# DISTRIBUTED PARETO-OPTIMAL SOLUTIONS VIA DIFFUSION ADAPTATION

*Jianshu Chen* and *Ali H. Sayed*

Department of Electrical Engineering
University of California, Los Angeles

## ABSTRACT

We consider solving multi-objective optimization problems in a distributed manner over a network of nodes. The problem is equivalent to optimizing a global cost that is the sum of individual components. Diffusion adaptation enables the nodes to cooperate locally through in-network processing in order to approach Pareto-optimality. We analyze the mean-square-error performance of the diffusion strategy and show that, at steady-state, all nodes can be made to approach a Pareto-optimal solution.

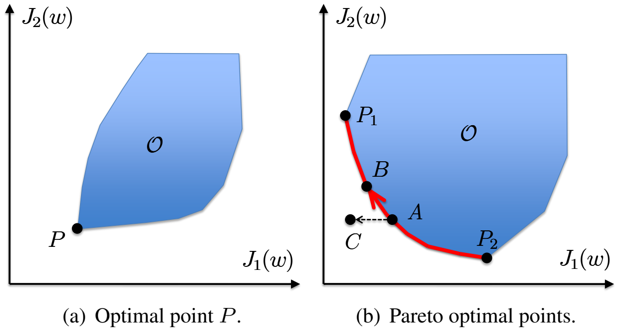
***Index Terms—*** Distributed optimization, diffusion adaptation, Pareto optimality, mean-square performance, convergence, stability.

## 1. INTRODUCTION

We consider solving a multi-objective optimization problem in a distributed manner over a network of $N$ nodes. An individual cost $J_k(w)$ is associated with each node $k = 1, 2, \dots, N$, where $w$ is $M \times 1$. These individual costs may not be minimized at the same location $w^o$. As such, we seek a solution $w^o$ that is "optimal" in some sense for all nodes. In these cases, a more general concept of optimality—*Pareto optimality*—is useful to characterize how good a solution is. A point $w^o$ is said to be Pareto optimal if there does not exist another point $w$ that is able to improve (reduce) any particular cost, say, $J_k(w)$, without hurting (increasing) all other costs $\{J_l(w)\}_{l \neq k}$. To illustrate the idea of Pareto optimality, let

$$\mathcal{O} \triangleq \{(J_1(w), \dots, J_N(w)) : \ w \in \mathbb{W}\} \subseteq \mathbb{R}^N \tag{1}$$

denote the set of achievable cost values, where $\mathbb{W}$ denotes the feasible set. Each point $P \in \mathcal{O}$ represents attained values of the cost functions $\{J_l(w)\}$ at a certain $w \in \mathbb{W}$. Let us consider the two-node case ($N = 2$) in Fig. 1, where the shaded areas are the sets of achievable cost values $\mathcal{O}$. In Fig. 1(a), both $J_1(w)$ and $J_2(w)$ achieve their minima at the same point $P = (J_1(w^o), J_2(w^o))$, where $w^o$ is the corresponding minimizer. However, in Fig. 1(b), $J_1(w)$ attains its minimum at point $P_1$, while $J_2(w)$ attains its minimum at point $P_2$, so that they do not have a common minimizer. Instead, all the points on the heavy red curve between points $P_1$ and $P_2$ are Pareto optimal points. For example, if we want to further reduce the value of $J_1(w)$ from point $A$ without increasing the value of $J_2(w)$, then we will need to move out of the achievable set $\mathcal{O}$ towards point $C$. The alternative choice that would keep us on the curve is to move to another Pareto optimal point $B$, which would increase the value of $J_2(w)$. In other words, we need to trade the value of $J_2(w)$ for $J_1(w)$. For this reason, the curve from $P_1$ to $P_2$ is called the optimal tradeoff curve (or optimal tradeoff surface if $N > 2$) [1, p.183].



(a) Optimal point $P$. (b) Pareto optimal points.

**Fig. 1**. Optimal and Pareto optimal points for $N = 2$.

To solve for the Pareto optimal points, a scalarization technique is usually used to form a cost function that is the weighted sum of the component costs:

$$J^{\text{glob}}(w) = \sum_{l=1}^{N} \pi_l J_l(w) \tag{2}$$

where $\pi_l$ is the positive weight attached to the $l$-th cost. It was shown in [1, pp.178–180] that the optimal solution to (2) is Pareto optimal for the original multi-objective optimization problem for any positive scalars $\{\pi_l\}$. By varying the values of $\{\pi_l\}$, we are able to get different Pareto optimal points on the tradeoff curve. Observing that we can always redefine $J_l(w) \leftarrow \pi_l J_l(w)$ to absorb the weight $\pi_l$, we only need to consider global costs of the form:

$$\boxed{J^{\text{glob}}(w) = \sum_{l=1}^{N} J_l(w)} \tag{3}$$

Furthermore, we can always use a barrier function to convert a constrained optimization problem to an unconstrained problem [1, 2]. For example, suppose there is a constraint $p_k^T w < b_k$ at node $k$, where $p_k$ is $M \times 1$ and $b_k$ is a scalar. Then, we can modify the cost $J_k(w)$ to be $J_k(w) \leftarrow J_k(w) + \phi(p_k^T w - b_k)$, where $\phi(x)$ is a barrier function that penalizes the values of $w$ that violate the constraint. Therefore, without loss of generality, we assume $\mathbb{W} = \mathbb{R}^M$ and only consider unconstrained optimizations. Moreover, we assume the $\{J_l(w)\}$ are differentiable and strictly convex so that $J^{\text{glob}}(w)$ in (3) is also strictly convex and the minimizer $w^o$ is unique [3].

There are already a few useful techniques for the solution of such optimization problems [4–9]. One notable technique is the incremental approach [4, 5]. In this approach, a cyclic path is defined over the nodes and data are processed in a cyclic manner through the network until optimization is achieved. However, determining a cyclic path that covers all nodes is an NP-hard problem [10] and, in

Email: {jshchen, sayed}@ee.ucla.edu. This work was supported in part by NSF grants CCF-1011918 and CCF-0942936.

addition, cyclic trajectories are vulnerable to link and node failures. Another useful distributed optimization approach relies on the use of consensus strategies [6–9]. In this approach, vanishing step-sizes are used to ensure that agents reach consensus and converge to the optimizer in steady-state. However, in time-varying environments, diminishing step-sizes prevent the network from continuous learning; when the step-sizes die out, the network stops learning. In [11], we generalized our earlier work on diffusion adaptation and learning over networks [12, 13] to optimize the same form of global cost functions in a decentralized manner. In the diffusion approach, information is processed locally at the nodes and then diffused through a real-time sharing mechanism. In the work [14], it is shown that diffusion networks outperform consensus networks in terms of mean-square-error stability and performance. In this paper, we show that the diffusion approach of [11] can also be used to solve such multi-objective optimization problems, which are common in the context of multi-agent decision making (see, for example, [2, 15]). With local cooperation and under constant step-sizes, each agent can approach global Pareto-optimal decisions within a small mean-square-error (MSE) bound.

## 2. DIFFUSION ADAPTATION STRATEGIES

In our previous work [11], we derived diffusion adaptation strategies to minimize (3) when each component function, $J_l(w)$, has a minimizer at the same $w^o$. Here, we consider the more general case when the individual minimizers of $\{J_l(w)\}$ may be different. These two scenarios arise in different contexts. The scenario of [11] occurs when the data processed at the nodes are generated by the same underlying model. On the other hand, the situation discussed here appears in problems where multiple agents have "conflicts of interest" but wish to coordinate with each other to reach "socially-optimal" (Pareto-optimal) solutions. This optimization problem can be solved in a distributed manner using the same class of adaptive diffusion algorithms derived in [11]. Specifically, for each node $k$, we implement:

$$\left\{\begin{aligned}
\phi_{k,i-1} &= \sum_{l=1}^{N} a_{1,lk} w_{l,i-1} && \text{(4a)}\\
\psi_{k,i} &= \phi_{k,i-1} - \mu_k \sum_{l=1}^{N} c_{lk} \nabla_w J_l(\phi_{k,i-1}) && \text{(4b)}\\
w_{k,i} &= \sum_{l=1}^{N} a_{2,lk} \psi_{l,i} && \text{(4c)}
\end{aligned}\right.$$

where $w_{k,i}$ is the local estimate for $w^o$ at node $k$ and time $i$, $\mu_k$ is the positive step-size parameter used by node $k$, and $\nabla_w J_l(w)$ is the (column) gradient vector of $J_l(\cdot)$ relative to $w$. Moreover, the nonnegative coefficients $\{a_{1,lk}\}$, $\{c_{lk}\}$, and $\{a_{2,lk}\}$ are the $(l,k)$-th entries of matrices $A_1$, $C$, and $A_2$, respectively, and are required to satisfy:

$$\boxed{\begin{array}{l} A_1^T \mathbb{1} = \mathbb{1},\ A_2^T \mathbb{1} = \mathbb{1},\ C\mathbb{1} = \mathbb{1} \\ a_{1,lk} = 0,\ a_{2,lk} = 0\ , c_{lk} = 0 \text{ if } l \notin \mathcal{N}_k \end{array}} \tag{5}$$

where $\mathbb{1}$ denotes a vector with all entries equal to one, and $\mathcal{N}_k$ denotes the neighborhood of node $k$. According to (4a)–(4c), each node $k$ in the network aggregates intermediate estimates from its neighbors via steps (4a) and (4c). Each node $k$ also adapts its intermediate estimate via (4b) by incorporating gradient information from its neighbors. Algorithm (4a)–(4c) is general and different choices for $\{A_1, A_2, C\}$ lead to different cooperation strategies [11, 13]. For example, choosing $A_1 = I$, $A_2 = A$, and $C = I$, we obtain the following adapt-then-combine (ATC) special case of (4a)–(4c):

$$\psi_{k,i} = w_{k,i-1} - \mu_k \nabla_w J_k(w_{k,i-1}) \tag{6}$$

$$w_{k,i} = \sum_{l\in\mathcal{N}_k} a_{lk} \psi_{l,i} \tag{7}$$

In this implementation, there is only one aggregation step and it follows the adaptation step. Moreover, the adaptation step relies only on the local gradient information at node $k$. For generality, we continue with form (4a)–(4c).

## 3. PERFORMANCE ANALYSIS

In many situations in practice, the true gradient vectors needed in (4b) may not be available. Instead, perturbed versions are available, which we model as

$$\widehat{\nabla}_w J_l(\boldsymbol{w}) = \nabla_w J_l(\boldsymbol{w}) + \boldsymbol{v}_{l,i}(\boldsymbol{w}) \tag{8}$$

where the noise term, $\boldsymbol{v}_{l,i}(\boldsymbol{w})$, may depend on $\boldsymbol{w}$ (and also on time $i$); it will be required to satisfy certain conditions given by (19)–(20). We refer to the perturbation in (8) as gradient noise; such noise is used to model the statistical fluctuations caused by using stochastic gradients or instantaneous approximations in the context of stochastic approximation and adaptive filter theory [3, 11, 16]. Note that we are denoting $\boldsymbol{w}$ in bold in Eq. (8) in order to highlight the fact that the estimates $\{\phi_{k,i-1}, \psi_{k,i}, w_{k,i}\}$ that are generated via (4a)–(4c) become random quantities $\{\boldsymbol{\phi}_{k,i-1}, \boldsymbol{\psi}_{k,i}, \boldsymbol{w}_{k,i}\}$ once gradient noise is present. We reserve the boldface notation for random variables. We now examine the effect of gradient noise on the mean-square performance of the diffusion strategy and establish convergence of all nodes towards a Pareto-optimal solution.

### 3.1. Error Recursions

Introduce the error vectors:

$$\tilde{\boldsymbol{\phi}}_{k,i} \triangleq w^o - \boldsymbol{\phi}_{k,i},\ \tilde{\boldsymbol{\psi}}_{k,i} \triangleq w^o - \boldsymbol{\psi}_{k,i},\ \tilde{\boldsymbol{w}}_{k,i} \triangleq w^o - \boldsymbol{w}_{k,i} \tag{9}$$

Then, we can establish that [11]:

$$\boxed{\tilde{\boldsymbol{w}}_i = \mathcal{A}_2^T [I_{MN} - \mathcal{M}\boldsymbol{\mathcal{R}}_{i-1}]\mathcal{A}_1^T \tilde{\boldsymbol{w}}_{i-1} + \mathcal{A}_2^T \mathcal{M}\mathcal{C}^T g^o + \mathcal{A}_2^T \mathcal{M}\boldsymbol{g}_i} \tag{10}$$

where

$$\begin{aligned}
\tilde{\boldsymbol{w}}_i &\triangleq \text{col}\{\tilde{\boldsymbol{w}}_{1,i}, \dots, \tilde{\boldsymbol{w}}_{N,i}\} && (11)\\
\mathcal{A}_1 &\triangleq A_1 \otimes I_M,\ \mathcal{A}_2 \triangleq A_2 \otimes I_M,\ \mathcal{C} \triangleq C \otimes I_M && (12)\\
\mathcal{M} &\triangleq \text{diag}\{\mu_1, \dots, \mu_N\} \otimes I_M && (13)\\
\boldsymbol{H}_{lk,i-1} &\triangleq \int_0^1 \nabla_w^2 J_l\Big(w^o - t\sum_{l=1}^{N} a_{1,lk}\tilde{\boldsymbol{w}}_{l,i-1}\Big)dt && (14)\\
\boldsymbol{\mathcal{R}}_{i-1} &\triangleq \sum_{l=1}^{N} \text{diag}\big\{c_{l1}\boldsymbol{H}_{l1,i-1}, \cdots, c_{lN}\boldsymbol{H}_{lN,i-1}\big\} && (15)\\
\boldsymbol{g}_i &\triangleq \sum_{l=1}^{N} \text{col}\big\{c_{l1}\boldsymbol{v}_{l,i}(\boldsymbol{\phi}_{1,i-1}), \cdots, c_{lN}\boldsymbol{v}_{l,i}(\boldsymbol{\phi}_{N,i-1})\big\} && (16)\\
g^o &\triangleq \text{col}\big\{\nabla_w J_1(w^o),\ \cdots,\ \nabla_w J_N(w^o)\big\} && (17)
\end{aligned}$$

and the symbol $\otimes$ denotes Kronecker products [17]. Compared to [11], there is an extra term (the second term) on the right-hand side of (10), which arises when $g^o \neq 0$; this happens because the $\{J_l(w)\}$ do not necessarily have their minimizers at the same $w^o$ — $\nabla_w J_l(w^o) \neq 0$ for some $l = 1, \ldots, N$. This term biases the solution and its effect needs to be examined closely. To proceed with the analysis, we introduce similar assumptions on the cost functions and gradient noise as in [11].

**Assumption 1** (Bounded Hessian)**.** *There exist nonnegative real numbers $\lambda_{l,\min}$ and $\lambda_{l,\max}$ such that, for each $l = 1, \ldots, N$:*

$$\lambda_{l,\min} I_M \leq \nabla_w^2 J_l(w) \leq \lambda_{l,\max} I_M, \quad \sum_{l=1}^{N} c_{lk} \lambda_{l,\min} > 0 \quad (18)$$

**Assumption 2** (Gradient noise)**.** *There exist $\alpha \geq 0$ and $\sigma_v^2 \geq 0$ such that, for all $\boldsymbol{w} \in \mathcal{F}_{i-1}$ and for all $i, l$:*

$$\mathbb{E}\left\{\boldsymbol{v}_{l,i}(\boldsymbol{w}) \mid \mathcal{F}_{i-1}\right\} = 0 \quad (19)$$

$$\mathbb{E}\left\{\|\boldsymbol{v}_{l,i}(\boldsymbol{w})\|^2\right\} \leq \alpha \mathbb{E}\|\nabla_w J_l(\boldsymbol{w})\|^2 + \sigma_v^2 \quad (20)$$

*where $\mathcal{F}_{i-1}$ denotes the past history ($\sigma-$field) of estimators $\{\boldsymbol{w}_{k,j}\}$ for $j \leq i-1$ and all $k$.*

### 3.2. Transient and Bias Analysis

Our strategy to analyze the mean-square performance of the diffusion strategy (4a)–(4c) is to show that, in the presence of gradient noise, the recursion will converge to the unique fixed point $w_\infty$ of the recursion (4a)–(4c) within a small MSE bound. Afterwards, we show that the fixed point $w_\infty$ can be made arbitrarily close to the Pareto-optimal solution $w^o$ for sufficiently small step-sizes. To begin with, we show the existence and uniqueness of the fixed point, and examine how close the estimators $\{\boldsymbol{w}_{k,i}\}$ get to $w_\infty$.

**Lemma 1** (Existence and Uniqueness of Fixed Point)**.** *Suppose the step-size parameters $\{\mu_k\}$ satisfy the following conditions*

$$0 < \mu_k < \frac{2}{\sigma_{k,\max}}, \qquad k = 1, \ldots, N \quad (21)$$

*where $\sigma_{k,\max} \triangleq \sum_{l=1}^{N} c_{lk} \lambda_{l,\max}$ Then, there exists a unique fixed point $w_\infty \triangleq \mathrm{col}\{w_{1,\infty}, \ldots, w_{N,\infty}\}$ for the deterministic iterations* (4a)–(4c) *(with true gradients), i.e.,*

$$\begin{cases} \phi_{k,\infty} = \sum_{l=1}^{N} a_{1,lk} w_{l,\infty} & (22a) \\ \psi_{k,\infty} = \phi_{k,\infty} - \mu_k \sum_{l=1}^{N} c_{lk} \nabla_w J_l(\phi_{k,\infty}) & (22b) \\ w_{k,\infty} = \sum_{l=1}^{N} a_{2,lk} \psi_{l,\infty} & (22c) \end{cases}$$

*Proof.* The idea is to show that iterations (4a)–(4c) lead to a contraction mapping when the step-sizes satisfy (21). ■

**Theorem 1** (Mean-Square Stability and Bounds)**.** *Suppose the step-size parameters satisfy the following condition:*

$$0 < \mu_k < \min\left\{\frac{\sigma_{k,\max}}{\sigma_{k,\max}^2 + 4\alpha\lambda_{\max}^2 \|C\|_1^2}, \ \frac{\sigma_{k,\min}}{\sigma_{k,\min}^2 + 4\alpha\lambda_{\max}^2 \|C\|_1^2}\right\} \quad (23)$$

*for $k = 1, \ldots, N$, where $\sigma_{k,\max}$ was defined earlier in Lemma 1, $\sigma_{k,\min} \triangleq \sum_{l=1}^{N} c_{lk} \lambda_{l,\min}$, and $\lambda_{\max} \triangleq \max_l \lambda_{l,\max}$. Then,*

$$\limsup_{i \to \infty} \|\mathrm{MSP}_i\|_\infty \leq \frac{\|C\|_1^2 \cdot \|b_v\|_\infty \cdot \mu_{\max}}{2\beta\sigma_{\min} - \mu_{\max}(\sigma_{\max}^2 + 4\alpha\lambda_{\max}\|C\|_1^2)} \quad (24)$$

*where $\sigma_{\min}$ and $\sigma_{\max}$ are the minimum of $\sigma_{k,\min}$ and the maximum of $\sigma_{k,\max}$, respectively, and*

$$\begin{aligned} \mathrm{MSP}_i &\triangleq \mathrm{col}\left\{\mathbb{E}\|\boldsymbol{w}_{1,i} - w_{1,\infty}\|^2, \ldots, \mathbb{E}\|\boldsymbol{w}_{N,i} - w_{N,\infty}\|^2\right\} & (25) \\ \mu_{\max} &\triangleq \max_{1 \leq k \leq N} \mu_k, \ \ \mu_{\min} \triangleq \min_{1 \leq k \leq N} \mu_k, \ \ \beta \triangleq \frac{\mu_{\min}}{\mu_{\max}} & (26) \\ b_v &\triangleq 4\alpha\lambda_{\max}^2 A_1^T \mathrm{col}\left\{\|w_{1,\infty} - w^o\|^2, \ldots, \|w_{N,\infty} - w^o\|^2\right\} \\ &\quad + \max_{1 \leq k \leq N} \{2\alpha\|\nabla_w J_k(w^o)\|^2 + \sigma_v^2\} & (27) \end{aligned}$$

*Proof.* Omitted for brevity. ■

$\mathrm{MSP}_i$ represents the $N \times 1$ mean-square-perturbation (MSP) vector at time $i$. The $k$-th entry of $\mathrm{MSP}_i$ characterizes how far the estimate $\boldsymbol{w}_{k,i}$ at node $k$ and time $i$ is from $w_{k,\infty}$ in the mean-square sense. The right-hand side of (24) can be made arbitrarily small for sufficiently small $\mu_{\max}$. It follows that the steady-state MSP can be made arbitrarily small for small step-sizes, and the estimators $\boldsymbol{w}_i = \mathrm{col}\{\boldsymbol{w}_{1,i}, \ldots, \boldsymbol{w}_{N,i}\}$ will be close to the fixed point $w_\infty$ (in the mean-square sense) even under gradient perturbations. Next, we examine how close the fixed point $w_\infty$ is to $w^o$.

**Theorem 2** (Bias at Small Step-sizes)**.** *Suppose that $A_2^T A_1^T$ is a primitive right-stochastic matrix, so that its eigenvalue of largest magnitude is one with multiplicity one, and all other eigenvalues are strictly smaller than one. Let $\theta^T$ denote the left eigenvector of $A_2^T A_1^T$ of eigenvalue one. Furthermore, assume the following condition holds:*

$$\theta^T A_2^T \Omega C^T = c_0 \mathbb{1}^T \quad (28)$$

*where $\Omega \triangleq \mathrm{diag}\{\mu_1, \ldots, \mu_N\}$, and $c_0$ is some constant. Then,*

$$\|\mathbb{1}_N \otimes w^o - w_\infty\|^2 \sim O(\mu_{\max}^2) \quad (29)$$

*Proof.* Omitted for brevity. ■

Therefore, as long as the network is connected (not necessarily fully connected) and condition (28) holds, the bias would become arbitrarily small. For condition (28) to hold, one choice is to require the matrices $A_1^T$ and $A_2^T$ to be doubly stochastic, and all nodes to use the same step-size $\mu$, namely, $\Omega = \mu I_N$. In that case, the matrix $A_1^T A_2^T$ is doubly-stochastic so that the left eigenvector of eigenvalue one is $\theta^T = \mathbb{1}^T$ and (28) holds with $c_0 = \mu$. Combining Theorems 1 and 2, we conclude that the steady-state mean-square-error will be small when the step-size is sufficiently small because by (24) and (29):

$$\begin{aligned} \mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 &= \mathbb{E}\|\boldsymbol{w}_{k,i} - w_{k,\infty} + w_{k,\infty} - w^o\|^2 \\ &\leq 2\mathbb{E}\|\boldsymbol{w}_{k,i} - w_{k,\infty}\|^2 + 2\|w_{k,\infty} - w^o\|^2 \\ &\sim O(\mu_{\max}) + O(\mu_{\max}^2) & (30) \end{aligned}$$

as $i \to \infty$. Therefore, diffusive adaptation enables each node to approach the same global Pareto-optimal solution via local interactions.

### 3.3. Mean-Square Performance

Next, we determine an expression (rather than a bound) for the MSE. To do this, we first establish some useful approximations at small step-sizes. Expression (30) implies that, for small step-sizes and after long enough time, the random variable $\tilde{\boldsymbol{w}}_{k,i}$ is highly concentrated around $w^o$. Using this fact in (14)–(16), we can approximate $\boldsymbol{H}_{lk,i-1}$, $\boldsymbol{\mathcal{R}}_{i-1}$ and $\boldsymbol{g}_i$ in (14)–(16) by

$$\boldsymbol{H}_{lk,i-1} \approx \int_0^1 \nabla_w^2 J_l(w^o) dt = \nabla_w^2 J_l(w^o) \tag{31}$$

$$\boldsymbol{\mathcal{R}}_{i-1} \approx \sum_{l=1}^{N} \text{diag}\big\{c_{l1}\nabla_w^2 J_l(w^o), \dots, c_{lN}\nabla_w^2 J_l(w^o)\big\} \triangleq \mathcal{R}_\infty \tag{32}$$

$$\boldsymbol{g}_i \approx \sum_{l=1}^{N} \text{col}\big\{c_{l1}\boldsymbol{v}_{l,i}(w^o), \cdots, c_{lN}\boldsymbol{v}_{l,i}(w^o)\big\} \tag{33}$$

so that the matrices $\boldsymbol{H}_{lk,i-1}$ and $\boldsymbol{\mathcal{R}}_{i-1}$ become deterministic at small step-sizes, and we use regular font to represent them from now on. As a result, the error recursion (10) can be approximated by

$$\tilde{\boldsymbol{w}}_i = \mathcal{A}_2^T[I_{MN} - \mathcal{M}\mathcal{R}_\infty]\mathcal{A}_1^T\tilde{\boldsymbol{w}}_{i-1} + \mathcal{A}_2^T\mathcal{M}\mathcal{C}^T g^o + \mathcal{A}_2^T\mathcal{M}\boldsymbol{g}_i \tag{34}$$

Taking expectation of both sides of (34), we obtain

$$\boxed{\mathbb{E}\tilde{\boldsymbol{w}}_i = \mathcal{A}_2^T[I_{MN} - \mathcal{M}\mathcal{R}_\infty]\mathcal{A}_1^T\mathbb{E}\tilde{\boldsymbol{w}}_{i-1} + \mathcal{A}_2^T\mathcal{M}\mathcal{C}^T g^o} \tag{35}$$

This recursion converges when the matrix $\mathcal{A}_2^T[I_{MN} - \mathcal{M}\mathcal{R}_\infty]\mathcal{A}_1^T$ is stable, which is guaranteed by (23) (see Appendix C of [11]). Denote $\mathbb{E}\tilde{\boldsymbol{w}}_\infty \triangleq \lim_{i\to\infty} \mathbb{E}\tilde{\boldsymbol{w}}_i$ and let $i \to \infty$ on both sides of (35) so that

$$\boxed{\mathbb{E}\tilde{\boldsymbol{w}}_\infty = \big[I_{MN} - \mathcal{A}_2^T\,(I_{MN} - \mathcal{M}\mathcal{R}_\infty)\,\mathcal{A}_1^T\big]^{-1}\mathcal{A}_2^T\mathcal{M}\mathcal{C}^T g^o} \tag{36}$$

Let $\Sigma$ denote an arbitrary positive semi-definite matrix that we are free to choose. Let $\sigma = \text{vec}(\Sigma)$ denote the vectorization operation that stacks the columns of $\Sigma$ on top of each other. We shall use the notation $\|x\|_\sigma^2$ and $\|x\|_\Sigma^2$ interchangeably. Let also $\mathcal{R}_v$ denote the covariance matrix of $\boldsymbol{g}_i$ evaluated at $w^o$, $\mathcal{R}_v \triangleq \mathbb{E}\{\boldsymbol{g}_i\boldsymbol{g}_i^T\}|_{\{\boldsymbol{\phi}_{k,i-1}=w^o\}}$. Then, equating the squared *weighted* Euclidean norm of both sides of (34) and applying the expectation operator with assumption (19), we can establish the following variance relation:

$$\mathbb{E}\|\tilde{\boldsymbol{w}}_i\|_\sigma^2 = \mathbb{E}\|\tilde{\boldsymbol{w}}_{i-1}\|_{\mathcal{F}\sigma}^2 + r^T\sigma + \sigma^T\mathcal{Q}\,\mathbb{E}\tilde{\boldsymbol{w}}_{i-1} \tag{37}$$

where

$$\boxed{\begin{aligned} r &\triangleq \text{vec}\Big(\mathcal{A}_2^T\mathcal{M}\mathcal{R}_v\mathcal{M}\mathcal{A}_2\Big) + \mathcal{A}_2^T\mathcal{M}\mathcal{C}^T g^o \otimes \mathcal{A}_2^T\mathcal{M}\mathcal{C}^T g^o \\ \mathcal{Q} &\triangleq 2\Big(\mathcal{A}_2^T(I_{MN} - \mathcal{M}\mathcal{R}_\infty)\mathcal{A}_1^T\Big) \otimes \Big(\mathcal{A}_2^T\mathcal{M}\mathcal{C}^T g^o\Big) \\ \mathcal{F} &\triangleq (\mathcal{A}_1[I_{MN} - \mathcal{M}\mathcal{R}_\infty]\mathcal{A}_2) \otimes (\mathcal{A}_1[I_{MN} - \mathcal{M}\mathcal{R}_\infty]\mathcal{A}_2) \end{aligned}} \tag{38}$$

It was shown in [16, pp.344-346] that recursions such as (37) converge to a steady-state value if the matrix $\mathcal{F}$ is stable, i.e., $\rho(\mathcal{F}) < 1$. This condition is guaranteed when the step-sizes are sufficiently small (or chosen according to (37)) — see the proof in Appendix C of [11]. Letting $i \to \infty$ on both sides of expression (37), we obtain:

$$\boxed{\mathbb{E}\|\tilde{\boldsymbol{w}}_\infty\|_{(I-\mathcal{F})\sigma}^2 \approx \big(r + \mathcal{Q}\,\mathbb{E}\tilde{\boldsymbol{w}}_\infty\big)^T\sigma} \tag{39}$$

We can now resort to (39) and use it to evaluate various performance metrics by choosing proper weighting matrices $\Sigma$ (or $\sigma$), as it was done in [11]. For example, the MSE of any node $k$ can be obtained by computing $\mathbb{E}\|\tilde{\boldsymbol{w}}_\infty\|_T^2$ with a block weighting matrix $T$ that has an identity matrix at block $(k, k)$ and zeros elsewhere:

$$\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,\infty}\|^2 = \mathbb{E}\|\tilde{\boldsymbol{w}}_\infty\|_T^2 \tag{40}$$

Denote the vectorized version of this matrix by $t_k \triangleq \text{vec}(\text{diag}(e_k) \otimes I_M)$, where $e_k$ is a vector whose $k$th entry is one and zeros elsewhere. Then, if we select $\sigma$ in (39) as $\sigma = (I - \mathcal{F})^{-1}t_k$, the term on the left-hand side becomes the desired $\mathbb{E}\|\tilde{\boldsymbol{w}}_{k,\infty}\|^2$ and the MSE for node $k$ is therefore given by:

$$\boxed{\text{MSE}_k \triangleq \lim_{i\to\infty} \mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 \approx \big(r + \mathcal{Q}\,\mathbb{E}\tilde{\boldsymbol{w}}_\infty\big)^T(I - \mathcal{F})^{-1}t_k} \tag{41}$$

If we are interested in the average network MSE, then it is given by $\overline{\text{MSE}} \triangleq \frac{1}{N}\sum_{k=1}^{N}\text{MSE}_k$.

## 4. REFERENCES

[1] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[2] Z. J. Towfic, J. Chen, and A. H. Sayed, "Distributed throughput optimization over P2P mesh networks using diffusion adaptation," in *Proc. International Conf. Commun. (ICC)*, Ottawa, Canada, June 2012, pp. 1–5.

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

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

[5] C. G. Lopes and A. H. Sayed, "Incremental adaptive strategies over distributed networks," *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 4064–4077, Aug. 2007.

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

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

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

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

[10] R. M. Karp, "Reducibility among combinational problems," *Complexity of Computer Computations (R. E. Miller and J. W. Thatcher, eds.)*, pp. 85–104, 1972.

[11] J. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," to appear in *IEEE Trans. Signal Process.,* 2012 [also available as Arxiv preprint arXiv:1111.0034, Oct. 2011].

[12] 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, July 2008.

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

[14] S.-Y. Tu and A. H. Sayed, "Diffusion networks outperform consensus networks," in *Proc. IEEE Statistical Signal Processing Workshop (SSP)*, Ann Arbor, MI, Aug. 2012, pp. 1–4.

[15] Y. He, I. Lee, and L. Guan, "Distributed throughput maximization in P2P VoD applications," *IEEE Trans. Multimedia,*, vol. 11, no. 3, pp. 509–522, April 2009.

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

[17] A. J. Laub, *Matrix Analysis for Scientists and Engineers*, Society for Industrial and Applied Mathematics (SIAM), PA, 2005.