Definition 1: Eigendecomposition & Diagonalization

Suppose an n by n matrix A has n linearly independent eigenvectors $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$. Put them into the columns of the **eigenvector matrix** $\mathbf{X}$. Then $\mathbf{\Lambda} \equiv \mathbf{X}^{-1}\mathbf{A}\mathbf{X}$ is the **eigenvalue matrix**. A has been **diagonalized**.

$$\mathbf{\Lambda} \equiv \mathbf{X}^{-1}\mathbf{A}\mathbf{X} \qquad \mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{\Lambda} \qquad \mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$$

$\mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$ is called the **eigendecomposition** of A. But what does the eigenvalue matrix $\mathbf{\Lambda}$ look like? Why do we say A has been "diagonalized"?

Lemma 1: Eigenvalue Matrix Is Diagonal With Eigenvalues In Diagonal

Suppose an n by n matrix A has n linearly independent eigenvectors $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$. Then the eigenvalue matrix $\mathbf{\Lambda} \equiv \mathbf{X}^{-1}\mathbf{A}\mathbf{X}$ is a diagonal matrix with the eigenvalues in the diagonal.

Proof:
AX and $X\Lambda$ are both equal to $\begin{bmatrix}\lambda_1\mathbf{x}_1 & \dots & \lambda_n\mathbf{x}_n\end{bmatrix}$:

$$AX = \begin{bmatrix}\mathbf{a}_1 \\ \vdots \\ \mathbf{a}_n\end{bmatrix}\begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_n\end{bmatrix} = \begin{bmatrix}\mathbf{a}_1\mathbf{x}_1 & \mathbf{a}_1\mathbf{x}_2 & \dots & \mathbf{a}_1\mathbf{x}_n \\ \mathbf{a}_2\mathbf{x}_1 & \mathbf{a}_2\mathbf{x}_2 & \dots & \mathbf{a}_2\mathbf{x}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_n\mathbf{x}_1 & \mathbf{a}_n\mathbf{x}_2 & \dots & \mathbf{a}_n\mathbf{x}_n\end{bmatrix}$$

$$= \begin{bmatrix}A\mathbf{x}_1 & \dots & A\mathbf{x}_n\end{bmatrix} = \begin{bmatrix}\lambda_1\mathbf{x}_1 & \dots & \lambda_n\mathbf{x}_n\end{bmatrix}$$

$$X\Lambda = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_n\end{bmatrix}\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n\end{bmatrix} = \begin{bmatrix}x_{1,1} & x_{2,1} & \dots & x_{n,1} \\ x_{1,2} & x_{2,2} & \dots & x_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,n} & x_{2,n} & \dots & x_{n,n}\end{bmatrix}\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n\end{bmatrix}$$

$$= \begin{bmatrix}x_{1,1}\lambda_1 & x_{2,1}\lambda_2 & \dots & x_{n,1}\lambda_n \\ x_{1,2}\lambda_1 & x_{2,2}\lambda_2 & \dots & x_{n,2}\lambda_n \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,n}\lambda_1 & x_{2,n}\lambda_2 & \dots & x_{n,n}\lambda_n\end{bmatrix} = \begin{bmatrix}\lambda_1\mathbf{x}_1 & \dots & \lambda_n\mathbf{x}_n\end{bmatrix}$$

The eigenvectors are linearly independent, so X is invertible and we have

$$\mathbf{X}^{-1}\mathbf{A}\mathbf{X} = \mathbf{\Lambda} \qquad \mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{\Lambda} \qquad \mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$$

■

Let's try an example:

$$A = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \to 0 = \begin{vmatrix} 1-\lambda & 2 \\ 4 & 3-\lambda \end{vmatrix} = (1-\lambda)(3-\lambda) - 8 = \lambda^2 - 4\lambda + 3 - 8$$

$$= \lambda^2 - 4\lambda - 5 = (\lambda+1)(\lambda-5) \implies \lambda_1 = -1 \quad \lambda_2 = 5$$

Let's check: $|A| = 1 \cdot 3 - 2 \cdot 4 = -5 = \lambda_1 \cdot \lambda_2 \qquad tr(A) = 4 = \lambda_1 + \lambda_2$

$$\lambda_1 = -1 \to \begin{bmatrix} 1-\lambda_1 & 2 \\ 4 & 3-\lambda_1 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 4 & 4 \end{bmatrix} \to \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \to \mathbf{x_1} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$A\mathbf{x_1} = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1+2 \\ -4+3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \lambda_1 \mathbf{x_1}$$

$$\lambda_2 = 5 \to \begin{bmatrix} 1-\lambda_2 & 2 \\ 4 & 3-\lambda_2 \end{bmatrix} = \begin{bmatrix} -4 & 2 \\ 4 & -2 \end{bmatrix} \to \begin{bmatrix} 1 & -\frac{1}{2} \\ 0 & 0 \end{bmatrix} \to \mathbf{x_2} = \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}$$

$$A\mathbf{x_2} = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}+2 \\ 2+3 \end{bmatrix} = \begin{bmatrix} 2\frac{1}{2} \\ 5 \end{bmatrix} = \lambda_2 \mathbf{x_2}$$

Let's check that our eigenvectors are independent:

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 & \frac{1}{2} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$

$$\begin{bmatrix} -1 & \frac{1}{2} \\ 1 & 1 \end{bmatrix} \to \begin{bmatrix} -1 & \frac{1}{2} \\ 0 & \frac{3}{2} \end{bmatrix} \to \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \to \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$\implies$ No free variables and nullspace $= \{\mathbf{0}\}$

And let's calculate X^{-1}:

$$X^{-1} = \frac{1}{|X|} adj(X) = \frac{1}{-1-\frac{1}{2}} \begin{bmatrix} 1 & -\frac{1}{2} \\ -1 & -1 \end{bmatrix} = -\frac{2}{3} \begin{bmatrix} 1 & -\frac{1}{2} \\ -1 & -1 \end{bmatrix}$$

And check the inverse calculation:

$$X^{-1}X = -\frac{2}{3} \begin{bmatrix} 1 & -\frac{1}{2} \\ -1 & -1 \end{bmatrix} \begin{bmatrix} -1 & \frac{1}{2} \\ 1 & 1 \end{bmatrix} = -\frac{2}{3} \begin{bmatrix} -1-\frac{1}{2} & \frac{1}{2}-\frac{1}{2} \\ 1-1 & -\frac{1}{2}-1 \end{bmatrix} = -\frac{2}{3} \begin{bmatrix} -\frac{3}{2} & 0 \\ 0 & -\frac{3}{2} \end{bmatrix} = I$$

And, finally, let's diagonlize A:

$$X\Lambda X^{-1} = \begin{bmatrix} -1 & \frac{1}{2} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \Big(-\frac{2}{3}\Big) \begin{bmatrix} 1 & -\frac{1}{2} \\ -1 & -1 \end{bmatrix}$$

$$= \Big(-\frac{2}{3}\Big) \begin{bmatrix} -1 & \frac{1}{2} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} \\ -1 & -1 \end{bmatrix}$$

$$= \Big(-\frac{2}{3}\Big) \begin{bmatrix} -1 & \frac{1}{2} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & \frac{1}{2} \\ -5 & -5 \end{bmatrix} = \Big(-\frac{2}{3}\Big) \begin{bmatrix} 1-\frac{5}{2} & -\frac{1}{2}-\frac{5}{2} \\ -1-5 & \frac{1}{2}-5 \end{bmatrix}$$

$$= \Big(-\frac{2}{3}\Big) \begin{bmatrix} -\frac{3}{2} & -\frac{6}{2} \\ -6 & -\frac{9}{2} \end{bmatrix} = \frac{2}{3} \begin{bmatrix} \frac{3}{2} & \frac{6}{2} \\ 6 & \frac{9}{2} \end{bmatrix} = A$$

So we can't just diagonalize any ol' square matrix? How do we know which matrices have independent eigenvectors? Please don't tell me that we have to compute the eigenvectors to see if they're independent.

Lemma 2: Independent Eigenvectors From Distinct Eigenvalues

The eigenvectors from distinct eigenvalues are linearly independent.

Proof:
Suppose $\sum_{i=1}^{n} c_i \mathbf{x_i} = \mathbf{0}$. Multiply by A and λ_n to get two equations:

$$\mathbf{0} = A\sum_{i=1}^{n} c_i \mathbf{x}_i = \sum_{i=1}^{n} c_i A\mathbf{x}_i = \sum_{i=1}^{n} c_i \lambda_i \mathbf{x}_i$$

$$\mathbf{0} = \lambda_n \sum_{i=1}^{n} c_i \mathbf{x}_i = \sum_{i=1}^{n} c_i \lambda_n \mathbf{x}_i$$

Subtract those two equations so that the coefficient of $\mathbf{x}_n$ is $\lambda_n - \lambda_n$ and $\mathbf{x}_n$ is gone. Now multiply by A and λ_{n-1} and subtract. $\mathbf{x}_{n-1}$ is gone.

$$\text{We reach } \; (\lambda_1 - \lambda_2)...(\lambda_1 - \lambda_n)c_1\mathbf{x_1} = \mathbf{0} \implies c_1 = 0$$

Similarly, $c_i = 0$ for all i's and the eigenvectors are independent. ■

Corollary 1: Diagonalizable Matrix

An n by n matrix with n distinct eigenvalues must be diagonalizable.

Symmetric Matrices

Symmetric matrices ($A^T = A$) are very important. And they have some very useful properties. And their ability to diagonalize is extraordinary. At first glance, their symmetry reveals:

$$X\Lambda X^{-1} = A = A^T = (X\Lambda X^{-1})^T = (X^{-1})^T \Lambda X^T$$

Does this suggest that $X^T = X^{-1}$? If so, then $X^T X = I$ and each eigenvector in X is orthogonal to the other eigenvectors. Even better, we could write $A = Q\Lambda Q^T$. It is much more efficient to compute a transpose than an inverse; so this is a big deal by itself. But we will see much greater benefits as well.

Lemma 3: Properties Of Eigenvalues And Eigenvectors For Symmetric Matrix

For a real, symmetric matrix:

(1) The eigenvalues are real

(2) The eigenvectors corresponding to distinct eigenvalues are orthogonal

Proof:

For (1), suppose there exists a complex λ such that $S\mathbf{x} = \lambda\mathbf{x}$. Then

$$S\mathbf{x} \overset{\phi}{=} \lambda\mathbf{x} \implies S\overline{\mathbf{x}} = \overline{\lambda}\overline{\mathbf{x}} \implies \mathbf{x}^H S = (S\overline{\mathbf{x}})^T \overset{\psi}{=} (\overline{\lambda}\overline{\mathbf{x}})^T = \mathbf{x}^H\overline{\lambda}$$

Now multiply ϕ on the left by $\mathbf{x}^H$ and multiply ψ on the right by $\mathbf{x}$:

$$\mathbf{x}^H S\mathbf{x} = \mathbf{x}^H\lambda\mathbf{x} \qquad \mathbf{x}^H S\mathbf{x} = \mathbf{x}^H\overline{\lambda}\mathbf{x}$$

So $\mathbf{x}^H\lambda\mathbf{x} = \mathbf{x}^H\overline{\lambda}\mathbf{x}$. Hence $\lambda = \overline{\lambda}$ and λ is real.

For (2), let $S\mathbf{x} = \lambda_1\mathbf{x}$ and $S\mathbf{y} = \lambda_2\mathbf{y}$ where $\lambda_1 \neq \lambda_2$.

$$\mathbf{x}^T\lambda_1\mathbf{y} = (\lambda_1\mathbf{x})^T\mathbf{y} = (S\mathbf{x})^T\mathbf{y} = \mathbf{x}^T S^T\mathbf{y} = \mathbf{x}^T S\mathbf{y} = \mathbf{x}^T\lambda_2\mathbf{y}$$

So $\mathbf{x}^T\mathbf{y} = 0$. ■

Corollary 2: Limited Spectral Theorem

Given a real, symmetric matrix S with distinct eigenvalues:

(1) Its eigenvector matrix Q is orthogonal and invertible and $Q^{-1} = Q^T$.

(2) It's diagonalizable: $S = Q\Lambda Q^{-1} = Q\Lambda Q^T = \sum_{i=1}^n \lambda_i \mathbf{q}_i\mathbf{q}_i^T$

For (1), since its eigenvectors are orthogonal, they are linearly independent and Q is invertible. We can normalize the eigenvectors so that Q has orthonormal columns and $Q^T = Q^{-1}$.

For (2), we assumed distinct eigenvalues, so Corollary 1 ensures it is diagonalizable with $S = Q\Lambda Q^{-1}$. Using (1), we have $Q^{-1} = Q^T$ so that $S = Q\Lambda Q^{-1} = Q\Lambda Q^T$. ■

You might be thinking... OK, that's all great. $Q^T = Q^{-1}$ makes the diagonalization more efficient to compute. Whoop-dee-doo. Didn't you say that symmetric matrices have some extraordinary ability to diagonalize? We already knew from Corollary 1 that any matrix with distinct eigenvalues is diagonalizable. Is the "distinct eigenvalues" disclaimer superfluous for symmetric matrices? Are you suggesting that any symmetric matrix is diagonalizable?

Spectral Theorem

Every symmetric matrix has the factorization $S = Q\Lambda Q^T$ with real eigenvalues in Λ and orthonormal eigenvectors in the columns of Q.

$$S = Q\Lambda Q^T = Q\Lambda Q^{-1} = \sum_{i=1}^{n} \lambda_i \mathbf{q}_i \mathbf{q}_i^T \text{ and } Q^{-1} = Q^T$$

Proof:
Many proofs exist. Inductive Proof. Another Proof. It is also possible to prove that you can get an orthonormal set of eigenvectors for any symmetric matrix, even for the case of repeated eigenvalues (basically apply Gram-Schmidt to the eigenspaces of repeated eigenvalues). So that's another proof. Later I will present another proof using the Rayleigh Quotient.

$\mathbf{A^TA}$ and $\mathbf{AA^T}$ Properties

Let A be a real m by n matrix of rank r. Then A^TA and AA^T

1. are symmetric
2. are positive semi-definite (and positive definite if A is square and non-singular)
3. share the same nullspace as A and A^T, respectively.
4. share rank r.
5. share the same determinant (=$|A|^2$ if A is square, haha)
6. share the same trace
7. share the same eigenvalues
8. share related eigenvectors: $\mathbf{v} = A^T\mathbf{u}$ and $\mathbf{u} = A\mathbf{v}$
9. are diagonalizable with orthonormal eigenvectors, $Q^{-1} = Q^T$, and $Q\Lambda Q^T = \sum_{i=1}^{n} \lambda_i \mathbf{q}_i \mathbf{q}_i^T$
10. have r positive eigenvalues

Proof:
To show (1):

$$(A^TA)^T = A^T(A^T)^T = A^TA$$

For (2), we already showed symmetry. We must also show non-negative eigenvalues. It is sufficient to show that $\mathbf{x}^TA^TA\mathbf{x} \geq 0$ for any vector $\mathbf{x}$ (the energy-based definition):

$$\mathbf{x}^TA^TA\mathbf{x} = (A\mathbf{x})^TA\mathbf{x} = \|A\mathbf{x}\|^2 \geq 0$$

$$\mathbf{y}^TAA^T\mathbf{y} = (A^T\mathbf{y})^TA^T\mathbf{y} = \left\|A^T\mathbf{y}\right\|^2 \geq 0$$

In particular, if A is non-singular, then A^TA and AA^T are positive definite.

For (3), suppose $\mathbf{x} \in N(A)$ so that $A\mathbf{x} = \mathbf{0}$. Then $A^TA\mathbf{x} = A^T\mathbf{0} = \mathbf{0}$. So $\mathbf{x} \in N(A^TA)$. Conversely, suppose that $\mathbf{x} \in N(A^TA)$ so that $A^TA\mathbf{x} = \mathbf{0}$. Left multiplying by $\mathbf{x}^T$, we get

$$0 = \mathbf{x}^T\mathbf{0} = \mathbf{x}^TA^TA\mathbf{x} = (A\mathbf{x})^T(A\mathbf{x}) = \|A\mathbf{x}\|^2$$

Hence $A\mathbf{x} = \mathbf{0}$ and $\mathbf{x} \in N(A)$.

We have shown that $N(A) \subset N(A^TA)$ and $N(A^TA) \subset N(A)$. Hence $N(A^TA) = N(A)$.

For AA^T, suppose $\mathbf{y} \in N(A^T)$ so that $A^T\mathbf{y} = \mathbf{0}$. Then $AA^T\mathbf{y} = A\mathbf{0} = \mathbf{0}$. So $\mathbf{y} \in N(AA^T)$. Conversely, suppose that $\mathbf{y} \in N(AA^T)$ so that $AA^T\mathbf{y} = \mathbf{0}$. Left multiplying by $\mathbf{y}^T$, we get

$$0 = \mathbf{y}^T\mathbf{0} = \mathbf{y}^TAA^T\mathbf{y} = (A^T\mathbf{y})^T(A^T\mathbf{y}) = \left\|A^T\mathbf{y}\right\|^2$$

Hence $A^T\mathbf{y} = \mathbf{0}$ and $\mathbf{y} \in N(A^T)$.

We have shown that $N(A^T) \subset N(AA^T)$ and $N(AA^T) \subset N(A^T)$. Hence $N(AA^T) = N(A^T)$.

To prove (4), (3) $\implies$ $dim(N(A^TA)) = dim(N(A)) = n - r$. Hence $dim(C(A^TA)) = n - (n - r) = r$ and the rank of A^TA is r. Similarly, we can show that r is the rank of AA^T.

For (5):

$$|A^TA| = |A^T||A| = |A||A^T| = |AA^T|$$

And if A is square: $|A^TA| = |AA^T| = |A^T||A| = |A||A| = |A|^2$

For (7) and (8), let λ be an eigenvalue of AA^T with eigenvector $\mathbf{u}$. And let $\mathbf{v} = A^T\mathbf{u}$. Then

$$A\mathbf{v} = AA^T\mathbf{u} = \lambda\mathbf{u}$$

$$\implies A^TA\mathbf{v} = A^T\lambda\mathbf{u} = \lambda A^T\mathbf{u} = \lambda\mathbf{v}$$

Hence λ is an eigenvalue of A^TA with eigenvector $\mathbf{v} = A^T\mathbf{u}$. On the other hand, let λ be an eigenvalue of A^TA with eigenvector $\mathbf{v}$. And let $\mathbf{u} = A\mathbf{v}$. Then

$$A^T\mathbf{u} = A^TA\mathbf{v} = \lambda\mathbf{v}$$

$$\implies AA^T\mathbf{u} = A\lambda\mathbf{v} = \lambda A\mathbf{v} = \lambda\mathbf{u}$$

Hence λ is an eigenvalue of AA^T with eigenvector $\mathbf{u} = A\mathbf{v}$.

(9) follows directly from the Spectral Theorem.

For (10), non-negativity of eigenvalues follows from positive semi-definiteness proven in (2). But now we want to know the number of positive eigenvalues vs. the number of zero eigenvalues. (9) shows A^TA is diagonalizable. So, for every eigenvalue of A^TA, the geometric multiplicity equals the algebraic multiplicity. So n minus the dimension of the eigenspace of zero will equal the number of positive eigenvalues. In (3), we showed that A^TA has the same nullspace as A. But the nullspace of A^TA is just the eigenspace of zero for A^TA. So

Number of positive eigenvalues for $A^TA = n -$ dimension of the eigenspace of zero for A^TA

$$= n - dim(N(A^TA)) = n - dim(N(A))$$

$$= n - (n - r) = r$$

A similar argument works for AA^T. ■

Theorem: Singular Value Decomposition

Let A be an m by n matrix of rank r. There exists a diagonlization $A = U\Sigma V^T$ such that:

1. The first r columns of U form an orthonormal basis for the column space of A.
2. The first r columns of V form an orthonormal basis for the row space of A.
3. The first r elements of the diagonal of Σ are square roots of eigenvalues of A^TA (and equivalently AA^T).
4. The remaining $m - r$ columns of U form an orthonormal basis for the left nullspace of A.
5. The remaining $n - r$ columns of V form an orthonormal basis for the nullspace of A.
6. $A = U\Sigma V^T$ separates A into r rank 1 matrices $\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$.
7. $A^TA = \sum_{i=1}^{r} \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^T$ and the right singular vectors of A are eigenvectors of A^TA with eigenvalues $\sigma_i^2 = \lambda_i$.
8. $AA^T = \sum_{i=1}^{r} \sigma_i^2 \mathbf{u}_i \mathbf{u}_i^T$ and the left singular vectors of A are eigenvectors of AA^T with eigenvalues $\sigma_i^2 = \lambda_i$.
9. The columns of U form an orthonormal basis for $\mathbb{R}^m$.
10. The columns of V form an orthonormal basis for $\mathbb{R}^n$.

There are at least a couple of ways to prove the SVD Theorem. I present two proofs that I find to be very instructive. "All at once" and "one at a time".

SVD Proof: All At Once

Suppose $r \leq n \leq m$. The proof for $m < n$ is similar.

We choose the $\mathbf{v}_i$'s (AKA the right singular vectors) to be orthonormal eigenvectors of A^TA with corresponding eigenvalues λ_i's. **$\mathbf{A^TA}$ and $\mathbf{AA^T}$ Properties (9)** guarantees we can get n such pairs of eigenvalues and orthonormal eigenvectors. We know from **$\mathbf{A^TA}$ and $\mathbf{AA^T}$ Properties (10)** that there are r positive eigenvalues and $n - r$ zero eigenvalues. Index the eigenvalues and eigenvectors so that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0$ and $\lambda_{r+1} = ... = \lambda_n = 0$. Define the singular values σ_i as

$$\sigma_i \equiv \|A\mathbf{v}_i\| = \sqrt{(A\mathbf{v}_i)^T A\mathbf{v}_i} = \sqrt{\mathbf{v}_i^T A^T A\mathbf{v}_i} = \sqrt{\mathbf{v}_i^T \lambda_i \mathbf{v}_i} = \sqrt{\lambda_i}\sqrt{\mathbf{v}_i^T \mathbf{v}_i} = \sqrt{\lambda_i}\,\|\mathbf{v}_i\| = \sqrt{\lambda_i}$$

and, for $\sigma_i > 0$, define $\mathbf{u}_i \equiv \frac{A\mathbf{v}_i}{\sigma_i}$. Then

$$\mathbf{u}_i^T \mathbf{u}_j = \left(\frac{A\mathbf{v}_i}{\sigma_i}\right)^T \left(\frac{A\mathbf{v}_j}{\sigma_j}\right) = \frac{\mathbf{v}_i^T A^T A\mathbf{v}_j}{\sigma_i\sigma_j} = \frac{\mathbf{v}_i^T \lambda_j \mathbf{v}_j}{\sigma_i\sigma_j} = 0 \qquad \text{for } i = 1, 2, ..., r$$

The $\mathbf{u}_i$'s inherit orthogonality from the $\mathbf{v}_i$'s. This is vital to SVD. And the $\mathbf{u}_i$'s are also unit vectors: $\|\mathbf{u}_i\| = \left\|\frac{A\mathbf{v}_i}{\sigma_i}\right\| = \frac{\|A\mathbf{v}_i\|}{\sigma_i} = 1$.

Now we wish to extend the $\mathbf{u}_i$'s (AKA the left singular vectors) to m vectors. And we wish to extend the singular values to a count of m. Choose $\sigma_{n+1} = ... = \sigma_m = 0$. Any choice for the remaining $\mathbf{u}_{r+1}$,...,$\mathbf{u}_m$ vectors will trivially satisfy $A\mathbf{v}_i = \sigma_i\mathbf{u}_i$ since $\sigma_{r+1} = ... = \sigma_m = 0$. We choose some orthonormal basis for the nullspace of A^T, which has dimension $m - r$. So choosing a basis from the nullspace of A^T will give us $m - r$ vectors, which is the number of vectors we want for $\mathbf{u}_{r+1}$,...,$\mathbf{u}_m$.

Define $U_{m\times m} \equiv \begin{bmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_m\end{bmatrix}$, $V_{n\times n} \equiv \begin{bmatrix}\mathbf{v}_1 & \dots & \mathbf{v}_n\end{bmatrix}$, and

$$\Sigma_{m\times n} \equiv \begin{bmatrix}\sigma_1 & & & & & & \\ & \sigma_2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_r & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0\end{bmatrix}$$

We now have all the pieces in place. Let's start proving stuff.

For (1), we already showed that the first r left singular vectors ($\mathbf{u}_1$,...,$\mathbf{u}_r$) are orthonormal. So they form an orthonormal basis for some r dimensional subspace of $\mathbb{R}^m$. But is that subspace the column space of A? Well, we chose $\mathbf{u}_i = \frac{A\mathbf{v}_i}{\sigma_i} = A\mathbf{w}_i$ where $\mathbf{w}_i = \frac{1}{\sigma_i}\mathbf{v}_i$. So, by definition, the $\mathbf{u}_1$,...,$\mathbf{u}_r$ are in the column space of A.

For (2), we chose orthonormal $\mathbf{v}_i$'s. And for $i \leq r$, we have

$$\lambda_i\mathbf{v}_i = A^TA\mathbf{v}_i = A^T\sigma_i\mathbf{u}_i \iff \mathbf{v}_i = A^T\mathbf{w}_i \quad \text{where } \mathbf{w}_i = \frac{\sigma_i}{\lambda_i}\mathbf{u}_i$$

Hence the $\mathbf{v}_1$, ..., $\mathbf{v}_r$'s are in the row space of A and form an orthonormal basis for the row space of A.

For (3), you can clearly see this.

For (4), note that we chose $\mathbf{u}_{r+1}$,...,$\mathbf{u}_m$ as an orthonormal basis for the nullspace of A^T.

For (5), we chose orthonormal $\mathbf{v}_i$'s. And we indexed the $\mathbf{v}_i$'s and λ_i's such that $\lambda_{r+1} = ... = \lambda_n = 0$. Hence $\mathbf{v}_{r+1}$,...,$\mathbf{v}_n$ are in the zero eigenspace of A^TA, which is the nullspace of A^TA, which is the nullspace of A, by **$\mathbf{A^TA}$ and $\mathbf{AA^T}$ Properties (3)**.

For (6), first let's check that $A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n}$:

$$U\Sigma V^T = \begin{bmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_m\end{bmatrix}\begin{bmatrix}\sigma_1 & & & & & & \\ & \sigma_2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_r & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0\end{bmatrix}\begin{bmatrix}\mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_n^T\end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_m \end{bmatrix} \begin{bmatrix} \sigma_1 v_{1,1} & \sigma_1 v_{1,2} & \dots & \sigma_1 v_{1,r} & \sigma_1 v_{1,r+1} & \dots & \sigma_1 v_{1,n} \\ \sigma_2 v_{2,1} & \sigma_2 v_{2,2} & \dots & \sigma_2 v_{2,r} & \sigma_2 v_{2,r+1} & \dots & \sigma_2 v_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \dots & \vdots \\ \sigma_r v_{r,1} & \sigma_r v_{r,2} & \dots & \sigma_r v_{r,r} & \sigma_r v_{r,r+1} & \dots & \sigma_r v_{r,n} \\ 0 & & & & 0 & & \\ & \ddots & & & & \ddots & \\ & & 0 & 0 & & & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^r \sigma_i u_{i,1} v_{i,1} & \sum_{i=1}^r \sigma_i u_{i,1} v_{i,2} & \dots & \sum_{i=1}^r \sigma_i u_{i,1} v_{i,r} & \sum_{i=1}^r \sigma_i u_{i,1} v_{i,r+1} & \dots & \sum_{i=1}^r \sigma_i u_{i,1} v_{i,n} \\ \sum_{i=1}^r \sigma_i u_{i,2} v_{i,1} & \sum_{i=1}^r \sigma_i u_{i,2} v_{i,2} & \dots & \sum_{i=1}^r \sigma_i u_{i,2} v_{i,r} & \sum_{i=1}^r \sigma_i u_{i,2} v_{i,r+1} & \dots & \sum_{i=1}^r \sigma_i u_{i,2} v_{i,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \dots & \vdots \\ \sum_{i=1}^r \sigma_i u_{i,r} v_{i,1} & \sum_{i=1}^r \sigma_i u_{i,r} v_{i,2} & \dots & \sum_{i=1}^r \sigma_i u_{i,r} v_{i,r} & \sum_{i=1}^r \sigma_i u_{i,r} v_{i,r+1} & \dots & \sum_{i=1}^r \sigma_i u_{i,r} v_{i,n} \\ \sum_{i=1}^r \sigma_i u_{i,r+1} v_{i,1} & \sum_{i=1}^r \sigma_i u_{i,r+1} v_{i,2} & \dots & \sum_{i=1}^r \sigma_i u_{i,r+1} v_{i,r} & \sum_{i=1}^r \sigma_i u_{i,r+1} v_{i,r+1} & \dots & \sum_{i=1}^r \sigma_i u_{i,r+1} v_{i,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \\ \sum_{i=1}^r \sigma_i u_{i,m} v_{i,1} & \sum_{i=1}^r \sigma_i u_{i,m} v_{i,2} & \dots & \sum_{i=1}^r \sigma_i u_{i,m} v_{i,r} & \sum_{i=1}^r \sigma_i u_{i,m} v_{i,r+1} & \dots & \sum_{i=1}^r \sigma_i u_{i,m} v_{i,n} \end{bmatrix} = \Phi$$

To show that $\Phi = A$, we must show that $A[j,k] = \sum_{i=1}^r \sigma_i u_{i,j} v_{i,k}$. Recall that $\mathbf{u}_i = \frac{A\mathbf{v}_i}{\sigma_i}$. Hence $u_{i,j} = \frac{\mathbf{A_j}\mathbf{v}_i}{\sigma_i}$ and

$$\sum_{i=1}^r \sigma_i u_{i,j} v_{i,k} = \sum_{i=1}^r \sigma_i \Big(\frac{\mathbf{A_j}\mathbf{v}_i}{\sigma_i} \Big) v_{i,k} = \mathbf{A_j} \sum_{i=1}^r \mathbf{v}_i v_{i,k}$$

So we must show that $\sum_{i=1}^r \mathbf{v}_i v_{i,k}$ is the kth standard unit vector in $\mathbb{R}^n$:

$$\sum_{i=1}^r \mathbf{v}_i v_{i,k} = \sum_{i=1}^r \begin{bmatrix} v_{i,1} \\ v_{i,2} \\ \vdots \\ v_{i,k} \\ \vdots \\ v_{i,n} \end{bmatrix} v_{i,k} = \sum_{i=1}^r \begin{bmatrix} v_{i,1} v_{i,k} \\ v_{i,2} v_{i,k} \\ \vdots \\ v_{i,k}^2 \\ \vdots \\ v_{i,n} v_{i,k} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^r v_{i,1} v_{i,k} \\ \sum_{i=1}^r v_{i,2} v_{i,k} \\ \vdots \\ \sum_{i=1}^r v_{i,k}^2 \\ \vdots \\ \sum_{i=1}^r v_{i,n} v_{i,k} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

The last equation follows from the fact that the transpose of an orthogonal matrix (V) is also an orthogonal matrix. Hence $\Phi = A$.

Now let's look at the dyadic sum $\sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T$. It is clear that each term in that sum is rank 1. It is instructive to write out the first couple of terms in this sum:

$$\sigma_1 \mathbf{u}_1 \mathbf{v}_1^T = \sigma_1 \begin{bmatrix} u_{1,1} \\ u_{1,2} \\ \vdots \\ u_{1,m} \end{bmatrix} \begin{bmatrix} v_{1,1} & v_{1,2} & \dots & v_{1,n} \end{bmatrix} = \begin{bmatrix} \sigma_1 u_{1,1} v_{1,1} & \sigma_1 u_{1,1} v_{1,2} & \dots & \sigma_1 u_{1,1} v_{1,n} \\ \sigma_1 u_{1,2} v_{1,1} & \sigma_1 u_{1,2} v_{1,2} & \dots & \sigma_1 u_{1,2} v_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1 u_{1,m} v_{1,1} & \sigma_1 u_{1,m} v_{1,2} & \dots & \sigma_1 u_{1,m} v_{1,n} \end{bmatrix}$$

$$\sigma_2 \mathbf{u}_2 \mathbf{v}_2^T = \sigma_2 \begin{bmatrix} u_{2,1} \\ u_{2,2} \\ \vdots \\ u_{2,m} \end{bmatrix} \begin{bmatrix} v_{2,1} & v_{2,2} & \dots & v_{2,n} \end{bmatrix} = \begin{bmatrix} \sigma_2 u_{2,1} v_{2,1} & \sigma_2 u_{2,1} v_{2,2} & \dots & \sigma_2 u_{2,1} v_{2,n} \\ \sigma_2 u_{2,2} v_{2,1} & \sigma_2 u_{2,2} v_{2,2} & \dots & \sigma_2 u_{2,2} v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_2 u_{2,m} v_{2,1} & \sigma_2 u_{2,m} v_{2,2} & \dots & \sigma_2 u_{2,m} v_{2,n} \end{bmatrix}$$

And summing them:

$$\sigma_1\mathbf{u}_1\mathbf{v}_1^T + \sigma_2\mathbf{u}_2\mathbf{v}_2^T = \begin{bmatrix} \sum_{i=1}^2 \sigma_i u_{i,1}v_{i,1} & \sum_{i=1}^2 \sigma_i u_{i,1}v_{i,2} & \dots & \sum_{i=1}^2 \sigma_i u_{i,1}v_{i,n} \\ \sum_{i=1}^2 \sigma_i u_{i,2}v_{i,1} & \sum_{i=1}^2 \sigma_i u_{i,2}v_{i,2} & \dots & \sum_{i=1}^2 \sigma_i u_{i,2}v_{i,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^2 \sigma_i u_{i,m}v_{i,1} & \sum_{i=1}^2 \sigma_i u_{i,m}v_{i,2} & \dots & \sum_{i=1}^2 \sigma_i u_{i,m}v_{i,n} \end{bmatrix}$$

And it is clear that $\Phi = \sum_{i=1}^r \sigma_i\mathbf{u}_i\mathbf{v}_i^T$.

For (7):

$$A^TA = \Big(\sum_{i=1}^r \sigma_i\mathbf{v}_i\mathbf{u}_i^T\Big)\Big(\sum_{j=1}^r \sigma_j\mathbf{u}_j\mathbf{v}_j^T\Big) = \sum_{i,j=1}^r \sigma_i\sigma_j\mathbf{v}_i\mathbf{u}_i^T\mathbf{u}_j\mathbf{v}_j^T = \sum_{i=1}^r \sigma_i^2\mathbf{v}_i\mathbf{v}_i^T$$

because the orthonormality of the $\mathbf{u}_i$'s means $\mathbf{u}_i^T\mathbf{u}_j = 0$ for $i \neq j$ and $\mathbf{u}_i^T\mathbf{u}_j = 1$ for $i = j$.

$$A^TA\mathbf{v}_j = \Big(\sum_{i=1}^r \sigma_i^2\mathbf{v}_i\mathbf{v}_i^T\Big)\mathbf{v}_j = \sum_{i=1}^r \sigma_i^2\mathbf{v}_i\mathbf{v}_i^T\mathbf{v}_j = \sigma_j^2\mathbf{v}_j$$

thanks again to orthonormality.

For (8):

$$AA^T = \Big(\sum_{i=1}^r \sigma_i\mathbf{u}_i\mathbf{v}_i^T\Big)\Big(\sum_{i=j}^r \sigma_j\mathbf{v}_j\mathbf{u}_j^T\Big) = \sum_{i,j=1}^r \sigma_i\sigma_j\mathbf{u}_i\mathbf{v}_i^T\mathbf{v}_j\mathbf{u}_j^T = \sum_{i=1}^r \sigma_i^2\mathbf{u}_i\mathbf{u}_i^T$$

$$AA^T\mathbf{u}_j = \Big(\sum_{i=1}^r \sigma_i^2\mathbf{u}_i\mathbf{u}_i^T\Big)\mathbf{u}_j = \sum_{i=1}^r \sigma_i^2\mathbf{u}_i\mathbf{u}_i^T\mathbf{u}_j = \sigma_j^2\mathbf{u}_j$$

For (9), there are m orthonormal $m \times 1$ column vectors in U. Or are there? We never showed that each of the $\mathbf{u}_{r+1}$,...,$\mathbf{u}_m$ is orthogonal to each of the $\mathbf{u}_1$,...,$\mathbf{u}_r$. But (1) showed that the $\mathbf{u}_1$,...,$\mathbf{u}_r$ are in the column space of A. And (4) showed that the $\mathbf{u}_{r+1}$,...,$\mathbf{u}_m$ are in the nullspace of A^T. And these two spaces are orthogonal complements. Hence the $\mathbf{u}_1$,...,$\mathbf{u}_m$ are orthonormal and they form an orthonormal basis for $\mathbb{R}^m$.

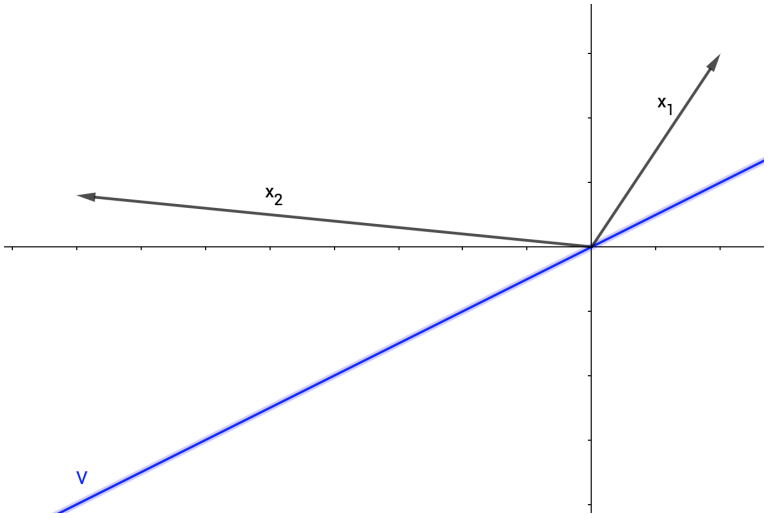
For (10), there are n orthonormal $n \times 1$ columns in V. This forms an orthonormal basis for $\mathbb{R}^n$. ■

You can see why I called this proof the "all at once" proof. We got all the singular values and all the right singular vectors in one shot.

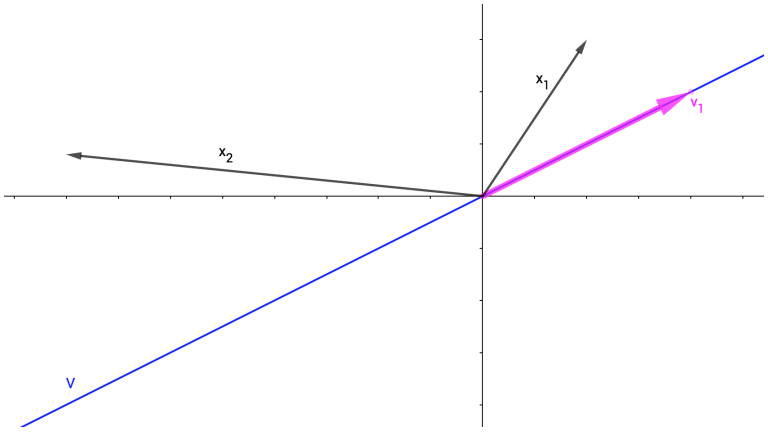
There is a different way to look at SVD. This will lead to a restatement of SVD. It will also lead to the "one at a time" proof of SVD that I mentioned before. This approach, sometimes called the **best least squares fit**, starts with a question: Say we have a set $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\}$ of m vectors in $\mathbb{R}^n$. Can we find the best k-dimensional subspace V of $\mathbb{R}^n$ with respect to this set of points? We define "best" to mean the shortest distance from X to the subspace V. More specifically, we define "best" to mean the minimum sum of the squares of the perpendicular distances from the points in X to the subspace V.

I know of at least two great resources that treat SVD as a **best least squares fit** problem. One resource is a pair of blog posts by Jeremy Kun (SVD 1 and SVD 2). I highly recommend both posts. He provides great intuition into a number of topics in these two posts. The other resource is here. But I'm sure there are other excellent treatments of SVD as a *best least squares fit* problem out there.

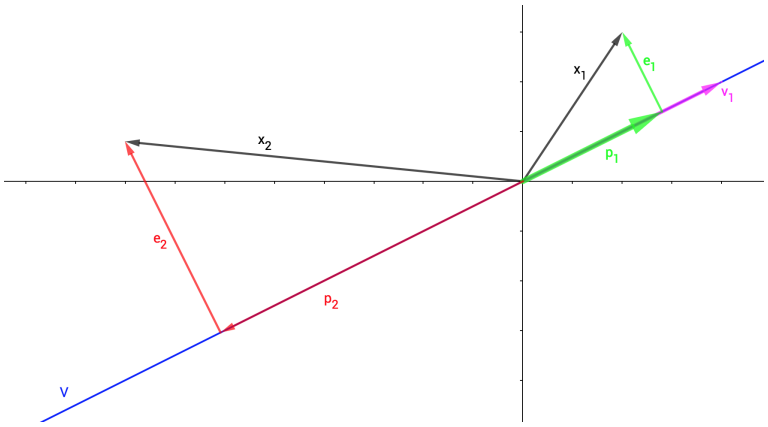
OK, let's start with some simple pictures. Once we get the geometry, the best least squares fit approach to SVD and the "one at a time" proof follow very quickly. Let's say that $m = n = 2$ and let's say $k = 1$. So $X = \{\mathbf{x}_1, \mathbf{x}_2\}$ consists of two vectors in $\mathbb{R}^2$. For the moment, the pictured V can be any line through the origin in $\mathbb{R}^2$. The pictured V is not yet the "best-fit" line:



If we want to minimize the sum of the squares of the perpendicular distances from the points in X to the subspace V, then we will need a basis for V. Let's call $\{\mathbf{v}_1\}$ an orthonormal basis for (the yet non-optimal subspace) V.



And now we want to see the projections of X onto V:



$\mathbf{p_1}$ is the projection of $\mathbf{x_1}$ onto $\mathbf{v_1}$ and $\mathbf{p_2}$ is the projection of $\mathbf{x_2}$ onto $\mathbf{v_1}$. $\mathbf{e_1} = \mathbf{x_1} - \mathbf{p_1}$ is the error from the projection of $\mathbf{x_1}$ onto $\mathbf{v_1}$ and $\mathbf{e_2} = \mathbf{x_2} - \mathbf{p_2}$ is the error from the projection of $\mathbf{x_2}$ onto $\mathbf{v_1}$. Recall that, for a given $\mathbf{v_1}$, the error $\mathbf{e_i}$ is minimized when we put $\mathbf{p_i} = \frac{\mathbf{v_1}^T\mathbf{x_i}}{\mathbf{v_1}^T\mathbf{v_1}}\mathbf{v_1}$. It is easy to see and compute that $\mathbf{e_1}$ and $\mathbf{e_2}$ are perpendicular to $\mathbf{v_1}$:

$$\mathbf{e_i} \cdot \mathbf{v_1} = \mathbf{x_i}^T\mathbf{v_1} - \mathbf{p_i}^T\mathbf{v_1} = \mathbf{x_i}^T\mathbf{v_1} - \left(\frac{\mathbf{v_1}^T\mathbf{x_i}}{\mathbf{v_1}^T\mathbf{v_1}}\mathbf{v_1}\right)^T\mathbf{v_1}$$

$$= \mathbf{x_i}^T\mathbf{v_1} - \left(\frac{\mathbf{x_i}^T\mathbf{v_1}}{\mathbf{v_1}^T\mathbf{v_1}}\mathbf{v_1}^T\right)\mathbf{v_1} = \mathbf{x_i}^T\mathbf{v_1} - \mathbf{x_i}^T\mathbf{v_1} = 0$$

Similarly, $\mathbf{e_i}$ is perpendicular to $\mathbf{p_i}$:

$$\mathbf{e_i} \cdot \mathbf{p_i} = \mathbf{x_i}^T\mathbf{p_i} - \mathbf{p_i}^T\mathbf{p_i} = \mathbf{x_i}^T\mathbf{p_i} - \left(\frac{\mathbf{v_1}^T\mathbf{x_i}}{\mathbf{v_1}^T\mathbf{v_1}}\mathbf{v_1}\right)^T\mathbf{p_i}$$

$$= \mathbf{x_i}^T\mathbf{p_i} - \left(\frac{\mathbf{x_i}^T\mathbf{v_1}}{\mathbf{v_1}^T\mathbf{v_1}}\mathbf{v_1}^T\right)\mathbf{p_i} = \mathbf{x_i}^T\mathbf{p_i} - \frac{\mathbf{x_i}^T\mathbf{v_1}}{\|\mathbf{v_1}\|^2}\mathbf{p_i}^T\mathbf{v_1}$$

$$= \mathbf{x_i}^T\mathbf{p_i} - \mathbf{x_i}^T\mathbf{v_1}\mathbf{p_i}^T\mathbf{v_1} = \mathbf{x_i}^T\mathbf{p_i} - \mathbf{x_i}^T\mathbf{p_i}\mathbf{v_1}^T\mathbf{v_1} = \mathbf{x_i}^T\mathbf{p_i} - \mathbf{x_i}^T\mathbf{p_i} = 0$$

Notice that

$$0 = \mathbf{e_i} \cdot \mathbf{p_i} = \mathbf{x_i}^T\mathbf{p_i} - \mathbf{p_i}^T\mathbf{p_i} = \mathbf{p_i}^T\mathbf{x_i} - \mathbf{p_i}^T\mathbf{p_i}$$

So that

$$\mathbf{x_i}^T\mathbf{p_i} = \mathbf{p_i}^T\mathbf{p_i} \text{ and } \mathbf{p_i}^T\mathbf{x_i} = \mathbf{p_i}^T\mathbf{p_i}$$

Hence, for a given $\mathbf{v_1}$, we have:

$$\|\mathbf{e_i}\|^2 + \|\mathbf{p_i}\|^2 = \mathbf{e_i}^T\mathbf{e_i} + \mathbf{p_i}^T\mathbf{p_i} = (\mathbf{x_i} - \mathbf{p_i})^T(\mathbf{x_i} - \mathbf{p_i}) + \mathbf{p_i}^T\mathbf{p_i}$$

$$= (\mathbf{x_i}^T - \mathbf{p_i}^T)(\mathbf{x_i} - \mathbf{p_i}) + \mathbf{p_i}^T\mathbf{p_i} = \mathbf{x_i}^T\mathbf{x_i} - \mathbf{x_i}^T\mathbf{p_i} - \mathbf{p_i}^T\mathbf{x_i} + \mathbf{p_i}^T\mathbf{p_i} + \mathbf{p_i}^T\mathbf{p_i} = \|\mathbf{x_i}\|^2$$

Succintly, for a given $\mathbf{v_1}$, we have the Pythagorean Theorem:

$$\|\mathbf{e_i}\|^2 + \|\mathbf{p_i}\|^2 = \|\mathbf{x_i}\|^2$$

Now remember that we haven't yet selected the "best-fit" subspace V nor its basis $\{\mathbf{v_1}\}$. The objective is to find the "optimal" 1-dimensional V such that we minimize the sum of the squares of the perpendicular distances from the points in X to any 1-dimensional subspace. In our simple pictures, the sum of the squares of the perpendicular distances is $\|\mathbf{e_1}\|^2 + \|\mathbf{e_2}\|^2$:

$$\|\mathbf{e_1}\|^2 + \|\mathbf{e_2}\|^2 = \|\mathbf{x_1}\|^2 + \|\mathbf{x_2}\|^2 - \|\mathbf{p_1}\|^2 - \|\mathbf{p_2}\|^2$$

So how can we choose V and its basis $\{\mathbf{v_1}\}$ so as to minimize $\|\mathbf{e_1}\|^2 + \|\mathbf{e_2}\|^2$? Well $\mathbf{x_1}$ and $\mathbf{x_2}$ are given and thus fixed. But $\mathbf{p_1} \equiv \frac{\mathbf{v_1}^T\mathbf{x_1}}{\mathbf{v_1}^T\mathbf{v_1}}\mathbf{v_1}$ and $\mathbf{p_2} \equiv \frac{\mathbf{v_1}^T\mathbf{x_2}}{\mathbf{v_1}^T\mathbf{v_1}}\mathbf{v_1}$ are dependent on $\mathbf{v_1}$, when we select $\mathbf{p_1}$ and $\mathbf{p_2}$ "correctly". Hence we can minimize the sum of the squares of the perpendicular distances by choosing $\mathbf{v_1}$ to maximize $\|\mathbf{p_1}\|^2 + \|\mathbf{p_2}\|^2$. More generally, we can minimize the sum of the squares of the perpendicular distances between X and V by maximizing the sum of the squared lengths of the projections of X onto V. That is, we have converted our original minimization problem into a maximization problem:

$$\mathbf{v_1} = \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1}\Big(\|\mathbf{p_1}\|^2 + \|\mathbf{p_2}\|^2\Big) = \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1}\Big(\sum_{i=1}^{2}\|\mathbf{p_i}\|^2\Big)$$

$$= \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1}\Big(\sum_{i=1}^{2}\left\|\frac{\mathbf{v}^T\mathbf{x_i}}{\mathbf{v}^T\mathbf{v}}\mathbf{v}\right\|^2\Big) = \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1}\Big(\sum_{i=1}^{2}\left\|\frac{\mathbf{v}^T\mathbf{x_i}}{1}\mathbf{v}\right\|^2\Big)$$

$$= \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1}\Big(\sum_{i=1}^{2}\left\|(\mathbf{v}^T\mathbf{x_i})\mathbf{v}\right\|^2\Big) = \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1}\Big(\sum_{i=1}^{2}(\mathbf{v}^T\mathbf{x_i})^2\|\mathbf{v}\|^2\Big)$$

$$= \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1}\sum_{i=1}^{2}(\mathbf{v}\cdot\mathbf{x_i})^2$$

But we can write this more compactly by putting the $\mathbf{x_i}$'s into the rows of a matrix, call it A:

$$A = \begin{bmatrix}\mathbf{x_1}^T\\ \mathbf{x_2}^T\end{bmatrix} = \begin{bmatrix}x_{1,1} & x_{1,2}\\ x_{2,1} & x_{2,2}\end{bmatrix}$$

So that

$$A\mathbf{v} = \begin{bmatrix}x_{1,1} & x_{1,2}\\ x_{2,1} & x_{2,2}\end{bmatrix}\begin{bmatrix}v_1\\ v_2\end{bmatrix} = \begin{bmatrix}\mathbf{x_1}\cdot\mathbf{v}\\ \mathbf{x_2}\cdot\mathbf{v}\end{bmatrix}$$

And

$$\|A\mathbf{v}\|^2 = (\mathbf{x_1}\cdot\mathbf{v})^2 + (\mathbf{x_2}\cdot\mathbf{v})^2$$

And

$$\mathbf{v_1} = \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1}\sum_{i=1}^{2}(\mathbf{x_i}\cdot\mathbf{v})^2 = \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1}\|A\mathbf{v}\|^2$$

$$= \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1}\|A\mathbf{v}\|$$

The last equation follows from the fact the squaring function is a monotonically increasing function.

Now let's take this up a notch. Suppose $X = \{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}\}$ with $\mathbf{x_i} \in \mathbb{R}^n$. We want to find the plane V in $\mathbb{R}^n$ that minimizes the sum of the squares of the perpendicular distances. That is, the objective is to find an orthonormal basis $\{\mathbf{v_1}, \mathbf{v_2}\}$ with $\mathbf{v_i} \in \mathbb{R}^n$ such that we minimize the sum of the squares of the perpendicular distances between $\{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}\}$ and $\{\mathbf{v_1}, \mathbf{v_2}\}$.

Or even more generally, let $X = \{\mathbf{x_1}, ..., \mathbf{x_m}\}$ with $\mathbf{x_i} \in \mathbb{R}^n$. The objective is to find some k-dimensional subspace of $\mathbb{R}^n$, call it V, with orthonormal basis $\{\mathbf{v_1}, ..., \mathbf{v_k}\}$ such that we minimize the sum of the squares of the perpendicular distances between $\{\mathbf{x_1}, ..., \mathbf{x_m}\}$ and $\{\mathbf{v_1}, ..., \mathbf{v_k}\}$. Equivalently, the objective is to find some k-dimensional subspace $V \subset \mathbb{R}^n$ with orthonormal basis $\{\mathbf{v_1}, ..., \mathbf{v_k}\}$ such that we maximize the sum of the squared lengths of the projections of $\{\mathbf{x_1}, ..., \mathbf{x_m}\}$ onto $\{\mathbf{v_1}, ..., \mathbf{v_k}\}$.

Well just how are we supposed to do that? Maybe "one at a time"? But we must remember to select orthogonal $\mathbf{v_i}$'s:

$$\mathbf{v_1} = \text{arg max}_{\mathbf{v}:\|\mathbf{v}\|=1} \|A\mathbf{v}\|$$

$$\mathbf{v_2} = \text{arg max}_{\mathbf{v}\perp\mathbf{v_1},\|\mathbf{v}\|=1} \|A\mathbf{v}\|$$

For $\mathbf{v_2}$, you see that I added a second condition to the arg max function: $\mathbf{v} \perp \mathbf{v_1}$. So this will get us orthonormal vectors. But will it minimize the sum of the squares of the perpendicular distances between $\{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}\}$ and $\{\mathbf{v_1}, \mathbf{v_2}\}$?

Before we answer that most important question, let's officially recognize a few things that you have probably noticed. By taking this "one at a time" approach, we are selecting the right singular vectors, the $\mathbf{v_i}$'s. We are also selecting the singular values and left singular vectors:

$$\sigma_1 \equiv \text{max}_{\mathbf{v}:\|\mathbf{v}\|=1} \|A\mathbf{v}\|$$

$$\sigma_i \equiv \text{max}_{\mathbf{v}\perp\mathbf{v_1},...,\mathbf{v}\perp\mathbf{v_{i-1}},\|\mathbf{v}\|=1} \|A\mathbf{v}\|$$

$$\mathbf{u_i} \equiv \frac{A\mathbf{v_i}}{\sigma_i}$$

It is clear that the singular values are monotonically decreasing: $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{k-1} \geq \sigma_k$. This follows directly from our definition for each σ_i. Notice that we specify $\|A\mathbf{v}\|$ (without the square) for the defintion of σ_1 and σ_i, rather than $\|A\mathbf{v}\|^2$. For the definition of $\mathbf{v_1}$ and $\mathbf{v_i}$, we can use $\|A\mathbf{v}\|$ and $\|A\mathbf{v}\|^2$ interchangeably. The monotonicity of the squaring function guarantees that the same $\mathbf{v_i}$ will satisfy both. But for the definition of the singular value σ_i, we specifically want $\|A\mathbf{v}\|$, without the square.

We these definitions, we can actually prove that our "one at a time" approach is equivalent to the "all at once" approach! Let's do that. If you wish to skip this and continue with the "one at a time approach", click here. First we define $r(\mathbf{x}) = \frac{\mathbf{x}^T S\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$ for a symmetric matrix S. This is called the Rayleigh Quotient and it comes in quite handy:

Lemma 4: Eigenvectors & The Rayleigh Quotient

Suppose S is a symmtric matrix and let $r(\mathbf{x}) = \frac{\mathbf{x}^T S\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$. Then $\mathbf{x} \neq \mathbf{0}$ solves $\mathbf{0} = \nabla r(\mathbf{x})$ if and only if $\mathbf{x}$ is an eigenvector of S with eigenvalue $r(\mathbf{x})$.

Proof

First we compute the partials:

$$\frac{\partial r(\mathbf{x})}{\partial x_j} = \frac{\frac{\partial}{\partial x_j}\left(\mathbf{x}^T S\mathbf{x}\right)}{\mathbf{x}^T\mathbf{x}} - \frac{\left(\mathbf{x}^T S\mathbf{x}\right)\frac{\partial}{\partial x_j}\left(\mathbf{x}^T\mathbf{x}\right)}{(\mathbf{x}^T\mathbf{x})^2}$$

$$= \frac{2(S\mathbf{x})_j}{\mathbf{x}^T\mathbf{x}} - \frac{(\mathbf{x}^T S\mathbf{x})2x_j}{(\mathbf{x}^T\mathbf{x})^2}$$

$$= \frac{2}{\mathbf{x}^T\mathbf{x}}(S\mathbf{x} - r(\mathbf{x})\mathbf{x})_j$$

So our gradient is $\nabla r(\mathbf{x}) = \frac{2}{\mathbf{x}^T\mathbf{x}}(S\mathbf{x} - r(\mathbf{x})\mathbf{x})$ and we find that

$$\mathbf{0} = \nabla r(\mathbf{x}) = \frac{2}{\mathbf{x}^T\mathbf{x}}(S\mathbf{x} - r(\mathbf{x})\mathbf{x})$$

$$\iff \mathbf{0} = S\mathbf{x} - r(\mathbf{x})\mathbf{x}$$

$$\iff S\mathbf{x} = r(\mathbf{x})\mathbf{x}$$

■

This lemma tells us is that the stationary points of $r(\mathbf{x}) = \frac{\mathbf{x}^T S\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$ are eigenvectors for S with eigenvalues equal to $r(\mathbf{x})$ at those points. Hence we can set $S = A^T A$ and get the largest eigenvalue by computing $\lambda_1 = \max_{\mathbf{x}} r(\mathbf{x}) = \max_{\mathbf{x}} \frac{\mathbf{x}^T A^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \max_{\mathbf{x}} \frac{\|A\mathbf{x}\|^2}{\|\mathbf{x}\|^2}$ and

$$\lambda_1 = \max_{\mathbf{v}:\|\mathbf{v}\|=1} \|A\mathbf{v}\|^2 = \sigma_1^2$$

$$\mathbf{v_1} = \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1} \|A\mathbf{v}\|^2$$

So $\lambda_1 = \sigma_1^2$ is the largest eigenvalue of $A^T A$ with corresponding eigenvector $\mathbf{v_1}$. Of course these are the same values we computed above to maximize the sum of the squared lengths of the projections of X onto V. Similarly, lemma 4 says that

$$\lambda_i = \max_{\mathbf{v}\in \text{span}(\mathbf{v_1},\ldots,\mathbf{v_{i-1}})^\perp, \|\mathbf{v}\|=1} \|A\mathbf{v}\|^2 = \sigma_i^2$$

$$\mathbf{v_i} = \arg\max_{\mathbf{v}\in \text{span}(\mathbf{v_1},\ldots,\mathbf{v_{i-1}})^\perp, \|\mathbf{v}\|=1} \|A\mathbf{v}\|^2$$

is the i[th] largest eigenpair for $A^T A$ since $\mathbf{v_i}$ is a stationary point of $r(\mathbf{x})$.

So how does this show that the "all at once" approach is equivalent to this "one at a time" approach? Well, in the first few lines of that proof, we picked the right singular vectors to be orthonormal eigenvectors of $A^T A$. And we defined the singulars values to be the square roots of the corresponding eigenvalues. Everything else in the proof follows from that. But now we have shown than the $\mathbf{v_i}$'s and σ_i^2's defined to maximize the sum of the squared lengths of the projections of X onto V are in fact orthonormal eigenpairs for $A^T A$. Hence equivalency.

Incidentally, this leads directly to another proof of the Spectral Theorem. By iteratively computing the λ_i's and orthonormal $\mathbf{v_i}$'s for any given symmetric matrix S using the max, argmax of the Rayleigh Quotient, we are also computing $S = \sum_{i=1}^{n} \lambda_i \mathbf{v_i}\mathbf{v_i}^T = V\Lambda V^T$.

OK, back to our "one at a time" approach. Now you might be asking, can we do this iterative step n times? Well, as long as $\|A\mathbf{v}\| > 0$, then you will have n positive singular values: $\sigma_i > 0$ for $i = 1, ..., n$. And you will have an orthonormal basis $\{\mathbf{v_i}, ..., \mathbf{v_n}\}$ for $\mathbb{R}^n$. As we proved in the "all at once" proof, A has rank n.

But what happens when $\sigma_k > 0$ but $\sigma_{k+1} = 0$ for some $k < n$? This must mean that the rank of A is k. Let's see why that must be the case. Here's what we have: For any $j > k$, any $\mathbf{v_j}$ satisfying $\mathbf{v_j} \perp \mathbf{v_i}$ for all $i = 1, ..., k$ must also satisfy $\|A\mathbf{v_j}\| = \sigma_j = 0$. Such $\mathbf{v_j}$'s must also satisfy $A\mathbf{v_j} = \mathbf{0}$. That is, the only remaining $\mathbf{v_j}$'s that are orthogonal to all $\{\mathbf{v_1}, ..., \mathbf{v_k}\}$ must be in the nullspace of A. Since the dimension of $\mathbb{R}^n$ is n, surely we can select $n - k$ such $\mathbf{v_j}$'s. Said another way, we can extend $\{\mathbf{v_1}, ..., \mathbf{v_k}\}$ to an orthonormal basis for $\mathbb{R}^n$ but only with $\mathbf{v_j}$'s that are in the nullspace of A. Hence the nullspace of A has dimension $n - k$ and hence A has rank k.

In this case where the rank of A is k for $k < n$, you have k positive singular values: $\sigma_i > 0$ for $i = 1, ..., k$. And you have k right singular vectors $\{\mathbf{v_i} : i = 1, ..., k\}$. Then you can put $\sigma_i = 0$ for $i > k$ to get n nonnegative singular values: $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{n-1} \geq \sigma_n = 0$. And you can extend $\{\mathbf{v_1}, ..., \mathbf{v_k}\}$ with any orthonormal basis $\{\mathbf{v_{k+1}}, ..., \mathbf{v_n}\}$ for the nullspace of A to get a full set of right singular vectors $\{\mathbf{v_1}, ..., \mathbf{v_k}, \mathbf{v_{k+1}}, ..., \mathbf{v_n}\}$.

One last piece of terminology: We will say a k-dimensional subspace V is **best-fit** or **optimal** for a matrix A if it minimizes the sum of the squares of the perpendicular distances between the rows of A and all k-dimensional subspaces. Equivalently, a subspace V is **best-fit** or **optimal** for A if it maximizes the sum of the squared lengths of the projections of the rows of A onto all k-dimensional subspaces.

OK, finally, let's state and prove the "alternate" form of SVD:

Theorem: Alternate Singular Value Decomposition

Let $A_{m \times n} = \begin{bmatrix} \mathbf{x_1}^T \\ \vdots \\ \mathbf{x_m}^T \end{bmatrix}$, $\mathbf{x_i} \in \mathbb{R}^n$, where the rank of A is r. Let $\{\mathbf{v_1}, ..., \mathbf{v_r}\}$ be its right singular vectors and let $\{\sigma_1, ..., \sigma_r\}$ be its singular values:

$$\mathbf{v_i},\ \sigma_i \equiv \text{argmax, max}_{\mathbf{v} \perp \mathbf{v_1}, ..., \mathbf{v} \perp \mathbf{v_{i-1}}, \|\mathbf{v}\|=1} \|A\mathbf{v}\|$$

For any $k \leq r$, $V_k \equiv \text{Span}\{\mathbf{v_1}, ..., \mathbf{v_k}\}$ is the best-fit k-dimensional vector subspace for A.

SVD Proof: One At A Time

For $k = 1$, we have already shown that this reduces to a least squares fitting problem for which $\mathbf{v_1}$ is the solution. For $k = 2$, let W be a best-fit 2-dimensional subspace for A. For any basis $\mathbf{w_1}$, $\mathbf{w_2}$ of W, $\|A\mathbf{w_1}\| + \|A\mathbf{w_2}\|$ is the sum of squared lengths of the projections of the rows of A onto W. Now,

choose a basis $\mathbf{w_1}$, $\mathbf{w_2}$ of W so that $\mathbf{w_2}$ is perpendicular to $\mathbf{v_1}$. If $\mathbf{v_1}$ is perpendicular to W, any unit vector in W will do as $\mathbf{w_2}$. If not, choose $\mathbf{w_2}$ to be the unit vector in W perpendicular to the projection of $\mathbf{v_1}$ onto W. Since $\mathbf{v_1}$ was chosen to maximize $\|A\mathbf{v}\|^2$, it follows that $\|A\mathbf{w_1}\|^2 \leq \|A\mathbf{v_1}\|^2$. Since $\mathbf{v_2}$ was chosen to maximize $\|A\mathbf{v}\|^2$ over all $\mathbf{v}$ perpendicular to $\mathbf{v_1}$, $\|A\mathbf{w_2}\|^2 \leq \|A\mathbf{v_2}\|^2$. Hence

$$\|A\mathbf{w_1}\|^2 + \|A\mathbf{w_2}\|^2 \leq \|A\mathbf{v_1}\|^2 + \|A\mathbf{v_2}\|^2$$

and V_2 is at least as good as W and so is a best-fit 2-dimensional subspace.

For general k, proceed by induction. By the induction hypothesis, V_{k-1} is a best-fit $k-1$ dimensional subspace. Suppose W is a best-fit k-dimensional subspace. Choose a basis $\mathbf{w_1}$, $\mathbf{w_2}$,..., $\mathbf{w_k}$ of W so that $\mathbf{w_k}$ is perpendicular to $\mathbf{v_1}$, $\mathbf{v_2}$,..., $\mathbf{v_{k-1}}$. Then

$$\|A\mathbf{w_1}\|^2 + \|A\mathbf{w_2}\|^2 + ... + \|A\mathbf{w_k}\|^2 \leq \|A\mathbf{v_1}\|^2 + \|A\mathbf{v_2}\|^2 + ... + \|A\mathbf{v_{k-1}}\|^2 + \|A\mathbf{w_k}\|^2$$

since V_{k-1} is an optimal $k-1$ dimensional subspace. Since $\mathbf{w_k}$ is perpendicular to $\mathbf{v_1}$, $\mathbf{v_2}$,..., $\mathbf{v_{k-1}}$, by the definition of $\mathbf{v_k}$, $\|A\mathbf{w_k}\|^2 \leq \|A\mathbf{v_k}\|^2$. Thus

$$\|A\mathbf{w_1}\|^2 + \|A\mathbf{w_2}\|^2 + ... + \|A\mathbf{w_{k-1}}\|^2 + \|A\mathbf{w_k}\|^2 \leq \|A\mathbf{v_1}\|^2 + \|A\mathbf{v_2}\|^2 + ... + \|A\mathbf{v_{k-1}}\|^2 + \|A\mathbf{v_k}\|^2$$

proving that V_k is at least as good as W and hence is optimal. ■

Now we would like to find a reasonably efficient algorithm for computing the singular vectors and singular values.

The Power Method

For (just about) any $\mathbf{x} \in \mathbb{R}^n$ and any $m \times n$ matrix A:

$$\lim_{s\to\infty} \frac{(A^TA)^s\mathbf{x}}{\|(A^TA)^s\mathbf{x}\|} = \mathbf{v_1} \quad \text{if and only if } \sigma_1 > \sigma_2$$

and

$$\lim_{s\to\infty} \frac{(AA^T)^s\mathbf{x}}{\|(AA^T)^s\mathbf{x}\|} = \mathbf{u_1} \quad \text{if and only if } \sigma_1 > \sigma_2$$

where $\mathbf{v_1}$ is the top right singular vector, $\mathbf{u_1}$ is the top left singular vector, and σ_1 and σ_2 are the first and second singular values of A.

Proof

First we need to find a convenient formula for $(A^TA)^s$. Suppose A has rank r and orthonormal right singular vectors $\mathbf{v_1}, ..., \mathbf{v_n}$. Then

$$(A^TA)^2 = \Big(\sum_{i=1}^{r} \sigma_i^2 \mathbf{v_i}\mathbf{v_i}^T\Big)^2 = \Big(\sum_{i=1}^{r} \sigma_i^2 \mathbf{v_i}\mathbf{v_i}^T\Big)\Big(\sum_{j=1}^{r} \sigma_j^2 \mathbf{v_j}\mathbf{v_j}^T\Big)$$

$$= \sum_{i,j=1}^{r} \sigma_i^2\sigma_j^2 \mathbf{v_i}\mathbf{v_i}^T\mathbf{v_j}\mathbf{v_j}^T = \sum_{i=1}^{r} \sigma_i^4 \mathbf{v_i}\mathbf{v_i}^T$$

Inductively, we find that

$$(A^TA)^s = \sum_{i=1}^{r} \sigma_i^{2s} \mathbf{v_i} \mathbf{v_i}^T$$

Aside: The outer product is more convenient here but we also could have found

$$A^TA = (U\Sigma V^T)^T(U\Sigma V^T) = V\Sigma^T U^T U\Sigma V^T = V\Sigma^2 V^T$$

The last equation follows from the orthonormality of U so that $U^TU = I$. Also Σ is diagonal and hence symmetric: $\Sigma^T = \Sigma$. Similarly, we can compute

$$(A^TA)^2 = V\Sigma^2 V^T V\Sigma^2 V^T = V\Sigma^4 V^T$$

And inductively find

$$(A^TA)^s = V\Sigma^{2s} V^T$$

Anyway, back to the proof: since $\{\mathbf{v_1}, ..., \mathbf{v_n}\}$ is an orthonormal basis for $\mathbb{R}^n$, there exists scalars $c_1, ..., c_n$ such that $\mathbf{x} = \sum_{i=1}^{n} c_i \mathbf{v_i}$. And we can compute

$$A^TA\mathbf{x} = \Big(\sum_{i=1}^{n} \sigma_i^2 \mathbf{v_i} \mathbf{v_i}^T\Big)\Big(\sum_{j=1}^{n} c_j \mathbf{v_j}\Big) = \sum_{i,j=1}^{n} \sigma_i^2 \mathbf{v_i} \mathbf{v_i}^T c_j \mathbf{v_j} = \sum_{i=1}^{n} c_i \sigma_i^2 \mathbf{v_i} = \sum_{i=1}^{r} c_i \sigma_i^2 \mathbf{v_i}$$

Similarly, using our handy formula for $(A^TA)^s$, we find that

$$(A^TA)^s\mathbf{x} = \sum_{i=1}^{r} c_i \sigma_i^{2s} \mathbf{v_i}$$

We also need a handy formula for $\big\|(A^TA)^s\mathbf{x}\big\|$:

$$\Big\|(A^TA)^s\mathbf{x}\Big\|^2 = ((A^TA)^s\mathbf{x})^T((A^TA)^s\mathbf{x}) = \Big(\sum_{i=1}^{r} c_i \sigma_i^{2s} \mathbf{v_i}\Big)^T \Big(\sum_{j=1}^{r} c_j \sigma_j^{2s} \mathbf{v_j}\Big)$$

$$= \Big(\sum_{i=1}^{r} c_i \sigma_i^{2s} \mathbf{v_i}^T\Big)\Big(\sum_{j=1}^{r} c_j \sigma_j^{2s} \mathbf{v_j}\Big) = \sum_{i,j=1}^{r} c_i c_j \sigma_i^{2s} \sigma_j^{2s} \mathbf{v_i}^T \mathbf{v_j} = \sum_{i=1}^{r} c_i^2 \sigma_i^{4s}$$

So we have a tractable formula for $\frac{(A^TA)^s\mathbf{x}}{\|(A^TA)^s\mathbf{x}\|}$ as well:

$$\frac{(A^TA)^s\mathbf{x}}{\|(A^TA)^s\mathbf{x}\|} = \frac{1}{\|(A^TA)^s\mathbf{x}\|} \sum_{i=1}^{r} c_i \sigma_i^{2s} \mathbf{v_i} = \frac{1}{\sqrt{\sum_{j=1}^{r} c_j^2 \sigma_j^{4s}}} \sum_{i=1}^{r} c_i \sigma_i^{2s} \mathbf{v_i}$$

$$= \sum_{i=1}^{r} \frac{c_i \sigma_i^{2s}}{\sqrt{\sum_{j=1}^{r} c_j^2 \sigma_j^{4s}}} \mathbf{v_i}$$

So let's compute $\lim_{s\to\infty} \frac{c_i \sigma_i^{2s}}{\sqrt{\sum_{j=1}^{r} c_j^2 \sigma_j^{4s}}}$:

$$\lim_{s\to\infty} \frac{c_i \sigma_i^{2s}}{\sqrt{\sum_{j=1}^{r} c_j^2 \sigma_j^{4s}}} = \lim_{s\to\infty} \sqrt{\frac{c_i^2 \sigma_i^{4s}}{\sum_{j=1}^{r} c_j^2 \sigma_j^{4s}}} = \sqrt{\lim_{s\to\infty} \frac{c_i^2 \sigma_i^{4s}}{\sum_{j=1}^{r} c_j^2 \sigma_j^{4s}}}$$

$$= \sqrt{\lim_{s\to\infty} \frac{1}{1 + \frac{1}{c_i^2\sigma_i^{4s}} \sum_{j=1, j\neq i}^{r} c_j^2\sigma_j^{4s}}} = \sqrt{\lim_{s\to\infty} \frac{1}{1 + \sum_{j=1, j\neq i}^{r} \frac{c_j^2\sigma_j^{4s}}{c_i^2\sigma_i^{4s}}}}$$

$$= \sqrt{\lim_{s\to\infty} \frac{1}{1 + \sum_{j=1, j\neq i}^{r} (\frac{c_j}{c_i})^2 (\frac{\sigma_j}{\sigma_i})^{4s}}}$$

For $i = 1$, we have

$$\lim_{s\to\infty} \frac{c_1\sigma_1^{2s}}{\sqrt{\sum_{j=1}^{r} c_j^2\sigma_j^{4s}}} = \begin{cases} 1 & \text{if } \sigma_1 > \sigma_2 \\ \sqrt{\frac{1}{1+(\frac{c_2}{c_1})^2}} & \text{if } \sigma_1 = \sigma_2 \end{cases}$$

Of course we are also assuming that the random vector $\mathbf{x}$ has a nonzero first compotent $c_1 \neq 0$ (this is the "just about" disclaimer in the theorem statement). For $i = 2$, we have

$$\lim_{s\to\infty} \frac{c_2\sigma_2^{2s}}{\sqrt{\sum_{j=1}^{r} c_j^2\sigma_j^{4s}}} = \begin{cases} 0 & \text{if } \sigma_1 > \sigma_2 \\ \sqrt{\frac{1}{1+(\frac{c_1}{c_2})^2}} & \text{if } \sigma_1 = \sigma_2 \end{cases}$$

And putting it together:

$$\lim_{s\to\infty} \frac{(A^TA)^s\mathbf{x}}{\|(A^TA)^s\mathbf{x}\|} = \lim_{s\to\infty} \sum_{i=1}^{r} \frac{c_i\sigma_i^{2s}}{\sqrt{\sum_{j=1}^{r} c_j^2\sigma_j^{4s}}} \mathbf{v_i} = \mathbf{v_1} \quad \text{so long as } \sigma_1 > \sigma_2$$

The proof for $\lim_{s\to\infty} \frac{(AA^T)^s\mathbf{x}}{\|(AA^T)^s\mathbf{x}\|} = \mathbf{u_1}$ if and only if $\sigma_1 > \sigma_2$ is similar. ■

Of course, once we have numerically computed $\mathbf{v_1}$, we can use the Rayleigh Quotient to get the corresponding eigenvalue of A^TA which is also the square of the singular value of A: $\sigma_1^2 = \lambda_1 = r(\mathbf{v_1}) = \frac{\mathbf{v_1}^TA^TA\mathbf{v_1}}{\mathbf{v_1}^T\mathbf{v_1}} = \mathbf{v_1}^TA^TA\mathbf{v_1} = (A\mathbf{v_1})^TA\mathbf{v_1} = \mathbf{u_1}^T\mathbf{u_1}$. And then we can normalize $\mathbf{u_1}$ to finish off the first pair of singular vectors.

Since we have computed σ_1, $\mathbf{v_1}$, and $\mathbf{u_1}$, we can repeat this procedure for $A - \sigma_1\mathbf{u_1}\mathbf{v_1}^T$. The top singular value and vectors of this matrix are σ_2, $\mathbf{v_1}$, and $\mathbf{u_2}$. Repeating this procedure k times will get us the top k singular values and vectors and the best-fit k-dimensional subspace $V_k = \text{Span}\{\mathbf{v_1}, ..., \mathbf{v_k}\}$.

So the question becomes, how quickly does $\frac{(A^TA)^s\mathbf{x}}{\|(A^TA)^s\mathbf{x}\|}$ converge to $\mathbf{v_1}$? Well, the Power Method isn't industrial strength, but it's still pretty good. You can find a detailed analysis here or at the end this. It turns out that the convergence is geometric with ratio $\frac{\lambda_2}{\lambda_1}$.

When we actually code this up, when we're computing an approximation for some $\mathbf{v_i}$, instead of iterating a calculated number of times, we will simply stop iterating when the angle between iteration results is sufficiently small. What does that mean? Let $\mathbf{t_s} = \frac{(A^TA)^s\mathbf{x}}{\|(A^TA)^s\mathbf{x}\|}$. Then $\mathbf{t_s}$ is a unit vector. So the cos of the angle θ_s between $\mathbf{t_s}$ and $\mathbf{t_{s-1}}$ can be computed as $cos(\theta_s) = \frac{\mathbf{t_s}\cdot\mathbf{t_{s-1}}}{\|\mathbf{t_s}\|\|\mathbf{t_{s-1}}\|} = \mathbf{t_s}\cdot\mathbf{t_{s-1}}$. We know that $\lim_{\theta\to 0} cos(\theta) = 1$ and that $cos(\theta) \leq 1$ for all θ. At each iteration s, we check whether $\mathbf{t_s}\cdot\mathbf{t_{s-1}} > 1 - \epsilon$ for some arbitrarily small ϵ. When this check is true for some iteration s, we have a sufficiently small angle between iteration results and we're not going to get a much better approximation for $\mathbf{v_i}$ than $\mathbf{t_s}$. So we stop.