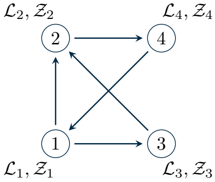
# Decentralized Optimization for ML

Presented on August 3, 2021

# Collaborative Statistical Machine Learning

**Empirical risk minimization:**

$$\theta^\star \in \operatorname*{argmin}_{\theta} \; \frac{1}{m}\sum_{i=1}^{m} f_i(\theta; \mathcal{Z}_i)$$




**Full Data:** realizations $(x, y) \in \mathcal{Z}$.

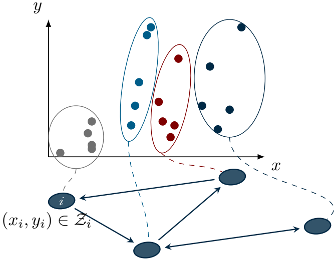
**Agent Partition:** $\mathcal{Z} = \mathcal{Z}_1 \cup \mathcal{Z}_2 \cdots \cup \mathcal{Z}_m$; $\mathcal{Z}_i$: data of agent $i$.

**Model:** $h_\theta$ such that $h_\theta(x) \approx y$.

**Local Loss:** $f_i(\theta) = \frac{1}{|\mathcal{Z}_i|} \sum_{(x,y)\in \mathcal{Z}_i} \ell(h_\theta(x), y)$

# Example: Decentralized Nonlinear Fitting

**Data:** $(x_i, y_i)$

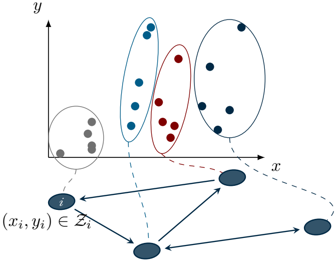


# Example: Decentralized Nonlinear Fitting

**Data:** $(x_i, y_i)$

**Model:**

$$h_\theta(x) = \theta_1 \cdot x^2 + \theta_2 \cdot x + \theta_3$$

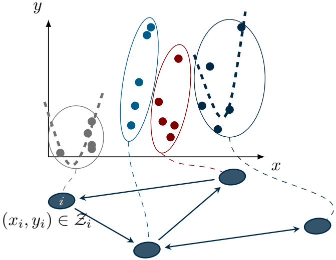


# Example: Decentralized Nonlinear Fitting

**Data:** $(x_i, y_i)$

**Model:**

$$h_\theta(x) = \theta_1 \cdot x^2 + \theta_2 \cdot x + \theta_3$$




**Loss Function:**

$$\ell(\theta) = \frac{1}{2}\big(y - h_\theta(x)\big)^2$$

**Local Loss:**

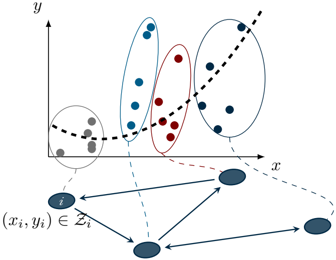
$f_{\color{red}i}(\theta, \mathcal{Z}_i) = \frac{1}{|\mathcal{Z}_i|} \sum_{(x,y)\in {\color{red}\mathcal{Z}_i}} \frac{1}{2}(y - h_\theta(x))^2$

# Example: Decentralized Nonlinear Fitting

**Data:** $(x_i, y_i)$

**Model:**

$$h_\theta(x) = \theta_1 \cdot x^2 + \theta_2 \cdot x + \theta_3$$




**Loss Function:**

$$\ell(\theta) = \frac{1}{2}\big(y - h_\theta(x)\big)^2$$

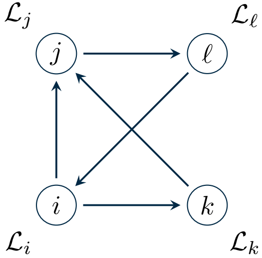
**Local Loss:**

$f_{i}(\theta, \mathcal{Z}_i) = \frac{1}{|\mathcal{Z}_i|} \sum_{(x,y)\in\mathcal{Z}_i} \frac{1}{2}(y - h_\theta(x))^2$

$$\theta^\star \in \operatorname*{argmin}_\theta \frac{1}{m} \sum_{i=1}^{m} f_{i}(\theta, \mathcal{Z}_i)$$

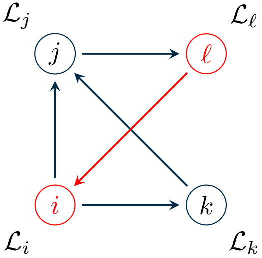
# Network Model

Dynamic network topology: Agents are embedded in a *time-varying directed* communication *graph* with general topology



# Network Model

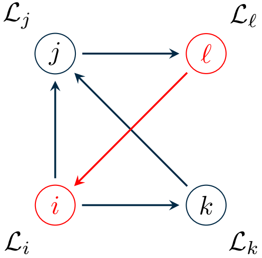
Dynamic network topology: Agents are embedded in a *time-varying directed* communication *graph* with general topology




$$\mathcal{N}_i^{in} \triangleq \{\text{agents send info. to } i\} \cup \{i\}$$

# Network Model

Dynamic network topology: Agents are embedded in a *time-varying directed* communication *graph* with general topology



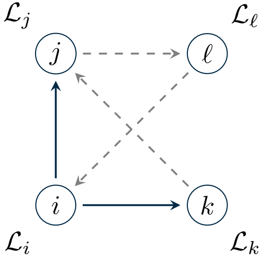

$$\mathcal{N}_i^{in} \triangleq \{\text{agents send info. to } i\} \cup \{i\}$$

**Assumptions on the network & agents' knowledge**

- Local information: each agent $i$ knows its $f_i$ but not $\sum_{j\neq i} f_j$
- Local communications: agent $i$ can receive information from its "neighbors"
- Long term connectivity: $T$-strongly connected digraphs

# Network Model

Dynamic network topology: Agents are embedded in a *time-varying directed* communication *graph* with general topology



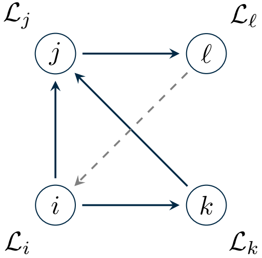

$$\mathcal{N}_i^{in} \triangleq \{\text{agents send info. to } i\} \cup \{i\}$$

**Assumptions on the network & agents' knowledge**

- Local information: each agent $i$ knows its $f_i$ but not $\sum_{j \neq i} f_j$
- Local communications: agent $i$ can receive information from its "neighbors"
- Long term connectivity: $T$-strongly connected digraphs

# Network Model

Dynamic network topology: Agents are embedded in a *time-varying directed* communication *graph* with general topology



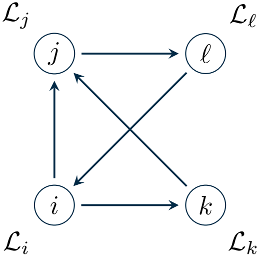

$$\mathcal{N}_i^{in} \triangleq \{\text{agents send info. to } i\} \cup \{i\}$$

**Assumptions on the network & agents' knowledge**

- Local information: each agent $i$ knows its $f_i$ but not $\sum_{j\neq i} f_j$
- Local communications: agent $i$ can receive information from its "neighbors"
- Long term connectivity: $T$-strongly connected digraphs

# Network Model

Dynamic network topology: Agents are embedded in a *time-varying directed* communication *graph* with general topology




$$\mathcal{N}_i^{in} \quad \triangleq \quad \{\text{agents send info. to } i\} \cup \{i\}$$

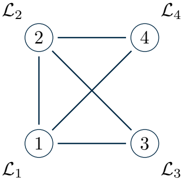
**Assumptions on the network & agents' knowledge**

- Local information: each agent $i$ knows its $f_i$ but not $\sum_{j \neq i} f_j$
- Local communications: agent $i$ can receive information from its "neighbors"
- Long term connectivity: $T$-strongly connected digraphs

# Decentralized Gradient Descent

Empirical risk minimization:

$$\min_{\theta} \left\{ f^N(\theta) = \frac{1}{m} \sum_{i=1}^{m} f_i(\theta) \right\} \qquad \text{(P)}$$

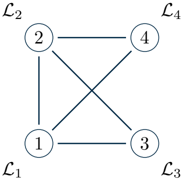



---

$\theta_i$: local copy of $\theta$

Two objectives: consensus and optimality

# Decentralized Gradient Descent




Empirical risk minimization:

$$\min_{\theta} \left\{ f^N(\theta) = \frac{1}{m} \sum_{i=1}^{m} f_i(\theta) \right\} \qquad \text{(P)}$$

---

$\theta_i$: local copy of $\theta$

Two objectives: consensus and optimality

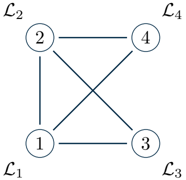
- consensus: $$\theta_i^{k+1} = \sum_{j \in \mathcal{N}_i^{in}} w_{ij} \theta_j^k$$

- perturbation:

# Decentralized Gradient Descent

Empirical risk minimization:

$$\min_{\theta} \left\{ f^N(\theta) = \frac{1}{m} \sum_{i=1}^{m} f_i(\theta) \right\} \qquad \text{(P)}$$




---

$\theta_i$: local copy of $\theta$

Two objectives: consensus and optimality

- consensus:
$$\theta_i^{k+1} = \sum_{j \in \mathcal{N}_i^{in}} w_{ij} \theta_j^k$$
- perturbation:
$$\theta_i^{k+\frac{1}{2}} = \sum_{j \in \mathcal{N}_i^{in}} w_{ij} \theta_j^k - \gamma^k \cdot \nabla f_i(\theta_i^k)$$
- dilemma: ($\gamma^k \downarrow 0$: sublinear rate) vs. ($\gamma^k \equiv \gamma$: linear rate but $\mathcal{N}_\epsilon(\theta^*)$).

# Speed Accuracy Dilemma

Assume for simplicity $d = 1$.

Notations:

- Consensus matrix: $W = \{w_{ij}\}$
- Stacked local variables: $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_m]^\top$
- Pseudo gradient: $\nabla F(\boldsymbol{\theta}) = [\nabla f_1(\theta_1), \ldots, \nabla f_m(\theta_m)]^\top$

DGD in matrix form: $\boldsymbol{\theta}^{k+1} = W\boldsymbol{\theta}^k - \gamma \cdot \nabla F(\boldsymbol{\theta}^k)$

Sanity check:

- suppose $\boldsymbol{\theta}^k \to \boldsymbol{\theta}^*$ (convergence) and $\theta_i^* = \theta_j^*$ (consensus)
- $\Rightarrow \nabla f_i(\theta^*) = 0$ for all $i = 1, \ldots, m$.
- cannot achieve both consensus and optimality with constant $\gamma$.

# Speed Accuracy Dilemma

Assume for simplicity $d = 1$.

Notations:

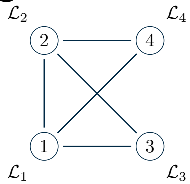
- Consensus matrix: $W = \{w_{ij}\}$
- Stacked local variables: $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_m]^\top$
- Pseudo gradient: $\nabla F(\boldsymbol{\theta}) = [\nabla f_1(\theta_1), \ldots, \nabla f_m(\theta_m)]^\top$

DGD in matrix form: $\boldsymbol{\theta}^{k+1} = W\boldsymbol{\theta}^k - \gamma \cdot \nabla F(\boldsymbol{\theta}^k)$ needs correction

Sanity check:

- suppose $\boldsymbol{\theta}^k \to \boldsymbol{\theta}^*$ (convergence) and $\theta_i^* = \theta_j^*$ (consensus)
- $\Rightarrow \nabla f_i(\theta^*) = 0$ for all $i = 1, \ldots, m$.
- cannot achieve both consensus and optimality with constant $\gamma$.

# Decentralized Gradient Tracking




Empirical risk minimization:

$$\min_{\theta} \left\{ f^N(\theta) = \frac{1}{m} \sum_{i=1}^{m} f_i(\theta) \right\} \qquad \text{(P)}$$

---

- correct direction:

$$\theta_i^{k+\frac{1}{2}} = \sum_{j \in \mathcal{N}_i^{in}} w_{ij} \theta_j^k - \gamma^k \cdot \cancel{\nabla f_i(\theta_i^k)} \quad g_i^k \to \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\theta_i^k)$$

- gradient tracking:

$$g_i^{k+1} = \sum_{j \in \mathcal{N}_i^{in}} w_{ij} (g_j^k + \nabla f_j(\theta_j^{k+1}) - \nabla f_j(\theta_j^k))$$

# How Tracking Works?

In vector form:

$$\mathbf{g}^{k+1} = W(\mathbf{g}^k + \nabla F(\boldsymbol{\theta}^{k+1}) - \nabla F(\boldsymbol{\theta}^k))$$

$W$ is doubly stochastic:

- Consensus forcing $W1 = 1$
- Sum preserving $1^\top W = 1^\top$

Taking sum:

$$\begin{aligned} 1^\top \mathbf{g}^{k+1} &= 1^\top W(\mathbf{g}^k + \nabla F(\boldsymbol{\theta}^{k+1}) - \nabla F(\boldsymbol{\theta}^k)) \\ &= 1^\top(\mathbf{g}^k + \nabla F(\boldsymbol{\theta}^{k+1}) - \nabla F(\boldsymbol{\theta}^k)) \end{aligned}$$

Initialize $\mathbf{g}^0 = \nabla F(\theta^0)$, then $1^\top \mathbf{g}^k = 1^\top \nabla F(\boldsymbol{\theta}^k)$.

If $\theta_i$'s and $g_i$'s are consensual, then $g_i^k \to \nabla f^N(\theta_i^k)$.

# Convergence Proof

**Assumption:** Each $\nabla f_i$ is $L$-smooth, $\rho \triangleq \sigma(W - J) \leq 1$.

DGT in matrix form:

$$\begin{aligned}
\boldsymbol{\theta}^{k+1} &= W\boldsymbol{\theta}^k - \gamma \cdot \mathbf{g}^k \\
\mathbf{g}^{k+1} &= W(\mathbf{g}^k + \nabla F(\boldsymbol{\theta}^{k+1}) - \nabla F(\boldsymbol{\theta}^k))
\end{aligned}$$

The average process:

$$\begin{aligned}
\bar{\theta}^{k+1} &= \bar{\theta}^k - \gamma \cdot \bar{g}^k \\
&= \bar{\theta}^k - \gamma \cdot \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\theta_i^k) \qquad \text{(tracking property)}
\end{aligned}$$

The average process can be viewed as the inexact centralized GD on $\bar{\theta}^k$

# GD Proof Recap

Gradient iteration:

$$\theta^{k+1} = \theta^k - \gamma \cdot \nabla f^N(\theta^k)$$

Apply descent lemma

$$\begin{aligned} f^N(\theta^{k+1}) &\leq f^N(\theta^k) + \nabla f^N(\theta^k)^\top(\theta^{k+1} - \theta^k) + \frac{L}{2}\|\theta^{k+1} - \theta^k\|^2 \\ &= f^N(\theta^k) - \underbrace{\gamma \cdot \|\nabla f^N(\theta^k)\|^2}_{O(\gamma)} + \underbrace{\frac{\gamma^2 L}{2}\|\nabla f^N(\theta^k)\|^2}_{O(\gamma^2)} \end{aligned}$$

By the monotone convergence theorem: if $\gamma < \frac{2}{L}$, then

- $\{f^N(\theta^k)\}$ converges
- $\|\nabla f^N(\theta^k)\| \to 0$

# Main Steps

**Step 1:** Descent on the average

$$f^N(\bar{\theta}^{k+1}) \leq f^N(\bar{\theta}^k) - \frac{1}{m}\gamma\left(1-\gamma L\right)\|\mathbf{g}^k\|^2 + \underbrace{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2L}\|\nabla f^N(\bar{\theta}^k) - g_i^k\|^2}_{\text{tracking error}}.$$

**Step 2:** Bounding tracking error

$$\|\nabla f^N(\bar{\theta}^k) - g_i^k\| \leq \frac{1}{m}\sum_{j=1}^{m} L\|\bar{\theta}^k - \theta_j^k\| + \|\bar{g}^k - g_i^k\|$$

**Step 3:** Bounding consensus error

$$\|\bar{\boldsymbol{\theta}}^{k+1} - \boldsymbol{\theta}^{k+1}\| \leq \rho\|\bar{\boldsymbol{\theta}}^k - \boldsymbol{\theta}^k\| + \gamma\|\bar{\mathbf{g}}^k - \mathbf{g}^k\|$$
$$\|\bar{\mathbf{g}}^{k+1} - \mathbf{g}^{k+1}\| \leq \rho\|\bar{\mathbf{g}}^k - \mathbf{g}^k\| + 2\rho L\|\bar{\boldsymbol{\theta}}^k - \boldsymbol{\theta}^k\| + \gamma\rho L\|\mathbf{g}^k\|$$

Consequence: tracking error $= O(\gamma^2\|\mathbf{g}^k\|^2) \Rightarrow$ descent if $\gamma$ is small enough

# Inexact Gradient Descent

Inexact gradient descent:

$$\bar{\theta}^{k+1} = \bar{\theta}^k - \gamma \cdot \bar{g}^k$$

By the descent lemma

$$\begin{aligned}
f^N(\bar{\theta}^{k+1}) &\leq f^N(\bar{\theta}^k) + \nabla f^N(\bar{\theta}^k)^\top(\bar{\theta}^{k+1} - \bar{\theta}^k) + \frac{L}{2}\|\bar{\theta}^{k+1} - \bar{\theta}^k\|^2 \\
&= f^N(\bar{\theta}^k) - \gamma \nabla f^N(\bar{\theta}^k)^\top \bar{g}^k + \frac{\gamma^2 L}{2}\|\bar{g}^k\|^2 \\
&\leq f^N(\bar{\theta}^k) - \gamma \cdot \frac{1}{m}\sum_{i=1}^m \nabla f^N(\bar{\theta}^k)^\top g_i^k + \frac{1}{m}\sum_{i=1}^m \frac{\gamma^2 L}{2}\|g_i^k\|^2
\end{aligned}$$

If $\nabla f^N(\bar{\theta}^k)$ were equal to $g_i^k$ then we are done. But it's not that bad...

Remember we are constructing $g_i$ to track $\nabla f^N(\bar{\theta}^k)$

# Inexact Gradient Descent (Cont.)

Descent Lemma

$$f^N(\bar{\theta}^{k+1}) \leq f^N(\bar{\theta}^k) - \gamma \cdot \frac{1}{m}\sum_{i=1}^{m}(\nabla f^N(\bar{\theta}^k) \pm g_i^k)^\top g_i^k + \frac{1}{m}\sum_{i=1}^{m}\frac{\gamma^2 L}{2}\|g_i^k\|^2$$

$$\leq \underbrace{f^N(\bar{\theta}^k) - \frac{1}{m}\sum_{i=1}^{m}\left(\gamma\|g_i^k\|^2 - \frac{\gamma^2 L}{2}\|g_i^k\|^2\right)}_{\text{seen before}}$$

$$\underbrace{-\gamma\frac{1}{m}\sum_{i=1}^{m}(\nabla f^N(\bar{\theta}^k) - g_i^k)^\top g_i^k}_{\text{error term}}$$

$$\leq f^N(\bar{\theta}^k) - \frac{1}{m}\left(\gamma\|\mathbf{g}^k\|^2 - \frac{\gamma^2 L}{2}\|\mathbf{g}^k\|^2\right) + \frac{\gamma}{m}\sum_{i=1}^{m}\|g_i^k\|\|\nabla f^N(\bar{\theta}^k) - g_i^k\|$$

Split the product $\left(2ab \leq a^2 + b^2\right)$

$$\frac{\gamma}{m}\sum_{i=1}^{m}\|g_i^k\|\|\nabla f^N(\bar{\theta}^k) - g_i^k\| \leq \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\gamma^2 L}{2}\|g_i^k\|^2 + \frac{1}{2L}\|\nabla f^N(\bar{\theta}^k) - g_i^k\|^2\right)$$

# Bounding Tracking Error

Inequality for descent:

$$f^N(\bar{\theta}^{k+1}) \leq f^N(\bar{\theta}^k) - \frac{1}{m}\gamma\,(1-\gamma L)\,\|\mathbf{g}^k\|^2 + \underbrace{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2L}\|\nabla f^N(\bar{\theta}^k) - g_i^k\|^2}_{\text{tracking error}}.$$

Facts:

- We used consensus to force $g_i^k \to \bar{g}^k$
- $\bar{g}^k = \frac{1}{m}\sum_{i=1}^m \nabla f_i(\theta_i^k)$
- We used consensus to force $\theta_i^k \to \bar{\theta}^k$

Split the terms accordingly

$$\begin{aligned}
\|\nabla f^N(\bar{\theta}^k) - g_i^k\| &\leq \|\nabla f^N(\bar{\theta}^k) - \bar{g}^k\| + \|\bar{g}^k - g_i^k\| \\
&= \left\|\frac{1}{m}\sum_{j=1}^{m}\nabla f_j(\bar{\theta}^k) - \nabla f_j(\theta_j^k))\right\| + \|\bar{g}^k - g_i^k\| \\
&\leq \frac{1}{m}\sum_{j=1}^{m} L\|\bar{\theta}^k - \theta_j^k\| + \|\bar{g}^k - g_i^k\|
\end{aligned}$$

# Bounding Consensus Error ($\theta$ part)

Introduce averaging matrix $J = \frac{1}{m}11^\top$

- $\bar{\boldsymbol{\theta}}^k = J\boldsymbol{\theta}^k$
- $JW = J$
- $\|J - W\|_2 \le \rho$ (graph is connected)

Then we can bound consensus error on $\theta$ as

$$\begin{aligned}
\bar{\boldsymbol{\theta}}^{k+1} - \boldsymbol{\theta}^{k+1} &= (J - I)\boldsymbol{\theta}^{k+1} \\
&= (J - I)(W\boldsymbol{\theta}^k - \gamma \cdot \mathbf{g}^k) \\
&= (J - W)\boldsymbol{\theta}^k - \gamma \quad \underbrace{(J - I)\mathbf{g}^k}_{\text{consensus error of } g_i^k}
\end{aligned}$$

Taking $\ell_2$ norm:

$$\|\bar{\boldsymbol{\theta}}^{k+1} - \boldsymbol{\theta}^{k+1}\| \le \rho\|\bar{\boldsymbol{\theta}}^k - \boldsymbol{\theta}^k\| + \gamma\|\bar{\mathbf{g}}^k - \mathbf{g}^k\|$$

# Bounding Consensus Error ($g$ part)

Now we need to bound $\bar{\mathbf{g}}^k - \mathbf{g}^k$.

Tracking dynamics:

$$\mathbf{g}^{k+1} = W(\mathbf{g}^k + \nabla F(\boldsymbol{\theta}^{k+1}) - \nabla F(\boldsymbol{\theta}^k))$$

Multiplying by $J - W$:

$$\begin{aligned}
\|\bar{\mathbf{g}}^{k+1} - \mathbf{g}^{k+1}\| &= \|(J - W)\left(\mathbf{g}^k + \nabla F(\boldsymbol{\theta}^{k+1}) - \nabla F(\boldsymbol{\theta}^k)\right)\| \\
&\leq \rho\|\bar{\mathbf{g}}^k - \mathbf{g}^k\| + \rho L\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\| \\
&\leq \rho\|\bar{\mathbf{g}}^k - \mathbf{g}^k\| + \rho L\|W\boldsymbol{\theta}^k - \gamma\mathbf{g}^k - \boldsymbol{\theta}^k\| \\
&\leq \rho\|\bar{\mathbf{g}}^k - \mathbf{g}^k\| + \rho L\|(W - J)\boldsymbol{\theta}^k\| + \gamma\rho L\|\mathbf{g}^k\| + \rho L\|(J - I)\boldsymbol{\theta}^k\| \\
&\leq \rho\|\bar{\mathbf{g}}^k - \mathbf{g}^k\| + 2\rho L\|\bar{\boldsymbol{\theta}}^k - \boldsymbol{\theta}^k\| + \gamma\rho L\|\mathbf{g}^k\|
\end{aligned}$$

HW: complete the proof

# Directed Graph

DGT in matrix form

$$\boldsymbol{\theta}^{k+1} = W\boldsymbol{\theta}^k - \gamma \cdot \mathbf{g}^k$$
$$\mathbf{g}^{k+1} = W(\mathbf{g}^k + \nabla F(\boldsymbol{\theta}^{k+1}) - \nabla F(\boldsymbol{\theta}^k))$$

Doubly stochastic $W$: generally requires graph undirected

Do we really need double stochasticity?

Properties we need are

- consensus of $\theta_i \Rightarrow W$ being row stochastic
- each $g_i \propto \frac{1}{m}\sum_{i=1}^m \nabla f_i(\theta_i) \Rightarrow W$ being column stochastic

But we can use two matrices to split the work!

# The Push-Pull Algorithm

Row stochastic $R$ and column stochastic $C$:

$$
\begin{aligned}
\boldsymbol{\theta}^{k+1} &= R\boldsymbol{\theta}^k - \gamma \cdot \mathbf{g}^k \\
\mathbf{g}^{k+1} &= C(\mathbf{g}^k + \nabla F(\boldsymbol{\theta}^{k+1}) - \nabla F(\boldsymbol{\theta}^k))
\end{aligned}
$$

Implementation:

- Pull $\theta_i$ from in-neighbors and then averages
- Split $g_i$ and push to out-neighbors
- can be adapted to time-varying network if knowing the $\#$out-neighbors

# Supplemental - DGD

DGD iterate: $\boldsymbol{\theta}^{k+1} = W\boldsymbol{\theta}^k - \gamma^k \cdot \nabla F(\boldsymbol{\theta}^k)$

**The average process:**

$$\bar{\theta}^{k+1} = \bar{\theta}^k - \gamma^k \cdot \underbrace{\frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\theta_i^k)}_{\approx \nabla F(\bar{\theta}^k)}$$

We almost have a centralized gradient step performed on the average.

**The consensus process:** $\boldsymbol{\theta}^{k+1} = W\boldsymbol{\theta}^k - \underbrace{\gamma^k \cdot \nabla F(\boldsymbol{\theta}^k)}_{\text{diminishing perturbation}}$

**A key assumption:** $\|\nabla f_i(\theta)\|$ is uniformly bounded

**Consequences:** Optimization and consensus can be analyzed separately

- Consensus is achieved as long as perturbation diminishes
- Optimality is achieved since the inexact error will vanish

# Supplemental - DGD

Decentralized reformulation

$$\begin{aligned} \min_{\{\theta_i\}_{i=1}^m} \quad & \frac{1}{m}\sum_{i=1}^m f_i(\theta_i) \\ \text{s.t.} \quad & \theta_i = \theta_j. \quad \Longleftrightarrow \quad W\boldsymbol{\theta} = \boldsymbol{\theta} \end{aligned}$$

The penalized problem

$$\min_{\boldsymbol{\theta}} \; \sum_{i=1}^m f_i(\theta_i) + \frac{1}{2\gamma}\|\boldsymbol{\theta}\|_{I-W}^2$$

GD with step size $\gamma$: $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \gamma \cdot (\nabla F(\boldsymbol{\theta}^k) + \gamma^{-1}(I - W)\boldsymbol{\theta}^k)$

Consequences:

- Convergence rate analysis of GD applies directly
- Converge to a neighborhood of $\theta^\star$