Orthogonal matrices have come up frequently in this course for a good reason:

- $\|Qx\|_2 = \|x\|_2$
- $Q^TQ = I$
- $Q_1 \in Q_2$ orthog $\Rightarrow Q_1Q_2 = \text{orth}$
- $\text{cond}(Q) = 1$

These properties make orthogonal matrices nice to work with. We've seen examples including:

1) permutation matrices
2) rotation matrices
3) reflection matrices
4) $Q$ from QR factorization

Question: is projection matrix $P = I - \frac{vv^T}{v^Tv}$ orthogonal? **No!**

not even invertible
Orthogonal matrices play a leading role in the singular value decomposition, one of the most celebrated tools in scientific computing.

The SVD can be used for constructing low-rank approximations to matrices, doing principal component analysis, computing matrix ranks & cond. #'s, and even solving linear systems & doing least squares... but it doesn't come cheap!

Motivating idea: the image of the unit sphere under any $m \times n$ matrix is a hyperellipse (ellipse in 2D, ellipsoid in 3D).

A hyperellipse is the surface obtained by
stretching the unit sphere by vectors $\sigma_1, \ldots, \sigma_n$ in orthonormal directions $u_1, \ldots, u_n$.

The vectors $\sigma_i u_i$ are the principal semiaxes of the hyperellipse, with lengths $\sigma_1, \ldots, \sigma_n$.

The vectors $u_i$ are called the left singular vectors, and $v_i$ are called the right singular vectors.

The numbers $\sigma_1, \ldots, \sigma_n$ are the singular values of $A$.

By convention, the $\sigma_i$ are ordered so that

$$0 \leq \sigma_1 \leq \cdots \leq \sigma_n \leq 0$$

If $\text{rank}(A) = r$, exactly $r$ of the $\sigma_i$ will be nonzero.

According to the picture above, the $u_i$ and $v_i$ are related by

$$Av_i = \sigma_i u_i$$

for each $i$. 
In matrix form, this may be written as
\[
A \left( \begin{array}{c}
\frac{1}{\sigma_1} \\
\vdots \\
\frac{1}{\sigma_n}
\end{array} \right) = \left( \begin{array}{c}
\frac{1}{\sigma_1} \\
\vdots \\
\frac{1}{\sigma_n}
\end{array} \right) \left( \begin{array}{c}
\hat{U} \\
\hat{V}^T
\end{array} \right)
\]

or
\[
A U V^T = \hat{U} \hat{\Sigma} V^T
\]

where \( \hat{U} \in \mathbb{C}^{m \times m} \) are orthogonal and \( \hat{\Sigma} \) is diagonal.

We have written the reduced SVD.

As in the QR factorization, there is also a full SVD obtained by extending \( \hat{U} \) to a square matrix and adding \( m - n \) silent rows to \( \hat{\Sigma} \).

Reduced
\[
A = \hat{U} \hat{\Sigma} V^T
\]

Full
\[
A = U \Sigma V^T
\]
Full \( V = UΣV^T \) in Matlab

\[
\hat{U} = \begin{pmatrix}
U_1 & U_2 & \ldots & U_r
\end{pmatrix}
\]

so that its columns form an orthonormal basis of \( \mathbb{R}^m \)

\[
Σ = \begin{pmatrix}
\hat{Σ} \\
0 & \ldots & 0
\end{pmatrix}
\]

SVD allows us to read off many properties of \( A \):

1) \( r = \text{rank}(A) = \# \text{ non-zero } Σ_i \)

2) \( \text{range} (A) = \text{span} (u_1, \ldots, u_r) \)

3) \( \text{null} (A) = \text{span} (u_{r+1}, \ldots, u_n) \)

4) \( \text{cond} (A) = Σ_1 / Σ_n \)

(examples)

a) \( R(θ) = \begin{pmatrix}
\cos θ & -\sin θ \\
\sin θ & \cos θ
\end{pmatrix} \)

\( R^TR = I \), so \( R \) is orthogonal \( u ∈ \mathbb{V} \)
\[ \implies \text{SVD of } R \text{ may be written } R = R U V^T \]

i.e. SVD is not unique.

However, the set of \( \sigma_i \) are unique and corresponding \( U_i \in \mathbb{R}^n \) are unique up to sign for non-degenerate \( \sigma_i \).

\[ b) \quad A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} I I^T = I I \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \]

\[ c) \quad \text{Note that } A^T A = (U \Sigma V^T)^T (U \Sigma V^T) = V \Sigma^T U^T U \Sigma V^T = V \Sigma^2 V^T \]

Similarly, \( A A^T = U \Sigma^2 U^T \)

These are eigenvalue decompositions of \( A A^T \) and \( A^T A \) with eigenvalues \( \sigma_i^2 \). Follows that

\[ \sigma_i = \sqrt{\lambda_i}, \quad \text{where } \lambda_i \text{ are the eigenvalues}. \]
of $A^TA$ or $AA^T$. Similarly, the right singular vectors $v_i$ are the corresponding eigenvectors of $A^TA$.

This can be used to compute the SVD of small matrices by hand. We see handout online.

We'll focus on the cases like (a) and (b) above where you can essentially guess the right factorization.

Don't forget about the fundamental equations

$$A v_i = \sigma_i u_i \quad \text{for } i \in \{1, 2, \ldots, k\}$$

orthonormal set, i.e. if $A$ is $2 \times 2$, just need to find $v_1, v_2$ such that $u_1 \perp u_2 \ L v_1, L v_2$

Low-rank approximation

It follows from the SVD that $A$ is a sum of rank-1 matrices

$$A \approx \sum \sigma_i u_i v_i^T$$
To derive this formula, note that
\[
\sum_{i=1}^{n} \begin{pmatrix} 0 \\ \sigma_i \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \sigma_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \sigma_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots + \begin{pmatrix} 0 \\ \sigma_n \\ 0 \\ \vdots \\ 0 \end{pmatrix}
\]
so
\[
\hat{\mathbf{U}} \hat{\Sigma} = \begin{pmatrix} 1 \\ \sigma_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \sigma_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots
\]
and
\[
\hat{\mathbf{U}} \hat{\Sigma} \mathbf{V}^T = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + \sigma_n \mathbf{u}_n \mathbf{v}_n^T
\]
since for the \(i\)th term in the above sum, all rows of \(\mathbf{V}^T\) are multiplied by \(\sigma_i\) except for row \(i\).

Since \(\sigma_1 \geq \sigma_2 \geq \cdots\), the first term in the sum has the largest influence on \(\mathbf{A}\). If we keep \(p\) terms in the sum, we have a rank-\(p\) approximation
\[
\hat{\mathbf{A}}_p = \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^T
\]
This is, however, not symmetrically. The idea is that
Thinking geometrically, the idea is that the best way to approximate a hyperellipse by a line segment is to take its longest axis, best way to approx. by an ellipse is to take its two longest axes, etc.

In fact, recalling that \( \|A\| \) is defined by \( \|A\| = \max \|Ax\|, \quad \|x\| = 1 \), i.e. the maximum amount by which \( A \) stretches the unit sphere,

\[
\|A\|_2 = \sigma_1
\]

by the geometric definition of the SVD.

Therefore \( \|A - A_p\|_2 = \sigma_{p+1} \),

so, the rank-\( p \) approximations get
closer to $A^\dagger$ as $p$ increases since the $0$'s are decreasing

If $p \ll n$, $A^\dagger$ requires much less storage than $A$

SVD can be used for data compression for this reason. The same principle can be used to extract the most essential features from a data set

**Principal Component Analysis**

Suppose we collect the fastest times on the 50m butterfly & 200m freestyle from a group of swimmers

$$A = \begin{pmatrix} 50 \text{ fly times} & 200 \text{ free times} \end{pmatrix}$$

We expect these times to be highly
correlated, since strong swimmers were generally have faster times in both events than weak swimmers (insert self-deprecating joke here)

Therefore, we expect a single underlying component, "swimming skill," to approximate both times.

Since $A$ is $m \times 2$, $A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$ and $A_1 = \sigma_1 u_1 v_1^T$ is the rank-1 approximation, which predicts

- 50 fly time $\approx v_{11} \cdot \text{skill}$, where
- 200 free time $\approx v_{21} \cdot \text{skill}$, \text{skill} = \sigma_1 u_1

Next: application of SVD to solving linear systems & doing least squares (the pseudo-inverse) & discussion of complexity & comparison of SVD to eigenvalue
Solving linear systems by the SVD

If $A$ is square and invertible,
\[
A^{-1} = \frac{1}{\sigma} (U \Sigma V^T)^{-1} = (V^T)^{-1} \Sigma^{-1} U^{-1} = V \Sigma^{-1} U^T,
\]

where 
\[
\Sigma^{-1} = \begin{pmatrix}
\frac{1}{\sigma_1} & & \\
& \ddots & \\
& & \frac{1}{\sigma_n}
\end{pmatrix}
\]

This gives a straightforward way to solve $A x = b$ via $x = V \Sigma^{-1} U^T b$

(Note that $\frac{1}{\sigma_n}$ is the largest singular value of $A^{-1}$, which is why $\text{cond}(A) = \|A^{-1}A\|_1 = \frac{1}{\frac{1}{\sigma_1} \cdots \frac{1}{\sigma_n}} = \frac{\sigma_1}{\sigma_n}$)

Solving least squares by the SVD

As mentioned before, the left singular vectors $u_i$ in $\hat{U}$ in the SVD $A = \hat{U} \Sigma \hat{V}^T$ form an orthonormal basis of the range of $A$

There the least squares solution satisfies $\hat{U}^T A x = \hat{U}^T b$. 
\[ \iff \hat{\Sigma} \hat{V} \hat{x} = \hat{U}^T \hat{b} \\iff \hat{x} = \hat{V} \hat{\Sigma}^{-1} \hat{U}^T \hat{b} \]

This is essentially the same equation as the one we had above for a square system.

For this reason, \( A^+ = \hat{V} \hat{\Sigma}^{-1} \hat{U}^T \) is called the pseudo inverse.

Formally, the above definition is mathematically equivalent to \( A^+ = (A^T A)^{-1} A^T \), but has better conditioning.

Note that if \( A \) is square and invertible, \( A^+ = (A^T A)^{-1} A^T = A^{-1} (A^T A)^{-1} = A^{-1} \).

The pseudo inverse allows us to generalize the definition of condition number to non-square matrices via

\[ \text{cond} (A) = \| A \| \| A^+ \| \]

As we've now seen, the SVD can be used to perform all the linear algebra we've discussed so far. Why not rely on it exclusively?
Computational cost!

Computing the SVD involves an eigenvalue problem for $A^T A$:

- Complexity of SVD $\approx 2mn \sqrt{\frac{2}{3}} n^3$
- $QR$ $\approx 2mn \sqrt{\frac{2}{3}} n^3$
- $LU$ $\approx \frac{2}{3}mn^3$

It's not the right tool for every problem.

Context: comparing the SVD & eigenvalue decomp.

There are closely related. If $A^T = A$, the $A$ has eigenval: $A = Q \Lambda Q^T$ for $Q$ orthog.

SVD: $A = U \Sigma V^T$, where $\sigma_i = |\lambda_i|$

$U_i = \pm \frac{q_i}{\sqrt{\xi}}$

But there are key differences:

**Eigenvalue decomposition**

$A = XXX^{-1}$ \text{ diag.} $\Lambda X$ inv.

**SVD**

$A = U \Sigma V^T$ \text{ diag.} $\Sigma$ \text{ orthog.}

Two bases (left, right)
one basis (eigenvectors) \( \ldots \)

- \( X \) not nec. orthog.
- \( U \& V \) orthog.
- always exists

\((\epsilon, \lambda) A = (0, 1) \; \lambda e_1 = \epsilon_1 \)
\( \lambda_1 = \lambda_2 = 1 \; \Rightarrow A (v_1) = (v_1) \Rightarrow v_2 \neq 0 \)

Finally, the singular values are less sensitive to the matrix entries than the eigenvalues.

Suppose \( \Lambda = X^{-1}AX \). For some perturbation \( A \rightarrow A + \delta A \), what is the corresponding change \( \Lambda \rightarrow \Lambda + \delta \Lambda \)?

\[ \delta \Lambda = X^{-1} \delta A X, \text{ so } \| \delta \Lambda \| \leq \| X^{-1} \| \| \delta A \| \| X \| = \text{cond}(X) \| \delta A \| \]

The sensitivity of \( \Lambda \) to perturbations depends on condition \# of the matrix of eigenvectors.

Contrast with the SVD:

\[ \delta \Sigma = \delta U^T \delta A \delta V \Rightarrow \| \delta \Sigma \| = \| \delta A \|, \]

since \( \| U \| = \| V \| = 1 \) by orthogonality.

Singular value problem is always well-conditioned.
We'll return to eigenvalue problems after the matrix...