CONVERGENCE ANALYSIS OF THE RANK-RESTRICTED SOFT SVD ALGORITHM *

MAHENDRA PANAGODA[†], TYRUS BERRY[†‡], AND HARBIR ANTIL[†]

Abstract. The soft SVD is a robust matrix decomposition algorithm and a key component of matrix completion methods. However, computing the soft SVD for large sparse matrices is often impractical using conventional numerical methods for the SVD due to large memory requirements. The Rank-Restricted Soft SVD (RRSS) algorithm introduced by Hastie et al. addressed this issue by sequentially computing low-rank SVDs that easily fit in memory. We analyze the convergence of the standard RRSS algorithm and we give examples where the standard algorithm does not converge. We show that convergence requires a modification of the standard algorithm, and is related to non-uniqueness of the SVD. Our modification specifies a consistent choice of sign for the left singular vectors of the low-rank SVDs in the iteration. Under these conditions, we prove linear convergence of the singular vectors using a technique motivated by alternating subspace iteration. We then derive a fixed point iteration for the evolution of the singular values and show linear convergence to the soft thresholded singular values of the original matrix. This last step requires a perturbation result for fixed point iterations which may be of independent interest.

Key words. low rank approximation, soft SVD, matrix completion, regularization

AMS subject classifications. 65F55, 65F22, 15A83

1. The Rank-Restricted Soft SVD.

In this paper we consider the following rank-restricted matrix decomposition problem,

$$\min_{A\in\mathbb{R}^{n\times r},B\in\mathbb{R}^{m\times r}} \quad \frac{1}{2}\|X-AB^\top\|_F^2+\frac{\lambda}{2}(\|A\|_F^2+\|B\|_F^2) \tag{1.1}$$

where $X\in\mathbb{R}^{n\times m}$ is considered the input to the problem, $r\le p\equiv\min\{m,n\}$ is the rank restriction, and λ is a regularization parameter. The product $AB^\top$ is an approximation of X in the Frobenius norm with rank at most r. In [11, 9] it was shown that when A,B solve (1.1) the product $AB^\top$ solves,

$$\min_{Z\,:\,\mathrm{rank}(Z)\le r} \quad \frac{1}{2}\|X-Z\|_F^2+\lambda\|Z\|_* \tag{1.2}$$

where the nuclear norm $\|Z\|_*$ is the sum of the singular values of Z. The relationship between these solutions suggests that $AB^\top$ is a robust low-rank approximation to X. This approximation is a key component of many matrix completion algorithms [9, 11, 4, 5]. In this paper we will analyze a numerical method for solving (1.1) proposed by Hastie et al. in [9]. We will show that a modification is required to obtain convergence, and we give the first complete proof of convergence.

The problem (1.1) is called the Rank-Restricted Soft SVD (RRSS) because the solution involves soft-thresholding of the singular value decomposition (SVD). Given the reduced SVD, $X=USV^\top$ ($U\in\mathbb{R}^{n\times p}$, $S\in\mathbb{R}^{p\times p}$, $V\in\mathbb{R}^{m\times p}$ where $p\equiv\min\{m,n\}$), the solution to (1.1) is found by first soft-thesholding the singular values,

$$D\equiv\sqrt{(S-\lambda I)^+}=\sqrt{\max\{0,S-\lambda I\}}$$

*Submitted to the editors DATE.

Funding: This work is partially supported by NSF grants DMS-1854204, DMS-2006808, DMS-1818772, DMS-1913004, the Air Force Office of Scientific Research (AFOSR) under Award NO: FA9550-19-1-0036, and Department of Navy, Naval PostGraduate School under Award NO: N00244-20-1-0005.

[†]Department of Mathematical Sciences, George Mason University, Fairfax, VA.

[‡]tberry@gmu.edu

and then defining $A_{\text{opt}} = UDI_{p\times r}$ and $B_{\text{opt}} = VD^\top I_{p\times r}$ [9]. When X is full matrix, a standard or partial SVD can be used to obtain this solution. However, in many applications such as matrix completion, X is a sparse matrix that is too large to be stored as a full matrix. Motivated by these applications, in [9] Hastie et al. introduced a fast and memory efficient alternating ridge regression algorithm shown as Algorithm 1 below.

Algorithm 1.1 Alternating Directions Optimization for (1.1)

```
Inputs: An n × m matrix X, rank restriction r, and regularization parameter λ
Outputs: An n × r matrix A and an m × r matrix B

Initialize A as a random n × r matrix and A_p = B_p = 0
while ||A−A_p||_max/||A||_max + ||B−B_p||_max/||B||_max > tol do
    A_p = A, B_p = B
    Update B leaving A fixed:
        B ← X^⊤ A(A^⊤ A + λ I_{r×r})^{−1}
    Update A leaving B fixed:
        A ← XB(B^⊤ B + λ I_{r×r})^{−1}
end while
```

We first consider a simplistic approach to solving (1.1) shown in Algorithm 1. This method is motivated by the alternating directions method of optimization [12, 2]. The objective function in (1.1) is not convex as a function of both A and B together, however, when either A or B is fixed the objective function is convex and quadratic in the other. For example when A is fixed, we can rewrite the objective function in (1.1) as,

$$\sum_{i=1}^{N} \frac{1}{2}||X_i - AB_i||_2^2 + \frac{\lambda}{2}||B_i||_2^2 + c_1 = \sum_{i=1}^{N} \frac{1}{2} B_i^\top (A^\top A + \lambda I_{r\times r})B_i - B_i^\top A^\top X_i + c_2$$

where X_i is the i-th column of X and B_i is the i-th column of $B^\top$ (c_1, c_2 are constants with respect to B). The optimization problems for each column of $B^\top$ are independent and the optimal solution is $B_i = (A^\top A + \lambda I_{r\times r})^{-1} A^\top X_i$. Combining these columns we find the optimal solution for B, when A is fixed, has the closed form solution, $X^\top A(A^\top A + \lambda I_{r\times r})^{-1}$. If we then hold B fixed, we have a similar optimization problem for A with optimal solution $XB(B^\top B + \lambda I_{r\times r})^{-1}$.

While the alternating directions method does converge, as shown in Figure 1(right panel) it has slow convergence even when X is approximately low-rank . Hastie et al. noticed that the Algorithm 1 looks like a power iteration method, since at each step we multiply the current A or B by either X or $X^\top$ respectively [9]. Thus, motivated by the idea of orthogonal power iteration, Hastie et al. introduced the idea of using an SVD between each alternation in order to orthogonalize the columns of A and B. Notice that A and B are $m\times r$ and $n\times r$ respectively, so for $r \ll \min\{m, n\}$ these SVDs will often be computable even when the full SVD of X is impractical. These insights led Hastie et al. to introduce Algorithm 1.2 in [9]. The authors in [9] suggested that the approach used to show convergence of orthogonal power iteration (see for example [8] Theorem 8.2.2, also [1]) could be applied to Algorithm 1.2. In Section 2 we will confirm that the method of [8] can indeed be adapted to show convergence of the

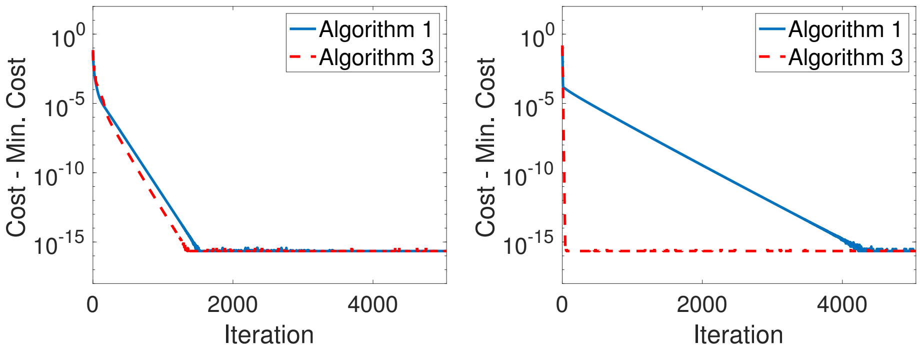

FIG. 1. *For a full-rank X matrix (left) Algorithms 1 and 1.3 have similar performance, however when X is approximately low-rank (right) Algorithm 1.3 is significantly faster. In both examples X is 500×500 and we set $r = 10$ and $\lambda = 0.5$ and the minimum cost is computed using $A_{\mathrm{opt}}, B_{\mathrm{opt}}$. In the left panel the entries of X are independent standard Gaussian random variables. In the right panel $\tilde{A}, \tilde{B}$ are 500×10 matrices with independent standard Gaussian entries and $X = \tilde{A}\tilde{B}^\top + 10\tilde{X}$ where $\tilde{X}$ is 500×500 with standard Gaussian entries.*

singular vectors. However, a more detailed analysis is required to show convergence of the singular values, as we will show in Section 3. Moreover, Algorithm 1.2 can fail to converge or converge to a non-optimal stationary point due to a subtle issue involving the non-uniqueness of the SVD.

Algorithm 1.2 Rank-Restricted Soft SVD [9]

Inputs: An $n \times m$ matrix X,
Rank restriction r, and
Regularization parameter λ
Outputs: An $n \times r$ matrix A and
An $m \times r$ matrix B

Initialize $D = I_{r \times r}$
Initialize $U \in \mathbb{R}^{n \times r}$ a random orthonormal matrix
Initialize $A = UD$ and $A_{\mathrm{p}} = B_{\mathrm{p}} = 0$
while $\frac{||A - A_{\mathrm{p}}||_{\max}}{||A||_{\max}} + \frac{||B - B_{\mathrm{p}}||_{\max}}{||B||_{\max}} > \mathrm{tol}$ **do**
  Set $A_{\mathrm{p}} = A$, $B_{\mathrm{p}} = B$
  Update B leaving A fixed:
    $B \leftarrow X^\top A(D^2 + \lambda I_{r \times r})^{-1}$
    Find the SVD: $BD = USV^\top$
    $D \leftarrow S^{\frac{1}{2}}$
    $B \leftarrow UD$
  Update A leaving B fixed:
    $A \leftarrow XB(D^2 + \lambda I_{r \times r})^{-1}$
    Find the SVD: $AD = USV^\top$
    $D \leftarrow S^{\frac{1}{2}}$
    $A \leftarrow UD$
end while

Algorithm 1.3 Modified Rank-Restricted Soft SVD

Inputs: An $n \times m$ matrix X,
Rank restriction r, and
Regularization parameter λ
Outputs: An $n \times r$ matrix A and
An $m \times r$ matrix B

Initialize $D = I_{r \times r}$
Initialize $U \in \mathbb{R}^{n \times r}$ a random orthonormal matrix
Initialize $A = UD$ and $A_{\mathrm{p}} = B_{\mathrm{p}} = 0$
while $\frac{||A - A_{\mathrm{p}}||_{\max}}{||A||_{\max}} + \frac{||B - B_{\mathrm{p}}||_{\max}}{||B||_{\max}} > \mathrm{tol}$ **do**
  Set $A_{\mathrm{p}} = A$, $B_{\mathrm{p}} = B$
  Update B leaving A fixed:
    $B \leftarrow X^\top A(D^2 + \lambda I_{r \times r})^{-1}$
    Find the SVD: $BD = USV^\top$
    $D \leftarrow S^{\frac{1}{2}}, W = \mathrm{diag}(\mathrm{sign}(V^\top \vec{1}))$
    $B \leftarrow UWD$
  Update A leaving B fixed:
    $A \leftarrow XB(D^2 + \lambda I_{r \times r})^{-1}$
    Find the SVD: $AD = USV^\top$
    $D \leftarrow S^{\frac{1}{2}}, W = \mathrm{diag}(\mathrm{sign}(V^\top \vec{1}))$
    $A \leftarrow UWD$
end while

1.1. Proposed Algorithm. Despite the similarity of Algorithm 1.2 to orthogonal power iteration, there is a key difference which can cause Algorithm 1.2 to fail to converge. Orthogonal power iteration uses the QR factorization, which is naturally unique when you specify that the the diagonal entries of R are non-negative. The SVD on the other hand does not have a natural choice of sign for the singular vectors [3]. The SVD is only unique up to a choice of sign since for any matrix W which is diagonal with diagonal entries in $\{-1, 1\}$ we have,

$$USV^\top = UWSWV^\top = \tilde{U}S\tilde{V}^\top.$$

This non-uniqueness means that many SVD algorithms will return different choices of W each time they are run (due to random initialization). This can lead to failure of Algorithm 1.2 to converge, simply due to oscillations in A and B caused by varying implicit choices of W in the SVD steps. Moreover, as we will show in Section 4, the different choices of W correspond to alternate stationary points of the cost function in (1.1).

To address these issues, we introduce Algorithm 1.3 which is a modification of Algorithm 1.2. The new aspect of Algorithm 1.3 is that, after each SVD, we make a unique choice of sign for the left singular vectors. This seemingly minor addition proves critical for convergence as shown in Figure 2 and as we will prove analytically in Section 3 below. In fact, we will show that this choice of sign insures that the matrices V of right singular vectors converge to the identity matrix and that this choice is required to obtain the optimal solution of (1.1).

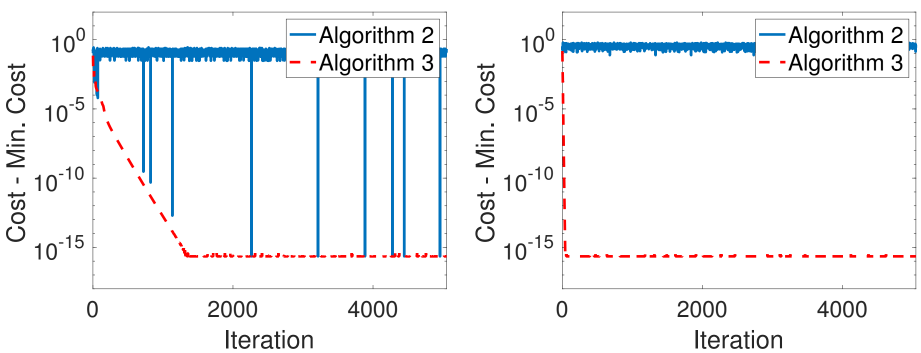

FIG. 2. *Comparison of Algorithm 1.2 from [9] with our new Algorithm 1.3 on the same full-rank (left) and approximately low-rank (right) examples from Figure 1.*

We will formalize Algorithm 1.3 mathematically since Algorithm 1.2 can then be obtained by simply redefining the choice of W. Based on Algorithm 1.3 we make the following recursive definitions,

$$B_{k+1} = X^\top U_k W_k D_k (D_k^2 + \lambda I)^{-1} \tag{1.3a}$$

$$\tilde{U}_k \tilde{W}_k \tilde{D}_k^2 \tilde{W}_k \tilde{V}_k^\top = B_{k+1} D_k \tag{1.3b}$$

$$A_{k+1} = X \tilde{U}_k \tilde{W}_k \tilde{D}_k (\tilde{D}_k^2 + \lambda I)^{-1} \tag{1.3c}$$

$$U_{k+1} W_{k+1} D_{k+1}^2 W_{k+1} V_{k+1}^\top = A_{k+1} \tilde{D}_k \tag{1.3d}$$

where (1.3b) and (1.3d) define all the quantities on the left hand side by computing the SVD of the right hand side. We initialize $\tilde{D}_{-1} = D_0 = W_0 = I$ and choose U_0 to be a random orthonormal $n \times r$ matrix and set $A_0 = U_0 W_0 D_0$.

The matrices $W_k, \tilde{W}_k$ are diagonal matrices where each diagonal entry is either 1 or -1. These matrices define the choice of signs for the left and right singular vectors resulting from the SVD computation. In fact, due to random initializations of most SVD algorithms, the matrices $W_k, \tilde{W}_k$ are typically random and will be different each time the SVD algorithm is run. As we will see, this will be the cause of the erratic behavior of the cost function in Algorithm 1.2 as shown in Figure 2.

A more concise iteration can be obtained by solving solving (1.3d) (at the previous step) for $U_k W_k D_k = A_k \tilde{D}_{k-1} V_k W_k D_k^{-1}$ and substituting into (1.3a) we have,

$$B_{k+1} = X^\top A_k \tilde{D}_{k-1} V_k W_k D_k^{-1} (D_k^2 + \lambda I)^{-1}. \tag{1.4}$$

Similarly, solving (1.3b) for $\tilde{U}_k \tilde{W}_k \tilde{D}_k = B_{k+1} D_k \tilde{V}_k \tilde{W}_k \tilde{D}_k^{-1}$ and by substituting into (1.3c) we can write,

$$A_{k+1} = X B_{k+1} D_k \tilde{V}_k \tilde{W}_k \tilde{D}_k^{-1} (\tilde{D}_k^2 + \lambda I)^{-1}. \tag{1.5}$$

Here we can immediately see that the product $A_{k+1} B_{k+1}^\top$ will not converge unless the signed right singular vectors $\tilde{V}_k \tilde{W}_k, W_k V_k^\top$ of (1.3b),(1.3d) converge since,

$$A_{k+1} B_{k+1}^\top = X B_{k+1} D_k \tilde{V}_k \tilde{W}_k \tilde{D}_k^{-1} (\tilde{D}_k^2 + \lambda I)^{-1} (D_k^2 + \lambda I)^{-1} D_k^{-1} W_k V_k^\top \tilde{D}_{k-1} A_k^\top X.$$

This explains the jumps of Algorithm 1.2 shown in Figure 2.

1.2. Overview. In Section 2 we will show that, in an appropriate sense, we have $U_k \to U$ and $\tilde{U}_k \to V$. Then, in Section 3, we turn to the singular values and show that $D_k, \tilde{D}_k$ both converge to $I_{r\times p} D I_{p\times r}$ given by the softmax function $D = \sqrt{(S - \lambda I)^+}$. Finally, in Section 4 we will show that $V_k, \tilde{V}_k$ converge to diagonal matrices determined by the choice of $W_k, \tilde{W}_k$. We will see that any convergent choice for the diagonal sign matrices $W_k, \tilde{W}_k$ will yield a convergent algorithm. These results will culminate in Theorem 4.2 which reveals that, assuming $\tilde{W}_k \to \tilde{W}_*$ and $W_k \to W_*$, we have the limiting matrices,

$$A_k \to A_* = USD(D^2 + \lambda I)^{-1} I_{p\times r} \tilde{W}_*$$
$$B_k \to B_* = VSD(D^2 + \lambda I)^{-1} I_{p\times r} W_*$$

for Algorithm 1.3. Moreover, the dependence of the first term of the cost function (1.1) on the sign matrices is given by,

$$||X - A_* B_*^\top||_F = ||S - S^2 D^2 (D^2 + \lambda I)^{-2} I_{p\times r} \tilde{W}_* W_* I_{r\times p}||_F \tag{1.6}$$

and only the choice $\tilde{W}_* W_* = I$ will minimize the cost. When $\lambda < S_{rr}$ the above cost simplifies to,

$$||X - A_* B_*^\top||_F = ||S - (S - \lambda)^+ I_{p\times r} \tilde{W}_* W_* I_{r\times p}||_F$$

which is optimal when $\tilde{W}_* W_* = I$. This explains the large cost values for Algorithm 1.2 shown in Figure 2 since the random $W_k, \tilde{W}_k$ essentially replace $\tilde{W}_*, W_*$ with random sign matrices. Of course, occasionally these random sign matrices yield $\tilde{W}_k W_k = I$, which explains why the cost sometimes jumps down to the optimal cost. This also justifies our choice in Algorithm 1.3 where $W_k, \tilde{W}_k$ are chosen to insure that the sum of the columns of $W_k V_k$ and $\tilde{W}_k \tilde{V}_k$ are positive. As $V_k, \tilde{V}_k$ converge to diagonal matrices, this choice will guarantee that $\tilde{W}_* W_* = I$, thereby obtaining the minimal cost solution.

2. Convergence of the Singular Vectors. The first part of proving the convergence of Algorithm 1.3 is showing that the sequences U_k and $\tilde{U}_k$ defined in (1.3b) and (1.3d) converge to the top r left and and right singular vectors of X respectively. In other words, if $X = USV^\top$ is the SVD of X then loosely speaking we have $U_k \to U_{(1:r)}$ and $V_k \to V_{(1:r)}$ where the subscript $(1:r)$ indicates the first through r-th columns of the matrix. The reason we say 'loosely speaking' is due to the non-uniqueness of sign in the singular vectors, even for unique singular values (for repeated singular values we only have uniqueness up to orthogonal linear transformations). Thus, the first column of U_k could alternate between that of U and its negative and this would still be considered convergence since we would have obtained the correct subspace.

We define convergence in terms of the norm of the matrix of inner products $||U_k^\top U_{(r+1:n)}||$ converging to 0 (any matrix norm can be used since this always implies $U_k^\top U_{(r+1:n)}$ is zero). Since $U_k U_k^\top = I_{r\times r}$, the columns of U_k span an r-dimensional subspace, so if $U_k^\top U_{(r+1:n)} = 0$ this subspace is orthogonal to the subspace spanned by the last $n-r$ columns of U. Thus, $||U_k^\top U_{(r+1:n)}||_{\max} \to 0$ implies that the subspace spanned by the columns of U_k is aligning with the subspace spanned by the first r columns of U. As shown in Figure 3 we have $||U_k^\top U_{(r+1:n)}||_{\max} \to 0$ for both Algorithm 1.2 and Algorithm 1.3.

In this section we will prove that this convergence is independent of the choice of $W_k, \tilde{W}_k$ and show that the convergence rate is determined by the ratio of the $(r+1)$-st and r-th squared singular values of X. In particular, when X is low-rank or approximately low-rank, this will imply the fast convergence observed in Figure 1. We first note that the iteration (1.3a)-(1.3d) is rank preserving in the generic case when X is full-rank.

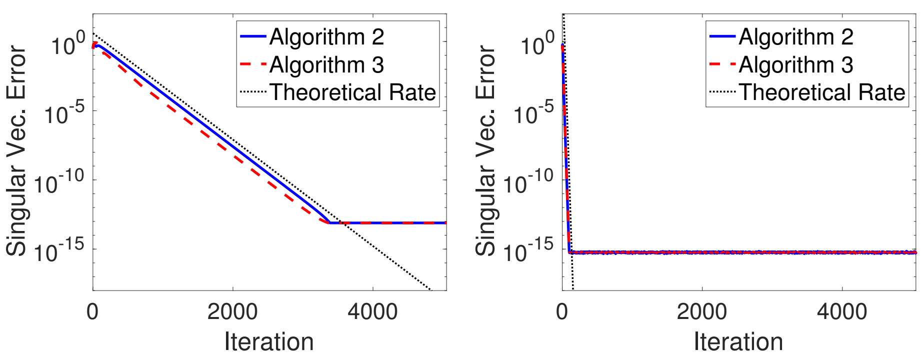

FIG. 3. *Comparison of the convergence of the singular vectors on the same full-rank (left) and approximately low-rank (right) examples from Figure 1. Error is measured by* $||U_k^\top U_{(r+1:n)}||_{\max}$, *where* $U_{(r+1:n)}$ *is the matrix containing the* $(r+1)$*-st through* n*-th columns of* U*. The theoretical convergence rate* $\left(\frac{s_{r+1}}{s_r}\right)^2$ *shown is proven in Theorem 2.3 . Notice that the singular vectors converge for both Algorithm 1.2 from [9] and our new Algorithm 1.3 .*

LEMMA 2.1. *Let* $X \in \mathbb{R}^{n\times m}$*, have full rank, namely* $\mathrm{rank}(X) = \min\{m,n\}$*, then for all* k *the matrices* $A_k, B_k, U_k, W_k, D_k, V_k, \tilde{U}_k, \tilde{W}_k, \tilde{D}_k, \tilde{V}_k$ *defined by the iteration* (1.3a)-(1.3d) *are all full rank.*

Proof. The algorithm is initialized with $A_0 = U_0 W_0 D_0$, where U_0 is a random matrix and thus generically full rank and $D_0 = W_0 = I$ is full rank. By (1.3a) we have $B_{k+1} = X^\top A_k (D_k^2 + \lambda I)^{-1}$ and since X and $D_k^2 + \lambda I$ are full rank, we have $\mathrm{rank}(B_{k+1}) = \mathrm{rank}(A_k)$. This establishes the base case, and if we inductively assume A_k, D_k are full rank we immediately find that B_{k+1} is full rank and thus $B_{k+1}D_k$ is also full rank. Since the right-hand-side of (1.3b) is full rank, all the matrices $\tilde{U}_k, \tilde{W}_k, \tilde{D}_k, \tilde{V}_k$ on the left-hand-side of (1.3b) are full rank since they are defined to be the SVD of a full rank matrix. By (1.3c) we have A_{k+1} written as a product of full rank matrices and thus A_{k+1} is full rank. Finally, the right-hand-side of (1.3d) is now full rank which implies that all the matrices on the left-hand-side, $U_{k+1}, W_{k+1}, D_{k+1}, V_{k+1}$ are all full rank. This completes the induction. ◻

When X is not full rank, generically the random initial matrix U_0 will not be orthogonal to the subspace spanned by the rows of X and since $B_1 = X^\top A_0/(1+\lambda)$ we find $\mathrm{rank}(B_1) = \min\{\mathrm{rank}(X), \mathrm{rank}(A_0)\}$. Note that since $D_0 = I$ we have $(D_0^2 + \lambda I)^{-1} = I/(1+\lambda)$. When $\mathrm{rank}(X) \geq r$ we expect all of the matrices in Lemma 2.1 to have rank r and when $\mathrm{rank}(X) < r$ they should all have rank equal to $\mathrm{rank}(X)$. However, showing that U_k does not evolve to become orthogonal to the span or the rows of X requires Theorem 2.3 below.

The next step is to make a connection between the iteration (1.3a)-(1.3d) and the SVD of X. In the next lemma we show how the (1.3a) followed by (1.3c) is related to multiplication by $XX^\top$ and similarly (1.3c) followed by (1.3a) is related to multiplication by $X^\top X$.

LEMMA 2.2. *Let $X \in \mathbb{R}^{n\times m}$, and using the notation of* (1.3a)-(1.3d) *define*

$$P_{k+1} \equiv D_{k+1}^2 V_{k+1}^\top (\tilde{D}_k^2 + \lambda I)\tilde{W}_k \tilde{V}_k^\top D_k^{-2}(D_k^2 + \lambda I)W_k$$
$$\tilde{P}_{k+1} \equiv \tilde{D}_{k+1}^2 \tilde{V}_{k+1}^\top (D_{k+1}^2 + \lambda I)W_{k+1} V_{k+1}^\top \tilde{D}_k^{-2}(\tilde{D}_k^2 + \lambda I)\tilde{W}_k$$

then

$$XX^\top U_k = U_{k+1}P_{k+1} \qquad (XX^\top)^k U_0 = U_k P_k \cdots P_1$$
$$X^\top X\tilde{U}_k = \tilde{U}_{k+1}\tilde{P}_{k+1} \qquad (X^\top X)^k \tilde{U}_0 = \tilde{U}_k \tilde{P}_k \cdots \tilde{P}_1$$

and the products

$$\begin{aligned} Q_k &\equiv D_k^{-2} P_k \cdots P_1 \\ &= V_k^\top \left(\prod_{i=1}^{k-1} (\tilde{D}_i^2 + \lambda I)\tilde{W}_i \tilde{V}_i^\top (D_i^2 + \lambda I) W_i V_i^\top \right) (\tilde{D}_0^2 + \lambda I)\tilde{W}_0 V_0^\top (1+\lambda) \\ \tilde{Q}_k &\equiv \tilde{D}_k^{-2}\tilde{P}_k \cdots \tilde{P}_1 = \tilde{V}_k^\top (D_k^2 + \lambda I) W_k Q_k \end{aligned}$$

are invertible with inverses bounded by $||Q_k^{-1}|| \leq \lambda^{1-2k}$, and $||\tilde{Q}_k^{-1}|| \leq \lambda^{2-2k}$.

Proof. We first solve (1.3a) for $X^\top U_k = B_{k+1}(D_k^2 + \lambda I)D_k^{-1}W_k$ to obtain,

$$\begin{aligned} XX^\top U_k &= XB_{k+1}(D_k^2 + \lambda I)D_k^{-1}W_k \\ &= A_{k+1}(\tilde{D}_k^2 + \lambda I)\tilde{D}_k \tilde{W}_k \tilde{V}_k^\top D_k^{-1}(D_k^2 + \lambda I)D_k^{-1}W_k \\ &= U_{k+1}D_{k+1}^2 V_{k+1}^\top (\tilde{D}_k^2 + \lambda I)\tilde{W}_k \tilde{V}_k^\top D_k^{-2}(D_k^2 + \lambda I)W_k \end{aligned} \tag{2.1}$$

where the second equality follows from (1.5) and the last follows from (1.3d) after rearranging the diagonal matrices. The definition of P_k then immediately yeilds $XX^\top U_k = U_{k+1}P_{k+1}$ and a similar computation shows $XX^\top \tilde{U}_k = \tilde{U}_{k+1}\tilde{P}_{k+1}$.

The formulas for Q_k and $\tilde{Q}_k$ follow by a simple induction using the formulas for $P_k, \tilde{P}_k$. Note that $Q_k, \tilde{Q}_k$ are products of diagonal matrices (with non-zero diagonal entries), sign matrices and orthogonal matrices and thus are both invertible. Moreover, since $\lambda > 0$ we have the upper bound,

$$||Q_k^{-1}|| \leq \left(\prod_{i=1}^{k-1} \frac{1}{||\tilde{D}_i^2 + \lambda I|| \, ||D_i^2 + \lambda I||} \right) \frac{1}{||\tilde{D}_0^2 + \lambda I||(1+\lambda)} \leq \lambda^{1-2k}$$

and $||\tilde{Q}_k^{-1}|| \leq \frac{||Q_k^{-1}||}{||D_k^2 + \lambda I||} \leq \lambda^{2-2k}$. ◻

In order to connect the iteration (1.3a)-(1.3d) to the singular vectors of X we will use the formulas,

$$(XX^\top)^k U_0 = U_k D_k^2 Q_k, \qquad (X^\top X)^k \tilde{U}_0 = \tilde{U}_k \tilde{D}_k^2 \tilde{Q}_k$$

which follow from Lemma 2.2. Substituting the SVD of $X = USV^\top$ results in,

$$US^{2k}U^\top U_0 = U_k D_k^2 Q_k, \qquad VS^{2k}V^\top \tilde{U}_0 = \tilde{U}_k \tilde{D}_k^2 \tilde{Q}_k$$

and using the invertibility of the $D_k, \tilde{D}_k, Q_k, \tilde{Q}_k$ matrices we have,

$$U^\top U_k = S^{2k} U^\top U_0 D_k^{-2} Q_k^{-1}, \qquad V^\top \tilde{U}_k = S^{2k} V^\top \tilde{U}_0 \tilde{D}_k^{-2} \tilde{Q}_k^{-1}. \tag{2.2}$$

Notice that we have again rearranged the diagonal matrices.

The key to leveraging (2.2) for analyzing the convergence of $U_k, \tilde{U}_k$ is to split the true singular vectors, U, into two groups by choosing an arbitrary $\ell \in \{1, ..., p-1\}$ where $p = \min\{m, n\}$. We then split $U = [U_{(1)} \, U_{(2)}]$ where $U_{(1)}$ contains the first ℓ columns of U, and similarly $V = [V_{(1)} \, V_{(2)}]$ and finally we split the diagonal matrix of singular values as $S = \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix}$ where S_1 is $\ell \times \ell$ and contains the first ℓ singular values.

Theorem 2.3. *Let $X \in \mathbb{R}^{n \times m}$ have SVD $X = USV^\top$ and set $p = \min\{m, n\}$ then, using the notation of Lemma 2.2, for any splitting of the singular vectors $\ell \in \{1, ..., p-1\}$ we have*

$$U_{(1)}^\top U_{k,\ell} = S_1^{2k} U_{(1)}^\top U_{0,\ell} Z_{k,\ell} \qquad U_{(2)}^\top U_{k,\ell} = S_2^{2k} U_{(2)}^\top U_{0,\ell} Z_{k,\ell} \tag{2.3}$$

$$V_{(1)}^\top \tilde{U}_{k,\ell} = S_1^{2k} V_{(1)}^\top \tilde{U}_{0,\ell} \tilde{Z}_{k,\ell} \qquad V_{(2)}^\top \tilde{U}_{k,\ell} = S_2^{2k} V_{(2)}^\top \tilde{U}_{0,\ell} \tilde{Z}_{k,\ell} \tag{2.4}$$

where $U_{k,\ell}, \tilde{U}_{k,\ell}$ are the first ℓ columns of $U_k, \tilde{U}_k$ respectively and $Z_{k,\ell}, \tilde{Z}_{k,\ell}$ are the first ℓ rows of $D_k^{-2} Q_k^{-1}, \tilde{D}_k^{-2} \tilde{Q}_k^{-1}$ respectively. Moreover, as $k \to \infty$, we have

$$\frac{||U_{(2)}^\top U_{k,\ell}||}{||U_{(1)}^\top U_{k,\ell}||} \leq c_\ell \left(\frac{s_{\ell+1}}{s_\ell} \right)^{2k} \to 0 \qquad \frac{||V_{(2)}^\top \tilde{U}_{k,\ell}||}{||V_{(1)}^\top \tilde{U}_{k,\ell}||} \leq \tilde{c}_\ell \left(\frac{s_{\ell+1}}{s_\ell} \right)^{2k} \to 0.$$

Proof. From (2.2) we have,

$$\begin{pmatrix} U_{(1)}^\top \\ U_{(2)}^\top \end{pmatrix} U_k = \begin{pmatrix} S_1^{2k} & 0 \\ 0 & S_2^{2k} \end{pmatrix} \begin{pmatrix} U_{(1)}^\top \\ U_{(2)}^\top \end{pmatrix} U_0 D_k^{-2} Q_k^{-1}$$

which immediately splits into the equations (2.3) and a similar splitting occurs for V which yields (2.4). Next we solve the left equation of (2.3) for $Z_{k,\ell}$ and substitute into the right equation of (2.3) to find,

$$U_{(2)}^\top U_{k,\ell} = S_2^{2k} U_{(2)}^\top U_{0,\ell} (U_{(1)}^\top U_{0,\ell})^{-1} S_1^{-2k} U_{(1)}^\top U_{k,\ell}$$

and obtain the upper bound,

$$||U_{(2)}^\top U_k|| \leq ||S_2^{2k}||\, c_\ell\, ||S_1^{-2k}||\, ||U_{(1)}^\top U_k|| = \left(\frac{s_{\ell+1}}{s_\ell}\right)^{2k} ||U_{(1)}^\top U_k||$$

where the constant c_ℓ is determined by the inner products with $U_{0,\ell}$ and is independent of k. ∎

The power of Theorem 2.3 is that the splitting ℓ was arbitrary. In the generic case of distinct singular values, $\ell = 1$ immediately implies that the first column of U_k becomes orthogonal to the last $p-1$ left singular vectors of X (columns of U) and hence must lie in the space spanned by the first left singular vector of X. Then, $\ell = 2$ implies that the second column of U_k must be orthogonal to the last $p-2$ left singular vectors. Moreover, the definition of U_k via the SVD in (1.3d) implies that the second column of U_k is orthogonal to the first column of U_k and hence must be in the subspace spanned by the second left singular vector of X. Inductively, this shows that the columns of U_k converge to lie in the subspaces spanned by the corresponding columns of U. In the generic case of distinct singular values, this means that the columns of U_k are converging to those of U up to sign. Moreover, in the non-generic case of a repeated singular value, Theorem 2.3 shows the convergence of the corresponding columns of U_k to the subspace spanned by the singular vectors corresponding to the repeated singular value. We can now turn to the convergence of the singular values.

3. Convergence of the Singular Values.

We can combine (1.3a) and (1.3b) into a single equation (and similarly for (1.3c) and (1.3d)),

$$\tilde{U}_k \tilde{W}_k \tilde{S}_k \tilde{W}_k \tilde{V}_k^\top = X^\top U_k W_k S_k (S_k + \lambda I)^{-1} \tag{3.1}$$

$$U_{k+1} W_{k+1} S_{k+1} W_{k+1} V_{k+1}^\top = X \tilde{U}_k \tilde{W}_k \tilde{S}_k (\tilde{S}_k + \lambda I)^{-1} \tag{3.2}$$

where $S_k = D_k^2$ and $\tilde{S}_k = \tilde{D}_k^2$ and the terms on the left-hand-side of (3.1) and (3.2) are defined to be the singular value decomposition of the right-hand-side. Substituting the singular value decomposition of $X = USV^\top$ we have,

$$\tilde{U}_k \tilde{W}_k \tilde{S}_k \tilde{W}_k \tilde{V}_k^\top = VSU^\top U_k W_k S_k (S_k + \lambda I)^{-1} \tag{3.3}$$

$$U_{k+1} W_{k+1} S_{k+1} W_{k+1} V_{k+1}^\top = USV^\top \tilde{U}_k \tilde{W}_k \tilde{S}_k (\tilde{S}_k + \lambda I)^{-1}. \tag{3.4}$$

We first consider the simplified iteration where the singular vectors are set equal to their limits, namely, $U_k = U_{(1:r)}$ and $\tilde{U}_k = V_{(1:r)}$. Since $U_k \to U_{(1:r)}$ and $\tilde{U}_k \to V_{(1:r)}$ we will be able to use a perturbation argument to extend this simplified case to the true $U_k, \tilde{U}_k$ sequences. In the simplified iteration, $U^\top U_k = V^\top \tilde{U}_k = I_{n\times r}$ where $I_{n\times r}$ is an r-by-r identity matrix concatenated with an $(n-r)$-by-r matrix of all zeros. In this case we obtain

$$\tilde{U}_k \tilde{W}_k \tilde{S}_k \tilde{W}_k \tilde{V}_k^\top = VI_{n\times r} W_k S S_k (S_k + \lambda I)^{-1} \tag{3.5}$$

$$U_{k+1} W_{k+1} S_{k+1} W_{k+1} V_{k+1}^\top = UI_{n\times r} \tilde{W}_k S \tilde{S}_k (\tilde{S}_k + \lambda I)^{-1}. \tag{3.6}$$

Note that the left-hand-sides of (3.5) and (3.6) are defined to be the unique SVD of the right-hand-sides. This implies that $\tilde{U}_k = VI_{n\times r}$ and $U_{k+1} = UI_{n\times r}$ and $\tilde{V}_k^\top = V_{k+1}^\top = I_{n\times r}$ which shows that this is a fixed point for the singular vectors. Moreover, we obtain the following iteration for the singular values,

$$\tilde{S}_k = SS_k(S_k + \lambda I)^{-1} \tag{3.7}$$

$$S_{k+1} = S\tilde{S}_k(\tilde{S}_k + \lambda I)^{-1}. \tag{3.8}$$

Since these are all diagonal matrices, we can focus on the fixed point iteration for a single diagonal entry $s_k = (S_k)_{ii}$ and $s = S_{ii}$ we find,

$$s_{k+1} = s^2 \frac{s_k}{s_k + \lambda}\left(\frac{ss_k}{s_k+\lambda} + \lambda\right)^{-1} = \frac{s^2 s_k}{s_k(s+\lambda)+\lambda^2} \tag{3.9}$$

for any $i \in \{1, ..., r\}$.

LEMMA 3.1. *For any $s, \lambda, s_0 \in \mathbb{R}$ with $s \neq \lambda$ the iteration (3.9) converges locally to the softmax function,*

$$s_k \to (s-\lambda)^+ \equiv \max\{0, s-\lambda\},$$

which is the only stable fixed point.

Proof. The fixed points of this iteration are the solutions $\hat{s}$ of $\hat{s} = \frac{s^2\hat{s}}{\hat{s}(s+\lambda)+\lambda^2}$ which implies

$$\hat{s}(\hat{s}(s+\lambda) + \lambda^2 - s^2) = 0$$

so the fixed points are $\hat{s} = 0$ and $\hat{s} = s - \lambda$. Next we analyze the stability of the fixed points by computing the derivative of the iteration,

$$\frac{d}{ds_k}\left(\frac{s^2 s_k}{s_k(s+\lambda)+\lambda^2}\right) = \frac{(s_k(s+\lambda)+\lambda^2)s^2 - s^2 s_k(s+\lambda)}{(s_k(s+\lambda)+\lambda^2)^2}$$

and evaluating at the fixed point $s_k = \hat{s} = 0$ we find

$$\frac{d}{ds_k}\left(\frac{s^2 s_k}{s_k(s+\lambda)+\lambda^2}\right)\bigg|_{s_k=0} = \frac{s^2}{\lambda^2}$$

and at the fixed point $s_k = \hat{s} = s - \lambda$ we find

$$\frac{d}{ds_k}\left(\frac{s^2 s_k}{s_k(s+\lambda)+\lambda^2}\right)\bigg|_{s_k=s-\lambda} = \frac{((s-\lambda)(s+\lambda)+\lambda^2)s^2 - s^2(s-\lambda)(s+\lambda)}{((s-\lambda)(s+\lambda)+\lambda^2)^2} = \frac{\lambda^2}{s^2}.$$

Thus we see that when $s < \lambda$ the fixed point $\hat{s} = 0$ is stable and when $s > \lambda$ the fixed points $\hat{s} = s - \lambda$ is stable. In other words, when $s - \lambda$ is positive the stable fixed point is $s - \lambda$ and when $s - \lambda$ is negative the stable fixed point is zero, thus we see that the iteration converges to the soft-max function,

$$s_k \to \max\{0, s-\lambda\}$$

This completes the proof. ∎

The case $\lambda \neq s$ is generic, however, we note that for the case of $s = \lambda$ we have the simplified iteration $s_{k+1} = \frac{\lambda s_k}{2s_k+\lambda}$ and inductively we have, $s_k = \frac{\lambda s_0}{(2k)s_0+\lambda}$ so unless $s_0 = -\frac{\lambda}{2k}$ for some $k \in \mathbb{N}$, we again have $s_k \to 0 = \max\{0, s-\lambda\}$.

Lemma 3.1 holds for any real $s \neq \lambda$ and any initial condition s_0 including negative numbers. Of course, in our current application, these are all constrained to be non-negative. When any of them are zero the iteration is trivial, so in the next lemma we consider the case when $s, \lambda, s_0 > 0$ and show a stronger convergence property that will be required for the perturbation result.

Lemma 3.2. *For any $s, \lambda, s_0 \in (0, \infty)$, with $s \neq \lambda$, there exists $c \in [0, 1)$ such that*

$$|s_{k+1} - (s-\lambda)^+| \leq c|s_k - (s-\lambda)^+|$$

and the iteration (3.9) converges globally on $(0, \infty)$ to the softmax function, $s_k \to (s-\lambda)^+$.

Proof. Note that $s, \lambda, s_0 > 0$ implies $s_k \geq 0$ for all k by a simple induction.

First consider the case when $\lambda > s$ so that $(s-\lambda)^+ = 0$. Setting $c_1 = \frac{s^2}{\lambda^2} < 1$ we have

$$|s_{k+1} - (s-\lambda)^+| = \frac{s^2 s_k}{s_k(s+\lambda)+\lambda^2} < \frac{s^2}{\lambda^2} s_k = c_1|s_k - (s-\lambda)^+|.$$

Next consider the case where $\lambda < s$ so that $(s-\lambda)^+ = (s-\lambda)$ and

$$(s_{k+1} - (s-\lambda)) = \frac{s^2 s_k - s_k(s^2-\lambda^2) - \lambda^2(s-\lambda)}{s_k(s+\lambda)+\lambda^2} = \frac{\lambda^2}{s_k(s+\lambda)+\lambda^2}(s_k - (s-\lambda)). \tag{3.10}$$

Since $\frac{\lambda^2}{s_k(s+\lambda)+\lambda^2} \leq 1$, (3.10) implies $|s_{k+1} - (s-\lambda)| \leq |s_k - (s-\lambda)|$ and inductively

$$|s_{k+1} - (s-\lambda)| \leq |s_0 - (s-\lambda)|$$

which means that the sequence can never move further away from $s-\lambda$. Moreover, the sequence can never move to the other side of $s-\lambda$, namely, since $\frac{\lambda^2}{s_k(s+\lambda)+\lambda^2} > 0$, if $s_0 \geq s-\lambda$ then (3.10) implies that $s_0 \geq s_k \geq s-\lambda$ for all k, and if $s_0 < s-\lambda$ then $s_0 \leq s_k < s-\lambda$ for all k.

Now if $s_0 < s-\lambda$ then we have $s_k \geq s_0$ for all k and setting $c_2 = \frac{\lambda^2}{s_0(s+\lambda)+\lambda^2} < 1$, (3.10) implies,

$$|s_{k+1} - (s-\lambda)^+| = \frac{\lambda^2|s_k - (s-\lambda)|}{s_k(s+\lambda)+\lambda^2} \leq \frac{\lambda^2|s_k - (s-\lambda)|}{s_0(s+\lambda)+\lambda^2} = c_2|s_k - (s-\lambda)^+|.$$

On the other hand, if $s_0 \geq s-\lambda$ then we have $s_0 \geq s_k \geq s-\lambda$ for all k, and setting $c_3 = \frac{\lambda^2}{s^2} < 1$, (3.10) implies

$$|s_{k+1} - (s-\lambda)^+| = \frac{\lambda^2|s_k - (s-\lambda)|}{s_k(s+\lambda)+\lambda^2} \leq \frac{\lambda^2|s_k - (s-\lambda)|}{(s-\lambda)(s+\lambda)+\lambda^2} = c_3|s_k - (s-\lambda)^+|.$$

So in each case we have $|s_{k+1} - (s-\lambda)^+| \leq c|s_k - (s-\lambda)^+|$ for some $c \in [0, 1)$. ∎

The above lemma establishes a linear convergence rate which is crucial when we consider the perturbed iteration below which will be critical to establishing convergence of the full iteration (3.3) and (3.4). We first establish a general perturbation results for convergent sequences.

Lemma 3.3. *Consider an iteration $x_{k+1} = f(x_k)$ with a fixed point x^* such that for some $c \in [0, 1)$ we have*

$$|f(x) - x^*| < c|x - x^*|$$

for all x. Consider a sequence of perturbations e_k such that for some $a \in [0, 1)$ we have $|e_{k+1}| < a|e_k|$ then the perturbed sequence $w_{k+1} = f(w_k) + e_k$ converges to x^* for any w_0.

Proof. First, since $x^* = f(x^*)$ we have,

$$|w_{k+1} - x^*| = |f(w_k) + e_k - x^*| \leq |f(w_k) - f(x^*)| + |e_k| < c|w_k - x^*| + |e_k|$$

and a simple induction shows that $|w_{k+1} - x^*| < \sum_{i=0}^{k} c^i |e_{k-i}|$. Since $|e_{k+1}| < a|e_k|$ for all k, we have $|e_{k-i}| < a^{k-i}|e_0|$ and thus,

$$|w_{k+1} - x^*| < \sum_{i=0}^{k} c^i |e_{k-i}| < |w_{k+1} - x^*| < |e_0| \sum_{i=0}^{k} c^i a^{k-i} = |e_0| \frac{a^{k+1} - c^{k+1}}{a - c} \to 0$$

since $c, a, \in [0, 1)$, so $w_k \to x^*$. ∎

Note that when applying Lemma 3.3 to the sequence s_k of singular values, the required inequality on f holds only on $(0, \infty)$, however the sequence of perturbations cannot cause the sequence to leave this set since the perturbed sequence is also a sequence of singular values.

3.1. Perturbation of Singular Values. We can now show that as $U_k \to U$, the singular values of (3.3) and (3.4) are a perturbation of the iteration in Lemma 3.1. This perturbed sequence will satisfy the assumptions of Lemma 3.3 and thus will still converge to the softmax, $(s - \lambda)^+$.

Returning to (3.3), when $U_k \neq U$ by Theorem 2.3 we can write $U_k = U + E_k$ where the perturbations E_k decay linearly to zero, $||E_{k+1}|| < a||E_k|| \to 0$ for some $a \in [0, 1)$. We can write (3.3) as

$$\begin{aligned} U_k \tilde{W}_k \tilde{S}_k \tilde{W}_k \tilde{V}_k^\top &= V S U^\top U_k W_k S_k (S_k + \lambda I)^{-1} \\ &= V S U^\top (U + E_k) W_k S_k (S_k + \lambda I)^{-1} \\ &= V S U^\top U W_k S_k (S_k + \lambda I)^{-1} + V S U^\top E_k W_k S_k (S_k + \lambda I)^{-1} \end{aligned}$$

The first term above will be same as right-hand-side of (3.5) and will simplify to give the right-hand-side of (3.7). The second term has bound

$$||V S U^\top E_k W_k S_k (S_k + \lambda I)^{-1}|| \leq ||S|| \, ||E_k|| \, ||S_k (S_k + \lambda I)^{-1}|| < ||S|| \, ||E_k||$$

since $V, U^\top, W_k$ are orthogonal and $S_k(S_k + \lambda I)^{-1}$ is diagonal with diagonal entries less than 1. By Weyl's law for the stability of singular values under perturbation (see for example Theorem 1 of [13]) the singular values $\tilde{s}_k$ on the left-hand-side of (3.5) are given by a perturbation e_k of the right-hand-side (3.7) bounded by $||S||||E_k||$. The iteration for the true singular values becomes,

$$\tilde{s}_k = s s_k (s_k + \lambda)^{-1} + e_k \tag{3.11}$$

$$s_{k+1} = s \tilde{s}_k (\tilde{s}_k + \lambda)^{-1} + \tilde{e}_k. \tag{3.12}$$

where $|e_k| < ||S||\,||E_k||$ and by a similar we find a perturbation argument we have $|\tilde{e}_k| < ||S||\,||\tilde{E}_k||$. Finally, the iteration (3.9) becomes,

$$s_{k+1} = \frac{s^2 s_k(s_k+\lambda)^{-1} + e_k}{s s_k(s_k+\lambda)^{-1} + e_k + \lambda} + \tilde{e}_k = \frac{s^2 s_k + e_k(s_k+\lambda)}{s_k(s+\lambda) + \lambda^2 + e_k(s_k+\lambda)} + \tilde{e}_k$$

$$= \frac{s^2 s_k}{s_k(s+\lambda)+\lambda^2} + \hat{e}_k \tag{3.13}$$

where

$$\hat{e}_k = e_k \frac{(s_k+\lambda)(s_k(s+\lambda-s^2)+\lambda^2)}{(s_k(s+\lambda)+\lambda^2)(s_k(s+\lambda)+\lambda^2+e_k(s_k+\lambda))} + \tilde{e}_k \tag{3.14}$$

Noting that $s_k(s+\lambda-s^2)+\lambda^2 \le s_k(s+\lambda)+\lambda^2$, we can estimate $\hat{e}_k$ as,

$$|\hat{e}_k| \le |e_k| \left| \frac{s_k+\lambda}{s_k(s+\lambda)+\lambda^2+e_k(s_k+\lambda)} \right| + |\tilde{e}_k|$$

Since $e_k \to 0$, for k sufficiently large we have $-\lambda < e_k < \lambda$. We can bound the above denominator by, $s_k(s+\lambda)+\lambda^2+e_k(s_k+\lambda) > s_k(s+\lambda)+\lambda^2-\lambda(s_k+\lambda) = s_k s$. Then,

$$|\hat{e}_k| \le |e_k| \left| \frac{s_k+\lambda}{s_k s} \right| + |\tilde{e}_k| \le c|e_k| + |\tilde{e}_k|$$

since s_k is bounded. Since e_k and $\tilde{e}_k$ have linear convergence, this implies that $\hat{e}_k$ has linear convergence as well. Thus, by Lemma 3.3 the true singular values, $s_k, \tilde{s}_k$ converge to the same limit as the unperturbed singular values, namely the soft max, $(s-\lambda)^+$.

4. Effect of sign matrices on the cost functional.

We can now show that the matrices of right singular vectors $V_k, \tilde{V}_k$ from the SVDs in (1.3b) and (1.3d), converge to diagonal sign matrices when $\lambda < S_{rr}$.

THEOREM 4.1. *Let $X \in \mathbb{R}^{n\times m}$ have SVD $X = USV^\top$. For $\lambda > 0$ let $V_k, \tilde{V}_k$ be the sequence of matrices defined by (1.3b) and (1.3d), then*

$$||\tilde{V}_k - I_{r\times p}((S-\lambda I)^+ + \lambda I)S^{-1} I_{p\times r} W_k||_{\max} \to 0$$

and when W_k converges to a limit W_* then $\tilde{V}_k \to I_{r\times p}((S-\lambda I)^+ + \lambda I)S^{-1} I_{p\times r} W_*$. When $\lambda < S_{rr}$ we have $||\tilde{V}_k - W_k||_{\max} \to 0$ and when $W_k \to W_*$ we have $V_k \to W_*$.

Proof. Substituting (1.3a) in (1.3b) we have,

$$\tilde{U}_k \tilde{W}_k \tilde{D}_k^2 \tilde{W}_k \tilde{V}_k^\top = X^\top U_k W_k D_k (D_k^2 + \lambda I)^{-1} D_k$$

where $X^\top U_k$ is $n \times r$ with $r \le p \equiv \min\{m, n\}$. In order to solve for $\tilde{V}_k^\top$ we multiply both sides by $U_k^\top X$ since $U_k^\top X X^\top U_k = U_k^\top U S^2 U^\top U_k$ is invertible so that,

$$U_k^\top X \tilde{U}_k \tilde{W}_k \tilde{D}_k^2 \tilde{W}_k = U_k^\top U S^2 U^\top U_k W_k D_k (D_k^2 + \lambda I)^{-1} D_k \tilde{V}_k$$

and solving for $\tilde{V}_k$ yields,

$$\tilde{V}_k = D_k^{-2}(D_k^2 + \lambda I) W_k (U_k^\top U S^2 U^\top U_k)^{-1} U_k^\top U S V^\top \tilde{U}_k \tilde{D}_k^2.$$

By Theorem 2.3 we have $U_k^\top U \to I_{r\times p}$ and $V^\top \tilde{U}_k \to I_{p\times r}$ as $k \to \infty$ and as shown in Section 3 we have $D_k \to I_{r\times p} D I_{p\times r} = I_{r\times p}(S-\lambda I)^+ I_{p\times r}$ and also $\tilde{D}_k \to I_{r\times p} D I_{p\times r}$. Substituting these limits into the above equation gives the desired result. Notice that when $\lambda < S_{rr}$ the maximum with zero has no effect and thus $((S-\lambda I)^+ + \lambda I)S^{-1} = I$ so that $||\tilde{V}_k - W_k||_{\max} \to 0$. ∎

A similar argument shows that when $\lambda < S_{rr}$ we have $||V_k - \tilde{W}_k||_{\max} \to 0$ so that both $V_k, \tilde{V}_k$ are converging to diagonal sign matrices. We can now characterize the convergence of Algorithm 1.3.

THEOREM 4.2. *Let $X \in \mathbb{R}^{n\times m}$ have SVD $X = USV^\top$. For $\lambda > 0$, the iteration* (1.3a)-(1.3d) *converges whenever the sign matrices $W_k, \tilde{W}_k$ are chosen so that they converge to limits $W_k \to W_*$ and $\tilde{W}_k \to \tilde{W}_*$. The cost* (1.1) *of the limiting matrices A_*, B_* of the iteration is*

$$||X - A_*B_*^\top||_F = ||S - (S-\lambda I)^+ S^2((S-\lambda I)^+ + \lambda I)^{-2} I_{p\times r}\tilde{W}_* W_* I_{r\times p}||_F$$

and when $\lambda < S_{rr}$ it is

$$||X - A_*B_*^\top||_F = ||S - (S-\lambda I)^+ I_{p\times r}\tilde{W}_* W_* I_{r\times p}||_F$$

and only $\tilde{W}_* W_* = I$ will minimize the cost.

Proof. If we make a convergent choice for the sign matrices $W_k \to W_*$ and $\tilde{W}_k \to \tilde{W}_*$ equation (1.3a) defines a steady state,

$$B_* = X^\top U_* W_* D_* (D_*^2 + \lambda I)^{-1} = VSD(D^2+\lambda I)^{-1} I_{p\times r} W_*$$

where $D_* = I_{r\times p} D I_{p\times r}$ as shown in Section 3. Similarly (1.3c) defines a steady state,

$$A_* = X\tilde{U}_*\tilde{W}_* D_*(D_*^2 + \lambda I)^{-1} = USD(D^2+\lambda I)^{-1} I_{p\times r}\tilde{W}_*.$$

Thus we find the low rank approximation of X to be given by,

$$A_*B_*^\top = US^2D^2(D^2+\lambda I)^{-2} I_{p\times r}\tilde{W}_* W_* I_{r\times p} V^\top$$

and when $\lambda < S_{rr}$ this reduces to

$$A_*B_*^\top = U(S-\lambda I)^+ I_{p\times r}\tilde{W}_* W_* I_{r\times p} V^\top.$$

Notice that when $\tilde{W}_* W_* = I$ this is the optimal solution of (1.1) and (1.2). In the general case, we find the first part of the cost functional is given by,

$$\begin{aligned} ||X - A_*B_*^\top||_F &= ||USV^\top - U(S-\lambda I)^+ I_{p\times r}\tilde{W}_* W_* I_{r\times p} V^\top||_F \\ &= ||S - S^2D^2(D^2+\lambda I)^{-2} I_{p\times r}\tilde{W}_* W_* I_{r\times p}||_F \end{aligned}$$

and when $\lambda < S_{rr}$ we have,

$$||X - A_*B_*^\top||_F = ||S - (S-\lambda)^+ I_{p\times r}\tilde{W}_* W_* I_{r\times p}||_F.$$

Since $\tilde{W}_*$ and W_* are diagonal sign matrices, so is $\tilde{W}_* W_*$ and any negative entries would change the subtraction to addition in the above cost functional, so the solution $A_*B_*^\top$ is optimal only when $\tilde{W}_* W_* = I$. ∎

Finally, since $\tilde{W}_*$ and W_* are both sign matrices, the way to insure $\tilde{W}_* W_* = I$ is to choose $W_* = \tilde{W}_*$. In other words, we need to ensure that the choice of sign matrices in (1.3b) and (1.3d) are the same. Algorithm 1.3 does this by choosing the diagonal entries of $\tilde{W}_k$ to be the signs of the sums of the columns of $\tilde{V}_k$ and similarly for W_k in terms of V_k. Since Theorem 4.1 show that $\tilde{V}_k, V_k$ are converging to diagonal matrices (independent of the choice of $\tilde{W}_k, W_k$) these choices of $\tilde{W}_k, W_k$ will insure that both $\tilde{W}_k\tilde{V}_k^\top$ and W_kV_k converge to the identity matrix. In fact, it does not matter which unique sign choice is made in the SVDs in (1.3b) and (1.3d) as long as the *same* choice is made for both SVDs. Effectively, the choice of sign matrices is how the right singular vectors of (1.3b) and (1.3d) contribute to the iteration in Algorithm 1.3, whereas they are not used at all in Algorithm 1.2.

5. Conclusions and Future Work. In this paper we introduced Algorithm 1.3 as a new rank-restricted soft SVD method and we have proven convergence to the optimal solution of (1.1). We have shown that the standard method, Algorithm 1.2, can fail to converge or can converge to a non-optimal stationary point. Moreover, we have derived the convergence rate of Algorithm 1.3 based on the singular values of the matrix X which shows how Algorithm 1.3 can obtain much faster convergence than the naive alternating directions approach of Algorithm 1. Since Algorithm 1.3 is only one component of the matrix completion method introduced in [9], an important future direction is analyzing the entire matrix completion algorithm. Moreover, the choice of the rank restriction, r, and regularization parameter λ are critical for obtaining the best matrix completion. Investigating methods of selecting these parameters, possibly based on cross-validation, is another critical direction for future research. Finally, while Algorithm 1.3 is of significant interest due to its use in matrix completion problems [9, 11, 4, 5], it could also be used as a partial SVD algorithm and comparison to state-of-the-art SVD methods [6, 7, 10] could yield future insights or improvements.

REFERENCES

[1] H. Antil, D. Chen, and S. Field, *A note on qr-based model reduction: Algorithm, software, and gravitational wave applications*, Computing in Science & Engineering, 20 (2018).

[2] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Now Publishers Inc, 2011.

[3] R. Bro, E. Acar, and T. G. Kolda, *Resolving the sign ambiguity in the singular value decomposition*, Journal of Chemometrics: A Journal of the Chemometrics Society, 22 (2008), pp. 135–140.

[4] J.-F. Cai, E. J. Candès, and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM Journal on optimization, 20 (2010), pp. 1956–1982.

[5] E. J. Candes and Y. Plan, *Matrix completion with noise*, Proceedings of the IEEE, 98 (2010), pp. 925–936.

[6] Z. Drmač and K. Veselić, *New fast and accurate jacobi svd algorithm. i*, SIAM Journal on matrix analysis and applications, 29 (2008), pp. 1322–1342.

[7] Z. Drmač and K. Veselić, *New fast and accurate jacobi svd algorithm. ii*, SIAM Journal on matrix analysis and applications, 29 (2008), pp. 1343–1362.

[8] G. H. Golub and C. F. Van Loan, *Matrix computations*, vol. 3, JHU press, 2013.

[9] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh, *Matrix completion and low-rank svd via fast alternating least squares*, The Journal of Machine Learning Research, 16 (2015), pp. 3367–3402.

[10] M. Kurucz, A. A. Benczúr, and K. Csalogány, *Methods for large scale svd with missing values*, in Proceedings of KDD cup and workshop, vol. 12, Citeseer, 2007, pp. 31–38.

[11] R. Mazumder, T. Hastie, and R. Tibshirani, *Spectral regularization algorithms for learning large incomplete matrices*, The Journal of Machine Learning Research, 11 (2010), pp. 2287–2322.

[12] N. Parikh and S. Boyd, *Proximal algorithms*, Foundations and Trends in optimization, 1 (2014), pp. 127–239.

[13] G. W. Stewart, *Perturbation theory for the singular value decomposition*, tech. report, 1998, http://hdl.handle.net/1903/552.