Note 16: Minimum-Energy Control and Singular Value Decomposition

NOTE: This note is long. It contains a lot of information about the singular value decomposition, which is one of the most important tools from linear algebra, and there are many perspectives to discuss it from. As a result, you probably should not expect to cover the whole note at once.

1 Overview and Motivation

In Note 12, we discussed the problem of controllability and reachability in discrete-time. Reachability analysis amounted to solving a linear system of the form

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

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 1** (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*. We discuss this conversion in Section 2.

To solve this problem in the abstract, we then introduce the singular value decomposition (SVD), whose properties are fleshed out in Sections 3 to 6.

> **Key Idea 2** (Singular Value Decomposition)
> The SVD is a particular decomposition of a matrix $A = U\Sigma V^\top$, where U and V are orthonormal, and Σ is a (possibly non-square) diagonal matrix.[a] Each of U, Σ, V has important linear-algebraic properties.
>
> [a]This just means that only the diagonal entries of Σ, i.e., Σ_{ii}, may be nonzero.

This decomposition will allow us to solve the minimum-energy control problem in Section 7.

2 Minimum-Energy Control

As previously mentioned, in reachability analysis we try to solve the linear system

$$\mathcal{C}_{i^\star}\begin{bmatrix}\vec{u}[0]\\ \vdots \\ \vec{u}[i^\star-1]\end{bmatrix} = \vec{x}^\star - A^{i^\star}\vec{x}_0 \tag{2}$$

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

$$C\vec{w} = \vec{z}. \tag{3}$$

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]

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 3 (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

$$\min_{\vec{w}} \quad \|\vec{w}\|^2 \tag{4}$$
$$\text{s.t.} \quad C\vec{w} = \vec{z}. \tag{5}$$

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

$$\min_{\vec{x}} \quad \|\vec{x}\|^2 \tag{6}$$
$$\text{s.t.} \quad A\vec{x} = \vec{b}, \tag{7}$$

are ubiquitous in engineering even outside control theory. In the subsequent sections, we will develop tools to think about and solve these problems.

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.

3 Singular Value Decomposition: Existence, Uniqueness, Computation

First, we will introduce the singular value decomposition (SVD) as a matrix factorization, and give an algorithm to efficiently compute it.

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

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

Proposition 4 (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 matrices of rank r. Each has exactly r nonzero eigenvalues, which are real and positive.

The proof of Proposition 4 is on the longer side and may distract from the overall flow of this note, so it is left to Appendix A.1. We fully expect you to read the proof and understand it. It is completely in-scope for the course.

3.1 Existence

Now we may define the SVD. Here we call it the *full SVD* to contrast with the *compact SVD* and the *outer product SVD* that will be introduced in Section 4. When we say "SVD" without further specification, we mean full SVD.

Definition 5 (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

$$A = U\Sigma V^\top = \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix} \begin{bmatrix} V_r^\top \\ V_{n-r}^\top \end{bmatrix} \tag{8}$$

where

(I) $U = \begin{bmatrix} U_r & U_{m-r} \end{bmatrix}$ is a matrix of *left singular vectors*; $U \in \mathbb{R}^{m\times m}$, $U_r \in \mathbb{R}^{m\times r}$, $U_{m-r} \in \mathbb{R}^{m\times(m-r)}$;

(II) $V = \begin{bmatrix} V_r & V_{n-r} \end{bmatrix}$ is a matrix of *right singular vectors*; $V \in \mathbb{R}^{n\times n}$, $V_r \in \mathbb{R}^{n\times r}$, $V_{n-r} \in \mathbb{R}^{n\times(n-r)}$;

(III) $\Sigma = \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix}$ is a diagonal matrix of *singular values*, where $\Sigma \in \mathbb{R}^{m\times n}$ and $\Sigma_r \in \mathbb{R}^{r\times r}$ is diagonal with positive diagonal entries;

such that the following holds:

(i) U is an orthonormal matrix of eigenvectors of $AA^\top$;

(ii) V is an orthonormal matrix of eigenvectors of $A^\top A$;

(iii) $\Sigma\Sigma^\top$ is the matrix of eigenvalues of $AA^\top$;

(iv) $\Sigma^\top\Sigma$ is the matrix of eigenvalues of $A^\top A$;

(v) $\mathrm{Col}(U_r) = \mathrm{Col}(A)$;

(vi) $\mathrm{Col}(U_{m-r}) = \mathrm{Null}(A^\top)$;

(vii) $\mathrm{Col}(V_r) = \mathrm{Col}(A^\top)$;

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

For notation's sake: the columns of U are $\{\vec{u}_1, \ldots, \vec{u}_m\}$; the columns of V (i.e., rows of $V^\top$) are $\{\vec{v}_1, \ldots, \vec{v}_n\}$; and the diagonal entries of Σ are $\{\sigma_1, \ldots, \sigma_{\min\{m,n\}}\}$. Moreover, the columns of U_r are $\vec{u}_1, \ldots, \vec{u}_r$, and the columns of U_{m-r} are $\vec{u}_{r+1}, \ldots, \vec{u}_m$; similarly, the columns of V_r are $\vec{v}_1, \ldots, \vec{v}_r$, and the columns of V_{n-r} are $\vec{v}_{r+1}, \ldots, \vec{v}_n$. Finally, the diagonal entries of Σ (the *singular values*) are $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_{\min\{m,n\}} = 0$, and the diagonal entries of Σ_r are $\sigma_1 \geq \cdots \geq \sigma_r > 0$.

Finally, the matrices $\Sigma^\top \Sigma$ and $\Sigma \Sigma^\top$ are square diagonal matrices which have the following structure:

$$\Sigma^\top \Sigma = \begin{bmatrix} \Sigma_r^2 & 0_{r \times (n-r)} \\ 0_{(n-r) \times r} & 0_{(n-r) \times (n-r)} \end{bmatrix} \qquad \Sigma \Sigma^\top = \begin{bmatrix} \Sigma_r^2 & 0_{r \times (m-r)} \\ 0_{(m-r) \times r} & 0_{(m-r) \times (m-r)} \end{bmatrix}. \tag{9}$$

So the singular values Σ are the *square roots* of the eigenvalues of $A^\top A$ and $AA^\top$, roughly speaking.

Now we are ready to show our main result regarding the existence of the SVD.

> **Theorem 6** (Existence of Full SVD)
> Let $A \in \mathbb{R}^{m \times n}$. There exists a full SVD $A = U\Sigma V^\top$.

The proof of Theorem 6 is on the longer side and may distract from the overall flow of this note, so it is left to Appendix A.2. We fully expect you to read the proof and understand it. It is completely in-scope for the course.

3.2 Computation

The proof of Theorem 6 is constructive, so we can give an algorithm to construct the SVD.

Algorithm 7 Constructing the SVD

function $\textsc{FullSVD}(A \in \mathbb{R}^{m \times n})$
  $r := \textsc{Rank}(A)$
  $(V, \Lambda) := \textsc{Diagonalize}(A^\top A)$ ▷ Sorted so that $\Lambda_{11} \geq \cdots \geq \Lambda_{nn}$
  Unpack $V := \begin{bmatrix} V_r & V_{n-r} \end{bmatrix}$
  Unpack $\Lambda := \begin{bmatrix} \Lambda_r & 0_{r \times (n-r)} \\ 0_{(n-r) \times r} & 0_{(n-r) \times (n-r)} \end{bmatrix}$
  $\Sigma_r := \Lambda_r^{1/2}$
  Pack $\Sigma := \begin{bmatrix} \Sigma_r & 0_{r \times (n-r)} \\ 0_{(m-r) \times r} & 0_{(m-r) \times (n-r)} \end{bmatrix}$
  $U_r := AV_r\Sigma_r^{-1}$
  $U := \textsc{ExtendBasis}(U_r, \mathbb{R}^m)$
  return (U, Σ, V)
end function

NOTE: Sometimes (if $m \ll n$), $A^\top A$ will be large, and $AA^\top$ will be small. In this case, it is more efficient to compute the SVD of $A^\top = XDY^\top$, and then take the transpose $A = YD^\top X^\top$. Letting $U := Y$, $\Sigma := D^\top$, and $V^\top := X^\top$, this recovers an SVD $A = U\Sigma V^\top$.

4 Singular Value Decomposition: Alternate Forms

4.1 Compact SVD

Earlier, we gave a way to find the full SVD, i.e., the decomposition

$$A = \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0_{r \times (n-r)} \\ 0_{(m-r) \times r} & 0_{(m-r) \times (n-r)} \end{bmatrix} \begin{bmatrix} V_r^\top \\ V_{n-r}^\top \end{bmatrix}. \tag{10}$$

But this decomposition seems a bit wasteful. Not only are we storing many zeros in Σ, we are also storing eigenvectors in U_{m-r} and V_{n-r} that never get used, because they match up with coefficients 0 in Σ. To figure out how to remove this inefficiency, we may try to *use the block matrix structure* to simplify the matrix product.

$$A = U\Sigma V^\top \tag{11}$$

$$= \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix} \begin{bmatrix} V_r^\top \\ V_{n-r}^\top \end{bmatrix} \tag{12}$$

$$= \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r V_r^\top \\ 0_{(m-r)\times n} \end{bmatrix} \tag{13}$$

$$= U_r \Sigma_r V_r^\top + U_{m-r} 0_{(m-r)\times n} \tag{14}$$

$$= U_r \Sigma_r V_r^\top. \tag{15}$$

It turns out that we only need the reduced matrices (U_r, Σ_r, V_r) to fully capture A. This decomposition is called the *compact SVD*, and we now give a definition, existence theorem, and algorithm.

Definition 8 (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

$$A = U_r \Sigma_r V_r^\top \tag{16}$$

where

(I) $U_r \in \mathbb{R}^{m\times r}$ is a matrix of *left singular vectors*;

(II) $V_r \in \mathbb{R}^{n\times r}$ is a matrix of *right singular vectors*;

(III) Σ_r is a square diagonal matrix of *singular values*, with positive diagonal entries;

such that the following holds:

(i) U_r is a matrix whose columns are orthonormal eigenvectors of $AA^\top$ corresponding to nonzero eigenvalues of $AA^\top$;

(ii) V_r is a matrix whose columns are orthonormal eigenvectors of $A^\top A$ corresponding to nonzero eigenvalues of $A^\top A$;

(iii) Σ_r^2 is the matrix of nonzero eigenvalues of $AA^\top$ and $A^\top A$;

(iv) $\mathrm{Col}(U_r) = \mathrm{Col}(A)$;

(v) $\mathrm{Col}(V_r) = \mathrm{Col}\left(A^\top\right)$;

Theorem 9 (Existence of Compact SVD)
Let $A \in \mathbb{R}^{m\times n}$ have rank $r \leq \min\{m, n\}$. There exists a compact SVD $A = U_r \Sigma_r V_r^\top$.

Algorithm 10 Constructing the Compact SVD

function $\textsc{CompactSVD}(A \in \mathbb{R}^{m\times n})$
  $r := \textsc{Rank}(A)$
  $(V, \Lambda) := \textsc{Diagonalize}(A^\top A)$ ▷ Sorted so that $\Lambda_{11} \geq \cdots \geq \Lambda_{nn}$
  Unpack $V := \begin{bmatrix} V_r & V_{n-r} \end{bmatrix}$
  Unpack $\Lambda := \begin{bmatrix} \Lambda_r & 0_{r\times(n-r)} \\ 0_{(n-r)\times r} & 0_{(n-r)\times(n-r)} \end{bmatrix}$
  $\Sigma_r := \Lambda_r^{1/2}$
  $U_r := A V_r \Sigma_r^{-1}$
  return (U_r, Σ_r, V_r)
end function

4.2 Outer Product Form of SVD

We are still being marginally wasteful with the compact SVD. In particular, Σ_r is diagonal, so we only need to store r entries, not r^2 for the whole matrix. To reduce our consumption, let us *use the diagonal structure of* Σ_r to simplify our matrix product:

$$\begin{aligned} A &= U_r \Sigma_r V_r^\top && (17) \\ &= \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_r \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} \vec{v}_1^\top \\ \vdots \\ \vec{v}_r^\top \end{bmatrix} && (18) \\ &= \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_r \end{bmatrix} \begin{bmatrix} \sigma_1 \vec{v}_1^\top \\ \vdots \\ \sigma_r \vec{v}_r^\top \end{bmatrix} && (19) \\ &= \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top . && (20) \end{aligned}$$

Writing A as the sum of rank-1 matrices, i.e., the outer products $\vec{u}_i \vec{v}_i^\top$, is called the *outer product form of the SVD*. This representation is maximally space-saving, since there are no more redundant entries in any matrix or vector to store. (This also hints at a way to do data compression – it tells us that we can mostly have the same data but save memory by throwing away rank-1 matrices with small σ – but more on that in the next note).

Definition 11 (Outer Product Form of 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

$$A = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top \tag{21}$$

where

(I) $\{\vec{u}_1, \ldots, \vec{u}_r\} \subset \mathbb{R}^m$ is a set of *left singular vectors*;

(II) $\{\vec{v}_1, \ldots, \vec{v}_r\} \subset \mathbb{R}^n$ is a set of *right singular vectors*;

(III) $\sigma_1, \ldots, \sigma_r > 0$ are positive scalars, i.e., *singular values*;

such that the following holds:

(i) $\vec{u}_1, \ldots, \vec{u}_r$ are orthonormal eigenvectors of $AA^\top$ corresponding to nonzero eigenvalues of $AA^\top$;

(ii) $\vec{v}_1, \ldots, \vec{v}_r$ are orthonormal eigenvectors of $A^\top A$ corresponding to nonzero eigenvalues of $A^\top A$;

(iii) $\sigma_1^2, \ldots, \sigma_r^2$ are the nonzero eigenvalues of $AA^\top$ and $A^\top A$, in descending order;

(iv) $\mathrm{Span}(\vec{u}_1, \ldots, \vec{u}_r) = \mathrm{Col}(A)$;

(v) $\mathrm{Span}(\vec{v}_1, \ldots, \vec{v}_r) = \mathrm{Col}(A^\top)$.

Theorem 12 (Existence of Outer Product Form of SVD)
Let $A \in \mathbb{R}^{m \times n}$ have rank $r \leq \min\{m, n\}$. There exists an outer product form of the SVD $A = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top$.

Algorithm 13 Constructing the Outer Product Form of SVD

```
function OUTERPRODUCTSVD(A ∈ ℝ^{m×n})
    r := RANK(A)
    (V, Λ) := DIAGONALIZE(AᵀA)                    ▷ Sorted so that Λ_11 ≥ ··· ≥ Λ_nn
    Unpack V := [v_1 ··· v_n]
    Unpack Λ := [[Λ_r, 0_{r×(n−r)}], [0_{(n−r)×r}, 0_{(n−r)×(n−r)}]]
    Unpack Λ_r := diag(λ_1, ..., λ_r)
    for i ∈ {1, ..., r} do
        σ_i := √λ_i
        u_i := A v_i / σ_i
    end for
    return ({u_1, ..., u_r}, {σ_1, ..., σ_r}, {v_1, ..., v_r})
end function
```

These new representations are theoretically useful (as in, they help us prove things), and also useful in practical applications (to save memory and computation).

5 Singular Value Decomposition: Uniqueness

We now deal with the question of uniqueness – that is, *to what degree is an SVD of a given matrix unique, if at all*. The answer to this question is difficult and tedious to formalize in general. To begin with, any matrix that "looks" like an SVD is actually an SVD.

Theorem 14
Let $X \in \mathbb{R}^{m\times m}$ and $Y \in \mathbb{R}^{n\times n}$ be square orthogonal matrices. Let $D \in \mathbb{R}^{m\times n}$ be diagonal and have non-negative diagonal entries which are sorted in non-increasing order. Let $A := XDY^\top$. Then $A = XDY^\top$ is an SVD of A.

This basically means that given a decomposition of A into factors that look like the factors of an SVD, we can more or less read off the linear algebraic properties of this matrix, at least those that are detailed in the definition of the Full SVD.

While useful, this theorem has unfortunate implications for uniqueness. This implies that, in our SVD construction algorithm, any basis of $\text{Null}(A)$ and $\text{Null}\left(A^\top\right)$ can be used to construct U_{m-r} and V_{n-r}, for instance. And if two eigenvalues of $A^\top A$ are equal, say λ, (even if $\lambda \neq 0$) then picking any pair of eigenvectors $\vec{v}_i, \vec{v}_j$ that span the corresponding eigenspace $\text{Null}\left(A^\top A - \lambda I\right)$ in the construction still leads to a valid SVD. Overall, there are a lot of degrees of freedom in the construction of the SVD, and so we should expect there to usually be infinitely many SVDs.

However, the ambiguity in choosing U_{m-r} and V_{n-r} does not exist for the compact SVD. And without repeated eigenvalues of $A^\top A$, we do not need to deal with arbitrariness in the eigenspace basis (since all eigenspaces are 1-dimensional), except up to sign. This yields the following theorem.

Theorem 15 (Uniqueness of Compact SVD Without Repeated Singular Values)
Let $A \in \mathbb{R}^{m\times n}$ have rank $r \leq \min\{m, n\}$ and no repeated nonzero singular values. Then the compact SVD of A is unique up to the signs of the $\vec{v}_i$, i.e., there are 2^r possible compact SVDs of A, one for each choice of sign for $\vec{v}_i$, $i \in \{1, \ldots, r\}$.

Finally, if A is square and symmetric, then it is orthonormally diagonalizable, i.e., we may write $A = V\Lambda V^\top$ where V is square orthonormal and Λ is square diagonal. If the diagonal of Λ is non-negative then Theorem 14 tells us that $A = V\Lambda V^\top$ is a diagonalization of A, i.e., $(U, \Sigma, V) = (V, \Lambda, V)$, after permutation of entries of Λ and columns of V to arrange the diagonal entries of Λ in non-increasing order. Of course, the diagonal of Λ may have negative entries; in this case, we can make the corresponding column of V swap sign, so that $V\Lambda$ remains the same, but now Λ has all non-negative entries. Stated more formally, we have the following theorem.

Theorem 16 (Relationship between SVD and Orthonormal Diagonalization)
Let $A \in \mathbb{R}^{n\times n}$ be a square symmetric matrix with orthonormal diagonalization $A = V\Lambda V^\top$. Define the matrices $\Sigma \in \mathbb{R}^{n\times n}$ and $U \in \mathbb{R}^{n\times n}$ by

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} \qquad \text{where} \qquad \sigma_i = |\lambda_i| \tag{22}$$

$$U = \begin{bmatrix} \vec{u}_1 & \cdots & \vec{u}_n \end{bmatrix} \qquad \text{where} \qquad \vec{u}_i = \operatorname{sign}(\lambda_i)\vec{v}_i \tag{23}$$

where in this case only, we define $\operatorname{sign}(0) = 1$. After sorting the diagonal entries of Σ in non-increasing order (simultaneously sorting the corresponding columns of U and V), we have that $A = U\Sigma V^\top$ is an SVD of A.

6 Singular Value Decomposition: Geometric Properties

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

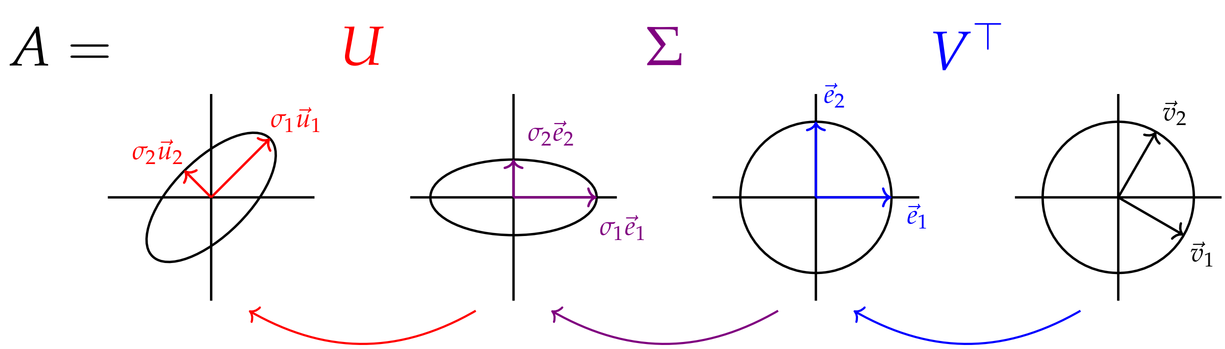
Recall from Note 13 that for any vector $\vec{x}$ and any orthogonal matrix U that $\|U\vec{x}\| = \|\vec{x}\|$. Thus we can interpret multiplication by an orthonormal matrix as a combination of operations that don't change length, such as rotations and reflections.

Since Σ_r is diagonal with entries $\sigma_1, \ldots, \sigma_r$, multiplying a vector by Σ stretches the first entry of the vector by σ_1, the second entry by σ_2, and so on.

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,
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

$$\|A\vec{x}\| = \left\|U\Sigma V^\top \vec{v}_1\right\| = \|U\Sigma\vec{e}_1\| = \|\sigma_1 U\vec{e}_1\| = \|\sigma_1\vec{u}_1\| = \sigma_1\|\vec{u}_1\| = \sigma_1. \tag{24}$$

7 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 17 (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.

Proposition 18 (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 19 (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) $A^\dagger AA^\dagger = A^\dagger$ (weak right inverse property);

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

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

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

Concept Check: Prove Proposition 19. 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 20 (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.

The proof of Theorem 20 is on the longer side and may distract from the overall flow of this note, so it is left to Appendix B.1. We fully expect you to read the proof and understand it. It is completely in-scope for the course.

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 21 (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 $\vec{x}^\star_{\mathrm{LS}} = A^\dagger\vec{b}$.

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

Proof. Using the notation of Theorem 20, if A has full column rank then S has exactly one element, which is the least squares solution $\vec{x}^\star_{\mathrm{LS}} = (A^\top A)^{-1}A^\top\vec{b}$. Hence $A^\dagger = (A^\top A)^{-1}A^\top$ as desired. □

Corollary 22 (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 $\vec{x}^\star_{\mathrm{LN}} = A^\dagger\vec{b}$.

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

Proof. Using the notation of Theorem 20, 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. □

8 Examples

8.1 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

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

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.

8.2 Numerical Example 1

Let's find the SVD for

$$A = \begin{bmatrix} 4 & 4 \\ -3 & 3 \end{bmatrix}. \quad (39)$$

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

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

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

$$\lambda_1 = 32 \qquad \lambda_2 = 18 \quad (41)$$

We can select the orthonormal eigenvectors:

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

The singular values are

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

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.

8.3 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 $[0, 2\pi)$.

8.4 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 (Note 12), 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

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

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, $[-2000, 2000]\,\mathrm{N\,m}$.

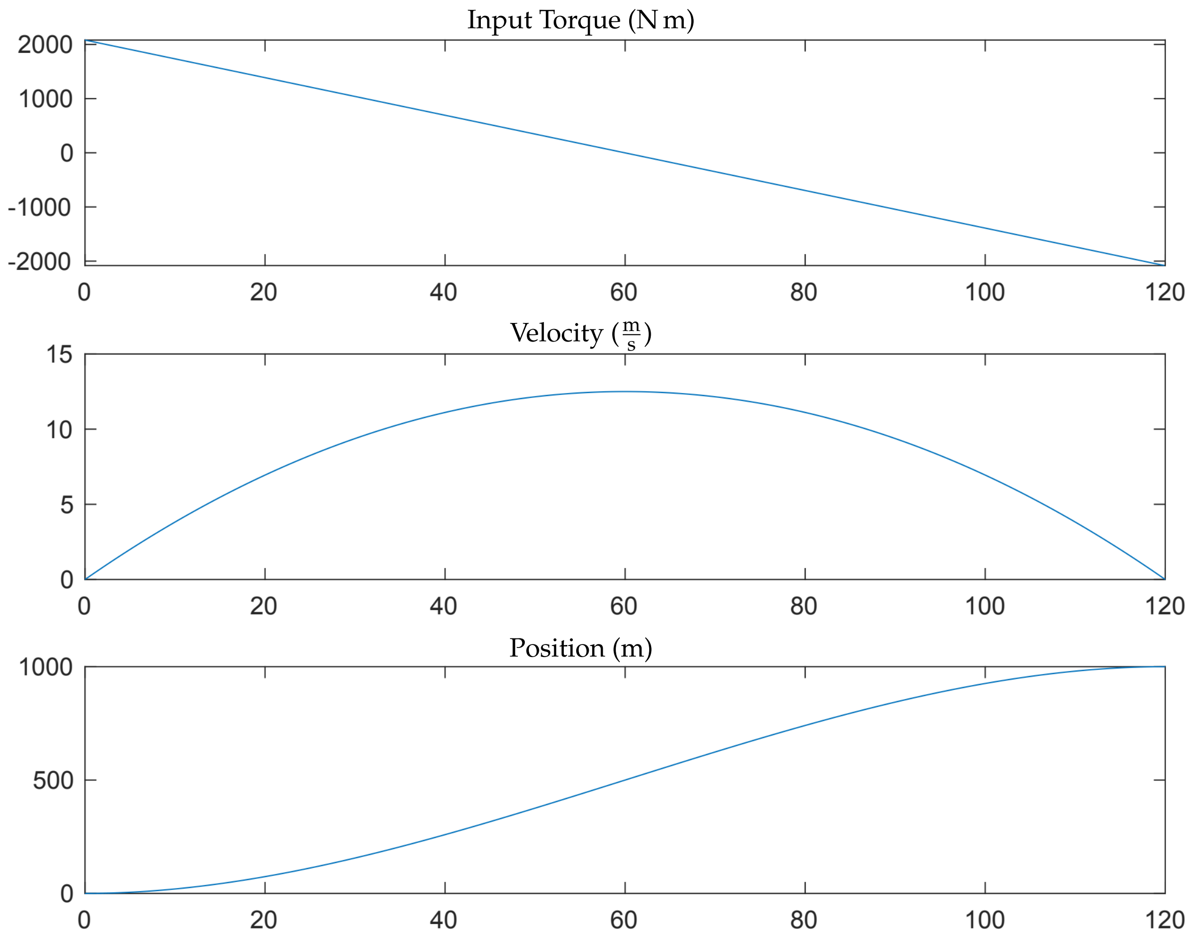


Figure 1: The minimum norm input torque sequence, and the resulting velocity and position profiles, for $RM = 5000\,\mathrm{kg\,m}$, $p^\star = 1000\,\mathrm{m}$, $\Delta = 0.1\,\mathrm{s}$, and $\ell = 1200$. The horizontal axis is time, which ranges from 0 to $\ell\Delta = 120\,\mathrm{s}$. The vehicle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

9 Final Comments

In this note, we discussed the problem of minimum energy control, then turned it into a more abstract least-norm problem. We introduced the singular value decomposition, proved existence and uniqueness, and showed its applicability to the least-norm and least-squares problems via the Moore-Penrose pseudoinverse.

The SVD is a very powerful method to do efficient and expressive data analysis. We will see more applications of this flavor in the next note.

A Proofs for Section 3

A.1 Proof of Proposition 4

Proof of Proposition 4. 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.

To show that $\mathrm{rank}(A^\top A) = r$, note that in EECS 16A Note 23, we showed that $\mathrm{Null}(A) = \mathrm{Null}\left(A^\top A\right)$, and in particular $\dim(\mathrm{Null}(A)) = \dim(\mathrm{Null}\left(A^\top A\right))$. By the Rank-Nullity theorem (EECS 16A Note 8) applied to A, we have that

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

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

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

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

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 (Note 15), and the fact that $A^\top A$ is symmetric, we know that $A^\top A$ is diagonalizable. Again by Note 15, we know that for a diagonalizable matrix, the geometric and algebraic multiplicities of each eigenvalue agree, e.g., $m^a_{A^\top A}(\lambda) = m^g_{A^\top A}(\lambda)$ 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.

We now show that all nonzero eigenvalues of $A^\top A$ are real and positive. Indeed, since $A^\top A$ is symmetric, the Spectral Theorem says that $A^\top A$ has real eigenvalues. It remains to show that they are all non-negative. 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 $\|\vec{v}\|$ is positive. Hence $\|A\vec{v}\|^2$ 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 $\mathrm{rank}(AA^\top) = r$, and that $AA^\top$ has exactly r nonzero eigenvalues, which are real and positive. □

A.2 Proof of Theorem 6

To prove this theorem we require the following lemma.

Lemma 23
Let $X \subseteq Y \subseteq \mathbb{R}^n$ be two subspaces of $\mathbb{R}^n$. If $\dim(X) = \dim(Y)$ then $X = Y$.

Proof. Since $X \subseteq Y$, every basis for X is a subset of a basis for Y. But since all bases for X and for Y have the same size (i.e., $\dim(X) = \dim(Y)$), we have that every basis for X is exactly a basis for Y, and vice versa. Thus $X = Y$. □

Now we may continue to the proof of Theorem 6.

Proof of Theorem 6. We prove this theorem constructively, that is, we give an explicit way to construct U, V, Σ, then we prove that our construction has the properties we want. Here is the construction we use.

1. By Proposition 4, we know that $A^\top A$ is symmetric, so by the Spectral Theorem, it is orthonormally diagonalizable, i.e., we may diagonalize $A^\top A = V\Lambda V^\top$ where V is square orthonormal and Λ is square diagonal.
2. By Proposition 4, all eigenvalues of $A^\top A$ are real and non-negative. Rearrange the columns of V (as well as the rows of $V^\top$) and entries of Λ so that the diagonal of Λ is sorted from greatest to least, while maintaining that $A = V\Lambda V^{-1}$. (This means that make the required swaps on the diagonal of Λ, and at the same time swap the *same* columns of V and the *same* rows of $V^\top$).
3. Partition the sorted V into $V = \begin{bmatrix} V_r & V_{n-r} \end{bmatrix}$.
4. By Proposition 4, $A^\top A$ has exactly r nonzero eigenvalues, which are non-negative. Thus this sorted Λ is zeros except for a diagonal sub-block $\Lambda_r \in \mathbb{R}^{r\times r}$, i.e., $\Lambda = \begin{bmatrix} \Lambda_r & 0_{r\times(n-r)} \\ 0_{(n-r)\times r} & 0_{(n-r)\times(n-r)} \end{bmatrix}$, and Λ_r has all positive entries on the diagonal. Finally, define $\Sigma_r := \Lambda_r^{1/2}$, i.e., since Λ_r is diagonal, Σ_r is the element-wise square root of Λ_r. Note that Σ_r is a diagonal matrix with positive entries on the diagonal.
5. Complete $\Sigma := \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix}$.
6. Since Σ_r has positive entries on its diagonal, Σ_r^{-1} exists; it is a diagonal matrix, such that $(\Sigma_r^{-1})_{ii} = \frac{1}{(\Sigma_r)_{ii}} = \frac{1}{\sigma_i}$. Define $U_r := AV_r\Sigma_r^{-1}$.
7. Find U as the basis completion of U_r to a basis of $\mathbb{R}^m$, i.e., $U := \textsc{ExtendBasis}(U_r, \mathbb{R}^m)$.

We now prove all the different items in Definition 5 for our construction. Note that we have to prove Equation (8) holds, i.e., $A = U\Sigma V^\top$, for our construction; we prove this once we have already proved a few facts that do not rely on this equality.

Note that we prove them in a different order than they are presented in Definition 5.

(ii) We defined V as the orthonormal matrix of eigenvectors of $A^\top A$ (with possible column permutations) via the diagonalization of $A^\top A$, so (ii) holds.

(iv) We defined $\Sigma_r := \Lambda_r^{1/2}$ where Λ_r is the diagonal matrix of nonzero (hence positive) eigenvalues of $A^\top A$. Thus

$$\Sigma^\top\Sigma = \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix}^\top \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix} \tag{71}$$

$$= \begin{bmatrix} \Sigma_r & 0_{r\times(m-r)} \\ 0_{(n-r)\times r} & 0_{(n-r)\times(m-r)} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix} \tag{72}$$

$$= \begin{bmatrix} \Sigma_r^2 & 0_{r\times(n-r)} \\ 0_{(n-r)\times r} & 0_{(n-r)\times(n-r)} \end{bmatrix} \tag{73}$$

$$= \begin{bmatrix} \Lambda_r & 0_{r\times(n-r)} \\ 0_{(n-r)\times r} & 0_{(n-r)\times(n-r)} \end{bmatrix} \tag{74}$$

$$= \Lambda. \tag{75}$$

Thus $\Sigma^\top \Sigma$ is exactly the eigenvalues Λ of $A^\top A$.

(viii) We know that, since V_r is an orthonormal basis of eigenvectors of $A^\top A$ corresponding to nonzero (thus positive) eigenvalues of $A^\top A$, V_{n-r} is an orthonormal basis of eigenvectors of $A^\top A$ corresponding to the eigenvalue 0 of $A^\top A$. Thus $A^\top A V_{n-r} = 0_{n\times(n-r)}$, so $\mathrm{Col}(V_{n-r}) \subseteq \mathrm{Null}(A^\top A)$. But we know that $\mathrm{Null}(A^\top A) = \mathrm{Null}(A)$, and so $\mathrm{Col}(V_{n-r}) \subseteq \mathrm{Null}(A)$.

Now by the rank-nullity theorem,

$$\dim(\mathrm{Null}(A)) = n - \mathrm{rank}(A) = n - r = \mathrm{rank}(V_{n-r}) = \dim(\mathrm{Col}(V_{n-r})) \tag{76}$$

so by Lemma 23, we have that $\mathrm{Null}(A) = \mathrm{Col}(V_{n-r})$.

(vii) We first show that $\mathrm{Col}(A^\top)$ and $\mathrm{Null}(A)$ are orthogonal. Indeed, take any $\vec{x} \in \mathrm{Col}(A^\top)$ such that $\vec{x} = A^\top \vec{w}$, and take any $\vec{y} \in \mathrm{Null}(A)$. Then

$$\langle \vec{x}, \vec{y} \rangle = \left\langle A^\top \vec{w}, \vec{y} \right\rangle = \langle \vec{w}, A\vec{y} \rangle = \left\langle \vec{w}, \vec{0}_m \right\rangle = 0. \tag{77}$$

Thus $\mathrm{Col}(A^\top)$ and $\mathrm{Null}(A) = \mathrm{Col}(V_{n-r})$ are orthogonal subspaces. Since V is a square orthonormal matrix, it has orthonormal columns, and so $\mathrm{Col}(A^\top) \subseteq \mathrm{Col}(V_r)$.

Now we have

$$\dim\left(\mathrm{Col}\left(A^\top\right)\right) = \mathrm{rank}(A^\top) = \mathrm{rank}(A) = r = \mathrm{rank}(V_r) = \dim(\mathrm{Col}(V_r)) \tag{78}$$

so by Lemma 23, we have that $\mathrm{Col}(A^\top) = \mathrm{Col}(V_r)$.

Equation (8) We prove that $A = U\Sigma V^\top$. Indeed, we have

$$\begin{aligned}
U\Sigma V^\top &= \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix} \begin{bmatrix} V_r^\top \\ V_{n-r}^\top \end{bmatrix} && (79)\\
&= U_r \Sigma_r V_r^\top && (80)\\
&= (A V_r \Sigma_r^{-1}) \Sigma_r V_r^\top && (81)\\
&= A V_r V_r^\top. && (82)
\end{aligned}$$

Here to progress, we have already shown that V_{n-r} is an orthonormal basis for $\mathrm{Null}(A)$. Thus

$$AV_{n-r} = A \begin{bmatrix} \vec{v}_{r+1} & \cdots & \vec{v}_n \end{bmatrix} = \begin{bmatrix} A\vec{v}_{r+1} & \cdots & A\vec{v}_n \end{bmatrix} = \begin{bmatrix} \vec{0}_m & \cdots & \vec{0}_m \end{bmatrix} = 0_{m\times(n-r)}. \tag{83}$$

Thus

$$\begin{aligned}
U\Sigma V^\top &= A V_r V_r^\top && (84)\\
&= A V_r V_r^\top + A V_{n-r} V_{n-r}^\top && (85)\\
&= A \begin{bmatrix} V_r & V_{n-r} \end{bmatrix} \begin{bmatrix} V_r^\top \\ V_{n-r}^\top \end{bmatrix} && (86)\\
&= AVV^\top && (87)\\
&= A && (88)
\end{aligned}$$

where the last equality is due to the orthonormality of V.

(i, iii) We first show that U is orthonormal. Indeed, U_{m-r} is constructed such that its columns are an orthonormal set which are orthogonal to each column of U_r, so we have $U_{m-r}^\top U_r = 0_{(m-r)\times r}$ and $U_{m-r}^\top U_{m-r} = I_{m-r}$. Then by our construction for U_r, we have

$$U^\top U = \begin{bmatrix} U_r & U_{m-r} \end{bmatrix}^\top \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \tag{89}$$

$$= \begin{bmatrix} U_r^\top \\ U_{m-r}^\top \end{bmatrix} \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \tag{90}$$

$$= \begin{bmatrix} U_r^\top U_r & U_r^\top U_{m-r} \\ U_{m-r}^\top U_r & U_{m-r}^\top U_{m-r} \end{bmatrix} \tag{91}$$

$$= \begin{bmatrix} (AV_r\Sigma_r^{-1})^\top (AV_r\Sigma_r^{-1}) & 0_{r\times(m-r)} \\ 0_{(m-r)\times r} & I_{m-r} \end{bmatrix}. \tag{92}$$

To calculate the top left term, we can use the identity $\Lambda_r = \Sigma_r^2$ to compute

$$(AV_r\Sigma_r^{-1})^\top (AV_r\Sigma_r^{-1}) = \Sigma_r^{-1} V_r^\top (A^\top A V_r)\Sigma_r^{-1} \tag{93}$$

$$= \Sigma_r^{-1} V_r^\top (V_r \Sigma_r^2)\Sigma_r^{-1} \tag{94}$$

$$= \Sigma_r^{-1} (V_r^\top V_r)(\Sigma_r^2 \Sigma_r^{-1}) \tag{95}$$

$$= \Sigma_r^{-1}\Sigma_r \tag{96}$$

$$= I_r. \tag{97}$$

Thus

$$U^\top U = \begin{bmatrix} (AV_r\Sigma_r^{-1})^\top (AV_r\Sigma_r^{-1}) & 0_{r\times(m-r)} \\ 0_{(m-r)\times r} & I_{m-r} \end{bmatrix} = \begin{bmatrix} I_r & 0_{r\times(m-r)} \\ 0_{(m-r)\times r} & I_{m-r} \end{bmatrix} = I_m \tag{98}$$

and so U is orthonormal. Now we compute

$$AA^\top = (U\Sigma V^\top)(U\Sigma V^\top)^\top \tag{99}$$

$$= U\Sigma V^\top V \Sigma^\top U^\top \tag{100}$$

$$= U\Sigma\Sigma^\top U^\top. \tag{101}$$

Since $\Sigma\Sigma^\top$ is diagonal and U is orthonormal, this is an orthonormal diagonalization of $AA^\top$. Thus $\Sigma\Sigma^\top$ is the matrix of eigenvalues of $AA^\top$ and U is the orthonormal matrix of corresponding eigenvectors of $AA^\top$. This proves (i) and (iii).

(v) The simplest way to show this is to compute A in terms of the reduced matrices U_r, V_r, Σ_r:

$$A = U\Sigma V^\top \tag{102}$$

$$= \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0_{r\times(n-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{bmatrix} \begin{bmatrix} V_r^\top \\ V_{n-r}^\top \end{bmatrix} \tag{103}$$

$$= \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} \Sigma_r V_r^\top \\ 0_{(m-r)\times n} \end{bmatrix} \tag{104}$$

$$= U_r\Sigma_r V_r^\top + U_{m-r} 0_{(m-r)\times n} \tag{105}$$

$$= U_r\Sigma_r V_r^\top. \tag{106}$$

Since for any $\vec{x} \in \mathbb{R}^n$ we have

$$A\vec{x} = U_r\Sigma_r V_r^\top \vec{x} = U_r\vec{y} \qquad \text{for } \vec{y} := \Sigma_r V_r^\top \vec{x} \in \mathbb{R}^r, \tag{107}$$

this implies $\mathrm{Col}(A) \subseteq \mathrm{Col}(U_r)$. Now

$$\dim(\mathrm{Col}(A)) = \mathrm{rank}(A) = r = \mathrm{rank}(U_r) = \dim(\mathrm{Col}(U_r)) \tag{108}$$

so by Lemma 23, we have $\mathrm{Col}(A) = \mathrm{Col}(U_r)$ as desired.

(vi) We first show that $\mathrm{Col}(A)$ and $\mathrm{Null}(A^\top)$ are orthogonal. Indeed, we have already shown that $\mathrm{Col}(A^\top)$ and $\mathrm{Null}(A)$ are orthogonal. Applying this fact to $A^\top$ (instead of A) shows that $\mathrm{Null}(A^\top)$ and $\mathrm{Col}(A) = \mathrm{Col}(U_r)$ are orthogonal subspaces. Since U is a square orthonormal matrix, it has orthonormal columns, so $\text{Null}(A^\top) \subseteq \text{Col}(U_{m-r})$. By the rank-nullity theorem and the fact that $AA^\top = (A^\top)^\top(A^\top)$, we have

$$\dim\left(\text{Null}\left(A^\top\right)\right) = \dim\left(\text{Null}\left(AA^\top\right)\right) = m - \text{rank}(AA^\top) = m - r = \text{rank}(U_{m-r}) = \dim(\text{Col}(U_{m-r})). \tag{109}$$

Hence by Lemma 23, $\text{Null}(A^\top) = \text{Col}(U_{m-r})$.

□

B Proofs for Section 7

B.1 Proof of Theorem 20

Proof of Theorem 20. We first show that $\vec{x}^\star \in S$, then that it is the unique member of S with minimum norm.

From the Orthogonality Principle of Note 13, we have that $\vec{x} \in S$ if and only if $A\vec{x} - \vec{b}$ is orthogonal to $\text{Col}(A)$. 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

$$A^\top A\vec{x} = A^\top\vec{b}. \tag{110}$$

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

$$\begin{aligned}
A^\top A\vec{x}^\star &= A^\top AA^\dagger\vec{b} && (111)\\
&= (U_r\Sigma_r V_r^\top)^\top(U_r\Sigma_r V_r^\top)(V_r\Sigma_r^{-1}U_r^\top)\vec{b} && (112)\\
&= V_r\Sigma_r U_r^\top U_r\Sigma_r V_r^\top V_r\Sigma_r^{-1}U_r^\top\vec{b} && (113)\\
&= V_r\Sigma_r U_r^\top\vec{b} && (114)\\
&= (U_r\Sigma_r V_r^\top)^\top\vec{b} && (115)\\
&= A^\top\vec{b}. && (116)
\end{aligned}$$

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 $A\vec{x} = \text{proj}_{\text{Col}(A)}\left(\vec{b}\right)$ is independent of $\vec{x}$. Suppose that there are two solutions $\vec{x}_1, \vec{x}_2 \in S$. Then since

$$A\vec{x}_1 = A\vec{x}_2 = \text{proj}_{\text{Col}(A)}\left(\vec{b}\right) \tag{117}$$

we see that

$$A(\vec{x}_1 - \vec{x}_2) = A\vec{x}_1 - A\vec{x}_2 = \text{proj}_{\text{Col}(A)}\left(\vec{b}\right) - \text{proj}_{\text{Col}(A)}\left(\vec{b}\right) = \vec{0}_m. \tag{118}$$

Thus $\vec{x}_1 - \vec{x}_2 \in \text{Null}(A)$. 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 $\vec{z} \in \text{Null}(A)$, and further there is exactly one $\vec{x}_0 \in S$ such that $\vec{x}_0$ is orthogonal to $\text{Null}(A)$.[2] For this $\vec{x}_0$ and any $\vec{x} \in S$, we would have

$$\|x\|^2 = \langle\vec{x}, \vec{x}\rangle \tag{119}$$

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

$$
\begin{aligned}
&= \langle \vec{x}_0 + (\vec{x} - \vec{x}_0), \vec{x}_0 + (\vec{x} - \vec{x}_0) \rangle && (120)\\
&= \langle \vec{x}_0, \vec{x}_0 \rangle + \langle \vec{x}_0, \vec{x} - \vec{x}_0 \rangle + \langle \vec{x} - \vec{x}_0, \vec{x}_0 \rangle + \langle \vec{x} - \vec{x}_0, \vec{x} - \vec{x}_0 \rangle && (121)\\
&= \|\vec{x}_0\|^2 + 2 \underbrace{\langle \vec{x}_0, \vec{x} - \vec{x}_0 \rangle}_{=0} + \|\vec{x} - \vec{x}_0\|^2 && (122)\\
&= \|\vec{x}_0\|^2 + \|\vec{x} - \vec{x}_0\|^2 && (123)\\
&\geq \|\vec{x}_0\|^2 && (124)
\end{aligned}
$$

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 $\vec{x} - \vec{x}_0 \in \text{Null}(A)$ and $\vec{x}_0$ is orthogonal to $\text{Null}(A)$. Thus $\vec{x}_0$ is the unique solution to the optimization problem

$$\min_{\vec{x} \in S} \|\vec{x}\|^2. \tag{125}$$

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 $\text{Null}(A)$. So we need to show that $\vec{x}^\star = A^\dagger \vec{b}$ is orthogonal to $\text{Null}(A)$. Indeed,

$$\text{Col}\left(A^\dagger\right) = \text{Col}\left(V_r \Sigma_r^{-1} U_r^\top\right) \subseteq \text{Col}(V_r) = \text{Col}\left(A^\top\right). \tag{126}$$

We have shown in the proof of Theorem 6 that $\text{Col}\left(A^\top\right)$ is orthogonal to $\text{Null}(A)$. Since $A^\dagger \vec{b} \in \text{Col}\left(A^\dagger\right)$, we have $A^\dagger \vec{b} \in \text{Col}\left(A^\top\right)$ and thus $\vec{x}^\star$ is orthogonal to $\text{Null}(A)$. Thus $\vec{x}_0 := \vec{x}^\star$ and the proof is complete. $\square$

Contributors:

- Druv Pai.
- Ashwin Vangipuram.
- Rahul Arya.
- Murat Arcak.
- Anant Sahai.

