Monash Workshop on Numerical PDEs

Nonlinearly Preconditioned Optimization on Grassmann Manifolds for Tucker Tensor Approximations

Alexander Howse [1] **Hans De Sterck** [2]

[1] University of Waterloo [2] Monash University

Outline

- Tensor Approximations and Decompositions
- The Higher Order Singular Value Decomposition
- Matrix Manifold Optimization
- Nonlinearly Preconditioned Conjugate Gradient
- Nonlinear GMRES
- Numerical Results

Tensor Preliminaries

- *Tensor:* a multidimensional array
- Tensor *order:* number of dimensions
- Notation: vector ($\mathbf{x}$), matrix ($\mathbf{X}$), tensor ($\mathcal{X}$)
- Frobenius norm:

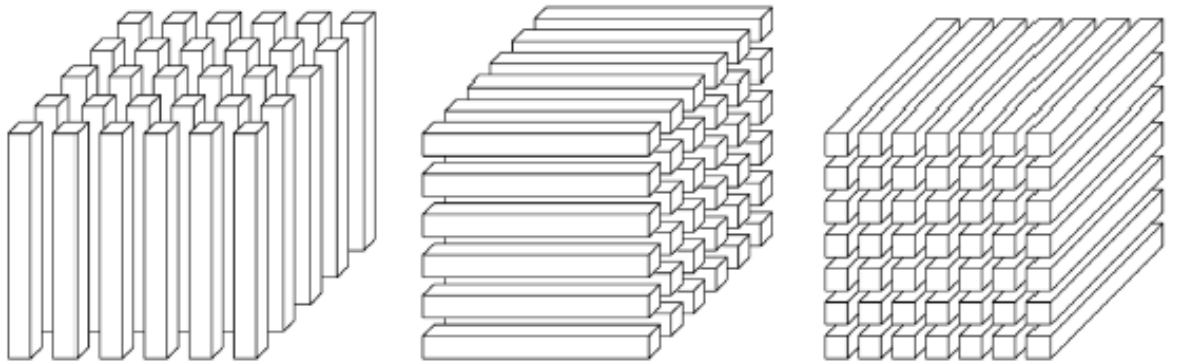
$$\|\mathcal{X}\|_F = \sqrt{\sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} \mathcal{X}_{i_1 \cdots i_N}^2}$$

- Inner product:

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1=1}^{I_1} \cdots \sum_{i_N=1}^{I_N} \mathcal{X}_{i_1 \cdots i_N} \mathcal{Y}_{i_1 \cdots i_N}$$

Tensor Preliminaries

- Tensor *fiber:* order-1 tensor obtained by fixing all indices but one
- *Mode-n matricization*, $\mathbf{X}_{(n)}$: has mode-n fibers of $\mathcal{X}$ as columns
- *Multilinear rank:* $\left(\text{rank}\left(\mathbf{X}_{(1)}\right), \ldots, \text{rank}\left(\mathbf{X}_{(N)}\right)\right)$



(a) Mode-1 (column) fibers: $\mathbf{x}_{:jk}$ (b) Mode-2 (row) fibers: $\mathbf{x}_{i:k}$ (c) Mode-3 (tube) fibers: $\mathbf{x}_{ij:}$

Figure: Image due to [Kolda & Bader, 2009].

Tensor Approximation Problem

Given $\mathcal{X}$, compute approximation $\widehat{\mathcal{X}}$ such that

1. $\widehat{\mathcal{X}}$ is much cheaper to store
2. $\widehat{\mathcal{X}}$ extracts relevant information from $\mathcal{X}$

Determine $\widehat{\mathcal{X}}$ by solving

$$\min_{\widehat{\mathcal{X}}} \quad \left\| \mathcal{X} - \widehat{\mathcal{X}} \right\|_F$$

Applications of Tensor Approximation

- Data Compression
- Handwritten Character Recognition
- Computer Vision
- Multi-way Statistical Analysis

Tensor Decompositions

- Express $\mathcal{X}$ as a sum or product of several components
- Specific decomposition for $\widehat{\mathcal{X}} \Rightarrow$ minimize over required components

Tucker Decomposition

- n-mode product of $\mathcal{S} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ and $\mathbf{U} \in \mathbb{R}^{J \times I_n}$ is $\mathcal{X} = (\mathbf{U})_n \cdot \mathcal{S} \in \mathbb{R}^{I_1 \times \cdots \times I_{n-1} \times J \times I_{n+1} \times \cdots \times I_N}$:

$$\mathcal{X}_{i_1 \ldots i_{n-1} j i_{n+1} \ldots i_N} = \sum_{i_n=1}^{I_n} \mathbf{U}_{j i_n} \mathcal{S}_{i_1 \ldots i_N}$$

- *Tucker decomposition:*

$$\mathcal{X} = (\mathbf{U}_1, \ldots, \mathbf{U}_N) \cdot \mathcal{S}$$

Singular Value Decomposition (SVD)

- $\mathbf{M} \in \mathbb{R}^{m \times n}$ has SVD $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal$
  - $\mathbf{U}$, $\mathbf{V}$ orthogonal
  - $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is diag $\left(\sigma_1, \ldots, \sigma_{\min(m,n)}\right)$, $\sigma_1 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0$
  - $\{\sigma_i\}_{i=1}^{\min(m,n)}$ are *singular values*
  - Best rank-r approx. in $\|\cdot\|_F$: set $\sigma_i = 0, \quad i = r+1, \ldots, \min(m,n)$

The Higher Order SVD

- $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ has HOSVD $\mathcal{X} = (\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(N)}) \cdot \mathcal{S}$
  - $\mathbf{U}^{(n)}$ orthogonal
  - $\mathcal{S} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ satisfies for all n

    (i) $\langle \mathcal{S}_{i_n=\alpha}, \mathcal{S}_{i_n=\beta} \rangle = 0, \quad \alpha \neq \beta$

    (ii) $\|\mathcal{S}_{i_n=1}\|_F \geq \|\mathcal{S}_{i_n=2}\|_F \geq \cdots \geq \|\mathcal{S}_{i_n=I_n}\|_F \geq 0$

Computing The HOSVD

- *Mode-n matricization*, $\mathbf{X}_{(n)}$: has mode-n fibers of $\mathcal{X}$ as columns
- *Multilinear rank:* $\left(\text{rank}\left(\mathbf{X}_{(1)}\right), \ldots, \text{rank}\left(\mathbf{X}_{(N)}\right)\right)$

1: **procedure** $\text{HOSVD}(\mathcal{X})$
2: **for** $n = 1, \ldots, N$ **do**
3: $\mathbf{U}^{(n)} \leftarrow$ left singular vectors of $\mathbf{X}_{(n)}$
4: **end for**
5: $\mathcal{S} \leftarrow \left(\mathbf{U}^{(1)\intercal}, \ldots, \mathbf{U}^{(N)\intercal}\right) \cdot \mathcal{X}$
6: **end procedure**

- The optimal $\mathcal{S}$ for given $\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(N)}$ is $\mathcal{S} = (\mathbf{U}^{(1)\intercal}, \ldots, \mathbf{U}^{(N)\intercal}) \cdot \mathcal{X}$

HOSVD Tensor Approximations

- May truncate $\mathcal{S}$ to $\hat{\mathcal{S}} \in \mathbb{R}^{R_1 \times \cdots \times R_N}$ and $\mathbf{U}^{(n)}$ to $\hat{\mathbf{U}}^{(n)} \in \mathbb{R}^{I_n \times R_n}$:

$$\hat{\mathcal{X}} = (\hat{\mathbf{U}}^{(1)}, \ldots, \hat{\mathbf{U}}^{(N)}) \cdot \hat{\mathcal{S}}$$

  - If $R_n \geq \text{rank}(\mathbf{X}_{(n)})$ for all n, $\hat{\mathcal{X}} = \mathcal{X}$
  - Otherwise $\hat{\mathcal{X}} \approx \mathcal{X}$ with multilinear rank $(R_1, \ldots, R_N)$
- Truncation does not give the best approximation. We may instead solve:

$$\begin{aligned} \min_{\mathcal{S}, \{\mathbf{U}^{(n)}\}} \quad & \tfrac{1}{2} \left\| \mathcal{X} - (\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(N)}) \cdot \mathcal{S} \right\|_F^2 \\ \text{s.t.} \quad & (\mathbf{U}^{(n)})^\intercal \mathbf{U}^{(n)} = \mathbf{I}_{R_n} \end{aligned}$$

$$\Updownarrow$$

$$\begin{aligned} \max_{\{\mathbf{U}^{(n)}\}} \quad & \tfrac{1}{2} \left\| (\mathbf{U}^{(1)\intercal}, \ldots, \mathbf{U}^{(N)\intercal}) \cdot \mathcal{X} \right\|_F^2 \\ \text{s.t.} \quad & (\mathbf{U}^{(n)})^\intercal \mathbf{U}^{(n)} = \mathbf{I}_{R_n} \end{aligned}$$

Higher Order Orthogonal Iteration

- Let $\mathcal{W}_n = (\mathbf{U}^{(1)\intercal}, \ldots, \mathbf{U}^{(n-1)\intercal}, \mathbf{I}, \mathbf{U}^{(n+1)\intercal}, \ldots \mathbf{U}^{(N)\intercal}) \cdot \mathcal{X}$
- n-mode matricization implies

$$\left\| (\mathbf{U}^{(1)\intercal}, \ldots, \mathbf{U}^{(N)\intercal}) \cdot \mathcal{X} \right\|_F = \left\| \mathbf{U}^{(n)\intercal} \mathbf{W}_{\mathbf{n}(n)} \right\|_F$$

- Solve for $\mathbf{U}^{(n)}$ in alternating fashion

procedure $\mathrm{HOOI}(\mathcal{X}, R_1, \ldots, R_N)$
  initialize each $\mathbf{U}^{(n)}$ using SVD of $\mathbf{X}_{(n)}$
  repeat
    for $n = 1, \ldots, N$ **do**
      Compute $\mathbf{W}_{\mathbf{n}(n)}$
      $\mathbf{U}^{(n)} \leftarrow R_n$ leading left singular vectors of $\mathbf{W}_{\mathbf{n}(n)}$
    end for
  until termination condition satisfied
end procedure

- HOOI can be slow to converge in practice

Optimization Challenges

- Norm invariance under orthogonal transformations $\mathbf{V}^{(n)}$:

$$\|\mathcal{X}\|_F = \left\|(\mathbf{V}^{(1)}, \ldots, \mathbf{V}^{(N)}) \cdot \mathcal{X}\right\|_F$$

  - Decomposition $(\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(N)}) \cdot \mathcal{S}$ not unique:

$$\hat{\mathcal{S}} = \left(\mathbf{V}^{(n)\,-1}\right)_n \cdot \mathcal{S} \text{ and } \hat{\mathbf{U}}^{(n)} = \mathbf{U}^{(n)}\mathbf{V}^{(n)}$$

- Constraints $(\mathbf{U}^{(n)})^\intercal \mathbf{U}^{(n)} = \mathbf{I}_{R_n}$

Our solution: optimize over matrix manifolds!

Grassmann Manifolds

- $\mathrm{Gr}(n, p)$: the set of p-dimensional linear subspaces of $\mathbb{R}^n$
- Specify $\mathcal{Y} \in \mathrm{Gr}(n, p)$ by $\mathrm{Col}(\mathbf{Y})$ for some orthonormal $\mathbf{Y} \in \mathbb{R}^{n \times p}$
- Orthonormal matrices with the same column space:

$$[\mathbf{Y}] := \{\mathbf{YM} | \mathbf{M} \text{ orthogonal}\}$$

- Identify $\mathrm{Gr}(n, p)$ with $\{[\mathbf{Y}] \, | \mathbf{Y}^\intercal \mathbf{Y} = \mathbf{I}_p\}$
- A *Riemannian manifold*: smoothly varying inner product $\langle [\mathbf{X}], [\mathbf{Y}] \rangle = \mathrm{tr}(\mathbf{X}^\intercal \mathbf{Y})$

Working on Manifolds

- *Tangent vector*, ξ: possible direction of movement along manifold
- *Retraction map*, $R_{\mathbf{X}}(t\xi)$: curve along manifold
  - $\text{Exp}_{\mathbf{Y}}(t\xi) = \mathbf{Y}\mathbf{V}\cos(\mathbf{\Sigma}t)\mathbf{V}^\intercal + \mathbf{U}\sin(\mathbf{\Sigma}t)\mathbf{V}^\intercal, \quad$ where $\xi = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal$
  - Use $R_{\mathbf{X}}(t\xi) = \text{qf}(\mathbf{X} + t\xi)$
- *Vector transport map*, $\mathcal{T}_{\mathbf{Y}}(\xi)$: maps ξ at $\mathbf{X}$ to representation $\mathcal{T}_{\mathbf{Y}}(\xi)$ at $\mathbf{Y}$
  - Cannot compare tangent vectors at different points directly!
  - Use $\mathcal{T}_{\mathbf{Y}}(\xi) = (\mathbf{I} - \mathbf{Y}\mathbf{Y}^\intercal)\xi$
- *Logarithmic map*, $\text{Log}_{\mathbf{X}}(\mathbf{Y})$: tangent vector at $\mathbf{X}$ indicating direction of $\mathbf{Y}$
  - $\text{Log}_{\mathbf{X}}(\mathbf{Y}) = \mathbf{U}\arctan(\mathbf{\Sigma})\mathbf{V}^\intercal, \quad$ where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal = (\mathbf{I} - \mathbf{X}\mathbf{X}^\intercal)\mathbf{Y}(\mathbf{X}^\intercal\mathbf{Y})^{-1}$.
- Gradient and Hessian operators are adapted to manifold
  - $\nabla f(\mathbf{x}) \Rightarrow \text{grad}\, f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x}) \Rightarrow \text{Hess}\, f(\mathbf{x})$
- N matrix variables $\Longrightarrow$ work with $\prod_{k=1}^{N} \text{Gr}(n_k, p_k)$
  - Above maps extend componentwise

HOSVD Problem Revisited

- Optimizing over a product of Grassmann manifolds:

$$\max_{\{\mathbf{U}^{(n)}\}} \quad \tfrac{1}{2} \left\| (\mathbf{U}^{(1)\intercal}, \ldots, \mathbf{U}^{(N)\intercal}) \cdot \mathcal{X} \right\|_F^2$$

- $\mathbf{U}^{(n)}$ represents $\left[\mathbf{U}^{(n)}\right]$
  - Unconstrained optimization
  - Isolated local minimizers

Existing Grassmann Manifold Methods for Tucker Approximations

- Nonlinear Conjugate Gradient
- Newton's Method
- (Limited Memory) Quasi-Newton Methods
- Trust Region Method

Nonlinearly Preconditioned Conjugate Gradient

- Solve $\mathbf{g}(\mathbf{x}) := \nabla f(\mathbf{x}) = \mathbf{0}$
- Introduce nonlinear preconditioner P and define $\overline{\mathbf{g}}(\mathbf{x}) = \mathbf{x} - P(\mathbf{x})$
- NPCG iteration:

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_k \mathbf{p}_k, \\ \mathbf{p}_{k+1} &= -\overline{\mathbf{g}}(\mathbf{x}_{k+1}) + \overline{\beta}_{k+1}\mathbf{p}_k, \quad \mathbf{p}_0 = -\overline{\mathbf{g}}(\mathbf{x}_0).\end{aligned}$$

- $\overline{\beta}_{k+1}$ obtained from NCG β_{k+1} formulae
  - $\tilde{\beta}$: Replace each $\mathbf{g}$ with $\overline{\mathbf{g}}$
  - $\hat{\beta}$: Replace only certain instances of $\mathbf{g}$

For Tensor Approximation Problem

- $\mathbf{x} = (\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(N)})$
- $f(\mathbf{x}) = -\frac{1}{2}\left\|(\mathbf{U}^{(1)\intercal}, \ldots, \mathbf{U}^{(N)\intercal}) \cdot \mathcal{X}\right\|_F^2$
- $P(\mathbf{x}) :=$ one iteration of HOOI

Manifold Adaptation

Manifold NPCG iteration:

$$\mathbf{x}_{k+1} = R_{\mathbf{x}_k}(\alpha \mathbf{p}_k),$$
$$\mathbf{p}_{k+1} = -\overline{\mathbf{g}}_{k+1} + \overline{\beta}_{k+1} \mathcal{T}_{\mathbf{x}_{k+1}}(\mathbf{p}_k), \quad \mathbf{p}_0 = -\overline{\mathbf{g}}_0$$

- $\mathbf{p}_k$ and $\overline{\mathbf{g}}(\mathbf{x_k})$ are *tangent vectors* at $\mathbf{x}_k$
- Line search uses *retraction* $R_{\mathbf{x}_k}(\alpha \mathbf{p}_k)$
- Define $\overline{\mathbf{g}}(\mathbf{x}) := -\operatorname{Log}_{\mathbf{x}}(P(\mathbf{x}))$
- $\mathbf{p}_{k+1}$ requires a representation of $\mathbf{p}_k$ at $\mathbf{x}_{k+1}$: $\mathcal{T}_{\mathbf{x}_{k+1}}(\mathbf{p}_k)$
- $\mathbf{g}(\mathbf{x}) = \operatorname{grad} f(\mathbf{x})$
- $\overline{\beta}_{k+1}$ corresponding to $\beta^{\mathsf{HS}}_{k+1} = \frac{\mathbf{g}^{\mathsf{T}}_{k+1}(\mathbf{g}_{k+1} - \mathbf{g}_k)}{\mathbf{p}^{\mathsf{T}}_k(\mathbf{g}_{k+1} - \mathbf{g}_k)}$:

$$\tilde{\beta}^{\mathsf{HS}}_{k+1} = \frac{\langle \overline{\mathbf{g}}_{k+1}, \overline{\mathbf{y}}_k \rangle}{\langle \mathcal{T}_{\mathbf{x}_{k+1}}(\mathbf{p}_k), \overline{\mathbf{y}}_k \rangle} \quad \text{or} \quad \hat{\beta}^{\mathsf{HS}}_{k+1} = \frac{\langle \mathbf{g}_{k+1}, \overline{\mathbf{y}}_k \rangle}{\langle \mathcal{T}_{\mathbf{x}_{k+1}}(\mathbf{p}_k), \mathbf{y}_k \rangle}$$

where $\overline{\mathbf{y}}_k = \overline{\mathbf{g}}_{k+1} - \mathcal{T}_{\mathbf{x}_{k+1}}(\overline{\mathbf{g}}_k)$ and $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathcal{T}_{\mathbf{x}_{k+1}}(\mathbf{g}_k)$

Nonlinear GMRES

- Accelerates iteration $P(\cdot)$ solving $\mathbf{g}(\mathbf{x}) := \nabla f(\mathbf{x}) = \mathbf{0}$
- Given $\overline{\mathbf{x}}_{k+1} = P(\mathbf{x}_k)$ and past iterates $\{\mathbf{x}_j\}_{j=1}^k$, let

$$\hat{\mathbf{x}}_{k+1} = \overline{\mathbf{x}}_{k+1} + \sum_{j=1}^k \alpha_j(\overline{\mathbf{x}}_{k+1} - \mathbf{x}_j)$$

- Determine α_j which approximately minimize $\|\mathbf{g}(\hat{\mathbf{x}}_{k+1})\|_2$
- Linearizing $\mathbf{g}(\hat{\mathbf{x}}_{k+1})$ about $\overline{\mathbf{x}}_{k+1}$ gives a least squares problem:

$$\left\| \mathbf{g}(\overline{\mathbf{x}}_{k+1}) + \sum_{j=1}^k \alpha_j(\mathbf{g}(\overline{\mathbf{x}}_{k+1}) - \mathbf{g}(\mathbf{x}_j)) \right\|_2$$

- Line search in direction of $\hat{\mathbf{x}}_{k+1}$ determines $\mathbf{x}_{k+1}$:

$$\mathbf{x}_{k+1} = \overline{\mathbf{x}}_{k+1} + \beta(\hat{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_{k+1})$$

For Tensor Approximation Problem

- Define $\mathbf{x}$, $f(\mathbf{x})$, and $P(\mathbf{x})$ as before

Manifold Adaptation

- $\mathbf{g}(\mathbf{x}) = \text{grad}\, f(\mathbf{x})$
- Linearization of $\mathbf{g}(\hat{\mathbf{x}}_{k+1})$ about $\overline{\mathbf{x}}_{k+1}$:

$$\mathbf{g}(\hat{\mathbf{x}}_{k+1}) \approx \mathbf{g}(\overline{\mathbf{x}}_{k+1}) + \sum_{j=0}^{k} \alpha_j \text{Hess}(\overline{\mathbf{x}}_{k+1})[\boldsymbol{\xi}_j], \tag{1}$$

  where $\boldsymbol{\xi}_j = -\operatorname{Log}_{\overline{\mathbf{x}}_{k+1}}(\mathbf{x}_j)$
- Further approximation

$$\text{Hess}(\overline{\mathbf{x}}_{k+1})[\boldsymbol{\xi}_j] \approx \mathbf{g}(\overline{\mathbf{x}}_{k+1}) - \mathcal{T}_{\overline{\mathbf{x}}_{k+1}}(\mathbf{g}(\mathbf{x}_j)) \tag{2}$$

- Do not need $\hat{\mathbf{x}}_{k+1}$, only the tangent vector $\mathbf{p}_{k+1} = \hat{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_{k+1}$:

$$\mathbf{p}_{k+1} = -\sum_{j=0}^{k} \alpha_j \operatorname{Log}_{\overline{\mathbf{x}}_{k+1}}(\mathbf{x}_j)$$

- Line search carried out using $R_{\overline{\mathbf{x}}_{k+1}}(\alpha \mathbf{p}_{k+1})$

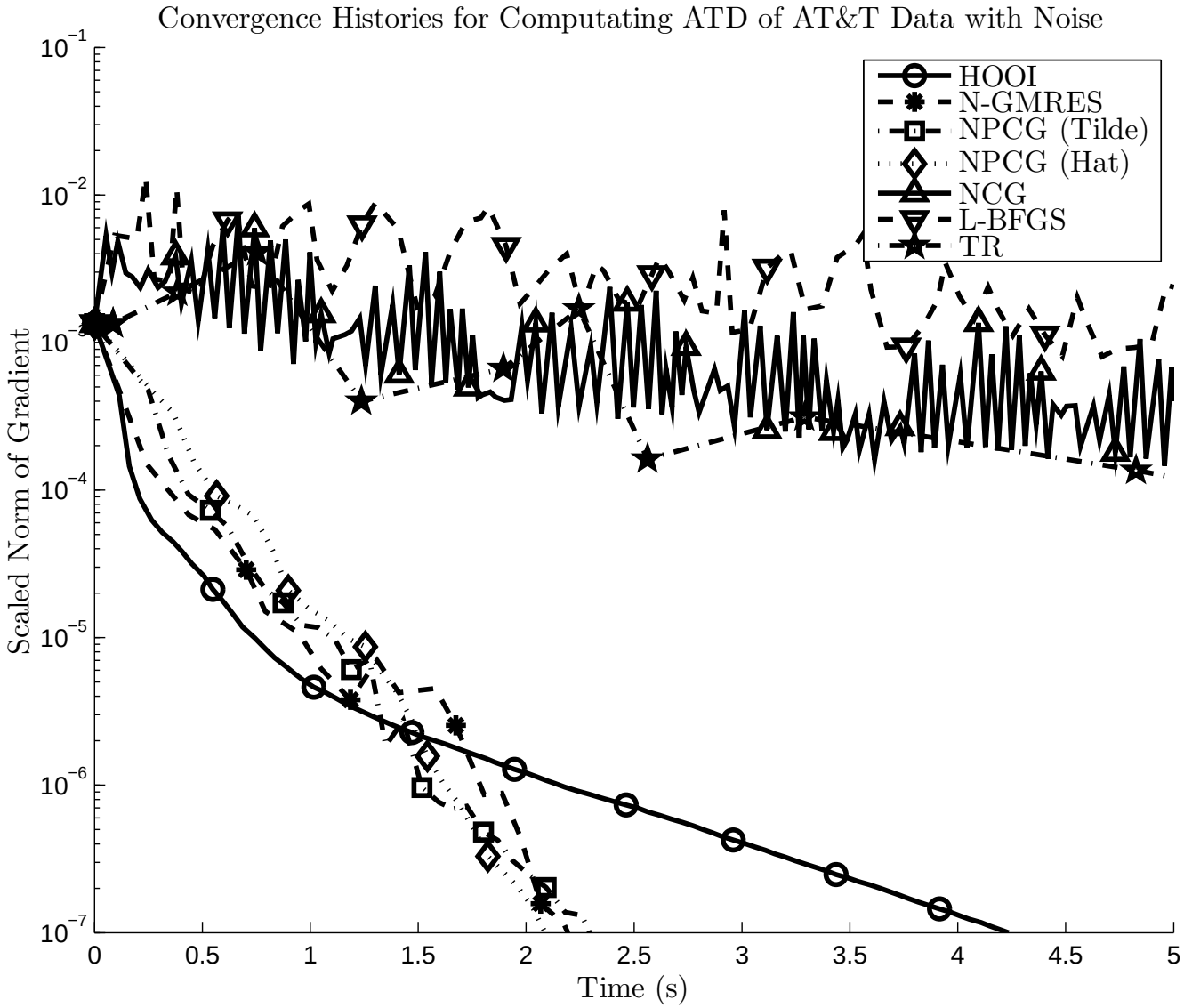


Figure: Convergence histories

- Tensor of size $92 \times 112 \times 10$, approximation with multilinear rank $(30, 35, 8)$.

Figure: From top to bottom: input tensor image, image from approx., noisy tensor image, and image from approx. of noisy tensor.

- Tensor of size $92 \times 112 \times 10$, approximation with multilinear rank $(30, 35, 8)$.

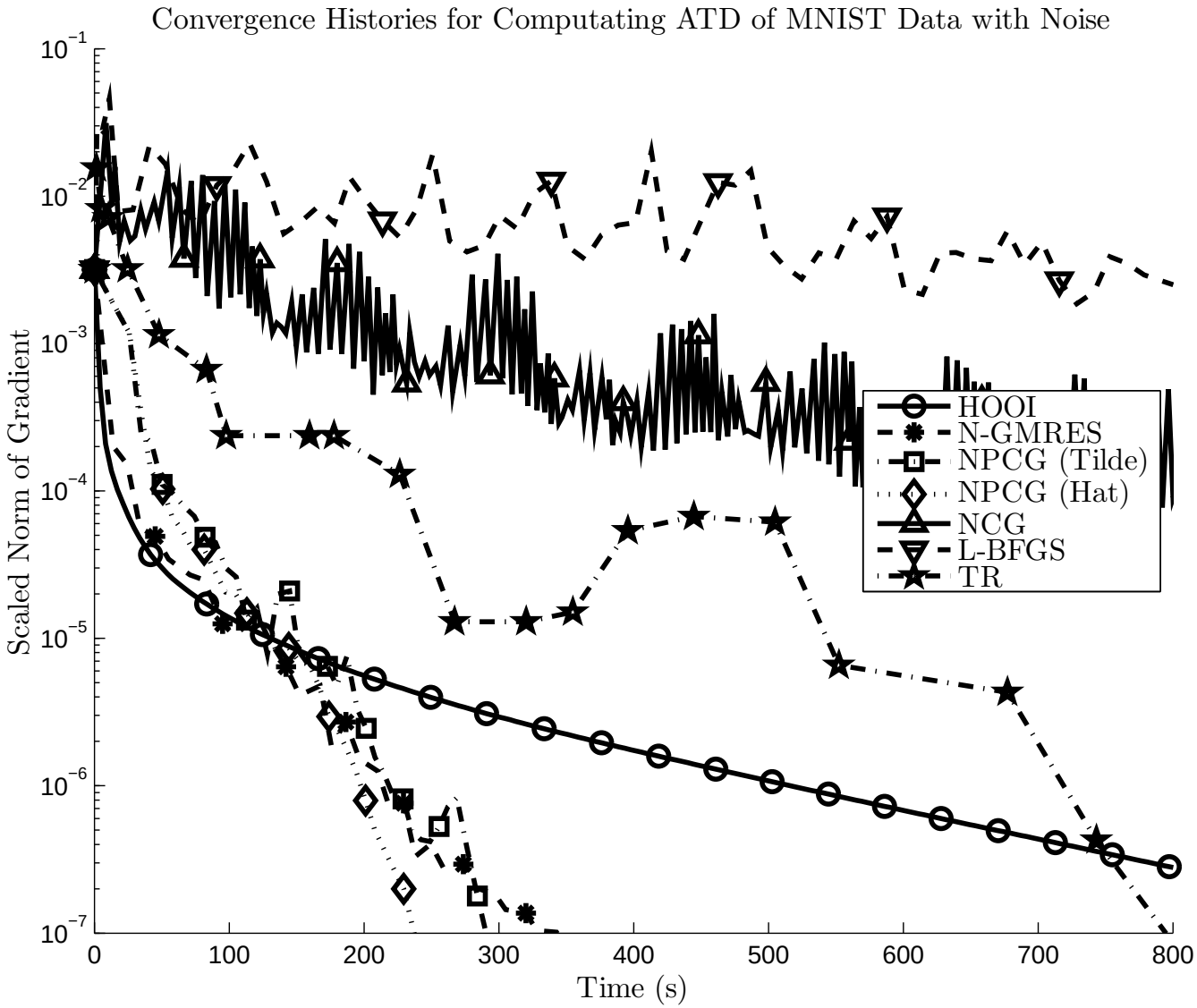


Figure: Convergence histories

- Tensor of size $28 \times 28 \times 5000$, approximation with multilinear rank $(14, 14, 100)$.

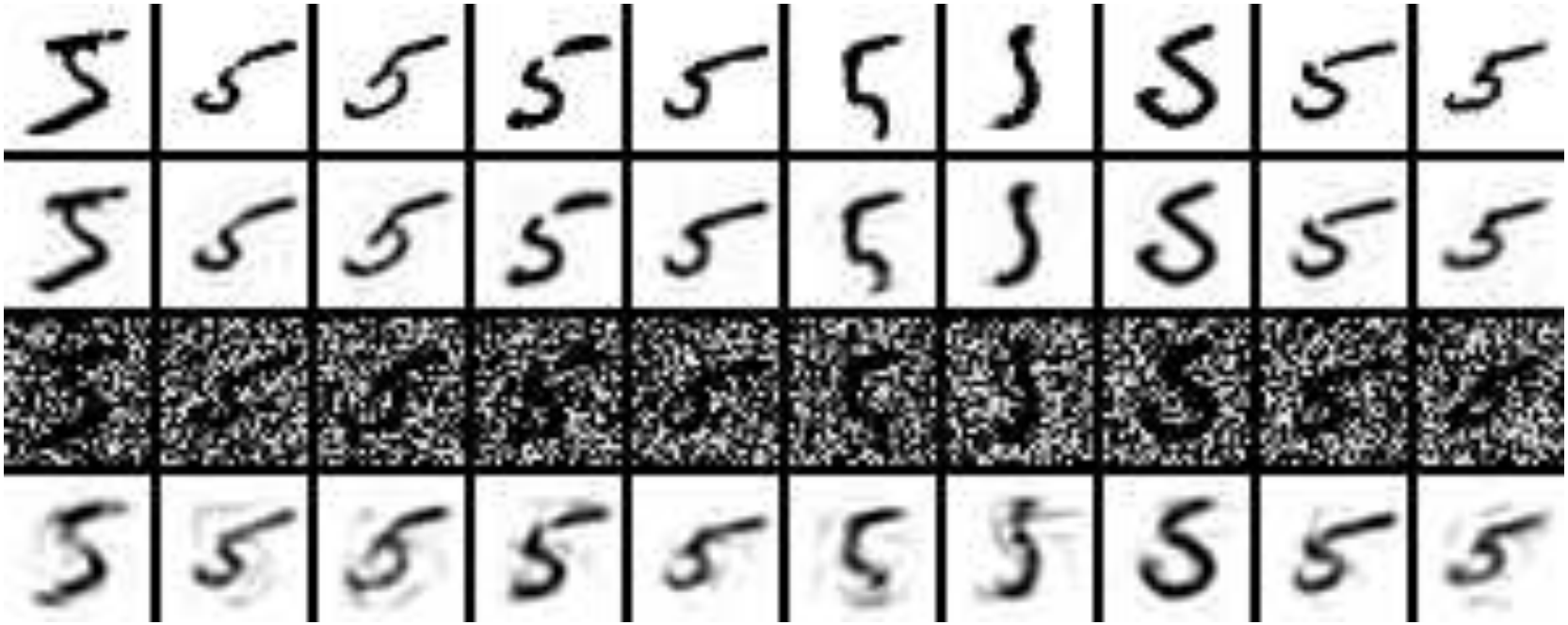


Figure: From top to bottom: input tensor image, image from approx., noisy tensor image, and image from approx. of noisy tensor.

- Tensor of size $28 \times 28 \times 5000$, approximation with multilinear rank $(14, 14, 100)$.

Thank you for listening