Average Consensus Problems

Mingyi Hong

University Of Minnesota

M. Hong would like to thank Mr. Siliang Zeng for helping preparing the slides.

Outline

- Network Model and Assumptions
- The Consensus Problem
- Weight / Consensus matrices
- Convergence of Transition Matrices to Averages
- Main reference [Nedich- Ozdaglar-09] [1]

[1] A. Nedich and A. Ozdaglar, "Cooperative Distributed multi-agent optimization", 2009

Unconstrained multiagent-optimization problem

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}} \quad & \sum_{i=1}^{m} f_i(x) \\ \text{subject to} \quad & x \in \mathbb{R}^n \end{aligned}$$

- Each $f_i : \mathbb{R}^n \to \mathbb{R}$ is a convex function, representing the local-objective function of agent i, which is known only to this agent.
- Do not assume differentiability of f_i . At the points where the function fails to be differentiable, assume a subgradient exists (cf. Lecture 2)
- For simplicity of discussion, will mostly assume that $n = 1$.

Multi-agent optimization method

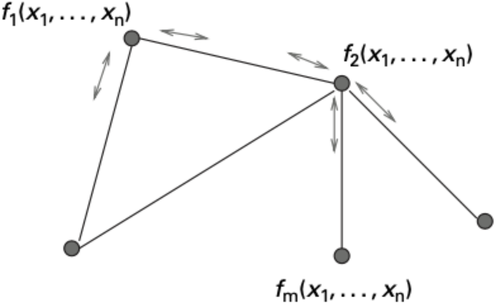

Figure 1.1: Illustration of the network with each agent having its local objective and communicating locally with its neighbors [Nedich-Ozdaglar-09].

[2]

[2] A. Nedich and A. Ozdaglar, "Cooperative Distributed multi-agent optimization", 2009

Distributed Subgradient Algorithm

- We introduce the distributed subgradient algorithm
- **Main idea**: Consensus Step + Gradient Descent
- Sometimes known as DGD (distributed gradient descent)
- First proposed in [Nedich-Ozdaglar-09] [3]
- Also closely related to other classical methods, such as [Tsitsiklis et al 86] [4]

[3] A. Nedich and A. Ozdaglar, "Distributed Subgradient Methods for Multi-Agent Optimization", TAC, 2009

[4] J. Tsitsiklis and D. P. Bertsekas and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms", TAC, 1986

Distributed Subgradient Algorithm

- Specifically, agent i updates its estimates by setting

$$x_i(k+1) = \underbrace{\sum_{j=1}^{m} w_{ij}(k)x_j(k)}_{\text{Consensus Step}} \quad - \quad \underbrace{\alpha \times d_i(k)}_{\text{Subgradient Step}} \tag{1.1}$$

  where the scalar $\alpha > 0$ is a stepsize and $d_i(k)$ is a subgradient of the agent i cost function $f_i(x)$ at $x = x_i(k)$.

- The scalars $w_{i1}(k), ..., w_{im}(k)$ are non-negative weights that agent i gives to the estimates $x_1(k), ..., x_m(k)$.

- Note, we use (k) to denote k-th iterations because the expression will be complicated in this lecture; later we will also use superscript ^{k} to denote iterations

Distributed Subgradient Algorithm

Considering the weight matrix $W(k) = [w_{ij}(k)]_{i,j=1,\cdots,m}$ where its ij-th element is $w_{ij}(k)$; Then (1.1) can be expressed as

$$\boldsymbol{x}(k+1) = W(k+1)\boldsymbol{x}(k) + \alpha \times \boldsymbol{d}(k)$$

where

$$\begin{aligned}\boldsymbol{x}(k+1) &:= [x_1(k+1); \cdots ; x_m(k+1)] \\ \boldsymbol{d}(k) &:= [d_1(k+1); \cdots ; d_m(k+1)]\end{aligned}$$

Distributed Subgradient Algorithm

The matrix $W(k)$ is an important object; it has the properties that:

- $w_{i,j}(k)$ characterizes the active link (j, i) at time k. Suppose the neighbors j communicates with agent i at time k, then $w_{ij}(k) > 0$ (including i itself).

- Suppose node j do not communicate with i at time k then $w_{ij}(k) = 0$.

- At each iteration k, such a matrix can change – indicating that the connectivity pattern can also change

Representation using transition matrices

- To understand the dynamics, we need to analyze the properties of matrices $W(k)$'s

- In particular, we define a "transition matrix" $\Phi(k, s)$ for any s and k with $k \geq s$, as follows:

$$\Phi(k, s) = W(k)W(k-1)\cdots W(s+1)W(s) \qquad (1.2)$$

- s: the starting index; k: the end index; $(k-s+1)$, # of multiplications

Representation using transition matrices

- Specifically, for the iterates generated by (1.1), we have for any i, and any s and k with $k \geq s$ [recurse back to iteration s],

$$x_i(k+1) \tag{1.3}$$
$$= \sum_{j=1}^{m} [\Phi(k,s)]_{ij} x_j(s) - \alpha \sum_{r=s}^{k-1} \sum_{j=1}^{m} [\Phi(k,r+1)]_{ij} d_j(r) - \alpha d_i(k)$$

- To study the asymptotic behavior of the estimates $x_i(k)$, we need to understand the behavior of the transition matrices $\Phi(k,s)$.

The Plan

- This seems to be a complicated dynamics to analyze
- We start from the simplest case: $f_i(x) \equiv 0, \ \forall \ i$, and just analyze the averaging iteration (the remaining lecture)
- **Key point:** the average iteration is able to make $\boldsymbol{x}(k)$'s converge to their averages linearly (i.e., very fast)
- Then go back to analyze the DGD iteration (next lecture)

Network model and transition matrices

- Definition 1: A vector a is said to be a "stochastic vector" when its components a_i are non-negative and $\sum_i a_i = 1$.

- Definition 2: A square matrix W is said to be "stochastic" when each row of W is a stochastic vector:

$$W\mathbf{1} = \mathbf{1}$$

- Definition 3: It is said to be "doubly stochastic" when both W and its transpose W' are stochastic matrices:

$$W\mathbf{1} = \mathbf{1}, \quad \mathbf{1}^T W = \mathbf{1}^T$$

Network model and transition matrices

- To understanding some properties and behaviors of the weight matrices $W(k)$, we need to further assumptions:
  - **Assumption 1** : For any time $k \geq 0$, the weight matrix $W(k)$ is doubly stochastic with positive diagonal. Additionally, there is a scalar $\eta > 0$ such that if $w_{ij}(k) > 0$, then $w_{ij}(k) \geq \eta$.
  - Note, the matrix does not need to be symmetric

Network model and transition matrices

The understanding of Assumption 1:

- The doubly stochasticity assumption on the weight matrix will guarantee that the function f_i of every agent i receives similar weight in the long run (after all, no one's gradients are more important than the other). We will see this fact very soon.

- The significant weight (characterized by using $\mu > 0$) is needed to ensure that new information is aggregated into the agent system persistently in time.

Network model and transition matrices

One example satisfying Assumption 3 when the agent communications are bidirectional (undirected graph):

- Metropolis-based weights: For all i and j with $j \neq i$,

$$w_{ij}(k) = \begin{cases} \frac{1}{1+\max\{n_i(k), n_j(k)\}} & \text{if j communicates with i at time k,} \\ 0 & \text{otherwise,} \end{cases}$$

  $n_i(k)$ is the number of neighbors communicating with agent i at time k. Using these, the weights $w_{ii}(k)$ for all $i = 1, \cdots, m$ are as follows

$$w_{ii}(k) = 1 - \sum_{j \neq i} w_{ij}(k) > 0$$

Network model and transition matrices

- Is **Assumption 1** enough? What will happen if $W(k)$'s only satisfy this assumption?

Network model and transition matrices

- Is **Assumption 1** enough? What will happen if $W(k)$'s only satisfy this assumption?

- Introduce the index set $\mathcal{N} = \{1, \cdots, m\}$ and define $\xi(W(k))$ to be the set of directed links at time k induced by the weight matrix $W(k)$.

$$\xi(W(k)) = \{(j,i) | w_{ij}(k) > 0, i, j = 1, \cdots, m\} \text{ for all } k.$$

That is, $\xi(W(k))$ contains all active edges at time k

Network model and transition matrices

- Then we make an assumption states that the agent network is frequently connected.

- **Assumption 2**: There exists an integer $B \geq 1$ such that the directed graph

$$G := (\mathcal{N}, \xi(W(kB)) \cup \cdots \cup \xi(W((k+1)B-1)))$$

  is strongly connected for all $k \geq 0$.
  Inuitively, every B instances the network will be "connected" (i.e., everyone can visit everyone)

- **Special Case:** If the network is fixed and do not change, then $B = 1$, and $W(k) = W(j)$ for all $k \neq j$.

Convergence of Transition Matrices to Averages

- Here, we study the behavior of the transition matrices $\Phi(k, s) = W(k)W(k-1)\cdots W(s+1)W(s)$.

- Bottom line: This product will converge to the following matrix very quickly

$$\frac{1}{n}\begin{bmatrix} 1, & \cdots & 1 \\ \vdots & \vdots & \vdots \\ 1, & \cdots & 1 \end{bmatrix} = \frac{1}{n}\mathbf{1}\mathbf{1}^T$$

- This understanding will be instrumental for analyzing the DGD method (with non-zero objective value)

Convergence of Transition Matrices to Averages

- To understand the convergence of the transition matrices $\Phi(k, s)$, we set $f_i(x) \equiv 0$, $\forall\ i$; DGD becomes the following "consensus-type" algorithm

$$\boldsymbol{x}(k+1) = W(k)\boldsymbol{x}(k) \tag{1.4}$$

where $\boldsymbol{x}(0) \in \mathbb{R}^m$ is an initial vector.

- We define

$$V(\boldsymbol{x}(k)) = \sum_{j=1}^{m} (x_j(k) - \bar{x}(k))^2 \text{ for all } k \geq 0, \tag{1.5}$$

where $\bar{x}(k) = \frac{1}{n} \sum_{i=1}^{n} x_i(n)$ is the average of x's ; $V(\boldsymbol{x}(k))$ will be abbreviated as $V(k)$.

- This quantity characterizes the consensus violation at time k

Convergence of Transition Matrices to Averages

- Under the doubly stochasticity of $W(k)$, the initial average $\bar{z}(0)$ is preserved by the update rule (1.4), (why?)

$$\bar{x}(k) = \bar{x}(0), \ \forall \ k$$

- Hence, we could define $V(k)$ as the measure of the "disagreement" among the agents.

- How do you think the iteration (1.4) will behave?

The Simple Case

- Let us first consider the simpler case to gain some intuition
- Assume for now that
  - $W(k) = W(j) = W$ for all i, j
  - The graph is connected
  - W is a symmetric and double stochastic matrix satisfying

$$\mathbf{1}^T W = \mathbf{1}^T, \quad W\mathbf{1} = \mathbf{1}.$$

- Then $\Phi(k, s) = W(k)W(k-1)\cdots W(s+1)W(s) = W^{k-s+1}$.
- Since the graph is connected, then by Perron - Frobenius theorem, $\rho(W) = 1$ and it is simple, where $\rho(A)$ is the spectral norm of the matrix W

The Simple Case

- The DGD iteration becomes

$$\boldsymbol{x}(k+1) = W\boldsymbol{x}(k) = \cdots W^K \boldsymbol{x}(0) \tag{2.1}$$

- Note $\mathbf{1}\bar{x}(k) = \frac{1}{m}\mathbf{1}\mathbf{1}^T \boldsymbol{x}(k), \ \forall\ k$

- Multiplying $\frac{1}{m}\mathbf{1}\mathbf{1}^T$ on both sides of DGD iteration shows

$$\bar{x}(k+1) = \bar{x}(k), \ k = 0, 1, \cdots$$

  Average are always the same!

- So we have the following key relation

$$\begin{aligned}\boldsymbol{x}(k+1) - \mathbf{1}\bar{x}(k+1) &= W\boldsymbol{x}(k) - \mathbf{1}\bar{x}(k) \\ &= (W - \frac{1}{m}\mathbf{1}\mathbf{1}^T)\boldsymbol{x}(k)\end{aligned} \tag{2.2}$$

The Simple Case

- On the other hand, notice that

$$(W - \frac{1}{m}\mathbf{1}\mathbf{1}^T)\bar{x}(k) = \frac{1}{m}W\mathbf{1}\mathbf{1}^T\boldsymbol{x}(k) - \frac{1}{m^2}\mathbf{1}\mathbf{1}^T\mathbf{1}\mathbf{1}^T\boldsymbol{x}(k) = 0$$

- Combine this with the previous inequality shows

$$\boldsymbol{x}(k+1) - \mathbf{1}\bar{x}(k+1) = (W - \frac{1}{m}\mathbf{1}\mathbf{1}^T)(\boldsymbol{x}(k) - \mathbf{1}\bar{x}(k)).$$

- Using the fact that the spectral radius of $W - \frac{1}{m}\mathbf{1}\mathbf{1}^T$ (denoted as θ) is strictly less than 1

$$\begin{aligned} \|\boldsymbol{x}(k+1) - \mathbf{1}\bar{x}(k+1)\| &\le \theta\|\boldsymbol{x}(k) - \mathbf{1}\bar{x}(k)\| \\ &\le \theta^k\|\boldsymbol{x}(0) - \mathbf{1}\bar{x}(0)\| \end{aligned} \qquad (2.3)$$

The Simple Case

- In summary, convergence to average exponentially quickly
- To analyze the property of the product we observe

$$
\begin{aligned}
&(W - \frac{1}{m}\mathbf{1}\mathbf{1}^T)(W - \frac{1}{m}\mathbf{1}\mathbf{1}^T) \\
&\quad = W^2 - \frac{1}{m}W\mathbf{1}\mathbf{1}^T - \frac{1}{m}\mathbf{1}\mathbf{1}^T W + \frac{1}{m}\mathbf{1}\mathbf{1}^T = W^2 - \frac{1}{m}\mathbf{1}\mathbf{1}^T
\end{aligned}
$$

- So the product of the matrix W also converges to the average exponentially

$$
\theta^k \geq \|(W - \frac{1}{m}\mathbf{1}\mathbf{1}^T)\|^k \geq \|(W - \frac{1}{m}\mathbf{1}\mathbf{1}^T)^k\| \geq \|W^k - \frac{1}{m}\mathbf{1}\mathbf{1}^T\|
$$

The General Case

- Next, we consider more general scenario
- We will present several lemmas and a theorem to identify the convergence of transition matrices $\Phi(k, s)$ to averages; Main results from [Nedic-09] [5]
- Mainly talks about proof steps; details see the paper, or the appendix

[5] Nedic, Angelia, et al. "On distributed averaging algorithms and quantization effects." IEEE Transactions on automatic control 54.11 (2009): 2506-2517.

Per-Iteration Descent on $V(\cdot)$

Lemma 3.1

Let A be a doubly stochastic matrix. Then, for all $\boldsymbol{x} \in \mathbb{R}^m$,

$$V(W\boldsymbol{x}) = V(\boldsymbol{x}) - \sum_{i<j} a_{ij}(x_i - x_j)^2$$

where a_{ij} is the (i,j)-th entry of the matrix $W^T W$.

Proof: Let $\mathbf{1}$ denote the vector in $\mathbb{R}^m$ with all entries equal to 1. Given the double stochasticity of W, we have

$$W\mathbf{1} = \mathbf{1}, \mathbf{1}^T W = \mathbf{1}^T,$$

$$\overline{W\boldsymbol{x}} = \frac{1}{n}\mathbf{1}^T W\boldsymbol{x} = \frac{1}{n}\mathbf{1}^T\boldsymbol{x} = \bar{x}.$$

[after applying W, the average does not change]

Proof of Lemma 3.1

Then we obtain that

$$\begin{aligned} V(\boldsymbol{x}) - V(W\boldsymbol{x}) &= (\boldsymbol{x} - \bar{x}\mathbf{1})^T(\boldsymbol{x} - \bar{x}\mathbf{1}) - (W\boldsymbol{x} - \overline{W\boldsymbol{x}}\mathbf{1})^T(W\boldsymbol{x} - \overline{W\boldsymbol{x}}\mathbf{1}) \\ &= (\boldsymbol{x} - \bar{x}\mathbf{1})^T(\boldsymbol{x} - \bar{x}\mathbf{1}) - (W\boldsymbol{x} - \bar{x}W\mathbf{1})^T(W\boldsymbol{x} - \bar{x}W\mathbf{1}) \\ &= (\boldsymbol{x} - \bar{x}\mathbf{1})^T(I - W^TW)(\boldsymbol{x} - \bar{x}\mathbf{1}). \qquad (3.1) \end{aligned}$$

Let a_{ij} be the (i,j)-th entry of W^TW. Note that W^TW is symmetric and stochastic, so that $a_{ij} = a_{ji}$ and $a_{ii} = 1 - \sum_{j \neq i} a_{ij}$. Then, it can be verified that

$$W^TW = I - \sum_{i<j} a_{ij}(e_i - e_j)(e_i - e_j)^T, \qquad (3.2)$$

where e_i is a unit vector with the i-th entry equal to 1, and all other entries equal to 0.

Proof of Lemma 3.1

By combing the equations (3.1) and (3.2), we obtain

$$\begin{aligned} V(\boldsymbol{x}) - V(A\boldsymbol{x}) &= (\boldsymbol{x} - \bar{x}\mathbf{1})^T \left(\sum_{i<j} a_{ij}(e_i - e_j)(e_i - e_j)^T \right) (\boldsymbol{x} - \bar{x}\mathbf{1}) \\ &= \sum_{i<j} a_{ij}(x_i - x_j)^2. \end{aligned} \tag{3.3}$$

$\square$

Proof of Lemma 3.1

- Lemma 3.1 implies that $V(k+1) \leq V(k)$ for all k.
- The amount of variance decrease is given by

$$V(k) - V(k+1) = \sum_{i<j} a_{ij}(k)(x_i(k) - x_j(k))^2. \tag{3.4}$$

- So intuitively, if there is no consensus, the algorithm will always continue
- But the thing is, at each iteration k, the graph may not be connected!
- We will further use this bound to provide a lower bound on the amount of decrease of $V(k)$.

Per-Period Descent on $V(\cdot)$

Lemma 3.2

Let Assumptions 1 and 2 hold. Let $\{\boldsymbol{x}(k)\}$ be generated by the update rule (1.1)*. Suppose that the components $x_i(kB)$ of the vector $\boldsymbol{x}(kB)$ have been ordered from largest to smallest, with ties broken arbitrarily. Then, we have*

$$V(kB) - V((k+1)B) \geq \frac{\eta}{2} \sum_{i=1}^{m-1} \left(x_i(kB) - x_{i+1}(kB)\right)^2$$

From ‘Descent’ to ‘Contraction’

- We have constructed a some useful Lemmas to analyze the convergence of the transition matrices.
- We next establish a bound on the variance contraction that plays a key role in our convergence analysis.

Per-Period Contraction on $V(\cdot)$

Lemma 3.3

Let Assumption 1 and 2 hold, and suppose that $V(kB) > 0$. Then,

$$V((k+1)B) \leq \left(1 - \frac{\eta}{2m^2}\right) V(kB) \text{ for all } k.$$

Note that this is a stronger result than the previous one; it shows that $V(\cdot)$ contracts after each full period B

Proof of Lemma 3.3

Proof: Without loss of generality, we assume that the components of $\boldsymbol{x}(kB)$ have been sorted in nonincreasing order. By Lemma 3.2, we have

$$V(kB) - V((k+1)B) \geq \frac{\eta}{2} \sum_{i=1}^{m-1} \left(x_i(kB) - x_{i+1}(kB)\right)^2 .$$

It implies that

$$\frac{V(kB) - V((k+1)B)}{V(kB)} \geq \frac{\eta}{2} \frac{\sum_{i=1}^{m-1} \left(x_i(kB) - x_{i+1}(kB)\right)^2}{\sum_{i=1}^{m} \left(x_i(kB) - \bar{x}(kB)\right)^2}$$

Proof of Lemma 3.3

- Observe that the right-hand side does not change when we add a constant to every $x_i(kB)$. So without loss of generality, assume $\bar{x}(kB) = 0$, and obtain

$$\frac{V(kB) - V((k+1)B)}{V(kB)} \geq \frac{\eta}{2} \min_{\substack{x_1 \geq x_2 \geq \cdots \geq x_m \\ \sum_i x_i = 0}} \frac{\sum_{i=1}^{m-1} (x_i - x_{i+1})^2}{\sum_{i=1}^{m} x_i^2}.$$

- Note, some indices have been removed for simplicity

- Also sine we assume $\bar{x}(kB) = 0$, it is equivalent to assuming that

$$\sum_i x_i(kB) = 0$$

Proof of Lemma 3.3

- Note that the RHS is unchanged if we multiply each x_i by the same constant. Therefore, we can assume, without loss of generality, that $\sum_{i=1}^{m} x_i^2 = 1$, so that

$$\frac{V(kB) - V((k+1)B)}{V(kB)} \geq \frac{\eta}{2} \min_{\substack{x_1 \geq x_2 \geq \cdots \geq x_m \\ \sum_i x_i = 0, \sum_{i=1} x_i^2 = 1}} \sum_{i=1}^{m-1} (x_i - x_{i+1})^2. \tag{3.5}$$

- Our goal is to get the lower bound of the RHS minimization problem.

Proof of Lemma 3.3

- $\sum_{i=1} x_i^2 = 1$ implies that the average value of x_i^2 is $\frac{1}{m}$
- So there exists some j such that $|x_j| \geq \frac{1}{\sqrt{m}}$. Without loss of generality, let us suppose that this x_j is positive.
- Let us define

$$z_i = x_i - x_{i+1} \text{ for } i < m, \text{ and } z_m = 0.$$

Proof of Lemma 3.3

- Note that $z_i \geq 0$ for all i and

$$\sum_{i=1}^{m} z_i = x_1 - x_m.$$

- Since $x_j \geq \frac{1}{\sqrt{m}}$ for some j, we have that $x_1 \geq \frac{1}{\sqrt{m}}$ ([x is arrange in a decreasing order])

- Since $\sum_{i=1}^{m} x_i = 0$, it follows that at least one x_i is negative, and therefore $x_m < 0$. This gives us

$$\sum_{i=1}^{m} z_i \geq \frac{1}{\sqrt{m}}$$

Proof of Lemma 3.3

- Combining with equation (3.5), we obtain

$$\frac{V(kB) - V((k+1)B)}{V(kB)} \geq \frac{\eta}{2} \min_{z_i \geq 0, \sum_{i=1}^m z_i \geq 1/\sqrt{m}} \sum_{i=1}^m z_i^2$$

- The minimization problem on the right-hand side is a symmetric convex optimization problem, and therefore has a symmetric optimal solution, namely $z_i = \frac{1}{m^{1.5}}$ for all i. This results in an optimal value of $\frac{1}{m}$.

- Therefore, we get the desired result

$$\frac{V(kB) - V((k+1)B)}{V(kB)} \geq \frac{\eta}{2m^2}.$$

$\square$

The Next Step

- Using Lemma 3.3, we can establish the convergence of $\Phi(k, s)$ in (1.3) to the matrix with all entries equal to $\frac{1}{m}$.

- We can further show that the difference between the entries of $\Phi(k, s)$ and $\frac{1}{m}$ converges to zero geometrically fast.

Convergence of Transition Matrices to Averages

Theorem 3.4

Let Assumption 1-2 hold, for all i, j and all k, s with $k \geq s$, we have

$$\left|[\Phi(k,s)]_{ij} - \frac{1}{m}\right| \leq \left(1 - \frac{\eta}{4m^2}\right)^{\lceil \frac{k-s+1}{B} \rceil - 2}.$$

Proof: By Lemma 3.3, we have for all $k \geq s$,

$$V(kB) \leq (1 - \frac{\eta}{2m^2})^{k-s} V(sB).$$

Let k and s be arbitrary with $k \geq s$, and let

$$\tau B \leq s < (\tau + 1)B, \quad tB \leq k < (t+1)B,$$

with $\tau \leq t$ (That is, s in 'period' τ and k in 'period' t).

Proof of Theorem 3.4

By the descent property of $V(k)$, we have

$$\begin{aligned} V(k) &\leq V(tB) \\ &\leq (1 - \frac{\eta}{2m^2})^{t-\tau-1} V((\tau+1)B) \\ &\leq (1 - \frac{\eta}{2m^2})^{t-\tau-1} V(s) \end{aligned}$$

Note that $k - s < (t-\tau)B + B$ implying that $\frac{k-s+1}{B} \leq t - \tau + 1$, where we used the fact that both sides of the inequality are integers. Therefore $\lceil \frac{k-s+1}{B} \rceil - 2 \leq t - \tau - 1$, and we have for all k and s with $k \geq s$,

$$V(k) \leq V(s)(1 - \frac{\eta}{2m^2})^{\lceil \frac{k-s+1}{B} \rceil - 2} \tag{3.6}$$

Proof of Theorem 3.4

- Recall that $V(k)$ represents the consensus violation at iteration k
- So the result we just derived:

$$V(k) \leq V(s)(1 - \frac{\eta}{2m^2})^{\lceil \frac{k-s+1}{B} \rceil - 2} \tag{3.7}$$

  implies that the consensus violation contracts and will converge to zero
- The next step is relatively simple, we will show that the entires of $\Phi(k, s) \to 1/m$ as $k \to \infty$

Proof of Theorem 3.4

- Define a new sequence $\mathbf{z}(k+1) = W(k)\mathbf{z}(k)$, we have

$$\mathbf{z}(k+1) = \Phi(k,s)\mathbf{z}(s), \ \forall k \geq s.$$

- Note that for this sequence, Lemma 3.3 still holds, because this lemma only depends on the properties of $\Phi(k,s)$

- Let $e_i \in \mathbb{R}^m$ denote the vector with entries all equal to 0, except for the ith entry which is equal to 1.

- Letting $\mathbf{z}(s) = e_i$ we obtain $\mathbf{z}(k+1) = [\Phi(k,s)']_i$, where $[\Phi(k,s)']_i$ denotes the ith row of the transpose of the matrix. Using the inequalities (3.6) and $V(e_i) \leq 1$, we obtain

$$V([\Phi(k,s)']_i) \leq (1 - \frac{\eta}{2m^2})^{\lceil \frac{k-s+1}{B} \rceil - 2}$$

Proof of Theorem 3.4

The matrix $\Phi(k,s)$ is doubly stochastic. Thus, the average entry of $[\Phi(k,s)']_i$ is $\frac{1}{m}$.
So by the definition of $V(\cdot)$, for all i and j,

$$\begin{aligned}\left([\Phi(k,s)]_{ji} - \frac{1}{m}\right)^2 &\leq V([\Phi(k,s)']_i)\\ &\leq (1-\frac{\eta}{2m^2})^{\lceil \frac{k-s+1}{B} \rceil - 2}.\end{aligned}$$

From the preceding relation and $\sqrt{1-\eta/(2m^2)} \leq 1-\eta/(4m^2)$, we obtain

$$\left|[\Phi(k,s)]_{ji} - \frac{1}{m}\right| \leq (1-\frac{\eta}{4m^2})^{\lceil \frac{k-s+1}{B} \rceil - 2}.$$

$\square$

Summary

- Up to now, we have analyzed the following iteration

$$\boldsymbol{x}(k+1) = W(k)\boldsymbol{x}(k), \quad k = 1, 2, \cdots$$

- Key results
  - The sequence converges to $\bar{x}(0)$
  - After each entire cycle/period B, the distance of the entries of $\Phi(k, s)$ to $1/m$ shrinks by a constant factor $(1 - \eta/(4m^2))$
- Key proof steps
  - Construct a "potential function" $V(k)$, related to consensus violation
  - Show that it shrinks linearly
  - Then translate the definition of $V(k)$ to Φ

Appendix and Additional Proofs

A useful Lemma

Lemma 4.1

Let A be a row-stochastic matrix with positive diagonal entries, and the smallest positive entry in A is at least η. Also, let (S^-, S^+) be a partition of the set $\{1, \cdots, m\}$ into two disjoint sets. If

$$\sum_{i \in S^-, j \in S^+} a_{ij} > 0,$$

then

$$\sum_{i \in S^-, j \in S^+} a_{ij} > \frac{\eta}{2}.$$

Proof of Lemma 4.1

Proof:

- Let $\sum_{i\in S^-,j\in S^+} a_{ij} > 0$. From the definition of the weights a_{ij}, we have $a_{ij} = \sum_k w_{ki}w_{kj}$, which shows that there exist $i \in S^-, j \in S^+$, and some k such that $a_{ki} > 0$ and $a_{kj} > 0$.

- For either case where k belongs to S^- or S^+, we see that there exists an edge in the set $\xi(A)$ that crosses the cut (S^-, S^+). Let (i^*, j^*) be such an edge.

- Without loss of generality, we assume that $i^* \in S^-$ and $j^* \in S^+$.

Proof of Lemma 4.1

We define

$$C_{j^*}^{+} := \sum_{i \in S^+} a_{j^* i}$$

$$C_{j^*}^{-} := \sum_{i \in S^-} a_{j^* i}$$

Since A is a row-stochastic matrix, we have

$$C_{j^*}^{+} + C_{j^*}^{-} = 1,$$

which implying that at least one of the following is true:

$$\text{Case (a)} : C_{j^*}^{-} \geq \frac{1}{2},$$

$$\text{Case (b)} : C_{j^*}^{+} \geq \frac{1}{2},$$

Then we consider these two cases separately.

Proof of Lemma 4.1

Case(a) : $C^-_{j^*} \geq \frac{1}{2}$.

- We focus on those a_{ij} with $i \in S^-$ and $j = j^*$. Indeed, since all a_{ij} are nonnegative, we have

$$\sum_{i \in S^-, j \in S^+} a_{ij} \geq \sum_{i \in S^-} a_{ij^*}. \tag{4.1}$$

- For each element in the sum on the right-hand side, we have

$$a_{ij^*} = \sum_{k=1}^{n} w_{ki} w_{kj^*} \geq w_{j^*i} w_{j^*j^*} \geq w_{j^*i} \eta, \tag{4.2}$$

since the diagonal entries of W are positive, and at least η.

Proof of Lemma 4.1

- Consequently, we have

$$\sum_{i \in S^-} a_{ij^*} \geq \eta \sum_{i \in S^-} w_{j^*i} = \eta C^-_{j^*}. \qquad (4.3)$$

- Combining equations (4.1) and (4.3), also recalling the assumption $C^-_{j^*} \geq \frac{1}{2}$, we get

$$\sum_{i \in S^-, j \in S^+} a_{ij} \geq \frac{\eta}{2}.$$

Proof of Lemma 4.1

Case(b) : $C^+_{j^*} \geq \frac{1}{2}$.

- We focus on those a_{ij} with $i = i^*$ and $j \in S^+$. We have

$$\sum_{i \in S^-, j \in S^+} a_{ij} \geq \sum_{j \in S^+} a_{i^* j}. \tag{4.4}$$

- For each element in the sum on the right-hand side, we have

$$a_{i^* j} = \sum_{k=1}^{n} w_{ki^*} w_{kj} \geq w_{j^* i^*} w_{j^* j} \geq \eta w_{j^* j}, \tag{4.5}$$

since the choice $(i^*, j^*) \in \xi(A)$ implies that $w_{j^* i^*} \geq \eta$.

Proof of Lemma 4.1

- Consequently,

$$\sum_{j \in S^+} a_{i^*j} \geq \eta \sum_{j \in S^+} w_{j^*j} = \eta C_{j^*}^+. \tag{4.6}$$

  Combining equations (4.4) and (4.6), and recalling the assumption $C_{j^*}^+ \geq \frac{1}{2}$, the result follows.

- Thus, the result holds when either case(a) or case(b) happens.

$\square$

Convergence of Transition Matrices to Averages

- To further construct the convergence analysis, we need some connectivity assumptions.

- **Assumption 3**: Given an integer $k \geq 0$, suppose that the components of $\boldsymbol{x}(kB)$ have been reordered so that they are in nonincreasing order. We assume that for every $d \in \{1, \cdots, m-1\}$, we either have $x_d(kB) = x_{d+1}(kB)$, or there exist some time $t \in \{kB, \cdots, (k+1)B-1\}$and some $i \in \{1, \cdots, d\}$, $j \in \{d+1, \cdots, n\}$ such that (i, j) or (j, i) belongs to $\xi(A(t))$.

Through introducing Assumption 3, the strong Assumption 2 can be relaxed.

Convergence of Transition Matrices to Averages

Lemma 4.2

Assumption 2 implies Assumption 3, with the same value of B.

Proof:

- If Assumption 3 does not hold, then there must exist an index d [for which $x_d(kB) \neq x_{d+1}(kB)$ holds] such that there are no edges between nodes $1, 2, \cdots, d$ and nodes $d+1, \cdots, m$ during times $t = kB, \cdots, (k+1)B - 1$.

- But this implies that the graph

$$(\mathcal{N}, \xi(W(kB)) \cup \cdots \cup \xi(W((k+1)B - 1)))$$

is disconnected, which violates Assumption 2. □

Convergence of Transition Matrices to Averages

- For our convergence time results, we will use the weaker Assumption 3, rather than the stronger Assumption 2.

- We now proceed to bound the decrease of the disagreement $V(k)$ during the interval $[kB, (k+1)B - 1]$.

Proof of Lemma 3.2

Proof of Lemma 3.2

- By Lemma 3.1, we have for all t,

$$V(t) - V(t+1) = \sum_{i<j} a_{ij}(t)\left(x_i(t) - x_j(t)\right)^2, \tag{4.7}$$

  where $a_{ij}(t)$ is the (i,j)-th entry of $W(t)^T W(t)$.

- Summing up the variance differences $V(t) - V(t+1)$ over different values of t, we obtain

$$V(kB) - V((k+1)B) = \sum_{t=kB}^{(k+1)B-1} \sum_{i<j} a_{ij}(t)\left(x_i(t) - x_j(t)\right)^2. \tag{4.8}$$

Proof of Lemma 3.2

We next introduce some notation.

- For all $d \in \{1, \cdots, m-1\}$, let t_d be the first time larger than or equal to kB (if it exists) at which there is a communication between two nodes belonging to the two sets $\{1, \cdots, d\}$ and $\{d+1, \cdots, m\}$.

- For all $t \in \{kB, \cdots, (k+1)B-1\}$, let $D(t) = \{d | t_d = t\}$, i.e., $D(t)$ consists of "cuts" $d \in \{1, \cdots, m-1\}$ such that time t is the first communication time larger than or equal to kB between nodes in the sets $\{1, ..., d\}$ and $\{d+1, ..., n\}$. Because of Assumption 3, the union of the sets $D(t)$ includes all indices $1, \cdots, m-1$, except possibly for indices for which $x_d(kB) = x_{d+1}(kB)$.

Proof of Lemma 3.2

- For all $d \in \{1, \cdots, n-1\}$, let $C_d = \{(i,j), (j,i) | i \leq d, d+1 \leq j\}$.
- For all $t \in \{kB, \cdots, (k+1)B - 1\}$, let

$$F_{ij}(t) = \{d \in D(t) | (i,j) \text{ or } (j,i) \in C_d\}$$

  That is, $F_{ij}(t)$ consists of all cuts d such that the edge (i,j) or (j,i) at time t is the first communication across the cut at a time larger than or equal to kB.

- To simplify notation, let $y_i = x_i(kB)$. By assumption, we have $y_1 \geq \cdots \geq y_m$.

Proof of Lemma 3.2

With these notations, we make two obsevations.
Observation 1: Suppose that $d \in D(t)$. Then, for some $(i,j) \in C_d$, we have either $a_{ij}(t) > 0$ or $a_{ji}(t) > 0$. Because $A(t)$ is nonnegative with positive diagonal entries, we have

$$a_{ij}(t) = \sum_{k=1}^{n} w_{ki} w_{kj} \geq w_{ii}(t) w_{ij}(t) + w_{ji}(t) w_{jj}(t) > 0,$$

and by Lemma 4.1, we obtain

$$\sum_{(i,j) \in C_d} a_{ij}(t) \geq \frac{\eta}{2}. \qquad (4.9)$$

Proof of Lemma 3.2

Observation 2:

- Fix some (i, j), with $i < j$, and time $t \in \{kB, \cdots, (k+1)B - 1\}$, and suppose that $F_{ij}(t)$ is nonempty. Let $F_{ij}(t) = \{d_1, \cdots, d_k\}$, where the d_j are arranged in increasing order.

- Since $d_1 \in F_{ij}(t)$, we have $d_1 \in D(t)$ and therefore $t_{d_1} = t$. By the definition of t_{d_1} , this implies that there has been no communication between a node in $\{1, \cdots, d_1\}$ and a node in $\{d_1 + 1, \cdots, n\}$ during the time interval $[kB, t - 1]$.
  It follows that $x_i(t) \geq y_{d_1}$.

- By a symmetrical argument, we also have $x_j(t) \leq y_{d_k+1}$.

Proof of Lemma 3.2

These relations imply that

$$x_i(t) - x_j(t) \geq y_{d_1} - y_{d_k+1} \geq \sum_{d \in F_{ij}(t)} \left(y_d - y_{d+1}\right).$$

Since the components of y are sorted in nonincreasing order, we have $y_d - y_{d+1}$, for every $d \in F_{ij}(t)$. For any nonnegative numbers z_i, we have

$$\left(z_1 + \cdots z_k\right)^2 \geq z_1^2 + \cdots + z_k^2,$$

which implies that

$$\left(x_i(t) - x_j(t)\right)^2 \geq \sum_{d \in F_{ij}(t)} \left(y_d - y_{d+1}\right)^2. \tag{4.10}$$

Proof of Lemma 3.2

We now use these two observations to provide a lower bound on the expression on the right-hand side of equation (4.7) at time t. We use equation (4.10) and then (4.9), to obtain

$$\begin{aligned}
\sum_{i<j} a_{ij}(t)\left(x_i(t)-x_j(t)\right)^2 &\geq \sum_{i<j} a_{ij}(t) \sum_{d\in F_{ij}(t)} \left(y_d - y_{d+1}\right)^2 \\
&= \sum_{d\in D(t)} \sum_{(i,j)\in C_d} a_{ij}(t)(y_d - y_{d+1})^2 \\
&\geq \frac{\eta}{2} \sum_{d\in D(t)} \left(y_d - y_{d+1}\right)^2 .
\end{aligned}$$

Proof of Lemma 3.2

We now sum both sides of the above inequality for different values of t, and use equation (4.8), to obtain

$$
\begin{aligned}
V(kB) - V((k+1)B) &= \sum_{t=kB}^{(k+1)B-1} \sum_{i<j} a_{ij}(t)\left(x_i(t) - x_j(t)\right)^2 \\
&\geq \frac{\eta}{2} \sum_{t=kB}^{(k+1)B-1} \sum_{d \in D(t)} \left(y_d - y_{d+1}\right)^2 \\
&= \frac{\eta}{2} \sum_{d=1}^{m-1} \left(y_d - y_{d+1}\right)^2,
\end{aligned}
$$

where the last inequality follows from the fact that the union of the sets $D(t)$ is only missing those d for which $y_d = y_{d+1}$. $\square$