Convex Combination of Diffusion Strategies over Networks

Danqi Jin, Jie Chen, Cédric Richard, Jingdong Chen, Ali Sayed

Convex Combination of Diffusion Strategies over Networks

Danqi Jin, *Student Member, IEEE*, Jie Chen, *Senior Member, IEEE*, Cédric Richard, *Senior Member, IEEE*, Jingdong Chen, *Senior Member, IEEE*, Ali H. Sayed, *Fellow, IEEE*

***Abstract*—Combining diffusion strategies with complementary properties enables enhanced performance when they can be run simultaneously. In this paper, we propose two convex combination schemes, the power-normalized one and the sign-regressor one. Without loss of generality, theoretical investigations are focused on the former. An analysis of universality shows that it cannot perform worse than any of its component strategies in terms of the excess mean-square-error (EMSE) at steady-state. A theoretical analysis of stability also reveals that it is more stable than affine combination schemes. Next, several adjustments are proposed to further improve the performance of convex combination schemes. Finally, simulation results are presented to demonstrate their effectiveness as well as the accuracy of the theoretical results.**

***Index Terms*—Distributed optimization, diffusion strategy, convex combination, adaptive fusion strategy, performance analysis.**

I. Introduction

Model and parameter selection problems are ubiquitous and challenging in signal processing and machine learning. In most situations, selecting an optimal model structure is a difficult task and requires deep knowledge of the problem domain. Instead, one can resort to training a collection of models and combining them in a manner that enhances performance. Combination strategies have already been successfully considered for traditional adaptive filters [2], [3], in multi-kernel learning [4], as well as with deep neural network structures [5].

An inspection of the existing literature on diffusion strategies reveals that they include diffusion LMS [6], [7], diffusion APA [8], diffusion Kalman filter [9], diffusion RLS [10], and others [11]–[13], in addition to multi-task learning counterparts [14]–[25]. These different strategies perform well in the conditions under which they were derived; some deliver better performance than others depending on the underlying model for the data. In this paper, we show how to take advantage of multiple schemes by combining them in a way that can lead to enhanced performance. We introduce several combination strategies and study how performance improvements occur.

Convex and affine combinations are two useful schemes for fusing adaptive schemes with different adaptation gains [26]–[31] or complementary capabilities [32]–[35]. In [36], we have thoroughly studied the affine combination of diffusion strategies over networks. Its universality at steady-state was established, and its stochastic behavior was analyzed in the mean and mean-square sense. Though affine combination can provide good performance, convex combination scheme is often preferred for stand-alone adaptive filters since it has a wider stability range [31]. The aim of this paper is to determine if these conclusions still hold for convex combinations of diffusion strategies compared to affine combinations ones. This question is particularly challenging in the context of adaptive networks due to the multiple agents that can interact. The contributions of this work are summarized as follows:

1) Convex combinations of diffusion strategies are introduced to address the model and parameter selection problem within the context of distributed estimation over adaptive networks. In particular, two convex combiners are considered, the power-normalized one and the sign-regressor one.
2) A theoretical analysis of the convex power-normalized combiner is conducted to illustrate universality at steady state, and to derive conditions that ensure the mean and mean-square stabilities of this scheme. Theoretical results reveal that, compared to affine combination, convex power-normalized combiners are more stable.
3) Extensions are discussed to further improve the performance of convex combination schemes.
4) The computation and communication complexities of convex combiners are discussed, and a comparison of convex and affine combination schemes is provided.

Though the convex combination algorithm shares a similar form with the affine combination strategy studied in [36], the universality and behavior analyses require original manipulations. Similar steps are voluntarily omitted in the presentation to avoid redundancy and focus on the main differences. The result shows that, as long as each component strategy is stable, any diffusion network with convex power-normalized combination scheme will be stable. This differs from [36] where additional conditions are required to ensure the stability of affine combiners.

The paper is organized as follows. Signal model and diffusion LMS algorithm are presented in Section II. Section III presents the convex combination framework and introduces two strategies to adapt the convex combination coefficients. Section IV analyzes the theoretical performance of the convex power-normalized combiner with two diffusion LMS. Extensions of this scheme are discussed to further improve the performance in Section V. Discussions about the computation and communication complexities are provided in Section VI, as well as a comparison with the affine combination scheme. Simulation results are provided in Section VII. Section VIII concludes this work.

Notation. Normal font x denotes scalars. Boldface lower-

A preliminary version of this work appeared in [1].

cased letters $\boldsymbol{x}$ and capital letters $\boldsymbol{X}$ denote column vectors and matrices, respectively. The superscript $(\cdot)^\top$ denotes the transpose operator. The inverse of a matrix is denoted by $(\cdot)^{-1}$. The mathematical expectation is denoted by $\mathbb{E}\{\cdot\}$. The operator $[\cdot]_a^b$ truncates its argument with lower bound a and upper bound b. The operator $\mathrm{diag}\{\cdot\}$ takes the diagonal elements of its matrix argument, or generates a diagonal matrix from its vector argument. $\boldsymbol{I}_N$ and $\boldsymbol{0}_N$ denote identity matrix and zero matrix of size $N \times N$, respectively. All-one vector of length N is denoted by $\mathbb{1}_N$. $\mathcal{N}_k$ denotes the neighbors of node k, including k.

II. Signal Model and Diffusion LMS

Consider a connected network consisting of N nodes. The problem is to estimate an unknown parameter vector $\boldsymbol{w}_k^\star$ of length $L \times 1$ at each agent k. Agent k has access to temporal measurement sequences $\{d_{k,n}, \boldsymbol{u}_{k,n}\}$, where $d_{k,n}$ denotes a reference signal, and $\boldsymbol{u}_{k,n}$ is an $L \times 1$ regression vector with positive definite covariance matrix $\boldsymbol{R}_{u,k}$. The data at agent k and time instant n are related according to the linear model:

$$d_{k,n} = \boldsymbol{u}_{k,n}^\top \boldsymbol{w}_k^\star + z_{k,n}, \tag{1}$$

where $z_{k,n}$ is an additive noise satisfying Assumption 1 below. Note that the network operates in the so-called multi-task setting when the unknown parameter vectors $\boldsymbol{w}_k^\star$ differ from each other. The single-task setting usually considered in the literature is a special case of the multi-task one considered in this paper. It is obtained by setting $\boldsymbol{w}_1^\star = \boldsymbol{w}_2^\star = \cdots = \boldsymbol{w}_N^\star$.

Assumption 1: $z_{k,n}$ is a zero-mean, stationary, independent and identically distributed (i.i.d.) additive noise with variance $\sigma_{z,k}^2$, and independent of any other signals.

A1[1] is widely adopted in the literature of adaptive filters and distributed online learning over networks. To determine $\boldsymbol{w}_k^\star$, we consider the MSE cost at agent k:

$$J_k(\boldsymbol{w}) = \mathbb{E}\{|d_{k,n} - \boldsymbol{u}_{k,n}^\top \boldsymbol{w}|^2\}. \tag{2}$$

Clearly, $J_k(\boldsymbol{w})$ is minimized at $\boldsymbol{w}_k^\star$. For single-task problems, each agent in the network estimates the same parameter vector, while for multi-task problems, agents may estimate distinct parameter vectors.

Diffusion LMS was derived in [6], [7], [37], [38] to minimize the global cost defined by:

$$J^{\mathrm{glob}}(\boldsymbol{w}) = \sum_{k=1}^{N} J_k(\boldsymbol{w}) \tag{3}$$

in a cooperative manner. The general diffusion LMS algorithm is given by:

$$\begin{cases} \boldsymbol{\phi}_{k,n} = \displaystyle\sum_{\ell \in \mathcal{N}_k} a_{1,\ell k}\, \boldsymbol{w}_{\ell,n} \\ \boldsymbol{\psi}_{k,n+1} = \boldsymbol{\phi}_{k,n} + \mu_k \displaystyle\sum_{\ell \in \mathcal{N}_k} c_{\ell k}\, \boldsymbol{u}_{\ell,n}\, (d_{\ell,n} - \boldsymbol{u}_{\ell,n}^\top \boldsymbol{\phi}_{k,n}) \\ \boldsymbol{w}_{k,n+1} = \displaystyle\sum_{\ell \in \mathcal{N}_k} a_{2,\ell k}\, \boldsymbol{\psi}_{\ell,n+1} \end{cases} \tag{4}$$

[1]In this paper, we adopt the acronym '**A**' for '**Assumption**'.

where the nonnegative coefficients $\{a_{1,\ell k}\}$, $\{a_{2,\ell k}\}$ and $\{c_{\ell k}\}$ are (ℓ, k)-th entries of two left stochastic matrices $\boldsymbol{A}_1, \boldsymbol{A}_2$ and a right stochastic matrix $\boldsymbol{C}$, respectively, satisfying:

$$\boldsymbol{A}_1^\top \mathbb{1}_N = \mathbb{1}_N,\ \boldsymbol{A}_2^\top \mathbb{1}_N = \mathbb{1}_N,\ \boldsymbol{C}\mathbb{1}_N = \mathbb{1}_N, \tag{5}$$

$$a_{1,\ell k} = 0,\ \ a_{2,\ell k} = 0,\ \ c_{\ell k} = 0 \ \text{ if } \ \ell \notin \mathcal{N}_k. \tag{6}$$

Setting $\boldsymbol{A}_1 = \boldsymbol{I}$ or $\boldsymbol{A}_2 = \boldsymbol{I}$ leads to the adapt-then-combine (ATC) and the combine-then-adapt (CTA) diffusion strategy, respectively. Note that algorithm (4) can be used to solve both single-task problems [39] and multi-task problems [15] by selecting appropriate matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$.

III. Convex Combination Framework

As illustrated in Fig. 1, the convex combination framework consists of two concurrent layers, namely, a diffusion strategy layer and a combination layer. In the diffusion strategy layer, the network simultaneously runs M candidate diffusion strategies, resulting in M groups of estimates for the optimal weight vector. We shall consider, without lack of generality, the case $M = 2$ in the rest of the paper.

Diffusion strategy $\mathrm{S}^{(i)}$ parameters are $\boldsymbol{A}_1^{(i)}$, $\boldsymbol{A}_2^{(i)}$, $\boldsymbol{C}^{(i)}$, $\mu_k^{(i)}$ for $i = 1, 2$, where the superscript $^{(i)}$ is the indicator for the i-th component strategy. Based on the signal model (1), we define the estimates of the reference signal, the a priori output estimation error, and the a priori estimation error, as follows:

$$y_{k,n}^{(i)} \triangleq \boldsymbol{u}_{k,n}^\top\, \boldsymbol{w}_{k,n}^{(i)} \tag{7}$$

$$e_{k,n}^{(i)} \triangleq d_{k,n} - \boldsymbol{u}_{k,n}^\top\, \boldsymbol{w}_{k,n}^{(i)} \tag{8}$$

$$\tilde{e}_{k,n}^{(i)} \triangleq \boldsymbol{u}_{k,n}^\top\, (\boldsymbol{w}_k^\star - \boldsymbol{w}_{k,n}^{(i)}). \tag{9}$$

respectively, at each agent k and time instant n. The input of the combination layer consists of the two outputs from the diffusion strategy layer. By assigning convex combination coefficients:

$$\gamma_{k,n}^{(1)} \triangleq \gamma_{k,n} \tag{10}$$

$$\gamma_{k,n}^{(2)} \triangleq 1 - \gamma_{k,n} \tag{11}$$

to the two component strategies at agent k and time instant n, we obtain the overall weight estimation $\boldsymbol{w}_{k,n}$ at the combination layer for agent k:

$$\begin{aligned} \boldsymbol{w}_{k,n} &= \sum_{i=1}^{2} \gamma_{k,n}^{(i)} \boldsymbol{w}_{k,n}^{(i)} \\ &= \gamma_{k,n} \boldsymbol{w}_{k,n}^{(1)} + (1 - \gamma_{k,n}) \boldsymbol{w}_{k,n}^{(2)}. \end{aligned} \tag{12}$$

Convex constraint on $\gamma_{k,n}^{(i)}$ requires that $\gamma_{k,n}^{(i)} \in [0, 1]$. By using model (1) and equations (7)–(12), we obtain the following relation between the overall quantities at the combination layer and the corresponding quantities at the diffusion strategy layer, at each node k and time instant n:

$$x_{k,n} = \sum_{i=1}^{2} \gamma_{k,n}^{(i)} x_{k,n}^{(i)} \tag{13}$$

where the quantities $x_{k,n}^{(i)}$ for each component strategy $\mathrm{S}^{(i)}$ and $x_{k,n}$ at the combination layer generically refer to the estimates

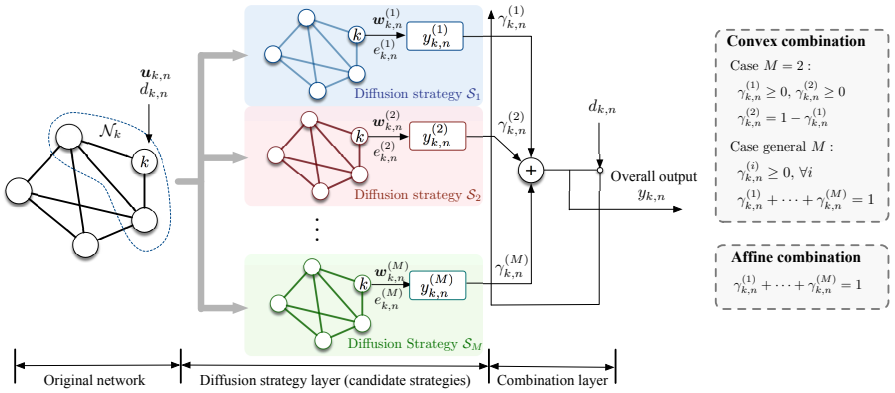


Fig. 1. Illustration of the combination framework with M component diffusion strategies. Section III is a special case by setting $M = 2$.

of the reference signal in (7), the error signals in (8), and the a prior estimation error in (9), respectively.

The goal of the combination layer is to learn which diffusion strategy performs better at each time instant and each agent, and to assign them with weights in order to optimize the overall network performance. The problem then reduces to designing a strategy to adjust $\gamma_{k,n}$. To avoid using hard-thresholding operators to satisfy the convex combination requirement, we introduce an auxiliary variable $\alpha_{k,n}$ to reparameterize $\gamma_{k,n}$ as follows:

$$\gamma_{k,n}^{(1)} = \gamma_{k,n} = \frac{1}{1+e^{-\alpha_{k,n}}} \tag{14}$$

$$\gamma_{k,n}^{(2)} = 1 - \gamma_{k,n}. \tag{15}$$

Then, we adjust $\alpha_{k,n}$ by minimizing the MSE at the combination layer, which is defined by:

$$J_n^{\text{MSE}} = \frac{1}{2}\sum_{k=1}^{N} \mathbb{E}\{e_{k,n}^2\}. \tag{16}$$

Adaptation of $\alpha_{k,n}$ can be conducted by performing stochastic gradient descent on (16), that is,

$$\begin{aligned}\alpha_{k,n+1} &= \alpha_{k,n} - v'_{\alpha_k}\frac{\partial J_n^{\text{MSE}}}{\partial \alpha_{k,n}} \\ &\approx \alpha_{k,n} + v'_{\alpha_k}\gamma_{k,n}(1-\gamma_{k,n})e_{k,n}\boldsymbol{u}_{k,n}^{\top}(\boldsymbol{w}_{k,n}^{(1)} - \boldsymbol{w}_{k,n}^{(2)})\end{aligned} \tag{17}$$

with v'_{α_k} a positive step-size. On the one hand, observe that iteration (17) will stop if $\alpha_{k,n}$ is allowed to have unbounded growth or decline, since $\gamma_{k,n}$ will get close to 0 or 1 and will make the term $\gamma_{k,n}(1-\gamma_{k,n})$ being 0. Then, to ensure continuous learning, we propose restricting $\alpha_{k,n}$ to be within the interval $[-\alpha^+, \alpha^+]$ [31]. On the other hand, in order to compensate the effect of large fluctuations in the power of $[\boldsymbol{u}_{k,n}^{\top}(\boldsymbol{w}_{k,n}^{(1)} - \boldsymbol{w}_{k,n}^{(2)})]$ at the adaptation level, we propose normalizing the step-sizes v'_{α_k} in the following two ways.

1) Convex power-normalized scheme: Setting $v'_{\alpha_k} = \frac{v_{\alpha_k}}{\varepsilon + p_{k,n}}$ with v_{α_k} the initial step-size, we obtain the convex power-normalized scheme. In this expression, parameter ε is a small positive constant to avoid dividing by zero, and $p_{k,n}$ is an estimate of the power of $\boldsymbol{u}_{k,n}^{\top}(\boldsymbol{w}_{k,n}^{(1)} - \boldsymbol{w}_{k,n}^{(2)})$ calculated as:

$$p_{k,n} = \eta\, p_{k,n-1} + (1-\eta)\big[\,\boldsymbol{u}_{k,n}^{\top}(\boldsymbol{w}_{k,n}^{(1)} - \boldsymbol{w}_{k,n}^{(2)})\,\big]^2, \tag{18}$$

with $0 \ll \eta < 1$ a temporal smoothing factor.

2) Convex sign-regressor scheme: To save computation and storage resources in evaluating $p_{k,n}$ in the power-normalized scheme, we introduce another normalization. Setting

$$v'_{\alpha_k} = \frac{v_{\alpha_k}}{|\boldsymbol{u}_{k,n}^{\top}(\boldsymbol{w}_{k,n}^{(1)} - \boldsymbol{w}_{k,n}^{(2)})|}$$

results in the convex sign-regressor scheme given by:

$$\begin{aligned}&\alpha_{k,n+1} = \\ &\alpha_{k,n} + v_{\alpha_k}\gamma_{k,n}(1\gamma_{k,n})e_{k,n}\text{sgn}\{\boldsymbol{u}_{k,n}^{\top}(\boldsymbol{w}_{k,n}^{(1)} - \boldsymbol{w}_{k,n}^{(2)})\}\end{aligned} \tag{19}$$

where $\text{sgn}\{x\}$ is the sign function. Compared with the power-normalized scheme, (19) requires only evaluating the sign function.

IV. Theoretical Analysis of Convex Power-Normalized Scheme

Due to space limitation and technical complexity, we shall only conduct the theoretical analysis for the power-normalized scheme. The theoretical analysis of the sign-regressor scheme can be derived by following the same routine. However, other techniques for dealing with the nonlinear sign function may be required.

To facilitate the theoretical analysis, we shall now introduce several assumptions and approximations as in [36].

Assumption 2: The regression vector $\boldsymbol{u}_{k,n}$, generated from a zero-mean random process, is temporally stationary, white (over n) and spatially independent (over k) with positive definite covariance matrix $\boldsymbol{R}_{u,k} = \mathbb{E}\{\boldsymbol{u}_{k,n}\boldsymbol{u}_{k,n}^{\top}\}$.

Approximation 1: At steady state, $\gamma_{k,n}$ is statistically independent of $\tilde{e}_{k,n}^{(i)}$ and $p_{k,n}$.

Approximation 2: For a large enough temporal smoothing factor η, $p_{k,n}$ is statistically independent of $\boldsymbol{u}_{k,n}^{\top}\boldsymbol{w}_{k,n}^{(i)}$, that is, it is independent of $\tilde{e}_{k,n}^{(i)}$.

Approximation 3: At each time instant n, $\gamma_{k,n}$ is statistically independent of $\boldsymbol{w}_{k,n}^{(i)}$ for $i = 1, 2$.

Approximation 4: $\boldsymbol{\Gamma}_{n+1}$ is statistically independent of $\boldsymbol{B}_n^{(i)}$, $\boldsymbol{v}_n^{(i)}$, $\boldsymbol{g}_n^{(i)}$ and $\boldsymbol{r}_n^{(i)}$ for $i=1,2$, where $\boldsymbol{\Gamma}_{n+1}$ is defined further ahead in (35), $\boldsymbol{v}_n^{(i)}$ is defined in (38), $\boldsymbol{g}_n^{(i)}$ and $\boldsymbol{r}_n^{(i)}$ are defined in (81) and (85) of Appendix B, respectively.

Approximation 5: Parameter $\alpha_{k,n}$ varies slowly enough so that $\mathbb{E}\{\alpha_{k,n}\tilde{e}_{k,n}^{(m)}\tilde{e}_{k,n}^{(n)}\} \approx \mathbb{E}\{\alpha_{k,n}\}\mathbb{E}\{\tilde{e}_{k,n}^{(m)}\tilde{e}_{k,n}^{(n)}\}$, $m,n=1,2$.

Approximation 6: The a priori estimation errors $\tilde{e}_{k,n}^{(1)}$ and $\tilde{e}_{k,n}^{(2)}$ are jointly Gaussian distributed, which implies [40]:

$$\mathbb{E}\{(\tilde{e}_{k,n}^{(i)})^4\} = 3\left[J_{\text{ex},k,n}^{(i)}\right]^2, \quad i=1,2, \tag{20}$$

$$\mathbb{E}\{(\tilde{e}_{k,n}^{(1)})^3(\tilde{e}_{k,n}^{(2)})^1\} = 3\,J_{\text{ex},k,n}^{(1)}J_{\text{ex},k,n}^{(1,2)}, \tag{21}$$

$$\mathbb{E}\{(\tilde{e}_{k,n}^{(1)})^1(\tilde{e}_{k,n}^{(2)})^3\} = 3\,J_{\text{ex},k,n}^{(1,2)}J_{\text{ex},k,n}^{(2)}, \tag{22}$$

$$\mathbb{E}\{(\tilde{e}_{k,n}^{(1)})^2(\tilde{e}_{k,n}^{(2)})^2\} = 2\left[J_{\text{ex},k,n}^{(1,2)}\right]^2 + J_{\text{ex},k,n}^{(1)}J_{\text{ex},k,n}^{(2)}, \tag{23}$$

where $J_{\text{ex},k,n}^{(i)}$ and $J_{\text{ex},k,n}^{(1,2)}$ are defined further in (24) and (26).

Approximation 7: At steady state when $n\to\infty$, the variance of $\gamma_{k,n}$ is small. □

Although not true in general, these assumptions and approximations are usually adopted to simplify the derivation without constraining the conclusions. Specifically, there are several results in the literature showing that performance results obtained under **A2** match well with actual performance when step-sizes are sufficiently small [39], [41]. As discussed in [2], approximation **Ap1**[2] is reasonable when adopting a decaying step-size ν_{α_k}, and is more justified when $\alpha_{k,n+1}$ approaches α^+ or $-\alpha^+$, in which case $\gamma_{k,n}(1-\gamma_{k,n})$ of (17) tends to zero. **Ap2** is justified when using a large temporal smoothing factor η. Though actually not hold, **Ap3** and **Ap4** do not affect the theoretical results heavily, as illustrated in simulation results. **Ap5** is widely adopted in the analysis of convex combinations of filters [42], [43], and it coincides with simulation result that $\alpha_{k,n}$ converges slowly compared to the variation of $\boldsymbol{u}_{k,n}$, thus to the variation of $\tilde{e}_{k,n}^{(i)}$. **Ap6** is also adopted and examined in the analysis of combinations of adaptive filters [29], [42]. **Ap7** has been adopted in the analysis of adaptive filters [2], and it is reasonable in that at steady state, $\gamma_{k,n}$ will converge to a fixed value or fluctuate within a small neighborhood of this value. Thus it is reasonable to assume that the variance of $\gamma_{k,n}$ is small. We shall challenge these assumptions and approximations in the simulations.

A. Universality at steady state

The EMSE of component strategy $\mathrm{S}^{(i)}$ and that after combination at node k and time instant n are defined by:

$$J_{\text{ex},k,n}^{(i)} \triangleq \mathbb{E}\{(\tilde{e}_{k,n}^{(i)})^2\} \tag{24}$$

$$J_{\text{ex},k,n} \triangleq \mathbb{E}\{[\gamma_{k,n}\tilde{e}_{k,n}^{(1)} + (1-\gamma_{k,n})\tilde{e}_{k,n}^{(2)}]^2\}, \tag{25}$$

respectively. Let us also introduce the cross-EMSE defined as:

$$J_{\text{ex},k,n}^{(1,2)} \triangleq \mathbb{E}\{(\tilde{e}_{k,n}^{(1)}\tilde{e}_{k,n}^{(2)}\} \tag{26}$$

[2]In this paper, we adopt the acronym '**Ap**' for '**Approximation**'.

The EMSEs of the entire network, for component strategy $\mathrm{S}^{(i)}$ and after combination, are respectively defined by:

$$J_{\text{ex,net},n}^{(i)} \triangleq \sum_{k=1}^{N} J_{\text{ex},k,n}^{(i)} \quad \text{for} \quad i=1,2 \tag{27}$$

$$J_{\text{ex,net},n} \triangleq \sum_{k=1}^{N} J_{\text{ex},k,n}. \tag{28}$$

By taking the limit as $n\to\infty$, we obtain the corresponding values at steady state: $J_{\text{ex},k,\infty}^{(i)}$, $J_{\text{ex},k,\infty}$, $J_{\text{ex,net},\infty}^{(i)}$ and $J_{\text{ex,net},\infty}$. Based on the above definitions and several approximations, we arrive at the following result.

Theorem 1 *(Universality at steady state): Assume data model* (1)*, assumptions **A1**, **A2** and approximations **Ap1**, **Ap2**, **Ap7** hold. Then for any initial conditions, the network with power-normalized scheme* (17) *is universal at steady state, which means that the EMSE of the diffusion network after combination cannot be worse than that of the best component strategy, namely,*

$$J_{\text{ex,net},\infty} \le \min\{J_{\text{ex,net},\infty}^{(1)}, J_{\text{ex,net},\infty}^{(2)}\}. \tag{29}$$

Further, when $J_{\text{ex},k,\infty}^{(i)} > J_{\text{ex},k,\infty}^{(1,2)}$ for $i=1,2$ and $\bar{\gamma}_{k,\infty}$ defined in (65) satisfies $\bar{\gamma}_{k,\infty} \in \left(1-\theta_k^+, \theta_k^+\right)$ for some node k, there is an improvement in the EMSE after combination, such that:

$$J_{\text{ex,net},\infty} < \min\{J_{\text{ex,net},\infty}^{(1)}, J_{\text{ex,net},\infty}^{(2)}\}. \tag{30}$$

Proof: See Appendix A. ■

Theorem 1 is meaningful in that it shows that, through local combination at each node, the convex power-normalized scheme reaches the minimal EMSE of the two component strategies, and can even do better.

B. Mean weight and mean-square behaviors analysis

The mean weight and mean-square behavior analysis follows a similar routine as in [36] for affine combination schemes. To avoid redundancy, we provide the main results without proving them to better highlight the main differences.

Define the following block vectors:

$$\boldsymbol{w}^\star \triangleq \mathrm{col}\{\boldsymbol{w}_1^\star, \cdots, \boldsymbol{w}_N^\star\} \tag{31}$$

$$\boldsymbol{w}_n^{(i)} \triangleq \mathrm{col}\{\boldsymbol{w}_{1,n}^{(i)}, \cdots, \boldsymbol{w}_{N,n}^{(i)}\} \tag{32}$$

$$\boldsymbol{w}_n \triangleq \mathrm{col}\{\boldsymbol{w}_{1,n}, \cdots, \boldsymbol{w}_{N,n}\}, \tag{33}$$

where $\boldsymbol{w}^\star$ is the block optimum weight vector, $\boldsymbol{w}_n^{(i)}$ and $\boldsymbol{w}_n$ are the block weight estimate of component strategy $\mathrm{S}^{(i)}$ and after combination at time instant n, respectively. Using definitions (12) and (32), (33), we arrive at:

$$\boldsymbol{w}_n = \boldsymbol{\Gamma}_n\boldsymbol{w}_n^{(1)} + (\boldsymbol{I}_{NL} - \boldsymbol{\Gamma}_n)\boldsymbol{w}_n^{(2)}, \tag{34}$$

where

$$\boldsymbol{\Gamma}_n \triangleq \mathrm{diag}\{\gamma_{1,n}, \cdots, \gamma_{N,n}\} \otimes \boldsymbol{I}_L \tag{35}$$

with symbol $\otimes$ denoting the Kronecker product. The weight error vector of node k for component strategy $\mathrm{S}^{(i)}$ and that for combination layer are defined by:

$$\boldsymbol{v}_{k,n}^{(i)} \triangleq \boldsymbol{w}_{k,n}^{(i)} - \boldsymbol{w}_k^\star \tag{36}$$

$$\boldsymbol{v}_{k,n} \triangleq \boldsymbol{w}_{k,n} - \boldsymbol{w}_k^\star, \tag{37}$$

respectively. By stacking $\boldsymbol{v}_{k,n}^{(i)}$ and $\boldsymbol{v}_{k,n}$ over the entire network into block vectors, we have:

$$\boldsymbol{v}_n^{(i)} \triangleq \text{col}\{\boldsymbol{v}_{1,n}^{(i)}, \cdots, \boldsymbol{v}_{N,n}^{(i)}\} \tag{38}$$

$$\boldsymbol{v}_n \triangleq \text{col}\{\boldsymbol{v}_{1,n}, \cdots, \boldsymbol{v}_{N,n}\}. \tag{39}$$

Using (31)–(39), we have:

$$\boldsymbol{v}_n = \boldsymbol{\Gamma}_n \boldsymbol{v}_n^{(1)} + (\boldsymbol{I}_{NL} - \boldsymbol{\Gamma}_n)\boldsymbol{v}_n^{(2)}. \tag{40}$$

1) Mean weight behavior analysis

Under approximation **Ap3** and using (40), the mean weight behavior at the combination layer satisfies:

$$\begin{aligned}\mathbb{E}\{\boldsymbol{v}_{n+1}\} = \mathbb{E}\{\boldsymbol{\Gamma}_{n+1}\}\mathbb{E}\{\boldsymbol{v}_{n+1}^{(1)}\} \\ + \mathbb{E}\{\boldsymbol{I}_{NL} - \boldsymbol{\Gamma}_{n+1}\}\mathbb{E}\{\boldsymbol{v}_{n+1}^{(2)}\}.\end{aligned} \tag{41}$$

We need to evaluate the iteration of $\mathbb{E}\{\boldsymbol{v}_{n+1}^{(i)}\}$ to analyze the mean weight behavior of $\boldsymbol{v}_{n+1}$; see Appendix B. Based on Appendix B, we obtain the following result.

Theorem 2 *(Stability in the mean)*: *Assume data model* (1)*, assumptions **A1**, **A2** and approximation **Ap3** hold. Then for any initial conditions, the network with power-normalized scheme* (17) *asymptotically converges in the mean if the step sizes in the network are chosen to satisfy:*

$$0 < \mu_k^{(i)} < \frac{2}{\lambda_{\max}\{\boldsymbol{R}_k^{(i)}\}}, \ k = 1, \cdots, N \text{ and } i = 1, 2 \tag{42}$$

with $\lambda_{\max}\{\cdot\}$ denoting the largest eigenvalue of its matrix argument. The asymptotic bias at steady state is given by:

$$\begin{aligned}\mathbb{E}\{\boldsymbol{v}_\infty\} = -\mathbb{E}\{\boldsymbol{\Gamma}_\infty\}\big(\boldsymbol{I}_{NL} - \boldsymbol{B}^{(1)}\big)^{-1}\boldsymbol{r}^{(1)} \\ - \big(\boldsymbol{I}_{NL} - \mathbb{E}\{\boldsymbol{\Gamma}_\infty\}\big)\big(\boldsymbol{I}_{NL} - \boldsymbol{B}^{(2)}\big)^{-1}\boldsymbol{r}^{(2)}.\end{aligned} \tag{43}$$

Proof: On the basis of Appendix B, arguments run along the lines of [36, Appendix C] where affine combinations of two diffusion LMS are considered. ■

Unlike the proof in [36] where we need to impose an additional condition on the step-size v_{α_k} of the combination layer to ensure that matrix $\mathbb{E}\{\boldsymbol{\Gamma}_{n+1}\}$ is bounded, we do not have such condition in the current work for convex power-normalized scheme. In that way, convex combination schemes are more stable than affine ones in the mean sense.

2) Mean-square behavior analysis

The concept of mean-square stability is one of the most attractive ones within the large branch of stability analysis, and is widely adopted in the analyses of adaptive filters [41], [44], [45] and adaptive networks [37]–[39]. Indeed, requiring only mean stability for adaptive filters and adaptive networks is not fully satisfactory since a filter can converge in the mean sense but may fluctuate around its mean value. Mean-square stability analyses provide complementary tools for understanding and predicting filters behavior.

We need to evaluate the evolution of $\mathbb{E}\{\|\boldsymbol{v}_{n+1}\|_{\boldsymbol{\Sigma}}^2\}$ over time, where $\boldsymbol{\Sigma}$ denotes an arbitrary positive semi-definite matrix, and $\|\boldsymbol{x}\|_{\boldsymbol{\Sigma}}^2 \triangleq \boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}$. The evolution of $\mathbb{E}\{\|\boldsymbol{v}_{n+1}\|_{\boldsymbol{\Sigma}}^2\}$ is depicted in Appendix C. On the basis of Appendix C, we obtain the following theorem.

Theorem 3 *(Mean-square stability)*: *Assume data model* (1)*, assumptions **A1**, **A2** and approximations **Ap3**, **Ap4** hold. Assume further that step-sizes $\mu_k^{(i)}$ are sufficiently small such that condition* (42) *is satisfied and approximations* (93), (100) *are justified by ignoring higher powers of step-size. Then for any initial conditions, any network with doubly stochastic matrices $\boldsymbol{A}_1^{(i)}, \boldsymbol{A}_2^{(i)}$, i.e., both columns and rows add up to one, and power-normalized scheme* (17) *is mean-square stable for sufficiently small step-sizes satisfying condition* (42)*.*

Proof: On the basis of the Appendix C, the proof of this theorem follows the same routine as in [36, Appendix E]. We omit it to avoid redundancy. ■

Unlike the proof in [36] where we impose an additional condition on the step-size v_{α_k} of the combination layer to ensure that matrix $\mathbb{E}\{\boldsymbol{\Gamma}_{n+1}^\top \otimes \boldsymbol{\Gamma}_{n+1}^\top\}$ is bounded, we do not have such condition here since it is bounded with convex power-normalized scheme.

Corollary 1 *(Transient MSD)*: Using (87) with $\boldsymbol{\Sigma} = \frac{1}{N}\boldsymbol{I}_{NL}$, we evaluate the mean-square-deviation (MSD) learning curve of the entire network, defined by $\xi_{n+1} \triangleq \mathbb{E}\{\|\boldsymbol{v}_{n+1}\|_{\frac{1}{N}\boldsymbol{I}_{NL}}^2\}$. All terms on the RHS of (87) can be evaluated recursively. See Appendix D for the explicit expressions of these recursions.

Corollary 2 *(Steady-state MSD)*: For sufficiently small step-sizes satisfying condition (42) to ensure stabilities in the mean and mean-square sense of the power-normalized scheme, the steady-state MSD is provided by (107) in Appendix E.

Proof: Following the same routine as in [36, Appendix G] leads to the result. ■

Corollaries 1 and **2** characterize the transient and steady-state MSD of the convex power-normalized scheme. These results help to tune the model parameters in practice.

C. Mean and mean-square behaviors of $\gamma_{k,n}$

To analyze the mean and mean-square behaviors of the power-normalized scheme, we need to study the mean behaviors of $\boldsymbol{\Gamma}_n$ and $\boldsymbol{\Gamma}_n^\top\boldsymbol{\Gamma}_n$. They are obtained by characterizing the mean and mean-square behaviors of $\gamma_{k,n}$ since $\boldsymbol{\Gamma}_n$ is diagonal. Following the lines in [42], [43], we can conduct a theoretical analysis for $\gamma_{k,n}$ by using a first-order Taylor series expansion. The derivations are depicted in Appendix F and Appendix G. We arrive at the following two theorems.

Theorem 4 *(Mean behavior of $\gamma_{k,n}$)*: *Assume data model* (1) *and approximations **Ap1**, **Ap5** hold. Then for any initial conditions, the convex combination coefficients $\gamma_{k,n}$ are stable in the mean sense. The mean behavior of $\gamma_{k,n}$ is described by* (123) *in Appendix F. The value of $\mathbb{E}\{\gamma_{k,n}\}$ at steady-state is given by* (65)*.*

Proof: See Appendix F. ■

Theorem 5 *(Mean-square behavior of $\gamma_{k,n}$)*: *Assume data model* (1) *and approximations **Ap1**, **Ap5**, **Ap6** hold. Then for any initial conditions, the convex combination coefficients $\gamma_{k,n}$ are stable in the mean-square sense. The mean-square*

behavior of $\gamma_{k,n}$ is described by (125) in Appendix G. The value of $\mathbb{E}\{\gamma_{k,n}^2\}$ at steady-state is given by (134).

Proof: See Appendix G. ■

V. Extensions of the scheme

We shall now introduce several extensions to further improve the performance of convex combiners, and generalize them to multiple strategies.

A. Performance improvements

Several adjustments can be implemented to further improve the performance of combination schemes in certain situations. We shall now extend the adjustment strategies already proposed for combinations of adaptive filters [3], [31], [46], [47] to combinations of diffusion strategies.

1) Weight transfer: At each time instant n and each node k, parameter $\gamma_{k,n}$ indicates which component strategy locally performs better. So, at each time instant n and each node k, the weight vector of the best component strategy can be transferred to the other component strategies in order to improve the overall performance of the combination layer. Depending on how the selected weight vector is shared, weight transfer strategies can be further divided into two categories, the copying one and the leakage one. Note that the sharing procedures described below have to be run after the combination step has been completed. The results obtained with these procedures will then be used by each node at the adaptation step of the next iteration.

Copying weights: The component strategy with the weakest performance copies the received weight vector for itself when the following conditions are satisfied simultaneously:

- $\gamma_{k,n} > \beta_1$ or $\gamma_{k,n} < 1-\beta_1$, where $0 \ll \beta_1 < 1$ is a pre-defined threshold value with typical value 0.95. This condition means that the weight transfer can occur when one component greatly outperforms another.
- $\mathrm{mod}(n, N_0) = 0$, where the $\mathrm{mod}(\cdot,\cdot)$ function returns the remainder after division, and $N_0 \geq 2$. This condition implies that weight transfers can occur periodically with period $N_0 \geq 2$.

By combining the above two conditions, we finally have:

$$\boldsymbol{w}_{k,n}^{(1)} = \begin{cases} \boldsymbol{w}_{k,n}^{(2)}, & \text{if } \gamma_{k,n} < 1-\beta_1 \text{ and } \mathrm{mod}(n,N_0)=0 \\ \boldsymbol{w}_{k,n}^{(1)}, & \text{otherwise} \end{cases} \tag{44}$$

and

$$\boldsymbol{w}_{k,n}^{(2)} = \begin{cases} \boldsymbol{w}_{k,n}^{(1)}, & \text{if } \gamma_{k,n} > \beta_1 \text{ and } \mathrm{mod}(n,N_0)=0 \\ \boldsymbol{w}_{k,n}^{(2)}, & \text{otherwise.} \end{cases} \tag{45}$$

Leakage transfer: The component strategy with the weakest performance partially absorbs the received weight vector when $\gamma_{k,n} > \beta_2$ or $\gamma_{k,n} < 1-\beta_2$, that is,

$$\boldsymbol{w}_{k,n}^{(1)} = \begin{cases} \rho\boldsymbol{w}_{k,n}^{(1)} + (1-\rho)\boldsymbol{w}_{k,n}^{(2)}, & \text{if } \gamma_{k,n} < 1-\beta_2 \\ \boldsymbol{w}_{k,n}^{(1)}, & \text{otherwise} \end{cases} \tag{46}$$

and

$$\boldsymbol{w}_{k,n}^{(2)} = \begin{cases} \rho\boldsymbol{w}_{k,n}^{(2)} + (1-\rho)\boldsymbol{w}_{k,n}^{(1)}, & \text{if } \gamma_{k,n} > \beta_2 \\ \boldsymbol{w}_{k,n}^{(2)} m & \text{otherwise} \end{cases} \tag{47}$$

where β_2 and ρ are two non-negative parameters satisfying $0 \ll \beta_2 < 1$ and $0 \ll \rho < 1$, with typical value of 0.95.

2) Weight feedback: Since the combined estimate at each node cannot be worse than the estimate of each component strategy, we can feedback the combined estimate to each component strategy periodically to improve the performance, that is,

$$\boldsymbol{w}_{k,n}^{(1)} = \begin{cases} \boldsymbol{w}_{k,n}, & \text{if } \mathrm{mod}(n,N_0')=0 \\ \boldsymbol{w}_{k,n}^{(1)}, & \text{otherwise} \end{cases} \tag{48}$$

$$\boldsymbol{w}_{k,n}^{(2)} = \begin{cases} \boldsymbol{w}_{k,n}, & \text{if } \mathrm{mod}(n,N_0')=0 \\ \boldsymbol{w}_{k,n}^{(2)}, & \text{otherwise,} \end{cases} \tag{49}$$

with period N_0' a large positive integer.

B. Convex combination of multiple strategies

We shall now extend the convex combination scheme to multiple component strategies. The general scheme follows the description in Section III except that we have M component diffusion strategies. We introduce M convex combination coefficients $\gamma_{k,n}^{(1)}, \gamma_{k,n}^{(2)}, \cdots, \gamma_{k,n}^{(M)}$ at each node k and time instant n, satisfying the non-negativity and sum-to-one constraints. By combining the M local estimates at each agent k, we obtain the overall system coefficients $\boldsymbol{w}_{k,n}$ and estimation error $e_{k,n}$ at the combination layer, defined as follows:

$$\boldsymbol{w}_{k,n} = \sum_{i=1}^{M} \gamma_{k,n}^{(i)} \boldsymbol{w}_{k,n}^{(i)} \tag{50}$$

$$e_{k,n} = \sum_{i=1}^{M} \gamma_{k,n}^{(i)} e_{k,n}^{(i)} \tag{51}$$

We adapt $\gamma_{k,n}^{(i)}$ by minimizing the MSE of the combination layer. To satisfy the non-negativity and sum-to-one constraints, we introduce a nonlinear modified softmax function to calculate $\gamma_{k,n}^{(i)}$ as:

$$\gamma_{k,n}^{(i)} = \frac{\exp(\alpha_{k,n}^{(i)}) + \delta}{\sum_{j=1}^{M} \exp(\alpha_{k,n}^{(j)}) + M\delta}, \, , \, i = 1, \cdots, M, \tag{52}$$

where $\delta \geq 0$, $\alpha_{k,n}^{(i)}$ are newly introduced auxiliary variables, and $\exp(\cdot)$ denotes the exponential function. Parameterization of $\gamma_{k,n}^{(i)}$ via (52) satisfies the non-negativity and the sum-to-one constraints.

We shall now directly update $\alpha_{k,n}^{(i)}$ instead of $\gamma_{k,n}^{(i)}$ by considering the following adaptation scheme. Using stochastic gradient descent to minimize (16), we obtain the multiple strategies LMS scheme as:

$$\begin{aligned} \alpha_{k,n+1}^{(i)} &= \alpha_{k,n}^{(i)} - v_{\alpha_k} \frac{\partial J_n^{\mathrm{MSE}}}{\partial \alpha_{k,n}^{(i)}} \\ &\approx \alpha_{k,n}^{(i)} + v_{\alpha_k} \frac{\exp(\alpha_{k,n}^{(i)})(e_{k,n} - e_{k,n}^{(i)})}{\sum_{j=1}^{M} \exp(\alpha_{k,n}^{(j)}) + M\delta} e_{k,n}. \end{aligned} \tag{53}$$

Specifically, when setting $\delta = 0$, (53) reduces to:

$$\alpha_{k,n+1}^{(i)} \approx \alpha_{k,n}^{(i)} + v_{\alpha_k} \gamma_{k,n}^{(i)} e_{k,n}(e_{k,n} - e_{k,n}^{(i)}) \tag{54}$$

To bound the dynamic of $\alpha_{k,n+1}^{(i)}$, parameters $\alpha_{k,n}^{(i)}$ are further required to be in interval $[-\alpha_0^+, \alpha_0^+]$ with $\alpha_0^+ > 0$.

VI. Discussion

A. Computation complexity and communication overhead

The computation complexity and communication overhead of the convex combination schemes result from those of the diffusion strategy layer and the combination layer. We shall consider the case of M component strategies $\mathrm{S}^{(i)}$.

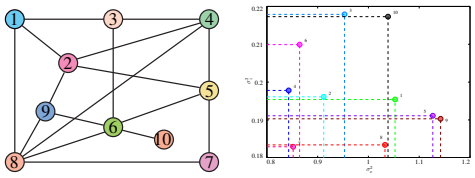
- Regarding the computation overhead, since convex combination schemes can be applied to any component strategies, we cannot describe all scenarios in a unified manner since the computation overhead of each component strategy is unknown. To alleviate this, we shall adopt the following description. We denote the computation cost of each component strategy by $q^{(i)}$, and we adopt the notation p_k to denote the computation overhead of node k at the combination layer. Then the total computational complexity q can be evaluated as:

$$q = \sum_{i=1}^{M} q^{(i)} + \sum_{k=1}^{N} p_k. \tag{55}$$

  Given any candidate diffusion strategy, $q^{(i)}$ is related to the filter length L and the total number of nodes N, as well as the complexity in evaluating the stochastic matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2$ and $\boldsymbol{C}$. Quantity p_k in convex combination schemes is related to the total number M of candidate diffusion strategies and the complexity in evaluating the exponential function in (52), or the sigmoid function (14). If the exponential function can be evaluated in advance and its values stored in a table, the computation overhead can be reduced greatly. As a conclusion, the total computation complexity q of convex combination schemes is larger than the sum of the computation complexities of all component strategies.
- We denote the communication cost of each component strategy by $h^{(i)}$. Since the combination schemes are conducted in a distributed manner at each node, there is no communication overhead at the combination layer. Therefore, the total communication cost h of the convex combination scheme can be evaluated as:

$$h = \sum_{i=1}^{M} h^{(i)}, \tag{56}$$

  which means that h is equal to the sum of the communication costs of all component strategies. Given any candidate diffusion strategy, $h^{(i)}$ is related to the filter length L and the network topology. The total communication complexity h can be reduced, for instance, by using compression coding methods before communicating with neighboring nodes [48], [49].



(a) (b)

Fig. 2. (a) Network topology; (b) Agent input and noise variances.

B. Comparison with the affine combination scheme

Without loss of generality, we focus on the case of two component strategies. For ease of comparison, we recall the update equation of the affine combination schemes in [36]:

$$\gamma_{k,n+1} = \gamma_{k,n} + v'_{\alpha_k} e_{k,n} \boldsymbol{u}_{k,n}^{\top} (\boldsymbol{w}_{k,n}^{(1)} - \boldsymbol{w}_{k,n}^{(2)}) \tag{57}$$

Compared with the affine combination scheme (57), the update equation of the convex combination scheme (17) is more complex, in particular because of the truncation operation and the nonlinear function used to evaluate coefficients $\gamma_{k,n+1}$. However, as derived in the current work, the convex combination scheme is more stable than the affine one.

VII. Simulation results

In this section, we present simulation results to illustrate the proposed convex combination schemes and theoretical results. All simulated curves were obtained by averaging over 100 Monte Carlo runs.

A. Validation of convex combination schemes

We considered a non-stationary system identification scenario where $\boldsymbol{w}_k^{\star}$ varies over time. The network consisted of 10 nodes with connection topology depicted in Fig. 2(a). The regressors were generated from a zero-mean multivariate Gaussian distribution with covariance matrix $\boldsymbol{R}_{u,k} = \sigma_{u,k}^2 \boldsymbol{I}_{60}$. The noise signals was generated from Gaussian distribution $\mathcal{N}(0, \sigma_{z,k}^2)$. Variances $\sigma_{u,k}^2$ and $\sigma_{z,k}^2$ were generated randomly as depicted in Fig. 2(b).

1) Combination of two diffusion LMS strategies: We considered two ATC diffusion LMS strategies with $\boldsymbol{C} = \boldsymbol{I}_N$ and $\boldsymbol{A}_1^{(i)} = \boldsymbol{I}_N$ as component strategies. For matrices $\boldsymbol{A}_2^{(i)}$, we considered two groups of settings: static matrices with $\boldsymbol{A}_2^{(1)} = \boldsymbol{I}_N$ and $\boldsymbol{A}_2^{(2)}$ generated from the averaging rule, and adaptive matrices with $\boldsymbol{A}_2^{(1)}, \boldsymbol{A}_2^{(2)}$ given by [15] and [50], respectively. The evolution of $\boldsymbol{w}_k^{\star}$ was divided into five stationary stages and four transient episodes. During the stationary stages, we set $\boldsymbol{w}_k^{\star}$ of each agent so that, from $n = 1$ to 1000, and from $n = 4500$ to 6000, the entire network pursued the same target. While from instant $n = 1500$ to 2500, from $n = 3000$ to 4000 and from $n = 6500$ to 8000, the network split to pursue 2, 4, and 6 targets, respectively. The transient episodes between two adjacent stationary stages were designed by using linear interpolation over 500 time instants. The results are plotted in Figs. 3 and 4.

In Fig. 3(a), as expected, both the power-normalized scheme and the sign-regressor scheme with static fusion matrices tend to the best component strategy at each stage. Their behavior

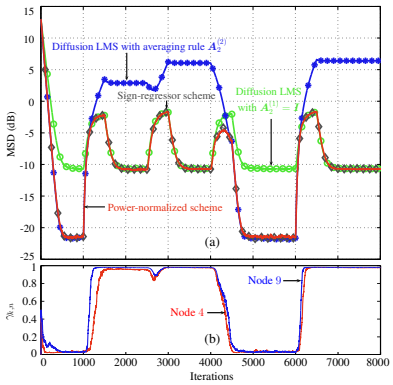


Fig. 3. Simulation results with static fusion matrices. (a) Network MSD learning curves; (b) Evolution of convex combination coefficient $\gamma_{k,n}$ for power-normalized scheme.

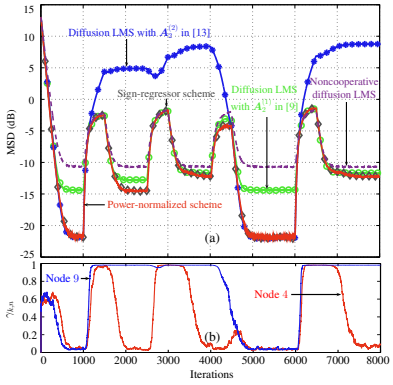


Fig. 4. Simulation results with adaptive fusion matrices.

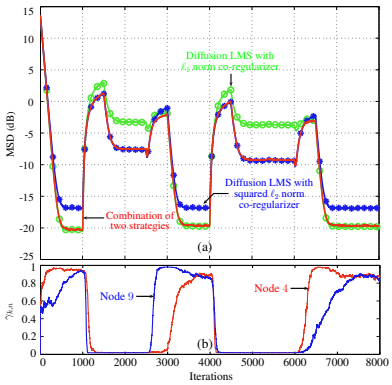


Fig. 5. Simulation results of the power-normalized scheme with the two component strategies in [14] and [21]. (a) Network MSD learning curves; (b) Evolution of the convex combination coefficients $\gamma_{k,n}$.

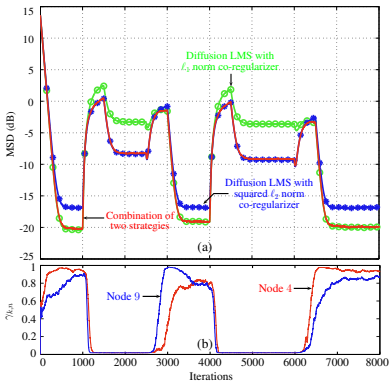


Fig. 6. Simulation results of the sign-regressor scheme with the two diffusion strategies in [14] and [21].

is similar to that of the non-cooperative diffusion LMS when estimating multiple targets, and similar to that of the diffusion LMS with averaging rule when pursuing the same target. The evolution of the convex combination coefficients in Fig. 3(b) validates the effectiveness of the combination scheme.

The results with adaptive fusion matrices are illustrated in Fig. 4. A similar learning behavior can be observed and a similar conclusion can be drawn.

2) Combination of two distinct diffusion strategies: We considered the multitask diffusion strategies for clustered networks proposed in [14] and [21], where the former uses squared ℓ_2-norm co-regularizer to promote cooperation within clusters, and the latter uses ℓ_1-norm co-regularizer. The simulation setting was the same to that in Section VII-A1, except that the 10 nodes were divided into three clusters to pursue three groups of different but related targets, and the targets for nodes within the same cluster were identical. For four stationary stages, the weight vectors $\boldsymbol{w}_{\mathcal{C}_i}^{\star}$ were generated according to $\boldsymbol{w}_{\mathcal{C}_i}^{\star} = \boldsymbol{w}_o + \delta_{\mathcal{C}_i}\boldsymbol{w}_{\mathcal{C}_i}$. When $\delta_{\mathcal{C}_i}$ for $i = 1, 2, 3$ are the same, the component strategy with ℓ_1-norm co-regularizer should perform better. Otherwise, the component strategy with squared ℓ_2-norm co-regularizer should have a better performance. We set the regularization strengths of both co-regularizers to 0.1, and a uniform $\boldsymbol{A}_2^{(i)}$ was used such that:

$$a_{2,\ell k}^{(i)} = |\mathcal{N}_k \cap \mathcal{C}(k)|^{-1}$$

The results are plotted in Figs. 5 and 6. Both the power-normalized scheme and the sign-regressor scheme have quite the same behavior as the optimum component strategy at the different stages.

3) Combination of multiple diffusion strategies: We shall now examine the performance of the combination scheme (53) via simulation. The simulation setting was similar to that in Section VII-A1, except that we changed the duration time of the stationary stages. We considered three component diffusion strategies for illustration purposes: $\boldsymbol{A}_2^{(2)}$ was obtained by the averaging rule with network step-sizes $\mu_k^{(2)}$ being set to 0.01, while $\boldsymbol{A}_2^{(1)}$ and $\boldsymbol{A}_2^{(3)}$ was set to the identity matrix corresponding to non-cooperative strategy, with the network step-sizes being set to $\mu_k^{(1)} = 0.01$ and $\mu_k^{(3)} = 0.0028$, respectively. When the entire network cooperates to pursue the same target, i.e., from time instant $n = 1$ to 800 and from instant $n = 5000$ to 6300 in Fig. 7(a), the diffusion strategy with averaging rule $\boldsymbol{A}_2^{(2)}$ should perform better. At other time instants when the network is split to pursue 2, 4 and 6 targets, the non-cooperative diffusion strategies with fusion matrices $\boldsymbol{A}_2^{(1)}$ and $\boldsymbol{A}_2^{(3)}$ should perform better. Moreover, since the step-sizes $\mu_k^{(i)}$ control the trade-off between convergence rate and steady-state performance, it should be helpful to adopt two different step-sizes and combine them, such as from time instant $n = 1000$ to 5000 when the network pursues multiple targets, to obtain a faster initial convergence speed and lower misadjustment at steady state simultaneously.

The results are illustrated in Fig. 7. As expected, at the different stages the proposed combination scheme tracks the

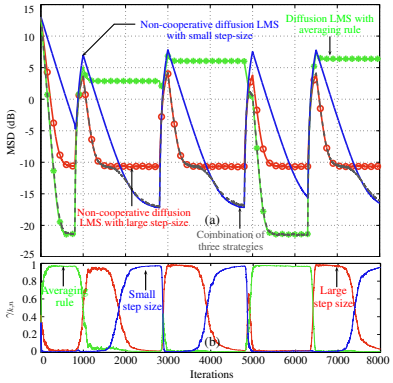


Fig. 7. Simulation results of combination scheme for three diffusion strategies. (a) Network MSD learning curves; (b) Evolution of combination coefficients $\gamma_{k,n}^{(i)}$ at agent 9.

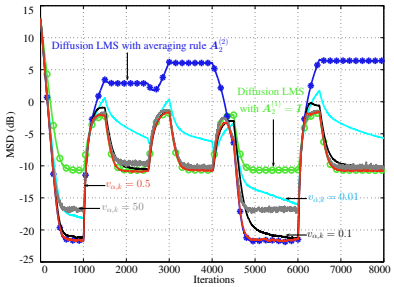


Fig. 8. Simulation results of the convex power-normalized scheme with different step-sizes $v_{\alpha,k}$.

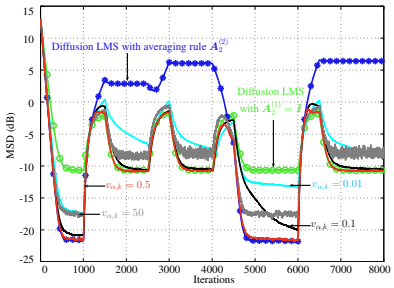


Fig. 9. Simulation results of the convex sign-regressor scheme with different step-sizes $v_{\alpha,k}$.

best component strategy, which is further validated from the evolution of affine combination coefficients of node 9 in Fig. 7(b). These results illustrate the effectiveness of combination framework in combining multiple component strategies.

4) Influence of the step-size $v_{\alpha,k}$: We shall now examine the influence of the parameters in the power-normalized scheme, sign-regressor scheme and the multiple strategies LMS scheme. Based on various experiments, we found that the performance of the combination schemes is not sensitive to the temporal smoothing factor η, to parameters α^+ or α_0^+, and to small-valued ε. We therefore set η to a typical value of 0.95 and α^+ to 4, α_0^+ to 2.5 and ε to 0.05. We shall also examine the influence of the step-size $v_{\alpha,k}$. The simulation settings were identical to those used in the first experiment, and we only considered static combination matrices.

The results are plotted in Figs. 8 to 10. For all these three schemes, a small-valued $v_{\alpha,k}$ results in a weak ability of tracking the best component, while a large-valued $v_{\alpha,k}$ leads

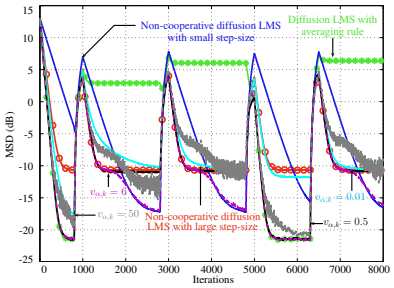


Fig. 10. Simulation results of the multiple strategies LMS scheme with different step-sizes $v_{\alpha,k}$.

to bias. Thus the value of $v_{\alpha,k}$ needs to be fine-tuned to ensure tracking performance and estimation accuracy.

B. Adjustments to improve performance

We shall now check the effectiveness of the three adjustment strategies proposed to improve the performance of the combination schemes. The simulation setting was similar to that in Section VII-A, except that we merely considered the first four stationary stages and the duration of stationary stages were set to different values. We considered the diffusion LMS with two static fusion matrices: non-cooperative $\boldsymbol{A}_2^{(1)} = \boldsymbol{I}$ with network step-size being set to 0.01, and the averaging rule for $\boldsymbol{A}_2^{(2)}$ with network step-size being set to 0.0015. Besides, for the three adjustment strategies, we set β_1, β_2 and ρ to 0.95, N_0 to 50 and N_0' to 170.

The results are illustrated in Fig. 11. As depicted by grey dashed lines in sub-figures (a), (b) and (c), the results with the three proposed adjustments are not worse than the original power-normalized scheme, and sometimes even offer a faster convergence rate and a lower misadjustment, such as from time instant $n = 1$ to $n = 2500$ and from $n = 9000$ to $n = 12000$. Besides, though obtaining similar steady-state MSDs, the convergence rate of the weight feedback adjustment is faster than those of the leakage adjustment and copying adjustment.

C. Validation of improvement with power-normalized scheme

We shall now validate the improvement in the EMSE obtained with the power-normalized scheme. As stated in **Theorem 1**, the improvement occurs when conditions $J_{\mathrm{ex},k,\infty}^{(1,2)} < J_{\mathrm{ex},k,\infty}^{(i)}$ and $\bar{\gamma}_{k,\infty} \in \left(1-\theta_k^+, \theta_k^+\right)$ are satisfied. We considered the same setting as that used in Fig. 4 of Section VII-A1, except that we only considered the second stationary stage. The simulation results are plotted in Fig. 12 and given in Table I. The EMSE learning curve of Fig. 12(a) validates the improvement in convex power-normalized scheme at steady state. The estimated steady-state EMSEs $J_{\mathrm{ex},k,\infty}^{(i)}$ with $i = 1, 2$ and steady-state cross-EMSE $J_{\mathrm{ex},k,\infty}^{(1,2)}$, averaged over the last 1500 iterations, are plotted in Fig. 12(b).As can be observed in Fig. 12(b), for several nodes the condition $J_{\mathrm{ex},k,\infty}^{(1,2)} < J_{\mathrm{ex},k,\infty}^{(i)}$ is satisfied. It can be further checked in Table I that $\bar{\gamma}_{k,\infty} \in \left(1-\theta_k^+, \theta_k^+\right)$ for nodes 1 and 6. All these results in Fig. 12 and Table I validate the improvement in the EMSE at steady state with the power-normalized scheme.

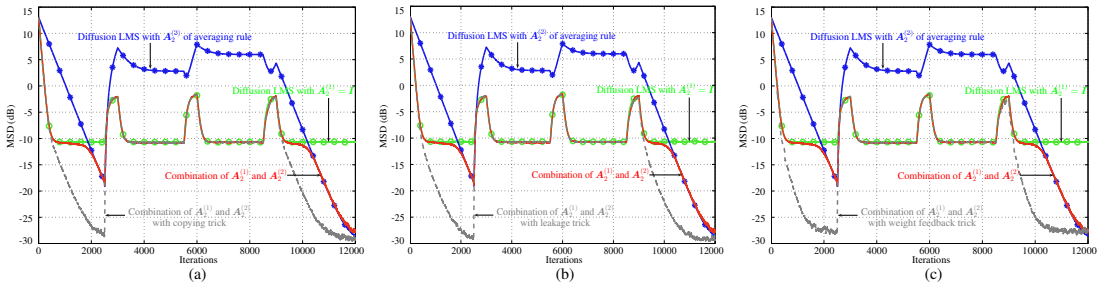


Fig. 11. Simulation results of convex power-normalized scheme with three adjustments. (a) Network MSD learning curves with copying weights; (b) Network MSD learning curves with leakage transfer; (c) Network MSD learning curves with weight feedback.

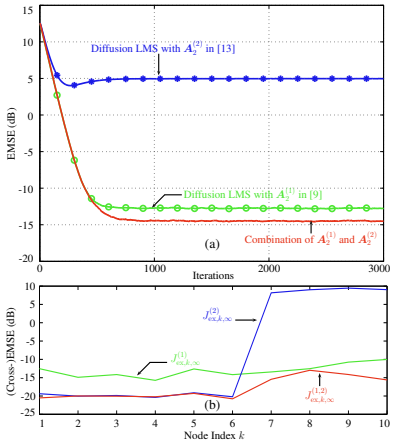


Fig. 12. Validation of the improvement obtained via the convex power-normalized scheme.

TABLE I

ESTIMATED STEADY-STATE VALUES $\bar{\gamma}_{k,\infty}$ AT EACH NODE. SINCE WE SET α^+ TO 4, $\theta_k^+ \triangleq \frac{1}{1+e^{-\alpha^+}} = 0.982$ AND $1-\theta_k^+ = 0.018$.

Node k	1	2–5	6	7–10
$\bar{\gamma}_{k,\infty}$	0.0541	0.018	0.0376	0.982

D. Theoretical model validation

As has been done in [36], to illustrate the theoretical results as well as to challenge the assumptions and approximations adopted in the theoretical analysis, we considered two networks with different connectivity parameters as described in Table II. Net1 consisted of 10 nodes with the network topology given in Fig. 2(a). Net2 was generated by splitting 20 nodes into seven fully connected clusters, with 3 nodes in each of the first six clusters and 2 nodes in the last cluster. These seven clusters were connected in chain, with a single edge connecting adjacent clusters: agent 3 (in cluster 1) was connected with agent 4 (in cluster 2), and agent 6 (in cluster 2) was connected with agent 7 (in cluster 3), and so on until agent 18 (in cluster 6) was connected to agent 19 (in cluster 7).

The unknown system coefficients to be estimated were of length $L = 2$. The regressors were generated from a zero-mean Gaussian distribution with covariance matrix $\boldsymbol{R}_{u,k} = \sigma_{u,k}^2 \boldsymbol{I}_L$ for white inputs and with

$$\boldsymbol{R}_{u,k} = \sigma_{u,k}^2 \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

TABLE II

NETWORK STATISTICS FOR THEORETICAL MODELS VALIDATION. $\boldsymbol{L}$ IS THE LAPLACIAN MATRIX ASSOCIATED WITH THE GRAPH (NETWORK), $\lambda_2(\boldsymbol{L})$ IS THE ALGEBRAIC CONNECTIVITY [51] OF GRAPH, SIZE IS THE NUMBER OF NODES, DENSITY IS THE NUMBER OF NON-ZERO ENTRIES OF THE ADJACENCY MATRIX OF GRAPH, AND DIAMETER IS THE MAXIMUM DISTANCE BETWEEN ANY TWO NODES [52].

Network	Size	Density	$\lambda_2(\boldsymbol{L})$	Diameter
Net1	10	44%	0.8576	3
Net2	20	17.25%	0.0439	13

for colored inputs. Variances $\sigma_{u,k}^2$ and $\sigma_{z,k}^2$ at each agent were generated randomly. For white inputs, by varying $\sigma_{z,k}^2$, we changed the signal-to-noise ratio (SNR) [53] to two levels as described in Table III. For illustration purpose, we plot $\sigma_{u,k}^2$ and $\sigma_{z,k}^2$ of each agent with SNR1 in Fig. 2(b). The power-normalized scheme was run with network step-sizes being set to 0.01 and 0.004, respectively.

TABLE III

TWO SNR LEVELS IN DECIBEL (DB) FOR THEORETICAL MODELS VALIDATION. SINCE SNRS VARY FROM ONE NODE TO ANOTHER, WE ENUMERATE THE MAXIMUM, MINIMUM AND MEAN VALUES.

SNR Level	Maximum	Minimum	Mean
SNR1	7.079	5.4439	6.2272
SNR2	-8.9416	-10.5767	-9.7934

We first validate the theoretical results for the mean and mean-square behaviors of $\gamma_{k,n}$, the transient and steady-state MSDs of each component strategies, and the cross-MSD of two component strategies over the entire network defined by $\text{MSD}_{\text{cross}} \triangleq \frac{1}{N}\mathbb{E}\{\boldsymbol{v}_n^{(1)\top}\boldsymbol{v}_n^{(2)}\}$. All these quantities are necessary in evaluating the MSD behavior of the convex power-normalized scheme. Then by using these results, we evaluate the theoretical MSD behaviors of the convex power-normalized scheme. All results are plotted in Figs. 13 to 19.

We observe in Figs. 13 to 15 that the simulated and theoretical transient values, and theoretical steady-state values of network cross-MSD and MSDs of two component diffusion strategies are superimposed, respectively, which illustrates the accuracy of the theoretical analysis for MSD at each component. Since the analysis of $\mathbb{E}\{\gamma_{k,n}\}$ and $\mathbb{E}\{\gamma_{k,n}^2\}$ are based on the Taylor series expansion, there are biases between simulated and theoretical transient values of $\mathbb{E}\{\gamma_{k,n}\}$ and $\mathbb{E}\{\gamma_{k,n}^2\}$. However the theoretical results for the power-normalized scheme are still acceptable and satisfying.

The results of the power-normalized diffusion for white Gaussian inputs with Net2 and SNR1 are plotted in Fig. 16.

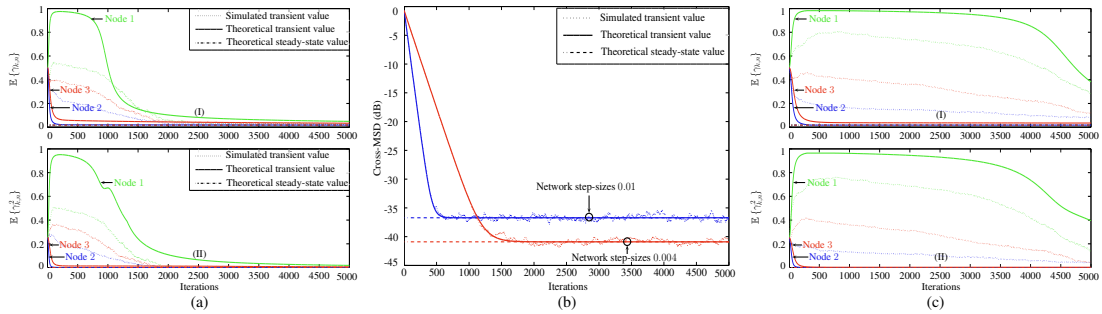


Fig. 13. Illustration of simulation results (model vs. Monte Carlo) for the convex power-normalized scheme. Transient and steady-state values of $\mathbb{E}\{\gamma_{k,n}\}$ derived in (123) and (65) (*top*), as well as these of $\mathbb{E}\{\gamma_{k,n}^2\}$ derived in (125) and (134) (*bottom*) for network step-size 0.01 (a) and 0.004 (c); (b) Transient and steady-state cross-MSDs derived in (104) and (107).

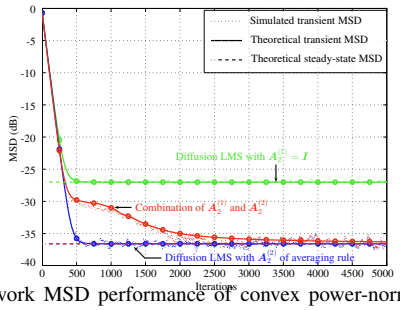


Fig. 14. Network MSD performance of convex power-normalized scheme (model vs. Monte Carlo) with network step-size 0.01 for white input in `Net1` and `SNR1`.

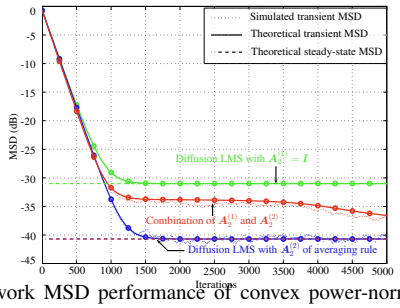


Fig. 15. Network MSD performance of convex power-normalized scheme (model vs. Monte Carlo) with network step-size 0.004 for white input in `Net1` and `SNR1`.

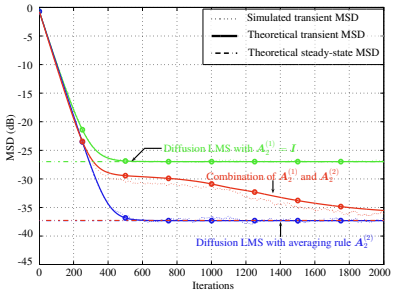


Fig. 16. Network performance with the power-normalized scheme (model vs. Monte Carlo) with network step-size 0.01 for white input in `Net2` and `SNR1`.

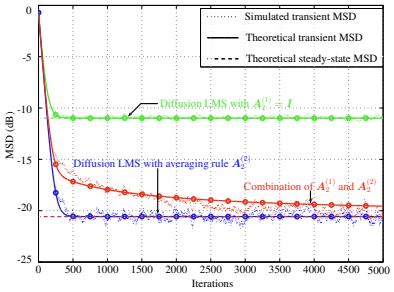


Fig. 17. Network performance with the power-normalized scheme (model vs. Monte Carlo) with network step-size 0.01 for white input in `Net1` and `SNR2`.

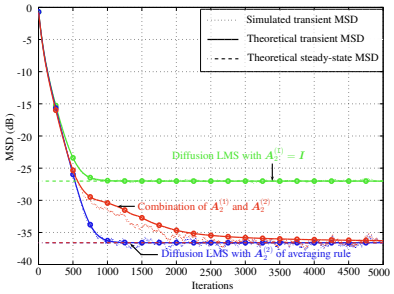


Fig. 18. Network MSD performance of convex power-normalized scheme (model vs. Monte Carlo) with network step-size 0.01 for colored input in `Net1` and `SNR1`.

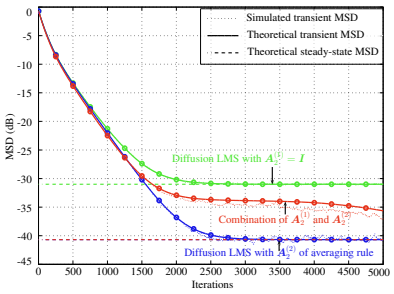


Fig. 19. Network MSD performance of convex power-normalized scheme (model vs. Monte Carlo) with network step-size 0.004 for colored input in `Net1` and `SNR1`.

The results with `Net1` and `SNR2` are plotted in Fig. 17. Together with Figs. 14 and 15, all these results validate the accuracy of theoretical analyses under different SNR conditions and network connectivity parameters.

The results for moderately colored inputs with `Net1` and `SNR1` are provided in Fig. 18–Fig. 19. Though assumption **A2** is violated, the superimposition of simulated and theoretical curves validates the accuracy of the theoretical results for sufficiently small step-sizes.

In addition, observe in Figs 14 to 19 that the combination scheme performs worse than the best component strategy. On the one hand, the performance of the power-normalized scheme is closely related to the system parameters, and when L is large, it is easier for the power-normalized scheme to track the best component strategy. Unfortunately, in the

simulation setting of Figs. 14 to 19, for the purpose of saving computations, we set parameter $L = 2$. On the other hand, we have proved in **Theorem 1** that the power-normalized scheme is universal at steady state. The Monte-Carlo curves coincide with the theoretical results.

VIII. Conclusions

Combining diffusion strategies enables a network to reach a better performance. In this paper, we proposed several schemes for the convex combination of multiple component diffusion strategies, as well as several adjustments to further improve the performance. We conducted theoretical analysis for the convex power-normalized scheme. Based on the theoretical results, we conclude that the convex power-normalized scheme is universal at steady state, meanwhile the mean and mean-square stabilities of power-normalized scheme require only the stability of its two component diffusion strategies. Thus the convex power-normalized schemes are more stable than affine combination schemes. Several open problems still have to be addressed. For instance, it would be interesting to conduct a theoretical analysis for the convex sign-regressor scheme. It would also be interesting to explore other combination frameworks and schemes.

Appendix A
Proof of Universality Analysis Result

We first consider two extreme situations of $\alpha_{k,n}$:

- **Situation 1**: If $\lim_{n\to\infty}\mathbb{E}\{\alpha_{k,n}\} = \alpha^+$, we conclude that $\alpha_{k,n} \to \alpha^+$ with $n \to \infty$ is almost sure. Then we have $\gamma_{k,n} \to \theta_k^+ \triangleq \frac{1}{1+e^{-\alpha^+}} \approx 1$, and $J_{\text{ex},k,\infty} \approx J_{\text{ex},k,\infty}^{(1)}$.
- **Situation 2**: If $\lim_{n\to\infty}\mathbb{E}\{\alpha_{k,n}\} = -\alpha^+$, we deduce that $\alpha_{k,n} \to -\alpha^+$ as $n \to \infty$ in a high probability. Then we have $\gamma_{k,n} \to \frac{1}{1+e^{\alpha^+}} = 1-\theta_k^+ \approx 0$, and $J_{\text{ex},k,\infty} \approx J_{\text{ex},k,\infty}^{(2)}$.

Based on the above results, to evaluate the EMSE after combination at steady state, it is necessary to examine the limiting behavior of $\mathbb{E}\{\alpha_{k,n}\}$. To do so, taking the expectation of (17), we have:

$$\mathbb{E}\{\alpha_{k,n+1}\} \approx \big[\mathbb{E}\{\alpha_{k,n} + \mu'_{\alpha_k}\gamma_{k,n}(1-\gamma_{k,n}) e_{k,n}\boldsymbol{u}_{k,n}^\top(\boldsymbol{w}_{k,n}^{(1)} - \boldsymbol{w}_{k,n}^{(2)})\}\big]_{-\alpha^+}^{\alpha^+}, \quad (58)$$

where we have exchanged the order of expectation and truncation operations to simplify the derivation. This approximation is reasonable, since the likelihood of $\alpha_{k,n+1}$ to be greater than α^+ or less than $-\alpha^+$ before truncation is small due to the existence of factor $\gamma_{k,n}(1-\gamma_{k,n})$ in the update equation. From (9) and (13), we have:

$$\boldsymbol{u}_{k,n}^\top(\boldsymbol{w}_{k,n}^{(1)} - \boldsymbol{w}_{k,n}^{(2)}) = \tilde{e}_{k,n}^{(2)} - \tilde{e}_{k,n}^{(1)} \quad (59)$$

$$e_{k,n} = \gamma_{k,n}\tilde{e}_{k,n}^{(1)} + (1-\gamma_{k,n})\tilde{e}_{k,n}^{(2)} + z_{k,n}. \quad (60)$$

Under assumptions **A1** and **A2**, substituting (59), (60) into (58), and taking the limit with $n \to \infty$, we obtain:

$$\mathbb{E}\{\alpha_{k,n+1}\} = \big[\mathbb{E}\{\alpha_{k,n}\} - \bar{v}_{\alpha_{k,n}}\mathbb{E}\{\gamma_{k,n}^2(1-\gamma_{k,n})\}\triangle J_{k,\infty}^{(1)} + \bar{v}_{\alpha_{k,n}}\mathbb{E}\{\gamma_{k,n}(1-\gamma_{k,n})^2\}\triangle J_{k,\infty}^{(2)}\big]_{-\alpha^+}^{\alpha^+} \text{ with } n\to\infty, \quad (61)$$

where $\bar{v}_{\alpha_{k,n}} \triangleq \mathbb{E}\{\frac{v_{\alpha_k}}{\varepsilon+p_{k,n}}\}$, and $\triangle J_{k,\infty}^{(i)} \triangleq J_{\text{ex},k,\infty}^{(i)} - J_{\text{ex},k,\infty}^{(1,2)}$ measures the difference between the steady-state EMSE and steady-state cross-EMSE, with the latter being defined by $J_{\text{ex},k,\infty}^{(1,2)} \triangleq \lim_{n\to\infty}\mathbb{E}\{\tilde{e}_{k,n}^{(1)}\tilde{e}_{k,n}^{(2)}\}$. From Cauchy-Schwartz inequality and according to the relations between $J_{\text{ex},k,\infty}^{(1,2)}$ and $J_{\text{ex},k,\infty}^{(i)}$, we further divide the problem of evaluating the limiting behavior of $\mathbb{E}\{\alpha_{k,n}\}$ into three cases:

- **Case 1**: $J_{\text{ex},k,\infty}^{(1)} \le J_{\text{ex},k,\infty}^{(1,2)} \le J_{\text{ex},k,\infty}^{(2)}$. We then have $\triangle J_{k,\infty}^{(1)} \le 0$ and $\triangle J_{k,\infty}^{(2)} \ge 0$. Since $\gamma_{k,n}$ and $1-\gamma_{k,n}$ lie in the interval $[1-\theta_k^+, \theta_k^+]$, both $\mathbb{E}\{\gamma_{k,n}^2(1-\gamma_{k,n})\}$ and $\mathbb{E}\{\gamma_{k,n}(1-\gamma_{k,n})^2\}$ are lower bounded by $\hat{b}_k \triangleq \theta_k^+(1-\theta_k^+)^2$. Define $\bar{b}_k \triangleq \lim_{n\to\infty}\bar{v}_{\alpha_{k,n}}$. Then (61) writes to:

$$\mathbb{E}\{\alpha_{k,n+1}\} \ge \big[\mathbb{E}\{\alpha_{k,n}\} + b_k\big]_{-\alpha^+}^{\alpha^+} \text{ with } n\to\infty \quad (62)$$

  where $b_k \triangleq \hat{b}_k\bar{b}_k(\triangle J_{k,\infty}^{(2)} - \triangle J_{k,\infty}^{(1)}) > 0$. It follows from (62) that the unique stationary point of $\mathbb{E}\{\alpha_{k,n+1}\}$ with $n \to \infty$ is α^+. According to the previous conclusion drawn in **Situation 1**, we conclude that $J_{\text{ex},k,\infty} \approx J_{\text{ex},k,\infty}^{(1)}$.
- **Case 2**: $J_{\text{ex},k,\infty}^{(1)} \ge J_{\text{ex},k,\infty}^{(1,2)} \ge J_{\text{ex},k,\infty}^{(2)}$. We then have $\triangle J_{k,\infty}^{(1)} \ge 0$ and $\triangle J_{k,\infty}^{(2)} \le 0$. Then (61) writes to:

$$\mathbb{E}\{\alpha_{k,n+1}\} \le \big[\mathbb{E}\{\alpha_{k,n}\} - b'_k\big]_{-\alpha^+}^{\alpha^+} \text{ with } n\to\infty \quad (63)$$

  where $b'_k \triangleq \hat{b}_k\bar{b}_k(\triangle J_{k,\infty}^{(1)} - \triangle J_{k,\infty}^{(2)}) > 0$. Thus the unique stationary point of $\mathbb{E}\{\alpha_{k,n+1}\}$ with $n \to \infty$ is $-\alpha^+$. According to the conclusion drawn in **Situation 2**, we have $J_{\text{ex},k,\infty} \approx J_{\text{ex},k,\infty}^{(2)}$.
- **Case 3**: $J_{\text{ex},k,\infty}^{(i)} > J_{\text{ex},k,\infty}^{(1,2)}$ for $i = 1, 2$. We then have $\triangle J_{k,\infty}^{(i)} > 0$. Iteration (61) converges to the stationary point if and only if:

$$\mathbb{E}\{\gamma_{k,n}^2(1-\gamma_{k,n})\}\triangle J_{k,\infty}^{(1)} = \mathbb{E}\{\gamma_{k,n}(1-\gamma_{k,n})^2\}\triangle J_{k,\infty}^{(2)} \text{ with } n\to\infty. \quad (64)$$

  To make equation (64) tractable, by using **Ap7** which assumes that the variance of $\gamma_{k,n}$ is small with $n \to \infty$, we obtain:

$$\bar{\gamma}_{k,\infty} \triangleq \lim_{n\to\infty}\mathbb{E}\{\gamma_{k,n}\} = \left[\frac{\triangle J_{k,\infty}^{(2)}}{\triangle J_{k,\infty}^{(1)} + \triangle J_{k,\infty}^{(2)}}\right]_{1-\theta_k^+}^{\theta_k^+}. \quad (65)$$

  Since the steady-state EMSE of node k after combination writes to:

$$J_{\text{ex},k,\infty} = \bar{\gamma}_{k,\infty}^2 J_{\text{ex},k,\infty}^{(1)} + (1-\bar{\gamma}_{k,\infty})^2 J_{\text{ex},k,\infty}^{(2)} + 2\bar{\gamma}_{k,\infty}(1-\bar{\gamma}_{k,\infty})J_{\text{ex},k,\infty}^{(1,2)}, \quad (66)$$

  by substituting $\bar{\gamma}_{k,\infty}$ of (65) without truncation operation into (66) and after some algebraic manipulations, we arrive at:

$$J_{\text{ex},k,\infty} = J_{\text{ex},k,\infty}^{(1,2)} + \frac{\triangle J_{k,\infty}^{(1)}\triangle J_{k,\infty}^{(2)}}{\triangle J_{k,\infty}^{(1)} + \triangle J_{k,\infty}^{(2)}}. \quad (67)$$

Since $J_{\text{ex},k,\infty}^{(i)} > J_{\text{ex},k,\infty}^{(1,2)}$ for $i = 1,2$, we have $\frac{\triangle J_{k,\infty}^{(1)} \triangle J_{k,\infty}^{(2)}}{\triangle J_{k,\infty}^{(1)} + \triangle J_{k,\infty}^{(2)}} < \triangle J_{k,\infty}^{(i)}$, and we conclude:

$$J_{\text{ex},k,\infty} < J_{\text{ex},k,\infty}^{(i)}, \tag{68}$$

which means that there is an improvement of EMSE after combination. Since the cross-EMSE is lower than EMSE of each component, we may extract extra information through combination, which brings the gain in EMSE. As for $\bar{\gamma}_{k,\infty} = \theta_k^+$ or $1-\theta_k^+$ obtained with truncation, we have derived in **Situation 1** and **Situation 2** that

$$J_{\text{ex},k,\infty} \approx \min\{J_{\text{ex},k,\infty}^{(1)}, J_{\text{ex},k,\infty}^{(2)}\}. \tag{69}$$

For the steady-state EMSE of the entire network defined by:

$$J_{\text{ex,net},\infty} \triangleq \sum_{k=1}^{N} J_{\text{ex},k,\infty}, \tag{70}$$

we have:

$$J_{\text{ex,net},\infty} \leq \min\left\{J_{\text{ex,net},\infty}^{(1)}, J_{\text{ex,net},\infty}^{(2)}\right\}, \tag{71}$$

which means that the EMSE of diffusion network after combination is no worse than that of the best component strategy, leading to the universality of convex power-normalized scheme at steady state.

Appendix B
Iteration of $\mathbb{E}\{\boldsymbol{v}_{n+1}^{(i)}\}$

Under assumption **A1** and along the lines developed in [15], we have:

$$\boldsymbol{v}_{n+1}^{(i)} = \boldsymbol{B}_n^{(i)} \boldsymbol{v}_n^{(i)} + \boldsymbol{g}_n^{(i)} - \boldsymbol{r}_n^{(i)}, \tag{72}$$

$$\mathbb{E}\{\boldsymbol{v}_{n+1}^{(i)}\} = \overline{\boldsymbol{B}}^{(i)} \mathbb{E}\{\boldsymbol{v}_n^{(i)}\} - \overline{\boldsymbol{r}}^{(i)}, \tag{73}$$

with quantities defined by:

$$\boldsymbol{B}_n^{(i)} = \boldsymbol{\mathcal{A}}_2^{(i)\top}\left(\boldsymbol{I}_{NL} - \boldsymbol{U}^{(i)}\boldsymbol{H}_n^{(i)}\right)\boldsymbol{\mathcal{A}}_1^{(i)\top}, \tag{74}$$

$$\overline{\boldsymbol{B}}^{(i)} = \boldsymbol{\mathcal{A}}_2^{(i)\top}\left(\boldsymbol{I}_{NL} - \boldsymbol{U}^{(i)}\overline{\boldsymbol{H}}^{(i)}\right)\boldsymbol{\mathcal{A}}_1^{(i)\top}, \tag{75}$$

$$\boldsymbol{\mathcal{A}}_j^{(i)} = \boldsymbol{A}_j^{(i)} \otimes \boldsymbol{I}_L, \forall j = 1,2, \tag{76}$$

$$\boldsymbol{U}^{(i)} = \text{diag}\left\{\mu_1^{(i)}, \cdots, \mu_N^{(i)}\right\} \otimes \boldsymbol{I}_L, \tag{77}$$

$$\boldsymbol{H}_n^{(i)} = \text{diag}\left\{\sum_{\ell\in\mathcal{N}_k} c_{\ell k}^{(i)} \boldsymbol{x}_{\ell,n}\boldsymbol{x}_{\ell,n}^\top\right\}_{k=1}^{N}, \tag{78}$$

$$\overline{\boldsymbol{H}}^{(i)} = \text{diag}\{\boldsymbol{R}_1^{(i)}, \cdots, \boldsymbol{R}_N^{(i)}\}, \tag{79}$$

$$\boldsymbol{R}_k^{(i)} \triangleq \sum_{\ell\in\mathcal{N}_k} c_{\ell k}^{(i)} \boldsymbol{R}_{u,\ell}, \tag{80}$$

$$\boldsymbol{g}_n^{(i)} = \boldsymbol{\mathcal{A}}_2^{(i)\top}\boldsymbol{U}^{(i)}\boldsymbol{p}_{zx,n}^{(i)}, \tag{81}$$

$$\boldsymbol{p}_{zx,n}^{(i)} = \text{col}\left\{\sum_{\ell\in\mathcal{N}_k} c_{\ell k}^{(i)} \boldsymbol{x}_{\ell,n} z_{\ell,n}\right\}_{k=1}^{N}, \tag{82}$$

$$\boldsymbol{h}_{u,n}^{(i)} = \text{col}\left\{\sum_{\ell\in\mathcal{N}_k} c_{\ell k}^{(i)} \boldsymbol{x}_{\ell,n}\boldsymbol{x}_{\ell,n}^\top\left(\boldsymbol{w}_k^\star - \boldsymbol{w}_\ell^\star\right)\right\}_{k=1}^{N}, \tag{83}$$

$$\overline{\boldsymbol{h}}_u^{(i)} = \text{col}\left\{\sum_{\ell\in\mathcal{N}_k} c_{\ell k}^{(i)} \boldsymbol{R}_{u,k}\left(\boldsymbol{w}_k^\star - \boldsymbol{w}_\ell^\star\right)\right\}_{k=1}^{N}, \tag{84}$$

$$\boldsymbol{r}_n^{(i)} \triangleq \underbrace{\boldsymbol{\mathcal{A}}_2^{(i)\top}\boldsymbol{U}^{(i)}\boldsymbol{h}_{u,n}^{(i)}}_{\boldsymbol{r}_{\boldsymbol{u},n}^{(i)}} - \underbrace{\left[\boldsymbol{\mathcal{A}}_2^{(i)\top}\left(\boldsymbol{I}_{NL} - \boldsymbol{U}^{(i)}\boldsymbol{H}_n^{(i)}\right)\left(\boldsymbol{\mathcal{A}}_1^{(i)\top} - \boldsymbol{I}_{NL}\right) + \left(\boldsymbol{\mathcal{A}}_2^{(i)\top} - \boldsymbol{I}_{NL}\right)\right]\boldsymbol{w}^\star}_{\boldsymbol{r}_{\boldsymbol{w},n}^{(i)}} \tag{85}$$

$$\overline{\boldsymbol{r}}^{(i)} \triangleq \mathbb{E}\{\boldsymbol{r}_n^{(i)}\} = \overline{\boldsymbol{r}}_{\boldsymbol{u}}^{(i)} - \overline{\boldsymbol{r}}_{\boldsymbol{w}}^{(i)}. \tag{86}$$

Expression (73) helps to evaluate the iteration of the mean behavior $\mathbb{E}\{\boldsymbol{v}_{n+1}^{(i)}\}$.

Appendix C
Evolution of $\mathbb{E}\{\|\boldsymbol{v}_{n+1}\|_{\boldsymbol{\Sigma}}^2\}$

Using (40), we have:

$$\mathbb{E}\{\|\boldsymbol{v}_{n+1}\|_{\boldsymbol{\Sigma}}^2\} = 2\mathbb{E}\{\boldsymbol{v}_{n+1}^{(1)\top}\boldsymbol{\Gamma}_{n+1}\boldsymbol{\Sigma}(\boldsymbol{I}_{NL} - \boldsymbol{\Gamma}_{n+1})\boldsymbol{v}_{n+1}^{(2)}\} + \mathbb{E}\{\|(\boldsymbol{I}_{NL} - \boldsymbol{\Gamma}_{n+1})\boldsymbol{v}_{n+1}^{(2)}\|_{\boldsymbol{\Sigma}}^2\} + \mathbb{E}\{\|\boldsymbol{\Gamma}_{n+1}\boldsymbol{v}_{n+1}^{(1)}\|_{\boldsymbol{\Sigma}}^2\}. \tag{87}$$

Define:

$$\boldsymbol{\Sigma}_{n+1}^{(1)} \triangleq \mathbb{E}\{\boldsymbol{\Gamma}_{n+1}^\top\boldsymbol{\Sigma}\boldsymbol{\Gamma}_{n+1}\} \tag{88}$$

$$\boldsymbol{\Sigma}_{n+1}^{(2)} \triangleq \mathbb{E}\{(\boldsymbol{I}_{NL} - \boldsymbol{\Gamma}_{n+1})^\top\boldsymbol{\Sigma}(\boldsymbol{I}_{NL} - \boldsymbol{\Gamma}_{n+1})\} \tag{89}$$

$$\boldsymbol{\sigma}_{n+1}^{(i)} = \text{vec}\{\boldsymbol{\Sigma}_{n+1}^{(i)}\}. \tag{90}$$

Under approximation **Ap3**, the last two terms on the RHS of (87) can be written in compact form as $\mathbb{E}\{\boldsymbol{v}_{n+1}^{(i)\top}\boldsymbol{\Sigma}_{n+1}^{(i)}\boldsymbol{v}_{n+1}^{(i)}\}$ for $i = 1,2$, which are evaluated in [36] under **Ap4** as:

$$\mathbb{E}\left\{\|\boldsymbol{v}_{n+1}^{(i)}\|_{\boldsymbol{\sigma}_{n+1}^{(i)}}^2\right\} = \mathbb{E}\left\{\|\boldsymbol{v}_n^{(i)}\|_{\boldsymbol{K}^{(i)}\boldsymbol{\sigma}_{n+1}^{(i)}}^2\right\} + \left[\text{vec}\{\boldsymbol{G}^{(i)\top}\}\right]^\top\boldsymbol{\sigma}_{n+1}^{(i)} + \boldsymbol{f}\left(\overline{\boldsymbol{r}}^{(i)}, \boldsymbol{\Sigma}_{n+1}^{(i)}, \mathbb{E}\{\boldsymbol{v}_n^{(i)}\}\right), \tag{91}$$

where $\|\cdot\|_{\boldsymbol{\Sigma}_{n+1}^{(i)}}^2$ and $\|\cdot\|_{\boldsymbol{\sigma}_{n+1}^{(i)}}^2$ are used interchangeably, with:

$$\boldsymbol{f}\left(\overline{\boldsymbol{r}}^{(i)}, \boldsymbol{\Sigma}_{n+1}^{(i)}, \mathbb{E}\{\boldsymbol{v}_n^{(i)}\}\right) \triangleq \|\overline{\boldsymbol{r}}^{(i)}\|_{\boldsymbol{\Sigma}_{n+1}^{(i)}}^2 - 2\overline{\boldsymbol{r}}^{(i)\top}\boldsymbol{\Sigma}_{n+1}^{(i)}\overline{\boldsymbol{B}}^{(i)}\mathbb{E}\{\boldsymbol{v}_n^{(i)}\} \tag{92}$$

$$\boldsymbol{K}^{(i)} \approx \overline{\boldsymbol{B}}^{(i)\top} \otimes \overline{\boldsymbol{B}}^{(i)\top} \tag{93}$$

$$\boldsymbol{G}^{(i)} \triangleq \mathbb{E}\{\boldsymbol{g}_n^{(i)}\boldsymbol{g}_n^{(i)\top}\}. \tag{94}$$

Expression (91) was derived in Appendix D of [36]. For the first term on RHS of (87), we have:

$$\mathbb{E}\{\boldsymbol{v}_{n+1}^{(1)\top}\boldsymbol{\Sigma}_{\text{x},n+1}\boldsymbol{v}_{n+1}^{(2)}\} = \mathbb{E}\{\boldsymbol{v}_n^{(1)\top}\boldsymbol{\Sigma}_{\text{xc},n+1}\boldsymbol{v}_n^{(2)}\} + \left[\text{vec}\{\boldsymbol{G}_{\text{x}}^\top\}\right]^\top\boldsymbol{\sigma}_{\text{x},n+1} + \boldsymbol{f}_{\text{x}}\left(\overline{\boldsymbol{r}}^{(1)}, \overline{\boldsymbol{r}}^{(2)}, \boldsymbol{\Sigma}_{\text{x},n+1}, \mathbb{E}\{\boldsymbol{v}_n^{(1)}\}, \mathbb{E}\{\boldsymbol{v}_n^{(2)}\}, \overline{\boldsymbol{B}}^{(1)}, \overline{\boldsymbol{B}}^{(2)}\right), \tag{95}$$

where

$$\boldsymbol{f}_{\text{x}}\left(\overline{\boldsymbol{r}}^{(1)}, \overline{\boldsymbol{r}}^{(2)}, \boldsymbol{\Sigma}_{\text{x},n+1}, \mathbb{E}\{\boldsymbol{v}_n^{(1)}\}, \mathbb{E}\{\boldsymbol{v}_n^{(2)}\}, \overline{\boldsymbol{B}}^{(1)}, \overline{\boldsymbol{B}}^{(2)}\right) \triangleq \overline{\boldsymbol{r}}^{(1)\top}\boldsymbol{\Sigma}_{\text{x},n+1}\overline{\boldsymbol{r}}^{(2)} - \mathbb{E}\{\boldsymbol{v}_n^{(1)\top}\}\overline{\boldsymbol{B}}^{(1)\top}\boldsymbol{\Sigma}_{\text{x},n+1}\overline{\boldsymbol{r}}^{(2)} - \overline{\boldsymbol{r}}^{(1)\top}\boldsymbol{\Sigma}_{\text{x},n+1}\overline{\boldsymbol{B}}^{(2)}\mathbb{E}\{\boldsymbol{v}_n^{(2)}\} \tag{96}$$

with

$$\boldsymbol{\Sigma}_{\mathrm{x},n+1} \triangleq \mathbb{E}\{\boldsymbol{\Gamma}_{n+1}\boldsymbol{\Sigma}(\boldsymbol{I}_{NL}-\boldsymbol{\Gamma}_{n+1})\}, \tag{97}$$

$$\boldsymbol{\Sigma}_{\mathrm{xc},n+1} \triangleq \mathrm{vec}^{-1}\{\boldsymbol{K}_{\mathrm{x}}\boldsymbol{\sigma}_{\mathrm{x},n+1}\}, \tag{98}$$

$$\boldsymbol{\sigma}_{\mathrm{x},n+1} \triangleq \mathrm{vec}\{\boldsymbol{\Sigma}_{\mathrm{x},n+1}\}, \tag{99}$$

$$\boldsymbol{K}_{\mathrm{x}} \approx \overline{\boldsymbol{B}}^{(2)\top} \otimes \overline{\boldsymbol{B}}^{(1)\top}, \tag{100}$$

$$\boldsymbol{G}_{\mathrm{x}} \triangleq \mathbb{E}\{\boldsymbol{g}_n^{(2)}\boldsymbol{g}_n^{(1)\top}\}, \tag{101}$$

and $\mathrm{vec}^{-1}\{\cdot\}$ is the inverse vectorization operator. Expression (87) helps evaluate the transient mean-square behavior $\mathbb{E}\{\|\boldsymbol{v}_{n+1}\|_{\boldsymbol{\Sigma}}^2\}$, which is important in setting model parameters in practice.

Appendix D
Recursions for Evaluating Transient MSD

Recursion for evaluating $\mathbb{E}\{\|\boldsymbol{v}_n^{(i)}\|^2_{\boldsymbol{K}^{(i)}\boldsymbol{\sigma}_{n+1}^{(i)}}\}$, that is the last two terms of (87) compactly, is given in [15] as:

$$\begin{aligned}\xi_{n+1}^{(i)} &= \xi_n^{(i)} + \big[\big(\mathrm{vec}\{\boldsymbol{G}^{(i)\top}\}\big)^{\top}(\boldsymbol{K}^{(i)})^n\boldsymbol{\sigma}^{(i)} \\ &\quad + \|\overline{\boldsymbol{r}}^{(i)}\|^2_{(\boldsymbol{K}^{(i)})^n\boldsymbol{\sigma}^{(i)}} - \|\boldsymbol{v}_0^{(i)}\|^2_{(\boldsymbol{I}-\boldsymbol{K}^{(i)})(\boldsymbol{K}^{(i)})^n\boldsymbol{\sigma}^{(i)}} \\ &\quad - 2\big(\boldsymbol{\Lambda}_n^{(i)} + (\overline{\boldsymbol{B}}^{(i)}\mathbb{E}\{\boldsymbol{v}_n^{(i)}\})^{\top} \otimes \overline{\boldsymbol{r}}^{(i)\top}\big)\boldsymbol{\sigma}^{(i)}\big],\end{aligned} \tag{102}$$

and

$$\boldsymbol{\Lambda}_{n+1}^{(i)} = \boldsymbol{\Lambda}_n^{(i)}\boldsymbol{K}^{(i)} + \big((\overline{\boldsymbol{B}}^{(i)}\mathbb{E}\{\boldsymbol{v}_n^{(i)}\})^{\top} \otimes \overline{\boldsymbol{r}}^{(i)\top}\big)\big(\boldsymbol{K}^{(i)}-\boldsymbol{I}\big) \tag{103}$$

with $\boldsymbol{\Lambda}_0^{(i)} = \boldsymbol{0}_{1\times(NL)^2}$, $\xi_{n+1}^{(i)} = \mathbb{E}\{\|\boldsymbol{v}_{n+1}^{(i)}\|^2_{\boldsymbol{K}^{(i)}\boldsymbol{\sigma}_{n+1}^{(i)}}\}$, $\xi_0^{(i)} = \|\boldsymbol{v}_0^{(i)}\|^2_{\boldsymbol{K}^{(i)}\boldsymbol{\sigma}_{n+1}^{(i)}}$, $\boldsymbol{\sigma}^{(i)} = \boldsymbol{K}^{(i)}\boldsymbol{\sigma}_{n+1}^{(i)}$ for $i = 1, 2$. Following the same routine, $\mathbb{E}\{\boldsymbol{v}_n^{(1)\top}\boldsymbol{\Sigma}_{\mathrm{xc},n+1}\boldsymbol{v}_n^{(2)}\}$ can be evaluated as follows:

$$\begin{aligned}\xi_{\mathrm{x},n+1} &= \xi_{\mathrm{x},n} + \big(\mathrm{vec}\{\boldsymbol{G}_{\mathrm{x}}^{\top}\}\big)^{\top}(\boldsymbol{K}_{\mathrm{x}})^n\boldsymbol{\sigma}_{\mathrm{x}} + (\boldsymbol{\Pi}_n^{(1)} + \boldsymbol{\Pi}_n^{(2)})\boldsymbol{\sigma}_{\mathrm{x}} \\ &\quad - \big[(\overline{\boldsymbol{B}}^{(2)}\mathbb{E}\{\boldsymbol{v}_n^{(2)}\})^{\top} \otimes \overline{\boldsymbol{r}}^{(1)\top} + \overline{\boldsymbol{r}}^{(2)\top} \otimes (\overline{\boldsymbol{B}}^{(1)}\mathbb{E}\{\boldsymbol{v}_n^{(1)}\})^{\top}\big]\boldsymbol{\sigma}_{\mathrm{x}} \\ &\quad - \boldsymbol{v}_0^{(1)\top}\mathrm{vec}^{-1}\{(\boldsymbol{I}-\boldsymbol{K}_{\mathrm{x}})\big((\boldsymbol{K}_{\mathrm{x}})^n\boldsymbol{\sigma}_{\mathrm{x}}\big)\}\boldsymbol{v}_0^{(2)} \\ &\quad + \overline{\boldsymbol{r}}^{(1)\top}\mathrm{vec}^{-1}\{(\boldsymbol{K}_{\mathrm{x}})^n\boldsymbol{\sigma}_{\mathrm{x}}\}\overline{\boldsymbol{r}}^{(2)},\end{aligned} \tag{104}$$

and

$$\boldsymbol{\Pi}_{n+1}^{(1)} = \boldsymbol{\Pi}_n^{(1)}\boldsymbol{K}_{\mathrm{x}} + \big[\overline{\boldsymbol{r}}^{(2)\top} \otimes (\overline{\boldsymbol{B}}^{(1)}\mathbb{E}\{\boldsymbol{v}_n^{(1)}\})^{\top}\big]\big(\boldsymbol{I}-\boldsymbol{K}_{\mathrm{x}}\big), \tag{105}$$

$$\boldsymbol{\Pi}_{n+1}^{(2)} = \boldsymbol{\Pi}_n^{(2)}\boldsymbol{K}_{\mathrm{x}} + \big[(\overline{\boldsymbol{B}}^{(2)}\mathbb{E}\{\boldsymbol{v}_n^{(2)}\})^{\top} \otimes \overline{\boldsymbol{r}}^{(1)\top}\big]\big(\boldsymbol{I}-\boldsymbol{K}_{\mathrm{x}}\big) \tag{106}$$

with $\boldsymbol{\Pi}_0^{(i)} = \boldsymbol{0}_{1\times(NL)^2}$, $\boldsymbol{\sigma}_{\mathrm{x}} = \boldsymbol{K}_{\mathrm{x}}\boldsymbol{\sigma}_{\mathrm{x},n+1}$, $\xi_{\mathrm{x},n+1} = \mathbb{E}\{\boldsymbol{v}_{n+1}^{(1)\top}\boldsymbol{\Sigma}_{\mathrm{xc},n+1}\boldsymbol{v}_{n+1}^{(2)}\}$, $\xi_{\mathrm{x},0} = \mathbb{E}\{\boldsymbol{v}_0^{(1)\top}\boldsymbol{\Sigma}_{\mathrm{xc},n+1}\boldsymbol{v}_0^{(2)}\}$. By substituting (102) and (104) into (87), we evaluate the transient MSD of the convex power-normalized scheme.

Appendix E
Steady-state MSD of diffusion network with power-normalized scheme

$$\begin{aligned}\mathrm{MSD}^{\mathrm{steady}} &= \big[\mathrm{vec}\{\boldsymbol{G}^{(1)\top}\}\big]^{\top}\boldsymbol{\sigma}_{\infty}^{(1)} + \boldsymbol{f}\big(\overline{\boldsymbol{r}}^{(1)}, \boldsymbol{\Sigma}_{\infty}^{(1)}, \mathbb{E}\{\boldsymbol{v}_{\infty}^{(1)}\}\big) + \\ &\quad \big[\mathrm{vec}\{\boldsymbol{G}^{(2)\top}\}\big]^{\top}\boldsymbol{\sigma}_{\infty}^{(2)} + \boldsymbol{f}\big(\overline{\boldsymbol{r}}^{(2)}, \boldsymbol{\Sigma}_{\infty}^{(2)}, \mathbb{E}\{\boldsymbol{v}_{\infty}^{(2)}\}\big) + \big[\mathrm{vec}\{\boldsymbol{G}_{\mathrm{x}}^{\top}\}\big]^{\top}\boldsymbol{\sigma}_{\mathrm{x},\infty} \\ &\quad + \boldsymbol{f}_{\mathrm{x}}\big(\overline{\boldsymbol{r}}^{(1)}, \overline{\boldsymbol{r}}^{(2)}, \boldsymbol{\Sigma}_{\mathrm{x},\infty}, \mathbb{E}\{\boldsymbol{v}_{\infty}^{(1)}\}, \mathbb{E}\{\boldsymbol{v}_{\infty}^{(2)}\}, \overline{\boldsymbol{B}}^{(1)}, \overline{\boldsymbol{B}}^{(2)}\big)\end{aligned} \tag{107}$$

with

$$\boldsymbol{\Sigma}_{\infty}^{(i)} = \mathrm{vec}^{-1}\{\boldsymbol{\sigma}_{\infty}^{(i)}\},\ \forall\ i = 1, 2 \tag{108}$$

$$\boldsymbol{\Sigma}_{\mathrm{x},\infty} = \mathrm{vec}^{-1}\{\boldsymbol{\sigma}_{\mathrm{x},\infty}\}. \tag{109}$$

Expression (107) characterizes the steady-state MSD of the convex power-normalized scheme.

Appendix F
Proof of Theorem 4

Since $\gamma_{k,n}$ is related to $\alpha_{k,n}$ via the mapping (14), we evaluate the behaviors of $\alpha_{k,n}$ first, then $\gamma_{k,n}$.

Substituting (59) and (60) into (17), we have:

$$\begin{aligned}\alpha_{k,n+1} \approx \big[\alpha_{k,n} + v'_{\alpha_k}\gamma_{k,n}(1-\gamma_{k,n})\big\{&(2\gamma_{k,n}-1)\tilde{e}_{k,n}^{(1)}\tilde{e}_{k,n}^{(2)} - \\ &\gamma_{k,n}(\tilde{e}_{k,n}^{(1)})^2 + (1-\gamma_{k,n})(\tilde{e}_{k,n}^{(2)})^2 + (\tilde{e}_{k,n}^{(2)} - \tilde{e}_{k,n}^{(1)})z_{k,n}\big\}\big]_{-\alpha^+}^{\alpha^+}.\end{aligned} \tag{110}$$

Define:

$$\zeta_1 \triangleq (\tilde{e}_{k,n}^{(1)})^2 \tag{111}$$

$$\zeta_2 \triangleq (\tilde{e}_{k,n}^{(2)})^2 \tag{112}$$

$$\zeta_3 \triangleq \tilde{e}_{k,n}^{(1)}\tilde{e}_{k,n}^{(2)} \tag{113}$$

$$\zeta_4 \triangleq (\tilde{e}_{k,n}^{(2)} - \tilde{e}_{k,n}^{(1)})z_{k,n} \tag{114}$$

$$f_1(\alpha_{k,n}) \triangleq -\gamma_{k,n}^2(1-\gamma_{k,n}) \tag{115}$$

$$f_2(\alpha_{k,n}) \triangleq \gamma_{k,n}(1-\gamma_{k,n})^2 \tag{116}$$

$$f_3(\alpha_{k,n}) \triangleq \gamma_{k,n}(1-\gamma_{k,n})(2\gamma_{k,n}-1) \tag{117}$$

$$f_4(\alpha_{k,n}) \triangleq \gamma_{k,n}(1-\gamma_{k,n}) \tag{118}$$

Using definitions (111)–(118), expression (110) becomes:

$$\alpha_{k,n+1} \approx \Big[\alpha_{k,n} + v'_{\alpha_k}\sum_{\ell=1}^{4} f_\ell(\alpha_{k,n})\zeta_\ell\Big]_{-\alpha^+}^{\alpha^+}. \tag{119}$$

Observing (110) and (119), the iterations for $\alpha_{k,n}$ are coupled with $\gamma_{k,n}$. It is difficult, if not impossible, to evaluate the behavior of $\alpha_{k,n}$ directly since the explicit probability distribution of $\alpha_{k,n}$ is unknown. We alleviate this problem by using the first-order Taylor series expansion. Though not accurate, the first-order Taylor series expansion is widely adopted in the analysis of adaptive filters to simplify the derivations [42], [43]. Define $\bar{\alpha}_{k,n} \triangleq \mathbb{E}\{\alpha_{k,n}\}$. Expanding $f_\ell(\alpha_{k,n})$ around $\bar{\alpha}_{k,n}$ to its first-order, we have:

$$f_\ell(\alpha_{k,n}) \approx f_\ell(\bar{\alpha}_{k,n}) + f'_\ell(\bar{\alpha}_{k,n})(\alpha_{k,n} - \bar{\alpha}_{k,n}) \tag{120}$$

for $\ell = 1, \cdots, 4$, where $f'_\ell(\bar{\alpha}_{k,n}) = \frac{\mathrm{d}f_\ell(\bar{\alpha}_{k,n})}{\mathrm{d}\alpha_{k,n}}$ is the first-order derivative of $f_\ell(\alpha_{k,n})$ over $\alpha_{k,n}$ and evaluated at $\alpha_{k,n} = \bar{\alpha}_{k,n}$.

Under approximations **Ap1**, **Ap5**, substituting (120) into (119) and taking the expectation, we get the mean behavior of $\alpha_{k,n}$:

$$\begin{aligned}\mathbb{E}\{\alpha_{k,n+1}\} \approx \big[&\bar{\alpha}_{k,n} + \bar{v}_{\alpha_{k,n}}f_1(\bar{\alpha}_{k,n})J_{\mathrm{ex},k,n}^{(1)} + \\ &\bar{v}_{\alpha_{k,n}}f_2(\bar{\alpha}_{k,n})J_{\mathrm{ex},k,n}^{(2)} + \bar{v}_{\alpha_{k,n}}f_3(\bar{\alpha}_{k,n})J_{\mathrm{ex},k,n}^{(1,2)}\big]_{-\alpha^+}^{\alpha^+},\end{aligned} \tag{121}$$

where the order of the expectation and truncation operations are changed to facilitate the analysis, and $J_{\mathrm{ex},k,n}^{(1,2)} \triangleq$

$\mathbb{E}\{e^{(1)}_{a,k,n}e^{(2)}_{a,k,n}\}$ is the cross-EMSE at node k and time instant n. Besides, the following approximation in (121) is adopted to simplify the derivation:

$$\bar{v}_{\alpha_{k,n}} = \mathbb{E}\Big\{\frac{v_{\alpha_k}}{\varepsilon + p_{k,n}}\Big\} \approx \frac{v_{\alpha_k}}{\varepsilon + \bar{p}_{k,n}} \quad \text{with}$$
$$\bar{p}_{k,n} = \eta\bar{p}_{k,n-1} + (1-\eta)(J^{(1)}_{\text{ex},k,n} + J^{(2)}_{\text{ex},k,n} - 2J^{(1,2)}_{\text{ex},k,n}). \quad (122)$$

Then, for the mean behavior of $\gamma_{k,n}$, we evaluate it via the approximation:

$$\mathbb{E}\{\gamma_{k,n}\} \approx \frac{1}{1 + e^{-\mathbb{E}\{\alpha_{k,n}\}}}. \quad (123)$$

For the steady-state value $\bar{\gamma}_{k,\infty} \triangleq \mathbb{E}\{\gamma_{k,\infty}\}$, taking the limit of (121) with $n \to \infty$ and solving for $\bar{\gamma}_{k,\infty}$, we obtain the same steady-state value as (65).

Transient value $\mathbb{E}\{\gamma_{k,n}\}$ of (123) and steady-state value $\mathbb{E}\{\gamma_{k,\infty}\}$ of (65) are used in evaluating the transient and steady-state mean behavior of the convex power-normalized scheme, respectively.

Appendix G
Proof of **Theorem 5**

Similarly, we evaluate the mean-square behavior of $\gamma_{k,n}$ based on the first-order Taylor series expansion. Define $f_5(\alpha_{k,n}) \triangleq \frac{1}{1+e^{-\alpha_{k,n}}}$. Using the first-order Taylor series expansion, we have:

$$f_5(\alpha_{k,n}) \approx f_5(\bar{\alpha}_{k,n}) + f_5'(\bar{\alpha}_{k,n})(\alpha_{k,n} - \bar{\alpha}_{k,n}), \quad (124)$$

where $f_5'(\bar{\alpha}_{k,n}) \triangleq \frac{\mathrm{d}f_5(\bar{\alpha}_{k,n})}{\mathrm{d}\alpha_{k,n}}$ is the first-order derivative of $f_5(\alpha_{k,n})$ and evaluated at $\alpha_{k,n} = \bar{\alpha}_{k,n}$. Substituting the explicit expression of $f_5(\alpha_{k,n})$, we have that $f_5'(\bar{\alpha}_{k,n}) = f_5(\bar{\alpha}_{k,n})\big[1 - f_5(\bar{\alpha}_{k,n})\big]$. Since $\gamma_{k,n} = f_5(\alpha_{k,n})$ and using (124), we have:

$$\mathbb{E}\{\gamma^2_{k,n}\} \approx \big[f_5(\bar{\alpha}_{k,n})\big]^2 + \big[f_5'(\bar{\alpha}_{k,n})\big]^2\sigma^2_{\alpha_{k,n}}, \quad (125)$$

where $\sigma^2_{\alpha_{k,n}}$ is the variance of $\alpha_{k,n}$, and we have:

$$\sigma^2_{\alpha_{k,n}} = \mathbb{E}\{\alpha^2_{k,n}\} - \big[\mathbb{E}\{\alpha_{k,n}\}\big]^2. \quad (126)$$

In order to use (126) in calculating $\mathbb{E}\{\gamma^2_{k,n}\}$ of (125), we need to evaluate $\mathbb{E}\{\alpha^2_{k,n}\}$. By first discarding the truncation operation, squaring both sides of (119) and taking the expectation, we have:

$$\mathbb{E}\{\alpha^2_{k,n+1}\} \approx \mathbb{E}\{\alpha^2_{k,n}\} + 2\bar{v}_{\alpha_{k,n}}\sum_{\ell=1}^{4}\mathbb{E}\big\{\alpha_{k,n}\zeta_\ell f_\ell(\alpha_{k,n})\big\}$$
$$+ \bar{v}^2_{\alpha_{k,n}}\sum_{\ell=1}^{4}\sum_{m=1}^{4}\mathbb{E}\big\{f_\ell(\alpha_{k,n})f_m(\alpha_{k,n})\zeta_\ell\zeta_m\big\}, \quad (127)$$

where in the derivation of (127), we used **Ap1**, (122) and approximation $\mathbb{E}\big\{(\frac{v_{\gamma_k}}{\varepsilon+p_{k,n}})^2\big\} \approx \bar{v}^2_{\alpha_{k,n}}$. According to the first-order Taylor series expansion, we have:

$$\alpha_{k,n}f_\ell(\alpha_{k,n}) \approx \bar{\alpha}_{k,n}f_\ell(\bar{\alpha}_{k,n}) + \big[f_\ell(\bar{\alpha}_{k,n}) + \bar{\alpha}_{k,n}f_\ell'(\bar{\alpha}_{k,n})\big](\alpha_{k,n} - \bar{\alpha}_{k,n}) \quad (128)$$

and

$$f_\ell(\alpha_{k,n})f_m(\alpha_{k,n}) \approx f_\ell(\bar{\alpha}_{k,n})f_m(\bar{\alpha}_{k,n}) + \big[f_\ell'(\bar{\alpha}_{k,n})f_m(\bar{\alpha}_{k,n}) + f_\ell(\bar{\alpha}_{k,n})f_m'(\bar{\alpha}_{k,n})\big](\alpha_{k,n} - \bar{\alpha}_{k,n}). \quad (129)$$

Using (128), (129) and approximation **Ap5** yields:

$$\mathbb{E}\big\{\alpha_{k,n}\zeta_\ell f_\ell(\alpha_{k,n})\big\} \approx \bar{\alpha}_{k,n}f_\ell(\bar{\alpha}_{k,n})\mathbb{E}\big\{\zeta_\ell\big\}, \ \forall \ell = 1,2,3 \quad (130)$$
$$\mathbb{E}\big\{\alpha_{k,n}\zeta_4 f_4(\alpha_{k,n})\big\} = 0, \quad (131)$$
$$\mathbb{E}\big\{f_\ell(\alpha_{k,n})f_m(\alpha_{k,n})\zeta_\ell\zeta_m\big\} \approx f_\ell(\bar{\alpha}_{k,n})f_m(\bar{\alpha}_{k,n})\mathbb{E}\big\{\zeta_\ell\zeta_m\big\} \ \ \forall \ell, m = 1,2,3,4. \quad (132)$$

Substituting (130)–(132) into (127) leads to:

$$\mathbb{E}\{\alpha^2_{k,n+1}\} \approx \mathbb{E}\{\alpha^2_{k,n}\} + 2\bar{v}_{\alpha_{k,n}}\sum_{\ell=1}^{3}\bar{\alpha}_{k,n}f_\ell(\bar{\alpha}_{k,n})\mathbb{E}\{\zeta_\ell\} + \bar{v}^2_{\alpha_{k,n}}\Big[\sum_{\ell=1}^{3}\sum_{m=1}^{3}f_\ell(\bar{\alpha}_{k,n})f_m(\bar{\alpha}_{k,n})\mathbb{E}\{\zeta_\ell\zeta_m\} + f_4(\bar{\alpha}_{k,n})f_4(\bar{\alpha}_{k,n})\mathbb{E}\{\zeta_4^2\}\Big] \quad (133)$$

Under approximation **Ap6** and substituting (115)–(118) into (133), we obtain the explicit expression of $\mathbb{E}\{\alpha^2_{k,n+1}\}$. Now, by taking the truncation operation of $\alpha_{k,n}$ into consideration and using the non-negative property of $\sigma^2_{\alpha_{k,n}}$ in (126), we obtain that $\mathbb{E}\{\alpha^2_{k,n}\}$ is constrained to be in the interval $\big[(\mathbb{E}\{\alpha_{k,n}\})^2, (\alpha^+)^2\big]$. Then by using (125), (126) and (133) together, we evaluate the mean-square value $\mathbb{E}\{\gamma^2_{k,n}\}$. Besides, the steady-state value $\mathbb{E}\{\gamma^2_{k,\infty}\}$ is approximated by:

$$\mathbb{E}\{\gamma^2_{k,\infty}\} \approx \big[\mathbb{E}\{\gamma_{k,\infty}\}\big]^2. \quad (134)$$

Transient value $\mathbb{E}\{\gamma^2_{k,n}\}$ of (125) and steady-state value $\mathbb{E}\{\gamma_{k,\infty}\}^2$ of (134) are used in evaluating the transient and steady-state mean-square behavior of the convex power-normalized scheme, respectively.

References

[1] D. Jin, J. Chen, and J. Chen, "Convex combination of diffusion strategies over distributed networks," in *Proc. Asia-Pacific Signal Inf. Process. Association*, Hawaii, USA, Nov. 2018, pp. 224–228.

[2] J. Arenas-Garcia, A. R. Figueiras-Vidal, and A. H. Sayed, "Mean-square performance of a convex combination of two adaptive filters," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 1078–1090, Mar. 2006.

[3] J. Arenas-Garcia, M. Martinez-Ramon, A. Navia-Vazquez, and A. R. Figueiras-Vidal, "Plant identification via adaptive combination of transversal filters," *Signal Process.*, vol. 86, no. 9, pp. 2430 – 2438, 2006.

[4] R. Alain, B. Francis, C. Stephane, and G. Yves, "Simple MKL," *Journal of Mach. Learn. Research*, vol. 9, no. 3, pp. 2491–2521, 2008.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. of CVPR*, Boston, MA, USA, June 2015, pp. 1–9.

[6] 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.

[7] 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.

[8] L. Li and J. Chambers, "Distributed adaptive estimation based on the APA algorithm over diffusion networks with changing topology," in *Proc. IEEE SSP*, 2009, pp. 757–760.

[9] F. Cattivelli and A. H. Sayed, "Diffusion strategies for distributed Kalman filtering and smoothing," *IEEE Trans. Autom. Control*, vol. 55, no. 9, pp. 2069–2084, 2010.

[10] F. Cattivelli, C. G. Lopes, and A. H. Sayed, "Diffusion recursive least-squares for distributed estimation over adaptive networks," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1865–1877, May 2008.

[11] Y. Liu, C. Li, and Z. Zhang, "Diffusion sparse least-mean squares over networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4480–4485, Aug. 2012.

[12] S. Vlaski, L. Vandenberghe, and A. H. Sayed, "Diffusion stochastic optimization with non-smooth regularizers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 4149–4153.

[13] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Proximal multitask learning over networks with sparsity-inducing coregularization," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6329–6344, Dec. 2016.

[14] J. Chen, C. Richard, and A. H. Sayed, "Multitask diffusion adaptation over networks," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4129–4144, Aug. 2014.

[15] J. Chen, C. Richard, and A. H. Sayed, "Diffusion LMS over multitask networks," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2733–2748, Jun. 2015.

[16] J. Chen, C. Richard, and A. H. Sayed, "Multitask diffusion adaptation over networks with common latent representations," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 3, pp. 563–579, 2017.

[17] X. Zhao and A. H. Sayed, "Distributed clustering and learning over networks," *IEEE Trans. Signal Process.*, vol. 63, no. 13, pp. 3285–3300, Jul. 2015.

[18] V. C. Gogineni and M. Chakraborty, "Diffusion affine projection algorithm for multitask networks," in *Proc. Asia-Pacific Signal Inf. Process. Assoc.*, 2018, pp. 201–206.

[19] V. C. Gogineni and M. Chakraborty, "Improving the performance of multitask diffusion APA via controlled inter-cluster cooperation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 3, pp. 903–912, 2020.

[20] V. C. Gogineni and M. Chakraborty, "Partial diffusion affine projection algorithm over clustered multitask networks," in *Proc. IEEE ISCAS*, 2019, pp. 1–5.

[21] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Multitask diffusion LMS with sparsity-based regularization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2015, pp. 3516–3520.

[22] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Multitask diffusion adaptation over asynchronous networks," *IEEE Trans. Signal Process.*, vol. 64, no. 11, pp. 2835–2850, 2016.

[23] D. Jin, J. Chen, C. Richard, and J. Chen, "Online proximal learning over jointly sparse multitask networks with $\ell_{\infty,1}$ regularization," *IEEE Trans. Signal Process., to appear*.

[24] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Diffusion LMS for multitask problems with local linear equality constraints," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 4979–4993, Oct. 2017.

[25] J. Chen, C. Richard, S. K. Ting, and A. H. Sayed, "Chapter # - Multitask learning over adaptive networks with grouping strategies," in *Cooperative and Graph Signal Processing*, pp. 107 – 129. Academic Press, 2018.

[26] M. Martinez-Ramon, J. Arenas-Garcia, A. Navia-Vazquez, and A. R. Figueiras-Vidal, "An adaptive combination of adaptive filters for plant identification," in *Proc. Int. Conf. on Digital Signal Process.*, 2002, vol. 2, pp. 1195–1198.

[27] N. J. Bershad, J. C. M. Bermudez, and J. Y. Tourneret, "An affine combination of two LMS adaptive filters—transient mean-square analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1853–1864, May 2008.

[28] L. A. Azpicueta-Ruiz, A. R. Figueiras-Vidal, and J. Arenas-Garcia, "A normalized adaptation scheme for the convex combination of two adaptive filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2008, pp. 3301–3304.

[29] R. Candido, M. T. M. Silva, and V. H. Nascimento, "Transient and steady-state analysis of the affine combination of two adaptive filters," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4064–4078, Aug. 2010.

[30] S. S. Kozat, A. T. Erdogan, A. C. Singer, and A. H. Sayed, "Steady-state MSE performance analysis of mixture approaches to adaptive filtering," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4050–4063, Aug. 2010.

[31] J. Arenas-Garcia, L. A. Azpicueta-Ruiz, M. T. M. Silva, V. H. Nascimento, and A. H. Sayed, "Combinations of adaptive filters: Performance and convergence properties," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 120–140, Jan. 2016.

[32] B. K. Das and M. Chakraborty, "Sparse adaptive filtering by an adaptive convex combination of the LMS and the ZA-LMS algorithms," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 61, no. 5, pp. 1499–1507, 2014.

[33] B. K. Das, G. V. Chakravarthi, and M. Chakraborty, "A convex combination of NLMS and ZA-NLMS for identifying systems with variable sparsity," *IEEE Trans. Circuits Syst. II, Express Briefs*, vol. 64, no. 9, pp. 1112–1116, 2017.

[34] V. C. Gogineni, B. K. Das, and M. Chakraborty, "An adaptive convex combination of APA and ZA-APA for identifying systems having variable sparsity and correlated input," *Digital Signal Process.*, vol. 82, pp. 118 – 132, 2018.

[35] B. K. Das and M. Chakraborty, "A block-based convex combination of NLMS and ZA-NLMS for identifying sparse systems with variable sparsity," in *Proc. IEEE ISCAS*, 2017, pp. 1–4.

[36] D. Jin, J. Chen, C. Richard, J. Chen, and A. H. Sayed, "Affine combination of diffusion strategies over networks," *IEEE Trans. Signal Process.*, vol. 68, no. 1, pp. 2087–2104, Dec. 2020.

[37] A. H. Sayed, "Diffusion adaptation over networks," in *Academic Press Libraray in Signal Processing*, R. Chellapa and S. Theodoridis, Eds., vol. 3, pp. 322–454. Elsevier, 2014.

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

[39] A. H. Sayed, *Adaptation, Learning, and Optimization over Networks*, vol. 7, Now Publishers Inc., Hanover, MA, USA, Jul. 2014.

[40] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill Higher Education, 4 edition, 2002.

[41] A. H. Sayed, *Adaptive Filters*, John Wiley & Sons, Inc., 2008.

[42] V. H. Nascimento, M. T. M. Silva, R. Candido, and J. Arenas-Garcia, "A transient analysis for the convex combination of adaptive filters," in *Proc. IEEE SSP*, Aug. 2009, pp. 53–56.

[43] M. T. M. Silva, V. H. Nascimento, and J. Arenas-Garcia, "A transient analysis for the convex combination of two adaptive filters with transfer of coefficients," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 3842–3845.

[44] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1985.

[45] S. Haykin, *Adaptive Filter Theory*, Pearson Education India, 4th edition, 2005.

[46] V. H. Nascimento and R. C. de Lamare, "A low-complexity strategy for speeding up the convergence of convex combinations of adaptive filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 3553–3556.

[47] L. F. O. Chamon, W. B. Lopes, and C. G. Lopes, "Combination of adaptive filters with coefficients feedback," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 3785–3788.

[48] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "Federated learning with quantization constraints," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 8851–8855.

[49] I. E. K. Harrane, R. Flamary, and C. Richard, "On reducing the communication cost of the diffusion LMS algorithm," *IEEE Trans. Signal Inf. Process. Networ.*, vol. 5, no. 1, pp. 100–112, 2019.

[50] X. Zhao and A. H. Sayed, "Clustering via diffusion adaptation over networks," in *Proc. Int. Workshop Cognitive Inf. Process.*, May. 2012, pp. 1–6.

[51] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Mathematical Journal*, vol. 23, no. 2, pp. 298–305, 1973.

[52] A. Simões and J. Xavier, "FADE: Fast and asymptotically efficient distributed estimator for dynamic networks," *IEEE Trans. Signal Process.*, vol. 67, no. 8, pp. 2080–2092, Apr. 2019.

[53] S. Das and J. M. F. Moura, "Distributed state estimation in multi-agent networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 4246–4250.