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}@oregonstate.edu

ABSTRACT

We present a momentum-based accelerated iterative hard thresholding (IHT) for low-rank matrix completion. We analyze the convergence of the proposed Heavy Ball (HB) accelerated IHT near the solution and provide optimal step size parameters that guarantee the fastest rate of convergence. Since the optimal step sizes depend on the unknown structure of the solution matrix, we further propose a heuristic for parameter selection that is inspired by recent results in random matrix theory. Our experiment on a simple matrix completion setting verifies our analysis and illustrates the competitive rate of convergence that can be obtained with the proposed algorithm.

Index Terms— Low-rank matrix completion, Heavy Ball method, Iterative hard thresholding.

1. INTRODUCTION

This paper studies the problem of low-rank matrix completion. Given an $m \times n$ matrix M with low rank r and a set $\mathcal{S} \subset [m] \times [n]$ of its observed entries, where $[m] = \{1, 2, \ldots, m\}$, the goal is to recover the remaining entries of M. Similar to sparse recovery, the matrix completion problem (MCP) is shown to be NP-hard [1], considering the non-convexity of the problem rooted in the rank constraint.

In 2009, Candès and Recht [2] achieved a major breakthrough in matrix completion. The authors presented a convex relaxation approach to MCP by replacing the non-convex rank minimization with a (convex) nuclear norm minimization. They showed that one can perfectly recover most low-rank matrices provided the cardinality of $\mathcal{S}$ is sufficiently large. Following this work, a plethora of algorithms have been proposed for low-rank matrix completion via nuclear norm minimization. Among which, first-order methods (e.g., proximal-type algorithms) have grown more attractive due to their simplicity and scalability. However, the conservative nature of the soft thresholding operator associated with such methods often results in slow convergence.

To improve convergence while maintaining scalability, the original non-convex formulation of the problem was revisited. Empirical evidence indicated that iterative approaches to the non-convex rank minimization are faster to converge compared to their convex counterparts. Notwithstanding, theoretical convergence guarantees for such methods are non-trivial and often rely on the Restricted Isometry Property (RIP) of the affine transformations in matrix sensing. Most known examples in this category include iterative hard thresholding (IHT) [3] and alternating minimization (AMMC) [4]. Unfortunately, RIP does not hold for matrix completion even though this problem is a special case of matrix sensing. Thus, recent efforts in understanding algorithms for MCP are limited to probabilistic convergence guarantees [5, 6] or local convergence analysis [7, 8]. Moreover, acceleration techniques have been introduced to improve the performance of IHT in matrix sensing [9, 10]. Under similar assumptions to matrix RIP, the authors provided an analysis of momentum behavior and proved the linear convergence of accelerated IHT. Empirically, the authors of [10] demonstrated a faster convergence of accelerated IHT relative to plain IHT. However, they stated that the sufficient conditions to guarantee such acceleration remain as an open question.

In this work, we develop an accelerated variant of IHT for solving MCP. While the aforementioned approaches to accelerating IHT employ Nesterov's Accelerated Gradient method, we utilize Heavy Ball method due to its faster local convergence. In particular, we provide a theoretical analysis on the local convergence of the proposed algorithm and identify the choice of step sizes that guarantees optimal acceleration. Since it is computationally expensive to perform line search for the momentum parameters, we propose a simple heuristic to approximate the optimal values based on recent results from random matrix theory. Our experiment verifies the convergence rates obtained in our analysis and illustrates the efficiency of the proposed algorithm.

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

2. NOTATIONS

Without loss of generality, assume $m \geq n$. Assume the solution matrix $M = U\Sigma V^T$ is a rank-r matrix with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0$. We partition U, Σ, V as follows:

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}$$

where $\Sigma_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$, $\Sigma_2 = \mathbf{0}$, and U_1, V_1, U_2, V_2 are semi-unitary matrices corresponding to the partition of Σ.

Let $X \in \mathbb{R}^{m\times n}$ be an arbitrary matrix. We define the rank-r projection $\mathcal{P}_r$ as $\mathcal{P}_r(X) = \sum_{i=1}^{r} \sigma_i(X) u_i(X) v_i(X)^T$, where $\sigma_i(X)$, $u_i(X)$, and $v_i(X)$ are the i-th singular value, column vector, and row vector, of X, respectively. This projection produces the best rank-r approximation of X [11] and it is unique if either $\sigma_r(X) > \sigma_{r+1}(X)$ or $\sigma_r(X) = 0$. Further, we denote the cardinality of $\mathcal{S}$ by s. The sampling operator $X_{\mathcal{S}}$ is given by

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

where $\mathcal{S}^c$ is the complement of $\mathcal{S}$. Let $\hat{\mathcal{S}}^c = \{i + m(j-1) \mid (i,j) \in \mathcal{S}^c\}$. We define $S_c \in \mathbb{R}^{(mn-s)\times mn}$ as the row selection matrix obtained by selecting a subset of rows corresponding to the elements of $\hat{\mathcal{S}}^c$ from the $mn \times mn$ identity matrix.

3. BACKGROUND

Iterative hard thresholding for matrix recovery was first introduced by Jain et. al. [3] and quickly became a very attractive method for solving this problem, thanks to its simplicity and efficiency over the proximal-type algorithms [12]. Despite the successful development in theoretical analyses of IHT for *matrix sensing* [10,13], there has been little progress in understanding the convergence of IHT for *low-rank matrix completion*. The lack of RIP guarantees for MCP leaves the global convergence of IHT for MCP as an open question. Nonetheless, empirical performance analysis of the algorithm often shows linear convergence of the approach. Hence, there have been efforts to establish local convergence guarantees [7, 8]. Notably, the authors of [7] showed that the local rate of convergence of MCP-IHT can be described in a closed-form. We review the IHT algorithm for matrix completion in Algorithm 1 and restate the local convergence results in Theorem 1 and Theorem 2, using our aforementioned notations for consistency.

Theorem 1. *(Rephrased from [7]) Let $\Delta \in \mathbb{R}^{m\times n}$ be a perturbation matrix such that $\|\Delta\|_F < \frac{\epsilon}{2}$, where $\epsilon = \min_{\sigma_i > \sigma_{i+1}} \{\sigma_i - \sigma_{i+1}\}$. Then the rank-$r$ projection of $M + \Delta$ is given by*

$$\mathcal{P}_r(M + \Delta) = M + \Delta - U_2 U_2^T \Delta V_2 V_2^T + Q(\Delta) \quad (1)$$

Algorithm 1 Iterative Hard Thresholding

1: $X^{(0)} = M_{\mathcal{S}}$
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad X^{(k)} = \mathcal{P}_r\big(X^{(k-1)} - \alpha_k [X^{(k-1)} - M]_{\mathcal{S}}\big)$

where $Q : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ satisfies $\|Q(\Delta)\|_F = O(\|\Delta\|_F^2)$.

Note that (1) can be vectorized as $\operatorname{vec}\big(\mathcal{P}_r(M+\Delta)\big) = \operatorname{vec}(M) + \big(I_{mn} - (V_2 \otimes U_2)(V_2 \otimes U_2)^T\big)\operatorname{vec}(\Delta) + q(\operatorname{vec}(\Delta))$, where $q(\operatorname{vec}(\Delta)) = \operatorname{vec}(Q(\Delta))$. Denote the error vector $e^{(k)} = S_c \operatorname{vec}(X^{(k)} - M)$. Then considering Algorithm 1 with a unit step size ($\alpha_k = 1$), one can show a recursion of the error vector as follows

$$e^{(k)} = (I_{mn-s} - H) e^{(k-1)} + q(e^{(k-1)})$$

where $H = S_c(V_2 \otimes U_2)(V_2 \otimes U_2)^T S_c^T$. Further, let $L = \lambda_{\max}(H)$ and $\mu = \lambda_{\min}(H)$ be the largest and smallest eigenvalues of H, respectively. Since H is positive semi-definite and S_c, V_2, U_2 are semi-unitary matrices, it holds that $0 \leq \mu \leq L \leq 1$.

Theorem 2. *(Rephrased from [7]) If $\mu > 0$, then Algorithm 1 with a unit step size converges to M locally at a linear rate $1 - \mu$. In other words, there exists a neighborhood $\mathcal{E}(M)$ of M and a constant C such that if $X^{(0)} \in \mathcal{E}(M)$, then*

$$\left\| X^{(k)} - M \right\|_F \leq C\big(1-\mu\big)^k \left\| X^{(0)} - M \right\|_F.$$

Interestingly, the convergence rate $1 - \mu$ depends only on the solution M and the set of observed entries $\mathcal{S}$. It is also noteworthy that similar local linear convergence has been studied later in [8]. However, there is no explicit formulation of the convergence rate specified by the authors.

To gain intuition into accelerated IHT, let us start with classic results on the convergence of first-order methods for minimizing *convex quadratic functions*. In Table 1, the parameter selection is optimal in the sense that no other choice of fixed step sizes achieves faster convergence rate (see details in [14]). We list methods in ascending order of the convergence rate. In fact, Heavy Ball method not only has the fastest rate but also achieves the lower bound on convergence rate for any first-order methods for minimizing μ-strongly convex, L-smooth functions [15]. Extending these results to study the local convergence of those algorithms for optimizing a non-convex function, one could argue that the objective function can be well approximated by a quadratic inside the region near the optimum. Hence, we consider an HB-variant of MCP-IHT and analyze its local convergence behavior.

4. MAIN RESULTS

We begin this section by a brief discussion on parameter selection for plain IHT. In [3], the authors suggested an empirical choice of $\alpha_k = \frac{mn}{(1+\delta)s}$, where δ is a constant determined

Table 1. Parameter selection and convergence rate of different first-order methods for minimizing a convex quadratic function $f(x) = \frac{1}{2}x^T Ax + b^T x + c$, where $x \in \mathbb{R}^d$ and $\mu I_d \preceq A \preceq LI_d$. 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., $\left\|x^{(k)} - x^*\right\|_2 \leq \epsilon \left\|x^{(0)} - x^*\right\|_2$. 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*	$x^{(k)} = x^{(k-1)} - \alpha\nabla f(x^{(k-1)}) + \beta(x^{(k-1)} - x^{(k-2)})$	$\alpha = \left(\frac{2}{\sqrt{L}+\sqrt{\mu}}\right)^2, \beta = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2$	$1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$	$\frac{1}{2}(\sqrt{\frac{L}{\mu}}+1)\log(1/\epsilon)$

Algorithm 2 HB-IHT

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

from experiments. To further investigate the step-size selection, we examine the local convergence rate for Algorithm 1 and obtain the optimal step size in the following theorem.

Theorem 3. *If $\mu > 0$, then Algorithm 1 with step size $\alpha_k = \frac{2}{L+\mu}$ converges to M locally at a linear rate $1 - \frac{2\mu}{L+\mu}$. In other words, there exists a neighborhood $\mathcal{E}(M)$ of M and a constant C such that if $X^{(0)} \in \mathcal{E}(M)$, then*

$$\left\|X^{(k)} - M\right\|_F \leq C\big(1 - \frac{2\mu}{L+\mu}\big)^k \left\|X^{(0)} - M\right\|_F.$$

Although the optimal step size in Theorem 3 is similar to the classical result in Table 1, we note that the analysis addresses the issue on the non-convex nature of the rank-r projection.

4.1. HB-IHT

Similar to the classic Heavy Ball method, we propose an accelerated algorithm that adds a momentum term to the update in plain IHT (see Algorithm 2). This simple modification to plain IHT maintains the computational complexity of the algorithm with one additional step of calculating the difference matrix. On the other hand, the local rate of convergence can be improved significantly. Theorem 4 characterizes the local convergence of HB-IHT by providing the optimal parameter selection that guarantees improvement over plain IHT.[1]

Theorem 4. *If $\mu > 0$, then Algorithm 2 with step sizes $\alpha_k = \left(\frac{2}{\sqrt{L}+\sqrt{\mu}}\right)^2, \beta_k = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2$ converges to M locally at a linear rate $1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$. In other words, there exists a neighborhood $\mathcal{E}(M)$ of M and a constant C such that if $X^{(0)} \in \mathcal{E}(M)$, then*

$$\left\|X^{(k)} - M\right\|_F \leq C\big(1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\big)^k \left\|X^{(0)} - M\right\|_F.$$

Further, this is the optimal rate among all fixed α, β.

It is noteworthy that despite the operation of non-convex rank-r projections, we still end up with the similar result given in Table 1, thanks to the approximation of $\mathcal{P}_r$ given in (1).

4.2. A practical guide to parameter selection

Step size selection is critical to the performance of HB-IHT in practice. In this section, we propose a simple heuristic to determine the values of α_k and β_k in Algorithm 2 with no prior knowledge about L and μ. The idea is to exploit the special structure of H in order to estimate its extreme eigenvalues. We express this matrix in form of $H = WW^T$, where

$$W = S_c(V_2 \otimes U_2) = S_c(V \otimes U)(S_{2V} \otimes S_{2U})^T,$$

and $S_{2U} \in \mathbb{R}^{(m-r)\times m}$, $S_{2V} \in \mathbb{R}^{(n-r)\times n}$ are row selection matrices. Note that W is a submatrix of the Kronecker product $V \otimes U$ with the row ratio $p = 1 - \frac{s}{mn}$ and the column ratio $q = (1 - \frac{r}{m})(1 - \frac{r}{n})$. In this representation, the structure of H is closely related to the MANOVA random matrix ensemble, and more interestingly, the limiting density of its eigenvalues is identified by Watcher [16], dating back to the early 1980s. In his study, Watcher showed that as the size of a MANOVA matrix with parameters (p, q) approaches infinity, its empirical spectral distribution (ESD) converges to a deterministic probability measure supported on the interval $[\lambda^-, \lambda^+] \cup \{0, 1\}$, where $\lambda^\pm = \left(\sqrt{p(1-q)} \pm \sqrt{q(1-p)}\right)^2$. Recently, similar result was also found by Raich and Kim [17] for the truncation of random unitary matrices. Moreover,

[1] The proofs of Theorem 1, 3 and 4 are given at the folowing link: http://web.engr.oregonstate.edu/∼vutru/hb_appendix.pdf

Farrell and Nadakuditi [18] extended the results from Haar (uniformly) distributed unitary matrices to Kronecker product case. The authors proved that the ESD of random matrices of the form $\Pi_1(U \otimes V)\Pi_2(U \otimes V)^*\Pi_1$, where Π_1, Π_2 are orthogonal projections of ranks pn and qn, respectively, also converges to the same limiting distribution. Considering H to be an instance of this case, we conjecture that its spectral distribution will be close to the aforementioned. In particular,

1. if $p < q$, then H has no zero eigenvalue and the smallest eigenvalue of H is close to λ^- with high probability,
2. if $p + q > 1$, then H has unit eigenvalue and the largest eigenvalue of H is 1.

It is worthwhile to note that both conditions usually hold in practice when q is rather close to 1. Hence, we propose the following estimation of L and μ:

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

Empirically, we observe this heuristic significantly outperforms plain IHT in terms of convergence. However, understanding when and how it works would involve the non-asymptotic theory of random matrices [19]. For instance, characterizing the variance of extreme eigenvalues, i.e., difference between μ and $\hat{\mu}$, in case of Kronecker unitary matrices is much more challenging than their Haar-distributed counterparts. Our experiments suggest that they tend to have wider fluctuations. We leave this analysis for future direction.

5. NUMERICAL EVALUATION

This section presents an empirical evaluation of several methods for low-rank matrix completion including the proposed approach. First, we generate a low-rank solution matrix $M \in \mathbb{R}^{m \times n}$ by taking the product of an $m \times r$ matrix and an $r \times n$ matrix, each having i.i.d. normally distributed entries. Next, we sample the observation set $\mathcal{S}$ uniformly at random. In our experiment, we choose $m = 50$, $n = 40$, $r = 3$, and $s = 1000$. For comparison, we consider the following methods: SVT [12], SVP [3], IHTSVD [7] and AMMC [5]. Although the convergence guarantee of SVP does not hold for MCP in general, it is interesting to compare its empirical performance with optimal step size given in Theorem 3. In our own implementation of these algorithms, we use the set of parameters as suggested by the authors. For the proximal-type SVT algorithm, we set the step size $\delta = 1.2\frac{mn}{s}$ and the threshold $\tau = 5\sqrt{mn}$. For SVP, we set the step size $\eta_t = \frac{mn}{1.2s}$. IHTSVD and AMMC are parameter-free. Finally, we add HB-IHT with the aforementioned theoretical optimal step sizes and heuristic step sizes for comparison.

Figure 1 illustrates the Frobenius norm of the error matrix as a function of the number of iterations. The dashed lines correspond to the theoretical convergence of IHTSVD (purple) at rate $1 - \mu$, optimal step size SVP (yellow) at rate $1 - \frac{2\mu}{L+\mu}$ and optimal step size HB-IHT (green) at rate

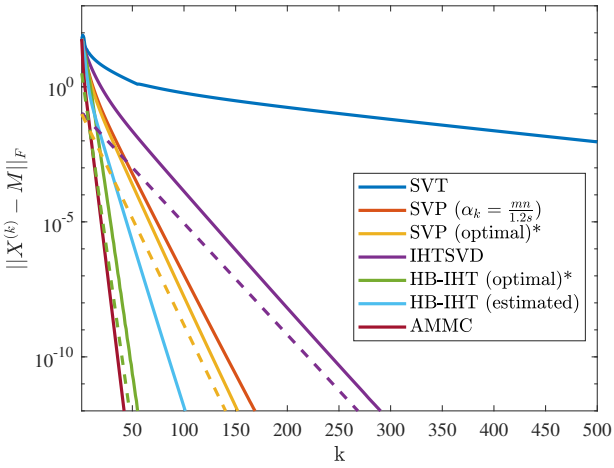

Fig. 1. The distance to the solution (in log-scale) as a function of the iteration number for various algorithms (solid) and their corresponding theoretical bounds up to a constant (dashed). Asterisks indicate algorithms using theoretical step sizes that are not available in practice. All algorithms share the same computational complexity per iteration except AMMC.

$1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$. These three algorithms certainly match the performance predicted in theory. SVT exhibits the slowest convergence as expected from our foregoing discussion. By contrast, all IHT algorithms enjoy the linear convergence. Without acceleration, SVP with step size either $\frac{mn}{1.2s}$ or $\frac{2}{L+\mu}$ is clearly faster than IHTSVD. Nevertheless, HB-IHT with estimated step sizes outperforms all plain IHT algorithms, yet still slower than HB-IHT with theoretically-optimal step sizes. Finally, we compare the performance of HB-IHT with optimal step sizes with AMMC, which is shown to converge linearly at rate faster than $1/4$ in [5]. While our accelerated algorithm obtains a comparable rate, it requires significantly less computation per iteration thanks to the recent breakthroughs in k-SVD algorithms [20], i.e., the iteration complexity for HB-IHT is $O(mnr + poly(1/\epsilon))$, compared to $O(sm^2r^2 + m^3r^3)$ for AMMC as claimed in [5].

6. CONCLUSION AND FUTURE WORK

To summarize, we introduced the use of Heavy Ball method to significantly accelerate IHT for low-rank matrix completion. We analyzed the local convergence of HB-IHT and established the optimal step sizes to guarantee better performance over plain IHT. We further provided evidence that these optimal values can be approximated by a simple calculation in practice. Our experiment verified the analysis and demonstrated the efficiency of the proposed algorithm. Study of our approach in the noisy case is left for an extended version of this paper.

7. REFERENCES

[1] A. L. Chistov and D. Yu. Grigoriev, "Complexity of quantifier elimination in the theory of algebraically closed fields," *Proceedings of the 11th Symposium on Mathematical Foundations of Computer Science*, vol. 176, pp. 17–31, 1984.

[2] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717, 2009.

[3] P. Jain, R. Meka, and I. Dhillon, "Guaranteed rank minimization via singular value projection," in *Advances in Neural Information Processing Systems (NIPS)*, 2010, pp. 937–945.

[4] C. Chen, B. He, and X. Yuan, "Matrix completion via an alternating direction method," *IMA Journal of Numerical Analysis*, vol. 32, no. 1, pp. 227–245, 2012.

[5] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, 2013, pp. 665–674.

[6] R. H. Keshavan, *Efficient algorithms for collaborative filtering*, Ph.D. thesis, Stanford University, 2012.

[7] 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.

[8] M.-J. Lai and A. Varghese, "On convergence of the alternating projection method for matrix completion and sparse recovery problems," *eprint arXiv:1711.02151*, 2017.

[9] A. Kyrillidis and V. Cevher, "Matrix recipes for hard thresholding methods," *Journal of mathematical imaging and vision*, vol. 48, no. 2, pp. 235–265, 2014.

[10] R. Khanna and A. Kyrillidis, "IHT dies hard: Provable accelerated iterative hard thresholding," in *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, 2018, vol. 84, pp. 188–198.

[11] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[12] J.-F. Cai, E. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[13] J. Tanner and K. Wei, "Normalized iterative hard thresholding for matrix completion," *SIAM Journal on Scientific Computing*, vol. 35, no. 5, 2013.

[14] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 57–95, 2014.

[15] A. Nemirovski and D. Yudin, *Problem complexity and method efficiency in optimization*, John Wiley & Sons, 1983.

[16] K. W. Wachter, "The limiting empirical measure of multiple discriminant ratios," *The Annals of Statistics*, vol. 8, pp. 937–957, 1980.

[17] R. Raich and J. Kim, "On the eigenvalue distribution of column sub-sampled semi-unitary matrices," in *Statistical Signal Processing Workshop (SSP), 2016 IEEE*. IEEE, 2016, pp. 1–5.

[18] 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.

[19] M. Rudelson and R. Vershynin, "Non-asymptotic theory of random matrices: extreme singular values," in *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010)*. World Scientific, 2010, pp. 1576–1602.

[20] Z. Allen-Zhu and Y. Li, "LazySVD: Even faster SVD decomposition yet without agonizing pain," in *Advances in Neural Information Processing Systems*, 2016, pp. 974–982.