Imperial College London

# Maths for Signals and Systems

## Session 3 [Week 4-5]

**Pier Luigi Dragotti**

EEE Department

Imperial College London

p.dragotti@imperial.ac.uk

# Outline and Motivation

- The goal of this group of lectures is to solve $\boldsymbol{Ax} = \boldsymbol{b}$
  - Specifically, given $\boldsymbol{b} \in \mathbb{C}^m$ and $\boldsymbol{A} \in \mathbb{C}^{m \times n}$, we want to understand when a solution $\boldsymbol{x}$ exists and if yes when it is unique.
  - If there is no solution we want to pick the $\boldsymbol{x}$ that *best approximate* $\boldsymbol{Ax} = \boldsymbol{b}$
- We want to answer these questions by checking the properties of $\boldsymbol{A}$ (range and null space) and by using the geometry of vector spaces and linear mapping
- The more interesting cases in engineering are when $m \neq n$ and when there is no solution or the solution is not unique…

# Motivation

- Estimating $\boldsymbol{x}$ from $\boldsymbol{Ax} = \boldsymbol{b}$ is an *inverse problem* which appears in many engineering applications
- Assume that the equation models a filtering operation, then the goal is to remove the distortion due to the filter. For example,
  - you take a picture and (motion) blur has distorted it and you want to remove the blur
  - $\boldsymbol{b}$ is some audio recorded and $\boldsymbol{A}$ models the reflections due to the room. Estimating $\boldsymbol{x}$ is like removing the echo experienced during the recording

**x →**

***b=Ax →***

# Ax=b – Geometrical Interpretation

- We can think of $\boldsymbol{Ax} = \boldsymbol{b}$ in pure maths and geometrical terms as a system of linear equations
  - if $m = n$, we have $n$ equations in $n$ unknowns
  - e.g. $\begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 1 \end{pmatrix}$ each equation is a line, two non-parallel lines intersect in one point which is the **unique** solution,
  - If the lines are parallel we either have no-solution or an infinite number of solutions (two overlapping lines). Clearly $rank(\boldsymbol{A})$ tells us when the solution is **unique**
  - if $m > n$, we have more equations than unknowns ($\boldsymbol{A}$ is '**tall**'), more then two lines normally (but not necessarily) intersect in more than one point unless several lines overlap
  - if $m < n$, we have more unknowns than equation ($\boldsymbol{A}$ is '**fat**'), e.g., when $m = 2, n = 3$, we have two planes in 3-D and unless they are parallel they intersect along a line (infinite number of solutions). So the solution, if it exists, is not unique.

# Ax=b – 'Vector Space' Interpretation

- We can think of the matrix $\boldsymbol{A}$ as describing a linear mapping from $\mathbb{C}^n$ to $\mathbb{C}^m$
- We can then think of $\boldsymbol{Ax}$ as computing linear combinations of columns of $\boldsymbol{A}$,
- **Thus a solution to $Ax = b$ exists if $b$ is in the range space of $A$**
- Any vector in $\mathbb{C}^m$ can be reached by $\boldsymbol{A}$ when $\Re(\boldsymbol{A}) = \mathbb{C}^m$ that is when $\boldsymbol{A}$ is full row rank
- While the range space characterizes existence, the null space characterize uniqueness.
- **When the null space of $A$ is non-trivial, if a solution to $Ax = b$ exists, the solution is not unique.**

# Ax=b - Existence

- We can summarize this discussion with the following theorem:

- **Theorem (existence):** Let $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ then the following statements are equivalent:
  1. For each $\boldsymbol{b} \in \mathbb{C}^m$ there exists at least one solution to $\boldsymbol{Ax} = \boldsymbol{b}$
  2. The range space of $\boldsymbol{A}$ is full: $\Re(\boldsymbol{A}) = \mathbb{C}^m$
  3. $rank(\boldsymbol{A}) = m$
  4. The row of $\boldsymbol{A}$ are linearly independent: $\boldsymbol{A}$ is full row rank

- Note that each of the above conditions implies that $m \leq n$ ('fat/square' matrix)

Imperial College London

# Existence (examples)

1. $\boldsymbol{A} = \begin{pmatrix} 4 & 8 & 6 \\ 0 & 4 & 7 \end{pmatrix}$. The rows are linearly independent so for every $\boldsymbol{b} \in \mathbb{C}^2$ there is at ***least*** one solution.

2. A=$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix}$. The matrix has three rows but only rank 2. The existence theorem is not satisfied and $\boldsymbol{b}$ exist for which $\boldsymbol{Ax} = \boldsymbol{b}$ has no solution. For example, $\boldsymbol{b} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^T$

3. $\boldsymbol{A} = \begin{pmatrix} 6 & 5 & 6 & 3 \\ 2 & 3 & 0 & 1 \\ 4 & 4 & 3 & 2 \end{pmatrix}$. Although this matrix has more columns than rows **it is not full row rank**. Hence, existence theorem is not satisfied and a solution to $\boldsymbol{Ax} = \boldsymbol{b}$ does not always exist! So remember 'more unknowns than equations' or 'underdetermined' does not imply existence. 'Full row rank' does.

# Ax=b – 'Vector Space' Interpretation

- **Thus a solution to $Ax = b$ exists if $b$ is in the range space of $A$**
- While the range space characterizes existence, the null space characterize uniqueness.
- **When the null space of $A$ is non-trivial, if a solution to $Ax = b$ exists, the solution is not unique.**

# Ax=b - Uniqueness

- **Theorem (uniqueness):** Let $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ then the following statements are equivalent:
    1. If there exist a solution to $\boldsymbol{Ax} = \boldsymbol{b}$, it is unique
    2. The null space of $\boldsymbol{A}$ is trivial
    3. $rank(\boldsymbol{A}) = n$
    4. The columns of $\boldsymbol{A}$ are linearly independent: $\boldsymbol{A}$ is full column rank

- Note that each of the above conditions implies that $m \geq n$ ('tall/square' matrix)

# Uniqueness (examples)

- $\boldsymbol{A}=\begin{pmatrix}0 & 1\\ 1 & 1\\ 1 & 0\end{pmatrix}$. The columns are linearly independent, if a solution exists it is unique.

- For example, if $\boldsymbol{b} = [4,6,2]^T$, the unique solution is $\boldsymbol{x} = [2,4]^{\mathrm{T}}$.

- However, if $\boldsymbol{b} = [1,0,0]^T$ no solution exists.

# Uniqueness (examples)

- $\boldsymbol{A}=\begin{pmatrix}1 & 0 & 1\\ 0 & 1 & 1\\ 1 & 1 & 2\end{pmatrix}$. The columns are linearly dependent. Null space $=$ $span\{(1 \quad 1 \quad -1)^T\}$. So uniqueness not satisfied nor existence.

- $\boldsymbol{A}=\begin{pmatrix}2 & 5 & 3\\ 3 & 5 & 2\\ 3 & 7 & 4\\ 7 & 8 & 1\end{pmatrix}$. The columns are linearly dependent so despite being overdetermined uniqueness cannot be guaranteed

# Ax=b – Existence and Uniqueness

- **Theorem (existence and uniqueness):** Let $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ then the following statements are equivalent:
  1. For each $\boldsymbol{b} \in \mathbb{C}^{m}$ there exists a solution to $\boldsymbol{Ax} = \boldsymbol{b}$ and it is unique
  2. The null space of $\boldsymbol{A}$ is trivial and dimension of range space is $m$
  3. $rank(\boldsymbol{A}) = n = m$
  4. $\boldsymbol{A}$ is square and nonsingular

- The case of $\boldsymbol{A}$ being square and nonsingular is of less interest, so we now focus on finding $\boldsymbol{x}$ given $\boldsymbol{b}$ and $\boldsymbol{A}$ for the other cases.

# Solutions to Ax=b

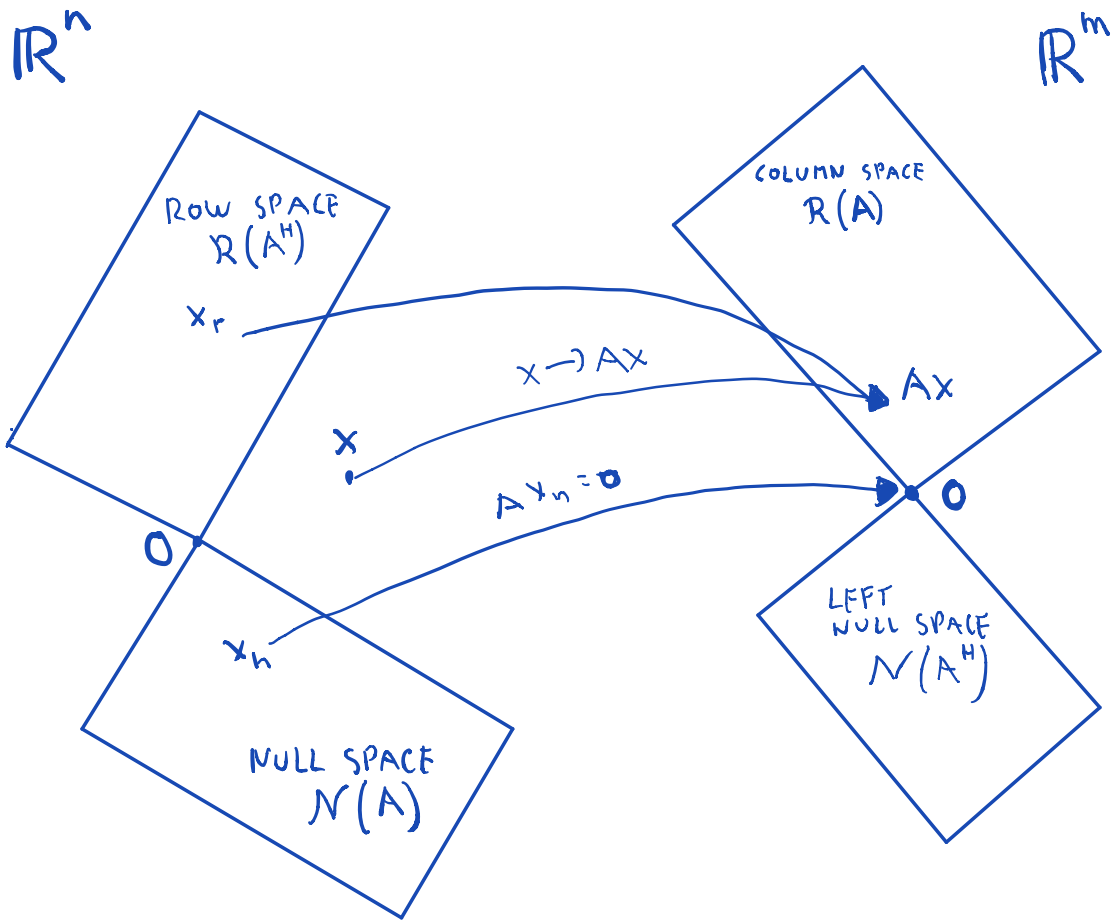
- Recall that we are given $\boldsymbol{A}$ and $\boldsymbol{b}$ and we want to find $\boldsymbol{x}$
- We begin by focusing on the full column rank case
- This means that $\boldsymbol{A}$ is tall (or square) and that if the solution exists it is unique
- So our goal is to develop a single method that either finds the $\boldsymbol{x}$ which is the solution to $\boldsymbol{Ax} = \boldsymbol{b}$ if it exists or finds the '*best approximation*' to $\boldsymbol{Ax} = \boldsymbol{b}$
- For *best approximation* we mean the $\boldsymbol{x}$ that minimizes $\|\boldsymbol{Ax} - \boldsymbol{b}\|_2$ (least-squares approximation)
- In order to achieve that we need to introduce a few new notions

# Four Fundamental Subspaces

- Given a linear mapping described by $\boldsymbol{A} \in \mathbb{C}^{\boldsymbol{m}\times\boldsymbol{n}}$, we have previously introduced the notion of range of $\boldsymbol{A}$, $\Re(\boldsymbol{A})$, and null space of $\boldsymbol{A}$, $\mathcal{N}(\boldsymbol{A})$
- Consider now the linear mapping described by $\boldsymbol{A}^H$, this maps vectors from $\mathbb{C}^m$ to $\mathbb{C}^n$ (however, remember that $\boldsymbol{A}^H$ is normally **not** the inverse of $\boldsymbol{A}$)
- We then have $\Re\left(\boldsymbol{A}^{\boldsymbol{H}}\right)$ and $\mathcal{N}(\boldsymbol{A}^{\boldsymbol{H}})$
- These four subspaces are related as follows (without proof):
  - $\mathbb{C}^m = \Re(\boldsymbol{A}) \oplus \mathcal{N}(\boldsymbol{A}^{\boldsymbol{H}})$, $\quad \Re^{\perp}(\boldsymbol{A}) = \mathcal{N}(\boldsymbol{A}^{\boldsymbol{H}})$, $\quad \mathcal{N}^{\perp}\left(\boldsymbol{A}^{\boldsymbol{H}}\right) = \Re(\boldsymbol{A})$,
  - $\mathbb{C}^n = \Re\left(\boldsymbol{A}^{\boldsymbol{H}}\right) \oplus \mathcal{N}(\boldsymbol{A})$, $\quad \Re^{\perp}\left(\boldsymbol{A}^{\boldsymbol{H}}\right) = \mathcal{N}(\boldsymbol{A})$, $\quad \mathcal{N}^{\perp}(\boldsymbol{A}) = \Re\left(\boldsymbol{A}^{\boldsymbol{H}}\right)$,

Imperial College London

# Four Fundamental Subspaces



Imperial College London

# Four Fundamental Subspaces – Worked Example

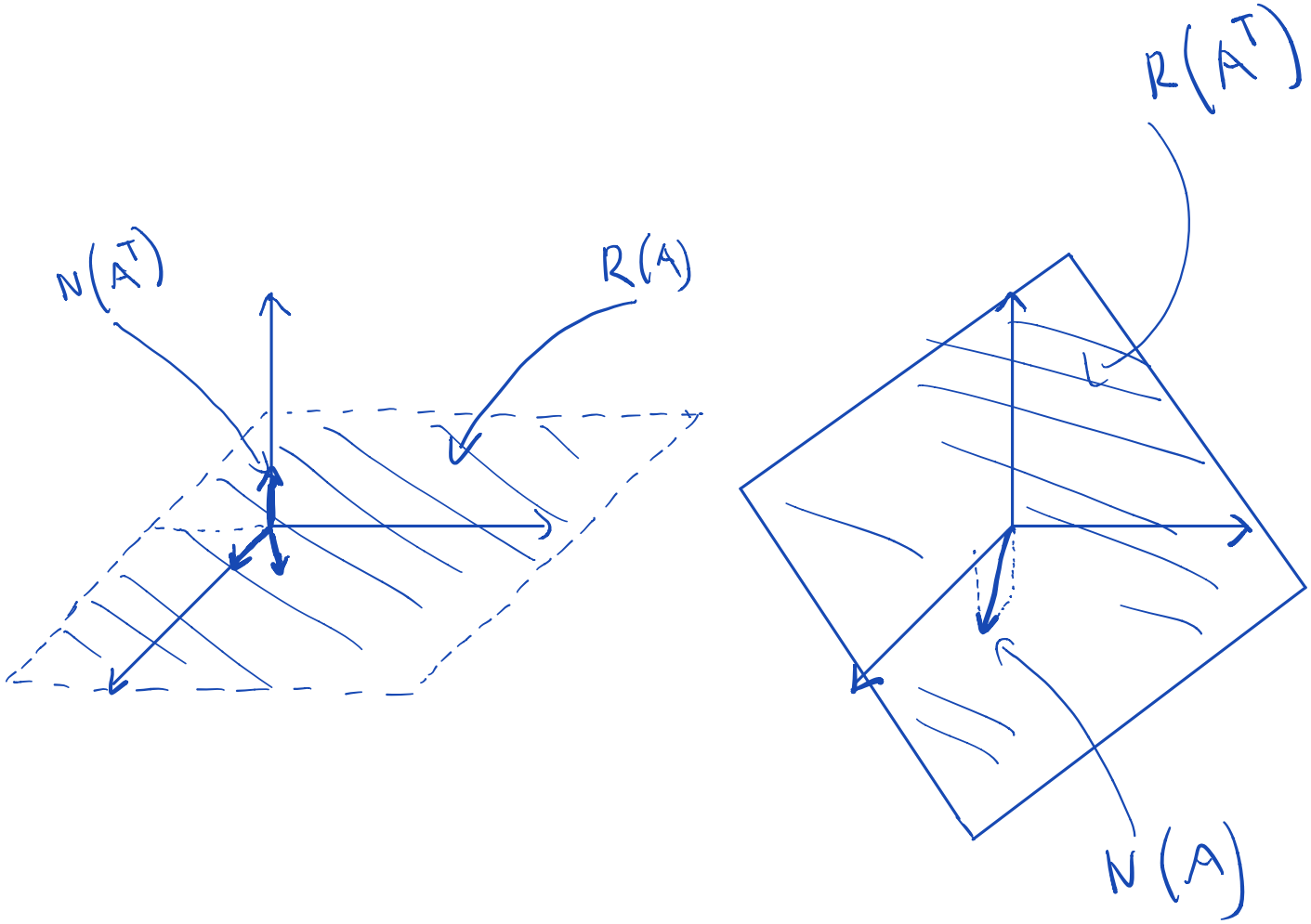
- **Question:** Find the four fundamental subspaces of $\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
- We first should row-reduce $\boldsymbol{A}$ to find the dimension of the range space and null space. $\boldsymbol{A}$ is already in a reduced form and clearly the range space has dimension 2 and a possible basis is given by the first two columns of $\boldsymbol{A}$, so $\Re(\boldsymbol{A}) = \text{span}\{[1,0,0]^{\text{T}}, [1,1,0]^{\text{T}}\}$,
- dim($\mathcal{N}(\boldsymbol{A})$)=3-dim($\Re(\boldsymbol{A})$)=1 and the basis is found by computing $\boldsymbol{An} = \boldsymbol{0}$ which in this case is trivially given by $[1,0,-1]^T$, so $\mathcal{N}(\boldsymbol{A}) = span\{[1,0,-1]^T\}$
- dim($\mathcal{N}(\boldsymbol{A^H})$)=3-rank($\boldsymbol{A}$)=1; and we find a basis by imposing $\mathcal{N}^{\perp}(\boldsymbol{A^H}) = \Re(\boldsymbol{A})$, so $\mathcal{N}(\boldsymbol{A^H})$=$span\{[0,0,1]^T\}$
- Finally a basis for $\boldsymbol{A}^H$ is given by the first two rows of $\boldsymbol{A}$ therefore $\Re(\boldsymbol{A^H}) = span\{[1,1,1]^T, [0,1,0]^T\}$

# Four Fundamental Subspaces – Worked Example

- **Question:** Find the four fundamental subspaces of $\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$
- **Answer:** In conclusion:

- $\Re(\boldsymbol{A}) = \text{span}\{[1,0,0]^{\text{T}}, [1,1,0]^{\text{T}}\}$, $\mathcal{N}(\boldsymbol{A}) = span\{[1,0,-1]^{T}\}$
- $\Re\left(\boldsymbol{A}^{\boldsymbol{H}}\right) = span\{[1,1,1]^{T}, [0,1,0]^{T}\}$, $\mathcal{N}(\boldsymbol{A}^{\boldsymbol{H}})$=$span\{[0,0,1]^{T}\}$

# Four Fundamental Subspaces - Example



# Matrix Inverses

- Assume $\boldsymbol{A}$ is nonsingular and square, then we know that it has an inverse $\boldsymbol{A^{-1}}$
- This means that $\boldsymbol{AA^{-1} = A^{-1}A = I}$
- If for example the matrix is not square, it does not have an inverse but it may have a:
  - **Left Inverse** if a matrix $\boldsymbol{B}$ is such that $\boldsymbol{BA = I}$ (but $\boldsymbol{AB \neq I}$)
  - **Right Inverse** if there exists a matric $\boldsymbol{C}$ such that $\boldsymbol{AC=I}$
- The left inverse exists if $\boldsymbol{A}$ is full column rank (tall matrix)
- The right inverse exists if $\boldsymbol{A}$ is full row rank (fat matrix)
- Note that these matrices might not be unique!

Imperial College London

# Matrix Inverses: Examples

Assume $\boldsymbol{A} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 7 & 0 \end{pmatrix}$, this matrix is full row-rank so it has a right inverse given

by: $\boldsymbol{A_R} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/7 \\ c_1 & c_2 \end{pmatrix}$

Clearly $\boldsymbol{AA_R} = \boldsymbol{I}$ but $\boldsymbol{A_R A} \neq \boldsymbol{I}$

**Exercise:** Find a left inverse of $\boldsymbol{A^T}$

# Projections and Approximations

Assume that we have a vector $\boldsymbol{x} \in H$ where $H$ is a vector space and assume we want to find an approximation $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}$ in a subspace $V$ with $V \subset H$. The projection theorem tells us how to achieve that:

**Projection Theorem:** *Let $V$ be a closed subspace of space $H$ and let $x$ be a vector in $H$.*

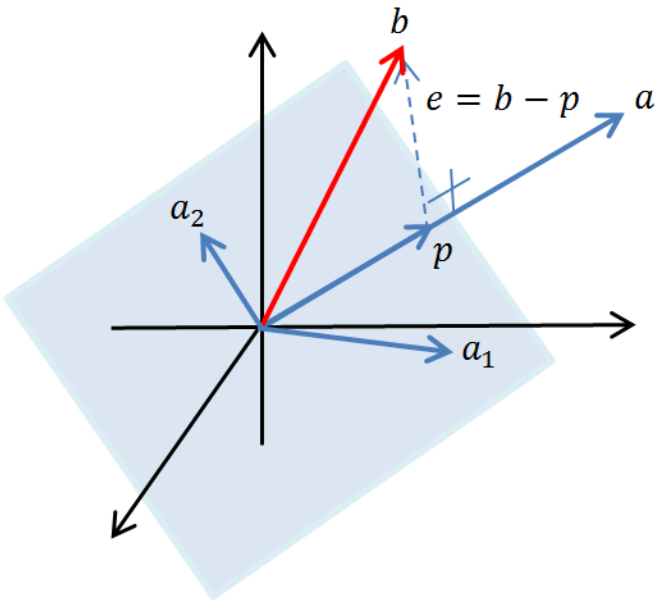
1. *Existence: There exists $\hat{x} \in V$ such that $\|x - \hat{x}\| \leq \|x - v\|$ for all $v \in V$.*
2. *Orthogonality: $x - \hat{x} \perp V$ is necessary and sufficient for determining $\hat{x}$.*
3. *Uniqueness: The vector $\hat{x}$ is unique.*
4. *Linearity: $\hat{x} = Px$ where $P$ is a linear operator that depends on $V$ and not on $x$.*
5. *Idempotency: $P(Px) = Px$ for all $x \in H$.*
6. *Self-adjointness: $P = P^H$.*

# Projections and Approximations

- Linearity of the projection means we can represent it using a matrix $\boldsymbol{P}$
- Idempotent means that the matrix satisfies $\boldsymbol{P}^2 = \boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$
- Self-Adjoint means $\boldsymbol{P} = \boldsymbol{P}^H$
- Idempotency really means that once you have computed the projection, if you apply the projection operator again you get the same result

# Projections and Approximations

- Assume we want to project a vector $\boldsymbol{x} \in \mathbb{C}^m$ onto the subspace $\mathbb{C}^n$ with $n < m$ (see the figure below for the case $m = 3$ and $n = 2$).
- Assume the vectors $\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_n}$ span $\mathbb{C}^n$, we stack them together to form the tall matrix $\boldsymbol{A} \in \mathbb{C}^{m \times n}$. This matrix is full column rank.
- We now want to find the projection matrix $\boldsymbol{P}$



# Projections and Approximations

Claim: the matrix $\boldsymbol{P} = \boldsymbol{A}(\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H$ is the orthogonal projection operator

Proof: First note that $\boldsymbol{A}^H\boldsymbol{A}$ is square and nonsingular because

$$\boldsymbol{A}^H\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \Rightarrow \boldsymbol{x}^H\boldsymbol{A}^H\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \Rightarrow \|\boldsymbol{A}\boldsymbol{x}\|^2 = \boldsymbol{0} \Rightarrow \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \Rightarrow \boldsymbol{x} = \boldsymbol{0}$$

since the columns of $\boldsymbol{A}$ are linearly independent and so only the zero vector can give us zero. This means that $\boldsymbol{P}$ is well defined. Moreover, $\boldsymbol{P}$ is idempotent since

$$\boldsymbol{P}^2 = \boldsymbol{A}(\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H \cdot \boldsymbol{A}(\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H = \boldsymbol{A}(\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H = \boldsymbol{P}$$

$\boldsymbol{P}$ is also self-adjoint so it is the projection operator.

It is also easy to verify that orthogonality is satisfied: $\boldsymbol{A}^H(\boldsymbol{x} - \boldsymbol{P}\boldsymbol{x}) = \boldsymbol{0}$

Imperial College London

# An Application of the Projection Theorem - The Gram-Schmidt orthogonalization process

- We introduced in the first week the notion of orthogonal matrices:
- Consider a set of vector $\boldsymbol{v}_i \; i = 1,2, \dots, n$ which forms an **orthonormal** basis and create a matrix $\boldsymbol{A}$ from stacking them one after the other:

$$\boldsymbol{A} = \begin{pmatrix} v_{1,1} & \cdots & v_{1,n} \\ \vdots & \ddots & \vdots \\ v_{n,1} & \cdots & v_{n,n} \end{pmatrix}$$

- Clearly $\boldsymbol{A}^H \boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^H = \boldsymbol{I}$, where $\boldsymbol{A}^H$ is the Hermitian transpose of $\boldsymbol{A}$ and $\boldsymbol{I}$ is the identity matrix. This means that $\boldsymbol{A}^H = \boldsymbol{A}^{-1}$.
- This type of matrices are called orthogonal matrices and **they preserve the induced norm**:
- If the set of vector $\boldsymbol{v}_i \; i = 1,2, \dots, n$ is not orthonormal, the process to make them orthogonal is known as Gram-Schmidt process.

Imperial College London

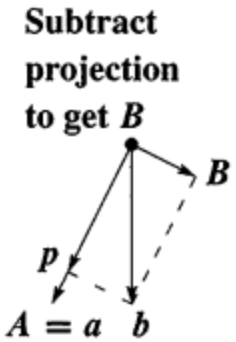
# The Gram-Schmidt orthogonalization process

- The goal here is to start with three independent vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ and construct three orthogonal vectors $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ and finally three orthonormal vectors.

$$\boldsymbol{q}_1 = \boldsymbol{A}/\|\boldsymbol{A}\|, \boldsymbol{q}_2 = \boldsymbol{B}/\|\boldsymbol{B}\|, \boldsymbol{q}_3 = \boldsymbol{C}/\|\boldsymbol{C}\|$$

- We begin by choosing $\boldsymbol{A} = \boldsymbol{a}$. This first direction is accepted.
- The next direction $\boldsymbol{B}$ must be perpendicular to $A$. We start with $\boldsymbol{b}$ and subtract its projection along $A$. This leaves the part of $\boldsymbol{b}$ which we call vector $\boldsymbol{B}$ (what we knew before as the error of projection), defined as (remember that the projection operator is $\boldsymbol{P} = \boldsymbol{A}(\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H$ ):

$$\boldsymbol{B} = \boldsymbol{b} - \frac{\boldsymbol{A}\boldsymbol{A}^T}{\boldsymbol{A}^T\boldsymbol{A}}\boldsymbol{b}$$



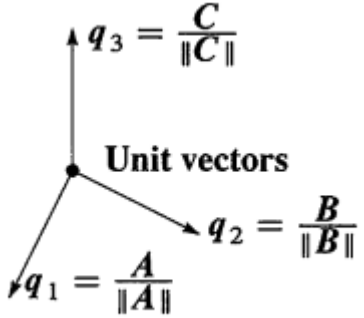

- By the projection theorem we know that the projection error is orthogonal to $\boldsymbol{a}$

# The Gram-Schmidt process

- The third direction starts with $\boldsymbol{c}$. This is not a combination of $\boldsymbol{A}$ and $\boldsymbol{B}$.
- Most likely $\boldsymbol{c}$ is not already perpendicular to $\mathbf{A}$ and $\mathbf{B}$ .
- Therefore, we subtract the projections of $\boldsymbol{c}$ along $\mathbf{A}$ and $\mathbf{B}$ to get $\boldsymbol{C}$:

$$\boldsymbol{C} = \boldsymbol{c} - \frac{\boldsymbol{A}\mathbf{A}^{T}}{\mathbf{A}^{\mathrm{T}}\mathbf{A}}\boldsymbol{c} - \frac{\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}}{\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}}\boldsymbol{c}$$

- **In general we subtract from every new vector its projections in the directions already set.**
- If we had a fourth vector $\boldsymbol{d}$, we would subtract three projections onto $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ to get $\boldsymbol{D}$.
- We make the resulting vectors orthonormal.
- This is done by dividing the vectors with their magnitudes.



Imperial College London

# The Gram-Schmidt process and QR Factorization

- Gram-Schimdt can be used to factorize a matrix $\boldsymbol{A}$ into the product of an orthogonal and an upper triangular matrix. We focus only on the case $\boldsymbol{A}$ is squared and invertible.
- Consider a squared matrix $\boldsymbol{A}$ made of $n$ linearly independent vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, .., \boldsymbol{a}_n$:

$$\boldsymbol{A} = \begin{pmatrix} | & | & \cdots & | \\ \boldsymbol{a_1} & \boldsymbol{a_2} & \cdots & \boldsymbol{a_n} \\ | & | & \cdots & | \end{pmatrix}$$

- Apply Gram-Schmidt:
  - $\boldsymbol{u}_1 = \boldsymbol{a}_1$; $\quad \boldsymbol{e}_1 = \frac{u_1}{||u_1||}$
  - $\boldsymbol{u}_2 = \boldsymbol{a}_2 - \frac{\langle u_1, a_2 \rangle}{||u_1||^2}\boldsymbol{u}_1$; $\quad \boldsymbol{e}_2 = \frac{u_2}{||u_2||}$
  - $\boldsymbol{u}_k = \boldsymbol{a}_k - \sum_{j=1}^{k-1} \frac{\langle u_j, a_k \rangle}{||u_j||^2}\boldsymbol{u}_\mathrm{j}$ $\quad \boldsymbol{e}_k = \frac{u_k}{||u_k||}$

Imperial College London

# The Gram-Schmidt process and QR Factorization

- We now have (note that: $\langle \boldsymbol{e}_i, \boldsymbol{a}_i \rangle = ||\boldsymbol{u}_i||$ ):

- $\boldsymbol{a}_1 = \langle \boldsymbol{e}_1, \boldsymbol{a}_1 \rangle \boldsymbol{e}_1$
- $\boldsymbol{a}_2 = \langle \boldsymbol{e}_1, \boldsymbol{a}_2 \rangle \boldsymbol{e}_1 + \langle \boldsymbol{e}_2, \boldsymbol{a}_2 \rangle \boldsymbol{e}_2$
- $\boldsymbol{a}_3 = \langle \boldsymbol{e}_1, \boldsymbol{a}_3 \rangle \boldsymbol{e}_1 + \langle \boldsymbol{e}_2, \boldsymbol{a}_3 \rangle \boldsymbol{e}_2 + \langle \boldsymbol{e}_3, \boldsymbol{a}_3 \rangle \boldsymbol{e}_3$
- $\vdots$

- This means we can write $\boldsymbol{A} = \boldsymbol{QR}$ where $\boldsymbol{Q}$ is orthogonal and $\boldsymbol{R}$ is upper triangular with:

- $\boldsymbol{Q} = (\boldsymbol{e}_1 \quad \cdots \quad \boldsymbol{e}_n)$ and $\boldsymbol{R} = \begin{pmatrix} \langle \boldsymbol{e}_1, \boldsymbol{a}_1 \rangle & \langle \boldsymbol{e}_1, \boldsymbol{a}_2 \rangle & \cdots & \langle \boldsymbol{e}_1, \boldsymbol{a}_n \rangle \\ 0 & \langle \boldsymbol{e}_2, \boldsymbol{a}_2 \rangle & \cdots & \langle \boldsymbol{e}_2, \boldsymbol{a}_n \rangle \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & \langle \boldsymbol{e}_n, \boldsymbol{a}_n \rangle \end{pmatrix}$

# Solution to Ax=b

- Let's go back to the problem of finding $\boldsymbol{x}$ such that $\boldsymbol{Ax} = \boldsymbol{b}$
- We are focusing on the full column rank case
- This means that $\boldsymbol{A}$ is tall and that if the solution exists it is unique
- We essentially need to find the left inverse of $\boldsymbol{A}$
- We will achieve that through a somewhat convoluted but geometrical approach
- We look for $\boldsymbol{x}_{LS}$ such that $\|\boldsymbol{Ax}_{\boldsymbol{LS}} - \boldsymbol{b}\| \leq \|\boldsymbol{Ax} - \boldsymbol{b}\|$ for any $\boldsymbol{x}$
- So that if a solution to $\boldsymbol{Ax} = \boldsymbol{b}$ exists $\boldsymbol{x}_{LS}$ is that solution

# Solution to Ax=b

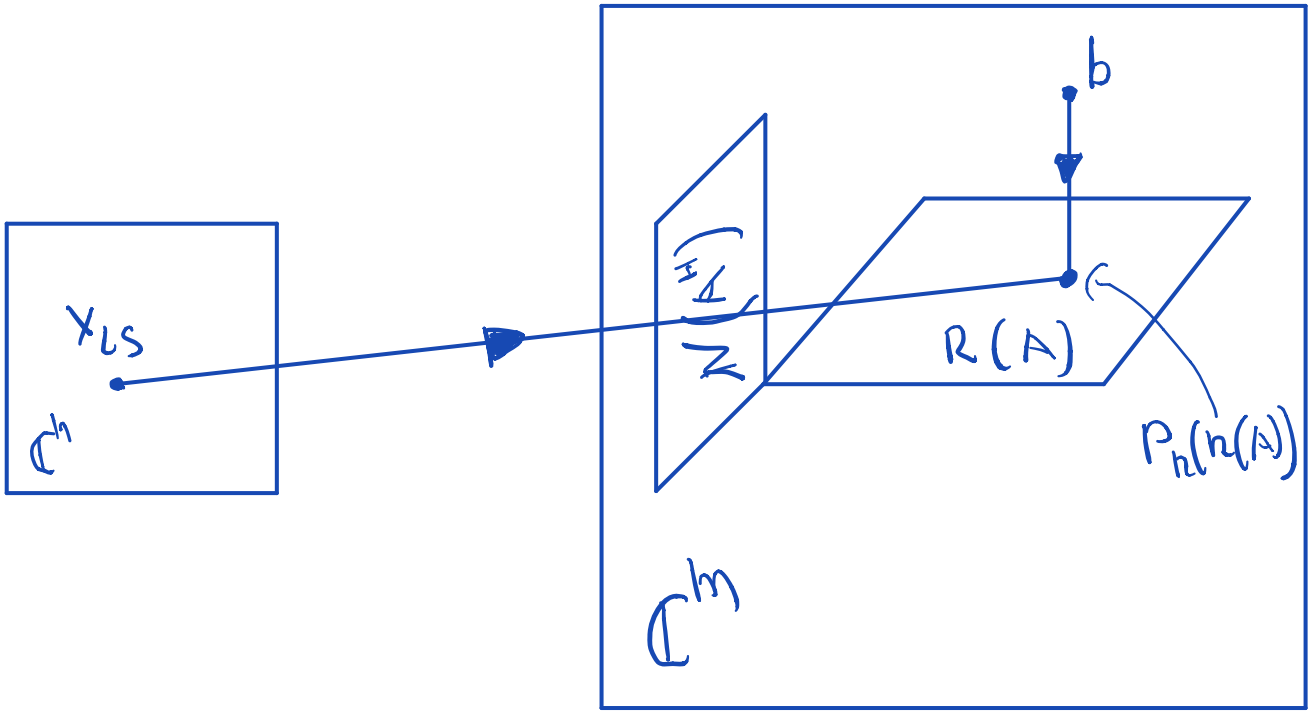
Claim: the vector $\hat{\boldsymbol{x}}$ minimizes $\|\boldsymbol{Ax}-\boldsymbol{b}\|$ if and only if $\boldsymbol{A^H A\hat{x}} = \boldsymbol{A^H b}$ (normal equation)

Proof:

- Minimizing $\|\boldsymbol{Ax}-\boldsymbol{b}\|$ is equivalent to minimizing $\|\hat{\boldsymbol{b}}-\boldsymbol{b}\|$ where $\hat{\boldsymbol{b}} \in \mathfrak{R}(\boldsymbol{A})$.
- By the projection theorem: $\boldsymbol{b}-\hat{\boldsymbol{b}} \in \mathfrak{R}(\boldsymbol{A})^{\perp} \Rightarrow \boldsymbol{b}-\hat{\boldsymbol{b}} \in \mathcal{N}(\boldsymbol{A^H})$
- Therefore, $\boldsymbol{A^H}(\boldsymbol{b}-\hat{\boldsymbol{b}}) = \boldsymbol{0} \Rightarrow \boldsymbol{A^H b} = \boldsymbol{A^H}\hat{\boldsymbol{b}} \Rightarrow \boldsymbol{A^H A\hat{x}} = \boldsymbol{A^H b}$
- Conversely $\boldsymbol{A^H A\hat{x}} = \boldsymbol{A^H b} \Rightarrow \boldsymbol{A^H}(\boldsymbol{b}-\boldsymbol{A\hat{x}}) = \boldsymbol{0}$

# Least-Squares Solution

- Geometric Interpretation to LS Solution

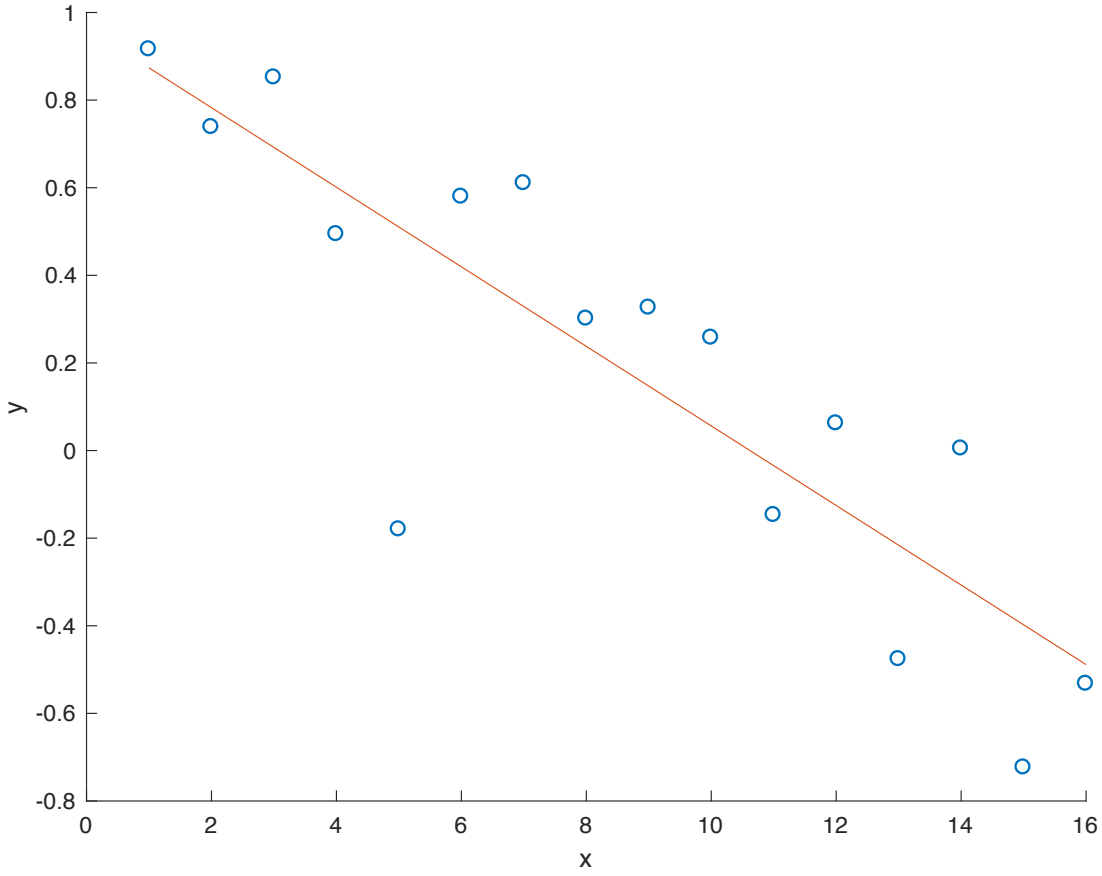


# Example: Linear Regression

- Assume you observe a function f(x) at instants $x_1, x_2, \dots, x_n$: $y_i = f(x_i) \quad i = 1,2, \dots, n$
- You want to approximate f(x) with a line, that is, $y_i \approx ax_i + b$
- In matrix/vector form: $\boldsymbol{y} = \boldsymbol{Ac} + \boldsymbol{e}$, where $\boldsymbol{y} = [y_1 \quad y_2 \quad \cdots \quad y_n]^T$, $\boldsymbol{c} = \begin{bmatrix} a \\ b \end{bmatrix}$, $\boldsymbol{e}$ is the approximation error and $\boldsymbol{A} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$.

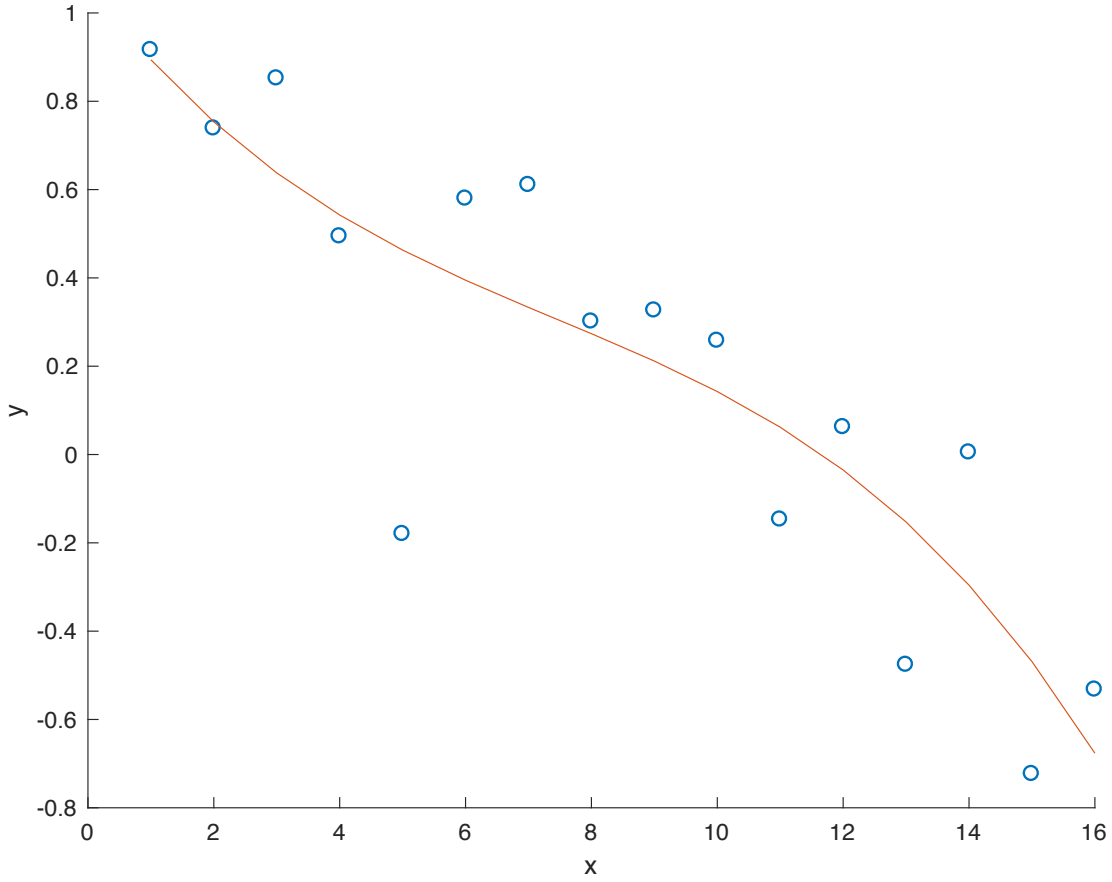
- You want to minimize the norm of the error. Note that $\boldsymbol{A}$ is tall and full column rank, so you can use the least-square solution discussed before: $\boldsymbol{A^H A\hat{c}} = \boldsymbol{A}^H \boldsymbol{y}$

# Example: Linear Regression




- Can you find $\boldsymbol{A}$ in the case we want to fit a higher order polynomial to the data?

# Example: High-order Regression




- Can you find $\boldsymbol{A}$ in the case we want to fit a higher order polynomial to the data?

# Example: least-square deconvolution

Assume the FIR deconvolution problem: the output of the filter is $\boldsymbol{y} \in \mathbb{C}^{n+K-1}$ and the unknown input signal is $\boldsymbol{x} \in \mathbb{C}^{n}$ and the two are related by $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{A} \in \mathbb{C}^{(n+k-1)\times n}$ is the convolution (Toeplitz) matrix associated with the known *K*-tap filter **h**.

Given noisy measurements $\boldsymbol{b} = \boldsymbol{y} + \boldsymbol{n}$ we want to find the $\boldsymbol{x}$ that is closest to the correct one and when there is no noise we want the solution to be correct.

$\boldsymbol{A}$ is tall and we can use the previous result, thus, $\boldsymbol{A}^{\boldsymbol{H}}\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{A}^{H}\boldsymbol{b}$

# Example: least-square deconvolution

***A*** is tall and we can use the previous result, thus, $\boldsymbol{A^H A\hat{x} = A^H b}$

***x*** →

***y*=Ax** →

***b*=*y*+*n*** →

$\boldsymbol{x_{LS}}$ →

$\boldsymbol{x_{LS}}$ →

# Least-Square Solution with Tikhonov Regularization

- In some cases you need to "regularize" the reconstruction process. This is an active research area. The most traditional regularization is known as "Tikhonov" regularization and tries to constrain the norm of $\boldsymbol{x}$:

$$\min_{\boldsymbol{x}} \left(\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \lambda\|\boldsymbol{x}\|^{\boldsymbol{2}}\right)$$

- This is equivalent to the modified least-square problem:

$$\min_{\boldsymbol{x}} \left(\left\|\widetilde{\boldsymbol{A}}\boldsymbol{x} - \widetilde{\boldsymbol{b}}\right\|^2\right) \text{ with } \widetilde{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A} \\ \sqrt{\lambda}\boldsymbol{I} \end{bmatrix} \text{ and } \widetilde{\boldsymbol{b}} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{bmatrix}$$

- This regularization essentially improves the condition number of $\boldsymbol{A}^H\boldsymbol{A}$.

Imperial College London

# Least-Square Solution with Tikhonov Regularization

Original Image

Blurred Image with noise

LS with Tikhonov

# Binary Classification Using Least-Squares

- Assume you have a set of "feature" vectors representing your training data: $\boldsymbol{v_i}$ $i = 1,2,..N$. These vectors belong to two different classes which you label with "1" or "-1". So we associate $y_i = \pm 1$ to each $\boldsymbol{v}_i$.
- We want to find a vector $\boldsymbol{x}$ such that $\langle \boldsymbol{x}, \boldsymbol{v}_i \rangle > 0$ if $y_i = 1$ and $\langle \boldsymbol{x}, \boldsymbol{v}_i \rangle < 0$ if $y_i = -1$
- We create a matrix $\boldsymbol{A}$ where each row is one feature vector and then solve:
- $\min_{\boldsymbol{x}} \left( \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 \right)$ where $\boldsymbol{y}$ is the vector with entries $y_i$.
- Once you have learned $\boldsymbol{x}$, then given a new "test" feature vector you expect the inner product to help classify this new vector.
- This is a simple classifier, but it is very simple to implement and is fast.

# Solution to Ax=b (full row rank case)

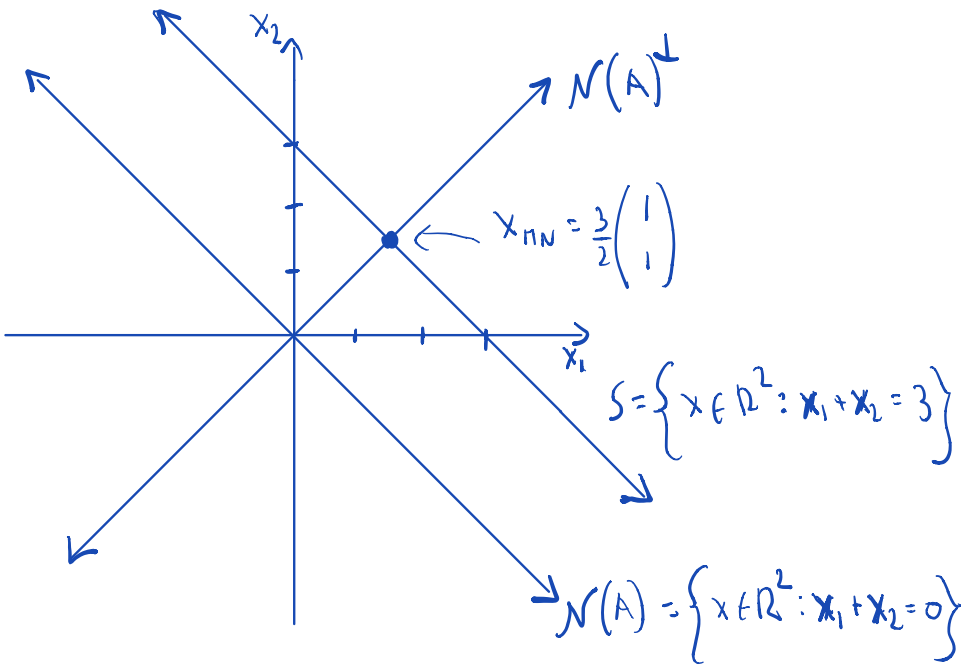
- We now understand how to find an approximate solution $\boldsymbol{Ax} = \boldsymbol{b}$ in the full column rank case where a solution may not exist.
- We now consider the full row rank case ($\boldsymbol{A}$ is 'fat') and we know that at least one solution exists.
- When the solution exists, but is not unique, the set of all solutions has a special structure.
- Let $\boldsymbol{x}_p$ be one solution to $\boldsymbol{Ax} = \boldsymbol{b}$**.** Then any other solution is $\boldsymbol{x} = \boldsymbol{x}_p + \boldsymbol{z}$ with $\boldsymbol{z} \in \mathcal{N}(\boldsymbol{A})$
- Then the solution to $\boldsymbol{Ax} = \boldsymbol{b}$ is: $S = \{\boldsymbol{x}_p\} + \mathcal{N}(\boldsymbol{A})$
- Since $\mathcal{N}(\boldsymbol{A})$ is a subspace $S$ is a *translated subspace* also called **affine subspace**

# Solution to Ax=b (full row rank case)

- We now consider the full row rank case ($\boldsymbol{A}$ is 'fat') and we know that at least one solution exists.
- We want to pick one of the (infinitely) many solutions
- We look for the ***minimum norm solution:***
  - minimize $\|\boldsymbol{x}\|$ subject to $\boldsymbol{Ax} = \boldsymbol{b}$

# Minimum Norm Solution - Example

- Let $\boldsymbol{A} = [1,1]$ and suppose $\boldsymbol{b}$=3.
- The set of all possible (real) solutions to $\boldsymbol{Ax}$=$\boldsymbol{b}$ is
- $T = \{\boldsymbol{x}\colon\ x_1 + x_2 = 3\}$
- So $\boldsymbol{x} = \begin{pmatrix} \alpha \\ 3-\alpha \end{pmatrix}$ and $\|x\|^2 = \alpha^2 + (3-\alpha)^2$
- By taking the derivative of the norm and setting it to zero we find the minimum norm solution which is: $\boldsymbol{x} = \begin{pmatrix} 3/2 \\ 3/2 \end{pmatrix}$



# Minimum Norm Solution

**Dual Projection Theorem** (without proof): Let $T = \{\boldsymbol{x}_p\} + \mathcal{S}$ be an affine subspace of $\mathbb{C}^n$. Then the element $\boldsymbol{t}_{MN}$ of T with minimum norm exists, is unique, and satisfies $\boldsymbol{t}_{MN} \in T \cap \mathcal{S}^{\perp}$ . Moreover, $\boldsymbol{t_{MN}} = P_{\mathcal{S}^{\perp}}\boldsymbol{t}$.

**Minimum Norm Theorem**: Let ***Ax***=***b*** and let ***A*** be a full-row rank matrix. Amongst all the $\boldsymbol{x} \in \mathbb{C}^n$ satisfying ***Ax***=***b***, there exists a unique solution with minimum norm. It lies in $\Re(\boldsymbol{A}^H)$ and is given by $\boldsymbol{x}_{MN} = \boldsymbol{A}_{MN}\boldsymbol{b}$ with $\boldsymbol{A}_{MN} = \boldsymbol{A}^{\boldsymbol{H}}(\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{H}})^{-1}$.

Proof: If $\boldsymbol{x_p}$ is a solution to ***Ax***=***b*** then all the solutions are in the affine subspace $T = \{\boldsymbol{x}_p\} + \mathcal{N}(\boldsymbol{A})$. Using the Dual Projection Theorem we conclude that

$$\boldsymbol{x}_{MN} \in T \cap \mathcal{N}(\boldsymbol{A})^{\perp} = T \cap \Re(\boldsymbol{A}^H)$$

Since $x_{MN} \in \Re(\boldsymbol{A}^H) \Rightarrow \boldsymbol{x}_{MN} = \boldsymbol{A}^H\boldsymbol{z} \Rightarrow \boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}_{MN} = \left(\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{H}}\right)\boldsymbol{z}$.

Therefore, $\boldsymbol{z} = (\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{H}})^{-1}\boldsymbol{b}$ which implies that $\boldsymbol{x}_{MN} = \boldsymbol{A}^{\boldsymbol{H}}(\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{H}})^{-1}\boldsymbol{b}$

# Minimum Norm Solution: worked example

**Question:** find the minimum norm solution to $\boldsymbol{Ax}=\boldsymbol{b}$ with

$$\boldsymbol{A}=\begin{pmatrix}1 & 0 & 1\\ 0 & 1 & 0\end{pmatrix} \text{ and } \boldsymbol{b}=[1,2]^T$$

Approach 1 (recommended): Compute directly $\boldsymbol{x}_{MN}=\boldsymbol{A}^{\boldsymbol{H}}(\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{H}})^{-1}\boldsymbol{b}$ which is in this case

$$\boldsymbol{x}_{MN}=\begin{bmatrix}1 & 0\\ 0 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}2 & 0\\ 0 & 1\end{bmatrix}^{-1}\begin{bmatrix}1\\ 2\end{bmatrix}=\begin{bmatrix}1/2 & 0\\ 0 & 1\\ 1/2 & 0\end{bmatrix}\begin{bmatrix}1\\ 2\end{bmatrix}=\begin{bmatrix}1/2\\ 2\\ 1/2\end{bmatrix}$$

Approach 2: the constraints are $x_1+x_3=1; x_2=2$, therefore $\boldsymbol{x}_p=[\alpha,2,1-\alpha]^T$

$\|\boldsymbol{x}_p\|^2=\alpha^2+4+(1-\alpha)^2$ which implies $\alpha=\frac{1}{2}$ and $\boldsymbol{x}_{MN}=\left[\frac{1}{2},2,\frac{1}{2}\right]^T$

# Minimum Norm Solution: worked example

**Question:** find the minimum norm solution to $\boldsymbol{Ax}=\boldsymbol{b}$ with

$$\boldsymbol{A}=\begin{pmatrix}1 & 0 & 1\\ 0 & 1 & 0\end{pmatrix} \text{ and } \boldsymbol{b}=[1,2]^T$$

Approach 3: $\mathcal{N}(\boldsymbol{A})$ =span{ $[1,0,-1]^T$} since the minimum norm solution $\boldsymbol{x}\in\mathcal{N}(\boldsymbol{A})^{\perp}$ we need that if $\boldsymbol{n}\in\mathcal{N}(\boldsymbol{A})$ then $\boldsymbol{n}^T\boldsymbol{x}=0$, we know that $x_1+x_3=1; x_2=2$, therefore $\boldsymbol{n}^T\boldsymbol{x}=0$ implies $x_1-x_3=0$ and $\boldsymbol{x}_{MN}=\left[\frac{1}{2},2,\frac{1}{2}\right]^T$

Approach 4: Given a possible solution $\boldsymbol{x}_p$ project it onto $\Re(\boldsymbol{A}^H)$ so in our case

$$\boldsymbol{x}_{MN}=\boldsymbol{A}^H\left(\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{H}}\right)^{-1}\boldsymbol{A}\boldsymbol{x}_p=\frac{1}{2}\begin{bmatrix}1 & 0 & 1\\ 0 & 2 & 0\\ 1 & 0 & 1\end{bmatrix}\begin{bmatrix}1\\ 2\\ 0\end{bmatrix}=\begin{bmatrix}1/2\\ 2\\ 1/2\end{bmatrix}$$

# Computing left and right inverses

Claim: Assume $\boldsymbol{A}$ is 'tall' and full column rank then $\boldsymbol{A}$ has a left inverse $\boldsymbol{A_{left}}$

Sketch of proof: $\boldsymbol{A^T A}$ is full rank so $(\boldsymbol{A^T A})^{-1}$ exists. Then $\boldsymbol{A_{left}} = \left(\boldsymbol{A^T A}\right)^{-1}\boldsymbol{A^T}$

Likewise: Assume $\boldsymbol{A}$ is 'fat' and full row rank then $\boldsymbol{A}$ has a right inverse $\boldsymbol{A_{rig}}$ and

$$\boldsymbol{A_{rig}} = \boldsymbol{A^T}\left(\boldsymbol{A A^T}\right)^{-1}$$

# Least-Square problems with linear constraints

- Let's revisit the minimum norm solution approach:
  - Remember that we are back to the full row rank case (***A*** is 'fat') and we know that at least one solution exists.
  - Since there are (infinitely) many solutions
  - We look for a specific one:

  - minimize $\|\boldsymbol{x}\|^2$ subject to $\boldsymbol{Ax} = \boldsymbol{b}$

- The idea now is to interpret the above equation differently: as a least square minimization with some (linear constraints)

- A more general form is: minimize $\|\boldsymbol{y} - \boldsymbol{Cx}\|^2$ subject to $\boldsymbol{Ax} = \boldsymbol{b}$

# Least-Square problems with linear constraints

- Problem: minimize $\|\boldsymbol{y} - \boldsymbol{Cx}\|^2$ subject to $\boldsymbol{Ax} = \boldsymbol{b}$
  - We use Lagrangian multipliers to solve the problem:
  - Minimize $L(\boldsymbol{x}, \boldsymbol{\lambda}) = \|\boldsymbol{y} - \boldsymbol{Cx}\|^2 + \lambda_1(\boldsymbol{a}_1^T\boldsymbol{x} - b_1) + \lambda_2(\boldsymbol{a}_2^T\boldsymbol{x} - b_2)$+...+ $\lambda_m(\boldsymbol{a}_m^T\boldsymbol{x} - b_m)$, where $\boldsymbol{a}_i^T$ are the rows of $\boldsymbol{A}$
  - Compute $\frac{\partial L(x,\lambda)}{\partial x_i} = 0 \Rightarrow \quad 2\boldsymbol{C}^T\boldsymbol{Cx} - 2\boldsymbol{C}^T\boldsymbol{y} + \boldsymbol{A}^T\boldsymbol{\lambda}$=**0**
  - Combining this set of linear equations with the feasibility conditions $\boldsymbol{Ax} = \boldsymbol{b}$ , we can write the optimality conditions as one set of m+$n$ linear equations in the variables ($\boldsymbol{x}$; $\boldsymbol{\lambda}$ ):

$$\begin{pmatrix} 2\boldsymbol{C}^T\boldsymbol{C} & \boldsymbol{A}^T \\ \boldsymbol{A} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} 2\boldsymbol{C}^T\boldsymbol{y} \\ \boldsymbol{b} \end{pmatrix}$$

# Least-Square problems with linear constraints

- Note that the square matrix $\begin{pmatrix} 2\boldsymbol{C}^{\boldsymbol{T}}\boldsymbol{C} & \boldsymbol{A}^{T} \\ \boldsymbol{A} & \boldsymbol{0} \end{pmatrix}$ is invertible if and only if $\boldsymbol{A}$ is full row rank and $\begin{pmatrix} \boldsymbol{C} \\ \boldsymbol{A} \end{pmatrix}$ is full column rank

# Minimum Norm Least Square Solution to Ax=b

- We have so far dealt with the full column and full row rank cases
- When $\boldsymbol{A}$ is neither full column or row rank, a solution to $\boldsymbol{Ax} = \boldsymbol{b}$ may not exist and if it does it is not unique
- Let $\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ and $\boldsymbol{b} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$, we note $\boldsymbol{b} \notin \Re(\boldsymbol{A})$, furthermore $\mathcal{N}(\boldsymbol{A}) = span \begin{pmatrix} 2 \\ -1 \end{pmatrix}$
- We may try to find $\boldsymbol{x}_{LS}$ by solving $\boldsymbol{A^H A x_{LS}} = \boldsymbol{A}^H \boldsymbol{b}$ which yields

$$5 \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \boldsymbol{x}_{LS} = 8 \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

- There are an infinity of solutions all in the affine subspace $T = \begin{pmatrix} 1 \\ 3/10 \end{pmatrix} + span \begin{pmatrix} 2 \\ -1 \end{pmatrix}$
- We may then pick the one with minimum norm

# Minimum Norm Least Square Solution to Ax=b

- It turns out that the minimum norm least square solution is unique and is given by the Moore-Penrose pseudoinverse of $\boldsymbol{A}$ which we denote with $\boldsymbol{A}^{+}$

- Given $\boldsymbol{A}$ its Moore-Penrose pseudoinverse $\boldsymbol{A}^{+}$ is defined as the matrix satisfying:

1. $\boldsymbol{A}\boldsymbol{A}^{+}\boldsymbol{A} = \boldsymbol{A}$
2. $\boldsymbol{A}^{+}\boldsymbol{A}\boldsymbol{A}^{+} = \boldsymbol{A}^{+}$
3. $(\boldsymbol{A}\boldsymbol{A}^{+})^{\boldsymbol{H}} = \boldsymbol{A}\boldsymbol{A}^{+}$
4. $(\boldsymbol{A}^{+}\boldsymbol{A})^{\boldsymbol{H}} = \boldsymbol{A}^{+}\boldsymbol{A}$

- The pseudo inverse is computed using the **Singular Value Decomposition** (SVD)