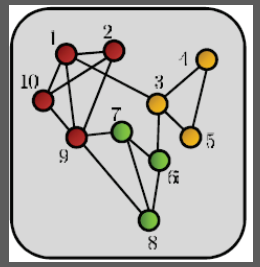

INFERENCE OVER NETWORKS

LECTURE #13: Centralized Adaptation & Learning

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.

Part III: Centralized Adaptation And Learning

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 5 (Centralized Adaptation & Learning, pp. 368-407-430):

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.

Performance Metrics

Theorem 4.7: For small-enough step-sizes and real data:

$$H \triangleq \nabla_w^2 J(w^o)$$

$$\begin{aligned} \text{MSD} &= \frac{\mu}{2}\,\text{Tr}\left(H^{-1}R_s\right) \\ \text{ER} &= \frac{\mu}{4}\,\text{Tr}\left(R_s\right) \end{aligned}$$

The rate at which $\mathbb{E}\,\|\widetilde{w}_i\|^2$ approaches its steady-state region is approximated to first-order in μ by:

$$\alpha = 1 - 2\mu\lambda_{\min}(H)$$

Recall #1 (Example #3.1)

Example 3.1 (LMS adaptation). Let $\boldsymbol{d}(i)$ denote a streaming sequence of zero-mean random variables with variance $\sigma_d^2 = \mathbb{E}\,\boldsymbol{d}^2(i)$. Let $\boldsymbol{u}_i$ denote a streaming sequence of $1 \times M$ independent zero-mean random vectors with covariance matrix $R_u = \mathbb{E}\,\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i > 0$. Both processes $\{\boldsymbol{d}(i), \boldsymbol{u}_i\}$ are assumed to be jointly wide-sense stationary. The cross-covariance vector between $\boldsymbol{d}(i)$ and $\boldsymbol{u}_i$ is denoted by $r_{du} = \mathbb{E}\,\boldsymbol{d}(i)\boldsymbol{u}_i^{\mathsf{T}}$. The data $\{\boldsymbol{d}(i), \boldsymbol{u}_i\}$ are assumed to be related via a linear regression model of the form:

$$\boldsymbol{d}(i) \;=\; \boldsymbol{u}_i w^o \;+\; \boldsymbol{v}(i) \tag{3.6}$$

Recall #1 (Example #3.1)

for some unknown parameter vector w^o, and where $\boldsymbol{v}(i)$ is a zero-mean white-noise process with power $\sigma_v^2 = \mathbb{E}\,\boldsymbol{v}^2(i)$ and assumed independent of $\boldsymbol{u}_j$ for all i, j. Observe that we are using parentheses to represent the time-dependency of a scalar variable, such as writing $\boldsymbol{d}(i)$, and subscripts to represent the time-dependency of a vector variable, such as writing $\boldsymbol{u}_i$. This convention will be used throughout this work. In a manner similar to Example 2.1, we again pose the problem of estimating w^o by minimizing the mean-square error cost

$$J(w) = \mathbb{E}\,(\boldsymbol{d}(i) - \boldsymbol{u}_i w)^2 \;\equiv\; \mathbb{E}\,Q(w; \boldsymbol{x}_i) \tag{3.7}$$

Recall #1 (Example #3.1)

$$r_{du} \approx \boldsymbol{d}(i)\boldsymbol{u}_i^{\mathsf{T}}, \quad R_u \approx \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i \tag{3.9}$$

By doing so, the true gradient vector is approximated by:

$$\widehat{\nabla_{w^{\mathsf{T}}} J(w)} = 2\left[\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i w - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{d}(i)\right] = \nabla_{w^{\mathsf{T}}} Q(w;\boldsymbol{x}_i) \tag{3.10}$$

Substituting (3.10) into (3.8) leads to the well-known least-mean-squares (LMS, for short) algorithm [107, 206, 262]:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} + 2\mu\boldsymbol{u}_i^{\mathsf{T}}[\boldsymbol{d}(i) - \boldsymbol{u}_i\boldsymbol{w}_{i-1}], \quad i \geq 0 \tag{3.13}$$

Recall #2 (Example #3.3)

Example 3.3 (Gradient noise). It is clear from the expressions in Examples 2.3 and 3.1 that the corresponding gradient noise process is given by:

$$\begin{aligned}
\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) &= \widehat{\nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}_{i-1})} - \nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}_{i-1}) \\
&= 2\left(\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i\right)\boldsymbol{w}_{i-1} - 2\boldsymbol{u}_i^{\mathsf{T}}[\boldsymbol{u}_i w^o + \boldsymbol{v}(i)] - 2R_u\boldsymbol{w}_{i-1} + 2R_u w^o \\
&= 2(R_u - \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i)\widetilde{\boldsymbol{w}}_{i-1} - 2\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{v}(i) \qquad (3.19)
\end{aligned}$$

$\longrightarrow$ $H = 2R_u$ and $R_s = 4\sigma_v^2 R_u$.

Example #4.3

Example 4.3 (Performance of LMS adaptation). We reconsider the LMS recursion (3.13). We know from Example 3.3 and (4.13) that this situation corresponds to $H = 2R_u$ and $R_s = 4\sigma_v^2 R_u$. Substituting into (4.100)–(4.101) leads to the following well-known expressions for the performance of the LMS filter for sufficiently small step-sizes — see [96, 97, 100, 107, 114, 130, 206, 261, 262]:

$$\text{MSD} = \mu M \sigma_v^2 = O(\mu) \tag{4.146}$$

$$\text{EMSE} = \mu \sigma_v^2 \,\text{Tr}(R_u) = O(\mu) \tag{4.147}$$

where we are replacing ER by the notation EMSE, which is more common in the adaptive filtering literature.

Setting

The discussion in the last two chapters established the mean-square stability of stand-alone adaptive agents for small constant step-sizes (Lemmas 3.1 and 3.5), and provided expressions for their MSD and ER metrics (Theorems 4.7 and 4.8) for both cases of real and complex-valued data. In this chapter, and in preparation for our treatment of networked agents in future chapters, we examine two situations involving a *multitude* of similar agents behaving in one of two modes [207].

Setting

In the first scenario, each agent senses data and analyzes it independently of the other agents. We refer to this mode of operation as *non-cooperative* processing. In the second scenario, the agents transmit the collected data for processing at a fusion center. We refer to this mode of operation as *centralized* or batch processing. We motivate the discussion by considering first the case of mean-square-error costs. Subsequently, we extend the results to more general costs.

Non-Cooperative Processing

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.

Non-Cooperative Processing

Thus, consider *separate* agents, labeled $k = 1, 2, \ldots, N$. Following the framework discussed in Examples 3.1 and 3.4 on LMS adaptation in the real and complex domains, each agent, k, receives streaming data $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i},\ i \geq 0\}$, where we are using the subscript k to index the data at agent k. We treat the real and complex data cases uniformly by using the data-type variable in the expressions that follow:

$$h \triangleq \begin{cases} 1 & \text{(real data)} \\ 2 & \text{(complex data)} \end{cases} \tag{5.1}$$

Non-Cooperative Processing

We assume the data at each agent satisfies the same statistical properties as in Examples 3.1 and 3.4, and the same linear regression model (3.119) with a common w^o albeit with noise $\boldsymbol{v}_k(i)$:

$$\boldsymbol{d}_k(i) \;=\; \boldsymbol{u}_{k,i}w^o + \boldsymbol{v}_k(i), \qquad k = 1, 2, \ldots, N \tag{5.2}$$

We denote the statistical moments of the data at agent k by

$$\sigma_{v,k}^2 = \mathbb{E}\,|\boldsymbol{v}_k(i)|^2 \tag{5.3}$$

and

$$R_{u,k} \;\triangleq\; \begin{cases} \mathbb{E}\,\boldsymbol{u}_{k,i}^{\mathsf{T}}\boldsymbol{u}_{k,i} > 0 & \text{(real data)} \\ \mathbb{E}\,\boldsymbol{u}_{k,i}^{*}\boldsymbol{u}_{k,i} > 0 & \text{(complex data)} \end{cases} \tag{5.4}$$

Non-Cooperative Processing

We further assume in this motivating section that the $R_{u,k}$ are uniform across the agents so that

$$R_{u,k} \equiv R_u, \qquad k = 1, 2, \ldots, N \tag{5.5}$$

In this way, the mean-square-error cost,

$$J_k(w) \triangleq \mathbb{E}\,|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}|^2 \tag{5.6}$$

which is associated with agent k, will satisfy a condition similar to (3.114), namely,

$$0 < \frac{\nu}{h} I_{hM} \leq \nabla_w^2 J_k(w) \leq \frac{\delta}{h} I_{hM} \tag{5.7}$$

Non-Cooperative Processing

with the corresponding parameters $\{\nu, \delta\}$ given by (cf. (2.19)):

$$\nu = 2\lambda_{\min}(R_u), \quad \delta = 2\lambda_{\max}(R_u) \tag{5.8}$$

Now, assume each agent estimates w^o by running the LMS learning rule, say, (3.13) for real data or (3.125) for complex data, which we can describe uniformly in terms of the single recursion:

$$\boldsymbol{w}_{k,i} = \boldsymbol{w}_{k,i-1} + \frac{2\mu}{h}\boldsymbol{u}_{k,i}^*[\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{k,i-1}], \quad i \geq 0 \tag{5.9}$$

Non-Cooperative Processing

using the data-type variable, h, and with the understanding that complex conjugation, $\boldsymbol{u}_{k,i}^{*}$, is replaced by real transposition, $\boldsymbol{u}_{k,i}^{\mathsf{T}}$, when the data are real. Then, according to (4.146) and (4.186), each agent k will attain an individual MSD level that is given by

$$\mathrm{MSD}_{\mathrm{ncop},k} \;=\; \frac{\mu}{h} M\,\sigma_{v,k}^{2}, \qquad k=1,2,\ldots,N \tag{5.10}$$

Moreover, according to (3.38) and (3.142), each agent k will converge towards this level at a rate dictated by:

$$\alpha_{\mathrm{ncop},k} \;=\; 1-\frac{4\mu}{h}\lambda_{\min}(R_u) \tag{5.11}$$

Non-Cooperative Processing

***N* non-cooperative LMS agents** labeled $k=1,2,\ldots,N$ with uniform $R_{u,k} \equiv R_u$:

$$\mathrm{MSD}_{\mathrm{ncop},k} = \frac{\mu}{h} M \sigma_{v,k}^2, \qquad k = 1,2,\ldots,N$$

$$\alpha_{\mathrm{ncop},k} = 1 - \frac{4\mu}{h}\lambda_{\min}(R_u)$$

→ Agents with noisier data have larger MSD levels than agents with cleaner data.

Non-Cooperative Processing

If we average the performance level (5.10) across the N agents, we find that the average MSD metric is given by

$$\mathrm{MSD}_{\mathrm{ncop,av}} = \frac{\mu}{h} M \left(\frac{1}{N} \sum_{k=1}^{N} \sigma_{v,k}^2 \right) \tag{5.12}$$

in terms of the average noise power across the agents.

Non-Cooperative Processing

The subscript "ncop" is used in (5.10)–(5.12) to indicate that these expressions are for the non-cooperative mode of operation. It is seen from (5.10) that agents with noisier data (i.e., larger $\sigma_{v,k}^2$) will perform worse and have larger MSD levels than agents with cleaner data. In other words, whenever adaptive agents act individually, the quality of their solution will be as good as the quality of their noisy data.

Non-Cooperative Processing

This is a sensible conclusion and it is illustrated numerically in Figure 5.1. The figure plots the ensemble-average learning curves, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i}\|^2$, for two agents. The curves are generated by averaging the trajectories $\{\|\widetilde{\boldsymbol{w}}_{k,i}\|^2\}$ over 2000 repeated experiments. The label on the vertical axis in the figure refers to the learning curves by writing MSD(i), with an iteration index i. Each experiment involves running the non-cooperative LMS recursion (5.9) on complex-valued data

Non-Cooperative Processing

$\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$ generated according to the model $\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i}w^o + \boldsymbol{v}_k(i)$ with $M = 10$, $R_u = 2I_M$, and $\mu = 0.005$. The noise variances are set to $\sigma^2_{v,1} = 0.032$ and $\sigma^2_{v,2} = 0.010$. The noise and regressor processes are both Gaussian distributed in this simulation. The unknown vector w^o is generated randomly and its norm is normalized to one. It is seen in the figure that the learning curves by the agents tend to the MSD levels predicted by the theoretical expression (5.10).

Non-Cooperative Processing

We are going to show in later chapters that cooperation among agents, whereby agents share information with their neighbors, can help enhance their individual performance levels. The analysis will show that both types of agents can benefit from cooperation: agents with "bad" data and agents with "good" data; this is because all data carry information about w^o. However, for these conclusions to hold, it is necessary for cooperation to be carried out in proper ways — see Chapter 12.

Non-Cooperative Processing

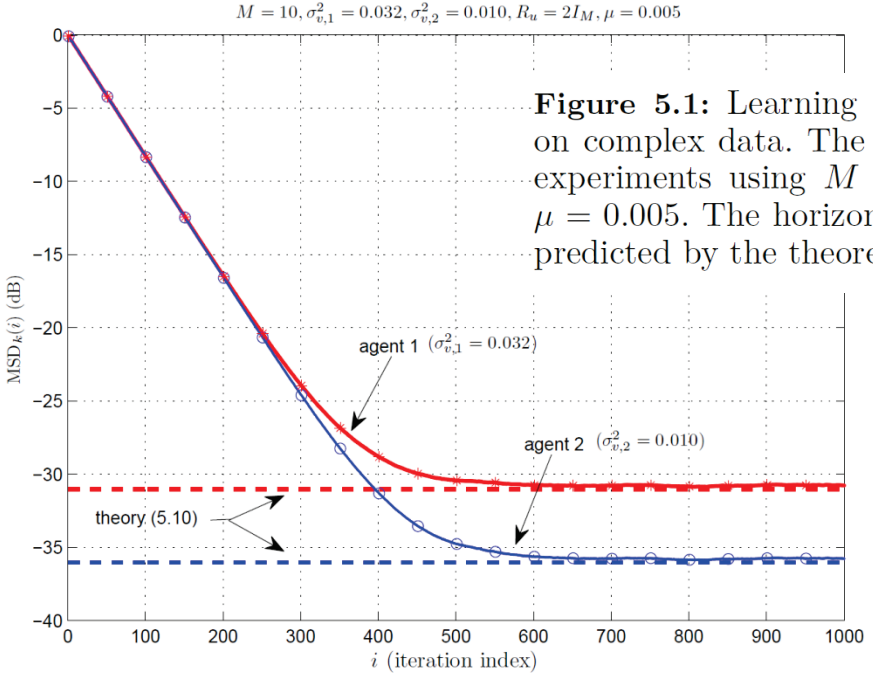

Figure 5.1: Learning curves for two non-cooperative agents running (5.9) on complex data. The curves are obtained by averaging over 2000 repeated experiments using $M = 10$, $\sigma^2_{v,1} = 0.032$, $\sigma^2_{v,2} = 0.010$, $R_u = 2I_M$ and $\mu = 0.005$. The horizontal dashed lines indicate the steady-state MSD levels predicted by the theoretical expression (5.10) for complex data ($h = 2$).

Centralized Processing

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.

Centralized Processing

Let us now contrast the above non-cooperative solution with a centralized implementation whereby, at every iteration i, the N agents transmit their raw data $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$ to a fusion center for processing. One could also consider situations where agents transmit processed data, e.g., as happens with useful techniques for combining adaptive filter outputs [10]. Once the fusion center receives the raw data, we assume it runs a stochastic-gradient update of the form:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} + \mu \left(\frac{1}{N} \sum_{k=1}^{N} \frac{2}{h} \boldsymbol{u}_{k,i}^* (\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i} \boldsymbol{w}_{i-1}) \right) \tag{5.13}$$

Centralized Processing

where the term between parentheses multiplying μ can be interpreted as corresponding to the sample average of several approximate gradient vectors; one for the data originating from each agent, since

$$\widehat{\nabla_{w^{\mathsf{T}}} J_k}(\boldsymbol{w}_{i-1}) = 2\boldsymbol{u}_{k,i}^{\mathsf{T}}(\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{i-1}) \quad \text{(real data)} \tag{5.14}$$

and

$$\widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{i-1}) = \boldsymbol{u}_{k,i}^{*}(\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}\boldsymbol{w}_{i-1}) \quad \text{(complex data)} \tag{5.15}$$

Centralized Processing

The analysis in the sequel will show that the MSD performance that results from implementation (5.13) is given by (using future expression (5.65) with the identifications $H_k = 2R_u/h$ and $R_{s,k} = 4\sigma_{v,k}^2 R_u/h^2$):

$$\text{MSD}_{\text{cent}} \;=\; \frac{\mu}{h} M \frac{1}{N} \left(\frac{1}{N} \sum_{k=1}^{N} \sigma_{v,k}^2 \right) \tag{5.16}$$

Moreover, using expression (5.60) given further ahead, this centralized solution will converge towards the above MSD level at the same rate (5.11) as the non-cooperative solution:

$$\alpha_{\text{cent}} \;=\; 1 - \frac{4\mu}{h} \lambda_{\min}(R_u) \tag{5.17}$$

Fusion Center

All ***N agents*** transfer data to a fusion center:

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

$$\begin{cases} \mathrm{MSD}_{\mathrm{cent}} = \dfrac{\mu}{h} M \dfrac{1}{N} \left(\dfrac{1}{N} \displaystyle\sum_{k=1}^{N} \sigma_{v,k}^2 \right) \\ \alpha_{\mathrm{cent}} = 1 - \dfrac{4\mu}{h} \lambda_{\min}(R_u) \end{cases}$$

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

Centralized Processing

Observe from (5.16) that the MSD level attained by the centralized solution is proportional to $1/N$ times the *average* noise power across all non-cooperative agents in (5.10). At least two conclusions follow from this observation.

First, comparing (5.16) with the average performance (5.12) in the non-cooperative case, we observe that the centralized solution provides an $N-$fold improvement in MSD performance in the mean-square-error case. Figure 5.2 illustrates this situation numerically.

Centralized Processing

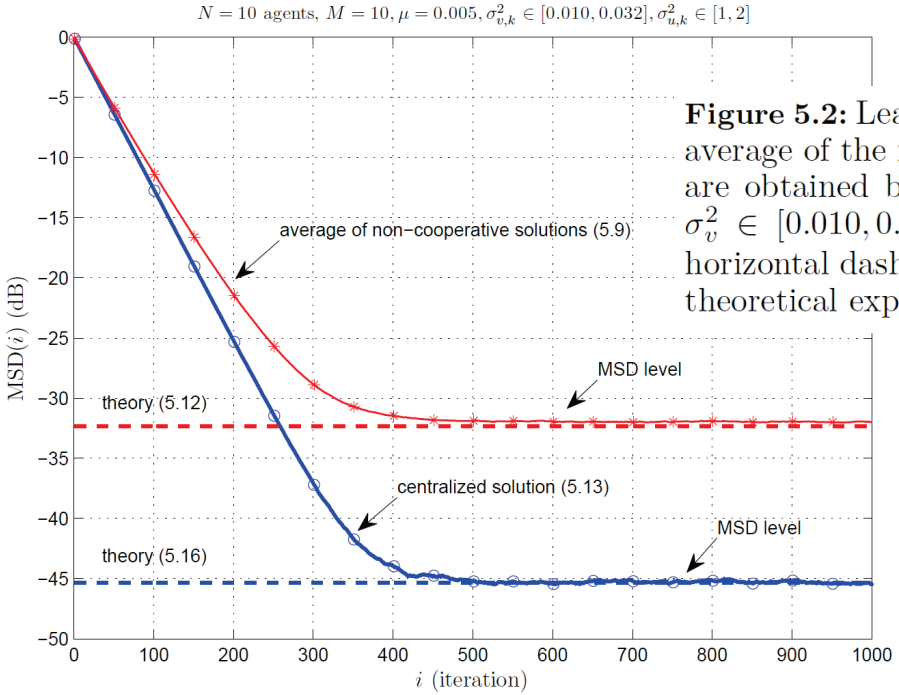

Figure 5.2: Learning curves for the centralized LMS solution (5.13) and for the average of the non-cooperative solution (5.9) over $N = 20$ agents. The curves are obtained by averaging over 2000 repeated experiments using $M = 10$, $\sigma_v^2 \in [0.010, 0.032]$, $R_u = \sigma_{u,k}^2 I_M$ with $\sigma_{u,k}^2 \in [1, 2]$, and $\mu = 0.005$. The horizontal dashed lines indicate the steady-state MSD levels predicted by the theoretical expressions (5.12) and (5.16) for complex data ($h = 2$).

Centralized Processing

The figure plots two ensemble-average learning curves. One curve represents the evolution of the variance $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ for the centralized solution and is generated by averaging the trajectories $\{\|\widetilde{\boldsymbol{w}}_i\|^2\}$ over 200 repeated experiments. The second ensemble-average curve is obtained by averaging the individual learning curves, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{k,i}\|^2$, of all N non-cooperative agents. Again, a total of 2000 repeated experiments are used to generate each individual learning curve. The label on the vertical axis in the figure refers to the learning curves by writing $\mathrm{MSD}(i)$, with an iteration index i. Each experiment involves running either the centralized LMS recursion (5.13) or the non-cooperative recursion (5.9)

Centralized Processing

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 $N = 20$ agents, $M = 10$, and $\mu = 0.005$. The noise variances, $\{\sigma^2_{v,k}\}$, are chosen randomly from within the range $[0.010, 0.032]$, while the covariance matrices are chosen of the form $R_{u,k} = \sigma^2_{u,k} I_M$ with $\sigma^2_{u,k}$ chosen randomly within the range $[1, 2]$. The noise and regressor processes are both Gaussian distributed in this simulation. The unknown vector w^o is generated randomly and its norm is normalized to one. It is seen in the figure that the learning curve by the centralized solution tends to an MSD level that is $N-$fold superior to the average non-cooperative solution; this translates into the difference of $10\log_{10}(N) \approx 13$dB seen in the figure between the two dashed horizontal lines.

Centralized Processing

The second observation that follows from (5.16) is that, although the centralized solution outperforms the averaged non-cooperative performance, it does not generally hold that the centralized solution outperforms each individual non-cooperative agent [276]. This is because the average noise power is scaled by $1/N$ in (5.16), 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,

Centralized Processing

$$\frac{1}{N}\left(\frac{1}{N}\sum_{k=1}^{N}\sigma_{v,k}^2\right) \;=\; 1.5\sigma_v^2 \tag{5.18}$$

which is larger than $\sigma_{v,1}^2$ and smaller than $\sigma_{v,2}^2$. In this case, the centralized solution (5.16) performs better than non-cooperative agent 2 (i.e., leads to a smaller MSD) but worse than non-cooperative agent 1.

Stochastic-Gradient Centralized Solution

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.

Aggregate Cost

The last two sections focused on mean-square-error adaptation. Next, we extend the conclusions to more general costs. Thus, consider a collection of N agents, each with an individual twice-differentiable convex cost function, $J_k(w)$. The objective is to determine the unique minimizer w^o of the aggregate cost:

$$J^{\text{glob}}(w) \triangleq \sum_{k=1}^{N} J_k(w) \tag{5.19}$$

Stochastic-Gradient Alg.

It is now the above aggregate cost, $J^{\text{glob}}(w)$, that will be required to satisfy conditions similar to (4.4) and (4.18) relative to some parameters $\{\nu_c, \delta_c, \kappa_c\}$, with the subscript "c" used to indicate that these factors correspond to the centralized implementation.

Conditions on Aggregate Cost

Assumption 5.1 (Conditions on aggregate cost function). The aggregate cost function, $J^{\text{glob}}(w)$, is twice-differentiable and satisfies

$$0 < \frac{\nu_c}{h} I_{hM} \;\leq\; \nabla_w^2 \, J^{\text{glob}}(w) \;\leq\; \frac{\delta_c}{h} I_{hM} \tag{5.20}$$

for some positive parameters $\nu_c \leq \delta_c$. Condition (5.20) is equivalent to requiring $J^{\text{glob}}(w)$ to be ν_c-strongly convex and for its gradient vector to be δ_c-Lipschitz. In addition, it is assumed that the aggregate cost is smooth enough so that its Hessian matrix is locally Lipschitz continuous in a small neighborhood around $w = w^o$, i.e.,

$$\left\| \nabla_w^2 \, J^{\text{glob}}(w^o + \Delta w) - \nabla_w^2 \, J^{\text{glob}}(w^o) \right\| \;\leq\; \kappa_c \, \|\Delta w\| \tag{5.21}$$

for small perturbations $\|\Delta w\| \leq \epsilon$ and for some $\kappa_c \geq 0$.

Conditions on Aggregate Cost

Under these conditions, the cost $J^{\text{glob}}(w)$ will have a unique minimizer, which we continue to denote by w^o. We will not be requiring each individual cost, $J_k(w)$, to be strongly convex. It is sufficient for at least one of these costs to be strongly convex while the remaining costs can be simply convex; this condition ensures the strong convexity of $J^{\text{glob}}(w)$. Moreover, minimizers of the individual costs $\{J_k(w)\}$ need not coincide with each other or with w^o; we shall write w_k^o to refer to a minimizer of $J_k(w)$.

Stochastic-Gradient Algorithm

There are many centralized solutions that can be used to determine the unique minimizer w^o of (5.19), with some solution techniques being more powerful than other techniques. Nevertheless, we shall focus on centralized implementations of the *stochastic gradient* type. The reason we consider the *same* class of stochastic gradient algorithms for non-cooperative, centralized, and distributed solutions in this work is to enable a *meaningful* comparison among the various implementations. Thus, we consider a centralized strategy of the following form:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \frac{\mu}{N}\sum_{k=1}^{N}\widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{i-1}), \quad i \geq 0 \tag{5.22}$$

Stochastic-Gradient Alg.

in terms of approximations for the individual gradient vectors at $\boldsymbol{w}_{i-1}$. Here, again, we will be treating the case of real and complex data jointly. For this reason, although we are computing the gradient vector relative to w^* in the above recursion, it is to be understood that this step should be replaced by differentiation relative to w^{T} in the real case; i.e., complex conjugation should be replaced by real transposition when the data are real in which case the update would take the form:

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \frac{\mu}{N}\sum_{k=1}^{N}\widehat{\nabla_{w^{\mathsf{T}}} J_k}(\boldsymbol{w}_{i-1}), \quad i \geq 0 \tag{5.23}$$

Gradient Noise Model

Continuing with the general form (5.22), we note that the sum multiplying μ/N is an approximation for the true gradient vector of $J^{\text{glob}}(w)$; the scaling of μ by N in (5.22) is meant to ensure similar convergence rates for the non-cooperative and centralized solutions — as explained further ahead in (5.78). We introduce the *individual* gradient noise processes:

$$\boldsymbol{s}_{k,i}(\boldsymbol{w}_{i-1}) \;\triangleq\; \widehat{\nabla_{w^*} J_k(\boldsymbol{w}_{i-1})} \;-\; \nabla_{w^*} J_k(\boldsymbol{w}_{i-1}) \tag{5.24}$$

for $k = 1, 2, \ldots, N$, and note that the overall gradient noise corresponding to (5.22) is given by:

Gradient Noise Model

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \;=\; \sum_{k=1}^{N} \boldsymbol{s}_{k,i}(\boldsymbol{w}_{i-1}) \tag{5.25}$$

We also introduce the covariance matrices of the individual noise processes. Specifically, for any $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$ and for every $k = 1, 2, \ldots, N$, we define the extended gradient noise vector of size $2M \times 1$:

$$\boldsymbol{s}^{e}_{k,i}(\boldsymbol{w}) \;\stackrel{\Delta}{=}\; \begin{bmatrix} \boldsymbol{s}_{k,i}(\boldsymbol{w}) \\ \left(\boldsymbol{s}^{*}_{k,i}(\boldsymbol{w})\right)^{\mathsf{T}} \end{bmatrix} \tag{5.26}$$

Gradient Noise Model

and denote its conditional covariance matrix by

$$R^{e}_{s,k,i}(\boldsymbol{w}) \triangleq \mathbb{E}\left[\boldsymbol{s}^{e}_{k,i}(\boldsymbol{w})\boldsymbol{s}^{e*}_{k,i}(\boldsymbol{w}) \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \tag{5.27}$$

We further assume that, in the limit, the following moment matrices tend to constant values when evaluated at w^o:

$$R_{s,k} \triangleq \lim_{i\to\infty} \mathbb{E}\left[\boldsymbol{s}_{k,i}(w^o)\boldsymbol{s}^{*}_{k,i}(w^o) \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \tag{5.28}$$

$$R_{q,k} \triangleq \lim_{i\to\infty} \mathbb{E}\left[\boldsymbol{s}_{k,i}(w^o)\boldsymbol{s}^{\mathsf{T}}_{k,i}(w^o) \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \tag{5.29}$$

Gradient Noise Model

We define similar quantities for the aggregate noise process (5.25) and denote them by

$$R_{s,i}^{e}(\boldsymbol{w}) \triangleq \mathbb{E}\,[\,\boldsymbol{s}_i^{e}(\boldsymbol{w})\boldsymbol{s}_i^{e*}(\boldsymbol{w})\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,] \tag{5.30}$$

$$R_s \triangleq \lim_{i\to\infty}\mathbb{E}\,[\,\boldsymbol{s}_i(w^o)\boldsymbol{s}_i^{*}(w^o)\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,] \tag{5.31}$$

$$R_q \triangleq \lim_{i\to\infty}\mathbb{E}\,\Big[\,\boldsymbol{s}_i(w^o)\boldsymbol{s}_i^{\mathsf{T}}(w^o)\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\Big] \tag{5.32}$$

Gradient Noise Model

Now since the centralized iteration (5.22) has the form of a stochastic gradient recursion, we should be able to infer its mean-square-error behavior from Lemma 3.5 and Theorem 4.8 if the aggregate noise process (5.25) satisfies conditions similar to Assumption 3.4. It is straightforward to verify that this is possible, for example, if the *individual* components satisfy conditions similar to Assumption 3.4 and condition (4.67) and when, additionally, these individual components are uncorrelated with each other and second-order circular as described by the following statement.

Gradient Noise Model

Assumption 5.2 (Conditions on gradient noise). It is assumed that the first and fourth-order conditional moments of the individual gradient noise processes, $\boldsymbol{s}_{k,i}(\boldsymbol{w})$, defined by (5.24) satisfy the following conditions for any iterates $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$ and for all $k, \ell = 1, 2, \ldots, N$:

$$\mathbb{E}\left[\,\boldsymbol{s}_{k,i}(\boldsymbol{w})\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\right] = 0 \tag{5.33}$$

$$\mathbb{E}\left[\boldsymbol{s}_{k,i}(\boldsymbol{w})\boldsymbol{s}_{\ell,i}^{*}(\boldsymbol{w})|\boldsymbol{\mathcal{F}}_{i-1}\right] = 0, \quad k \neq \ell \tag{5.34}$$

$$\mathbb{E}\left[\boldsymbol{s}_{k,i}(\boldsymbol{w})\boldsymbol{s}_{\ell,i}^{\mathsf{T}}(\boldsymbol{w})|\boldsymbol{\mathcal{F}}_{i-1}\right] = 0, \quad k \neq \ell \tag{5.35}$$

$$\mathbb{E}\left[\,\|\boldsymbol{s}_{k,i}(\boldsymbol{w})\|^{4}\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\right] \leq \left(\bar{\beta}_k/h\right)^{4}\|\boldsymbol{w}\|^{4} + \bar{\sigma}_{s,k}^{4} \tag{5.36}$$

Gradient Noise Model

almost surely, for some nonnegative scalars $\bar{\beta}_k^4$ and $\bar{\sigma}_{s,k}^4$ and where $h = 2$ for complex data and $h = 1$ for real data. We also assume that the conditional second-order moments of the aggregate noise process satisfies a smoothness condition similar to (4.166), namely,

$$\left\| R_{s,i}^e(w^o + \Delta w) - R_{s,i}^e(w^o) \right\| \leq \kappa_{c,2} \left\| \Delta w \right\|^\gamma \tag{5.37}$$

in terms of the extended covariance matrix, for small perturbations $\|\Delta w\| \leq \epsilon$, and for some constants $\kappa_{c,2} \geq 0$ and exponent $0 < \gamma \leq 4$.

Gradient Noise Model

It is straightforward to verify from conditions (5.34)–(5.35) that

$$R_s = \sum_{k=1}^{N} R_{s,k} \tag{5.38}$$

$$R_q = \sum_{k=1}^{N} R_{q,k} \tag{5.39}$$

Gradient Noise Model

Moreover, in a manner similar to (3.134), we conclude from (5.36) that the second-order moments of the individual gradient noise processes satisfy:

$$\mathbb{E}\left[\,\|\boldsymbol{s}_{k,i}(\boldsymbol{w})\|^2\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\right] \;\leq\; \left(\bar{\beta}_k/h\right)^2\,\|\boldsymbol{w}\|^2 \;+\; \bar{\sigma}_{s,k}^2 \tag{5.40}$$

Using this condition, along with (5.33), it is again straightforward to verify that the aggregate noise satisfies

$$\mathbb{E}\,[\;\boldsymbol{s}_i(\boldsymbol{w})\;|\;\boldsymbol{\mathcal{F}}_{i-1}\;] \;=\; 0 \tag{5.41}$$

$$\mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{w})\|^2\,|\,\boldsymbol{\mathcal{F}}_{i-1}\right] \;\leq\; \frac{1}{h^2}\left(\sum_{k=1}^{N}\bar{\beta}_k^2\right)\|\boldsymbol{w}\|^2 + \sum_{k=1}^{N}\bar{\sigma}_{s,k}^2 \tag{5.42}$$

Gradient Noise Model

so that repeating argument (3.28) we can deduce that

$$\mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^2 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \;\leq\; (\beta_c/h)^2\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \;+\; \sigma_s^2 \qquad (5.43)$$

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

$$\beta_c^2 \;\triangleq\; 2\left(\sum_{k=1}^{N}\bar{\beta}_k^2\right) \qquad (5.44)$$

$$\sigma_s^2 \;\triangleq\; 2\left(\sum_{k=1}^{N}\bar{\beta}_k^2\right)\|w^o\|^2 \;+\; \sum_{k=1}^{N}\bar{\sigma}_{s,k}^2 \qquad (5.45)$$

Gradient Noise Model

Likewise, we can conclude from (5.36) that

$$\mathbb{E}\left[\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{i-1})\|^4 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \leq (\beta_{4,k}/h)^4 \|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + \sigma_{s4,k}^4 \tag{5.46}$$

in terms of the scalars

$$\beta_{4,k}^4 \triangleq 8\bar{\beta}_k^4 \tag{5.47}$$

$$\sigma_{s4,k}^4 \triangleq 8(\bar{\beta}_{4,k}^4/h^4)\|w^o\|^4 + \bar{\sigma}_{s,k}^4 \tag{5.48}$$

By extrapolation, we also conclude that the fourth-order moment of the aggregate noise, $\boldsymbol{s}_i(\boldsymbol{w}_{i-1})$, is similarly bounded. More explicitly, it will hold that

Gradient Noise Model

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}\right] &\leq N^3\left(\sum_{k=1}^{N}(\beta_{4,k}^4/h^4)\right)\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + N^3\left(\sum_{k=1}^{N}\sigma_{s4,k}^4\right)\\ &\triangleq (\beta_a/h)^4\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + \sigma_a^4 \qquad (5.49)\end{aligned}$$

for some nonnegative constants β_a^4 and σ_a^4. This can be seen as follows. Exploiting the convexity of the norm function $f(x) = \|x\|^4$ and using Jensen's inequality (F.26) we can write

Gradient Noise Model

$$
\begin{aligned}
\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4 &\overset{(5.25)}{=} \left\|\sum_{k=1}^{N} \boldsymbol{s}_{k,i}(\boldsymbol{w}_{i-1})\right\|^4 \\
&= \left\|\sum_{k=1}^{N} \frac{1}{N} N\, \boldsymbol{s}_{k,i}(\boldsymbol{w}_{i-1})\right\|^4 \\
&\overset{(\mathrm{F}.26)}{\leq} \frac{1}{N}\sum_{k=1}^{N} N^4 \|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{i-1})\|^4 \\
&\leq N^3 \left(\sum_{k=1}^{N} \|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{i-1})\|^4\right) \qquad (5.50)
\end{aligned}
$$

Gradient Noise Model

from which we conclude that

$$\mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1})\|^4 \,|\, \boldsymbol{\mathcal{F}}_{i-1}\right] \leq N^3 \left(\sum_{k=1}^{N} \mathbb{E}\left[\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{i-1})\|^4 \,|\boldsymbol{\mathcal{F}}_{i-1}\right]\right) \quad (5.51)$$

and result (5.49) follows.

MSE Performance

Motivated by the discussion that led to expressions (4.94) and (4.95) for the MSD and ER metrics in the single agent case, we similarly define the MSD and ER performance measures for the centralized solution as follows:

$$\mathrm{MSD}_{\mathrm{cent}} \triangleq \mu \cdot \left(\lim_{\mu \to 0} \limsup_{i \to \infty} \frac{1}{\mu} \mathbb{E} \, \|\widetilde{\boldsymbol{w}}_i\|^2 \right) \tag{5.52}$$

$$\mathrm{ER}_{\mathrm{cent}} \triangleq \frac{\mu}{N} \cdot \left(\lim_{\mu \to 0} \limsup_{i \to \infty} \frac{1}{\mu} \mathbb{E} \left\{ J^{\mathrm{glob}}(\boldsymbol{w}_{i-1}) - J^{\mathrm{glob}}(w^o) \right\} \right) \tag{5.53}$$

MSE Performance

where the scaling by $1/N$ in (5.53) is meant to ensure that $\mathrm{ER}_{\mathrm{cent}}$ is compatible with the definition used for non-cooperative agents in (4.95) and later for multi-agent networks in (11.34). For example, when the individual costs happen to coincide, say, $J_k(w) \equiv J(w)$ for $k = 1, 2, \ldots, N$, then the aggregate cost (5.19) reduces to $J^{\mathrm{glob}}(w) = N\,J(w)$ and expression (5.53) becomes consistent with the earlier expression (4.95). Note that we are adding the subscript "cent" to indicate that the above MSD and ER measures are associated with the centralized solution. As explained earlier in Sec. 4.5, we sometimes rewrite the above definitions for the MSD and ER measures more compactly (but less rigorously) as

MSE Performance

$$\mathrm{MSD}_{\mathrm{cent}} = \lim_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \tag{5.54}$$

$$\mathrm{ER}_{\mathrm{cent}} = \lim_{i\to\infty} \frac{1}{N}\mathbb{E}\left\{J^{\mathrm{glob}}(\boldsymbol{w}_{i-1}) - J^{\mathrm{glob}}(w^o)\right\} \tag{5.55}$$

with the understanding that the limits on the right-hand side in the above two expressions are computed according to the definitions (5.52)–(5.53).

Hessian and Moment Matrices

Definition 5.1 (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 minimizer $w = w^o$. The matrices are defined as follows:

$$H_k \triangleq \nabla_w^2 J_k(w^o), \qquad G_k \triangleq \begin{cases} R_{s,k} & \text{(real case)} \\ \begin{bmatrix} R_{s,k} & R_{q,k} \\ R_{q,k}^* & R_{s,k}^{\mathsf{T}} \end{bmatrix} & \text{(complex case)} \end{cases} \tag{5.56}$$

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

Hessian Matrix

In view of the lower bound condition in (5.20), it follows that

$$\sum_{k=1}^{N} H_k \;>\; 0 \tag{5.57}$$

so that the sum of the $\{H_k\}$ matrices is invertible. This matrix sum appears in the performance expressions below.

Centralized Solution

Theorem 5.1: For small-enough step-sizes:

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

The rate at which $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$ approaches its steady-state region is approximated to first-order in μ by:

$$\alpha_{\mathrm{cent}} = 1 - 2\nu_c\left(\frac{\mu}{hN}\right)$$

MSE Performance

Theorem 5.1 (Performance of centralized solution). Assume the aggregate cost (5.19) satisfies condition (5.20) for some parameters $0 < \nu_c \leq \delta_c$. Assume also that the gradient noise processes satisfy conditions (5.40)–(5.33). For any μ satisfying

$$\frac{\mu}{hN} < \frac{2\nu_c}{\delta_c^2 + \beta_c^2} \tag{5.58}$$

it holds that

$$\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;\leq\; \alpha\,\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 \;+\; \left(\frac{\mu}{N}\right)^2 \sigma_s^2 \tag{5.59}$$

MSE Performance

where the parameters $\{\sigma_s^2, \beta_c^2\}$ are defined by (5.44)–(5.45), and where the scalar α satisfies $0 \leq \alpha < 1$ and is given by

$$\alpha \;=\; 1 - 2\nu_c \left(\frac{\mu}{hN}\right) \;+\; (\delta_c^2 + \beta_c^2)\left(\frac{\mu}{hN}\right)^2 \tag{5.60}$$

It follows from (5.59) that for sufficiently small step-sizes:

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;=\; O(\mu) \tag{5.61}$$

Moreover, under the additional smoothness conditions (5.21) on $J^{\text{glob}}(w)$ and (5.37) on the individual noise covariance matrices, it holds that

MSE Performance

$$\limsup_{i\to\infty} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 = \mathrm{MSD}_{\mathrm{cent}} + O\left(\mu^{1+\gamma_m}\right) \tag{5.62}$$

$$\limsup_{i\to\infty} \frac{1}{N}\mathbb{E}\left\{J^{\mathrm{glob}}(\boldsymbol{w}_{i-1}) - J^{\mathrm{glob}}(w^o)\right\} = \mathrm{ER}_{\mathrm{cent}} + O\left(\mu^{1+\gamma_m}\right) \tag{5.63}$$

where

$$\gamma_m \triangleq \frac{1}{2}\min\{1,\gamma\} > 0 \tag{5.64}$$

with $\gamma \in (0,4]$ from (5.37), and where

MSE Performance

$$\mathrm{MSD}_{\mathrm{cent}} = \frac{\mu}{2hN} \mathrm{Tr}\left[\left(\sum_{k=1}^{N} H_k\right)^{-1} \left(\sum_{k=1}^{N} G_k\right)\right] \tag{5.65}$$

$$\mathrm{ER}_{\mathrm{cent}} = \frac{\mu h}{4N^2} \mathrm{Tr}\left(\sum_{k=1}^{N} R_{s,k}\right) \tag{5.66}$$

MSE Performance

The N^2 factor in the denominator of (5.66) is because of the normalization by $1/N$ in the definition (5.53). Moreover, for $i \gg 1$, the rate at which the error variance, $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$, approaches its steady-state region (5.62) is well-approximated to first-order in μ by

$$\alpha = 1 - \frac{2\mu}{N}\lambda_{\min}\left(\sum_{k=1}^{N} H_k\right) \tag{5.67}$$

Conditions on Gradient Noise

If desired, we can relax conditions (5.33)–(5.36) and replace them by requirements on the aggregate noise process (5.25) directly, such as requiring:

$$\mathbb{E}\,[\,\boldsymbol{s}_i(\boldsymbol{w})\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,] \;=\; 0 \tag{5.68}$$

$$\mathbb{E}\,\Big[\,\|\boldsymbol{s}_i(\boldsymbol{w})\|^4\,|\,\boldsymbol{\mathcal{F}}_{i-1}\,\Big] \;\leq\; (\beta_c/h)^4\,\|\boldsymbol{w}\|^4 \;+\; \sigma_s^4 \tag{5.69}$$

for some nonnegative constants β_c^4 and σ_s^4. Note in particular that these assumptions do not impose the uncorrelatedness and circularity conditions (5.34)–(5.35) on the individual noise processes. We also replace

Conditions on Gradient Noise

condition (5.37), which involves the individual agents, by the requirement

$$\left\| R_{s,i}^{e}(w^o + \Delta w) - R_{s,i}^{e}(w^o) \right\| \leq \kappa_{c,2} \left\| \Delta w \right\|^{\gamma} \tag{5.70}$$

in terms of the covariance matrix of the extended aggregate noise vector, $\boldsymbol{s}_i^e(\boldsymbol{w})$. Then, the conclusions of Theorem 5.1 will continue to hold using $\{\beta_c^2, \sigma_s^2\}$ from (5.69), and with the sum of the $\{G_k\}$ appearing in (5.65) replaced by

Conditions on Gradient Noise

$$G_c \triangleq \begin{cases} R_s & \text{(real case)} \\ \begin{bmatrix} R_s & R_q \\ R_q^* & R_s^{\mathsf{T}} \end{bmatrix} & \text{(complex case)} \end{cases} \tag{5.71}$$

in terms of the moment matrices (5.31)–(5.32) for the aggregate noise process. More specifically, let

$$H_c \triangleq \sum_{k=1}^{N} H_k \tag{5.72}$$

denote the aggregate Hessian matrix. It will then hold that

Conditions on Gradient Noise

$$\mathrm{MSD}_{\mathrm{cent}} = \frac{\mu}{2hN}\,\mathrm{Tr}\left(H_c^{-1}G_c\right) \tag{5.73}$$

$$\mathrm{ER}_{\mathrm{cent}} = \frac{\mu h}{8N^2}\,\mathrm{Tr}\left(G_c\right) \tag{5.74}$$

When the individual gradient noise processes satisfy conditions (5.34)–(5.35), it is easy to verify that the moment matrix G_c will be given by

$$G_c = \sum_{k=1}^{N} G_k \tag{5.75}$$

so that the above MSD and ER expressions reduce to (5.65)–(5.66).

Comparison with Single Agents

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.

Non-Cooperative Processing

Continuing with the conditions in Assumption 5.2, we now compare the performance of the centralized solution (5.22) to that of non-cooperative processing where agents act independently of each other and run the recursion:

$$\boldsymbol{w}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu \widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{k,i-1}), \quad i \geq 0 \tag{5.76}$$

This comparison is *meaningful* only when all agents share the same minimizer, i.e., when

$$w_k^o = w^o, \quad k = 1, 2, \ldots, N \tag{5.77}$$

Non-Cooperative Processing

so that we can compare how well the individual agents are able to recover the same w^o as the centralized solution. For this reason, we need to re-introduce in this section only the requirement that all individual costs $\{J_k(w)\}$ are $\nu-$strongly convex with a uniform parameter ν. Since $J^{\text{glob}}(w)$ is the aggregate sum of the individual costs, then we can set the lower bound ν_c for the Hessian of $J^{\text{glob}}(w)$ in (5.20) at $\nu_c = N\nu$. From expressions (3.142) and (5.60) we then conclude that, for a sufficiently small μ, the convergence rates of the non-cooperative and centralized solutions will be similar to first-order in μ:

Non-Cooperative Processing

$$\begin{aligned}\alpha_{\text{cent}} &\overset{(5.60)}{\approx} 1 - 2\nu_c\left(\frac{\mu}{hN}\right) \\ &= 1 - 2\nu\left(\frac{\mu}{h}\right) \\ &\overset{(3.142)}{\approx} \alpha_{\text{ncop},k}\end{aligned} \tag{5.78}$$

where the symbol $\approx$ signifies (here and elsewhere) that we are ignoring higher-order terms in μ. Moreover, we observe from (4.170) that the average MSD level across N non-cooperative agents is given by

Non-Cooperative Processing

$$\begin{aligned}
\mathrm{MSD}_{\mathrm{ncop,av}} &\triangleq \frac{1}{N}\sum_{k=1}^{N}\mathrm{MSD}_{\mathrm{ncop},k} \\
&= \frac{1}{N}\sum_{k=1}^{N}\frac{\mu}{2h}\mathrm{Tr}\left(H_k^{-1}G_k\right) \\
&= \frac{\mu}{2hN}\,\mathrm{Tr}\left(\sum_{k=1}^{N}H_k^{-1}G_k\right) \qquad (5.79)
\end{aligned}$$

so that comparing with (5.65), some simple algebra allows us to conclude the following statement.

No-Cooperation vs Centralized

$$\boldsymbol{w}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu \widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{k,i-1}), \quad i \geq 0$$

(non-cooperative)

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

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

$$\longrightarrow \quad \mathrm{MSD}_{\mathrm{cent}} < \mathrm{MSD}_{\mathrm{ncop,av}}$$

Non-Cooperative Processing

Lemma 5.2 (Centralized MSD is superior to non-cooperative MSD)**.** Comparing the MSD performance levels (5.79) and (5.65) it holds that for sufficiently small step-sizes:

$$\mathrm{MSD}_{\mathrm{cent}} \;<\; \mathrm{MSD}_{\mathrm{ncop,av}} \tag{5.80}$$

Proof

Proof. First recall that $H_k > 0$ and $G_k \geq 0$ for each k; note that the individual $\{H_k\}$ are now positive-definite in view of the strong convexity assumption on the individual costs in this section. Let

$$G_k = L_k L_k^*, \quad k = 1, 2, \ldots, N \tag{5.81}$$

denote a square-root factorization for G_k where the L_k are full-rank matrices. Then, using the property $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ for any matrices A and B of compatible dimensions, the MSD expressions can be re-written as (using H_c from (5.72):

$$G_k = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1^* \\ U_2^* \end{bmatrix} = U_1 \Lambda_1 U_1^* = \left(U_1 \Lambda_1^{1/2}\right) \left(U_1 \Lambda_1^{1/2}\right)^*$$

Proof

$$\mathrm{MSD}_{\mathrm{ncop,av}} = \frac{\mu}{2Nh}\,\mathrm{Tr}\left[\sum_{k=1}^{N} L_k^* H_k^{-1} L_k\right] \tag{5.82}$$

$$\mathrm{MSD}_{\mathrm{cent}} = \frac{\mu}{2Nh}\,\mathrm{Tr}\left[\sum_{k=1}^{N} L_k^* H_c^{-1} L_k\right] \tag{5.83}$$

so that

$$\mathrm{MSD}_{\mathrm{ncop,av}} - \mathrm{MSD}_{\mathrm{cent}} = \frac{\mu}{2Nh}\,\mathrm{Tr}\left[\sum_{k=1}^{N} L_k^*(H_k^{-1} - H_c^{-1})L_k\right] \tag{5.84}$$

The result follows by noting that $H_c^{-1} < H_k^{-1}$ for any k.

□

Centralized vs. Non-Cooperative

That is, while the centralized solution need not outperform every individual non-cooperative agent in general, its performance outperforms the average performance across all non-cooperative agents. The next example illustrates the above result by considering the scenario where all agents have the same Hessian matrices at $w = w^o$, namely,

$$H_k \equiv H, \quad k = 1, 2, \ldots, N \tag{5.85}$$

This situation occurs, for example, when the individual costs are identical across the agents, say, $J_k(w) \equiv J(w)$, as is common in machine

Centralized vs. Non-Cooperative

learning applications. This situation also occurs for mean-square-error costs of the form described by (5.5)–(5.6), when the regression covariance matrices, $\{R_{u,k}\}$, are uniform across all agents. In these cases when the Hessian matrices H_k are uniform, the example below establishes that the centralized solution actually improves over the average MSD performance of the non-cooperative solution by a factor of N [207].

Example #5.1

Example 5.1 (N-fold improvement in performance). Consider a collection of N agents whose individual cost functions, $J_k(w)$, are $\nu-$strongly convex and are minimized at the same location $w = w^o$. The costs are also assumed to have identical Hessian matrices at $w = w^o$, i.e., $H_k \equiv H$. Then, using (5.65), the MSD of the centralized implementation is given by

$$\text{MSD}_{\text{cent}} = \frac{1}{N}\left(\frac{\mu}{2Nh}\sum_{k=1}^{N}\text{Tr}(H^{-1}G_k)\right) \overset{(5.79)}{=} \frac{1}{N}\,\text{MSD}_{\text{ncop,av}} \qquad (5.86)$$

■

Example #5.2

Example 5.2 (Multi-fold improvement in performance). Assume in this example that all data are real-valued, and consider a situation in which the matrices $\{R_{s,k}\}$ are uniform across all agents so that $R_{s,k} \equiv R_s$, while $H_k = \alpha_k I_M > 0$ for some scalars $\{\alpha_k\}$. This situation arises, for instance, in the mean-square-error case (5.6) when $R_{u,k} = \sigma_{u,k}^2 I_M$ and the noise variances $\sigma_{v,k}^2$ across the agents are such that the product $\sigma_{v,k}^2 \sigma_{u,k}^2 \equiv \sigma^2/4$ remains invariant over the agents. Then, in this case,

$$H_k \overset{(2.8)}{=} 2R_{u,k} = 2\sigma_{u,k}^2 I_M \equiv \alpha_k I_M \tag{5.87}$$

$$R_{s,k} \overset{(4.14)}{=} 4\sigma_{v,k}^2 R_{u,k} = 4\sigma_{v,k}^2 \sigma_{u,k}^2 I_M = \sigma^2 I_M \equiv R_s \tag{5.88}$$

Example #5.2

Let α_A and α_H denote the arithmetic and harmonic means of the scalars $\{\alpha_k\}$:

$$\alpha_A \triangleq \frac{1}{N}\sum_{k=1}^{N}\alpha_k, \quad \alpha_H \triangleq \left(\frac{1}{N}\sum_{k=1}^{N}\alpha_k^{-1}\right)^{-1} \tag{5.89}$$

Then, expressions (5.79) and (5.65) give

$$\text{MSD}_{\text{ncop,av}} = \mu\,\frac{1}{\alpha_H}\,M\sigma^2, \qquad \text{MSD}_{\text{cent}} = \frac{\mu}{N}\,\frac{1}{\alpha_A}\,M\sigma^2 \tag{5.90}$$

so that

$$\frac{\text{MSD}_{\text{cent}}}{\text{MSD}_{\text{ncop,av}}} = \frac{1}{N}\left(\frac{\alpha_H}{\alpha_A}\right) \tag{5.91}$$

Example #5.2

in terms of the ratio of the harmonic mean to the arithmetic mean of the $\{\alpha_k\}$. Recall that the harmonic mean of a set of numbers is always smaller than or equal to the arithmetic mean of these numbers (and, moreover, its value tends to be close to the smaller numbers), it then holds that, for sufficiently small step-sizes:

$$\frac{\mathrm{MSD}_{\mathrm{cent}}}{\mathrm{MSD}_{\mathrm{ncop,av}}} \leq \frac{1}{N} \tag{5.92}$$

■

Example #5.3

Example 5.3 (Centralized learner). We revisit Example 4.5 and consider now a collection of N learners labeled $k = 1, 2, \ldots, N$. As before, each learner k receives a streaming sequence of real-valued vector samples $\{\boldsymbol{x}_{k,i}, i = 1, 2, \ldots\}$ arising from some fixed distribution $\mathcal{X}$. The goal is to determine the $M \times 1$ minimizer w^o of the $\nu-$strongly convex risk function $J(w)$ in (4.151). In Example 4.5 we examined the non-cooperative solution (4.152) where agents worked independently of each other to estimate w^o. In this example, we examine a centralized solution of the following stochastic-gradient form:

$$\boldsymbol{w}_i \;=\; \boldsymbol{w}_{i-1} \;-\; \frac{\mu}{N}\sum_{k=1}^{N} \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{i-1}; \boldsymbol{x}_{k,i}), \quad i \geq 0 \tag{5.93}$$

Example #5.3

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

$$\boldsymbol{s}_{k,i}(\boldsymbol{w}_{i-1}) \;=\; \nabla_{w^{\mathsf{T}}} Q(\boldsymbol{w}_{i-1};\boldsymbol{x}_{k,i}) \;-\; \nabla_{w^{\mathsf{T}}} J(\boldsymbol{w}_{i-1}) \tag{5.94}$$

so that evaluating the expression for $\boldsymbol{s}_{k,i}(\boldsymbol{w})$ at $\boldsymbol{w} = w^o$, and using the fact that $\nabla_w J(w^o) = 0$, we get

$$\boldsymbol{s}_{k,i}(w^o) \;=\; \nabla_{w^{\mathsf{T}}} Q(w^o;\boldsymbol{x}_{k,i}) \tag{5.95}$$

Since we are assuming the distribution of the random process $\boldsymbol{x}_{k,i}$ is stationary and fixed across all agents, it follows that the covariance matrix of $\boldsymbol{s}_{k,i}(w^o)$ is constant across all agents:

$$R_{s,k} \;\triangleq\; \mathbb{E}\,\boldsymbol{s}_{k,i}(w^o)\boldsymbol{s}_{k,i}^{\mathsf{T}}(w^o) \;\equiv\; R_s, \qquad k = 1, 2, \ldots, N \tag{5.96}$$

Example #5.3

Moreover, since all data are real-valued, it follows that the moment matrix G_k is $M \times M$ and given by

$$G_k = R_s, \quad k = 1, 2, \ldots, N \tag{5.97}$$

Substituting into (5.66), and using $h = 1$ for real data, we conclude that the excess-risk of the centralized solution (per unit agent) is given by

$$\mathrm{ER}_{\mathrm{cent}} = \frac{\mu}{4N^2} \mathrm{Tr}(NR_s) = \frac{\mu}{4N} \mathrm{Tr}(R_s) \tag{5.98}$$

Example #5.3

which is $N-$fold superior to the performance of the non-cooperative agent given by (4.155) when $\mu_k \equiv \mu$. Similarly, using (5.65) we find that the MSD performance of the centralized solution is given by

$$\text{MSD}_{\text{cent}} \;=\; \frac{\mu}{2N}\,\text{Tr}(H^{-1}R_s) \tag{5.99}$$

■

Example #5.4

Example 5.4 (Fully-connected networks). In preparation for the discussion on networked agents, it is useful to describe one extreme situation where a collection of N agents are fully connected to each other — see Figure 5.3. In this case, each agent is able to access the data from all other agents and, therefore, each agent can run a centralized implementation of the same form as (5.22), namely,

$$\boldsymbol{w}_{k,i} = \boldsymbol{w}_{k,i-1} - \frac{\mu}{N}\sum_{\ell=1}^{N}\widehat{\nabla_{w^*} J_\ell}(\boldsymbol{w}_{k,i-1}), \quad i \geq 0 \tag{5.100}$$

Example #5.4

In a fully-connected network, each agent can run the centralized algorithm:

$$\boldsymbol{w}_{k,i} = \boldsymbol{w}_{k,i-1} - \frac{\mu}{N}\sum_{\ell=1}^{N}\widehat{\nabla_{w^*} J_\ell}(\boldsymbol{w}_{k,i-1}),$$

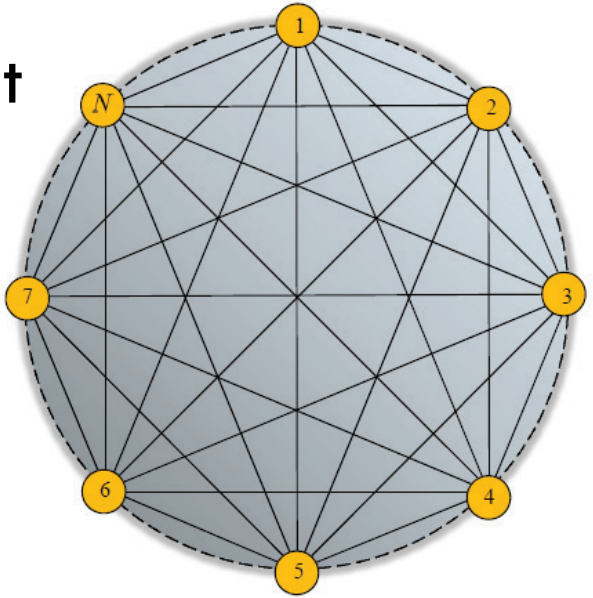

Figure 5.3

- Other pieces of information can be shared.
- Combination coefficients do not need to be uniform at $1/N$.
- No need to have access to information from all agents.

Example #5.4

When this happens, each agent will attain the same performance level as that of the centralized solution. Two observations are in place [207]. First, note from (5.100) that the information that agent k is receiving from all other agents is their gradient vector approximations. Obviously, other pieces of information could be shared among the agents, such as their iterates $\{\boldsymbol{w}_{\ell,i-1}\}$. Second, note that the right-most term multiplying μ in (5.100) corresponds to a convex combination of the approximate gradients from the various agents, with the combination coefficients being uniform and equal to $1/N$. In general, there is no need for these combination weights to be identical. Even more importantly, agents do not need to have access to information from all other agents in the network. We are going to see in the future chapters that interactions with a limited number of neighbors is sufficient for the agents to attain performance that is comparable to that of the centralized solution.

Example #5.4

Figure 5.4 shows a sample selection of connected topologies for five agents. The panels in the first row correspond to the non-cooperative case (left) and the fully-connected case (right). The panels in the bottom row illustrate some other topologies. In the coming chapters, we are going to present results that allow us to answer useful questions about such networked agents such as [207]: (a) which topology has best performance in terms of mean-square error and convergence rate? (b) Given any connected topology, can it be made to approach the performance of the centralized stochastic-gradient solution?

Example #5.4

(c) Which aspects of the topology influence performance? (d) Which aspects of the combination weights (policy) influence performance? (e) Can different topologies deliver similar performance levels? (f) Is cooperation always beneficial? (g) If the individual agents are able to solve the inference task individually in a stable manner, does it follow that the connected network will remain stable regardless of the topology and regardless of the cooperation strategy?

■

Example #5.4

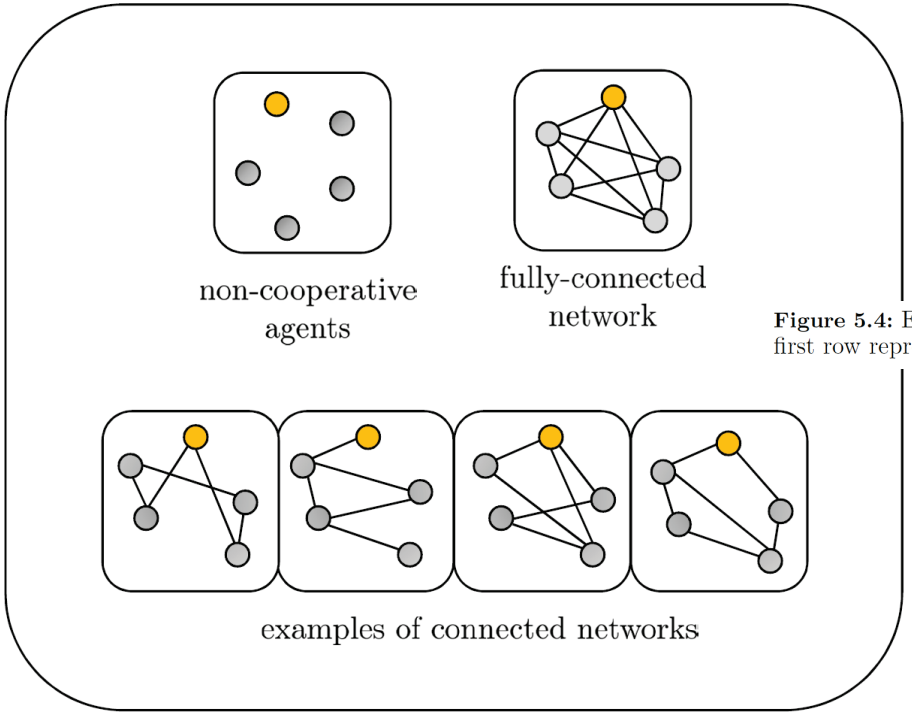

Figure 5.4: Examples of connected networks, with the left-most panel on the first row representing a collection of non-cooperative agents.

Example #5.4

- Which topology has best performance?
- Which aspects of the topology influence performance?
- Which aspects of combination policy influence performance?
- Can different topologies deliver same performance?
- Is cooperation always beneficial?
- Can networks match performance of centralized solutions?
- Does stability of individual agents ensure stability of network?

Decaying Step-Sizes

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.

Decaying Step-Sizes

We finally examine the convergence and performance of the centralized solution (5.22) with a decaying step-size sequence, namely,

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} - \frac{\mu(i)}{N} \sum_{k=1}^{N} \widehat{\nabla_{w^*} J_k}(\boldsymbol{w}_{i-1}), \quad i \geq 0 \tag{5.101}$$

where $\mu(i) > 0$ satisfies either of the following two sets of conditions:

Decaying Step-Sizes

$$\sum_{i=0}^{\infty} \mu(i) = \infty, \qquad \lim_{i\to\infty} \mu(i) = 0 \tag{5.102}$$

or

$$\sum_{i=0}^{\infty} \mu(i) = \infty, \qquad \sum_{i=0}^{\infty} \mu^2(i) < \infty \tag{5.103}$$

The following statement follows from the results of Lemmas 3.7 and 3.8 applied to the stochastic-gradient recursion (5.101).

Decaying Step-Sizes

Lemma 5.3 (Performance with decaying step-size). Assume the aggregate cost (5.19) satisfies condition (5.20) for some parameters $0 < \nu_c \leq \delta_c$. Assume also that the individual gradient noise processes defined by (5.24) satisfy conditions (5.40)–(5.33). Then, the following convergence properties hold for (5.101):

(a) If the step-size sequence $\mu(i)$ satisfies (5.103), then $\boldsymbol{w}_i$ converges almost surely to w^o, written as $\boldsymbol{w}_i \to w^o$ a.s.

(b) If the step-size sequence $\mu(i)$ satisfies (5.102), then $\boldsymbol{w}_i$ converges in the mean-square-error sense to w^o, i.e., $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \to 0$.

Decaying Step-Sizes

(c) If the step-size sequence is selected as $\mu(i) = \tau_c/(i+1)$, where $\tau_c > 0$, then three convergence rates are possible. Specifically, for large enough i, it holds that:

$$\begin{cases} \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;\leq\; \left(\frac{(\tau_c/N)^2\sigma_s^2}{(\nu_c/h)(\tau_c/N)-1}\right)\frac{1}{i} \;+\; o\left(\frac{1}{i}\right), & \nu_c\tau_c/hN > 1 \\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;=\; O\left(\frac{\log i}{i}\right), & \nu_c\tau_c/hN = 1 \\ \mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2 \;=\; O\left(\frac{1}{i(\nu_c/h)(\tau_c/N)}\right), & \nu_c\tau_c/hN < 1 \end{cases} \tag{5.104}$$

where $h = 2$ for complex data and $h = 1$ for real data. The fastest convergence rate occurs when $\nu_c\tau_c/hN > 1$ (i.e., for large enough τ_c) and is in the order of $O(1/i)$.

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.