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

Outline

The Netflix Prize Problem

Movies

Users

			
	4	?	?
	?	?	4
	?	2	?
	4	?	4

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

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

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

Low-Rank Matrix Completion Problem

$$\overbrace{\begin{bmatrix} \mathbf{4} & ? & ? \\ ? & ? & \mathbf{4} \\ ? & \mathbf{2} & ? \\ \mathbf{4} & ? & \mathbf{4} \end{bmatrix}}^{M} \xrightarrow{\text{Given } r=1} \overbrace{\begin{bmatrix} \mathbf{4} & 2 & 4 \\ 4 & 2 & \mathbf{4} \\ 4 & \mathbf{2} & 4 \\ \mathbf{4} & 2 & \mathbf{4} \end{bmatrix}}^{X^*} \underline{\underline{SVD}} \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} \cdot 6 \cdot \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

$$\begin{array}{rl} \text{find} & X_{ij}, \qquad (i,j) \in \mathcal{S}^c \\ \text{subject to} & \text{rank}(X) \leq r \text{ and } X_{ij} = M_{ij} \text{ for } (i,j) \in \mathcal{S}. \end{array}$$

$$(r < n \leq m)$$

Notations

- **Sampling operator $X_{\mathcal{S}}$**

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

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

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

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

- 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

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

Several Formulations of Low-Rank Matrix Completion

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

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

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

Outline

Iterative Hard Thresholding for Matrix Completion

$$\min_{X \in \mathbb{R}^{m \times n}} \frac{1}{2} \|X_{\mathcal{S}} - M_{\mathcal{S}}\|_F^2 \qquad \text{s.t. } \operatorname{rank}(X) \le r \qquad (*)$$

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

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

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

$$0 \cdot \|X\|_F^2 \le \|[X]_{\mathcal{S}}\|_F^2 \le 1 \cdot \|X\|_F^2$$

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

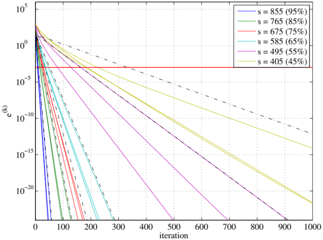
Local Convergence of IHT

Algorithm 1 IHTSVD

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

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

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



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

Source: *[Chunikhina, Raich, and Nguyen 2014]*

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

Linearization of the Rank-r Projection

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

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

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

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

$$\underbrace{\mathsf{S}_{(\mathcal{S}^c)} \operatorname{vec}(E^{(k+1)})}_{e^{(k+1)}} \overset{1}{=\!=} \underbrace{\left(I_s - \mathsf{S}_{(\mathcal{S}^c)}(V_2 \otimes U_2)(V_2 \otimes U_2)^T \mathsf{S}_{(\mathcal{S}^c)}^T\right)}_{A} \underbrace{\mathsf{S}_{(\mathcal{S}^c)} \operatorname{vec}(E^{(k)})}_{e^{(k)}}$$

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

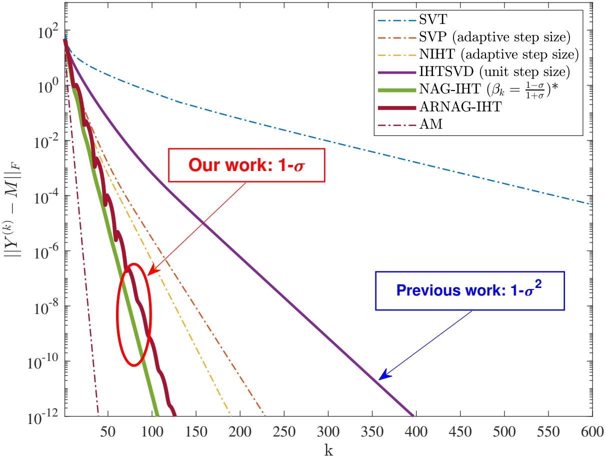


Figure 1: The distance to the solution (in log-scale) as a function of the iteration number for various algorithms. $m = 50$, $n = 40$, $r = 3$, and $s = 1000$. All algorithms share the same computational complexity per iteration $\left(O(mnr)\right)$ except SVT $\left(O(mn^2)\right)$ *[Cai, Candès,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

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

- 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

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

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

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 $\beta_k = \frac{1-\sigma}{1+\sigma}$	$O(mnr)$	$1 - \sigma$	$\frac{1}{\sigma}\log(1/\epsilon)$

* $\sigma = \sigma_{\min}\big(\mathsf{S}_{(\mathcal{S}^c)}(V_2 \otimes U_2)\big)$

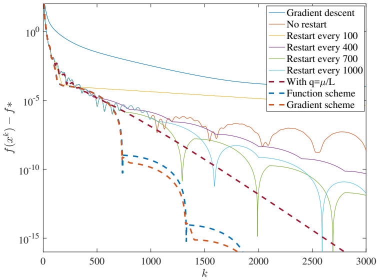
A Practical Method for Step Size Selection

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

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



Algorithm 3 ARNAG-IHT

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

Numerical Evaluation

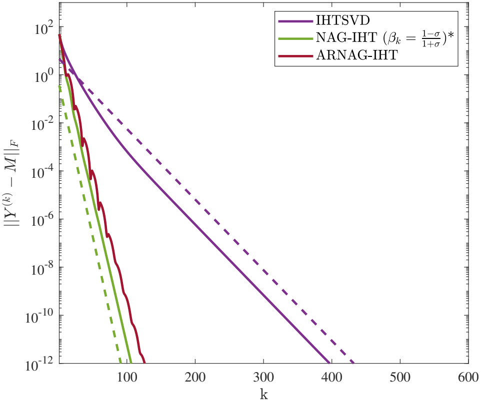


Figure 2: The distance to the solution (in log-scale) as a function of the iteration number for IHT algorithms (solid) and their corresponding theoretical bounds up to a constant (dashed). $m = 50$, $n = 40$, $r = 3$, and $s = 1000$. *NAG-IHT using optimal step size is not applicable in practice.

Outline

Conclusions and Future Work

- Conclusions
  - Propose Nesterov's Accelerated Gradient for iterative hard thresholding for matrix completion.
  - Analyze NAG-IHT with optimal step size and prove that the iteration complexity improves from $O(1/\sigma^2)$ to $O(1/\sigma)$ after acceleration.
  - Propose adaptive restart for sub-optimal step size selection that recovers the optimal rate of convergence in practice.
- Future work
  - Extend 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.