E0 206: Theorist's Toolkit, Fall 2020 **Dec 8**

Lectures 16-17: Singular Value Decomposition

Instructors: Arindam Khan, Anand Louis *Scribe: Aditya Lonkar, Manish Kesarwani*

Singular Value Decomposition(SVD) is a notion related to decomposing a matrix into constituent matrices so as to gain some summarized valuable information regarding the matrix. Matrices occur almost everywhere in computation, for instance in many computational problems where the input is a graph and is represented as its adjacency matrix. This serves as a a very strong motivation for studying the spectrum of a matrix, the spectrum being its eigenvalues, eigenvectors(in the case of a square matrix), singular values and singular vectors. Please note, this lecture is based on Chapter 3 of Book [1] and Chapter 1 of Book [2].

1 Eigenvalues, Eigenvectors, Singular Values and Singular Vectors

For an $n \times n$ matrix A, x is an eigenvector with eigenvalue λ if $Ax = \lambda x$. To stress an important point here, eigenvalues and eigenvectors are only defined for square matrices. A related notion of singular values and singular vectors for rectangular matrices exists, which we will look at, further in this lecture.

To show that eigenvalues and eigenvectors indeed exist for a matrix A, observe that

$$Ax = \lambda x \text{ iff } (A - \lambda I)x = 0$$

where I is the $n \times n$ identity matrix. Thus, immediately we can see that if $A - \lambda I$ is invertible, the only solution to this equation is the 0-vector. Otherwise if $A - \lambda I$ is non-invertible, there clearly exist solutions for x since A is not a full rank matrix and hence has a non-zero null space. Equivalently,

$$det(A - \lambda I) = 0$$

Since A is an $n \times n$ matrix, the above equation is an n-degree polynomial in λ. Thus it has n complex roots(not necessarily distinct).

1.1 Eigenvalues and Eigenvectors of symmetric matrices

We will focus our attention on symmetric matrices since a lot of this course involves the same. Recall that a matrix A is symmetric iff $A^T = A$ and hence it has to be a square matrix. Graph adjacency matrix(in undirected graphs) is an example of a symmetric matrix. Some useful claims regarding symmetric matrices are stated below.

Claim 1.1 *If A is a symmetric matrix, all its eigenvalues are real.*

Proof: Consider some eigenvector x os A with eigenvalue λ. Assume that λ is not real(In general, x may not be real as well).

$$Ax = \lambda x \tag{1}$$

Taking the complex conjugates,

$$A\bar{x} = \bar{\lambda}\bar{x} \tag{2}$$

Multiply 1 by $\bar{x}^T$ from the left. Similarly Multiply 2 by x^T from the left and subtract these two equations.

$$\bar{x}^T Ax - x^T A\bar{x} = (\lambda - \bar{\lambda})x^T\bar{x}$$

The quantity on the L.H.S is 0 since A is symmetric whereas on the R.H.S the quantity $x^T\bar{x}$ can never be 0 unless x is a 0-vector. Thus, $\lambda = \bar{\lambda}$ and hence, any eigenvalue is real. ■

Claim 1.2 *If A is a symmetric matrix and has eigenvectors x_1 and x_2 with eigenvalues λ_1 and λ_2 respectively and $\lambda_1 \neq \lambda_2$, then $\langle x_1, x_2\rangle = 0$.*

Proof: Assume here that the eigen vectors x_1 and x_2 are scaled so that their norm is 1. Consider the inner product between λx_1 and x_2. That is,

$$\lambda_1\langle x_1, x_2\rangle = \langle Ax_1, x_2\rangle = x_1^T A^T x_2 = x_1^T(\lambda_2 x_2) = \lambda_2\langle x_1, x_2\rangle$$

To explain each of these equalities; first equality follows from the fact that x_1 is an eigen vector with eigenvalue λ_1. Second equality follows from writing inner product in the vector notation. Third equality follows from A being a symmetric matrix and x_2 is an eigenvector of A with eigenvalue λ_2. The final equality follows from the change of vector notation to inner product. Thus, $\lambda_1\langle x_1, x_2\rangle = \lambda_2\langle x_1, x_2\rangle$ implies that $\langle x_1, x_2\rangle = 0$ since the two eigenvalues are unequal. ■

Claim 1.3 *If $\lambda_1 = \lambda_2$ then $c_1x_1 + c_2x_2$ is an eigenvector $\forall\ c_1, c_2 \in \mathbb{R}$.*

The proof is as follows.

$$A(c_1x_1 + c_2x_2) = c_1Ax_1 + c_2Ax_2 = \lambda_1(c_1x_1 + c_2x_2)$$

A general useful claim in proving equality of matrices is as follows.

Claim 1.4 *Two matrices $A, B \in \mathbb{R}^{m\times n}$ are equal iff $Av = Bv$ for all vectors $v \in \mathbb{R}^n$.*

Proof: If $A = B$ then clearly $Av = Bv$ for all $v \in \mathbb{R}^n$. For the reverse direction if $Av = Bv$ for all vectors $v \in \mathbb{R}^n$ then $(A - B)v = 0$ for all vectors of n dimensions. For the sake of contradiction assume that $A - B = C$ is not a 0-matrix. Then at least one row of C has non-zero entries. Consider any such row C_i. If we take $v = C_i$, then the ith element of $(A - B)v = C_i^T v = ||C_i||^2 > 0$. Thus for such a vector $(A - B)v \neq 0$ which is a contradiction to our assumption that $A \neq B$. Thus, the statement follows. ■

Claim 1.5 *$A = V\Lambda V^T$, where columns of V are eigenvectors of A and Λ is a diagonal matrix with Λ_{ii} being the eigenvalue of v_i (ith column of V).*

Proof: Here we will make use of Claim 1.4. Let $V\Lambda V^T = B$ for sake of notation. First we will prove that A and B have the same eigenvectors with same respective eigen values. Consider any eigenvector v_i of A which is the ith eigenvector in terms of its eigenvalue. Then,

$$Av_i = V\Lambda V^T v_i = V\Lambda e_i = V\Lambda_{ii}e_i = \lambda_{ii}v_i$$

Here $e_i \in \mathbb{R}^n$ is the vector whose ith co-ordinate is 1 and all other co-ordinates are 0s. In the above equation, the second equality follows from the fact that all eigenvectors are orthogonal. Now, any vector x can be written as a linear combination of the eigenvectors along with another term $v_\perp$ which is to account for the

component of x orthogonal to the rowspace of A and hence by implication orthogonal to all of its eigenvectors. Then, let $x = (\sum_i c_i v_i) + c_\perp v_\perp$ where c_i s along with $c_\perp$ are appropriate constants. Hence,

$$Ax = A((\sum_i c_i v_i) + c_\perp v_\perp) = \sum_i \lambda_i c_i v_i + 0$$

Similarly,

$$V\Lambda V^T x = V\Lambda V^T(((\sum_i c_i v_i) + c_\perp v_\perp)) = \sum_i \lambda_i c_i v_i + 0$$

The 0 at the end of both of the equations occurs as a result of $v_\perp$ being orthogonal to the rowspace of A as well as the eigenvectors of A. The above two equations imply that $Av = Bv$ for all n-dimensional vectors and hence they are equal. ■

1.2 Singular values and vectors

For a matrix $A \in \mathbb{R}^{m\times n}$, σ is a singular value with corresponding singular vectors $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ if they satisfy the following two conditions,

$$Av = \sigma u \qquad \text{and} \qquad u^T A = \sigma v^T$$

In this case, u is called the left singular vector and v is the corresponding right singular vector of A with singular value σ. Without loss of generality we can assume $||u|| = ||v|| = 1$ since

$$\sigma||u||^2 = u^T\sigma u = u^T Av = \sigma v^T v = \sigma||v||^2$$

These equalities are quite straightforward and from now on, we will always assume that singular vectors are normalized to have unit norm.

1.2.1 Singular values vs Eigenvalues

Claim 1.6 *Right singular vectors of $A \equiv$ Eigenvectors of A^TA.*

Proof: Here we prove the statement for the right singular vectors of A^TA and a similar statement holds for the left singular vectors of A and the eigenvectors of AA^T. The proof for that goes exactly as the proof presented below. For the forward direction, let u and v be left and right singular vectors of A for a singular value σ. Thus,

$$A^TAv = A^T\sigma u = (u^TA)^T\sigma = (\sigma v)^T\sigma = \sigma^2 v$$

For the reverse direction, let v be an eigenvector of A^TA with eigenvalue λ. Then $A^TAv = \lambda v$. First, we prove an important property of the eigenvalues of A^TA which is that they are positive.

$$\lambda||v||^2 = v^T(\lambda v) = v^T(A^TAv) = (Av)^T(Av) = ||Av||^2$$

Notice that then $\lambda = ||Av||^2/||v||^2$ which is the division of two positive quantities and hence, is positive. Now, set $\sigma = \sqrt{\lambda}$ and we can do this since λ is positive(σ is a real number). Set Av/σ which implies $Av = \sigma u$. Now we need to prove the other property of singular vectors u and v for A.

$$u^TA = \Big(\frac{Av}{\sigma}\Big)^T A = \frac{v^TA^TA}{\sigma} = \frac{(A^TAv)^T}{\sigma} = \frac{(\lambda v)^T}{\sigma} = \sigma v^T$$

■

Thus, all the right singular vectors of A are equivalently the the eigenvectors of A^TA with singular values the square roots of the corresponding eigenvalues for A^TA. σ is a singular value of A iff σ^2 is an eigenvalue of A^TA.

1.2.2 Top singular value

Now, we give some characterizations for the singular values of a matrix.

Theorem 1.1

$$v_1' \stackrel{\text{def}}{=} \arg\max_{x \in \mathbb{R}^n} \frac{||Ax||}{||x||}$$

is a singular vector of A, and $\frac{||Av_1'||}{||v_1'||}$ is the largest singular value of A.

Proof: Let $v_1, v_2, .., v_n$ be the orthonormal eigenvectos of A^TA with eigenvalues $\sigma_1^2 \geq ... \geq \sigma_n^2$. This can be assumed because A^TA is a symmetric matrix and the eigenvectors can be assumed to be orthonormal and its eigenvalues are all non-negative. Now, since $v_1, ..., v_n$ form an orthonormal basis, any vector $x \in \mathbb{R}^n$ can be written as a linear combination of the basis. That is, $x = c_1v_1 + c_2v_2 + ... + c_nv_n$. Here, the c_is are basically $\langle x, v_i \rangle$ which can be easily observed as the singular vectors are orthogonal to each other.

$$||Ax||^2 = (Ax)^T(Ax) = x^T(A^TA)x = \Big(\sum_j c_jv_j\Big)^T(A^TA)\Big(\sum_i c_iv_i\Big)$$

$$= \Big(\sum_j c_jv_j\Big)^T\Big(\sum_i c_i\sigma_i^2v_i\Big) = \sum_j\sum_i c_ic_j\sigma_i^2v_iv_j^T = \sum_i c_i^2\sigma_i^2$$

Therefore,

$$\frac{||Ax||}{||x||} = \sqrt{\frac{\sum_i c_i^2\sigma_i^2}{\sum_i c_i^2}} \leq \sigma_1$$

The first equality follows from the last set of equalities and the fact that x is written as a linear combination of orthonormal vectors. The second equality follows because σ_1 is the largest singular value. This just shows an upper bound on the above quantity. But we can indeed achieve this value by setting $c_1 = 1$ and $c_i = 0$ for every other index $i \in [n]$ which is equivalent to setting $x = v_1$. But notice that v_1 which is an eigenvector of A^TA is also a singular vector of A with corresponding singular value σ_1. Therefore, $||Av_1||$ is the largest singular value of A since $\sigma_1^2 = ||Av_1||^2$ is the largest eigenvalue of A^TA. ■

1.2.3 Best fit line

The above characterization of singular vectors and singular values, in particular the largest singular value is useful in solving problems like the one discussed here which is "The Best fit line". The problem is that we are given a set of points $a_1, ..., a_m$ and we have to find the "best fit" line. There can be many notions for the Best fit line. Here we define it as follows:

Find the direction v such that the sum of the squared lengths of the projections of the points on v is maximized.

$$\arg\max_{v:||v||=1} \sum_i \langle a_i, v \rangle^2$$

An equivalent notion of defining the Best fit line is to find the direction v such that the sum of squared length of distances of the points to v is minimized. This actually turns out to be an equivalent definition to the one before it. That is because by *Pythagoras Theorem*,

$$Projection^2 + Distance^2 = Length^2$$

$$\implies \sum_{i \in [m]} Projection_i^2 + \sum_{i \in [m]} Distance_i^2 = \sum_{i \in [m]} a_i^2$$

Thus maximizing the projections is equivalent to minimizing the distances. Now, let A be the matrix with rows $a_1, a_2, ..., a_m$. Then

$$\sum_{i\in[m]} \langle a_i, v\rangle^2 = ||Av||^2$$

This just follows from the definition of inner product. Therefore,

$$\arg\max_{v:||v||=1} \sum_{i\in[m]} \langle a_i, v\rangle^2 = \arg\max_{v:||v||=1} \sum_{i\in[m]} ||Av||^2 = v_1$$

Where v_1 is the singular vector with the highest singular value for A. Thus, the top singular vector gives the best fir line for a set of points.

Now, we define another problem very closely related to the problem of maximizing the quantity $||Ax||/||x||$ which yields the answer(for the argmax) as the top singular vector. Given a matrix A as before, define

$$v_1 = \arg\max_{x\in\mathbb{R}^n} \frac{||Ax||}{||x||} \text{ and } v_i' \stackrel{\text{def}}{=} \arg\max_{x\perp v_1', ..., v_{i-1}'} \frac{||Ax||}{||x||}$$

The theorem statement says that v_k' is the ith singular vector.

Proof: This proof is essentially similar to the proof for the top singular value theorem that we showed before. This proof uses induction on k. Base case is for $k = 1$ which we have already proved before. Suppose then this statement holds for $i \leq k-1$, i.e., $v_i' = v_i$ for $i \in [k-1]$. Fix an $x \perp v_1, ..., v_{k-1}$. Then $x = \sum_i c_i v_i$ where $c_1, ..., c_{k-1} = 0$. This is because x is perpendicular to the first $k-1$ singular vectors. As shown before,

$$||Ax||^2 = \sum_i c_i^2\sigma_i^2 = \sum_{i\geq k} c_i^2\sigma_i^2$$

Therefore,

$$\max_{x\perp v_1, ..., v_{k-1}} \frac{||Ax||}{||x||} = \sqrt{\frac{\sum_{i\geq k} c_i^2\sigma_i^2}{\sum_{i\geq k} c_i^2}} \leq \sigma_k$$

Thus, v_k' is indeed the kth singular vector since it is also orthogonal to the first $k-1$ singular vectors by the induction hypothesis and $||Av_k||$ is the kth largest singular value of A. ■

1.2.4 Best fit subspace

We define now a problem which is a generalization of the Best fit line problem seen in the last part, which is the Best fit subspace problem. Given a set of points A(the matrix A has rows as the points, i.e., $A_i = a_i$ where $a_i \in \mathbb{R}^n$ are the points), compute a k-dimensional subspace V_k' such that the sum of squared lengths of projections of the points on V_k' is maximized. To define the sum of squared lengths of projections on this subspace V_k', let $w_1, ..., w_k$ be an orthonormal bases for V_k'. Sum of squared lengths of projections

$$= \sum_{i\in[m]} \Big(\sum_{j\in[k]} \langle a_i, w_j\rangle^2 \Big) = \sum_{j\in[k]} \Big(\sum_{i\in[m]} \langle a_i, w_j\rangle^2 \Big) = \sum_{j\in[k]} ||Aw_j||^2$$

An important point to note here is that for a given subspace $V_k\prime$, the choice of the orthogonal basis does not matter. The proof is simple and follows from *Pythagoras Theorem*. To sketch the proof, fix any subspace V_k'. Then consider the subspace orthogonal to this space, lets call it X_{n-k} and consider any orthogonal basis of this subspace $x_1, ..., x_{n-k}$. Then for any orthogonal basis $w_1, ..., w_k$ of V_k' and for any given point a_i, the

total sum of the squared lengths of its projections on any n-dimensional space is the square of its norm. Thus the basis $w_1, ..., w_k$ along with $x_1, ..., x_{n-k}$ form a orthonormal basis for the whole vector space.

$$\sum_{j\in[k]} \langle a_i, w_j\rangle^2 + \sum_{l\in[n-k]} \langle a_i, x_l\rangle^2 = ||a_i||^2$$

Clearly the second quantity on the L.H.S of the equation is fixed once the subspace is fixed regardless of the basis for V_k' and the quantity on the R.H.S is a constant. Hence, the choice of basis does not matter for a fixed subspace.

Theorem 1.2 *Given a set of points A, the best fit k-dimensional subspace is given by span of the top k singular vectors(V_k).*

Proof: Proof follows by induction on k. Base case($k = 1$) is just the best fit line case in which the statement is clearly true. By the induction hypothesis, the best fit $k-1$-dimensional subspace is V_{k-1} which is given by the span of top $k-1$ singular vectors. For the sake of contradiction, assume that V_k' is the best fit k-dimensional subspace which is not the span of the top k singular vectors. Let $w_1, ..., w_k$ be an orthonormal basis for V_k' such that $w_k \perp V_{k-1}$. This can always be done because the first subspace has a rank which exceeds the rank of V_{k-1} by 1.

Now, by optimality of V_{k-1}, we have that

$$||Aw_1||^2 + ... + ||Aw_{k-1}||^2 \leq ||Av_1||^2 + ... + ||Av_{k-1}||^2 \tag{1}$$

And

$$v_k = \arg\max_{x\perp V_{k-1}} \frac{||Ax||}{||x||} \tag{2}$$

This implies that since w_k and v_k are unit norm vectors and the fact that w_k is orthogonal to the subspace V_{k-1}, $||Aw_k||^2 \leq ||Av_k||^2$. Thus, summing the two inequalities above,

$$||Aw_1||^2 + ... + ||Aw_{k-1}||^2 + ||Aw_k||^2 \leq ||Av_1||^2 + ... + ||Av_{k-1}||^2 + ||Av_k||^2 \tag{3}$$

Therefore, the subspace V_k is at least as good as V_k' in terms of being the best fit k-dimensional subspace. To show an example where there can be more than best fit subspace, consider again the problem of best fit line which is essentially the best fit subspace problem but for $k = 1$. Here, if the top singular values are equal, we will get two best fit lines which would be the corresponding singular vectors(or it can be a linear combination of the two singular vectors). This idea can be extended to the best fit subspace problem as well. ■

Finally, using these definitions of singular values we can define the *Singular Value Decomposition* for a matrix A as

$$A = \sum_{i\in[r]} \sigma_i u_i v_i^T = U\Sigma V^T$$

Here r is the rank of A, u_i and v_i are the ith left and right singular vectors respectively of A. Further, columns of U are the u_is, columns of V are the v_is and Σ is a diagonal matrix with $\Sigma_{ii} = \sigma_i$ and of course, $i \in [r]$. The proof of this is very similar to the proof of Claim 1.5.

1.3 Norms

Recall that for a vector $x \in \mathbb{R}^n$, we define its Euclidean norm as $\|x\| \stackrel{\text{def}}{=} \left(\sum_i x_i^2\right)^{1/2}$.
So, for an $m \times n$ matrix A, its Frobenius norm is defined as follows (analogous to the Euclidean norm of a

vector):

$$\|A\|_F \stackrel{\text{def}}{=} \left(\sum_{i\in[m]} \sum_{j\in[n]} A_{ij}^2 \right)^{1/2}$$

Further, this is not the only norm associated with matrices, there are many norms. Lets see another norm called the spectral norm of a matrix. Think of a matrix as a function, which takes an n dimensional vector x and maps it to an m dimensional vector Ax. We want to understand what is the change in the length of the vector x once it is acted upon by A. So we look at the ratio of $\|Ax\|$ divided by $\|x\|$. We define the spectral norm of a matrix to be the largest such value over all x. So formally, we define spectral norm as

$$\|A\| \stackrel{\text{def}}{=} \max_{x\in\mathbb{R}^n} \frac{\|Ax\|}{\|x\|}$$

We have already seen till now that spectral norm of a matrix is equal to its top singular value, i.e. $\|A\| = \sigma_1(A)$

Theorem 1.3 *Frobenius norm also has a close relationship with the singular value, and is given by*

$$\|A\|_F^2 = \sum_j \sigma_j^2$$

Proof: Let A be the matrix and let a_i denotes the rows of the matrix, then $\|A\|_F^2 = \sum_{i\in[m]} \|a_i\|^2$. Since the right singular vectors, v_js form an orthonormal basis for the row-space of A, hence $\|a_i\|^2 = \sum_j \langle a_i, v_j\rangle^2$. Therefore,

$$\|A\|_F^2 = \sum_{i\in[m]} \|a_i\|^2 = \sum_{i\in[m]} \sum_j \langle a_i, v_j\rangle^2 = \sum_j \sum_{i\in[m]} \langle a_i, v_j\rangle^2 = \sum_j \|Av_j\|^2 = \sum_j \sigma_j^2$$

■

1.4 Low Rank Matrix Approximation

Given a matrix A and a number k the goal is to compute a rank k matrix D that minimizes $\|A - D\|_F^2$

Theorem 1.4 *Best rank k approximation is given by the top k singular vectors, i.e. $A_k \stackrel{\text{def}}{=} \sum_{i\in[k]} \sigma_i u_i v_i^T$. Moreover,*

$$\left\| A - \sum_{i\in[k]} \sigma_i u_i v_i^T \right\|_F^2 = \sum_{i>k} \sigma_i^2$$

Proof: Let D be the optimal rank k matrix. Then,

$$\|A - D\|_F^2 = \sum_{i\in[m]} \|A_i - D_i\|^2$$

In the above equation, A_i and D_i are the i^{th} rows of A and D respectively. We may assume without loss of generality that D_i is the projection of A_i on a rank k subspace (If this is not the case then we can show

that D is a sub-optimal matrix). Therefore, $\|A_i - D_i\|^2 = \|A_i\|^2 - \|D_i\|^2$.
Now we can write the error as,

$$\sum_{i\in[m]} \|A_i - D_i\|^2 = \sum_{i\in[m]} (\|A_i\|^2 - \|D_i\|^2) = \|A\|_F^2 - \sum_{i\in[m]} \|D_i\|^2$$

Therefore, to minimize the error, our goal is to maximize $\sum_{i\in[m]} \|D_i\|^2$. We have already seen earlier that, if we want to maximize the projection on a rank k subspace, then the optimal subspace is given by top k singular vectors, i.e. $D_i = \sum_{j\in[k]} (A_i v_j) \cdot v_j^T = A_i \sum_{j\in[k]} v_j v_j^T$

$$D = A\sum_{j\in[k]} v_j v_j^T = \sum_i \sigma_i u_i v_i^T \sum_{j\in[k]} v_j v_j^T = \sum_i \sum_{j\in[k]} \sigma_i u_i v_i^T v_j v_j^T = \sum_{i\in[k]} \sigma_i u_i v_i^T$$

■

1.4.1 Spectral Norm Approximation

Given a matrix A and a number k, find the best rank k matrix D that minimizes $\|A - D\|_2^2$.

Theorem 1.5 *A_k gives the best rank k approximation, and $\|A - A_k\|_2^2 = \sigma_{k+1}^2$*

Proof: Let D be any optimal matrix. Let z be a unit vector in the span$\{v_1, \cdots v_{k+1}\}$ such that $z \perp$ span D. Write $z = \sum_{i\in[k+1]} c_i v_i$

$$\|A - D\|_2^2 \geq \frac{\|(A-D)z\|^2}{\|z\|^2} = \frac{\|Az\|^2}{\|z\|^2} = \frac{z^T(A^TA)z}{\|z\|^2} = \frac{\sum_{i\in[k+1]} c_i^2\sigma_i^2}{\sum_{i\in[k+1]} c_i^2} \geq \sigma_{k+1}^2$$

Therefore, the matrix A_k gives the best rank k approximation in the spectral norm approximation also. ■

There are several application of Low Rank Approximation. The error is sufficiently small depending on the application. We can work with the low rank approximation of the matrix instead of the original matrix, since the low rank approximation would be smaller in general, the algorithms working on the approximate matrix might work faster and moreover the low rank matrix might require less space to store it in the memory.

1.5 Power Iteration

Power iteration is one of the simplest method to compute eigenvalues, eigenvectors, singular values and singular vectors of a matrix.
Given any matrix M, computing its singular values and singular vectors is equivalent to computing the eigenvalues and eigenvectors of the matrix M^TM. So, let A be a symmetric matrix with eigenvalues $\sigma_1 \geq \cdots \sigma_n \geq 0$ and $v_1, \cdots, v_n$ are the corresponding eigenvectors. Here, just for simplification, the eigenvalues are considered non-negative, but the analysis in this section can be extended for the case when the eigenvalues are negative. Now, given this we want to compute the eigenvalues, because eigenvalues and eigenvectors can be irrational; therefore cannot be computed exactly in general.
So, our goal is that given A and an error parameter ϵ, compute an "approximate" top eigenvector.
Let x_0 be a "random" unit vector. Write $x_0 = c_1v_1 + \cdots + c_nv_n$. Then,

$$\frac{Ax_0}{\|Ax_0\|} = \frac{c_1Av_1 + \cdots + c_nAv_n}{\|Ax\|} = \frac{c_1\sigma_1v_1 + \cdots + c_n\sigma_nv_n}{\sqrt{c_1^2\sigma_1^2 + \cdots + c_n^2\sigma_n^2}}$$

Notice that each of the component along $v_i s$ are getting scaled, for example the component along v_1 is scaled by σ_1 after appropriate normalization, similarly the component along v_n is scaled by σ_n. So the component along v_1 is getting scaled by the largest value and all other components are getting scaled by smaller values. Therefore after normalization in a relative sense, the component along v_1 is increasing whereas the component along the other eigenvectors is decreasing. So, formally,

$$\frac{A^k x}{\|A^k x\|} = \frac{c_1\sigma_1^k v_1 + c_2\sigma_2^k v_2 + \cdots + c_n\sigma_n^k v_n}{\sqrt{c_1^2\sigma_1^{2k} + \cdots + c_n^2\sigma_n^{2k}}}$$

Define $x_k \stackrel{\text{def}}{=} A^k x_0/\|A^k x_0\|$. Now, if we look at the limit as $k \to \infty$, then $x_k \to v_1$, because the coefficient of v_1 is increasing at a much faster rate than the coefficients of the other eigenvectors. So, observe that if we take any random vector and keep multiplying it by A, we will eventually get to v_1. Also if there is a large gap between σ_1 and σ_2, i.e. $\sigma_2 \ll \sigma_1$, then we can show "fast" convergence.
But in general, there need not be a large gap between the top two singular vectors. Then probably we cannot guarantee that we will converge to v_1, but we can observe following: Given an error parameter ϵ, let p be the index such that $\sigma_p \geq (1-\epsilon)\sigma_1 > \sigma_{p+1}$, i.e. p is the number of eigenvalues which are greater than or equal to $(1-\epsilon)\sigma_1$. Let V_p be the subspace spanned by $v_1, \cdots, v_p$. Compute a unit vector x whose projection on V_p is at least $1-\epsilon$. Informally, we don't want the vector that we are computing to be close to v_1, all we want is that it should have a large projection on the eigenvectors corresponding to the eigenvalues which are close to σ_1.
As we have seen before, $\|A^k x_0\|^2 = \sum_i \sigma_i^{2k} c_i^2 \geq \sigma_1^{2k} c_1^2$
So, now lets try to upper bound the component orthogonal to v_p,

$$\sum_{i \geq p+1} \sigma_i^{2k} c_i^2 \leq (1-\epsilon)^{2k}\sigma_1^{2k} \sum_{i \geq p+1} c_i^2 \leq (1-\epsilon)^{2k}\sigma_1^{2k}$$

Therefore, the component of x_k orthogonal to V_p has squared length

$$\frac{\sum_{i \geq p+1} \sigma_i^{2k} c_i^2}{\|A^k x_0\|^2} \leq \frac{(1-\epsilon)^{2k}\sigma_1^{2k}}{\sigma_1^{2k} c_1^2} \leq \frac{e^{-2k\epsilon}}{c_1^2}$$

Taking $k \geq \frac{1}{2\epsilon}\left(\ln \frac{1}{c_1^2 \epsilon}\right)$ suffices to ensure that $\frac{e^{-2k\epsilon}}{c_1^2} \leq \epsilon$
Taking a "random" unit vector will ensure that with high probability $c_1 = \Omega\left(\frac{1}{\sqrt{n}}\right)$ (Proof in Lemma 3.12 of [1]). Therefore, given this bound, taking $k = \Theta\left(\frac{1}{\epsilon}\left(\log \frac{n}{\epsilon}\right)\right)$ will suffice.

References

[1] Avrim Blum, John Hopcroft, and Ravindran Kannan. Foundations of data science. *Vorabversion eines Lehrbuchs*, 5, 2016.

[2] Ravindran Kannan and Santosh Vempala. *Spectral algorithms*. Now Publishers Inc, 2009.