模和内积II

Norms and Inner Products

Baobin Li

Email:libb@ucas.ac.cn

School of Computer and Control Engineering, UCAS

Orthogonal Reduction

- A matrix $\mathbf{A}$ can be reduced to row echelon form by elementary row operation by Gaussian elimination.
- Gaussian elimination is not the only way to reduce a matrix.
- Elementary reflector $\mathbf{R}_k$ can accomplish the same purpose, which called **Householder reduction**. It proceeds as follows.
- For $\mathbf{A}_{m\times n} = [\mathbf{A}_{*1}|\mathbf{A}_{*2}|\cdots|\mathbf{A}_{*n}]$, use $\mathbf{x} = \mathbf{A}_{*1}$ to construct the elementary reflector

$$\mathbf{R}_1 = \mathbf{I} - 2\frac{\mathbf{u}\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}} \quad where \quad \mathbf{u} = \mathbf{A}_{*1} \pm \mu\|\mathbf{A}_{*1}\|\mathbf{e}_1,$$

- So that $\mathbf{R}_1\mathbf{A}_{*1} = \mp\mu\|\mathbf{A}_{*1}\|\mathbf{e}_1 = (t_{11},\ 0, \cdots,\ 0)^T$.
- Applying $\mathbf{R}_1$ to $\mathbf{A}$ yields

$$\mathbf{R}_1\mathbf{A} = [\mathbf{R}_1\mathbf{A}_{*1}|\mathbf{R}_1\mathbf{A}_{*2}|\cdots|\mathbf{R}_1\mathbf{A}_{*n}] = \begin{pmatrix} t_{11} & \mathbf{t}_1^T \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix},$$

where $\mathbf{A}_2$ is $m-1\times n-1$.

- Thus all entries below the (1,1)-position are annihilated.
- Now apply the same procedure to $\mathbf{A}_2$ to construct an elementary reflector $\hat{\mathbf{R}}_2$ that annihilates all entries below the (1,1)-position in $\mathbf{A}_2$.
- Set $\mathbf{R}_2 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_2 \end{pmatrix}$, then $\mathbf{R}_2\mathbf{R}_1$ is an orthogonal matrix such that

$$\mathbf{R}_2\mathbf{R}_1\mathbf{A} = \begin{pmatrix} t_{11} & \mathbf{t}_1^T \\ \mathbf{0} & \hat{\mathbf{R}}_2\mathbf{A}_2 \end{pmatrix}$$

- The result after $k-1$ steps is $\mathbf{R}_{k-1}\cdots\mathbf{R}_2\mathbf{R}_1\mathbf{A} = \begin{pmatrix} \mathbf{T}_{k-1} & \widetilde{\mathbf{T}}_{k-1} \\ \mathbf{0} & \mathbf{A}_k \end{pmatrix}$.
- Eventually, all of the rows or all of the columns will be exhausted, so the final result is one of the two following upper-trapezoidal forms:

$$\mathbf{R}_n\cdots\mathbf{R}_2\mathbf{R}_1\mathbf{A}_{m\times n} = \left.\begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & * \\ \hline 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}\right\}{\scriptstyle n\times n} \qquad \text{when} \quad m > n,$$

$$\mathbf{R}_{m-1}\cdots\mathbf{R}_2\mathbf{R}_1\mathbf{A}_{m\times n} = \left(\begin{array}{cccc|ccc} * & * & \cdots & * & * & \cdots & * \\ 0 & * & \cdots & * & * & \cdots & * \\ \vdots & & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & * & * & \cdots & * \end{array}\right) \quad \text{when} \quad m < n.$$

$m\times m$

- If $m = n$, then the final form is an upper-triangular matrix.
- The elementary reflectors $\mathbf{R}_i$ described above are unitary matrices, so every product $\mathbf{R}_k\cdots\mathbf{R}_1$ is a unitary matrix.

Orthogonal Reduction

- For every $\mathbf{A}\in\mathcal{C}^{m\times n}$, there exists a unitary matrix $\mathbf{P}$ such that

$$\mathbf{PA} = \mathbf{T}$$

has an upper-trapezoidal form. When $\mathbf{P}$ is constructed as a product of elementary reflectors as described above, the process is called ***Householder reduction.***

- If $\mathbf{A}$ is square, then $\mathbf{T}$ is upper triangular, and if $\mathbf{A}$ is real, then the $\mathbf{P}$ can be taken to be an orthogonal matrix.

- **Problem:** Use Householder reduction to find an orthogonal matrix $\mathbf{P}$ such that $\mathbf{PA} = \mathbf{T}$ is upper triangular with positive diagonal entries, where

$$\mathbf{A} = \begin{pmatrix} 0 & -20 & -14 \\ 3 & 27 & -4 \\ 4 & 11 & -2 \end{pmatrix}.$$

- **Solution:** To annihilate the entries below the (1,1)-position and to guarantee that t_{11} is positive, we set

$$\mathbf{u}_1 = \mathbf{A}_{*1} - \|\mathbf{A}_{*1}\|\mathbf{e}_1 = \mathbf{A}_{*1} - 5\mathbf{e}_1 = (-5\ 3\ 4)^T \quad \textit{and} \quad \mathbf{R}_1 = \mathbf{I} - 2\frac{\mathbf{u}_1\mathbf{u}^T}{\mathbf{u}_1^T\mathbf{u}_1}.$$

We obtain

$$\mathbf{R}_1\mathbf{A} = [\mathbf{R}_1\mathbf{A}_{*1} \,|\, \mathbf{R}_1\mathbf{A}_{*2} \,|\, \mathbf{R}_1\mathbf{A}_{*3}] = \left(\begin{array}{c|cc} 5 & 25 & -4 \\ \hline 0 & 0 & -10 \\ 0 & -25 & -10 \end{array}\right).$$

To annihilate the entry below the (2,2)-position, set

$$\mathbf{A}_2 = \begin{pmatrix} 0 & -10 \\ -25 & -10 \end{pmatrix} \quad and \quad \mathbf{u}_2 = [\mathbf{A}_2]_{*1} - \|[\mathbf{A}_2]_{*1}\|\mathbf{e}_1 = 25\begin{pmatrix} -1 \\ -1 \end{pmatrix}.$$

If $\hat{\mathbf{R}}_2 = \mathbf{I} - 2\frac{\mathbf{u}_2\mathbf{u}_2^T}{\mathbf{u}_2^T\mathbf{u}_2}$ and $\mathbf{R}_2 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_2 \end{pmatrix}$ then

$$\hat{\mathbf{R}}_2\mathbf{A}_2 = \begin{pmatrix} 25 & 10 \\ 0 & 10 \end{pmatrix} \quad and \quad \mathbf{R}_2\mathbf{R}_1\mathbf{A} = \begin{pmatrix} 5 & 25 & -4 \\ 0 & 25 & 10 \\ 0 & 0 & 10 \end{pmatrix}$$

and

$$\mathbf{P} = \mathbf{R}_2\mathbf{R}_1 = \frac{1}{25}\begin{pmatrix} 0 & 15 & 20 \\ -20 & 12 & -9 \\ -15 & -16 & 12 \end{pmatrix}.$$

- Elementary reflectors are not the only type of orthogonal matrices that can be used to reduce a matrix to an upper-trapezoidal form.
- Plane rotation matrices are also orthogonal and can be used to selectively annihilate any component in a given column.

- A sequence of plane rotations can be used to annihilate all elements below a particular pivot.
- This means that a matrix $\mathbf{A}$ can be reduced to an upper-trapezoidal form by using plane rotations.
- Such a process is usually called a **Givens reduction**
- Householder and Givens reductions are closely related to the results produced by applying the Gram – Schmidt process to the columns of $\mathbf{A}$.
- When $\mathbf{A}$ is nonsingular, Householder, Givens, and Gram – Schmidt each produce an orthogonal matrix $\mathbf{Q}$ and an upper-triangular matrix $\mathbf{R}$ such that $\mathbf{A} = \mathbf{QR}$.

QR Factorization

For each nonsingular $\mathbf{A} \in \Re^{n \times n}$, there is a unique orthogonal matrix $\mathbf{Q}$ and a unique upper-triangular matrix $\mathbf{R}$ with positive diagonal entries such that

$$\mathbf{A} = \mathbf{QR}.$$

This "square" QR factorization is a special case of the more general "rectangular" QR factorization

Orthogonal Reduction and Least Squares

- Orthogonal reduction can be used to solve the least squares problem with an inconsistent system $\mathbf{Ax} = \mathbf{b}$ in which $\mathbf{A} \in \Re^{m \times n}$ and $m \geq n$.
- If ε denotes the difference $\varepsilon = \mathbf{Ax} - \mathbf{b}$, then, the general least square problem is to find a vector $\mathbf{x}$ that minimizes the quantity

$$\sum_{i=1}^{m} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = \|\boldsymbol{\varepsilon}\|^2 ,$$

where $\|\star\|$ is the standard euclidean vector norm. Suppose that $\mathbf{A}$ is reduced to an upper-trapezoidal matrix $\mathbf{T}$ by an orthogonal matrix $\mathbf{P}$, and write

$$\mathbf{PA} = \mathbf{T} = \begin{pmatrix} \mathbf{R}_{n \times n} \\ \mathbf{0} \end{pmatrix} \quad \text{and} \quad \mathbf{Pb} = \begin{pmatrix} \mathbf{c}_{n \times 1} \\ \mathbf{d} \end{pmatrix}$$

in which $\mathbf{R}$ is an upper-triangular matrix. An orthogonal matrix is an isometry——so that

$$\|\boldsymbol{\varepsilon}\|^2 = \|\mathbf{P}\boldsymbol{\varepsilon}\|^2 = \|\mathbf{P}(\mathbf{Ax} - \mathbf{b})\|^2 = \left\| \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix} \mathbf{x} - \begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix} \right\|^2 = \left\| \begin{pmatrix} \mathbf{Rx} - \mathbf{c} \\ \mathbf{d} \end{pmatrix} \right\|^2$$
$$= \|\mathbf{Rx} - \mathbf{c}\|^2 + \|\mathbf{d}\|^2 .$$

Consequently, $\|\boldsymbol{\varepsilon}\|^2$ is minimized when $\mathbf{x}$ is a vector such that $\|\mathbf{Rx} - \mathbf{c}\|^2$ is minimal or, in other words, $\mathbf{x}$ is a least squares solution for $\mathbf{Ax} = \mathbf{b}$ if and only if $\mathbf{x}$ is a least squares solution for $\mathbf{Rx} = \mathbf{c}$.

- If $\mathbf{A}$ has linearly independent columns, then the least squares solution for $\mathbf{Ax} = \mathbf{b}$ is obtained by solving the nonsingular triangular system $\mathbf{Rx} = \mathbf{c}$.
- We now have four different ways to reduce a matrix to an upper-triangular form
  (1) Gaussian elimination
  (2) Gram-Schmidt procedure
  (3) Householder reduction
  (4) Givens reduction
- It's natural to try to compare them and to sort out the advantages and disadvantages of each. First consider numerical stability.
- Strictly speaking, an algorithm is considered to be **numerically stable** if, under floating-point arithmetic, it always returns an answer that is the exact solution of a nearby problem.
- The Householder or Givens reduction is a stable algorithm for producing the $\mathbf{QR}$ factorization of $\mathbf{A}_{n\times n}$.
  - Suppose that floating-point arithmetic produces an orthogonal matrix $\mathbf{Q} + \mathbf{E}$ and upper triangular matrix $\mathbf{R} + \mathbf{F}$

$$\tilde{\mathbf{A}} = (\mathbf{Q}+\mathbf{E})(\mathbf{R}+\mathbf{F}) = \mathbf{QR}+\mathbf{QF}+\mathbf{ER}+\mathbf{EF} = \mathbf{A}+\mathbf{QF}+\mathbf{ER}+\mathbf{EF}.$$

- If $\mathbf{E}$ and $\mathbf{F}$ account for the roundoff errors, and if their entries are small relative to those in $\mathbf{A}$, then the entries in $\mathbf{EF}$ are negligible, and

$$\tilde{\mathbf{A}} \approx \mathbf{A} + \mathbf{QF} + \mathbf{ER}.$$

- Since $\mathbf{Q}$ is orthogonal, $\|\mathbf{QF}\|_F = \|\mathbf{F}\|_F$, and $\|\mathbf{A}\|_F = \|\mathbf{QR}\|_F = \|\mathbf{R}\|_F$.
- This means that neither $\mathbf{QF}$ nor $\mathbf{ER}$ can contain entries that are large relative to those in $\mathbf{A}$.
- Hence $\tilde{\mathbf{A}} \approx \mathbf{A}$, which says that the algorithm is stable.

- Gaussian elimination is not a stable algorithm.
  - Consider the LU factorization of $\mathbf{A} = \mathbf{LU}$.
  - Suppose that floating-point Gaussian elimination with no pivoting returns matrices $\mathbf{L} + \mathbf{E}$ and $\mathbf{U} + \mathbf{F}$.

$$\tilde{\mathbf{A}} = (\mathbf{L} + \mathbf{E})(\mathbf{U} + \mathbf{F}) = \mathbf{LU} + \mathbf{LF} + \mathbf{EU} + \mathbf{EF} = \mathbf{A} + \mathbf{LF} + \mathbf{EU} + \mathbf{EF}.$$

  - If $\mathbf{E}$ and $\mathbf{F}$ account for the roundoff errors, and if their entries are small relative to those in $\mathbf{A}$, the the entries in $\mathbf{EF}$ are negligible,

$$\tilde{\mathbf{A}} \approx \mathbf{A} + \mathbf{LF} + \mathbf{EU}.$$

- If $\mathbf{L}$ or $\mathbf{U}$ contains entries that are large relative to those in $\mathbf{A}$, then $\mathbf{LF}$ or $\mathbf{EU}$ can contain entries that are significant.
- If partial pivoting is employed, then no multiplier can exceed 1. $\mathbf{L}$ can not greatly magnify the entries of $\mathbf{F}$, so $\tilde{\mathbf{A}} \approx \mathbf{A} + \mathbf{EU}$.
- Numerical stability rests on the magnitude of the entries in $\mathbf{U}$. Unfortunately, partial pivoting may not be enough to control the growth of all entries in $\mathbf{U}$.
- In general, it has been proven that if complete pivoting is used on a matrix $\mathbf{A}_{n\times n}$ for which $max|a_{ij}| = 1$, then no entry of $\mathbf{U}$ can exceed

$$\gamma = n^{1/2}(2^1 3^{1/2} \cdots n^{1/n-1})^{1/2}.$$

- Gaussian elimination with complete pivoting is stable, but Gaussian elimination with partial pivoting is not.

- Algorithms based on the Gram-Schmidt procedure are more complicated.
  - As an algorithm to return the QR factorization of $\mathbf{A}$, the modified Gram-Schmidt procedure has been proven to be unstable.
  - But as an algorithm used to solve the least squares problem, it is stable.

Summary of Numerical Stability

- Gaussian elimination with scaled partial pivoting is theoretically unstable, but it is "practically stable"—i.e., stable for most practical problems.
- Complete pivoting makes Gaussian elimination unconditionally stable.
- For the QR factorization, the Gram–Schmidt procedure (classical or modified) is not stable. However, the modified Gram–Schmidt procedure is a stable algorithm for solving the least squares problem.
- Householder and Givens reductions are unconditionally stable algorithms for computing the QR factorization.

Summary of Computational Effort

The approximate number of multiplications/divisions required to reduce an $n \times n$ matrix to an upper-triangular form is as follows.

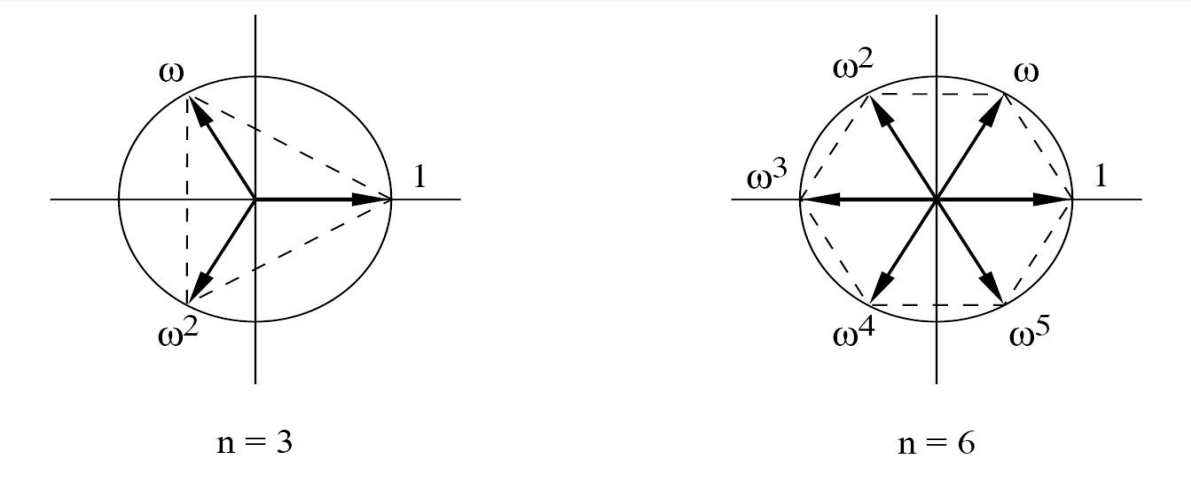
- Gaussian elimination (scaled partial pivoting) $\approx n^3/3$.
- Gram–Schmidt procedure (classical and modified) $\approx n^3$.
- Householder reduction $\approx 2n^3/3$.
- Givens reduction $\approx 4n^3/3$.

Discrete Fourier Transform

- For a positive integer n, the complex numbers $\{1, \omega, \omega^2, \cdots, \omega^{n-1}\}$, where

$$\omega = e^{2\pi i/n} = \cos\frac{2\pi}{n} + i\sin\frac{2\pi}{n}$$

  are called the n^{th} roots of unity.

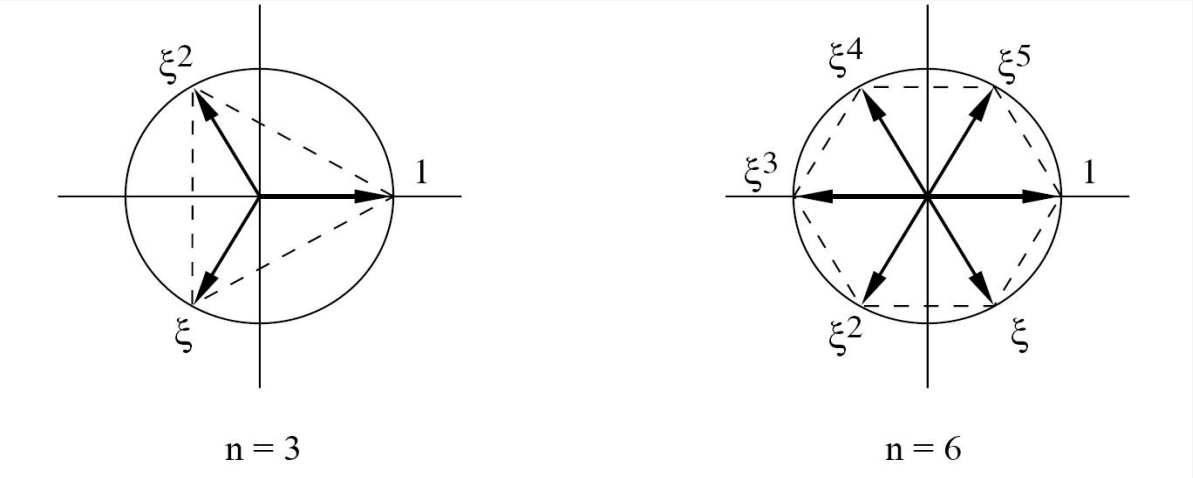


- They represent all solution to $z^n = 1$.
- Geometrically, they are the vertices of a regular polygon of n sides.
- The roots of unity are cyclic: $\omega^k = \omega^{k(mod\ n)}$

- The numbers $\{1, \xi, \xi^2, \cdots \xi^{n-1}\}$, where

$$\xi = e^{-2\pi i/n} = \cos\frac{2\pi}{n} - i\sin\frac{2\pi}{n} = \bar{\omega}$$

are also the n^{th} roots of unity.



- If k is an integer, $\xi^{-k} = \omega^k$, and $1 + \xi^k + \xi^{2k} + \cdots + \xi^{(n-1)k} = 0$.
- The Fourier matrix is a special case of the Vandermonde matrix.
- the columns in $\mathbf{F}_n$ are mutually orthogonal, and each column has norm $\sqrt{n}$. This means that $(1/\sqrt{n})\mathbf{F}_n$ is a unitary matrix.

Fourier Matrix

The $n \times n$ matrix whose (j,k)-entry is $\xi^{jk} = \omega^{-jk}$ for $0 \le j,k \le n-1$ is called the ***Fourier matrix*** of order n, and it has the form

$$\mathbf{F}_n = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \xi & \xi^2 & \cdots & \xi^{n-1} \\ 1 & \xi^2 & \xi^4 & \cdots & \xi^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \xi^{n-1} & \xi^{n-2} & \cdots & \xi \end{pmatrix}_{n\times n}.$$

Note. Throughout this section entries are indexed from 0 to $n-1$. For example, the upper left-hand entry of $\mathbf{F}_n$ is considered to be in the $(0,0)$ position (rather than the $(1,1)$ position), and the lower right-hand entry is in the $(n-1,n-1)$ position. When the context makes it clear, the subscript n on $\mathbf{F}_n$ is omitted.

- $\mathbf{F}_n^{-1} = \frac{1}{n}\bar{\mathbf{F}}_n = \frac{1}{n}\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n-1} & \omega^{n-2} & \cdots & \omega \end{pmatrix}_{n\times n}$

- For example

The Fourier matrices of orders 2 and 4 are given by

$$\mathbf{F}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{F}_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \end{pmatrix},$$

and their inverses are

$$\mathbf{F}_2^{-1} = \frac{1}{2}\overline{\mathbf{F}}_2 = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{F}_4^{-1} = \frac{1}{4}\overline{\mathbf{F}}_4 = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \end{pmatrix}.$$

Discrete Fourier Transform

Given a vector $\mathbf{x}_{n\times 1}$, the product $\mathbf{F}_n\mathbf{x}$ is called the ***discrete Fourier transform*** of $\mathbf{x}$, and $\mathbf{F}_n^{-1}\mathbf{x}$ is called the ***inverse transform*** of $\mathbf{x}$. The k^{th} entries in $\mathbf{F}_n\mathbf{x}$ and $\mathbf{F}_n^{-1}\mathbf{x}$ are given by

$$[\mathbf{F}_n\mathbf{x}]_k = \sum_{j=0}^{n-1} x_j \xi^{jk} \quad \text{and} \quad [\mathbf{F}_n^{-1}\mathbf{x}]_k = \frac{1}{n}\sum_{j=0}^{n-1} x_j \omega^{jk}.$$

Problem: Computing the Inverse Transform. Explain why any algorithm or program designed to compute the discrete Fourier transform of a vector $\mathbf{x}$ can also be used to compute the *inverse* transform of $\mathbf{x}$.

Solution: Call such an algorithm FFT . The fact that

$$\mathbf{F}_n^{-1}\mathbf{x} = \frac{\overline{\mathbf{F}}_n\mathbf{x}}{n} = \frac{\overline{\mathbf{F}_n\overline{\mathbf{x}}}}{n}$$

means that FFT will return the inverse transform of $\mathbf{x}$ by executing the following three steps:

(1) $\mathbf{x} \longleftarrow \overline{\mathbf{x}}$ (compute $\overline{\mathbf{x}}$).

(2) $\mathbf{x} \longleftarrow \mathrm{FFT}(\mathbf{x})$ (compute $\mathbf{F}_n\overline{\mathbf{x}}$).

(3) $\mathbf{x} \longleftarrow (1/n)\overline{\mathbf{x}}$ (compute $n^{-1}\overline{\mathbf{F}_n\overline{\mathbf{x}}} = \mathbf{F}_n^{-1}\mathbf{x}$).

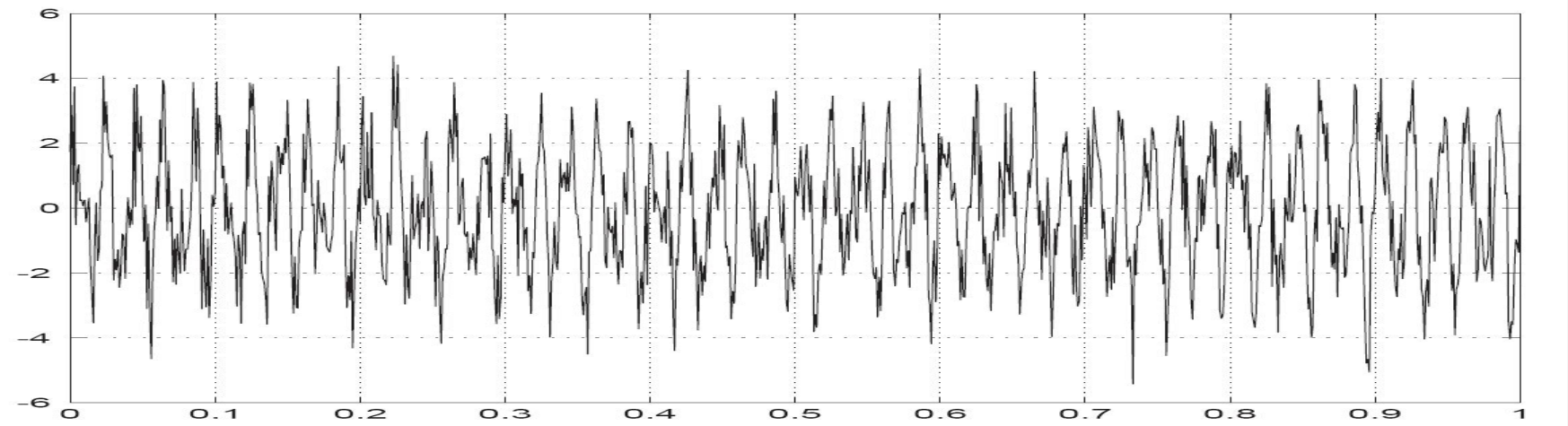
For example, computing the inverse transform of $\mathbf{x} = (\,\mathrm{i} \quad 0 \quad -\mathrm{i} \quad 0\,)^T$ is accomplished as follows—recall that $\mathbf{F}_4$

$$\overline{\mathbf{x}} = \begin{pmatrix} -\mathrm{i} \\ 0 \\ \mathrm{i} \\ 0 \end{pmatrix}, \quad \mathbf{F}_4\overline{\mathbf{x}} = \begin{pmatrix} 0 \\ -2\mathrm{i} \\ 0 \\ -2\mathrm{i} \end{pmatrix}, \quad \frac{1}{4}\overline{\mathbf{F}_4\overline{\mathbf{x}}} = \frac{1}{4}\begin{pmatrix} 0 \\ 2\mathrm{i} \\ 0 \\ 2\mathrm{i} \end{pmatrix} = \mathbf{F}_4^{-1}\mathbf{x}.$$

You may wish to check that this answer agrees with the result obtained by directly multiplying $\mathbf{F}_4^{-1}$ times $\mathbf{x}$.

Signal Processing

- Suppose that a microphone is placed under a hovering helicopter. The following is the sound signal recorded during 1 second of time.



- It seems reasonable to expect that the signal should have oscillatory components together with some random noise contamination. That is

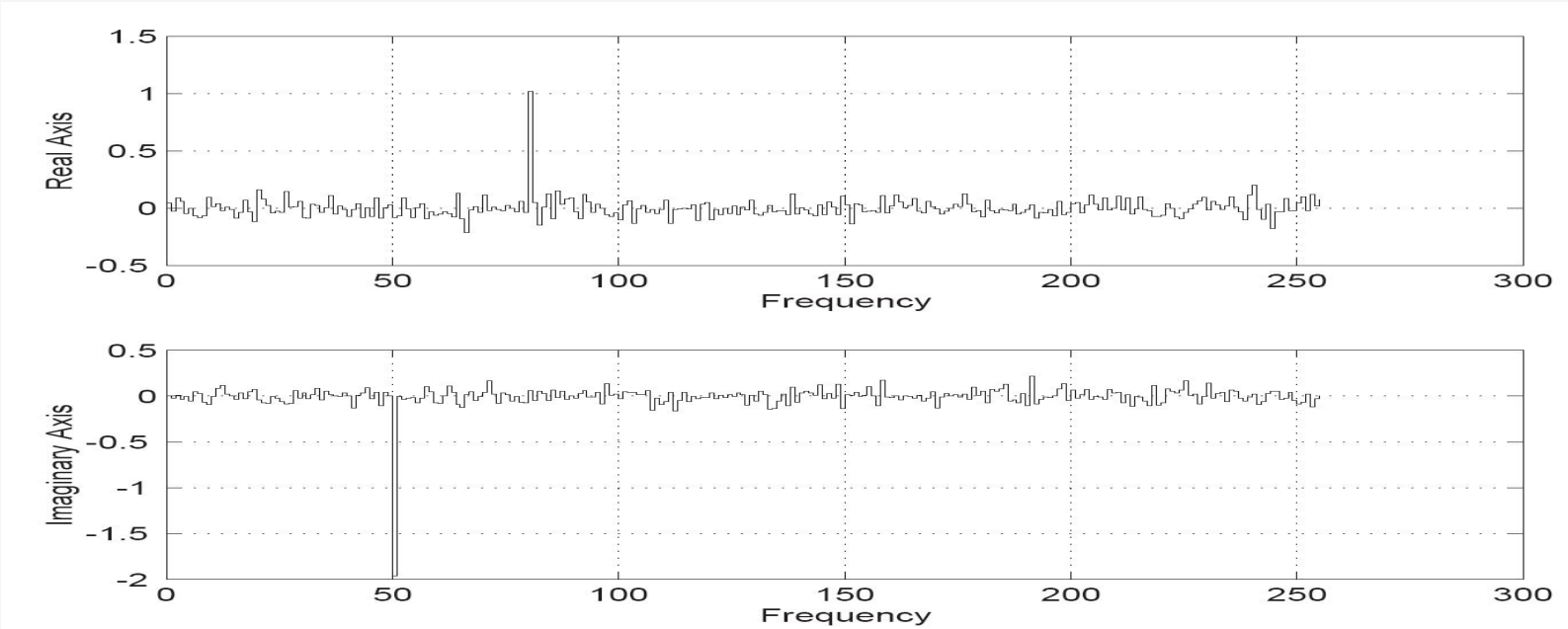
$$y(\tau) = \left(\sum_k \alpha_k \cos 2\pi f_k \tau + \beta_k \sin 2\pi f_k \tau \right) + Noise.$$

- But due to the noise contamination, the oscillatory nature of the signal is only barely apparent.

- To reveal the oscillatory components, the magic of the Fourier transform is employed.
- Let $\mathbf{x}$ be the vector obtained by sampling the signal at n equally spaced points between time $\tau = 0$ and $\tau = 1 (n = 512$ in this case).
- Let $\mathbf{y} = (2/n)\mathbf{F}_n\mathbf{x} = \mathbf{a} + i\mathbf{b}$, where

$$\mathbf{a} = (2/n)Re(\mathbf{F}_n\mathbf{x}), \quad \mathbf{b} = (2/n)Im(\mathbf{F}_n\mathbf{x}).$$

- Using only the first 256 entries in $\mathbf{a}$ and $i\mathbf{b}$, plot the points



- Now there are some obvious characteristics-the plot of $\mathbf{a}$ has a spike of height approximately 1 at entry 80, and the plot of $i\mathbf{b}$ in the bottom graph has a spike of height approximately -2 at entry 50.
- These two spikes indicate that the signal is made up primarily of two oscillatory components:
- The spike in the real vector $\mathbf{a}$ indicates that one of the oscillatory components is a cosine of frequency 80 Hz (or period $= 1/80$) whose amplitude is approximately 1.
- The spike in the imaginary vector $i\mathbf{b}$ indicates there is a sine component with frequency 50 Hz and amplitude of about 2.
- In other words, the Fourier transform indicates the signal is

$$y(\tau) = (\cos 2\pi(80\tau) + 2\sin 2\pi(50\tau)) + Noise.$$

If

$$\mathbf{a} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{pmatrix}_{n\times 1} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{n-1} \end{pmatrix}_{n\times 1},$$

then the vector

$$\mathbf{a} \odot \mathbf{b} = \begin{pmatrix} \alpha_0\beta_0 \\ \alpha_0\beta_1 + \alpha_1\beta_0 \\ \alpha_0\beta_2 + \alpha_1\beta_1 + \alpha_2\beta_0 \\ \vdots \\ \alpha_{n-2}\beta_{n-1} + \alpha_{n-1}\beta_{n-2} \\ \alpha_{n-1}\beta_{n-1} \\ 0 \end{pmatrix}_{2n\times 1}$$

is called the **convolution** of $\mathbf{a}$ and $\mathbf{b}$.

- The 0 in the last position is for convenience only-it makes the size of the convolution twice the size of the original vectors.

- Furthermore, it is sometimes convenient to pad $\mathbf{a}$ and $\mathbf{b}$ with n additional zeros to consider them to be vectors with $2n$ components.
- Setting $\alpha_n = \cdots = \alpha_{2n-1} = \beta_n = \cdots = \beta_{2n-1} = 0$ allows us to write

$$[\mathbf{a} \odot \mathbf{b}]_K = \sum_{j=0}^{k} \alpha_j \beta_{k-j} \quad for \quad k = 0, 1, \cdots, 2n-1.$$

- A visual way to form $\mathbf{a} \odot \mathbf{b}$ is to "slide" the reversal of $\mathbf{b}$ "against" $\mathbf{a}$.

$$\begin{array}{ccccccc}
\beta_{n-1} & & \beta_{n-1} & & \beta_{n-1} & & \alpha_0 \\
\vdots & & \vdots & & \vdots & & \vdots \\
\beta_1 & & \alpha_0 \times \beta_1 & & \alpha_0 \times \beta_2 & & \alpha_{n-2} \\
\alpha_0 \times \beta_0 & & \alpha_1 \times \beta_0 & & \alpha_1 \times \beta_1 & \cdots & \alpha_{n-1} \times \beta_{n-1} \\
\alpha_1 & & \vdots & & \alpha_2 \times \beta_0 & & \beta_{n-2} \\
\vdots & & \alpha_{n-1} & & \vdots & & \vdots \\
\alpha_{n-1} & & & & \alpha_{n-1} & & \beta_0
\end{array}$$

- The convolution operation is a natural occurrence in a variety of situations, and polynomial multiplication is one such example.

Polynomial Multiplication. For $p(x) = \sum_{k=0}^{n-1} \alpha_k x^k$, $q(x) = \sum_{k=0}^{n-1} \beta_k x^k$, let $\mathbf{a} = \begin{pmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_{n-1} \end{pmatrix}^T$ and $\mathbf{b} = \begin{pmatrix} \beta_0 & \beta_1 & \cdots & \beta_{n-1} \end{pmatrix}^T$. The product $p(x)q(x) = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \cdots + \gamma_{2n-2} x^{2n-2}$ is a polynomial of degree $2n-2$ in which γ_k is simply the k^{th} component of the convolution $\mathbf{a} \odot \mathbf{b}$ because

$$p(x)q(x) = \sum_{k=0}^{2n-2} \left[\sum_{j=0}^{k} \alpha_j \beta_{k-j} \right] x^k = \sum_{k=0}^{2n-2} [\mathbf{a} \odot \mathbf{b}]_k x^k.$$

Convolution Theorem

Let $\mathbf{a} \times \mathbf{b}$ denote the entry-by-entry product

$$\mathbf{a} \times \mathbf{b} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{n-1} \end{pmatrix} = \begin{pmatrix} \alpha_0\beta_0 \\ \alpha_1\beta_1 \\ \vdots \\ \alpha_{n-1}\beta_{n-1} \end{pmatrix}_{n \times 1},$$

and let $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ be the padded vectors

$$\hat{\mathbf{a}} = \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_{n-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{2n \times 1} \quad \text{and} \quad \hat{\mathbf{b}} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{n-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{2n \times 1}.$$

If $\mathbf{F} = \mathbf{F}_{2n}$ is the Fourier matrix of order $2n$, then

$$\mathbf{F}(\mathbf{a} \odot \mathbf{b}) = (\mathbf{F}\hat{\mathbf{a}}) \times (\mathbf{F}\hat{\mathbf{b}}) \quad \text{and} \quad \mathbf{a} \odot \mathbf{b} = \mathbf{F}^{-1}\left[(\mathbf{F}\hat{\mathbf{a}}) \times (\mathbf{F}\hat{\mathbf{b}})\right].$$

- According to the convolution theorem, the convolution of two $n \times 1$ vectors can be computed by executing three discrete Fourier transforms of order $2n$.
- But it is still not clear that much has been accomplished.
- Using matrix – vector multiplication to perform the computations on the right-hand side of requires at least 12 times the number of scalar multiplications demanded by the definition of convolution.
- It is necessary to be able to perform a discrete Fourier transform in far fewer scalar multiplications than that required by standard matrix – vector multiplication.
- Two Americans, J. W. Cooley and J. W. Tukey, introduced a fast Fourier transform (FFT) algorithm that requires only on the order of $(n/2)log_2 n$ scalar multiplications to compute $\mathbf{F}_n\mathbf{x}$.
- The magic of the fast Fourier transform algorithm emanates from the fact that if n is a power of 2, then a discrete Fourier transform of order n can be executed by performing two transforms of order $n/2$.

The magic of the fast Fourier transform algorithm emanates from the fact that if n is a power of 2, then a discrete Fourier transform of order n can be executed by performing two transforms of order $n/2$. To appreciate exactly how this comes about, observe that when $n = 2^r$ we have $\left(\xi^j\right)^n = \left(\xi^{2j}\right)^{n/2}$, so

$$\left\{1, \xi, \xi^2, \xi^3, \ldots, \xi^{n-1}\right\} = \text{ the } n^{th} \text{ roots of unity}$$

$$\text{if and only if}$$

$$\left\{1, \xi^2, \xi^4, \xi^6, \ldots, \xi^{n-2}\right\} = \text{ the } (n/2)^{th} \text{ roots of unity.}$$

This means that the (j,k)-entries in the Fourier matrices $\mathbf{F}_n$ and $\mathbf{F}_{n/2}$ are

$$[\mathbf{F}_n]_{jk} = \xi^{jk} \quad \text{and} \quad [\mathbf{F}_{n/2}]_{jk} = (\xi^2)^{jk} = \xi^{2jk}.$$

If the columns of $\mathbf{F}_n$ are permuted so that columns with even subscripts are listed before those with odd subscripts, and if $\mathbf{P}_n^T$ is the corresponding permutation matrix, then we can partition $\mathbf{F}_n\mathbf{P}_n^T$ as

$$\mathbf{F}_n\mathbf{P}_n^T = [\mathbf{F}_{*0}\,\mathbf{F}_{*2}\,\cdots\,\mathbf{F}_{*n-2}\,|\,\mathbf{F}_{*1}\,\mathbf{F}_{*3}\,\cdots\,\mathbf{F}_{*n-1}] = \begin{pmatrix} \mathbf{A}_{\frac{n}{2}\times\frac{n}{2}} & \mathbf{B}_{\frac{n}{2}\times\frac{n}{2}} \\ \mathbf{C}_{\frac{n}{2}\times\frac{n}{2}} & \mathbf{G}_{\frac{n}{2}\times\frac{n}{2}} \end{pmatrix}.$$

we see that the entries in $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{G}$ are

$$\mathbf{A}_{jk} = \mathbf{F}_{j,2k} = \xi^{2jk} = [\mathbf{F}_{n/2}]_{jk},$$

$$\mathbf{B}_{jk} = \mathbf{F}_{j,2k+1} = \xi^{j(2k+1)} = \xi^j \xi^{2jk} = \xi^j [\mathbf{F}_{n/2}]_{jk},$$

$$\mathbf{C}_{jk} = \mathbf{F}_{\frac{n}{2}+j,\,2k} = \xi^{(\frac{n}{2}+j)2k} = \xi^{nk}\xi^{2jk} = \xi^{2jk} = [\mathbf{F}_{n/2}]_{jk},$$

$$\mathbf{G}_{jk} = \mathbf{F}_{\frac{n}{2}+j,\,2k+1} = \xi^{(\frac{n}{2}+j)(2k+1)} = \xi^{nk}\xi^{n/2}\xi^j\xi^{2jk} = -\xi^j\xi^{2jt} = -\xi^j[\mathbf{F}_{n/2}]_{jk}.$$

In other words, if $\mathbf{D}_{n/2}$ is the diagonal matrix

$$\mathbf{D}_{n/2} = \begin{pmatrix} 1 & & & & \\ & \xi & & & \\ & & \xi^2 & & \\ & & & \ddots & \\ & & & & \xi^{\frac{n}{2}-1} \end{pmatrix},$$

then

$$\mathbf{F}_n \mathbf{P}_n^T = \begin{pmatrix} \mathbf{A}_{(n/2)\times(n/2)} & \mathbf{B}_{(n/2)\times(n/2)} \\ \mathbf{C}_{(n/2)\times(n/2)} & \mathbf{G}_{(n/2)\times(n/2)} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_{n/2} & \mathbf{D}_{n/2}\mathbf{F}_{n/2} \\ \mathbf{F}_{n/2} & -\mathbf{D}_{n/2}\mathbf{F}_{n/2} \end{pmatrix}.$$

Decomposing the Fourier Matrix

If $n = 2^r$, then

$$\mathbf{F}_n = \begin{pmatrix} \mathbf{F}_{n/2} & \mathbf{D}_{n/2}\mathbf{F}_{n/2} \\ \mathbf{F}_{n/2} & -\mathbf{D}_{n/2}\mathbf{F}_{n/2} \end{pmatrix} \mathbf{P}_n,$$

where

$$\mathbf{D}_{n/2} = \begin{pmatrix} 1 & & & & \\ & \xi & & & \\ & & \xi^2 & & \\ & & & \ddots & \\ & & & & \xi^{\frac{n}{2}-1} \end{pmatrix}$$

contains half of the n^{th} roots of unity and $\mathbf{P}_n$ is the "even–odd" permutation matrix defined by

$$\mathbf{P}_n^T = [\mathbf{e}_0\ \mathbf{e}_2\ \mathbf{e}_4\ \cdots\ \mathbf{e}_{n-2} \mid \mathbf{e}_1\ \mathbf{e}_3\ \mathbf{e}_5\ \cdots\ \mathbf{e}_{n-1}].$$

- The decomposition says that a discrete Fourier transform of order $n = 2^r$ can be accomplished by two Fourier transforms of order $n/2 = 2^{r-1}$, and this leads to the FFT algorithm.
- To get a feel for how the FFT works, consider the case when $n = 8$, and proceed to "divide and conquer."

$$\mathbf{x}_8 = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix}, \quad \text{then} \quad \mathbf{P}_8\mathbf{x}_8 = \begin{pmatrix} x_0 \\ x_2 \\ x_4 \\ x_6 \\ \hline x_1 \\ x_3 \\ x_5 \\ x_7 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_4^{(0)} \\ \hline \mathbf{x}_4^{(1)} \end{pmatrix},$$

so

$$\mathbf{F}_8\mathbf{x}_8 = \begin{pmatrix} \mathbf{F}_4 & \mathbf{D}_4\mathbf{F}_4 \\ \mathbf{F}_4 & -\mathbf{D}_4\mathbf{F}_4 \end{pmatrix} \begin{pmatrix} \mathbf{x}_4^{(0)} \\ \mathbf{x}_4^{(1)} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_4\mathbf{x}_4^{(0)} + \mathbf{D}_4\mathbf{F}_4\mathbf{x}_4^{(1)} \\ \mathbf{F}_4\mathbf{x}_4^{(0)} - \mathbf{D}_4\mathbf{F}_4\mathbf{x}_4^{(1)} \end{pmatrix}.$$

But

$$\mathbf{P}_4\mathbf{x}_4^{(0)} = \begin{pmatrix} x_0 \\ x_4 \\ \hline x_2 \\ x_6 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_2^{(0)} \\ \hline \mathbf{x}_2^{(1)} \end{pmatrix} \quad \text{and} \quad \mathbf{P}_4\mathbf{x}_4^{(1)} = \begin{pmatrix} x_1 \\ x_5 \\ \hline x_3 \\ x_7 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_2^{(2)} \\ \hline \mathbf{x}_2^{(3)} \end{pmatrix},$$

so

$$\mathbf{F}_4\mathbf{x}_4^{(0)} = \begin{pmatrix} \mathbf{F}_2 & \mathbf{D}_2\mathbf{F}_2 \\ \mathbf{F}_2 & -\mathbf{D}_2\mathbf{F}_2 \end{pmatrix} \begin{pmatrix} \mathbf{x}_2^{(0)} \\ \mathbf{x}_2^{(1)} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_2\mathbf{x}_2^{(0)} + \mathbf{D}_2\mathbf{F}_2\mathbf{x}_2^{(1)} \\ \mathbf{F}_2\mathbf{x}_2^{(0)} - \mathbf{D}_2\mathbf{F}_2\mathbf{x}_2^{(1)} \end{pmatrix}$$

and

$$\mathbf{F}_4\mathbf{x}_4^{(1)} = \begin{pmatrix} \mathbf{F}_2 & \mathbf{D}_2\mathbf{F}_2 \\ \mathbf{F}_2 & -\mathbf{D}_2\mathbf{F}_2 \end{pmatrix} \begin{pmatrix} \mathbf{x}_2^{(2)} \\ \mathbf{x}_2^{(3)} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_2\mathbf{x}_2^{(2)} + \mathbf{D}_2\mathbf{F}_2\mathbf{x}_2^{(3)} \\ \mathbf{F}_2\mathbf{x}_2^{(2)} - \mathbf{D}_2\mathbf{F}_2\mathbf{x}_2^{(3)} \end{pmatrix}.$$

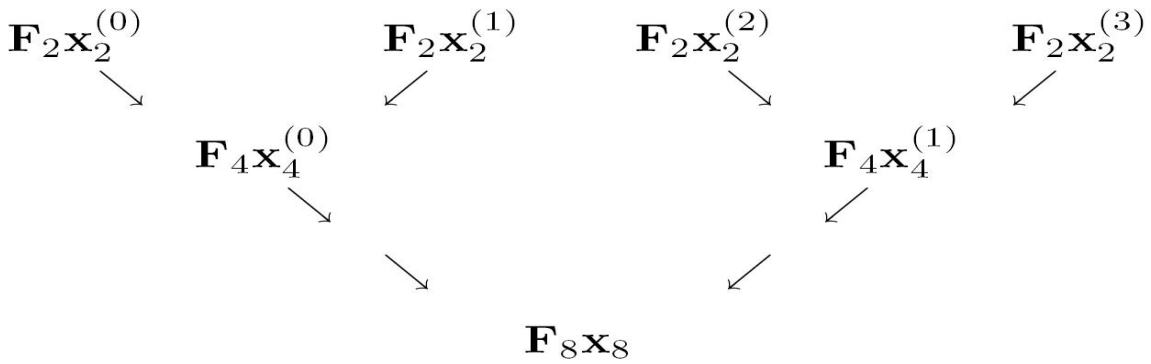
Now, since $\mathbf{F}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, it is a trivial matter to compute the terms

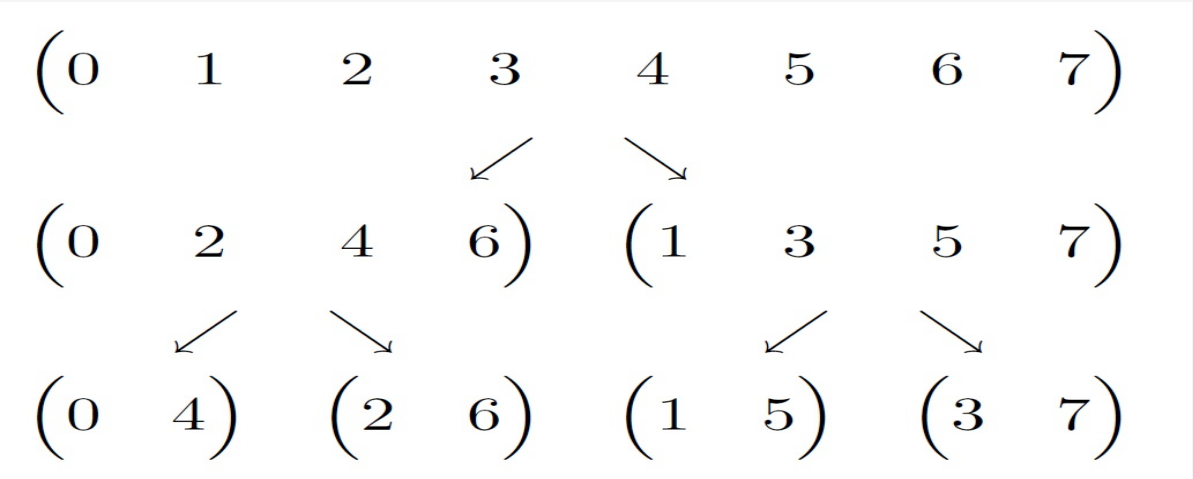
$$\mathbf{F}_2\mathbf{x}_2^{(0)}, \quad \mathbf{F}_2\mathbf{x}_2^{(1)}, \quad \mathbf{F}_2\mathbf{x}_2^{(2)}, \quad \mathbf{F}_2\mathbf{x}_2^{(3)}.$$

Of course, to actually carry out the computation, we need to work backward through the preceding sequence of steps. That is, we start with

$$\tilde{\mathbf{x}}_8 = \begin{pmatrix} \mathbf{x}_2^{(0)} \\ \hline \mathbf{x}_2^{(1)} \\ \hline \mathbf{x}_2^{(2)} \\ \hline \mathbf{x}_2^{(3)} \end{pmatrix} = \begin{pmatrix} x_0 \\ x_4 \\ \hline x_2 \\ x_6 \\ \hline x_1 \\ x_5 \\ \hline x_3 \\ x_7 \end{pmatrix},$$

.





Natural order	First level	Second level
$0 \leftrightarrow 000$	$0 \leftrightarrow 00\mathbf{0}$	$0 \leftrightarrow 0\mathbf{0}0$
$1 \leftrightarrow 001$	$2 \leftrightarrow 01\mathbf{0}$	$4 \leftrightarrow 1\mathbf{0}0$
$2 \leftrightarrow 010$	$4 \leftrightarrow 10\mathbf{0}$	$2 \leftrightarrow 0\mathbf{1}0$
$3 \leftrightarrow 011$	$6 \leftrightarrow 11\mathbf{0}$	$6 \leftrightarrow 1\mathbf{1}0$
$4 \leftrightarrow 100$	$1 \leftrightarrow 00\mathbf{1}$	$1 \leftrightarrow 0\mathbf{0}1$
$5 \leftrightarrow 101$	$3 \leftrightarrow 01\mathbf{1}$	$5 \leftrightarrow 1\mathbf{0}1$
$6 \leftrightarrow 110$	$5 \leftrightarrow 10\mathbf{1}$	$3 \leftrightarrow 0\mathbf{1}1$
$7 \leftrightarrow 111$	$7 \leftrightarrow 11\mathbf{1}$	$7 \leftrightarrow 1\mathbf{1}1$

Fast Fourier Transform

For a given input vector $\mathbf{x}$ containing $n = 2^r$ components, the discrete Fourier transform $\mathbf{F}_n\mathbf{x}$ is the result of successively creating the following arrays.

$$\mathbf{X}_{1\times n} \longleftarrow rev(\mathbf{x}) \qquad \text{(bit reverse the subscripts)}$$

For $j = 0, 1, 2, 3, \ldots, r-1$

$$\mathbf{D} \longleftarrow \begin{pmatrix} 1 \\ e^{-\pi i/2^j} \\ e^{-2\pi i/2^j} \\ e^{-3\pi i/2^j} \\ \vdots \\ e^{-(2^j-1)\pi i/2^j} \end{pmatrix}_{2^j\times 1} \quad \begin{matrix}\text{(Half of the } (2^{j+1})^{th} \text{ roots of 1,} \\ \text{perhaps from a lookup table)}\end{matrix}$$

$$\mathbf{X}^{(0)} \longleftarrow \begin{pmatrix} \mathbf{X}_{*0} & \mathbf{X}_{*2} & \mathbf{X}_{*4} & \cdots & \mathbf{X}_{*2^{r-j}-2} \end{pmatrix}_{2^j\times 2^{r-j-1}}$$

$$\mathbf{X}^{(1)} \longleftarrow \begin{pmatrix} \mathbf{X}_{*1} & \mathbf{X}_{*3} & \mathbf{X}_{*5} & \cdots & \mathbf{X}_{*2^{r-j}-1} \end{pmatrix}_{2^j\times 2^{r-j-1}}$$

$$\mathbf{X} \longleftarrow \begin{pmatrix} \mathbf{X}^{(0)} + \mathbf{D}\times\mathbf{X}^{(1)} \\ \mathbf{X}^{(0)} - \mathbf{D}\times\mathbf{X}^{(1)} \end{pmatrix}_{2^{j+1}\times 2^{r-j-1}} \quad \begin{pmatrix} \text{Define } \times \text{ to mean} \\ [\mathbf{D}\times\mathbf{M}]_{ij} = d_i m_{ij} \end{pmatrix}$$

Problem: Perform the FFT on $\mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}$.

Solution: Start with $\mathbf{X} \longleftarrow rev(\mathbf{x}) = (\, x_0 \quad x_2 \quad x_1 \quad x_3 \,)$.

For $j = 0$:

$\mathbf{D} \longleftarrow (1)$ (Half of the square roots of 1)

$\mathbf{X}^{(0)} \longleftarrow (\, x_0 \quad x_1 \,)$

$\mathbf{X}^{(1)} \longleftarrow (\, x_2 \quad x_3 \,)$ and $\mathbf{D} \times \mathbf{X}^{(1)} \longleftarrow (\, x_2 \quad x_3 \,)$

$$\mathbf{X} \longleftarrow \begin{pmatrix} \mathbf{X}^{(0)} + \mathbf{D} \times \mathbf{X}^{(1)} \\ \mathbf{X}^{(0)} - \mathbf{D} \times \mathbf{X}^{(1)} \end{pmatrix} = \begin{pmatrix} x_0 + x_2 & x_1 + x_3 \\ x_0 - x_2 & x_1 - x_3 \end{pmatrix}$$

For $j = 1$:

$$\mathbf{D} \longleftarrow \begin{pmatrix} 1 \\ -\mathrm{i} \end{pmatrix}$$ (Half of the 4^{th} roots of 1)

$$\mathbf{X}^{(0)} \longleftarrow \begin{pmatrix} x_0 + x_2 \\ x_0 - x_2 \end{pmatrix}$$

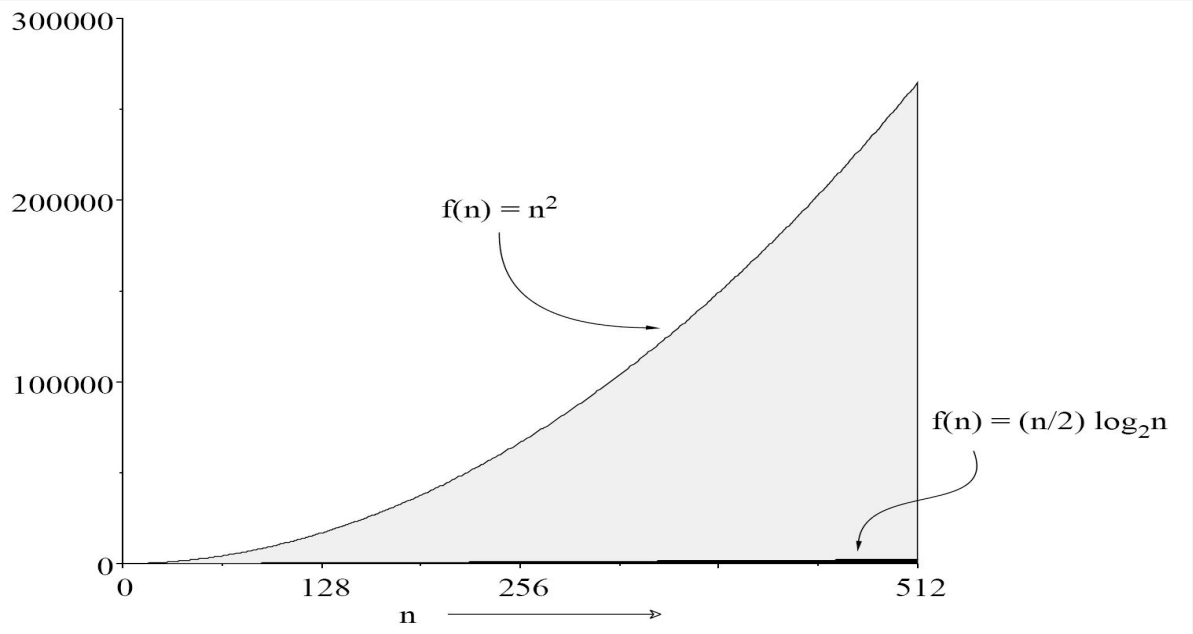
$$\mathbf{X}^{(1)} \longleftarrow \begin{pmatrix} x_1 + x_3 \\ x_1 - x_3 \end{pmatrix} \quad \text{and} \quad \mathbf{D} \times \mathbf{X}^{(1)} \longleftarrow \begin{pmatrix} x_1 + x_3 \\ -\mathrm{i}x_1 + \mathrm{i}x_3 \end{pmatrix}$$

$$\mathbf{X} \longleftarrow \begin{pmatrix} \mathbf{X}^{(0)} + \mathbf{D} \times \mathbf{X}^{(1)} \\ \mathbf{X}^{(0)} - \mathbf{D} \times \mathbf{X}^{(1)} \end{pmatrix} = \begin{pmatrix} x_0 + x_2 + x_1 + x_3 \\ x_0 - x_2 - \mathrm{i}x_1 + \mathrm{i}x_3 \\ x_0 + x_2 - x_1 - x_3 \\ x_0 - x_2 + \mathrm{i}x_1 - \mathrm{i}x_3 \end{pmatrix} = \mathbf{F}_4\mathbf{x}$$

FFT Multiplication Count

If n is a power of 2, then applying the FFT to a vector of n components requires at most $(n/2)\log_2 n$ multiplications.

- The $(n/2)log_2n$ count represents a tremendous advantage over the n^2 factor demanded by a direct matrix - vector product.
- To appreciate the magnitude of the difference between n^2 and $(n/2)log_2n$, look at the following figure



- For $n = 512$, $n^2 = 262144$ and $(n/2)log_2n = 2304$.

Complementary Subspaces

Complementary Subspaces

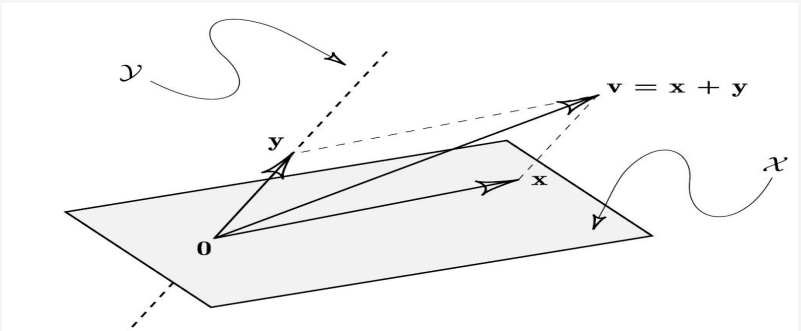
Subspaces $\mathcal{X}$, $\mathcal{Y}$ of a space $\mathcal{V}$ are said to be ***complementary*** whenever

$$\mathcal{V} = \mathcal{X} + \mathcal{Y} \quad \text{and} \quad \mathcal{X} \cap \mathcal{Y} = \mathbf{0},$$

in which case $\mathcal{V}$ is said to be the ***direct sum*** of $\mathcal{X}$ and $\mathcal{Y}$, and this is denoted by writing $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$.

- For a vector space $\mathcal{V}$ with subspaces $\mathcal{X}, \mathcal{Y}$ having respective bases $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{Y}}$, the following statements are equivalent.
  - $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$.
  - For each $\mathbf{v} \in \mathcal{V}$ there are *unique* vectors $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ such that $\mathbf{v} = \mathbf{x} + \mathbf{y}$.
  - $\mathcal{B}_{\mathcal{X}} \cap \mathcal{B}_{\mathcal{Y}} = \phi$ and $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is a basis for $\mathcal{V}$.

- For example, consider the two subspaces of $\Re^3$, $\mathcal{X}$ is a plane through the origin, and $\mathcal{Y}$ is a line through the origin.



Projection

Suppose that $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$ so that for each $\mathbf{v} \in \mathcal{V}$ there are unique vectors $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ such that $\mathbf{v} = \mathbf{x} + \mathbf{y}$.

- The vector $\mathbf{x}$ is called the ***projection*** of $\mathbf{v}$ onto $\mathcal{X}$ along $\mathcal{Y}$.
- The vector $\mathbf{y}$ is called the projection of $\mathbf{v}$ onto $\mathcal{Y}$ along $\mathcal{X}$.

- It's clear that if $\mathcal{X} \perp \mathcal{Y}$, then the notion of projection agrees with the concept of orthogonal projection.
- The phrase oblique projection is sometimes used to emphasize the fact that $\mathcal{X}$ and $\mathcal{Y}$ are not orthogonal subspaces.
- Given a pair of complementary subspaces $\mathcal{X}$ and $\mathcal{Y}$ of $\Re^n$ and an arbitrary vector $\mathbf{v} \in \Re^n$, how can the projection of $\mathbf{v}$ onto $\mathcal{X}$ be computed?
- One way is to build a projector that is a matrix $\mathbf{P}_{n \times n}$ with the property that for each $\mathbf{v} \in \Re^n$, the product $\mathbf{Pv}$ is the projection of $\mathbf{v}$ onto $\mathcal{X}$ along $\mathcal{Y}$.

Projectors

Let $\mathcal{X}$ and $\mathcal{Y}$ be complementary subspaces of a vector space $\mathcal{V}$ so that each $\mathbf{v} \in \mathcal{V}$ can be uniquely resolved as $\mathbf{v} = \mathbf{x} + \mathbf{y}$, where $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$. The unique linear operator $\mathbf{P}$ defined by $\mathbf{Pv} = \mathbf{x}$ is called the ***projector onto*** $\mathcal{X}$ ***along*** $\mathcal{Y}$, and $\mathbf{P}$ has the following properties.

- $\mathbf{P}^2 = \mathbf{P}$ ($\mathbf{P}$ is idempotent).
- $\mathbf{I} - \mathbf{P}$ is the complementary projector onto $\mathcal{Y}$ along $\mathcal{X}$.
- $R(\mathbf{P}) = \{\mathbf{x} \mid \mathbf{Px} = \mathbf{x}\}$ (the set of "fixed points" for $\mathbf{P}$).
- $R(\mathbf{P}) = N(\mathbf{I} - \mathbf{P}) = \mathcal{X}$ and $R(\mathbf{I} - \mathbf{P}) = N(\mathbf{P}) = \mathcal{Y}$.
- If $\mathcal{V} = \Re^n$ or $\mathcal{C}^n$, then $\mathbf{P}$ is given by

$$\mathbf{P} = [\mathbf{X} \mid \mathbf{0}][\mathbf{X} \mid \mathbf{Y}]^{-1} = [\mathbf{X} \mid \mathbf{Y}] \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} [\mathbf{X} \mid \mathbf{Y}]^{-1},$$

where the columns of $\mathbf{X}$ and $\mathbf{Y}$ are respective bases for $\mathcal{X}$ and $\mathcal{Y}$.

- It is easy to get $\mathbf{P}$ is similar to the diagonal matrix, which must be the matrix representation of the linear operator that when restricted to $\mathcal{X}$ is the identity operator and when restricted to $\mathcal{Y}$ is the zero operator.
- Statement says that $\mathbf{P}$ is a projector, then $\mathbf{P}$ is idempotent.
- Is every idempotent linear operator necessarily a projector?

Projectors and Idempotents

A linear operator $\mathbf{P}$ on $\mathcal{V}$ is a projector if and only if $\mathbf{P}^2 = \mathbf{P}$.

Proof. The fact that every projector is idempotent was proven . The proof of the converse rests on the fact that

$$\mathbf{P}^2 = \mathbf{P} \implies R(\mathbf{P}) \text{ and } N(\mathbf{P}) \text{ are complementary subspaces.}$$

To prove this, observe that $\mathcal{V} = R(\mathbf{P}) + N(\mathbf{P})$ because for each $\mathbf{v} \in \mathcal{V}$,

$$\mathbf{v} = \mathbf{P}\mathbf{v} + (\mathbf{I} - \mathbf{P})\mathbf{v}, \quad \text{where} \quad \mathbf{P}\mathbf{v} \in R(\mathbf{P}) \text{ and } (\mathbf{I} - \mathbf{P})\mathbf{v} \in N(\mathbf{P}).$$

Furthermore, $R(\mathbf{P}) \cap N(\mathbf{P}) = \mathbf{0}$ because

$$\mathbf{x} \in R(\mathbf{P}) \cap N(\mathbf{P}) \implies \mathbf{x} = \mathbf{P}\mathbf{y} \text{ and } \mathbf{P}\mathbf{x} = \mathbf{0} \implies \mathbf{x} = \mathbf{P}\mathbf{y} = \mathbf{P}^2\mathbf{y} = \mathbf{0},$$

Problem: Let $\mathcal{X}$ and $\mathcal{Y}$ be the subspaces of $\Re^3$ that are spanned by

$$\mathcal{B}_{\mathcal{X}} = \left\{ \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix} \right\} \quad \text{and} \quad \mathcal{B}_{\mathcal{Y}} = \left\{ \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \right\},$$

respectively. Explain why $\mathcal{X}$ and $\mathcal{Y}$ are complementary, and then determine the projector onto $\mathcal{X}$ along $\mathcal{Y}$. What is the projection of $\mathbf{v} = \begin{pmatrix} -2 & 1 & 3 \end{pmatrix}^T$ onto $\mathcal{X}$ along $\mathcal{Y}$? What is the projection of $\mathbf{v}$ onto $\mathcal{Y}$ along $\mathcal{X}$?

Solution: $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{Y}}$ are linearly independent, so they are bases for $\mathcal{X}$ and $\mathcal{Y}$, respectively. The spaces $\mathcal{X}$ and $\mathcal{Y}$ are complementary because

$$rank\,[\mathbf{X} \,|\, \mathbf{Y}] = rank \left(\begin{array}{cc|c} 1 & 0 & 1 \\ -1 & 1 & -1 \\ -1 & -2 & 0 \end{array} \right) = 3$$

insures that $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is a basis for $\Re^3$. The projector onto $\mathcal{X}$ along $\mathcal{Y}$ is

$$\mathbf{P} = [\mathbf{X} \,|\, \mathbf{0}]\,[\mathbf{X} \,|\, \mathbf{Y}]^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & -2 & 0 \end{pmatrix} \begin{pmatrix} -2 & -2 & -1 \\ 1 & 1 & 0 \\ 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} -2 & -2 & -1 \\ 3 & 3 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Range-Nullspace Decomposition

- Since there are infinitely many different pairs of complementary subspaces in $\Re^n$ (or $\mathcal{C}^n$), is some pair more "natural" than the rest?
- If we start with a given matrix $\mathbf{A}_{n\times n}$, then there is a very natural direct sum decomposition of $\Re^n$ defined by fundamental subspaces associated with powers of $\mathbf{A}$.
- The rank plus nullity theorem says that $dimR(\mathbf{A}) + dimN(\mathbf{A}) = n$, so it's reasonable to ask about the possibility of $R(\mathbf{A})$ and $N(\mathbf{A})$ being complementary subspaces.
- If $\mathbf{A}$ is nonsingular, then it's trivially true that $R(\mathbf{A})$ and $N(\mathbf{A})$ are complementary.
- But when $\mathbf{A}$ is singular, this need not be the case because $R(\mathbf{A})$ and $N(\mathbf{A})$ need not be disjoint.
- For example

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \Longrightarrow \begin{pmatrix} 1 \\ 0 \end{pmatrix} \in R(\mathbf{A}) \bigcap N(\mathbf{A}).$$

Range-Nullspace Decomposition

For every singular matrix $\mathbf{A}_{n\times n}$, there exists a positive integer k such that $R\left(\mathbf{A}^k\right)$ and $N\left(\mathbf{A}^k\right)$ are complementary subspaces. That is,

$$\Re^n = R\left(\mathbf{A}^k\right) \oplus N\left(\mathbf{A}^k\right).$$

- The smallest positive integer k is called the **index** of $\mathbf{A}$. For nonsingular matrices we define $index(\mathbf{A}) = 0$.

Index

The index of a square matrix $\mathbf{A}$ is the smallest nonnegative integer k such that any one of the three following statements is true.

- $rank\left(\mathbf{A}^k\right) = rank\left(\mathbf{A}^{k+1}\right)$.
- $R\left(\mathbf{A}^k\right) = R\left(\mathbf{A}^{k+1}\right)$—i.e., the point where $R\left(\mathbf{A}^k\right)$ stops shrinking.
- $N\left(\mathbf{A}^k\right) = N\left(\mathbf{A}^{k+1}\right)$—i.e., the point where $N\left(\mathbf{A}^k\right)$ stops growing.

For nonsingular matrices, $index(\mathbf{A}) = 0$. For singular matrices, $index(\mathbf{A})$ is the smallest *positive* integer k such that either of the following two statements is true.

- $R\left(\mathbf{A}^k\right) \cap N\left(\mathbf{A}^k\right) = \mathbf{0}$.
- $\Re^n = R\left(\mathbf{A}^k\right) \oplus N\left(\mathbf{A}^k\right)$.

Problem: Determine the index of $\mathbf{A} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & -1 \end{pmatrix}$.

Solution: $\mathbf{A}$ is singular (because $rank\,(\mathbf{A}) = 2$), so $index(\mathbf{A}) > 0$. Since

$$\mathbf{A}^2 = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{A}^3 = \begin{pmatrix} 8 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

we see that $rank\,(\mathbf{A}) > rank\,(\mathbf{A}^2) = rank\,(\mathbf{A}^3)$, so $index(\mathbf{A}) = 2$. Alternately,

$$R\,(\mathbf{A}) = span\left\{ \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \right\},\ R\,(\mathbf{A}^2) = span \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix},\ R\,(\mathbf{A}^3) = span \begin{pmatrix} 8 \\ 0 \\ 0 \end{pmatrix},$$

so $R\,(\mathbf{A}) \supset R\,(\mathbf{A}^2) = R\,(\mathbf{A}^3)$ implies $index(\mathbf{A}) = 2$.

Nilpotent Matrices

- $\mathbf{N}_{n\times n}$ is said to be ***nilpotent*** whenever $\mathbf{N}^k = \mathbf{0}$ for some positive integer k.
- $k = index(\mathbf{N})$ is the smallest positive integer such that $\mathbf{N}^k = \mathbf{0}$. (Some authors refer to $index(\mathbf{N})$ as the ***index of nilpotency***.)

Problem: Verify that

$$\mathbf{N} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

is a nilpotent matrix, and determine its index.

Solution: Computing the powers

$$\mathbf{N}^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{N}^3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

reveals that $\mathbf{N}$ is indeed nilpotent, and it shows that $index(\mathbf{N}) = 3$ because $\mathbf{N}^3 = \mathbf{0}$, but $\mathbf{N}^2 \neq \mathbf{0}$.

- Anytime $\Re^n$ can be written as the direct sum of two complementary subspaces such that one of them is an invariant subspace for a given square matrix $\mathbf{A}$.
- We have a block-triangular representation for $\mathbf{A}$.
- And if both complementary spaces are invariant under $\mathbf{A}$, this block-triangular representation is actually block diagonal.

Core-Nilpotent Decomposition

If $\mathbf{A}$ is an $n \times n$ singular matrix of index k such that $rank\left(\mathbf{A}^k\right) = r$, then there exists a nonsingular matrix $\mathbf{Q}$ such that

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{pmatrix}$$

in which $\mathbf{C}$ is nonsingular, and $\mathbf{N}$ is nilpotent of index k. In other words, $\mathbf{A}$ is *similar* to a 2×2 block-diagonal matrix containing a nonsingular "core" and a nilpotent component. The block-diagonal matrix is called a ***core-nilpotent decomposition*** of $\mathbf{A}$.

Note: When $\mathbf{A}$ is nonsingular, $k = 0$ and $r = n$, so $\mathbf{N}$ is not present, and we can set $\mathbf{Q} = \mathbf{I}$ and $\mathbf{C} = \mathbf{A}$ (the nonsingular core is everything).

- **Drazin Inverse**: if

$$\mathbf{A} = \mathbf{Q}\begin{pmatrix} \mathbf{C} & 0 \\ 0 & \mathbf{N} \end{pmatrix}\mathbf{Q}^{-1}, \quad then \quad \mathbf{A}^D = \mathbf{Q}\begin{pmatrix} \mathbf{C}^{-1} & 0 \\ 0 & 0 \end{pmatrix}\mathbf{Q}^{-1}$$

defines the Drazin inverse of $\mathbf{A}$.

  - $\mathbf{A}^D$ is unique, and $\mathbf{A}^D = \mathbf{A}^{-1}$ when $\mathbf{A}$ is nonsingular.
  - $\mathbf{A}^D\mathbf{A}\mathbf{A}^D = \mathbf{A}^D$,$\mathbf{A}\mathbf{A}^D = \mathbf{A}^D\mathbf{A}$, $\mathbf{A}^{k+1}\mathbf{A}^D = \mathbf{A}^k$, where $k = index(\mathbf{A})$.

Orthogonal Decomposition

Orthogonal Complement

For a subset $\mathcal{M}$ of an inner-product space $\mathcal{V}$, the ***orthogonal complement*** $\mathcal{M}^{\perp}$ (pronounced "$\mathcal{M}$ perp") of $\mathcal{M}$ is defined to be the set of all vectors in $\mathcal{V}$ that are orthogonal to every vector in $\mathcal{M}$. That is,

$$\mathcal{M}^{\perp} = \{\mathbf{x} \in \mathcal{V} \mid \langle \mathbf{m}|\mathbf{x}\rangle = 0 \text{ for all } \mathbf{m} \in \mathcal{M}\}.$$

- For example, if $\mathcal{M} = \{x\}$ is a single vector in $\Re^2$, $\mathcal{M}^{\perp}$ is the line through the origin that is perpendicular to x.
- If $\mathcal{M}$ is a plane through the origin in $\Re^3$, then $\mathcal{M}^{\perp}$ is the line through the origin that is perpendicular to the plane.

Orthogonal Complementary Subspaces

If $\mathcal{M}$ is a subspace of a finite-dimensional inner-product space $\mathcal{V}$, then

$$\mathcal{V} = \mathcal{M} \oplus \mathcal{M}^{\perp}.$$

Furthermore, if $\mathcal{N}$ is a subspace such that $\mathcal{V} = \mathcal{M} \oplus \mathcal{N}$ and $\mathcal{N} \perp \mathcal{M}$ (every vector in $\mathcal{N}$ is orthogonal to every vector in $\mathcal{M}$), then

$$\mathcal{N} = \mathcal{M}^{\perp}.$$

Perp Operation

If $\mathcal{M}$ is a subspace of an n-dimensional inner-product space, then the following statements are true.

- $\dim \mathcal{M}^{\perp} = n - \dim \mathcal{M}$.
- $\mathcal{M}^{\perp^{\perp}} = \mathcal{M}$.

Orthogonal Decomposition Theorem

For every $\mathbf{A} \in \Re^{m \times n}$,

$$R(\mathbf{A})^{\perp} = N(\mathbf{A}^T) \quad \text{and} \quad N(\mathbf{A})^{\perp} = R(\mathbf{A}^T).$$

this means that every matrix $\mathbf{A} \in \Re^{m \times n}$ produces an orthogonal decomposition of $\Re^m$ and $\Re^n$ in the sense that

$$\Re^m = R(\mathbf{A}) \oplus R(\mathbf{A})^{\perp} = R(\mathbf{A}) \oplus N(\mathbf{A}^T),$$

and

$$\Re^n = N(\mathbf{A}) \oplus N(\mathbf{A})^{\perp} = N(\mathbf{A}) \oplus R(\mathbf{A}^T).$$

- The orthogonal decomposition theorem holds for all matrices.
- It tells us how to decompose $\Re^m$ and $\Re^n$ in terms of the four fundamental subspaces of $\mathbf{A}$.
- It also tells us how to decompose $\mathbf{A}$ itself into more basic components.

URV Factorization

For each $\mathbf{A} \in \Re^{m\times n}$ of rank r, there are orthogonal matrices $\mathbf{U}_{m\times m}$ and $\mathbf{V}_{n\times n}$ and a nonsingular matrix $\mathbf{C}_{r\times r}$ such that

$$\mathbf{A} = \mathbf{URV}^T = \mathbf{U}\begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{m\times n} \mathbf{V}^T.$$

- The first r columns in $\mathbf{U}$ are an orthonormal basis for $R(\mathbf{A})$.
- The last $m-r$ columns of $\mathbf{U}$ are an orthonormal basis for $N(\mathbf{A}^T)$.
- The first r columns in $\mathbf{V}$ are an orthonormal basis for $R(\mathbf{A}^T)$.
- The last $n-r$ columns of $\mathbf{V}$ are an orthonormal basis for $N(\mathbf{A})$.

Each different collection of orthonormal bases for the four fundamental subspaces of $\mathbf{A}$ produces a different URV factorization of $\mathbf{A}$. In the complex case, replace $(\star)^T$ by $(\star)^*$ and "orthogonal" by "unitary."

- Support that $rank(\mathbf{A}) = r$, and let

$$\mathcal{B}_{R(\mathbf{A})} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\} \quad \text{and} \quad \mathcal{B}_{N\left(\mathbf{A}^T\right)} = \{\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \ldots, \mathbf{u}_m\}$$

be orthonormal bases for $R(\mathbf{A})$ and $N(\mathbf{A}^T)$, respectively, and let

$$\mathcal{B}_{R\left(\mathbf{A}^T\right)} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\} \quad \text{and} \quad \mathcal{B}_{N(\mathbf{A})} = \{\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, \ldots, \mathbf{v}_n\}$$

be orthonormal bases for $R(\mathbf{A}^T)$ and $N(\mathbf{A})$, respectively. It follows that $\mathcal{B}_{R(\mathbf{A})} \cup \mathcal{B}_{N(\mathbf{A}^T)}$ and $\mathcal{B}_{R(\mathbf{A}^T)} \cup \mathcal{B}_{N(\mathbf{A})}$ are orthonormal bases for $\Re^m$ and $\Re^n$, respectively, and hence

$$\mathbf{U}_{m\times m} = \begin{pmatrix} \mathbf{u}_1 \,|\, \mathbf{u}_2 \,|\, \cdots \,|\, \mathbf{u}_m \end{pmatrix} \quad \text{and} \quad \mathbf{V}_{n\times n} = \begin{pmatrix} \mathbf{v}_1 \,|\, \mathbf{v}_2 \,|\, \cdots \,|\, \mathbf{v}_n \end{pmatrix}$$

are orthogonal matrices. Now consider the product $\mathbf{R} = \mathbf{U}^T\mathbf{A}\mathbf{V}$, and notice that $r_{ij} = \mathbf{u}_i^T\mathbf{A}\mathbf{v}_j$. However, $\mathbf{u}_i^T\mathbf{A} = \mathbf{0}$ for $i = r+1, \ldots, m$ and $\mathbf{A}\mathbf{v}_j = \mathbf{0}$ for $j = r+1, \ldots, n$, so

$$\mathbf{R} = \mathbf{U}^T\mathbf{A}\mathbf{V} = \begin{pmatrix} \mathbf{u}_1^T\mathbf{A}\mathbf{v}_1 & \cdots & \mathbf{u}_1^T\mathbf{A}\mathbf{v}_r & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ \mathbf{u}_r^T\mathbf{A}\mathbf{v}_1 & \cdots & \mathbf{u}_r^T\mathbf{A}\mathbf{v}_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

In other words, $\mathbf{A}$ can be factored as

$$\mathbf{A} = \mathbf{U}\mathbf{R}\mathbf{V}^T = \mathbf{U}\begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^T.$$

- The range-nullspace decomposition and orthogonal decomposition theorem produces a decomposition of $\mathbf{A}$.
- The range-nullspace decomposition decomposes $\Re^n$ with square matrices while the orthogonal decomposition theorem does it with rectangular matrices.
- So does this mean that the range-nullspace decomposition is a special case of the orthogonal decomposition theorem?
- No! Even for square matrices they are not very comparable because each says something that the other doesn't.

- The core-nilpotent decomposition is obtained by a similarity transformation,
- Orthogonal decomposition has the advantage whenever orthogonality is naturally built into a problem-such as least squares applications.
- Orthogonal methods often produce numerically stable algorithms for floating-point computation, whereas similarity transformations are generally not well suited for numerical computations.
- The value of similarity is mainly on the theoretical side of the coin.

Range Perpendicular to Nullspace

For $rank\,(\mathbf{A}_{n\times n}) = r$, the following statements are equivalent:

- $R(\mathbf{A}) \perp N(\mathbf{A})$,
- $R(\mathbf{A}) = R\left(\mathbf{A}^T\right)$,
- $N(\mathbf{A}) = N\left(\mathbf{A}^T\right)$,
- $\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T$

in which $\mathbf{U}$ is orthogonal and $\mathbf{C}$ is nonsingular. Such matrices will be called ***RPN matrices,*** short for "range perpendicular to nullspace." Some authors call them *range-symmetric* or *EP* matrices. Nonsingular matrices are trivially RPN because they have a zero nullspace. For complex matrices, replace $(\star)^T$ by $(\star)^*$ and "orthogonal" by "unitary."

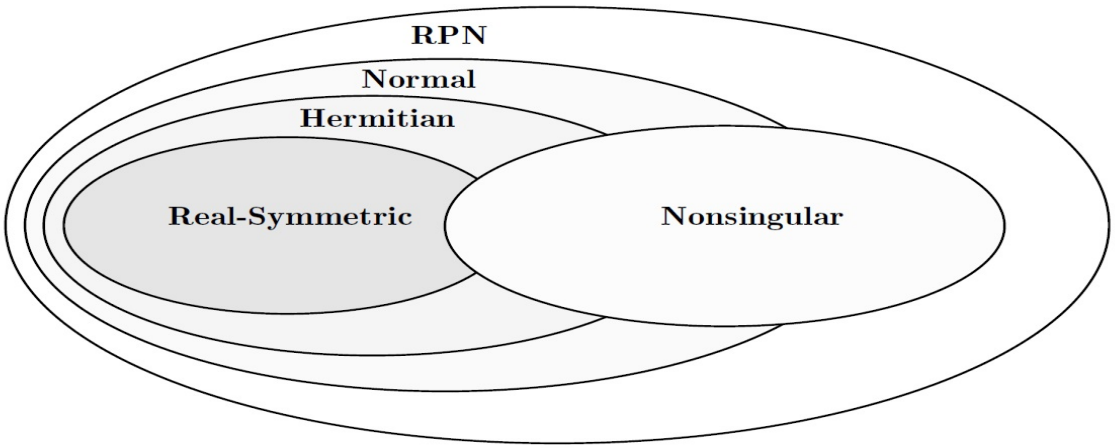
Problem: Explain how to make $\mathbf{C}$ lower triangular.

Solution: Apply Householder (or Givens) reduction to produce an orthogonal matrix $\mathbf{P}_{m\times m}$ such that $\mathbf{PA} = \begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix}$, where $\mathbf{B}$ is $r \times n$ of rank r. Householder (or Givens) reduction applied to $\mathbf{B}^T$ results in an orthogonal matrix $\mathbf{Q}_{n\times n}$ and a nonsingular upper-triangular matrix $\mathbf{T}$ such that

$$\mathbf{QB}^T = \begin{pmatrix} \mathbf{T}_{r\times r} \\ \mathbf{0} \end{pmatrix} \implies \mathbf{B} = (\mathbf{T}^T \,|\, \mathbf{0})\mathbf{Q} \implies \begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{T}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q},$$

so $\mathbf{A} = \mathbf{P}^T \begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix} = \mathbf{P}^T \begin{pmatrix} \mathbf{T}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q}$ is a URV factorization.

Note: $\mathbf{C}$ can in fact be made diagonal.



- $\mathbf{A}$ is called a **normal matrix** whenever $\mathbf{AA}^* = \mathbf{A}^*\mathbf{A}$.

Singular Value Decomposition

Singular Value Decomposition

For each $\mathbf{A} \in \Re^{m \times n}$ of rank r, there are orthogonal matrices $\mathbf{U}_{m \times m}$, $\mathbf{V}_{n \times n}$ and a diagonal matrix $\mathbf{D}_{r \times r} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$ such that

$$\mathbf{A} = \mathbf{U} \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{m \times n} \mathbf{V}^T \quad \text{with} \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$

The σ_i's are called the nonzero ***singular values*** of $\mathbf{A}$. When $r < p = \min\{m, n\}$, $\mathbf{A}$ is said to have $p - r$ additional zero singular values.

- The above factorization is called a **singular value decomposition** of $\mathbf{A}$, and the columns in $\mathbf{U}$ and $\mathbf{V}$ are called left-hand and right-hand singular vectors for $\mathbf{A}$, respectively.
- While the constructive method used to derive the SVD can be used as an algorithm, more sophisticated techniques exist.
- All good matrix computation packages contain numerically stable SVD implementations.
- The SVD is valid for complex matrices when $(\star)^T$ is replaced by $(\star)^*$.

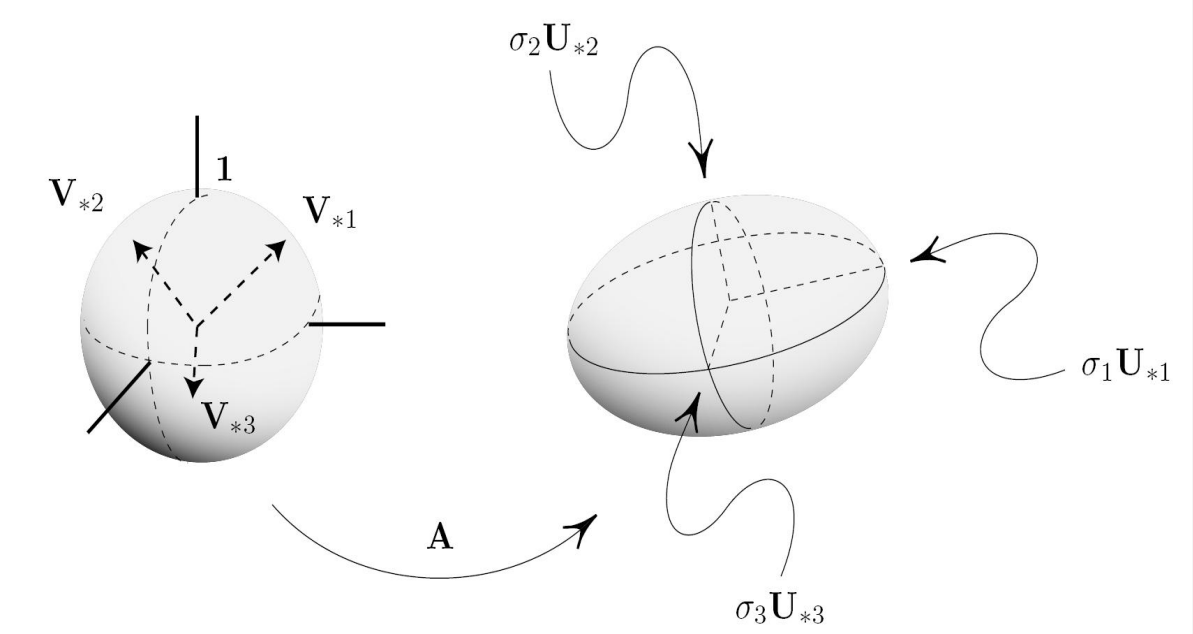
- Singular values reveal something about the geometry of linear transformations.
- Because the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ of a matrix $\mathbf{A}$ tell us how much distortion can occur under transformation by $\mathbf{A}$.
- They give us an explicit picture of how $\mathbf{A}$ distorts the unit sphere.
- Suppose $\mathbf{A} \in \Re^{n \times n}$ is nonsingular, and let $\mathcal{S}_2 = \{\mathbf{x} | \|\mathbf{x}\|_2 = 1\}$ be the unit sphere in $\Re^n$.
- The nature of the image $\mathbf{A}(\mathcal{S}_2)$ is revealed by considering the singular value decompositions

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T \quad and \quad \mathbf{A}^{-1} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T \quad with \quad \mathbf{D} = diag(\sigma_1, \sigma_2, \cdots, \sigma_n)$$

  where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices.
- For each $\mathbf{y} \in \mathbf{A}(\mathcal{S}_2)$ there is an $\mathbf{x} \in \mathcal{S}_2$ such that $\mathbf{y} = \mathbf{A}\mathbf{x}$, so, with $\mathbf{w} = \mathbf{U}^T\mathbf{y}$,

$$\begin{aligned} 1 &= \|\mathbf{x}\|_2^2 = \|\mathbf{A}^{-1}\mathbf{A}\mathbf{x}\|_2^2 = \|\mathbf{A}^{-1}\mathbf{y}\|_2^2 = \|\mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T\mathbf{y}\|_2^2 \\ &= \|\mathbf{D}^{-1}\mathbf{w}\|_2^2 = \frac{\omega_1^2}{\sigma_1^2} + \frac{\omega_2^2}{\sigma_2^2} + \cdots + \frac{\omega_n^2}{\sigma_n^2}. \end{aligned}$$

- This means that $\mathbf{U}^T\mathbf{A}(\mathcal{S}_2)$ is an ellipsoid whose k^{th} semiaxis has length σ_k.
- Because orthogonal transformations are isometries, $\mathbf{A}(\mathcal{S}_2)$ is also an ellipsoid whose k^{th} semiaxis has length σ_k.



- The degree of distortion of the unit sphere under transformation by $\mathbf{A}$ is therefore measured by $\kappa_2 = \sigma_1/\sigma_n$.
- On the other hand, $\sigma_1 = \|\mathbf{A}\|_2$ and $\sigma_n = 1/\|\mathbf{A}^{-1}\|_2$, so $\kappa_2 = \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2$. This is called the 2-norm condition number of $\mathbf{A}$.

Image of the Unit Sphere

For a nonsingular $\mathbf{A}_{n\times n}$ having singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ and an SVD $\mathbf{A} = \mathbf{UDV}^T$ with $\mathbf{D} = \text{diag}\,(\sigma_1, \sigma_2, \ldots, \sigma_n)$, the image of the unit 2-sphere is an ellipsoid whose k^{th} semiaxis is given by $\sigma_k \mathbf{U}_{*k}$ Furthermore, $\mathbf{V}_{*k}$ is a point on the unit sphere such that $\mathbf{AV}_{*k} = \sigma_k \mathbf{U}_{*k}$ In particular,

- $\sigma_1 = \|\mathbf{AV}_{*1}\|_2 = \max\limits_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 = \|\mathbf{A}\|_2,$
- $\sigma_n = \|\mathbf{AV}_{*n}\|_2 = \min\limits_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 = 1/\|\mathbf{A}^{-1}\|_2.$

The degree of distortion of the unit sphere under transformation by $\mathbf{A}$ is measured by the 2-norm ***condition number***

- $\kappa_2 = \dfrac{\sigma_1}{\sigma_n} = \|\mathbf{A}\|_2 \left\|\mathbf{A}^{-1}\right\|_2 \geq 1.$

Notice that $\kappa_2 = 1$ if and only if $\mathbf{A}$ is an orthogonal matrix.

- The amount of distortion of the unit sphere under transformation by $\mathbf{A}$ determines the degree to which uncertainties in a linear system $\mathbf{Ax} = \mathbf{b}$ can be magnified.
- **Problem:** Let $\mathbf{Ax} = \mathbf{b}$ be a nonsingular system in which $\mathbf{A}$ is known exactly, but $\mathbf{b}$ is subject to an uncertainty $\mathbf{e}$, and consider $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{B} - \mathbf{e} = \tilde{\mathbf{b}}$. Estimate the relative uncertainty $\|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$ in terms of the relative uncertainty $\|\mathbf{b} - \tilde{\mathbf{b}}\|/\|\mathbf{b}\| = \|\mathbf{e}\|/\|\mathbf{b}\|$.

Solution: Use $\|\mathbf{b}\| = \|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\|\,\|\mathbf{x}\|$ with $\mathbf{x} - \tilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{e}$ to write

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} = \frac{\left\|\mathbf{A}^{-1}\mathbf{e}\right\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}\|\left\|\mathbf{A}^{-1}\right\|\|\mathbf{e}\|}{\|\mathbf{b}\|} = \kappa\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|},$$

where $\kappa = \|\mathbf{A}\|\left\|\mathbf{A}^{-1}\right\|$ is a *condition number* as discussed earlier ($\kappa = \sigma_1/\sigma_n$ if the 2-norm is used). Furthermore, $\|\mathbf{e}\| = \|\mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}})\| \leq \|\mathbf{A}\|\,\|(\mathbf{x} - \tilde{\mathbf{x}})\|$ and $\|\mathbf{x}\| \leq \left\|\mathbf{A}^{-1}\right\|\|\mathbf{b}\|$ imply

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \geq \frac{\|\mathbf{e}\|}{\|\mathbf{A}\|\,\|\mathbf{x}\|} \geq \frac{\|\mathbf{e}\|}{\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|\,\|\mathbf{b}\|} = \frac{1}{\kappa}\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|}.$$

This yields the following bounds on the relative uncertainty:

$$\kappa^{-1}\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|} \leq \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \kappa\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|}, \quad \text{where} \quad \kappa = \|\mathbf{A}\|\left\|\mathbf{A}^{-1}\right\|.$$

- In other words, when $\mathbf{A}$ is well conditioned (i.e. when κ is small), small relative uncertainties in $\mathbf{b}$ can not greatly affect the solution.
- When $\mathbf{A}$ is ill conditioned (when κ is large). a relatively small uncertainty in $\mathbf{b}$ might result in a relatively large uncertainty in $\mathbf{x}$.

- In addition to measuring the distortion of the unit sphere and gauging the sensitivity of linear systems, singular values provide a measure of how close A is to a matrix of lower rank.

Distance to Lower-Rank Matrices

If $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are the nonzero singular values of $\mathbf{A}_{m\times n}$, then for each $k < r$, the distance from $\mathbf{A}$ to the closest matrix of rank k is

$$\sigma_{k+1} = \min_{rank(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_2.$$

- The SVD can be a useful tool in applications involving the need to sort through noisy data and lift out relevant information.
- Suppose that $\mathbf{A}_{m\times n}$ is a matrix containing data that are contaminated with a certain level of noise.
- The SVD resolves the data in $\mathbf{A}$ into r mutually orthogonal components by writing

$$\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{D}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \sum_{i=1}^{r} \sigma_i \mathbf{Z}_i,$$

where $\mathbf{Z}_i = \mathbf{u}_i \mathbf{v}_i^T$.

- In other words, the SVD can be regarded as a Fourier expansion, and $\sigma_i =< \mathbf{Z}_i|\mathbf{A} >$ can be interpreted as the proportion of $\mathbf{A}$ lying in the direction of $\mathbf{Z}_i$.
- In many applications the noise contamination in $\mathbf{A}$ is random (or nondirectional) in the sense that the noise is distributed more or less uniformly across the $\mathbf{Z}_i$'s.
- That is, there is about as much noise in the "direction" of one $\mathbf{Z}_i$ as there is in the "direction" of any other.
- This means that if SNR($\sigma_i\mathbf{Z}_i$) denotes the signal-to-noise ratio then

$$SNR(\sigma_1\mathbf{Z}_1) \geq SNR(\sigma_2\mathbf{Z}_2) \geq \cdots \geq SNR(\sigma_r\mathbf{Z}_r),$$

  more or less.
- If some of singular values, say $\sigma_{k+1}, \cdots, \sigma_r$ are small relative to (total noise)/r, then the terms $\sigma_{k+1}\mathbf{Z}_{k+1}, \cdots, \sigma_r\mathbf{Z}_r$ have small signal-to-noise ratios.
- Therefore, if we delete these term, then we lose a small part of the total signal, but we remove a disproportionately large component of the total noise in $\mathbf{A}$.
- This explains why a truncated SVD $\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i\mathbf{Z}_i$ can in many instances, filter out some of the noise without losing significant information about the signal.

- Just as the Drazin inverse of a square matrix, a URV factorization or an SVD can be used to define a generalized inverse for rectangular matrices.

Moore–Penrose Pseudoinverse

- In terms of URV factors, the Moore–Penrose pseudoinverse of

$$\mathbf{A}_{m\times n}=\mathbf{U}\begin{pmatrix}\mathbf{C}_{r\times r} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T \text{ is } \mathbf{A}^{\dagger}_{n\times m}=\mathbf{V}\begin{pmatrix}\mathbf{C}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T.$$

- When $\mathbf{Ax}=\mathbf{b}$ is consistent, $\mathbf{x}=\mathbf{A}^{\dagger}\mathbf{b}$ is the solution of minimal euclidean norm.
- When $\mathbf{Ax}=\mathbf{b}$ is inconsistent, $\mathbf{x}=\mathbf{A}^{\dagger}\mathbf{b}$ is the least squares solution of minimal euclidean norm.
- When an SVD is used, $\mathbf{C}=\mathbf{D}=\operatorname{diag}(\sigma_1,\ldots,\sigma_r)$, so

$$\mathbf{A}^{\dagger}=\mathbf{V}\begin{pmatrix}\mathbf{D}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T=\sum_{i=1}^{r}\frac{\mathbf{v}_i\mathbf{u}_i^T}{\sigma_i} \quad \text{and} \quad \mathbf{A}^{\dagger}\mathbf{b}=\sum_{i=1}^{r}\frac{(\mathbf{u}_i^T\mathbf{b})}{\sigma_i}\mathbf{v}_i.$$

- It can be proven that $\mathbf{A}^{\dagger}$ is unique by arguing that $\mathbf{A}^{\dagger}$ is the unique solution to the four Penrose equations

$$\mathbf{AA}^{\dagger}\mathbf{A}=\mathbf{A},\quad \mathbf{A}^{\dagger}\mathbf{AA}^{\dagger}=\mathbf{A}^{\dagger},\quad (\mathbf{AA}^{\dagger})^T=\mathbf{AA}^{\dagger},\quad (\mathbf{A}^{\dagger}\mathbf{A})^T=\mathbf{A}^{\dagger}\mathbf{A}.$$

Orthogonal Projection

- As discussed in the above section, every pair of complementary subspaces defines a projector.
- When the complementary subspaces happen to be orthogonal, the resulting projector has some particularly nice properties.
- This section is to develop this special case in more detail.
- Discussion are in the context of real space, but generalizations to complex spaces are straightforward by replacing $(\star)^T$ by $(\star)^*$ and "orthogonal matrix" by "unitary matrix."

Orthogonal Projection

For $\mathbf{v} \in \mathcal{V}$, let $\mathbf{v} = \mathbf{m} + \mathbf{n}$, where $\mathbf{m} \in \mathcal{M}$ and $\mathbf{n} \in \mathcal{M}^{\perp}$.

- $\mathbf{m}$ is called the ***orthogonal projection*** of $\mathbf{v}$ onto $\mathcal{M}$.
- The projector $\mathbf{P}_{\mathcal{M}}$ onto $\mathcal{M}$ along $\mathcal{M}^{\perp}$ is called the ***orthogonal projector*** onto $\mathcal{M}$.
- $\mathbf{P}_{\mathcal{M}}$ is the unique linear operator such that $\mathbf{P}_{\mathcal{M}}\mathbf{v} = \mathbf{m}$.

- Given an arbitrary pair of complementary subspace $\mathcal{M}$, $\mathcal{N}$ of $\Re^n$, the projector $\mathbf{P}$ onto $\mathcal{M}$ along $\mathcal{N}$ is given by

$$\mathbf{P} = (\mathbf{M}|\mathbf{N}) \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} (\mathbf{M}|\mathbf{N})^{-1} = (\mathbf{M}|\mathbf{0})(\mathbf{M}|\mathbf{N})^{-1}.$$

- How does this expression simplify when $\mathcal{N} = \mathcal{M}^{\perp}$?

Constructing Orthogonal Projectors

Let $\mathcal{M}$ be an r-dimensional subspace of $\Re^n$, and let the columns of $\mathbf{M}_{n\times r}$ and $\mathbf{N}_{n\times n-r}$ be bases for $\mathcal{M}$ and $\mathcal{M}^{\perp}$, respectively. The orthogonal projectors onto $\mathcal{M}$ and $\mathcal{M}^{\perp}$ are

- $\mathbf{P}_{\mathcal{M}} = \mathbf{M}\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T$ and $\mathbf{P}_{\mathcal{M}^{\perp}} = \mathbf{N}\left(\mathbf{N}^T\mathbf{N}\right)^{-1}\mathbf{N}^T$.

If $\mathbf{M}$ and $\mathbf{N}$ contain *orthonormal* bases for $\mathcal{M}$ and $\mathcal{M}^{\perp}$, then

- $\mathbf{P}_{\mathcal{M}} = \mathbf{M}\mathbf{M}^T$ and $\mathbf{P}_{\mathcal{M}^{\perp}} = \mathbf{N}\mathbf{N}^T$.
- $\mathbf{P}_{\mathcal{M}} = \mathbf{U}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T$, where $\mathbf{U} = (\mathbf{M}\,|\,\mathbf{N})$.
- $\mathbf{P}_{\mathcal{M}^{\perp}} = \mathbf{I} - \mathbf{P}_{\mathcal{M}}$ in all cases.

- **Problem:** Let $\mathbf{u}_{n\times 1} \neq 0$, and consider the line $\mathcal{L} = span\{\mathbf{u}\}$. Construct the orthogonal projector onto $\mathcal{L}$, and then determine the orthogonal projection of a vector $\mathbf{x}_{n\times 1}$ onto $\mathcal{L}$.
- **Solution:** The vector $\mathbf{u}$ by itself is a basis for $\mathcal{L}$, so

$$\mathbf{P}_{\mathcal{L}} = \mathbf{u}(\mathbf{u}^T\mathbf{u})^{-1}\mathbf{u}^T = \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}$$

  is the orthogonal projector ont $\mathcal{L}$. The orthogonal projection of a vector $\mathbf{x}$ onto $\mathcal{L}$ is given by

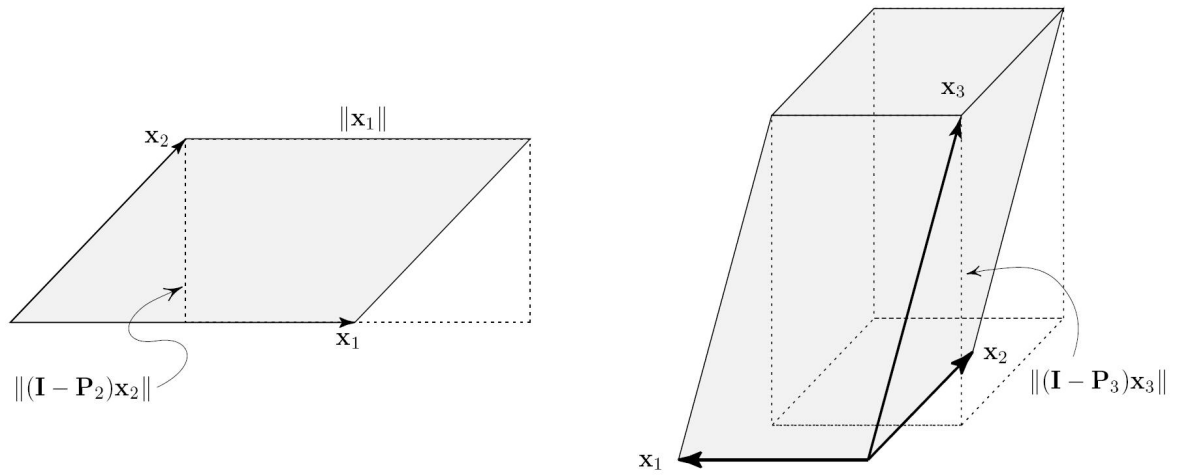
$$\mathbf{P}_{\mathcal{L}}\mathbf{x} = \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}\mathbf{x} = \frac{\mathbf{u}^T\mathbf{x}}{\mathbf{u}^T\mathbf{u}}\mathbf{u}.$$

- Note: if $\|\mathbf{u}\|_2 = 1$ then $\mathbf{P}_{\mathcal{L}} = \mathbf{u}\mathbf{u}^T$, so $\mathbf{P}_{\mathcal{L}}\mathbf{x} = (\mathbf{u}^T\mathbf{x})\mathbf{u}$, and

$$\|\mathbf{P}_{\mathcal{L}}\mathbf{x}\|_2 = |\mathbf{u}^T\mathbf{x}|.$$

- Since $\mathbf{P}_{\mathcal{L}} = \mathbf{u}\mathbf{u}^T$ is the orthogonal projector onto $\mathcal{L}$, it must be the case that $\mathbf{P}_{\mathcal{L}^\perp} = \mathbf{I} - \mathbf{u}\mathbf{u}^T$ is the orthogonal projection onto $\mathcal{L}^\perp$.
- This was called an elementary orthogonal projector.

- **Problem:** Determine the volumes of a two-dimensional and a three -dimensional parallelepiped, and then make the natural extension to define the volume of an n-dimensional parallelepiped.



- **Solution:** In the two-dimensional case, volume is area, and it's evident that the area of the shaded parallelogram is the same as the area of the dotted rectangle.
  - The width of the dotted rectangle is $\nu_1 = \|\mathbf{x}_1\|_2$, and the height is $\nu_2 = \|(\mathbf{I}-\mathbf{P}_2)\mathbf{x}_2\|_2$, where $\mathbf{P}_2$ is the orthogonal projector onto the space spanned by $\mathbf{x}_1$.
  - $V_2 = \|\mathbf{x}_1\|_2\|(\mathbf{I}-\mathbf{P}_2)\mathbf{x}_2\|_2 = \nu_1\nu_2$.

- Similarly, the volume of a three-dimensional parallelepiped is the area of its base times its projected height.
  - The area of the base was just determined to be $V_2 = \|\mathbf{x}_1\|_2\|(\mathbf{I}-\mathbf{P}_2)\mathbf{x}_2\|_2 = \nu_1\nu_2$.
  - The projected height is $\nu_3 = \|(\mathbf{I}-\mathbf{P}_3)\mathbf{x}_3\|_2$, where $\mathbf{P}_3$ is the orthogonal projector onto $span\{\mathbf{x}_1, \mathbf{x}_2\}$.
  - The volume of the parallelepiped generated by $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is $V_3 = \nu_1\nu_2\nu_3$.
- It's now clear how to inductively define V_4, V_5, etc. In general, the volume of the parallelepiped generated by a linearly independent set $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ is

$$V_n = \|\mathbf{x}_1\|_2\|(\mathbf{I}-\mathbf{P}_2)\mathbf{x}_2\|_2\|(\mathbf{I}-\mathbf{P}_3)\mathbf{x}_3\|_2\cdots\|(\mathbf{I}-\mathbf{P}_n)\mathbf{x}_n\|_2 = \nu_1\nu_2\cdots\nu_n,$$

where $\mathbf{P}_k$ is the orthogonal projector onto $span\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{k-1}\}$, and where

$$\nu_1 = \|\mathbf{x}_1\|_2 \quad and \quad \nu_k = \|(\mathbf{I}-\mathbf{P}_k)\mathbf{x}_k\|_2 \quad for \quad k > 1.$$

Note that if $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ is an orthogonal set, $V_n = \|\mathbf{x}_1\|_2\|\mathbf{x}_2\|_2\cdots\|\mathbf{x}_n\|_2$.

- If $\mathbf{A}_{m\times n} = \mathbf{Q}_{m\times n}\mathbf{R}$ it the QR factorization of a matrix with linearly independent columns, then the volume of the n-dimensional parallelepiped generated by the columns of $\mathbf{A}$ is $V_n = \nu_1\nu_2\cdots\nu_n$, where the ν_k's are the diagonal elements of $\mathbf{R}$.
- What characteristic features distinguish orthogonal projectors from more general oblique projectors?

Orthogonal Projectors

Suppose that $\mathbf{P} \in \Re^{n\times n}$ is a projector—i.e., $\mathbf{P}^2 = \mathbf{P}$. The following statements are equivalent to saying that $\mathbf{P}$ is an *orthogonal* projector.

- $R(\mathbf{P}) \perp N(\mathbf{P})$.
- $\mathbf{P}^T = \mathbf{P}$ (i.e., orthogonal projector $\Longleftrightarrow \mathbf{P}^2 = \mathbf{P} = \mathbf{P}^T$).
- $\|\mathbf{P}\|_2 = 1$ for the matrix 2-norm.

- The notion of orthogonal projection in higher-dimensional spaces is consistent with the visual geometry in $\Re^2$ and $\Re^3$.

- In particular, if $\mathcal{M}$ is a subspace of $\Re^3$, and if $\mathbf{b}$ is a vector outside of $\mathcal{M}$, then the point in $\mathcal{M}$ that is closest to $\mathbf{b}$ is $\mathbf{p} = \mathbf{P}_{\mathcal{M}}\mathbf{b}$, the orthogonal projection of b onto $\mathcal{M}$.
- The situation is exactly the same in higher dimensions.

Closest Point Theorem

Let $\mathcal{M}$ be a subspace of an inner-product space $\mathcal{V}$, and let $\mathbf{b}$ be a vector in $\mathcal{V}$. The unique vector in $\mathcal{M}$ that is closest to $\mathbf{b}$ is $\mathbf{p} = \mathbf{P}_{\mathcal{M}}\mathbf{b}$, the orthogonal projection of $\mathbf{b}$ onto $\mathcal{M}$. In other words,

$$\min_{\mathbf{m}\in\mathcal{M}} \|\mathbf{b} - \mathbf{m}\|_2 = \|\mathbf{b} - \mathbf{P}_{\mathcal{M}}\mathbf{b}\|_2 = dist\,(\mathbf{b}, \mathcal{M}).$$

This is called the ***orthogonal distance*** between $\mathbf{b}$ and $\mathcal{M}$.

- We are now in a position to replace the classical calculus-based theory of least squares presented in with a more modern vector space development.
- For an inconsistent system $\mathbf{Ax} = \mathbf{b}$, the object of the least squares problem is to find vectors $\mathbf{x}$ that minimize the quantity

$$(\mathbf{Ax} - \mathbf{b})^T(\mathbf{Ax} - \mathbf{b}) = \|\mathbf{Ax} - \mathbf{b}\|_2^2.$$

- The classical development relies on calculus to argue that the set of vectors $\mathbf{x}$ is exactly the set that solves the (always consistent) system of normal equations $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$.
- In the context of the closest point theorem the least squares problem asks for vectors $\mathbf{x}$ such that $\mathbf{A}\mathbf{x}$ is as close to $\mathbf{b}$ as possible.
- $\mathbf{A}\mathbf{x}$ is always a vector in $R(\mathbf{A})$, and the closest point theorem says that the vector closest to $\mathbf{b}$ is $\mathbf{P}_{R(\mathbf{A})}\mathbf{b}$.

Least Squares Solutions

Each of the following four statements is equivalent to saying that $\widehat{\mathbf{x}}$ is a least squares solution for a possibly inconsistent linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$.

- $\|\mathbf{A}\widehat{\mathbf{x}} - \mathbf{b}\|_2 = \min_{\mathbf{x}\in\Re^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$.
- $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{P}_{R(\mathbf{A})}\mathbf{b}$.
- $\mathbf{A}^T\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^T\mathbf{b}$ $\quad$ ($\mathbf{A}^*\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^*\mathbf{b}$ when $\mathbf{A} \in \mathcal{C}^{m\times n}$).
- $\widehat{\mathbf{x}} \in \mathbf{A}^\dagger\mathbf{b} + N(\mathbf{A})$ ($\mathbf{A}^\dagger\mathbf{b}$ is the minimal 2-norm LSS).

- **Caution:**These are valuable theoretical characterizations, but none is recommended for floating-point computation.

Exercises

1. Using Housholder reduction and Givens reduction, compute the QR factors of
$$\mathbf{A} = \begin{pmatrix} 1 & 19 & -34 \\ -2 & -5 & 20 \\ 2 & 8 & 37 \end{pmatrix}.$$
2. By using Housholder reduction, find an orthonormal basis for $R(\mathbf{A})$, where
$$\mathbf{A} = \begin{pmatrix} 4 & -3 & 4 \\ 2 & -14 & -3 \\ -2 & 14 & 0 \\ 1 & -7 & 15 \end{pmatrix}.$$
3. Evaluate the following convolutions
(a) $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \odot \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}$ (b) $\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \odot \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$.

4. Evaluate the discrete Fourier transform of $\begin{pmatrix} 1 \\ -i \\ -1 \\ i \end{pmatrix}$.
5. Prove that $\mathbf{a} \odot \mathbf{b} = \mathbf{b} \odot \mathbf{a}$ for all $\mathbf{a}, \mathbf{b} \in \mathcal{C}^n$.
6. Apply the FFT algorithm to the vector $\mathbf{x}_8 = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_7 \end{pmatrix}$, and then verify that your answer agrees with the result obtained by computing $\mathbf{F}_8\mathbf{x}_8$.
7. Let $\mathcal{X}$ and $\mathcal{Y}$ be subspaces of $\Re^3$ whose respective bases are

$$\mathcal{B}_{\mathcal{X}} = \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \right\} \quad and \quad \mathcal{B}_{\mathcal{Y}} = \left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \right\}.$$

   (a) Explain why $\mathcal{X}$ and $\mathcal{Y}$ are complementary subspaces of $\Re^3$.
   (b) Determine the projector $\mathbf{P}$ onto $\mathcal{X}$ along $\mathcal{Y}$ as well as the complementary projector $\mathbf{Q}$ onto $\mathcal{Y}$ along $\mathcal{X}$.
   (c) Determine the projection of $\mathbf{v} = (2 \ -1 \ 1)^T$ onto $\mathcal{Y}$ along $\mathcal{X}$.

(d) Verify that $\mathbf{P}$ and $\mathbf{Q}$ are both idempotent.
(e) Verify that $R(\mathbf{P}) = \mathcal{X} = N(\mathbf{Q})$ and $N(\mathbf{P}) = \mathcal{Y} = R(\mathbf{Q})$.

8. Construct an example to show that if $\mathcal{V} = \mathcal{X} + \mathcal{Y}$ but $\mathcal{X} \bigcap \mathcal{Y} \neq 0$, then a vector $\mathbf{v} \in \mathcal{V}$ can have two different representations as

$$\mathbf{v} = \mathbf{x}_1 + \mathbf{y}_1 \quad and \quad \mathbf{v} = \mathbf{x}_2 + \mathbf{y}_2,$$

where $\mathbf{x}_1$, $\mathbf{x}_2 \in \mathcal{X}$ and $\mathbf{y}_1$, $\mathbf{y}_2 \in \mathcal{Y}$, but $\mathbf{x}_1 \neq \mathbf{x}_2$ and $\mathbf{y}_1 \neq \mathbf{y}_2$.

9. If $\mathbf{A}$ is a square matrix of index $k > 0$, prove that $index(\mathbf{A}^k) = 1$.

10. Find the Drazin inverse of $\mathbf{A} = \begin{pmatrix} -2 & 0 & -4 \\ 4 & 3 & 4 \\ 3 & 2 & 2 \end{pmatrix}$.

11. For every inner-product space $\mathcal{V}$, prove that if $\mathcal{M} \subseteq \mathcal{V}$, then $\mathcal{M}^{\perp}$ is a subspace of $\mathcal{V}$.

12. Following the derivation in the text, find an SVD for $\mathbf{C} = \begin{pmatrix} -4 & -6 \\ 3 & -8 \end{pmatrix}$.

13. Consider the matrix $\mathbf{A} = \begin{pmatrix} -4 & -2 & -4 & -2 \\ 2 & -2 & 2 & 1 \\ -4 & 1 & -4 & -2 \end{pmatrix}$, Use the URV factorization to determine $\mathbf{A}^{\dagger}$.
14. Find the orthogonal projection of $\mathbf{b}$ onto $\mathcal{M} = span\{\mathbf{u}\}$, and then determine the orthogonal projection of $\mathbf{b}$ onto $\mathcal{M}^{\perp}$, where $\mathbf{b} = (4\ 8)^T$ and $\mathbf{u} = (3\ 1)^T$.
15. Let $\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 4 & 1 \\ 1 & 2 & 0 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.
    (a) Compute the orthogonal projectors onto each of the four fundamental subspaces associated with $\mathbf{A}$.
    (b) Find the point in $N(\mathbf{A})^{\perp}$ that is closest to $\mathbf{b}$.
16. Determine the orthogonal projection of $\mathbf{b}$ onto $\mathcal{M}$, where

$$\mathbf{b} = \begin{pmatrix} 5 \\ 2 \\ 5 \\ 3 \end{pmatrix} \quad and \quad \mathcal{M} = span\left\{ \begin{pmatrix} -3/5 \\ 0 \\ 4/5 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 4/5 \\ 0 \\ 3/5 \\ 0 \end{pmatrix} \right\}$$