# Row-Stochastic Matrices Can Provably Outperform Doubly Stochastic Matrices in Decentralized Learning

Bing Liu, Boao Kong, Limin Lu, Kun Yuan, Chengcheng Zhao†

ICML International Conference On Machine Learning

Paper

## Motivation

### General Problem Formulation:

$$\min_{\theta\in\mathbb{R}^d} F(\theta) = \frac{1}{n}\sum_{i=1}^{n} F_i(\theta) := \mathbb{E}_{\xi_{i,j}\sim\mathcal{D}_i}\left[f_i(\theta,\xi_{i,j})\right]$$

- Doubly–stochastic matrix is the standard paradigm
- Row–stochastic matrix will harm the optimization

### The Weighted Problem:

$$\min_{\theta\in\mathbb{R}^d} F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i F_i(\theta) := \mathbb{E}_{\xi_{i,j}\sim\mathcal{D}_i}\left[f_i(\theta,\xi_{i,j})\right]$$

### Two Strategies:

- **Strategy I: weighted local losses**
  Absorb $\lambda_i$ into local losses –> doubly stochasticity

$$\frac{\mathbf{1}^\top}{n}W^{\mathrm{ds}} = \frac{\mathbf{1}^\top}{n} \Longrightarrow \bar{\theta}^{(t+1)} = \bar{\theta}^{(t)} - \frac{\alpha}{n}\sum_{i=1}^{n}\lambda_i\nabla F_i(\theta_i^{(t)})$$

- **Strategy II: weighted mixing matrix**
  Absorb $\lambda_i$ into mixing matrix –> row–stochasticity

$$\frac{\lambda^\top}{n}W = \frac{\lambda^\top}{n} \Longrightarrow \bar{\theta}_\lambda^{(t+1)} = \bar{\theta}_\lambda^{(t)} - \frac{\alpha}{n}\sum_{i=1}^{n}\lambda_i\nabla F_i(\theta_i^{(t)})$$

### Questions and Contributions:

- **Is Euclidean analysis tight?**
  **No.** With heterogeneous node weights, the natural geometry is the weighted Hilbert space $L^2(\lambda;\mathbb{R}^d)$, which gives strictly tighter convergence bounds.
- **Is the gap only due to spectral gaps?**
  **No.** In $L^2(\lambda;\mathbb{R}^d)$, the $\lambda$–induced row–stochastic matrix is self–adjoint, while the doubly stochastic matrix is not. This creates extra consensus penalty terms for Strategy I.
- **When should we prefer weighted mixing?**
  **Strategy II** can be better even with a smaller spectral gap. Our conditions translate into a simple topology rule: $d_i \propto \lambda_i$.

### Row–Stochasitc Mixing:

$$W_{i,j} = \begin{cases} \frac{1-\varepsilon}{d_i}\min\left(1, \frac{\lambda_j d_i}{\lambda_i d_j}\right), & \text{if } i\neq j \text{ and } j\in\mathcal{N}_i,\\ 1-\sum_{k\in\mathcal{N}_i} W_{i,k}, & \text{if } i=j,\\ 0, & \text{otherwise.}\end{cases}$$

Metropolis–Hastings (MH) rule

Setting $\lambda = \mathbf{1}^\top$ recovers doubly stochastic mixing.

## Algorithm and Framework

### Standard Single–Loop Gradient Tracking:

**Algorithm 1** Weighted Decentralized Gradient Tracking

**Input:** step size $\alpha$, batch sizes $\{b_i\}$, total iterations $T$, and system matrices $W_\lambda, G_\lambda$.
**Initialize:** initial states $\Theta^{(0)}$; $y_i^{(0)} = g_i^{(0)}$ for Strategy II, or $y_i^{(0)} = \lambda_i g_i^{(0)}$ for Strategy I.
**for** $t = 0$ **to** $T-1$ **do**
  **for** each node $i \in \mathcal{V}$ **do**
    Receive $\{\theta_j^{(t)}, y_j^{(t)}\}_{j\in\mathcal{N}_i}$ from neighbors.
    **Model update:**
    $$\theta_i^{(t+1)} = \sum_{j=1}^{n}[W_\lambda]_{i,j}(\theta_j^{(t)} - \alpha y_j^{(t)}).$$
    **Tracker update:**
    $$y_i^{(t+1)} = \sum_{j=1}^{n}[W_\lambda]_{i,j}y_j^{(t)} + [G_\lambda]_{i,i}(g_i^{(t+1)} - g_i^{(t)}).$$
  **end for**
**end for**

### The $L^2(\lambda;\mathbb{R}^d)$–Hilbert Space:

- **Inner product**

$$\langle X, Y\rangle_{\lambda,d} = \sum_{i=1}^{n}\lambda_i\langle x_i, y_i\rangle, \quad X, Y \in (\mathbb{R}^d)^n.$$

- **Weighted Frobenius norm**

$$\|X\|_{\lambda,F} = \Big(\sum_i \lambda_i\|x_i\|^2\Big)^{1/2}.$$

- **Key Spectral Difference**
  Consensus projections: $J := \frac{\mathbf{1}\mathbf{1}^\top}{n}$ and $\Lambda := \frac{\mathbf{1}\lambda^\top}{n}$

$$\|W^{\mathrm{ds}} - J\|_\lambda = \|W_J\|_\lambda = \|D_\lambda^{1/2}W_J D_\lambda^{-1/2}\|_2 \le \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\rho_J := \kappa_\lambda\rho_J.$$

$$\|W - \Lambda\|_\lambda = \|W_\Lambda\|_\lambda = \|D_\lambda^{1/2}W_\Lambda D_\lambda^{-1/2}\|_2 = \|\widetilde{W}_\Lambda\|_2 := \rho_\Lambda.$$

**Takeaway:** In the weighted space, $W$ is self–adjoint and has no distortion penalty; $W^{\mathrm{ds}}$ is generally non–self–adjoint and pays $\kappa_\lambda$.

### Why $L^2(\lambda;\mathbb{R}^d)$ ?

- **Descent Lemma**

$$\mathbb{E}[F(\bar{\theta}_*^{(t+1)})] \le \mathbb{E}[F(\bar{\theta}_*^{(t)})] - \frac{\alpha}{2}\mathbb{E}\|\nabla F(\bar{\theta}_*^{(t)})\|^2 + \frac{\alpha\beta^2}{2n}\mathbb{E}\underbrace{\|(I-M_*)\Theta^{(t)}\|_{F,\lambda}^2}_{\text{weighted consensus error}} + \frac{\alpha^2 c_\lambda\beta v^2}{2}.$$

$$(\bar{\theta}_*^{(t)}, M_*) = \begin{cases}(\bar{\theta}^{(t)}, J), & \text{Strategy I},\\ (\bar{\theta}_\lambda^{(t)}, \Lambda), & \text{Strategy II}.\end{cases}$$

Consensus error is measured in $\|\cdot\|_{F,\lambda}$, so the natural space is $L^2(\lambda;\mathbb{R}^d)$.

## Main Results

### Consensus Error: Where the Penalty Appears

**Strategy I**:

$$\sum_{t=0}^{T-1}\mathbb{E}[\|E_{\mathrm{I}}^{(t)}\|_{F,\lambda}^2] \sim \kappa_\lambda^2\cdot\mathrm{Init}_J + \lambda_{\max}^2\cdot\mathrm{Grad}_J + \lambda_{\max}^2\cdot\mathrm{Noise}_J,$$

**Strategy II**:

$$\sum_{t=0}^{T-1}\mathbb{E}[\|E_{\mathrm{II}}^{(t)}\|_{F,\lambda}^2] \sim \mathrm{Init}_\Lambda + \mathrm{Grad}_\Lambda + \mathrm{Noise}_\Lambda.$$

**Doubly stochastic mixing pays extra factors; weighted mixing avoids them.**

### Convergence Rate:

**Strategy I**:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{\theta}^{(t)})\|^2] \lesssim \frac{1}{\alpha T} + \alpha c_\lambda\beta v^2 + \kappa_\lambda\frac{\mathrm{Init}_J}{nT} + \lambda_{\max}^2\alpha^2\beta^2 v^2\mathrm{Net}_J,$$

**Strategy II**:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\bar{\theta}_\lambda^{(t)})\|^2] \lesssim \frac{1}{\alpha T} + \alpha c_\lambda\beta v^2 + \frac{\mathrm{Init}_\Lambda}{nT} + \alpha^2\beta^2 v^2\mathrm{Net}_\Lambda.$$

**Even if $W$ has a smaller spectral gap: $1-\rho_\Lambda < 1-\rho_J$,**
**Strategy II can still achieve a sharper bound because it avoids $\kappa_\lambda^2$ and $\lambda_{\max}^2$ penalties.**

### Head–to–Head Comparison and Topology Design:

- **Sufficient spectral condition:**

$$1-\rho_\Lambda \gtrsim \lambda_{\max}^{-1/2}(1-\rho_J).$$

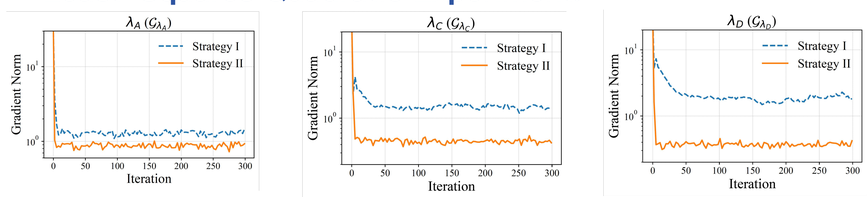
- **Topology design rule:**

$$d_i \propto \lambda_i$$

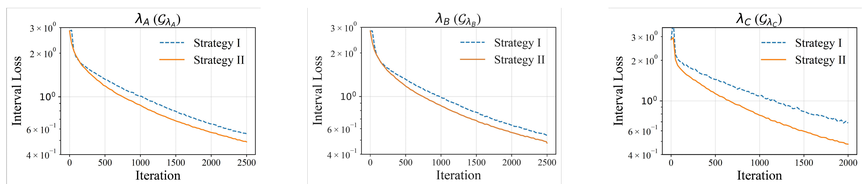
**Nodes with larger weights should have higher connectivity.**

## Experiments

### Least–Squares Quadratic Experiment:




### ResNet–18 on CIFAR–10:




### Nodes Number:

$$\lambda_A, \lambda_b : 16, \lambda_C : 32, \lambda_D : 64$$