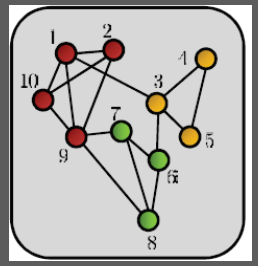

INFERENCE OVER NETWORKS

LECTURE #22: Benefits of Cooperation

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 12 (Benefits of Cooperation, pp. 623-645):

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: 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.

Recall#3 (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 $k = 1, 2, \ldots, N$.

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

Recall#3 (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

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

MSE networks

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

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

Individual Agent Performance

However, it does not generally hold that the distributed solution outperforms each individual non-cooperative agent [276]. This is because the average noise power is scaled by $1/N$, and this scaled power can be larger than some of the individual noise variances and smaller than the remaining noise variances. For example, consider a situation with $N = 2$ agents, $\sigma_{v,2}^2 = 5\sigma_v^2$ and $\sigma_{v,1}^2 = \sigma_v^2$. Then,

$$\frac{1}{N}\left(\frac{1}{N}\sum_{k=1}^{N}\sigma_{v,k}^2\right) = 1.5\sigma_v^2$$

which is larger than $\sigma_{v,1}^2$ and smaller than $\sigma_{v,2}^2$.

Setting

Example 11.5 focused on MSE networks with quadratic costs and showed that for adaptation and learning under doubly-stochastic combination policies, it is not necessarily the case that every agent will benefit from cooperation with its neighbors. Some agents can see their performance degraded relative to what they would have attained had they operated independently of the other agents and in a non-cooperative manner. We verify in this chapter that the same conclusion holds for more general costs: doubly-stochastic combination policies enhance the average network performance albeit at the possible expense of some individual agents having their performance degrade relative to the non-cooperative scenario.

Can we do better?

One useful question to consider is whether it is possible to select combination matrices, A, that ensure that distributed (consensus or diffusion) networks will outperform the non-cooperative strategy both in terms of the overall average performance *and* the individual agent performance. The choice of A will generally need to be left-stochastic. We again recall that in order to carry a meaningful comparison with non-cooperative implementations, it is necessary to assume that all individual costs, $J_k(w)$, share the *same* global minimizer so that $w^\star = w^o$. It is also necessary to assume uniform step-sizes across all agents since the performance of the non-cooperative agents is influenced by the step-sizes.

Setting

Similarly, a meaningful comparison between distributed and centralized implementations requires that they employ the same step-size parameter and that both implementations approach the same limit point and, therefore, we also need to have $w^\star = w^o$. For these reasons, we shall assume in the sequel that

$$\mu_k \equiv \mu, \quad k = 1, 2, \ldots, N \tag{12.1}$$

For ease of reference we recall the expressions for the MSD performance of distributed (consensus and diffusion), centralized, and non-cooperative strategies for sufficiently small step-sizes, for both individual agents (when applicable) and for the average network performance:

Meaningful Comparison

For a **meaningful comparison** of non-cooperative, distributed, and centralized strategies, we assume:

- All costs $J_k(w)$ are strongly-convex.
- All costs share the same minimizer, so that $w^\star = w^o$.
- Uniform step-sizes $\mu_k \equiv \mu$.

MSD Performance Expressions

$$\theta_k^2 \triangleq \mathrm{Tr}(H_k^{-1} G_k)$$

$$\mathrm{MSD}_{\mathrm{cent}} = \frac{\mu}{2Nh}\,\mathrm{Tr}\left[\left(\sum_{k=1}^{N} H_k\right)^{-1}\left(\sum_{k=1}^{N} G_k\right)\right] \tag{12.2}$$

$$\mathrm{MSD}_{\mathrm{ncop},k} = \frac{\mu}{2h}\,\mathrm{Tr}\left(H_k^{-1} G_k\right) \tag{12.3}$$

$$\mathrm{MSD}_{\mathrm{ncop,av}} = \frac{\mu}{2Nh}\,\mathrm{Tr}\left[\sum_{k=1}^{N} H_k^{-1} G_k\right] \tag{12.4}$$

$$\mathrm{MSD}_{\mathrm{dist},k} = \mathrm{MSD}_{\mathrm{dist,av}} = \frac{\mu}{2h}\mathrm{Tr}\left[\left(\sum_{k=1}^{N} p_k H_k\right)^{-1}\left(\sum_{k=1}^{N} p_k^2 G_k\right)\right] \tag{12.5}$$

Setting

In the analysis that follows, we assume that the various strategies are employing the *same* construction for their gradient vectors and that the moment matrices $\{G_k\}$ can be taken to be the same in all implementations. The matrices $\{H_k, G_k\}$ are defined by (11.12) in terms of the Hessian matrices of the individual costs, evaluated at $w^\star = w^o$, and in terms of the second-order moments of the gradient noise processes across the agents.

Doubly-Stochastic Policies

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.

Doubly-Stochastic Policies

Consider first the case in which the combination matrix, A, used by the consensus strategy (7.9) and the diffusion strategies (7.18) and (7.19) is doubly stochastic. Then, the Perron eigenvector p defined by (11.136) is given by $p = \mathbb{1}/N$ so that all its entries are equal to $1/N$. In this case, expressions (12.2) and (12.5) lead to the conclusion that:

$$\mathrm{MSD}_{\mathrm{dist},k} \;=\; \mathrm{MSD}_{\mathrm{dist,av}} \;=\; \mathrm{MSD}_{\mathrm{cent}} \tag{12.6}$$

Doubly-Stochastic Policies

That is, the distributed consensus and diffusion strategies are able to attain the same MSD performance level as the centralized solution. Since we already showed in (5.80) that the centralized solution outperforms the non-cooperative solution, we conclude that the distributed solutions also outperform the non-cooperative solution:

$$\mathrm{MSD}_{\mathrm{dist,av}} \;=\; \mathrm{MSD}_{\mathrm{cent}} \;\leq\; \mathrm{MSD}_{\mathrm{ncop,av}} \qquad (12.7)$$

Doubly-Stochastic Policies

Result (12.7) is in terms of the average network performance (obtained by averaging the MSD levels of the individual agents). In this way, the result establishes that the average MSD performance of the distributed solution is superior (i.e., lower) than the average MSD performance attained by the agents in a non-cooperative implementation. This conclusion motivates the following inquiry: is the improvement in network performance attained at the expense of deterioration in the performance of some of the agents?

Doubly-Stochastic Policies

In other words, will the performance of some agents in the distributed solution become worse than what it would be if they operate independently? If this is the case, then result (12.7) would mean that in moving from non-cooperation to cooperation, some agents see their performance improve while other agents see their performance degrade in such a manner that the net effect for the network is a better (i.e., lower) average MSD value. We now verify that this is indeed the case for doubly-stochastic combination policies.

Doubly-Stochastic Policies

From (12.3) and (12.5) we observe that, to first-order in the step-size parameter, the MSD of the individual agents in the distributed implementation will be smaller (and, hence, better) than the MSD of the individual agents in the non-cooperative implementation only when for each $k = 1, 2, \ldots, N$:

$$\frac{1}{N}\,\mathrm{Tr}\left[\left(\sum_{k=1}^{N} H_k\right)^{-1}\left(\sum_{k=1}^{N} G_k\right)\right] \;\leq\; \mathrm{Tr}(H_k^{-1}G_k) \qquad (12.8)$$

Doubly-Stochastic Policies

Unfortunately, this condition may or may not hold as illustrated by the next example. Agents for which the condition is violated would experience deterioration in their MSD level from cooperation. Before presenting the example, though, we mention that there are situations where condition (12.8) holds for all agents, in which case all agents will benefit from cooperation. This happens, for example, when the Hessian matrices, H_k, and the gradient noise covariances, G_k, are uniform across the agents, namely, when

$$H_k \equiv H, \quad G_k \equiv G, \quad k = 1, 2, \ldots, N \tag{12.9}$$

Doubly-Stochastic Policies

The condition also holds when the following two requirements hold for each $k = 1, 2, \ldots, N$:

$$H_k \equiv H \tag{12.10}$$

$$\frac{1}{N}\,\mathrm{Tr}\left[\sum_{k=1}^{N} H^{-1}G_k\right] \leq N\,\mathrm{Tr}(H^{-1}G_k) \tag{12.11}$$

We summarize the main conclusion so far in the following statement. We illustrated this conclusion earlier in Example 11.5.

Doubly-Stochastic Policies

Lemma 12.1 (Doubly-stochastic combination policies). Assume all agents employ the same step-size parameter and that the individual costs are strongly-convex and their minimizers coincide with each other. For doubly stochastic combination matrices it holds that

$$\mathrm{MSD}_{\mathrm{dist,av}} \;=\; \mathrm{MSD}_{\mathrm{cent}} \;\leq\; \mathrm{MSD}_{\mathrm{ncop,av}} \tag{12.12}$$

Benefits of Cooperation

Lemma 12.1: For doubly-stochastic combination policies and small step-sizes, it holds that

$$\mathrm{MSD}_{\mathrm{dist,av}} = \mathrm{MSD}_{\mathrm{cent}} \leq \mathrm{MSD}_{\mathrm{ncop,av}}$$

However, the performance of individual agents need not satisfy the same relation!

Recall#4: Metropolis Rule

is doubly-stochastic:

$$a_{\ell k} = \begin{cases} \dfrac{1}{\max\{n_k, n_\ell\}}, & \ell \in \mathcal{N}_k \backslash \{k\} \\ 1 - \displaystyle\sum_{m \in \mathcal{N}_k \backslash \{k\}} a_{mk}, & \ell = k \end{cases} \tag{8.100}$$

Example #12.1

Example 12.1 (Doubly-stochastic policies over MSE networks). We reconsider the setting of Example 11.4, which deals with MSE networks operating on real-valued data and refer to the strongly-connected network of Figure 11.1 with $N = 20$ agents. We assume uniform step-sizes, $\mu_k \equiv \mu = 6 \times 10^{-4}$, and uniform regression covariance matrices of the form $R_{u,k} = \sigma_u^2 I_M$ where $\sigma_u^2 = 2$. In this setting, we have

$$H_k = 2\sigma_u^2 I_M \equiv H, \qquad G_k = 4\sigma_{v,k}^2 \sigma_u^2 I_M, \quad \theta_k^2 = 2M\sigma_{v,k}^2 \tag{12.13}$$

Example #12.1

We consider two scenarios. In the first case, the agents run the ATC diffusion strategy (7.23) with the Metropolis combination weights (8.100), namely,

$$\left\{\begin{array}{rcl} \boldsymbol{\psi}_{k,i} & = & \boldsymbol{w}_{k,i-1} + 2\mu \boldsymbol{u}_{k,i}^{\mathsf{T}} \left[\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{k,i-1}\right] \\ \boldsymbol{w}_{k,i} & = & \displaystyle\sum_{\ell \in \mathcal{N}_k} a_{\ell k}\, \boldsymbol{\psi}_{\ell,i} \end{array}\right. \tag{12.14}$$

The Metropolis weights result in a doubly-stochastic combination matrix, A, so that $p_k = 1/N$. In the second case, the agents transfer the data to a fusion center running the centralized strategy (5.13), i.e.,

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} + \mu \left(\frac{1}{N} \sum_{k=1}^{N} 2\boldsymbol{u}_{k,i}^{\mathsf{T}} (\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{i-1}) \right) \tag{12.15}$$

Example #12.1

The resulting MSD performance levels are given by expressions (12.2) and (12.5), which in the current setting reduce to (using $h = 1$ for real data):

$$\text{MSD}_{\text{cent}} = \text{MSD}_{\text{dist,av}} = \frac{\mu M}{N}\left(\frac{1}{N}\sum_{k=1}^{N}\sigma_{v,k}^2\right) \tag{12.16}$$

We illustrate these results numerically in Figure 12.1 for the two algorithms listed above running 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 $M = 10$ and where the noise profile is the same one shown earlier in the left plot of Figure 11.2. The unknown vector w^o is generated randomly and its norm is normalized to one. Figure 12.1 plots

Example #12.1

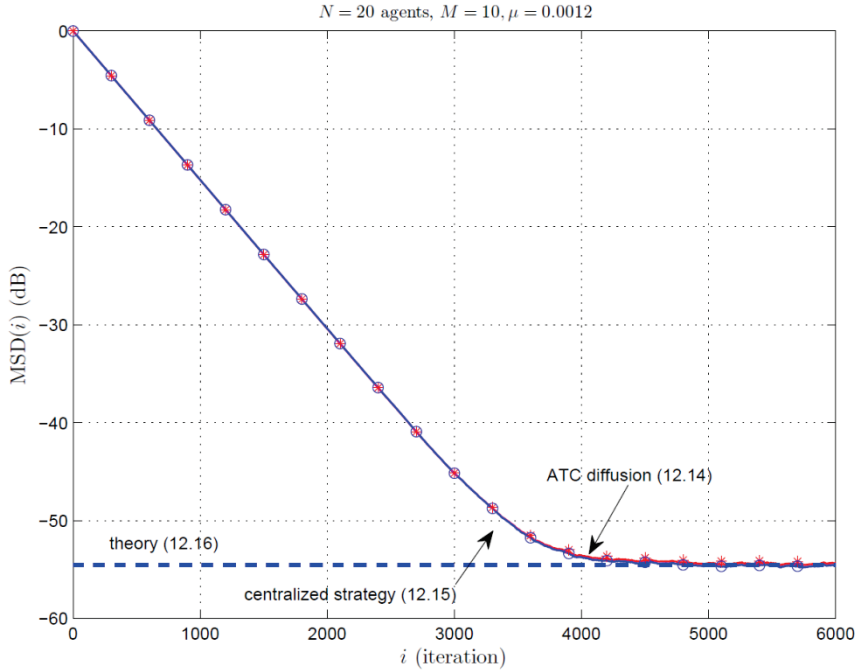

Figure 12.1

Example #12.1

the evolution of the ensemble-average learning curves, $\frac{1}{N}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ for diffusion and $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ for centralized and weighted centralized. The curves are obtained by averaging simulated trajectories over 100 repeated experiments. The label on the vertical axis in the figure refers to the learning curves by writing $\mathrm{MSD}(i)$, with an iteration index i. It is observed both strategies tend towards the same MSD level that is predicted by the theoretical expression (12.16).

■

Left-Stochastic Policies

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.

Left-Stochastic Policies

The previous analysis shows that under doubly-stochastic combination policies, cooperation among the agents enhances the *network* MSD performance albeit possibly at the expense of deterioration in the performance of some *individual* agents. A useful question to consider is whether it is possible to select combination matrices A that will ensure that distributed (consensus or diffusion) networks will outperform the non-cooperative strategy *both* in terms of the overall network performance *and* the individual agent performance. We need to search over the larger set of left-stochastic matrices A since we already know that doubly-stochastic matrices A may not be sufficient to guarantee this property.

Left-Stochastic Policies

From expression (12.3) we observe that the performance of each agent in the non-cooperative mode of operation is dependent on its Hessian matrix, H_k. We therefore focus on the important special case in which these Hessian matrices are uniform across the agents:

$$H_k \equiv H, \quad k = 1, 2, \ldots, N \tag{12.17}$$

$$\mathrm{MSD}_{\mathrm{ncop},k} = \frac{\mu}{2h} \mathrm{Tr}\left(H_k^{-1} G_k\right)$$

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

Left-Stochastic Policies

As explained earlier, this scenario is common in important situations of interest such as the MSE networks of Example 6.3 and in machine learning applications where all agents minimize the same cost function as in Examples 7.4 and 11.9. For a given network topology, we then consider the problem of minimizing the MSD level of the distributed strategies under these conditions, namely,

$$A^o \triangleq \arg\min_{A\in\mathbb{A}} \ \mathrm{Tr}\left(\sum_{k=1}^{N} p_k^2 H^{-1} G_k\right) \qquad (12.18)$$

$$\text{subject to} \ \ Ap = p, \ \ \mathbb{1}^{\mathsf{T}} p = 1, \ \ p_k > 0$$

Left-Stochastic Policies

where the symbol $\mathbb{A}$ denotes the set of all $N \times N$ primitive left-stochastic matrices A whose entries $\{a_{\ell k}\}$ satisfy conditions (7.10). To solve the above problem, we start by introducing the nonnegative scalars:

$$\theta_k^2 \triangleq \mathrm{Tr}(H^{-1}G_k), \quad k = 1, 2, \ldots, N \tag{12.19}$$

and refer to them as gradient-noise factors (since they incorporate information about the gradient noise moments, G_k). Comparing with (12.3), the scalar θ_k^2 is seen to be proportional to the MSD level at agent k in the non-cooperative mode of operation. Interpreting every

Left-Stochastic Policies

$A \in \mathbb{A}$ as the probability transition matrix of an irreducible aperiodic Markov chain [169, 186], and using a construction procedure developed in [42, 106], it was argued in [276] that one choice for an optimal A^o that solves optimization problems of the form (12.18) is the following left-stochastic matrix (which we refer to as the Hastings combination rule).

Hastings Rule

Lemma 12.2 (Hastings rule). The following combination matrix, denoted by A^o with a superscript o, is a solution to the optimization problem (12.18):

$$a_{\ell k}^o = \begin{cases} \dfrac{\theta_k^2}{\max\{\, n_k\theta_k^2,\ n_\ell\theta_\ell^2 \,\}}, & \ell \in \mathcal{N}_k \backslash \{k\} \\ 1 - \displaystyle\sum_{m \in \mathcal{N}_k \backslash \{k\}} a_{mk}^o, & \ell = k \end{cases} \tag{12.20}$$

where $n_k = |\mathcal{N}_k|$ denotes the cardinality of $\mathcal{N}_k$ or the degree of agent k (i.e., number of its neighbors). The entries of the corresponding Perron eigenvector are given by

$$p_k^o = \frac{1}{\theta_k^2}\left(\sum_{\ell=1}^{N}\frac{1}{\theta_\ell^2}\right)^{-1}, \quad k = 1, 2, \ldots, N \tag{12.21}$$

Proof

Proof. We first consider the optimization problem (12.18) without the eigenvector constraint, $Ap = p$, and minimize instead over the positive scalars $\{p_k\}$:

$$p_k^o \triangleq \arg\min_{p_k} \sum_{k=1}^{N} p_k^2 \theta_k^2 \quad \text{subject to } \mathbb{1}^{\mathsf{T}} p = 1, \;\; p_k > 0 \tag{12.22}$$

It is easy to verify that the solution to this problem is given by (12.21). Next, we verify that the matrix A^o defined by (12.20) is a left-stochastic primitive matrix that has $p^o = \text{col}\{p_k^o\}$ as its Perron eigenvector.

$$L(p, \lambda) \;=\; p^{\mathsf{T}} D p + \lambda(\mathbb{1}^{\mathsf{T}} p - 1)$$

Proof

To begin with, it is straightforward to verify from (12.20) that A^o is left-stochastic. We now establish that $A^o p^o = p^o$, i.e., for every $1 \leq \ell \leq N$:

$$\sum_{k=1}^{N} a_{\ell k}^o p_k^o \;=\; p_\ell^o \tag{12.23}$$

For this purpose, we note first that for any $\ell \neq k$, the following balanced relation holds:

Proof

$$
\begin{aligned}
a_{\ell k}^{o}\, p_k^{o} &= \left(\frac{\theta_k^2}{\max\{\, n_k\theta_k^2,\ n_\ell\theta_\ell^2 \,\}}\right) \frac{1}{\theta_k^2} \left(\sum_{\ell=1}^{N} \frac{1}{\theta_\ell^2}\right)^{-1} \\
&= \left(\frac{1}{\max\{\, n_k\theta_k^2,\ n_\ell\theta_\ell^2 \,\}}\right) \left(\sum_{\ell=1}^{N} \frac{1}{\theta_\ell^2}\right)^{-1} \\
&= a_{k\ell}^{o}\, p_\ell^{o} \qquad (12.24)
\end{aligned}
$$

so that

Proof

$$
\begin{aligned}
\sum_{k=1}^{N} a_{\ell k}^{o} p_{k}^{o} &= \sum_{k\neq\ell} a_{\ell k}^{o} p_{k}^{o} + a_{\ell\ell} p_{\ell}^{o} \\
&\overset{(12.24)}{=} \sum_{k\neq\ell} a_{k\ell}^{o} p_{\ell}^{o} + a_{\ell\ell} p_{\ell}^{o} \\
&= \sum_{k=1} a_{k\ell}^{o} p_{\ell}^{o} \\
&= \left(\sum_{k=1} a_{k\ell}^{o}\right) p_{\ell}^{o} \\
&= p_{\ell}^{o} \quad (\text{since } A^{o} \text{ is left-stochastic}) \qquad (12.25)
\end{aligned}
$$

Proof

It remains to show that A^o is primitive. To do so, and in view of Lemma 6.1, it is sufficient to show that $a^o_{kk} > 0$ for some k. This property actually holds for all diagonal entries a^o_{kk} in this case. Indeed, note that since

$$a^o_{\ell k} = \frac{\theta_k^2}{\max\{\, n_k\theta_k^2,\ n_\ell\theta_\ell^2 \,\}} \leq \frac{\theta_k^2}{n_k\theta_k^2} \leq \frac{1}{n_k} \qquad (12.26)$$

we get

Proof

$$\begin{aligned}\sum_{k\neq\ell} a_{\ell k}^{o} &= \sum_{\ell\in\mathcal{N}_k\backslash\{k\}} a_{\ell k}^{o} \\ &\leq \sum_{\ell\in\mathcal{N}_k\backslash\{k\}} \frac{1}{n_k} \\ &= \frac{n_k-1}{n_k} \end{aligned} \tag{12.27}$$

Proof

which implies that

$$\begin{aligned} a_{kk}^o &= 1 - \sum_{\ell \in \mathcal{N}_k \backslash \{k\}} a_{\ell k}^o \\ &\geq 1 - \frac{n_k - 1}{n_k} \\ &= \frac{1}{n_k} \\ &> 0 \end{aligned} \tag{12.28}$$

□

Hastings Rule

The Hastings rule is a fully-distributed solution — each agent k only needs to obtain the products $\{n_\ell \theta_\ell^2\}$ from its neighbors to compute the combination weights $\{a_{\ell k}^o\}$. Substituting (12.21) into (12.18), we find that the resulting optimal value for the distributed network MSD is:

$$\text{MSD}_{\text{dist,av}}^o = \frac{\mu}{2h} \left(\sum_{\ell=1}^{N} \frac{1}{\theta_\ell^2} \right)^{-1} \tag{12.29}$$

Hastings Rule

At the same time, it follows from (12.5) that the MSD performance of the distributed network for any doubly-stochastic (d.s.) matrix A is:

$$\mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{d.s.}} = \frac{\mu}{2N^2h}\left(\sum_{\ell=1}^{N}\theta_\ell^2\right) \tag{12.30}$$

Now, using the following algebraic property [206], which is valid for any scalars $\{\theta_\ell^2\}$:

$$N^2 \le \left(\sum_{\ell=1}^{N}\theta_\ell^2\right)\left(\sum_{\ell=1}^{N}\frac{1}{\theta_\ell^2}\right) \tag{12.31}$$

Hastings Rule

we conclude that

$$\mathrm{MSD}^{o}_{\mathrm{dist,av}} \;\leq\; \mathrm{MSD}^{\mathrm{d.s.}}_{\mathrm{dist,av}} \;\leq\; \mathrm{MSD}_{\mathrm{ncop,av}} \qquad (12.32)$$

so that, as expected, the MSD of the distributed (consensus or distributed) network with the optimal left-stochastic matrix, A^o, is also superior to the MSD of the non-cooperative network. More importantly, though, this optimal choice for A leads to the following performance level at the individual agents in the distributed solution:

Hastings Rule

$$\begin{aligned}
\mathrm{MSD}^{o}_{\mathrm{dist},k} &= \frac{\mu}{2h}\left(\sum_{\ell=1}^{N}\frac{1}{\theta_\ell^2}\right)^{-1} \\
&\leq \frac{\mu}{2h}\left(\frac{1}{\theta_k^2}\right)^{-1} \\
&\overset{(12.3)}{=} \mathrm{MSD}_{\mathrm{ncop},k}, \quad k=1,2,\ldots,N \qquad (12.33)
\end{aligned}$$

Hastings Rule

so that, to first-order in the step-size parameter, the individual agent performance in the optimized distributed network is improved across all agents relative to the non-cooperative case:

$$\mathrm{MSD}^{o}_{\mathrm{dist},k} \;\leq\; \mathrm{MSD}_{\mathrm{ncop},k}, \quad k = 1, 2, \ldots, N \tag{12.34}$$

We summarize the main conclusion in the following statement.

Left-Stochastic Policies

Lemma 12.3 (Left-stochastic combination policies). Assume all agents employ the same step-size parameter and that the individual costs are strongly-convex and their minimizers coincide with each other. Assume further that the Hessian matrices evaluated at the optimal solution, w^o, are uniform across all agents as in (12.17). For the left-stochastic Hastings policy (12.20) it holds that

$$\mathrm{MSD}^{o}_{\mathrm{dist,av}} \leq \mathrm{MSD}^{\mathrm{d.s.}}_{\mathrm{dist,av}} \leq \mathrm{MSD}_{\mathrm{ncop,av}} \tag{12.35}$$

$$\mathrm{MSD}^{o}_{\mathrm{dist},k} \leq \mathrm{MSD}_{\mathrm{ncop},k}, \quad k = 1, 2, \ldots, N \tag{12.36}$$

Example #12.2

Example 12.2 (Optimal combination policy for MSE networks). Let us reconsider the setting of Example 11.3, which deals with MSE networks. We assume uniform step-sizes and uniform regression covariances, i.e., $\mu_k \equiv \mu$ and $R_{u,k} \equiv R_u$ for $k = 1, 2, \ldots, N$. In this setting we have

$$H_k = \begin{bmatrix} R_u & 0 \\ 0 & R_u^{\mathsf{T}} \end{bmatrix} \equiv H, \qquad G_k = \sigma_{v,k}^2 \begin{bmatrix} R_u & \times \\ \times & R_u^{\mathsf{T}} \end{bmatrix}, \quad \theta_k^2 = 2M\sigma_{v,k}^2 \tag{12.37}$$

For these values of $\{H_k, G_k\}$, the optimization problem (12.18) reduces to

Example #12.2

$$A^o \triangleq \arg\min_{A\in\mathbb{A}} \sum_{k=1}^{N} p_k^2 \sigma_{v,k}^2 \qquad (12.38)$$

$$\text{subject to } Ap = p, \quad \mathbb{1}^{\mathsf{T}} p = 1, \quad p_k > 0$$

which is of course the same problem we would be motivated to optimize had we started from the MSD expression (11.147). Using (12.20), an optimal solution is given by

Example #12.2

$$a_{\ell k}^{o} = \begin{cases} \dfrac{\sigma_{v,k}^2}{\max\{\, n_k\sigma_{v,k}^2,\ n_\ell\sigma_{v,\ell}^2 \,\}}, & \ell \in \mathcal{N}_k \backslash \{k\} \\ 1 - \displaystyle\sum_{m \in \mathcal{N}_k \backslash \{k\}} a_{mk}^{o}, & \ell = k \end{cases} \tag{12.39}$$

with

$$\mathrm{MSD}_{\mathrm{dist},k}^{o} = \mathrm{MSD}_{\mathrm{dist,av}}^{o} = \frac{\mu M}{2} \left(\sum_{\ell=1}^{N} \frac{1}{\sigma_{v,\ell}^2} \right)^{-1} \tag{12.40}$$

Example #12.2

Note that

$$\mathrm{MSD}^{o}_{\mathrm{dist},k} \leq \frac{\mu M}{2}\left(\frac{1}{\sigma^2_{v,k}}\right)^{-1} \overset{(12.3)}{=} \mathrm{MSD}_{\mathrm{ncop},k} \qquad (12.41)$$

so that the individual agent performance in the optimized distributed network is improved across all agents relative to the non-cooperative case.

■

Example #12.3

Example 12.3 (Optimal MSD combination policy for online learning). We revisit Example 11.9, which deals with a collection of N learners. Using the notation of that example we have that, in this case, the gradient-noise factors $\{\theta_k^2\}$ are now uniform:

$$\theta_k^2 \equiv \theta^2 \;=\; \mathrm{Tr}(H^{-1}R_s) \tag{12.42}$$

Substituting into expression (12.20) for Hastings rule, we find that the optimal combination coefficients reduce to the following so-called Metropolis rule, which we encountered earlier in Example 8.9:

Example #12.3

$$a_{\ell k}^{o} = \begin{cases} \dfrac{1}{\max\{\, n_k,\ n_\ell \,\}}, & \ell \in \mathcal{N}_k \backslash \{k\} \\ 1 - \displaystyle\sum_{m \in \mathcal{N}_k \backslash \{k\}} a_{mk}^{o}, & \ell = k \end{cases} \tag{12.43}$$

Therefore, the optimal combination policy happens to be doubly-stochastic in this case. Observe that the above combination coefficients now depend solely on the degrees of the agents (i.e., the extent of their connectivity). Moreover, from (12.29) and using $h = 1$ for real data, the optimal MSD value is given by

Example #12.3

$$\mathrm{MSD}^{o}_{\mathrm{dist,av}} = \frac{\mu}{4}\left(\frac{1}{N}\right)\mathrm{Tr}(H^{-1}R_s) \tag{12.44}$$

which, in this case, agrees with the performance of the centralized solution given by (12.2). On the other hand, for arbitrary left-stochastic combination matrices A, the MSD performance of the distributed (consensus and diffusion) solutions can be deduced from (12.5) and would be given by

$$\mathrm{MSD}_{\mathrm{dist,av}} = \frac{\mu}{4}\left(\sum_{k=1}^{N} p_k^2\right)\mathrm{Tr}(H^{-1}R_s) \tag{12.45}$$

■

Comparison with Centralized Solutions

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

The third question we consider in this chapter is to compare the optimal MSD performance of the distributed consensus and diffusion solutions (resulting from the use of the Hastings rule (12.20)), with the MSD performance of the centralized solution under the same condition (12.17) of uniform Hessian matrices. In this case, from expressions (12.2) and (12.29), the MSD levels of the centralized and (optimized) distributed solutions are given by:

Comparison

$$\mathrm{MSD}_{\mathrm{cent}} = \frac{\mu}{2N^2h}\left(\sum_{\ell=1}^{N}\theta_\ell^2\right) \tag{12.46}$$

$$\mathrm{MSD}^{o}_{\mathrm{dist,av}} = \frac{\mu}{2h}\left(\sum_{\ell=1}^{N}\frac{1}{\theta_\ell^2}\right)^{-1} \tag{12.47}$$

Using the inequality (12.31) again, we readily conclude that, to first-order in the step-size parameter,

$$\mathrm{MSD}^{o}_{\mathrm{dist,av}} \leq \mathrm{MSD}_{\mathrm{cent}} \tag{12.48}$$

Comparison

so that the optimized distributed network running the consensus strategy (7.9) or the diffusion strategies (7.18) or (7.19) with the Hasting combination rule (12.20) *outperforms* the centralized solution (5.22), which we repeat below for ease of reference

$$\boldsymbol{w}_i \;=\; \boldsymbol{w}_{i-1} \;-\; \mu\left(\frac{1}{N}\sum_{k=1}^{N}\widehat{\nabla_{w^*}J_k}(\boldsymbol{w}_{i-1})\right), \quad i\geq 0 \qquad (12.49)$$

Comparison

The conclusion that the distributed solutions outperform the centralized solution may seem puzzling at first. However, this result follows from the fact that the optimized combination coefficients (12.20) for the distributed implementations exploit information about the gradient noise factors, $\{\theta_\ell^2\}$. This information is not used by the centralized algorithm (12.49). We can of course modify (12.49) to include information about the gradient noise factors as well.

Weighted Centralized Strategy

One way to modify the centralized solution (12.49) is as follows [279]. We incorporate the positive weighting coefficients $\{p_k^o\}$ into the centralized update equation:

$$\boldsymbol{w}_i \;=\; \boldsymbol{w}_{i-1} \;-\; \mu \left(\sum_{k=1}^{N} p_k^o \, \widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{i-1}) \right), \quad i \geq 0 \qquad (12.50)$$

where the p_k^o were defined earlier in (12.21):

$$p_k^o \;\triangleq\; \frac{1}{\theta_k^2} \left(\sum_{\ell=1}^{N} \frac{1}{\theta_\ell^2} \right)^{-1}, \quad k = 1, 2, \ldots, N \qquad (12.51)$$

Weighted Centralized Strategy

The MSD performance of the weighted centralized solution (12.50) can be verified to match that of the optimized distributed solution (12.47). Indeed, compared with (12.49), we can interpret algorithm (12.50) as corresponding to the centralized stochastic gradient implementation that would result from minimizing instead the following modified global cost

$$J^{\text{glob,b}}(w) \triangleq \sum_{k=1}^{N} J_k^b(w) \tag{12.52}$$

where each individual cost is a scaled version of the original cost:

$$J_k^b(w) \triangleq N p_k^o \, J_k(w) \tag{12.53}$$

Weighted Centralized Strategy

In this way, the gradient noise vectors that result from using the modified costs $\{J_k^b(w)\}$ will be scaled by the same factors $\{Np_k^o\}$ relative to the gradient noise vectors that result from using the original costs $\{J_k(w)\}$. Specifically, if we denote the individual gradient noise process corresponding to implementation (12.49) by

$$\boldsymbol{s}_{k,i}(\boldsymbol{w}_{i-1}) = \widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{i-1}) - \nabla_{w^*} J_k(\boldsymbol{w}_{i-1}) \tag{12.54}$$

Weighted Centralized Strategy

then the gradient noise process that corresponds to implementation (12.50) will be given by

$$\begin{aligned} \boldsymbol{s}_{k,i}^{b}(\boldsymbol{w}_{i-1}) &\triangleq \widehat{\nabla_{w^*} J_k^{b}}(\boldsymbol{w}_{i-1}) - \nabla_{w^*} J_k^{b}(\boldsymbol{w}_{i-1}) \\ &= N p_k^o \, \boldsymbol{s}_{k,i}(\boldsymbol{w}_{i-1}) \end{aligned} \tag{12.55}$$

under the reasonable expectation that the gradient vector approximation, $\widehat{\nabla_{w^*} J_k^{b}}(\boldsymbol{w}_{i-1})$, is similarly scaled by Np_k^o. Consequently, the limiting moment matrices corresponding to the new gradient noise vectors, $\{\boldsymbol{s}_{k,i}^{b}(w^o)\}$, will be scaled multiples of the moment matrices corresponding to the previous gradient noise vectors $\{\boldsymbol{s}_{k,i}(w^o)\}$, i.e.,

Weighted Centralized Strategy

$$R_{s,k}^{b} = (Np_k^o)^2 R_{s,k} \tag{12.56}$$

$$R_{q,k}^{b} = (Np_k^o)^2 R_{q,k}, \quad k = 1, 2, \ldots, N \tag{12.57}$$

It follows from definition (5.56) that the matrices $\{H^b, G_k^b\}$ for the weighted centralized solution (12.50) are related to the matrices $\{H, G_k\}$ for the original centralized solution (12.49) as follows:

$$H^b = Np_k^o H \tag{12.58}$$

$$G_k^b = (Np_k^o)^2 G_k, \quad k = 1, 2, \ldots, N \tag{12.59}$$

Weighted Centralized Strategy

and, therefore, the corresponding gradient noise factors $\{\theta_k^2, (\theta_k^b)^2\}$ are related as

$$\left(\theta_k^b\right)^2 = N p_k^o \theta_k^2, \quad k = 1, 2, \ldots, N \tag{12.60}$$

Substituting into (12.46) we find that the MSD level for the weighted centralized solution, denoted by MSD_{wcen} is given by

Weighted Centralized Strategy

$$\begin{aligned}
\mathrm{MSD}_{\mathrm{wcen}} &= \frac{\mu}{2N^2h}\sum_{\ell=1}^{N}\left(\theta_\ell^b\right)^2 \\
&= \frac{\mu}{2N^2h}\sum_{\ell=1}^{N} N p_\ell^o \theta_\ell^2 \\
&\overset{(12.51)}{=} \frac{\mu}{2h}\left(\sum_{\ell=1}^{N}\frac{1}{\theta_\ell^2}\right)^{-1} \\
&\overset{(12.47)}{=} \mathrm{MSD}^o_{\mathrm{dist,av}} \qquad (12.61)
\end{aligned}$$

Weighted Centralized Strategy

We conclude that it is possible to modify the centralized solution into the weighted form (12.50) such that the MSD performance of the optimal distributed solution matches the MSD performance of the weighted centralized solution.

Example #12.4

Example 12.4 (Comparing distributed and centralized solutions). We reconsider the setting of Example 11.3, which deals with MSE networks. We assume uniform step-sizes, $\mu_k \equiv \mu = 0.001$, and real-valued data with uniform regression covariance matrices of the form $R_{u,k} = \sigma_u^2 I_M$ where σ_u^2 is chosen randomly from within the range $[1, 2]$. In this setting, we have

$$H_k = 2\sigma_u^2 I_M \equiv H, \qquad G_k = 4\sigma_{v,k}^2 \sigma_u^2 I_M, \quad \theta_k^2 = 2M\sigma_{v,k}^2 \tag{12.62}$$

Example #12.4

We consider three scenarios. In the first case, the agents run the ATC diffusion strategy (7.23), namely,

$$\begin{cases} \boldsymbol{\psi}_{k,i} &= \boldsymbol{w}_{k,i-1} + 2\mu \boldsymbol{u}_{k,i}^{\mathsf{T}} \left[\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i} \boldsymbol{w}_{k,i-1} \right] \\ \boldsymbol{w}_{k,i} &= \displaystyle\sum_{\ell \in \mathcal{N}_k} a_{\ell k}^o \, \boldsymbol{\psi}_{\ell,i} \end{cases} \tag{12.63}$$

where the combination weights $\{a_{\ell k}^o\}$ are the Hastings weights from (12.39). In the second case, the agents transfer the data to a fusion center running the

Example #12.4

conventional (un-weighted) centralized strategy (5.13), i.e.,

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} + \mu \left(\frac{1}{N} \sum_{k=1}^{N} 2\boldsymbol{u}_{k,i}^{\mathsf{T}} (\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{i-1}) \right) \tag{12.64}$$

In the third case, the fusion center employs a weighted centralized solution of the form:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} + \mu \left(\sum_{k=1}^{N} 2p_k^o \boldsymbol{u}_{k,i}^{\mathsf{T}} (\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{i-1}) \right) \tag{12.65}$$

Example #12.4

where the $\{p_k^o\}$ are the entries of the Perron vector given by (12.21), which in the current setting reduces to:

$$p_k^o = \frac{1}{\sigma_{v,k}^2}\left(\sum_{\ell=1}^{N}\frac{1}{\sigma_{v,\ell}^2}\right)^{-1}, \quad k=1,2,\ldots,N \tag{12.66}$$

The resulting MSD performance levels are given by expressions (12.46)–(12.47) and (12.61) using $h=1$:

Example #12.4

$$\mathrm{MSD}_{\mathrm{cent}} = \frac{\mu}{2N^2}\left(\sum_{\ell=1}^{N}\theta_\ell^2\right) \tag{12.67}$$

$$\mathrm{MSD}_{\mathrm{wcent}} = \mathrm{MSD}^{o}_{\mathrm{dist,av}} = \frac{\mu}{2}\left(\sum_{\ell=1}^{N}\frac{1}{\theta_\ell^2}\right)^{-1} \tag{12.68}$$

where $\theta_\ell^2 = 2M\sigma_{v,\ell}^2$.

Example #12.4

We illustrate these results numerically for the connected network topology shown in Figure 12.2 with $N = 20$ agents. The measurement noise variances, $\{\sigma_{v,k}^2\}$, and the power of the regression data, are shown in the plots of Figure 12.3, 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 #12.4

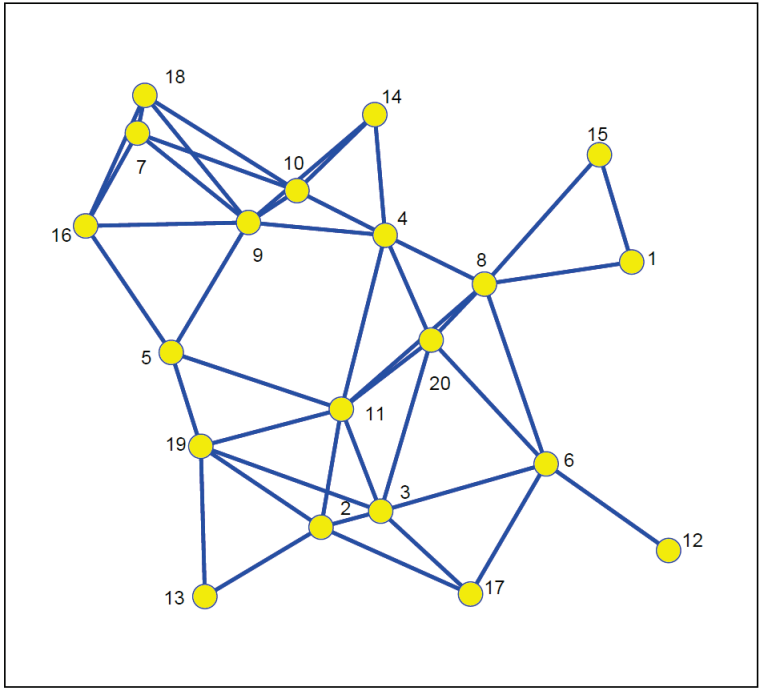

Figure 12.2

Example #12.4

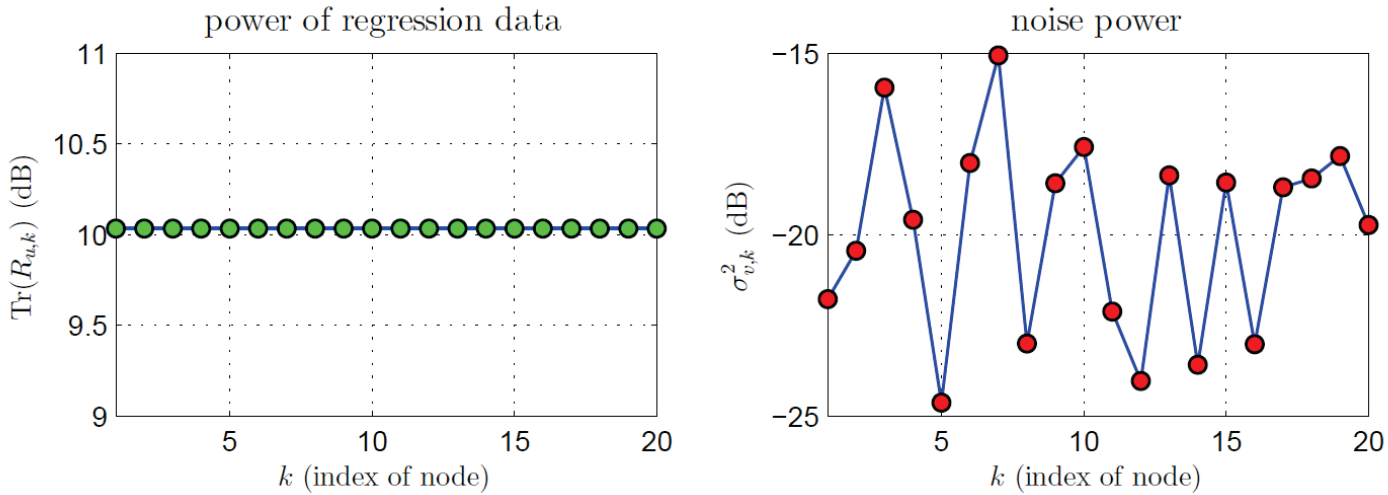

Figure 12.3: Regression data power (left) and measurement noise profile (right) across all agents in the network. The covariance matrices are assumed to be of the form $R_{u,k} = \sigma_u^2 I_M$, and the noise and regression data are Gaussian distributed in this simulation.

Example #12.4

Figure 12.4 plots the resulting learning curves for the three algorithms listed above: ATC diffusion, centralized, and weighted centralized running on real-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. The figure plots the evolution of the ensemble-average learning curves, $\frac{1}{N}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ for diffusion and $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ for centralized and weighted centralized. The curves are obtained by averaging simulated trajectories over 100 repeated experiments. The labels on the vertical axes in the figures refer to the learning curves by writing MSD(i), with an iteration index i. It is seen in the figure that the MSD level that is attained

Example #12.4

by the diffusion strategy is better (lower) than the MSD level that is attained by the un-weighted centralized strategy, in agreement with the theoretical result (12.48). On the other hand, the same figure shows that the weighted centralized solution (12.65) eliminates the degradation in performance, again in agreement with the theoretical result (12.61).

■

Example #12.4

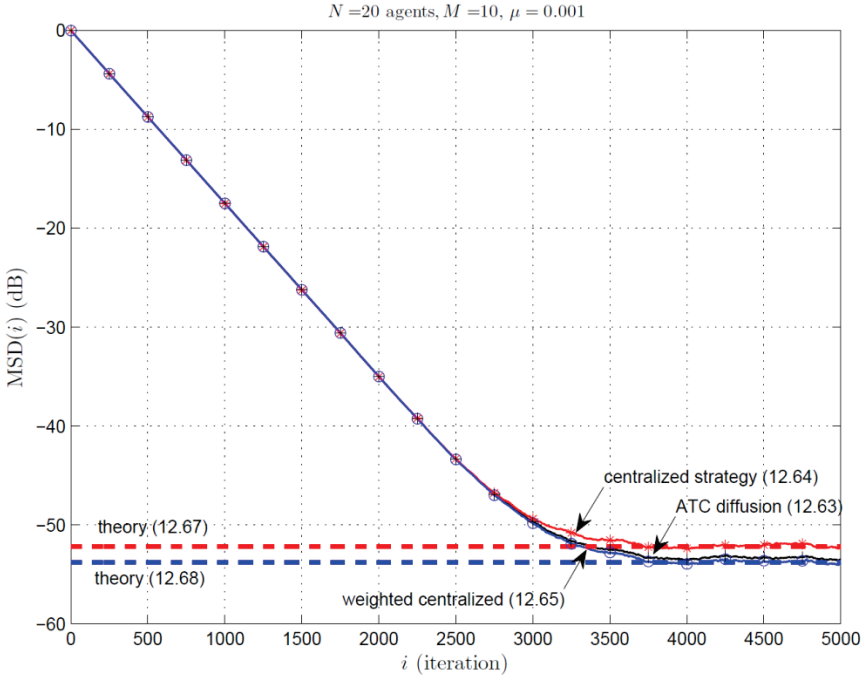

Figure 12.4

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

We focused in the previous sections on the MSD performance measure. The same conclusions extend to the ER performance measure and, therefore, we shall be brief. To begin with, for a meaningful comparison with the non-cooperative solution, we shall assume in this section that all cost functions are uniform across the agents, namely,

$$J_k(w) \equiv J(w), \quad k = 1, 2, \ldots, N \tag{12.69}$$

The ER performance levels for the non-cooperative, centralized, and distributed strategies are then given by

ER Performance

$$\mathrm{ER}_{\mathrm{cent}} = \frac{\mu h}{4}\left(\frac{1}{N^2}\right)\mathrm{Tr}\left(\sum_{k=1}^{N} R_{s,k}\right) \tag{12.70}$$

$$\mathrm{ER}_{\mathrm{ncop},k} = \frac{\mu h}{4}\mathrm{Tr}\left(R_{s,k}\right) \tag{12.71}$$

$$\mathrm{ER}_{\mathrm{ncop,av}} = \frac{\mu h}{4}\left(\frac{1}{N}\right)\mathrm{Tr}\left(\sum_{k=1}^{N} R_{s,k}\right) \tag{12.72}$$

$$\mathrm{ER}_{\mathrm{dist},k} = \mathrm{ER}_{\mathrm{dist,av}} = \frac{\mu h}{4}\mathrm{Tr}\left(\sum_{k=1}^{N} p_k^2 R_{s,k}\right) \tag{12.73}$$

ER Performance

For doubly-stochastic combination matrices, and to first-order in the step-size parameter, it again holds that

$$\mathrm{ER}_{\mathrm{dist,av}} = \mathrm{ER}_{\mathrm{cent}} = \frac{1}{N}\mathrm{ER}_{\mathrm{ncop,av}} \tag{12.74}$$

ER Performance

This result is in terms of the average network performance (obtained by averaging the ER levels of the individual agents). In this way, the result establishes that the average ER performance of the distributed solution is $N-$fold superior (i.e., lower) than the average ER performance attained by the agents in a non-cooperative solution. However, from (12.71) and (12.73) we observe that the ER of the individual agents in the distributed implementation will be smaller (and, hence, better) than the ER of the individual agents in the non-cooperative implementation only when for each $k = 1, 2, \ldots, N$:

ER Performance

$$\frac{1}{N}\sum_{k=1}^{N}\mathrm{Tr}\left(R_{s,k}\right) \quad \leq \quad N\mathrm{Tr}(R_{s,k}) \tag{12.75}$$

Unfortunately, this condition may or may not hold. For example, if all the $\{R_{s,k}\}$ are uniform across the agents, then the condition is clearly satisfied and the performance of all individual agents will improve through cooperation. On the other hand, if we consider the example $N = 2$, $R_{s,1} = rI_M$ and $R_{s,2} = 9rI_M$ for some $r > 0$. Then,

$$\frac{1}{N}\sum_{k=1}^{N}\mathrm{Tr}\left(R_{s,k}\right) \;=\; 5rI_M \tag{12.76}$$

ER Performance

which is larger than $2R_{s,1}$ but smaller than $2R_{s,2}$. In this case, agent 2 will benefit from cooperation while agent 1 will not.

We can then seek a left-stochastic policy that optimizes the ER level by solving

$$A^o \triangleq \arg\min_{A\in\mathbb{A}} \ \mathrm{Tr}\left(\sum_{k=1}^{N} p_k^2 R_{s,k}\right) \qquad (12.77)$$

$$\text{subject to } Ap = p, \ \ \mathbb{1}^{\mathsf{T}} p = 1, \ \ p_k > 0$$

ER Performance

The solution to (12.77) can be obtained in a manner similar to the solution of the earlier problem (12.18). The only difference is that the parameters θ_k^2 should now be defined as follows:

$$\theta_k^2 \triangleq \mathrm{Tr}(R_{s,k}), \quad k = 1, 2, \ldots, N \tag{12.78}$$

in terms of the moment matrices $\{R_{s,k}\}$ alone — compare with (12.19). These parameters can then be used in (12.20) to construct the corresponding Hastings combination rule. The resulting (optimized) ER value will be

ER Performance

$$\mathrm{ER}^{o}_{\mathrm{dist,av}} = \frac{\mu h}{4}\left(\sum_{\ell=1}^{N}\frac{1}{\theta_\ell^2}\right)^{-1} \tag{12.79}$$

and it again holds that

$$\mathrm{ER}^{o}_{\mathrm{dist,av}} \leq \mathrm{ER}^{\mathrm{d.s.}}_{\mathrm{dist,av}} = \frac{1}{N}\mathrm{ER}_{\mathrm{ncop,av}} \tag{12.80}$$

so that, as expected, the ER of the distributed (consensus or distributed) network with an optimal left-stochastic matrix, A^o, is also superior to the ER of the non-cooperative scenario. More importantly, though, this optimal choice for A leads again to

$$\mathrm{ER}^{o}_{\mathrm{dist},k} \leq \mathrm{ER}_{\mathrm{ncop},k}, \quad k = 1, 2, \ldots, N \tag{12.81}$$

Example #12.5

Example 12.5 (Comparing distributed and centralized learners). We reconsider the numerical example at the end of Example 11.11, which deals with logistic networks operating on real data $\{\boldsymbol{\gamma}_k(i), \boldsymbol{h}_{k,i}\}$ originating from the alpha data set [223]. We consider the same network topology shown earlier in Figure 11.5 with $N = 20$ agents employing uniform step-sizes, $\mu_k \equiv \mu$. We already know from the result of Example 12.3 that the (optimal) Hastings rule reduces to the Metropolis rule (12.43), which is doubly-stochastic. Therefore, the entries of the corresponding Perron eigenvector are $p_k^o = 1/N$.

Example #12.5

In this example, we compare the performance of two algorithms, ATC diffusion and the weighted centralized strategy, for the minimization of the (regularized) logistic risk function (11.205). The algorithms take the following form in this case:

$$\left\{\begin{array}{rcl} \boldsymbol{\psi}_{k,i} & = & (1-\rho\mu_k)\boldsymbol{w}_{k,i-1} + \mu\boldsymbol{\gamma}_k(i)\boldsymbol{h}_{k,i}\left(\dfrac{1}{1+e^{\boldsymbol{\gamma}_k(i)\boldsymbol{h}_{k,i}^{\mathsf{T}}\boldsymbol{w}_{k,i-1}}}\right) \\ \boldsymbol{w}_{k,i} & = & \displaystyle\sum_{\ell\in\mathcal{N}_k} a_{\ell k}\,\boldsymbol{\psi}_{\ell,i} \qquad (\underline{\text{ATC diffusion}}) \end{array}\right. \tag{12.82}$$

and

$$\boldsymbol{w}_i = (1-\rho\mu)\boldsymbol{w}_{i-1} + \frac{\mu}{N}\sum_{k=1}^{N}\boldsymbol{\gamma}_k(i)\boldsymbol{h}_{k,i}\left(\frac{1}{1+e^{\boldsymbol{\gamma}_k(i)\boldsymbol{h}_{k,i}^{\mathsf{T}}\boldsymbol{w}_{i-1}}}\right) \quad (\underline{\text{weigh. centr.}})$$

Example #12.5

In this case, and since the combination policy is doubly-stochastic, the ER performance of both algorithms will tend towards similar values. Using expression (12.79) with $h = 1$ for real data, this level is given by

$$\mathrm{ER}_{\mathrm{cent}} = \mathrm{ER}^{o}_{\mathrm{dist,av}} = \frac{\mu}{4}\left(\sum_{\ell=1}^{N}\frac{1}{\theta_\ell^2}\right)^{-1} = \frac{\mu}{4N}\mathrm{Tr}(R_s) \tag{12.84}$$

where we used (12.78) to note that

$$\theta_\ell^2 \equiv \theta^2 = \mathrm{Tr}(R_s) \tag{12.85}$$

Example #12.5

Figure 12.5 plots the evolution of the ensemble-average learning curves, $\mathbb{E}\,\{J(\boldsymbol{w}_{i-1}) - J(w^o)\}$, for the above ATC diffusion and weighted centralized strategies using $\mu = 1 \times 10^{-4}$. The curves are obtained by averaging the trajectories $\{J(\boldsymbol{w}_{i-1}) - J(w^o)\}$ over 100 repeated experiments. The label on the vertical axis in the figure refers to the learning curves by writing $\mathrm{ER}(i)$, with an iteration index i. Each experiment involves running the diffusion strategy (12.82) or the weighted centralized strategy (12.83) with $\rho = 10$. 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 $\mathrm{Tr}(R_s) \approx 131.48$. It is observed in the figure that the learning curves tend towards the ER value predicted by the theoretical expression (12.84).

Example #12.5

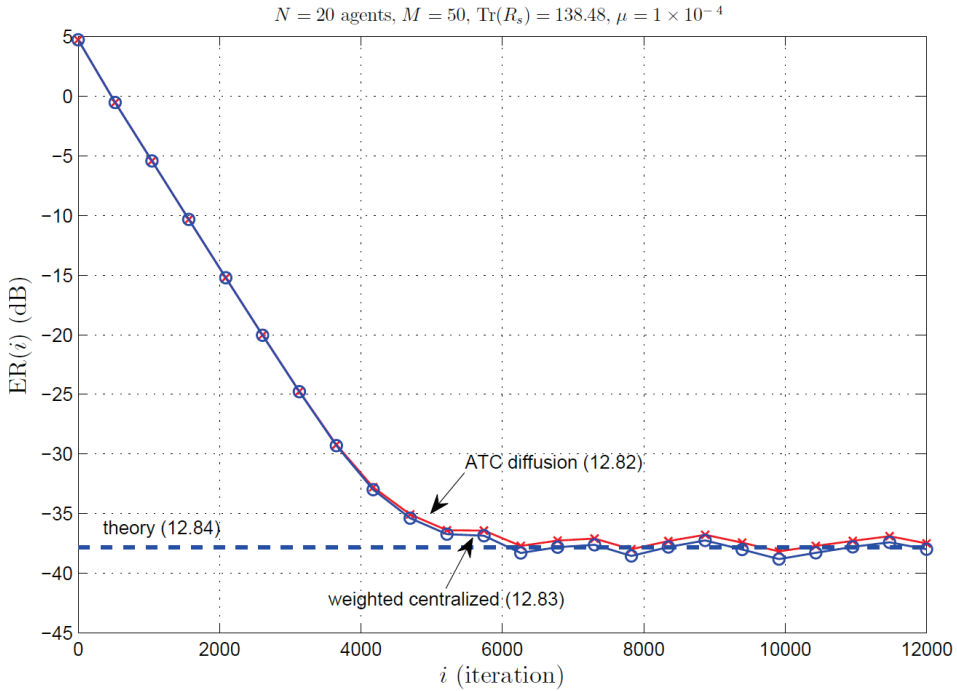

Figure 12.5

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.