Large-scale Structured Low-rank Matrix Learning

Pratik Jawanpuria and Bamdev Mishra

amazon

IIT Madras Presentation

Applications of low-rank matrix completion

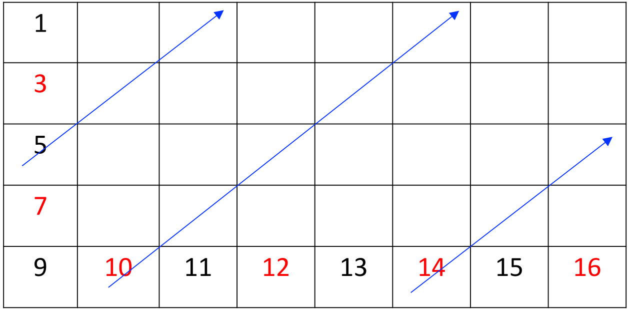
- Collaborative filtering (customers rank desirable items)
- Netflix problem (customers give ratings to movies)
  - 17 770 movies, 480 189 customers [Recht and Ré, 2013]
  - Predict missing entries ($\sim$ 99% entries are missing)

Applications of low-rank matrix completion with structural constraints

- Robust matrix completion – matrix is superimposition of a low-rank and a sparse matrix
  - Few "noisy" entries
- 0-1 matrix completion - entries are 0 or 1
  - Customers give "thumbs up" or "thumbs down" to products
- Bounded matrix completion - entries lie within a range
  - Amazon product ratings are between 5 and 1
  - Image completion (all entries are non-negative)
- Group of entries having linear constraints
  - "Godfather" should be among the highest rated movies
  - Probability constraints

Application: low-rank Hankel matrix learning

A rank 5 Hankel matrix



- All anti-diagonal entries are same
- System identification problems

Other applications of structured low-rank matrix learning

- Multi-task/multi-class learning
- Metric learning, distance matrix learning
- Sparse coding for image classification

Outline

Existing large-scale approaches for learning low-rank matrices

- Iterative Hard Thresholding (a.k.a. Singular Value Projection)

IHT [Jain et al., 2010]

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times T}} \quad \|\mathbf{W}_\Omega - \mathbf{Y}_\Omega\|$$
$$s.t \quad rank(\mathbf{W}) \leq r$$

Existing large-scale approaches for learning low-rank matrices

- Iterative Hard Thresholding (a.k.a. Singular Value Projection)
- Alternating Least Squares

ALS [Jain et al., 2013]

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times T}} \min_{\mathbf{U},\mathbf{V}} \|\mathbf{W}_\Omega - \mathbf{Y}_\Omega\|$$

$$s.t \quad \mathbf{W} = \mathbf{U}\mathbf{V}^\top, \mathbf{U} \in \mathbb{R}^{d\times r},\ \mathbf{V} \in \mathbb{R}^{T\times r}$$

Existing large-scale approaches for learning low-rank matrices

- Iterative Hard Thresholding (a.k.a. Singular Value Projection)
- Alternating Least Squares, Riemannian approaches

RTRMC [Boumal and Absil, 2011, 2015]

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times T}} \min_{\mathbf{U},\mathbf{V}} \|\mathbf{W}_\Omega - \mathbf{Y}_\Omega\| + \lambda\|\mathbf{W}_{\Omega^c}\|$$

$$s.t \quad \mathbf{W} = \mathbf{U}\mathbf{V}^\top, \mathbf{U}\in\mathbb{R}^{d\times r},\ \mathbf{V}\in\mathbb{R}^{T\times r}$$

$$\mathbf{U}\mathbf{V}^\top = (\mathbf{U}\mathbf{Z})(\mathbf{Z}^{-1}\mathbf{V}^\top), \mathbf{Z}\in\mathbb{R}^{r\times r}$$

Existing large-scale approaches for learning low-rank matrices

- Iterative Hard Thresholding (a.k.a. Singular Value Projection)
- Alternating Least Squares, Riemannian approaches
- Singular value thresholding, Active subspace algorithm

APGL [Toh and Yun, 2010], Active ALT [Hsieh and Olsen, 2014]

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times T}} \quad \|\mathbf{W}_\Omega - \mathbf{Y}_\Omega\| \ + \lambda \|\mathbf{W}\|_*$$

$$s.t \quad rank(\mathbf{W}) \leq r$$

$$\|\mathbf{W}\|_* = trace(\sqrt{\mathbf{W}^\top \mathbf{W}}) = \sum_i \sigma_i(\mathbf{W})$$

Existing large-scale approaches for learning low-rank matrices

- Iterative Hard Thresholding (a.k.a. Singular Value Projection)
- Alternating Least Squares, Riemannian approaches
- Singular value thresholding, Active subspace algorithm
- Hankel constraint – ADMM

DADM [Fazel et al., 2013]

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times T}} \quad \|\mathbf{W}_\Omega - \mathbf{Y}_\Omega\| + \lambda\|\mathbf{W}\|_*$$

$$s.t \quad \mathbf{AW} = \mathbf{0}$$

$$\|\mathbf{W}\|_* = trace(\sqrt{\mathbf{W}^\top \mathbf{W}}) = \sum_i \sigma_i(\mathbf{W})$$

Existing large-scale approaches for learning low-rank matrices

- Iterative Hard Thresholding (a.k.a. Singular Value Projection)
- Alternating Least Squares, Riemannian approaches
- Singular value thresholding, Active subspace algorithm
- Hankel constraint – ADMM, Conditional gradients (Frank-Wolfe)

GCG [Yu et al., 2014]

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times T}} \quad \|\mathbf{W}_\Omega - \mathbf{Y}_\Omega\| \; + \lambda\|\mathbf{W}\|_* + \beta\|\mathbf{AW}\|$$

$$s.t \quad \mathbf{AW} = \mathbf{0}$$

$$\|\mathbf{W}\|_* = trace(\sqrt{\mathbf{W}^\top\mathbf{W}}) = \sum_i \sigma_i(\mathbf{W})$$

Existing large-scale approaches for learning low-rank matrices

- Iterative Hard Thresholding (a.k.a. Singular Value Projection)
- Alternating Least Squares, Riemannian approaches
- Singular value thresholding, Active subspace algorithm
- Hankel constraint – ADMM, Conditional gradients (Frank-Wolfe)

Robust loss function

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times T}} \quad \|\mathbf{W}_\Omega - \mathbf{Y}_\Omega\|_1 + \lambda\|\mathbf{W}\|_* + \beta\|\mathbf{AW}\|$$

$$s.t \quad \mathbf{AW} = \mathbf{0}$$

$$\|\mathbf{W}\|_* = trace(\sqrt{\mathbf{W}^\top\mathbf{W}}) = \sum_i \sigma_i(\mathbf{W})$$

Existing large-scale approaches for learning low-rank matrices

- Iterative Hard Thresholding (a.k.a. Singular Value Projection)
- Alternating Least Squares, Riemannian approaches
- Singular value thresholding, Active subspace algorithm
- Hankel constraint – ADMM, Conditional gradients (Frank-Wolfe)

Robust loss function

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times T}} \quad \|\mathbf{W}_\Omega - \mathbf{Y}_\Omega\|_1 + \lambda\|\mathbf{W}\|_* + \beta\|\mathbf{A}\mathbf{W}\|$$

$$s.t \quad \mathbf{A}\mathbf{W} = \mathbf{0}$$

$$\|\mathbf{W}\|_* = trace(\sqrt{\mathbf{W}^\top\mathbf{W}}) = \sum_i \sigma_i(\mathbf{W})$$

Unified large-scale framework still missing!

Main contributions of our work

- Propose a novel decomposition of matrix $\mathbf{W} = \mathbf{U}\mathbf{U}^\top(\mathbf{Z} + \mathbf{S})$
  - $\mathbf{U}$: low-rank constraint, $\mathbf{S}$: structural constraints, and $\mathbf{Z}$: loss specific constraints
  - Makes optimization simpler and scalable
- General convex loss functions and subspace constraints
  - Existing large-scale approaches admit only a specific smooth loss
- Comparable against state-of-the-art in multiple applications
  - Standard/robust matrix completion
  - Hankel matrix learning
  - Multi-task learning
- Readily scale to the Netflix data set with non-smooth ℓ_1-loss and ϵ-SVR loss

Outline

Proposed framework

Primal formulation

$$\min_{\mathbf{W}} \|\mathbf{W} - \mathbf{Y}\|_F^2 \quad + \quad \lambda \|\mathbf{W}\|_*^2$$

$$s.t\ \mathbf{A}_t w_t = \mathbf{0}, \qquad \mathbf{W} = [w_1, \dots, w_T], \mathbf{Y} = [y_1, \dots, y_T]$$

Proposed framework

Primal formulation

$$\min_{\mathbf{W}} \|\mathbf{W} - \mathbf{Y}\|_F^2 \quad + \quad \lambda trace(\sqrt{\mathbf{W}^\top \mathbf{W}})^2$$

$$s.t \ \mathbf{A}_t w_t = \mathbf{0}, \qquad \mathbf{W} = [w_1, \ldots, w_T], \mathbf{Y} = [y_1, \ldots, y_T]$$

Proposed framework

Primal formulation

$$\min_{\Theta \succeq 0, \|\Theta\|_* = 1} \min_{\mathbf{W}} \|\mathbf{W} - \mathbf{Y}\|_F^2 + \lambda trace(\mathbf{W}^\top \Theta^\dagger \mathbf{W})$$

$$s.t\ \mathbf{A}_t w_t = \mathbf{0}, \qquad \mathbf{W} = [w_1, \dots, w_T], \mathbf{Y} = [y_1, \dots, y_T]$$

Proposed framework

Primal formulation

$$\min_{\Theta \succeq 0, \|\Theta\|_* = 1} \min_{\mathbf{W}} \|\mathbf{W} - \mathbf{Y}\|_F^2 + \lambda trace(\mathbf{W}^\top \Theta^\dagger \mathbf{W})$$

$$s.t\ \mathbf{A}_t w_t = \mathbf{0}, \qquad \mathbf{W} = [w_1, \ldots, w_T], \mathbf{Y} = [y_1, \ldots, y_T]$$

Proposed Dual formulation

$$\overbrace{\min_{\Theta \succeq 0, \|\Theta\|_* = 1} \underbrace{\max_{s_t, z_t \forall t} \sum_t y_t^\top z_t - \frac{\lambda}{2} z_t^\top z_t - \frac{1}{2}(z_t + \mathbf{A}_t^\top s_t)^\top \Theta (z_t + \mathbf{A}_t^\top s_t)}_{g(\mathbf{U}\mathbf{U}^\top)\text{ subproblem: application specific algorithm}}}^{\text{Optimization over } \Theta}$$

Proposed framework

Primal formulation

$$\min_{\Theta \succeq 0, \|\Theta\|_* = 1} \min_{\mathbf{W}} \|\mathbf{W} - \mathbf{Y}\|_F^2 + \lambda trace(\mathbf{W}^\top \Theta^\dagger \mathbf{W})$$

$$s.t\ \mathbf{A}_t w_t = \mathbf{0}, \qquad \mathbf{W} = [w_1, \ldots, w_T], \mathbf{Y} = [y_1, \ldots, y_T]$$

Proposed Dual formulation

$$\overbrace{\min_{\Theta \succeq 0, \|\Theta\|_* = 1} \underbrace{\max_{s_t, z_t \forall t} \sum_t y_t^\top z_t - \frac{\lambda}{2} z_t^\top z_t - \frac{1}{2}(z_t + \mathbf{A}_t^\top s_t)^\top \Theta (z_t + \mathbf{A}_t^\top s_t)}_{g(\mathbf{U}\mathbf{U}^\top) \text{ subproblem: application specific algorithm}}}^{\text{Optimization over } \Theta}$$

$\Theta = \mathbf{U}\mathbf{U}^\top, \mathbf{U} \in \mathbb{R}^{d \times r}, \|\Theta\|_* = 1 \Leftrightarrow \|\mathbf{U}\|_F = 1 \Leftrightarrow$ Spectrahedron Manifold

Proposed framework

Primal formulation

$$\min_{\Theta \succeq 0, \|\Theta\|_* = 1} \; \min_{\mathbf{W}} \; \|\mathbf{W} - \mathbf{Y}\|_F^2 \; + \; \lambda trace(\mathbf{W}^\top \Theta^\dagger \mathbf{W})$$

$$s.t \; \mathbf{A}_t w_t = \mathbf{0}, \qquad \mathbf{W} = [w_1, \ldots, w_T], \mathbf{Y} = [y_1, \ldots, y_T]$$

Proposed Dual formulation

$$\overbrace{\min_{\mathbf{U}:\|\mathbf{U}\|_F = 1} \; \underbrace{\max_{s_t, z_t \forall t} \; \sum_t y_t^\top z_t - \frac{\lambda}{2} z_t^\top z_t - \frac{1}{2}(z_t + \mathbf{A}_t^\top s_t)^\top \mathbf{U}\mathbf{U}^\top (z_t + \mathbf{A}_t^\top s_t)}_{g(\mathbf{U}\mathbf{U}^\top) \text{ subproblem: application specific algorithm}}}^{\text{Optimization over } \mathbf{U}}$$

$\Theta = \mathbf{U}\mathbf{U}^\top, \mathbf{U} \in \mathbb{R}^{d \times r}, \|\Theta\|_* = 1 \Leftrightarrow \|\mathbf{U}\|_F = 1 \Leftrightarrow$ Spectrahedron Manifold

Proposed framework

Primal formulation

$$\min_{\Theta \succeq 0, \|\Theta\|_* = 1} \; \min_{\mathbf{W}} \; \|\mathbf{W} - \mathbf{Y}\|_F^2 \; + \; \lambda trace(\mathbf{W}^\top \Theta^\dagger \mathbf{W})$$

$$s.t \; \mathbf{A}_t w_t = \mathbf{0}, \qquad \mathbf{W} = [w_1, \ldots, w_T], \mathbf{Y} = [y_1, \ldots, y_T]$$

Proposed Dual formulation

$$\overbrace{\min_{\mathbf{U}: \|\mathbf{U}\|_F = 1} \; \underbrace{\max_{s_t, z_t \forall t} \; \sum_t y_t^\top z_t - \frac{\lambda}{2} z_t^\top z_t - \frac{1}{2}(z_t + \mathbf{A}_t^\top s_t)^\top \mathbf{U}\mathbf{U}^\top (z_t + \mathbf{A}_t^\top s_t)}_{g(\mathbf{U}\mathbf{U}^\top) \text{ subproblem: application specific algorithm}}}^{\text{Optimization over } \mathbf{U}}$$

$\Theta = \mathbf{U}\mathbf{U}^\top, \mathbf{U} \in \mathbb{R}^{d \times r}, \|\Theta\|_* = 1 \Leftrightarrow \|\mathbf{U}\|_F = 1 \Leftrightarrow$ Spectrahedron Manifold

Efficient Riemannian first- and second-order algorithms

Subproblem and its solver for various applications

Application	$g(\mathbf{U}\mathbf{U}^\top)$ subproblem	Algorithm
Hankel Matrix Learning	$\max_{s_t \in \mathbb{R}^d \forall t, z \in \mathbb{R}^{d+T-1}} \langle y, z \rangle - \frac{1}{4C} \langle z, z \rangle - \frac{1}{2} \sum_{t=1}^{T} \left\langle \mathbf{U}^\top s_t, \mathbf{U}^\top s_t \right\rangle$ subject to : $z_k - \sum_{\substack{(i,t): i+t=k, \\ 1 \leq i \leq d, 1 \leq t \leq T}} s_{ti} = 0 \ \forall k = 2, \ldots, d+T$	Preconditioned conjugate gradients
Robust Matrix Completion	$\sum_{t=1}^{T} \max_{z_t \in [-C, C]^{m_t}} \langle y_t, z_t \rangle - \frac{1}{2} \left\langle \mathbf{U}^\top z_t, \mathbf{U}^\top z_t \right\rangle$	Dual coordinate descent§
Matrix Completion	$\sum_{t=1}^{T} \max_{z_t} \langle y_t, z_t \rangle - \frac{1}{4C} \langle z_t, z_t \rangle - \frac{1}{2} \left\langle \mathbf{U}^\top z_t, \mathbf{U}^\top z_t \right\rangle$	Least square solver§
Multi-task Feature Learning	$\sum_{t=1}^{T} \max_{z_t} \langle y_t, z_t \rangle - \frac{1}{4C} \langle z_t, z_t \rangle - \frac{1}{2} \left\langle \mathbf{U}^\top \mathbf{X}_t z_t, \mathbf{U}^\top \mathbf{X}_t z_t \right\rangle$	Least square solver§

§ Problem for each t can be solved in parallel

Outline of the proposed algorithm

Algorithm 1 Proposed first- and second-order algorithms

Input: $\{y_t\}_{t=1}^{T}$, rank r, regularization parameter C.
Initialize $\mathbf{U}$.
repeat
  1: Solve for $\{\mathbf{Z}, \mathbf{S}\}$ by solving the subproblem $g(\mathbf{U}\mathbf{U}^\top)$.
  2: Compute $\nabla_{\mathbf{U}} g(\mathbf{U}\mathbf{U}^\top)$.
  3: Riemannian conjugate-gradients step. | **3:** Riemannian trust-regions step.
  4: Update $\mathbf{U}$ (retraction step).
until convergence
Output: $\{\mathbf{U}, \mathbf{Z}, \mathbf{S}\}$ and $\mathbf{W} = \mathbf{U}\mathbf{U}^\top(\mathbf{Z} + \mathbf{S})$.

Salient features of our framework

- Algorithms for both batch and online setting
  - Columns (customers) are added daily/periodically
  - Learn the missing entries only for these additional columns
- Parallelizable across columns, with column-wise structural constraints
- Highly efficient for rectangular matrices ($d \ll T$)

Duality gap criterion for verifying global optimality

Using the duality framework, we can efficiently compute the duality gap of a feasible solution, $\mathbf{W} = \mathbf{U}\mathbf{U}^\top(\mathbf{Z} + \mathbf{S})$, with respect to the (unknown) optimal $\mathbf{W}^*$.

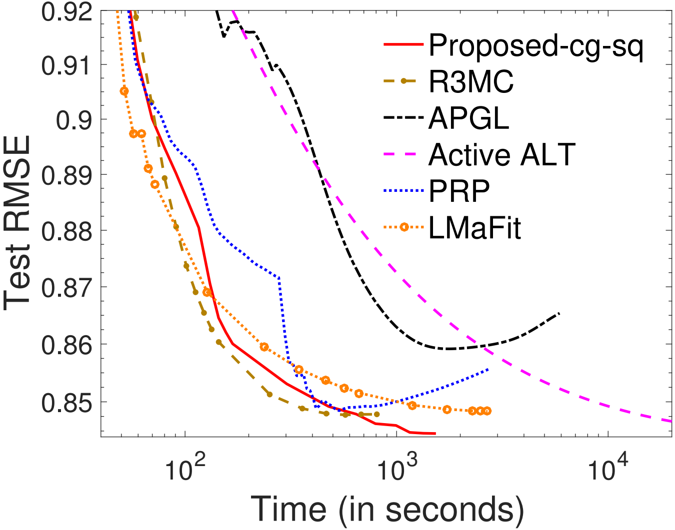
Outline

Experimental setup for matrix completion

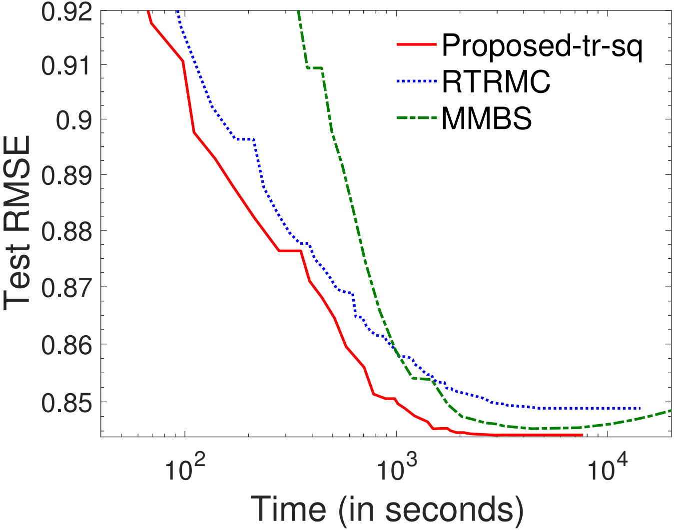
Data set	d	T	$\|\Omega\|$
Netflix	17 770	480 189	100 198 805
MovieLens10m	10 677	71 567	10 000 054
MovieLens20m	26 744	138 493	20 000 263

- Compared approaches
  - APGL [Toh and Yun, 2010] (Singular value thresholding)
  - Active ALT [Hsieh and Olsen, 2014] (Active subspace selection)
  - MMBS [Mishra et al., 2013], R3MC [Mishra and Sepulchre, 2014], RTRMC [Boumal and Absil, 2011, 2015], PRP [Tan et al., 2016] (Riemannian methods)
  - LMaFit [Wen et al., 2012] (Alternate least-squares)
  - Proposed-cg-sq, Proposed-tr-sq

First-order approaches: Proposed approach achieves the best generalization result on the Netflix dataset



Second-order approaches: Proposed approach achieves the best generalization result on the Netflix dataset



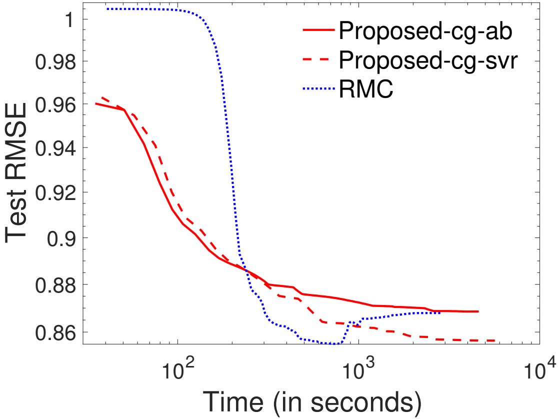
Proposed approach achieves the best generalization result in matrix completion experiments

Table: Generalization error (test RMSE) on matrix completion datasets.

	Netflix	ML10m	ML20m
Proposed-cg-sq	**0.8449**	**0.8026**	**0.7963**
Proposed-tr-sq	**0.8443**	**0.8026**	**0.7962**
R3MC	0.8478	0.8070	0.7982
RTRMC	0.8489	0.8161	0.8044

Proposed approach scales on the Netflix data set even with non-smooth ℓ_1-norm loss and ϵ-SVR loss

Robust matrix completion experiment



RMC [Cambier and Absil, 2016] optimizes smooth pseudo-Huber loss

Proposed approach learns closest low-rank approximation of the given Hankel matrix

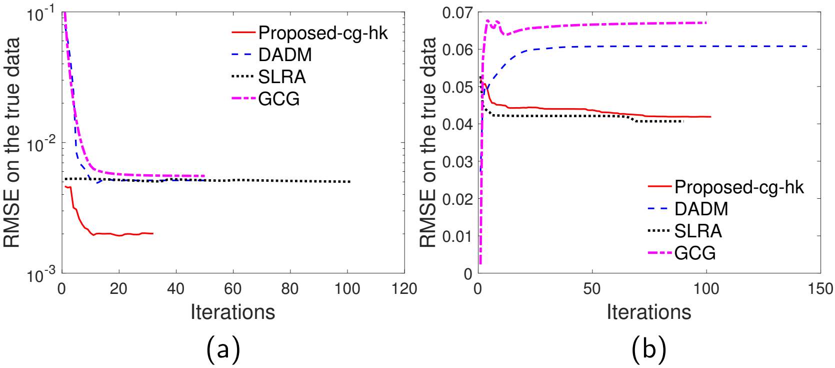


Figure: (a) System identification, and (b) Stochastic system realization

GCG [Yu et al., 2014]

SLRA [Markovsky, 2014, Markovsky and Usevich, 2014]

DADM [Fazel et al., 2013]

Proposed approach achieves better generalization result at lower ranks on multi-task regression data sets

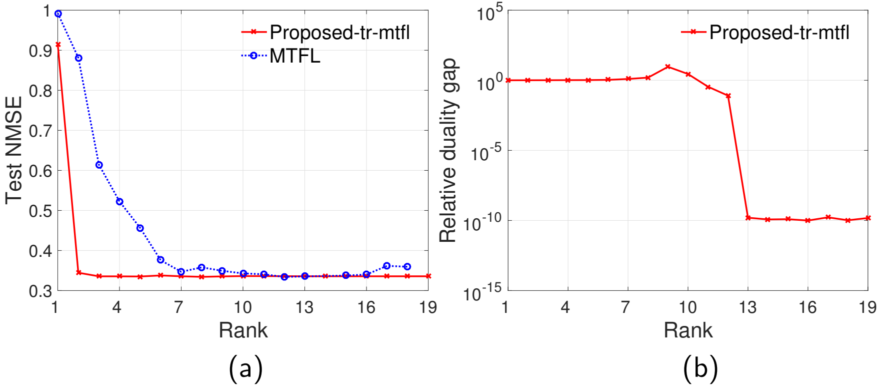


Figure: (a) Normalized mean squared error (NMSE) vs rank, and (b) Relative duality gap vs rank.

MTFL [Argyriou et al., 2008]

Summary

- Generic large-scale framework for learning low-rank matrices with structural constraints
- Novel decomposition for structured low-rank matrix learning: $\mathbf{W} = \mathbf{U}\mathbf{U}^\top(\mathbf{Z} + \mathbf{S})$
- Decoupling of factors - leads to simple and scalable optimization approach
- Local/global optimality guarantees
- State-of-the-art generalization performance on different applications
- ArXiv preprint and codes available: www.pratikjawanpuria.com

Future Work

- Non-linear constraints such as **W** being positive semi-definite matrix
  - Employed in metric learning, distance matrix learning
- We have recently extended the saddle point approach to Tensors. A manuscript is in progress
- Our approach gives novel insights in following the solution path of **W** with respect to rank r. A manuscript is in progress
- Recover structured probability matrices, given
  - Number of times customer i buy item j
  - Number of times word i is followed by word j

Large-scale Structured Low-rank Matrix Learning

Pratik Jawanpuria and Bamdev Mishra

IIT Madras Presentation

Results on global optimality of the obtained solution

Lemma

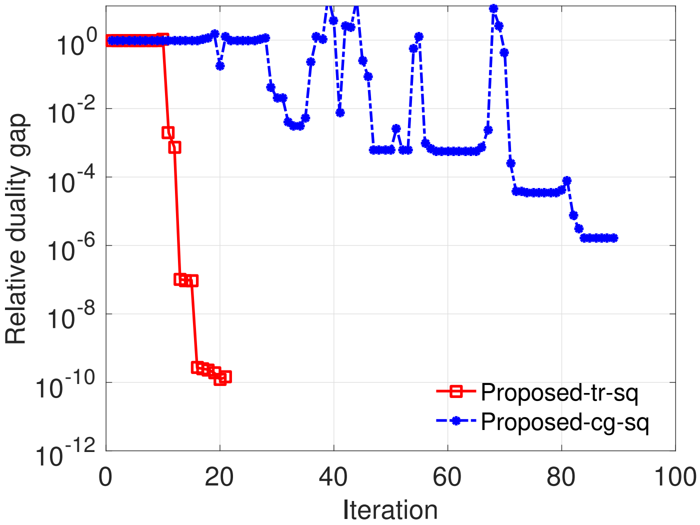
Let $\mathbf{U}^*$ be a local minimizer of the proposed fixed-rank $(= r)$ non-convex formulation. If $rank(\mathbf{U}^*) < r$ or $r = d$, then $\Theta^* = \mathbf{U}^*(\mathbf{U}^*)^\top$ is a stationary point for the convex trace-norm squared minimization problem.

Proposition

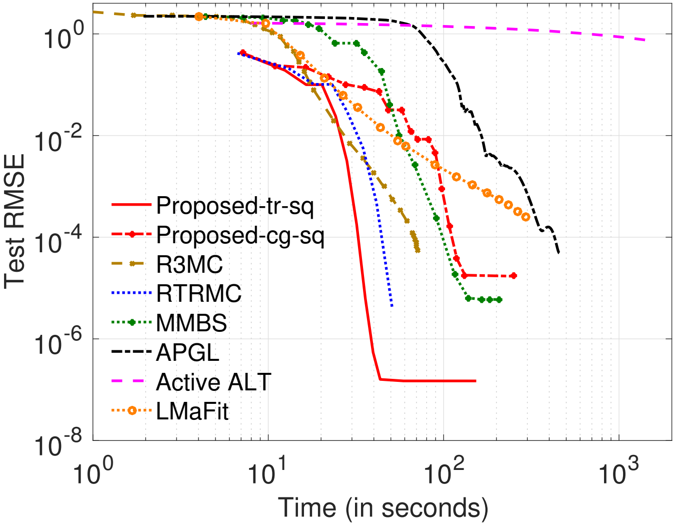
Let $\hat{\mathbf{U}}$ be a feasible solution. Let $\{\hat{\mathbf{Z}}, \hat{\mathbf{S}}\}$ be an optimal solution of the convex problem $g(\mathbf{U}\mathbf{U}^\top)$ at $\hat{\mathbf{U}}$. In addition, let σ_1 be the maximum singular of the the matrix $\hat{\mathbf{Z}} + \hat{\mathbf{S}}$. The duality gap (Δ) associated with $\{\hat{\mathbf{U}}, \hat{\mathbf{Z}}, \hat{\mathbf{S}}\}$ is given by

$$\Delta = \frac{1}{2}\left(\sigma_1^2 - \left\|\hat{\mathbf{U}}^\top(\hat{\mathbf{Z}} + \hat{\mathbf{S}})\right\|^2\right).$$

Second-order algorithm usually shows faster and more stable convergence than first-order algorithm



Second-order algorithm usually shows faster and more stable convergence than first-order algorithm



A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73:243–272, 2008.

N. Boumal and P.-A. Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 406–414, 2011.

N. Boumal and P.-A. Absil. Low-rank matrix completion via preconditioned optimization on the Grassmann manifold. *Linear Algebra and its Applications*, 475:200–239, 2015.

L. Cambier and P. A. Absil. Robust low-rank matrix completion by Riemannian optimization. *SIAM J. Sci. Comput.*, 38(5):S440—S460, 2016.

M. Fazel, P. T. Kei, D. Sun, and P. Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977, 2013.

C.-J. Hsieh and P. A. Olsen. Nuclear norm minimization via active subspace selection. In *International Conference on Machine learning (ICML)*, 2014.

P. Jain, R. Meka, and I. Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems 23 (NIPS)*, pages 937–945, 2010.

P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *ACM Symposium on Theory of Computing (STOC)*, 2013.

I. Markovsky. Recent progress on variable projection methods for structured low-rank approximation. *Signal Processing*, 96(Part B): 406–419, 2014.

I. Markovsky and K. Usevich. Software for weighted structured low-rank approximation. *J. Comput. Appl. Math.*, 256:278–292, 2014.

B. Mishra and R. Sepulchre. R3MC: A Riemannian three-factor algorithm for low-rank matrix completion. In *Proceedings of the 53rd IEEE Conference on Decision and Control (CDC)*, pages 1137–1142, 2014.

B. Mishra, G. Meyer, F. Bach, and R. Sepulchre. Low-rank optimization with trace norm penalty. *SIAM Journal on Optimization*, 23(4): 2124–2149, 2013.

B. Recht and C Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2): 201–226, 2013.

M. Tan, S. Xiao, J. Gao, D. Xu, A. van den Hengel, and Q. Shi. Proximal riemannian pursuit for large-scale trace-norm minimization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

K. C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pacific Journal of Optimization*, 6(3):615–640, 2010.

Z. Wen, W. Yin, and Y. Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361, 2012.

A. W. Yu, Ma W., Yu Y., Carbonell J. G., and Sra S. Efficient structured matrix rank minimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.