CPSC 532D — A. THE SINGULAR VALUE DECOMPOSITION

Danica J. Sutherland

University of British Columbia, Vancouver

Fall 2025

These notes give a quick overview of the singular value decomposition, a concept that when I started grad school I was vaguely aware of and by the end of grad school became basically the main way I think about linear algebra.

(For more, check out the Wikipedia article, Section 4.5 of [DFO20], Section 7E of [Axl25], or Chapter 17 of Roman [Rom07]. The version discussed here is the "compact SVD," which I find way nicer to work with; when you use the "regular" SVD, you very often have to write stuff in block matrices where the part that matters is exactly the compact SVD.)

Let X be an $m \times n$ matrix of rank r, which among other things means that we can write $X = a_1 b_1^\mathsf{T} + \cdots + a_r b_r^\mathsf{T}$ for some vectors $a_i \in \mathbb{R}^m$, $b_i \in \mathbb{R}^n$. The SVD is a particular decomposition like that: it's

$$X = \sigma_1 u_1 v_1^\mathsf{T} + \cdots + \sigma_r u_r v_r^\mathsf{T}$$

where each $\sigma_i > 0$ and the $\{u_i\} \subset \mathbb{R}^m$ and the $\{v_i\} \subset \mathbb{R}^n$ are each orthonormal: $\|u_i\| = 1$, $u_i^\mathsf{T} u_j = 0$ for $i \neq j$. The σ_i are the (nonzero) *singular values*, the u_i are the *left singular vectors*, and the v_i are the *right singular vectors*. This decomposition always exists, for any X.

We can collect these into three matrices, called the *compact SVD*:

$$\overbrace{X}^{m\times n} = \overbrace{U}^{m\times r} \overbrace{\Sigma}^{r\times r} \overbrace{V^\mathsf{T}}^{r\times n}$$

$$U^\mathsf{T} U = I_r \qquad \Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_r) \qquad V^\mathsf{T} V = I_r.$$

This always exists; if we sort the Σ in nonincreasing order, $\sigma_1 \geq \cdots \geq \sigma_r > 0$, then Σ is unique. U and V are not unique, though; you could always e.g. replace u_i by $-u_i$ and v_i by $-v_i$ and still get a valid SVD.

A.1 PROJECTIONS

In this version of the SVD, UU^T is an orthogonal projection, since $(UU^\mathsf{T})^\mathsf{T} = UU^\mathsf{T}$ and $(UU^\mathsf{T})^2 = U(U^\mathsf{T}U)U^\mathsf{T} = UU^\mathsf{T}$. In fact, it is the projection onto the image of X: for any $y \in \mathbb{R}^n$, $UU^\mathsf{T}(Xy) = UU^\mathsf{T}U\Sigma V^\mathsf{T} y = U\Sigma V^\mathsf{T} y = Xy$. On the other hand, suppose that $X^\mathsf{T} z = 0$, i.e. z is in the left null space of X (the orthogonal complement of the image); then $V\Sigma U^\mathsf{T} z = 0$. But if we left-multiply that equation by $\Sigma^{-1} V^\mathsf{T}$, then we get that $U^\mathsf{T} z = 0$, and thus $UU^\mathsf{T} z = 0$.

$\Sigma^{-1} = \operatorname{diag}(1/\sigma_i)$ *always exists.*

Similarly, VV^T is the orthogonal projection onto the row space (the orthogonal complement of the null space).

For more, visit https://cs.ubc.ca/~dsuth/532D/25w1/.

If, say, $r = m$, then the image of X is all of $\mathbb{R}^m$ and we do in fact have $UU^\top = I_m$; similarly if $r = n$ then $VV^\top = I_n$; but if $r < \min(m, n)$ then neither holds.

The SVD is useful for understanding lots of things related to a matrix. For instance,

$$X^\top X = V\Sigma U^\top U\Sigma V^\top = V\Sigma^2 V^\top$$

tells us that the nonzero singular values are the nonzero squared eigenvalues of $X^\top X$, and similarly they are the nonzero squared eigenvalues of $XX^\top$; likewise, the left singular vectors are the eigenvectors of $XX^\top$ and the right singular vectors of $X^\top X$.

A.2 (PSEUDO)INVERSES

If X is invertible, then $m = n = r$ and we know that

$$I = XX^{-1} = U\Sigma V^\top X^{-1};$$

let's left-multiply by $V\Sigma^{-1}U^\top$, getting

$$V\Sigma^{-1}U^\top = V\Sigma^{-1}U^\top U\Sigma V^\top X^{-1} = VV^\top X^{-1}.$$

Because X must be full-rank if it's invertible, in this case we have that $VV^\top = I$, and so this means that if it exists, $X^{-1} = V\Sigma^{-1}U^\top$.

A neat thing is that this formula still makes sense even if X is not invertible – even if it's not square. We call that matrix $X^\dagger$, the *pseudoinverse* of X:

$$X^\dagger = V\Sigma^{-1}U^\top.$$

We just saw that if X is invertible then $X^{-1} = X^\dagger$. But it's also useful much more generally.

For example, in Section 1.1, we saw that ERM for linear regression is equivalent to the condition $\mathbf{X}^\top\mathbf{X}w = \mathbf{X}^\top\mathbf{y}$. Using the SVD, we can write this as

$$V\Sigma^2 V^\top w = V\Sigma U^\top \mathbf{y};$$

left-multiplying by $V\Sigma^{-2}V^\top$ gives the condition

$$VV^\top w = V\Sigma^{-1}U^\top\mathbf{y} = X^\dagger\mathbf{y}.$$

Now, anything of the form $w = X^\dagger\mathbf{y} + (I - VV^\top)z$, where $z \in \mathbb{R}^d$ is arbitrary, satisfies this condition, since $VV^\top X^\dagger = X^\dagger$ and $VV^\top(I - VV^\top) = 0$. If X is rank d, then $VV^\top = I$, so no matter the choice of z we have $w = X^\dagger\mathbf{y}$. Otherwise, there are infinitely many solutions, taking $X^\dagger\mathbf{y}$ and then allowing an arbitrary component in the null space of X.

It's worth noting that of these solutions, $X^\dagger\mathbf{y}$ is the one with minimum norm. Note that

$$\langle VV^\top a, (I - VV^\top)b\rangle = a^\top VV^\top(I - VV^\top)b = a^\top(VV^\top - VV^\top)b = 0,$$

because indeed the row space and its orthogonal complement are orthogonal. Since $VV^\top X^\dagger = X^\dagger$, we have that

$$\begin{aligned}\|X^\dagger\mathbf{y} + (I - VV^\top)z\|^2 &= \|X^\dagger\mathbf{y}\|^2 + \|(I - VV^\top)z\|^2 + 2\langle X^\dagger\mathbf{y}, (I - VV^\top)z\rangle \\ &= \|X^\dagger\mathbf{y}\|^2 + \|(I - VV^\top)z\|^2;\end{aligned}$$

the first term here doesn't depend on the choice of z, while the second term is always nonnegative and is zero iff $(\mathrm{I} - \mathrm{V}\mathrm{V}^\mathsf{T})z = 0$. So, we uniquely minimize the norm with $w = \mathrm{X}^\dagger \mathbf{y}$.

There are potentially many choices of z here, but they all yield the same w.

This indeed agrees with the usual way of writing the solution that you might remember; if $m \geq d$ and X is full rank ($r = d$), then $\mathrm{X}^\mathsf{T}\mathrm{X} = \mathrm{V}\Sigma^2\mathrm{V}^\mathsf{T}$ has inverse $\mathrm{V}\Sigma^{-2}\mathrm{V}^\mathsf{T}$, because

$$(\mathrm{V}\Sigma^{-2}\mathrm{V}^\mathsf{T})(\mathrm{V}\Sigma^2\mathrm{V}^\mathsf{T}) = \mathrm{V}\mathrm{V}^\mathsf{T} = \mathrm{I}_d,$$

and thus

$$(\mathrm{X}^\mathsf{T}\mathrm{X})^{-1}\mathrm{X}^\mathsf{T} = \mathrm{V}\Sigma^{-2}\mathrm{V}^\mathsf{T}\mathrm{V}\Sigma\mathrm{U}^\mathsf{T} = \mathrm{V}\Sigma^{-1}\mathrm{U}^\mathsf{T} = \mathrm{X}^\dagger.$$

This helps us understand better the sense in which $\mathrm{X}^\dagger$ is a "pseudo-"inverse:

$$\mathrm{X}\mathrm{X}^\dagger = \mathrm{U}\Sigma\mathrm{V}^\mathsf{T}\mathrm{V}\Sigma^{-1}\mathrm{U}^\mathsf{T} = \mathrm{U}\mathrm{U}^\mathsf{T}$$

is the identity if $r = m$, but otherwise is the projection onto the image of X; similarly $\mathrm{X}^\dagger\mathrm{X} = \mathrm{V}\mathrm{V}^\mathsf{T}$ is the identity if $r = n$, but otherwise is the projection onto the row space of X.

A.3 NORMS

Recall that the Frobenius inner product is given by

$$\langle \mathrm{A}, \mathrm{B}\rangle_\mathrm{F} = \sum_{ij} \mathrm{A}_{ij}\mathrm{B}_{ij} = \sum_i \left(\sum_j \mathrm{A}_{ij}(\mathrm{B}^\mathsf{T})_{ji} \right) = \sum_i (\mathrm{A}\mathrm{B}^\mathsf{T})_{ii} = \mathrm{Tr}(\mathrm{A}\mathrm{B}^\mathsf{T}),$$

and so the squared Frobenius norm of X, $\|\mathrm{X}\|_\mathrm{F}^2 = \sum_{ij} \mathrm{X}_{ij}^2$, can also be computed as

$$\|\mathrm{X}\|_\mathrm{F}^2 = \mathrm{Tr}(\mathrm{X}^\mathsf{T}\mathrm{X}) = \mathrm{Tr}(\mathrm{V}\Sigma^2\mathrm{V}^\mathsf{T}) = \mathrm{Tr}(\mathrm{V}^\mathsf{T}\mathrm{V}\Sigma^2) = \mathrm{Tr}(\Sigma^2) = \sum_i \sigma_i^2.$$

In the middle here we used trace rotation,

$$\mathrm{Tr}(\mathrm{A}\mathrm{B}) = \sum_i \sum_j \mathrm{A}_{ij}\mathrm{B}_{ji} = \sum_j \sum_i \mathrm{B}_{ji}\mathrm{A}_{ij} = \mathrm{Tr}(\mathrm{B}\mathrm{A}).$$

The operator norm, $\|\mathrm{X}\| = \sup_{y : \|y\| \leq 1} \|\mathrm{X}y\|$, also has a nice expression in terms of the singular values: we can write

$$\|\mathrm{X}\|^2 = \sup_{y:\|y\|\leq 1} \|\mathrm{U}\Sigma\mathrm{V}^\mathsf{T}y\|^2 = \sup_{y:\|y\|\leq 1} y^\mathsf{T}\mathrm{V}\Sigma^2\mathrm{V}^\mathsf{T}y.$$

Now,

$$\|\mathrm{V}^\mathsf{T}y\|^2 = y^\mathsf{T}\mathrm{V}\mathrm{V}^\mathsf{T}y = \|(\mathrm{V}\mathrm{V}^\mathsf{T})y\|^2 \leq \|(\mathrm{V}\mathrm{V}^\mathsf{T})y\|^2 + \|(\mathrm{I} - \mathrm{V}\mathrm{V}^\mathsf{T})y\|^2 = \|y\|^2,$$

and so $\{\mathrm{V}^\mathsf{T}y : \|y\| \leq 1\} \subseteq \{z : \|z\| \leq 1\}$. Thus we have that

$$\|\mathrm{X}\|^2 \leq \sup_{z:\|z\|\leq 1} z\Sigma^2 z = \sup_{z:\|z\|\leq 1} \sum_i \sigma_i^2 z_i^2 \leq \sup_{z:\|z\|\leq 1} \sigma_1^2 \sum_i z_i^2 = \sigma_1^2,$$

since $\sigma_1 \geq \sigma_i$ for all i. Moreover, we can achieve this upper bound by picking $y = v_1$, getting $\mathrm{V}^\mathsf{T}y = e_1$. Thus $\|\mathrm{X}\| = \sigma_1$.

REFERENCES

[Axl25] Sheldon Axler. *Linear Algebra Done Right*. 4th edition. 2025.

[DFO20] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. *Mathematics for Machine Learning*. 2020.

[Rom07] Steven Roman. *Advanced Linear Algebra*. 3rd edition. 2007.