Gamma_e)^{-1}\left(\gamma_c^i I-\Gamma_e^i\right)\mathcal{W}_{e,0}\right\|\\
&\leq\lim_{\mu_{\max}\to 0}|h_c(\mu_{\max})|\cdot\|\mathbb{1}\|\cdot\|(\gamma_c I-\Gamma_e)^{-1}\|\\
&\quad\cdot\left(\|\gamma_c^i I\|+\|\Gamma_e^i\|\right)\cdot\|\mathcal{W}_{e,0}\|\\
&\overset{(a)}{\leq}\lim_{\mu_{\max}\to 0}|h_c(\mu_{\max})|\cdot N\cdot\|(\gamma_c I-\Gamma_e)^{-1}\|\\
&\quad\cdot\left(1+C_e\cdot\big(\rho(\Gamma_e)+\epsilon\big)^i\right)\cdot\|\mathcal{W}_{e,0}\|\\
&\overset{(b)}{\leq}\lim_{\mu_{\max}\to 0}|h_c(\mu_{\max})|\cdot N\cdot\|(\gamma_c I-\Gamma_e)^{-1}\|\cdot(1+C_e)\cdot\|\mathcal{W}_{e,0}\|\\
&\overset{(c)}{=}|h_c(0)|\cdot N\cdot\|(I-\Gamma_e)^{-1}\|\cdot[1+C_e]\cdot\|\mathcal{W}_{e,0}\|\\
&=\text{constant} \qquad (277)
\end{aligned}
$$

where step (a) uses $\gamma_c=1-\mu_{\max}\lambda_L+\frac{1}{2}\mu_{\max}^2\lambda_L^2<1$ for sufficiently small step-sizes, and uses the property that for any small $\epsilon>0$ there exists a constant C such that $\|X^i\|\leq C\cdot[\rho(X)+\epsilon]^i$ for all $i\geq 0$ [49, p.38], step (b) uses the fact that $\rho(\Gamma_e)=|\lambda_2(A)|<1$ so that $\rho(\Gamma_e)+\epsilon<1$ for small ϵ (e.g., $\epsilon=(1-\rho(\Gamma_e))/2$), and step (c) uses $\gamma_c=1-\mu_{\max}\lambda_L+\frac{1}{2}\mu_{\max}^2\lambda_L^2\to 1$ when $\mu_{\max}\to 0$.

It remains to prove that condition (129) guarantees the stability of the matrix Γ, i.e., $\rho(\Gamma)<1$. First, we introduce the diagonal matrices $D_{\epsilon,0}\triangleq\mathrm{diag}\{\epsilon,\cdots,\epsilon^{N-1}\}$ and $D_\epsilon=\mathrm{diag}\{1,D_{\epsilon,0}\}$, where ϵ is chosen to be

$$
\epsilon\triangleq\frac{1}{4}(1-|\lambda_2(A)|)^2\leq\frac{1}{4} \qquad (278)
$$

It holds that $\rho(\Gamma)=\rho(D_\epsilon^{-1}\Gamma D_\epsilon)$ since similarity transformations do not alter eigenvalues. By the definition of Γ in (116), we have

$$
D_\epsilon^{-1}\Gamma D_\epsilon=D_\epsilon^{-1}\Gamma_0 D_\epsilon+\mu_{\max}^2\psi_0\cdot D_\epsilon^{-1}\mathbb{1}\mathbb{1}^T D_\epsilon \qquad (279)
$$

We now recall that the spectral radius of a matrix is upper bounded by any of its matrix norms. Thus, taking the $1-$norm (the maximum absolute column sum of the matrix) of both sides of the above expression and using the triangle inequality and the fact that $0<\epsilon\leq 1/4$, we get

$$
\begin{aligned}
\rho(\Gamma)&=\rho(D_\epsilon^{-1}\Gamma D_\epsilon)\\
&\leq\|D_\epsilon^{-1}\Gamma_0 D_\epsilon\|_1+\|\mu_{\max}^2\psi_0\cdot D_\epsilon^{-1}\mathbb{1}\mathbb{1}^T D_\epsilon\|_1\\
&\leq\|D_\epsilon^{-1}\Gamma_0 D_\epsilon\|_1+\mu_{\max}^2\psi_0\cdot\|D_\epsilon^{-1}\mathbb{1}\mathbb{1}^T D_\epsilon\|_1\\
&=\|D_\epsilon^{-1}\Gamma_0 D_\epsilon\|_1+\mu_{\max}^2\psi_0\cdot\left(1+\epsilon^{-1}+\cdots+\epsilon^{-(N-1)}\right)\\
&=\|D_\epsilon^{-1}\Gamma_0 D_\epsilon\|_1+\mu_{\max}^2\psi_0\cdot\frac{1-\epsilon^{-N}}{1-\epsilon^{-1}}\\
&=\|D_\epsilon^{-1}\Gamma_0 D_\epsilon\|_1+\mu_{\max}^2\psi_0\cdot\frac{\epsilon(\epsilon^{-N}-1)}{1-\epsilon}\\
&\leq\|D_\epsilon^{-1}\Gamma_0 D_\epsilon\|_1+\mu_{\max}^2\psi_0\cdot\frac{\frac{1}{4}(\epsilon^{-N}-1)}{1-\frac{1}{4}}\\
&\leq\|D_\epsilon^{-1}\Gamma_0 D_\epsilon\|_1+\frac{1}{3}\mu_{\max}^2\psi_0\epsilon^{-N} \qquad (280)
\end{aligned}
$$

Moreover, we can use (117) to write:

$$
D_\epsilon^{-1}\Gamma_0 D_\epsilon=\begin{bmatrix}\gamma_c & \mu_{\max}h_c(\mu_{\max})\mathbb{1}^T D_{\epsilon,0}\\ 0 & D_{\epsilon,0}^{-1}\Gamma_e D_{\epsilon,0}\end{bmatrix} \qquad (281)
$$

where (recall the expression for Γ_e from (171) where we replace d_2 by $\lambda_2(A)$):

$$
D_{\epsilon,0}^{-1}\Gamma_e D_{\epsilon,0}=\begin{bmatrix}|\lambda_2(A)| & \frac{1-|\lambda_2(A)|}{2} & & \\ & \ddots & \ddots & \\ & & \ddots & \frac{1-|\lambda_2(A)|}{2}\\ & & & |\lambda_2(A)|\end{bmatrix} \qquad (282)
$$

$$
\mu_{\max}h_c(\mu_{\max})\mathbb{1}^T D_{\epsilon,0}=\mu_{\max}h_c(\mu_{\max})\begin{bmatrix}\epsilon & \cdots & \epsilon^{N-1}\end{bmatrix} \qquad (283)
$$

Therefore, the 1-norm of $D_\epsilon^{-1}\Gamma_0 D_\epsilon$ can be evaluated as

$$
\begin{aligned}
\|D_{\epsilon,0}^{-1}\Gamma_0 D_{\epsilon,0}\|_1=\max\Big\{&\gamma_c,\ |\lambda_2(A)|+\mu_{\max}h_c(\mu_{\max})\epsilon,\\
&\frac{1+|\lambda_2(A)|}{2}+\mu_{\max}h_c(\mu_{\max})\epsilon^2,\cdots,\\
&\frac{1+|\lambda_2(A)|}{2}+\mu_{\max}h_c(\mu_{\max})\epsilon^{N-1}\Big\}
\end{aligned} \qquad (284)
$$

Since $0 < \epsilon \leq 1/4$, we have $\epsilon > \epsilon^2 > \cdots > \epsilon^{N-1} > 0$. Therefore,

$$\|D_{\epsilon,0}^{-1}\Gamma_0 D_{\epsilon,0}\|_1 = \max\Big\{\gamma_c,\ |\lambda_2(A)| + \mu_{\max}h_c(\mu)\epsilon, \frac{1+|\lambda_2(A)|}{2} + \mu_{\max}h_c(\mu_{\max})\epsilon^2\Big\} \quad (285)$$

Substituting the above expression for $\|D_{\epsilon,0}^{-1}\Gamma_0 D_{\epsilon,0}\|_1$ into (280) leads to

$$\rho(\Gamma) \leq \max\Big\{\gamma_c,\ |\lambda_2(A)| + \mu_{\max}h_c(\mu_{\max})\epsilon, \frac{1+|\lambda_2(A)|}{2} + \mu_{\max}h_c(\mu_{\max})\epsilon^2\Big\} + \frac{1}{3}\mu_{\max}^2\psi_0\epsilon^{-N} \quad (286)$$

We recall from (167) that $\gamma_c > 0$. To ensure $\rho(\Gamma) < 1$, it suffices to require that $\mu_{\max}$ is such that the following conditions are satisfied:

$$\gamma_c + \frac{1}{3}\mu_{\max}^2\psi_0\epsilon^{-N} < 1 \quad (287)$$

$$|\lambda_2(A)| + \mu_{\max}h_c(\mu_{\max})\epsilon + \frac{1}{3}\mu_{\max}^2\psi_0\epsilon^{-N} < 1 \quad (288)$$

$$\frac{1+|\lambda_2(A)|}{2} + \mu_{\max}h_c(\mu_{\max})\epsilon^2 + \frac{1}{3}\mu_{\max}^2\psi_0\epsilon^{-N} < 1 \quad (289)$$

We now solve these three inequalities to get a condition on $\mu_{\max}$. Substituting the expression for γ_c from (166) into (287), we get

$$1 - \mu_{\max}\lambda_L + \mu_{\max}^2\Big(\frac{1}{3}\psi_0\epsilon^{-N} + \frac{1}{2}\|p\|_1^2\lambda_U^2\Big) < 1 \quad (290)$$

the solution of which is given by

$$0 < \mu_{\max} < \frac{\lambda_L}{\frac{1}{3}\psi_0\epsilon^{-N} + \frac{1}{2}\|p\|_1^2\lambda_U^2} \quad (291)$$

For (288)–(289), if we substitute the expression for $h_c(\mu_{\max})$ from (121) into (288)–(289), we get a third-order inequality in $\mu_{\max}$, which is difficult to solve in closed-form. However, inequalities (288)–(289) can be guaranteed by the following conditions:

$$\mu_{\max}h_c(\mu_{\max})\epsilon < \frac{(1-|\lambda_2(A)|)^2}{4},\ \frac{\mu_{\max}^2\psi_0\epsilon^{-N}}{3} < \frac{1-|\lambda_2(A)|}{4} \quad (292)$$

This is because we would then have:

$$\begin{aligned}|\lambda_2(A)| + \mu_{\max}h_c(\mu_{\max})\epsilon + \frac{1}{3}\mu_{\max}^2\psi_0\epsilon^{-N} &< |\lambda_2(A)| + \frac{(1-|\lambda_2(A)|)^2}{4} + \frac{1-|\lambda_2(A)|}{4}\\ &\leq |\lambda_2(A)| + \frac{1-|\lambda_2(A)|}{4} + \frac{1-|\lambda_2(A)|}{4}\\ &= \frac{1+|\lambda_2(A)|}{2} < 1\end{aligned} \quad (293)$$

Likewise, by the fact that $0 < \epsilon \leq 1/4 < 1$,

$$\begin{aligned}\frac{1+|\lambda_2(A)|}{2} + \mu_{\max}h_c(\mu_{\max})\epsilon^2 + \frac{1}{3}\mu_{\max}^2\psi_0\epsilon^{-N} &< \frac{1+|\lambda_2(A)|}{2} + \mu_{\max}h_c(\mu_{\max})\epsilon + \frac{1}{3}\mu_{\max}^2\psi_0\epsilon^{-N}\\ &< \frac{1+|\lambda_2(A)|}{2} + \frac{(1-|\lambda_2(A)|)^2}{4} + \frac{1-|\lambda_2(A)|}{4}\\ &\leq \frac{1+|\lambda_2(A)|}{2} + \frac{1-|\lambda_2(A)|}{4} + \frac{1-|\lambda_2(A)|}{4}\\ &= 1\end{aligned} \quad (294)$$

Substituting (121) and (278) into (292), we find that the latter conditions are satisfied for

$$\begin{cases}0 < \mu_{\max} < \dfrac{\lambda_L}{\|p\|_1^2\lambda_U^2\big(\|\bar{P}[\mathcal{A}^T\mathcal{U}_2]\|_\infty^2 + \frac{1}{2}\big)}\\[2ex] 0 < \mu_{\max} < \sqrt{\dfrac{3(1-|\lambda_2(A)|)^{2N+1}}{2^{2N+2}\psi_0}}\end{cases} \quad (295)$$

Combining (287), (289) and (295) , we arrive at condition (129).

APPENDIX K
PROOF OF LEMMA 9

Applying the matrix inversion lemma [52] to (116), we get

$$\begin{aligned}(I-\Gamma)^{-1} &= (I - \Gamma_0 - \mu_{\max}^2\psi_0\cdot\mathbb{1}\mathbb{1}^T)^{-1}\\ &= (I-\Gamma_0)^{-1} + \frac{\mu_{\max}^2\psi_0\cdot(I-\Gamma_0)^{-1}\mathbb{1}\mathbb{1}^T(I-\Gamma_0)^{-1}}{1-\mu_{\max}^2\psi_0\cdot\mathbb{1}^T(I-\Gamma_0)^{-1}\mathbb{1}}\end{aligned} \quad (296)$$

so that

$$\mathbb{1}^T(I-\Gamma)^{-1} = \frac{1}{1-\mu_{\max}^2\psi_0\cdot\mathbb{1}^T(I-\Gamma_0)^{-1}\mathbb{1}}\cdot\mathbb{1}^T(I-\Gamma_0)^{-1} \quad (297)$$

By (117), the matrix Γ_0 is a 2×2 block upper triangular matrix whose inverse is given by

$$(I-\Gamma_0)^{-1} = \begin{bmatrix}(1-\gamma_c)^{-1} & \frac{\mu_{\max}h_c(\mu_{\max})}{1-\gamma_c}\mathbb{1}^T(I-\Gamma_e)^{-1})\\ 0 & (I-\Gamma_e)^{-1}\end{bmatrix} \quad (298)$$

Substituting (166) into the above expression leads to (265). Furthermore, from (265), we have

$$\begin{aligned}\mathbb{1}^T(I-\Gamma_0)^{-1}\mathbb{1} &= \frac{\mu_{\max}^{-1}}{\lambda_L - \frac{\mu_{\max}}{2}\|p\|_1^2\lambda_U^2}\\ &\quad + \left(1 + \frac{h_c(\mu_{\max})}{\lambda_L - \frac{\mu}{2}\|p\|_1^2\lambda_U^2}\right)\mathbb{1}^T(I-\Gamma_e)^{-1}\mathbb{1}\end{aligned} \quad (299)$$

Substituting (299) into (297), we obtain (264).

APPENDIX L
PROOF OF THEOREM 5

Taking the squared Euclidean norm of both sides of (74) and applying the expectation operator, we obtain

$$\begin{aligned}\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 &= \|\tilde{w}_{c,i}\|^2 + \mathbb{E}\,\|\check{\boldsymbol{w}}_{c,i} + (u_{L,k}\otimes I_M)\boldsymbol{w}_{e,i}\|^2\\ &\quad - 2\tilde{w}_{c,i}^T\left[\mathbb{E}\check{\boldsymbol{w}}_{c,i} + (u_{L,k}\otimes I_M)\boldsymbol{w}_{e,i}\right]\end{aligned} \quad (300)$$

which means that, for all $i \geq 0$,

$$\begin{aligned}&\big|\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 - \|\tilde{w}_{c,i}\|^2\big|\\ &= \Big|\mathbb{E}\,\|\check{\boldsymbol{w}}_{c,i} + (u_{L,k}\otimes I_M)\boldsymbol{w}_{e,i}\|^2\end{aligned}$$

$$
\begin{aligned}
&\quad - 2\tilde{w}_{c,i}^T\big[\mathbb{E}\check{\boldsymbol{w}}_{c,i} + (u_{L,k}\otimes I_M)\mathbb{E}\boldsymbol{w}_{e,i}\big]\Big| \\
&\overset{(a)}{\leq} \mathbb{E}\,\|\check{\boldsymbol{w}}_{c,i} + (u_{L,k}\otimes I_M)\boldsymbol{w}_{e,i}\|^2 \\
&\quad + 2\|\tilde{w}_{c,i}\|\cdot\big[\,\|\mathbb{E}\check{\boldsymbol{w}}_{c,i}\| + \|u_{L,k}\otimes I_M\|\cdot\|\mathbb{E}\boldsymbol{w}_{e,i}\|\big] \\
&\overset{(b)}{\leq} 2\mathbb{E}\|\check{\boldsymbol{w}}_{c,i}\|^2 + 2\|u_{L,k}\otimes I_M\|^2\cdot\mathbb{E}\|\boldsymbol{w}_{e,i}\|^2 \\
&\quad + 2\|\tilde{w}_{c,i}\|\cdot\big[\mathbb{E}\|\check{\boldsymbol{w}}_{c,i}\| + \|u_{L,k}\otimes I_M\|\cdot\mathbb{E}\|\boldsymbol{w}_{e,i}\|\big] \\
&\overset{(c)}{\leq} 2\mathbb{E}\|\check{\boldsymbol{w}}_{c,i}\|^2 + 2\|u_{L,k}\otimes I_M\|^2\cdot\mathbb{E}\|\boldsymbol{w}_{e,i}\|^2 \\
&\quad + 2\|\tilde{w}_{c,i}\|\cdot\big[\sqrt{\mathbb{E}\|\check{\boldsymbol{w}}_{c,i}\|^2} + \|u_{L,k}\otimes I_M\|\cdot\sqrt{\mathbb{E}\|\boldsymbol{w}_{e,i}\|^2}\big] \\
&\overset{(d)}{=} 2\mathbb{E}P[\check{\boldsymbol{w}}_{c,i}] + 2\|u_{L,k}\otimes I_M\|^2\cdot\mathbb{1}^T\mathbb{E}P[\boldsymbol{w}_{e,i}] \\
&\quad + 2\|\tilde{w}_{c,i}\|\cdot\Big[\sqrt{\mathbb{E}P[\check{\boldsymbol{w}}_{c,i}]} + \|u_{L,k}\otimes I_M\|\cdot\sqrt{\mathbb{1}^T\mathbb{E}P[\boldsymbol{w}_{e,i}]}\Big] \\
&\overset{(e)}{\leq} O(\mu_{\max}) + 2\|u_{L,k}\otimes I_M\|^2\cdot\big(\mathbb{1}^T\Gamma_e^i\mathcal{W}_{e,0} + \mathbb{1}^T\check{\mathcal{W}}_{e,\infty}^{\mathrm{ub}'}\big) \\
&\quad + 2\gamma_c^i\|\tilde{w}_{c,0}\|\cdot\Big[O(\mu_{\max}^{\frac{1}{2}}) \\
&\qquad + \|u_{L,k}\otimes I_M\|\cdot\sqrt{\mathbb{1}^T\Gamma_e^i\mathcal{W}_{e,0} + \mathbb{1}^T\check{\mathcal{W}}_{e,\infty}^{\mathrm{ub}'}}\Big] \\
&\overset{(f)}{\leq} O(\mu_{\max}) + 2\|u_{L,k}\otimes I_M\|^2\cdot\big(\mathbb{1}^T\Gamma_e^i\mathcal{W}_{e,0} + O(\mu_{\max}^2)\big) \\
&\quad + 2\gamma_c^i\|\tilde{w}_{c,0}\|\cdot\Bigg[O(\mu_{\max}^{\frac{1}{2}}) \\
&\qquad + \|u_{L,k}\otimes I_M\|\cdot\Big(\sqrt{\mathbb{1}^T\Gamma_e^i\mathcal{W}_{e,0}} + \sqrt{O(\mu_{\max}^2)}\Big)\Bigg] \\
&= 2\|u_{L,k}\otimes I_M\|^2\cdot\mathbb{1}^T\Gamma_e^i\mathcal{W}_{e,0} \\
&\qquad + 2\gamma_c^i\cdot\|\tilde{w}_{c,0}\|\cdot\|u_{L,k}\otimes I_M\|\cdot\sqrt{\mathbb{1}^T\Gamma_e^i\mathcal{W}_{e,0}} \\
&\quad + 2\gamma_c^i\|\tilde{w}_{c,0}\|\cdot\Big[O(\mu_{\max}^{\frac{1}{2}}) + \|u_{L,k}\otimes I_M\|\cdot O(\mu_{\max})\Big] \\
&\quad + O(\mu_{\max}) + O(\mu_{\max}^2) \\
&\overset{(g)}{\leq} 2\|u_{L,k}\otimes I_M\|^2\cdot\mathbb{1}^T\Gamma_e^i\mathcal{W}_{e,0} \\
&\quad + 2\|\tilde{w}_{c,0}\|\cdot\|u_{L,k}\otimes I_M\|\cdot\sqrt{\mathbb{1}^T\Gamma_e^i\mathcal{W}_{e,0}} \\
&\quad + \gamma_c^i\cdot O(\mu_{\max}^{\frac{1}{2}}) + O(\mu_{\max})
\end{aligned}
\tag{301}
$$

where step (a) used Cauchy-Schwartz inequality, step (b) used $\|x+y\|^2 \leq 2\|x\|^2+2\|y\|^2$, step (c) applied Jensen's inequality to the concave function $\sqrt{\cdot}$, step (d) used property (157), step (e) substituted the non-asymptotic bounds (124) and (125) and the fact that $\mathbb{E}P[\check{\boldsymbol{w}}_{c,i}] \leq O(\mu_{\max})$ for *all* $i \geq 0$ from (128), step (f) used (127) and the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x, y \geq 0$, and step (g) used $\gamma_c < 1$ for sufficiently small step-sizes (guaranteed by (112)).

REFERENCES

[1] J. Chen and A. H. Sayed, "On the limiting behavior of distributed optimization strategies," in *Proc. Allerton Conf.*, Monticello, IL, Oct. 2012, pp. 1535–1542.

[2] S. Barbarossa and G. Scutari, "Bio-inspired sensor network design," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 26–35, May 2007.

[3] L. Li and J. A. Chambers, "A new incremental affine projection-based adaptive algorithm for distributed networks," *Signal Processing*, vol. 88, no. 10, pp. 2599–2603, Oct. 2008.

[4] A. Nedic and D. P. Bertsekas, "Incremental subgradient methods for nondifferentiable optimization," *SIAM J. Optim.*, vol. 12, no. 1, pp. 109–138, 2001.

[5] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, no. 9, pp. 803–812, 1986.

[6] S. Kar and J. M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE J. Sel. Topics. Signal Process.*, vol. 5, no. 4, pp. 674–690, Aug. 2011.

[7] S. Kar, J. M. F. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3575–3605, Jun. 2012.

[8] S. Kar, J. M. F. Moura, and H. V. Poor, "Distributed linear parameter estimation: Asymptotically efficient adaptive strategies," *SIAM Journal on Control and Optimization*, vol. 51, no. 3, pp. 2200–2229, 2013.

[9] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1864, Nov. 2010.

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

[11] A. Nedic and A. Ozdaglar, "Cooperative distributed multi-agent optimization," *Convex Optimization in Signal Processing and Communications,* Y. Eldar and D. Palomar (Eds.), Cambridge University Press, pp. 340–386, 2010.

[12] C. Eksin and A. Ribeiro, "Distributed network optimization with heuristic rational agents," *IEEE Trans. Signal Proc.*, vol. 60, no. 10, pp. 5396–5411, Oct. 2012.

[13] C. Eksin, P. Molavi, A. Ribeiro, and A. Jadbabaie, "Learning in network games with incomplete information: Asymptotic analysis and tractable implementation of rational behavior," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 30–42, May 2013.

[14] S. Theodoridis, K. Slavakis, and I. Yamada, "Adaptive learning in a world of projections," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 97–123, Jan. 2011.

[15] D. H. Dini and D. P. Mandic, "Cooperative adaptive estimation of distributed noncircular complex signals," in *Proc. Asilomar Conf. Signals, Syst. and Comput.*, Pacific Grove, CA, Nov. 2012, pp. 1518–1522.

[16] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, Jul. 2008.

[17] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, Mar. 2010.

[18] J. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4289–4305, Aug. 2012.

[19] X. Zhao and A. H. Sayed, "Performance limits for distributed estimation over LMS adaptive networks," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5107–5124, Oct. 2012.

[20] J. Chen and A. H. Sayed, "Distributed Pareto optimization via diffusion adaptation," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 2, pp. 205–220, Apr. 2013.

[21] A. H. Sayed, "Diffusion adaptation over networks," *in* Academic Press Library in Signal Processing, vol. 3, R. Chellapa and S. Theodoridis, *editors,* pp. 323–454, Elsevier, 2014.

[22] A. H. Sayed, "Adaptive networks," *Proc. IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.

[23] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning,* vol. 7, issue 4–5, NOW Publishers, Boston-Delft, Jul. 2014., pp. 311–801.

[24] J. Chen, Z. J. Towfic, and A. H. Sayed, "Dictionary learning over distributed models," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 1001–1016, Feb. 2015.

[25] S. V. Macua, J. Chen, S. Zazo, and A. H. Sayed, "Distributed policy evaluation under multiple behavior strategies," *IEEE Trans. Autom. Control*, vol. 60, no. 5, May 2015.

[26] S. Chouvardas, K. Slavakis, and S. Theodoridis, "Adaptive robust distributed learning in diffusion sensor networks," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4692–4707, Oct. 2011.

[27] O. N. Gharehshiran, V. Krishnamurthy, and G. Yin, "Distributed energy-aware diffusion least mean squares: Game-theoretic learning," *IEEE Journal Sel. Topics Signal Process.*, vol. 7, no. 5, pp. 821–836, Jun. 2013.

[28] S. S. Ram, A. Nedic, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *J. Optim. Theory Appl.*, vol. 147, no. 3, pp. 516–545, 2010.

[29] K. Srivastava and A. Nedic, "Distributed asynchronous constrained stochastic optimization," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 772–790, Aug. 2011.

[30] S. Lee and A. Nedic, "Distributed random projection algorithm for convex optimization," *IEEE Journal Sel. Topics Signal Process.*, vol. 7, no. 2, pp. 221–229, Apr. 2013.

[31] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, "Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning," in *Proc. Annual Allerton Conference on Commun. Control and Comput.*, Monticello, IL, Oct. 2012, pp. 1543–1550.

[32] D.P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

[33] V. Saligrama, M. Alanyali, and O. Savas, "Distributed detection in sensor networks with packet losses and finite capacity links," *IEEE Trans. Signal Proc.*, vol. 54, no. 11, pp. 4118–4132, Oct. 2006.

[34] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "A collaborative training algorithm for distributed learning," *IEEE Trans. Inf. Theory*, vol. 55, no. 4, pp. 1856–1871, Apr. 2009.

[35] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2508–2530, Jun. 2006.

[36] W. Ren and R. W. Beard, "Consensus seeking in multiagent systems under dynamically changing interaction topologies," *IEEE Trans. Autom. Control*, vol. 50, no. 5, pp. 655–661, May 2005.

[37] S. Sardellitti, M. Giona, and S. Barbarossa, "Fast distributed average consensus algorithms based on advection-diffusion processes," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 826–842, Feb. 2010.

[38] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Trans. Autom. Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[39] R. Olfati-Saber, J.A. Fax, and R.M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proc. IEEE*, vol. 95, no. 1, pp. 215–233, Jan. 2007.

[40] P. Di Lorenzo and S. Barbarossa, "A bio-inspired swarming algorithm for decentralized access in cognitive radio," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6160–6174, Dec. 2011.

[41] F. S. Cattivelli and A. H. Sayed, "Self-organization in bird flight formations using diffusion adaptation," in *Proc. 3rd International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, Aruba, Dutch Antilles, Dec. 2009, pp. 49–52.

[42] S-Y. Tu and A. H. Sayed, "Mobile adaptive networks with self-organization abilities," in *Proc. 7th International Symposium on Wireless Communication Systems*, York, United Kingdom, Sep. 2010, pp. 379–383.

[43] Y.-W. Hong and A. Scaglione, "A scalable synchronization protocol for large scale sensor networks and its applications," *IEEE J. Sel. Areas Comm.*, vol. 23, no. 5, pp. 1085–1099, May 2005.

[44] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks — Part II: Performance analysis," *to appear in IEEE Trans. Inf. Theory*, 2015 [also available as arXiv:1312.7580, Dec. 2013].

[45] P. Bianchi, G. Fort, and W. Hachem, "Performance of a distributed stochastic approximation algorithm," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7405–7418, Nov. 2013.

[46] B. Johansson, T. Keviczky, M. Johansson, and K.H. Johansson, "Subgradient methods and consensus algorithms for solving convex optimization problems," in *Proc. IEEE Conf. Decision and Control (CDC)*, Cancun, Mexico, Dec. 2008, IEEE, pp. 4185–4190.

[47] P. Braca, S. Marano, and V. Matta, "Running consensus in wireless sensor networks," in *Proc. 11th IEEE Int. Conf. on Information Fusion*, Cologne, Germany, June 2008, pp. 1–6.

[48] S. S. Stankovic, M. S. Stankovic, and D. M. Stipanovic, "Decentralized parameter estimation by consensus based stochastic approximation," *IEEE Trans. Autom. Control*, vol. 56, no. 3, pp. 531–543, Mar. 2011.

[49] B. Polyak, *Introduction to Optimization*, Optimization Software, NY, 1987.

[50] A. Tahbaz-Salehi and A. Jadbabaie, "A necessary and sufficient condition for consensus over random networks," *IEEE Trans. Autom. Control*, vol. 53, no. 3, pp. 791–795, Apr. 2008.

[51] D. P. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM J. Optim.*, vol. 10, no. 3, pp. 627–642, 2000.

[52] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990.

[53] F. Iutzeler, C. Philippe, and W. Hachem, "Analysis of sum-weight-like algorithms for analysis of sum-weight-like algorithms for averaging in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 61, no. 11, pp. 2802–2814, Jun. 2013.

[54] B. Widrow, J. M. McCool, M. G. Larimore, and C. R. Johnson Jr, "Stationary and nonstationary learning characterisitcs of the LMS adaptive filter," *Proc. IEEE*, vol. 64, no. 8, pp. 1151—1162, Aug. 1976.

[55] S. Jones, R. Cavin III, and W. Reed, "Analysis of error-gradient adaptive linear estimators for a class of stationary dependent processes," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 318–329, Mar. 1982.

[56] W. A. Gardner, "Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis, and critique," *Signal Process.*, vol. 6, no. 2, pp. 113–133, Apr. 1984.

[57] A. Feuer and E. Weinstein, "Convergence analysis of LMS filters with uncorrelated gaussian data," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 1, pp. 222–230, Feb. 1985.

[58] A. J. Laub, *Matrix Analysis for Scientists and Engineers*, SIAM, PA, 2005.

[59] E. Kreyszig, *Introductory Functional Analysis with Applications*, Wiley, NY, 1989.

[60] A. H. Sayed, *Adaptive Filters*, Wiley, NJ, 2008.