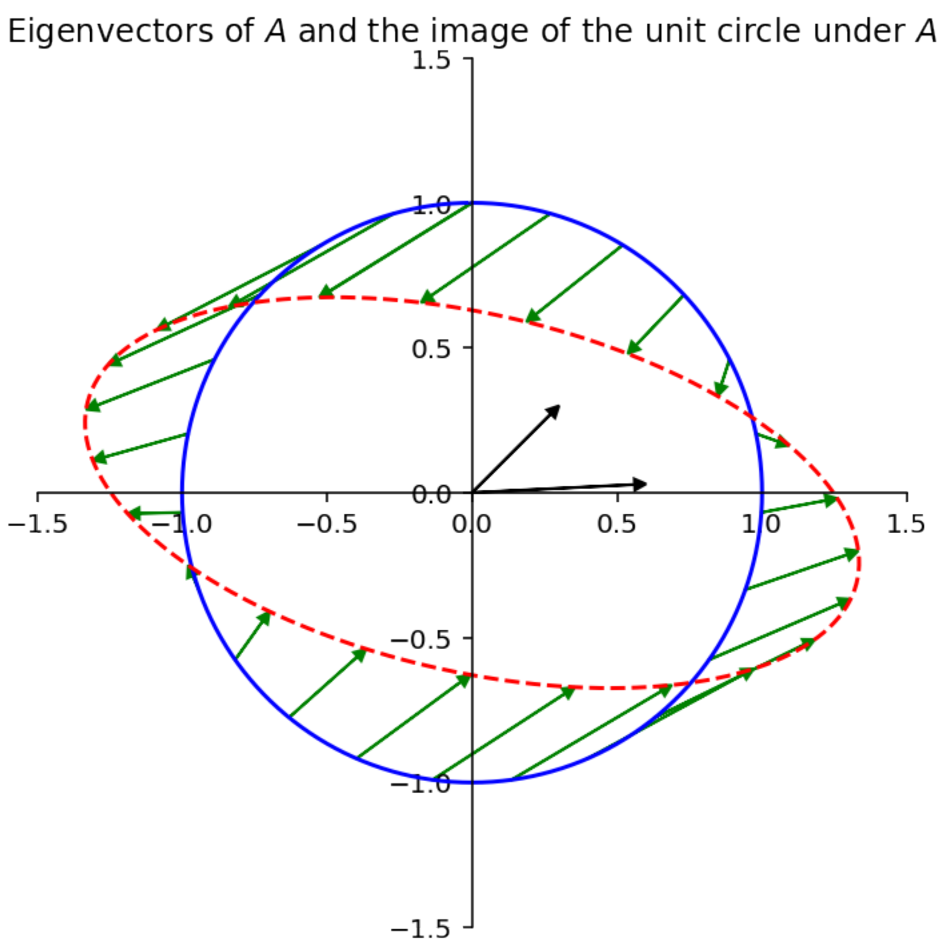
The Singular Value Decomposition

Today we'll begin our study of the most useful decomposition in applied Linear Algebra.
Pretty exciting, eh?

> The Singular Value Decomposition is the **"Swiss Army Knife"** and the **"Rolls Royce"** of matrix decompositions.

– Diane O'Leary
The singular value decomposition is a matrix factorization.
Now, the first thing to know is that **EVERY** matrix has a singular value decomposition.
The singular value decomposition (let's just call it SVD) is based on a very simple question:
Let's say you are given an arbitrary matrix A, which does not need to be square.
Here is the question:
Among all unit vectors, what is the vector $\mathbf{x}$ that maximizes $\|A\mathbf{x}\|$?
In other words, in which direction does A create the largest output vector from a unit input?
To set the stage to answer this question, let's review a few facts.
You recall that the eigenvalues of a **square** matrix A measure the amount that A "stretches or shrinks" certain special vectors (the eigenvectors).
For example, for a square A, if $A\mathbf{x} = \lambda\mathbf{x}$ and $\|\mathbf{x}\| = 1$, then

$$\|A\mathbf{x}\| = \|\lambda\mathbf{x}\| = |\lambda|\,\|\mathbf{x}\| = |\lambda|.$$



The **largest** value of $\|A\mathbf{x}\|$ is the long axis of the ellipse. Clearly there is some $\mathbf{x}$ that is mapped to that point by A. That $\mathbf{x}$ is what we want to find.

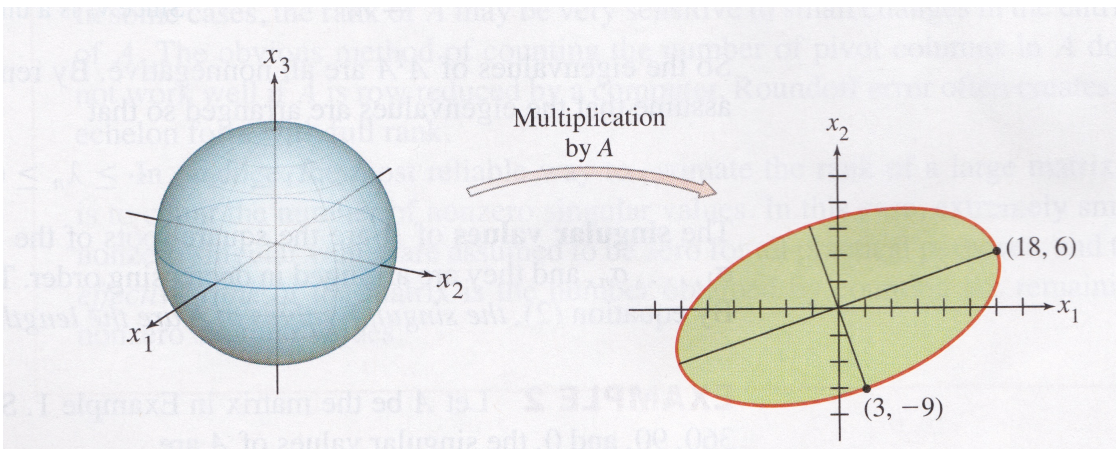
And let's make clear that we can apply this idea to **arbitrary** (non-square) matrices.

Here is an example that shows that we can still ask the question of what unit $\mathbf{x}$ maximizes $\|A\mathbf{x}\|$ even when A is not square.

For example:

If $A = \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix}$,

then the linear transformation $\mathbf{x} \mapsto A\mathbf{x}$ maps the unit sphere $\{\mathbf{x} : \|\mathbf{x}\| = 1\}$ in $\mathbb{R}^3$ onto an ellipse in $\mathbb{R}^2$, as shown here:



Now, here is a way to answer our question:

Problem. Find the unit vector $\mathbf{x}$ at which the length $\|A\mathbf{x}\|$ is maximized, and compute this maximum length.

Solution.

The quantity $\|A\mathbf{x}\|^2$ is maximized at the same $\mathbf{x}$ that maximizes $\|A\mathbf{x}\|$, and $\|A\mathbf{x}\|^2$ is easier to study.

So let's ask to find the unit vector $\mathbf{x}$ at which $\|A\mathbf{x}\|^2$ is maximized.

Observe that

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

$$= \mathbf{x}^T A^T A\mathbf{x}$$

$$= \mathbf{x}^T (A^T A)\mathbf{x}$$

Now, A^TA is a symmetric matrix.

So we see that $\|A\mathbf{x}\|^2 = \mathbf{x}^T A^T A\mathbf{x}$ is a quadratic form!

... and we are seeking to maximize it subject to the constraint $\|\mathbf{x}\| = 1$.

As we learned in the last lecture, the maximum value of a quadratic form, subject to the constraint that $\|\mathbf{x}\| = 1$, is the largest eigenvalue of the symmetric matrix.

So the maximum value of $\|A\mathbf{x}\|$ subject to $\|\mathbf{x}\| = 1$ is λ_1, the largest eigenvalue of A^TA.

Also, the maximum is attained at a unit eigenvector of A^TA corresponding to λ_1.

For the matrix A in the 2×3 example,

$$A^T A = \begin{bmatrix} 4 & 8 \\ 11 & 7 \\ 14 & -2 \end{bmatrix} \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix} = \begin{bmatrix} 80 & 100 & 40 \\ 100 & 170 & 140 \\ 40 & 140 & 200 \end{bmatrix}.$$

The eigenvalues of $A^T A$ are $\lambda_1 = 360, \lambda_2 = 90$, and $\lambda_3 = 0$.

The corresponding unit eigenvectors are, respectively,

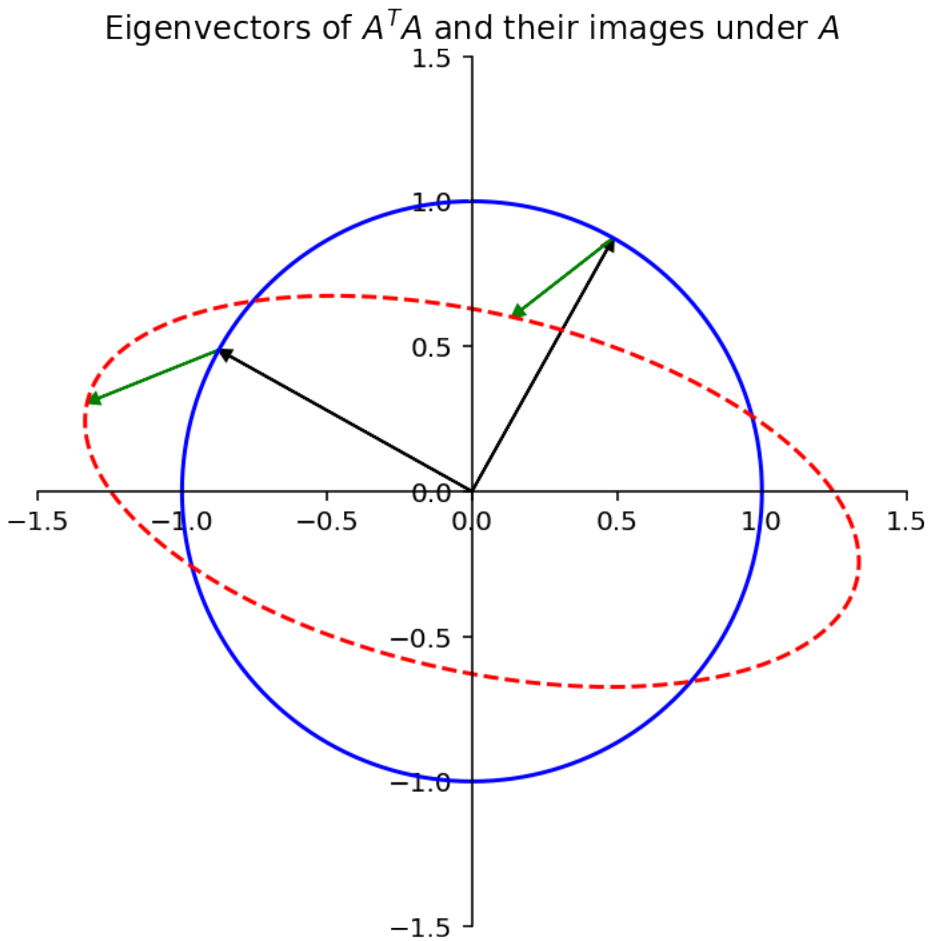
$$\mathbf{v}_1 = \begin{bmatrix} 1/3 \\ 2/3 \\ 2/3 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} -2/3 \\ -1/3 \\ 2/3 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} 2/3 \\ -2/3 \\ 1/3 \end{bmatrix}.$$

For $\|\mathbf{x}\| = 1$, the maximum value of $\|A\mathbf{x}\|$ is $\|A\mathbf{v}_1\| = \sqrt{360}$.

This example shows that the key to understanding the effect of A on the unit sphere in $\mathbb{R}^3$ is to examime the quadratic form $\mathbf{x}^T(A^T A)\mathbf{x}$.

We can also go back to our 2×2 example.

Let's plot the eigenvectors of $A^T A$.



We see that the eigenvector corresponding to the largest eigenvalue of $A^T A$ indeed shows us where $\|A\mathbf{x}\|$ is maximized – where the ellipse is longest.

Also, the other eigenvector of $A^T A$ shows us where the ellipse is narrowest.

In fact, the entire geometric behavior of the transformation $\mathbf{x} \mapsto A\mathbf{x}$ is captured by the quadratic form $\mathbf{x}^T A^T A\mathbf{x}$.

The Singular Values of a Matrix

Let's continue to consider A to be an arbitrary $m \times n$ matrix.

Notice that even though A is not square in general, A^TA is square and **symmetric.**

So, there is a lot we can say about A^TA.

In particular, since A^TA is symmetric, it can be **orthogonally diagonalized** (as we saw in the last lecture).

So let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be an orthonormal basis for $\mathbb{R}^n$ consisting of eigenvectors of A^TA, and let $\lambda_1, \ldots, \lambda_n$ be the corresponding eigenvalues of A^TA.

Then, for any eigenvector $\mathbf{v}_i$,

$$\|A\mathbf{v}_i\|^2 = (A\mathbf{v}_i)^T A\mathbf{v}_i = \mathbf{v}_i^T A^T A\mathbf{v}_i$$

$$= \mathbf{v}_i^T(\lambda_i)\mathbf{v}_i$$

(since $\mathbf{v}_i$ is an eigenvector of A^TA)

$$= \lambda_i$$

(since $\mathbf{v}_i$ is a unit vector.)

Now any expression $\|\cdot\|^2$ is nonnegative.

So the eigenvalues of A^TA are all nonnegative.

That is: A^TA is **positive semidefinite.**

We can therefore renumber the eigenvalues so that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0.$$

Definition. The **singular values** of A are the square roots of the eigenvalues of A^TA. They are denoted by $\sigma_1, \ldots, \sigma_n$, and they are arranged in decreasing order.

That is, $\sigma_i = \sqrt{\lambda_i}$ for $i = 1, \ldots, n$.

By the above argument, **the singular values of** A **are the lengths of the vectors** $A\mathbf{v}_1, \ldots, A\mathbf{v}_n$.

The Eigenvectors of A^TA are an orthogonal basis for $\operatorname{Col} A$

Now: we know that vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are an orthogonal set because they are eigenvectors of the symmetric matrix A^TA.

However, it's **also** the case that $A\mathbf{v}_1, \ldots, A\mathbf{v}_n$ are an orthogonal set.

This fact is key to the SVD.

This fact is not obvious at first!

But it is true – let's prove it (and a bit more).

Theorem. Suppose $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of A^TA, arranged so that the corresponding eigenvalues of A^TA satisfy $\lambda_1 \geq \cdots \geq \lambda_n$, and suppose A has r nonzero singular values.

Then $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_r\}$ is an orthogonal basis for $\operatorname{Col} A$, and rank $A = r$.

Proof. What we need to do is establish that $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_r\}$ is an orthogonal linearly independent set whose span is $\operatorname{Col} A$.

Because $\mathbf{v}_i$ and $\mathbf{v}_j$ are orthogonal for $i \neq j$,

$$(A\mathbf{v}_i)^T(A\mathbf{v}_j) = \mathbf{v}_i^T A^T A\mathbf{v}_j = \mathbf{v}_i^T(\lambda_j\mathbf{v}_j) = 0.$$

So $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_n\}$ is an orthogonal set.

Furthermore, since the lengths of the vectors $A\mathbf{v}_1, \ldots, A\mathbf{v}_n$ are the singular values of A, and since there are r nonzero singular values, $A\mathbf{v}_i \neq \mathbf{0}$ if and only if $1 \leq i \leq r$.

So $A\mathbf{v}_1, \ldots, A\mathbf{v}_r$ are a linearly independent set (because they are orthogonal and all nonzero), and clearly they are each in $\operatorname{Col} A$.

Finally, we just need to show that Span $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_r\}$ = $\operatorname{Col} A$.

To do this we'll show that for any $\mathbf{y}$ in $\operatorname{Col} A$, we can write $\mathbf{y}$ in terms of $\{A\mathbf{v}_1, \ldots, A\mathbf{v}_r\}$:

Say $\mathbf{y} = A\mathbf{x}$.

Because $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is a basis for $\mathbb{R}^n$, we can write $\mathbf{x} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$, so

$$\mathbf{y} = A\mathbf{x} = c_1A\mathbf{v}_1 + \cdots + c_rA\mathbf{v}_r + \cdots + c_nA\mathbf{v}_n.$$

$$= c_1A\mathbf{v}_1 + \cdots + c_rA\mathbf{v}_r.$$

(because $A\mathbf{v}_i = \mathbf{0}$ for $i > r$).

In summary: $\{A\mathbf{v}_1, \dots, A\mathbf{v}_n\}$ is an (orthogonal) linearly independent set whose span is Col A, so it is an (orthogonal) basis for Col A.

Notice that we have also proved that rank $A = \dim \text{Col } A = r$.

In other words, if A has r nonzero singular values, A has rank r.

The Singular Value Decomposition

What we have just proved is that the eigenvectors of A^TA are rather special.

Note that, thinking of A as a linear operator: * its domain is $\mathbb{R}^n$, and * its range is Col A.

So we have just proved that * the set $\{\mathbf{v}_i\}$ is an orthogonal basis for the domain of A, and * the set $\{A\mathbf{v}_i\}$ is an orthogonal basis for the range of A.

Now we can define the SVD.

Theorem. Let A be an $m \times n$ matrix with rank r. Then there exists an $m \times n$ matrix Σ whose diagonal entries are the first r singular values of A, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, and there exists an $m \times m$ orthogonal matrix U and an $n \times n$ orthogonal matrix V such that

$$A = U\Sigma V^T$$

Any factorization $A = U\Sigma V^T$, with U and V orthogonal and Σ a diagonal matrix is called a **singular value decomposition (SVD)** of A.

The columns of U are called the **left singular vectors** and the columns of V are called the **right singular vectors** of A.

Aside: regarding the "Rolls Royce" property, consider how elegant this structure is.

In particular:

- A is an arbitrary matrix
- U and V are both **orthogonal** matrices
- Σ is a **diagonal** matrix
- all singular values are **positive or zero**
- there are as many **positive** singular values as the rank of A
  - (not part of the theorem but we'll see it is true)

We have built up enough tools now that the proof is quite straightforward.

Proof. Let λ_i and $\mathbf{v}_i$ be the eigenvalues and eigenvectors of A^TA, and $\sigma_i = \sqrt{\lambda_i}$.

As we have seen, $\{A\mathbf{v}_1, \dots, A\mathbf{v}_r\}$ is an orthogonal basis for Col A.

Normalize each $A\mathbf{v}_i$ to obtain an orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_r\}$, where

$$\mathbf{u}_i = \frac{1}{\|A\mathbf{v}_i\|}A\mathbf{v}_i = \frac{1}{\sigma_i}A\mathbf{v}_i$$

Then

$$A\mathbf{v}_i = \sigma_i\mathbf{u}_i \quad (1 \leq i \leq r)$$

Now add additional orthonormal vectors $\{\mathbf{u}_{r+1} \dots \mathbf{u}_m\}$ to the set so that they span $\mathbb{R}^m$.

Now collect the vectors into matrices.

$$U = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_m\end{bmatrix}$$

and

$$V = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{bmatrix}$$

Recall that these matrices are orthogonal because the $\{\mathbf{v_i}\}$ are orthogonal and the $\{A\mathbf{v_i}\}$ are orthogonal, as we previously proved.

So

$$AV = [A\mathbf{v}_1 \ \cdots \ A\mathbf{v}_r \ \overbrace{\mathbf{0} \cdots \mathbf{0}}^{n-r}]$$

$$= [\sigma_1 \mathbf{u}_1 \ \cdots \ \sigma_r \mathbf{u}_r \ \mathbf{0} \ \cdots \ \mathbf{0}] = U\Sigma.$$

So

$$AV = U\Sigma$$

Now, V is an orthogonal matrix, so multiplying both sides on the right by A^T:

$$U\Sigma V^T = AVV^T = A.$$

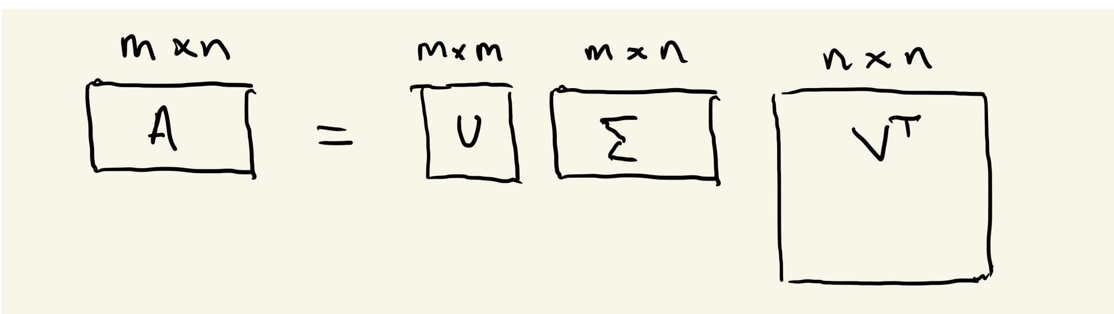
The Reduced SVD and the Pseudoinverse

Let's step back to get a sense of how the SVD decomposes a matrix.

Let's say A is $m \times n$ with $m < n$.

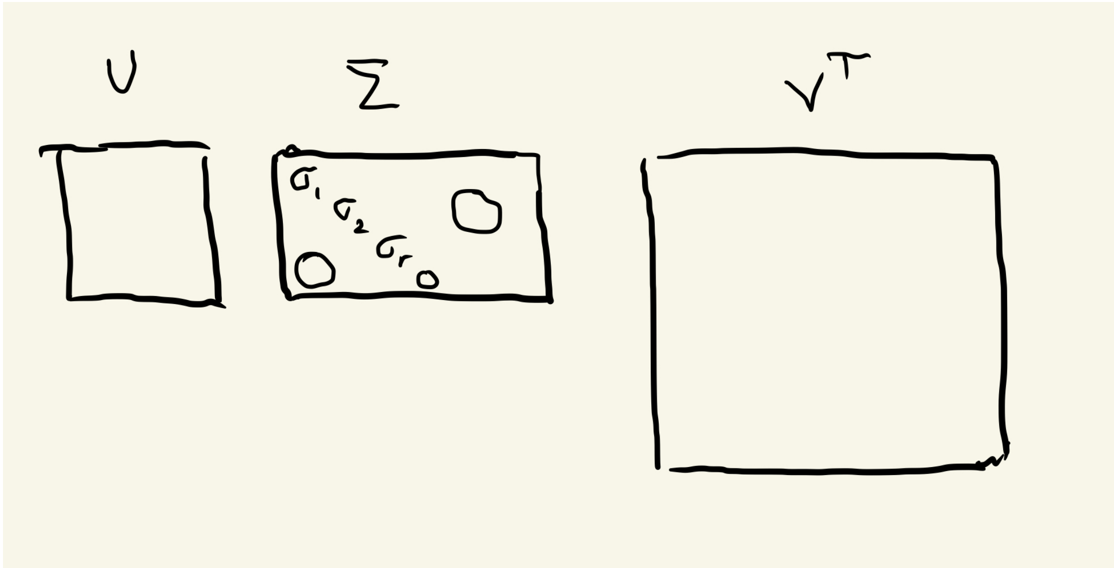
(The situation when $m > n$ follows similarly).

The SVD looks like this, with singular values on the diagonal of Σ:

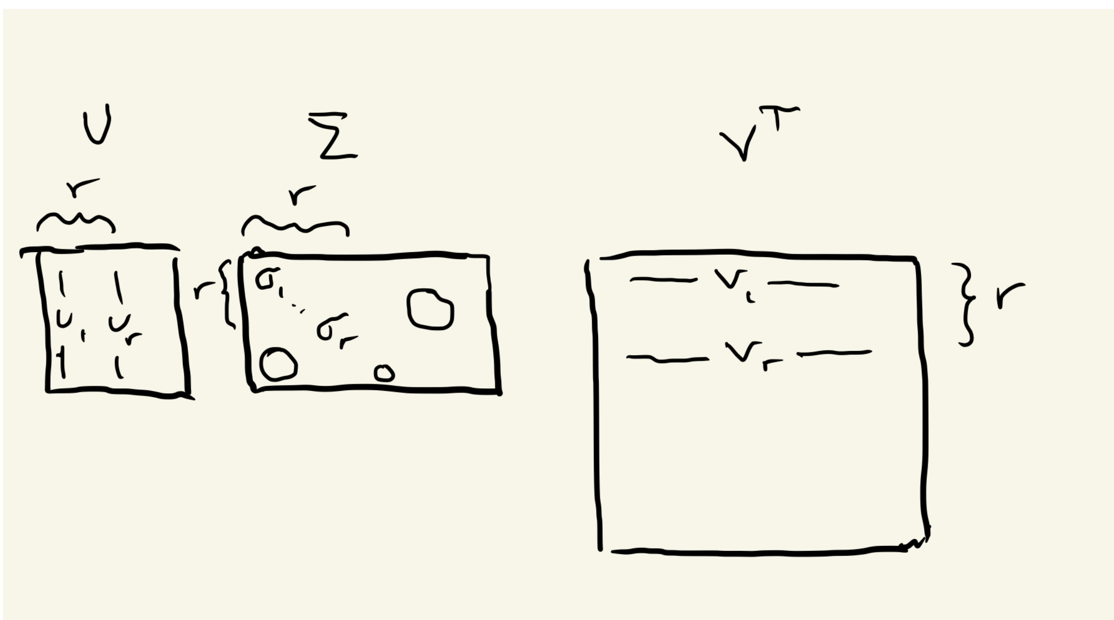


Now, let's assume that the number of nonzero singular values r is less than m.

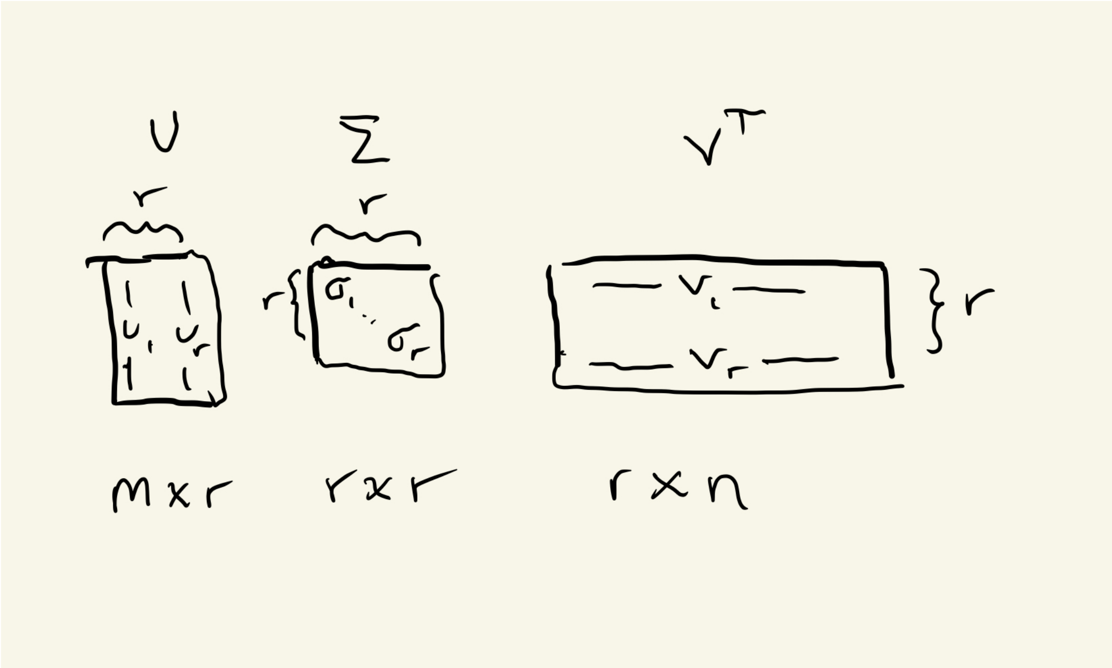
Again, other cases would be similar.



In many cases we are only concerned with representing A.
That is, we don't need U or V to be orthogonal (square) matrices.
Then, to compute A, we only need the r leftmost columns of U, and the r upper rows of V^T.
That's because all the other values on the diagonal of Σ are zero, so they don't contribute anything to A.



So we often work with the **reduced SVD** of A:



Note that in the reduced SVD, Σ has all nonzero entries on its diagonal, so it can be inverted.
However, we still have that $A = U\Sigma V^T$.

The Pseudoinverse

Consider the case where we are working with the reduced SVD of A:

$$A = U\Sigma V^T.$$

In the reduced SVD, Σ is invertible (it is a diagonal matrix with all positive entries on the diagonal).
Using this decomposition we can define an important matrix corresponding to A.

$$A^+ = V\Sigma^{-1}U^T$$

This matrix A^+ is called the **pseudoinverse** of A.
(Sometimes called the Moore-Penrose pseudoinverse).
Obviously, A cannot have an inverse, because it is not even square (let alone invertible) in general.
So why is A^+ called the pseudoinverse?
Let's go back to our favorite equation, $A\mathbf{x} = \mathbf{b}$, specifically in the case where there are no solutions.
In that case, we can find least-squares solutions by finding $\hat{\mathbf{x}}$ such that $A\hat{\mathbf{x}}$ is the projection of $\mathbf{b}$ onto Col A.
And, **if A^TA is invertible,** that $\hat{\mathbf{x}}$ is given by

$$\hat{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b}$$

But, what if A^TA is not invertible?
There are still least-square solutions, but now there are an infinite number.
What if we just want to find **one** of them?
Let's use the pseudoinverse:

$$\hat{\mathbf{x}} = A^+\mathbf{b}$$

Then:

$$A\hat{\mathbf{x}} = AA^{+}\mathbf{b}$$

$$= (U\Sigma V^T)(V\Sigma^{-1}U^T)\mathbf{b}$$

$$= U\Sigma\Sigma^{-1}U^T\mathbf{b}$$

$$= UU^T\mathbf{b}$$

Now, U is an orthonormal basis for Col A.

And, $U^T\mathbf{b}$ are the coefficients of the projection of $\mathbf{b}$ onto each column of U, since the columns are unit length.

So, $UU^T\mathbf{b}$ is the projection of $\mathbf{b}$ onto Col A.

So, $\hat{\mathbf{x}} = A^{+}\mathbf{b}$ is a least squares solution of $A\mathbf{x} = \mathbf{b}$,

even when A^TA is not invertible,

ie, this formula works for **any** A.

Remember, any A has an SVD, and so any A has a pseudoinverse!