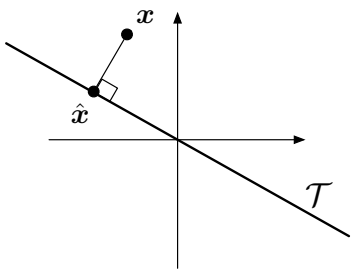
Linear approximation in a Hilbert space

Consider the following problem:

Let $\mathcal{S}$ be a Hilbert space, and let $\mathcal{T}$ be a subspace of $\mathcal{S}$.
Given a $\boldsymbol{x} \in \mathcal{S}$, what is the **closest point** $\hat{\boldsymbol{x}} \in \mathcal{T}$?



In other words, find $\hat{\boldsymbol{x}} \in \mathcal{T}$ that minimizes $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|$; given $\boldsymbol{x}$, we want to solve the following optimization program

$$\underset{\boldsymbol{y} \in \mathcal{T}}{\text{minimize}} \ \|\boldsymbol{x} - \boldsymbol{y}\|, \tag{1}$$

where the norm above is the one induced by the inner product. This problem has a unique solution which is characterized by the **orthogonality principle**.

Theorem: Let $\mathcal{S}$ be a Hilbert space, and let $\mathcal{T}$ be a finite dimensional subspace[1]. Given an arbitrary $\boldsymbol{x} \in \mathcal{S}$,

1. there is exactly one $\hat{\boldsymbol{x}} \in \mathcal{T}$ such that

$$\boldsymbol{x} - \hat{\boldsymbol{x}} \ \perp \ \mathcal{T}, \tag{2}$$

meaning $\langle \boldsymbol{x} - \hat{\boldsymbol{x}}, \boldsymbol{y} \rangle = 0$ for all $\boldsymbol{y} \in \mathcal{T}$, and

[1]The same results hold when $\mathcal{T}$ is infinite dimensional and is *closed*. We do not prove the infinite dimensional case just because it requires some analysis of infinite sequences which, while not really that difficult, kind of distract from the overall geometrical picture we are trying to paint here.

2. this $\hat{\boldsymbol{x}}$ is the closest point in $\mathcal{T}$ to $\boldsymbol{x}$; that is, $\hat{\boldsymbol{x}}$ is the unique minimizer to (1).

Proof: We will show that the first part is true in the next section of the notes, where we show how to explicitly calculate such an $\hat{x}$.

For the second part, let $\hat{\boldsymbol{x}}$ be the vector which obeys

$$\hat{\boldsymbol{e}} = \boldsymbol{x} - \hat{\boldsymbol{x}} \ \perp \ \mathcal{T}.$$

Let $\boldsymbol{y}$ be any other vector in $\mathcal{T}$, and set

$$\boldsymbol{e} = \boldsymbol{x} - \boldsymbol{y}.$$

We will show that

$$\|\boldsymbol{e}\| \ > \ \|\hat{\boldsymbol{e}}\| \quad (\text{i.e. that } \|\boldsymbol{x} - \boldsymbol{y}\| \ > \ \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|)\,.$$

Note that

$$\begin{aligned}\|\boldsymbol{e}\|^2 &= \|\boldsymbol{x} - \boldsymbol{y}\|^2 = \|\hat{\boldsymbol{e}} - (\boldsymbol{y} - \hat{\boldsymbol{x}})\|^2 \\ &= \langle \hat{\boldsymbol{e}} - (\boldsymbol{y} - \hat{\boldsymbol{x}}) \ , \ \hat{\boldsymbol{e}} - (\boldsymbol{y} - \hat{\boldsymbol{x}}) \rangle \\ &= \|\hat{\boldsymbol{e}}\|^2 + \|\boldsymbol{y} - \hat{\boldsymbol{x}}\|^2 - \langle \hat{\boldsymbol{e}}, \boldsymbol{y} - \hat{\boldsymbol{x}} \rangle - \langle \boldsymbol{y} - \hat{\boldsymbol{x}}, \hat{\boldsymbol{e}} \rangle.\end{aligned}$$

Since $\boldsymbol{y} - \hat{\boldsymbol{x}} \in \mathcal{T}$ and $\hat{\boldsymbol{e}} \perp \mathcal{T}$,

$$\langle \hat{\boldsymbol{e}}, \boldsymbol{y} - \hat{\boldsymbol{x}} \rangle = 0, \quad \text{and} \quad \langle \boldsymbol{y} - \hat{\boldsymbol{x}}, \hat{\boldsymbol{e}} \rangle = 0,$$

and so

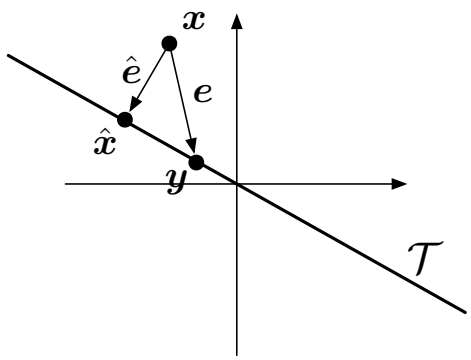
$$\|\boldsymbol{e}\|^2 = \|\hat{\boldsymbol{e}}\|^2 + \|\boldsymbol{y} - \hat{\boldsymbol{x}}\|^2.$$

Since all three quantities in the expression above are positive and

$$\|\boldsymbol{y} - \hat{\boldsymbol{x}}\| > 0 \quad \Leftrightarrow \quad \boldsymbol{y} \neq \hat{\boldsymbol{x}},$$

we see that

$$\boldsymbol{y} \neq \hat{\boldsymbol{x}} \quad \Leftrightarrow \quad \|\boldsymbol{e}\| > \|\hat{\boldsymbol{e}}\|.$$



Computing the best approximation

The orthogonality principle gives us a concrete procedure for actually **computing** the optimal point $\hat{\boldsymbol{x}}$.

Let N be the dimension of $\mathcal{T}$, and let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ be a basis for $\mathcal{T}$. We want to find coefficients $a_1, \ldots, a_N \in \mathbb{C}$ such that

$$\hat{\boldsymbol{x}} = a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2 + \cdots + a_N\boldsymbol{v}_N.$$

The orthogonality principle tells us that

$$\langle \boldsymbol{x} - \hat{\boldsymbol{x}}, \boldsymbol{v}_n \rangle = 0 \quad \text{for} \quad n = 1, \ldots, N.$$

This means the a_n must obey

$$\langle \boldsymbol{x} - \sum_{k=1}^{N} a_k \boldsymbol{v}_k \ , \ \boldsymbol{v}_n \rangle = 0 \quad \text{for} \quad n = 1, \ldots, N,$$

or moving things around,

$$\langle \boldsymbol{x}, \boldsymbol{v}_n \rangle = \sum_{k=1}^{N} a_k \langle \boldsymbol{v}_k, \boldsymbol{v}_n \rangle \quad \text{for} \quad n = 1, \ldots, N.$$

Since $\boldsymbol{x}$ and the $\{\boldsymbol{v}_n\}$ are given, we know both the $\langle \boldsymbol{x}, \boldsymbol{v}_n \rangle$ and the $\langle \boldsymbol{v}_k, \boldsymbol{v}_n \rangle$. We are left with a set of N **linear equations** with N unknowns:

$$\begin{bmatrix} \langle \boldsymbol{v}_1, \boldsymbol{v}_1 \rangle & \langle \boldsymbol{v}_2, \boldsymbol{v}_1 \rangle & \cdots & \langle \boldsymbol{v}_N, \boldsymbol{v}_1 \rangle \\ \langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle & \langle \boldsymbol{v}_2, \boldsymbol{v}_2 \rangle & & \langle \boldsymbol{v}_N, \boldsymbol{v}_2 \rangle \\ \vdots & & \ddots & \vdots \\ \langle \boldsymbol{v}_1, \boldsymbol{v}_N \rangle & \cdots & & \langle \boldsymbol{v}_N, \boldsymbol{v}_N \rangle \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} = \begin{bmatrix} \langle \boldsymbol{x}, \boldsymbol{v}_1 \rangle \\ \langle \boldsymbol{x}, \boldsymbol{v}_2 \rangle \\ \vdots \\ \langle \boldsymbol{x}, \boldsymbol{v}_N \rangle \end{bmatrix}$$

The matrix on the left hand side above is called the **Gram matrix** or **Grammian** of the basis $\{\boldsymbol{v}_n\}$.

In more compact notation, we want to find $\boldsymbol{a} \in \mathbb{C}^N$ such that

$$\boldsymbol{G}\boldsymbol{a} = \boldsymbol{b},$$

where $b_n = \langle \boldsymbol{x}, \boldsymbol{v}_n \rangle$ and $G_{k,n} = \langle \boldsymbol{v}_n, \boldsymbol{v}_k \rangle$.

Two notes on the structure of $\boldsymbol{G}$:

- $\boldsymbol{G}$ is guaranteed to be invertible because the $\{\boldsymbol{v}_n\}$ are linearly independent. We can comfortably write

$$\boldsymbol{a} = \boldsymbol{G}^{-1}\boldsymbol{b}.$$

- $\boldsymbol{G}$ is **conjugate symmetric** ("Hermitian"):

$$\boldsymbol{G} = \boldsymbol{G}^{\mathrm{H}},$$

  where $\boldsymbol{G}^{\mathrm{H}}$ is the conjugate transpose of $\boldsymbol{G}$ (take the transpose, then take the complex conjugate of all the entries). This fact has algorithmic implications when it comes time to actually solve the system of equations.

Uniqueness

It should be clear that if $\langle \boldsymbol{e}, \boldsymbol{v}_k \rangle = 0$ for all of the basis vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$, then $\langle \boldsymbol{e}, \boldsymbol{y} \rangle = 0$ for all $\boldsymbol{y} \in \mathcal{T}$. The converse is also true: if $\langle \boldsymbol{e}, \boldsymbol{y} \rangle \neq 0$ for some $\boldsymbol{y} \in \mathcal{T}$ not equal to $\boldsymbol{0}$, then $\langle \boldsymbol{e}, \boldsymbol{v}_k \rangle \neq 0$ for at least one of the $\boldsymbol{v}_k$.

With the work above, this means that a necessary and sufficient condition for $\langle \boldsymbol{x} - \hat{\boldsymbol{x}}, \boldsymbol{y} \rangle = 0$ for all $\boldsymbol{y} \in \mathcal{T}$ is to have

$$\hat{\boldsymbol{x}} = \sum_{n=1}^{N} a_n \boldsymbol{v}_n, \quad \text{where } \boldsymbol{a} \text{ satisfies} \quad \boldsymbol{G}\boldsymbol{a} = \boldsymbol{b}.$$

Since $\boldsymbol{G}$ is square and invertible, there is exactly one such $\boldsymbol{a}$, and hence exactly one $\hat{\boldsymbol{x}}$ that obeys the condition

$$\boldsymbol{x} - \hat{\boldsymbol{x}} \perp \mathcal{T}.$$

Example: Let

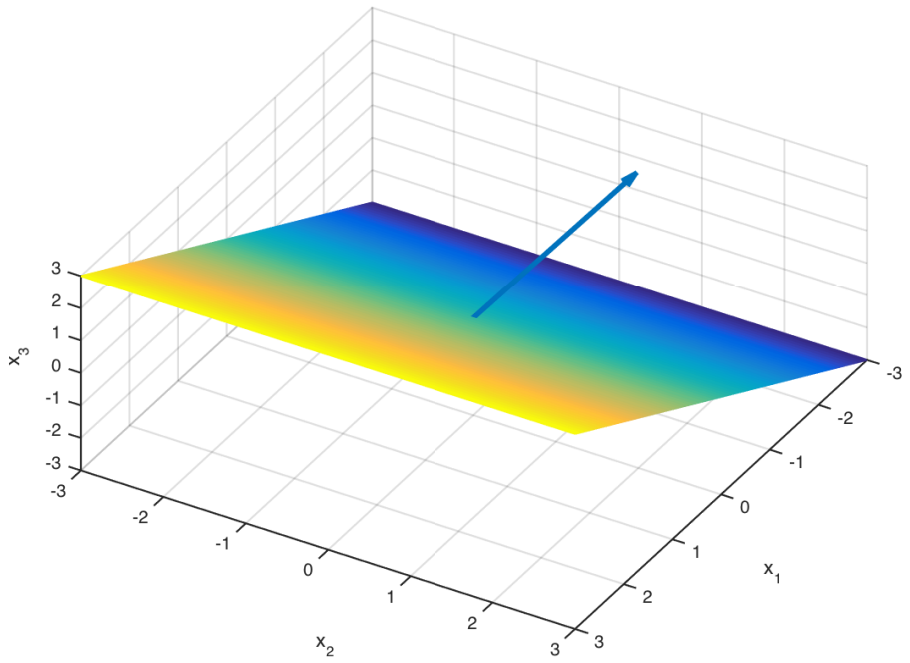
$$\mathcal{T} = \text{Span}\left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \right), \qquad \boldsymbol{x} = \begin{bmatrix} -2 \\ 1 \\ 3 \end{bmatrix}$$

Find the solution to

$$\underset{\boldsymbol{y} \in \mathcal{T}}{\text{minimize}} \, \|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

(Recall that $\|\cdot\|_2$ in $\mathbb{R}^3$ is induced by the standard inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{n=1}^{3} x_n y_n$.)

Here is a plot of $\mathcal{T}$ and $\boldsymbol{x}$:



Solution: We have

$$\boldsymbol{G} = \qquad\qquad \boldsymbol{b} =$$

and so

$$\boldsymbol{G}^{-1} =$$

This means that

$$\boldsymbol{a} =$$

from which we synthesize the answer

$$\hat{\boldsymbol{x}} =$$

We can also check that "the error is orthogonal to the approximation"

$$\langle \boldsymbol{x} - \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}} \rangle =$$

Example: Polynomial approximation of e^t

1. We want to calculate a quadratic approximation of $x(t) = e^t$ over the interval $[0, 1]$.

   (a) Sketch $x(t)$ on $[0, 1]$.

   (b) One way to do the approximation is to truncate the Taylor expansion

   $$e^t = 1 + t + \frac{t^2}{2} + \frac{t^3}{6} + \frac{t^4}{24} + \cdots$$

   In this case the approximation is

   $$\tilde{x}_{\text{taylor}}(t) = 1 + t + t^2/2.$$

   Sketch $\tilde{x}_{\text{taylor}}(t)$.

   (c) Is the truncated Taylor approximation the best possible quadratic approximation? The answer is no if we are interested in minimizing the energy of the error. We want to find the second order polynomial

   $$\tilde{x}(t) = a_1 + a_2 t + a_3 t^2$$

   that minimizes

   $$\|e^t - \tilde{\boldsymbol{x}}\|_{L_2([0,1])} = \sqrt{\int_0^1 |e^t - \tilde{x}(t)|^2 \ dt}.$$

   We set this up as a subspace approximation problem. Set

   $$v_1(t) = 1, \quad v_2(t) = t, \quad v_3(t) = t^2,$$

   and set $\mathcal{T} = \text{Span}\left(\{v_1, v_2, v_3\}\right)$. We of course now know a systematic way to find the best $\tilde{x} \in \mathcal{T}$. Start by calculating the Gram matrix $\boldsymbol{G}$ (recall that $G_{ij} = \langle v_i, v_j \rangle$).

   (d) Calculate the right-hand-side $\boldsymbol{b}$ (recall that $b_i = \langle x, v_i \rangle$). Write down the system of equations that need to be solved. You can use MATLAB to solve this system.

   (e) Calculate the error for both the Taylor approximation and the optimal approximation computed above. Plot $x(t), \tilde{x}_{\text{taylor}}(t)$, and $\tilde{x}(t)$ on the same set of axes.

2. We consider the same problem, but calculate the error in a different way. The norm we use to measure the error is

$$\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_S^2 = \int_0^1 w(t)|x(t) - \tilde{x}(t)|^2 \, dt$$

where

$$w(t) = 16(t - 1/2)^2.$$

(a) Plot $w(t)$ and argue that measuring the error in this way will penalize mismatch at the ends of the interval than in the middle.

(b) Write down the inner product that induces $\|\cdot\|_S$.

(c) Find the second order polynomial that is the best approximation to e^t in the $\|\cdot\|_S$ norm.