Local Convergence of the Heavy Ball method in Iterative Hard Thresholding for Low-Rank Matrix Completion

Trung Vu and Raviv Raich

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

{vutru,raich}@eecs.oregonstate.edu

1. Low-Rank Matrix Completion

Movies

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

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

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

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

2. Problem Formulation

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

$\lVert X \rVert_* = \sum_{i=1}^n \sigma_i(X)$

3. Iterative Hard Thresholding (IHT)

Algorithm 1 IHTSVD

1: **for** $k = 1, 2, \dots$ **do**

2: $\quad X^{(k)} = \mathcal{P}_r\big(Y^{(k-1)}\big)$

3: $\quad Y^{(k)} = \mathcal{P}_{M,\mathcal{S}}\big(X^{(k)}\big)$

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

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

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

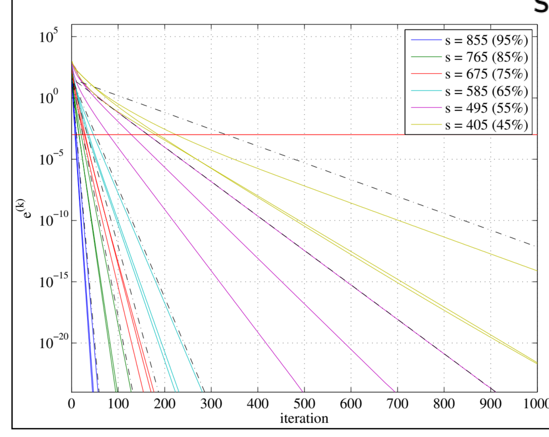
$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 0 & 4 \\ 0 & 2 & 0 \\ 4 & 0 & 4 \end{bmatrix} \xrightarrow{\mathcal{P}_r} \begin{bmatrix} 2 & 0 & 2 \\ 2 & 0 & 2 \\ 0 & 0 & 0 \\ 4 & 0 & 4 \end{bmatrix} \xrightarrow{\mathcal{P}_{M,\mathcal{S}}} \begin{bmatrix} 4 & 0 & 2 \\ 2 & 0 & 4 \\ 0 & 2 & 0 \\ 4 & 0 & 4 \end{bmatrix}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

4. Local Convergence of IHT

[1] Let $H = S_{(\mathcal{S}^c)}(V_2 \otimes U_2)(V_2 \otimes U_2)^T S_{(\mathcal{S}^c)}$ and L, μ are $\lambda_{\max}(H)$ and $\lambda_{\min}(H)$, respectively. If $\mu > 0$, then IHTSVD converges to M *locally* at a linear rate $1 - \mu$.

where

- the **row selection matrix** $S_{(\mathcal{S}^c)} \in \mathbb{R}^{(mn-s) \times mn}$ corresponding to $\mathcal{S}^c$
- the **SVD** of the matrix M can be *partitioned* based on the signal subspace and the orthogonal subspace



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

Source: [1] E. Chunikhina et. al., 2014.

5. First-Order Methods in Optimization

Table 1. Parameter selection and convergence rate of different first-order methods for minimizing a convex quadratic function $f(x) = \frac{1}{2}x^T A x + b^T x + c$, where $x \in \mathbb{R}^d$ and $\mu I_d \preceq A \preceq L I_d$. Asterisks indicate algorithms with optimal fixed step sizes. The last column describes the proportional numbers of iterations needed to reach a relative accuracy ϵ, i.e., $\lVert x^{(k)} - x^* \rVert_2 \le \epsilon \lVert x^{(0)} - x^* \rVert_2$. All algorithms share the same computational complexity per iteration.

Method	Update at each iteration	Step size selection	Rate	#Iters. needed
Gradient	$x^{(k)} = x^{(k-1)} - \alpha \nabla f(x^{(k-1)})$	$\alpha = \frac{1}{L}$	$1 - \frac{\mu}{L}$	$\frac{L}{\mu} \log(1/\epsilon)$
Gradient*		$\alpha = \frac{2}{L+\mu}$	$1 - \frac{2\mu}{L+\mu}$	$\frac{1}{2}(\frac{L}{\mu} + 1) \log(1/\epsilon)$
Nesterov	$y^{(k)} = x^{(k-1)} - \alpha \nabla f(x^{(k-1)})$	$\alpha = \frac{1}{L}, \beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$	$1 - \frac{\sqrt{\mu}}{\sqrt{L}}$	$\sqrt{\frac{L}{\mu}} \log(1/\epsilon)$
Nesterov*	$x^{(k)} = y^{(k-1)} + \beta(y^{(k-1)} - y^{(k-2)})$	$\alpha = \frac{4}{3L+\mu}, \beta = \frac{\sqrt{3L+\mu}-2\sqrt{\mu}}{\sqrt{3L+\mu}+2\sqrt{\mu}}$	$1 - 2\frac{\sqrt{\mu}}{\sqrt{3L+\mu}}$	$\frac{1}{2}\sqrt{3\frac{L}{\mu} + 1} \log(1/\epsilon)$
Heavy Ball*	$x^{(k)} = x^{(k-1)} - \alpha \nabla f(x^{(k-1)}) + \beta(x^{(k-1)} - x^{(k-2)})$	$\alpha = (\frac{2}{\sqrt{L}+\sqrt{\mu}})^2, \beta = (\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}})^2$	$1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$	$\frac{1}{2}(\sqrt{\frac{L}{\mu}} + 1) \log(1/\epsilon)$

6. Accelerated IHT

Algorithm 2 HB-IHT

1: $X^{(0)} = X^{(1)} = M_{\mathcal{S}}$

2: **for** $k = 1, 2, \dots$ **do**

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

Method	# Ops. / Iter.	Local conv. rate	#Iters. needed
IHTSVD ($\alpha_k = 1$)	$O(mnr)$	$1 - \mu$	$\frac{1}{\mu} \log(1/\epsilon)$
IHT with $\alpha_k = \frac{2}{L+\mu}$	$O(mnr)$	$1 - \frac{2\mu}{L+\mu}$	$\frac{1+L/\mu}{2} \log(1/\epsilon)$
HB-IHT with $\alpha_k = (\frac{2}{\sqrt{L}+\sqrt{\mu}})^2, \beta = (\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}})^2$	$O(mnr)$	$1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$	$\frac{1+\sqrt{L/\mu}}{2} \log(1/\epsilon)$

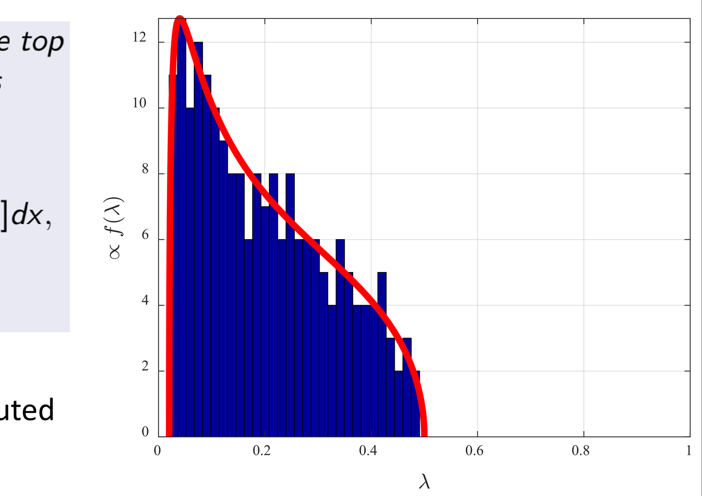
7. A Practical Guide to Parameter Selection

[2] Let U be a Haar distributed unitary matrix on $\mathbb{U}(n)$ and U_{pq} be the top left $pn \times qn$ minors of U. Consider the matrix $C_n = U_{pq} U_{pq}^*$. Then as $n \to \infty$, the ESD of C_n converges almost surely to the distribution

$$(1 - \frac{q}{p})_+ \delta_0 + (\frac{p+q-1}{p})_+ \delta_1 + \frac{\sqrt{(\lambda^+ - x)(x - \lambda^-)}}{2\pi p x (1-x)} \mathbb{I}[\lambda^- \le x \le \lambda^+] dx,$$

where $\lambda^{\pm} = (\sqrt{p(1-q)} \pm \sqrt{q(1-p)})^2$ and $x_+ = \max\{0, x\}$.

[3] extends the result to the case of Kronecker product of Haar-distributed unitary matrices

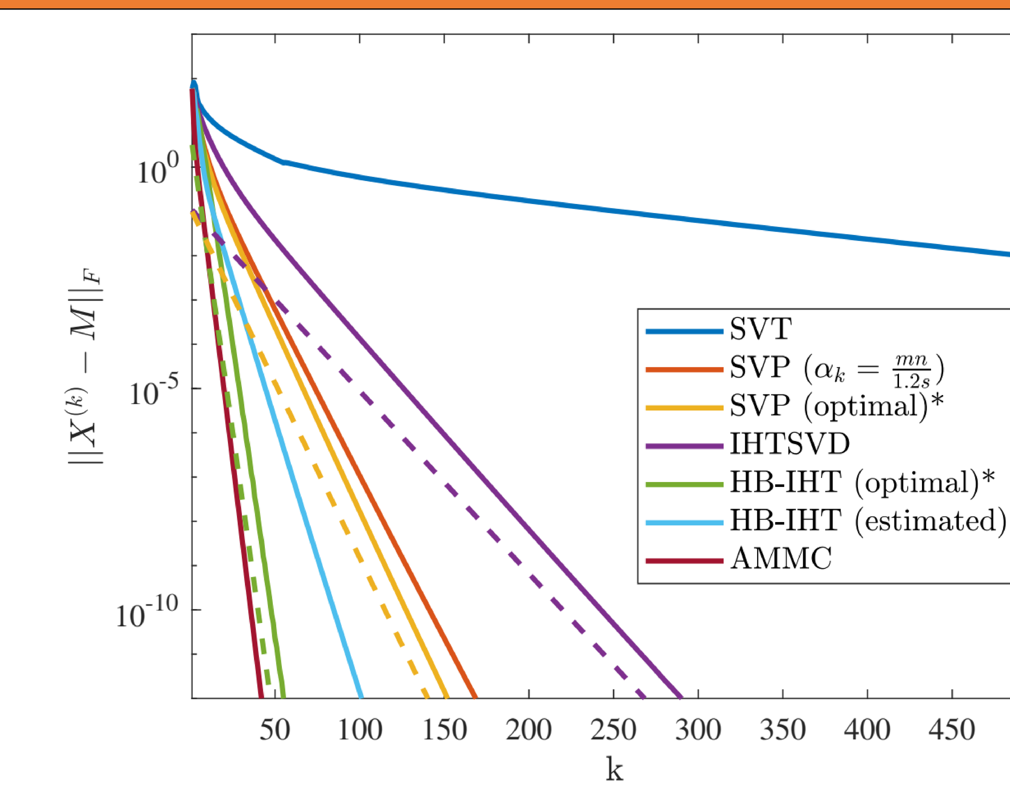


$$\hat{L} = 1, \quad \hat{\mu} = \big(\sqrt{q(1-p)} - \sqrt{p(1-q)}\big)^2.$$

$$p = 1 - \frac{s}{mn}$$

$$q = (1 - \frac{r}{m})(1 - \frac{r}{n})$$

8. Numerical Result



$m = 50, n = 40$
$r = 3, s = 1000$

9. Conclusions

- The local convergence of IHT for matrix completion can be characterized by the eigenvalues of the linearized update operator.
- As the size of the matrix grows, the eigenvalue distribution approaches the limiting ESD of the MANOVA ensemble in random matrix theory.
- Heavy Ball method can be applied to improve the local convergence of IHT.
- Future works:
  - Extend the analysis to the case where inputs are noisy or close to being low-rank.
  - Convergence under a simple initialization suggests potential analysis of global convergence.

References

1. E. Chunikhina, R. Raich, and T. Nguyen, "Performance analysis for matrix completion via iterative hard-thresholded SVD," in 2014 IEEE Workshop on Statistical Signal Processing (SSP), 2014, pp. 392–395.
2. K. W. Wachter, "The limiting empirical measure of multiple discriminant ratios," The Annals of Statistics, vol. 8, pp. 937–957, 1980.
3. B. Farrell and R. R. Nadakuditi, "Local spectrum of truncations of Kronecker products of Haar-distributed unitary matrices," Random Matrices: Theory and Applications, vol. 4, no. 1, 2013.

Acknowledgements: This work is partially supported by the National Science Foundation grants CCF-1254218, DBI-1356792.