Convex Optimization

Review Sessions 5 & 6

Question 1 (Estimation with Conditional Independence Constraints)

[1]Consider N independent samples $\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(N)} \in \mathbb{R}^n$ of a vector $\boldsymbol{X}$. Assume $\boldsymbol{X}$ has a multivariate Gaussian distribution.

Assume also that you are given a list $\mathscr{N}$ of tuples $(i, j) \in \{1, \cdots, n\} \times \{1, \cdots, n\}$. If $(i, j) \in \mathscr{N}$, then X_i and X_j are conditionally independent in the said distribution (in other words, the distribution of the vector $(X_i, X_j) \in \mathbb{R}^2$ conditional on every other component of $\boldsymbol{X}$ has a diagonal covariance matrix).

Write a program to find the maximum-likelihood estimates of the mean and covariance matrix of the distribution given the observed data and the conditional independence assumptions.

Solution

Let us first develop some notation.

- The subscript $_{i,j}$ implies that we only need components concerning variables i and j. For example, $\boldsymbol{\mu}_{i,j} = (\mu_i, \mu_j)^\top$. Similarly,
$$\Sigma_{i,j} = \left[\begin{array}{cc} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{array} \right]$$
- The subscript $_{i,j}$ on a *barred* variable implies every component concerning variables *other than* i and j.

In this question, we will, without loss of generality, re-order the components of $\boldsymbol{X}$ so that i and j are the *last* two components.

We'll begin by proving a lemma, followed by a theorem

> **Theorem 1. (Completing the Square)**
>
> $$\boldsymbol{x}^\top A \boldsymbol{x} + \boldsymbol{b} \cdot \boldsymbol{x} = (\boldsymbol{x} + A^{-1}\boldsymbol{b})^\top A (\boldsymbol{x} + A^{-1}\boldsymbol{b}) - \boldsymbol{b}^\top A^{-1} \boldsymbol{b}$$

Proof. This is none other than completing the square in multiple

[1] Based on additional exercise 6.2 in Boyd

dimensions.

$$\begin{aligned}
\boldsymbol{x}^\top A\boldsymbol{x} + \boldsymbol{b}\cdot\boldsymbol{x} &= \|A^{1/2}\boldsymbol{x}\|_2^2 + 2\boldsymbol{b}\cdot\boldsymbol{x} \\
&= \|A^{1/2}\boldsymbol{x} + A^{-1/2}\boldsymbol{b}\|_2^2 - \boldsymbol{b}^\top A^{-1}\boldsymbol{b} \\
&= \|A^{1/2}(\boldsymbol{x} + A^{-1}\boldsymbol{b})\|_2^2 - \boldsymbol{b}^\top A^{-1}\boldsymbol{b} \\
&= (\boldsymbol{x} + A^{-1}\boldsymbol{b})^\top A(\boldsymbol{x} + A^{-1}\boldsymbol{b}) - \boldsymbol{b}^\top A^{-1}\boldsymbol{b}
\end{aligned}$$

□

Theorem 2. (Conditional Marginal Density of Multivariate Gaussian) Suppose $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$. Then

$$(\boldsymbol{X}_{i,j}|\bar{\boldsymbol{X}}_{i,j} = \bar{\boldsymbol{x}}_{i,j}) \sim N\Big(\boldsymbol{\mu}_{i,j} + B^\top\bar{\Sigma}_{i,j}^{-1}(\bar{\boldsymbol{x}}_{i,j} - \bar{\boldsymbol{\mu}}_{i,j}), S\Big)$$

where S is the Schurr complement of $\bar{\Sigma}_{i,j}$ in Σ.

Proof. To show this, consider the following decomposition of Σ

$$\Sigma = \begin{bmatrix} \bar{\Sigma}_{i,j} & B \\ B^\top & \Sigma_{i,j} \end{bmatrix}$$

and note that by the discussion on page 650 of Boyd,

$$\Sigma^{-1} = \begin{bmatrix} \bar{\Sigma}_{i,j}^{-1} + \bar{\Sigma}_{i,j}^{-1}BS^{-1}B^\top\bar{\Sigma}_{i,j}^{-1} & -\bar{\Sigma}_{i,j}^{-1}BS^{-1} \\ -S^{-1}B^\top\bar{\Sigma}_{i,j}^{-1} & S^{-1} \end{bmatrix}$$

where S is the Schur complement in the statement of the theorem. Now, let $\boldsymbol{\chi} = \boldsymbol{x} - \boldsymbol{\mu}$ and consider that[1]

$$\begin{aligned}
\mathbb{P}(\boldsymbol{X}_{i,j} = \boldsymbol{x}_{i,j}|\bar{\boldsymbol{X}}_{i,j} = \bar{\boldsymbol{x}}_{i,j}) &\propto \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) \\
&\propto \exp\left\{-\frac{1}{2}\boldsymbol{\chi}^\top\Sigma^{-1}\boldsymbol{\chi}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\chi}_{i,j}^\top S^{-1}\boldsymbol{\chi}_{i,j} - \boldsymbol{\chi}_{i,j}\cdot(2S^{-1}B^\top\bar{\Sigma}_{i,j}^{-1}\bar{\boldsymbol{\chi}}_{i,j})\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\chi}_{i,j} - B^\top\bar{\Sigma}_{i,j}^{-1}\bar{\boldsymbol{\chi}}_{i,j})^\top S^{-1}(\boldsymbol{\chi}_{i,j} - B^\top\bar{\Sigma}_{i,j}^{-1}\bar{\boldsymbol{\chi}}_{i,j})\right]\right\}
\end{aligned}$$

This is clearly in the form of a multivariate distribution of the form stated in the theorem. □

In some sense, the theorem above is slight overkill. All we'll really need for this question is the following theorem

Theorem 3. Suppose $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$. Then X_i and X_j are independent given the values of every other variable if and only if $(\Sigma^{-1})_{i,j} = 0$.

[1] Note that it is customary, when working with conditional distributions, to ignore all multiplicative factors that do not depend on the quantities of interest (in this case $\boldsymbol{x}_{i,j}$. By Bayes' Theorem, the constants will then 'work out' to ensure our result is a proper distribution.

Proof. We showed in Theorem 2 that the *inverse* of the conditional covariance matrix of $\boldsymbol{X}_{i,j}$ conditional on $\bar{\boldsymbol{X}}_{i,j}$ is the matrix formed by the i^{th} and j^{th} rows and columns of Σ^{-1}.

Now, the conditional covariance matrix is diagonal if and only if its inverse is diagonal, which happens if and only if the matrix formed by the i^{th} and j^{th} rows and columns of Σ^{-1} is diagonal. Thus, we require

$$(\Sigma^{-1})_{i,j} = (\Sigma^{-1})_{j,i} = 0$$

But since Σ is symmetric (since it's a covariance matrix), the condition in the theorem is enough. □

After this lengthy preliminary, let's get going. Our decision variables in this case are $\boldsymbol{\mu}$ and Σ, the maximum-likelihood parameters. Define the following

- $\boldsymbol{m} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}^{(N)}$
- ψ is the matrix whose *columns* are the vectors $\boldsymbol{y}^{(i)} - \boldsymbol{m}$.
- $S = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y}^{(i)} - \boldsymbol{m})(\boldsymbol{y}^{(i)} - \boldsymbol{m})^\top = \frac{1}{N}\psi\psi^\top$

And now, note that given $\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(N)}$, the log-likelihood is

$$\begin{aligned}
\ell &= \sum_{i=1}^{N}\log\left\{\frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}}\exp\left[-\frac{1}{2}(\boldsymbol{y}^{(i)}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{y}^{(i)}-\boldsymbol{\mu})\right]\right\} \\
&= -\frac{N}{2}\log\det\Sigma - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{y}^{(i)}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{y}^{(i)}-\boldsymbol{\mu}) + C \\
&= -\frac{N}{2}\log\det\Sigma - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{y}^{(i)}-\boldsymbol{\mu}+\boldsymbol{m}-\boldsymbol{m})^\top\Sigma^{-1}(\boldsymbol{y}^{(i)}-\boldsymbol{\mu}+\boldsymbol{m}-\boldsymbol{m}) + C \\
&= -\frac{N}{2}\log\det\Sigma - \frac{N}{2}(\boldsymbol{\mu}-\boldsymbol{m})^\top\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{m}) - \frac{1}{2}\mathrm{tr}(\psi^\top\Sigma^{-1}\psi) + C \\
&= -\frac{N}{2}\log\det\Sigma - \frac{N}{2}(\boldsymbol{\mu}-\boldsymbol{m})^\top\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{m}) - \frac{N}{2}\mathrm{tr}(\Sigma^{-1}\frac{1}{N}\psi\psi^\top) + C \\
&= \frac{N}{2}\Big(-\log\det\Sigma - (\boldsymbol{\mu}-\boldsymbol{m})^\top\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{m}) - \mathrm{tr}(\Sigma^{-1}S)\Big) + C \\
&= \frac{N}{2}\Big(\log\det\Sigma^{-1} - (\boldsymbol{\mu}-\boldsymbol{m})^\top\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{m}) - \mathrm{tr}(\Sigma^{-1}S)\Big) + C
\end{aligned}$$

We need to minimize this with respect to both $\boldsymbol{\mu}$ and Σ. Clearly, the optimal solution for $\boldsymbol{\mu}$ will be $\boldsymbol{\mu} = \boldsymbol{m}$. This leaves us with the following quantity to maximize

$$\log\det\Sigma^{-1} - \mathrm{tr}(\Sigma^{-1}S)$$

Finally, we need to use the last theorem above to add constraints to ensure the independence assumptions in $\mathscr{N}$ are satisfied.

Our final program is

$$\begin{array}{rll} \max & \log\det\Sigma^{-1} - \operatorname{tr}(\Sigma^{-1}S) & \\ \text{s.t.} & (\Sigma^{-1})_{ij} = 0 & \forall (i,j) \in \mathscr{N} \\ & \Sigma \succeq 0 & \end{array}$$

This is convex in Σ.

■ □ ■

Question 2 (Image Interpolation)

A grayscale image is represented as a matrix of pixels $A \in \mathbb{R}^{m \times n}$. Some of the pixels are known to you, but others are not.

Find a way to reconstruct the image using just the known pixels.

Solution

Suppose $\mathscr{K}$ is the set of pixels we know. Our decision variables are A_{ij} for $(i,j) \notin \mathcal{K}$. Our aim will be to set these pixels so as to make the picture as 'smooth' as possible.

We'll experiment with two different 'kinds' of smoothness

ℓ_2 **variation** simply measures roughness as

$$\sum_{i=2}^{m}\sum_{j=2}^{n}\left\{(A_{ij} - A_{i-1,j})^2 + (A_{ij} - A_{i,j-1})^2\right\}$$

Total variation simply measures roughness as

$$\sum_{i=2}^{m}\sum_{j=2}^{n}\left\{|A_{ij} - A_{i-1,j}| + |A_{ij} - A_{i,j-1}|\right\}$$

■ □ ■

Question 3 (Theory of Hilbert Spaces)

In this "question", we'll be discussing Hilbert spaces.

Note that much of the discussion here is analogous to the discussion we'll be having when we consider Banach spaces, and in many cases the proofs and definitions are identical. Whenever this is the case, I've annotated the relevant section with a * – those items will not be repeated when we consider Banach spaces.

Solution

[2] See `http://bit.ly/Hdf9nj` for the definition of a vector space. One of the properties required of a set S for it to be a vector space is that it be closed under addition (ie: for any $\boldsymbol{x}, \boldsymbol{y} \in S$, we have $\boldsymbol{x} + \boldsymbol{y} \in S$). But there are many more.

[3] To be an inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ must satisfy

- $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle^*$ (where x^* is the complex conjugate of x)
- $\langle \boldsymbol{x} + \boldsymbol{y}, \boldsymbol{z} \rangle = \langle \boldsymbol{x}, \boldsymbol{z} \rangle + \langle \boldsymbol{y}, \boldsymbol{z} \rangle$
- $\langle \lambda \boldsymbol{x}, \boldsymbol{y} \rangle = \lambda \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ for all λ
- $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$, with equality if and only if $\boldsymbol{x} = \boldsymbol{0}$.

Before we study Hilbert Spaces, we ought to define them.

> **Definition 1. (Pre-Hilbert Space)** A *pre-Hilbert space* is a vector space[2] V over $\mathbb{C}$ equipped with an inner product[3] $\langle \boldsymbol{x}, \boldsymbol{y} \rangle : V \times V \to \mathbb{C}$

Of course, one common pre-Hilbert space is $\mathbb{R}^n$, equipped with the standard inner product. However, this is only one example of many vector spaces. Indeed, if you've never encountered more general vector spaces before, it might require a pretty radical shift in mindset to stop viewing vectors as arrows in space, and instead to see them as objects that satisfy certain properties. In some ways, it's quite liberating!

Here are a few examples of vector spaces you might not have encountered before.

Example

- The space of all sequences that are square summable is a vector space (quick check: the sum of two such sequences is square summable as well, so it is also in the space). Each element $\boldsymbol{x}$ is an infinite sequence $\{x_i\}_{i=1}^{\infty}$ that satisfies $\sum_{i=1}^{\infty} x_i^2 < \infty$.

  This vector space can be made into a pre-Hilbert space by defining the following inner product.

  $$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{\infty} x_i y_i$$

- The space of all polynomials of $[a, b]$ is a vector space (quick check: the sum of two polynomials is also a polynomial and therefore also in the space). Each element $\boldsymbol{x}$ is a polynomial on $[a, b]$. We can make this into a pre-Hilbert space by defining the following inner product

  $$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \int_a^b \boldsymbol{x}(t)\boldsymbol{y}(t)\mathrm{d}t$$

Note also the following useful theorem

> **Theorem 4.** The *induced norm*
>
> $$\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$$
>
> satisfies all the properties required of norms. Furthermore, the inner product is continuous in both its arguments with respect to this nor.

Proof. Luenberger, pp 49. □

Now that we understand the concept of a pre-Hilbert space, we are ready to define a Hilbert space.

> **Definition 2. (Hilbert Space)**
>
> A *Hilbert Space* is a pre-Hilbert space in which every Cauchy sequence[4] converges in the space, with respect to the induced norm $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$.
>
> When this happens, we say the underlying vector space is *completes* with respect to the induced norm.

The idea of completeness (which is loosely related to the idea of closure) may seem utterly mystifying – so let's look at an example.

Example

Consider the vector space $C[0,1]$ of all continuous functions on the interval [0,1] (quick check: the sum of two such functions is also continuous and therefore in the space). Consider the following two norms

$$\|f\|_1 = \max_{\boldsymbol{t}\in[0,1])} |f(t)| \qquad\qquad \|f\|_2 = \int_0^1 |f(t)|\,\mathrm{d}t$$

(It turns out that neither norms are induced by an inner product, so we haven't really defined a pre-Hilbert space. Regardless, we can still explore the completeness $C[0,1]$.)

Initially, you might be worried that $C[0,1]$ is not complete. Indeed, it's easy to find a set of continuous functions that tend to the step function (which is not in $C[0,1]$).

It actually makes no sense, however, to talk of completeness without quoting a norm. So let's consider the 'counterexample' for each of the norms above.

- For $\|f\|_2$, we do indeed have a problem. As our sequence of functions approaches the step function, the integral of these functions does indeed approach the integral of the step function. Thus, the sequence is Cauchy, and the fact it does not converge in the space implies our space is incomplete.
- For $\|f\|_1$, we don't have this problem! Indeed, note that $\|f_n - f_m\|$ finds the *maximum* difference between the two functions f_n and f_m, and

[4] A sequence $\{x_n\}$ is *Cauchy* in a pre-Hilbert space V is *Cauchy* if and only if

$$\|\boldsymbol{x}_n - \boldsymbol{x}_m\| \to 0$$

as $n, m \to \infty$.

we will never find a continuous function that gets arbitrarily close to the discontinuous step function at every single point. Thus, the sequence is not Cauchy, and this counterexample doesn't mean $C[0,1]$ is incomplete. (Indeed, it turns out that $C[0,1]$ is complete with respect to $\|f\|_1$).

Before we complete this section, it is worth mentioning some common Hilbert spaces, for future reference

> **Theorem 5. (Common Hilbert Spaces)**
>
> The following spaces are Hilbert spaces:
>
> - $\mathbb{R}^n$, with the standard dot product.
> - ℓ_2 (the set of square-summable sequences) with inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum x_i y_i$.
> - L_2 (the set of square-integrable functions) with inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \int x(t)y(t)\mathrm{d}t$.

The Projection Theorem

The time has come to look at our first optimization problem in (possibly infinite-dimensional) Hilbert space. We will consider the following problem

> Given an afine set $\mathscr{V}$ in a Hilbert space H, find the vector $\boldsymbol{v} \in \mathscr{V}$ of minimum norm (ie: the vector closest to $\mathbf{0}$).

This is a simpler problem that most you've encountered so far in this course – indeed, it insists that the objective function be a norm, and it only allows the feasible region to be an affine space (in other words, the result of equality constraints). Nevertheless, we will find that this framework is useful in solving many problems of interest.[5]

Let us first derive some optimality conditions for such a problem.

> **Theorem 6. (Projection)** We give three (equivalent) statements of this important theorem
>
> - Let H be a Hilbert space and $\mathscr{V}$ a closed affine set in H (which can be expressed as $\mathscr{V} = \boldsymbol{v}_0 + M$, where M is a closed subspace of H).
>
>   Then there exists a unique vector $\boldsymbol{v}_{opt} \in \mathscr{V}$ of minimum norm. A necessary and sufficient condition for $\boldsymbol{v}_{opt}$ to be

[5] In fact, the next theorem can be extended to convex sets quite easily – see Luenberger pp 69. That said, most of the problems we'll consider in this course will only involve equality constraints, and this discussion here will be more than sufficient in dealing with those problems.

this vector is that it be orthogonal to M.

- Let H be a Hilbert space, M a closed subspace of H, and $\boldsymbol{x}$ an arbitrary vector in H.

  Then there exists a unique vector $\boldsymbol{m}_{opt} \in M$ that is closer to $\boldsymbol{x}$ than any other vector in M. A necessary and sufficient condition for $\boldsymbol{m}_{opt}$ to be this vector is that $\boldsymbol{x} - \boldsymbol{m}_{opt}$ be orthogonal to M.

- Let H be a Hilbert space, $\boldsymbol{x}$ an arbitrary vector in H, and $\mathscr{V}$ a closed affine set in H (which can be expressed as $\mathscr{V} = \boldsymbol{v}_0 + M$, where M is a closed subspace of H).

  Then there exists a unique vector $\boldsymbol{v}_{opt} \in \mathscr{V}$ that is closer to $\boldsymbol{x}$ than any other vector in $\mathscr{V}$. A necessary and sufficient condition for $\boldsymbol{v}_{opt}$ to be this vector is that $\boldsymbol{x} - \boldsymbol{v}_{opt}$ be orthogonal to M.

Proof. See Luenberger, pp 50-51 and pp 64. □

This theorem is of enormous practical importance, as will become obvious in the forthcoming exercises. It also allows us to derive the following extremely useful theorem:

Theorem 7. Let M be a closed subspace of a Hilbert space H. Then

$$H = M \oplus M^{\perp}$$

and

$$M = M^{\perp\perp}$$

where $M^{\perp}$ is the set of vectors orthogonal to (all vectors in) M.

Proof. The first statement follows directly from the Projection Theorem. By the second statement above, for every $\boldsymbol{x}$, there is a point $\boldsymbol{m}_{opt} \in M$ such that $\boldsymbol{m}_{\perp} = \boldsymbol{x} - \boldsymbol{m}_{opt} \in M^{\perp}$. Thus,

$$\boldsymbol{x} = \overbrace{\boldsymbol{m}_{opt}}^{\in M} + \overbrace{\boldsymbol{x} - \boldsymbol{m}_{opt}}^{\in M^{\perp}}$$

We can therefore express any $\boldsymbol{x} \in H$ as the sum of a vector in M and one in $M^{\perp}$.[6]

For the second part, it is obvious that $M \subseteq M^{\perp\perp}$, because every vector in $M^{\perp}$ is orthogonal to every vector in M, and so every vector in M is orthogonal to every vector in $M^{\perp}$ and therefore in $M^{\perp\perp}$. To show the other direction:

- Suppose we find a vector $\boldsymbol{x} \in M^{\perp\perp}$ but $\boldsymbol{x} \notin M$.

[6] We can show that this representation is unique by noting that if $\boldsymbol{x} = \boldsymbol{m}_1 + \boldsymbol{m}_1^{\perp}$ and $\boldsymbol{x} = \boldsymbol{m}_2 + \boldsymbol{m}_2^{\perp}$, then $(\boldsymbol{m}_1 - \boldsymbol{m}_2) + (\boldsymbol{m}_1^{\perp} - \boldsymbol{m}_2^{\perp}) = \boldsymbol{x} - \boldsymbol{x} = \boldsymbol{0}$. However, since each of the bracketed terms are orthogonal (being in M and $M^{\perp}$), we can take inner products of each side with itself to conclude that

$$\|\boldsymbol{m}_1 - \boldsymbol{m}_2\|^2 + \|\boldsymbol{m}_1^{\perp} - \boldsymbol{m}_2^{\perp}\|^2 = 0$$

which implies that $\boldsymbol{m}_1 = \boldsymbol{m}_2$ and $\boldsymbol{m}_1^{\perp} = \boldsymbol{m}_2^{\perp}$. Thus, the representation is unique.

- By the first part, we can write $\boldsymbol{x} = \boldsymbol{m} + \boldsymbol{m}^{\perp}$ (with $\boldsymbol{m} \in M$ and $\boldsymbol{m}^{\perp} \in M^{\perp}$).
- However, since $M \subseteq M^{\perp\perp}$, we have that $\boldsymbol{m} \in M^{\perp\perp}$. As such, since by assumption $\boldsymbol{x} \in M^{\perp\perp}$, we have that $\boldsymbol{x} - \boldsymbol{m} = \boldsymbol{m}^{\perp} \in M^{\perp\perp}$.
- Thus, $\boldsymbol{m}^{\perp} \in M^{\perp}$ and $\boldsymbol{m}^{\perp} \in M^{\perp\perp}$. Thus, the vector is perpendicular to itself and must be equal to 0; $\boldsymbol{m}^{\perp} = \mathbf{0}$.
- As such, $\boldsymbol{x} = \boldsymbol{m} \in M$. This is a contradiction.

□

Linear Functionals

Having finally defined Hilbert spaces, we now turn to functions on elements of these spaces.

> **Definition 3. (Linear Functional*)** A function $\phi : V \to \mathbb{C}$ is a *linear functional* on V if for any $\boldsymbol{x}, \boldsymbol{y} \in V$ and $\alpha, \beta \in \mathbb{C}$
>
> $$\phi(\alpha\boldsymbol{x} + \beta\boldsymbol{y}) = \alpha\phi(\boldsymbol{x}) + \beta\phi(\boldsymbol{y})$$
>
> We further say that
>
> - ϕ is continuous if for every $\epsilon > 0$, there exists a δ such that
> $$|\phi(\boldsymbol{y} - \boldsymbol{x})| \leq \epsilon \qquad \forall \boldsymbol{y} : \|\boldsymbol{x} - \boldsymbol{y}\| \leq \delta$$
> - ϕ is bounded if there exists some constant M such that
> $$|\phi(\boldsymbol{y})| \leq M\|\boldsymbol{y}\| \qquad \forall \boldsymbol{y} \in V$$
>   We define the *norm* of the functional to be the smallest such constant
> $$\begin{aligned} \|\phi\| &= \inf\{M : |\varphi(\boldsymbol{y})| \leq M\|\boldsymbol{y}\|\} \\ &= \inf\{M : \|\boldsymbol{y}\||\varphi(\hat{\boldsymbol{y}})| \leq M\|\boldsymbol{y}\|\} \\ &= \sup_{\|\boldsymbol{y}\|=1} |\phi(\boldsymbol{y})\| \end{aligned}$$

We begin by proving a useful theorem

> **Theorem 8. (*)** Let ϕ be a linear functional. The following three statements are equivalent
>
> - ϕ is continuous at some point.
> - ϕ is continuous everywhere.

- ϕ is bounded.

Proof. Let's prove each step:

1 $\rightarrow$ **2** : Let ϕ be continuous at $\boldsymbol{x}_0$. Note that by linearity,

$$|\phi(\boldsymbol{y}) - \phi(\boldsymbol{x})| = |\phi(\boldsymbol{y} - \boldsymbol{x} + \boldsymbol{x}_0) - \phi(\boldsymbol{x})0)|$$

Continuity at $\boldsymbol{x}_0$ then implies continuity everywhere.

2 $\rightarrow$ **3** : If ϕ is continuous everywhere, it is continuous at 0. As such, there exists a δ such that

$$|\phi(\boldsymbol{y})| \leq 1 \qquad \forall \|\boldsymbol{y}\| \leq \delta$$

Now, consider that for any $\boldsymbol{z}$

$$|\phi(\boldsymbol{z})| = \frac{\|\boldsymbol{z}\|}{\delta} \left| \phi \left(\delta \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|} \right) \right| \leq \frac{\|\boldsymbol{z}\|}{\delta}$$

3 $\rightarrow$ **1** : If ϕ is bounded (with norm M), then

$$|\phi(\boldsymbol{z})| \leq \epsilon \qquad \forall \|\boldsymbol{z}\| \leq \frac{\epsilon}{M}$$

It is therefore continuous at 0.

□

We now prove a somewhat astounding and rather beautiful theorem relating to linear functionals in Hilbert spaces.

Theorem 9. (Riesz-Frechet)

Let ϕ be a bounded linear functional on a Hilbert space H with inner product $\langle \cdot, \cdot \rangle$. Then there exists an element $\boldsymbol{z} \in H$ such that $\|\boldsymbol{z}\| = \|\phi\|$ (where the norm of a linear functional was defined above) and

$$\phi(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{z} \rangle \qquad \forall \boldsymbol{x} \in H$$

Proof. Let

$$M = \{\boldsymbol{y} : \phi(\boldsymbol{y}) = 0\}$$

Since the functional is continuous, M is closed. Now, if $M = H$ (ie: ϕ is 0 for every element in the space), simply set $\boldsymbol{z} = \mathbf{0}$ and we're done.

If not, pick some $\boldsymbol{\gamma} \in M^{\perp}$, and consider

$$\phi\left(\boldsymbol{x} - \frac{\phi(\boldsymbol{x})}{\phi(\boldsymbol{\gamma})}\boldsymbol{\gamma}\right) = \phi(\boldsymbol{x}) - \frac{\phi(\boldsymbol{x})}{\phi(\boldsymbol{\gamma})}\phi(\boldsymbol{\gamma}) = \phi(\boldsymbol{x}) - \phi(\boldsymbol{x}) = 0$$

As such, we have that

$$\begin{aligned}\boldsymbol{x} - \frac{\phi(\boldsymbol{x})}{\phi(\boldsymbol{\gamma})}\boldsymbol{\gamma} \in M \quad &\Rightarrow \quad \left\langle \boldsymbol{x} - \frac{\phi(\boldsymbol{x})}{\phi(\boldsymbol{\gamma})}\boldsymbol{\gamma}, \boldsymbol{\gamma} \right\rangle = 0 \\ &\Rightarrow \quad \langle \boldsymbol{x}, \boldsymbol{\gamma} \rangle = \frac{\phi(\boldsymbol{x})}{\phi(\boldsymbol{\gamma})}\langle \boldsymbol{\gamma}, \boldsymbol{\gamma} \rangle \\ &\Rightarrow \quad \phi(\boldsymbol{x}) = \frac{\phi(\boldsymbol{\gamma})}{\|\boldsymbol{\gamma}\|^2}\langle \boldsymbol{x}, \boldsymbol{\gamma} \rangle \\ &\Rightarrow \quad \phi(\boldsymbol{x}) = \left\langle \boldsymbol{x}, \frac{\bar{\phi}(\boldsymbol{\gamma})}{\|\boldsymbol{\gamma}\|^2}\boldsymbol{\gamma} \right\rangle \\ &\Rightarrow \quad \phi(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{z} \rangle\end{aligned}$$

Furthermore, by the Cauchy-Schwarz inequality,

$$\phi(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{z} \rangle \leq \|\boldsymbol{z}\| \cdot \|\boldsymbol{x}\| \Rightarrow \|\phi\| = \|\boldsymbol{z}\|$$

As required.

□

Why does this theorem qualify as 'astounding and rather beautiful'? Simply because it confirms something we already knew intuitively for $\mathbb{R}^n$ – that every linear function(al) in a Hilbert Space corresponds to another element in that same space; in other words, the space is self-dual. In $\mathbb{R}^n$, this is obvious – we already know that every hyperplane can be represented by a vector in $\mathbb{R}^n$.

While we're on the topic, let's formally define a hyperplane.

Definition 4. (Hyperplane*)

A *hyperplane* $\mathscr{H}$ in a normed linear space X (for example, a Hilbert space) is a maximal proper affine set. In other words, if $\mathscr{H} \subseteq \mathscr{A}$ and $setA$ is affine, then either $\mathscr{A} = \mathscr{H}$ or $\mathscr{A} = X$.

Theorem 10. (Characterizing Hyperplanes*)

A set $\mathscr{H}$ is a hyperplane in a normed linear space X if and only if it is of the form

$$\mathscr{H} = \{\boldsymbol{x} \in X : \phi(\boldsymbol{x}) = c\}$$

where ϕ is a non-zero linear functional, and c is a scalar.

Furthermore, $\mathscr{H}$ is closed if and only if f is continuous.

Proof. See Luenberger, pp 129-130. □

The Hahn-Banach Theorem

We now prove two theorems which, in Hilbert space, are either trivial or already known. However, by proving these theorems for Hilbert spaces, we will familiarize ourselves with the rather difficult proofs involved, which will hopefully make the proofs easier to understand when we move on to the significantly more difficult case of Banach spaces.

First, a definition

> **Definition 5. (Seminorm*)** A real-valued function p defined on a real vector space X is said to be a *sublinear functional* or *seminorm* on X if it satisfies
>
> - $p(\boldsymbol{x}_1 + \boldsymbol{x}_2) \leq p(\boldsymbol{x}_1) + p(\boldsymbol{x}_2)$.
> - $p(\alpha \boldsymbol{x}) = \alpha p(\boldsymbol{x})$ for all $\alpha \geq 0$.
>
> In other words, it satisfies all the properties of a norm except for $p(\boldsymbol{x}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{0}$.

> **Theorem 11. (Special Case of Hahn-Banach)**
>
> Let $M \subseteq H$ be a closed subspace of a Hilbert space H, and let p be a seminorm on X.
>
> Let ϕ be a continuous linear functional on M satisfying $\phi(\boldsymbol{m}) \leq p(\boldsymbol{m})$ for all $\boldsymbol{m} \in M$. Then there exists a continuous linear extension of ϕ on H, Φ, such that $\Phi(\boldsymbol{x}) \leq p(\boldsymbol{x})$ for all $\boldsymbol{x} \in H$.

Proof. In a Hilbert space, the proof is almost trivial (though it took me many hours to realize just how trivial!) It simply relies on two facts

- p can trivially be shown to be convex[7]. By the first-order conditions for convexity, this means that there is always a hyperplane that lies below this function.
- By the Riesz-Frechet Theorem, both ϕ and Φ are hyperplanes that pass through $\boldsymbol{0}$:

  $$\phi(\boldsymbol{m}) = \langle \boldsymbol{m}, \boldsymbol{f} \rangle, \text{ for some } \boldsymbol{f} \in M$$

  $$\Phi(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{F} \rangle, \text{ for some } \boldsymbol{F} \in X$$

  Since ϕ underestimates p in M, its defining vector $\boldsymbol{f}$ must lie in the projection of $\partial p(\boldsymbol{0})$ on M.

  Thus, it is possible to find an $\boldsymbol{F} \in \partial p(\boldsymbol{0})$ such that $\boldsymbol{F} = \boldsymbol{f} + \boldsymbol{f}_\perp$, where $\boldsymbol{f}_\perp \in M^\perp$. The resulting Φ clearly agrees with $\boldsymbol{\phi}$ on M,[8] and also globally underestimates p.

[7] By the two properties of seminorms,

$$\begin{aligned} p(\lambda \boldsymbol{x} + \bar{\lambda} \boldsymbol{y}) &\leq p(\lambda \boldsymbol{x}) + p(\bar{\lambda} \boldsymbol{y}) \\ &= \lambda p(\boldsymbol{x}) + \bar{\lambda} p(\boldsymbol{y}) \end{aligned}$$

[8] ... because for any $\boldsymbol{m} \in M$,

$$\begin{aligned} \Phi(\boldsymbol{m}) &= \langle \boldsymbol{f} + \boldsymbol{f}_\perp, \boldsymbol{m} \rangle \\ &= \langle \boldsymbol{f}, \boldsymbol{m} \rangle \\ &= \phi(\boldsymbol{m}) \end{aligned}$$

□

Let us now get one final piece of mileage out of the H-B Theorem by using it to prove the Separating Hyperplane Theorem. To do this, we will need to introduce some new notation.

> **Definition 6. (Minkowski Functional*)** Let $\mathscr{K}$ be a convex set in a normed linear space X and suppose $\mathbf{0}$ is an interior point of $\mathscr{K}$. Then the *Minkowski Functional* p of $\mathscr{K}$ is defined on X by
>
> $$p(\boldsymbol{x}) = \inf_{r \geq 0} \left\{ r : \frac{\boldsymbol{x}}{r} \in \mathscr{K} \right\}$$
>
> Intuitively, $p(\boldsymbol{x})$ is the factor by which $\mathscr{K}$ must be expanded to include the vector $\boldsymbol{x}$.

We now prove some results relating to the Minkowski Functional.

> **Theorem 12. (Properties of the Minkowski Functional*)** The Minkowski Functional satisfies the following properties
>
> 1. $0 \leq p(\boldsymbol{x}) \leq \infty$ for all $\boldsymbol{x} \in X$.
> 2. $p(\alpha \boldsymbol{x}) = \alpha p(\boldsymbol{x})$ for all $\alpha > 0$.
> 3. $p(\boldsymbol{x}_1 + \boldsymbol{x}_2) \leq p(\boldsymbol{x}_1) + p(\boldsymbol{x}_2)$
> 4. p is continuous.
> 5. The closure of $\mathscr{K}$ and the interior of $\mathscr{K}$ are given by
>
> $$\mathrm{cl}(\mathscr{K}) = \{\boldsymbol{x} : p(\boldsymbol{x}) \leq 1\}$$
>
> $$\mathrm{int}(\mathscr{K}) = \{\boldsymbol{x} : p(\boldsymbol{x} < 1\}$$
>
> Note that points 2 and 3 above imply that p is a seminorm.

Proof. Most of these results are rather intuitive, and not particularly difficult to prove. See pp 131-132 of the Luenberger for the full proof. □

And finally...

> **Theorem 13. (Geometric Hahn-Banach*)**
>
> Let $\mathscr{K}$ be a convex set with non-empty interior in a Hilbert space H. Suppose $\mathscr{V}$ is an affine set in X (which could be a single point $\boldsymbol{x}_0$) that contains no interior points of $\mathscr{K}$. Then there is a closed hyperplane in H containing $\mathscr{V}$ but containing no interior point of $\mathscr{K}$.

> In other words, there exists an element $\boldsymbol{h}^* \in H$ such that
>
> $$\langle \boldsymbol{v}, \boldsymbol{h}^* \rangle = c \qquad \forall \boldsymbol{v} \in \mathscr{V}$$
>
> $$\langle \boldsymbol{k}, \boldsymbol{h}^* \rangle < c \qquad \forall \boldsymbol{k} \in \operatorname{int}(\mathscr{K})$$

Proof. First, consider that by translating our entire space, we can assume $\mathbf{0} \in \operatorname{int}(\mathscr{K})$ and $\mathbf{0} \notin \mathscr{V}$.

Then, let M be the smallest subspace that contains the affine space $\mathscr{V}$. Then in that subspace, the set $\mathscr{V}$ is a hyperplane (because it is a minimally affine set) and it doesn't contain 0 (because $\mathbf{0} \notin \mathscr{V}$). Thus, by our characterisation of hyperplanes (theorem 10), we have found a functional ϕ on M such that

$$\mathscr{V} = \{\boldsymbol{x} : \phi(\boldsymbol{x}) = 1\}$$

Now consider the Minkowski Functional of $\mathscr{K}$. Since no point in $\mathscr{V}$ is in the interior of $\mathscr{K}$, we have that $p(\boldsymbol{v}) \geq 1$ for all $\boldsymbol{v} \in \mathscr{V}$, and so[9]

$$p(\alpha \boldsymbol{v}) = \alpha p(\boldsymbol{v}) \geq \alpha \cdot 1 = \alpha \phi(\boldsymbol{v}) = \phi(\alpha \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathscr{V}$$

So in other words, since all vectors in M have the form $\alpha \boldsymbol{v}$ for some $\boldsymbol{v} \in V$, we find that

$$\phi(\boldsymbol{m}) \leq p(\boldsymbol{m}) \qquad \forall \boldsymbol{m} \in M$$

Now, apply the Hahn-Banach Theorem to find a functional Φ on X such that

$$\Phi(\boldsymbol{x}) \leq p(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in H$$

and let

$$\mathscr{H} = \{\boldsymbol{x} : \Phi(\boldsymbol{x}) = 1\}$$

Clearly, $\Phi(\boldsymbol{v}) = 1$ for all $\boldsymbol{v} \in \mathscr{V}$ (since Φ is an extension of ϕ), and since $\Phi(\boldsymbol{x}) \leq p(\boldsymbol{x})$, we have that $\Phi(\boldsymbol{k}) < 1$ for all $\boldsymbol{k} \in \operatorname{int}(\mathscr{K})$.

Finally, note that by Theorem 12, p is continuous, and so $\mathscr{H}$ is closed. □

■ □ ■

Question 4 (Primal and Dual Norm Minimization)

Consider the following two problems

[9] Strictly speaking, this argument only applies for $\alpha > 0$. However, if $\alpha < 0$, then $\phi(\alpha \boldsymbol{x}_0) = \alpha < 0$, and since $p(\boldsymbol{x}) \geq 0$, we still find that $\phi \leq p$.

1. Given a Hilbert space H, a vector $\boldsymbol{x} \in H$, and a set of vectors $\left\{\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(N)}\right\} \subset H$, find the best approximation of $\boldsymbol{x}$ as a linear combination of the $\boldsymbol{y}$.

2. Given a Hilbert space H, a set of vectors $\left\{\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(N)}\right\} \subset H$, and a set of constants $\{c_1, \cdots, c_N\}$, find the vector $\boldsymbol{x}$ of smallest norm that that satisfies $\langle \boldsymbol{x}, \boldsymbol{y}^{(i)} \rangle = c_i$ for all i.

Characterise the solution to both these problems.

Solution

Consider each problem

1. Let M be the subspace generated by the $\boldsymbol{y}^{(i)}$:

$$M = \left\{ \boldsymbol{m} = \sum_{i=1}^{N} \alpha_i \boldsymbol{y}^{(i)} \right\}$$

   This problem simply seeks the vector in M that is closest to $\boldsymbol{x}$. By the second part of the projection theorem (6), the unique optimal vector $\boldsymbol{m}_{opt}$ satisfies $\boldsymbol{x} - \boldsymbol{m}_{opt}$ is in $M^{\perp}$. This is equivalent to insisting that $\boldsymbol{x} - \boldsymbol{m}_{opt}$ be orthogonal to every $\boldsymbol{y}^{(i)}$.

   As such, letting

$$\boldsymbol{m}_{opt} = \sum_{i=1}^{N} \alpha_i \boldsymbol{y}^{(i)}$$

   We can characterize our solution $\boldsymbol{\alpha}$ by solving the following N equations

$$\left\langle \boldsymbol{x} - \textstyle\sum_{i=1}^{N} \alpha_i \boldsymbol{y}^{(i)}, \boldsymbol{y}^{(i)} \right\rangle = 0 \qquad \forall i$$

   Or equivalently

$$\begin{bmatrix} \langle \boldsymbol{y}^{(1)}, \boldsymbol{y}^{(1)} \rangle & \cdots & \langle \boldsymbol{y}^{(1)}, \boldsymbol{y}^{(N)} \rangle \\ \vdots & & \vdots \\ \langle \boldsymbol{y}^{(N)}, \boldsymbol{y}^{(1)} \rangle & \cdots & \langle \boldsymbol{y}^{(N)}, \boldsymbol{y}^{(N)} \rangle \end{bmatrix}^{\top} \boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{x} \cdot \boldsymbol{y}^{(1)} \\ \vdots \\ \boldsymbol{x} \cdot \boldsymbol{y}^{(N)} \end{bmatrix}$$

   The matrix is called the *Gram matrix* and is denoted $G(\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(N)})$. Provided G is invertible[10], the problem has a unique solution, which is characterized by this set of linear equations.

2. Let M be the subspace generated by the $\boldsymbol{y}^{(i)}$:

$$M = \left\{ \boldsymbol{m} = \sum_{i=1}^{N} \alpha_i \boldsymbol{y}^{(i)} \right\}$$

[10] This happens if and only if the vectors $\boldsymbol{y}^{(i)}$ are linearly independent – see Luenberger pp 56, Proposition 1

and note that

$$M^{\perp} = \Big\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{y}^{(i)} \rangle = 0 \text{ for all } i\Big\}$$

Now, consider the affine space

$$\mathscr{V} = \Big\{\boldsymbol{x} : \langle \boldsymbol{x}, \boldsymbol{y}^{(i)} \rangle = c_i \text{ for all } i\Big\}$$

Our problem is to find the vector $\boldsymbol{v}_{opt} \in \mathscr{V}$ of minimum norm. It shouldn't be too difficult, though, to convince yourself that

$$\mathscr{V} = \boldsymbol{v}_0 + M^{\perp}$$

where $\boldsymbol{v}_0$ satisfies $\langle \boldsymbol{v}_0, \boldsymbol{y}^{(i)} \rangle = c_i$ for all i.[11]

By the first form of the projection theorem (Theorem 6), the unique vector $\boldsymbol{v}_{opt}$ satisfies[12]

$$\boldsymbol{v}_{opt} \in (M^{\perp})^{\perp} = M$$

Thus, our solution has the form

$$\boldsymbol{v}_{opt} = \sum_{i=1}^{N} \alpha_i \boldsymbol{y}^{(i)}$$

Requiring this vector to be in $\mathscr{V}$ leads the following set of linear equations

$$G(\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(N)})^{\top} \boldsymbol{\alpha} = \boldsymbol{c}$$

These equations characterize our solution.

■ □ ■

Question 5 (Polynomial Optimization) ___

Part A

[2]Find the function $x(t) = a + bt$ that minimizes $\int_{-1}^{1} [t^2 - x(t)]^2 \mathrm{d}t$.

Solution

We will cast this problem as a minimum-norm problem in Hilbert space, and then use the projection theorem.

We first need to choose a Hilbert space in which to cast this problem. We choose $L_2[-1, 1]$, the space of square-integrable functions

[11] This works if the number of linearly independent $\boldsymbol{y}$ is less than n. If this isn't true, though, the $\boldsymbol{y}$ span the space, and this problem becomes silly.

[12] Equality follows by Theorem 7.

[2] Luenberger Chapter 3, problem 5.

on $[-1,1]$. This is a Hilbert space (see top of pp 49 in the Luenberger). The inner product in this space is

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \int_{-1}^{1} x(t)y(t)\mathrm{d}t$$

Re-stated in a more user-friendly way, our problem basically requires us to find the vector

$$\boldsymbol{y}(t) = t^2 + bt + a \in L^2[-1,1]$$

that has minimum norm.[13]

The set of vectors we seek is, in fact, the affine space $\mathscr{V}$ given by[14]

$$\mathscr{V} = (t^2) + M$$

where M is the subspace

$$M = \{(a + bt)\}$$

By the first form of the projection theorem (Theorem 6), the unique vector $\boldsymbol{v}_{opt} \in \mathscr{V}$ of minimum norm is orthogonal to M. In other words, it is a vector such that

$$\langle t^2 + b_{opt}t + a_{opt}, a + bt \rangle = 0 \qquad \forall a, b$$

Using our particular inner product, this becomes

$$\begin{aligned} \langle t^2 + b_{opt}t + a, a + bt \rangle &= \int_{-1}^{1} (t^2 + b_{opt}t + a_{opt})(a + bt)\mathrm{d}t \\ &= \int_{-1}^{1} bt^3 + (a + bb_{opt})t^2 + (ab_{opt} + a_{opt}b)t + aa_{opt}\mathrm{d}t \\ &= \frac{2}{3}(a + bb_{opt}) + 2aa_{opt} \end{aligned}$$

We need this to be 0 for all a and b. Clearly, this immediately leads to $b_{opt} = 0$. We then get

$$\frac{2}{3}a + 2aa_{opt} = 0 \qquad \forall a$$

which immediately leads to $a_{opt} = -1/3$.

Thus, the linear function that solves our problem is[15]

$$x(t) = \frac{1}{3}$$

We can verify this result explicitly. Consider that

$$\begin{aligned} \int_{-1}^{1} [t^2 - a - bt]^2 \mathrm{d}t &= \int_{-1}^{1} t^4 - 2bt^3 + (b^2 - 2a)t^2 + 2abt + a^2\mathrm{d}t \\ &= \frac{2}{5} + \frac{2}{3}(b^2 - 2a) + 2a^2 \end{aligned}$$

[13] In fact, this is not quite true – the norm has the form $\int_{-1}^{1}[t^2 + x(t)]^2\mathrm{d}t$, whereas we seek the quantity with a negative sign before x. Dealing with this is trivial, though – we'll simply flip the sign on the a and b we obtain by minimizing the norm.

[14] Remember here that $\boldsymbol{x}(t) = t^2$ is such a vector in our space, $L_2[-1,1]$.

[15] Note that we're flipping the sign on a_{opt}, as discussed above.

Clearly, setting $b = 0$ minimizes this expression with respect to b. We're then left to minimize

$$2a^2 - \frac{4}{3}a + \frac{2}{5}$$

Differentiating and setting to 0, we find that this occurs when

$$4a = \frac{4}{3} \Rightarrow a = \frac{1}{3}$$

as deduced above.[16]

[16] Those of you familiar with the Euler-Lagrange equations can also verify that in this case, the E-L equations

$$\frac{\partial f}{\partial x} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial f}{\partial x'}\right)$$

reduce to

$$2x(t) - 2t^2 = \text{constant}$$

. This immediately implies that $b = 0$, and optimizing over a gives the required result.

Part B

[3]Consider a function $\boldsymbol{x} \in L_2[0,1]$. Suppose that we wish to find a polynomial p of degree n or less which minimizes

$$\int_0^1 |x(t) - p(t)|^2 \mathrm{d}t$$

while satisfying

$$\int_0^1 p(t)\mathrm{d}t = 0$$

Show that this problem has a unique solution, and that it can be solved in the following two steps

- Find the polynomial q of degree n or less which minimizes

  $$\int_0^1 |x(t) - q(t)|^2 \mathrm{d}t$$

  (ignoring the constraint).

- Find the polynomial p of degree n or less which minimizes

  $$\int_0^1 |p(t) - q(t)|^2 \mathrm{d}t$$

  and satisfies

  $$\int_0^1 p(t)\mathrm{d}t = 0$$

Briefly describe how you could achieve each part.

Solution

First, consider that this is simply an norm-minimization problem in $L_2[0,1]$, with inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \int_0^1 x(t)y(t)\mathrm{d}t$$

Now, consider our two constraints

[3] Based on Luenberger Chapter 3, problem 6

- By requiring that our polynomial p be of degree n or less, we are requiring it to be in the subspace M generated by the elements $\left\{1, t, t^2, \cdots, t^n\right\}$ of $L_2[0,1]$:

  $$M = \left\{ \boldsymbol{m} = \sum_{i=0}^{n} \alpha_i t^i \right\}$$

  This is obviously closed.

- By requiring the integral of p to be 1 (and remembering that $1 \in L_2[0,1]$), we are requiring it to be in the subspace

  $$M' = \{\boldsymbol{p} : \langle \boldsymbol{p}, 1 \rangle = 0\}$$

  This is also closed, since the inner product is continuous.

As such, our optimization problem is simply

$$\begin{array}{ll} \min & \|\boldsymbol{p} - \boldsymbol{x}\| \\ \text{s.t.} & \boldsymbol{p} \in K = M \cap M' \end{array}$$

Now, consider that K is non-empty ($p = 1 \in K$), closed and affine (it is an intersection of two closed affine sets). Thus, by the projection theorem, there is a unique solution $\boldsymbol{p}_{opt}$ to this problem, which must satisfy

$$\boldsymbol{p}_{opt} - \boldsymbol{x} \in [M \cap M']^{\perp} \tag{1}$$

All we need to show is that a $\boldsymbol{p}$ that solves the successive problems in the question also satisfies this condition. To do this, consider that

- By the second form of the projection theorem, a vector $\boldsymbol{q}$ that solves the first sub-problem satisfies

  $$\boldsymbol{x} - \boldsymbol{q} \in M^{\perp}$$

- By the second form of the projection theorem, a vector $\boldsymbol{p}$ that solves the second sub-problem satisfies

  $$\boldsymbol{p} - \boldsymbol{q} \in (M')^{\perp} \Rightarrow \boldsymbol{p} - \boldsymbol{x} + (\boldsymbol{x} - \boldsymbol{q}) \in (M')^{\perp}$$

  And since we know that $\boldsymbol{x} - \boldsymbol{q} \in M^{\perp}$, this means that

  $$\boldsymbol{p} - \boldsymbol{x} \in [M \cap M']^{\perp}$$

  As in equation 1. Thus, solving each problem individually leads to an optimal solution for the global problem.

■ □ ■

Question 6 (A Control Problem)

[4]The angular shaft velocity ω of a d-c motor driven by a variable current source u is governed by the following first-order differential equation:

$$\dot{\omega}(t) + \omega(t) = u(t) \tag{2}$$

where $u(t)$ is the field current at time t. The energy expended by the field u in a time T is assumed to be proportional to

$$\int_0^T u(t)^2 \mathrm{d}t$$

The angular position θ of the motor is the time time integral of ω.

Initially, the motor is at rest and at a starting angle of 0:

$$\omega(0) = \theta(0) = 0$$

We seek to find the function u of minimum energy which rotates the shaft to $\theta = 1$ and brings it back to rest, within one second:

$$\theta(1) = 1 \qquad \omega(1) = 0$$

Solution

This point of this problem is to find the optimal $u(t)$. Since we're trying to minimize the energy, it makes sense to restrict ourselves to u's with finite energies – in other words, we can cast this problem as an optimization problem in $L_2[0,1]$, with norm

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \int_0^1 x(t)y(t)\mathrm{d}t$$

Clearly, we wish to minimize the norm. However, the constraints on θ and ω at $t = 1$ constrain the particular u we can choose. To write this constraints explicitly, we begin by solving differential equation 2). This can easily be done by separability. Multiplying the equation throughout by the integrating factor e^t, we obtain

$$\dot{\omega}(t)e^t + \omega(t)e^t = u(t)e^t$$

this can be re-written as

$$\frac{\mathrm{d}}{\mathrm{d}t}(\omega(t)e^t) = u(t)e^t$$

Solving, and noting that e^{t-1} is square integrable between 0 and 1 and is therefore in $L_2[0,1]$, we obtain

$$\omega(1) - \omega(0) = \int_0^1 u(t)e^{t-1}\mathrm{d}t = \langle u(t), e^{t-1} \rangle$$

[4] Based on Luenberger, pp 66 Example 1.

or, since $\omega(0) = 0$ and that we requir $\omega(1) = 0$

$$\langle u(t), e^{t-1} \rangle = \omega(1) = 0$$

We have therefore succeeded in expression our first constraint in a recognizable form.

For the second constraint, you may be tempted to write $\theta(1) - \theta(0) = \int_0^1 \omega(t)\mathrm{d}t$, but in this case this leads to a dead-end (because ω is itself expressed in integral form). Instead, use equation 2 directly to relate $\dot{\omega}$ and $\dot{\theta}$ as follows

$$\dot{\omega}(t) + \dot{\theta}(t) = u(t)$$

From the results above, we know that $\dot{\omega}(t) = u(t)e^{t-1}$, and so we obtain

$$\dot{\theta}(t) = u(t)(1 - e^{t-1})$$

As such,

$$\theta(1) - \theta(0) = \int_0^1 u(t)(1 - e^{t-1}) = \langle u(t), 1 - e^{t-1} \rangle$$

Bearing in mind that $\theta(0) = 0$ and that we require $\theta(1) = 1$, this becomes

$$\langle u(t), 1 - e^{t-1} \rangle = \theta(1) = 1$$

As such, our problem is simply to minimize the norm of u subject to the two constraints

$$\langle u(t), e^{t-1} \rangle = 0$$
$$\langle u(t), 1 - e^{t-1} \rangle = 1$$

By the result of an earlier question, the result will have the form of a linear combination of the two vectors in the inner products above. In other words,

$$u(t) = \alpha e^{t-1} + \alpha'(1 - e^{t-1})$$

or, more simply,

$$u(t) = \alpha_1 + \alpha_2 e^t$$

Evaluating the two constraints (this is trivial – it simply evolves evaluating each integral and obtaining two linear constraints in α_1 and α_2), we conclude that

$$u(t) = \frac{1}{3 - e}\left[1 + e - 2e^t\right]$$

■ □ ■

Daniel Guetta (daniel.guetta.com), January 2012

http://creativecommons.org/licenses/by/3.0.