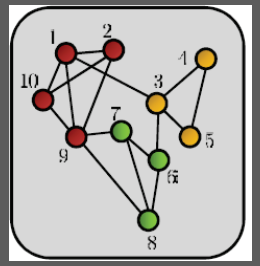


INFERENCE OVER NETWORKS

LECTURE #21: Performance of Multi-Agent Networks, Part II

Professor Ali H. Sayed
UCLA Electrical Engineering

Course EE210B
Spring Quarter 2015

Proc. IEEE, vol. 102, no. 4, pp. 460-497, April 2014.
Foundations and Trends in Machine Learning, vol. 7, no. 4-5, pp. 311-801, July 2014.

Reference

Chapter 11 (Performance of Multi-Agent Networks, pp. 574-621):

A. H. Sayed, ``Adaptation, learning, and optimization over networks," ***Foundations and Trends in Machine Learning***, vol. 7, issue 4-5, pp. 311-801, NOW Publishers, 2014.

Recall#1: Definition

Definition 11.1 (Hessian and moment matrices). We associate with each agent k a pair of matrices $\{H_k, G_k\}$, both of which are evaluated at the location of the limit point $w = w^\star$. The matrices are defined as follows:

$$H_k \triangleq \nabla_w^2 J_k(w^\star), \quad G_k \triangleq \begin{cases} R_{s,k} & \text{(real case)} \\ \begin{bmatrix} R_{s,k} & R_{q,k} \\ R_{q,k}^* & R_{s,k}^\mathsf{T} \end{bmatrix} & \text{(complex case)} \end{cases} \tag{11.12}$$

Both matrices are dependent on the data type (whether real or complex); in particular, each H_k is $2M \times 2M$ for complex data and $M \times M$ for real data. Note that $H_k \geq 0$ and $G_k \geq 0$.

Recall#2: MSE Performance

Theorem 11.2 (Network limiting performance). Consider a network of N interacting agents running the distributed strategy (8.46) with a primitive matrix $P = A_1 A_o A_2$. Assume the aggregate cost (9.10) and the individual costs, $J_k(w)$, satisfy the conditions in Assumptions 6.1 and 10.1. Assume further that the first and fourth-order moments of the gradient noise process satisfy the conditions of Assumption 8.1 with the second-order moment condition (8.115) replaced by the fourth-order moment condition (8.121). Assume also (11.10). Let

$$\gamma_m \triangleq \frac{1}{2}\min\{1, \gamma\} > 0 \tag{11.44}$$

Recall#2: MSE Performance

with $\gamma \in (0, 4]$ from (11.10). Then, it holds that

$$\limsup_{i\to\infty} \frac{1}{2}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i}^{e}\|^2 = \frac{1}{h}\mathrm{Tr}(\mathcal{J}_k\mathcal{X}) + O\left(\mu_{\max}^{1+\gamma_m}\right) \tag{11.45}$$

$$\limsup_{i\to\infty} \frac{1}{2N}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i}^{e}\|^2 = \frac{1}{hN}\mathrm{Tr}(\mathcal{X}) + O\left(\mu_{\max}^{1+\gamma_m}\right) \tag{11.46}$$

and, for large enough i, the convergence rate of the error variances, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i}\|^2$, towards the steady-state region (11.45) is given by

Recall#2: MSE Performance

$$\alpha \;=\; 1 - 2\lambda_{\min}\left(\sum_{k=1}^{N} q_k H_k\right) + O\left(\mu_{\max}^{(N+1)/N}\right) \tag{11.47}$$

where q_k is defined by (9.7) and $\alpha \in (0,1)$; the smaller the value of α is, the faster the convergence of $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i}\|^2$ towards (11.45). Moreover, the matrix $\mathcal{X}$ that appears in (11.45)–(11.46) is Hermitian non-negative definite and corresponds to the unique solution of the (discrete-time) Lypaunov equation:

$$\mathcal{X} - \mathcal{B}\mathcal{X}\mathcal{B}^* = \mathcal{Y} \tag{11.48}$$

where the quantities $\{\mathcal{Y}, \mathcal{B}, \mathcal{J}_k\}$ are defined by:

Recall#2: MSE Performance

$$\mathcal{A}_o = A_o \otimes I_{hM}, \quad \mathcal{A}_1 = A_1 \otimes I_{hM}, \quad \mathcal{A}_2 = A_2 \otimes I_{hM} \tag{11.49}$$

$$\mathcal{M} = \text{diag}\{\, \mu_1 I_{hM}, \mu_2 I_{hM}, \ldots, \mu_N I_{hM} \} \tag{11.50}$$

$$\mathcal{H} = \text{diag}\{\, H_1, H_2, \ldots, H_N \,\} \tag{11.51}$$

$$H_k = \nabla_w^2 J_k(w^\star) \tag{11.52}$$

$$\mathcal{S} = \text{diag}\{\, G_1,\ G_2, \ldots, G_N \,\} \tag{11.53}$$

$$\mathcal{Y} = \mathcal{A}_2^{\mathsf{T}} \mathcal{M} \mathcal{S} \mathcal{M} \mathcal{A}_2 \tag{11.54}$$

$$\mathcal{B} = \mathcal{A}_2^{\mathsf{T}} \left(\mathcal{A}_o^{\mathsf{T}} - \mathcal{M}\mathcal{H} \right) \mathcal{A}_1^{\mathsf{T}} \tag{11.55}$$

$$\mathcal{F} = \mathcal{B}^{\mathsf{T}} \otimes_b \mathcal{B}^* \tag{11.56}$$

$$\mathcal{J}_k = \text{diag}\{\, 0_{hM}, \ldots, 0_{hM}, I_{hM}, 0_{hM}, \ldots, 0_{hM} \,\} \tag{11.57}$$

Recall#2: MSE Performance

with $\mathcal{J}_k$ having an identity matrix at the $k-$th diagonal block, and $h = 1$ for real data and $h = 2$ for complex data. Furthermore, the following are equivalent characterizations for the matrix $\mathcal{X}$ or its trace:

$$\mathcal{X} = \sum_{n=0}^{\infty} \mathcal{B}^n \mathcal{Y} (\mathcal{B}^*)^n \tag{11.58}$$

$$\text{bvec}(\mathcal{X}) = (I - \mathcal{F}^*)^{-1} \text{bvec}(\mathcal{Y}) \tag{11.59}$$

$$\text{Tr}(\mathcal{X}) = \left(\text{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}} (I - \mathcal{F})^{-1} \text{bvec}\left(I_{hMN}\right) \tag{11.60}$$

$$\text{Tr}(\mathcal{J}_k \mathcal{X}) = \left(\text{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}} (I - \mathcal{F})^{-1} \text{bvec}\left(\mathcal{J}_k\right) \tag{11.61}$$

First-Order MSD Expression

Course EE210B
Spring Quarter 2015

Proc. IEEE, vol. 102, no. 4, pp. 460-497, April 2014.
Foundations and Trends in Machine Learning, vol. 7, no. 4-5, pp. 311-801, July 2014.

MSD Performance

We now use the result of Theorem 11.2 to derive an expression for the MSD performance of each agent and for the entire network. We will do so by appealing to the useful low-rank approximation (9.244). Two observations are in place in relation to the forthcoming result (11.118). First, observe from (11.118) the interesting conclusion that the consensus and diffusion strategies represented by (8.46) are able to equalize the MSD performance across all agents for sufficiently small step-sizes.

MSD Performance

This is a reassuring property since it means that all agents, regardless of the quality of their data, will end up achieving similar performance levels. At the same time, we remark that although expression (11.118) suggests that the performance of consensus and diffusion strategies match to first-order in $\mu_{\max}$, differences in performance actually occur for larger step-sizes with ATC diffusion exhibiting superior performance. These differences are illustrated and explained further ahead in Example 11.4, and also Examples 11.11–11.13.

Recall#3: Low-Rank Approximation

Lemma 9.5 (Low-rank approximation). Assume the matrix P is primitive. For sufficiently small step-sizes, it holds that

$$(I-\mathcal{F})^{-1} = \left[(p\otimes p)(\mathbb{1}\otimes\mathbb{1})^{\mathsf{T}}\right]\otimes Z^{-1} + O(1) \tag{9.244}$$

$$Z \triangleq \sum_{k=1}^{N} q_k \left[(I_{hM}\otimes H_k) + (H_k^{\mathsf{T}}\otimes I_{hM})\right] \tag{9.245}$$

Recall#4: Lyapunov Equation

A similar analysis applies to the following continuous-time Lyapunov equation (also called a Sylvester equation):

$$XA^* + AX + Q = 0 \qquad \text{(F.43)}$$

In the continuous-time case, a stable matrix A is one whose eigenvalues lie in the open left-half plane (i.e., they have strictly negative real parts).

Recall#4: Lyapunov Equation

Lemma F.3 (Continuous-time Lyapunov equation). Consider the Lyapunov equation (F.43). The following facts hold:

(a) The solution X is unique if, and only if, $\lambda_k(A) + \lambda_\ell^*(A) \neq 0$ for all $k, \ell = 1, 2, \ldots, N$. In this case, the unique solution X is Hermitian.

(b) When A is stable (i.e., all its eigenvalues lie in the open left-half plane), the solution X is unique, Hermitian, and nonnegative-definite.

MSD Performance

Lemma 11.3 (Network MSD performance). Under the same conditions of Theorem 11.2, it holds that

$$\mathrm{MSD}_{\mathrm{dist},k} \;=\; \mathrm{MSD}_{\mathrm{dist,av}} = \frac{1}{2h}\mathrm{Tr}\left[\left(\sum_{k=1}^{N} q_k H_k\right)^{-1}\left(\sum_{k=1}^{N} q_k^2 G_k\right)\right] \qquad (11.118)$$

where $h = 1$ for real data and $h = 2$ for complex data.

Example #A (Real Data)

Lemma 11.3: For sufficiently small step-sizes:

$$\mathrm{MSD}_{\mathrm{dist},k} \;=\; \mathrm{MSD}_{\mathrm{dist,av}} = \;\; \frac{1}{2}\mathrm{Tr}\left[\left(\sum_{k=1}^{N} q_k H_k\right)^{-1}\left(\sum_{k=1}^{N} q_k^2 R_{s,k}\right)\right]$$

$$\alpha_{\mathrm{dist}} \;=\; 1 - 2\lambda_{\min}\left(\sum_{k=1}^{N} q_k H_k\right) \;+\; O\left(\mu_{\max}^{(N+1)/N}\right)$$

Proof

Proof. We establish the result for $h = 2$ without loss of generality by extending the argument from [71, 278] to the current context. According to definition (11.37), and expressions (11.45) and (11.61), we need to evaluate the following limit:

$$\mathrm{MSD}_{\mathrm{dist},k} = \mu_{\max} \cdot \left(\lim_{\mu_{\max} \to 0} \limsup_{i \to \infty} \frac{1}{\mu_{\max}} \frac{1}{h} \left(\mathrm{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}} (I - \mathcal{F})^{-1} \mathrm{bvec}\left(\mathcal{J}_k\right) \right) \tag{11.119}$$

We focus on the rightmost factor inside the above expression. Using (9.244), along with the first line in (9.275), we get:

$$\left(\mathrm{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}} (I - \mathcal{F})^{-1} \mathrm{bvec}(\mathcal{J}_k) \;=\; O(\mu_{\max}^2) + \tag{11.120}$$

$$\left(\mathrm{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}} (p \otimes I_{2M}) \otimes_b (p \otimes I_{2M}) \, Z^{-1} \left(\mathbb{1}^{\mathsf{T}} \otimes I_{2M}\right) \otimes_b \left(\mathbb{1}^{\mathsf{T}} \otimes I_{2M}\right) \mathrm{bvec}\left(\mathcal{J}_k\right)$$

Proof

Using the Kronecker product property (11.86), it is straightforward to verify that the last three terms combine into the following result, where the bvec operation is relative to blocks of size $2M \times 2M$:

$$\left[\left(\mathbb{1}^{\mathsf{T}} \otimes I_{2M}\right) \otimes_b \left(\mathbb{1}^{\mathsf{T}} \otimes I_{2M}\right)\right] \mathrm{bvec}(\mathcal{J}_k) = \mathrm{vec}(I_{2M}) \tag{11.121}$$

with the rightmost term involving the traditional (not block) vec operator. Let us therefore evaluate the matrix vector product:

$$x \triangleq Z^{-1}\mathrm{vec}(I_{2M}) \tag{11.122}$$

Proof

This vector is the unique solution to the linear system of equations

$$Zx = \text{vec}(I_{2M}) \tag{11.123}$$

or, equivalently, by using definition (9.245) for Z:

$$\left(\sum_{k=1}^{N} q_k(I_{2M} \otimes H_k)\right) x \;+\; \left(\sum_{k=1}^{N} q_k(H_k^{\mathsf{T}} \otimes I_{2M})\right) x \;=\; \text{vec}(I_{2M}) \tag{11.124}$$

Proof

Let $X = \text{unvec}(x)$ denote the $2M \times 2M$ matrix whose vector representation is x. Applying to each of the terms appearing on the left-hand side of the above expression the Kronecker product property (11.87), albeit using vec instead of bvec operations, namely,

$$\text{vec}(UCW) \;=\; (W^{\mathsf{T}} \otimes U)\text{vec}(C) \tag{11.125}$$

we find that

Proof

$$\left(\sum_{k=1}^{N} q_k(I_{2M} \otimes H_k)\right) x = \text{vec}\left\{\left(\sum_{k=1}^{N} q_k H_k\right) X\right\} \qquad (11.126)$$

$$\left(\sum_{k=1}^{N} q_k(H_k^{\mathsf{T}} \otimes I_{2M})\right) x = \text{vec}\left\{X\left(\sum_{k=1}^{N} q_k H_k\right)\right\} \qquad (11.127)$$

We conclude from these equalities and from (11.124) that X is the unique solution to the (continuous-time) Lyapunov equation (cf. Lemma F.3 from the appendix):

Proof

$$\left(\sum_{k=1}^{N} q_k H_k\right) X \;+\; X \left(\sum_{k=1}^{N} q_k H_k\right) \;=\; I_{2M} \tag{11.128}$$

It is straightforward to verify that the solution X is given by

$$X = \frac{1}{2}\left(\sum_{k=1}^{N} q_k H_k\right)^{-1} \tag{11.129}$$

Therefore, substituting into (11.120) gives

Proof

$$\left(\text{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}}\left(I-\mathcal{F}\right)^{-1}\text{bvec}(\mathcal{J}_k) \;=\; \tag{11.130}$$
$$\left(\text{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}}\left[(p\otimes I_{2M})\otimes_{\text{b}}(p\otimes I_{2M})\right]\text{vec}(X) \;+\; O(\mu_{\max}^2)$$

Using the Kronecker product properties (11.87) and (11.125) again, we obtain

Proof

Since X is $2M \times 2M$, we have $\mathrm{bvec}(X) = \mathrm{vec}(X)$.

$$
\begin{aligned}
&\left(\mathrm{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}} \left[(p \otimes I_{2M}) \otimes_{\mathsf{b}} (p \otimes I_{2M})\right] \mathrm{vec}(X) \\
&= \mathrm{Tr}\left[\mathrm{unbvec}\left\{(p \otimes I_{2M}) \otimes_{\mathsf{b}} (p \otimes I_{2M})\, \mathrm{vec}(X)\right\} \mathcal{Y}\right] \\
&= \mathrm{Tr}\left[(p \otimes I_{2M})\, X\, \left(p^{\mathsf{T}} \otimes I_{2M}\right) \mathcal{Y}\right] \\
&= \mathrm{Tr}\left[\left(p^{\mathsf{T}} \otimes I_{2M}\right) \mathcal{A}_2^{\mathsf{T}} \mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A}_2\, (p \otimes I_{2M})\, X\right] \\
&= \mathrm{Tr}\left[\left(q^{\mathsf{T}} \otimes I_{2M}\right) \mathcal{S}\, (q \otimes I_{2M})\, X\right] \\
&\overset{(11.129)}{=} \frac{1}{2}\mathrm{Tr}\left[\left(\sum_{k=1}^{N} q_k H_k\right)^{-1} \left(\sum_{k=1}^{N} q_k^2 G_k\right)\right] \qquad (11.131)
\end{aligned}
$$

Proof

Grouping terms we conclude that:

$$\left(\text{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}}\left(I-\mathcal{F}\right)^{-1}\text{bvec}(\mathcal{J}_k)$$

$$=\frac{1}{2}\text{Tr}\left[\left(\sum_{k=1}^{N}q_kH_k\right)^{-1}\left(\sum_{k=1}^{N}q_k^2G_k\right)\right]+O(\mu_{\max}^2) \qquad (11.132)$$

Proof

We know from the definition of the scalars $\{q_k\}$ in (9.7) that each q_k is proportional to $\mu_{\max}$. Therefore, the first term on the right-hand side of the above expression is linear in $\mu_{\max}$. Now substituting (11.132) into the right-hand side of (11.119) and computing the limit as $\mu_{\max} \to 0$, we arrive at expression (11.118) for the performance of the individual agents. Since this expression is independent of the index of the agent, by averaging over all agents, we find that the network performance is given by the same expression.

□

MSD Performance

Lemma 11.3 (Network MSD performance). Under the same conditions of Theorem 11.2, it holds that

$$\mathrm{MSD}_{\mathrm{dist},k} \;=\; \mathrm{MSD}_{\mathrm{dist,av}} = \frac{1}{2h}\mathrm{Tr}\left[\left(\sum_{k=1}^{N} q_k H_k\right)^{-1}\left(\sum_{k=1}^{N} q_k^2 G_k\right)\right] \qquad (11.118)$$

where $h = 1$ for real data and $h = 2$ for complex data.

Example #A (Real Data)

Lemma 11.3: For sufficiently small step-sizes:

$$\mathrm{MSD}_{\mathrm{dist},k} = \mathrm{MSD}_{\mathrm{dist,av}} = \frac{1}{2}\mathrm{Tr}\left[\left(\sum_{k=1}^{N} q_k H_k\right)^{-1}\left(\sum_{k=1}^{N} q_k^2 R_{s,k}\right)\right]$$

$$\alpha_{\mathrm{dist}} = 1 - 2\lambda_{\min}\left(\sum_{k=1}^{N} q_k H_k\right) + O\left(\mu_{\max}^{(N+1)/N}\right)$$

Example #11.2

Example 11.2 (MSD performance of consensus and diffusion networks). We specialize the main result of Lemma 11.3 to the consensus and diffusion strategies, which correspond to the choices $\{A_o, A_1, A_2\}$ shown earlier in (8.7)–(8.10) in terms of a single combination matrix A, namely,

$$\text{consensus:} \quad A_o = A, \;\; A_1 = I_N = A_2 \tag{11.133}$$

$$\text{CTA diffusion:} \quad A_1 = A, \;\; A_2 = I_N = A_o \tag{11.134}$$

$$\text{ATC diffusion:} \quad A_2 = A, \;\; A_1 = I_N = A_o \tag{11.135}$$

Example #11.2

In these cases, the Perron eigenvector p defined by (9.9) will correspond to the Perron eigenvector associated with A:

$$Ap = p, \quad \mathbb{1}^{\mathsf{T}} p = 1, \quad p_k > 0 \tag{11.136}$$

Consequently, the entries q_k defined by (9.7) will reduce to

$$q_k \;=\; \mu_k p_k \tag{11.137}$$

Using these facts in (11.118) we obtain

Example #11.2

where $h = 1$ for real data and $h = 2$ for complex data. Moreover, the convergence rate of the error variances, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i}\|^2$, towards this MSD value is determined by

$$\alpha_{\text{dist}} \;=\; 1 - 2\lambda_{\min}\left(\sum_{k=1}^{N}\mu_k p_k H_k\right) \;+\; O\left(\mu_{\max}^{(N+1)/N}\right) \qquad (11.139)$$

where $\alpha_{\text{dist}} \in (0,1)$. When A is doubly-stochastic, and the step-sizes are uniform across the agents so that $\mu_k \equiv \mu$, the above expressions reduce to

Example #11.2

$$\mathrm{MSD}_{\mathrm{dist,av}} = \frac{\mu}{2hN}\mathrm{Tr}\left[\left(\sum_{k=1}^{N} H_k\right)^{-1}\left(\sum_{k=1}^{N} G_k\right)\right] \tag{11.140}$$

$$\alpha_{\mathrm{dist}} = 1 - \frac{2\mu}{N}\lambda_{\min}\left(\sum_{k=1}^{N} H_k\right) + o(\mu) \tag{11.141}$$

Comparing these expressions with (5.65) and (5.67) we observe that, to first-order in μ, the distributed solution is able to match the performance of the centralized solution for doubly-stochastic policies.

Example #11.2

Observe further from (11.138) that, for sufficiently small step-sizes, the consensus and diffusion strategies are able to equalize the MSD performance across all agents. It is also instructive to compare expression (11.138) with (5.79) and (5.65) in the non-cooperative and centralized cases. Note that the effect of distributed cooperation results in the appearance of the scaling coefficients $\{p_k\}$; these factors are determined by the combination policy A. ■

Example #11.3

Example 11.3 (MSD performance of MSE networks — Case I). We revisit the setting of Example 6.3, where the data $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$ satisfy the linear regression model (6.14) and where the cost associated with each agent is the mean-square-error cost, $J_k(w) = \mathbb{E}\,|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}w|^2$. As mentioned earlier, we already know from Example 6.1 that, in this case, the reference vectors w^o and $w^\star$ coincide. We assume the agents employ uniform step-sizes and sense regression data with uniform covariance matrices, i.e., $\mu_k \equiv \mu$ and $R_{u,k} \equiv R_u$ for $k = 1, 2, \ldots, N$. We can assess the performance of the resulting consensus network (cf. Example 7.2) or diffusion network (cf. Example 7.3) as follows. In the current setting, and assuming complex data for generality, we know from (8.15) that

Example #11.3

$$R_{s,k} \triangleq \lim_{i\to\infty} \mathbb{E}\left[\boldsymbol{s}_{k,i}(w^o)\boldsymbol{s}^*_{k,i}(w^o)\,|\boldsymbol{\mathcal{F}}_{i-1}\right] = \sigma^2_{v,k}R_{u,k} \tag{11.142}$$

Therefore, using the definitions (11.12), we have:

$$H_k = \begin{bmatrix} R_u & 0 \\ 0 & R_u^{\mathsf{T}} \end{bmatrix} \equiv H, \qquad G_k = \sigma^2_{v,k}\begin{bmatrix} R_u & \times \\ \times & R_u^{\mathsf{T}} \end{bmatrix} \tag{11.143}$$

where the off-diagonal block entries of G_k are not needed since H_k is block-diagonal. Substituting into (11.138), and using $h = 2$ for complex data, we conclude that the MSD performance of consensus or diffusion LMS networks is given by:

$$\text{MSD}_{\text{dist},k} = \text{MSD}_{\text{dist,av}} = \frac{\mu M}{2}\left(\sum_{k=1}^{N} p_k^2\sigma^2_{v,k}\right) \tag{11.144}$$

Example #11.3

If the combination matrix A happens to be doubly stochastic, then $p = \mathbb{1}/N$. Substituting $p_k = 1/N$ into (11.144) gives

$$\text{MSD}_{\text{dist},k} = \text{MSD}_{\text{dist,av}} = \frac{\mu M}{2} \frac{1}{N^2} \left(\sum_{k=1}^{N} \sigma_{v,k}^2 \right) \tag{11.145}$$

which agrees with the expression that would result from (5.65) for the centralized LMS solution in the complex case, namely,

$$\text{MSD}_{\text{cent}} = \frac{\mu M}{2} \frac{1}{N} \left(\frac{1}{N} \sum_{k=1}^{N} \sigma_{v,k}^2 \right) \tag{11.146}$$

Example #11.3

Therefore, the distributed strategies are able to match the performance of the centralized solution for doubly stochastic combination policies. Observe though that, more generally, when A is not doubly-stochastic, the scaling factors $\{p_k^2\}$ appear in (11.144).

If the step-sizes were different across the agents, then we would instead obtain from (11.138) the following expression for the network performance:

$$\mathrm{MSD}_{\mathrm{dist},k} = \mathrm{MSD}_{\mathrm{dist,av}} = \frac{M}{2}\left(\sum_{k=1}^{N}\mu_k p_k\right)^{-1}\left(\sum_{k=1}^{N}\mu_k^2 p_k^2 \sigma_{v,k}^2\right) \quad (11.147)$$

Example #11.3

Another situation of interest is when the combination weights $\{a_{\ell k}\}$ are selected according to the averaging (or uniform) rule we encountered earlier in (8.89), namely,

$$a_{\ell k} = \begin{cases} 1/n_k, & \ell \in \mathcal{N}_k \\ 0, & \text{otherwise} \end{cases} \tag{11.148}$$

where

$$n_k \triangleq |\mathcal{N}_k| \tag{11.149}$$

denotes the size of the neighborhood of agent k (or its degree). In this case, the matrix A will be left-stochastic and the entries of the corresponding Perron eigenvector are given by:

Example #11.3

$$p_k = n_k \left(\sum_{m=1}^{N} n_m \right)^{-1} \tag{11.150}$$

Then, expression (11.144) gives

$$\text{MSD}_{\text{dist},k} = \text{MSD}_{\text{dist,av}} = \frac{\mu M}{2} \left(\sum_{k=1}^{N} n_k \right)^{-2} \left(\sum_{k=1}^{N} n_k^2 \sigma_{v,k}^2 \right) \tag{11.151}$$

which would reduce to (11.145) when the degrees of all agents are uniform, i.e., $n_k \equiv n$.

■

Example #11.4

Example 11.4 (MSD performance of MSE networks — Case II). We continue with the scenario of Example 11.3 for MSE networks except that we now assume that the regression covariance matrices are not necessarily uniform but chosen of the form $R_{u,k} = \sigma_{u,k}^2 I_M$. In this case, the expressions for $\{H_k, G_k\}$ in (11.143) become

$$H_k = \sigma_{u,k}^2 \begin{bmatrix} I_M & 0 \\ 0 & I_M \end{bmatrix}, \qquad G_k = \sigma_{v,k}^2 \, \sigma_{u,k}^2 \begin{bmatrix} I_M & \times \\ \times & I_M \end{bmatrix} \tag{11.152}$$

We can assess the performance of the resulting consensus network (cf. Example 7.2) or diffusion network (cf. Example 7.3) by substituting these values into (11.138), and using $h = 2$ for complex data, to get:

Example #11.4

$$\mathrm{MSD}_{\mathrm{dist},k} = \mathrm{MSD}_{\mathrm{dist,av}} = \frac{M}{2}\left(\sum_{k=1}^{N}\mu_k^2 p_k^2 \sigma_{v,k}^2 \sigma_{u,k}^2\right)\left(\sum_{k=1}^{N}\mu_k p_k \sigma_{u,k}^2\right)^{-1} \tag{11.153}$$

If the combination matrix A happens to be doubly stochastic, then $p = \mathbb{1}/N$. Substituting $p_k = 1/N$ into (11.153) gives

$$\mathrm{MSD}_{\mathrm{dist},k} = \mathrm{MSD}_{\mathrm{dist,av}} = \frac{M}{2N}\left(\sum_{k=1}^{N}\mu_k^2 \sigma_{v,k}^2 \sigma_{u,k}^2\right)\left(\sum_{k=1}^{N}\mu_k \sigma_{u,k}^2\right)^{-1} \tag{11.154}$$

Example #11.4

On the other hand, if the combination weights $\{a_{\ell k}\}$ are selected according to the averaging rule (11.148), we would then substitute (11.150) into (11.153) to give

$$\begin{aligned}\mathrm{MSD}_{\mathrm{dist},k} &= \mathrm{MSD}_{\mathrm{dist,av}} \\ &= \frac{M}{2}\left(\sum_{k=1}^{N} n_k\right)^{-1}\left(\sum_{k=1}^{N} \mu_k^2 n_k^2 \sigma_{v,k}^2 \sigma_{u,k}^2\right)\left(\sum_{k=1}^{N} \mu_k n_k \sigma_{u,k}^2\right)^{-1} \qquad (11.155)\end{aligned}$$

Example #11.4

If the step-sizes are uniform across all agents, the above expression becomes

$$\begin{aligned}\mathrm{MSD}_{\mathrm{dist},k} &= \mathrm{MSD}_{\mathrm{dist,av}} \\ &= \frac{\mu M}{2}\left(\sum_{k=1}^{N} n_k\right)^{-1}\left(\sum_{k=1}^{N} n_k^2\sigma_{v,k}^2\sigma_{u,k}^2\right)\left(\sum_{k=1}^{N} n_k\sigma_{u,k}^2\right)^{-1} \qquad (11.156)\end{aligned}$$

Example #11.4

We illustrate these results numerically for the case of the averaging rule (11.148) with uniform step-sizes across the agents. Figure 11.1 shows the connected network topology with $N = 20$ agents used for this simulation, with the measurement noise variances, $\{\sigma_{v,k}^2\}$, and the power of the regression data, assumed of the form $R_{u,k} = \sigma_{u,k}^2 I_M$, shown in the plots of Figure 11.2, respectively. All agents are assumed to have a non-trivial self-loop so that the neighborhood of each agent includes the agent itself as well. The resulting network is therefore strongly-connected.

Example #11.4

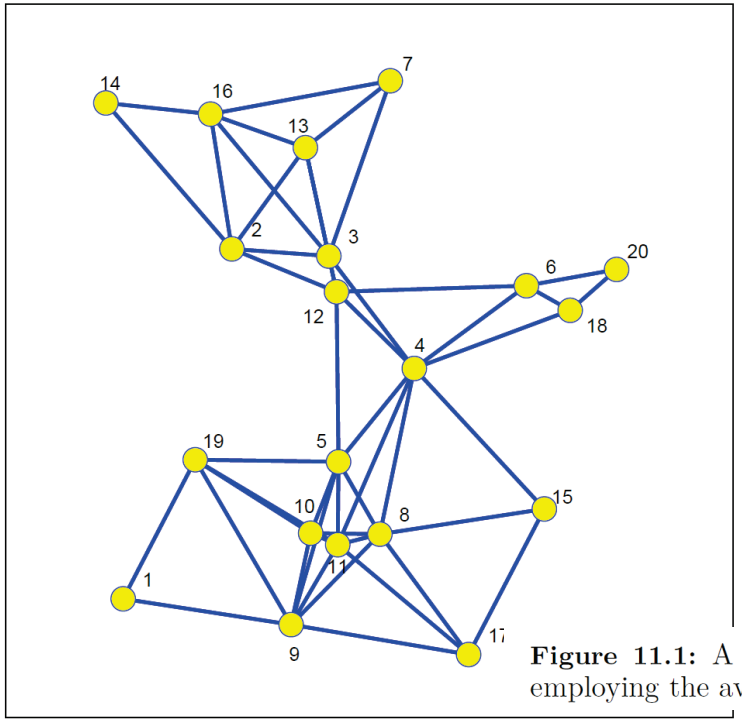


Figure 11.1: A connected network topology consisting of $N = 20$ agents employing the averaging rule (11.148).

Example #11.4

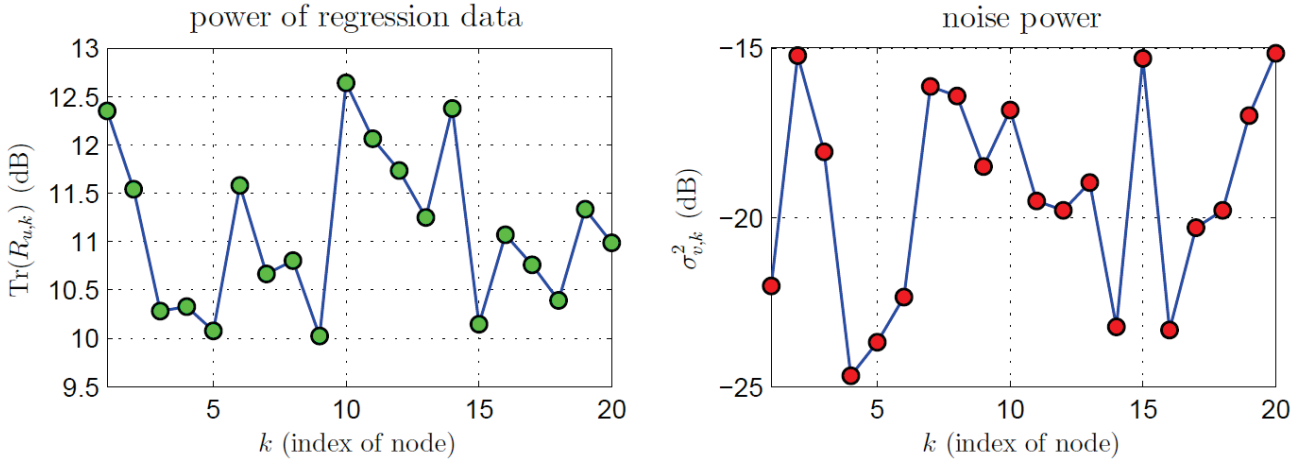


Figure 11.2: Regression data power (left) and measurement noise profile (right) across all agents in the network. The covariance matrices are assumed to be of the form $R_{u,k} = \sigma^2_{u,k} I_M$, and the noise and regression data are Gaussian distributed in this simulation.

Example #11.4

Figures 11.3 and 11.4 plot the evolution of the ensemble-average learning curves, $\frac{1}{N}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$, for consensus, ATC diffusion, and CTA diffusion for two choices of the step-size parameter: a smaller value at $\mu = 0.002$ and a second larger value at $\mu = 0.01$. The curves are obtained by averaging the trajectories $\{\frac{1}{N}\|\widetilde{\boldsymbol{w}}_i\|^2\}$ over 100 repeated experiments. The labels on the vertical axes in the figures refer to the learning curve $\frac{1}{N}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ by writing $\mathrm{MSD}_{\mathrm{dist,av}}(i)$, with an iteration index i. Each experiment involves running the consensus (7.14) or diffusion (7.22)–(7.23) LMS recursions with $h = 2$ on complex-valued data $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$ generated according to the model $\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i}w^o + \boldsymbol{v}_k(i)$, with $M = 10$. The unknown vector w^o is generated randomly and its norm is normalized to one.

Example #11.4

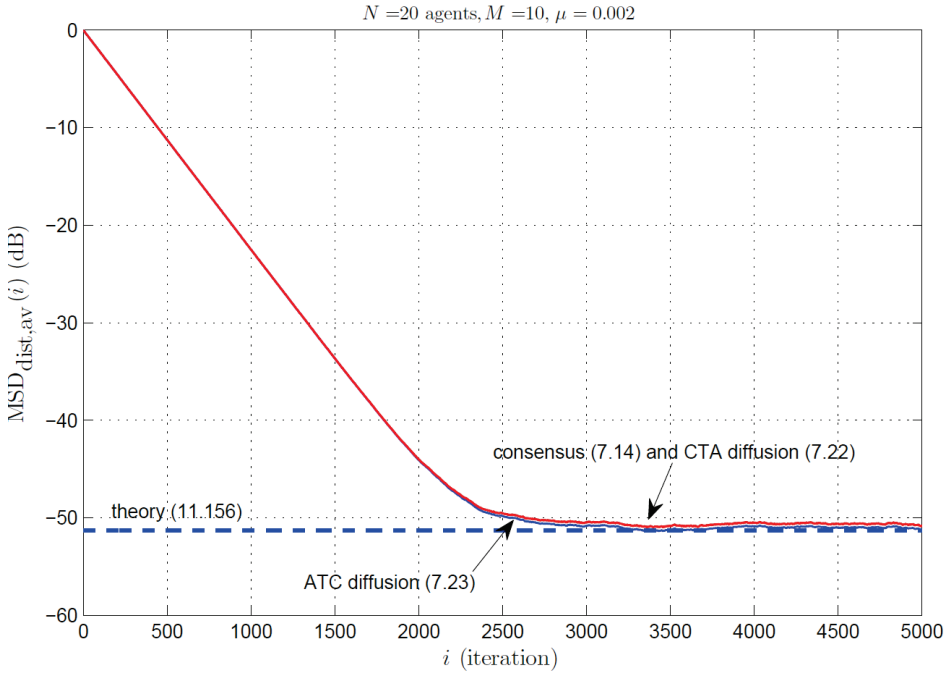


Figure 11.3

Example #11.4

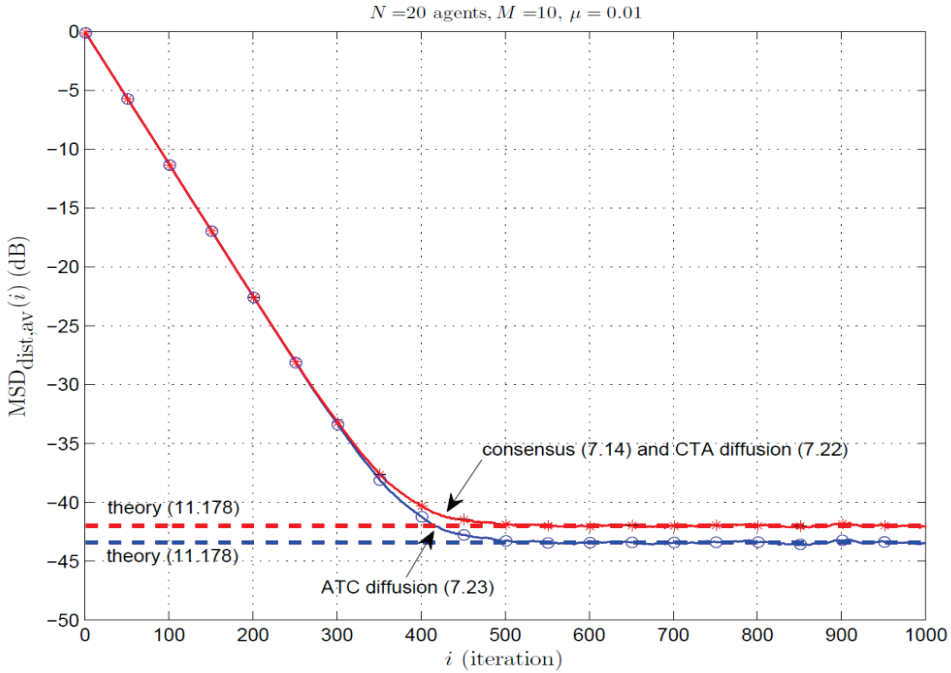


Figure 11.4

Example #11.4

Table 11.1: MSD values predicted by expressions (11.178) and (11.156) at the larger step-size value, $\mu = 0.01$.

algorithm	**result (11.178)**	**result (11.156)**
consensus strategy (7.14)	−42.00 dB	−44.34 dB
CTA diffusion strategy (7.22)	−42.00 dB	−44.34 dB
ATC diffusion strategy (7.23)	−43.42 dB	−44.34 dB

Example #11.4

It is observed in Figure 11.3 that the learning curves tend to the same MSD value predicted by the theoretical expression (11.156), which provides a good approximation for the performance of distributed strategies for small step-sizes. However, it is observed in Figure 11.4 that once the step-size value is increased, differences in MSD performance arise among the algorithms, with ATC diffusion exhibiting the lowest (i.e., best) MSD value. The horizontal lines in this second figure represent the MSD levels that are predicted by future expression (11.178). This latter expression reflects the effect of higher-order terms in $\mu_{\max}$ and generally leads to an enhanced representation for the error variance of the distributed strategies, while expression (11.156), which

Example #11.4

is the basis for the results in this example, is an expression for the MSD that is accurate to first-order in $\mu_{\max}$. Table 11.1 lists the MSD values that are predicted by expressions (11.178) and (11.156) at the larger step-size value, $\mu = 0.01$.

■

Example #11.5

Example 11.5 (Is cooperation always beneficial?). We continue with the discussion from Example 11.3 over MSE networks. If each agent in the network were to estimate w^o on its own in a non-cooperative manner by running its individual LMS learning rule (3.125), then we know from (4.186) that each agent will attain the MSD level shown below:

$$\text{MSD}_{\text{ncop},k} = \frac{\mu M}{2}\sigma_{v,k}^2 \tag{11.157}$$

along with the average performance across all N agents given by:

$$\text{MSD}_{\text{ncop,av}} = \frac{\mu M}{2}\left(\frac{1}{N}\sum_{k=1}^{N}\sigma_{v,k}^2\right) \tag{11.158}$$

Example #11.5

Now assume A is doubly stochastic. Comparing (11.145) with (11.158), it is obvious that

$$\mathrm{MSD}_{\mathrm{dist,av}} = \frac{1}{N}\,\mathrm{MSD}_{\mathrm{ncop,av}} \tag{11.159}$$

which shows that, for MSE networks, the consensus and diffusion strategies outperform the average performance of the non-cooperative strategy by a factor of N. But how do the performance metrics of an agent compare to

Example #11.5

each other in the distributed and non-cooperative modes of operation? From (11.145) and (11.157) we observe that if the noise variance is uniform across all agents, i.e., $\sigma_{v,k}^2 \equiv \sigma_v^2$, then the MSD of each individual agent in the distributed solution will be smaller by the same factor N than their non-cooperative performance. However, when the noise profile varies across the agents, then the performance metrics of an individual agent in the distributed and non-cooperative solutions cannot be compared directly: one can be larger than the other depending on the noise profile. For example, for $N = 2$, $\sigma_{v,1}^2 = 1$, and $\sigma_{v,2}^2 = 9$, agent 1 will not benefit from cooperation while agent 2 will. ■

Example #11.6

Example 11.6 (MSD performance of MSE networks — Case III). We reconsider the setting of Examples 8.8 and 8.11, which deals with a *variation* of MSE networks where the data model at each agent is instead assumed to be given by

$$\boldsymbol{d}_k(i) \;=\; \boldsymbol{u}_{k,i} w_k^o \;+\; \boldsymbol{v}_k(i) \tag{11.160}$$

with the model vectors, w_k^o, being possibly different at the various agents. We explained in Example 8.11 that the gradient noise process at agent k is given by expression (8.127), namely,

$$\boldsymbol{s}_{k,i}(\boldsymbol{\phi}_{k,i-1}) \;=\; \frac{2}{h}\left(R_{u,k} - \boldsymbol{u}_{k,i}^* \boldsymbol{u}_{k,i}\right)\left(w_k^o - \boldsymbol{\phi}_{k,i-1}\right) \;-\; \frac{2}{h}\boldsymbol{u}_{k,i}^* \boldsymbol{v}_k(i) \tag{11.161}$$

Example #11.6

By repeating the arguments of Example 8.8 for the general distributed strategy (8.5), we can similarly show that the limit point, $w^\star$, of the network is given by a relation similar to (8.86), namely,

$$w^\star = \left(\sum_{k=1}^{N} q_k R_{u,k}\right)^{-1} \left(\sum_{k=1}^{N} q_k R_{u,k} w_k^o\right) \tag{11.162}$$

where the positive scalars $\{q_k\}$ are the entries of the vector q defined by (8.50). Using (11.161) we can evaluate the second-order moment $R_{s,k}$ defined by (11.8) as follows. We introduce the difference

$$z_k \stackrel{\Delta}{=} w_k^o - w^\star, \quad k = 1, 2, \ldots, N \tag{11.163}$$

Example #11.6

It is clear that $z_k = 0$ when all w_k^o coincide at the same location w^o, in which case we get $w^\star = w^o$. In general though, the perturbation vectors, $\{z_k\}$ need not be zero. From (11.161), and using the conditions imposed on the regression data and noise processes across the agents from Example 6.3, we find that

$$R_{s,k} = \frac{4}{h^2}\mathbb{E}\left(R_{u,k} - \boldsymbol{u}_{k,i}^*\boldsymbol{u}_{k,i}\right) z_k z_k^* \left(R_{u,k} - \boldsymbol{u}_{k,i}^*\boldsymbol{u}_{k,i}\right) + \frac{4}{h^2}\sigma_{v,k}^2 R_{u,k} \tag{11.164}$$

The first term on the right-hand side involves a fourth-order moment in the regression data. To evaluate this term in closed-form, we assume that the regression data is circular and Gaussian-distributed. In that case, it is known that for any $M \times M$ Hermitian matrix W_k it holds that [206, p.11]:

Example #11.6

$$\mathbb{E}\left(\boldsymbol{u}_{k,i}\boldsymbol{u}_{k,i}^{*}W_k\boldsymbol{u}_{k,i}\boldsymbol{u}_{k,i}^{*}\right) = R_{u,k}\mathrm{Tr}(W_kR_{u,k}) + \frac{2}{h}R_{u,k}W_kR_{u,k} \qquad (11.165)$$

This expression shows how the (weighted) fourth-order moment of the process $\boldsymbol{u}_{k,i}$ is determined by its second-order moment, $R_{u,k}$. Let

$$W_k = z_kz_k^{*} \qquad (11.166)$$

which is a rank-one nonnegative definite Hermitian matrix. Expanding the first term on the right-hand side of (11.164) and using (11.165), we conclude that

Example #11.6

$$R_{s,k} = \frac{4}{h^2}\sigma_{v,k}^2 R_{u,k} \;+\; \frac{4}{h^2}R_{u,k}\mathrm{Tr}(W_k R_{u,k}) \;+\; \frac{4}{h^2}\left(\frac{2}{h}-1\right)R_{u,k}W_k R_{u,k} \tag{11.167}$$

In particular, for complex data, the above result evaluates to the following using $h = 2$:

$$R_{s,k} \;=\; \sigma_{v,k}^2 R_{u,k} \;+\; R_{u,k}\|z_k\|_{R_{u,k}}^2 \quad \text{(complex data)} \tag{11.168}$$

Each agent k in the network is associated with an individual cost of the form $J_k(w) = \mathbb{E}\,|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}w|^2$. We now assume that the regression covariance matrices are of the form $R_{u,k} = \sigma_{u,k}^2 I_M$. In this case, expression (11.168) for $R_{s,k}$ simplifies to

Example #11.6

$$\begin{aligned} R_{s,k} &= \left(\sigma_{v,k}^2 + \sigma_{u,k}^2 \|z_k\|^2\right) \sigma_{u,k}^2 I_M \\ &\triangleq \bar{\sigma}_{v,k}^2 \sigma_{u,k}^2 I_M \quad \text{(complex data)} \end{aligned} \tag{11.169}$$

where we introduced the modified noise variance

$$\bar{\sigma}_{v,k}^2 \triangleq \sigma_{v,k}^2 + \sigma_{u,k}^2 \|z_k\|^2 \tag{11.170}$$

Example #11.6

Consequently, the expressions for $\{H_k, G_k\}$ become (compare with (11.152)):

$$H_k = \sigma_{u,k}^2 \begin{bmatrix} I_M & 0 \\ 0 & I_M \end{bmatrix}, \qquad G_k = \bar{\sigma}_{v,k}^2 \sigma_{u,k}^2 \begin{bmatrix} I_M & \times \\ \times & I_M \end{bmatrix} \tag{11.171}$$

We can assess the performance of the resulting consensus network (cf. Example 7.2) or diffusion network (cf. Example 7.3) by substituting these values into (11.138), and using $h = 2$ for complex data, to get:

$$\mathrm{MSD}_{\mathrm{dist},k} = \mathrm{MSD}_{\mathrm{dist,av}} = \frac{M}{2} \left(\sum_{k=1}^{N} \mu_k^2 p_k^2 \bar{\sigma}_{v,k}^2 \sigma_{u,k}^2 \right) \left(\sum_{k=1}^{N} \mu_k p_k \sigma_{u,k}^2 \right)^{-1} \tag{11.172}$$

Example #11.6

If the combination matrix A happens to be doubly stochastic, then $p = \mathbb{1}/N$. Substituting $p_k = 1/N$ into (11.172) gives

$$\mathrm{MSD}_{\mathrm{dist},k} = \mathrm{MSD}_{\mathrm{dist,av}} = \frac{M}{2N}\left(\sum_{k=1}^{N}\mu_k^2\bar{\sigma}_{v,k}^2\sigma_{u,k}^2\right)\left(\sum_{k=1}^{N}\mu_k\sigma_{u,k}^2\right)^{-1} \tag{11.173}$$

On the other hand, if the combination weights $\{a_{\ell k}\}$ are selected according to the averaging rule (11.148), we would then substitute (11.150) into (11.153) to give

Example #11.6

$$\mathrm{MSD}_{\mathrm{dist},k} = \mathrm{MSD}_{\mathrm{dist,av}}$$

$$= \frac{M}{2}\left(\sum_{k=1}^{N} n_k\right)^{-1}\left(\sum_{k=1}^{N} \mu_k^2 n_k^2 \bar{\sigma}_{v,k}^2 \sigma_{u,k}^2\right)\left(\sum_{k=1}^{N} \mu_k n_k \sigma_{u,k}^2\right)^{-1} \tag{11.174}$$

If the step-sizes are uniform across all agents, the above expression becomes

$$\mathrm{MSD}_{\mathrm{dist},k} = \mathrm{MSD}_{\mathrm{dist,av}} = \frac{\mu M}{2}\left(\sum_{k=1}^{N} n_k\right)^{-1}\left(\sum_{k=1}^{N} n_k^2 \bar{\sigma}_{v,k}^2 \sigma_{u,k}^2\right)\left(\sum_{k=1}^{N} n_k \sigma_{u,k}^2\right)^{-1} \tag{11.175}$$

We illustrated this result numerically earlier in Figure 8.5 while discussing the convergence of the network towards its Pareto limit point.

■

Example #11.6

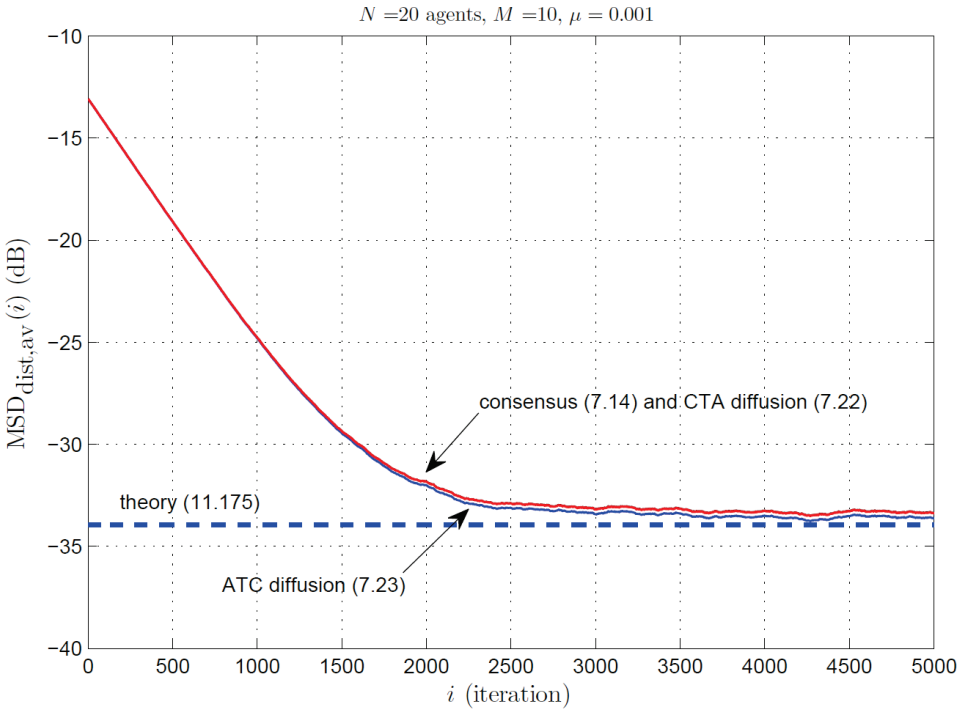


Figure 8.5

Example #11.7

Example 11.7 (Higher-order MSD terms). We explained earlier in Sec. 4.5, while motivating the definition of the MSD metric, that expressions of the form (11.37) help assess the size of the error variance, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i}\|^2$, in steady-state and for sufficiently small step-sizes (i.e., in the slow adaptation regime).

The computation leads to an expression for the MSD that is *first-order* in $\mu_{\max}$, as can be ascertained from (11.118).

Recall#5: First-Order Expression

$$\text{MSD} \triangleq \mu \cdot \left(\lim_{\mu \to 0} \limsup_{i \to \infty} \frac{1}{\mu} \mathbb{E} \, \|\widetilde{\boldsymbol{w}}_i\|^2 \right)$$

Recall #6: Ignoring $O(\mu_{\max}^2)$ Term

$$\mathrm{MSD}_{\mathrm{dist},k} = \mu_{\max} \cdot \left(\lim_{\mu_{\max} \to 0} \limsup_{i \to \infty} \frac{1}{\mu_{\max}} \frac{1}{h} \left(\mathrm{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right) \right)^{\mathsf{T}} (I - \mathcal{F})^{-1} \mathrm{bvec}\left(\mathcal{J}_k\right) \right) \tag{11.119}$$

$$\left(\mathrm{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right) \right)^{\mathsf{T}} (I - \mathcal{F})^{-1} \mathrm{bvec}(\mathcal{J}_k) = O(\mu_{\max}^2) + \tag{11.120}$$

$$\left(\mathrm{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right) \right)^{\mathsf{T}} (p \otimes I_{2M}) \otimes_b (p \otimes I_{2M}) \, Z^{-1} \left(\mathbb{1}^{\mathsf{T}} \otimes I_{2M} \right) \otimes_b \left(\mathbb{1}^{\mathsf{T}} \otimes I_{2M} \right) \mathrm{bvec}\left(\mathcal{J}_k\right)$$

Example #11.7

If we revisit the derivation of (11.118) in the proof of Lemma 11.3, we will observe that this expression was obtained by eliminating the contribution of the higher-order term, $O(\mu_{\max}^2)$, which appears in the expansion (11.120). We can motivate an alternative expression for assessing the size of the error variance, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i}\|^2$, by retaining the higher-order term that is available (i.e., known) rather than neglecting it. It is expected that, by doing so, the resulting performance expression will generally provide a more accurate representation for the error variance, especially at larger step-sizes; we illustrated this behavior already in the simulations of Example 11.4 — recall Figure 11.4. The alternative performance expression can be motivated as follows.

Example #11.7

Similarly to (4.83)–(4.84), the argument that led to (11.45) would establish the following two expressions for the limit superior and limit inferior of the error variance at each agent k (see, e.g., (11.107) and (11.109)):

$$\limsup_{i\to\infty} \frac{1}{2}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i}^{e}\|^2 = \frac{1}{h}\mathrm{Tr}(\mathcal{J}_k\mathcal{X}) + O\left(\mu_{\max}^{1+\gamma_m}\right) \tag{11.176}$$

$$\liminf_{i\to\infty} \frac{1}{2}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i}^{e}\|^2 = \frac{1}{h}\mathrm{Tr}(\mathcal{J}_k\mathcal{X}) - O\left(\mu_{\max}^{1+\gamma_m}\right) \tag{11.177}$$

Example #11.7

with the same common positive constant $\mathrm{Tr}(\mathcal{J}_k\mathcal{X})$; this constant is equal to the quantity that appears on the left-hand side of (11.120). Relations (11.176)–(11.177) indicate that we can also employ the quantity $\frac{1}{h}\mathrm{Tr}(\mathcal{J}_k\mathcal{X})$ to assess the size of the error variance, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i}\|^2$, in steady-state for small step-sizes. Subsequently, by averaging over all agents, we can similarly use the quantity $\frac{1}{hN}\mathrm{Tr}(\mathcal{X})$ to assess the size of the network error variance, $\frac{1}{N}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$, also in steady-state and for small step-sizes. If we recall (11.58), then this argument suggests the following alternative expressions for evaluating the network error variance:

Recall #7: Series Expression

$$\mathcal{X} = \sum_{n=0}^{\infty} \mathcal{B}^n \mathcal{Y} \left(\mathcal{B}^*\right)^n \tag{11.58}$$

$$\mathrm{bvec}(\mathcal{X}) = (I - \mathcal{F}^*)^{-1} \mathrm{bvec}(\mathcal{Y}) \tag{11.59}$$

$$\mathrm{Tr}(\mathcal{X}) = \left(\mathrm{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}} (I - \mathcal{F})^{-1} \mathrm{bvec}\left(I_{hMN}\right) \tag{11.60}$$

$$\mathrm{Tr}(\mathcal{J}_k \mathcal{X}) = \left(\mathrm{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}} (I - \mathcal{F})^{-1} \mathrm{bvec}\left(\mathcal{J}_k\right) \tag{11.61}$$

Example #11.7

$$\mathrm{MSD}_{\mathrm{dist,av}} = \frac{1}{hN}\sum_{n=0}^{\infty}\mathrm{Tr}\left[\mathcal{B}^n\mathcal{Y}\left(\mathcal{B}^*\right)^n\right] \tag{11.178}$$

$$= \frac{1}{hN}\left(\mathrm{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}}\left(I-\mathcal{F}\right)^{-1}\mathrm{bvec}\left(I_{hMN}\right) \tag{11.179}$$

where we continue to use the notation MSD to represent this value. As we already know from the proof of Lemma 11.3, if we expand the right-hand side of (11.179) in terms of powers of $\mu_{\max}$, then the first term in this expansion (i.e., the one that is linear in $\mu_{\max}$) will be given by expression (11.118).

■

Excess-Risk Performance

Course EE210B
Spring Quarter 2015

Proc. IEEE, vol. 102, no. 4, pp. 460-497, April 2014.
Foundations and Trends in Machine Learning, vol. 7, no. 4-5, pp. 311-801, July 2014.

ER Performance

Theorem 11.4 (Network ER performance). Consider a network of N interacting agents running the distributed strategy (8.46) with a primitive matrix $P = A_1 A_o A_2$. Assume the aggregate cost (9.10) and the individual costs, $J_k(w)$, satisfy the conditions in Assumptions 6.1 and 10.1. Assume further that the first and fourth-order moments of the gradient noise process satisfy the conditions of Assumption 8.1 with the second-order moment condition (8.115) replaced by the fourth-order moment condition (8.121). Assume also (11.11). Then, it holds that

ER Performance

$$\limsup_{i\to\infty} \frac{1}{2}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i-1}^{e}\|_{\bar{H}}^{2} = \frac{1}{2}\mathrm{Tr}(\mathcal{Q}_k\mathcal{X}) + O\left(\mu_{\max}^{1+\gamma_m}\right) \quad (11.180)$$

$$\limsup_{i\to\infty} \frac{1}{2N}\left(\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}^{e}\|_{(I_N\otimes\bar{H})}^{2}\right) = \frac{1}{2N}\mathrm{Tr}(\bar{\mathcal{H}}\mathcal{X}) + O\left(\mu_{\max}^{1+\gamma_m}\right) \quad (11.181)$$

for the same quantities defined earlier in Theorem 11.2 and where

$$\bar{\mathcal{H}} = I_N \otimes \bar{H} = \mathrm{diag}\{\bar{H},\bar{H},\ldots,\bar{H}\} \quad (11.182)$$

$$\mathcal{Q}_k = \mathrm{diag}\{\,0_{hM},\ldots,0_{hM},\bar{H},0_{hM},\ldots,0_{hM}\,\} \quad (11.183)$$

ER Performance

with the matrix $\bar{H}$ defined by (11.36) appearing in the $k-$block location of $\mathcal{Q}_k$. Moreover, it further holds that

$$\mathrm{Tr}(\mathcal{Q}_k\mathcal{X}) = \left(\mathrm{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}}(I-\mathcal{F})^{-1}\mathrm{bvec}\left(\mathcal{Q}_k\right) \tag{11.184}$$

$$\mathrm{Tr}(\bar{\mathcal{H}}\mathcal{X}) = \left(\mathrm{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}}(I-\mathcal{F})^{-1}\mathrm{bvec}\left(\bar{\mathcal{H}}\right) \tag{11.185}$$

ER Performance

and, for large enough i, the convergence rate of the excess-risk measure towards its steady-state region (11.180) is given by the same expression (11.47). Furthermore, the ER performance for the individual agents and for the network are given by:

$$\mathrm{ER}_{\mathrm{dist},k} = \mathrm{ER}_{\mathrm{dist,av}} = \frac{h}{4}\left(\sum_{k=1}^{N} q_k\right)^{-1} \mathrm{Tr}\left(\sum_{k=1}^{N} q_k^2 R_{s,k}\right) \tag{11.186}$$

Example #B (Real Data)

Theorem 11.4: For sufficiently small step-sizes:

$$\mathrm{ER}_{\mathrm{dist},k} = \mathrm{ER}_{\mathrm{dist,av}} = \frac{1}{4}\left(\sum_{k=1}^{N} q_k\right)^{-1} \mathrm{Tr}\left(\sum_{k=1}^{N} q_k^2 R_{s,k}\right)$$

Proof

Proof. We start from relation (11.84) but select Σ now as the solution to the following Lyapunov equation:

$$\Sigma - \mathcal{B}^{*}\Sigma\mathcal{B} = \bar{\mathcal{H}} \tag{11.187}$$

and repeat the argument that led to (11.106)–(11.107) to conclude that expressions (11.180)–(11.181) hold.

With regards to expression (11.186), we first note from (11.35) and (11.180) that we need to evaluate the limit:

$$\mathrm{ER}_{\mathrm{dist},k} = \mu_{\max} \cdot \left(\lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} \frac{1}{\mu_{\max}} \left(\mathrm{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}} (I-\mathcal{F})^{-1} \mathrm{bvec}(\mathcal{Q}_k) \right) \tag{11.188}$$

Proof

We focus on the right-most factor inside the above expression. Using the low-rank factorization (9.244), we have

$$\left(\text{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}}(I-\mathcal{F})^{-1}\text{bvec}(\mathcal{Q}_k) \;=\; O(\mu_{\max}^2)+ \tag{11.189}$$

$$\left(\text{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}}(p\otimes I_{2M})\otimes_b(p\otimes I_{2M})\,Z^{-1}\left(\mathbb{1}^{\mathsf{T}}\otimes I_{2M}\right)\otimes_b\left(\mathbb{1}^{\mathsf{T}}\otimes I_{2M}\right)\text{bvec}\left(\mathcal{Q}_k\right)$$

Using the block Kronecker product property (11.86), it can be verified that

$$\left(\mathbb{1}^{\mathsf{T}}\otimes I_{2M}\right)\otimes_b\left(\mathbb{1}^{\mathsf{T}}\otimes I_{2M}\right)\text{bvec}(\mathcal{Q}_k) \;=\; \text{vec}(\bar{H}) \tag{11.190}$$

Proof

Let $x = Z^{-1}\text{vec}(\bar{H})$. Then, the same argument that led to (11.128) will show that the $2M \times 2M$ matrix $X = \text{unvec}(x)$ is the unique solution to the Lyapunov equation

$$\left(\sum_{k=1}^{N} q_k\right) \bar{H} X \;+\; X \bar{H} \left(\sum_{k=1}^{N} q_k\right) \;=\; \bar{H} \tag{11.191}$$

so that

$$X \;=\; \frac{1}{2}\left(\sum_{k=1}^{N} q_k\right)^{-1} I_{2M} \tag{11.192}$$

Proof

Repeating the derivation that led to (11.132) we arrive at

$$\left(\text{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}}(I-\mathcal{F})^{-1}\text{bvec}(\mathcal{H}) = \frac{1}{2}\left(\sum_{k=1}^{N} q_k\right)^{-1}\text{Tr}\left(\sum_{k=1}^{N} q_k^2 G_k\right) + O(\mu_{\max}^2) \tag{11.193}$$

Substituting into the right-hand side of (11.188) and evaluating the limit we arrive at (11.186) after recalling from (11.12) that

$$\text{Tr}(G_k) = \begin{cases} R_{s,k} & \text{(real data)} \\ 2R_{s,k} & \text{(complex data)} \end{cases} \tag{11.194}$$

□

Example #11.8

Example 11.8 (ER performance of consensus and diffusion networks). We specialize the result of Theorem 11.4 to the same consensus and diffusion strategies from Example 11.2. In this case we get

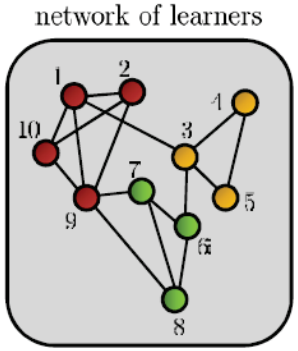
$$\mathrm{ER}_{\mathrm{dist},k} = \mathrm{ER}_{\mathrm{dist,av}} = \frac{h}{4}\mathrm{Tr}\left[\left(\sum_{k=1}^{N}\mu_k p_k\right)^{-1}\left(\sum_{k=1}^{N}\mu_k^2 p_k^2 R_{s,k}\right)\right] \qquad (11.195)$$

where $h = 1$ for real data and $h = 2$ for complex data. When the step-sizes are uniform across all agents, $\mu_k \equiv \mu$, and using the fact that the entries p_k add up to one, the above expression simplifies to

$$\mathrm{ER}_{\mathrm{dist},k} = \mathrm{ER}_{\mathrm{dist,av}} = \frac{\mu h}{4}\left(\sum_{k=1}^{N} p_k^2 R_{s,k}\right) \qquad (11.196)$$

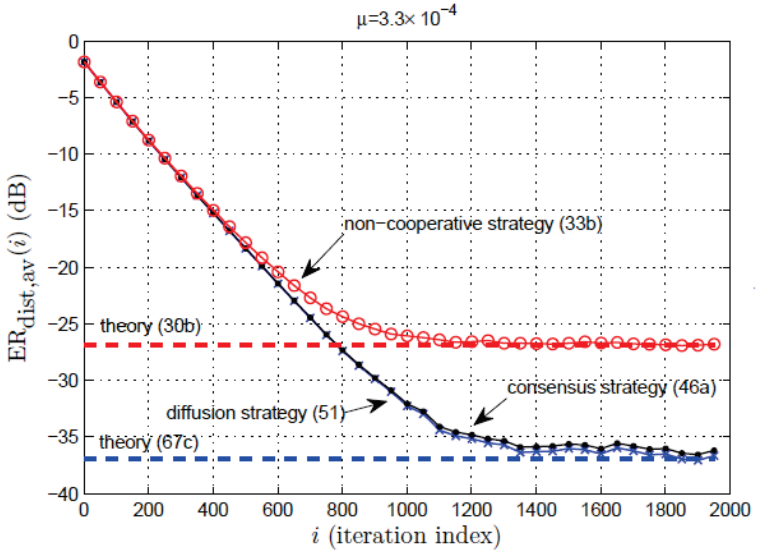
■

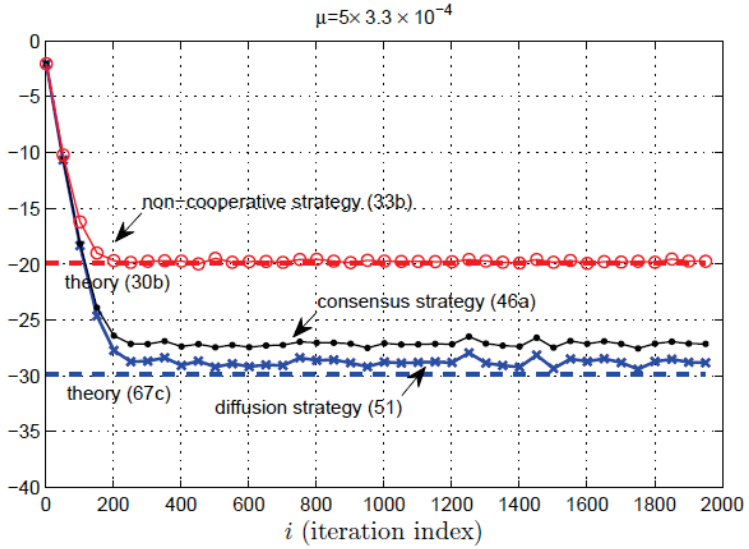
Example #C (Network of Learners)



(Metropolis rule)

$$\rho = 10$$





Example #11.9

Example 11.9 (Performance of diffusion learner). We generalize the scenario of Example 7.4 and consider a collection of N learners cooperating to minimize some arbitrary strongly-convex function $J(w)$ over a strongly-connected network, namely,

$$w^o \triangleq \arg\min_w J(w) \tag{11.197}$$

where $J(w)$ is the average of some loss measure, say, $J(w) = \mathbb{E}\, Q(w; \boldsymbol{x}_{k,i})$. As before, each learner k receives a streaming sequence of real-valued data vectors $\{\boldsymbol{x}_{k,i}, i = 1, 2, \ldots\}$ that arise from some fixed distribution $\mathcal{X}$. We assume the agents run a consensus or diffusion strategy, say, the ATC diffusion strategy (7.19):

Example #11.9

$$\left\{\begin{array}{rcl} \boldsymbol{\psi}_{k,i} &=& \boldsymbol{w}_{k,i-1} - \mu_k \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{k,i-1}; \boldsymbol{x}_{k,i}) \\ \boldsymbol{w}_{k,i} &=& \displaystyle\sum_{\ell \in \mathcal{N}_k} a_{\ell k}\, \boldsymbol{\psi}_{\ell,i} \end{array}\right. \tag{11.198}$$

The gradient noise vector corresponding to each individual agent k is given by

$$\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) \;=\; \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{k,i-1}; \boldsymbol{x}_{k,i}) \;-\; \nabla_{w^{\mathsf{T}}} \mathbb{E}\, Q(\boldsymbol{w}_{k,i-1}; \boldsymbol{x}_{k,i}) \tag{11.199}$$

so that

$$\boldsymbol{s}_{k,i}(w^o) \;=\; \nabla_{w^{\mathsf{T}}} Q(w^o; \boldsymbol{x}_{k,i}) \tag{11.200}$$

Example #11.9

Since we are assuming the distribution of the random process $\boldsymbol{x}_{k,i}$ is stationary and fixed across all agents, it follows that

$$R_{s,k} \;=\; \mathbb{E}\,\nabla_{w^{\mathsf{T}}} Q(w^o;\boldsymbol{x}_{k,i})\,[\nabla_{w^{\mathsf{T}}} Q(w^o;\boldsymbol{x}_{k,i})]^{\mathsf{T}} \;\equiv\; R_s, \qquad k=1,2,\ldots,N \tag{11.201}$$

Substituting into (11.186), and using $h=1$ for real data, we conclude that the excess-risk of the diffusion solution (and of consensus as well) is given by

$$\mathrm{ER}_{\mathrm{dist,av}} \;=\; \frac{1}{4}\left(\sum_{k=1}^{N}\mu_k p_k\right)^{-1}\left(\sum_{k=1}^{N}\mu_k^2 p_k^2\right)\mathrm{Tr}\,(R_s) \tag{11.202}$$

Example #11.9

If we assume uniform step-sizes, $\mu_k \equiv \mu$ for $k = 1, 2, \ldots, N$, and use the fact that the $\{p_k\}$ add up to one, then expression (11.202) reduces to

$$\mathrm{ER}_{\mathrm{dist,av}} = \frac{\mu}{4}\left(\sum_{k=1}^{N} p_k^2\right)\mathrm{Tr}\,(R_s) \tag{11.203}$$

For comparison purposes, we reproduce below ER expression (5.98) for the centralized solution from Example 5.3:

$$\mathrm{ER}_{\mathrm{cent}} = \frac{\mu}{4}\left(\frac{1}{N}\right)\mathrm{Tr}(R_s) \tag{11.204}$$

Example #11.9

For doubly-stochastic combination matrices A, it holds that $p_k = 1/N$ so that (11.203) reduces to (11.204).

We illustrate these results numerically for the logistic risk function (7.24) from Example 7.4, namely,

$$J(w) \triangleq \frac{\rho}{2}\|w\|^2 + \mathbb{E}\left\{\ln\left(1+e^{-\boldsymbol{\gamma}_k(i)\boldsymbol{h}_{k,i}^{\mathsf{T}}w}\right)\right\} \tag{11.205}$$

Figure 11.5 shows the connected network topology with $N = 20$ agents used for this simulation. All agents are assumed to employ the same step-size parameter, i.e., $\mu_k \equiv \mu$, and they have non-trivial self-loops so that the neighborhood of each agent includes the agent itself. The resulting network is therefore strongly-connected.

Example #11.9

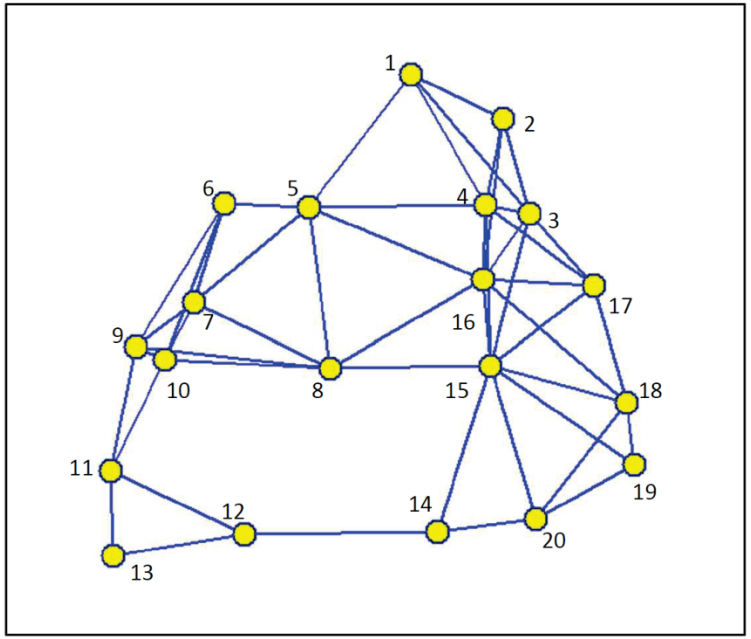


Figure 11.5: A connected network topology consisting of $N = 20$ agents employing the Metropolis rule (8.100). Each agent k is assumed to belong its neighborhood $\mathcal{N}_k$.

Example #11.9

$$\begin{cases} \boldsymbol{\psi}_{k,i-1} &= \displaystyle\sum_{\ell\in\mathcal{N}_k} a_{\ell k}\ \boldsymbol{w}_{\ell,i} \quad (\underline{\text{consensus}}) \\ \boldsymbol{w}_{k,i} &= (1-\rho\mu)\boldsymbol{\psi}_{k,i-1} + \mu\boldsymbol{\gamma}_k(i)\boldsymbol{h}_{k,i}\left(\dfrac{1}{1+e^{\boldsymbol{\gamma}_k(i)\boldsymbol{h}_{k,i}^{\mathsf{T}}\boldsymbol{w}_{k,i-1}}}\right) \end{cases} \tag{11.206}$$

and

$$\begin{cases} \boldsymbol{\psi}_{k,i-1} &= \displaystyle\sum_{\ell\in\mathcal{N}_k} a_{\ell k}\ \boldsymbol{w}_{\ell,i} \quad (\underline{\text{CTA diffusion}}) \\ \boldsymbol{w}_{k,i} &= (1-\rho\mu)\boldsymbol{\psi}_{k,i-1} + \mu\boldsymbol{\gamma}_k(i)\boldsymbol{h}_{k,i}\left(\dfrac{1}{1+e^{\boldsymbol{\gamma}_k(i)\boldsymbol{h}_{k,i}^{\mathsf{T}}\boldsymbol{\psi}_{k,i-1}}}\right) \end{cases} \tag{11.207}$$

Example #11.9

$$\left\{\begin{array}{rcl} \boldsymbol{\psi}_{k,i} & = & (1-\rho\mu)\boldsymbol{w}_{k,i-1} + \mu\boldsymbol{\gamma}_k(i)\boldsymbol{h}_{k,i}\left(\dfrac{1}{1+e^{\boldsymbol{\gamma}_k(i)\boldsymbol{h}_{k,i}^{\mathsf{T}}\boldsymbol{w}_{k,i-1}}}\right) \\ \boldsymbol{w}_{k,i} & = & \displaystyle\sum_{\ell\in\mathcal{N}_k} a_{\ell k}\,\boldsymbol{\psi}_{\ell,i} \qquad (\underline{\text{ATC diffusion}}) \end{array}\right. \tag{11.208}$$

where the combination weights $\{a_{\ell k}\}$ arise from the Metropolis rule (8.100). This rule leads to a doubly-stochastic matrix, A, so that the entries of the Perron eigenvector are given by $p_k = 1/N$. In this way, the ER performance level (11.203) for the above distributed strategies reduces to

$$\mathrm{ER}_{\mathrm{dist,av}} = \frac{\mu}{4}\left(\frac{1}{N}\right)\mathrm{Tr}(R_s) \tag{11.209}$$

Example #11.9

Figures 11.6 and 11.7 plot the evolution of the ensemble-average learning curves, $\mathbb{E}\,\{J(\boldsymbol{w}_{i-1}) - J(w^o)\}$, for consensus, ATC diffusion, and CTA diffusion for two choices of the step-size parameter: a smaller value at $\mu = 1 \times 10^{-4}$ and a second value that is three times larger at $\mu = 3 \times 10^{-4}$. The curves are obtained by averaging the trajectories $\{J(\boldsymbol{w}_{i-1}) - J(w^o)\}$ over 100 repeated experiments. The labels on the vertical axes in the figures refer to the learning curves by writing $\mathrm{ER}_{\mathrm{dist,av}}(i)$, with an iteration index i. Each experiment involves running the consensus (11.206) or diffusion (11.207)–(11.208) logistic recursions with $\rho = 10$ and $h = 1$ for real data $\{\boldsymbol{\gamma}_k(i), \boldsymbol{h}_{k,i}\}$, where the dimension of the feature vectors $\{\boldsymbol{h}_{k,i}\}$ is $M = 50$. The data used for the

Example #11.9

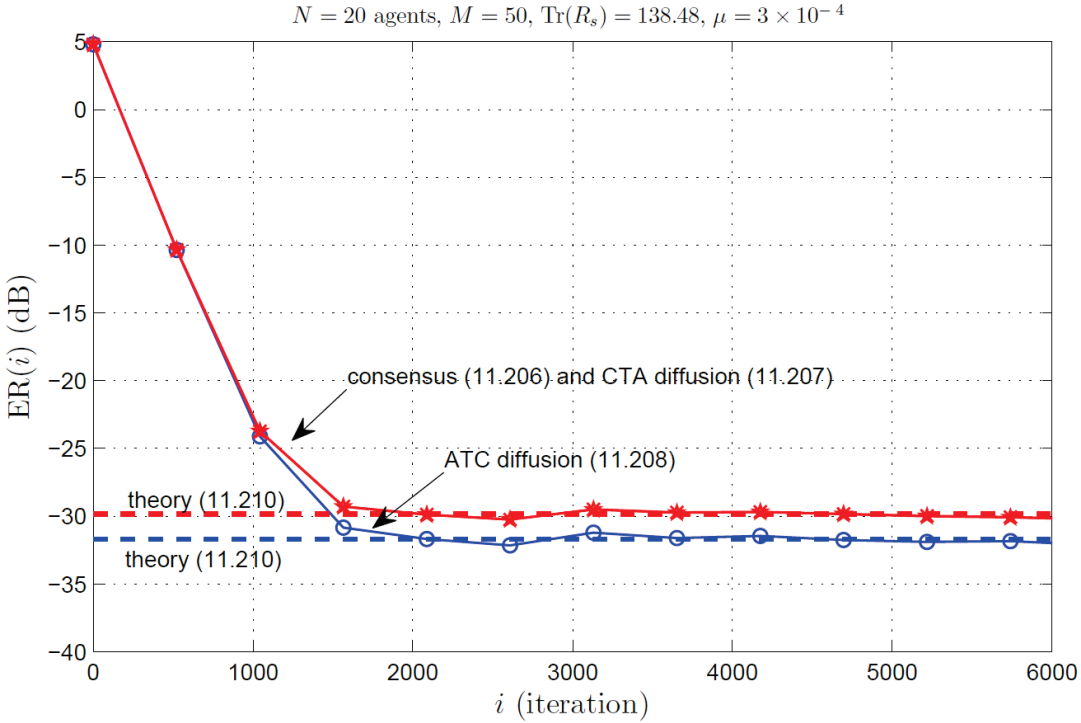


Figure 11.7

Example #11.9

simulation originate from the alpha data set [223]; we use the first 50 features for illustration purposes so that $M = 50$. To generate the trajectories for the experiments in this example, the optimal w^o and the gradient noise covariance matrix, R_s, are first estimated off-line by applying a batch algorithm to all data points. For the data used in this experiment we have $\mathrm{Tr}(R_s) \approx 131.48$.

Example #11.9

It is observed in Figure 11.6 that the learning curves tend towards the ER value predicted by the theoretical expression (11.209), which provides a good approximation for the performance of distributed strategies for small step-sizes. However, it is observed in Figure 11.7 that once the step-size value is increased, differences in ER performance arise among the algorithms, with ATC diffusion exhibiting the lowest (i.e., best) ER value. The horizontal lines in the second figure represent the ER levels that are predicted by the future expression (11.210). This latter expression reflects the effect of higher-order terms in $\mu_{\max}$ and generally leads to an enhanced representation for the mean excess cost, while expression (11.209), which is the basis for the results in this example, is an expression for the ER that is accurate to first-order in $\mu_{\max}$.

Example #11.10

Example 11.10 (Higher-order ER terms). We explained earlier following (11.39) that the ER metric (11.33) assesses the size of the mean fluctuation of the normalized aggregate cost, $\mathbb{E}\left\{\bar{J}^{\text{glob},\star}(\boldsymbol{w}_{k,i-1}) - \bar{J}^{\text{glob},\star}(w^\star)\right\}$, in steady-state and for sufficiently small step-sizes (i.e., in the slow adaptation regime). The computation leads to an expression for the ER that is *first-order* in $\mu_{\max}$, as can be ascertained from (11.186).

Example #11.10

If we revisit the derivation of (11.186) in the proof of Theorem 11.3, we will observe that this expression was obtained by eliminating the contribution of the higher-order term, $O(\mu_{\max}^2)$, which appears in the expansion (11.189). We can motivate an alternative expression for assessing the size of the mean cost fluctuation by retaining the higher-order term that is available (i.e., known) rather than neglecting it. It is expected that, by doing so, the resulting performance expression will generally provide a more accurate representation for the mean cost fluctuation, especially at larger step-sizes; we illustrated this behavior in Figure 11.7.

Example #11.10

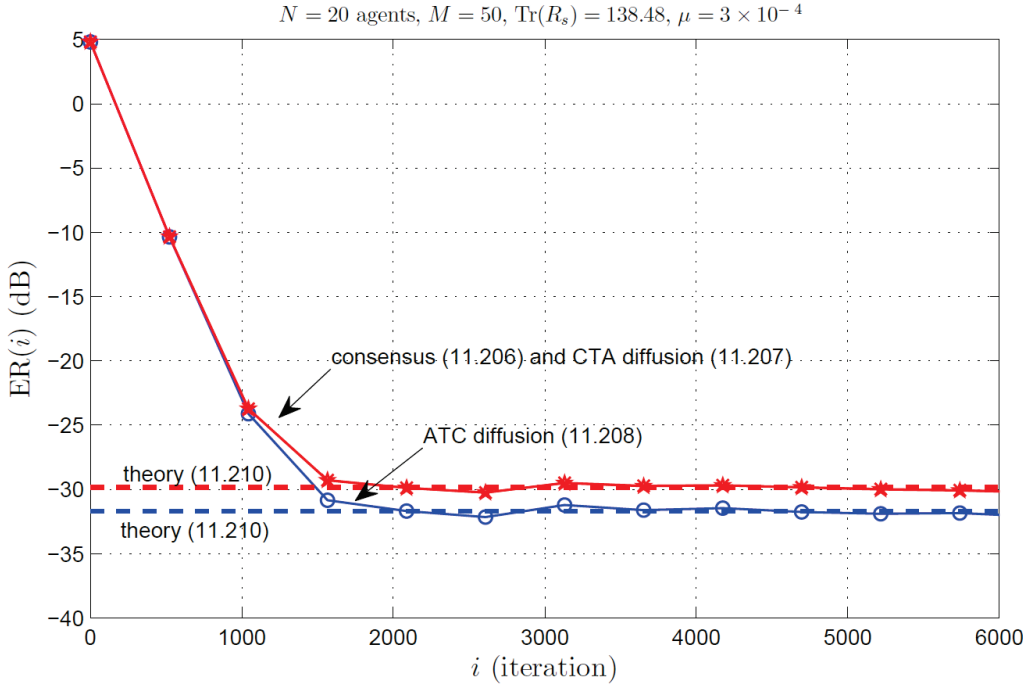


Figure 11.7

Example #11.10

In a manner similar to Example 11.7, we can motivate the following enhanced expression for the excess mean cost, which reflects contributions from higher-order powers of $\mu_{\max}$ as well:

$$\mathrm{ER}_{\mathrm{dist,av}} = \frac{1}{2N}\left(\mathrm{bvec}\left(\mathcal{Y}^{\mathsf{T}}\right)\right)^{\mathsf{T}}(I-\mathcal{F})^{-1}\mathrm{bvec}\left(\bar{\mathcal{H}}\right) \tag{11.210}$$

where we continue to use the notation ER to represent this value. As we already know from the proof of Theorem 11.3, if we expand the right-hand side of (11.210) in terms of powers of $\mu_{\max}$, then the first term in this expansion (i.e., the one that is linear in $\mu_{\max}$) will be given by expression (11.186).

■

Comparing Consensus & Diffusion

Course EE210B
Spring Quarter 2015

Proc. IEEE, vol. 102, no. 4, pp. 460-497, April 2014.
Foundations and Trends in Machine Learning, vol. 7, no. 4-5, pp. 311-801, July 2014.

Comparison

Using results from the previous sections, we can compare some performance properties of diffusion and consensus networks. Recall from (8.7)–(8.10) that the consensus and diffusion strategies correspond to the following choices for $\{A_o, A_1, A_2\}$ in terms of a single combination matrix A in the general description (8.46):

$$\text{consensus:} \quad A_o = A, \quad A_1 = I_N = A_2 \tag{11.211}$$

$$\text{CTA diffusion:} \quad A_1 = A, \quad A_2 = I_N = A_o \tag{11.212}$$

$$\text{ATC diffusion:} \quad A_2 = A, \quad A_1 = I_N = A_o \tag{11.213}$$

Example #11.11

Example 11.11 (Diffusion outperforms consensus over MSE networks). Expression (11.138) indicates that the MSD performance of the consensus and diffusion strategies are identical to first-order in the step-size parameters, as already anticipated by the results in Figures 11.3 and 11.4. We now examine the MSD performance level more closely by considering higher-order terms as well. More specifically, we resort to the alternative expression (11.178).

Example #11.11

The following example is a generalization of a similar discussion from [248]. Let us consider a situation in which all agents in a strongly-connected network employ the same step-size, i.e., $\mu_k \equiv \mu$, and that the diffusion and consensus strategies from (8.46) are implemented with the same combination matrix, A. Without loss in generality, we consider the case of real-valued data. Let us assume further that the Hessian matrices of all individual costs, $J_k(w)$, evaluate to the same value at the reference point $w^\star$, namely,

$$\nabla_w^2 J_k(w^\star) \;\equiv\; H, \quad k = 1, 2, \ldots, N \tag{11.214}$$

Example #11.11

for some constant matrix H. We also assume that the gradient noise variances $\{G_k\}$ approach the same value in steady-state apart from some scaling to account for the possibility of different noise power levels across the agents, i.e., we assume that the $\{G_k\}$ have the form:

$$G_k \equiv \sigma_{v,k}^2 G, \quad k = 1, 2, \ldots, N \tag{11.215}$$

for some constant matrix G. For example, these two conditions on $\{\nabla_w^2 J_k(w^\star),\ G_k\}$ are readily satisfied by the class of MSE networks defined earlier in Example 6.3 when the regression covariance matrices are uniform across all agents, $R_{u,k} \equiv R_u$ for $k = 1, 2, \ldots, N$. Indeed, if we write down an expression similar to (8.15) for the gradient noise process at each agent k, namely,

Example #11.11

$$\boldsymbol{s}_{k,i}(\boldsymbol{\phi}_{k,i-1}) \;=\; 2\left(R_u - \boldsymbol{u}_{k,i}^{\mathsf{T}}\boldsymbol{u}_{k,i}\right)\widetilde{\boldsymbol{\phi}}_{k,i-1} \;-\; 2\boldsymbol{u}_{k,i}^{\mathsf{T}}\boldsymbol{v}_k(i) \tag{11.216}$$

then we conclude that

$$R_{s,k} \;\triangleq\; \lim_{i\to\infty} \mathbb{E}\left[\,\boldsymbol{s}_{k,i}(w^\star)\boldsymbol{s}_{k,i}^{\mathsf{T}}(w^\star)\,|\boldsymbol{\mathcal{F}}_{i-1}\right] \;=\; 4\sigma_{v,k}^2 R_u \tag{11.217}$$

so that, using the definitions (11.12), we obtain for the case of real-data:

$$\nabla_w^2 J_k(w^\star) \;=\; 2R_u \;\equiv\; H, \quad G_k \;=\; 4\sigma_{v,k}^2 R_u \;\equiv\; \sigma_{v,k}^2 G \tag{11.218}$$

with $G = 2H$ in this case.

Example #11.11

We are interested in comparing the MSD performance of diffusion and consensus networks under conditions (11.214)–(11.215). If desired, we can also compare against the performance of the non-cooperative solution. For this latter comparison to be meaningful, we would need to assume that all individual costs, $J_k(w)$, have the same minimizer so that the distributed and the non-cooperative implementations would be seeking the same minimizer. If we were only interested in comparing the consensus and diffusion strategies, then there is no need to assume that the individual costs have the same minimizer; the argument given below would still apply.

Example #11.11

We collect the noise power scalings into an $N \times N$ diagonal matrix

$$R_v = \text{diag}\{\sigma_{v,1}^2,\ \sigma_{v,2}^2, \ldots,\ \sigma_{v,N}^2\} \tag{11.219}$$

Then, it holds from (11.53) and (11.215) that $\mathcal{S}$ can be expressed as the Kronecker product:

$$\mathcal{S} \ = \ R_v \otimes G \tag{11.220}$$

Using the series representation (11.178) we have

$$\text{MSD}_{\text{dist,av}} \ = \ \frac{1}{hN} \sum_{n=0}^{\infty} \text{Tr}\left[\mathcal{B}^n \mathcal{Y} \left(\mathcal{B}^*\right)^n\right] \tag{11.221}$$

$$\mathcal{S} = \text{diag}\{G_1, G_2, \ldots, G_N\}$$

Example #11.11

where $h = 1$ for real data and, from the expressions in Theorem 11.2, the matrices $\mathcal{B}$ and $\mathcal{Y}$ are given by the following relations for the various strategies:

$$\left\{\begin{array}{lcllcl} \mathcal{B}_{\rm ncop} &=& I_N \otimes (I_{hM} - \mu H), & \mathcal{Y}_{\rm ncop} &=& \mu^2 (R_v \otimes G) \\ \mathcal{B}_{\rm cons} &=& A^{\mathsf{T}} \otimes I_{hM} - \mu (I_{hM} \otimes H), & \mathcal{Y}_{\rm cons} &=& \mu^2 (R_v \otimes G) \\ \mathcal{B}_{\rm atc} &=& A^{\mathsf{T}} \otimes (I_{hM} - \mu H), & \mathcal{Y}_{\rm atc} &=& \mu^2 (A^{\mathsf{T}} R_v A \otimes G) \\ \mathcal{B}_{\rm cta} &=& A^{\mathsf{T}} \otimes (I_{hM} - \mu H), & \mathcal{Y}_{\rm cta} &=& \mu^2 (R_v \otimes G) \end{array}\right. \tag{11.222}$$

$$\mathcal{Y} = \mathcal{A}_2^{\mathsf{T}} \mathcal{M} \mathcal{S} \mathcal{M} \mathcal{A}_2$$

Example #11.11

We already know from Example 10.1 that, in general, $\rho(\mathcal{B}_{\text{diff}}) \leq \rho(\mathcal{B}_{\text{ncop}})$ so that diffusion strategies have a stabilizing effect. For the current data structure, it holds that these spectral radii are equal. Indeed, since A is a left-stochastic matrix, its spectral radius is given by $\rho(A) = 1$. Then,

$$\begin{aligned} \rho(\mathcal{B}_{\text{diff}}) &= \rho[A^{\mathsf{T}} \otimes (I_{hM} - \mu H)] \\ &= \rho(A)\,\rho(I_{hM} - \mu H) \\ &= \rho(I_{hM} - \mu H) \\ &= \rho(\mathcal{B}_{\text{ncop}}) \end{aligned} \tag{11.223}$$

On the other hand, let $\lambda_\ell(A)$ denote any of the eigenvalues of A. Since we know that $1 \in \{\lambda_\ell(A)\}$, it then follows:

Example #11.11

$$\begin{aligned}\rho(\mathcal{B}_{\text{ncop}}) &= \max_{1\le m\le 2M} |1-\mu\lambda_m(H)| \\ &\le \max_{1\le \ell\le N}\ \max_{1\le m\le 2M} |\lambda_\ell(A)-\mu\lambda_m(H)| \\ &\overset{(8.40)}{=} \rho(\mathcal{B}_{\text{cons}}) \end{aligned} \tag{11.224}$$

In other words, we arrive at the following conclusion for the scenario under study:

$$\rho(\mathcal{B}_{\text{diff}}) = \rho(\mathcal{B}_{\text{ncop}}) \le \rho(\mathcal{B}_{\text{cons}}) \tag{11.225}$$

It follows from this result that the convergence rate of the diffusion network is generally superior to the convergence rate of the consensus network.

Example #11.11

Not only the convergence rate is superior, but the MSD performance of the diffusion network is also superior. To see this, we first note that for consensus implementations, it is customary to employ a doubly-stochastic matrix A (see Appendix E in [208]). For example, a left-stochastic A that is also symmetric will be doubly-stochastic. For the derivation that follows, we shall therefore assume that A is symmetric, i.e., $A = A^{\mathsf{T}}$; the argument can be extended to matrices A that are "close-to-symmetric" (i.e., diagonalizable with left-eigenvectors $\{x_k\}$ that are practically orthogonal to each other) [248]. It is sufficient for this example to consider the case of symmetric combination policies, A.

Example #11.11

Since A is now diagonalizable, it admits a Jordan canonical decomposition of the form [27, 99, 104, 113]:

$$A^{\mathsf{T}} = XDX^{-1} \tag{11.226}$$

where D is a diagonal matrix with the eigenvalues of A, and X is a similarity transformation. Let $\{x_n\}$ denote the columns of X and let $\{y_n^*\}$ denote the rows of X^{-1}. Then, it follows from (11.226) and the fact that $XX^{-1} = I_N$ that

$$\begin{cases} A^{\mathsf{T}}x_n = \lambda_n(A)x_n \\ y_\ell^* A^{\mathsf{T}} = \lambda_\ell(A)y_\ell^* \\ y_\ell^* x_k = \delta_{\ell k} \\ \ell, k = 1, 2, \ldots, N \end{cases} \tag{11.227}$$

Example #11.11

so that the $\{x_n\}$ correspond to the right eigenvectors of A^{T} and the $\{y_m^*\}$ correspond to the left eigenvectors of A^{T}. We assume the eigenvectors $\{x_n\}$ are normalized to satisfy

$$\|x_n\|^2 = 1, \quad n = 1, 2, \ldots, N \tag{11.228}$$

Since A is symmetric, then X is an orthonormal matrix, i.e.,

$$x_\ell^* x_k = \delta_{\ell k} \tag{11.229}$$

Example #11.11

Under conditions (11.214)–(11.215), and for sufficiently small step-size μ to ensure mean-square stability, we now verify that diffusion networks lead to better MSD performance (i.e., smaller MSD values) than consensus networks. In particular, we verify that the ATC diffusion strategy achieves the lowest network MSD in comparison to the other strategies:

$$\mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{atc}} \leq \mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{cta}} \leq \mathrm{MSD}_{\mathrm{ncop,av}} \tag{11.230}$$

$$\mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{atc}} \leq \mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{cons}} \tag{11.231}$$

Furthermore, if it holds that

$$1 \leq \mu\lambda_{\min}(H) < 2 \tag{11.232}$$

Example #11.11

then we verify that the consensus strategy is the worst even in comparison to the non-cooperative strategy:

$$\mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{atc}} \leq \mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{cta}} \leq \mathrm{MSD}_{\mathrm{ncop,av}} \leq \mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{cons}} \tag{11.233}$$

To see this, we introduce the eigen-decompositions of the matrices A and H into (11.221) and compare the resulting MSD expressions for the various strategies. Let $\{\lambda_m(H) > 0\}$ denote the eigenvalues of the Hermitian and positive-definite matrix H with orthonormal eigenvectors denoted by $\{z_m\}$ $(m = 1, 2, \ldots, hM)$:

$$Hz_m = \lambda_m(H)z_m, \quad m = 1, 2, 3, \ldots, hM \tag{11.234}$$

Substituting the eigen-decompositions of A from (11.227) and H from (11.234) into (11.221) gives, after some algebra:

Example #11.11

$$\mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{atc}} = \frac{\mu^2}{hN} \sum_{k=1}^{N} \sum_{m=1}^{hM} \frac{|\lambda_k(A)|^2 \, \|y_k\|_{R_v}^2 \, \|z_m\|_G^2}{1 - |\lambda_k(A)|^2 \, [1 - \mu\lambda_m(H)]^2} \tag{11.235}$$

$$\mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{cta}} = \frac{\mu^2}{hN} \sum_{k=1}^{N} \sum_{m=1}^{hM} \frac{\|y_k\|_{R_v}^2 \, \|z_m\|_G^2}{1 - |\lambda_k(A)|^2 \, [1 - \mu\lambda_m(H)]^2} \tag{11.236}$$

$$\mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{cons}} = \frac{\mu^2}{hN} \sum_{k=1}^{N} \sum_{m=1}^{hM} \frac{\|y_k\|_{R_v}^2 \, \|z_m\|_G^2}{1 - |\lambda_k(A) - \mu\lambda_m(H)|^2} \tag{11.237}$$

$$\mathrm{MSD}_{\mathrm{ncop,av}} = \frac{\mu^2}{hN} \sum_{k=1}^{N} \sum_{m=1}^{hM} \frac{\|y_k\|_{R_v}^2 \, \|z_m\|_G^2}{1 - (1 - \mu\lambda_m(H))^2} \tag{11.238}$$

Example #11.11

Now note that since $|\lambda_k(A)| \leq 1$, it is obvious that

$$\text{MSD}_{\text{dist,av}}^{\text{atc}} \leq \text{MSD}_{\text{dist,av}}^{\text{cta}} \leq \text{MSD}_{\text{ncop,av}} \tag{11.239}$$

To compare ATC diffusion and consensus, it can be verified that the ratio of each term on the right-hand side of (11.235) to the corresponding term in (11.237) is smaller or equal to one [248]:

$$\frac{|\lambda_k(A)|^2 \left(1 - |\lambda_k(A) - \mu\lambda_m(H)|^2\right)}{1 - |\lambda_k(A)|^2 \left(1 - \mu\lambda_m(H)\right)^2} \leq 1 \tag{11.240}$$

so that

$$\text{MSD}_{\text{dist,av}}^{\text{atc}} \leq \text{MSD}_{\text{dist,av}}^{\text{cons}} \tag{11.241}$$

Example #11.11

We can further verify that the performance of the consensus strategy is worse than the non-cooperative strategy when the step-size satisfies $1 \leq \mu\lambda_{\min}(H) < 2$. This result is established by verifying that the ratio of the individual terms appearing in the sums (11.237)-(11.238) is upper bounded by one [248]:

$$\frac{1-|\lambda_k(A)-\mu\lambda_m(H)|^2}{1-(1-\mu\lambda_m(H))^2} \leq 1 \qquad (11.242)$$

■

Example #11.12

Example 11.12 (MSD performance of consensus and diffusion networks). The following example specializes the results of Example 11.11 to the case of MSE networks from Example 6.3. We reconsider the two-agent network from Example 10.2 with both agents running either the LMS consensus strategy (7.13) or the LMS diffusion strategies (7.22)–(7.23) albeit on real data (for which $h = 1$). We assume

$$\mu_1 = \mu_2 \equiv \mu \tag{11.243}$$

$$R_{u,1} = R_{u,2} \equiv \sigma_u^2 I_{hM} \tag{11.244}$$

$$0 < \mu\sigma_u^2 < 1 \tag{11.245}$$

Example #11.12

The second condition (11.244) ensures that $H = 2\sigma_u^2 I_M$. The third condition (11.245) ensures that both agents are individually stable in the mean since the matrix $\mathcal{B}_{\text{ncop}} = I_N \otimes (I_{hM} - \mu H)$ from Example 11.11 will be stable.

The eigenvalues of A defined by (10.129) are at $\lambda_1(A) = 1$ and $\lambda_2(A) = 1 - a - b$. Using the notation of Example 11.11, this situation corresponds to the case

$$\left\{ \begin{array}{rcl} R_v & = & \text{diag}\{\sigma_{v,1}^2, \sigma_{v,2}^2\} \\ G & = & 4\sigma_u^2 \, I_M \\ H & = & 2\sigma_u^2 \, I_M \end{array} \right. \tag{11.246}$$

In this case, expressions (11.235)–(11.238) reduce to (using $h = 1$ for real data):

Example #11.12

$$\mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{atc}} = 2\mu^2\sigma_u^2 M \left[\frac{y_1^* R_v y_1}{1-(1-2\mu\sigma_u^2)^2} + \frac{y_2^* R_v y_2 (1-a-b)^2}{1-(1-a-b)^2(1-2\mu\sigma_u^2)^2} \right] \tag{11.247}$$

$$\mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{cta}} = 2\mu^2\sigma_u^2 M \left[\frac{y_1^* R_v y_1}{1-(1-2\mu\sigma_u^2)^2} + \frac{y_2^* R_v y_2}{1-(1-a-b)^2(1-2\mu\sigma_u^2)^2} \right] \tag{11.248}$$

$$\mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{cons}} = 2\mu^2\sigma_u^2 M \left[\frac{y_1^* R_v y_1}{1-(1-2\mu\sigma_u^2)^2} + \frac{y_2^* R_v y_2}{1-(1-a-b-2\mu\sigma_u^2)^2} \right] \tag{11.249}$$

$$\mathrm{MSD}_{\mathrm{ncop,av}} = 2\mu^2\sigma_u^2 M \left[\frac{y_1^* R_v y_1}{1-(1-2\mu\sigma_u^2)^2} + \frac{y_2^* R_v y_2}{1-(1-2\mu\sigma_u^2)^2} \right] \tag{11.250}$$

Example #11.12

Note that the first terms inside the brackets of (11.247)-(11.250) are the same. Then, it can be verified that these MSD values are related as follows depending on the region in space where the parameters (a, b) lie:

$$\begin{cases} \text{MSD}_{\text{dist,av}}^{\text{cons}} \leq \text{MSD}_{\text{dist,av}}^{\text{cta}}, & \text{if} \quad 0 \leq a+b \leq \frac{1-2\mu\sigma_u^2}{1-\mu\sigma_u^2} \\ \text{MSD}_{\text{dist,av}}^{\text{cons}} \geq \text{MSD}_{\text{dist,av}}^{\text{cta}}, & \text{if} \quad \frac{1-2\mu\sigma_u^2}{1-\mu\sigma_u^2} \leq a+b < 2(1-\mu\sigma_u^2) \\ \text{MSD}_{\text{dist,av}}^{\text{cons}} \leq \text{MSD}_{\text{ncop, av}}, & \text{if} \quad 0 \leq a+b \leq 2(1-2\mu\sigma_u^2) \\ \text{MSD}_{\text{dist,av}}^{\text{cons}} \geq \text{MSD}_{\text{ncop, av}}, & \text{if} \quad 2(1-2\mu\sigma_u^2) \leq a+b < 2(1-\mu\sigma_u^2) \end{cases} \tag{11.251}$$

Example #11.12

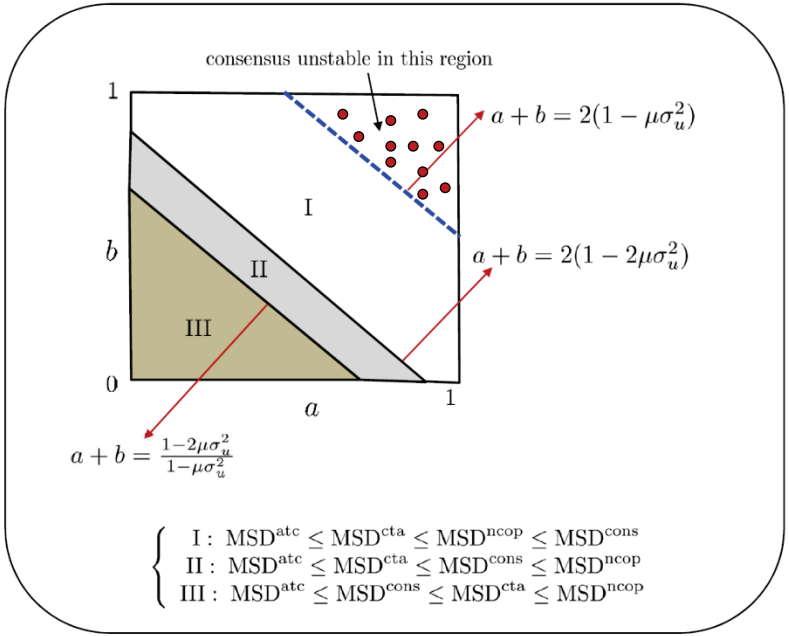


Figure 11.8: Comparison of the network MSD for $N = 2$ agents operating on complex-valued data. The consensus strategy is unstable when a and b lie above the dashed line in region I; it performs well in region III. ATC diffusion is superior in all three regions.

Example #11.12

For example, the first relation can be established as follows:

$$\begin{aligned}
&\mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{cons}} \;\leq\; \mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{cta}} \\
&\Leftrightarrow \quad (1-a-b-2\mu\sigma_u^2)^2 \leq (1-a-b)^2(1-2\mu\sigma_u^2)^2 \\
&\Leftrightarrow \quad (a+b)^2 - 2(a+b)(1-2\mu\sigma_u^2) \leq [-2(a+b)+(a+b)^2](1-2\mu\sigma_u^2)^2 \\
&\Leftrightarrow \quad (a+b)^2\left[1-(1-2\mu\sigma_u^2)^2\right] - 2(a+b)(1-2\mu\sigma_u^2)[1-(1-2\mu\sigma_u^2)] \leq 0 \\
&\Leftrightarrow \quad 0 \leq (a+b) \leq \frac{4(1-2\mu\sigma_u^2)\mu\sigma_u^2}{1-(1-2\mu\sigma_u^2)^2} \\
&\Leftrightarrow \quad 0 \leq (a+b) \leq \frac{1-2\mu\sigma_u^2}{1-\mu\sigma_u^2} \qquad (11.252)
\end{aligned}$$

Example #11.12

and similarly for the other inequalities. We can therefore divide the $a \times b$ plane into three regions I, II, and III, as shown in Figure 11.8, where each region represents one possible relation among the MSD levels of the various strategies. The ATC diffusion strategy is seen to be superior in all regions, while the consensus strategy is worse than the non-cooperative strategy in region I and is also unstable in the mean for values of (a, b) lying above the dashed line in that region, i.e., for $a + b > 2(1 - \mu\sigma_u^2)$, as can be verified by following an argument similar to (10.135).

■

Example #11.13

Example 11.13 (Higher-order terms in the MSD expression). Continuing with Example 11.12, we can rework expression (11.247) for $\text{MSD}_{\text{dist,av}}^{\text{atc}}$ into a more familiar form (and similarly for the other expressions). Thus, consider the eigenvectors $\{x_n, y_m\}$ defined by (11.227). Since A is left-stochastic, we have $A^{\mathsf{T}}\mathbb{1} = \mathbb{1}$. Note, however, from the definition of the eigenvectors $\{x_n\}$ that they need to satisfy the normalization condition (11.228). This means that we can select the first eigenvector as

$$x_1 = \frac{1}{\sqrt{N}}\mathbb{1} \tag{11.253}$$

Example #11.13

It then follows from the condition $y_1^* x_1 = 1$ that

$$y_1^* \mathbb{1} = \sqrt{N} \tag{11.254}$$

so that the entries of the right-eigenvector y_1 add up to $\sqrt{N}$. Now recall from definition (11.136) for the Perron eigenvector p that its entries must add up to one. Both p and y_1 are right-eigenvectors for A associated with the eigenvalue at one. Therefore, p and y_1 are related as follows:

$$p = \frac{1}{\sqrt{N}} y_1 \tag{11.255}$$

Example #11.13

Using this result, and the fact that μ is sufficiently small and that we are dealing with a two-agent network in this example (so that $N = 2$), we can rewrite (11.247) to first-order in μ as follows:

Example #11.13

$$
\begin{aligned}
\mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{atc}} &= 2\mu^2\sigma_u^2 M\,\frac{y_1^* R_v y_1}{4\mu\sigma_u^2 - 4\mu^2\sigma_u^4} \\
&= 2\mu M\,\frac{N p^* R_v p}{4 - 4\mu\sigma_u^2} \\
&\approx \frac{\mu M}{2}\,2\left(\sum_{k=1}^{2} p_k^2\sigma_{v,k}^2\right), \quad \text{since } N = 2 \text{ and } \mu \text{ is small} \\
&= \mu M \sum_{k=1}^{2} p_k^2\sigma_{v,k}^2 \qquad (11.256)
\end{aligned}
$$

and we recover the analogue of expression (11.144) for real-data.

■

End of Lecture

Proc. IEEE, vol. 102, no. 4, pp. 460-497, April 2014.
Foundations and Trends in Machine Learning, vol. 7, no. 4-5, pp. 311-801, July 2014.