

Accelerating Iterative Hard Thresholding For Low-rank Matrix Completion Via Adaptive Restart

Trung Vu and Raviv Raich

School of EECS, Oregon State University, Corvallis, OR 97331-5501, USA








{vutru,raich}@oregonstate.edu

May 16, 2019

- 1 Problem Formulation
- 2 Background
- 3 Main Results
- 4 Conclusions and Future Work

The Netflix Prize Problem

Movies

				
Users		4	?	?
		?	?	4
		?	2	?
		4	?	4

Known: $S = \{(i, j) \mid M_{ij} \text{ is observed}\}$

Unknown: $S^c = \{(i, j) \mid M_{ij} = ?\}$

A partially known rating matrix $M \in \mathbb{R}^{m \times n}$ with $\text{rank}(M) \leq r$

Low-Rank Matrix Completion Problem

$$\begin{array}{c} \underbrace{\hspace{10em}}_M \\ \begin{bmatrix} 4 & ? & ? \\ ? & ? & 4 \\ ? & 2 & ? \\ 4 & ? & 4 \end{bmatrix} \end{array} \xrightarrow{\text{Given } r=1} \begin{array}{c} \underbrace{\hspace{10em}}_{X^*} \\ \begin{bmatrix} 4 & 2 & 4 \\ 4 & 2 & 4 \\ 4 & 2 & 4 \\ 4 & 2 & 4 \end{bmatrix} \end{array} \stackrel{\text{SVD}}{=} \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} \cdot 6 \cdot \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

find $X_{ij}, \quad (i, j) \in \mathcal{S}^c$
subject to $\text{rank}(X) \leq r$ and $X_{ij} = M_{ij}$ for $(i, j) \in \mathcal{S}$.
 $(r < n \leq m)$

Notations

- Sampling operator X_S

$$[X_S]_{ij} = \begin{cases} X_{ij} & \text{if } (i,j) \in \mathcal{S} \\ 0 & \text{if } (i,j) \in \mathcal{S}^c \end{cases}$$

$$\begin{bmatrix} 4 & 2 & 4 \\ 4 & 2 & 4 \\ 4 & 2 & 4 \\ 4 & 2 & 4 \end{bmatrix} \xrightarrow{S} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 0 & 4 \\ 0 & 2 & 0 \\ 4 & 0 & 4 \end{bmatrix}$$

- Row selection matrix $S_{(\mathcal{S})} \in \mathbb{R}^{s \times mn}$ corresponding to \mathcal{S}

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}}_{S_{(\mathcal{S})}} \begin{bmatrix} 4 \\ 2 \\ 4 \\ 4 \\ 2 \\ 4 \\ 4 \\ 2 \\ 4 \\ 4 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 2 \\ 4 \\ 4 \end{bmatrix}$$

- The **rank- r projection** of an *arbitrary* matrix $X \in \mathbb{R}^{m \times n}$ is obtained by hard-thresholding singular values of X

$$\mathcal{P}_r(X) = \sum_{i=1}^r \sigma_i(X) u_i(X) v_i(X)^T$$

- The **SVD** of the matrix M can be *partitioned* based on the signal subspace and its orthogonal subspace

$$M = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} \quad \Sigma_1 \in \mathbb{R}^{r \times r}$$

Several Formulations of Low-Rank Matrix Completion

$$\text{find } X_{ij}, \quad (i, j) \in \mathcal{S}^c \quad \text{s.t.} \quad \text{rank}(X) \leq r \text{ and } X_{\mathcal{S}} = M_{\mathcal{S}}$$

Approach	Problem formulation	Property
Convex relaxation	$\min \ X\ _* \quad \text{s.t. } X_{\mathcal{S}} = M_{\mathcal{S}}$	✓ Rigorous guarantees ✗ Slow convergence
	$\min \lambda \ X\ _* + \frac{1}{2} \ X_{\mathcal{S}} - M_{\mathcal{S}}\ _F^2$	
	$\min \tau \ X\ _* + \frac{1}{2} \ X\ _F^2 \quad \text{s.t. } X_{\mathcal{S}} = M_{\mathcal{S}}$	
Non-convex	$\min \text{rank}(X) \quad \text{s.t. } X_{\mathcal{S}} = M_{\mathcal{S}}$	✓ Fast convergence ✗ Hard to analyze
	$\min \ X_{\mathcal{S}} - M_{\mathcal{S}}\ _F^2 \quad \text{s.t. } \text{rank}(X) \leq r \quad (*)$	
	$\min \ [XY^T]_{\mathcal{S}} - M_{\mathcal{S}}\ _F^2 \quad X \in \mathbb{R}^{m \times r}, Y \in \mathbb{R}^{n \times r}$	

$$\|X\|_* = \sum_{i=1}^n \sigma_i(X)$$

Outline

- 1 Problem Formulation
- 2 Background
- 3 Main Results
- 4 Conclusions and Future Work

Iterative Hard Thresholding for Matrix Completion

$$\min_{X \in \mathbb{R}^{m \times n}} \frac{1}{2} \|X_S - M_S\|_F^2 \quad \text{s.t. } \text{rank}(X) \leq r \quad (*)$$

- Iterative hard thresholding (IHT) is a variant of **non-convex** projected gradient descent

$$X^{(k+1)} = \mathcal{P}_r(X^{(k)} - \alpha_k[X^{(k)} - M]_S)$$

- Unlike *matrix sensing*, the matrix RIP does not hold for MCP

$$0 \cdot \|X\|_F^2 \leq \|[X]_S\|_F^2 \leq 1 \cdot \|X\|_F^2$$

- ▶ Global convergence is *non-trivial!* [Jain, Meka, and Dhillon 2010]

Local Convergence of IHT

Algorithm 1 IHTSVD

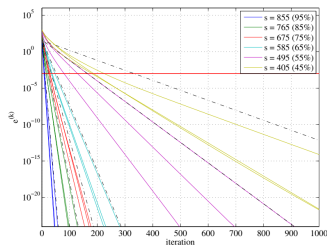
- 1: **for** $k = 0, 1, 2, \dots$ **do**
 - 2: $X^{(k+1)} = \mathcal{P}_r(Y^{(k)})$
 - 3: $Y^{(k+1)} = \mathcal{P}_{M,S}(X^{(k+1)})$
-

$$*\mathcal{P}_{M,S}(X) = X_{S^c} + M_S$$

► IHT with unit step size $\underline{\alpha_k = 1}$

[*ibid.*] If $\sigma = \sigma_{\min}(S_{(S^c)}(V_2 \otimes U_2)) > 0$, then IHTSVD converges to M locally at a linear rate $1 - \sigma^2$.

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 0 & 4 \\ 0 & 2 & 0 \\ 4 & 0 & 4 \end{bmatrix} \xrightarrow{\mathcal{P}_r} \begin{bmatrix} 2 & 0 & 2 \\ 2 & 0 & 2 \\ 0 & 0 & 0 \\ 4 & 0 & 4 \end{bmatrix} \xrightarrow{\mathcal{P}_{M,S}} \begin{bmatrix} 4 & 0 & 2 \\ 2 & 0 & 4 \\ 0 & 2 & 0 \\ 4 & 0 & 4 \end{bmatrix} \xrightarrow{\mathcal{P}_r} \dots$$



Source: [Chunikhina, Raich, and Nguyen 2014]

Linearization of the Rank- r Projection

$$\mathcal{P}_r(M + \Delta) = M + \Delta - U_2 U_2^T \Delta V_2 V_2^T + O(\|\Delta\|_F^2)$$

- Local convergence analysis assumes $Y^{(k)}$ is a perturbed matrix of M

$$M + E^{(k+1)} = Y^{(k+1)} = \mathcal{P}_{M,S}(\mathcal{P}_r(Y^{(k)})) = \mathcal{P}_{M,S}(\mathcal{P}_r(M + E^{(k)}))$$

- The recursion on the error matrix $E^{(k+1)} = [\mathcal{P}_r(M + E^{(k)}) - M]_{S^c}$ can be approximated by

$$\underbrace{S_{(S^c)} \text{vec}(E^{(k+1)})}_{e^{(k+1)}} \stackrel{1}{=} \underbrace{(I_s - S_{(S^c)}(V_2 \otimes U_2)(V_2 \otimes U_2)^T S_{(S^c)}^T)}_A \underbrace{S_{(S^c)} \text{vec}(E^{(k)})}_{e^{(k)}}$$

- Stable if $\lambda_{\max}(A) = 1 - \left(\sigma_{\min}(S_{(S^c)}(V_2 \otimes U_2))\right)^2 < 1$

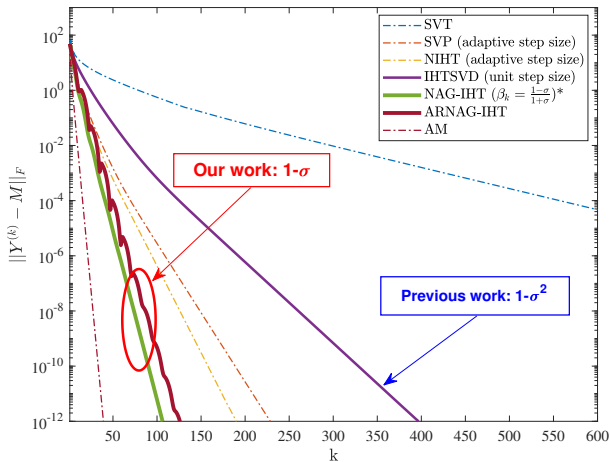


Figure 1: The distance to the solution (in log-scale) as a function of the iteration number for various algorithms. $m = 50$, $n = 40$, $r = 3$, and $s = 1000$. All algorithms share the same computational complexity per iteration ($O(mnr)$) except SVT ($O(mn^2)$) [Cai, Candès, and Shen 2010] and AM ($O(sm^2r^2 + m^3r^3)$) [Jain, Netrapalli, and Sanghavi 2013].

Outline

- 1 Problem Formulation
- 2 Background
- 3 Main Results**
- 4 Conclusions and Future Work

Our Contribution

- 1 Analyze the local convergence of accelerated IHTSVD for solving the rank constrained least squares problem (*).
- 2 Propose a practical way to select momentum step size that enables us to recovers the optimal rate of convergence near the solution.

Nesterov's Accelerated Gradient

- Nesterov's Accelerated Gradient (NAG) is a simple modification to gradient descent that **provably** accelerates the convergence

$$\begin{aligned}x^{(k+1)} &= y^{(k)} - \alpha_k \nabla f(y^{(k)}) \\y^{(k+1)} &= x^{(k+1)} + \beta_k (x^{(k+1)} - x^{(k)})\end{aligned}$$

- If f is μ -strongly convex, L -smooth function, NAG can improve the **linear convergence rate** from $1 - \mu/L$ to $1 - \sqrt{\mu/L}$ by setting

$$\alpha_k = \frac{1}{L}, \quad \beta_k = \frac{1 - \sqrt{\mu/L}}{1 + \sqrt{\mu/L}}. \quad [Nesterov 2004]$$

- Iteration complexity: $O(\sqrt{\kappa})$, compared to $O(\kappa)$ for gradient descent, where $\kappa = \frac{L}{\mu}$ is the condition number.

The Proposed NAG-IHT

Algorithm 2 NAG-IHT

- 1: **for** $k = 0, 1, 2, \dots$ **do**
 - 2: $X^{(k+1)} = \mathcal{P}_r(Y^{(k)})$
 - 3: $Y^{(k+1)} = \mathcal{P}_{M,S}(X^{(k+1)} + \beta_k(X^{(k+1)} - X^{(k)}))$
-

Method	# Ops./Iter.	Local conv. rate	#Iters. needed ϵ -acc.
IHTSVD	$O(mnr)$	$1 - \sigma^2$	$\frac{1}{\sigma^2} \log(1/\epsilon)$
NAG-IHT with $\beta_k = \frac{1-\sigma}{1+\sigma}$	$O(mnr)$	$1 - \sigma$	$\frac{1}{\sigma} \log(1/\epsilon)$

$$* \sigma = \sigma_{\min}(S_{(S^c)}(V_2 \otimes U_2))$$

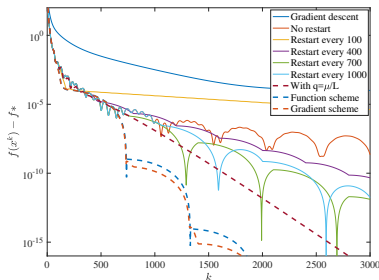
A Practical Method for Step Size Selection

- Practical issue: fast convergence requires **prior knowledge of global parameters** related to the objective function ($\beta_k = \frac{1-\sigma}{1+\sigma}$).
- Solution: **adaptive restart** [O'Donoghue and Candès 2015]

- Use an incremental momentum

$$\beta_k = \frac{t-1}{t+2} \text{ starting at } t = 1$$

- When $f(x^{(k+1)}) > f(x^{(k)})$, reset $t = 1$



The Proposed Adaptive Restart Scheme for NAG-IHT

Algorithm 3 ARNAG-IHT

- 1: $t = 1$
 - 2: $f_0 = \left\| X_S^{(0)} - M_S \right\|_F^2$
 - 3: **for** $k = 0, 1, 2, \dots$ **do**
 - 4: $X^{(k+1)} = \mathcal{P}_r(Y^{(k)})$
 - 5: $Y^{(k+1)} = \mathcal{P}_{M,S}(X^{(k+1)} + \frac{t-1}{t+2}(X^{(k+1)} - X^{(k)}))$
 - 6: $f_{k+1} = \left\| X_S^{(k+1)} - M_S \right\|_F^2$
 - 7: **if** $f_{k+1} > f_k$ **then** $t = 1$ **else** $t = t + 1$ ▷ function scheme
-

Numerical Evaluation

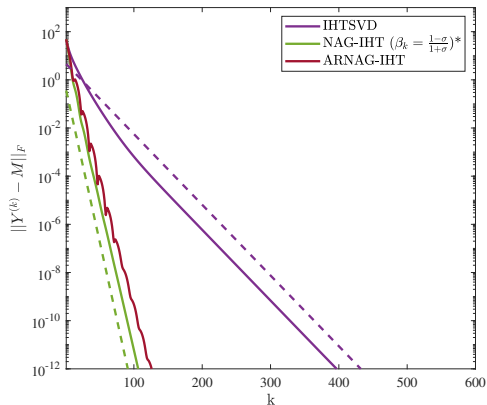


Figure 2: The distance to the solution (in log-scale) as a function of the iteration number for IHT algorithms (solid) and their corresponding theoretical bounds up to a constant (dashed). $m = 50$, $n = 40$, $r = 3$, and $s = 1000$. *NAG-IHT using optimal step size is not applicable in practice.

Outline

- 1 Problem Formulation
- 2 Background
- 3 Main Results
- 4 Conclusions and Future Work**

Conclusions and Future Work

- Conclusions

- The local convergence of IHT for low-rank matrix completion can be characterized through the linearization of the rank projection.
- Convex optimization concepts such as strong convexity can be exploited to analyze convergence property and accommodate acceleration.
- Adaptive restart is an efficient way to accommodate Nesterov's Accelerated Gradient in plain IHT in practice.

- Future work

- Extending the local convergence analysis to the real-world cases when the underlying matrix is noisy and/or not close to being low rank.
- Convergence under a simple initialization suggests potential analysis of global convergence of our algorithm.

References I

- Cai, J.-F., E. Candès, and Z. Shen (2010). “A Singular Value Thresholding Algorithm for Matrix Completion”. In: *SIAM Journal on Optimization* 20.4, pp. 1956–1982.
- Chunikhina, E., R. Raich, and T. Nguyen (2014). “Performance analysis for matrix completion via iterative hard-thresholded SVD”. In: *2014 IEEE Workshop on Statistical Signal Processing (SSP)*, pp. 392–395.
- Jain, P., R. Meka, and I. Dhillon (2010). “Guaranteed Rank Minimization via Singular Value Projection”. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 937–945.
- Jain, P., P. Netrapalli, and S. Sanghavi (2013). “Low-rank Matrix Completion Using Alternating Minimization”. In: *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, pp. 665–674.
- Nesterov, Y. (2004). *Introductory lectures on convex optimization: a basic course*. Kluwer Academic Publishers.
- O’Donoghue, B. and E. Candès (2015). “Adaptive Restart for Accelerated Gradient Schemes”. In: *Foundations of Computational Mathematics* 15.3, pp. 715–732.