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

1. Problem Formulation
2. Background
3. Main Results
4. Conclusions and Future Work
The Netflix Prize Problem

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

Given $r=1$, find $X_{ij}, \quad (i,j) \in S^c$

subject to

\[
\text{rank}(X) \leq r \quad \text{and} \quad X_{ij} = M_{ij} \quad \text{for} \quad (i,j) \in S.
\]

\[
(r < n \leq m)
\]
Notations

- **Sampling operator** $X_S$

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

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

$$S(S) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \ \end{bmatrix} \quad \rightarrow \quad \begin{bmatrix} 4 & 2 & 4 \\ 4 & 2 & 4 \\ 4 & 2 & 4 \end{bmatrix} \quad S \quad \begin{bmatrix} 4 & 0 & 0 \\ 0 & 0 & 4 \\ 0 & 2 & 0 \\ 4 & 0 & 4 \end{bmatrix}$$
The rank-\( r \) projection of an arbitrary matrix \( X \in \mathbb{R}^{m \times n} \) is obtained by hard-thresholding singular values of \( X \):

\[
P_r(X) = \sum_{i=1}^{r} \sigma_i(X) u_i(X) v_i(X)^T
\]

The SVD of the matrix \( M \) can be partitioned based on the signal subspace and its orthogonal subspace:

\[
M = \begin{bmatrix}
U_1 & U_2
\end{bmatrix}
\begin{bmatrix}
\Sigma_1 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
V_1^T \\
V_2^T
\end{bmatrix}
\]

\( \Sigma_1 \in \mathbb{R}^{r \times r} \)
Several Formulations of Low-Rank Matrix Completion

\begin{align*}
\text{find } X_{ij}, \quad (i, j) \in S^c \quad \text{s.t.} \quad \text{rank}(X) \leq r \text{ and } X_S = M_S
\end{align*}

<table>
<thead>
<tr>
<th>Approach</th>
<th>Problem formulation</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convex relaxation</td>
<td>min $|X|_*$ s.t. $X_S = M_S$</td>
<td>✓ Rigorous guarantees</td>
</tr>
<tr>
<td></td>
<td>min $\lambda |X|_* + \frac{1}{2} |X_S - M_S|_F^2$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>min $\tau |X|_* + \frac{1}{2} |X|_F^2$ s.t. $X_S = M_S$</td>
<td>✓ Slow convergence</td>
</tr>
<tr>
<td>Non-convex</td>
<td>min $\text{rank}(X)$ s.t. $X_S = M_S$</td>
<td>✓ Fast convergence</td>
</tr>
<tr>
<td></td>
<td>min $|X_S - M_S|_F^2$ s.t. $\text{rank}(X) \leq r$ (*)</td>
<td>X Hard to analyze</td>
</tr>
<tr>
<td></td>
<td>min $|[(XY^T)_S - M_S]|_F^2$ $X \in \mathbb{R}^{m \times r}, Y \in \mathbb{R}^{n \times r}$</td>
<td></td>
</tr>
</tbody>
</table>

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

1. Problem Formulation
2. Background
3. Main Results
4. Conclusions and Future Work
Iterative Hard Thresholding for Matrix Completion

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

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

\[ \mathcal{P}_{M,S}(X) = X_{Sc} + MS \]

\[ \text{IHT with unit step size } \alpha_k = 1 \]

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

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

\[
S(S_c) \text{vec}(E^{(k+1)}) \overset{1}{=} \left( I_S - S(S_c)(V_2 \otimes U_2)(V_2 \otimes U_2)^T S(S_c)^T \right) S(S_c) \text{vec}(E^{(k)})
\]

- Stable if \( \lambda_{\text{max}}(A) = 1 - \left( \sigma_{\text{min}}(S(S_c)(V_2 \otimes U_2)) \right)^2 < 1 \)
Our work: 1-σ

Previous work: 1-σ²

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
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 recover 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 \( \mu \)-strongly convex, \( L \)-smooth function, NAG can improve the linear convergence rate from \( 1 - \mu/L \) to \( 1 - \sqrt{\mu/L} \) by setting:

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

  \[ [\text{Nesterov 2004}] \]

- 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: \[ X^{(k+1)} = \mathcal{P}_r(Y^{(k)}) \]
3: \[ Y^{(k+1)} = \mathcal{P}_{M,S}(X^{(k+1)} + \beta_k(X^{(k+1)} - X^{(k)})) \]

<table>
<thead>
<tr>
<th>Method</th>
<th># Ops./Iter.</th>
<th>Local conv. rate</th>
<th># Iters. needed $\epsilon$-acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>IHTSVD</td>
<td>$O(mnr)$</td>
<td>$1 - \sigma^2$</td>
<td>$\frac{1}{\sigma^2} \log(1/\epsilon)$</td>
</tr>
<tr>
<td>NAG-IHT with $\beta_k = \frac{1-\sigma}{1+\sigma}$</td>
<td>$O(mnr)$</td>
<td>$1 - \sigma$</td>
<td>$\frac{1}{\sigma} \log(1/\epsilon)$</td>
</tr>
</tbody>
</table>

\[ * \quad \sigma = \sigma_{\min}(S_{(S^c)}(V_2 \otimes U_2)) \]
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
  
  \[
  \beta_k = \frac{t-1}{t+2}
  \]
  
  starting at $t = 1$

- When $f(x^{(k+1)}) > f(x^{(k)})$, reset $t = 1$
Algorithm 3 ARNAG-IHT

1: $t = 1$
2: $f_0 = \|X^{(0)}_S - M_S\|_F^2$
3: for $k = 0, 1, 2, \ldots$ do
4: \hspace{0.5cm} $X^{(k+1)} = \mathcal{P}_r(Y^{(k)})$
5: \hspace{0.5cm} $Y^{(k+1)} = \mathcal{P}_{M,S}(X^{(k+1)} + \frac{t-1}{t+2}(X^{(k+1)} - X^{(k)}))$
6: $f_{k+1} = \|X^{(k+1)}_S - M_S\|_F^2$
7: if $f_{k+1} > f_k$ then $t = 1$ else $t = t + 1$  \hspace{0.5cm} △ function scheme
**Numerical Evaluation**

![Graph showing 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).](image)

**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.


