MATH7502

The abstract least squares problem

Phil Isaac
1 Vector spaces

1.1 Definition

Let $V$ be a nonempty set on which are defined operations "+" (addition) and "," (scalar multiplication). $V$ is a vector space (over $\mathbb{R}$) if the following hold for all $u, v, w \in V$ and all $k, \ell \in F$:

(V1) $u + v \in V$ (closure)

(V2) $u + v = v + u$ (additive commutativity)

(V3) $u + (v + w) = (u + v) + w$ (additive associativity)

(V4) $\exists 0 \in V$ such that $u + 0 = u$ (zero vector, or additive identity)

(V5) For each $u \in V$, $\exists (-u) \in V$ such that $u + (-u) = 0$ (additive inverse)

(V6) $k \cdot u \in V$

(V7) $k \cdot (u + v) = k \cdot u + k \cdot v$ (multiplicative-additive distributivity)

(V8) $(k + \ell) \cdot u = k \cdot u + \ell \cdot u$ (additive-multiplicative distributivity)

(V9) $k \cdot (\ell \cdot u) = (k\ell) \cdot u$ (multiplicative-multiplicative distributivity)

(V10) $1 \cdot u = u$ (multiplicative identity)

The scalar multiplication symbol is often omitted. Elements of a vector space are usually called vectors.
1.2 Examples

- $\mathbb{R}^n$ – set of $n$-tuples

- $M_{m,n}(\mathbb{R})$ – set of $m \times n$ matrices

- $C[a, b]$ – set of continuous real-valued functions on $[a, b]$

- $P_n(\mathbb{R})$ – set of polynomials of degree at most $n$

- Set of solutions to a homogeneous linear ODE
2 Real inner product spaces

2.1 Dot product

The familiar dot product, $\mathbf{u} \cdot \mathbf{v}$, of the two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$ is given by

$$\mathbf{u} \cdot \mathbf{v} = u_1v_1 + u_2v_2 + u_3v_3;$$

where $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$. This is readily generalised to $\mathbb{R}^n$ by

$$\mathbf{u} \cdot \mathbf{v} = u_1v_1 + \ldots + u_nv_n, \quad \text{where} \quad \mathbf{u} = (u_1, \ldots, u_n), \quad \mathbf{v} = (v_1, \ldots, v_n).$$

This dot product has the following key properties:

$\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$

$(\mathbf{u} + \mathbf{v}) \cdot \mathbf{w} = \mathbf{u} \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w}$

$(k\mathbf{u}) \cdot \mathbf{v} = k(\mathbf{u} \cdot \mathbf{v})$

$\mathbf{u} \cdot \mathbf{u} \geq 0$

$\mathbf{u} \cdot \mathbf{u} = 0$ iff $\mathbf{u} = \mathbf{0}$

2.2 Inner product space

Inspired by the dot product on $\mathbb{R}^n$, we define a so-called inner product on a general real vector space by elevating the key properties of the dot product to axioms.

An inner product on $V$ is a function that takes each ordered pair $(\mathbf{u}, \mathbf{v})$ of elements of $V$ to a real number, denoted by $\langle \mathbf{u}, \mathbf{v} \rangle$, such that for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and all $k \in \mathbb{R}$:

(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) $\langle k\mathbf{u}, \mathbf{v} \rangle = k\langle \mathbf{u}, \mathbf{v} \rangle$

(4) $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$

(5) $\langle \mathbf{u}, \mathbf{u} \rangle = 0$ iff $\mathbf{u} = \mathbf{0}$ (where $\mathbf{0}$ is the unique zero vector)

A vector space with an inner product associated to it is called an inner product space.
2.3 Examples

- $M_{m,n}(\mathbb{R})$: $\langle u, v \rangle = \text{tr}(v^Tu)$

- $C[a, b]$: $\langle u, v \rangle = \int_a^b u(x)v(x) \, dx$

- $P_n(\mathbb{R})$:

\[
p(x) = p_0 + p_1 x + \cdots + p_n x^n
\]
\[
q(x) = q_0 + q_1 x + \cdots + q_n x^n
\]

$\Rightarrow \langle p, q \rangle = p_0q_0 + p_1q_1 + \cdots + p_nq_n$

- Set of solutions to a (second order) homogeneous linear ODE:

$\langle f, g \rangle = f(0)g(0) + f'(0)g'(0)$
3  Magnitude and direction

The norm (or magnitude or length) of an element $\mathbf{v} = (v_1, \ldots, v_n)$ of $\mathbb{R}^n$ is given by the familiar expression

$$||\mathbf{v}|| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{v_1^2 + \ldots + v_n^2}.$$

There is a similar notion for any real inner product space $V$. The norm of a vector $\mathbf{v} \in V$, denoted by $||\mathbf{v}||$, is thus defined by

$$||\mathbf{v}|| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}.$$

A vector with norm 1 is called a unit vector.

How would we define the distance, $d(\mathbf{u}, \mathbf{v})$, between two vectors $\mathbf{u}, \mathbf{v} \in V$? A natural notion of distance between two vectors should be independent of the order we happen to be viewing them. That is, we want the distance measure to be symmetric: $d(\mathbf{u}, \mathbf{v}) = d(\mathbf{v}, \mathbf{u})$. Again using $\mathbb{R}^n$ as inspiration, we now define the distance between two vectors $\mathbf{u}, \mathbf{v} \in V$ as

$$d(\mathbf{u}, \mathbf{v}) = ||\mathbf{u} - \mathbf{v}||.$$

Note that the notions of norm and distance are relative to the inner product used!

We may also talk about the angle between two vectors in an inner product space:

$$\theta = \cos^{-1} \left( \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{u}|| \ ||\mathbf{v}||} \right).$$

3.1 Orthogonality and Pythagoras

As in $\mathbb{R}^n$ with inner product given by the usual dot product, we say that two vectors $\mathbf{u}, \mathbf{v} \in V$ are orthogonal if

$$\langle \mathbf{u}, \mathbf{v} \rangle = 0.$$

There is an analogue of Pythagoras' Theorem for inner product spaces:

$$||\mathbf{u} + \mathbf{v}||^2 = ||\mathbf{u}||^2 + ||\mathbf{v}||^2 \quad \iff \quad \langle \mathbf{u}, \mathbf{v} \rangle = 0.$$
3.2 Orthogonal complement

Let $U$ be a subset of the real inner product space $V$. The orthogonal complement of $U$, denoted by $U^\perp$, is the set of all vectors in $V$ that are orthogonal to every vector in $U$. That is,

$$U^\perp = \{ v \in V | \langle v, u \rangle = 0 \text{ for every } u \in U \}.$$  

This is a vector space with addition and scalar multiplication inherited from $V$.

3.3 Orthogonal set

Let $V$ be a real inner product space. A nonempty set of vectors in $V$ is orthogonal if each vector in the set is orthogonal to all the other vectors in the set. That is, the set $\{v_1, \ldots, v_n\} \subseteq V$ is orthogonal if

$$\langle v_i, v_j \rangle = 0, \quad i \neq j.$$

3.4 Orthonormal basis

An orthogonal set of vectors in $V$ is called orthonormal if all the vectors in the set are unit vectors. That is, the set $\{e_1, \ldots, e_n\} \subseteq V$ is orthonormal if

$$\langle e_i, e_j \rangle = \delta_{i,j},$$

where the Kronecker delta is defined by

$$\delta_{i,j} = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$
3.5 Orthogonal projection

Let $U$ be a finite-dimensional subspace of the real inner product space $V$. Then, each $v \in V$ can be written in a unique way as

$$v = u + w, \quad u \in U, \quad w \in U^\perp.$$

In the proof, we will assume that $U$ has an orthonormal basis $S = \{e_1, \ldots, e_k\}$.

The vector $u \in U$ is called the orthogonal projection of $v$ onto $U$ and is given by

$$\text{Proj}_U(v) = \langle v, e_1 \rangle e_1 + \cdots + \langle v, e_k \rangle e_k.$$

Likewise, the vector $w \in U^\perp$ is called the orthogonal projection of $v$ onto $U^\perp$ and is given by

$$\text{Proj}_{U^\perp}(v) = v - \text{Proj}_U(v).$$
3.6 Gram-Schmidt process

Let $\beta = \{v_1, \ldots, v_n\}$ be a linearly independent set of vectors in an inner product space $V$. The following algorithm, called the Gram-Schmidt process, converts $\beta$ into an orthonormal set.

Step 1: Set $e_1 = \frac{v_1}{\|v_1\|}$

Step $i + 1$: Let $U_i = \text{span}\{e_1, \ldots, e_i\}$.
- Set $w_{i+1} = v_{i+1} - \text{Proj}_{U_i}(v_{i+1})$
- $\Rightarrow w_{i+1} \in U_i^\perp$, $w_{i+1} \neq 0$.
- Set $e_{i+1} = \frac{w_{i+1}}{\|w_{i+1}\|}$

Outcome: $\{e_1, \ldots, e_n\}$ is an orthonormal set.
4 Least squares problem - minimising distance to a subspace

A recurring problem in linear algebra, and in its myriad of applications, is the following:

- Given a vector \( \mathbf{v} \) in a real inner product space \( V \), give the best approximation to \( \mathbf{v} \) in a finite-dimensional subspace \( U \) of \( V \).

**Question:** What do we mean by "best approximation"?

**Answer:** Seek \( \mathbf{u} \in U \) that minimises \( ||\mathbf{v} - \mathbf{u}|| \). Equivalently, find a vector in a subspace (for example, corresponding to a point on a plane in \( \mathbb{R}^3 \)), of minimal distance to a given vector in the ambient vector space (in this example, corresponding to a point in \( \mathbb{R}^3 \)). Concretely, let \( \mathbf{v} \in V \). Then, the problem is to

\[
\text{find } \mathbf{u} \in U \text{ such that } d(\mathbf{u}, \mathbf{v}) \text{ is as small as possible.}
\]

This problem is called the "least squares problem."

**Theorem** (Best Approximation Theorem). If \( U \) is a finite-dimensional subspace of a real inner product space \( V \), and if \( \mathbf{v} \in V \), then \( \text{Proj}_U(\mathbf{v}) \) is the best approximation to \( \mathbf{v} \) from \( U \) in the sense that

\[
||\mathbf{v} - \text{Proj}_U(\mathbf{v})|| < ||\mathbf{v} - \mathbf{u}|| \quad \forall \mathbf{u} \in U : \mathbf{u} \neq \text{Proj}_U(\mathbf{v}).
\]
In practice, rather than work with minimising $||v - u||$, we minimise $||v - u||^2$ (same outcome, avoid square root). Then Best Approximation Theorem $\implies$ Proj$_U(v)$ is the best approximation

$\iff u = \text{Proj}_U(v)$ is the vector that minimises $||v - u||^2$.

4.1 Examples

1. Inconsistent linear systems: Let $A \in M_{m,n}(\mathbb{R})$ and $b \in M_{m,1}(\mathbb{R})$.
   
   For $m > n$, the linear system described by
   
   $$Ax = b$$
   
   is over-determined and does not in general have a solution. We can look at the least squares solution.

   Important observation: $Ax \in \text{Col}(A)$

   In the context of the inconsistent system and least squares, we seek the closest vector to $b$ in the column space of $A$, then solve
   
   $$A\hat{x} = \text{Proj}_{\text{Col}(A)}(b).$$

   We could use Gram-Schmidt and the orthogonal projection explicitly, but a more efficient approach for this case is to solve the normal equation

   $$A^TA\hat{x} = A^Tb \implies \hat{x} = (A^TA)^{-1}A^Tb.$$ 

   Such equations arise in polynomial fitting to data.
2. Least squares function approximation

Given a function $f \in C[a, b]$, find the best approximation to $f$ using only functions from a specified subspace $U$ of $C[a, b]$.

Interpret “best possible” in the sense of least squares.

Consider $g$ as an approximation to $f$.

At point $x_0$ the error is $|f(x_0) - g(x_0)|$. For the entire interval, define error as $\int_a^b |f(x) - g(x)| \, dx$.

This is area between curves.

An easier definition (and one more amenable to calculations) is the mean squared error (MSE)

$$\text{MSE} = \int_a^b (f(x) - g(x))^2 \, dx.$$ 

Recall the integral inner product on $C[a, b]$;

$$\langle p, q \rangle = \int_a^b p(x)q(x) \, dx$$

$$\Rightarrow \quad \text{MSE} = \|f - g\|^2 = \langle f - g, f - g \rangle = \int_a^b (f(x) - g(x))^2 \, dx.$$
e.g. $\sin(x)$
Find the least squares approximation for $\sin x$ in the subspace of $C[0, \pi]$ spanned by $\{1, x, x^2\}$. Use the inner product

$$\langle p, q \rangle = \int_0^\pi p(x)q(x)\, dx.$$ 

Solution is

$$y = \frac{12(\pi^2 - 10)}{\pi^3} + \frac{60(12 - \pi^2)}{\pi^4}x + \frac{60(\pi^2 - 12)}{\pi^5}x^2.$$ 

---

e.g. Fourier coefficients

In $C[0, 2\pi]$, the set

$$\beta_n = \left\{ \frac{1}{\sqrt{2\pi}} \right\} \cup \left\{ \frac{1}{\sqrt{\pi}} \cos kx \mid k = 1, \ldots, n \right\} \cup \left\{ \frac{1}{\sqrt{\pi}} \sin kx \mid k = 1, \ldots, n \right\},$$

where $n \in \mathbb{N}_0$, is orthonormal with respect to the inner product

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

It follows that $\beta_n$ is an orthonormal basis for the $(2n + 1)$-dimensional subspace $W_n = \text{span}(\beta_n)$ of $C[0, 2\pi]$. The orthogonal projection of $f \in C[0, 2\pi]$ onto $W_n$ is given by $\text{Proj}_{W_n}(f)$. In the limit $n \to \infty$, the corresponding approximation of $f(x)$ yields the **Fourier series** of $f(x)$ over the interval $[0, 2\pi]$:

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left( a_k \cos kx + b_k \sin kx \right),$$

where

$$a_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos kx \, dx, \quad b_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin kx \, dx,$$

are the associated **Fourier coefficients**.