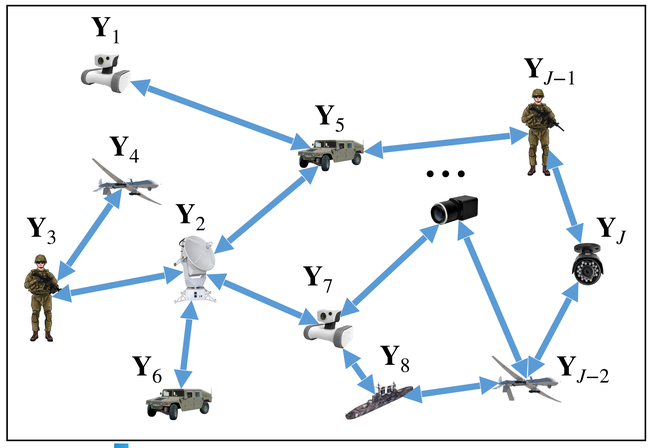
Distributed Low-rank Matrix Factorization With Exact Consensus

Zhihui Zhu*, Qiuwei Li†, Xinshuo Yang†, Gongguo Tang†, Michael B. Wakin†

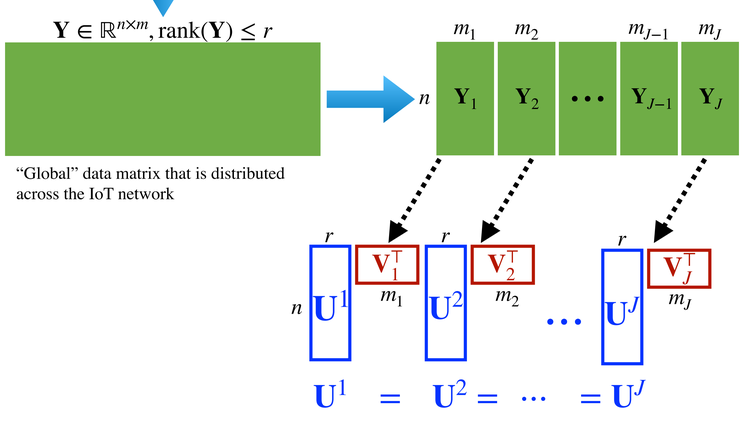
* Mathematical Institute for Data Science, Johns Hopkins University. † Department of Electrical Engineering, Colorado School of Mines.

Motivation

– Imagine data Y distributed across J agents in a connected network.



$\mathbf{Y} \in \mathbb{R}^{n\times m}, \text{rank}(\mathbf{Y}) \le r$



"Global" data matrix that is distributed across the IoT network

– This problem is referred to as the distributed matrix factorization (DMF) problem.

– Mathematically, we consider formulating DMF as a global consensus optimization problem:

$$\begin{aligned}\underset{\substack{\mathbf{U}^1\in\mathbb{R}^{n\times r},\cdots,\mathbf{U}^J\in\mathbb{R}^{n\times r},\\ \mathbf{V}_1\in\mathbb{R}^{m_1\times r},\cdots,\mathbf{V}_J\in\mathbb{R}^{m_J\times r}}}{\text{minimize}} & \sum_{j=1}^{J}\|\mathbf{U}^j\mathbf{V}_j{}^T-\mathbf{Y}_j\|_F^2 \\ \text{s.t. } & \mathbf{U}^1=\mathbf{U}^2=\cdots=\mathbf{U}^J\end{aligned} \qquad \textbf{(DMF)}$$

– In addition to having global consensus variables, (**DMF**) involves local variables. Therefore:

  – general distributed algorithms like distributed gradient descent (DGD) fail to apply

  – although certain distributed methods like ADMM apply to this scenario, there is no existing guarantee for exact recovery

– This work aims to extend the most simple distributed algorithm DGD such that it can achieve both exact consensus and globally optimal convergence.

Distributed Gradient Descent (DGD)+LOCAL

– **Centralized Problem:**

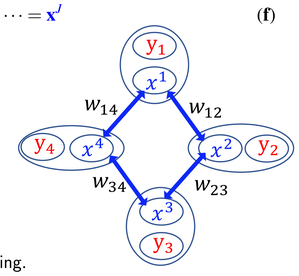
$$\underset{\mathbf{x},\mathbf{y}_1,\cdots,\mathbf{y}_J}{\text{minimize}}\ f(\mathbf{x},\mathbf{y}_1,\cdots,\mathbf{y}_J)=\sum_{j=1}^{J}f_j(\mathbf{x},\mathbf{y}_j) \qquad \textbf{(c)}$$

– **Distributed/Decentralized Problem:** involves common variables and local variables

$$\underset{\mathbf{x}^1,\ldots,\mathbf{x}^J,\mathbf{y}_1,\cdots,\mathbf{y}_J}{\text{minimize}}\ \sum_{j=1}^{J}f_j(\mathbf{x}^j,\mathbf{y}_j),\ \text{s.t. } \mathbf{x}^1=\cdots=\mathbf{x}^J \qquad \textbf{(f)}$$

– DGD + LOCAL update:

$$\begin{aligned}\mathbf{x}^j(k+1) &= \sum_{i=1}^{J}\left(\omega_{ji}\mathbf{x}^i(k)\right)-\mu\nabla_{\mathbf{x}}f_j\left(\mathbf{x}^j(k),\mathbf{y}_j(k)\right)\\ \mathbf{y}_j(k+1) &= \mathbf{y}_j(k)-\mu\nabla_{\mathbf{y}}f_j\left(\mathbf{x}^j(k),\mathbf{y}_j(k)\right)\end{aligned}$$



– $[\omega_{ji}]$ is a symmetric weight matrix, playing a role in local averaging.

– Standard DGD only involves common variables $\mathbf{x}^j$.

Consensus and convergence analysis

Proof Ideas

1. DGD+LOCAL $\iff$ applying Gradient Descent (GD) to (**g**).
2. Any critical point of (**g**) is in the consensus space.
3. Critical points of (**g**) and (**c**) correspond one-to-one.
4. GD converges to 2nd-order critical points $\Rightarrow$ DGD+LOCAL converges to 2nd-order critical points.

1. DGD+LOCAL $\iff$ applying GD with stepsize μ to (g)

$$\underset{\mathbf{x}^1,\ldots,\mathbf{x}^J,\mathbf{y}_1,\cdots,\mathbf{y}_J}{\text{minimize}}\ g(\mathbf{x}^1,\ldots,\mathbf{x}^J,\mathbf{y}_1,\cdots,\mathbf{y}_J)=\sum_{j=1}^{J}\left(f_j\left(\mathbf{x}^j,\mathbf{y}_j\right)+\sum_{i=1}^{J}\frac{\omega_{ji}}{4\mu}\left\|\mathbf{x}^j-\mathbf{x}^i\right\|_2^2\right) \qquad \textbf{(g)}$$

2. Any critical point of (g) is in the consensus space

Theorem 1. *Suppose any f_j satisfies the "symmetric gradient property" that $\langle\nabla_{\mathbf{x}}f_j(\mathbf{x}^j,\mathbf{y}_j),\mathbf{x}^j\rangle=\langle\nabla_{\mathbf{y}}f_j(\mathbf{x}^j,\mathbf{y}_j),\mathbf{y}_j\rangle$ for any $\mathbf{x}^j,\mathbf{y}_j$. Then any critical point of* (**g**) *satisfies $\mathbf{x}^1=\mathbf{x}^2=\ldots=\mathbf{x}^J$.*

	Local variables?	Exact consensus?	
DGD	✗	✗	consensus error $\propto$ step size
DGD+LOCAL	✓	✓	symmetric gradient property

3. Critical points of (g) and (c) correspond one-to-one

Theorem 2. *If $(\mathbf{x}^1,\ldots,\mathbf{x}^J,\mathbf{y}_1,\cdots,\mathbf{y}_J)$ is a 1st/2nd-order critical point of* (**g**) *and $\mathbf{x}^1=\mathbf{x}^2=\ldots=\mathbf{x}^J$ for some $\mathbf{x}$, then $(\mathbf{x},\mathbf{y}_1,\cdots,\mathbf{y}_J)$ is also a 1st/2nd-order critical point of* (**c**).

4. GD converges to 2nd-order critical points $\Rightarrow$ DGD+LOCAL...

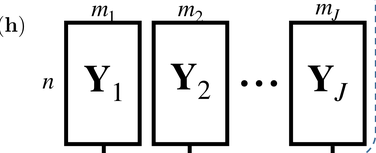
Theorem 3. *Assume every f_j satisfies the "symmetric gradient property," is globally lower-bounded, and has bounded gradient and hessian in any bounded set. Then any bounded sequence generated by DGD+LOCAL with a sufficiently small stepsize μ almost surely converges to a 2nd-order critical point of* (**g**)*, and therefore **corresponds to a 2nd-order critical point of*** (**c**).

Remark: If, furthermore, all 2nd-order critical points of (**c**) are global minima, then DGD+LOCAL converges to a global minimum of (**c**).

Why is exact consensus achieved for DMF?

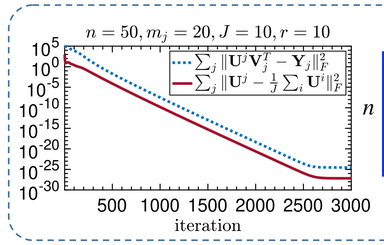
$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r},\mathbf{V}_j\in\mathbb{R}^{m_j\times r}}{\text{minimize}}\ \sum_{j=1}^{J}\|\mathbf{U}\mathbf{V}_j{}^T-\mathbf{Y}_j\|_F^2 \qquad \textbf{(h)}$$

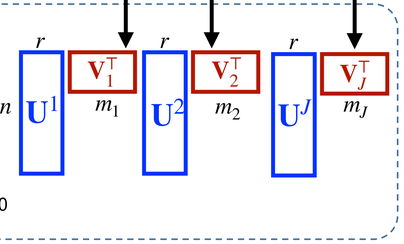
- satisfy **symmetric gradient property**
- every 2nd-order critical point is global optimal when Y is rank r



Theorem 4. *If* $\text{rank}(\mathbf{Y}) \le r$*, then any bounded sequence generated by applying DGD+LOCAL to* (**DMF**) *almost surely corresponds to a global minimizer of* (**h**).

DGD+LOCAL



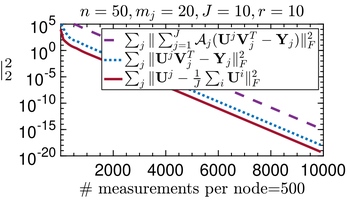


Distributed matrix completion/sensing

Distributed matrix sensing

$$\underset{\mathbf{U}^1,\cdots,\mathbf{U}^J,\mathbf{V}_1,\cdots,\mathbf{V}_J}{\text{minimize}}\ \sum_{j=1}^{J}\|\mathcal{A}_j(\mathbf{U}^j\mathbf{V}_j{}^T-\mathbf{Y}_j)\|_2^2 \quad \text{s.t. } \mathbf{U}^1=\mathbf{U}^2=\cdots=\mathbf{U}^J$$

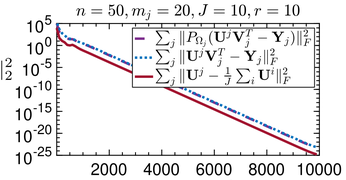
- satisfy symmetric gradient property.
- $\mathcal{A}_j$ is the local sensing operator at node j.



Distributed matrix completion

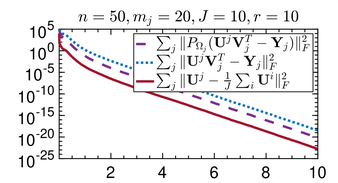
$$\underset{\mathbf{U}^1,\cdots,\mathbf{U}^J,\mathbf{V}_1,\cdots,\mathbf{V}_J}{\text{minimize}}\ \sum_{j=1}^{J}\|P_{\Omega_j}(\mathbf{U}^j\mathbf{V}_j{}^T-\mathbf{Y}_j)\|_2^2 \quad \text{s.t. } \mathbf{U}^1=\mathbf{U}^2=\cdots=\mathbf{U}^J$$

- satisfy symmetric gradient property.
- P_{Ω_j} is the local sampling operator at node j.



Distributed matrix completion

- solve SGD version of DGD+LOCAL.



Acknowledgement

This work was supported by the DARPA Lagrange Program under ONR/SPAWAR contract N660011824020. Thanks to Waheed Bajwa, Haroon Raja, Clement Royer, and Stephen Wright for many informative discussions.