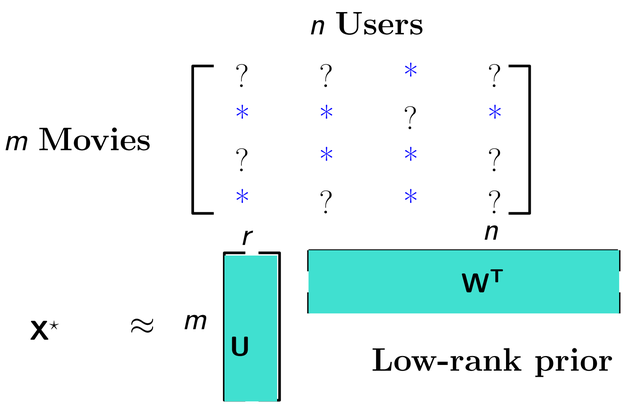
A Riemannian gossip approach to decentralized matrix completion

Bamdev Mishra*, Hiroyuki Kasai†, and Atul Saroop*

*Amazon Development Centre India, Bangalore, Karnataka 560055, India (Bamdevm, Asaroop@amazon.com)
†The University of Electro-Communications, Tokyo, 182-8585, Japan (kasai@is.uec.ac.jp)

Motivation



$(n+m-r)r,\ r \ll (m,n)$

Formulation

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} \ \frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{X}) - \mathcal{P}_\Omega(\mathbf{X}^\star)\|_F^2 \quad \text{subject to} \quad \text{rank}(\mathbf{X}) = r.$$

- The rank constrained is parameterized as $\mathbf{X} = \mathbf{U}\mathbf{W}^T$.

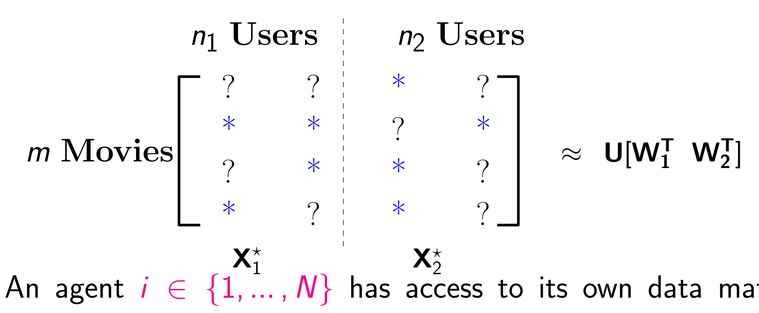
$\min_{\mathbf{U}\in\text{St}(r,m)} \min_{\mathbf{W}\in\mathbb{R}^{n\times r}} \|\mathcal{P}_\Omega(\mathbf{U}\mathbf{W}^T) - \mathcal{P}_\Omega(\mathbf{X}^\star)\|_F^2$
$\equiv$
$\min_{\mathbf{U}\in\text{Gr}(r,m)} f(\mathbf{U}, \mathbf{W}_\mathbf{U})$,
a Grassmann optimization problem.

- Ω is the set of known entries.
- The inner problem has a closed-form solution.

Contributions

- We develop a nonlinear gossip algorithm with minimal communication between agents which individually have only a part of data.
- The optimization formulation is based on a weighted combination of matrix completion and consensus terms.
- We develop the parallel and preconditioned variants of the proposed gossip algorithm.
- Code: www.bamdevmishra.com/codes/gossipMC.

Decentralized formulation



An agent $i \in \{1,\dots,N\}$ has access to its own data matrix $\mathbf{X}^\star = [\mathbf{X}_1^\star, \mathbf{X}_2^\star, \dots, \mathbf{X}_N^\star]$.

$\sum_i \min_{\mathbf{U}\in\text{St}(r,m),\mathbf{W}_i\in\mathbb{R}^{n_i\times r}} \frac{1}{2}\|\mathcal{P}_{\Omega_i}(\mathbf{U}\mathbf{W}_i^T) - \mathcal{P}_{\Omega_i}(\mathbf{X}_i^\star)\|_F^2$
$= \min_{\mathbf{U}\in\text{Gr}(r,m)} \frac{1}{2}\sum_i \|\mathcal{P}_{\Omega_i}(\mathbf{U}\mathbf{W}_{i\mathbf{U}}^T) - \mathcal{P}_{\Omega_i}(\mathbf{X}_i^\star)\|_F^2$,

where $\mathbf{W}_{i\mathbf{U}}$ is computed by agent i independently.

Although the problem is distributed, we still need to learn a common $\mathbf{U}$.

Proposed gossip approach decouples learning of $\mathbf{U}$

Key idea: introduce multiple copies of $\mathbf{U}$ among N agents, but allow them to reach consensus.

$$\min_{\mathbf{U}_1,\dots,\mathbf{U}_N\in\text{St}(r,m)} \frac{1}{2}\sum_i \underbrace{\|\mathcal{P}_{\Omega_i}(\mathbf{U}_i\mathbf{W}_{i\mathbf{U}_i}^T) - \mathcal{P}_{\Omega_i}(\mathbf{X}_i^\star)\|_F^2}_{\text{completion task handled by agent } i} + \frac{\rho}{2}\underbrace{\left(d(\mathbf{U}_1,\mathbf{U}_2)^2 + d(\mathbf{U}_2,\mathbf{U}_3)^2 + \dots + d(\mathbf{U}_{N-1},\mathbf{U}_N)^2\right)}_{\text{consensus among agents}}.$$

d is the Riemannian distance on the Grassmann manifold.

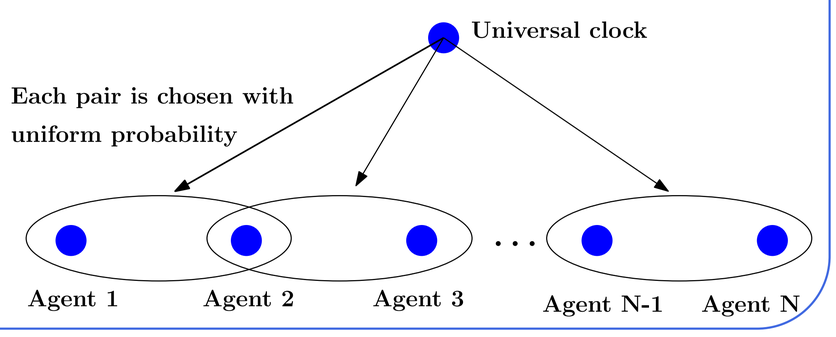
A large ρ trades-off completion with consensus.

Minimizing only consensus $\Rightarrow \mathbf{U}_1 = \mathbf{U}_2 = \dots = \mathbf{U}_{N-1} = \mathbf{U}_N$.

The consensus term exploits a particular topology, where the agents are connected in a linear fashion and each has only one neighbor.

Stochastic gradient descent on Grassmann manifold

1. Agents i and $i+1$ are neighbors for all $i \leqslant N-1$. (ordering of agents)
2. At each time slot, say t, we pick an agent $i \leqslant N-1$ randomly with uniform probability. (SGD updates)
   - Equivalently, we also pick agent $i+1$ (the neighbor of agent i).
   - Agents i and $i+1$ update $\mathbf{U}_i$ and $\mathbf{U}_{i+1}$, respectively, by taking a gradient descent step with stepsize γ_t on Grassmann manifold.
   - $\sum \gamma_t^2 < \infty$ and $\sum \gamma_t = +\infty$.
3. The Grassmann update equation is $\mathbf{U}_+ = \text{Exp}_\mathbf{U}(-\gamma\xi_\mathbf{U})$, where $\xi_\mathbf{U}$ is the gradient and Exp is the exponential map.



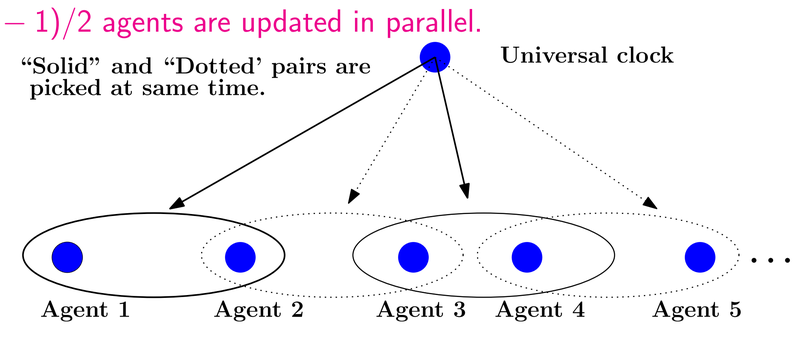
Preconditioned variant: scale the gradient

$$\xi_{\mathbf{U}_i} \mapsto \xi_{\mathbf{U}_i}\big(\underbrace{\mathbf{W}_{i\mathbf{U}_i}^T\mathbf{W}_{i\mathbf{U}_i}}_{\text{from completion}} + \underbrace{\rho\mathbf{I}}_{\text{from consensus}}\big)^{-1},$$

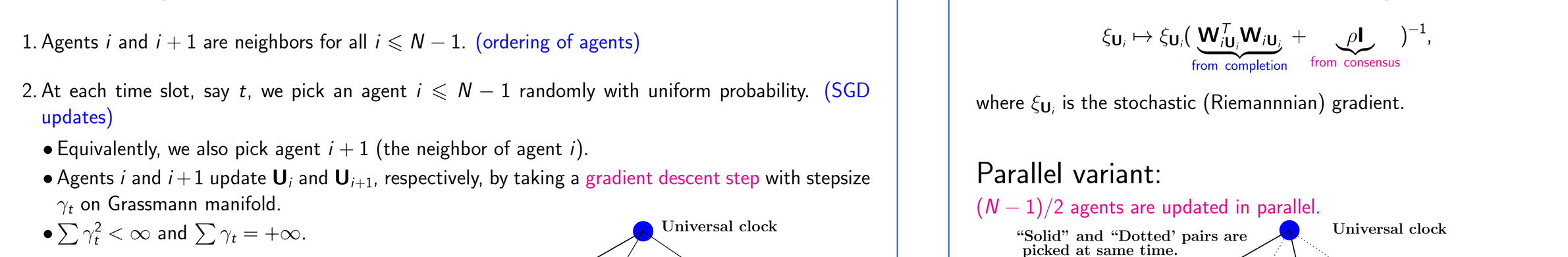
where $\xi_{\mathbf{U}_i}$ is the stochastic (Riemannnian) gradient.

Parallel variant:

$(N-1)/2$ agents are updated in parallel.



Numerical comparisons



(a) Effect of ρ. (b) Online versus parallel. (c) Preconditioning. (d) Online versus D-LMaFit. (e) Netflix: consensus.

Performance of Online Gossip on the Netflix dataset at rank 10 with different number of agents

	$N=2$	$N=5$	$N=10$	$N=15$	$N=20$	RTRMC-1 (batch method)
Test RMSE	0.877	0.885	0.891	0.894	0.900	0.873

References

[1] B. Mishra, H. Kasai, and A Saroop. A Riemannian gossip approach to decentralized matrix completion. arXiv preprint arXiv:1605.06968, 2016.
[2] S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. IEEE TAC, 2013.