Orthogonality

Dhruv Kohli

Department of Mathematics
Indian Institute of Technology, Guwahati

Outline

Motivation

- We need a basis to convert geometric calculations into algebraic calculations. An othogonal basis would make those calculations simple.
- What is the geometry of the four fundamental subspaces? It turns out that $C(A) \perp N(A^T)$ and $C(A^T) \perp N(A)$.
- If $Ax = b$ has no solution, what x should be chosen? The one which minimizes the squared error $\|Ax - b\|_2$. What is the geometric and algebraic interpretation of this least squares problem.
- How to convert any basis into orthogonal basis?
- What is the the workhorse of digital signal processing?

Orthogonal Vectors and Subspaces

- Length $\|x\|$ in $\mathbb{R}^n$ is the positive square root of $x^T x$. Proof by applying Pythagoras $n-1$ times.
- Orthogonal vectors $x^T y = 0$. Proof by applying Pythagoras on length of sides of a right angled triangle.
- The inner/scalar/dot product $x^T y = 0 \iff x \perp y$. If $x^T y > 0$ then the angle between them is < 90 and if $x^T y < 0$ then angle between them is > 90.

Result 1

If $v_1, v_2, \ldots, v_k$ are mutually orthogonal then those vectors are linearly independent.
Proof - Hints: Take dot product of $\sum_{i=1}^{k} c_i v_i = 0$ with v_j and conclude that $c_j = 0$.

Orthogonal Vectors and Subspaces contd.

Orthogonal Subspaces

Subspaces V and W are orthogonal if

$$v^T w = 0, \forall v \in V, \forall w \in W$$

OR

$$v^T w = 0, \forall v \in \text{Basis}(V), \forall w \in \text{Basis}(W)$$

- The subspace $\{0\}$ is orthogonal to all subspaces. A line can be orthogonal to a line or a plane but a plane cannot be orthogonal to a plane (are front and side walls of a room orthogonal?).

Orthogonal Vectors and Subspaces contd.

Result 2 - Fundamental Theorem of Orthogonality

For a matrix A, $C(A) \perp N(A^T)$ and $C(A^T) \perp N(A)$.

Proof - Hints: Let $x \in N(A)$ then,

$Ax = 0 \Rightarrow (\dots \ \text{row}_j \ \dots)^T x = 0 \Rightarrow \text{row}_j \perp x \Rightarrow C(A^T) \perp N(A)$

OR

Let $y = A^T x$ (L.C. of columns of A^T) and $z \in N(A)$ then,

$$y^T z = x^T A z = x^T 0 = 0 \Rightarrow C(A^T) \perp N(A)$$

Orthogonal Complement of a Subspace

Given a subspace V of $\mathbb{R}^n$. The space of all vectors orthogonal to V is called orthogonal complement of V, denoted by $V^\perp$. Also,

$$\dim V + \dim V^\perp = n$$

Orthogonal Vectors and Subspaces contd.

Result 3 - Fundamental Theorem of Linear Algebra, Part 2

Given $A_{m\times n}$, $C(A)^{\perp} = N(A^T)$ and $C(A^T)^{\perp} = N(A)$. As a result, $\dim C(A) + \dim N(A^T) = n, \dim C(A^T) + \dim N(A) = m$.

Proof - Hints: We must show the following,

$$b \in C(A) \iff y^T b = 0 \text{ whenever } y^T A = 0$$

$(\Rightarrow)$ Let $b = Ax$, then $y^T b = y^T Ax = 0x = 0$.

$(\Leftarrow)$?

Orthogonal Vectors and Subspaces contd.

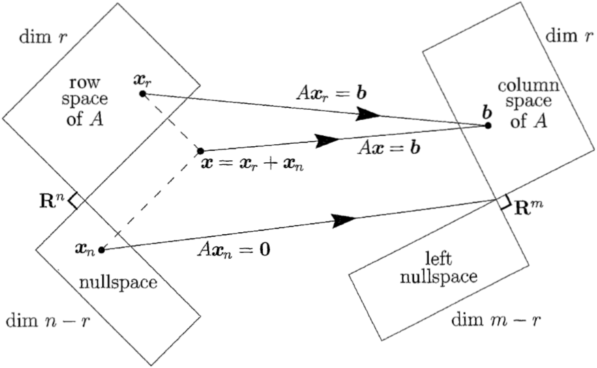

Figure: The true action $Ax = A(x_{\text{row}} + x_{\text{null}})$ of any m by n matrix.

Orthogonal Vectors and Subspaces contd.

Result 4

From the row space to the column space, A is actually invertible. Every vector b in the column space comes from exactly one vector x_r in the row space.

Proof - Hints:

$$Ax_{r_1} = b, Ax_{r_2} = b \Rightarrow A(x_{r_1} - x_{r_2}) = 0$$
$$\Rightarrow (x_{r_1} - x_{r_2}) \in N(A) \text{ and } (x_{r_1} - x_{r_2}) \in C(A^T)$$
$$\Rightarrow x_{r_1} - x_{r_2} = 0$$

- Every matrix transforms its row space onto its column space.

Cosines and Projections onto Lines

Result 5

The cosine of angle between any nonzero vectors a and b is,

$$\cos\theta = \frac{a^T b}{\|a\|\,\|b\|}$$

Proof - Hints: Proof by Law of Cosines

$$\|b-a\| = \|b\| + \|a\| - 2\,\|b\|\,\|a\|\cos\theta$$

Result 6

The projection of vector b onto the line in the direction of a is,

$$p = \hat{x}a = \frac{a^T b}{a^T a}a$$

Proof - Hints: $(b - \hat{x}a) \perp a$

Cosines and Projections onto Lines contd.

Result 7 - Schwarz inequality: $|a^T b| \leq \|a\| \, \|b\|$

Proof - Hints: $\|e\| = \|b - p\| \geq 0$ or $|\cos\theta| \leq 1$

- Equality holds iff b is a multiple of a i.e. $\theta = 0$ or 180.

Projection Matrix

From result 6, matrix that projects b to a is given by,

$$P = \frac{aa^T}{a^T a}$$

- $P = P^T$ and $P^2 = P$ (Pb already lies on the line along a).
- $C(P)$ is line through a and $N(P)$ is the plane perpendicular to a. Note: $N(P) \perp C(P)$ because $C(P) = C(P^T)$.
- $\text{Rank}(P) = 1$ (Why?).

Projection and Least Squares

- For system $A_{m\times n}x = b$, if number m of observations (rows) is larger than the number n of unknowns, it must be expected that $Ax = b$ will be inconsistent.
- Probably, there will not exist a choice of x that perfectly fits data b. In other words, b probably will not be in $C(A)$.
- The problem reduces to finding $\hat{x}$ that minimizes error $E = \|Ax - b\|$. This is exactly the distance between b and the point Ax in the column space.
- Need to locate $p = A\hat{x}$ that is closer to b than any other point in C(A). The error vector $e = b - A\hat{x}$ must be perpendicular to $C(A)$ i.e. must lie in $N(A^T)$.

$$A^T(A\hat{x} - b) = 0 \Rightarrow A^TA\hat{x} = A^Tb$$

Projection and Least Squares contd.

- Calculus way to prove is by taking derivative of $(Ax - b)^T(Ax - b)$ wrt x and equating to 0.

Least Squares Problems with Several Variables

When $Ax = b$ is inconsistent, its least-squares solution minimizes $\|Ax - b\|^2$:

$$A^T A\hat{x} = A^T b$$

$A^T A$ is invertible $\iff$ the columns of A are linearly independent[1]. Then,

$$\hat{x} = (A^T A)^{-1} A^T b$$

The projection of b onto the $C(A)$ is the nearest point $A\hat{x}$:

$$p = A\hat{x} = A(A^T A)^{-1} A^T b$$

[1] $\text{rank}(A) + \text{rank}(B) - n \leq \text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$

Projection and Least Squares contd.

- If $b \in C(A)$, $(b = Ay)$, then $p = A(A^TA)^{-1}A^Tb = Ay$.
- If $b \in N(A^T)$, then, $p = 0$.
- If A is invertible, then, $p = b$.

Result 8

The cross product matrix A^TA has same null space as A.
Proof - Hints:

$$A^TAx = 0 \Rightarrow x^TA^TAx = x^T0 = 0 \Rightarrow \|Ax\| = 0 \Rightarrow Ax = 0$$

Result 9

If $A_{m\times n}$ has independent columns then A^TA is square, symmetric, invertible and positive definite.
Proof - Hints: $\text{Rank}(A^TA) = n$.

Projection and Least Squares contd.

Result 10

Matrix $P = A(A^T A)^{-1} A^T$ projects onto $C(A)$ and $I - P$ projects onto $N(A^T)$. Two properties:

1. $P = P^T$
2. $P^2 = P$

Also, any matrix with above properties is a projection matrix. *Proof - Hints*: For converse, show that Pb is the projection of b in $C(P)$ or $(I - P)b$ is the projection of b onto a space orthogonal to $C(P)$. For a general vector Pc in $C(P)$, the dot product of it with $(I - P)b$ is,

$$(Pc)^T(I - P)b = c^T(P^T - P^T P)b = c^T(P - P^2)b = 0$$

Therefore, $(I - P)b$ is in a space orthogonal to $C(P)$, and Pb is the projection of b in $C(P)$.

Projection and Least Squares contd.
Weighted Least Squares

- Idea: some observations are more reliable than others. The new error looks like $E^2 = \sum_{i=1}^{m} w_i^2(a_{r_i}^T x_i - b_i)^2$.
- The solution $\hat{x_w}$ minimizes this error and is a solution of the system $WAx = Wb$.

Weighted Normal Equation

The least square solution to $WAX = Wb$ is $\hat{x_w}$:

$$A^T W^T WA\hat{x_w} = A^T W^T Wb$$

- The point $A\hat{x_w}$ still point in $C(A)$ that is closest to b. But the term "closest" has new meaning - all inner products $a^T b$ are replaced by $(Wa)^T(Wb) = a^T W^T Wb$. In this new sense, $A\hat{x_w} \perp b - A\hat{x_w}$.

Projection and Least Squares contd.

Weighted Least Squares

- Note: if W is orthogonal then $W^T W = I$.
- The important question is the choice of $C = W^T W$. The best answer comes from the statisticians:
  1. If the errors in b_i are independent of each other then the right weights are $w_i = \frac{1}{\sigma_i}$ where σ_i is the variance of the error in b_i. Higher the variance, lesser is the reliability and hence lesser is the weight.
  2. If the errors in b_i are coupled then the best unbiased matrix C is the inverse of the covariance matrix - whose i, j entry is the expected value of (error in b_i) times (error in b_j).

Orthogonal Basis and Gram-Schmidt

Orthonormal Vectors

The vectors $q_1, q_2, \ldots, q_n$ are orthonormal if

$$q_i^T q_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

Orthogonal Matrices

If Q has orthonormal columns, then $Q^T Q = I$. If Q is square, then it is called orthogonal matrix and $Q^T = Q^{-1}$.

- $Q^T Q = I$ even if Q is a rectangular matrix but then Q^T is only a left-inverse.
- Examples of orthogonal matrices - rotation matrix, permultation matrix, reflection matrix. **Every orthogonal matrix is a product of a rotation and a reflection**.

Orthogonal Basis and Gram-Schmidt contd.

- For square Q, $Q^T = Q^{-1} \Rightarrow QQ^T = I$ which means that **the rows of a square matrix are orthonormal whenever the columns are** even though the rows and columns point in a completely different direction.
- For square Q of size n, the columns span $\mathbb{R}^n$ so as rows.

Result 11

Multiplication by any Q (square or rectanlge) preserves length and inner produt.
Proof - Hints: $x^T Q^T Q y = x^T y$.

- All inner products and lengths are preserved when the space is rotated or reflected.

Orthogonal Basis and Gram-Schmidt contd.

Result 12

If vectors $q_1, q_2, \ldots, q_n \in \mathbb{R}^n$ form an orthonormal basis of $\mathbb{R}^n$ (or Q is orthogonal) then every $b \in \mathbb{R}^n$ can be written as:

$$b = \sum_{i=1}^{k} (q_i^T b) q_i \text{ or } b = Q(Q^T b)$$

Proof - Hints:

$$b = \sum_{i=1}^{n} x_i q_i \Rightarrow q_j^T b = q_j^T (\sum_{i=1}^{n} x_i q_i) \Rightarrow x_j = q_j^T b$$

OR

$$Qx = b \Rightarrow x = Q^T b \Rightarrow b = QQ^T b$$

- Every b is a sum of its one-dimesnional projections onto the lines through q's $\left(\frac{q_i^T b}{q_i^T q_i} q_i = (q_i^T b) q_i\right)$.

Orthogonal Basis and Gram-Schmidt contd.

Least Squares with Orthogonal Columns

If $Q_{m\times n}$ has orthonormal columns, the least-squares solution is:

$$Qx = b, \text{ rectangular system with no solutions for most } b$$

$$Q^TQ\hat{x} = Q^Tb, \text{ normal equation for best } \hat{x} \text{ - } Q^Q = I$$

$$\hat{x} = Q^Tb, \ \hat{x}_i = q_i^Tb$$

$$p = Q\hat{x}, \text{ projection of } b \text{ is } (q_1^Tb)q_1 + \ldots + (q_n^Tb)q_n$$

$$p = QQ^Tb, \text{ the projection matrix is } QQ^T$$

- $m = n \Rightarrow p = b$ and $m > n \Rightarrow p$ may or may not equal b.
- For $Ax = b$, $P = A(A^TA^{-1})A^T \xrightarrow{A=Q} P = QIQ^T = QQ^T$.
- P projects $q \in C(Q)$ to q and $q' \in N(Q^T)$ to 0 (Why)?

Orthogonal Basis and Gram-Schmidt contd.

Gram-Schmidt Orthogonalization

Input: independent vectors $a_1, a_2, \ldots, a_n$.
Output: orthonormal vectors $q_1, q_2, \ldots, q_n$.
At step j, it subtracts from a_j its components in the directions of $q_1, q_2, \ldots, q_{j-1}$ that are already settled:

$$A_j = a_j - (q_1^T a_j)q_1 - (q_2^T a_j)q_2 - \ldots - (q_{j-1}^T a_j)q_{j-1}$$

$$q_j = \frac{A_j}{\|A_j\|}$$

- A_j's may be normalized at the end without affecting the resulting q's (Why?).

Orthogonal Basis and Gram-Schmidt contd.

QR Factorization

Based on Gram-Schmidt Orthogonalization, every $m \times n$ matrix with independent columns can be factored - $A = Q_{m\times n}R_{n\times n}$. The columns of Q are orthonormal and R is upper triangular and invertible given by:

$$R_j = \begin{bmatrix} q_1^T a_j & q_2^T a_j & . & q_j^T a_j & 0 & . & 0 \end{bmatrix}^T \Rightarrow R_{ij} = q_i^T a_j$$

Note: a_j has no component in the direction of $q_{j+1}, \ldots, q_n$.

Least Squares using QR Factorization

If the columns of A are independent then $A = QR$ and

$$A^T A = R^T Q^T QR = R^T R, \; A^T b = R^T Q^T b$$

$$A^T A\hat{x} = A^T b \Rightarrow R^T R\hat{x} = R^T Q^T b \Rightarrow R\hat{x} = Q^T b$$

Note: Computational cost is mn^2 operations of Gram Schmidt.

Hilbert Space and Function Space

- All vectors in $\mathbb{R}^\infty$ which have finite length form a vector space called **Hilbert space**.
- A function defined on an interval can be imagined as a vector with a whole continuum of components. All those functions that have a finite length form **function space**.
- The inner product of f and g defined on $[a, b]$ and $[c, d]$ respectively, is defined in an analogous way as:

$$(f, g) = \int_{[a,b]\cap[c,d]} f(x)g(x)dx \text{ and } (f, f) = \int_{[a,b]} f(x)^2 dx$$

- Orthogonality condition - $v^T w = 0, (f, g) = 0$. Schwarz inequality - $|(f, g)| \leq \|f\| \, \|g\|$, $(f, f) = \|f\|^2$ (Why?).

Orthogonal Basis and Gram-Schmidt contd.

Function Spaces and Fourier Series

- Note: $\sin x$ and $\cos x, x \in [0, 2\pi]$ are orthogonal.

Fourier Series

$(*)$ *sines and cosines defined on* $[0, 2\pi]$ *are mutually orthogonal.*
Fourier series of $f(x)$ is its expansion into sines and cosines:

$$f(x) = a_0 + a_1 \cos x + b_1 \sin x + a_2 \cos 2x + b_2 \sin 2x + \dots$$

$$a_0 = \frac{(f, 1)}{(1, 1)}, a_k = \frac{(f, \cos kx)}{(\cos kx, \cos kx)}, b_k = \frac{(f, \sin kx)}{(\sin kx, \sin kx)}, k \neq 0$$

- Inner products are computed over $[0, 2\pi]$.
- Those coefficients are obtained by using $(*)$.
- Fourier series is projecting $f(x)$ onto orthogonal sines and cosines. *It gives the coordinates of the "vector" $f(x)$ with respect to a set of (infinitely many) perpendicular axes.*

- Suppose an approximation of a function $f(x)$ is required as a linear combination of $g_1(x), \ldots, g_k(x)$. For example, $f(x)$ is to be approximated with the closest polynomial of degree 2 i.e. linear combination of $\{1, x, x^2\}$ on $[0, 1]$.
- Since 1 and x^2 are never orthogonal, $f(x)$ cannot be written as a sum of its projections on $1, x$ and x^2.
- It is virtually hopeless to solve following for 10 degrees:

  $Ay = b$ where $A = [1, x, x^2], y = [y_1, y_2, y_3]^T, b = [f(x)]$

$$A^T A = \begin{bmatrix} (1,1) & (1,x) & (1,x^2) \\ (x,1) & (x,x) & (x,x^2) \\ (x^2,1) & (x^2,x) & (x^2,x^2) \end{bmatrix} = \begin{bmatrix} 1 & 1/2 & 1/1 \\ 1/2 & 13 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix}$$

- $A^T A$ (called Hilbert Matrix) is ill-conditioned - Gaussian Elimination amplifies roundoff error by 10^{13}. The right idea is to switch to orthogonal axis by Gram-Schmidt.

Gram-Schmidt for Functions

The process is same as the Gram-Schmidt for vectors except that the inner products will be those of functions.

- Example: Consider the functions $1, x, x^2$ defined on $[-1, 1]$ (it is easier to work with symmetric intervals).
- G-S process can start by accepting $v_1 = 1$ and $v_2 = x$ as first two perpendicular axes (because odd powers are perpendicular to even powers on symmetric interval.)
- $v_3 = x^2 - \frac{(1,x^2)}{(1,1)} - \frac{(x,x^2)}{(x,x)} = x^2 - \frac{1}{3}$ will then be third axis perpendicular to v_1 and v_2.
- The polynomials constructed in this way are called **Legendre Polynomials** and they are orthogonal to each other on the interval $[-1, 1]$.

Bibliography

Gilbert Strang. *Linear algebra and its applications*. Belmont, CA: Thomson, Brooks/Cole, 2006.