Information Theory for Machine Learning

Chapter 3: Orthogonality

Jin-Ho Chung
School of IT Convergence
University of Ulsan

Table of Contents

Norm

Definition

For any $\mathbf{x} \in \mathbb{C}^n$ and $c \in \mathbb{C}$, a real-valued function $\|\mathbf{x}\|$ is said to be a norm if it satisfies the followings:

- $\|\mathbf{x}\| \geq 0$ with equality only for $\mathbf{x} = \mathbf{0}$.
- $\|c\mathbf{x}\| = |c|\|\mathbf{x}\|$.
- $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

- l_1-norm: $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$.
- l_2-norm: $\|\mathbf{x}\|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2}$.
- l_∞-norm: $\|\mathbf{x}\|_\infty = \max |x_i|$.

Matrix Norm

Definition

The norm of a matrix A is defined as

$$\|A\| = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|.$$

- $\|A\mathbf{x}\| \leq \|A\| \|\mathbf{x}\|$.
- $\|A + B\| \leq \|A\| + \|B\|$.
- $\|AB\| \leq \|A\| \|B\|$.
- $\|(A + B)\mathbf{x}\| \leq (\|A\| + \|B\|)\|\mathbf{x}\|$.
- $\|AB\mathbf{x}\| \leq \|A\| \|B\mathbf{x}\| \leq \|A\| \|B\| \|\mathbf{x}\|$.

Inner Product

Definition

Let $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and $c \in \mathbb{C}$, the operation $(\cdot, \cdot)$ is said to be an inner product if it satisfies the followings:

- $(\mathbf{u} + \mathbf{v}, \mathbf{w}) = (\mathbf{u}, \mathbf{w}) + (\mathbf{v}, \mathbf{w})$.
- $(c\mathbf{v}, \mathbf{w}) = c(\mathbf{v}, \mathbf{w})$.
- $(\mathbf{v}, \mathbf{w}) = (\mathbf{w}, \mathbf{v})^*$.
- $(\mathbf{v}, \mathbf{v}) \geq 0$ with equality only for $\mathbf{v} = \mathbf{0}$..

Note that the norm $\mathbf{x}$ is the inner product $\langle \mathbf{x}, \mathbf{x} \rangle$.

Example of Inner Products

1. Let $F(0, 2\pi)$ be the set of all square-integrable functions defined on $[0, 2\pi]$. Then, it is a vector space over $\mathbb{R}$ with the inner product

$$(f(x), g(x)) = \int_0^{2\pi} f(x)g(x)dx.$$

2. Let P_∞ be the set of all polynomial functions defined on $-1 \leq x \leq 1$, that is ,

$$P_\infty = \left\{ \sum_{i=0}^{n} a_i x^i \,\middle|\, a_i \in \mathbb{R}, \quad n \in \mathbb{N} \cup \{0\} \right\}$$

Then, it is a vector space over $\mathbb{R}$ with the inner product

$$(f(x), g(x)) = \int_{-1}^{+1} f(x)g(x)dx.$$

Table of Contents

The Length of a Vector

Definition

For $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, the length (or norm) of $\mathbf{x}$ is defined as

$$\|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}.$$

- Clearly, $\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x}$.
- Properties
  - $\|\mathbf{x}\| \geq 0$ with equality only for $\mathbf{x} = \mathbf{0}$.
  - $\|c\mathbf{x}\| = |c|\|\mathbf{x}\|$.
  - $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

Orthogonal Vectors

Definition

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. The vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if

$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2,$$

or equivalently

$$\mathbf{x}^T\mathbf{y} = 0.$$

- $\mathbf{x}^T\mathbf{y}$ is called the inner product of $\mathbf{x}$ and $\mathbf{y}$.
- In the standard basis $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$, any two vectors are mutually orthogonal.

Linear Independence

Theorem

If the nonzero vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ in $\mathbb{R}^n$ are pairwise orthogonal, then they are linearly independent.

Proof. Suppose $c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \mathbf{0}$ for some $c_1, \ldots, c_k \in \mathbb{R}$. Then,

$$\begin{aligned} 0 &= \mathbf{v}_j^T \left(\sum_{i=1}^{k} c_i \mathbf{v}_i \right) \\ &= \sum_{i=1}^{k} c_i \mathbf{v}_j^T \mathbf{v}_i \\ &= c_j \mathbf{v}_j^T \mathbf{v}_j. \end{aligned}$$

Therefore, $c_j = 0$.

Orthogonal Spaces

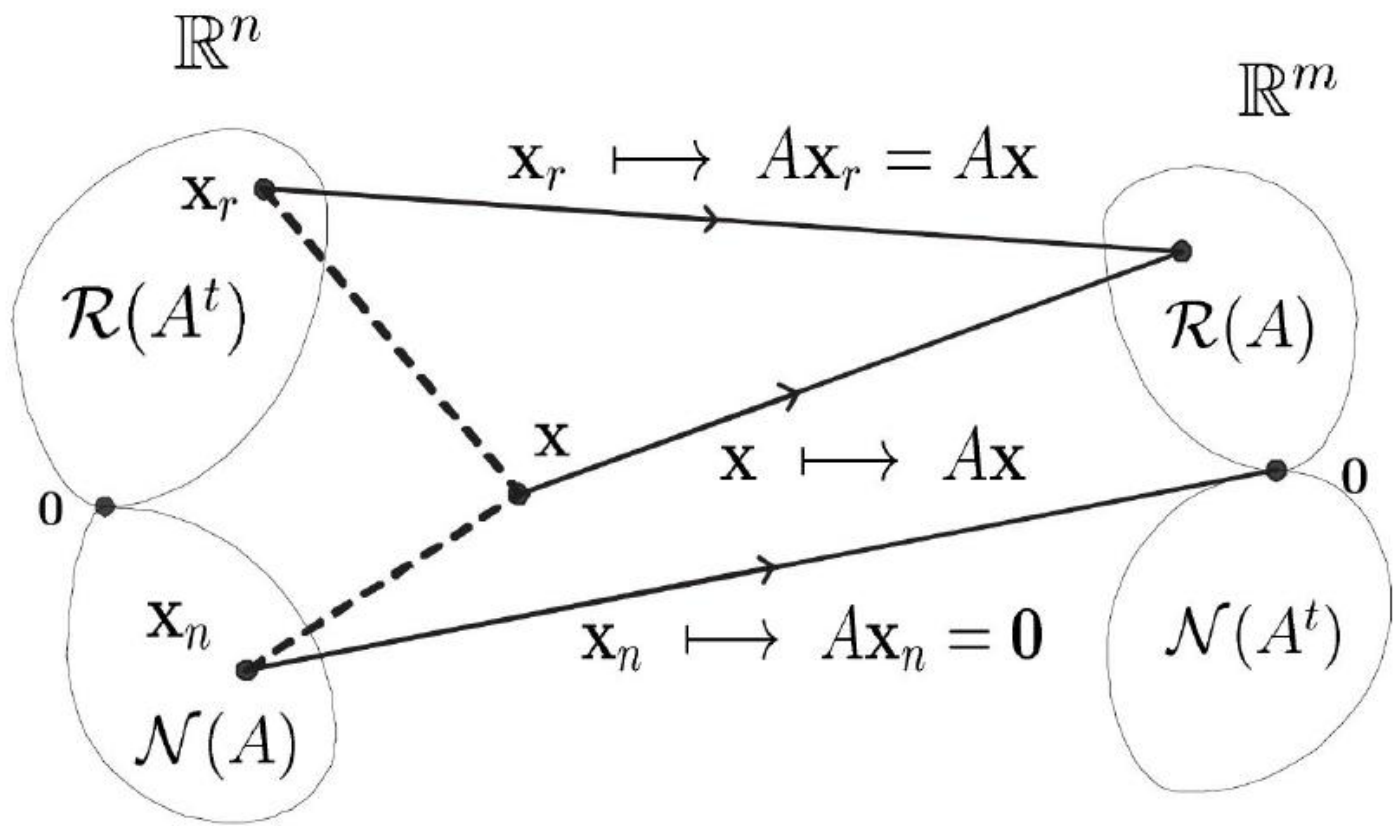
Definition

Let V and W be two subspaces of $\mathbb{R}^n$.

1) V and W are said to be *orthogonal* if $\mathbf{v}^T\mathbf{w} = 0$ for any $\mathbf{v} \in V$ and $\mathbf{w} \in W$. $(V \perp W)$
2) The set of all vectors in $\mathbb{R}^n$ which are orthogonal to all the vectors in V is called the *orthogonal complement* of V. $(V^{\perp})$

- $V^{\perp}$ is a subspace of $\mathbb{R}^n$.
- $(V^{\perp})^{\perp} = V$ and $(V_1 + V_2)^{\perp} = V_1^{\perp} \cap V_2^{\perp}$.
- The nullspace of an $m \times n$ matrix A is the orthogonal complement of its row space in $\mathbb{R}^n$.

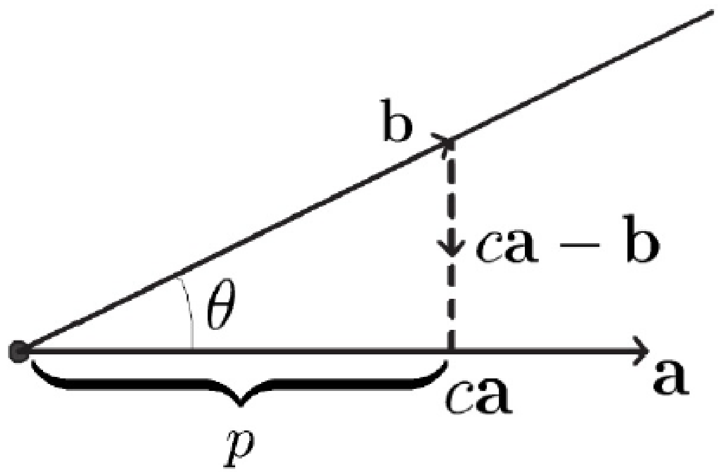
Orthogonal Spaces



$$\left[\begin{array}{lcl} \mathbb{R}^n & = & \mathcal{R}(A^t) + \mathcal{N}(A) \\ \mathbb{R}^m & = & \mathcal{R}(A) + \mathcal{N}(A^t). \end{array}\right.$$

Table of Contents

What Is Projection?



- Clearly, $(c\mathbf{a} - \mathbf{b}) \perp \mathbf{a}$.

$$\begin{aligned} \Rightarrow \quad & (c\mathbf{a} - \mathbf{b})^T \mathbf{a} = 0 \\ \Rightarrow \quad & c\mathbf{a}^T \mathbf{a} - \mathbf{b}^T \mathbf{a} = 0 \\ \Rightarrow \quad & c = \frac{\mathbf{b}^T \mathbf{a}}{\|\mathbf{a}\|^2}. \end{aligned}$$

- Cosine Function:

$$\begin{aligned} & \|\mathbf{p}\| = c\|\mathbf{a}\| \triangleq \|\mathbf{b}\| \cos\theta \\ \Rightarrow \quad & \cos\theta = \frac{\mathbf{b}^T \mathbf{a}}{\|\mathbf{a}\|\|\mathbf{b}\|} \end{aligned}$$

Projection Matrix

- Projected Vector $\mathbf{p}$

$$\mathbf{p} = c\,\mathbf{a} = \frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}}\,\mathbf{a} = \left(\frac{\mathbf{a}}{\|\mathbf{a}\|}\right)^T \mathbf{b}\,\frac{\mathbf{a}}{\|\mathbf{a}\|}.$$

- Projection Matrix P

$$\mathbf{p} = \mathbf{a}\,c = \mathbf{a}\frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}} = \frac{\mathbf{a}\mathbf{a}^T}{\mathbf{a}^T\mathbf{a}}\mathbf{b} \triangleq P\mathbf{b}.$$

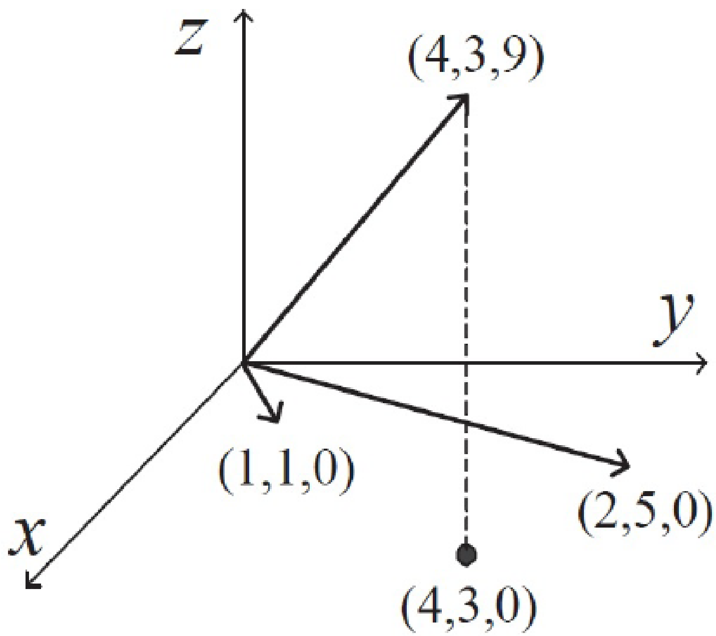
- Schwarts Inequality: For any $\mathbf{a},\ \mathbf{b} \in \mathbb{R}^n$,

$$|\mathbf{a}^T\mathbf{b}| \le \|\mathbf{a}\|\,\|\mathbf{b}\| \quad (\text{or } |\cos\theta| \le 1).$$

Table of Contents

Least Squares Approximation

- **Goal**: Given an $m \times 1$ vector $\mathbf{b}$ and an $m \times n$ matrix A, find $\mathbf{x}$ which minimizes $\|A\mathbf{x} - \mathbf{b}\|^2$.
- Example of the $n = 2$ case:



$$\begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \cong \begin{bmatrix} 4 \\ 3 \\ 9 \end{bmatrix}$$

$$\Rightarrow \quad x_1 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 5 \\ 0 \end{bmatrix} \cong \begin{bmatrix} 4 \\ 3 \\ 0 \end{bmatrix}$$

Orthogonality Principle

- **Statement**: The vector $\mathbf{x}_{\text{opt}}$ that minimizes $\|\mathbf{b} - A\mathbf{x}\|$ can be obtained by solving

$$A^T(\mathbf{b} - A\mathbf{x}_{\text{opt}}) = \mathbf{0}. \qquad (*)$$

- **Proof**: Let $\mathbf{x}_{\text{opt}}$ satisfy $(*)$. For any vector $\mathbf{x} \in \mathbb{R}_n$, we have

$$\begin{aligned}
\|\mathbf{b} - A\mathbf{x}\|^2 &= \|\mathbf{b} - A\mathbf{x}_{\text{opt}} + A\mathbf{x}_{\text{opt}} - A\mathbf{x}\|^2 \\
&= \|\mathbf{b} - A\mathbf{x}_{\text{opt}}\|^2 + \|A\mathbf{x}_{\text{opt}} - A\mathbf{x}\|^2 \\
&\qquad +2(\mathbf{x}_{\text{opt}} - \mathbf{x})^T A^T(\mathbf{b} - A\mathbf{x}_{\text{opt}}) \\
&= \|\mathbf{b} - A\mathbf{x}_{\text{opt}}\|^2 + \|A\mathbf{x}_{\text{opt}} - A\mathbf{x}\|^2 \\
&\geq \|\mathbf{b} - A\mathbf{x}_{\text{opt}}\|^2.
\end{aligned}$$

Remark

- **Question 1**: Does $A\mathbf{x}_{\text{opt}}$ depend on A? (Yes, but depends on the column space rather than A.)
- **Question 2**: Does $\mathbf{x}_{\text{opt}}$ depend on A? (Yes)
- **Exercise:** Let $\mathbf{b} = [4\ 3\ 7]^T$. Calculate $\mathbf{x}_{\text{opt}}$ and $A_i\mathbf{x}_{\text{opt}}$ for

$$A_1 = \begin{bmatrix} 2 & 2 \\ 2 & 5 \\ 0 & 0 \end{bmatrix}$$

and for

$$A_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix},$$

respectively.

Summary of LSA

- Normal Equations:

$$A^T A\mathbf{x}_{\text{opt}} = A^T\mathbf{b}.$$

- Special case: If the columns of A are linear independent, we have

$$\mathbf{x}_{\text{opt}} = (A^T A)^{-1} A^T\mathbf{b}$$

which is called best estimate, and the nearest point is

$$\mathbf{p} = A\mathbf{x}_{\text{opt}} = A(A^T A)^{-1} A^T\mathbf{b}.$$

The Cross-Product Matrix $A^T A$

Properties of $A^T A$

If A has independent columns, then $A^T A$ is square, symmetric, and invertible.

- $A^T A$ has the same nullspace as A.
  - $\mathcal{N}(A) \subseteq \mathcal{N}(A^T A)$:

$$A\mathbf{x} = \mathbf{0} \quad \Rightarrow \quad A^T(A\mathbf{x}) = \mathbf{0}.$$

  - $\mathcal{N}(A^T A) \subseteq \mathcal{N}(A)$:

$$A^T A\mathbf{x} = \mathbf{0} \;\Rightarrow\; \mathbf{x}^T A^T(A\mathbf{x}) = 0 \;\Leftrightarrow\; \|A\mathbf{x}\| = 0 \;\Rightarrow\; A\mathbf{x} = \mathbf{0}.$$

- $\dim(\mathcal{C}(A)) = n \quad \Rightarrow \quad \dim(\mathcal{C}(A^T A)) = n.$

Table of Contents

Orthonormal Vectors

- A set of vectors $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ are said to be orthonormal if

$$\mathbf{q}_i^T \mathbf{q}_j = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases}$$

- A square matrix Q is an orthogonal matrix if its columns are orthonormal (Note that $Q^T = Q^{-1}$).
- Examples of Q:

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

Properties of Orthogonal Matrices

- Q preserves inner products:

$$(Q\mathbf{x}, Q\mathbf{y}) = \mathbf{x}^T Q^T Q\mathbf{y} = \mathbf{x}^T\mathbf{y} = (\mathbf{x}, \mathbf{y}).$$

- Q preserves norms:

$$\|Q\mathbf{x}\|^2 = (Q\mathbf{x}, Q\mathbf{x}) = (\mathbf{x}, \mathbf{x}) = \|\mathbf{x}\|^2.$$

- If an $n \times n$ matrix A preserves inner products, then A should be orthogonal:

$$(A^T A)_{i,j} = \mathbf{e}_i^T A^T A\mathbf{e}_j = (A\mathbf{e}_i, A\mathbf{e}_j) = (\mathbf{e}_i, \mathbf{e}_j) = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j \end{cases}$$

where $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is the standard bases.

Table of Contents

Gram-Schmidt Orthogonalization

- Start: There is a set of linearly independent vectors

$$\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}.$$

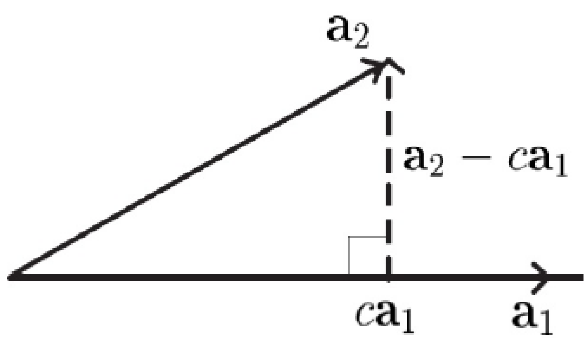
- Processing:

$$\begin{array}{c} \{\mathbf{a}_1, \ldots, \mathbf{a}_n\} \ \text{linearly independent} \\ \downarrow \\ \{\mathbf{b}_1, \ldots, \mathbf{b}_n\} \ \text{orthogonal} \\ \downarrow \\ \{\mathbf{q}_1, \ldots, \mathbf{q}_n\} \ \text{orthonormal} \end{array}$$

- If $Q = \{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ is a basis for a vector space V, it is convenient to express $\mathbf{v} \in V$ as

$$\mathbf{v} = v_1\mathbf{q}_1 + \cdots + v_n\mathbf{q}_n.$$

Process of G-S Orthogonalization (1)



- Step 1:

$$\begin{aligned}
\mathbf{b}_1 &= \mathbf{a}_1 \\
\mathbf{b}_2 &= \mathbf{a}_2 - c\mathbf{a}_1 \\
&= \mathbf{a}_2 - \mathbf{a}_1 \left(\frac{\mathbf{a}_2^T \mathbf{a}_1}{\|\mathbf{a}\|^2} \right) \\
&= \mathbf{a}_2 - \frac{\mathbf{a}_2^T \mathbf{b}_1}{\|\mathbf{b}_1\|^2} \mathbf{b}_1 .
\end{aligned}$$

- It is removing the $\mathbf{b}_1$-direction component of $\mathbf{a}_2$.

Process of G-S Orthogonalization (2)

- Step 2: Let $A = [\mathbf{b}_1 \ \mathbf{b}_2]$.

$$
\begin{aligned}
A\mathbf{x}_{\text{opt}} &= A(A^T A)^{-1} A^T \mathbf{a}_3 \\
&= [\mathbf{b}_1 \ \mathbf{b}_2] \begin{bmatrix} \frac{1}{\|\mathbf{b}_1\|^2} & 0 \\ 0 & \frac{1}{\|\mathbf{b}_2\|^2} \end{bmatrix} \begin{bmatrix} \mathbf{b}_1^T \\ \mathbf{b}_2^T \end{bmatrix} \mathbf{a}_3 \\
&= \frac{\mathbf{b}_1^T \mathbf{a}_3}{\|\mathbf{b}_1\|^2} \mathbf{b}_1 - \frac{\mathbf{b}_2^T \mathbf{a}_3}{\|\mathbf{b}_2\|^2} \mathbf{b}_2, \\
\mathbf{b}_3 &= \mathbf{a}_3 - \left(\frac{\mathbf{b}_1^T \mathbf{a}_3}{\|\mathbf{b}_1\|^2} \right) \mathbf{b}_1 - \left(\frac{\mathbf{b}_2^T \mathbf{a}_3}{\|\mathbf{b}_2\|^2} \right) \mathbf{b}_2.
\end{aligned}
$$

- Step k:

$$
\mathbf{b}_k = \mathbf{a}_k - \sum_{i=1}^{k-1} \left(\frac{\mathbf{b}_i^T \mathbf{a}_k}{\|\mathbf{b}_i\|^2} \right) \mathbf{b}_i. \qquad (1 \le k \le n)
$$

Process of G-S Orthogonalization (3)

- Normalization:

$$\mathbf{q}_i = \frac{\mathbf{b}_i}{\|\mathbf{b}_i\|}. \qquad (1 \le k \le n)$$

- Example

$$\mathbf{a}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{a}_2 = \begin{bmatrix} 3 \\ 1 \\ 3 \\ 1 \end{bmatrix}, \quad \mathbf{a}_3 = \begin{bmatrix} 3 \\ 1 \\ 1 \\ -1 \end{bmatrix}.$$

$$\Rightarrow \quad \mathbf{q}_1 = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{q}_2 = \frac{1}{2}\begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}, \quad \mathbf{q}_3 = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}.$$

Equivalent Representation for G-S Orthogonalization

- It is possible to use $\mathbf{q}_i = \frac{\mathbf{b}_i}{\|\mathbf{b}_i\|}$ in the intermediate process.

$$\mathbf{b}_k = \mathbf{a}_k - \sum_{i=1}^{k-1} \left(\mathbf{q}_i^T \mathbf{a}_k\right) \mathbf{q}_i. \quad (1 \le k \le n)$$

- Moreover, we have

$$\mathbf{a}_k = \sum_{i=1}^{k} \left(\mathbf{q}_i^T \mathbf{a}_k\right) \mathbf{q}_i. \quad (1 \le k \le n)$$

since

$$\mathbf{b}_k = \|\mathbf{b}_k\| \mathbf{q}_k = \frac{\mathbf{b}_k^T \mathbf{b}_k}{\|\mathbf{b}_k\|} \mathbf{q}_k = \mathbf{q}_k^T \mathbf{b}_k \mathbf{q}_k = \mathbf{q}_k^T \mathbf{a}_k \mathbf{q}_k.$$

QR Decomposition

- Every $m \times n$ matrix with independent columns can be factored into $A = QR$, that is,

$$
\begin{aligned}
A &= \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{bmatrix} \\
&= \underbrace{\begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_n \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} \mathbf{q}_1^T\mathbf{a}_1 & \mathbf{q}_1^T\mathbf{a}_2 & \cdots & \mathbf{q}_1^T\mathbf{a}_n \\ & \mathbf{q}_2^T\mathbf{a}_2 & \cdots & \mathbf{q}_2^T\mathbf{a}_n \\ & & \ddots & \vdots \\ & & & \mathbf{q}_n^T\mathbf{a}_n \end{bmatrix}}_{R}.
\end{aligned}
$$