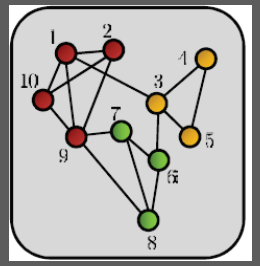


INFERENCE OVER NETWORKS

LECTURE #24: Combination Policies

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 14 (Combination Policies, pp. 662-682):

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.

Setting

We end our exposition by commenting on the selection of the combination policy, A. Although unnecessary, we assume in this chapter that all agents are informed so that their step-sizes are strictly positive. It is clear from the performance expression (11.118) that the combination weights $\{a_{\ell k}\}$ that are used by the consensus (7.9) and diffusion strategies (7.18) and (7.19) influence the performance of the distributed solution in a direct manner. Their influence is reflected by the entries $\{p_k\}$, defined earlier through (11.136), namely,

Setting

$$\mathrm{MSD}_{\mathrm{dist},k} = \mathrm{MSD}_{\mathrm{dist,av}} = \frac{1}{2h}\mathrm{Tr}\left[\left(\sum_{k=1}^{N}\mu_k p_k H_k\right)^{-1}\left(\sum_{k=1}^{N}\mu_k^2 p_k^2 G_k\right)\right] \quad (14.1)$$

There are several ways by which the coefficients $\{a_{\ell k}\}$ can be selected. On one hand, many existing combination policies rely on static selections for these coefficients, i.e., selections that are fixed during the adaptation and learning process and do not change with time. On the other hand, the discussion will reveal that it is important to consider selections where these coefficients are also adapted over time, and are allowed to evolve dynamically alongside the learning mechanism.

Static 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.

Static Policies

To begin with, Table 14.1 is extracted from [208] and lists some common static choices for selecting the combination weights $\{a_{\ell k}\}$ for a network with N agents. In the table, the symbol $n_k = |\mathcal{N}_k|$ denotes the degree of agent k, which is equal to the size of its neighborhood, and the symbol $n_{\max}$ denotes the maximum degree across the network:

$$n_{\max} \stackrel{\Delta}{=} \max_{1 \le k \le N} n_k \tag{14.2}$$

Static Policies

The Laplacian rule, which appears in the second line of the table, relies on the use of the Laplacian matrix of the network and a positive scalar, β. The Laplacian matrix is a symmetric matrix whose entries are constructed as follows [41, 82, 143, 208]:

$$[\mathcal{L}]_{\ell k} = \begin{cases} n_\ell - 1, & \text{if } k = \ell \\ -1, & \text{if } k \neq \ell \text{ and } \ell \in \mathcal{N}_k \\ 0, & \text{otherwise} \end{cases} \tag{14.3}$$

Static Policies

The Laplacian matrix has several useful properties and conveys important information about the network topology [208, App. B]. For example, (a) $\mathcal{L}$ is always nonnegative-definite; (b) the entries on each of its rows add up to zero; and (c) its smallest eigenvalue is zero. Moreover, (d) the multiplicity of zero as an eigenvalue for $\mathcal{L}$ is equal to the number of connected subgraphs of the network topology. Accordingly, a graph is connected if, and only if, the second smallest eigenvalue of $\mathcal{L}$ (also called the algebraic connectivity of the graph) is nonzero.

Static Policies

Entries of combination matrix A	Type of A
1. Averaging rule [39]: $a_{\ell k} = \begin{cases} 1/n_k, & \text{if } \ell \in \mathcal{N}_k \\ 0, & \text{otherwise} \end{cases}$	left-stochastic
2. Laplacian rule [215, 265]: $a_{\ell k} = 1 - \beta[\mathcal{L}]_{\ell k},\ \beta > 0$	symmetric and doubly-stochastic
3. Laplacian rule using $\beta = 1/n_{\max}$: $a_{\ell k} = \begin{cases} 1/n_{\max}, & \text{if } k \neq \ell \text{ are neighbors} \\ 1 - \frac{(n_k - 1)}{n_{\max}}, & k = \ell \\ 0, & \text{otherwise} \end{cases}$	symmetric and doubly-stochastic
4. Laplacian rule using $\beta = 1/N$ (or maximum-degree rule [266]) : $a_{\ell k} = \begin{cases} 1/N, & \text{if } k \neq \ell \text{ are neighbors} \\ 1 - (n_k - 1)/N, & k = \ell \\ 0, & \text{otherwise} \end{cases}$	symmetric and doubly-stochastic
5. Metropolis rule [106, 167, 265]: $a_{\ell k} = \begin{cases} \frac{1}{\max\{n_k, n_\ell\}}, & \text{if } k \neq \ell \text{ are neighbors} \\ 1 - \sum_{m \in \mathcal{N}_k \backslash \{k\}} a_{mk}, & k = \ell \\ 0, & \text{otherwise} \end{cases}$	symmetric and doubly-stochastic
6. Relative-degree rule [58]: $a_{\ell k} = \begin{cases} n_\ell \left(\sum_{m \in \mathcal{N}_k} n_m \right)^{-1}, & \text{if } \ell \in \mathcal{N}_k \\ 0, & \text{otherwise} \end{cases}$	left-stochastic

Static Policies

It is observed from the constructions in Table 14.1 that the values of the combination weights $\{a_{\ell k}\}$ are solely determined by the degrees (and, hence, the extent of connectivity) of the agents. As explained in [208], while such selections may be appropriate in some applications, they can nevertheless lead to degraded performance in the context of adaptation and learning over networks [232]. This is because these weighting schemes ignore the gradient noise profile across the network.

Need for Adaptive Policies

One way to capture the gradient noise profile across the network is by means of the factors $\{\theta_k^2\}$ defined earlier in (12.19) and (12.78):

$$\theta_k^2 \triangleq \begin{cases} \mathrm{Tr}(H^{-1}G_k) & \text{(for MSD performance)} \\ \mathrm{Tr}(R_{s,k}) & \text{(for ER performance)} \end{cases} \tag{14.4}$$

where G_k is also dependent on the gradient noise variance, $R_{s,k}$, in view of definition (11.12). Now, since some agents can be noisier (with larger θ_k^2) than others, it becomes important to take into account the amount

Need for Adaptive Policies

of noise that is present at the agents and to assign more or less weights to interactions with neighbors in accordance to their noise level. For example, if some agent k can determine which of its neighbors are the noisiest, then it can assign smaller combination weights to its interaction with these neighbors. One difficulty in employing this strategy is that the noise factors $\{\theta_\ell^2\}$ are unknown beforehand since their values depend on the unknown noise moments $\{G_\ell, R_{s,\ell}\}$. It therefore becomes necessary to devise noise-aware schemes that enable agents to estimate

Need for Adaptive Policies

the noise factors $\{\theta_\ell^2\}$ of their neighbors in order to assist them in the process of selecting proper combination coefficients. It is also desirable for these schemes to be adaptive so that they can track variations in the noise moments over time. The techniques described in this chapter are motivated by the procedures developed in [208, 244, 280]; variations appear in [95, 270]. We first consider an example to illustrate the idea.

Example #14.1

Example 14.1 (Noise variance estimation over MSE networks). We continue with the MSE network from Example 12.1 where we assumed uniform step-sizes and uniform regression covariance matrices, i.e., $\mu_k \equiv \mu$ and $R_{u,k} \equiv R_u > 0$ for $k = 1, 2, \ldots, N$. Recall that for these networks, the data $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$ are assumed to be related via the linear regression model:

$$\boldsymbol{d}_k(i) \;=\; \boldsymbol{u}_{k,i} w^o \;+\; \boldsymbol{v}_k(i), \qquad k = 1, 2, \ldots, N \tag{14.5}$$

where the variance of the noise is denoted by $\sigma_{v,k}^2 = \mathbb{E}\,|\boldsymbol{v}_k(i)|^2$. We derived in Example 12.2 the (optimal) combination coefficients in the form of the

Example #14.1

Hastings rule (12.39), namely,

$$a_{\ell k}^{o} = \begin{cases} \dfrac{\sigma_{v,k}^2}{\max\{\, n_k\sigma_{v,k}^2,\ n_\ell\sigma_{v,\ell}^2 \,\}}, & \ell \in \mathcal{N}_k \backslash \{k\} \\ 1 - \displaystyle\sum_{m \in \mathcal{N}_k \backslash \{k\}} a_{mk}^{o}, & \ell = k \end{cases} \tag{14.6}$$

and noted that the gradient noise factors in this case are given by $\theta_k^2 = 2M\sigma_{v,k}^2$; they are therefore proportional to the measurement noise power, $\sigma_{v,k}^2$. It is clear that rule (14.6) takes into account the size of the noise powers, $\{\sigma_{v,\ell}^2\}$, at the agents. Moreover, in this particular construction, only the noise

Example #14.1

levels of the two interacting agents are directly involved in the computation of their combination weights; no other agents from the neighborhood of agent k are involved in the calculation.

A second combination construction is motivated in [280] for MSE networks by solving an alternative optimization problem than the one that led to the Hastings rule (12.39) or (14.6). We shall describe this alternative construction further ahead in (14.27). For now, we simply state that the resulting combination rule for the case under study in this example, and which we shall refer to as the relative-variance rule [206], takes the following form:

Example #14.1

$$a_{\ell k}^{o} = \begin{cases} \dfrac{1}{\sigma_{v,\ell}^2}\left(\displaystyle\sum_{m\in\mathcal{N}_k}\frac{1}{\sigma_{v,m}^2}\right)^{-1}, & \ell\in\mathcal{N}_k \\ 0, & \text{otherwise} \end{cases} \tag{14.7}$$

Comparing with (14.6), we note that in this second rule, the interaction between agents k and ℓ is more broadly dependent on the noise profile across the *entire* neighborhood of agent k. In particular, neighbors with smaller noise power relative to the neighborhood are assigned larger weights.

Example #14.1

For every agent k, both rules (14.6) and (14.7) still require knowledge of the noise variances $\{\sigma_{v,\ell}^2\}$. This information is generally unavailable but can be estimated by agent k as follows — see the derivation that leads to (14.53) in the next section. Assume, for illustration purposes, that the agents are running the ATC LMS diffusion strategy (7.23):

$$\left\{\begin{array}{rcl} \boldsymbol{\psi}_{k,i} & = & \boldsymbol{w}_{k,i-1} + \mu \boldsymbol{u}_{k,i}^{*} \left[\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{k,i-1}\right] \\ \boldsymbol{w}_{k,i} & = & \displaystyle\sum_{\ell \in \mathcal{N}_k} a_{\ell k}\, \boldsymbol{\psi}_{\ell,i} \end{array}\right. \tag{14.8}$$

Example #14.1

Then, agent k can estimate the noise variance, $\sigma^2_{v,\ell}$, by running the recursion:

$$\boldsymbol{\gamma}^2_{\ell k}(i) \;=\; (1-\zeta)\,\boldsymbol{\gamma}^2_{\ell k}(i-1) \;+\; \zeta\,\|\boldsymbol{\psi}_{\ell,i} - \boldsymbol{w}_{k,i-1}\|^2, \qquad \ell \in \mathcal{N}_k \tag{14.9}$$

where $0 < \zeta \ll 1$ is a small positive coefficient, e.g., $\zeta = 0.1$. This recursion relies on smoothing the energy of the difference between the intermediate iterate, $\boldsymbol{\psi}_{\ell,i}$, received from neighbor ℓ and the existing iterate $\boldsymbol{w}_{k,i-1}$ at agent k. The resulting energy measure provides an indication of the amount of noise that is present at agent ℓ since it can be verified that asymptotically [208] — see also (14.55):

$$\mathbb{E}\,\boldsymbol{\gamma}^2_{\ell k}(i) \;\approx\; \mu^2\sigma^2_{v,\ell}\mathrm{Tr}(R_u), \qquad i \gg 1 \tag{14.10}$$

Example #14.1

with the limit being proportional to $\sigma_{v,\ell}^2$. Therefore, the running variables $\{\boldsymbol{\gamma}_{\ell k}^2(i)\}$ can be used by agent k as scaled estimates for the noise variances. These variables can then be used in place of the noise variances in rules (14.6) and (14.7) to adapt the combination weights over time. Under this construction, each agent k ends up running n_k recursions of the form (14.9), one for each of its neighbors, in order to update the necessary variables $\{\boldsymbol{\gamma}_{\ell k}^2(i),\ \ell \in \mathcal{N}_k\}$.

■

Noise-Aware 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.

Hastings Policy

Before discussing adaptive constructions for the combination weights, we present two combination policies that are noise-aware. We already encountered one such policy when we derived the Hastings rule earlier in Sec. 12.2 — see expression (12.20). Here we review it briefly before discussing the second policy, known as the relative variance rule. Recall that the Hastings rule was derived under the condition of uniform step-sizes and uniform Hessian matrices, namely,

$$\mu_k \equiv \mu, \quad H_k \equiv H, \quad k = 1, 2, \ldots, N \tag{14.11}$$

Hastings Policy

Optimal MSD level:

$$A^o \triangleq \arg\min_{A\in\mathbb{A}} \ \mathrm{Tr}\left(\sum_{k=1}^{N} p_k^2 H^{-1} G_k\right) \quad \text{subject to} \ \ Ap = p, \ \ \mathbb{1}^{\mathsf{T}} p = 1, \ \ p_k > 0 \tag{12.18}$$

Hastings Policy

The rule followed from the solution to the optimization problem (12.18) and led to

$$a_{\ell k}^o = \begin{cases} \dfrac{\theta_k^2}{\max\{\, n_k\theta_k^2,\ n_\ell\theta_\ell^2 \,\}}, & \ell \in \mathcal{N}_k \backslash \{k\} \\ 1 - \displaystyle\sum_{m \in \mathcal{N}_k \backslash \{k\}} a_{mk}^o, & \ell = k \end{cases} \tag{14.12}$$

Observe how the entries of this policy are dependent on the gradient-noise factors:

$$\theta_k^2 \triangleq \mathrm{Tr}(H^{-1}G_k), \quad k = 1, 2, \ldots, N \tag{14.13}$$

Hastings Policy

Observe also that these factors are not only dependent on G_k but that they also depend on the Hessian matrix information, H. In comparison, the relative-variance policy described in the next section will be independent of H. Recall from the derivation in Sec. 12.2 that the above Hastings rule is a solution to the optimization problem (12.18); it therefore minimizes the network MSD. While deriving the Hastings rule in Sec. 12.2, we formulated the problem in the context of cost

Hastings Policy

functions, $\{J_k(w)\}$, that share a common minimizer. In this case, the minimizer, w^o, of the aggregate cost, $J^{\text{glob}}(w)$, defined by (8.44) will be invariant under the combination policy, A. For this reason, we can interpret Hastings rule (14.12) as providing a combination policy that results in the smallest possible MSD relative to the same fixed limit point w^o.

Relative-Variance Policy

We now describe a second noise-aware policy to select the combination weights; this second rule will be independent of the Hessian matrix information, H.

Recall:

$$\mathrm{MSD}_{\mathrm{dist},k} = \mathrm{MSD}_{\mathrm{dist,av}} = \frac{1}{2h}\mathrm{Tr}\left[\left(\sum_{k=1}^{N}\mu_k p_k H_k\right)^{-1}\left(\sum_{k=1}^{N}\mu_k^2 p_k^2 G_k\right)\right] \quad (14.1)$$

Relative-Variance Policy

Recall that the Hastings rule was derived by working with the MSD expression (12.5), which results from keeping the first-order term in the MSD expression (11.178). The second policy that we shall derive here, and which we refer to as the relative-variance policy, is instead based on working with the alternative MSD expression (11.178). The derivation of this second policy does not require the uniformity conditions (14.11).

Relative-Variance Policy

To begin with, we know from (11.178) that the MSD performance of the ATC diffusion network (7.19) can be evaluated by means of the following series expression for sufficiently small step-sizes:

$$\mathrm{MSD}_{\mathrm{dist,av}}^{\mathrm{atc}} = \frac{1}{hN}\sum_{n=0}^{\infty}\mathrm{Tr}\left[\mathcal{B}_{\mathrm{atc}}^{n}\mathcal{Y}_{\mathrm{atc}}\left(\mathcal{B}_{\mathrm{atc}}^{*}\right)^{n}\right] \tag{14.14}$$

where $h = 1$ for real data and $h = 2$ for complex data, and where the

Relative-Variance Policy

matrix quantities $\{\mathcal{B}_{\rm atc}, \mathcal{Y}_{\rm atc}\}$ are defined as follows:

$$\mathcal{B}_{\rm atc} = \mathcal{A}^{\mathsf{T}}\left(I_{hMN} - \mathcal{M}\mathcal{H}\right) \tag{14.15}$$

$$\mathcal{Y}_{\rm atc} = \mathcal{A}^{\mathsf{T}}\mathcal{M}\mathcal{S}\mathcal{M}\mathcal{A} \tag{14.16}$$

which in turn are defined in terms of the quantities:

$$\mathcal{M} = \mathrm{diag}\left\{\mu_1 I_{hM},\ \mu_2 I_{hM}, \ldots, \mu_N I_{hM}\right\} \tag{14.17}$$

$$\mathcal{S} = \mathrm{diag}\{G_1, G_2, \ldots, G_N\} \tag{14.18}$$

$$\mathcal{R} = \mathrm{diag}\left\{H_1, H_2, \ldots, H_N\right\} \tag{14.19}$$

$$\mathcal{A} = A \otimes I_{hM} \tag{14.20}$$

Relative-Variance Policy

Starting from (14.14), we pose the problem of seeking a left-stochastic combination matrix A that solves:

$$A^o \triangleq \arg\min_{A\in\mathbb{A}} \sum_{n=0}^{\infty} \mathrm{Tr}\left[\mathcal{B}_{\mathrm{atc}}^n \mathcal{Y}_{\mathrm{atc}} \left(\mathcal{B}_{\mathrm{atc}}^*\right)^n\right] \quad \text{subject to } A^{\mathsf{T}}\mathbb{1} = \mathbb{1},\ a_{\ell k} \geq 0,\ a_{\ell k} = 0 \text{ if } \ell \notin \mathcal{N}_k \tag{14.21}$$

However, solving problem (14.21) is generally non-trivial and we replace it by a more tractable problem. Specifically, we replace the cost in (14.21) by an upper bound and minimize this upper bound instead. Indeed, it is shown in [208, Sec. 8.2] that the following inequality holds for a stable matrix $\mathcal{B}_{\mathrm{atc}}$:

Relative-Variance Policy

$$\sum_{n=0}^{\infty} \mathrm{Tr}\left[\mathcal{B}_{\mathrm{atc}}^{n}\mathcal{Y}_{\mathrm{atc}}\left(\mathcal{B}_{\mathrm{atc}}^{*}\right)^{n}\right] \leq c\,\mathrm{Tr}(\mathcal{Y}_{\mathrm{atc}}) \qquad (14.22)$$

for some finite positive constant c that is *independent* of A. In other words, the series is upper bounded by a multiple of the trace of $\mathcal{Y}_{\mathrm{atc}}$, which happens to be the first term of the series itself. Therefore, instead of minimizing the series in (14.21), we replace the problem by that of minimizing its first term, namely,

Relative-Variance Policy

$$\begin{aligned} \min_{A \in \mathbb{A}} \quad & \mathrm{Tr}(\mathcal{Y}_{\mathrm{atc}}) \\ \text{subject to} \quad & A^{\mathsf{T}}\mathbb{1} = \mathbb{1}, \;\; a_{\ell k} \geq 0, \;\; a_{\ell k} = 0 \text{ if } \ell \notin \mathcal{N}_k \end{aligned} \tag{14.23}$$

Using definition (14.16), the trace of $\mathcal{Y}_{\mathrm{atc}}$ can be expressed in terms of the combination coefficients $\{a_{\ell k}\}$ as follows:

$$\mathrm{Tr}(\mathcal{Y}_{\mathrm{atc}}) \;=\; \sum_{k=1}^{N} \sum_{\ell=1}^{N} \mu_{\ell}^{2} \, a_{\ell k}^{2} \, \mathrm{Tr}(G_{\ell}) \tag{14.24}$$

and it is seen that problem (14.23) can be decoupled into N separate optimization problems, one for each row of A:

Relative-Variance Policy

$$\min_{\{a_{\ell k}\}_{\ell=1}^{N}} \quad \sum_{\ell=1}^{N} \mu_\ell^2 \, a_{\ell k}^2 \, \mathrm{Tr}(G_\ell), \quad k = 1, \ldots, N$$

$$\text{subject to } \sum_{\ell=1}^{N} a_{\ell k} = 1, \; a_{\ell k} \geq 0, \; a_{\ell k} = 0 \text{ if } \ell \notin \mathcal{N}_k \tag{14.25}$$

With each agent ℓ, we associate the following nonnegative scalar, which is proportional to the trace of the gradient noise moment matrix G_ℓ:

$$\gamma_\ell^2 \triangleq \mu_\ell^2 \, \mathrm{Tr}(G_\ell), \quad \ell = 1, 2, \ldots, N \tag{14.26}$$

$a^{\mathsf{T}} \Gamma a$

Relative-Variance Policy

The factor γ_ℓ^2 so defined plays a role similar to the factor θ_ℓ^2 defined earlier in (14.13) for the Hastings rule; note that both factors contain information about the noise moment matrix, G_ℓ.

Lemma 14.1 (Relative-variance rule). The following combination matrix, denoted by A^o with a superscript o, is a solution to the optimization problem (14.25):

$$a_{\ell k}^o = \begin{cases} \dfrac{1}{\gamma_\ell^2}\left(\displaystyle\sum_{m\in\mathcal{N}_k}\frac{1}{\gamma_m^2}\right)^{-1}, & \text{if } \ell\in\mathcal{N}_k \\ 0, & \text{otherwise} \end{cases} \tag{14.27}$$

Relative-Variance Policy

In the above construction, agent k combines the iterates from its neighbors in proportion to $1/\gamma_\ell^2$. The result is physically meaningful. Agents with smaller noise power, relative to the neighborhood noise power, are assigned larger weights.

Example #14.2

Example 14.2 (Relative-variance rule for MSE networks). We return to the setting of Example 14.1, which deals with MSE networks. The agents employ uniform step-sizes and the data have uniform regression covariance matrices, i.e., $\mu_k \equiv \mu$ and $R_{u,k} \equiv R_u$ for $k = 1, 2, \ldots, N$. In this case,

$$G_k = \sigma_{v,k}^2 \begin{bmatrix} R_u & \times \\ \times & R_u^{\mathsf{T}} \end{bmatrix} \tag{14.28}$$

so that expression (14.27) reduces to expression (14.7), namely,

$$a_{\ell k}^o = \frac{1}{\sigma_{v,\ell}^2} \left(\sum_{m \in \mathcal{N}_k} \frac{1}{\sigma_{v,m}^2} \right)^{-1}, \quad \ell \in \mathcal{N}_k \tag{14.29}$$

Example #14.2

If the step-sizes are not uniform across the agents, then expression (14.27) would instead reduce to

$$a_{\ell k}^{o} = \frac{1}{\mu_\ell^2 \sigma_{v,\ell}^2} \left(\sum_{m \in \mathcal{N}_k} \frac{1}{\mu_m^2 \sigma_{v,m}^2} \right)^{-1}, \quad \ell \in \mathcal{N}_k \tag{14.30}$$

If both the step-sizes and the covariance matrices are not uniform across the agents, then expression (14.27) would lead to:

$$a_{\ell k}^{o} = \frac{1}{\mu_\ell^2 \sigma_{v,\ell}^2 \mathrm{Tr}(R_{u,\ell})} \left(\sum_{m \in \mathcal{N}_k} \frac{1}{\mu_m^2 \sigma_{v,m}^2 \mathrm{Tr}(R_{u,m})} \right)^{-1}, \quad \ell \in \mathcal{N}_k \tag{14.31}$$

Adaptive 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.

Adaptive Policy

To evaluate the relative-variance weights (14.27), the agents still need to know the gradient noise factors, $\{\gamma_\ell^2\}$, defined by (14.26). We motivate in this section a procedure for estimating these factors in an adaptive manner.

To begin with, we recall the definitions of the original and weighted aggregate cost functions:

$$J^{\text{glob}}(w) \triangleq \sum_{k=1}^{N} J_k(w) \tag{14.32}$$

$$J^{\text{glob},\star}(w) \overset{(8.53)}{\triangleq} \sum_{k=1}^{N} q_k J_k(w) \tag{14.33}$$

Adaptive Policy

whose unique minima are denoted by w^o and $w^\star$, respectively. The individual costs, $\{J_k(w)\}$, are assumed to share a common minimizer and, hence, $w^o = w^\star$, i.e.,

$$\nabla_w \, J_k(w^\star) = 0, \quad k = 1, 2, \ldots, N \tag{14.34}$$

The common minimizer assumption ensures that the location of the global solution, w^o or $w^\star$, is fixed and invariant under A. This is a useful condition especially when A is implemented in an adaptive manner and varies with time.

Adaptive Policy

We illustrate the construction of the adaptive combination policy by considering the ATC diffusion strategy (7.19), which is repeated here for ease of reference:

$$\left\{\begin{array}{rcl} \boldsymbol{\psi}_{k,i} & = & \boldsymbol{w}_{k,i-1} - \mu_k \widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{k,i-1}) \\ \boldsymbol{w}_{k,i} & = & \displaystyle\sum_{\ell \in \mathcal{N}_k} a_{\ell k}\, \boldsymbol{\psi}_{\ell,i} \end{array}\right. \tag{14.35}$$

A similar construction applies to the CTA diffusion strategy (7.18) and the consensus strategy (7.9). The following result forms the basis for the procedure developed in this section for estimating the factors $\{\gamma_\ell^2\}$.

Adaptive Policy

Lemma 14.2 (Useful expression for γ_ℓ^2). Consider a network of N interacting agents running the distributed strategy (14.35) with a primitive left-stochastic matrix A. Under the same conditions in the statement of Theorem 9.2, it holds that

$$\mathbb{E}\,\|\boldsymbol{\psi}_{\ell,i}^{e} - \boldsymbol{w}_{\ell,i-1}^{e}\|^2 \;=\; \gamma_\ell^2 \;+\; o(\mu_{\max}^2), \quad \text{for } i \gg 1 \qquad (14.36)$$

Proof

Proof. Using the mean-value theorem (D.20) from the appendix and (14.34) we note that we can write at an arbitrary agent ℓ:

$$\begin{bmatrix} \nabla_{w^*} J_\ell(\boldsymbol{w}_{\ell,i-1}) \\ \nabla_{w^\mathsf{T}} J_\ell(\boldsymbol{w}_{\ell,i-1}) \end{bmatrix} = -\left(\int_0^1 \nabla_w^2 J_\ell(w^\star - r\widetilde{\boldsymbol{w}}_{\ell,i-1})dr\right) \widetilde{\boldsymbol{w}}^e_{\ell,i-1}$$

$$\overset{(8.138)}{=} -\boldsymbol{H}_{\ell,i-1}\widetilde{\boldsymbol{w}}^e_{\ell,i-1} \tag{14.37}$$

where $\widetilde{\boldsymbol{w}}_{\ell,i-1} = w^\star - \boldsymbol{w}_{\ell,i-1}$ and

$$\widetilde{\boldsymbol{w}}^e_{\ell,i-1} \triangleq \begin{bmatrix} \widetilde{\boldsymbol{w}}_{\ell,i-1} \\ \left(\widetilde{\boldsymbol{w}}^*_{\ell,i-1}\right)^\mathsf{T} \end{bmatrix} \tag{14.38}$$

Proof

Therefore, in terms of the extended vectors and replacing the approximate gradient in terms of the sum of the true gradient and the gradient noise process, we can write for any arbitrary agent ℓ:

$$
\begin{aligned}
&\|\boldsymbol{\psi}_{\ell,i}^{e} - \boldsymbol{w}_{\ell,i-1}^{e}\|^2 \\
&\overset{(14.35)}{=} \mu_\ell^2 \left\| \begin{bmatrix} \boldsymbol{s}_{\ell,i}^{e}(\boldsymbol{w}_{e,i-1}) \\ \left(\boldsymbol{s}_{\ell,i}^{e*}(\boldsymbol{w}_{e,i-1})\right)^{\mathsf{T}} \end{bmatrix} + \begin{bmatrix} \nabla_{w^*} J_\ell(\boldsymbol{w}_{\ell,i-1}) \\ \nabla_{w^{\mathsf{T}}} J_\ell(\boldsymbol{w}_{\ell,i-1}) \end{bmatrix} \right\|^2 \\
&\overset{(14.37)}{=} \mu_\ell^2 \|\boldsymbol{s}_{\ell,i}^{e}(\boldsymbol{w}_{\ell,i-1}) - \boldsymbol{H}_{\ell,i-1}\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\|^2 \\
&\overset{(14.34)}{=} \mu_\ell^2 \|\boldsymbol{s}_{\ell,i}^{e}(\boldsymbol{w}_{\ell,i-1})\|^2 + \mu_\ell^2 \|\boldsymbol{H}_{\ell,i-1}\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\|^2 - \\
&\qquad 2\mu_\ell^2 \operatorname{Re}\left[\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e*}\boldsymbol{H}_{\ell,i-1}\boldsymbol{s}_{\ell,i}^{e}(\boldsymbol{w}_{\ell,i-1})\right] \qquad (14.39)
\end{aligned}
$$

Proof

Now, we can deduce from an argument similar to (11.30) and from (11.8) that, for $i \gg 1$, and for sufficiently small step-sizes:

$$\mathbb{E}\,\|\boldsymbol{s}_{\ell,i}^{e}(\boldsymbol{w}_{\ell,i-1})\|^2 \;=\; \mathrm{Tr}(G_{s,\ell}) \;+\; O(\mu_{\max}^{\gamma'/2}) \tag{14.40}$$

where $\gamma' = \min\{\gamma, 2\}$ and $\gamma \in (0, 4]$. Likewise, we can deduce from an argument similar to (9.280) that, for small step-sizes and for $i \gg 1$:

$$\mathbb{E}\,\|\boldsymbol{H}_{\ell,i-1}\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\|^2 \;\le\; a\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\|^4 \;\overset{(9.107)}{=}\; O(\mu_{\max}^2) \tag{14.41}$$

for some constant a that is independent of $\mu_{\max}$. Moreover, using the inequalities $|x^*y| \le \|x\|\,\|y\|$ for any vectors x and y, and $(\mathbb{E}\,\boldsymbol{a})^2 \le \mathbb{E}\,\boldsymbol{a}^2$ for any scalar real-valued random variable $\boldsymbol{a}$, we have

Proof

$$
\begin{aligned}
&\mathbb{E}\,\Big[\,\big|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e*}\boldsymbol{H}_{\ell,i-1}\boldsymbol{s}_{\ell,i}^{e}(\boldsymbol{w}_{\ell,i-1})\big|\;\big|\,\boldsymbol{\mathcal{F}}_{i-1}\Big] \\
&\leq \quad \big\|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e*}\boldsymbol{H}_{\ell,i-1}\big\|\;\mathbb{E}\,\Big[\,\|\boldsymbol{s}_{\ell,i}^{e}(\boldsymbol{w}_{\ell,i-1})\|\,\big|\,\boldsymbol{\mathcal{F}}_{i-1}\Big] \\
&\leq \quad \sqrt{\big\|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e*}\boldsymbol{H}_{\ell,i-1}\big\|^2}\;\sqrt{\mathbb{E}\,\Big[\,\|\boldsymbol{s}_{\ell,i}^{e}(\boldsymbol{w}_{\ell,i-1})\|^2\,\big|\,\boldsymbol{\mathcal{F}}_{i-1}\Big]} \\
&\overset{(9.280)}{\leq} \quad \sqrt{a\,\big\|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\big\|^4}\;\sqrt{\mathbb{E}\,\Big[\,\|\boldsymbol{s}_{\ell,i}^{e}(\boldsymbol{w}_{\ell,i-1})\|^2\,\big|\,\boldsymbol{\mathcal{F}}_{i-1}\Big]} \\
&\overset{(8.118)}{\leq} \quad \sqrt{a\,\big\|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\big\|^4}\;\sqrt{(\beta_\ell^2/h^2)\|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\|^2+2\sigma_{s,\ell}^2} \\
&\leq \quad \sqrt{a}\,\big\|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\big\|^2\;\Big[(\beta_\ell/h)\|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\|+\sqrt{2}\sigma_{s,\ell}\Big] \\
&= \quad \frac{\sqrt{a}\beta_\ell}{h}\,\big\|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\big\|^3+\sqrt{2a\sigma_{s,\ell}^2}\,\|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\|^2 \qquad (14.42)
\end{aligned}
$$

Proof

$$\mathbb{E}\,\left|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e*}\boldsymbol{H}_{\ell,i-1}\boldsymbol{s}_{\ell,i}^{e}(\boldsymbol{w}_{\ell,i-1})\right|$$

$$\begin{aligned}
&\leq \frac{\sqrt{a}\beta_\ell}{h}\,\mathbb{E}\,\left\|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\right\|^3 + \sqrt{2a\sigma_{s,\ell}^2}\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\|^2 \\
&\leq \frac{\sqrt{a}\beta_\ell}{h}\left(\mathbb{E}\,\left\|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\right\|^4\right)^{3/4} + \sqrt{2a\sigma_{s,\ell}^2}\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e}\|^2 \\
&= \frac{\sqrt{a}\beta_\ell}{h}\left(O(\mu_{\max}^2)\right)^{3/4} + \sqrt{2a\sigma_{s,\ell}^2}\,O(\mu_{\max}) \\
&= \frac{\sqrt{a}\beta_\ell}{h}O(\mu_{\max}^{3/2}) + \sqrt{2a\sigma_{s,\ell}^2}\,O(\mu_{\max}) \\
&= O(\mu_{\max}) \qquad (14.43)
\end{aligned}$$

Proof

Using the fact that $|\mathrm{Re}(z)| \leq |z|$ for any complex number, we deduce from (14.43) that

$$\mathbb{E}\,\left|\mathrm{Re}\left[\widetilde{\boldsymbol{w}}_{\ell,i-1}^{e*}\boldsymbol{H}_{\ell,i-1}\boldsymbol{s}_{\ell,i}^{e}(\boldsymbol{w}_{\ell,i-1})\right]\right| \;=\; O(\mu_{\max}) \tag{14.44}$$

Substituting these results into (14.39) we conclude that for $i \gg 1$ we can write:

$$\begin{aligned}
\mathbb{E}\,\|\boldsymbol{\psi}_{\ell,i}^{e} - \boldsymbol{w}_{\ell,i-1}^{e}\|^2 \;&=\; \mu_\ell^2\mathrm{Tr}(G_{s,\ell}) \;+\; O\left(\mu_{\max}^{\min\{3,2+\frac{\gamma'}{2}\}}\right) \\
&\overset{(14.26)}{=}\; \gamma_\ell^2 \;+\; O\left(\mu_{\max}^{\min\{3,2+\frac{\gamma}{2}\}}\right) \\
&=\; \gamma_\ell^2 \;+\; o(\mu_{\max}^2) \qquad (14.45)
\end{aligned}$$

□

Adaptive Policy

Result (14.45) shows that, for sufficiently small step-sizes, if we can approximate the limiting value of the variance that appears on the left-hand side of (14.36), after sufficient iterations have elapsed, then we would be able to estimate the desired factor γ_ℓ^2. We can estimate this variance iteratively by using at least one of two constructions.

Agent-Centered Calculation

First, observe that

$$\mathbb{E}\left\|\boldsymbol{\psi}_{\ell,i}^{e}-\boldsymbol{w}_{\ell,i-1}^{e}\right\|^{2} \;=\; 2\,\mathbb{E}\left\|\boldsymbol{\psi}_{\ell,i}-\boldsymbol{w}_{\ell,i-1}\right\|^{2} \tag{14.46}$$

where the extended $2M\times 1$ vectors $\{\boldsymbol{\psi}_{\ell,i}^{e},\boldsymbol{w}_{\ell,i-1}^{e}\}$ are replaced by the regular $M\times 1$ vectors $\{\boldsymbol{\psi}_{\ell,i},\boldsymbol{w}_{\ell,i-1}\}$. Then, agent ℓ can estimate its variance parameter by running a smoothing filter of the following form:

$$\widehat{\gamma}_{\ell}^{2}(i)=(1-\zeta_{\ell})\,\widehat{\gamma}_{\ell}^{2}(i-1)\;+\;\zeta_{\ell}\,\|\boldsymbol{\psi}_{\ell,i}\;-\;\boldsymbol{w}_{\ell,i-1}\|^{2} \tag{14.47}$$

Agent-Centered Calculation

where the quantities $\{\boldsymbol{\psi}_{\ell,i}, \boldsymbol{w}_{\ell,i-1}\}$ that are needed to run the recursion are available at agent k. In this recursion, the notation $\widehat{\boldsymbol{\gamma}}_{\ell}^{2}(i)$ denotes the estimator for γ_{ℓ}^{2} that is computed by agent ℓ at iteration i. Moreover, $0 < \zeta_{\ell} \ll 1$ is a positive scalar much smaller than one. Note that under expectation, expression (14.47) gives

$$\mathbb{E}\,\widehat{\boldsymbol{\gamma}}_{\ell}^{2}(i) = (1-\zeta_{\ell})\,\mathbb{E}\,\widehat{\boldsymbol{\gamma}}_{\ell}^{2}(i-1) \;+\; \zeta_{\ell}\,\mathbb{E}\,\|\boldsymbol{\psi}_{\ell,i} - \boldsymbol{w}_{\ell,i-1}\|^{2} \qquad (14.48)$$

so that after sufficient iterations and using (14.36):

$$\mathbb{E}\,\widehat{\boldsymbol{\gamma}}_{\ell}^{2}(i) \;\approx\; \gamma_{\ell}^{2}/2, \quad \text{for } i \gg 1 \qquad (14.49)$$

Agent-Centered Calculation

That is, the estimator $\widehat{\gamma}_\ell^2(i)$ converges on average to the desired measure γ_ℓ^2 (scaled by $1/2$); the scaling is irrelevant because it will appear in both the numerator and denominator of the expression for $a_{\ell k}^o$ in the relative-variance rule (14.27) and will therefore cancel out. Each agent ℓ can then share the estimator $\widehat{\gamma}_\ell^2(i)$ with its neighbors. That is, in this implementation, agent ℓ shares both $\boldsymbol{\psi}_{\ell,i}$ and $\widehat{\gamma}_\ell^2(i)$ with its neighbors. Using the iterates $\widehat{\gamma}_\ell^2(i)$, we can then replace the relative-variance weights (14.27) by their adaptive counterparts and write:

Agent-Centered Calculation

$$\boldsymbol{a}_{\ell k}^{o}(i) \;=\; \frac{1}{\widehat{\gamma}_{\ell}^{2}(i)}\left(\sum_{m\in\mathcal{N}_k}\frac{1}{\widehat{\gamma}_{m}^{2}(i)}\right)^{-1}, \quad \ell\in\mathcal{N}_k \qquad (14.50)$$

Equations (14.47) and (14.50) provide one adaptive construction for the relative-variance combination weights $\{a_{\ell k}^{o}\}$. These adaptive weights would be used in (14.35) to evaluate $\boldsymbol{w}_{k,i}$, and the process continues. The above procedure is valid for both real and complex data.

Agent-Centered Calculation

Adaptive relative-variance rule (agent-centered)
(individual costs have a common minimizer)

for each time instant $i \geq 0$ repeat:
  for each neighbor ℓ of agent $k = 1, 2, \ldots, N$ do :

$$\boldsymbol{y}_{\ell,i} \stackrel{\Delta}{=} \boldsymbol{\psi}_{\ell,i} - \boldsymbol{w}_{\ell,i-1} \quad \text{(ATC diffusion)}$$

$$\widehat{\gamma}_{\ell}^{2}(i) = (1-\zeta_{\ell})\,\widehat{\gamma}_{\ell}^{2}(i-1) + \zeta_{\ell}\,\|\boldsymbol{y}_{\ell,i}\|^{2}$$

$$\boldsymbol{a}_{\ell k}^{o}(i) = \frac{1}{\widehat{\gamma}_{\ell}^{2}(i)}\left(\sum_{m\in\mathcal{N}_k}\frac{1}{\widehat{\gamma}_{m}^{2}(i)}\right)^{-1}, \quad \ell \in \mathcal{N}_k$$

  end
end

Neighbor-Centered Calculation

There is an alternative implementation where we move the estimation of the parameter γ_ℓ^2 into the neighbors of agent ℓ; this mode of operation removes the need for transmitting $\widehat{\gamma}_\ell^2(i)$ from agent ℓ to its neighbors. This advantage, however, comes at the expense of added computations as follows. Note that agent k now only has access to the iterate $\boldsymbol{\psi}_{\ell,i}$ that it receives from its neighbor ℓ. Agent k does not have access to $\boldsymbol{w}_{\ell,i-1}$ in the ATC diffusion implementation. To overcome this difficulty, we can, for example, replace $\boldsymbol{w}_{\ell,i-1}$ by $\boldsymbol{w}_{k,i-1}$ since for $i \gg 1$, the iterates at the various agents approach $w^\star$ within $O(\mu_{\max})$ with high probability and, hence,

Neighbor-Centered Calculation

$$\mathbb{E}\,\left\|\boldsymbol{\psi}_{\ell,i}-\boldsymbol{w}_{\ell,i-1}\right\|^2 \;\approx\; \mathbb{E}\,\left\|\boldsymbol{\psi}_{\ell,i}-\boldsymbol{w}_{k,i-1}\right\|^2 \qquad (14.52)$$

With this substitution, agent k can now estimate the variance γ_ℓ^2 of its neighbor locally by running a smoothing filter of the following form:

$$\boldsymbol{\gamma}_{\ell k}^2(i) = (1-\zeta_k)\,\boldsymbol{\gamma}_{\ell k}^2(i-1) \;+\; \zeta_k\,\|\boldsymbol{\psi}_{\ell,i} \;-\; \boldsymbol{w}_{k,i-1}\|^2 \qquad (14.53)$$

Neighbor-Centered Calculation

where the quantities $\{\boldsymbol{\psi}_{\ell,i}, \boldsymbol{w}_{k,i-1}\}$ that are needed to run the recursion are available at agent k. In this recursion, we are employing the notation $\boldsymbol{\gamma}^2_{\ell k}(i)$, with two subscripts, to denote the estimator for γ^2_ℓ that is computed by agent k at iteration i. Thus, observe that now several estimators for the same quantity γ^2_ℓ are being computed: one by each neighbor of agent ℓ. Again, under expectation, expression (14.53) gives

$$\mathbb{E}\,\boldsymbol{\gamma}^2_{\ell k}(i) = (1-\zeta_k)\,\mathbb{E}\,\boldsymbol{\gamma}^2_{\ell k}(i-1) \;+\; \zeta_k\,\mathbb{E}\,\|\boldsymbol{\psi}_{\ell,i} - \boldsymbol{w}_{k,i-1}\|^2 \qquad (14.54)$$

Neighbor-Centered Calculation

so that, again, after sufficient iterations and using (14.36):

$$\mathbb{E}\boldsymbol{\gamma}_{\ell k}^2(i) \;\approx\; \gamma_\ell^2/2, \quad \text{for } i \gg 1 \tag{14.55}$$

That is, the estimator $\boldsymbol{\gamma}_{\ell k}^2(i)$ converges on average to the desired measure γ_ℓ^2 (scaled by $1/2$); the scaling is again irrelevant. Using the iterates $\boldsymbol{\gamma}_{\ell k}^2(i)$, we can replace the relative-variance weights (14.27) by their adaptive counterparts and write:

$$\boldsymbol{a}_{\ell k}^o(i) \;=\; \frac{1}{\boldsymbol{\gamma}_{\ell k}^2(i)} \left(\sum_{m \in \mathcal{N}_k} \frac{1}{\boldsymbol{\gamma}_{mk}^2(i)} \right)^{-1}, \quad \ell \in \mathcal{N}_k \tag{14.56}$$

Neighbor-Centered Calculation

Adaptive relative-variance rule (neighbor-centered)
(individual costs have a common minimizer)

for each time instant $i \geq 0$ repeat:
  for each neighbor ℓ of agent $k = 1, 2, \ldots, N$ do :

$$\boldsymbol{y}_{\ell k,i} \stackrel{\Delta}{=} \boldsymbol{\psi}_{\ell,i} - \boldsymbol{w}_{k,i-1} \quad \text{(ATC diffusion)}$$

$$\gamma_{\ell k}^2(i) = (1-\zeta_k)\,\gamma_{\ell k}^2(i-1) + \zeta_k \,\|\boldsymbol{y}_{\ell k,i}\|^2$$

$$a_{\ell k}^o(i) = \frac{1}{\gamma_{\ell k}^2(i)} \left(\sum_{m \in \mathcal{N}_k} \frac{1}{\gamma_{mk}^2(i)} \right)^{-1}, \quad \ell \in \mathcal{N}_k$$

  end
end

Example #14.3

Example 14.3 (Detecting intruders and agent clustering). The following example is extracted from [214]. Allowing diffusion networks to adjust their combination coefficients in real-time enables the agents to assign smaller or larger weights to their neighbors depending on how well they contribute to the inference task. This capability can be exploited by the network to exclude harmful neighbors (such as intruders) [273]. For example, over MSE networks, the ATC diffusion strategy (7.23) with the adaptive combination weights (14.57) will take the following form.

Example #14.3

ATC diffusion with adaptive combination weights

set $\gamma_{\ell k}^2(-1) = 0$ for all $k = 1, 2, \ldots, N$ and $\ell \in \mathcal{N}_k$.

for $i \geq 0$ and for every agent k do :

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} + \frac{2\mu}{h}\boldsymbol{u}_{k,i}^{*}\left[\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{k,i-1}\right]$$

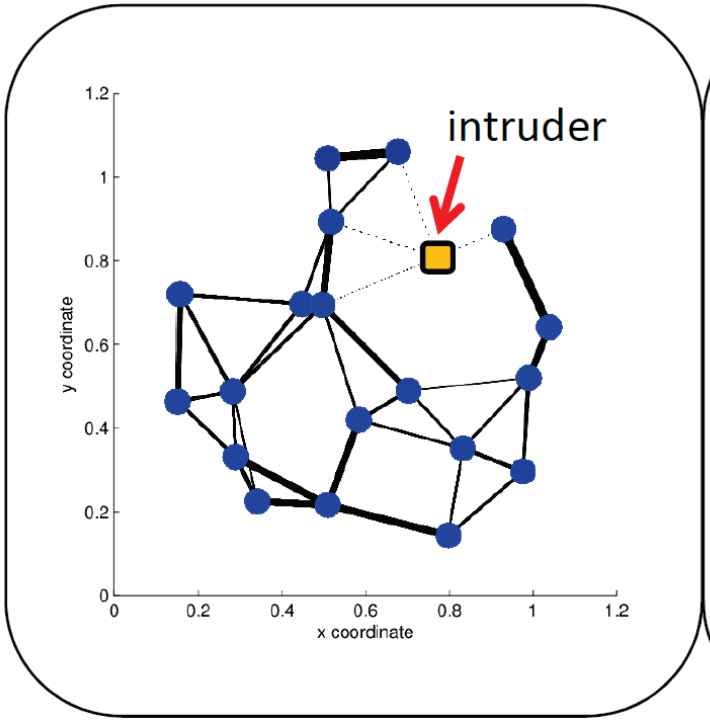
$$\gamma_{\ell k}^2(i) = (1-\zeta)\,\gamma_{\ell k}^2(i-1) + \zeta\,\|\boldsymbol{\psi}_{\ell,i} - \boldsymbol{w}_{k,i-1}\|^2, \quad \ell \in \mathcal{N}_k$$

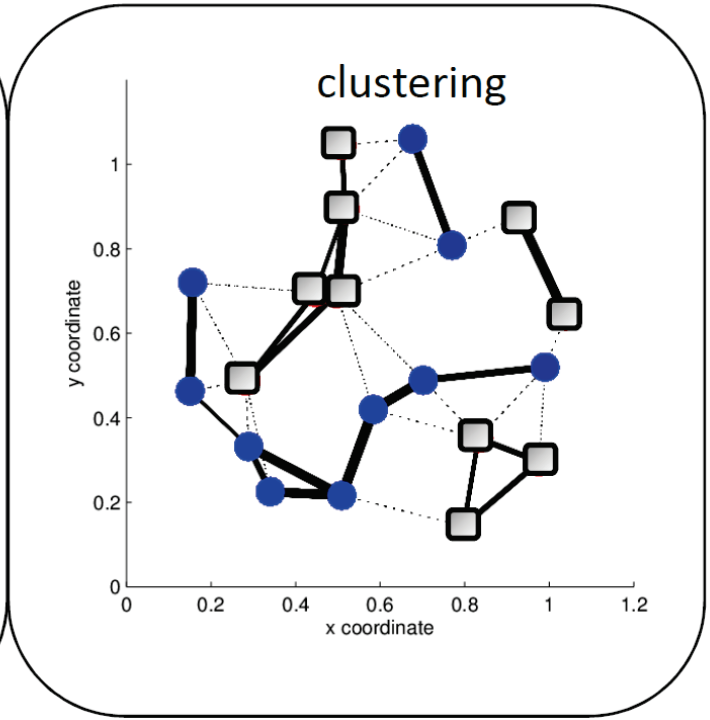
$$\boldsymbol{a}_{\ell k}(i) = \frac{\gamma_{\ell k}^{-2}(i)}{\sum_{m\in\mathcal{N}_k}\gamma_{mk}^{-2}(i)}, \quad \ell \in \mathcal{N}_k$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell\in\mathcal{N}_k}\boldsymbol{a}_{\ell k}(i)\boldsymbol{\psi}_{\ell,i}$$

end

Example #14.3


intruder
y coordinate
x coordinate
0
0.2
0.4
0.6
0.8
1
1.2


clustering
y coordinate
x coordinate
0
0.2
0.4
0.6
0.8
1
1.2

Example #14.3

Figure 14.1 illustrates the ability of networks running algorithm (14.58) to detect intrusion, and also to perform agent clustering. The figure shows a network with $N = 20$ agents. One of the agents, say, agent ℓ_o, is an intruder and it feeds its neighbors irrelevant data such as sending them wrong iterates $\boldsymbol{\psi}_{\ell_o,i}$. In some other applications, agent ℓ_o may not be an intruder but is simply subject to measurements $\{\boldsymbol{d}_{\ell_o}, \boldsymbol{u}_{\ell_o,i}\}$ that arise from a different model, $w^{\blacktriangle}$, than the model w^o. The figure on the left shows the state of the combination weights after 300 diffusion iterations: the thickness of the edges reflect the size of the combination weights assigned to them; thicker edges correspond to larger weights. Observe how the edges connecting to the intruder

Example #14.3

are essentially cut-off by the algorithm. The figure on the right illustrates the ability of diffusion strategies to perform agent clustering (i.e., to separate into groups agents that are influenced by two different models, $w^{\blacktriangle}$ and w^{o}). Agents do not know beforehand which of their neighbors are influenced by which model. They also do not know which model is influencing their own data. By allowing agents to adapt their combination coefficients on the fly, it becomes possible for the agents to cut their links over time to neighbors that are sensing a different model than their own. The net effect is that agents end up being clustered in two groups. Cooperation between the members of the same group then leads to the estimation of $\{w^{\blacktriangle}, w^{o}\}$.

Example #14.4

Example 14.4 (Adapting combination weights over MSE networks). We illustrate the performance of adaptive combination rules over MSE networks of the form described earlier in Example 6.3. We employ uniform step-sizes across the agents, $\mu_k = \mu = 0.001$. Figure 14.2 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 left and right plots of Figure 14.3, respectively.

Example #14.4

Figure 14.4 plots the evolution of the ensemble-average learning curves, $\frac{1}{N}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$, for the ATC diffusion strategy (14.58) using four different combination rules: the left-stochastic uniform or averaging rule (11.148), the doubly-stochastic Metropolis rule (12.43), the relative-variance rule (14.31), and the adaptive combination rule (14.58) with uniform $\zeta_k = \zeta = 0.01$. The curves are obtained by averaging the trajectories $\{\frac{1}{N}\|\widetilde{\boldsymbol{w}}_i\|^2\}$ over 100 repeated experiments. The label on the vertical axis in the figure refers to the learning curves $\frac{1}{N}\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ by writing $\mathrm{MSD}_{\mathrm{dist,av}}(i)$, with an iteration index i. Each experiment involves running the diffusion strategy 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 #14.4

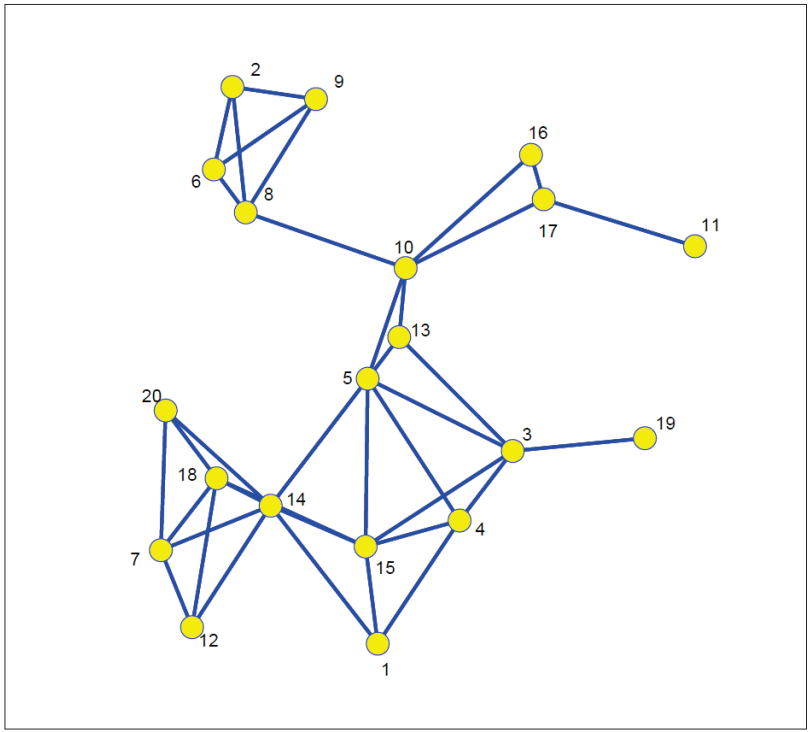


Figure 14.2

Example #14.4

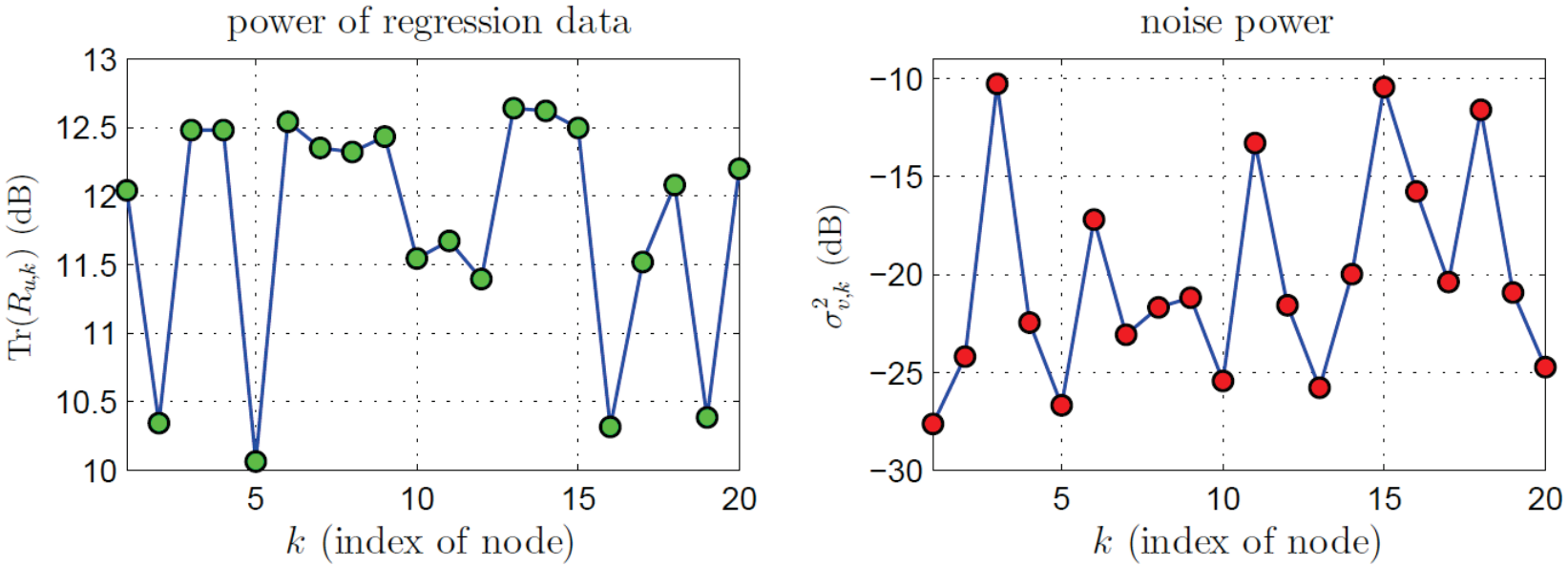


Figure 14.3

Example #14.4

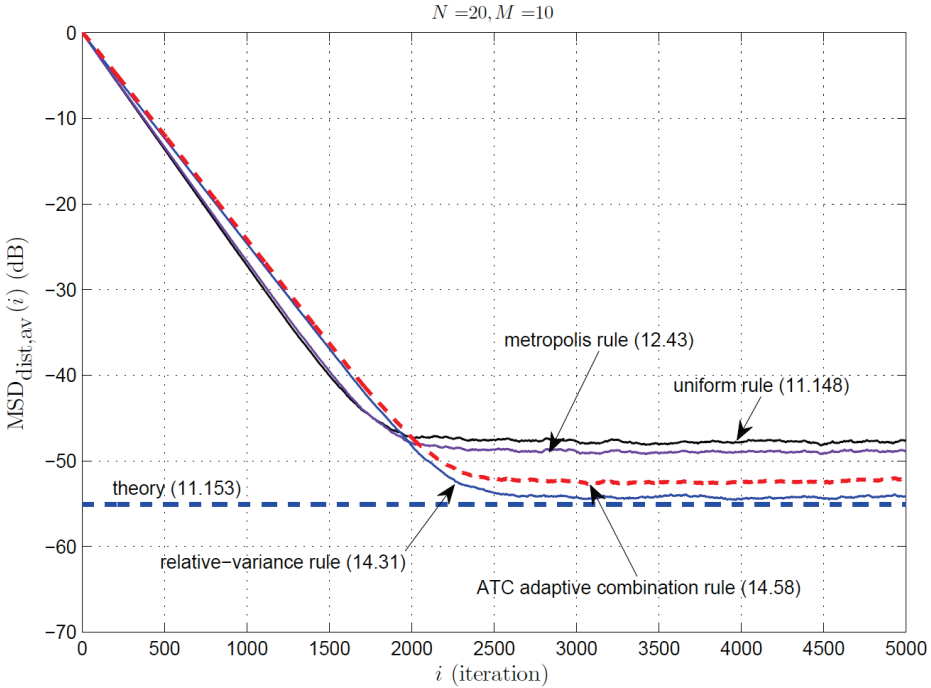


Figure 14.4

Example #14.4

It is further observed in the figure that the learning curve of the relative-variance rule tends to the MSD value predicted by the theoretical expression (11.153) with the entries $\{p_k\}$ corresponding to the Perron eigenvector that is associated with the combination policy (14.31), which reduces to the following expression in the example under consideration:

$$a_{\ell k}^{o} = \frac{1}{\sigma_{v,\ell}^2 \sigma_{u,\ell}^2} \left(\sum_{m \in \mathcal{N}_k} \frac{1}{\sigma_{v,m}^2 \sigma_{u,m}^2} \right)^{-1}, \quad \ell \in \mathcal{N}_k \tag{14.59}$$

Example #14.4

It is also observed from Figure 14.4 that the adaptive rule is able to learn the noise factors $\{\gamma_\ell^2\}$ and to attain a performance level that is expected from the relative-variance rule. However, the convergence rate of the adaptive rule is clearly slower than the uniform and Metropolis rules: this is because of the additional adaptation process that is involved in learning the noise factors $\{\gamma_\ell^2\}$ and the combination coefficients $\{\boldsymbol{a}_{\ell k}(i)\}$. Schemes for speeding up the

Example #14.4

convergence of the adaptive combination rule are proposed in [270] and [95]. One idea is based on training the network initially by using a static rule, such as the uniform rule, while the combination weights are being adapted and subsequently switch to the adaptive combination rule. Criteria for selecting the switching time is developed in these references. Figure 14.5 illustrates this construction where the switching time occurs at $i = 1000$. It is seen that the adaptive combination rule is able to recover the faster convergence rate of the uniform rule.

■

Example #14.4

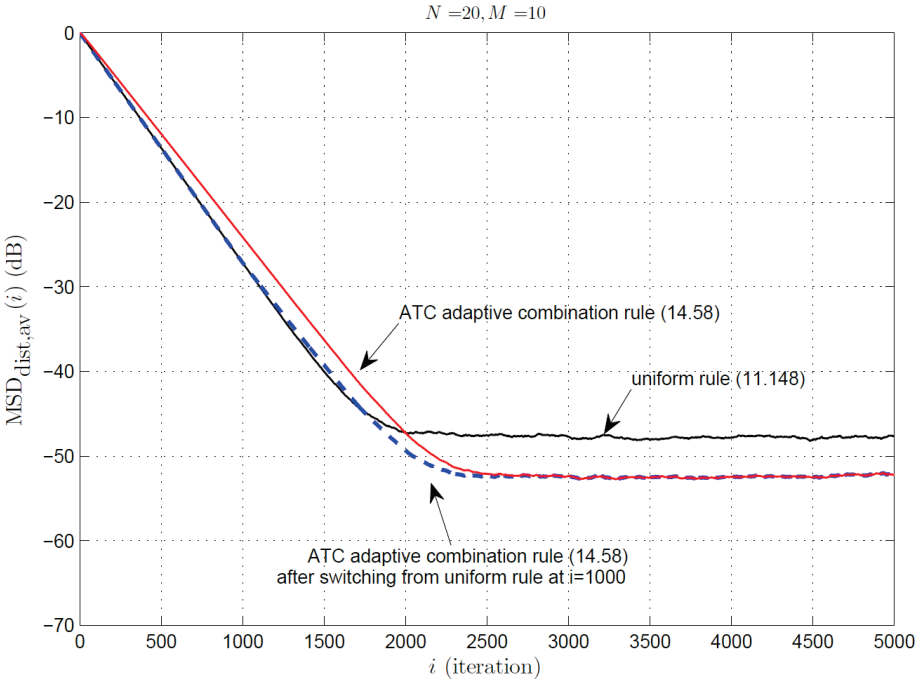


Figure 14.5

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.