1 Orthogonality and Least Sqaures

1.1 Inner Product, Length, and Orthogonality

Inner Product (generalization of the dot product of vectors in $\mathbf{R}^n$)

Inner Product/Dot Product of vectors in $\mathbf{R}^n$: $\vec{u} \cdot \vec{v} = \vec{u}^T \vec{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n$

Theorem 1.1

Let $\mathbf{u}, \mathbf{v}, \mathbf{w}$ be vectors, let c be a scalar, then an inner product is a function assigns a scalar to each pair of vector $\mathbf{u}$ and $\mathbf{v}$ satisfies:

- $\vec{u} \cdot \vec{v} = \vec{v} \cdot \vec{u}$
- $(\vec{u} + \vec{v}) \cdot \vec{w} = \vec{u} \cdot \vec{w} + \vec{v} \cdot \vec{w}$
- $(c\vec{u}) \cdot \vec{v} = c(\vec{u} \cdot \vec{v}) = \vec{u} \cdot (c\vec{v})$
- $\vec{u} \cdot \vec{u} = 0$ iff $\vec{u} = \vec{0}$

The Length/Norm of a vector of $\mathbf{v}$ is the nonnegative scalar $||\vec{v}|| = \sqrt{\vec{v} \cdot \vec{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$ and $||\vec{v}||^2 = \vec{v} \cdot \vec{v}$

- A vector whose length is 1 is called a unit vector. If we divide a nonzero vector $\mathbf{v}$ by its length, that is, multiply by $1/||\mathbf{v}||$, we obtain a unit vector this process is called normalizing (direction is preserved)

Distance in $\mathbf{R}^n$: for $\mathbf{u}$ and $\mathbf{v}$ in $\mathbf{R}^n$, the distance between u and v, written dist$(\mathbf{u}, \mathbf{v})$ is the length of the vector $\mathbf{u} - \mathbf{v}$, that is dist$(\mathbf{u}, \mathbf{v}) = ||\mathbf{u} - \mathbf{v}||$

- In $\mathbf{R}^2$ and $\mathbf{R}^3$ this definition coincides with the usual formulas for Euclidean distance between 2 points

Orthogonality of vectors in $\mathbf{R}^n$ is the generalization of the concept of perpendicular lines in ordinary Euclidean geometry.

Def: two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbf{R}^n$ are orthogonal (to each other) if $\mathbf{u} \cdot \mathbf{v} = 0$.

Note the zero vector is ortohgonal to every vector in $\mathbf{R}^n$

Pythagorean theorem: 2 vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if and only if $||\mathbf{u} + \mathbf{v}||^2 = ||\mathbf{u}||^2 + ||\mathbf{v}||^2$.

Orthogonal Complements: used in SVD

If a vector $\mathbf{z}$ is orthogonal to every vector in a subspace W of $\mathbf{R}^n$, then $\mathbf{z}$ is said to be ortohgonal to W

The set of all vectors $\mathbf{z}$ that are orthogonal to W is called the ortohgonal complement of W denoted $W^\perp$ read as "W perpendicular" or "W perp"

Properties/facts

- a vector $\mathbf{x}$ is in W perp if and only if $\mathbf{x}$ is orthogonal to every vector in a set that spans W
- W perp is a subspace of $\mathbf{R}^n$

Theorem 1.2

Let A be an $m \times n$ matrix. The orthogonal complement of the row space of A is the null space of A, and the orthogonal complement of the column space of A is the null space of A^T. (Row $A)^\perp =$ Nul A and (Col $A)^\perp =$ Nul A^T.

1.2 Orthogonal Sets

Orthogonal Sets: a set of vectors $\{\mathbf{u}_1, \mathbf{u}_2, dots, \mathbf{u}_p\}$ in $\mathbf{R}^n$ is said to be an orthogonal set if each pair of distinct vectors from theset is orthogonal.

Theorem 1.3

If $S = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p\}$ is an orthogonal set of non-zero vectors in $\mathbf{R}^n$, then S is linearly independent and hence is a basis for the subspace spanned by S.

Definition

An Orthogonal Basis for a subspace W of $\mathbf{R}^n$ is a basis for W that is also an orthogonal set.

Theorem 1.4

Let $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p\}$ be an orthogonal basis for a subspace W of $\mathbf{R}^n$. For each vector $\mathbf{y}$ in W, the weights in the linear combination $\mathbf{y} = c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + \cdots + c_p\mathbf{u}_p$ are:

$$c_j = \frac{\vec{v} \cdot \vec{u}_j}{\vec{u}_j \cdot \vec{u}_j} \qquad j = 1, 2, \ldots, p$$

This formula is why an orthogonal basis is much nicer than others.

Orthogonal projection: given a nonzero vector $\mathbf{u}$ in $\mathbf{R}^n$, consider the problem of decomposing a vector $\mathbf{y}$ in $\mathbf{R}^n$ into the sum of two vectors, one a multiple of $\mathbf{u}$ and the other orthogonal to $\mathbf{u}$.

$\hat{y}$ ios the orthogonal projection of $\mathbf{y}$ onto $\mathbf{u}$, the vector $|textbfz$ is the component of $\mathbf{y}$ orthogonal to $\mathbf{u}$.

Geometric Interpreation THeorem for finding coordinates of an orthogonal basis: The theorem decomposes vector $\mathbf{y}$ into a sum of orthogonal productions onto one-dimensiona subspaces.

Decomposing a Force into Component Forces: occurs in physics

Orthonormal Sets: a set $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p\}$ is an orthonormal set if it is an orthogonal set of unit vectors. If W is the subspace spanned by such a set, then $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p\}$ is an orthonormal basis for W¡ since the set is automatically linearly independent.

Matrices whose columns form an orthonormal set are important in applications and computer algorithms for matrix computations. Their main properties are given in the following two theorems:

Theorem 1.5

An $m \times n$ matrix U has orthonormal columns if and only if $U^TU = I$.

Theorem 1.6

Let U be an $m \times n$ matrix with orthonormal columns, and let $\mathbf{x}$ and $\mathbf{y}$ be in R^n then:

- $||U\vec{x}|| = ||\vec{x}||$
- $(U\vec{x}) \cdot (U\vec{y}) = \vec{x} \cdot \vec{y}$
- $(U\vec{x}) \cdot (U\vec{y}) = 0$ iff $\vec{x} \cdot \vec{y} = 0$

An orthogonal matrix is a square invertible matrix U, $U^{-1} = U^T$.

1.3 Orthogonal Projections

The Orthogonal Projection: Given a vector $\mathbf{y}$ and a subspace W in $\mathbf{R}^n$ there is a vector $\hat{\mathbf{y}}$ in W such that

- $\hat{\mathbf{y}}$ is the unique vector in W closest to $\mathbf{y}$
- $\hat{\mathbf{y}}$ is the unique vector for which $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to W.

These two properties of $\vec{\mathbf{y}}$ provide the key to finding the least squares solution to linear systems.

Theorem 1.7

Let W be a subspace of $\mathbf{R}^n$. Then each $\mathbf{y}$ in $\mathbf{R}^n$ can be written uniquely in the form: $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{z}$ where $\vec{\mathbf{y}}$ is in W and z is in $W^{\perp}$. In fact, if $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p\}$ is any orthogonal basis of W, then

$$\hat{y} = \frac{\vec{v} \cdot \vec{u}_1}{\vec{u}_1 \cdot \vec{u}_1}\vec{u}_1 + \cdots + \frac{\vec{v} \cdot \vec{u}_p}{\vec{u}_p \cdot \vec{u}_p}\vec{u}_p \text{ and } \vec{z} = \vec{y} \cdot \hat{y}$$

The theorem tells us decomposition of $\mathbf{y} = \mathbf{z}_1 + \mathbf{z}_2$ can be computed without having an orthogonal basis for $\mathbf{R}^n$. It is enough to haev an orthogonal basis only for W.

Geometric Interpretation of the Orthogonal Projection: the orthogonal projection $\hat{y}$ of $\mathbf{y}$ onto W is the sum of the projections of $\mathbf{y}$ onto one-dimensional subspaces that are orthogonal to each other.

Properties of Orthogonal Projections

Theorem 1.8

Let W be a subspace of $\mathbf{R}^n$, let $\mathbf{y}$ be any vector in $\mathbf{R}^n$ and $\hat{y}$ be the orthogonal projection of $\mathbf{y}$ onto W. Then $\hat{y}$ is the closest point in W to $\mathbf{y}$ in the sense that $||\mathbf{y} - \hat{\mathbf{y}}|| < ||\mathbf{y} - \mathbf{v}||$ for all $\mathbf{v}$ in W distinct from $\hat{\mathbf{y}}$.

Theorem 1.9

If $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p\}$ is an orthonormal basis for a subspace W of $\mathbf{R}^n$, then $\text{proj}_w\mathbf{y} = (\vec{y} \cdot \vec{u}_1)\vec{u}_1 + (\vec{y} \cdot \vec{u}_2)\vec{u}_2 + \cdots + (\vec{y} \cdot \vec{u}_p)\vec{u}_p$.

If $U = [\mathbf{u}_1\mathbf{u}_2 \ldots \mathbf{u}_p]$ then $\text{proj}_w\mathbf{y} = UU^T\mathbf{y}$ for all $\mathbf{y}$ in $\mathbf{R}^n$.

1.4 The Gram-Schmidt Process

The Gram-Schmidt process is a simple algorithm for producing an orthogonal basis for any nonzero subspace of $\mathbf{R}^n$.

Theorem 1.10

Given a basis $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p\}$ for a nonzero subspace of $\mathbf{R}^n$, define:

- $\vec{v}_1 = \vec{x}_1$
- $\vec{v}_2 = \vec{x}_2 - \frac{\vec{x}_2 \cdot \vec{v}_1}{\vec{v}_1 \cdot \vec{v}_1}\vec{v}_1$
- $\vec{v}_3 = \vec{x}_3 - \frac{\vec{x}_3 \cdot \vec{v}_1}{\vec{v}_1 \cdot \vec{v}_1}\vec{v}_1 - \frac{\vec{x}_2 \cdot \vec{v}_2}{\vec{v}_2 \cdot \vec{v}_2}\vec{v}_2$

Orthonormal Basis: when working problems by hand, it is easier to normalize each $\mathbf{v}_k$ as they are foud.

QR Factorization of Matrices: if an $m \times n$ matrix A has linearly independent columns $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ then applying the Gram-Schmidt process with normalizations to $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$ amounts to factoring A as described in the following theorem and is used widely in computer algorithms.

Theorem 1.11

If A is an $m \times n$ matrix with linearly independent columns then A can be factored as $A = QR$ where Q is an $m \times n$ matrix whose columns form an orthonormal basis for Col A and R is an $n \times n$ upper triangular matrix with positive entities on the diagonal.

When the Gram Schmidt process is run on the computer, a round off error can build up as the vectors are calculated, one by one. For j and k large but unequal, the innter product may not be sufficiently close to zero. A different computer based QR factorization is usually preferred to the modified Graham Schmidt Method because it yields a more accurate orthogonal basis, even though the factorization requires about twice as much arithmetic.

1.5 Least-Squares Problems

The Least Squares Problem: given $A\mathbf{x} = \mathbf{b}$ that is possibly inconsistent, find an $\mathbf{x}$ that makes $||\mathbf{b} - A\mathbf{x}||$ as small as possible.

Definition

If A is an $m \times n$ matrix and $\mathbf{b}$ is in $\mathbf{R}^n$, a least-squares solution of $A\mathbf{x} = \mathbf{b}$ is $\mathbf{x}$ in $\mathbf{R}^n$ such that $||\mathbf{b} - A\mathbf{x}|| \leq ||\mathbf{b} - A\mathbf{x}'||$ for all $\mathbf{x}'$ in $\mathbf{R}^n$

Notice: $A\mathbf{x}$ is in the column space of A, Col A, so we seek an $\mathbf{x}$ that makes $A\mathbf{x}$ the closest point to in Col A to $\mathbf{b}$.

Theorem 1.12

The set of least squares solutions of $A\mathbf{x} = \mathbf{b}$ coincides with the nonempty set of solutions of the normal equation $A^T A\hat{\mathbf{x}} = A^T\hat{\mathbf{b}}$

Note: if there is a free variable, the least squares solution may not be unique

Deriving the Normal Equations for $A\mathbf{x} = \mathbf{b}$:

1. Note $A\mathbf{x}$ is in the Col A, therefore the $\mathbf{b}$ associated least squares solution of $A\mathbf{x} = \mathbf{b}$ is in Col A
2. Use the Best Approximation Theorem to determine the solution of the Least Square Problem is the orthogonal projection of $\mathbf{b}$ onto Col A, $\hat{b} = \text{proj}_{\text{Col}A}\mathbf{b}$
3. Note: for the Least Squares Problem: we are looking for $\hat{\mathbf{x}}$ that satisfies $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$
4. use the Orthogonal Decomposition Theorem to find the vector $\mathbf{z}$ orthogonal to $\hat{\mathbf{b}}$, $\mathbf{z} = \mathbf{b} - \hat{\mathbf{b}}$
5. use the fact that $\mathbf{z}$ must be orthogonal to the Col A, and therefore any column vector of A, $\mathbf{a}_i$ is orthogonal to $\mathbf{z} = \mathbf{b} - \hat{\mathbf{b}} = \mathbf{b} - A\hat{\mathbf{x}} \implies a_i \cdot (\mathbf{b} - A\hat{\mathbf{x}} = 0)$

Theorem 1.13

Let A be an $m \times n$ matrix. The following statements are logically equivalent

1. the equation $A\mathbf{x} = \mathbf{b}$ has a unique least-square solution for each $\mathbf{b}$ in $\mathbf{R}^n$
2. the columns of A are linearly independent
3. the matrix $A^T A$ is invertible

When these statements are true, the least squares solution $\hat{\mathbf{x}}$ is given by $\hat{\mathbf{x}} = (A^T A)^{-1} A^T\mathbf{b}$

1.6 Machine Learning and Linear Models

Machine learning uses linear models in sitautions where the machine is being trained to predict the outcome based on the values of hte inputs.

The machine is given a set of training data where the values of the independent and dependent variables are known.

The least squares line is the line $y = \beta_0 + \beta_1 x$ that minimizes the sum of the squares of the residuals.

This line is also called a line of regression of y on x. The coefficients β_0, β_1 of the line are called regression coefficients.

In general a linear model will arise whenever y is to be predicted by an equation of the form

$$y = \beta_0 f(0)(u, v) + \beta_1 f_1(u, v) + \cdots + \beta_k f_k(u, v)$$

with $f_0, \ldots, f_k$ any sort of known functions and $\beta_0, \beta_1, \ldots, \beta_k$ unknown weights.

1.7 Inner Product Spaces

Definition

An inner product on a vector space V is a function that, to each pair of vectors **u** and **v** in V, associated a real number $\langle \mathbf{u}, \mathbf{v} \rangle$ and satisfies the following axioms, for all **u**, **v**, and **w** in V and all scalars c:

1. $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$
2. $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u} + \mathbf{w} \rangle + \langle \mathbf{v} + \mathbf{w} \rangle$
3. $\langle c\mathbf{u}, \mathbf{v} \rangle = c\langle \mathbf{u}, \mathbf{v} \rangle$
4. $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$ and $\langle \mathbf{u}, \mathbf{u} \rangle = 0$ if and only if $\mathbf{u} = \mathbf{0}$

A vector space with an innter product is called an inner product space.