Convex Matrix Completion: A Trace-Ball Optimization Perspective

Guangxiang Zeng[1], Ping Luo[2], Enhong Chen[1], Hui Xiong[3], Hengshu Zhu[4] and Qi Liu[1]

[1]University of Science and Technology of China

[2]Institute of Computing Technology, Chinese Academy of Sciences

[3]Rutgers University

[4]Baidu Research-Big Data Lab

Outline

- Background
- Motivation
- Problem Statement
- From Local Optimum to Global Optimum
- Trace Ball Optimization
- Running Example
- Discussion on the Parameter
- Experiment on Stable Parameter
- Conclusion

Background

- Matrix Completion

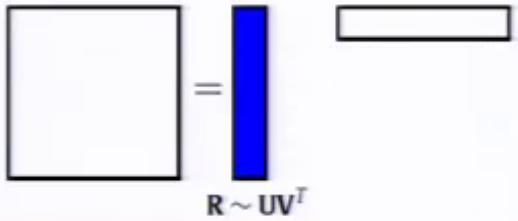
1	3	5	?	2
?	2	?	1	?
1	?	5	?	3
?	3	?	1	?

Background

- Matrix Completion

1	3	5	?	2
?	2	?	1	?
1	?	5	?	3
?	3	?	1	?

SVD based methods



e.g., SVD, PMF, SVD++

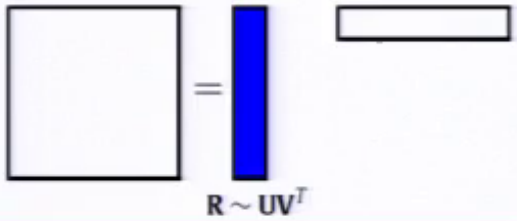
Non-convex low rank matrix completion

Background

- Matrix Completion

1	3	5	?	2
?	2	?	1	?
1	?	5	?	3
?	3	?	1	?

SVD based methods



e.g., SVD, PMF, SVD++

Non-convex low rank matrix completion

Trace norm (nuclear norm) regularization methods

e.g., Maximum-margin matrix factorization

Semi-Definite Programming (SDP) convex matrix completion

Motivation

- Non-convex low rank matrix completion (e.g., SVD, PMF)
  - **Strengths:** 1) First-order optimization method for searching solution
    2) Easy to be generalized to large scale problems
  - **Weakness:** 1) Local optimum
    2) The solution relies on the initial value

Motivation

- Non-convex low rank matrix completion (e.g., SVD, PMF)
  - **Strengths:** 1) First-order optimization method for searching solution
    2) Easy to be generalized to large scale problems
  - **Weakness:** 1) Local optimum
    2) The solution relies on the initial value
- Traditional SDP convex matrix completion (e.g., Maximum-margin matrix factorization)
  - **Strengths:** 1) Global optimum
    2) Robustness
  - **Weakness:** 1) Second-order method (e.g., interior point method)
    2) Can not be generalized to large scale problem due to storage inefficient
- Can we combine these two methods to keep their strengths and abandon the weakness?

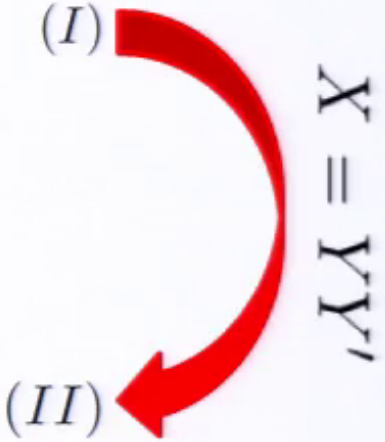
Problem Statement

- Bounding trace norm

$$\begin{aligned} \min \quad & F(X) = \sum_{(i,j)\in\Omega} (\widehat{R}_{ij} - R_{ij})^2 \\ s.t. \quad & X = \begin{bmatrix} W_1 & \widehat{R} \\ \widehat{R}' & W_2 \end{bmatrix} \succeq 0 \\ & tr(X) \le \gamma, \end{aligned}$$

- Non-convex relaxation

$$\begin{aligned} \min \quad & F(YY') \\ s.t. \quad & tr(YY') \le \gamma. \end{aligned}$$



- We aim at solving Problem (I) through solving Problem (II)

From Local Optimum to Global Optimum

- First-order optimum conditions (KKT conditions)

LEMMA 3.1. *Provided that $F(X)$ is a convex function of X. X is a optimal solution of Problem (I) if and only if there exists a $\sigma \in \mathbb{R}_+$ and a symmetric matrix $S \in \mathbb{S}^N$ such that the following conclusions hold:*

$$
\begin{array}{lrcl}
1) & tr(X) & \leq & \gamma, \\
2) & X & \succeq & 0, \\
3) & \sigma(tr(X)-\gamma) & = & 0, \\
4) & S & \succeq & 0, \\
5) & SX & = & 0,
\end{array} \tag{3.1}
$$

where $S = \nabla_X F(X) + \sigma I$.

LEMMA 3.2. *If Y is a local optimum of Problem (II), then there exists an $\alpha \in \mathbb{R}_+$ such that*

$$
\begin{array}{lrcl}
1) & tr(YY') & \leq & \gamma, \\
2) & \alpha(tr(YY')-\gamma) & = & 0, \\
3) & S_Y Y & = & 0.
\end{array}
$$

where $S_Y = \nabla_X F(YY') + \alpha I$

THEOREM 3.1. *Provided that $F(X)$ is a convex function of X. An Y which satisfies the conclusions of Lemma 3.2 provides a global minimum point YY' of Problem (I) if the matrix*

$$S_Y = \nabla_X F(YY') + \alpha I \tag{3.2}$$

is positive semi-definite, where $\alpha = -\frac{\nabla_X F(YY') \circ YY'}{tr(YY')}$.

From Local Optimum to Global Optimum

- First-order optimum conditions (KKT conditions)

LEMMA 3.1. *Provided that $F(X)$ is a convex function of X. X is a optimal solution of Problem (I) if and only if there exists a $\sigma \in \mathbb{R}_+$ and a symmetric matrix $S \in \mathbb{S}^N$ such that the following conclusions hold:*

$$
\begin{array}{lrcl}
1) & tr(X) & \leq & \gamma, \\
2) & X & \succeq & 0, \\
3) & \sigma(tr(X)-\gamma) & = & 0, \\
4) & S & \succeq & 0, \\
5) & SX & = & 0,
\end{array}
\qquad (3.1)
$$

where $S = \nabla_X F(X) + \sigma I$.

LEMMA 3.2. *If Y is a local optimum of Problem (II), then there exists an $\alpha \in \mathbb{R}_+$ such that*

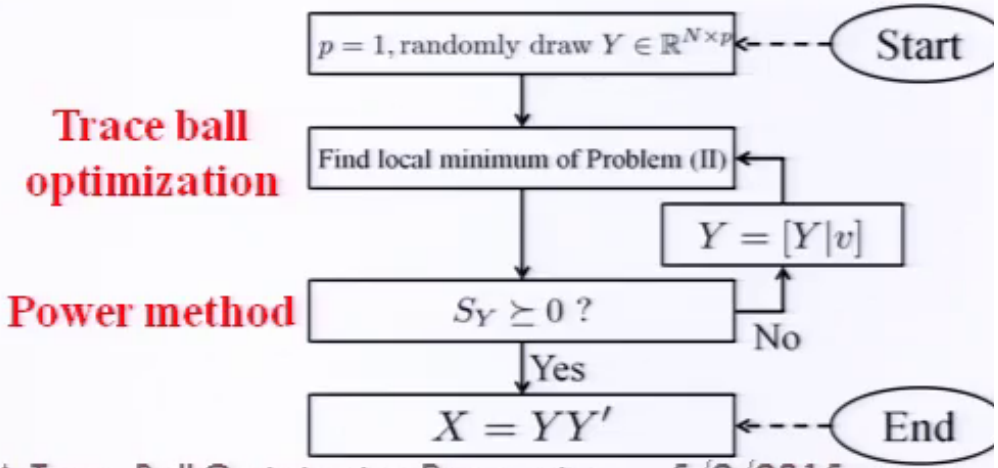
$$
\begin{array}{lrcl}
1) & tr(YY') & \leq & \gamma, \\
2) & \alpha(tr(YY')-\gamma) & = & 0, \\
3) & S_Y Y & = & 0.
\end{array}
$$

where $S_Y = \nabla_X F(YY') + \alpha I$

THEOREM 3.1. *Provided that $F(X)$ is a convex function of X. An Y which satisfies the conclusions of Lemma 3.2 provides a global minimum point YY' of Problem (I) if the matrix*

$$S_Y = \nabla_X F(YY') + \alpha I \qquad (3.2)$$

is positive semi-definite, where $\alpha = -\frac{\nabla_X F(YY') \circ YY'}{tr(YY')}$.

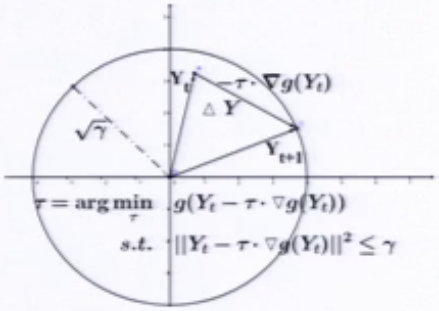


Trace Ball Optimization

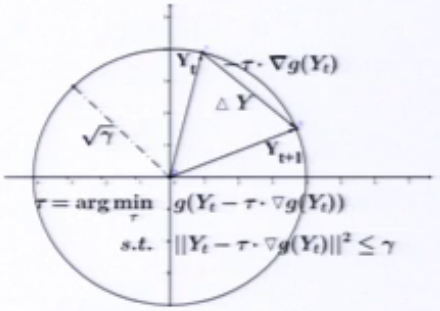
- Aim at solving Problem (II)

$$\begin{aligned} \min \quad & G(Y) = F(YY') \\ s.t. \quad & tr(YY') \le \gamma \end{aligned}$$

- Gradient decent with projection, search local minimum in a ball with radius $\sqrt{\gamma}$.

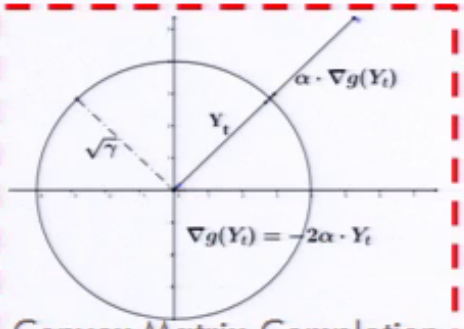


Case 1

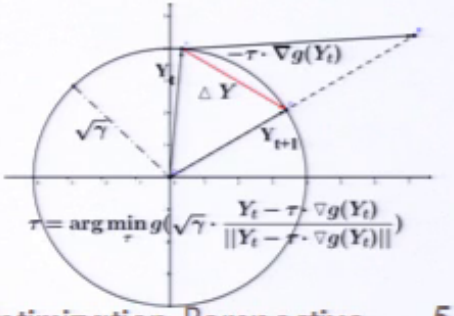


Case 2

KKT conditions is met!



Case 3



Case 4

Trace Ball Optimization

- When $S_Y \not\succeq 0$, we get the minimum eigenvalue of S_Y, and its corresponding eigenvector, $\{\rho_{min}, v\}$.

- We can search a better point for Problem (II) on point $Y = [Y|0_1]$ by direction $Z = [0_p|v]$, and solve Problem (II) again.

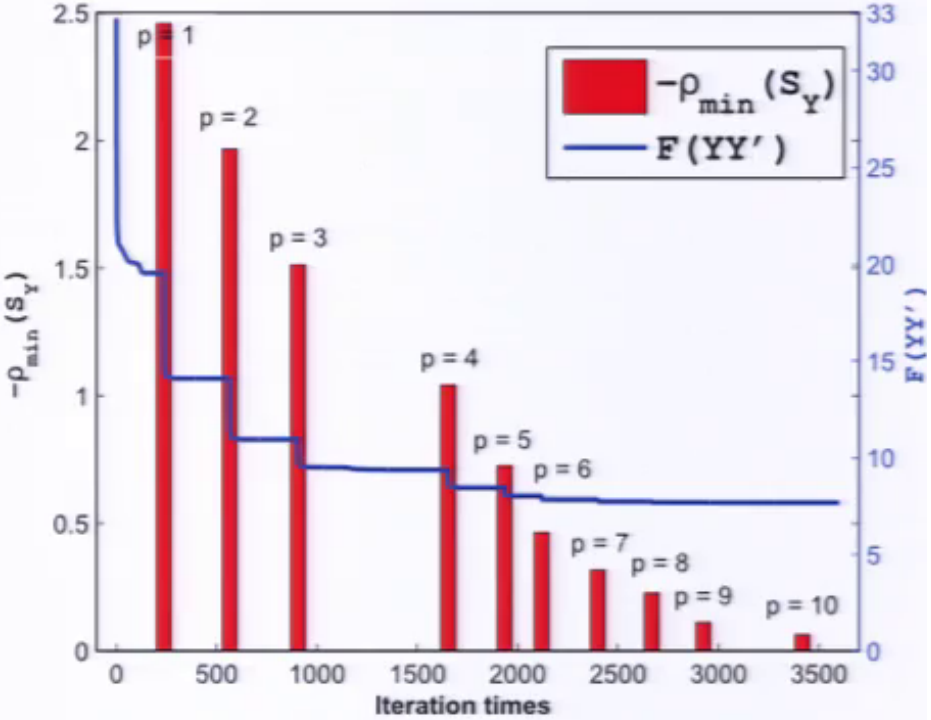
- Shifting S_Y for power method
  - Let C be a constant
  - Let $D = C \cdot I - S_Y$
  - Apply power method to D, get the dominant eigen value-vector pair $\{\rho_d, v_d\}$
  - Let $\{\rho_{min} = C - \rho_d, v = v_d\}$

Running Example

$\rho_{min} \geq 0$ implies $S_Y \succeq 0$

- An running example
  - Data set:

    35 users, 43 movies with 1000 ratings
  - With $\gamma = 25.0$, accuracy level is 10^{-5}
  - p denote the number of columns the current Y used.



Discussion on the Parameter

- Convex Formulation for Error-Free Matrix Completion

$$
\begin{aligned}
\min \quad & tr(X) \\
s.t. \quad & \widehat{R}_{ij} = R_{ij}, (i,j) \in \Omega, \\
& X = \begin{bmatrix} W_1 & \widehat{R} \\ \widehat{R}' & W_2 \end{bmatrix} \succeq 0.
\end{aligned} \qquad (III)
$$

- Let γ_b be the optimum objective of Problem (III), and denote γ as $\gamma = \eta \cdot \gamma_b$

- We consider three situations, namely $\eta = 1, \eta < 1$, and $\eta > 1$

Discussion on the Parameter

- When $\eta = 1$:

> THEOREM 4.1. *Problem (I) with $\gamma = \gamma_b$ is equivalent to Problem (III). It means the solution of Problem (I) with $\eta = 1$ is also the solution of Problem (III), and vice versa.*

$$\begin{aligned} \min \quad & F(X) = \sum_{(i,j)\in\Omega} (\widehat{R}_{ij} - R_{ij})^2 \\ s.t. \quad & X = \begin{bmatrix} W_1 & \widehat{R} \\ \widehat{R}' & W_2 \end{bmatrix} \succeq 0 \\ & tr(X) \le \gamma, \end{aligned}$$

(I) ⟺ When $\gamma = \gamma_b$

$$\begin{aligned} \min \quad & tr(X) \\ s.t. \quad & \widehat{R}_{ij} = R_{ij}, (i,j) \in \Omega, \\ & X = \begin{bmatrix} W_1 & \widehat{R} \\ \widehat{R}' & W_2 \end{bmatrix} \succeq 0. \end{aligned} \qquad (III)$$

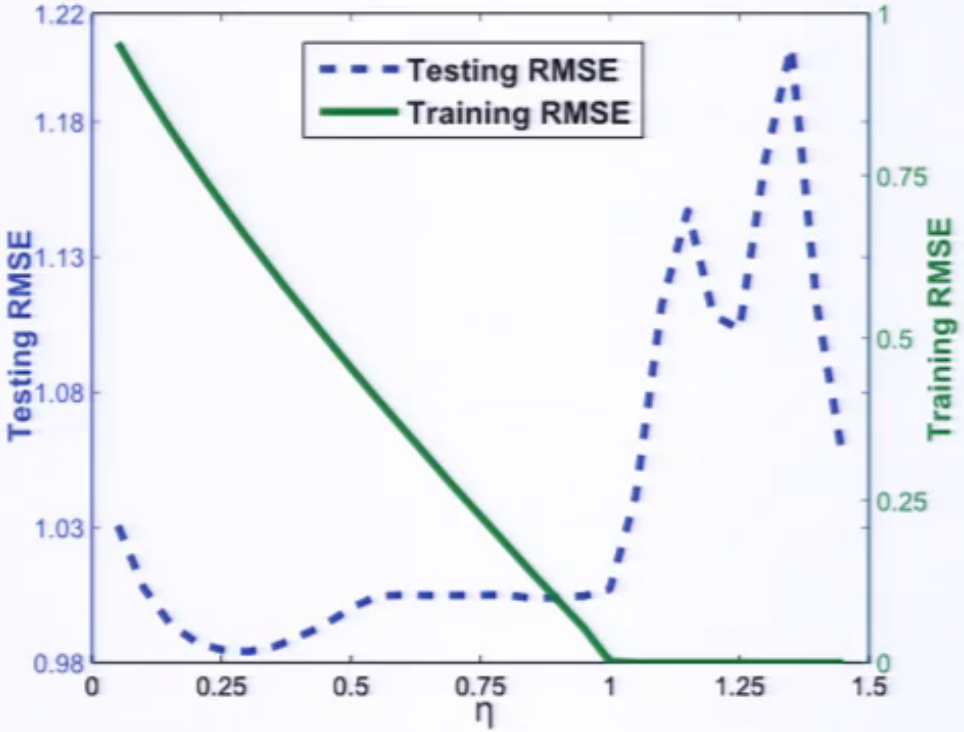
Discussion on the Parameter

- When $\eta = 1$:

THEOREM 4.1. *Problem (I) with $\gamma = \gamma_b$ is equivalent to Problem (III). It means the solution of Problem (I) with $\eta = 1$ is also the solution of Problem (III), and vice versa.*

- When $\eta < 1$:

- When $\eta > 1$:

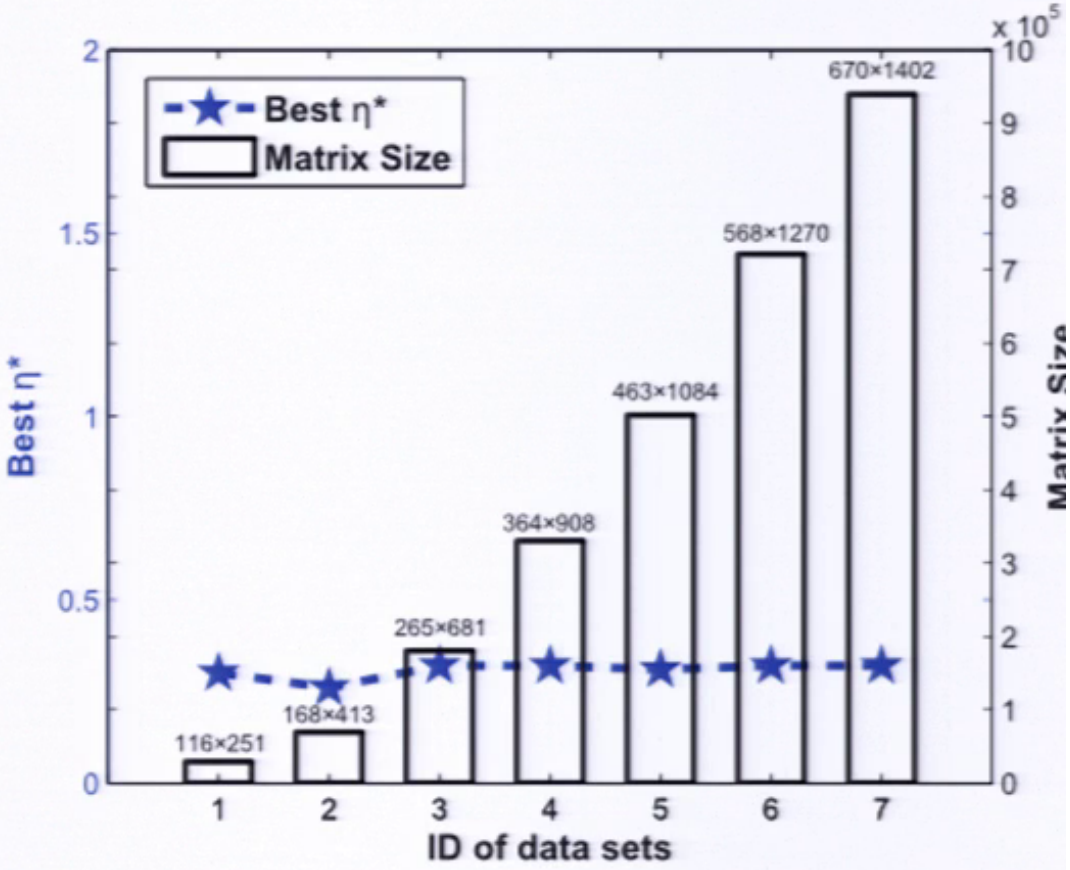


Discussion on the Parameter

- The Stable η for High Performance (Empirical Result)
  - Sample data sets: $R_i (i = 1, ..., 7)$
  - Compute $\gamma_b(R_i)$
  - Tune best η for each data set
  - Empirically found that

$$\eta_i \approx \eta_j (i \neq j)$$

  - It offers us an opportunity for tuning hyper-parameter efficiently.



Experiment on Stable Parameter

- Baseline (TReg): $\min \sum_{(i,j)\in\Omega} (\widehat{R}_{ij} - R_{i,j})^2 + \lambda||\widehat{R}||_*. \qquad (IV)$

$$\min \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} (\widehat{R}_{ij} - R_{i,j})^2 + \mu||\widehat{R}||_*. \qquad (V)\ (\mu = \lambda/|\Omega|)$$

- Data sets:

Table 1: Basic statistics of data sets.

	id	m	n	$\|\Omega\|$	$\|\Omega\|/(m \times n)$
MovieLens	M_1	116	251	7,699	0.2644
	M_2	168	413	14,485	0.2088
	M_3	265	681	28,496	0.1579
	M_4	364	908	43,389	0.1313
	M_5	463	1,084	58,550	0.1167
	M_6	568	1,270	75,813	0.1051
	M_7	670	1,402	91,727	0.0977
Jester	J_1	223	120	16,984	0.6347
	J_2	438	140	33,798	0.5512
	J_3	867	140	66,693	0.5495
	J_4	1,283	140	98,205	0.5467
	J_5	1,566	140	116,735	0.5325
	J_6	1,804	140	133,390	0.5282

Experiment on Stable Parameter

- Tuning the best parameter

Table 2: Tuning the best parameter on *MovieLens*

	TBall		TReg		
data set	η^*	RMSE	λ^*	$\lambda^*/\|\Omega\|$	RMSE
M_1	0.30	0.9641	2.9	4.6782E-04	0.9601
M_2	0.26	0.9267	3.8	3.2624E-04	0.9229
M_3	0.32	0.9293	4.2	1.8325E-04	0.9255
M_4	0.32	0.9289	4.6	1.3200E-04	0.9209
M_5	0.31	0.9102	4.8	1.0235E-04	0.9000
M_6	0.32	0.9089	5.0	8.2319E-05	0.9008
M_7	0.32	0.9120	5.4	7.3394E-05	0.9060
ave	0.31	0.9257	4.4	1.9534E-04	0.9194
std	0.0205	0.0182	0.7754	1.3723E-04	0.0193
std/ave	**0.0668**	0.0197	0.1768	0.7025	0.0210

Table 3: Tuning the best parameter on *Jester*

	TBall		TReg		
data set	η^*	RMSE	λ^*	$\lambda^*/\|\Omega\|$	RMSE
J_1	0.25	4.3273	4.5	3.3040E-04	4.2939
J_2	0.24	4.4100	5.5	2.0405E-04	4.3843
J_3	0.26	4.3654	6.8	1.2687E-04	4.3413
J_4	0.24	4.3408	7.4	9.3937E-05	4.3120
J_5	0.26	4.3168	7.7	8.2548E-05	4.2921
J_6	0.26	4.3305	8.6	8.0585E-05	4.3110
ave	0.25	4.3485	6.8	1.5306E-04	4.3225
std	0.0090	0.0314	1.3769	8.9865E-05	0.0320
std/ave	**0.0357**	0.0072	0.2040	0.5871	0.0074

The best hyper-parameters of TBall are much more stable than that of TReg.

Experiment on Stable Parameter

- Apply the best tuned parameter on the smallest data set

Table 4: RMSE of different sized datasets on *MovieLens*

	TBall	TReg	
data set	$\eta = 0.30$	$\lambda = 2.9$	$\lambda/\|\Omega\| = 4.6782E-04$
M_2	**0.9303±0.0053**	0.9499 ± 0.0039	0.9361 ± 0.0047
M_3	**0.9279±0.0025**	0.9591 ± 0.0038	0.9778 ± 0.0028
M_4	**0.9222±0.0035**	0.9632 ± 0.0043	1.0141 ± 0.0029
M_5	**0.9144±0.0028**	0.9605 ± 0.0050	1.0401 ± 0.0025
M_6	**0.9118±0.0033**	0.9658 ± 0.0038	1.0635 ± 0.0029
M_7	**0.9095±0.0034**	0.9679 ± 0.0052	1.0815 ± 0.0028
avg	**0.9193±0.0074**	0.9611 ± 0.0058	1.0189 ± 0.0499

Table 5: RMSE of different sized datasets on *Jester*

	TBall	TReg	
data set	$\eta = 0.25$	$\lambda = 4.5$	$\lambda/\|\Omega\| = 3.3040E-04$
J_2	**4.3939±0.0171**	4.4137 ± 0.0206	4.4917 ± 0.0180
J_3	**4.3515±0.0070**	4.5022 ± 0.0122	4.6101 ± 0.0099
J_4	**4.3434±0.0091**	4.5803 ± 0.0190	4.7092 ± 0.0033
J_5	**4.3285±0.0180**	4.5929 ± 0.0137	4.7484 ± 0.0138
J_6	**4.3124±0.0136**	4.6082 ± 0.0173	4.7979 ± 0.0139
avg	**4.3459±0.0274**	4.5394 ± 0.0727	4.6715 ± 0.1090

When inherit hyper-parameters tuned on the smallest dataset, the TBall method performs much better than TReg.

Conclusion

- Conclusion
  - We proposed a novel first-order low-rank approach for solving the convex MC problem in the form of trace norm bounding.
  - We gave both the theoretical and empirical analysis on the correctness of the proposed approach.
  - We discussed how the model parameter controls how well the model fits the training data.
  - Extensive experiments validated the parameter robustness properties of our method.