Machine Learning

LINEAR ALGEBRA FOR ML – PART 2

Dr. Panagiotis (Panos) Markopoulos
panos@utsa.edu

Linear Subspaces

- **Set of linearly independent vectors:** A set of vectors $\{\mathbf{x}_i\}_{i=1}^{N} \subset \mathbb{C}^M$ are linearly independent set (LI) iff: for any $i \in [N]$, $\nexists \mathbf{y} \in \mathbb{C}^{N-1}$ such that $\mathbf{x}_i = [\mathbf{X}]_{:,[N]\setminus i}\, \mathbf{y}$.

- **Linear subspace:** $\mathcal{S} \subset \mathbb{C}^M$ is a linear subspace iff for any $\mathbf{x}, \mathbf{y} \in \mathcal{S}$ and $a, b \in \mathbb{C}$, $\mathbf{x}a + \mathbf{y}b \in \mathcal{S}$.

- **Dimensionality:** $\dim(\mathcal{S})$ is the cardinality of the largest linearly independent subset in $\mathcal{S}$. It's a way to measure the "size" of $\mathcal{S}$. $\dim(\emptyset) = \dim(\{\mathbf{0}_M\}) = 0$.

Linear Subspaces (cont'd)

- **Span** or **Range** or **Column Space:** $\mathrm{span}(\mathbf{X}) = \{\mathbf{x} \in \mathbb{C}^M : \mathbf{x} = \mathbf{X}\mathbf{y}, \mathbf{y} \in \mathbb{C}^N\}$.
  - $\mathrm{span}(\mathbf{X})$ is a linear subspace.

- **Basis:** $\mathbf{X}$ is a basis for linear subspace $\mathcal{S}$ iff $\mathcal{S} = \mathrm{span}(\mathbf{X})$.
  - A subspace can be spanned by infinitely many distinct bases.
  - Each matrix spans a unique subspace.
  - If $\mathbf{X}^H\mathbf{X} = \mathbf{I}_N$, then $\mathbf{X}$ is an orthonormal basis for $\mathrm{span}(\mathbf{X})$.

Linear Subspaces (cont'd)

- **Orthogonal subspace:** $\mathcal{S}^{\perp} = \{\mathbf{x} \in \mathbb{C}^{M} : \mathbf{x}^{H}\mathbf{y} = 0 \ \forall \mathbf{y} \in \mathcal{S}\}$.
  - It holds $\dim(\mathcal{S}) = M - \dim(\mathcal{S}^{\perp})$.
  - Consider $\mathbf{X} \in \mathbb{C}^{M \times N}$ and $\mathbf{Y} \in \mathbb{C}^{M \times L}$. Then, $\mathbf{X}^{H}\mathbf{Y} = \mathbf{0}_{N,L} \Leftrightarrow \mathrm{span}(\mathbf{X}) = \mathrm{span}(\mathbf{Y})^{\perp}$.

- **Null-space** or **Kernel:** $\mathcal{N}(\mathbf{X}) = \{y \in \mathbb{C}^{N} : \mathbf{X}\mathbf{y} = \mathbf{0}\}$

Fundamental Theorem of Linear Algebra:

- $\dim(\mathrm{span}(\mathbf{X})) = M - \dim(\mathrm{span}(\mathbf{X})^{\perp})$
- $\mathcal{N}(\mathbf{X}^{H}) = \mathrm{span}(\mathbf{X})^{\perp}$

Matrix Rank

- **Matrix rank:** For $\mathbf{X} = [\mathbf{x}_1, \ldots \mathbf{x}_N] \in \mathbb{C}^{M \times N}$, $\text{rank}(\mathbf{X})$ is the size of the largest linearly independent subset among the columns of $\mathbf{X}$, $\{\mathbf{x}_i\}_{i=1}^{N}$.

Remarks:

- $\dim(\text{span}(\mathbf{X})) = \text{rank}(\mathbf{X})$.
- $\text{rank}(\mathbf{X}) \leq \min\{M, N\}$.
- If $\text{rank}(\mathbf{X}) = M$, $\mathbf{X}$ is full row rank.
- If $\text{rank}(\mathbf{X}) = N$, $\mathbf{X}$ is full column-rank.
- If $\text{rank}(\mathbf{X}) = M = N$, $\mathbf{X}$ is square full-rank.
- If $\mathbf{X} \in \mathbb{S}_{M,N}$, then $\text{rank}(\mathbf{X}) = N$.

Inverse and Pseudo-Inverse

- **Inverse:** If $\mathbf{X}$ is square and full rank, then $\mathbf{X}^{-1}$ exists, such that $\mathbf{X}^{-1}\mathbf{X} = \mathbf{X}\mathbf{X}^{-1} = \mathbf{I}_M$.

- **Moore-Penrose Pseudoinverses:**
  - Iff $\mathbf{X} \in \mathbb{C}^{M \times N}$ is full row rank (thus, wide), then the right-hand pseudoinverse $\mathbf{X}^{\dagger R} = \mathbf{X}^H(\mathbf{X}\mathbf{X}^H)^{-1}$ exists, such that $\mathbf{X}\mathbf{X}^{\dagger R} = \mathbf{I}_M$.
  - Iff $\mathbf{X} \in \mathbb{C}^{M \times N}$ is full column rank (thus, tall), then left-hand MP pseudoinverse $\mathbf{X}^{\dagger L} = (\mathbf{X}^H\mathbf{X})^{-1}\mathbf{X}^H$ exists, such that $\mathbf{X}^{\dagger L}\mathbf{X} = \mathbf{I}_N$.
  - If $\mathbf{X}$ is square full rank, then $\mathbf{X}^{\dagger R} = \mathbf{X}^{\dagger L} = \mathbf{X}^{-1}$.

Low-Rank Subspaces

It is often the case that high dimensional data largely reside on lower-dimensional subspaces. Thus, they can be compressed, denoised, visualized, and ML-processed within those subspaces with significant computational/storage gains and limited information loss.

Projection Matrix

❑ **Projection matrix:** $\mathbf{P}$ is a projection matrix iff $\mathbf{P} = \mathbf{P}\mathbf{P}$ and $\mathbf{P} = \mathbf{P}^H$.

Remarks:

- The mapping from projection $\mathbf{P}$ to $\text{span}(\mathbf{P})$ is 1-to-1.
- If $\mathbf{P}$ is projection, then $\mathbf{I}_M - \mathbf{P}$ is also projection with $\text{span}\,(\mathbf{I}_M - \mathbf{P}) = \text{span}(\mathbf{P})^{\perp}$.
- $\text{rank}(\mathbf{P}) = M - \text{rank}(\mathbf{I}_M - \mathbf{P})$.
- For any $\mathbf{x} \in \mathbb{C}^M$,

$$\mathbf{P}\mathbf{x} = \underset{\mathbf{y} \in \text{span}(\mathbf{P})}{\text{argmin}} \parallel \mathbf{y} - \mathbf{x} \parallel_2^2.$$

- If $\mathbf{U} \in \mathbb{S}_{M,K}$, then $\mathbf{U}\mathbf{U}^H$ is a projection matrix on $\text{span}(\mathbf{U}) = \text{span}(\mathbf{U}\mathbf{U}^H)$.

Singular-Value Decomposition

- **Singular-Value Decomposition (SVD):** Every $\mathbf{X} \in \mathbb{C}^{M \times N}$ can be decomposed as $\mathbf{X} = \mathbf{U} \quad \mathbf{\Sigma} \mathbf{V}^H$, where $\mathbf{U}\mathbf{U}^H = \mathbf{U}^H\mathbf{U} = \mathbf{I}_M, \mathbf{V}\mathbf{V}^H = \mathbf{V}^H\mathbf{V} = \mathbf{I}_N$, and $\forall i \neq j, [\mathbf{\Sigma}]_{i,i} = \sigma_i \geq 0$ and $[\mathbf{\Sigma}]_{i,j} = 0$.
  - The columns of $\mathbf{U}$ and $\mathbf{V}$ are the left-hand and right-hand **singular vectors** of $\mathbf{X}$, respectively. $\mathbf{\Sigma}$ is the matrix of **singular values**.
  - Typically, for $i \leq j, \sigma_i \geq \sigma_j$. The number of non-zero singular values, equals $\text{rank}(\mathbf{X})$.

- **Reduced SVD (RSVD):** Let $\rho = rank(X) \leq \min\{M, N\}$. We can decompose $\mathbf{X} = \mathbf{U} \quad \mathbf{\Sigma} \mathbf{V}^H$, so that $\mathbf{U}^H\mathbf{U} = \mathbf{I}_\rho,\ \mathbf{V}^H\mathbf{V} = \mathbf{I}_\rho$, and $[\mathbf{\Sigma}]_{i,i} = \sigma_i > 0$ for $i \in [\rho]$.
  - $\mathbf{U}\mathbf{U}^H$ is a projection matrix for $\text{span}(\mathbf{X}) = \text{span}(\mathbf{U}) = \text{span}(\mathbf{X})$.
  - $\mathcal{N}(\mathbf{X}) = \mathcal{N}(\mathbf{V}^H)$.
  - $\mathbf{\Sigma}$ is Hermitian/Symmetric and invertible.

SVD as a Sum of Rank-1 Factors

Consider $\boldsymbol{X} \in \mathbb{R}^{M \times N}$ admitting reduced SVD $\boldsymbol{X} = \boldsymbol{U\Sigma V^T}$ where $\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_R] \in \mathbb{S}_{M,R}$, $\boldsymbol{\Sigma} = \mathrm{diag}([\sigma_1, \dots, \sigma_R])$, $V = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_R] \in \mathbb{S}_{N,R}$. It holds that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_R > 0$.

Then, $\boldsymbol{X} = \sum_{r=1}^{R} \boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T$.

We observe that $\mathrm{rank}(u_r \sigma_r v_r^T) = 1$ and $\left(u_j \sigma_j v_j^T\right) = 0_{M \times M}$ for all $r \neq j$. [shown in class]

That is, SVD factorizes $\boldsymbol{X}$ into additive component matrices that are rank-1 and orthogonal.

SVD as a Sum of Rank-1 Factors

The "magnitude" of matrix $\boldsymbol{X}$ can be quantified by its Frobenius norm as:

$$||\, \boldsymbol{X} \,||_F^2 = Tr(\boldsymbol{X}^T\boldsymbol{X}) = Tr\left(\left(\boldsymbol{U\Sigma V^T}\right)^T \boldsymbol{U\Sigma V^T}\right) = Tr\left(\boldsymbol{V\Sigma^T U^T U\Sigma V^T}\right) = Tr\left(\boldsymbol{\Sigma}^2\right) = \sum_{r=1}^{R} \sigma_r^2.$$

Alternative derivation:

$$||\, \boldsymbol{X} \,||_F^2 = \left\| \sum_{r=1}^{R} \boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T \right\|_F^2 = Tr\left(\left(\sum_{r=1}^{R} \boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T \right)^T \sum_{j=1}^{R} \boldsymbol{u}_j \sigma_j \boldsymbol{v}_j^T \right) = Tr\left(\sum_{r=1}^{R} \sum_{j=1}^{R} \boldsymbol{v}_r \sigma_r \boldsymbol{u}_r^T\, \boldsymbol{u}_j \sigma_j \boldsymbol{v}_j^T \right)$$

$$= Tr\left(\sum_{r=1}^{R} \boldsymbol{v}_r \sigma_r \boldsymbol{u}_r^T\, \boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T \right) = \sum_{r=1}^{R} Tr(\boldsymbol{v}_r \sigma_r \boldsymbol{u}_r^T \boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T) = \sum_{r=1}^{R} \|\boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T\|_F^2$$

SVD as a Sum of Rank-1 Factors

$|| \boldsymbol{X} ||_F^2 = \sum_{r=1}^{R} ||\boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T||_F^2$, where $||\boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T||_F^2 = Tr\left((\boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T)^T \boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T\right) = Tr(\boldsymbol{v}_r \sigma_r \boldsymbol{u}_r^T \boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T) = \sigma_r^2$

Therefore, again, $|| \boldsymbol{X} ||_F^2 = \sum_{r=1}^{R} \sigma_r^2$.

- Moreover, we notice that the singular value σ_r captures the contribution of $\boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T$ to $\boldsymbol{X}$.
- Since span$(\boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T) = span(\boldsymbol{u}_r)$, term $\boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T$ captures content presence in span$(\boldsymbol{u}_r)$ and the amount of content in $span(\boldsymbol{u}_r)$ is quantified by σ_r.
- Since $\sigma_1 \geq \cdots \geq \sigma_r$, span$(\boldsymbol{u}_r)$ captures more-or-equal data content than span$(\boldsymbol{u}_j)$ for all $r \leq j$.
- The singular vectors are sorted in the amount of data content their subspaces capture.

Low-Rank Approximation with SVD

- Consider $M \times N$ matrix $\boldsymbol{X}$ with reduced SVD: $\boldsymbol{X} = \boldsymbol{U}\,\mathrm{diag}([\sigma_1, \dots, \sigma_R])\boldsymbol{V}^T = \sum_{r=1}^{R} \boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T$.

- For some $D < R$, define matrix $\widetilde{\boldsymbol{X}} = \boldsymbol{U}\,\mathrm{diag}([\sigma_1, \dots, \sigma_D, 0, \dots, 0])\boldsymbol{V}^T = \sum_{r=1}^{D} \boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T$.

- Note that we can write $\widetilde{\boldsymbol{X}}$ in the reduced SVD form $\widetilde{\boldsymbol{X}} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_D]\mathrm{diag}([\sigma_1, \dots, \sigma_D])[\boldsymbol{v}_1, \dots, \boldsymbol{v}_D]^T$.

- $\widetilde{\boldsymbol{X}}$ is a rank-D (low-rank) approximation of $\boldsymbol{X}$.
- The retained "information" can be quantified as $\left\|\widetilde{\boldsymbol{X}}\right\|_F^2 = \sum_{r=1}^{D} \sigma_r^2$ $\left(< \sum_{r=1}^{R} \sigma_r^2 = \|\boldsymbol{X}\|_F^2\right)$
- The discarded "information" in $\boldsymbol{E} = \boldsymbol{X} - \widetilde{\boldsymbol{X}}$ can be quantified as: $\|\boldsymbol{E}\|_F^2 = \sum_{r=D+1}^{R} \sigma_r^2$

- SVD allows for structured low-rank approximation of $\boldsymbol{X}$ with controlled information reduction.

Dimensionality Reduction with SVD

- Another way to see low-rank approximation is through the lens of projection and **dimensionality reduction**.
- Define $\boldsymbol{P} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_D][\boldsymbol{u}_1, \dots, \boldsymbol{u}_D]^T$. This is a projection matrix on S_D=span($[\boldsymbol{u}_1, \dots, \boldsymbol{u}_D]$).
- Then, let us project every column of X on S_D as:

$$[\boldsymbol{P}\boldsymbol{x}_1, \dots, \boldsymbol{P}\boldsymbol{x}_N] = \boldsymbol{P}\boldsymbol{X} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_D][\boldsymbol{u}_1, \dots, \boldsymbol{u}_D]^T[\boldsymbol{u}_1, \dots, \boldsymbol{u}_{\boldsymbol{R}}]\mathrm{diag}([\sigma_1, \dots, \sigma_R])[\boldsymbol{v}_1, \dots, \boldsymbol{v}_D]^T$$

- We notice that

$$[\boldsymbol{u}_1, \dots, \boldsymbol{u}_D]^T[\boldsymbol{u}_1, \dots, \boldsymbol{u}_{\boldsymbol{R}}] = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_D]^T\big[[\boldsymbol{u}_1, \dots, \boldsymbol{u}_D], \boldsymbol{u}_{D+1}, \dots, \boldsymbol{u}_R\big] = [\boldsymbol{I}_D, \boldsymbol{0}_{D\times(R-D)}]$$

$$[\boldsymbol{I}_D, \boldsymbol{0}_{D\times(R-D)}]\mathrm{diag}([\sigma_1, \dots, \sigma_R]) = \mathrm{diag}([\sigma_1, \dots, \sigma_D])[\boldsymbol{I}_D, \boldsymbol{0}_{D\times(R-D)}]$$

$$[\boldsymbol{I}_D, \boldsymbol{0}_{D\times(R-D)}]\,[\boldsymbol{v}_1, \dots, \boldsymbol{v}_R]^T = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_D]^T$$

Dimensionality Reduction with SVD

- Therefore, we find the low-rank approximation is projection on the low-rank span:

$$\boldsymbol{P}\boldsymbol{X} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_D]\mathrm{diag}([\sigma_1, \ldots, \sigma_D])[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_D]^T = \sum_{r=1}^{D} \boldsymbol{u}_r \sigma_r \boldsymbol{v}_r^T = \widetilde{\boldsymbol{X}}.$$

- Also, you should be able to show that $span(\widetilde{\boldsymbol{X}}) = span(\boldsymbol{P}) = span([\boldsymbol{u}_1, \ldots, \boldsymbol{u}_D]) = S_D \subset span(\boldsymbol{X})$

- Finally, you should be able to show that $\boldsymbol{E} = \boldsymbol{X} - \widetilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{P}\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X}$, where $(\boldsymbol{I} - \boldsymbol{P})$ is a projection matrix for $S_D^{\perp}$ and $span(\boldsymbol{E}) = span([\boldsymbol{u}_{D+1}, \ldots, \boldsymbol{u}_R]) \subset span(\boldsymbol{X})$. $span(\boldsymbol{E}) \perp span(\widetilde{\boldsymbol{X}})$ and $span(\widetilde{\boldsymbol{X}}) \cup span(\boldsymbol{E}) = span(\boldsymbol{X})$
- Data dimensionality reduced from $dim(span(\boldsymbol{X})) = R$ **to** $dim(S_D) = D$

Gram Schmidt and QR Decomposition

Given a full-rank matrix $\boldsymbol{X}$ in $\mathbb{R}^{m \times n}$ (with $m \geq n$ and rank L)

Gram-Schmidt Process:

$\boldsymbol{v}_1 = \boldsymbol{x}_1;$
$r = 1;$
$\boldsymbol{q}_r = \boldsymbol{v}_1 \frac{1}{\|\boldsymbol{v}_1\|}$]

For $k = 2$ to n:

$$\boldsymbol{v}_k = \boldsymbol{x}_k - \sum_{j=1}^{r} \boldsymbol{q}_j \boldsymbol{q}_j^T \boldsymbol{x}_k = (I - P_r)x_k \text{ where } P_r = [q_1, \ldots, q_r][q_1, \ldots, q_r]^T$$

If $\|\boldsymbol{v}_k\| > \boldsymbol{0}$, $r = r + 1$,

$$\boldsymbol{q}_r = \boldsymbol{v}_k \frac{1}{\|\boldsymbol{v}_k\|}.$$

Return $L = r$ and $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_L]$.

Gram Schmidt and QR Decomposition

It holds that $L = rank(\boldsymbol{X})$ and $\boldsymbol{Q}^T\boldsymbol{Q} = I_L$

Moreover, $span(\boldsymbol{X}) = span(\boldsymbol{Q})$. That is, $\boldsymbol{Q}$ is an orthonormal basis for $span(\boldsymbol{X})$.

By Gram Schmidt, for all $j = 1,2, \dots, n$, $\boldsymbol{x}_j \in span\left([\boldsymbol{Q}]_{:,1:j}\right)$. Therefore, $\boldsymbol{q}_i^T\boldsymbol{x}_j = 0$ for all $i > j$.

Define $\boldsymbol{R} = \boldsymbol{Q}^T\boldsymbol{X}$. The above mean that if $i > j$ then $[\boldsymbol{R}]_{i,j} = 0$.

Also, $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$.

That is, via GS we created a new decomposition, named “QR”, such that $\boldsymbol{Q}$ is orthonormal basis for the span of $\boldsymbol{X}$ and $\boldsymbol{R}$ has an upper triangular-like structure ($L \times L$ upper triangular appended to the right-hand-side with a $L \times (n - L)$ rectangular block).

SVD vs QR

SVD: $\boldsymbol{X} = \boldsymbol{U\Sigma V}^T$

1) Orthonormal basis (for both rows and columns)
2) Singular values to rank important of rank-1 subspaces that comprise the basis (explainability)
3) By 2, SVD allows for controlled dimensionality reduction; that is, for any desired max loss e, I can determine $D \leq rank(\boldsymbol{X})$ such that $\left\| \boldsymbol{X} - [\boldsymbol{U}]_{:,1:D}[\boldsymbol{U}]_{:,1:D}^T \boldsymbol{X} \right\|_F^2 \leq e$

QR: $\boldsymbol{X} = \boldsymbol{QR}^T$

1) Orthonormal basis (for columns)
2) Structured/sparse subspace coordinates in $\boldsymbol{R}$
3) No controlled DR
4) Faster than SVD

System of Linear Equations

Consider $\mathbf{A} \in \mathbb{C}^{M \times N}$. We want to **solve** the system of equations (M equations, N unknowns):

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

That is, we want to find $\mathbf{x}$ such that the above equation is satisfied, given $\mathbf{A}$ and $\mathbf{y}$.

Case 1: **Inconsistent** (no solution). The system is solvable iff $\mathbf{y} \in \text{span}(\mathbf{A})$. A sufficient (but not necessary) condition is $\text{rank}(\mathbf{A}) = M$.

SLE – General SVD Solution

Case 2: **Consistent**. If $\mathbf{y} \in \text{span}(\mathbf{A}) = \text{span}(\mathbf{U})$, the system $\mathbf{y} = \mathbf{A}\mathbf{x}$ is solvable.

For RSVD $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H$, the general solution is given by

$$\mathbf{x} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^H\mathbf{y} + \mathbf{b}$$

for any $\mathbf{b} \in \mathcal{N}(\mathbf{A}) = \mathcal{N}(\mathbf{V}^H) = \text{span}(\mathbf{V})^\perp = \{\mathbf{z} \in \mathbb{C}^N : \mathbf{z}^H\mathbf{y} = 0 \ \forall \mathbf{y} \in \text{span}(\mathbf{V})\}$.

Verify: $\mathbf{A}\mathbf{x} = \mathbf{A}(\mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^H\mathbf{y} + \mathbf{b}) = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H(\mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^H\mathbf{y} + \mathbf{b}) = \mathbf{U}\mathbf{U}^H\mathbf{y} = \mathbf{y}$.

Thus, there are as many solutions as $|\mathcal{N}(\mathbf{A})|$. Subspaces come in two cardinalities: 0 and infinity.

SLE – Full Row-Rank

<u>Special case 1:</u> $\rho = \text{rank}(\mathbf{A}) = M < N$ (full-row rank). This implies that $\mathbf{A}$ is wide. From FTLA, $\mathcal{N}(\boldsymbol{A}) = \text{span}(\boldsymbol{A}^H)^\perp$ and $\dim(\mathcal{N}(\boldsymbol{A}))$=$\dim(\text{span}(\boldsymbol{A}^H)^\perp) = N - \dim\big(\text{span}(\boldsymbol{A}^H)\big) = N - M > 0$. Since $\dim\big(\mathcal{N}(\boldsymbol{A})\big) > 0$, there are$|\mathcal{N}(\mathbf{A})| = \infty$ solutions to the system. Moreover,

$$\begin{aligned}\mathbf{A}^{\dagger R} &= \mathbf{A}^H(\mathbf{A}\mathbf{A}^H)^{-1} \\ &= \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^H(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^H)^{-1} \\ &= \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^H(\mathbf{U}^H)^{-1}\boldsymbol{\Sigma}^{-2}(\mathbf{U})^{-1} \\ &= \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^H.\end{aligned}$$

More unknowns (N) than independent equations ($\rho = M$).

In this special case, the system solution takes the special form:

$$\mathbf{x} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^H\mathbf{y} + \mathbf{b} \;=\; \mathbf{A}^{\dagger R}\mathbf{y} + \mathbf{b}, \qquad \forall \mathbf{b} \in \mathcal{N}(\mathbf{A})$$

and no decomposition of $\mathbf{A}$ (e.g., SVD) is needed for solving the system.

SLE – Full Column-Rank

<u>*Special case 2:*</u> $\rho = \text{rank}(\mathbf{A}) = N < M$ (full-column rank). This implies that $\dim(\mathcal{N}(\boldsymbol{A})) = 0$ and in turn $\mathcal{N}(\mathbf{A}) = \mathbf{0}_N$ (null-space is trivial). Therefore, the system has $|\mathcal{N}(\mathbf{A})| = 1$ (unique) solution. Moreover,

$$
\begin{aligned}
\mathbf{A}^{\dagger L} &= (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H \\
&= (\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^H\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H)^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^H \\
&= (\mathbf{V}^H)^{-1}\boldsymbol{\Sigma}^{-2}(\mathbf{V})^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^H \\
&= \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^H.
\end{aligned}
$$

More independent equations to satisfy (M) than unknowns to tune (N).

In this special case,

$$\mathbf{x} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^H\mathbf{y} + \mathbf{0}_N = \mathbf{A}^{\dagger L}\mathbf{y}$$

and no decomposition of $\mathbf{A}$ (e.g., SVD) is needed for solving the system.

SLE – Square, Full Rank

Special case 3: $\rho = \text{rank}(\mathbf{A}) = M = N$ (square full rank). This implies that $\mathcal{N}(\mathbf{A}) = \mathbf{0}_N$ and the system again has a single solution. Moreover,

$$\begin{aligned} \mathbf{A}^{\dagger L} = \mathbf{A}^{\dagger R} &= \mathbf{A}^{-1} \\ &= (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^H)^{-1} \\ &= (\mathbf{V}^H)^{-1}\mathbf{\Sigma}^{-1}(\mathbf{U})^{-1} \\ &= \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^H \end{aligned}$$

In this special case,

$$\mathbf{x} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^H\mathbf{y} + \mathbf{0}_N \;\; = \mathbf{A}^{-1}\mathbf{y}$$

and, again, no decomposition of $\mathbf{A}$ (e.g., SVD) is needed for solving the system.

Summary and Common Terminology

SLE: $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$, where $\mathbf{A} \in \mathbb{C}^{M \times N}$.

If $M > N$, the system has more equations than unknowns and it is called **overdetermined**.

If $N > M$, the system has more unknowns than equations and it is called **underdetermined**.

If $\boldsymbol{y} \in \mathrm{span}(\boldsymbol{A})$ the system is called **consistent**; else, it is called **inconsistent**.

A system has no solution iff it is inconsistent.

If a system is consistent, whether over- or under-determined, it can have either a unique solution or infinitely many solutions, depending on $\dim(\mathcal{N}(\boldsymbol{A}))$. **General solution** format: $\mathbf{x} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^{H}\mathbf{y} + \mathbf{b}, \forall \mathbf{b} \in \mathcal{N}(\mathbf{A})$. $\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^{H}\mathbf{y}$ is known as the **particular solution**.

Standard Textbook Simplification

That is, whether the system is overdetermined or underdetermined does not determine by itself if it will have 0, 1, or infinite solutions.

The number of solutions depends on the span of ***A*** and the dimensionality of its null-space.

In the special case where ***A*** is full-rank (i.e., full-column-rank if tall and full-row-rank if wide), the familiar simplification below holds true:

- Overdetermined ($M > N$) with $N = \text{rank}(\mathrm{A})$ $\Rightarrow$ System can be consistent or inconsistent. One unique solution (if consistent) or no solution (if inconsistent).
- Underdetermined ($N > M$) with $M = \text{rank}(\mathrm{A})$ $\Rightarrow$ System can only be consistent. Infinitely many solutions.