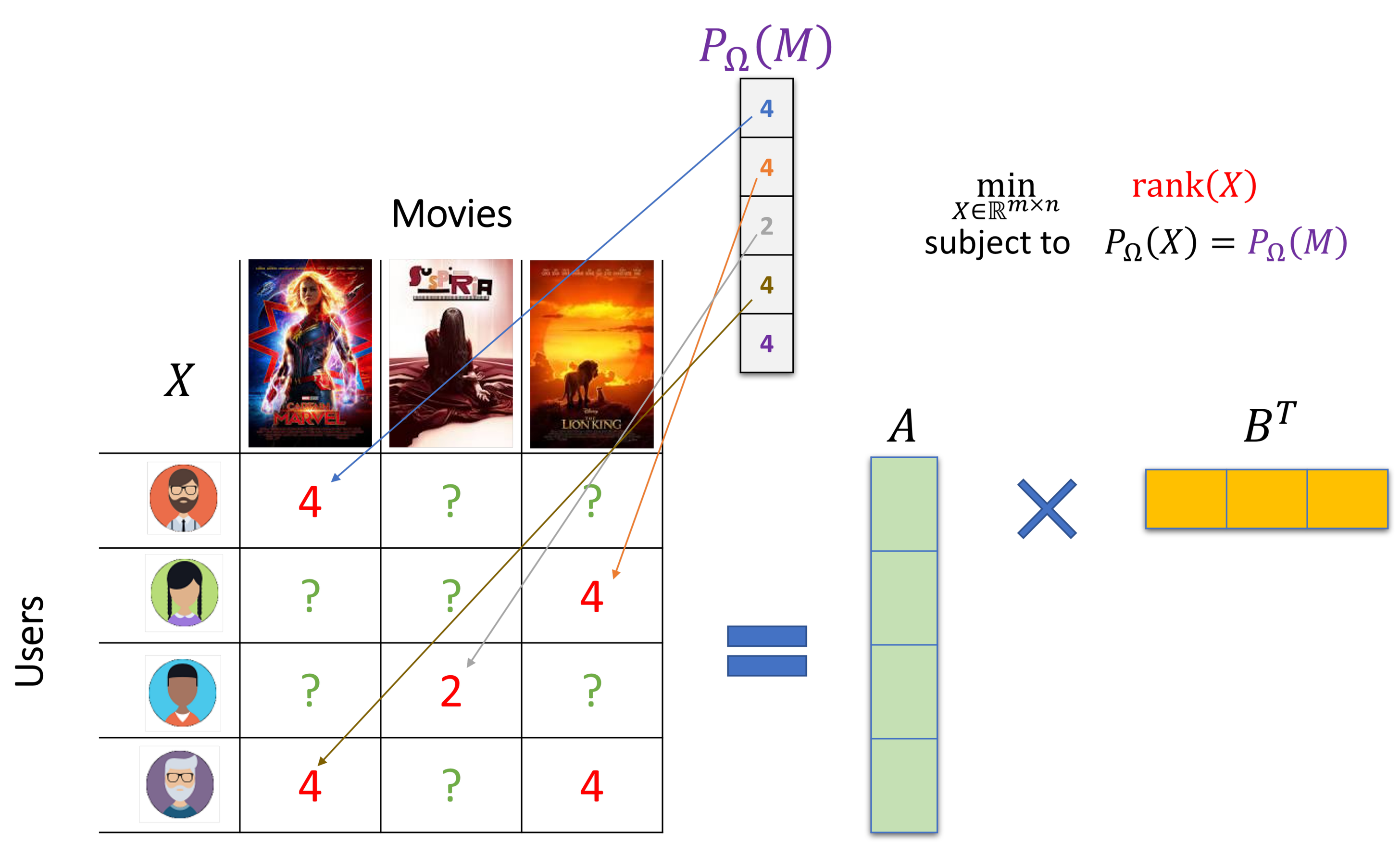
ON THE ASYMPTOTIC LINEAR CONVERGENCE OF GRADIENT DESCENT FOR NON-SYMMETRIC MATRIX COMPLETION

Trung Vu[1,2] and Raviv Raich[1]

[1] School of EECS, Oregon State University, Corvallis, OR 97331-5501, USA
[2] Department of CSEE, University of Maryland, Baltimore County, MD 21250-0002, USA

Matrix Completion Problem (MCP)



- MCP is in general **NP-hard**
- **Low-rank factorization** formulation is both *memory and computationally efficient*
- Gradient descent (GD) is a **simple and scalable** method
- **Convergence analysis** in *non-symmetric* MCP is challenging

Global versus Local Convergence Analysis

Global analysis

- requires assumptions on underlying model in asymptotic settings
- is powerful in proving convergence to a unique global optimum
- provides conservative upper bounds on the linear convergence rate

Local analysis

- identifies the *deterministic* conditions in a broad range of settings
- complementary to global analysis
- provides an exact estimate of the linear rate

Contributions

- ✓ Analyze the local convergence of GD for **non-symmetric** MCP
- ✓ Establish the **first-known** exact linear convergence rate
- ✓ Illustrate the correctness and tightness via numerical simulation

Gradient Descent for Non-Symmetric MCP

Proposed objective function with regularization:

$$\min_{\boldsymbol{A}\in\mathbb{R}^{m\times r},\boldsymbol{B}\in\mathbb{R}^{n\times r}} f(\boldsymbol{A},\boldsymbol{B}) = \frac{1}{2}\|\mathcal{P}_\Omega(\boldsymbol{A}\boldsymbol{B}^\top - \boldsymbol{M})\|_F^2 + \frac{1}{4}\|\boldsymbol{A}^\top\boldsymbol{A} - c\boldsymbol{I}_r\|_F^2$$

- Orthogonality regularization for $\boldsymbol{A}$ to ensure uniqueness
- Scaling factor c to improve the convergence speed

Algorithm 1 Factorization-based Gradient Descent

Input: $\boldsymbol{A}_0, \boldsymbol{B}_0, \mathcal{P}_\Omega(\boldsymbol{M}), \eta$
Output: $\{\boldsymbol{A}_k, \boldsymbol{B}_k\}$

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: $\quad \boldsymbol{P}_k = \mathcal{P}_\Omega(\boldsymbol{A}_k\boldsymbol{B}_k^\top - \boldsymbol{M})$
3: $\quad \boldsymbol{A}_{k+1} = \boldsymbol{A}_k - \eta\big(\boldsymbol{P}_k\boldsymbol{B}_k + \boldsymbol{A}_k(\boldsymbol{A}_k^\top\boldsymbol{A}_k - c\boldsymbol{I}_r)\big)$ ▷ $\boldsymbol{A}$-update
4: $\quad \boldsymbol{B}_{k+1} = \boldsymbol{B}_k - \eta\boldsymbol{P}_k^\top\boldsymbol{A}_k$ ▷ $\boldsymbol{B}$-update

Local Convergence Analysis

Establishing a recursion on the error

- Consider the SVD $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{A}_*\boldsymbol{B}_*^\top$. Define the error matrix

$$\boldsymbol{E}_k = \begin{bmatrix}\boldsymbol{A}_k\\ \boldsymbol{B}_k\end{bmatrix}\begin{bmatrix}\boldsymbol{A}_k^\top & \boldsymbol{B}_k^\top\end{bmatrix} - \begin{bmatrix}\boldsymbol{A}_*\\ \boldsymbol{B}_*\end{bmatrix}\begin{bmatrix}\boldsymbol{A}_*^\top & \boldsymbol{B}_*^\top\end{bmatrix} = \begin{bmatrix}\boldsymbol{E}_k^{AA} & \boldsymbol{E}_k^{AB}\\ \boldsymbol{E}_k^{BA} & \boldsymbol{E}_k^{BB}\end{bmatrix}$$

- Using the $\boldsymbol{A}$ and $\boldsymbol{B}$ updates to represent

$$\text{vec}(\boldsymbol{E}_{k+1}) = \boldsymbol{Z}(\boldsymbol{I}_{(m+n)^2} - \eta\boldsymbol{H})\boldsymbol{Z}^\top\text{vec}(\boldsymbol{E}_k) + \mathcal{O}(\|\boldsymbol{E}_k\|_F^2) \qquad (1)$$

  - $\boldsymbol{H}$ depends *only* on the solution matrix $\boldsymbol{M}$ and the sampling set Ω
  - $\boldsymbol{Z}$ is a permutation matrix

Integrating structural constraints

- $\boldsymbol{E}_k$ is the difference between 2 rank-r PSD matrices in $\mathbb{R}^{(m+n)\times(m+n)}$

$$\text{vec}(\boldsymbol{E}_k) = \boldsymbol{P}_{\text{sym}}\boldsymbol{P}_{T_r}\text{vec}(\boldsymbol{E}_k) + \mathcal{O}(\|\boldsymbol{E}_k\|_F^2) \qquad (2)$$

  - $\boldsymbol{P}_{\text{sym}}$ is the projection onto the set of symmetric matrices
  - $\boldsymbol{P}_{T_r}$ is the projection onto the *tangent space* of the set of rank-r matrices

- Decompose $\boldsymbol{P}_{\text{sym}}\boldsymbol{P}_{T_r} = \boldsymbol{Q}\boldsymbol{Q}^\top$. Substituting (2) back into (1) yields

$$\boldsymbol{Q}^\top\text{vec}(\boldsymbol{E}_{k+1}) = \boldsymbol{Q}^\top\boldsymbol{Z}(\boldsymbol{I}_{(m+n)^2} - \eta\boldsymbol{H})\boldsymbol{Z}^\top\boldsymbol{Q}\boldsymbol{Q}^\top\text{vec}(\boldsymbol{E}_k) + \mathcal{O}(\|\boldsymbol{E}_k\|_F^2)$$

Convergence of fixed-point iterations

If $\boldsymbol{Q}^\top\boldsymbol{Z}(\boldsymbol{I}_{(m+n)^2} - \eta\boldsymbol{H})\boldsymbol{Z}^\top\boldsymbol{Q}$ is a contraction map, then starting with $\boldsymbol{E}_0$ sufficiently small (in norm), we have

$$\|\boldsymbol{E}_k\|_F \leq C\|\boldsymbol{E}_0\|_F \cdot \rho^k,$$

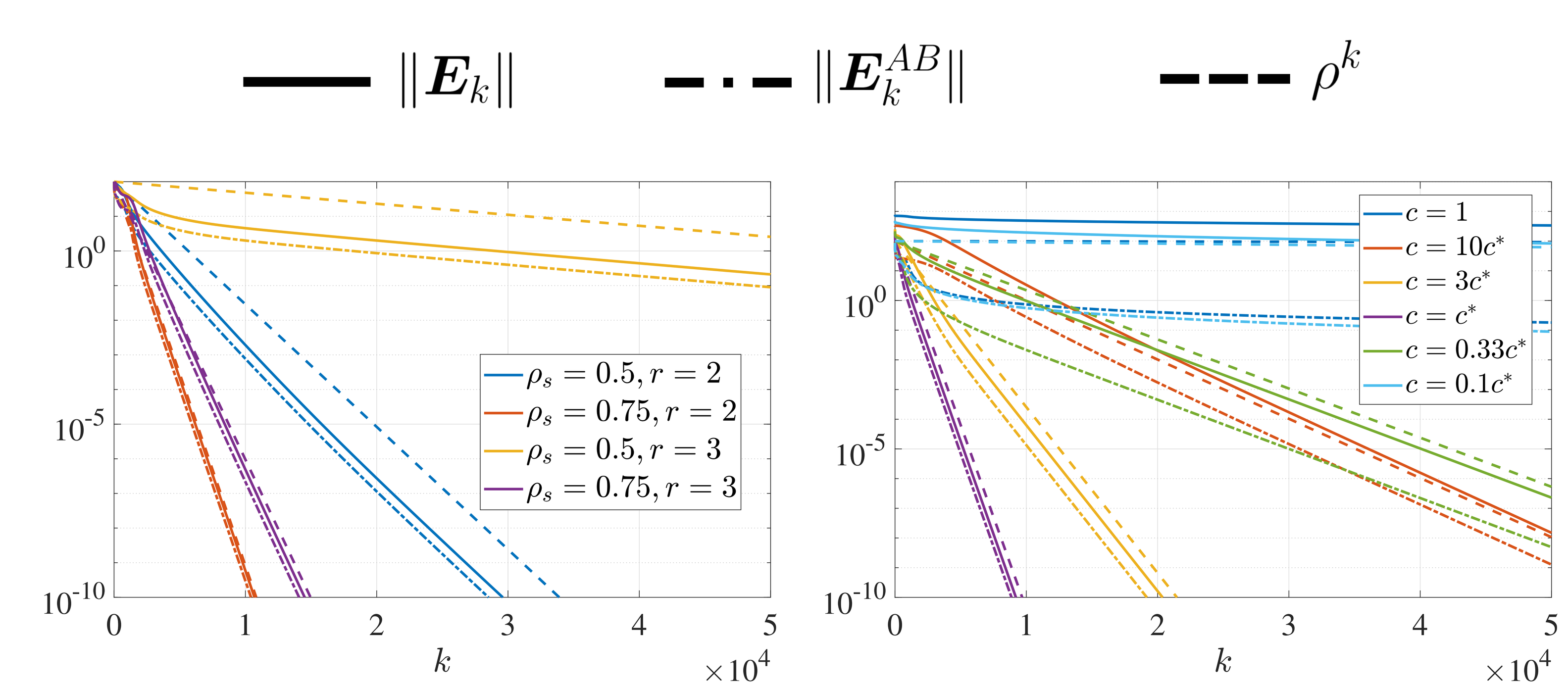
where ρ is the *spectral radius* of $\boldsymbol{Q}^\top\boldsymbol{Z}(\boldsymbol{I}_{(m+n)^2} - \eta\boldsymbol{H})\boldsymbol{Z}^\top\boldsymbol{Q}$ and $C > 0$ is some numerical constant

Main Theorem

If $\hat{\boldsymbol{H}} = \boldsymbol{Q}^\top\boldsymbol{Z}\boldsymbol{H}\boldsymbol{Z}^\top\boldsymbol{Q}$ is non-singular and $\boldsymbol{A}_0\boldsymbol{B}_0^\top$ is sufficiently close to $\boldsymbol{M}$, then Algorithm 1 produces a sequence of matrices $\boldsymbol{A}_k\boldsymbol{B}_k^\top$ converging to $\boldsymbol{M}$ at an asymptotic linear rate

$$\rho = \max\{|1 - \eta\lambda_{\max}(\hat{\boldsymbol{H}})|, |1 - \eta\lambda_{\min}(\hat{\boldsymbol{H}})|\}$$

Numerical Results



Summary

- (Left) The empirical rate at which $\|\boldsymbol{E}_k\|_F$ and $\|\boldsymbol{E}_k^{AB}\|_F$ decrease to zero **matches** that of our exact analytical rate ρ^k
- (Right) $c = 1$ results in **slow** convergence since $\|\boldsymbol{A}\|_F$ and $\|\boldsymbol{B}\|_F$ are significantly different
- (Right) $c^* = \sqrt{\frac{mn}{r|\Omega|}}\|\mathcal{P}_\Omega(\boldsymbol{M})\|_F$ yields the **fastest** empirical convergence with $\|\boldsymbol{A}\|_F \approx \|\boldsymbol{B}\|_F$

More about exact linear convergence rate analysis ⇒

trungvv@umbc.edu