Implicit Regularization in Matrix Factorization

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli,
Behnam Neyshabur, Nathan Srebro (May 2017)

https://arxiv.org/abs/1705.09280

UBC MLRG: 6 – Nov – 2019

Betty Shea

MLRG Theme: Good Solutions

- In the overparameterized setting, why do we tend to end up with the good solutions?
    - Sharp/ flat minima
    - Implicit regularization
        - Last week: geometry of optimization
        - This week: similar theme

Implicit Regularization

- Our choice of optimization algorithm biases us towards certain types of solution (without explicit regularization)
    - Coordinate descent: L1 norm
    - Gradient descent: L2 norm
    - Matrix factorization: this presentation
    - Logistic regression: stay tuned!

- But we don’t always get to exactly this norm in practice

Matrix Factorization

- Can be modeled by a 2-layer neural network with linear transfer. Gradient descent on the entries of the factor matrices (on the weights of the network).

- Example: matrix completion/ collaborative filtering problems

- PCA is also called a matrix factorization model

- Chi et al. Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview. https://www.princeton.edu/~yc5/publications/NcxOverview_Arxiv.pdf

Main conjecture (informal)

- Gradient descent on a full dimensional matrix factorization converges to the minimum nuclear norm solution with
    - Small enough step sizes
    - Initialization close enough to the origin

- Implicit regularization occurs with full gradient descent

- Nuclear norm of X = sum of the singular values of X

Problem Setting

- Least squares objective over symmetric positive semidefinite matrix X

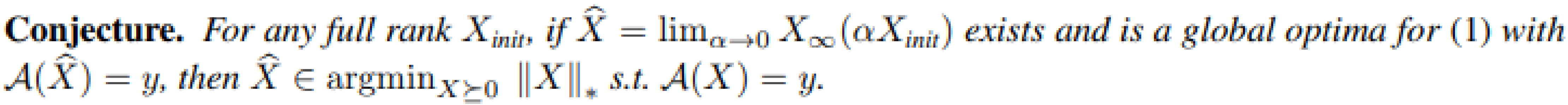
$$\min_{X \succeq 0} F(X) = \|A(X) - y\|_2^2$$

$$X \in \mathbb{R}^{n \times n} \quad y \in \mathbb{R}^m \quad \text{linear } A : \mathbb{R}^{n \times n} \to \mathbb{R}^m \quad A(X)_i = \langle A_i, X \rangle, A_i \in \mathbb{R}^{n \times n}$$

- Minimize on factorization of X instead with d=n

$$\min_{U \in \mathbb{R}^{n \times d}} F(U) = \|A(UU^\intercal) - y\|_2^2$$

Main Results: Conjecture



Conjecture. *For any full rank* X_{init}, *if* $\widehat{X} = \lim_{\alpha\to 0} X_\infty(\alpha X_{init})$ *exists and is a global optima for* (1) *with* $\mathcal{A}(\widehat{X}) = y$, *then* $\widehat{X} \in \operatorname{argmin}_{X\succeq 0} \|X\|_*$ *s.t.* $\mathcal{A}(X) = y$.

where (1) refers to the least square objective over X in a previous slide,

where $\lim\limits_{\alpha\to 0}$ ensures that the initial point gets close to the origin,

where $X_\infty(\alpha X_{init})$ is the limit point and is defined as $lim_{t\to\infty}X_t$
for the factorized gradient flow $\frac{dX_t}{dt} = -A^*(r_t)X_t - X_tA^*(r_t)$ initialized at
$X_0 = X_{init}$, residuals at time t $r_t = A(X_t) - y$, A* is the adjoint of A $A^* : \mathbb{R}^m \to \mathbb{R}^{n\times n}$

Some Details

- Factorized objective is non-convex optimization problem

- Analysis uses gradient flow instead of gradient descent

- Factorized gradient flow equation: $\frac{dX_t}{dt} = -A^*(r_t)X_t - X_t A^*(r_t)$

- Regular gradient flow equation: $\frac{dX_t}{dt} = -A^*(r_t)$

Main Results: Supporting Evidence

- Simulation of matrix reconstruction problem

- Theoretical analysis

- Empirical results of matrix completion problem

Simulation of Matrix Reconstruction

- Pick $X^* \succeq 0$, generate $m \ll n^2$ random measurement matrices and set $y = A(X^*)$
  - Chose three different 50 x 50 X*
- Run gradient descent on U until convergence
  - Different step sizes, initialization
- Measure reconstruction error $\|X - X^*\|_F$
  - Across different values of d

Simulation Results: Matrix Reconstruction

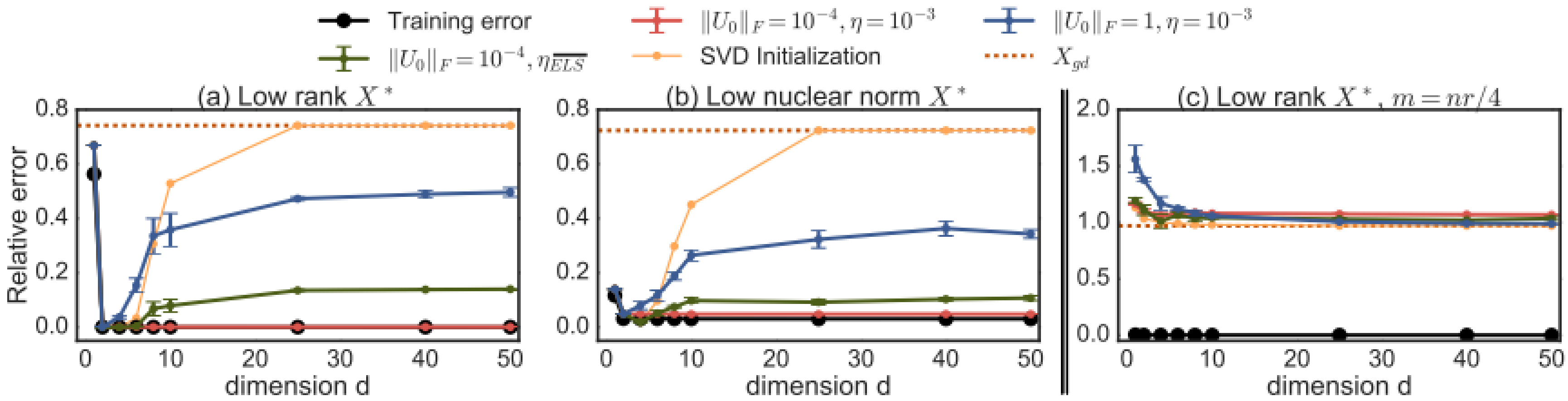


Figure 1: Reconstruction error of the solutions for the planted 50×50 matrix reconstruction problem. In (a) X^* is of rank $r = 2$ and $m = 3nr$, in (b) X^* has a spectrum decaying as $O(1/k^{1.5})$ normalized to have $\|X^*\|_* = \sqrt{r}\|X^*\|_F$ for $r = 2$ and $m = 3nr$, and in (c) we look at a non-reconstructable setting where the number of measurements $m = nr/4$ is much smaller than the requirement to reconstruct a rank $r = 2$ matrix. The plots compare the reconstruction error of gradient descent on U for different choices initialization U_0 and step size η, including fixed step-size and exact line search clipped for stability ($\eta_{\overline{ELS}}$). Additonaly, the orange dashed reference line represents the performance of X_{gd} — a rank unconstrained global optima obtained by projected gradient descent on X space for (1), and 'SVD-Initialization' is an example of an alternate rank d global optima, where initialization U_0 is picked based on SVD of X_{gd} and gradient descent with small stepsize is run on factor space. The results are averaged across 3 random initialization and (nearly zero) errorbars indicate the standard deviation.

Simulated Results: Matrix Reconstruction

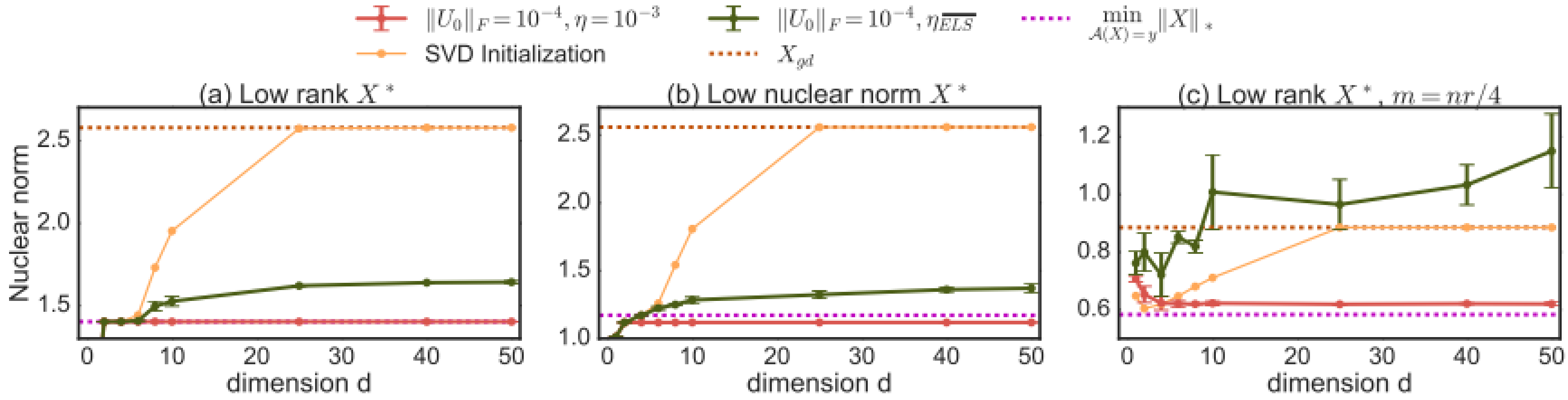


Figure 2: Nuclear norm of the solutions from Figure 1. In addition to the reference of X_{gd} from Figure 1, the magenta dashed line (almost overlapped by the plot of $\|U\|_F = 10^{-4}, \eta = 10^{-3}$) is added as a reference for the (rank unconstrained) minimum nuclear norm global optima. The error bars indicate the standard deviation across 3 random initializations. We have dropped the plot for $\|U\|_F = 1, \eta = 10^{-3}$ to reduce clutter.

Summary of Matrix Reconstruction Results

- Minimizing to the factorized objective performs better generally
  - Except in the case where reconstruction is impossible

- The solution with the lowest reconstruction error comes from initializing close to the origin and taking very small fixed step sizes

- There is some evidence that gradient descent on the factorized objective biases us towards minimum nuclear norm solution

Theoretical Results

- Two warmup cases
  - gradient descent on original convex objective: zero-error solution minimizes the Frobenius norm
  - proved informally that their main conjecture holds for m=1
- Case where A_i commute
  - Formal proof of their main conjecture
- Case where A_i do not commute
  - Solution is no longer simple and is a "time ordered exponential"

Theoretical Analysis: A_i commute

- Goal: show that a minimum nuclear norm solution with zero training error satisfies Karush-Kuhn-Tucker (KKT) conditions where $\exists v \in \mathbb{R}^m$ such that $A(X) = y$ $\quad X \succeq 0$ $\quad A^*(v) \preceq I$ $\quad (I - A^*(v))X = 0$
- The solution to $\frac{dX_t}{dt} = -A^*(r_t)X_t - X_tA^*(r_t)$ is

$$X_t = \exp(A^*(s_t))X_0 \exp(A^*(s_t)) \quad \text{where} \quad s_T = -\int_0^T r_t dt$$

- A_i commute mean that they have at least one shared eigenbasis $v_1, \ldots, v_n$.

Theoretical Analysis: A_i commute

- For any α, $X_\infty(\alpha I)$ and $\hat{X}$ are both diagonalizable by $v_1, \ldots, v_n$.
  This means that $\lambda_k(X_\infty(\alpha I))$ converges to $\lambda_k(\hat{X})$.

- Introduce $\beta = -\log \alpha$ and $\nu(\beta) = s_\infty(\beta)/\beta$. Show that $\lambda_k(A^*(\nu(\beta))) < 1$ as β goes to infinity

- Enough to satisfy $\lim_{\beta \to \infty} A^*(\nu(\beta)) \preceq I$ $\lim_{\beta \to \infty} A^*(\nu(\beta))\hat{X} = \hat{X}$

Theoretical Analysis: A_i commute

- Proof does not rely on any particular form of the residuals r_t

- Proof relies on showing that gradient flow stays within the manifold

$$\mathcal{M} = \{X = \exp\left(\mathcal{A}^*(s)\right) X_{\text{init}} \exp\left(\mathcal{A}^*(s)\right) \mid s \in \mathbb{R}^m\}$$

Theoretical Analysis: A_i do not commute

- The solution to the gradient flow equation is

$$X_t = \lim_{\epsilon \to 0} \left(\prod_{\tau = t/\epsilon}^{1} \exp\left(-\epsilon \mathcal{A}^*(r_{\tau\epsilon})\right) \right) X_0 \left(\prod_{\tau = 1}^{t/\epsilon} \exp\left(-\epsilon \mathcal{A}^*(r_{\tau\epsilon})\right) \right)$$

- A “time-ordered exponential” where the order of multiplication matters.

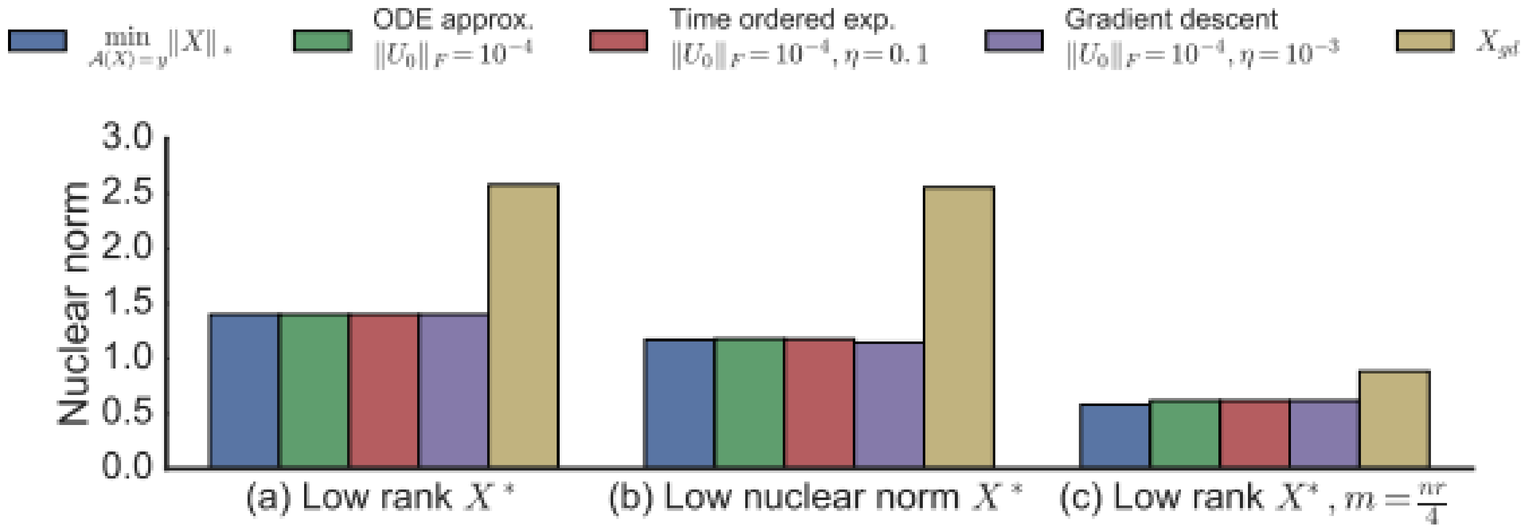
- Hard to characterize a manifold that the solution stays within without additional restrictions on the form of the residuals

Empirical Results: Matrix Completion

- Simulated data sampled from different probability distribution
  - Gaussian, uniform and power-law
  - Again, the same 3 planted X*
  - Compared the nuclear norm of solutions from different methods

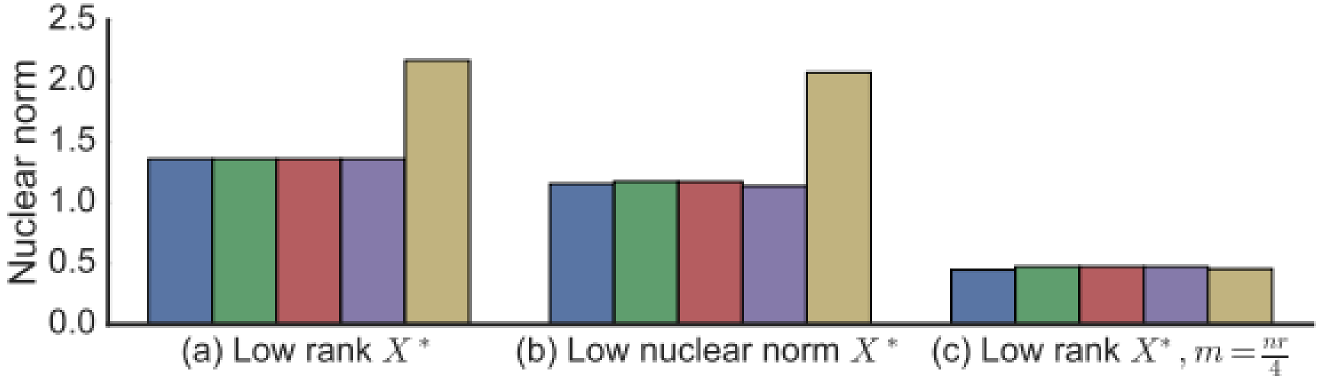
- Used data from Movielens
  - About 100,000 ratings from about 950 users and 1700 movies.

Simulated Results: Matrix Completion

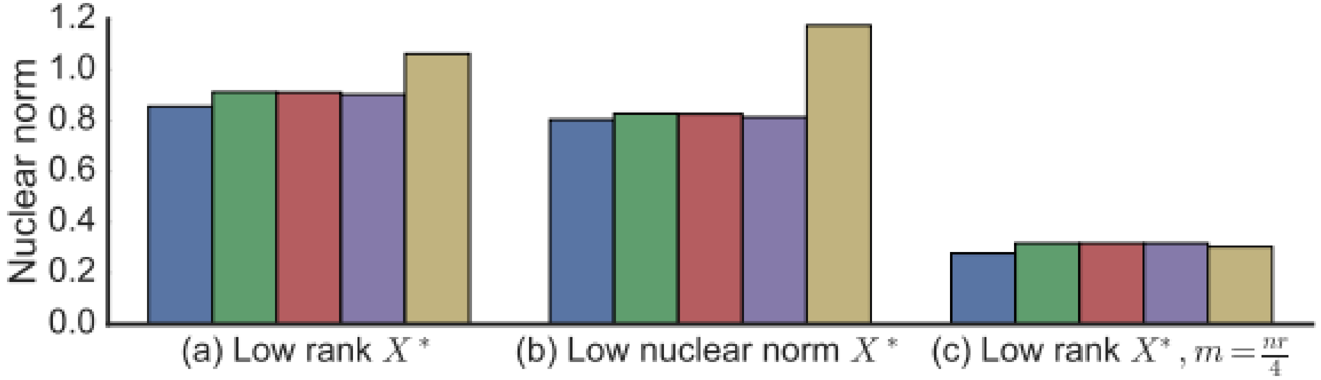


(i) Gaussian random measurements. We report the nuclear norm of the gradient flow solutions from three different approximations to (3) – numerical ODE solver (*ODE approx.*), time ordered exponential specified in (12) (*Time ordered exp.*) and standard gradient descent with small step size (*Gradient descent*). The nuclear norm of the solution from gradient descent on X space – X_{gd} and the minimum nuclear norm global minima are provided as references. In (a) X^* is rank r and $m = 3nr$, in (b) X^* has a decaying spectrum with $\|X^*\|_* = \sqrt{r}\|X^*\|_F$ and $m = 3nr$, and in (c) X^* is rank r with $m = nr/4$, where $n = 50$, $r = 2$.

Simulated Results: Matrix Completion

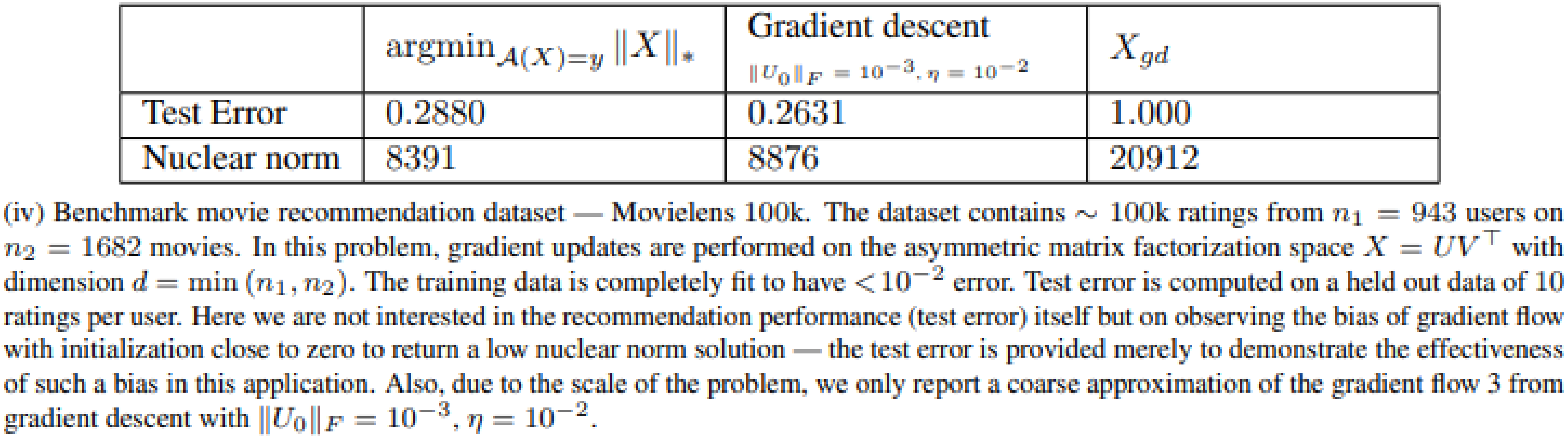


(ii) Uniform matrix completion: $\forall i$, A_i measures a uniform random entry of X^*. Details on X^*, number of measurements, and the legends follow Figure3-(i).



(iii) Power law matrix completion: $\forall i$, A_i measures a random entry of X^* chosen according to a power law distribution. Details on X^*, number of measurements, and the legends follow Figure3-(i).

Non-Simulated Results: Matrix Completion



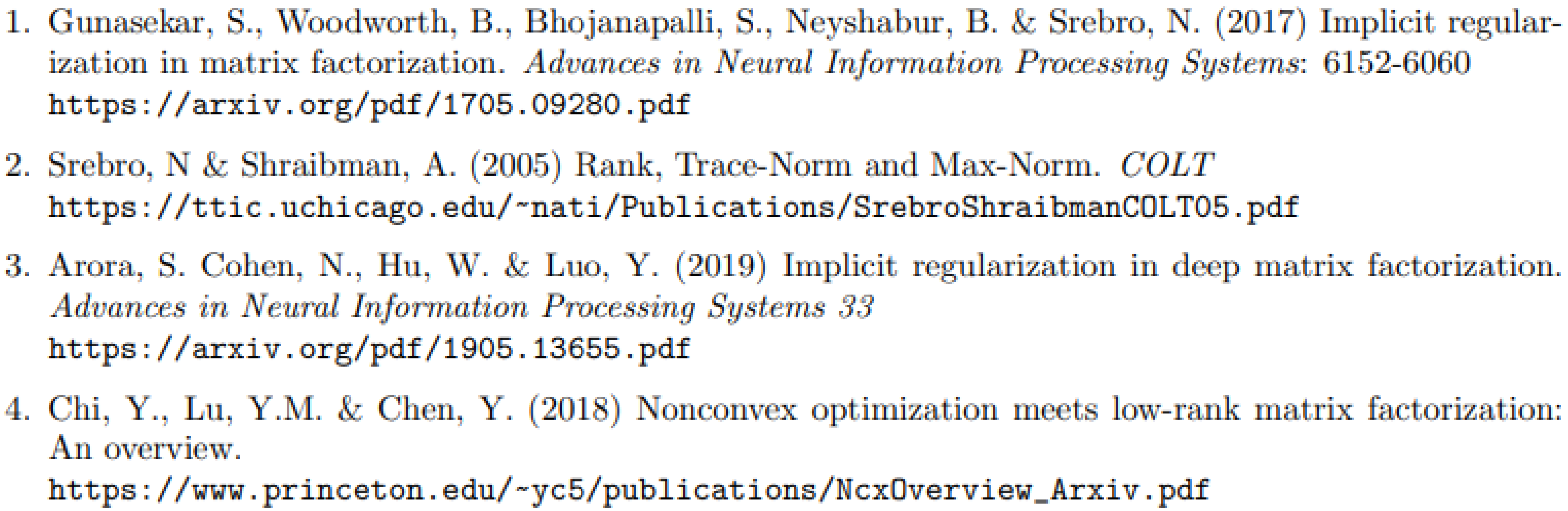
	$\text{argmin}_{\mathcal{A}(X)=y} \|X\|_*$	Gradient descent $\|U_0\|_F = 10^{-3}, \eta = 10^{-2}$	X_{gd}
Test Error	0.2880	0.2631	1.000
Nuclear norm	8391	8876	20912

(iv) Benchmark movie recommendation dataset — Movielens 100k. The dataset contains $\sim$ 100k ratings from $n_1 = 943$ users on $n_2 = 1682$ movies. In this problem, gradient updates are performed on the asymmetric matrix factorization space $X = UV^\top$ with dimension $d = \min(n_1, n_2)$. The training data is completely fit to have $< 10^{-2}$ error. Test error is computed on a held out data of 10 ratings per user. Here we are not interested in the recommendation performance (test error) itself but on observing the bias of gradient flow with initialization close to zero to return a low nuclear norm solution — the test error is provided merely to demonstrate the effectiveness of such a bias in this application. Also, due to the scale of the problem, we only report a coarse approximation of the gradient flow 3 from gradient descent with $\|U_0\|_F = 10^{-3}, \eta = 10^{-2}$.

Summary

- Paper argues that there is implicit regularization in gradient descent over matrix factorization
- Implicit regularization/ bias is towards a minimum nuclear norm solution
- Hard to prove theoretically for the general case of matrix factorization
- Newer paper “Implicit Regularization in Deep Matrix Factorization”. (Arora et al). https://arxiv.org/pdf/1905.13655.pdf

References



1. Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B. & Srebro, N. (2017) Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*: 6152-6060 https://arxiv.org/pdf/1705.09280.pdf
2. Srebro, N & Shraibman, A. (2005) Rank, Trace-Norm and Max-Norm. *COLT* https://ttic.uchicago.edu/~nati/Publications/SrebroShraibmanCOLT05.pdf
3. Arora, S. Cohen, N., Hu, W. & Luo, Y. (2019) Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems 33* https://arxiv.org/pdf/1905.13655.pdf
4. Chi, Y., Lu, Y.M. & Chen, Y. (2018) Nonconvex optimization meets low-rank matrix factorization: An overview. https://www.princeton.edu/~yc5/publications/NcxOverview_Arxiv.pdf