Riemannian preconditioning for tensor completion

Hiroyuki Kasai * and Bamdev Mishra †

* The university of Electro-Communications, Japan and Technische Universität München, Germany (kasai@is.uec.ac.jp)
† University of Liège, Belgium and University of Cambridge, UK (B.Mishra@ulg.ac.be)

Introduction

- We propose a novel Riemannian preconditioning approach for the tensor completion problem with rank constraint.
- The nuclear norm regularization approach from the matrix case to the tensor case leads to good results. But, its applicability to large-scale instances is not trivial, especially due to the necessity of high-dimensional SVD computations.
- A different approach exploits *Tucker decomposition* of a low-rank tensor $\mathcal{X}$ for large-scale algorithms.
- We propose *Riemannian preconditioning* that exploits the least-squares structure of the cost function and takes into account the structured symmetry in Tucker decomposition.

Tensor completion problem

- This research addresses the problem of low-rank tensor completion when the rank is a priori known or estimated below;

$$\min_{\mathcal{X}\in\mathbb{R}^{n_1\times n_2\times n_3}} \frac{1}{|\Omega|}\|\mathcal{P}_\Omega(\mathcal{X})-\mathcal{P}_\Omega(\mathcal{X}^\star)\|_F^2 \quad \text{subject to } \mathrm{rank}(\mathcal{X})=\mathbf{r}, \qquad (1)$$

- The operator $\mathcal{P}_\Omega(\mathcal{X})_{i_1i_2i_3}=\mathcal{X}_{i_1i_2i_3}$ if $(i_1,i_2,i_3)\in\Omega$ and $\mathcal{P}_\Omega(\mathcal{X})_{i_1i_2i_3}=0$ otherwise.
- $\mathrm{rank}(\mathcal{X})$ $(=\mathbf{r}=(r_1,r_2,r_3))$, called the *multilinear rank* of $\mathcal{X}$, is the set of the ranks of for each of mode-d unfolding matrices.
- $r_d \ll n_d$ enforces a low-rank structure.

Contributions

- We propose a novel *Riemannian metric* that exploits both the *symmetry* present in Tucker decomposition and the *least-squares* structure of problem (1).
- The specific metric allows to use the versatile framework of *Riemannian optimization* on *quotient manifolds* to develop a *Riemannian preconditioned* nonlinear conjugate gradient algorithm for the problem.
- To this end, concrete matrix representations of various optimization-related ingredients are listed.
- Numerical comparisons suggest that our proposed algorithm robustly outperforms state-of-the-art algorithms across different problem instances encompassing various synthetic and real-world datasets.

The quotient structure of Tucker decomposition

- The Tucker decomposition of a tensor $\mathcal{X}\in\mathbb{R}^{n_1\times n_2\times n_3}$ of rank $\mathbf{r}$ $(=(r_1,r_2,r_3))$ is $\mathcal{X}=\mathcal{G}\times_1\mathbf{U}_1\times_2\mathbf{U}_2\times_3\mathbf{U}_3$, where $\mathbf{U}_d\in\mathrm{St}(r_d,n_d)$ belongs to the *Stiefel manifold* of matrices of size $n_d\times r_d$ with orthogonal columns and $\mathcal{G}\in\mathbb{R}^{r_1\times r_2\times r_3}$.
- Because Tucker decomposition is *not unique* as $\mathcal{X}$ remains unchanged for all $\mathbf{O}_d\in\mathcal{O}(r_d)$, we define *equivalence classes* as,

$$[(\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3,\mathcal{G})]:=\{(\mathbf{U}_1\mathbf{O}_1,\mathbf{U}_2\mathbf{O}_2,\mathbf{U}_3\mathbf{O}_3,\mathcal{G}\times_1\mathbf{O}_1^T\times_2\mathbf{O}_2^T\times_3\mathbf{O}_3^T):\mathbf{O}_d\in\mathcal{O}(r_d)\}.$$

- The set of equivalence classes is the quotient manifold

$$\mathcal{M}/\sim\;:=\mathcal{M}/(\mathcal{O}(r_1)\times\mathcal{O}(r_2)\times\mathcal{O}(r_3)),$$

where $\mathcal{M}$ is called the *total space* (computational space) that is the product space $\mathcal{M}:=\mathrm{St}(r_1,n_1)\times\mathrm{St}(r_2,n_2)\times\mathrm{St}(r_3,n_3)\times\mathbb{R}^{r_1\times r_2\times r_3}$.

- Problem (1) is conceptually transformed into an unconstrained optimization problem on a *Riemannian quotient manifold* by endowing $\mathcal{M}/\sim$ with a Riemannian structure.

The least-squares structure of the cost function

- An induced metric resolves convergence issues of first-order optimization algorithms.
- Problem (1) is *convex and quadratic* in $\mathcal{X}$, and it is also convex and quadratic in the arguments $(\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3,\mathcal{G})$ individually.
- Block diagonal approximation of the Hessian as shown below of the simplified cost function $\|\mathcal{X}-\mathcal{X}^\star\|_F^2$ can be a good trade-off ingredients for a new metric.

$$((\mathbf{G}_1\mathbf{G}_1^T)\otimes\mathbf{I}_{n_1},(\mathbf{G}_2\mathbf{G}_2^T)\otimes\mathbf{I}_{n_2},(\mathbf{G}_3\mathbf{G}_3^T)\otimes\mathbf{I}_{n_3},\mathbf{I}_{r_1r_2r_3}).$$

A novel Riemannian metric

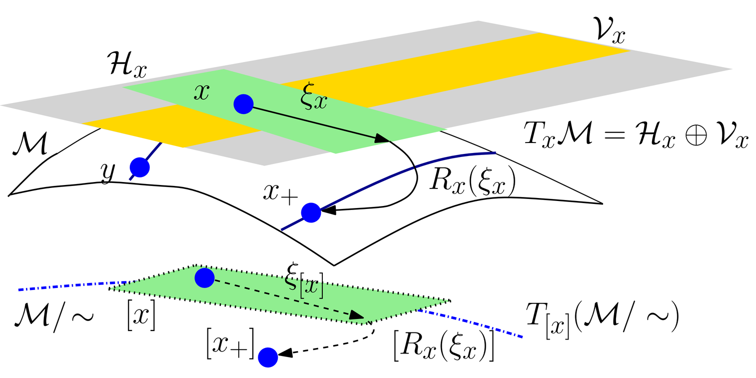
- From the discussions above on symmetry and least-squares structure, we propose the novel metric $g_x: T_x\mathcal{M}\times T_x\mathcal{M}\to\mathbb{R}$

$$g_x(\xi_x,\eta_x)=\langle\xi_{\mathbf{U}_1},\eta_{\mathbf{U}_1}(\mathbf{G}_1\mathbf{G}_1^T)\rangle+\langle\xi_{\mathbf{U}_2},\eta_{\mathbf{U}_2}(\mathbf{G}_2\mathbf{G}_2^T)\rangle+\langle\xi_{\mathbf{U}_3},\eta_{\mathbf{U}_3}(\mathbf{G}_3\mathbf{G}_3^T)\rangle+\langle\xi_{\mathcal{G}},\eta_{\mathcal{G}}\rangle,$$

where $\xi_x,\eta_x\in T_x\mathcal{M}$ are tangent vectors with matrix characterizations, $(\xi_{\mathbf{U}_1},\xi_{\mathbf{U}_2},\xi_{\mathbf{U}_3},\xi_{\mathcal{G}})$ and $(\eta_{\mathbf{U}_1},\eta_{\mathbf{U}_2},\eta_{\mathbf{U}_3},\eta_{\mathcal{G}})$, and $\langle\cdot,\cdot\rangle$ is the Euclidean inner product.

Quotient manifold and horizontal lift

- x and y belong to the same equivalence class and they represent a single point $[x]:=\{y\in\mathcal{M}:y\sim x\}$ on $\mathcal{M}/\sim$.
- Tangent space $T_x\mathcal{M}=\mathcal{V}_x\oplus\mathcal{H}_x$, where the vertical space $\mathcal{V}_x$ is the tangent space of the equivalence class $[x]$, and the horizontal space $\mathcal{H}_x$ is the *orthogonal subspace* to $\mathcal{V}_x$.



- $\xi_{[x]}\in T_{[x]}(\mathcal{M}/\sim)$ at $[x]$ has a unique element $\xi_x\in\mathcal{H}_x$ called its *horizontal lift*.

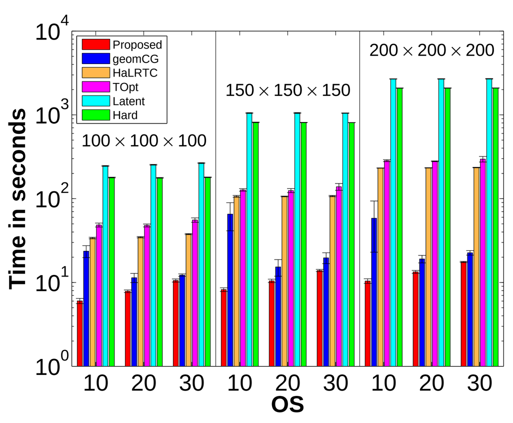
Ingredients to implement nonlinear CG with Manopt

$\Psi(\cdot)$ projects an ambient vector $(\mathbf{Y}_{\mathbf{U}_1},\mathbf{Y}_{\mathbf{U}_2},\mathbf{Y}_{\mathbf{U}_3},\mathbf{Y}_{\mathcal{G}})$ onto $T_x\mathcal{M}$	$(\mathbf{Y}_{\mathbf{U}_1}-\mathbf{U}_1\mathbf{S}_{\mathbf{U}_1}(\mathbf{G}_1\mathbf{G}_1^T)^{-1},\mathbf{Y}_{\mathbf{U}_2}-\mathbf{U}_2\mathbf{S}_{\mathbf{U}_2}(\mathbf{G}_2\mathbf{G}_2^T)^{-1},$ $\mathbf{Y}_{\mathbf{U}_3}-\mathbf{U}_3\mathbf{S}_{\mathbf{U}_3}(\mathbf{G}_3\mathbf{G}_3^T)^{-1},\mathbf{Y}_{\mathcal{G}})$, where $\mathbf{S}_{\mathbf{U}_d}$ are symmetric matrices.
$\Pi(\cdot)$ projects a tangent vector ξ onto $\mathcal{H}_x$	$(\xi_{\mathbf{U}_1}-\mathbf{U}_1\mathbf{\Omega}_1,\xi_{\mathbf{U}_2}-\mathbf{U}_2\mathbf{\Omega}_2,\xi_{\mathbf{U}_3}-\mathbf{U}_3\mathbf{\Omega}_3,$ $\xi_{\mathcal{G}}-(-(\mathcal{G}\times_1\mathbf{\Omega}_1+\mathcal{G}\times_2\mathbf{\Omega}_2+\mathcal{G}\times_3\mathbf{\Omega}_3)))$, $\mathbf{\Omega}_d$ are skew matrices.
First-order derivative of $f(x)$	$(\mathbf{S}_1(\mathbf{U}_3\otimes\mathbf{U}_2)\mathbf{G}_1^T,\mathbf{S}_2(\mathbf{U}_3\otimes\mathbf{U}_1)\mathbf{G}_2^T,\mathbf{S}_3(\mathbf{U}_2\otimes\mathbf{U}_1)\mathbf{G}_3^T),$ $\mathcal{S}\times_1\mathbf{U}_1^T\times_2\mathbf{U}_2^T\times_3\mathbf{U}_3^T)$, $\mathcal{S}=\frac{2}{\lvert\Omega\rvert}(\mathcal{P}_\Omega(\mathcal{G}\times_1\mathbf{U}_1\times_2\mathbf{U}_2\times_3\mathbf{U}_3)-\mathcal{P}_\Omega(\mathcal{X}^\star))$.
Retraction $R_x(\xi_x)$	$(\mathrm{uf}(\mathbf{U}_1+\xi_{\mathbf{U}_1}),\mathrm{uf}(\mathbf{U}_2+\xi_{\mathbf{U}_2}),\mathrm{uf}(\mathbf{U}_3+\xi_{\mathbf{U}_3}),\mathcal{G}+\xi_{\mathcal{G}})$
Horizontal lift of the vector transport $\mathcal{T}_{\eta_{[x]}}\xi_{[x]}$	$\Pi_{R_x(\eta_x)}(\Psi_{R_x(\eta_x)}(\xi_x))$

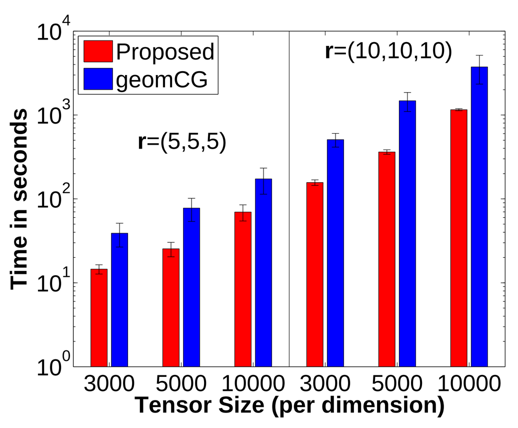
Experiments on synthetic datasets

We compare with Tucker decomposition based algorithms (geomCG, TOpt), and nuclear norm minimization algorithms (HaLRTC, Latent, Hard).

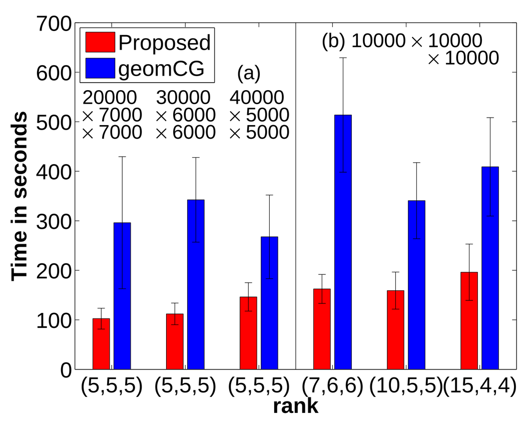
- **S1: small-scale instances.** Small-scale tensors of size $100\times100\times100$, $150\times150\times150$, and $200\times200\times200$ and rank $\mathbf{r}=(10,10,10)$. OS is $\{10,20,30\}$.
- **S2: large-scale instances.** We consider large-scale tensors of size $3000\times3000\times3000$, $5000\times5000\times5000$, and $10000\times10000\times10000$ and ranks $\mathbf{r}=(5,5,5)$ and $(10,10,10)$. OS is 10.
- **S3: asymmetric instances.** Case (3.a) considers $20000\times7000\times7000$, $30000\times6000\times6000$, and $40000\times5000\times5000$ with rank $\mathbf{r}=(5,5,5)$. Case (3.b) considers a tensor of size $10000\times10000\times10000$ with ranks $(7,6,6)$, $(10,5,5)$, and $(15,4,4)$.



S1:small-scale instances.



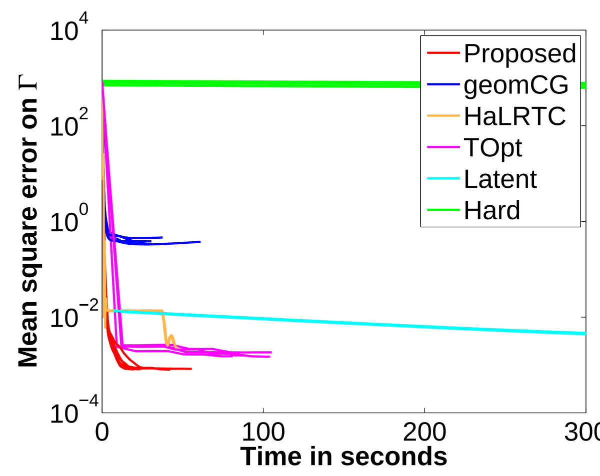
S2:large-scale isntances.



S3:asymmetric instances.

Experiments on Real-world datasets

- **R1: Hyperspectral image "Ribeira"** The tensor size is $203\times268\times33$ of the rank $\mathbf{r}$=(15, 15, 6), and OS are 11 and 22.
- **R2: MovieLens-10M** 71567 users and 10681 movies with 10000054 ratings by splitting into 7-days wide bins results in the tensor of size $71567\times10681\times731$.



R1: "Ribeira"

R1: Ribeira	OS = 11		OS = 22	
Algorithm	Time	MSE on Γ	Time	MSE on Γ
Proposed	**33 ± 13**	**$8.2095\cdot10^{-4}\pm1.7\cdot10^{-5}$**	67 ± 43	**$6.9516\cdot10^{-4}\pm1.1\cdot10^{-5}$**
geomCG	36 ± 14	$3.8342\cdot10^{-1}\pm4.2\cdot10^{-2}$	150 ± 48	$6.2590\cdot10^{-3}\pm4.5\cdot10^{-3}$
HaLRTC	46 ± 0	$2.2671\cdot10^{-3}\pm3.6\cdot10^{-5}$	48 ± 0	$1.3880\cdot10^{-3}\pm2.7\cdot10^{-5}$
TOpt	80 ± 32	$1.7854\cdot10^{-3}\pm3.8\cdot10^{-4}$	**27 ± 21**	$2.1259\cdot10^{-3}\pm3.8\cdot10^{-4}$
Latent	553 ± 3	$2.9296\cdot10^{-3}\pm6.4\cdot10^{-5}$	558 ± 3	$1.6339\cdot10^{-3}\pm2.3\cdot10^{-5}$
Hard	400 ± 5	$6.5090\cdot10^{2}\pm6.1\cdot10^{1}$	402 ± 4	$6.5989\cdot10^{2}\pm9.8\cdot10^{1}$
R2: MovieLens-10M	Proposed		geomCG	
r	Time	MSE on Γ	Time	MSE on Γ
(4, 4, 4)	**1748 ± 441**	**$0.6762\pm1.5\cdot10^{-3}$**	2981 ± 40	$0.6956\pm2.8\cdot10^{-3}$
(6, 6, 6)	**6058 ± 47**	**$0.6913\pm3.3\cdot10^{-3}$**	6554 ± 655	$0.7398\pm7.1\cdot10^{-3}$
(8, 8, 8)	**11370 ± 103**	**$0.7589\pm7.1\cdot10^{-3}$**	13853 ± 118	$0.8955\pm3.3\cdot10^{-2}$
(10, 10, 10)	**32802 ± 52**	**$1.0107\pm2.7\cdot10^{-2}$**	38145 ± 36	$1.6550\pm8.7\cdot10^{-2}$

This poster presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors. Bamdev Mishra is supported as an F.R.S.-FNRS research fellow (Belgian Fund for Scientific Research).