Note 15: Singular Value Decomposition

1 Singular Value Decomposition

First, we will introduce the singular value decomposition (SVD) as a particular matrix factorization. We will do this in *stages* of increasing levels of complexity and utility; each stage corresponds to a different *form* of the SVD. We will show how each form of of the decomposition allows us to read off and manipulate the linear algebraic properties of a matrix. Finally, we will give algorithms to compute each form of the SVD.

1.1 Preliminaries

In order to introduce the SVD, we introduce a result without which the SVD properties do not make sense.

Proposition 1 (Eigenvalues of $A^\top A$ and $AA^\top$)
Let $A \in \mathbb{R}^{m \times n}$ have rank $r \leq \min\{m, n\}$. Then $A^\top A \in \mathbb{R}^{n \times n}$ and $AA^\top \in \mathbb{R}^{m \times m}$ are symmetric **positive semi-definite matrices**. By definition, the eigenvalues of positive semi-definite matrices are all real and greater than or equal to zero. $A^\top A$ and $AA^\top$ are matrices of rank r, so they have exactly r positive eigenvalues, and $(n - r)$ and $(m - r)$ zero-valued eigenvalues respectively.

See Appendix A.1 for the proof.

1.2 Outer Product Form of the SVD

Definition 2 (Outer Product Form of the SVD)
Let $A \in \mathbb{R}^{m \times n}$ have rank $r \leq \min\{m, n\}$. An *outer product form of an SVD* of A is a decomposition

$$A = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top \tag{1}$$

where

(i) $\{\vec{u}_1, \ldots, \vec{u}_r\} \subseteq \mathbb{R}^m$ is an orthonormal set of vectors, and are so-called *left singular vectors* of A.

(ii) $\{\vec{v}_1, \ldots, \vec{v}_r\} \subseteq \mathbb{R}^n$ is an orthonormal set of vectors, and are so-called *right singular vectors* of A.

(iii) $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are positive ordered scalars, and are so-called *singular values* of A.

While this form looks quite unassuming, it actually reveals several linear-algebraic properties of A, as we will see in the following theorem.

Theorem 3 (Linear Algebra of the Outer Product SVD)
Let $A \in \mathbb{R}^{m \times n}$ have rank $r \leq \min\{m, n\}$. Let $A = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top$ be an outer product form SVD of A.

(i) $\{\vec{u}_1, \ldots, \vec{u}_r\}$ is an orthonormal basis for $\text{Col}(A)$.

(ii) $\{\vec{v}_1, \ldots, \vec{v}_r\}$ is an orthonormal basis for $\text{Col}(A^\top) = \text{Row}(A)$.

(iii) For each i, $1 \leq i \leq r$, we have $(\sigma_i^2, \vec{u}_i)$ is an eigenvalue-eigenvector pair for $AA^\top$.

(iv) For each i, $1 \leq i \leq r$, we have $(\sigma_i^2, \vec{v}_i)$ is an eigenvalue-eigenvector pair for $A^\top A$.

(v) For each i, $1 \leq i \leq r$, we have $A\vec{v}_i = \sigma_i \vec{u}_i$.

See Appendix A.2 for the proof.

Now, to prove that the SVD exists, we give an algorithm that constructs it, and prove that the algorithm's construction is valid.

Algorithm 4 Construction of the Outer Product Form of SVD

1: **function** OuterProductSVD($A \in \mathbb{R}^{m \times n}$)
2: $\quad r := \text{rank}(A)$
3: $\quad (\vec{v}_1, \lambda_1), \ldots, (\vec{v}_n, \lambda_n) :=$ OrthonormalEigenvectorsWithSortedEigenvalues($A^\top A$)
4: $\quad$ **for** $i \in \{1, \ldots, r\}$ **do**
5: $\quad\quad \sigma_i := \sqrt{\lambda_i}$
6: $\quad\quad \vec{u}_i := \dfrac{1}{\sigma_i} A\vec{v}_i$
7: $\quad$ **end for**
8: $\quad$ **return** $\{\vec{u}_1, \ldots, \vec{u}_r\}, \{\sigma_1, \ldots, \sigma_r\}, \{\vec{v}_1, \ldots, \vec{v}_r\}$
9: **end function**

To clarify, in line 3, the eigenvalues are sorted such that $\lambda_1 \geq \cdots \geq \lambda_n$.
See Appendix A.3 for the proof of correctness.

1.3 Compact SVD

Repackaging the outer product form as a product of matrices results in the so-called *compact SVD*.

Definition 5 (Compact SVD)
Let $A \in \mathbb{R}^{m \times n}$ have rank $r \leq \min\{m, n\}$. A *compact SVD* of A is a decomposition

$$A = U_r \Sigma_r V_r^\top \tag{2}$$

where

(i) $U_r \in \mathbb{R}^{m \times r}$ is a matrix with orthonormal columns, which are so-called *left singular vectors* of A.

(ii) $V_r \in \mathbb{R}^{n \times r}$ is a matrix with orthonormal columns, which are so-called *right singular vectors* of A.

(iii) $\Sigma_r \in \mathbb{R}^{r \times r}$ is a diagonal matrix whose diagonal entries $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are positive ordered real numbers, and are so-called *singular values* of A.

The compact SVD is connected to the outer product SVD by the following calculation.

$$\begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_r \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} \vec{v}_1 & \cdots & \vec{v}_r \end{bmatrix}^\top = \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_r \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} \vec{v}_1^\top \\ \vdots \\ \vec{v}_r^\top \end{bmatrix} \tag{3}$$

$$= \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_r \end{bmatrix} \begin{bmatrix} \sigma_1 \vec{v}_1^\top \\ \vdots \\ \sigma_r \vec{v}_r^\top \end{bmatrix} \tag{4}$$

$$= \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top. \tag{5}$$

Since the compact and outer product forms of the SVD are mathematically identical, the same theorems hold, and can be presented in matrix form as follows.

Theorem 6 (Linear Algebra of the Compact SVD)
Let $A \in \mathbb{R}^{m \times n}$ have rank $r \leq \min\{m, n\}$. Let $A = U_r \Sigma_r V_r^\top$ be an outer product form SVD of A.

(i) $\mathrm{Col}(U_r) = \mathrm{Col}(A)$.

(ii) $\mathrm{Col}(V_r) = \mathrm{Col}\left(A^\top\right) = \mathrm{Row}(A)$.

(iii) $AA^\top U_r = U_r \Sigma_r^2$.

(iv) $A^\top A V_r = V_r \Sigma_r^2$.

(v) $AV_r = U_r \Sigma_r$.

Concept Check: Prove Theorem 6.

We can also use the same algorithm to construct the compact SVD.

Algorithm 7 Construction of the Compact SVD

```
function CompactSVD(A ∈ ℝ^{m×n})
    r := rank(A)
    (v⃗_1, λ_1), ..., (v⃗_n, λ_n) := OrthonormalEigenvectorsWithSortedEigenvalues(AᵀA)
    for i ∈ {1, ..., r} do
        σ_i := √λ_i
        u⃗_i := (1/σ_i) A v⃗_i
    end for
    U_r := [u⃗_1 ··· u⃗_r]
    V_r := [v⃗_1 ··· v⃗_r]
    Σ_r := diag(σ_1, ..., σ_r)
    return U_r, Σ_r, V_r
end function
```

To clarify, in line 3, the eigenvalues are sorted such that $\lambda_1 \geq \cdots \geq \lambda_n$.

Concept Check: Prove correctness of Algorithm 7.

1.4 Full SVD

In the compact SVD, U_r and V_r are tall matrices with orthonormal columns. In contrast, Σ_r is a square diagonal invertible matrix. We can *trade off* the niceness of Σ_r for additional niceness for U_r and V_r, by extending U_r and V_r into fully square orthonormal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ while padding Σ_r with 0s to get $\Sigma \in \mathbb{R}^{m\times n}$ to make the matrix multiplication work out. Here U and V are square and orthonormal, but Σ is no longer square or invertible. This is called the *full SVD*.

Definition 8 (Full SVD)

Let $A \in \mathbb{R}^{m\times n}$ have rank $r \leq \min\{m, n\}$. A *(full) SVD* of A is a decomposition

$$A = U\Sigma V^\top \tag{6}$$

where

(i) $U = \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \in \mathbb{R}^{m\times m}$ is a square orthonormal matrix whose columns are the so-called *left singular vectors* of A; here $U_r \in \mathbb{R}^{m\times r}$ and $U_{m-r} \in \mathbb{R}^{m\times(m-r)}$ are tall matrices with orthonormal columns.

(ii) $V = \begin{bmatrix} V_r & V_{n-r} \end{bmatrix} \in \mathbb{R}^{n\times n}$ is a square orthonormal matrix whose columns are the so-called *right singular vectors* of A; here $V_r \in \mathbb{R}^{n\times r}$ and $V_{n-r} \in \mathbb{R}^{n\times(n-r)}$ are tall matrices with orthonormal columns.

(iii) $\Sigma = \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix} \in \mathbb{R}^{m\times n}$ is a non-square diagonal matrix whose diagonal entries $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_{\min\{m,n\}} = 0$ are non-negative ordered real numbers, and are the so-called *singular values* of A.

The full SVD is connected to the compact SVD via the following calculation.

$$\begin{aligned}
U\Sigma V^\top &= \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix} \begin{bmatrix} V_r & V_{n-r} \end{bmatrix}^\top &(7)\\
&= \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix} \begin{bmatrix} V_r^\top \\ V_{n-r}^\top \end{bmatrix} &(8)\\
&= \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \left(\begin{bmatrix} \Sigma_r \\ 0_{(m-r)\times r} \end{bmatrix} V_r^\top + \begin{bmatrix} 0_{r\times(n-r)} \\ 0_{(m-r)\times(n-r)} \end{bmatrix} V_{n-1}^\top \right) &(9)\\
&= \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r \\ 0_{(m-r)\times r} \end{bmatrix} V_r^\top &(10)\\
&= \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r V_r^\top \\ 0_{(m-r)\times n} \end{bmatrix} &(11)\\
&= U_r \Sigma_r V_r^\top + U_{m-r} 0_{(m-r)\times n} &(12)\\
&= U_r \Sigma_r V_r^\top. &(13)
\end{aligned}$$

Thus, the sub-matrices of a full SVD are a compact SVD.

We can show the same things as for the outer product and compact forms of the SVD, with further interpretation of U_{m-r} and V_{n-r}.

Theorem 9 (Linear Algebra of the Full SVD)
Let $A \in \mathbb{R}^{m \times n}$ have rank $r \leq \min\{m, n\}$. Let $A = U\Sigma V^\top$ be an SVD of A.

(i) $\mathrm{Col}(U_r) = \mathrm{Col}(A)$.

(ii) $\mathrm{Col}(V_r) = \mathrm{Col}\left(A^\top\right) = \mathrm{Row}(A)$.

(iii) $\mathrm{Col}(U_{m-r}) = \mathrm{Null}\left(A^\top\right)$.

(iv) $\mathrm{Col}(V_{n-r}) = \mathrm{Null}(A)$.

(v) $AA^\top = U\Sigma\Sigma^\top U^\top$.

(vi) $A^\top A = V\Sigma^\top\Sigma V^\top$.

(vii) $AV_r = U_r\Sigma_r$.

See Appendix A.4 for proof.

We can also use the same algorithm to construct the full SVD.

Algorithm 10 Construction of the Full SVD

1: **function** FULLSVD($A \in \mathbb{R}^{m \times n}$)
2: $\quad r := \mathrm{rank}(A)$
3: $\quad (\vec{v}_1, \lambda_1), \ldots, (\vec{v}_n, \lambda_n) :=$ ORTHONORMALEIGENVECTORSWITHSORTEDEIGENVALUES($A^\top A$)
4: $\quad$ **for** $i \in \{1, \ldots, r\}$ **do**
5: $\quad\quad \sigma_i := \sqrt{\lambda_i}$
6: $\quad\quad \vec{u}_i := \dfrac{1}{\sigma_i} A\vec{v}_i$
7: $\quad$ **end for**
8: $\quad U_r := \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_r \end{bmatrix}$
9: $\quad V_r := \begin{bmatrix} \vec{v}_1 & \cdots & \vec{v}_r \end{bmatrix}$
10: $\quad \Sigma_r := \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}$
11: $\quad U :=$ EXTENDBASIS($U_r, \mathbb{R}^m$)
12: $\quad V :=$ EXTENDBASIS($V_r, \mathbb{R}^n$)
13: $\quad \Sigma := \begin{bmatrix} \Sigma_r & 0_{r \times (n-r)} \\ 0_{(m-r) \times r} & 0_{(m-r) \times (n-r)} \end{bmatrix}$
14: $\quad$ **return** U, Σ, V
15: **end function**

To clarify, in line 3, the eigenvalues are sorted such that $\lambda_1 \geq \cdots \geq \lambda_n$.

See Appendix A.5 for proof of correctness.

1.5 Comparison Between the SVD Forms

We identify weaknesses and strengths of using the SVD forms, so you can decide which one to use.

- Outer Product SVD:
  - Weaknesses: The summation notation is messy and sometimes tedious to work with. Also, there is no characterization of null spaces, as in the full SVD.
  - Strengths: It is the most computationally efficient to construct by computer, and also saves the most memory. It is also easier to construct by hand than the full SVD.
- Compact SVD:
  - Weaknesses: The matrices U_r and V_r are not square, so they do not have an inverse. Also, there is no characterization of null spaces, as in the full SVD.
  - Strengths: The matrix Σ_r is square and invertible. It is also easier to construct by hand than the full SVD.
- Full SVD:
  - Weaknesses: It is the most computationally intensive to compute, either by computer or by hand. It requires running Gram-Schmidt twice to find U_{m-r} and V_{n-r}. Also Σ is non-square and thus not invertible.
  - Strengths: The matrices U and V are square orthonormal, and thus invertible. There is also a characterization of null spaces of $A^\top$ and A as the column spaces of U_{m-r} and V_{n-r} respectively.

When using the SVD to compute things, one potential rule to use is to use the compact SVD unless there is a need to analyze null spaces, in which case the full SVD is essentially required.

2 Singular Value Decomposition: Geometric Properties

Let $A = U\Sigma V^\top$ be the SVD of some matrix $A \in \mathbb{R}^{m\times n}$.

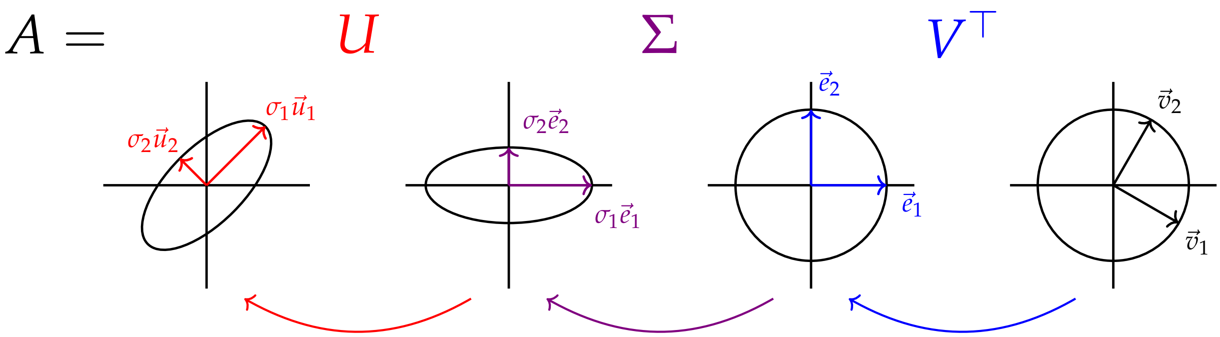
Recall that for any vector $\vec{x}$ and any orthogonal matrix U, it holds that $\|U\vec{x}\| = \|\vec{x}\|$. Additionally, if we consider a second vector $\vec{y}$, the inner product between $\vec{x}$ and $\vec{y}$ remains unchanged when both are multiplied by the same orthogonal matrix, that is, $\langle U\vec{x}, U\vec{y}\rangle = \langle \vec{x}, \vec{y}\rangle$. This property shows that orthogonal matrices do not change the lengths of vectors or the angles between them, implying that multiplication by an orthogonal matrix can be interpreted as performing operations such as rotations and reflections that do not alter lengths or inner products.

Since Σ_r is diagonal with entries $\sigma_1, \ldots, \sigma_r$, multiplying a vector by Σ stretches or compresses each component of the vector according to the corresponding σ value, effectively stretching the first entry of the vector by σ_1, the second entry by σ_2, and so on. In addition to stretching, Σ can also alter the dimensionality of the space by adding or collapsing dimensions.

Combining these observations, we interpret $A\vec{x}$ as the composition of three operations:

1. $V^\top \vec{x}$ which rotates $\vec{x}$ without changing its length.
2. $\Sigma V^\top \vec{x}$ which stretches the resulting vector along each axis with the corresponding singular value and adds or subtracts dimensions.
3. $U\Sigma V^\top \vec{x}$ which again rotates the resulting vector without changing its length.

The following figure illustrates these three operations moving from the right to the left.



Here as usual $\vec{e}_1, \vec{e}_2$ are the first and second standard basis vectors.

The geometric interpretation above reveals that σ_1 is the largest amplification factor a vector can experience upon multiplication by A. More specifically, if $\|\vec{x}\| \leq 1$ then $\|A\vec{x}\| \leq \sigma_1$. We achieve equality at $\vec{x} = \vec{v}_1$, because then

$$\|A\vec{x}\| = \left\|U\Sigma V^\top \vec{v}_1\right\| = \|U\Sigma\vec{e}_1\| = \|\sigma_1 U\vec{e}_1\| = \|\sigma_1\vec{u}_1\| = \sigma_1\|\vec{u}_1\| = \sigma_1. \tag{14}$$

3 The SVD and Orthonormal Diagonalization

We attempt to compare the results of the full SVD to orthonormal diagonalization. The spectral theorem says that, if $A \in \mathbb{R}^{n\times n}$ is symmetric, then there exists some orthonormal square matrix $P \in \mathbb{R}^{n\times n}$ and diagonal matrix $\Lambda \in \mathbb{R}^{n\times n}$ such that $A = P\Lambda P^\top$. It turns out that (up to sign issues and ordering of the singular values) that this is a valid SVD, and thus inherits the linear algebraic properties of the SVD.

Theorem 11 (Orthonormal Diagonalization is SVD)
Let $A \in \mathbb{R}^{n\times n}$ be a square symmetric matrix with orthonormal diagonalization $A = P\Lambda P^\top$, where

$$P = \begin{bmatrix} \vec{p}_1 & \cdots & \vec{p}_n \end{bmatrix} \qquad \Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}. \tag{15}$$

Suppose that $|\lambda_1| \geq \cdots \geq |\lambda_n|$. Define

$$U = \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_n \end{bmatrix} \quad \text{where} \quad \vec{u}_i = \begin{cases} \vec{p}_i & \lambda_i \geq 0 \\ -\vec{p}_i & \lambda_i < 0 \end{cases} \tag{16}$$

$$\Sigma = \begin{bmatrix} |\lambda_1| & & \\ & \ddots & \\ & & |\lambda_n| \end{bmatrix} \tag{17}$$

$$V = P. \tag{18}$$

Then $A = U\Sigma V^\top$ is an SVD of A.

Concept Check: Prove Theorem 11.

4 Minimum-Energy Control

Let's switch gears to focus on an application of SVD related to a topic we have already discussed: the problem of controllability and reachability in discrete-time. Reachability analysis amounted to solving a linear system of the form

$$\mathcal{C}_{i^\star} \begin{bmatrix} \vec{u}[0] \\ \vdots \\ \vec{u}[i^\star - 1] \end{bmatrix} = \vec{x}^\star - A^{i^\star} \vec{x}_0 \tag{19}$$

for the vector quantities $\vec{u}[0], \ldots, \vec{u}[i^\star - 1]$.

There could be many solutions to this vector system, and this translates to many choices for $\vec{u}[0], \ldots, \vec{u}[i^\star - 1]$. To pick the best one, we use the principle of minimum-energy control.

Key Idea 12 (Minimum-Energy Control)
The principle of minimum-energy control says that, when picking one of many choices of inputs, we should pick the one which causes the system to consume the least energy.

This principle turns the problem of reachability into a *constrained optimization problem*. To solve this problem in the abstract, we can use the singular value decomposition (SVD), whose properties have been fleshed out in Sections 1 to 3.

Cleaning up notation in 19, let us fix $i^\star$, let $C := \mathcal{C}_{i^\star}$, let $\vec{z} := \vec{x}^\star - A^{i^\star}\vec{x}_0$, and let $\vec{w} := \begin{bmatrix} \vec{u}[0] \\ \vdots \\ \vec{u}[i^\star - 1] \end{bmatrix}$. Then this linear system becomes

$$C\vec{w} = \vec{z}. \tag{20}$$

This system may have zero solutions, exactly one solution, or infinitely many solutions, depending on the rank and shape of C. In some sense, we already have a good idea of what to do when the system has no solutions or one solution.

- If the system $C\vec{w} = \vec{z}$ has one solution $\vec{w}_0$ for $\vec{w}$, then $\vec{x}[i^\star] = \vec{x}^\star$ if and only if our control inputs $\vec{w}$ are exactly that solution $\vec{w}_0$.
- If the system $C\vec{w} = \vec{z}$ has no solutions in $\vec{w}$, then there is no input $\vec{w}$ which makes $\vec{x}[i^\star] = \vec{x}^\star$. Moreover, $\|\vec{x}[i^\star] - \vec{x}^\star\|$ is minimized if our control inputs $\vec{w}$ are the least squares solution $\vec{w}_{\text{LS}} = (C^\top C)^{-1}C^\top \vec{z}$.[1]

[1]Here there is a caveat regarding invertibility of $C^\top C$. We omit this discussion now, since by the end of the note we will have a

If we have infinitely many solutions for $\vec{w}$, then any of them will make $\vec{x}[i^\star] = \vec{x}^\star$. We will distinguish between them by their energy.

Definition 13 (Energy of an Input)
The *energy* of an input $\vec{w} = \begin{bmatrix} \vec{u}[0] \\ \vdots \\ \vec{u}[i^\star - 1] \end{bmatrix}$ is its squared norm $\|\vec{w}\|^2 = \sum_{i=0}^{i^\star-1} \|\vec{u}[i]\|^2$.

So now we pick an input $\vec{w}$ which minimizes $\|\vec{w}\|^2$ while still solving $C\vec{w} = \vec{z}$, in essence solving the optimization problem

$$\min_{\vec{w}} \quad \|\vec{w}\|^2 \tag{21}$$
$$\text{s.t.} \quad C\vec{w} = \vec{z}. \tag{22}$$

More generically, so-called *minimum-norm problems* of the form

$$\min_{\vec{x}} \quad \|\vec{x}\|^2 \tag{23}$$
$$\text{s.t.} \quad A\vec{x} = \vec{b}, \tag{24}$$

are ubiquitous in engineering even outside control theory. In the next section, we will use the SVD as a tool to think about and solve this problem.

NOTE: From now on, we switch from the control-theoretic reachability notation $(C, \vec{w}, \vec{z})$ to the generic linear algebraic notation $(A, \vec{x}, \vec{b})$. Note that this A is not necessarily the same as the control system state transition matrix A.

5 Moore-Penrose Pseudoinverse

Now that we have the SVD, we may use it to define a *pseudoinverse*, i.e., an object which has some of the properties of an inverse, but is defined for non-invertible matrices. This will help us solve least-squares and least-norm problems.

Definition 14 (Moore-Penrose Psuedoinverse)
Suppose $A \in \mathbb{R}^{m\times n}$ has rank $r \leq \min\{m, n\}$. Let $A = U\Sigma V^\top$ be an SVD of A. The *Moore-Penrose pseudoinverse* of A is a matrix $A^\dagger \in \mathbb{R}^{n\times m}$ given by

$$A^\dagger := V\Sigma^\dagger U^\top \qquad \text{where} \qquad \Sigma^\dagger = \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix}^\dagger = \begin{bmatrix} \Sigma_r^{-1} & 0_{r\times(m-r)} \\ 0_{(n-r)\times r} & 0_{(n-r)\times(m-r)} \end{bmatrix}. \tag{25}$$

Since Σ_r is a diagonal matrix ordered in non-increasing order, Σ_r^{-1} is a diagonal matrix ordered in non-decreasing order. Thus $V\Sigma^\dagger U^\top$ is not an SVD of $A^\dagger$, but one can sort the diagonal entries of $\Sigma^\dagger$ and the corresponding columns of U and V in order to make it an SVD of $A^\dagger$.

We can also compactify this pseudoinverse, using the same derivation as that of the compact SVD.

more unified treatment of these solutions which does not require invertibility of $C^\top C$.

Proposition 15 (Compact Moore-Penrose Pseudoinverse)
Suppose $A \in \mathbb{R}^{m \times n}$ has rank $r \leq \min\{m, n\}$. Let $A = U_r \Sigma_r V_r^\top$ be a compact SVD of A. Then A's pseudoinverse $A^\dagger \in \mathbb{R}^{n \times m}$ can be expressed in terms of the compact SVD as

$$A^\dagger = V_r \Sigma_r^{-1} U_r^\top. \tag{26}$$

We also have some simple identities of the pseudoinverse.

Proposition 16 (Pseudoinverse Identities)

(i) If A is invertible, i.e., A^{-1} exists, then $A^\dagger = A^{-1}$ (inverse is pseudoinverse);

(ii) $(A^\dagger)^\dagger = A$ (taking pseudoinverse twice does nothing);

(iii) $(A^\top)^\dagger = (A^\dagger)^\top$ (pseudoinverse commutes with transpose);

(iv) For $\alpha \neq 0$, $(\alpha A)^\dagger = \alpha^{-1} A^\dagger$ (scalar distributivity);

(v) $AA^\dagger A = A$ (weak left inverse property);

(vi) $A^\dagger A A^\dagger = A^\dagger$ (weak right inverse property);

(vii) $AA^\dagger = U_r U_r^\top$ ($AA^\dagger$ is projection onto $\mathrm{Col}(A)$);

(viii) $A^\dagger A = V_r V_r^\top$ ($A^\dagger A$ is projection onto $\mathrm{Col}(A^\top)$).

Note that all these properties hold for the regular inverse, as well.

Concept Check: Prove Proposition 16. The proofs should entirely be writing out A and $A^\dagger$ in terms of the SVD of A, then simplifying the given expression as much as possible.

Now we can get onto our main theorem of the pseudoinverse, which is one of many reasons we should care about it.

Theorem 17 (Pseudoinverse Solves Least-Norm Least-Squares)
Let $A \in \mathbb{R}^{m \times n}$ have rank $r \leq \min\{m, n\}$, and let $\vec{b} \in \mathbb{R}^m$. Let S be the set of least squares solutions:[a]

$$S := \operatorname*{argmin}_{\vec{z} \in \mathbb{R}^n} \left\| A\vec{z} - \vec{b} \right\|^2. \tag{27}$$

The solution of the optimization problem

$$\min_{\vec{x} \in S} \|\vec{x}\|^2 \tag{28}$$

is unique and given by $\vec{x}^\star = A^\dagger \vec{b}$.

[a] Note that even though $\mathrm{proj}_{\mathrm{Col}(A)}\left(\vec{b}\right)$ is unique, the $\vec{z}$ such that $A\vec{z} = \mathrm{proj}_{\mathrm{Col}(A)}\left(\vec{b}\right)$ is not necessarily unique, so this set S may have more than one element.

See Appendix B.1 for proof.

Now this theorem seems a little abstract, but it has corollaries which are grounded in solving the least-squares and least-norm problems we are familiar with.

Corollary 18 (Pseudoinverse Solves Least-Squares)
Let $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ have full column rank, and let $\vec{b} \in \mathbb{R}^m$.

(i) The solution to the least-squares problem

$$\min_{\vec{x} \in \mathbb{R}^n} \left\| A\vec{x} - \vec{b} \right\|^2 \tag{29}$$

is given by $\vec{x}^{\star}_{\text{LS}} = A^{\dagger}\vec{b}$.

(ii) The pseudoinverse $A^{\dagger}$ has the formula $A^{\dagger} = (A^{\top}A)^{-1}A^{\top}$.

Proof. Using the notation of Theorem 17, if A has full column rank then S has exactly one element, which is the least squares solution $\vec{x}^{\star}_{\text{LS}} = (A^{\top}A)^{-1}A^{\top}\vec{b}$. Hence, $A^{\dagger} = (A^{\top}A)^{-1}A^{\top}$ as desired. □

Corollary 19 (Pseudoinverse Solves Least-Norm)
Let $A \in \mathbb{R}^{m \times n}$ with $m \leq n$ have full row rank, and let $\vec{b} \in \mathbb{R}^m$.

(i) The solution to the least-norm problem

$$\min_{\vec{x} \in \mathbb{R}^n} \quad \|\vec{x}\|^2 \tag{30}$$

$$\text{s.t.} \quad A\vec{x} = \vec{b} \tag{31}$$

is given by $\vec{x}^{\star}_{\text{LN}} = A^{\dagger}\vec{b}$.

(ii) The pseudoinverse $A^{\dagger}$ has the formula $A^{\dagger} = A^{\top}(AA^{\top})^{-1}$.

Proof. Using the notation of Theorem 17, if A has full row rank then S has infinitely many elements $\vec{x}$ such that $\left\| A\vec{x} - \vec{b} \right\| = 0$, i.e., $A\vec{x} = \vec{b}$. To show that $A^{\dagger} = A^{\top}(AA^{\top})^{-1}$, we compute

$$\begin{aligned}
A^{\top}(AA^{\top})^{-1} &= (U\Sigma V^{\top})^{\top}((U\Sigma V^{\top})(U\Sigma V^{\top})^{\top})^{-1} && (32)\\
&= V\Sigma^{\top}U^{\top}\left(U\Sigma V^{\top}V\Sigma^{\top}U^{\top}\right)^{-1} && (33)\\
&= V\Sigma^{\top}U^{\top}U\left(\Sigma\Sigma^{\top}\right)^{-1}U^{\top} && (34)\\
&= V\Sigma^{\top}\left(\Sigma\Sigma^{\top}\right)^{-1}U^{\top} && (35)\\
&= V\Sigma^{\dagger}U^{\top} && (36)\\
&= A^{\dagger} && (37)
\end{aligned}$$

as desired. □

6 Examples

6.1 Example SVD Interpretation

Suppose we have an $m \times n$ matrix A, of rank r, that contains the ratings of m viewers for n movies. Write

$$A = U\Sigma V^\top = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top. \tag{38}$$

We can interpret each rank 1 matrix $\sigma_i \vec{u}_i \vec{v}_i^\top$ to be due to a particular attribute, e.g., comedy, action, sci-fi, or romance content. Then σ_i determines how strongly the ratings depend on the i^{th} attribute; the entries of $\vec{v}_i^\top$ score each movie with respect to this attribute, and the entries of $\vec{u}_i$ evaluate how much each viewer cares about this particular attribute. Interestingly, the $(r+1)^{\text{th}}$ attributes onwards don't influence the ratings, according to our analysis.

6.2 Numerical Example 1

Let's find the SVD for

$$A = \begin{bmatrix} 4 & 4 \\ -3 & 3 \end{bmatrix}. \tag{39}$$

We find the SVD for $A^\top$ first and then take the transpose. We calculate

$$(A^\top)^\top(A^\top) = AA^\top = \begin{bmatrix} 4 & 4 \\ -3 & 3 \end{bmatrix} \begin{bmatrix} 4 & -3 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 32 & 0 \\ 0 & 18 \end{bmatrix}. \tag{40}$$

This happens to be diagonal, so we can read off the eigenvalues:

$$\lambda_1 = 32 \qquad \lambda_2 = 18 \tag{41}$$

We can select the orthonormal eigenvectors:

$$\vec{u}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \vec{u}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{42}$$

The singular values are

$$\sigma_1 = \sqrt{\lambda_1} = \sqrt{32} = 4\sqrt{2} \qquad \sigma_2 = \sqrt{\lambda_2} = \sqrt{18} = 3\sqrt{2}. \tag{43}$$

Then to find $\vec{v}_1, \vec{v}_2$, we do

$$\vec{v}_1 = \frac{A^\top \vec{u}_1}{\sigma_1} = \frac{1}{4\sqrt{2}} \begin{bmatrix} 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \tag{44}$$

$$\vec{v}_2 = \frac{A^\top \vec{u}_2}{\sigma_2} = \frac{1}{3\sqrt{2}} \begin{bmatrix} -3 \\ 3 \end{bmatrix} = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}. \tag{45}$$

Thus our SVD is

$$A = U\Sigma V^\top = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 4\sqrt{2} & 0 \\ 0 & 3\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}. \tag{46}$$

Note that we can change the signs of $\vec{u}_1, \vec{u}_2$ and they are still orthonormal eigenvectors, and produce a valid SVD. However, changing the sign of $\vec{u}_i$ requires us to change the sign of $\vec{v}_i = A^\top \vec{u}_i$, so therefore the product of $\vec{u}_i \vec{v}_i^\top$ remains unchanged.

Another source of non-uniqueness arises when we have repeated singular values, as seen in the next example.

6.3 Numerical Example 2.

We want to find an SVD for

$$A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{47}$$

Again, we find the SVD for $A^\top$ and then take the transpose. Note that $AA^\top = I_2$, which has repeated eigenvalues at $\lambda_1 = \lambda_2 = 1$. In particular, *any* pair of orthonormal vectors is a set of orthonormal eigenvectors for $I_2 = AA^\top$. We can parameterize all such pairs as

$$\vec{u}_1 = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \qquad \vec{u}_2 = \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix} \tag{48}$$

where θ is a free parameter. Since $\sigma_1 = \sigma_2 = 1$, we obtain

$$\vec{v}_1 = \frac{A^\top \vec{u}_1}{\sigma_1} = \begin{bmatrix} \cos(\theta) \\ -\sin(\theta) \end{bmatrix} \qquad \vec{v}_2 = \frac{A^\top \vec{u}_2}{\sigma_2} = \begin{bmatrix} -\sin(\theta) \\ -\cos(\theta) \end{bmatrix}. \tag{49}$$

Thus an SVD is

$$A = U\Sigma V^\top = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ -\sin(\theta) & -\cos(\theta) \end{bmatrix} \tag{50}$$

for any value of θ. Thus this matrix has *infinite* valid SVDs, one for each value of θ in the interval $[0, 2\pi)$.

6.4 Long-Form Example

In this example we review discretization, controllability, and minimum-norm solutions. Consider the model of a car moving in a lane

$$\frac{\mathrm{d}p(t)}{\mathrm{d}t} = v(t) \tag{51}$$

$$\frac{\mathrm{d}v(t)}{\mathrm{d}t} = \frac{1}{RM}u(t) \tag{52}$$

where $p(t)$ is position, $v(t)$ is velocity, $u(t)$ is wheel torque, R is wheel radius, and M is mass.

First we discretize this continuous-time model. If we apply the constant input $u(t)$ from $u_d[i]$ from $t = i\Delta$ to $t = (i+1)\Delta$, then by integration

$$p(t) = p_d[i] + (t - i\Delta)v_d[i] + \frac{1}{2}(t - i\Delta)^2 \frac{1}{RM} u_d[i] \tag{53}$$

$$v(t) = v_d[i] + (t - i\Delta)\frac{1}{RM} u_d[i] \tag{54}$$

for $t \in [i\Delta, (i+1)\Delta)$. In particular, at $t = (i+1)\Delta$,

$$p_d[i+1] = p((i+1)\Delta) = p_d[i] + \Delta v_d[i] + \frac{\Delta^2}{2RM} u_d[i] \tag{55}$$

$$v_d[i+1] = v((i+1)\Delta) = v_d[i] + \frac{\Delta}{RM} u_d[i]. \tag{56}$$

Putting these equations in matrix/vector form, we get

$$\begin{bmatrix} p_d[i+1] \\ v_d[i+1] \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix}}_{:=A} \begin{bmatrix} p_d[i] \\ v_d[i] \end{bmatrix} + \underbrace{\frac{1}{RM} \begin{bmatrix} \frac{1}{2}\Delta^2 \\ \Delta \end{bmatrix}}_{:=\vec{b}} u_d[i] \tag{57}$$

Now suppose the vehicle is at rest with $p_d[0] = v_d[0] = 0$ and the goal is to reach a target position $p^\star$ and stop there (i.e., $v^\star = 0$). From reachability analysis, if we can find a sequence $u_d[0], u_d[1], \ldots, u_d[\ell-1]$ such that

$$\begin{bmatrix} p^\star \\ 0 \end{bmatrix} = \begin{bmatrix} A^{\ell-1}\vec{b} & \cdots & A\vec{b} & \vec{b} \end{bmatrix} \begin{bmatrix} u_d[0] \\ \vdots \\ u_d[\ell-2] \\ u_d[\ell-1] \end{bmatrix} \tag{58}$$

then at time $t = \ell\Delta$, i.e., in ℓ timesteps we reach the desired state.

Since we have $n = 2$ state variables the controllability test we learned checks whether C_ℓ with $\ell = 2$ spans $\mathbb{R}^2$. This is indeed the case, since

$$\mathcal{C}_2 = \begin{bmatrix} A\vec{b} & \vec{b} \end{bmatrix} = \frac{1}{RM} \begin{bmatrix} \frac{3}{2}\Delta^2 & \frac{1}{2}\Delta^2 \\ \Delta & \Delta \end{bmatrix} \tag{59}$$

has linearly independent columns.

Although this test suggests we can reach the target state in two steps, the resulting values of $u_d[0]$ and $u_d[1]$ will likely exceed physical limits. If we take the values $RM = 5000\,\mathrm{kg\,m}$, $T = 0.1\,\mathrm{s}$, $p_\star = 1000\,\mathrm{m}$, then

$$\begin{bmatrix} u_d[0] \\ u_d[1] \end{bmatrix} = \mathcal{C}_2^{-1} \begin{bmatrix} p^\star \\ 0 \end{bmatrix} = \begin{bmatrix} 5 \times 10^8 \, \frac{\mathrm{kg\,m^2}}{\mathrm{s^2}} \\ -5 \times 10^8 \, \frac{\mathrm{kg\,m^2}}{\mathrm{s^2}} \end{bmatrix} \tag{60}$$

which exceeds the torque and breaking limits of a typical car by 5 orders of magnitude.

Therefore, in practice we need to select a sufficiently large number of time steps ℓ. This leads to a wide controllability matrix $\mathcal{C}_\ell$ and allows for infinitely many input sequences that reach our target state. Among them, we can select the minimum norm solution so we spend the least control energy. Using the minimum-norm formula

$$\begin{bmatrix} u_d[0] \\ \vdots \\ u_d[\ell-2] \\ u_d[\ell-1] \end{bmatrix} = C_\ell^\top (C_\ell C_\ell^\top)^{-1} \begin{bmatrix} p^\star \\ 0 \end{bmatrix} \tag{61}$$

and some algebra, one obtains the input sequence

$$u_d[i] = \frac{6RM(\ell - 1 - 2i)}{\Delta^2 \ell(\ell^2 - 1)} p_\star \qquad i \in \{0, 1, \ldots, \ell - 1\}. \tag{62}$$

In the plot below we show this input sequence, as well as the resulting velocity and position profiles for $RM = 5000\,\mathrm{kg\,m}$, $p_\star = 1000\,\mathrm{m}$, $\Delta = 0.1\,\mathrm{s}$, and $\ell = 1200$. With these parameters we allow $\ell\Delta = 120\,\mathrm{s}$ (2 minutes) to travel 1 km. Note that the vehicle accelerates in the first half of this period and decelerates in the second half, reaching the maximum velocity $12.5\,\frac{\mathrm{m}}{\mathrm{s}}$ ($\approx$ 28 mph) in the middle. The acceleration and deceleration are hardest at the very beginning and at the very end, respectively. The corresponding torque is within a physically reasonable range, $[-2000, 2000]\mathrm{N\,m}$.

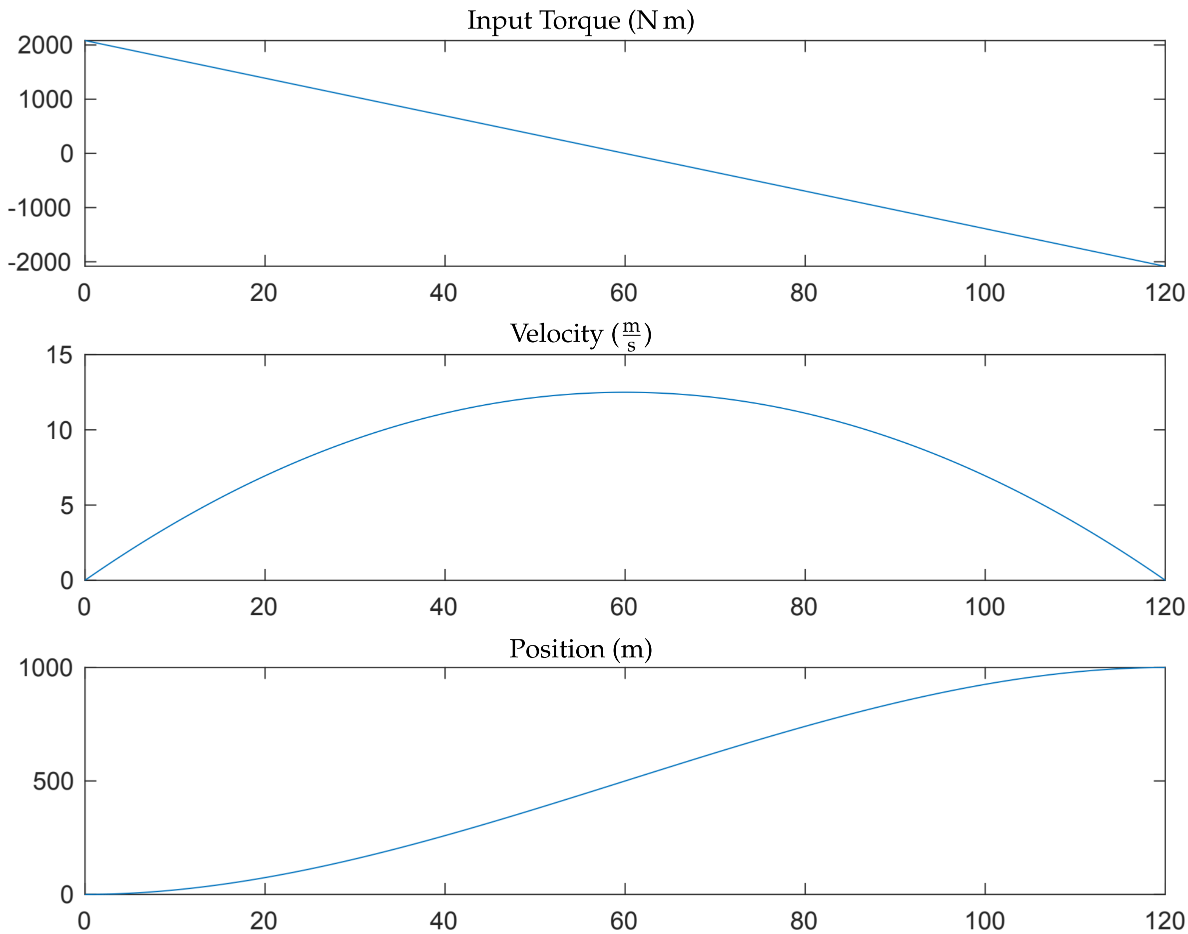


Figure 1: The minimum norm input torque sequence, and the resulting velocity and position profiles, for $RM = 5000\,\text{kg}\,\text{m}$, $p^\star = 1000\,\text{m}$, $\Delta = 0.1\,\text{s}$, and $\ell = 1200$. The horizontal axis is time, which ranges from 0 to $\ell\Delta = 120\,\text{s}$. The vehicle accelerates in the first half of this period and decelerates in the second half, reaching the maximum velocity $12.5\,\frac{\text{m}}{\text{s}}$ in the middle.

A Proofs for Section 1

A.1 Proof of Proposition 1

Proof of Proposition 1. We have

$$(A^\top A)^\top = (A)^\top (A^\top)^\top = A^\top A \tag{63}$$

so $(A^\top A)^\top = A^\top A$, and thus $A^\top A$ is symmetric.

Part 1 To show that $\operatorname{rank}(A^\top A) = r$, we can use the fact that $\operatorname{Null}(A) = \operatorname{Null}(A^\top A)$, which implies $\dim(\operatorname{Null}(A)) = \dim(\operatorname{Null}(A^\top A))$. Applying the Rank-Nullity theorem to A, we have that

$$r = \operatorname{rank}(A) = n - \dim(\operatorname{Null}(A)) = n - \dim\left(\operatorname{Null}\left(A^\top A\right)\right) = \operatorname{rank}\left(A^\top A\right). \tag{64}$$

Thus $\operatorname{rank}(A^\top A) = r$.

Part 2 Now we show that $A^\top A$ has exactly r nonzero eigenvalues. Indeed,

$$\dim\left(\operatorname{Null}\left(A^\top A\right)\right) = n - \operatorname{rank}\left(A^\top A\right) = n - r. \tag{65}$$

Thus $A^\top A$ has an $(n-r)$-dimensional null space, corresponding to an eigenvalue 0 with geometric multiplicity $m^g_{A^\top A}(0) = n - r$. By the Spectral Theorem, and the fact that $A^\top A$ is symmetric, we know that $A^\top A$ is diagonalizable.

We know that for a diagonalizable matrix, the geometric and algebraic multiplicities of each eigenvalue agree, i.e., $m^a_{A^\top A}(\lambda) = m^g_{A^\top A}(\lambda)$ for each eigenvalue λ of $A^\top A$. So $m^a_{A^\top A}(0) = n - r$. Since $\sum_\lambda m^a_{A^\top A}(\lambda) = n$, this implies that $A^\top A$ has r nonzero eigenvalues.

Part 3 We now show that all nonzero eigenvalues of $A^\top A$ are real and positive. Since $A^\top A$ is symmetric, the Spectral Theorem says that $A^\top A$ has real eigenvalues. Now we must show that they are all nonnegative. In other words, that $A^T A$ is, by definition, **positive semi-definite**. Let λ be a nonzero eigenvalue. Let λ be a nonzero eigenvalue of $A^\top A$ with eigenvector $\vec{v}$. Then

$$A^\top A\vec{v} = \lambda\vec{v} \tag{66}$$

$$\vec{v}^\top A^\top A\vec{v} = \lambda\vec{v}^\top\vec{v} \tag{67}$$

$$\langle A\vec{v},\ A\vec{v}\rangle = \lambda\langle\vec{v},\ \vec{v}\rangle \tag{68}$$

$$\|A\vec{v}\|^2 = \lambda\|\vec{v}\|^2. \tag{69}$$

We know that λ is nonzero, and $\vec{v}$ is nonzero so $\|\vec{v}\| > 0$. Hence, $\|A\vec{v}\|$ is nonzero and thus positive. Thus

$$\lambda = \frac{\|A\vec{v}\|^2}{\|\vec{v}\|^2} \tag{70}$$

is the quotient of positive numbers and thus positive.

Now let $B := A^\top$ and note that $AA^\top = (A^\top)^\top(A^\top) = B^\top B$. Thus applying the same calculation to the rank-r matrix $B = A^\top$ obtains that $AA^\top$ is symmetric, that $\operatorname{rank}(AA^\top) = r$, and that $AA^\top$ has exactly r nonzero eigenvalues, which are real and positive. □

A.2 Proof of Theorem 3

Proof of Theorem 3.

(i) Since $\{\vec{u}_1, \ldots, \vec{u}_r\}$ is already assumed to be orthonormal, it is left to show that this set spans $\mathrm{Col}(A)$. Indeed, for $\vec{x} \in \mathbb{R}^n$, we have

$$A\vec{x} = \left(\sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top\right)\vec{x} \tag{71}$$

$$= \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top \vec{x} \tag{72}$$

$$= \sum_{i=1}^{r} \sigma_i \langle \vec{x},\ \vec{v}_i \rangle \vec{u}_i. \tag{73}$$

This is a linear combination of the $\vec{u}_i$, so we have $\mathrm{Col}(A) \subseteq \mathrm{Span}(\vec{u}_1, \ldots, \vec{u}_r)$. Since $\dim(\mathrm{Col}(A)) = r$ and $\dim(\mathrm{Span}(\vec{u}_1, \ldots, \vec{u}_r)) = r$, they are the same dimension and thus are equal. Hence, $\mathrm{Col}(A) = \mathrm{Span}(\vec{u}_1, \ldots, \vec{u}_r)$ as desired.

(ii) Since $A^\top = \sum_{i=1}^{r} \sigma_i \vec{v}_i \vec{u}_i^\top$, applying part (i) to this matrix obtains that $\mathrm{Span}(\vec{v}_1, \ldots, \vec{v}_r) = \mathrm{Col}\left(A^\top\right)$.

(iii) We have

$$AA^\top \vec{u}_k = \left(\sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top\right)\left(\sum_{j=1}^{r} \sigma_i \vec{v}_i \vec{u}_i^\top\right)\vec{v}_k \tag{74}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{r} \sigma_i \sigma_j \vec{u}_i \vec{v}_i^\top \vec{v}_j \vec{u}_j^\top \vec{u}_k \tag{75}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{r} \sigma_i \sigma_j \left\langle \vec{v}_j,\ \vec{v}_i \right\rangle \left\langle \vec{u}_k,\ \vec{u}_j \right\rangle \vec{u}_i \tag{76}$$

$$= \sigma_k^2 \vec{u}_k \tag{77}$$

as desired; the mass cancellation is due to orthonormality.

(iv) We have

$$A^\top A\vec{v}_k = \left(\sum_{i=1}^{r} \sigma_i \vec{v}_i \vec{u}_i^\top\right)\left(\sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top\right)\vec{v}_k \tag{78}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{r} \sigma_i \sigma_j v_i u_i^\top \vec{u}_j \vec{v}_i^\top \vec{v}_k \tag{79}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{r} \sigma_i \sigma_j \left\langle \vec{u}_j,\ \vec{u}_i \right\rangle \langle \vec{v}_k,\ \vec{v}_i \rangle\, v_i \tag{80}$$

$$= \sigma_k^2 \vec{v}_k \tag{81}$$

where again the mass cancellation is due to orthonormality.

(v) We have

$$A\vec{v}_k = \left(\sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top\right)\vec{v}_k \tag{82}$$

$$= \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top \vec{v}_k \tag{83}$$

$$= \sum_{i=1}^{r} \sigma_i \langle \vec{v}_k, \vec{v}_i \rangle \vec{u}_i \tag{84}$$

$$= \sigma_k \vec{u}_k \tag{85}$$

where again the mass cancellation is due to orthonormality.

□

A.3 Proof of Correctness of Algorithm 4

Proof of Correctness of Algorithm 4. We first claim that $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are real. Indeed, by Proposition 1, the top r eigenvalues $\lambda_1 \geq \cdots \geq \lambda_r > 0$ of $A^\top A$ are real and positive, so their square roots $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are real and positive.

Now we claim that $\{\vec{v}_1, \ldots, \vec{v}_r\}$ is an orthonormal set. By the Spectral Theorem, the eigenvalues of $A^\top A$ are real and thus may be ordered. Furthermore, there is an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors $\vec{v}_1, \ldots, \vec{v}_n$ of $A^\top A$, ordered by the eigenvalues. Thus, the construction in Line 3 is valid and produces orthonormal $\vec{v}_1, \ldots, \vec{v}_n$, and so $\{\vec{v}_1, \ldots, \vec{v}_r\}$ is an orthonormal set.

Now we claim that $\{\vec{u}_1, \ldots, \vec{u}_r\}$ is an orthonormal set. Let $1 \leq i < j \leq r$. Then

$$\langle \vec{u}_i, \vec{u}_j \rangle = \left\langle \frac{1}{\sigma_i} A\vec{v}_i, \frac{1}{\sigma_j} A\vec{v}_j \right\rangle \tag{86}$$

$$= \frac{1}{\sigma_i \sigma_j} \langle A\vec{v}_i, A\vec{v}_j \rangle \tag{87}$$

$$= \frac{1}{\sigma_i \sigma_j} \left\langle A^\top A\vec{v}_i, \vec{v}_j \right\rangle \tag{88}$$

$$= \frac{1}{\sigma_i \sigma_j} \langle \sigma_i \vec{v}_i, \vec{v}_j \rangle \tag{89}$$

$$= \frac{1}{\sigma_j} \langle \vec{v}_i, \vec{v}_j \rangle \tag{90}$$

$$= 0 \tag{91}$$

by orthonormality of the $\vec{v}_i$.

Now let $1 \leq i \leq r$. Then

$$\|\vec{u}_i\|^2 = \|A\vec{v}_i\|^2 \tag{92}$$

$$= \frac{1}{\sigma_i^2} \|A\vec{v}_i\|^2 \tag{93}$$

$$= \frac{1}{\sigma_i^2} \langle A\vec{v}_i, A\vec{v}_i \rangle \tag{94}$$

$$= \frac{1}{\sigma_i^2} \left\langle A^\top A\vec{v}_i, \vec{v}_i \right\rangle \tag{95}$$

$$= \frac{1}{\sigma_i^2} \langle \lambda_i \vec{v}_i, \vec{v}_i \rangle \tag{96}$$

$$= \frac{\lambda_i}{\sigma_i^2} \langle \vec{v}_i, \vec{v}_i \rangle \tag{97}$$

$$= \frac{\sigma_i^2}{\sigma_i^2} \|\vec{v}_i\|^2 \tag{98}$$

$$= \|\vec{v}_i\|^2 \tag{99}$$

$$= 1. \tag{100}$$

Thus $\{\vec{u}_1, \dots, \vec{u}_r\}$ is an orthonormal set.

Finally, we claim that $A = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top$. This equality holds if and only if, for every $\vec{x} \in \mathbb{R}^n$, we have $A\vec{x} = \left(\sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top\right)$. We show the latter condition.

Since $\{\vec{v}_1, \dots, \vec{v}_n\}$ is an orthonormal basis of $\mathbb{R}^n$, we may let $\vec{x} = \sum_{i=1}^{n} \alpha_i \vec{v}_i$ for some constants $\alpha_1, \dots, \alpha_n \in \mathbb{R}$. The eigenvectors corresponding to the 0 eigenvalues of $A^\top A$ — that is, $\vec{v}_{r+1}, \dots, \vec{v}_n$ — are an orthonormal basis for $\text{Null}(A^\top A)$. We know that $\text{Null}(A^\top A) = \text{Null}(A)$, so $\{\vec{v}_{r+1}, \dots, \vec{v}_n\}$ is an orthonormal basis for $\text{Null}(A)$. Thus $A\vec{v}_{r+1} = \cdots = A\vec{v}_n = \vec{0}_m$. With this, we can compute

$$\left(\sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top\right) \vec{x} = \left(\sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top\right) \left(\sum_{i=1}^{n} \alpha_i \vec{v}_i\right) \tag{101}$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{n} \sigma_i \alpha_j \vec{u}_i \vec{v}_i^\top \vec{v}_j \tag{102}$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{n} \sigma_i \alpha_j \langle \vec{v}_j, \vec{v}_i \rangle \vec{u}_i \tag{103}$$

$$= \sum_{i=1}^{r} \alpha_i \sigma_i \vec{u}_i \tag{104}$$

$$= \sum_{i=1}^{r} \alpha_i A \vec{v}_i \tag{105}$$

$$= \sum_{i=1}^{n} \alpha_i A \vec{v}_i \tag{106}$$

$$= A \left(\sum_{i=1}^{n} \alpha_i \vec{v}_i\right) \tag{107}$$

$$= A\vec{x}. \tag{108}$$

Thus $A = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top$ as desired. □

A.4 Proof of Theorem 9

Proof of Theorem 9.

(i) Follows from the equivalence of compact and full SVDs: $A = U\Sigma V^\top = U_r \Sigma_r V_r^\top$.

(ii) Follows from the equivalence of compact and full SVDs: $A = U\Sigma V^\top = U_r \Sigma_r V_r^\top$.

(iii) Since $\{\vec{u}_1, \dots, \vec{u}_m\}$ is the orthonormal set of columns of U, we have that $\{\vec{u}_1, \dots, \vec{u}_r\}$ and $\{\vec{u}_{r+1}, \dots, \vec{u}_m\}$ are orthonormal bases for orthogonal subspaces. Since $\text{Span}(\vec{u}_1, \dots, \vec{u}_r) = \text{Col}(A)$, the set of vectors orthogonal to $\text{Col}(A)$ is given by $\text{Span}(\vec{u}_{r+1}, \dots, \vec{u}_m) = \text{Col}(U_{m-r})$. We are left to show that $\text{Null}(A^\top) = \text{Col}(U_{m-r})$. Indeed, let $\vec{y} \in \mathbb{R}^m$ such that $\vec{y}$ is orthogonal to $\text{Col}(A)$. Then

$$\langle \vec{y}, \vec{z} \rangle = 0 \qquad \text{for all } \vec{z} \in \text{Col}(A) \tag{109}$$

Since $\vec{z} \in \mathrm{Col}(A)$, we can also write $\vec{z}$ as $A\vec{x}$, giving us the equivalent equation

$$\langle \vec{y},\, A\vec{x} \rangle = 0 \qquad \text{for all } \vec{x} \in \mathbb{R}^n \tag{110}$$

We can further rewrite $\langle \vec{y},\, A\vec{x} \rangle$ as $(A\vec{x})^\top \vec{y} = \vec{x}^\top A^\top \vec{y} = \vec{x}^\top (A^\top \vec{y})$, giving us

$$\left\langle A^\top \vec{y},\, \vec{x} \right\rangle = 0 \qquad \text{for all } \vec{x} \in \mathbb{R}^n \tag{111}$$

$$\iff A^\top \vec{y} = \vec{0}_n \tag{112}$$

$$\iff \vec{y} \in \mathrm{Null}\left(A^\top\right). \tag{113}$$

Thus, $\mathrm{Col}(U_{m-r}) \subseteq \mathrm{Null}\left(A^\top\right)$. We can show the sets are exactly equal by applying Rank-Nullity Theorem to get that $\dim(\mathrm{Null}\left(A^\top\right)) = m - r$, which is also the dimension of $\mathrm{Col}(U_{m-r})$. Therefore, $\mathrm{Null}\left(A^\top\right) = \mathrm{Col}(U_{m-r})$.

(iv) Follows from applying (iii) to the SVD of $A^\top$, i.e., $A^\top = V\Sigma U^\top$.

(v) We have

$$AA^\top = (U\Sigma V^\top)(U\Sigma V^\top)^\top \tag{114}$$

$$= U\Sigma V^\top V\Sigma U^\top \tag{115}$$

$$= U\Sigma\Sigma U^\top. \tag{116}$$

(vi) We have

$$A^\top A = (U\Sigma V^\top)^\top (U\Sigma V^\top) \tag{117}$$

$$= V\Sigma^\top U^\top U\Sigma V^\top \tag{118}$$

$$= V\Sigma^\top \Sigma V^\top. \tag{119}$$

(vii) Follows from the equivalence of compact and full SVDs: $A = U\Sigma V^\top = U_r \Sigma_r V_r^\top$.

□

A.5 Proof of Correctness of Algorithm 10

Proof of Correctness of Algorithm 10. The only thing that has not already been shown from the proof of correctness of the outer product SVD algorithm is that U and V are square orthonormal matrices. But this is straightforward from Gram-Schmidt. □

B Proofs for Section 5

B.1 Proof of Theorem 17

Proof of Theorem 17. We first show that $\vec{x}^\star \in S$, then that it is the unique member of S with minimum norm.

We have that $\vec{x} \in S$ if and only if $A\vec{x} - \vec{b}$ is orthogonal to $\mathrm{Col}(A)$. This means that $\vec{x} \in S$ if and only if, for any $\vec{w} \in \mathbb{R}^n$ we have that $\left\langle A\vec{x} - \vec{b},\, A\vec{w} \right\rangle = 0$. This is equivalent to saying that, $\vec{x} \in S$ if and only if, for any $\vec{w} \in \mathbb{R}^n$ we have that $\left\langle A^\top (A\vec{x} - \vec{b}),\, \vec{w} \right\rangle = 0$. Since the left-hand argument of the inner product is not dependent on $\vec{w}$, $\vec{x} \in S$ if and only if $A^\top(A\vec{x} - \vec{b}) = \vec{0}_n$ (the so-called *normal equations*). Algebraically this is the equation

$$A^\top A\vec{x} = A^\top \vec{b}. \tag{120}$$

We verify that $\vec{x}^\star = A^\dagger \vec{b}$ fulfills this equation. Indeed,

$$\begin{aligned}
A^\top A\vec{x}^\star &= A^\top A A^\dagger \vec{b} && (121)\\
&= (U_r\Sigma_r V_r^\top)^\top (U_r\Sigma_r V_r^\top)(V_r\Sigma_r^{-1}U_r^\top)\vec{b} && (122)\\
&= V_r\Sigma_r U_r^\top U_r\Sigma_r V_r^\top V_r\Sigma_r^{-1}U_r^\top \vec{b} && (123)\\
&= V_r\Sigma_r U_r^\top \vec{b} && (124)\\
&= (U_r\Sigma_r V_r^\top)^\top \vec{b} && (125)\\
&= A^\top \vec{b}. && (126)
\end{aligned}$$

Thus $\vec{x}^\star$ fulfills the normal equations, so $\vec{x}^\star \in S$.

Now because the projection is unique, for any $\vec{x} \in S$, we have that $A\vec{x} = \mathrm{proj}_{\mathrm{Col}(A)}\left(\vec{b}\right)$ is independent of $\vec{x}$. Suppose that there are two solutions $\vec{x}_1, \vec{x}_2 \in S$. Then since

$$A\vec{x}_1 = A\vec{x}_2 = \mathrm{proj}_{\mathrm{Col}(A)}\left(\vec{b}\right) \tag{127}$$

we see that

$$A(\vec{x}_1 - \vec{x}_2) = A\vec{x}_1 - A\vec{x}_2 = \mathrm{proj}_{\mathrm{Col}(A)}\left(\vec{b}\right) - \mathrm{proj}_{\mathrm{Col}(A)}\left(\vec{b}\right) = \vec{0}_m. \tag{128}$$

Thus $\vec{x}_1 - \vec{x}_2 \in \mathrm{Null}(A)$. This implies that, for any $\vec{x} \in S$, every other vector $\vec{y} \in S$ can be written as $\vec{y} = \vec{x} + \vec{z}$ for $\vec{z} \in \mathrm{Null}(A)$, and further there is exactly one $\vec{x}_0 \in S$ such that $\vec{x}_0$ is orthogonal to $\mathrm{Null}(A)$.[2] For this $\vec{x}_0$ and any $\vec{x} \in S$, we would have

$$\begin{aligned}
\|x\|^2 &= \langle \vec{x},\, \vec{x}\rangle && (129)\\
&= \langle \vec{x}_0 + (\vec{x} - \vec{x}_0),\, \vec{x}_0 + (\vec{x} - \vec{x}_0)\rangle && (130)\\
&= \langle \vec{x}_0,\, \vec{x}_0\rangle + \langle \vec{x}_0,\, \vec{x} - \vec{x}_0\rangle + \langle \vec{x} - \vec{x}_0,\, \vec{x}_0\rangle + \langle \vec{x} - \vec{x}_0,\, \vec{x} - \vec{x}_0\rangle && (131)\\
&= \|\vec{x}_0\|^2 + 2\underbrace{\langle \vec{x}_0,\, \vec{x} - \vec{x}_0\rangle}_{=0} + \|\vec{x} - \vec{x}_0\|^2 && (132)\\
&= \|\vec{x}_0\|^2 + \|\vec{x} - \vec{x}_0\|^2 && (133)\\
&\geq \|\vec{x}_0\|^2 && (134)
\end{aligned}$$

with equality if and only if $\vec{x} = \vec{x}_0$. Here $\langle \vec{x}_0,\, \vec{x} - \vec{x}_0\rangle = 0$ due to the fact that $\vec{x} - \vec{x}_0 \in \mathrm{Null}(A)$ and $\vec{x}_0$ is orthogonal to $\mathrm{Null}(A)$. Thus $\vec{x}_0$ is the unique solution to the optimization problem

$$\min_{\vec{x} \in S} \|\vec{x}\|^2. \tag{135}$$

We need to show that $\vec{x}_0 := \vec{x}^\star$. Recall that we defined $\vec{x}_0$ as the unique element of S which is orthogonal to $\mathrm{Null}(A)$. So we need to show that $\vec{x}^\star = A^\dagger \vec{b}$ is orthogonal to $\mathrm{Null}(A)$. Indeed,

$$\mathrm{Col}\left(A^\dagger\right) = \mathrm{Col}\left(V_r\Sigma_r^{-1}U_r^\top\right) \subseteq \mathrm{Col}(V_r) = \mathrm{Col}\left(A^\top\right). \tag{136}$$

We have shown in the proof of Theorem 9 that $\mathrm{Col}\left(A^\top\right)$ is orthogonal to $\mathrm{Null}(A)$. Since $A^\dagger\vec{b} \in \mathrm{Col}\left(A^\dagger\right)$, we have $A^\dagger\vec{b} \in \mathrm{Col}\left(A^\top\right)$ and thus $\vec{x}^\star$ is orthogonal to $\mathrm{Null}(A)$. Thus $\vec{x}_0 := \vec{x}^\star$ and the proof is complete. □

[2]For any $\vec{x} \in S$, let $\vec{x}_0 := \vec{x} - \mathrm{proj}_{\mathrm{Null}(A)}(\vec{x})$. Then $A\vec{x} = A\vec{x}_0$ so $\vec{x}_0 \in S$, so there exists an $\vec{x}_0 \in S$ which is orthogonal to $\mathrm{Null}(A)$. And for any $\vec{x} \in S \setminus \{\vec{x}_0\}$, we have that $\vec{x} - \vec{x}_0 \in \mathrm{Null}(A)$ and is nonzero, so $\vec{x}$ is not orthogonal to $\vec{x} - \vec{x}_0$. Thus $\vec{x}$ is not orthogonal to $\mathrm{Null}(A)$, and so $\vec{x}_0$ is the unique element of S which is orthogonal to $\mathrm{Null}(A)$.

Contributors:

- Druv Pai.
- Ashwin Vangipuram.
- Rahul Arya.
- Murat Arcak.
- Anant Sahai.
- Matteo Guarrera.
- Aakarsh Vermani.
- Chancharik Mitra.

