ECE1655 Optimal Control

Bruce Francis

Course notes, Revised September 7, 2010

Contents

Chapter 1

Introduction

1.1 History

Optimal control is the subject within control theory where the control signals, or the controllers that generate them, optimize (minimize or maximize) some performance criterion. Let's look at some key developments, the first few being optimization problems that came before optimal control but influenced its development.

1. The **isoperimetric problem** is to find a closed curve of fixed length and maximum enclosed area: The solution is of course the circle. The study of this problem goes back to the ancient Greeks. The problem has the form

   $$\max_u J(u) \text{ such that } P(u) = l,$$

   where u is a closed curve, $J(u)$ denotes the enclosed area, $P(u)$ denotes the perimeter of u, and l is the given length. Thus J and P are functions that map curves to $\mathbb{R}$.

2. The **brachistochrone problem**, formulated by Bernoulli in 1696, is to find the shape of the curve down which a bead sliding from rest and accelerated by gravity will slip (without friction) from one point to another in the least time. This has the form

   $$\min_u J(u),$$

   where u denotes the curve and $J(u)$ is the time it takes for the bead to slide from the start point to the end point.

3. In the 1940s Wiener developed and solved an **optimal filtering problem**. A signal $s(t)$ is corrupted by additive noise $n(t)$ to produce the measured signal $y(t) = s(t) + n(t)$. In the simplest formulation the signals are zero-mean, stationary random processes and the goal is to find a filter with input $y(t)$ and output $\hat{s}(t)$, an estimate of $s(t)$, such that the variance of the error $\hat{s}(t) - s(t)$ is minimized. There is an equivalent deterministic problem known as $\mathcal{H}_2$-filtering,

4. Also in the 1940s, R. S. Phillips and colleagues extended Wiener's filtering problem to a control problem. The variance of a tracking error was the object to be minimized. This work initiated the development of $\mathcal{H}_2$-optimal control, the linear-quadratic regulator (LQR) problem being a special case.

5. The term **dynamic programming** was originally used in the 1940s by Richard Bellman to describe a process of solving sequential decision problems; a typical problem is to find a minimum-cost path through a graph. Bellman wrote an influential book, *Dynamic Programming*, published in 1957. The procedure is widely used, for example in Viterbi decoding.

6. In 1962 the very influential book *The Mathematical Theory of Optimal Processes*, by L. S. Pontryagin et al., appeared. The approach to constrained optimization problems is called the **maximum principle**.

7. In the late 1970s Zames posed the question, are classical frequency-domain feedback design methods (e.g., lead/lag compensation) optimal for some appropriate criterion? From this came $\mathcal{H}_\infty$-optimal control.

8. Many interesting problems can be formulated as **distance problems**. As an example, given a matrix A and a vector b, solve $Ax = b$, or, if it is not solvable, minimize the error $\|b - Ax\|$; that is, find y closest to b where y must lie in the set

$$\mathcal{V} = \{y; (\exists x) y = Ax\}.$$

This set (a subspace) equals the span of the columns of A. So the problem is to find the vector in $\mathcal{V}$ that is closest to b. A more interesting problem, the Nehari problem, is to find the stable LTI system that is closest to a given unstable LTI system.

1.2 What We Study

The course notes are divided into three parts:

Part I is a review, first of linear algebra and linear systems. These notes are taken from ECE557. You are expected to know this material and we won't do it in class. We begin with Chapter 3, the part of calculus concerning optimization in $\mathbb{R}^n$. We'll review the method of Lagrange multipliers.

Part II presents three interesting, old topics. First, the calculus of variations, in particular the brachistochrone problem. Second, the maximum principle, though just the rudiments. And finally, dynamic programming.

Part III is the meat of the course, $\mathcal{H}^2$ and $\mathcal{H}^\infty$ optimal control.

When we solve problems and get control laws that are optimal for a certain criterion, ask yourself these questions: How can the control signal be implemented? What sensors would be required? Is it a feedback controller? Is this controller robust to sensor noise and modeling errors?

1.3 Theorems and Proofs

The course is mathematical: The language is mathematical, the concepts are stated as definitions, the results are stated as theorems, and proofs are rigorous. Engineering students aren't used to this. They take calculus and linear algebra etc. as undergraduates, they see epsilon and delta, but they have difficulty working with them. One of the goals of this course is to improve that situation. Please download and read **Elements of Mathematical Style**:

http://sites.google.com/site/brucefranciscontact/

Why is control theory so mathematical? Why is it in the theorem/proof format? The answer is, for clarity. Results are written formally and proved so that they can be understood by everyone and so that there can be no doubt about what they mean. The following (real) email correspondence between a physicist and me emphasizes this point.

Me: On skimming through [your paper], I don't see any formal theorem statements, with proofs. So may I ask, are there mathematical results? That would help an outsider like me (of course, you didn't write the paper for outsiders).

Physicist: The paper is not written in the formal language of "theorems" and "proofs." Certainly, though, there are mathematical results reported that describe the physical properties of complex modes. ... In general, though, in the physics literature papers are not written as mathematical papers in the form of theorems, lemmas, and proofs.

Me: I wonder how physicists check each other's work without explicit assumptions, mathematical statements, and proofs. For example, your paper talks about "casual suggestions in the literature." Had they instead been rigorous statements, one could have checked them to be true or not. ...Your model in the appendices involves a limit (homogeneous limit), so I guess something converges and can be proved to converge in a precise way. Or is it that one doesn't actually prove convergence but instead verifies by experiment?

Physicist: This is a philosophical question really; for example one cannot prove Maxwell's equations mathematically. Mathematics is a tool, it cannot describe physical reality. So in the narrow area of Electromagnetics, whether something is true or not boils down to whether it satisfies Maxwell's equations.

1.4 Problems

1. Give an example of a subset of $\mathbb{R}^2$ that is the graph of a function $f : \mathbb{R} \longrightarrow \mathbb{R}$. Give an example of a subset of $\mathbb{R}^2$ that is not the graph of any function $f : \mathbb{R} \longrightarrow \mathbb{R}$.

2. Let U, V be two sets and let S be a subset of $U \times V$. Write in logic notation the condition for S to be the graph of a function from U to V.

3. Write the truth tables for $(P \wedge Q) \Rightarrow R$ and $P \wedge (Q \Rightarrow R)$.

4. Consider the differential equation $\dot{x} = f(x)$, where x is a vector. Assume $x = 0$ is an equilibrium point, i.e., $f(0) = 0$. We say the origin is a **stable** equilibrium point if

   $$(\forall \varepsilon > 0)(\exists \delta > 0)(\forall x(0))\|x(0)\| < \delta \Rightarrow (\forall t \geq 0)\|x(t)\| < \varepsilon.$$

   In words, for every $\varepsilon > 0$ there exists $\delta > 0$ such that if the state starts in the δ-ball, it will remain forever in the ε-ball. Write in logic notation the definition that the origin is not stable. Say this in words. Using your definition, prove that for $\dot{x} = x$ the origin is not stable.

5. Consider the linear equation $Ax = b$, where A is a matrix, not assumed to be square, and b, x are vectors. A necessary and sufficient condition for the equation to be solvable (for x to exist) is

   $$\text{rank} \begin{bmatrix} A & b \end{bmatrix} = \text{rank}\, A.$$

That is (necessity)

$$(\exists x) Ax = b \implies \text{rank} \begin{bmatrix} A & b \end{bmatrix} = \text{rank } A$$

and (sufficiency)

$$\text{rank} \begin{bmatrix} A & b \end{bmatrix} = \text{rank } A \implies (\exists x) Ax = b.$$

Write these two logic statements in contrapositive form.

6. Write a logic statement that there is a unique solution to the equation $Ax = b$, i.e., there exists a solution and it is unique.

7. Consider the polynomial $p(s) = s^3 + a_2 s^2 + a_1 s + a_0$, with real coefficients. Consider the two conditions: 1) The roots of $p(s)$ have negative real parts; 2) The coefficients a_i are all positive. Which condition is necessary for the other? Is it sufficient?

8. The set of integers, $0, \pm 1, \pm 2, \pm 3, \dots$, is denoted $\mathbb{Z}$. One way to say an integer is even is that it is a multiple of 2. Thus, if x is an integer, the following says it is even:

   $$(\exists y \in \mathbb{Z}) x = 2y.$$

   Write the following statements in logic notation, using only the set $\mathbb{Z}$:

   (a) Not every integer is even.

   (b) Not every even integer is a multiple of 4.

   (c) Every integer that is a multiple of 2 and a multiple of 3 is also a multiple of 6.

9. Prove rigorously that, if m and n are coprime integers (their greatest common divisor equals 1), then every integer that is a multiple of m and a multiple of n is also a multiple of mn.

10. If a single-input, single-output plant has a transfer function with a right half-plane zero, there's a maximum achievable gain margin. Make this into a theorem statement.

11. For each of the following statements, state if it is true or false.

    (a) A discrete-time signal $u[k]$ that converges to zero is bounded.

    (b) A discrete-time signal $u[k]$ converges to zero only if it is bounded.

    (c) A continuous-time signal $u(t)$ that converges to zero is bounded.

    (d) If e^{At} converges to zero, then it's bounded.

    (e) The negation of

    if $u(t)$ converges to zero, then $u(t)$ is bounded

    is

    if $u(t)$ is bounded, then $u(t)$ converges to zero.

    (f) The negation of

    Alice can pass ECE1655 only if she's brilliant and she works hard

    is

> Alice can pass ECE1655 and either she's not brilliant or she doesn't work hard.

(g) The statement

> Alice can pass ECE1655 only if she's brilliant and she works hard

is equivalent to

> if either Alice is not brilliant or she doesn't work hard, then she can't pass ECE1655.

(h) A necessary condition for the origin of $\dot{x} = Ax$ to be stable is that all eigenvalues of A satisfy $\text{Re}\,\lambda \leq 0$.

(i) A sufficient condition for the origin of $\dot{x} = Ax$ to be stable is that all eigenvalues of A satisfy $\text{Re}\,\lambda \leq 0$.

(j) A necessary condition for the origin of $\dot{x} = Ax$ to be asymptotically stable is that all eigenvalues of A satisfy $\text{Re}\,\lambda < 0$.

(k) A sufficient condition for the origin of $\dot{x} = Ax$ to be asymptotically stable is that all eigenvalues of A satisfy $\text{Re}\,\lambda < 0$.

(l) Every bounded input to the system with transfer function $\dfrac{s-1}{s^2-1}$ results in a bounded output.

(m) Every nonzero bounded input to the system with transfer function $\dfrac{1}{s-1}$ results in an unbounded output.

12. Many results in optimal control are in the form of providing a necessary condition for optimality. For example, the maximum principle and the Hamilton-Jacobi-Bellman equation. This problem urges you to understand what necessary condition means.

(a) Suppose x is a variable in an optimization problem, $P(x)$ is the proposition that x is optimal, and $Q(x)$ is some other proposition (it's going to be the necessary condition). Then

$$(\exists x)P(x)$$

is true means there exists an optimal solution. Discuss the meanings of the following statements. Is any the right one, i.e., the one we want in a theorem statement about optimality?

$$(\exists x)P(x) \Longrightarrow Q(x)$$

$$(\exists x)[P(x) \Longrightarrow Q(x)]$$

$$(\exists x)P(x), \ (\forall x)[P(x) \Longrightarrow Q(x)].$$

(b) Now consider the function $f(u) = -u^2$. Suppose $P(x)$ means that f is maximized at the point x and $Q(x)$ means that the derivative of f equals zero at x. Which if any of the three statements is true?

(c) Repeat for $f(u) = -u^3$.

13. The physicist's final email is contradictory: Mathematics can't describe physical reality, yet Maxwell's equations, which are mathematics, do. I believe physicists sometimes forget that all of physics is based on models. The models of electromagnetics—the concepts of electric and magnetic fields, charged particles and the forces on them, the differential equations and boundary conditions, and so on—are abstract concepts and relationships that we use to help us understand how we perceive phenomena. In fact all these concepts are mathematical: A force is a vector-valued function of space and time, etc. We have used these models to build technology, and in that sense the models have been wildly successful. But one, and especially a control engineer, must not forget that models are not real.

    Imagine a real battery connected to a small real DC motor sitting on some real lab bench somewhere on the planet Earth. Write a brief essay that distinguishes between a model and reality. In particular, answer this question: Does there exist a sequence of models, of ever increasing complexity, that converges to reality?

Chapter 2

Linear Algebra

This is a background chapter on linear algebra: subspaces, matrix representations, linear matrix equations, and invarianft subspaces. The material is from ECE557. If you took ECE410, then some of this material will be new.

2.1 Brief Review

In this brief section we review these concepts/results: $\mathbb{R}^n$, linear independence of a set of vectors, span of a set of vectors, subspace, basis for a subspace, rank of a matrix, existence and uniqueness of a solution to $Ax = b$ where A is not necessarily square, inverse of a matrix, invertibility. If you remember them (and I hope you do), skip to the next section.

The symbol $\mathbb{R}^n$ stands for the vector space of n-tuples, i.e., ordered lists of n real numbers.

A set of vectors $\{v_1, \ldots, v_k\}$ in $\mathbb{R}^n$ is **linearly independent** if none is a linear combination of the others. One way to check this is to write the equation

$$c_1 v_1 + \cdots + c_k v_k = 0$$

and then try to solve for the c_i's. The set is linearly independent iff the only solution is $c_i = 0$ for every i.

The **span** of $\{v_1, \ldots, v_k\}$, denoted $\mathrm{Span}\{v_1, \ldots, v_k\}$, is the set of all linear combinations of these vectors.

A **subspace** $\mathcal{V}$ of $\mathbb{R}^n$ is a subset of $\mathbb{R}^n$ that is also a vector space in its own right. This is true iff these two conditions hold: If x, y are in $\mathcal{V}$, then so is $x + y$; if x is in $\mathcal{V}$ and c is a scalar, then cx is in $\mathcal{V}$. Thus $\mathcal{V}$ is closed under the operations of addition and scalar multiplication. In $\mathbb{R}^3$ the subspaces are the lines through the origin, the planes through the origin, the whole of $\mathbb{R}^3$, and the set consisting of only the zero vector.

A **basis** for a subspace is a set of linearly independent vectors whose span equals the subspace. The number of elements in a basis is the **dimension** of the subspace.

The **rank** of a matrix is the dimension of the span of its columns. This can be proved to equal the dimension of the span of its rows.

The equation $Ax = b$ has a solution iff b belongs to the span of the columns of A, equivalently

$$\mathrm{rank}\, A = \mathrm{rank} \begin{bmatrix} A & b \end{bmatrix}.$$

When a solution exists, it is unique iff the columns of A are linearly independent, that is, the rank of A equals its number of columns.

The **inverse** of a square matrix A is a matrix B such that $BA = I$. If this is true, then $AB = I$. The inverse is unique and we write A^{-1}. A square matrix A is invertible iff its rank equals its dimension (we say "A has full rank"); equivalently, its determinant is nonzero. The inverse equals the adjoint divided by the determinant.

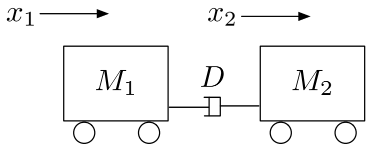
2.2 Eigenvalues and Eigenvectors

Now we turn to $\dot{x} = Ax$. The time evolution of $x(t)$ can be understood from the eigenvalues and eigenvectors of A—a beautiful connection between dynamics and algebra. Recall that the eigenvalue equation is

$$Av = \lambda v.$$

Here λ is a real or complex number and v is a nonzero real or complex vector; λ is an eigenvalue and v a corresponding eigenvector. The eigenvalues of A are unique but the eigenvectors are not: If v is an eigenvector, so is cv for any real number $c \neq 0$. The **spectrum** of A, denoted $\sigma(A)$, is its set of eigenvalues. The spectrum consists of n numbers, in general complex, and they are equal to the zeros of the characteristic polynomial $\det(sI - A)$.

Example Consider two carts and a dashpot like this:



Take $D = 1$, $M_1 = 1$, $M_2 = 1/2$, $x_3 = \dot{x}_1$, $x_4 = \dot{x}_2$. You can derive that the model is $\dot{x} = Ax$, where

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 2 & -2 \end{bmatrix}.$$

The characteristic polynomial of A is $s^3(s+3)$, and therefore

$$\sigma(A) = \{0, 0, 0, -3\}.$$

□

The equation $Av = \lambda v$ says that the action of A on an eigenvector is very simple—just multiplication by the eigenvalue. Likewise, the motion of $x(t)$ starting at an eigenvector is very simple.

Lemma 2.1 *If $x(0)$ is an eigenvector v of A and λ the corresponding eigenvalue, then $x(t) = \mathrm{e}^{\lambda t} v$. Thus $x(t)$ is an eigenvector too for every t.*

Proof The initial-value problem

$$\dot{x} = Ax, \quad x(0) = v$$

has a unique solution—this is from differential equation theory. So all we have to do is show that $\mathrm{e}^{\lambda t}v$ satisfies both the initial condition and the differential equation, for then $\mathrm{e}^{\lambda t}v$ must be the solution $x(t)$. The initial condition is easy:

$$\mathrm{e}^{\lambda t}v\Big|_{t=0} = v.$$

And for the differential equation,

$$\frac{d}{dt}(\mathrm{e}^{\lambda t}v) = \mathrm{e}^{\lambda t}\lambda v = \mathrm{e}^{\lambda t}Av = A(\mathrm{e}^{\lambda t}v).$$

□

The result of the lemma extends to more than one eigenvalue. Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of A and let $v_1, \dots, v_n$ be corresponding eigenvectors. Suppose the initial state $x(0)$ can be written as a linear combination of the eigenvectors:

$$x(0) = c_1 v_1 + \cdots + c_n v_n.$$

This is certainly possible for every $x(0)$ if the eigenvectors are linearly independent. Then the solution satisfies

$$x(t) = c_1 \mathrm{e}^{\lambda_1 t} v_1 + \cdots + c_n \mathrm{e}^{\lambda_n t} v_n.$$

This is called a **modal expansion** of $x(t)$.

Example

$$A = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix}, \quad \lambda_1 = 0, \ \lambda_2 = -3, \quad v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad v_2 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

Let's say $x(0) = (0, 1)$. The equation

$$x(0) = c_1 v_1 + c_2 v_2$$

is equivalent to

$$x(0) = Vc,$$

where V is the 2×2 matrix with columns v_1, v_2 and c is the vector (c_1, c_2). Solving gives $c_1 = c_2 = 1/3$. So

$$x(t) = \frac{1}{3}v_1 + \frac{1}{3}\mathrm{e}^{-3t}v_2$$

□

The case of complex eigenvalues is only a little complicated. If λ_1 is a complex eigenvalue, some other, say λ_2, is its complex conjugate: $\lambda_2 = \overline{\lambda_1}$. The two eigenvectors, v_1 and v_2, can be taken to be complex conjugates too (easy proof). Then if $x(0)$ is real and we solve

$$x(0) = c_1 v_1 + c_2 v_2,$$

we'll find that c_1, c_2 are complex conjugates as well. Thus the equation will look like

$$x(0) = c_1 v_1 + \overline{c_1 v_2} = 2\Re\ (c_1 v_1),$$

where $\Re$ denotes real part.

Example

$$A = \left[\begin{array}{rr} 0 & -1 \\ 1 & 0 \end{array} \right], \quad \lambda_1 = j,\ \lambda_2 = -j, \quad v_1 = \left[\begin{array}{c} 1 \\ -j \end{array} \right], \quad v_2 = \left[\begin{array}{c} 1 \\ j \end{array} \right]$$

Suppose $x(0) = (0, 1)$. Then $c_1 = j/2$, $c_2 = -j/2$ and

$$x(t) = 2\Re\ \left(c_1 e^{\lambda_1 t} v_1 \right) = \Re\ \left(j e^{jt} \left[\begin{array}{c} 1 \\ -j \end{array} \right] \right) = \left[\begin{array}{c} -\sin t \\ \cos t \end{array} \right].$$

□

2.3 The Jordan Form

Now we turn to the structure theory of a matrix related to its eigenvalues. It's convenient to introduce a term, the **kernel** of a matrix A. Kernel is another name for nullspace. Thus Ker A is the set of all vectors x such that $Ax = 0$; that is, Ker A is the solution space of the homogeneous equation $Ax = 0$. Notice that the zero vector is always in the kernel. If A is square, then Ker A is the zero subspace, and we write Ker $A = 0$, iff 0 is not an eigenvalue of A, equivalently, A is invertible. If 0 *is* an eigenvalue, then Ker A equals the span of all the eigenvectors corresponding to this eigenvalue; we say Ker A is the **eigenspace** corresponding to the eigenvalue 0. More generally, if λ is an eigenvalue of A the corresponding eigenspace is the solution space of $Av = \lambda v$, that is, of $(A - \lambda I)v = 0$, that is, Ker $(A - \lambda I)$.

Let's begin with the simplest case, where A is 2×2 and has 2 distinct eigenvalues, λ_1, λ_2. You can show (this is a good exercise) that there are then 2 linearly independent eigenvectors, say v_1, v_2 (maybe complex vectors). The equations

$$Av_1 = \lambda_1 v_1, \quad Av_2 = \lambda_2 v_2$$

are equivalent to the matrix equation

$$A \left[\begin{array}{cc} v_1 & v_2 \end{array} \right] = \left[\begin{array}{cc} v_1 & v_2 \end{array} \right] \left[\begin{array}{cc} \lambda_1 & 0 \\ 0 & \lambda_2 \end{array} \right],$$

that is, $AV = VA_{JF}$, where

$$V = \left[\begin{array}{cc} v_1 & v_2 \end{array} \right], \quad A_{JF} = \text{diag}\ (\lambda_1, \lambda_2).$$

The latter matrix is the **Jordan form** of A. It is unique up to reordering of the eigenvalues. The mapping $A \longmapsto A_{JF} = V^{-1}AV$ is called a similarity transformation. Example:

$$A = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & -1 \\ 1 & 2 \end{bmatrix}, \quad A_{JF} = \begin{bmatrix} 0 & 0 \\ 0 & -3 \end{bmatrix}.$$

Corresponding to the eigenvalue $\lambda_1 = 0$ is the eigenvector $v_1 = (1, 1)$, the first column of V. All other eigenvectors corresponding to λ_1 have the form cv_1, $c \neq 0$. We call the subspace spanned by v_1 the eigenspace corresponding to λ_1. Likewise, $\lambda_2 = -3$ has a one-dimensional eigenspace.

These results extend from $n = 2$ to general n. Note that in the preceding result we didn't actually need distinctness of the eigenvalues — only linear independence of the eigenvectors.

Theorem 2.1 *The Jordan form of A is diagonal, i.e., A is diagonalizable by similarity transformation, iff A has n linearly independent eigenvectors. A sufficient condition is n distinct eigenvalues.*

The great thing about diagonalization is that the equation $\dot{x} = Ax$ can be transformed via $w = V^{-1}x$ into $\dot{w} = A_{JF}w$, that is, n **decoupled** equations:

$$\dot{w}_i = \lambda_i w_i, \quad i = 1, \ldots, n.$$

The latter equations are trivial to solve:

$$w_i(t) = \mathrm{e}^{\lambda_i t} w_i(0), \quad i = 1, \ldots, n.$$

Now we look at how to construct the Jordan form when there are not n linearly independent eigenvectors. We start where A has only 0 as an eigenvalue.

Nilpotent matrices

Consider

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}. \tag{2.1}$$

For both of these matrices, $\sigma(A) = \{0, 0, 0\}$. For the first matrix, the eigenspace Ker A is two-dimensional and for the second matrix, one-dimensional. These are examples of nilpotent matrices: A is **nilpotent** if $A^k = 0$ for some $k \geq 1$. The following statements are equivalent:

1. A is nilpotent.

2. All its eigs are 0.

3. Its characteristic polynomial is s^n.

4. It is similar to a matrix of the form (2.1), where all elements are 0's, except 0's or 1's on the first diagonal above the main one. This is called the Jordan form of the nilpotent matrix.

Example Suppose A is 3×3 and $A = 0$. Then of course it's already in Jordan form,

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

□

Example Here we do an example of transforming a nilpotent matrix to Jordan form. Take

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -1 & -1 \end{bmatrix}.$$

The rank of A is 3 and hence the kernel has dimension 2. We can compute that

$$A^2 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad A^3 = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad A^4 = 0.$$

Take any vector v_5 in Ker $A^4 = \mathbb{R}^5$ that is not in Ker A^3, for example,

$$v_5 = (0,0,0,0,1).$$

Then take

$$v_4 = Av_5, \quad v_3 = Av_4, \quad v_2 = Av_3.$$

We get

$$v_4 = (0,0,0,1,-1) \in \text{Ker } A^3, \quad \notin \text{Ker } A^2$$

$$v_3 = (0,1,0,0,0) \in \text{Ker } A^2, \quad \notin \text{Ker } A$$

$$v_2 = (1,-1,0,0,0) \in \text{Ker } A.$$

Finally, take $v_1 \in$ Ker A, linearly independent of v_2, for example,

$$v_1 = (0,0,1,0,0).$$

Assemble $v_1, \ldots, v_5$ into the columns of V. Then

$$V^{-1}AV = A_{JF} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

This is block diagonal, like this:

$$\left[\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

□

In general, the Jordan form of a nilpotent matrix has 0 in each entry except possibly in the first diagonal above the main diagonal which may have some 1s.

A nilpotent matrix has only the eigenvalue 0. Now consider a matrix A that has only one eigenvalue, λ, i.e.,

$$\det(sI - A) = (s - \lambda)^n.$$

To simplify notation, suppose $n = 3$. Letting $r = s - \lambda$, we have

$$\det[rI - (A - \lambda I)] = r^3,$$

i.e., $A - \lambda I$ has only the zero eigenvalue, and hence $A - \lambda I =: N$, a nilpotent matrix. So the Jordan form of N must look like

$$\begin{bmatrix} 0 & \star & 0 \\ 0 & 0 & \star \\ 0 & 0 & 0 \end{bmatrix},$$

where each star can be 0 or 1, and hence the Jordan form of A is

$$\begin{bmatrix} \lambda & \star & 0 \\ 0 & \lambda & \star \\ 0 & 0 & \lambda \end{bmatrix}, \tag{2.2}$$

To recap, if A has just one eigenvalue, λ, then its Jordan form is $\lambda I + N$, where N is a nilpotent matrix in Jordan form.

An extension of this analysis results in the **Jordan form** in general. Suppose A is $n \times n$ and $\lambda_1, \ldots, \lambda_p$ are the distinct eigenvalues of A and $m_1, \ldots, m_p$ are their multiplicities; that is, the characteristic polynomial is

$$\det(sI - A) = (s - \lambda_1)^{m_1} \cdots (s - \lambda_p)^{m_p}.$$

Then A is similar to

$$A_{JF} = \begin{bmatrix} A_1 & & \\ & \ddots & \\ & & A_p \end{bmatrix},$$

where A_i is $m_i \times m_i$ and it has only the eigenvalue λ_i. Thus A_i has the form $\lambda_i I + N_i$, where N_i is a nilpotent matrix in Jordan form. Example:

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 2 & -2 \end{bmatrix}$$

As we saw, the spectrum is $\sigma(A) = \{0, 0, 0, -3\}$. Thus the Jordan form must be of the form

$$A_{JF} = \begin{bmatrix} 0 & \star & 0 & 0 \\ 0 & 0 & \star & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -3 \end{bmatrix}.$$

Since A has rank 2, so does A_{JF}. Thus only one of the stars is 1. Either is possible, for example,

$$A_{JF} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -3 \end{bmatrix}.$$

This has two "Jordan blocks":

$$A_{JF} = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_2 = -3.$$

2.4 The Transition Matrix

For a square matrix M, the exponential e^M is defined as

$$\mathrm{e}^M := I + M + \frac{1}{2!}M^2 + \frac{1}{3!}M^3 + \cdots .$$

The matrix e^M is not the same as the component-wise exponential of M. Facts:

1. e^M is invertible for every M, and $(\mathrm{e}^M)^{-1} = \mathrm{e}^{-M}$.
2. $\mathrm{e}^{M+N} = \mathrm{e}^M \mathrm{e}^N$ if M and N commute, i.e., $MN = NM$.

The matrix function $t \longmapsto \mathrm{e}^{tA} : \mathbb{R} \to \mathbb{R}^{n\times n}$ is then defined and is called the **transition matrix** associated with A. It has the properties

1. $\mathrm{e}^{tA}|_{t=0} = I$
2. e^{tA} and A commute.
3. $\mathrm{e}^{t(A+B)} = \mathrm{e}^{tA}\mathrm{e}^{tB}$ iff A and B commute.
4. $\dfrac{d}{dt}\mathrm{e}^{tA} = A\mathrm{e}^{tA} = \mathrm{e}^{tA}A$.

Moreover, the solution of

$$\dot{x} = Ax, \quad x(0) = x_0$$

is $x(t) = \mathrm{e}^{tA}x_0$. So e^{tA} maps the state at time 0 to the state at time t. In fact, it maps the state at any time t_0 to the state at time $t_0 + t$.

On computing the transition matrix

via the Jordan form If one can compute the Jordan form of A, then e^{tA} can be written in closed form, as follows. The equation

$$AV = VA_{JF}$$

implies

$$A^2V = AVA_{JF} = VA_{JF}^2.$$

Continuing in this way gives

$$A^kV = VA_{JF}^k,$$

and then

$$\mathrm{e}^{At}V = V\mathrm{e}^{A_{JF}t},$$

so finally

$$\mathrm{e}^{At} = V\mathrm{e}^{A_{JF}t}V^{-1}.$$

The matrix exponential $\mathrm{e}^{A_{JF}t}$ is easy to write down. For example, suppose there's just one eigenvalue, so $A_{JF} = \lambda I + N$, N nilpotent, $n \times n$. Then

$$\begin{aligned} \mathrm{e}^{A_{JF}t} &= \mathrm{e}^{\lambda t}\mathrm{e}^{Nt} \\ &= \mathrm{e}^{\lambda t}\left(I + Nt + N^2\frac{t^2}{2!} + \cdots + N^{n-1}\frac{t^{n-1}}{(n-1)!}\right). \end{aligned}$$

via Laplace transforms Taking Laplace transforms of

$$\dot{x} = Ax, \quad x(0) = x_0$$

gives

$$sX(s) - x_0 = AX(s).$$

This yields

$$X(s) = (sI - A)^{-1}x_0.$$

Comparing

$$x(t) = \mathrm{e}^{tA}x_0, \quad X(s) = (sI - A)^{-1}x_0$$

shows that e^{tA}, $(sI - A)^{-1}$ are Laplace transform pairs. So one can get e^{tA} by finding the matrix $(sI - A)^{-1}$ and then taking the inverse Laplace transform of each element.

2.5 Stability

The concept of stability is fundamental in control engineering. Here we look at the scenario where the system has no input, but its state has been perturbed and we want to know if the system will recover.

The maglev example is a good one to illustrate this point. Suppose a feedback controller has been designed to balance the ball's position at 1 cm below the magnet. Suppose if the ball is placed at **precisely** 1 cm it will stay there; that is, the 1 cm location is a closed-loop equilibrium point. Finally, suppose there is a temporary wind gust that moves the ball away from the 1 cm position. The stability questions are, will the ball move back to the 1 cm location; if not, will it at least stay near that location?

So consider

$$\dot{x} = Ax.$$

Obviously if $x(0) = 0$, then $x(t) = 0$ for all t. We say the origin is an **equilibrium point**—if you start there, you stay there. Equilibrium points can be stable or not. While there are more elaborate and formal definitions of stability for the above homogeneous system, we choose the following two: The origin is **asymptotically stable** if $x(t) \longrightarrow 0$ as $t \longrightarrow \infty$ for all $x(0)$. The origin is **stable** if $x(t)$ remains bounded as $t \longrightarrow \infty$ for all $x(0)$. Since $x(t) = \mathrm{e}^{At}x(0)$, the origin is asymptotically stable iff every element of the matrix e^{At} converges to zero, and is stable iff every element of the matrix e^{At} remains bounded as $t \longrightarrow \infty$. Of course, asymptotic stability implies stability.

Asymptotic stability is relatively easy to characterize. Using the Jordan form, one can prove this very important result, where $\Re$ denotes "real part":

Theorem 2.2 *The origin is asymptotically stable iff the eigenvalues of A all satisfy $\Re\,\lambda < 0$.*

Let's say the matrix A is **stable** if its eigenvalues satisfy $\Re\,\lambda < 0$. Then the origin is asymptotically stable iff A is stable.

Now we turn to the more subtle property of stability. We'll do some examples, and we may as well have A in Jordan form.

Consider the nilpotent matrix

$$A = N = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Obviously, $x(t) = x(0)$ for all t and so the origin is stable. By contrast, consider

$$A = N = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Then

$$\mathrm{e}^{Nt} = I + tN,$$

which is unbounded and so the origin is not stable. This example extends to the $n \times n$ case: If A is nilpotent, the origin is stable iff $A = 0$.

Here's the test for stability in general in terms of the Jordan form of A:

$$A_{JF} = \begin{bmatrix} A_1 & & \\ & \ddots & \\ & & A_p \end{bmatrix}.$$

Recall that each A_i has just one eigenvalue, λ_i, and that $A_i = \lambda_i I + N_i$, where N_i is a nilpotent matrix in Jordan form.

Theorem 2.3 *The origin is stable iff the eigenvalues of A all satisfy $\Re\,\lambda \leq 0$ and for any eigenvalue with $\Re\,\lambda_i = 0$, the nilpotent matrix N_i is zero, i.e., A_i is diagonal.*

Here's an example with complex eigenvalues:

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad A_{JF} = \begin{bmatrix} j & 0 \\ 0 & -j \end{bmatrix}.$$

The origin is stable since there are two 1×1 Jordan blocks. Now consider

$$A = \begin{bmatrix} 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

The eigenvalues are $j, j, -j, -j$ and so the Jordan form must look like

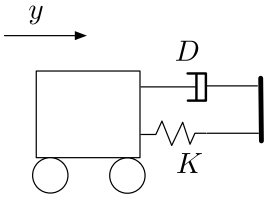
$$A_{JF} = \begin{bmatrix} j & \star & 0 & 0 \\ 0 & j & 0 & 0 \\ 0 & 0 & -j & \star \\ 0 & 0 & 0 & -j \end{bmatrix}.$$

Since the rank of $A - jI$ equals 3, the upper star is 1; since the rank of $A + jI$ equals 3, the lower star is 1. Thus

$$A_{JF} = \begin{bmatrix} j & 1 & 0 & 0 \\ 0 & j & 0 & 0 \\ 0 & 0 & -j & 1 \\ 0 & 0 & 0 & -j \end{bmatrix}.$$

Since the Jordan blocks are not diagonal, the origin is not stable.

Example Consider the cart-spring-damper system



The equation is

$$M\ddot{y} + D\dot{y} + Ky = 0.$$

Defining $x = (y, \dot{y})$, we have $\dot{x} = Ax$ with

$$A = \left[\begin{array}{cc} 0 & 1 \\ -K/M & -D/M \end{array} \right].$$

Assume $M > 0$ and $K, D \geq 0$. If $D = K = 0$, the eigenvalues are $\{0, 0\}$ and A is a nilpotent matrix in Jordan form. The origin is an unstable equilibrium. If only $D = 0$ or $K = 0$ but not both, the origin is stable but not asymptotically stable. And if both D, K are nonzero, the origin is asymptotically stable. □

Example Two points move on the line $\mathbb{R}$. The positions of the points are x_1, x_2. They move toward each other according to the control laws

$$\dot{x}_1 = x_2 - x_1, \quad \dot{x}_2 = x_1 - x_2.$$

Thus the state is $x = (x_1, x_2)$ and the state equation is

$$\dot{x} = Ax, \quad A = \left[\begin{array}{rr} -1 & 1 \\ 1 & -1 \end{array} \right].$$

The eigenvalues are $\lambda_1 = 0, \lambda_2 = -2$, so the origin is stable but not asymptotically stable. Obviously, the two points tend toward each other; that is, the state $x(t)$ tends toward the subspace

$$\mathcal{V} = \{x : x_1 = x_2\}.$$

This is the eigenspace for the zero eigenvalue. To see this convergence, write the initial condition as a linear combination of eigenvectors:

$$x(0) = c_1 v_1 + c_2 v_2, \quad v_1 = \left[\begin{array}{c} 1 \\ 1 \end{array} \right], \quad v_2 = \left[\begin{array}{r} -1 \\ 1 \end{array} \right].$$

Then

$$x(t) = c_1 e^{\lambda_1 t} v_1 + c_2 e^{\lambda_2 t} v_2 = c_1 v_1 + c_2 e^{-2t} v_2 \to c_1 v_1.$$

So $x_1(t)$ and $x_2(t)$ both converge to c_1, the same point. □

2.6 Subspaces

Let $\mathcal{X} = \mathbb{R}^n$ and let $\mathcal{V}, \mathcal{W}$ be subspaces of $\mathcal{X}$. Then $\mathcal{V} + \mathcal{W}$ denotes the set

$$\{v + w : v \in \mathcal{V}, w \in \mathcal{W}\},$$

and it is a subspace of $\mathcal{X}$. The set union $\mathcal{V} \cup \mathcal{W}$ is not a subspace in general unless one is contained in the other. The intersection $\mathcal{V} \cap \mathcal{W}$ is however a subspace. As an example:

$$\mathcal{X} = \mathbb{R}^3, \quad \mathcal{V} \text{ a line}, \quad \mathcal{W} \text{ a plane}.$$

Then $\mathcal{V} + \mathcal{W} = \mathbb{R}^3$ if $\mathcal{V}$ does not lie in $\mathcal{W}$. If $\mathcal{V} \subset \mathcal{W}$, then of course $\mathcal{V} + \mathcal{W} = \mathcal{W}$.[1]

It is a fact that

$$\dim(\mathcal{V} + \mathcal{W}) = \dim(\mathcal{V}) + \dim(\mathcal{W}) - \dim(\mathcal{V} \cap \mathcal{W}).$$

For example, think of $\mathcal{V}, \mathcal{W}$ as two planes in $\mathbb{R}^3$ that intersect in a line. Then the dimension equation evaluates to

$$3 = 2 + 2 - 1.$$

Two subspaces $\mathcal{V}, \mathcal{W}$ are **independent** if $\mathcal{V} \cap \mathcal{W} = 0$. This is not the same as being orthogonal. For example two lines in $\mathbb{R}^2$ are independent iff they are not colinear (i.e., the angle between them is not 0), while they are orthogonal iff the angle is 90°.

Every vector x in $\mathcal{V} + \mathcal{W}$ can be written as

$$x = v + w, \quad v \in \mathcal{V}, \ w \in \mathcal{W}.$$

If $\mathcal{V}, \mathcal{W}$ are independent, then v, w are unique. Think of v as the component of x in $\mathcal{V}$ and w as its component in $\mathcal{W}$. Let's prove uniqueness. Suppose

$$x = v + w = v_1 + w_1.$$

Then

$$v - v_1 = w_1 - w.$$

The left-hand side is in $\mathcal{V}$ and the right-hand side in $\mathcal{W}$. Since the intersection of these two subspaces is zero, both sides equal 0.

Clearly, $\mathcal{V}, \mathcal{W}$ are independent iff

$$\dim(\mathcal{V} + \mathcal{W}) = \dim(\mathcal{V}) + \dim(\mathcal{W}).$$

Three subspaces $\mathcal{U}, \mathcal{V}, \mathcal{W}$ are **independent** if $\mathcal{U}, \mathcal{V} + \mathcal{W}$ are independent, $\mathcal{V}, \mathcal{U} + \mathcal{W}$ are independent, and $\mathcal{W}, \mathcal{U} + \mathcal{V}$ are independent. This is not the same as being pairwise independent. As an example, let $\mathcal{U}, \mathcal{V}, \mathcal{W}$ be 1-dimensional subspaces of $\mathbb{R}^3$, i.e., three lines. When are they independent? Pairwise independent?

Every vector x in $\mathcal{U} + \mathcal{V} + \mathcal{W}$ can be written as

$$x = u + v + w, \quad u \in \mathcal{U}, \ v \in \mathcal{V}, \ w \in \mathcal{W}.$$

If $\mathcal{U}, \mathcal{V}, \mathcal{W}$ are independent, then u, v, w are unique. Also, $\mathcal{U}, \mathcal{V}, \mathcal{W}$ are independent iff

$$\dim(\mathcal{U} + \mathcal{V} + \mathcal{W}) = \dim(\mathcal{U}) + \dim(\mathcal{V}) + \dim(\mathcal{W}).$$

If $\mathcal{V}, \mathcal{W}$ are independent subspaces, we write their sum as $\mathcal{V} \oplus \mathcal{W}$. This is called a **direct sum**. Likewise for more than two.

Let's finish this section with a handy fact: Every subspace has an independent complement, i.e.,

$$\mathcal{V} \subset \mathcal{X} \quad \Longrightarrow \quad (\exists \mathcal{W} \subset \mathcal{X}) \ \mathcal{X} = \mathcal{V} \oplus \mathcal{W}.$$

Think of $\mathcal{X}$ as $\mathbb{R}^3$ and $\mathcal{V}$ as a plane. Then $\mathcal{W}$ can be any line not in the plane.

[1] In this chapter when we speak of lines we mean lines through 0. Similarly for planes.

2.7 Linear Transformations

We now introduce linear transformations. The important point is that a linear transformation is not the same as a matrix, but every linear transformation has a matrix representation once you choose a basis.

Let $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^p$. A linear function $\mathbf{A} : \mathcal{X} \to \mathcal{Y}$ defines a **linear transformation** (LT); $\mathcal{X}$ is called its **domain** and $\mathcal{Y}$ its **co-domain**. Thus

$$\mathbf{A}(x_1 + x_2) = \mathbf{A}x_1 + \mathbf{A}x_2, \quad x_1, x_2 \in \mathcal{X}$$

$$\mathbf{A}(ax) = a\mathbf{A}x, \quad a \in \mathbb{R}, \ x \in \mathcal{X}.$$

It is an important fact that an LT is uniquely determined by its action on a basis. That is, if $\mathbf{A} : \mathcal{X} \to \mathcal{Y}$ is an LT and if $\{e_1, \ldots, e_n\}$ is a basis for $\mathcal{X}$, then if we know the vectors $\mathbf{A}e_i$, we can compute $\mathbf{A}x$ for every $x \in \mathcal{X}$, by linearity.

Example For us, the most important example is an LT **generated by a matrix**. Let $A \in \mathbb{R}^{m \times n}$. For each vector x in $\mathbb{R}^n$, Ax is a vector in $\mathbb{R}^m$. The mapping $x \mapsto Ax$ is an LT $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$. Linearity is easy to check. □

Example Take a vector in the plane and rotate it counterclockwise by 90°. This defines an LT $\mathbf{A} : \mathbb{R}^2 \to \mathbb{R}^2$. Note that $\mathbf{A}$ is not given as a matrix; it's given by its domain, its co-domain, and its action on vectors. If we take a vector to be represented by its Cartesian coordinates, $x = (x_1, x_2)$, then we've chosen a basis for $\mathbb{R}^2$. In that case $\mathbf{A}$ maps $x = (x_1, x_2)$ to $\mathbf{A}x = (-x_2, x_1)$, and so there's an associated rotation matrix

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

We'll return to matrix representation later. □

Example Let $\mathcal{X} = \mathbb{R}^n$ and let $\{e_1, \ldots, e_n\}$ be a basis. Every vector x in $\mathcal{X}$ has a unique expansion

$$x = a_1 e_1 + \cdots + a_n e_n, \quad a_i \in \mathbb{R}.$$

Let a denote the vector $(a_1, \ldots, a_n)$, the n**-tuple of coordinates** of x with respect to the basis. The function $x \longmapsto a$ defines an LT $\mathbf{Q} : \mathcal{X} \to \mathbb{R}^n$. The equation

$$x = a_1 e_1 + \cdots + a_n e_n$$

can be written compactly as $x = Ea$, where E is the matrix with columns $e_1, \ldots, e_n$ and a is the vector with components $a_1, \ldots, a_n$. Therefore $a = E^{-1}x$ and so $\mathbf{Q}x = E^{-1}x$, that is, the action of $\mathbf{Q}$ is to multiply by the matrix E^{-1}.

For example, let $\mathcal{X} = \mathbb{R}^2$. Take the natural basis

$$e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

In this case $E = I$ and $\mathbf{Q}x = x$. If the basis instead is

$$e_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad e_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix},$$

then

$$E = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

and $\mathbf{Q}x = E^{-1}x$. □

Every LT on finite-dimensional vector spaces has a **matrix representation**. Let's do this very important construction carefully. Let $\mathbf{A}$ be an LT $\mathcal{X} \to \mathcal{Y}$,

$$\mathcal{X} = \mathbb{R}^n,\ \text{basis } \{e_1, \ldots, e_n\};\ \ \mathcal{Y} = \mathbb{R}^p,\ \text{basis } \{f_1, \ldots, f_p\}.$$

Bring in the coordinate LTs:

$$\mathbf{Q} : \mathcal{X} \to \mathbb{R}^n,\ \ \mathbf{R} : \mathcal{Y} \to \mathbb{R}^p.$$

So now we have the setup

$$\begin{tikzcd}
\mathcal{X} \arrow[r, "\mathbf{A}"] \arrow[d, "\mathbf{Q}"'] & \mathcal{Y} \arrow[d, "\mathbf{R}"] \\
\mathbb{R}^n & \mathbb{R}^p
\end{tikzcd}$$

The left downward arrow gives us the n-tuple, say a, that represents a vector x in the basis $\{e_1, \ldots, e_n\}$. The right downward arrow gives us the p-tuple, say b, that represents a vector y in the basis $\{f_1, \ldots, f_n\}$. It's possible to add a fourth LT to complete the square:

$$\begin{tikzcd}
\mathcal{X} \arrow[r, "\mathbf{A}"] \arrow[d, "\mathbf{Q}"'] & \mathcal{Y} \arrow[d, "\mathbf{R}"] \\
\mathbb{R}^n \arrow[r, "M"'] & \mathbb{R}^p
\end{tikzcd}$$

This is called a **commutative diagram.** The object M in the diagram is the matrix representation of $\mathbf{A}$ with respect to these two bases. Notice that the bottom arrow represents the LT generated by the matrix M; we write M in the diagram for simplicity, but you should understand that really the object is an LT. The matrix M is the $p \times n$ matrix that makes the diagram commute, that is, for every $x \in \mathcal{X}$

$$Ma = b, \quad \text{where } a = \mathbf{Q}x,\ b = \mathbf{R}\mathbf{A}x.$$

In particular, take $x = e_i$, the i^{th} basis vector in $\mathcal{X}$. Then a is the n-vector with 1 in the i^{th} entry and 0 otherwise. So Ma equals the i^{th} column of the matrix M. Thus, we have the following recipe for constructing the matrix M:

1. Take the 1^{st} basis vector e_1 of $\mathcal{X}$.

2. Apply the LT $\mathbf{A}$ to get $\mathbf{A}e_1$.

3. Find b, the coordinate vector of $\mathbf{A}e_1$ in the basis for $\mathcal{Y}$.

4. Enter this b as column 1 of M.

5. Repeat for the other columns.

Recall that $\mathbf{Q}$ is the LT generated by E^{-1}, where the columns of E are the basis in the domain of $\mathbf{A}$. Likewise, $\mathbf{R}$ is the LT generated by F^{-1}, where the columns of F are the basis in the co-domain of $\mathbf{A}$. Thus the equation $Ma = b$ reads

$$ME^{-1}x = F^{-1}\mathbf{A}x. \tag{2.3}$$

Example Let $\mathbf{A} : \mathbb{R}^2 \to \mathbb{R}^2$ be the LT that rotates a vector counterclockwise by 90°. Let's first take the standard bases: $e_1 = (1,0), e_2 = (0,1)$ for the domain and $f_1 = (1,0), f_2 = (0,1)$ for the co-domain. Following the steps we first apply $\mathbf{A}$ to e_1, that is, we rotate e_1 counterclockwise by 90°; the result is $\mathbf{A}e_1 = (0,1)$. Then we express this vector in the basis $\{f_1, f_2\}$:

$$\mathbf{A}e_1 = 0 \times f_1 + 1 \times f_2.$$

Thus the first column of M is $(0,1)$, the vector of coefficients. Now for the second column, rotate e_2 to get $(-1,0)$ and represent this in the basis $\{f_1, f_2\}$:

$$\mathbf{A}e_2 = -1 \times f_1 + 0 \times f_2.$$

So the second column of M is $(-1,0)$. Thus

$$M = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Suppose we had different bases:

$$e_1 = (1,1),\ e_2 = (-1,2),\ f_1 = (1,2),\ f_2 = (1,0).$$

Apply the recipe again. Get $\mathbf{A}e_1 = (-1,1)$. Expand it in the basis $\{f_1, f_2\}$:

$$(-1,1) = \frac{1}{2}f_1 - \frac{3}{2}f_2.$$

Get $\mathbf{A}e_2 = (-2,-1)$. Expand it in the basis $\{f_1, f_2\}$:

$$(-2,-1) = -\frac{1}{2}f_1 - \frac{3}{2}f_2.$$

Thus

$$M = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{3}{2} & -\frac{3}{2} \end{bmatrix}.$$

□

Example Let $A \in \mathbb{R}^{m \times n}$ and let $\mathbf{A} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be the generated LT. It is easy to check that A itself is then the matrix representation of $\mathbf{A}$ with respect to the standard bases. Let's do it.

Let $\{e_1, \dots, e_n\}$ be the standard basis on $\mathbb{R}^n$ and $\{f_1, \dots, f_m\}$ the standard basis on $\mathbb{R}^m$. Then $\mathbf{A}e_1 = Ae_1$ equals the first column, $(a_{11}, a_{21}, \dots, a_{m1})$, of A. This column can be written as

$$a_{11}f_1 + \cdots + a_{m1}f_m,$$

and hence $(a_{11}, a_{21}, \dots, a_{m1})$ is the first column of the matrix representation of $\mathbf{A}$.

Suppose instead that we have general bases, $\{e_1, \dots, e_n\}$ on $\mathbb{R}^n$ and $\{f_1, \dots, f_m\}$ on $\mathbb{R}^m$. Form the matrices E and F from these basis vectors. From (2.3) we get that the matrix representation M with respect to these bases satisfies

$$ME^{-1} = F^{-1}A,$$

or equivalently

$$AE = FM.$$

A very interesting special case of this is where A is square and the same basis $\{e_1, \dots, e_n\}$ is taken for both the domain and co-domain. Then

$$AE = EM,$$

or $M = E^{-1}AE$; the matrix M is a similarity transformation of the given matrix A.

Finally, suppose we start with a square A and take the basis $\{v_1, \dots, v_n\}$ of generalized eigenvectors. The new matrix representation is our familiar Jordan form $A_{JF} = V^{-1}AV$. Thus the two matrices A and A_{JF} represent the same LT: A in the given standard basis and A_{JF} in the basis of generalized eigenvectors. □

An LT has two important associated subspaces. Let $\mathbf{A} : \mathcal{X} \to \mathcal{Y}$ be an LT. The **kernel** (or nullspace) of $\mathbf{A}$ is the subspace of $\mathcal{X}$ on which $\mathbf{A}$ is zero:

$$\text{Ker } \mathbf{A} := \{x : \mathbf{A}x = 0\}.$$

The LT $\mathbf{A}$ is said to be **one-to-one** if Ker $\mathbf{A} = 0$, equivalently, the homogeneous equation $\mathbf{A}x = 0$ has only the trivial solution $x = 0$. The **image** (or range space) of $\mathbf{A}$ is the subspace of $\mathcal{Y}$ that $\mathbf{A}$ can reach:

$$\text{Im } \mathbf{A} := \{y : (\exists x \in \mathcal{X})y = \mathbf{A}x\}.$$

We say $\mathbf{A}$ is **onto** if Im $\mathbf{A} = \mathcal{Y}$, equivalently, the equation $\mathbf{A}x = y$ has a solution x for every y.

Whether $\mathbf{A}$ is one-to-one or onto (or both) can be easily checked by examining any matrix representation A:

$$\mathbf{A} \text{ is one-to-one} \iff A \text{ has full column rank;}$$

$$\mathbf{A} \text{ is onto} \iff A \text{ has full row rank.}$$

If A is a matrix, we will write Im A for the image of the generated LT—it's the column span of the matrix; and we'll write Ker A for the kernel of the LT.

Example Let $\mathbf{A} : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ map a vector to its projection on the horizontal plane. Then the kernel equals the vertical axis, the image equals the horizontal plane, $\mathbf{A}$ is neither onto nor one-to-one, and its matrix with respect to the standard basis is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

We could modify the co-domain to have $\mathbf{A} : \mathbb{R}^3 \longrightarrow \mathbb{R}^2$, again mapping a vector to its projection on the horizontal plane. Then the kernel equals the vertical axis, the image equals the horizontal plane, $\mathbf{A}$ is onto but not one-to-one, and its matrix with respect to the standard basis is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

□

Example Let $\mathcal{V} \subset \mathcal{X}$ (think of $\mathcal{V}$ as a plane in 3-dimensional space $\mathcal{X}$). Define the function $\mathbf{V} : \mathcal{V} \to \mathcal{X}$, $\mathbf{V}x = x$. This is an LT called the **insertion** LT. Clearly $\mathbf{V}$ is one-to-one and $\text{Im } \mathbf{V} = \mathcal{V}$. Suppose we have a basis for $\mathcal{V}$,

$$\{e_1, \ldots, e_k\},$$

and we extend it to get a basis for $\mathcal{X}$,

$$\{e_1, \ldots, e_k, \ldots, e_n\}.$$

Then the matrix rep. of $\mathbf{V}$ is

$$V = \begin{bmatrix} I_k \\ 0 \end{bmatrix}.$$

Clearly, rank $V = k$. □

Example Let $\mathcal{X}$ be 3-dimensional space, $\mathcal{V}$ a plane (2-dimensional subspace), and $\mathcal{W}$ a line not in $\mathcal{V}$. Then $\mathcal{V}, \mathcal{W}$ are independent subspaces and

$$\mathcal{X} = \mathcal{V} \oplus \mathcal{W}.$$

Every x in $\mathcal{X}$ can be written $x = v + w$ for unique v in $\mathcal{V}$ and w in $\mathcal{W}$. Define the function $\mathbf{P} : \mathcal{X} \to \mathcal{V}$ mapping x to v. This is an LT called the **natural projection** onto $\mathcal{V}$. Check that

$$\text{Im } \mathbf{P} = \mathcal{V}, \qquad \text{Ker } \mathbf{P} = \mathcal{W}.$$

Suppose $\{e_1, e_2\}$ is a basis for $\mathcal{V}$, $\{e_3\}$ a basis for $\mathcal{W}$. The induced matrix representation is

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

□

Example Let $\mathbf{A} : \mathcal{X} \to \mathcal{Y}$ be an LT. Its kernel, Ker $\mathbf{A}$, is a subspace of $\mathcal{X}$; let $\{e_{k+1}, \dots, e_n\}$ be a basis for Ker $\mathbf{A}$ and extend it to get a basis for $\mathcal{X}$:

$$\{e_1, \dots, e_k, \dots, e_n\} \text{ for } \mathcal{X}.$$

Then

$$\{\mathbf{A}e_1, \dots, \mathbf{A}e_k\}$$

is a basis for Im $\mathbf{A}$. Extend it to get a basis for $\mathcal{Y}$:

$$\{\mathbf{A}e_1, \dots, \mathbf{A}e_k, f_{k+1}, \dots, f_p\}.$$

Then the matrix representation of A is

$$A = \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}.$$

□

2.8 Matrix Equations

We already reviewed the linear equation

$$Ax = b, \ A \in \mathbb{R}^{n \times m}, \ x \in \mathbb{R}^m, \ b \in \mathbb{R}^n.$$

The equation is another way of saying b is a linear combination of the columns of A. Thus the equation has a solution iff $b \in$ column span of A, i.e., $b \in \mathrm{Im}A$. Then the solution is unique iff rank $A = m$, i.e., Ker $A = 0$.

These results extend to the matrix equation

$$AX = B, \ A \in \mathbb{R}^{n \times m}, \ X \in \mathbb{R}^{m \times p}, \ B \in \mathbb{R}^{n \times p}$$

In this section we study this and similar equations. We could work with LTs but we'll use matrices instead.

The first equation is $AX = I$. Such an X is called a **right-inverse** of A.

Lemma 2.2 *$A \in \mathbb{R}^{n \times m}$ has a right-inverse iff it's onto, i.e. the rank of A equals n.*

Proof ($\Longrightarrow$) If $AX = I$, then, for every $y \in \mathbb{R}^n$,

$$AXy = y.$$

Thus for every $y \in \mathbb{R}^n$, there exists $x \in \mathbb{R}^m$ such that $Ax = y$. Thus A is onto.
($\Longleftarrow$) Let $\{f_1, \dots, f_n\}$ be the standard basis for $\mathbb{R}^n$. Since A is onto

$$(\forall i)(\exists x_i \in \mathbb{R}^m) f_i = Ax_i.$$

Now define X to be the matrix whose i^{th} column is x_i, i.e., via $Xf_i = x_i$. Then $AXf_i = f_i$. This implies $AX = I$. □

The second equation is the dual situation $XA = I$. Obviously, such an X is a **left-inverse**.

Lemma 2.3 *$A \in \mathbb{R}^{n\times m}$ has a left-inverse iff it's one-to-one, i.e., A has rank m.*

Lemma 2.4 *1. There exists X such that $AX = B$ iff* $\text{Im } B \subset \text{Im } A$*, that is,*

$$\text{rank } A = \text{rank } \begin{bmatrix} A & B \end{bmatrix}.$$

2. There exists X such that $XA = B$ iff $\text{Ker } A \subset \text{Ker } B.$*, that is,*

$$\text{rank } A = \text{rank } \begin{bmatrix} A \\ B \end{bmatrix}.$$

2.9 Invariant Subspaces

Example Let

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$$

and let $\mathbf{A} : \mathbb{R}^2 \to \mathbb{R}^2$ be the generated LT. Clearly, Ker $\mathbf{A}$ is the 1-dimensional subspace spanned by $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. Also,

$$x \in \text{Ker } \mathbf{A} \Rightarrow \mathbf{A}x = 0 \in \text{Ker } \mathbf{A},$$

or equivalently,

$$\mathbf{A}\text{Ker } \mathbf{A} \subset \text{Ker } \mathbf{A}.$$

□

In general, if $\mathbf{A} : \mathcal{X} \to \mathcal{X}$ is an LT, a subspace $\mathcal{V} \subset \mathcal{X}$ is **A-invariant** if $\mathbf{A}\mathcal{V} \subset \mathcal{V}$. The zero subspace, $\mathcal{X}$ itself, Ker $\mathbf{A}$, and Im $\mathbf{A}$ are all $\mathbf{A}$-invariant. Now Ker $\mathbf{A}$ is the eigenspace for the zero eigenvalue, assuming $\lambda = 0$ **is** an eigenvalue (as in the example above).

More generally, suppose λ is an eigenvalue of $\mathbf{A}$. Assume $\lambda \in \mathbb{R}$. Then $\mathbf{A}x = \lambda x$ for some $x \neq 0$. Then $\mathcal{V} = \text{Span } \{x\}$ is $\mathbf{A}$-invariant. So is the **eigenspace**

$$\{x : \mathbf{A}x = \lambda x\} = \{x : (\mathbf{A} - \lambda\mathbf{I})x = 0\} = \text{Ker } (\mathbf{A} - \lambda\mathbf{I}).$$

Let $\mathcal{V}$ be an $\mathbf{A}$-invariant subspace. Take a basis for $\mathcal{V}$,

$$\{e_1, \ldots, e_k\},$$

and extend it to a basis for $\mathcal{X}$:

$$\{e_1, \ldots, e_k, \ldots, e_n\}.$$

Then the matrix representation of $\mathbf{A}$ has the form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}.$$

Notice that the lower-left block of A equals zero; this is because $\mathcal{V}$ is $\mathbf{A}$-invariant.

Example Let $\mathcal{X} = \mathbb{R}^3$, let $\mathcal{V}$ be the (x_1, x_2)-plane, and let $\mathbf{A} : \mathcal{X} \to \mathcal{X}$ be the LT that rotates a vector 90° about the x_3-axis using the right-hand rule. Thus $\mathcal{V}$ is $\mathbf{A}$-invariant. Let us take the bases

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \; e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \text{ for } \mathcal{V}$$

$$e_1, \; e_2, \; e_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \text{ for } \mathcal{X}.$$

The matrix representation of $\mathbf{A}$ with respect to the latter basis is

$$A = \left[\begin{array}{cc|c} 0 & -1 & -2 \\ 1 & 0 & 0 \\ \hline 0 & 0 & 1 \end{array}\right].$$

So, in particular, the restriction of $\mathbf{A}$ to $\mathcal{V}$ is represented by the rotation matrix

$$A_{11} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

□

Finally, let A be an $n \times n$ matrix. Suppose V is an $n \times k$ matrix. Then Im V is a subspace of $\mathbb{R}^n$. How can we know if this subspace is invariant under A, or more precisely, under the LT generated by A? The answer is this:

Lemma 2.5 *The subspace* Im V *is A-invariant iff the linear equation $AV = VA_1$ has a solution A_1.*

Proof If $AV = VA_1$, then Im $AV \subset$ Im V, that is, A Im $V \subset$ Im V, which says Im V is A-invariant.

Conversely, if Im $AV \subset$ Im V, then the equation $AV = VA_1$ is solvable, by Lemma 2.4. □.

2.10 Problems

1. Are the following vectors linearly independent?

$$v_1 = (1, 1, 2, 0), \quad v_2 = (1, 0, 2, -2), \quad v_3 = (-1, 2, -2, 6).$$

2. Continuing with the same vectors, find a basis for Span $\{v_1, v_2, v_3\}$.

3. What kind of geometric object is $\{x : Ax = b\}$ when $A \in \mathbb{R}^{m \times n}$? That is, is it a sphere, a point—what?

4. Show that $\mathrm{e}^{A+B} = \mathrm{e}^A \mathrm{e}^B$ does not imply that A and B commute, but $\mathrm{e}^{(A+B)t} = \mathrm{e}^{At}\mathrm{e}^{Bt}$ does.

5. (a) Let A be an 8×8 real matrix with eigenvalues

$$2, 2, -3, -3, -3, 8, 4, 4.$$

Assume

$$\text{rank}(A - 2I) = 7, \ \text{rank}(A + 3I) = 6, \ \text{rank}(A - 4I) = 6.$$

Write down the Jordan form of A.

(b) The matrix

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ -1 & 0 & 0 & -1 \end{bmatrix}$$

is nilpotent. Write down its Jordan form.

6. Take

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 2 & -2 \end{bmatrix}.$$

Show that the matrix V constructed as follows satisfies $V^{-1}AV = A_{JF}$:

Select v_3 in Ker A^2 but not in Ker A.

Set $v_2 = Av_3$.

Select v_1 in Ker A such that $\{v_1, v_2\}$ is linearly independent.

Select an eigenvector v_4 corresponding to the eigenvalue -3.

Set $V = [v_1 \ \ v_2 \ \ v_3 \ \ v_4]$.

(The general construction of the basis for the Jordan form is along these lines.)

7. Let

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -2 & 1 & 0 & 2 \end{bmatrix}.$$

Write down the Jordan form of A.

8. Consider

$$A = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix},$$

where σ and $\omega \neq 0$ are real. Find the Jordan form and the transition matrix.

9. In the previous problem, we saw that when

$$A = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}$$

its transition matrix is easy to write down. This problem demonstrates that a matrix with distinct complex eigenvalues can be transformed into the above form using a nonsingular transformation. Let

$$A = \begin{bmatrix} -1 & -4 \\ 1 & -1 \end{bmatrix}.$$

Determine the eigenvalues and eigenvectors of A, noting that they form complex conjugate pairs. Let the first eigenvalue be written as $a+jb$ with the corresponding eigenvector v_1+jv_2. Take v_1 and v_2 as the columns of a matrix V. Find $V^{-1}AV$.

10. Consider the homogeneous state equation $\dot{x} = Ax$ with

$$A = \begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}$$

and $x_0 = (3, 2)$. Find a modal expansion of $x(t)$.

11. Show that the origin is asymptotically stable for $\dot{x} = Ax$ iff all poles of every element of $(sI - A)^{-1}$ are in the open left half-plane. Show that the origin is stable iff all poles of every element of $(sI - A)^{-1}$ are in the closed left half-plane and those on the imaginary axis have multiplicity 1.

12. Consider the linear system

$$\begin{aligned} \dot{x} &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} x + \begin{bmatrix} -1 \\ 1 \end{bmatrix} u \\ y &= \begin{bmatrix} 0 & 1 \end{bmatrix} x \end{aligned}$$

   (a) If $u(t)$ is the unit step and $x(0) = 0$, is $y(t)$ bounded?

   (b) If $u(t) = 0$ and $x(0)$ is arbitrary, is $y(t)$ bounded?

13. (a) Suppose that $\sigma(A) = \{-1, -3, -3, -1+j2, -1-j2\}$ and the rank of $(A - \lambda I)_{\lambda=-3}$ is 4. Determine A_{JF}.

   (b) Suppose that $\sigma(A) = \{-1, -2, -2, -2\}$ and the rank of $(A - \lambda I)_{\lambda=-2}$ is 3. Determine A_{JF}.

   (c) Suppose that $\sigma(A) = \{-1, -2, -2, -2, -3\}$ and the rank of $(A - \lambda I)_{\lambda=-2}$ is 3. Determine A_{JF}.

14. Find A_{JF} for

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -2 & -4 & -3 \end{bmatrix}.$$

15. Summarize all the ways to find $\exp(At)$. Then find $\exp(At)$ for

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}.$$

16. Consider the set

$$\{cv : c \geq 0\},$$

where $v \neq 0$ is a given vector in $\mathbb{R}^2$. This set is called a **ray** from the origin in the direction of v. More generally,

$$\{x_0 + cv : c \geq 0\}$$

is a ray from x_0 in the direction of v. Find a 2×2 matrix A and a vector x_0 such that the solution $x(t)$ of $\dot{x} = Ax,\ x(0) = x_0$ is a ray.

17. Consider the following system:

$$\begin{aligned} \dot{x}_1 &= -x_2 \\ \dot{x}_2 &= x_1 - 3x_2 \end{aligned}$$

Do a phase portrait using Scilab or MATLAB. Interpret the phase portrait in terms of the modal decomposition of the system. Do lots more examples of this type.

18. Prove the following facts about subspaces:

    (a) $\mathcal{V} + \mathcal{V} = \mathcal{V}$
    Hint: You have to show $\mathcal{V}+\mathcal{V} \subset \mathcal{V}$ and $\mathcal{V} \subset \mathcal{V}+\mathcal{V}$. Similarly for other subspace equalities.

    (b) If $\mathcal{V} \subset \mathcal{W}$, then $\mathcal{V} + \mathcal{W} = \mathcal{W}$.

    (c) If $\mathcal{V} \subset \mathcal{W}$, then $\mathcal{W} \cap (\mathcal{V} + \mathcal{T}) = \mathcal{V} + \mathcal{W} \cap \mathcal{T}$.

19. Show that $\mathcal{W} \cap (\mathcal{V}+\mathcal{T}) = \mathcal{W}\cap\mathcal{V}+\mathcal{W}\cap\mathcal{T}$ is false in general by giving an explicit counterexample.

20. Let $\mathbf{A}$ be the identity LT on $\mathbb{R}^2$. Take

$$\left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\} = \text{ basis for domain, } \quad \left\{ \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 3 \end{bmatrix} \right\} = \text{ basis for co-domain.}$$

Find the matrix A.

21. Let $\mathbf{A}$ denote the LT $\mathbb{R}^4 \to \mathbb{R}^5$ with the action

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \mapsto \begin{bmatrix} x_4 \\ 0 \\ 2x_4 \\ x_2 + x_3 + 2x_4 \\ x_2 + x_3 \end{bmatrix}.$$

Find bases for $\mathbb{R}^4$ and $\mathbb{R}^5$ so that the matrix representation is

$$A = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}.$$

22. Let $\mathbf{A}$ be an LT. Show that if $\{\mathbf{A}e_1, \ldots, \mathbf{A}e_n\}$ is linearly independent, so is $\{e_1, \ldots, e_n\}$. Give an example where the converse is false.

23. Find all right-inverses of the matrix

$$A = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

24. Let $\mathcal{X}$ denote the 4-dimensional vector space with basis

$$\{\sin t, \cos t, \sin 2t, \cos 2t\}.$$

Thus vectors in $\mathcal{X}$ are time-domain signals of frequency 1 rad/s, 2 rad/s, or a combination of both. Suppose an input $x(t)$ from $\mathcal{X}$ is applied to a lowpass RC-filter, producing the output $y(t)$. The equation for the circuit is

$$RC\dot{y}(t) + y(t) = x(t).$$

For simplicity, take $RC = 1$. From circuit theory, we know that $y(t)$ belongs to $\mathcal{X}$ too. (This is steady-state analysis; transient response is neglected.) So the mapping from $x(t)$ to $y(t)$ defines a linear transformation $\mathbf{A} : \mathcal{X} \longrightarrow \mathcal{X}$. Find the matrix representation of $\mathbf{A}$ with respect to the given basis.

25. Consider the vector space $\mathbb{R}^3$. Let x_1, x_2, and x_3 denote the components of a vector x in $\mathbb{R}^3$. Now let $\mathcal{V}$ denote the subspace of $\mathbb{R}^3$ of all vectors x where

$$x_1 + x_2 - x_3 = 0,$$

and let $\mathcal{W}$ denote the subspace of $\mathbb{R}^3$ of all vectors x where

$$2x_1 - 3x_3 = 0.$$

Find a basis for the intersection $\mathcal{V} \cap \mathcal{W}$.

26. Let $\mathbf{A} : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ be the LT defined by

$$\mathbf{A} : \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \mapsto \begin{bmatrix} 8x_1 - 2x_3 \\ x_1 + 7x_2 - 2x_3 \\ 4x_1 - x_3 \end{bmatrix}.$$

Find bases for Ker $\mathbf{A}$ and Im $\mathbf{A}$.

27. Find all solutions of the matrix equation $XA = I$ where

$$A = \begin{bmatrix} 1 & 2 \\ 1 & 0 \\ 2 & -1 \end{bmatrix}.$$

28. For a square matrix X, let diag$X$ denote the vector formed from the elements on the diagonal of X.

    Let $\mathbf{A} : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^n$ be the LT defined by

    $$\mathbf{A} : X \mapsto \text{diag}X.$$

    Does $\mathbf{A}$ have a left inverse? A right inverse?

29. Consider the two matrices:

    $$\begin{bmatrix} 4 & 1 & -1 \\ 3 & 2 & -3 \\ 1 & 3 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 1 & 2 \\ 3 & 4 & 5 & 0 & 0 \end{bmatrix}.$$

    For each matrix, find its rank, a basis for its image, and a basis for its kernel.

30. Let $A, U \in \mathbb{R}^{n\times n}$ with U nonsingular. True or false:

    (a) Ker $(A) =$ Ker (UA).

    (b) Ker $(A) =$ Ker (AU).

    (c) Ker $(A^2) \subseteq$ Ker (A).

31. Is $\{(x_1, x_2, x_3) : 2x_1 + 3x_2 + 6x_3 - 5 = 0\}$ a subspace of $\mathbb{R}^3$?

32. You are given the n eigenvalues of a matrix in $\mathbb{R}^{n\times n}$. Can you determine the rank of the matrix? If no, can you give bounds on the rank?

33. Suppose that $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times m}$ with $m \leq n$ and rank $A =$ rank $B = m$. Find a necessary and sufficient condition that AB be invertible.

34. Let $\mathbf{A}$ be an LT from $\mathcal{X}$ to $\mathcal{X}$, a finite-dimensional vector space. Fix a basis for $\mathcal{X}$ and let A denote the matrix representation of $\mathbf{A}$ with respect to this basis. Show that A^2 is the matrix representation of $\mathbf{A}^2$.

35. Consider the following "result:"

    Lemma *If A is a matrix with full column rank, then the equation $Ax = y$ is solvable for every vector y.*

    Proof Let y be arbitrary. Multiply the equation $Ax = y$ by the transpose of A:

    $$A^T Ax = A^T y.$$

    Since A has full column rank, $A^T A$ is invertible. Thus

    $$x = (A^T A)^{-1} A^T y.$$

    □

    (a) Give a counterexample to the lemma.

    (b) What is the mistake in logic in the proof?

36. Let $\mathcal{L}$ denote the line in the plane that passes through the origin and makes an angle $+\pi/6$ radians with the positive x-axis. Let $\mathbf{A} : \mathbb{R}^2 \to \mathbb{R}^2$ be the LT that maps a vector to its reflection about $\mathcal{L}$.

    (a) Find the matrix representation of $\mathbf{A}$ with respect to the basis

    $$e_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad e_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

    (b) Show that $\mathbf{A}$ is invertible and find its inverse.

37. Fix a vector $v \neq 0$ in $\mathbb{R}^3$ and consider the LT $\mathbf{A} : \mathbb{R}^3 \to \mathbb{R}^3$ that maps x to the cross product $v \times x$.

    (a) Find $\mathrm{Ker}(\mathbf{A})$ and $\mathrm{Im}(\mathbf{A})$.

    (b) Is $\mathbf{A}$ invertible?

38. **Preamble**. This problem requires some notation. Suppose $f(x, y)$ is a function of two real variables, say, $f(x, y) = x^2 + xy$. Then $f(\cdot, y)$ denotes the function $x \mapsto f(x, y)$ where y is temporarily held constant; that is, $x^2 + xy$ considered as a function of x alone, with y constant. So for each y, $f(\cdot, y)$ is a function, indicated by the dot acting as a placemarker. But then $y \mapsto f(\cdot, y)$ is another function, namely, it maps y to the function of x given by $x^2 + xy$. Therefore, $f(\cdot, y)$ is a function that maps a real number to a real number, whereas $y \mapsto f(\cdot, y)$ is a function that maps a real number to a function.

    Another example where this situation comes up is a cart of mass M, input force u, and output position y. Then y is a function of M and u; let's write $y = G(M, u)$. Then $G(M, \cdot)$ is the input-output map for a given M, and $M \mapsto G(M, \cdot)$ is the map from the mass to the input-output system.

    The problem Fix the $n \times n$ matrix A and consider the equation

    $$\dot{x} = Ax, \quad x(0) = x_0.$$

    As you know, the state at time t starting from x_0 at time 0 is $x(t, x_0) = \mathrm{e}^{At} x_0$.

    (a) What are the domain and co-domain of the mapping

    $$(t, x_0) \mapsto x(t, x_0)?$$

    (b) What are the domain and co-domain of $x(\cdot, x_0)$? Of $x(t, \cdot)$?

    (c) Is the map $x_0 \mapsto x(\cdot, x_0)$ a linear transformation? Prove true or give a counterexample.

    (d) Is the map $t \mapsto x(t, \cdot)$ a linear transformation? Prove true or give a counterexample.

Chapter 3

Calculus

In this chapter we review some calculus, including the method of Lagrange multipliers.

3.1 Jacobians

Suppose $f : \mathbb{R} \longrightarrow \mathbb{R}$ is a function of class C^2, twice continuously differentiable. The Taylor series expansion of f at x is

$$f(x+\varepsilon) = f(x) + \frac{df}{dx}(x)\varepsilon + \frac{1}{2!}\frac{d^2f}{dx^2}(x)\varepsilon^2 + \frac{1}{3!}\frac{d^3f}{dx^3}(x)\varepsilon^3 + \cdots .$$

This extends to a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. Thus, in the expression $f(x)$, x is a vector with n components and $f(x)$ is a scalar. The Jacobian of f, denoted f_x, is the $1 \times n$ matrix (row vector) whose j^{th} element is $\partial f/\partial x_j$. We shall write the transpose of f_x as ∇f, the gradient of f. Thus ∇f is a column vector.

Another way to think of the Jacobian is via the directional derivative. Let x and h be vectors and ε a scalar. Consider $f(x+\varepsilon h)$ as a function of ε and think of its Taylor series at 0:

$$f(x+\varepsilon h) = f(x) + \varepsilon \frac{d}{d\varepsilon} f(x+\varepsilon h)\bigg|_{\varepsilon=0} + \frac{\varepsilon^2}{2}\frac{d^2}{d\varepsilon^2} f(x+\varepsilon h)\bigg|_{\varepsilon=0} + \cdots .$$

By the chain rule,

$$\frac{d}{d\varepsilon} f(x+\varepsilon h) = f_x(x+\varepsilon h)h$$

and so

$$\frac{d}{d\varepsilon} f(x+\varepsilon h)\bigg|_{\varepsilon=0} = f_x(x)h.$$

Thus

$$f(x+\varepsilon h) = f(x) + \varepsilon f_x(x)h + \cdots .$$

The third term in the expansion is this:

$$\frac{d^2}{d\varepsilon^2} f(x+\varepsilon h)\bigg|_{\varepsilon=0} \frac{\varepsilon^2}{2}.$$

Now

$$\frac{d^2}{d\varepsilon^2} f(x+\varepsilon h) = \frac{d}{d\varepsilon} f_x(x+\varepsilon h)h.$$

In more manageable terms, the right-hand side is (with the argument dropped and by use of the chain rule again)

$$\frac{\partial}{\partial x}\left(h_1\frac{\partial f}{\partial x_1}+\cdots+h_n\frac{\partial f}{\partial x_n}\right)h.$$

This in turn equals

$$h^T f_{xx}(x+\varepsilon h)h,$$

where $f_{xx}(x)$ is the Hessian matrix, whose ij^{th} element is

$$\frac{\partial^2 f}{\partial x_i \partial x_j}.$$

Thus the first three terms in the Taylor series become

$$f(x+\varepsilon h) = f(x) + \varepsilon f_x(x)h + \frac{\varepsilon^2}{2}h^T f_{xx}(x)h + \cdots .$$

This generalizes to a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$. Thus, in the expression $f(x)$, x is a vector with n components and $f(x)$ is a vector with m components. So we can write

$$x = (x_1, \ldots, x_n), \quad f = (f_1, \ldots, f_m).$$

The Jacobian of f, still denoted f_x, is the $m \times n$ matrix whose ij^{th} element is $\partial f_i/\partial x_j$. The derivative of f at the point x in the direction of the vector h is defined to be

$$\left.\frac{d}{d\varepsilon} f(x+\varepsilon h)\right|_{\varepsilon=0}.$$

This turns out to be a linear function of the vector h, and it must therefore equal Mh for some matrix M. In fact, M equals the Jacobian of f at x.

Example

$$m=1,\ n=2,\ f(x)=c_1x_1+c_2x_2, \quad f_x(x) = \left[\begin{array}{cc} c_1 & c_2 \end{array}\right]$$

More generally, if $f(x) = c^T x$, then $f_x(x) = c^T$. This can be derived like this:

$$\begin{aligned} f(x+\varepsilon h) &= c^T(x+\varepsilon h) \\ &= c^T x + \varepsilon c^T h \end{aligned}$$

$$\frac{d}{d\varepsilon} f(x+\varepsilon h) = c^T h$$

$$\left.\frac{d}{d\varepsilon} f(x+\varepsilon h)\right|_{\varepsilon=0} = c^T h$$

$$f_x(x) = c^T.$$

□

Example If

$$f(x) = \|x\|^2 = x_1^2 + \cdots + x_n^2,$$

then $f_x(x) = 2x^T$. More generally, consider $f(x) = x^T Q x$, where Q is a symmetric matrix. You can derive that $f_x(x) = 2x^T Q$. If Q is not symmetric, then

$$f_x(x) = x^T(Q^T + Q).$$

□

Example If $f(x) = x(\|x\|^2 - 1)$, $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$, then

$$\begin{aligned} f(x+\varepsilon h) &= (x+\varepsilon h)(\|x+\varepsilon h\|^2 - 1) \\ &= (x+\varepsilon h)(\|x\|^2 - 1 + 2\varepsilon x^T h + \varepsilon^2\|h\|^2) \\ &= x(\|x\|^2-1) + \varepsilon\|x\|^2 h + 2\varepsilon x x^T h - \varepsilon h + \text{HOT}. \end{aligned}$$

Thus

$$f_x(x) = (\|x\|^2 - 1)I + 2xx^T.$$

□

3.2 Optimization over an Open Set

A subset $\mathcal{V}$ of $\mathbb{R}^n$ is **open** if every point in $\mathcal{V}$ has the property that "it lies inside $\mathcal{V}$," that is,

$$(\forall x \in \mathcal{V})(\exists \varepsilon > 0)(\forall y)\|y - x\| < \varepsilon \Longrightarrow y \in \mathcal{V}.$$

Given a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ and an open set $\mathcal{V}$ in $\mathbb{R}^n$, the problem is to maximize $f(x)$ subject to $x \in \mathcal{V}$. Of course, minimizing f is the same as maximizing $-f$, so we're solving that problem too.

We say x^o is a **global maximizer** if

$$x^o \in \mathcal{V},\ (\forall x \in \mathcal{V})\ f(x^o) \geq f(x).$$

We say x^o is a **local maximizer** if

$$x^o \in \mathcal{V},\ (\exists \varepsilon > 0)(\forall x \in \mathcal{V})\|x - x^o\| < \varepsilon \Longrightarrow f(x^o) \geq f(x).$$

Minimizer has the obvious definition, and **optimizer** refers to a point that's either a maximizer of minimizer. Finally, we say f is of class C^r, or f is C^r, if all partial derivatives of f of order up to r exist and are continuous.

First, the necessary condition:

Lemma 3.1 *If f is C^1 and x^o is a global or local maximizer, then $f_x(x^o) = 0$.*

Proof Let h be arbitrary. For every ε,

$$f(x^o + \varepsilon h) = f(x^o) + \varepsilon f_x(x^o)h + o(\varepsilon),$$

where $o(\varepsilon)/\varepsilon$ converges to 0 as ε converges to 0. That's what little o means. Since x^o is a local or global maximizer and $\mathcal{V}$ is open,

$$f(x^o) \geq f(x^o + \varepsilon h)$$

for every ε sufficiently small. Thus

$$f_x(x^o)h + \frac{o(\varepsilon)}{\varepsilon} \leq 0$$

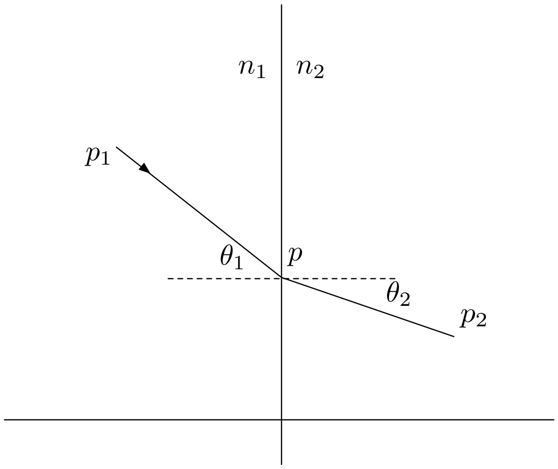
for sufficiently small $\varepsilon > 0$. Thus

$$f_x(x^o)h \leq 0.$$

Since h was arbitrary, $f_x(x^o) = 0$ □

Example

Fermat's principle (1662) is that the path between two points taken by a beam of light is the one that is traversed in the least time. **Snell's law** of refraction follows directly from this statement. To prove this, consider a light ray from a fixed point p_1 to a fixed point p_2. The two points are in different media, where the speeds of light in the media are respectively $c/n_1, c/n_2$.



Let p denote the point where the ray from p_1 to p_2 passes through the interface of the media. We want to find p according to Fermat's principle. Orient an (x, y) coordinate system as shown by the axes. The time for a ray to pass from p_1 to p_2 is

$$J = \frac{n_1}{c}\|p_1 - p\| + \frac{n_2}{c}\|p - p_2\|.$$

Let

$$p_1 = (x_1, y_1), \quad p_2 = (x_2, y_2), \quad p = (0, y) = y(0, 1) = ye_2$$

(e_2 is the unit vector along the y-axis). So J is a function of y,

$$J(y) = \frac{n_1}{c}\|p_1 - ye_2\| + \frac{n_2}{c}\|ye_2 - p_2\|,$$

and the problem is to find y to minimize J.

It can be shown that $J(y)$ is a continuously differentiable function with a unique minimum. Using the chain rule we have

$$\frac{dJ}{dy}(y) = -\frac{n_1}{c}\frac{1}{\|p_1 - ye_2\|}(p_1 - ye_2)^T e_2 - \frac{n_2}{c}\frac{1}{\|ye_2 - p_2\|}(p_2 - ye_2)^T e_2,$$

which simplifies to

$$\frac{dJ}{dy}(y) = \frac{n_1}{c}\frac{y - y_1}{\|p_1 - ye_2\|} + \frac{n_2}{c}\frac{y - y_2}{\|ye_2 - p_2\|}$$

and further to

$$\frac{dJ}{dy}(y) = -\frac{n_1}{c}\sin\theta_1 + \frac{n_2}{c}\sin\theta_2,$$

where the angles are shown in the figure. Setting the derivative to 0 gives

$$n_1 \sin\theta_1 - n_2 \sin\theta_2 = 0.$$

Thus

$$\frac{n_1}{n_2} = \frac{\sin\theta_2}{\sin\theta_1},$$

which is Snell's law. □

For the sufficient condition, we need the concept of positive definite matrix. Let Q be a real, square matrix. It is **positive semi-definite** (written $Q \geq 0$) if $x^TQx \geq 0$ for all x. It is **positive definite** (written $Q > 0$) if $x^TQx > 0$ for all $x \neq 0$. Negative definite and semi-definite are defined in the obvious way. If Q is symmetric, then it is positive semi-definite iff all its eigenvalues are ≥ 0, and positive definite iff they're all positive.

Let $H(x)$ denote the Hessian of f.

Lemma 3.2 *If f is C^2 and x^o satisfies*

$$x^o \in \mathcal{V}, \quad f_x(x^o) = 0, \quad H(x^o) < 0,$$

then x^o is a local maximizer.

Proof Fix $h \neq 0$. Then

$$f(x^o + \varepsilon h) = f(x^o) + \frac{\varepsilon^2}{2}h^T H(x^o)h + o(\varepsilon^2),$$

This can be written

$$f(x^o + \varepsilon h) = f(x^o) + \frac{\varepsilon^2}{2}\left[h^T H(x^o)h + \frac{o(\varepsilon^2)}{\varepsilon^2}\right].$$

The first term inside the brackets is negative, while the second term goes to zero as $\varepsilon \to 0$. Thus for ε sufficiently small

$$f(x^o + \varepsilon h) \leq f(x^o).$$

Since h was arbitrary, x^o is a local maximizer. □

Example

$$f(x) = \frac{1}{2}x^T Ax + b^T x + c, \quad A \text{ symmetric}$$

We have

$$f_x(x) = x^T A + b^T, \quad f_{xx}(x) = A.$$

Thus a local optimum exists only if the equation

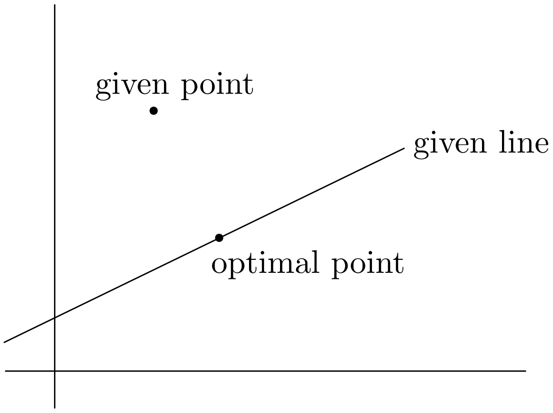
$$x^T A + b^T = 0$$

has a solution; that is, b belongs to the span of the columns of A. Then, if x^o is a solution of this equation and if A is negative definite, then x^o is a local maximizer. □

3.3 Optimizing a Quadratic Function with Equality Constraints

Let's begin with a very simple example:

Example In the plane, find the point on a given line that is closest to a given point:



This is a distance problem. Obviously, you can get the closest point by drawing the perpendicular from the given point to the given line.

Before we solve this problem, let's clarify some notation. The norm of $x = (x_1, x_2)$ is

$$\|x\| = \left(x_1^2 + x_2^2\right)^{1/2},$$

and this can also be written $\|x\| = (x^T x)^{1/2}$, that is,

$$x^T x = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + x_2^2.$$

To develop a solution method, suppose the given point is $v = (1, 2)$ and the equation of the given line is

$$x_2 = 0.5x_1 + 0.2.$$

Let $x = (x_1, x_2)$ be the point being sought. Define

$$c^T = \begin{bmatrix} -0.5 & 1 \end{bmatrix}, \quad b = 0.2.$$

Then x is on the line iff $c^T x = b$. Also, the distance from v to x is $\|v - x\|$. Note that $\|v - x\|$ is minimum iff $\|v - x\|^2$ is minimum. Thus we have arrived at the following equivalent problem: minimize the quadratic function $\|v - x\|^2$ of x subject to the equality constraint $c^T x = b$. Notice that

$$\|v - x\|^2 = (v - x)^T (v - x) = v^T v - v^T x - x^T v + x^T x.$$

The right-hand side is a quadratic function of x. Since $x^T v = v^T x$ (dot product of real vectors is symmetric), we have

$$\|v - x\|^2 = v^T v - 2v^T x + x^T x.$$

So we've reduced the problem to

$$\min_{x,\ c^T x = b} v^T v - 2v^T x + x^T x.$$

We'll return to this after we review some calculus.

Aside: This specific problem is easy to solve this way: Substitute the constraint $x_2 = 0.5x_1 + 0.2$ into

$$(1 - x_1)^2 + (2 - x_2)^2,$$

to get a function $f(x_1)$. Set the derivative of f to zero, solve for x_1, then get x_2. The answers are $x_1 = 1.52$, $x_2 = 0.96$. □

Lagrange Multipliers

Now we return to the first example in this section. It had the form

$$\min_{x,\ c^T x = b} f(x), \quad f(x) = v^T v - 2v^T x + x^T x, \quad c^T = \begin{bmatrix} -0.5 & 1 \end{bmatrix}, \quad b = 0.2.$$

We are going to use the method of **Lagrange multipliers**. The idea is to absorb the constraint $c^T x = b$, or equivalently $c^T x - b = 0$, into the function being minimized, leaving an unconstrained problem. Define the **Lagrangian**

$$L(x, \lambda) = f(x) + \lambda(c^T x - b).$$

Here λ is an unknown that multiplies the constraint equation. It turns out a necessary condition for optimality of x is that L should be stationary with respect to both x and λ, that is,

$$L_x = 0, \quad L_\lambda = 0.$$

These two equations give

$$f_x + \lambda c^T = 0, \quad c^T x - b = 0,$$

or, using the form of f,

$$-2v^T + 2x^T + \lambda c^T = 0, \quad c^T x - b = 0.$$

Finally, taking transpose and rearranging, we have

$$2x + \lambda c = 2v, \quad c^T x = b.$$

These can be assembled into one equation:

$$\begin{bmatrix} 2I & c \\ c^T & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} 2v \\ b \end{bmatrix}.$$

Let's put in our values for v, c, b:

$$\begin{bmatrix} 2 & 0 & -0.5 \\ 0 & 2 & 1 \\ -0.5 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \lambda \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 0.2 \end{bmatrix}.$$

This has a unique solution because the matrix is invertible:

$$x = (1.52, 0.96), \quad \lambda = 2.08.$$

The x is the optimal x, the closest point, and the λ can be discarded—it was introduced only to solve the problem. □

Let's look at a somewhat more general problem by the Lagrange multiplier method.

Example We'll solve the problem

$$\text{minimize}_x \|c - Ax\|$$

subject to the constraint $Bx = d$. Here x, c, d are vectors and A, B matrices. Assume A has full column rank and B has full row rank.

Define

$$\begin{aligned} J(x) = \|c - Ax\|^2 &= (c - Ax)^T (c - Ax) \\ &= c^T c - c^T Ax - x^T A^T c + x^T A^T Ax \\ &= c^T c - 2c^T Ax + x^T A^T Ax \end{aligned}$$

and

$$L(x, \lambda) = J(x) + \lambda^T (Bx - d).$$

Here the Lagrange multiplier has to be a vector. Differentiating with respect to x then λ, we get

$$-2c^TA + 2x^TA^TA + \lambda^TB = 0, \quad Bx - d = 0.$$

Transposing the first gives

$$-2A^Tc + 2A^TAx + B^T\lambda = 0, \quad Bx - d = 0.$$

Collect as one equation:

$$\begin{bmatrix} 2A^TA & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} 2A^Tc \\ b \end{bmatrix}.$$

If it can be proved that the matrix on the left is invertible, then the optimal x is

$$x = \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} 2A^TA & B^T \\ B & 0 \end{bmatrix}^{-1} \begin{bmatrix} 2A^Tc \\ b \end{bmatrix}.$$

So let's see that the matrix

$$\begin{bmatrix} 2A^TA & B^T \\ B & 0 \end{bmatrix}$$

is invertible. It suffices to prove that the only solution to the homogeneous equation

$$\begin{bmatrix} 2A^TA & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = 0$$

is the trivial solution. So start with

$$\begin{bmatrix} 2A^TA & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = 0.$$

Thus

$$2A^TAx + B^T\lambda = 0, \quad Bx = 0.$$

Since A has full column rank, the matrix A^TA is positive definite, hence invertible. Thus

$$x + (2A^TA)^{-1}B^T\lambda = 0, \quad Bx = 0.$$

Multiply the first equation by B and use the second:

$$B(2A^TA)^{-1}B^T\lambda = 0.$$

Pre-multiply by λ^T:

$$\lambda^TB(2A^TA)^{-1}B^T\lambda = 0.$$

Since $(2A^TA)^{-1}$ is positive definite, it follows that $B^T\lambda = 0$. Then, since B^T has full column rank, $\lambda = 0$. Finally, from the equation

$$x + (2A^TA)^{-1}B^T\lambda = 0,$$

we get that $x = 0$. Thus $x = 0, \lambda = 0$ is the only solution of

$$\begin{bmatrix} 2A^TA & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = 0.$$

□

Why the Lagrange multiplier method works

Consider the problem of minimizing a function $f(x)$ subject to an equality constraint $g(x) = 0$.

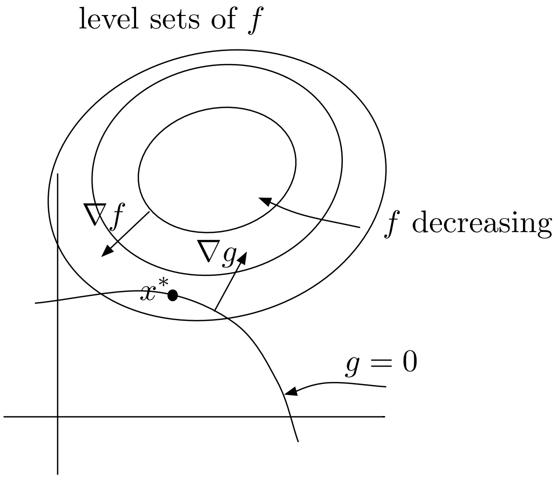
To be able to draw pictures, let's suppose

$$f, g : \mathbb{R}^2 \longrightarrow \mathbb{R}.$$

The set of all x satisfying the constraint $g(x) = 0$ typically is a curve. For a given constant c, the set of all x satisfying $f(x) = c$ is called a level set of f. Now assume x^o is a locally optimal point for the problem $\min_{g(x)=0} f(x)$. That is, if x is nearby x^o and $g(x) = 0$, then $f(x) > f(x^o)$.

Claim The gradients $\nabla f(x^o)$, $\nabla g(x^o)$ are collinear.

Proof The picture near x^o looks like this:



From this, the claim is clear. □

Thus there is a scalar λ^o such that $\nabla f(x^o) + \lambda^o \nabla g(x^o) = 0$. This implies the gradient of the function

$$f(x) + \lambda^o g(x)$$

equals zero at x^o. Finally, this implies the Lagrangian

$$L(x, \lambda) = f(x) + \lambda g(x)$$

satisfies

$$L_x(x^o, \lambda^o) = 0, \quad L_\lambda(x^o, \lambda^o) = 0.$$

In conclusion, a necessary condition for a point x^o to be a local optimum for the problem $\min_{g(x)=0} f(x)$ is that there exist a point λ^o such that the derivative of the Lagrangian $L(x, \lambda)$ equals zero at x^o, λ^o.

3.4 Optimization with Equality Constraints

Now we take a more general view. Namely, the problem is to maximize $f(x)$ over all x satisfying the constraint $g(x) = 0$. Here

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}, \quad g : \mathbb{R}^n \longrightarrow \mathbb{R}^m.$$

We begin with some definitions. A subset $\mathcal{A}$ of $\mathbb{R}^n$ is **closed** if it contains the limit of every convergent sequence of points in $\mathcal{A}$; that is, if $\{x_k\}$ is a sequence in $\mathcal{A}$ that converges to a point in $\mathbb{R}^n$, then that limit is actually in $\mathcal{A}$. A subset $\mathcal{A}$ of $\mathbb{R}^n$ is **bounded** if there exists $r > 0$ such that $\|x\| \leq r$ for every $x \in \mathcal{A}$. A closed and bounded set is said to be **compact**. (This is not actually the definition of a compact set, but in finite dimensional space it's equivalent.)

Now we look at the constraint set $\mathcal{C} := \{x : g(x) = 0\}$. If g is continuous, $\mathcal{C}$ is closed. This is pretty immediate: If $\{x_k\}$ is a sequence such that $g(x_k) = 0$ for all k and if the sequence converges to, say, x, then $g(x) = 0$ by continuity.

Now it's a fact from analysis that a continuous function on a compact set achieves its maximum. Thus if f and g in the given problem are continuous and if $\mathcal{C}$ is bounded (and therefore compact), then the problem

$$\max_{x \in \mathcal{C}} f(x)$$

is solvable—there is a maximizer. The trouble is that frequently $\mathcal{C}$ isn't bounded; of course, this doesn't imply there isn't a maximizer.

Now we see the mathematical justification of the method of Lagrange multipliers. For the problem at hand, define the Lagrangian

$$L(x, \lambda) = f(x) + \lambda^T g(x).$$

Theorem 3.1 *Suppose f, g are C^2, the problem $\max_{\mathcal{C}} f(x)$ has a local solution x^o, and $g_x(x^o)$ is surjective. Then there exists a vector λ^o such that*

$$L_x(x^o, \lambda^o) = 0, \quad L_\lambda(x^o, \lambda^o) = 0.$$

Proof We'll do only the simpler case where g is linear, $g(x) = Ax$. Then $g_x(x) = A$ and the hypothesis is that A has rank m. Suppose without loss of generality that

$$A = \left[\begin{array}{cc} B & C \end{array} \right],$$

B invertible. Partition x correspondingly:

$$x = (y, z), \quad y \in \mathbb{R}^m.$$

Then

$$\begin{aligned} & g(y,z) = 0 \\ \iff \ & By + Cz = 0 \\ \iff \ & y = -B^{-1}Cz. \end{aligned}$$

So the constrained problem

$$\max_{g(x)=0} f(x)$$

is equivalent to the unconstrained problem

$$\max_{z} f(-B^{-1}Cz, z).$$

Define

$$h(z) = f(-B^{-1}Cz, z).$$

Since x^o is an optimizer,

$$z^o = \left[\begin{array}{cc} 0 & I \end{array} \right] x^o$$

is an optimizer of h. Thus we have in turn

$$\begin{aligned} & h_z(z^o) = 0 \\ & -f_y(y^o, z^o)B^{-1}C + f_z(y^o, z^o) = 0 \\ & f_x(x^o) \left[\begin{array}{c} -B^{-1}C \\ I \end{array} \right] = 0. \end{aligned} \tag{3.1}$$

Define

$$\lambda^{*T} = f_x(x^o) \left[\begin{array}{c} -B^{-1} \\ 0 \end{array} \right].$$

Then

$$\begin{aligned} L_x(x^o, \lambda^o) &= f_x(x^o) + \lambda^{*T} g_x(x^o) \\ &= f_x(x^o) + \lambda^{*T} A \\ &= f_x(x^o) + \lambda^{*T} \left[\begin{array}{cc} B & C \end{array} \right] \\ &= f_x(x^o) + f_x(x^o) \left[\begin{array}{c} -B^{-1} \\ 0 \end{array} \right] \left[\begin{array}{cc} B & C \end{array} \right] \\ &= f_x(x^o) + f_x(x^o) \left[\begin{array}{cc} -I & -B^{-1}C \\ 0 & 0 \end{array} \right] \\ &= f_x(x^o) \left[\begin{array}{cc} 0 & -B^{-1}C \\ 0 & I \end{array} \right]. \end{aligned}$$

The last right-hand side equals 0 from (3.1).

Finally, the other equation, $L_\lambda(x^o, \lambda^o) = 0$ is satisfied because

$$L_\lambda = g(x).$$

□

3.5 Application: Sensor Placement

Where should we place sensors to optimize coverage?

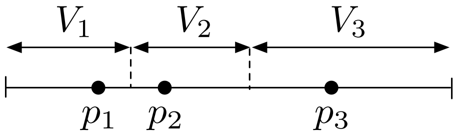
Lloyd's algorithm in 1D

This algorithm was originally developed for the problem of quantizing data. Let r be a real number that could take any value in the interval $[0, 1]$. We want to partition $[0, 1]$ into a finite number, n, of subintervals, $\{V_i\}_{i=1,\dots,n}$, and then, for each i, designate one point p_i in V_i as the codeword. Then the quantization function would be to map r to p_i if $r \in V_i$. The partition and code book are optimal in a certain sense.

There's a minor but annoying difficulty with the boundaries of $\{V_i\}_{i=1,\dots,n}$. Strictly speaking these intervals should not overlap; that is, every point should be in one and only one V_i. But this complicates the derivation to the point where it obscures the ideas. So we'll take the subintervals to be closed and ignore the case where a point lies on the boundaries of two subintervals.

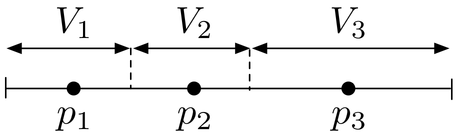
The algorithm is illustrated by an example.

Example ($n = 3$) Let $p_1 < p_2 < p_3$ be three arbitrary points in $[0, 1]$. Construct a partition $\{V_1, V_2, V_3\}$ as shown here:

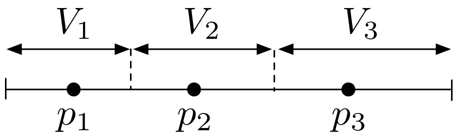


So V_1 is from 0 to the midpoint between p_1 and p_2, $(p_1+p_2)/2$; V_2 is from $(p_1+p_2)/2$ to $(p_2+p_3)/2$; and V_2 is from $(p_2 + p_3)/2$ to 1. This is called the *Voronoi partition*[1] $\mathcal{V}$ generated by $\{p_i\}$; the intervals are uniquely defined by this property: V_i is the set of all points q whose distance from p_i is less than or equal to the distances from all other p_j:

Continuing with the algorithm, update p_i to be the centre c_i of V_i:



Then update V_i to be the Voronoi partition:



[1] Named after the Ukrainian mathematician Georgy Voronoi (1868–1908).

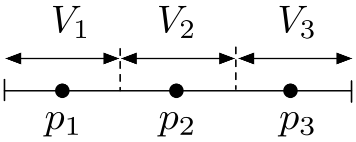
And so on. Does this procedure converge? Let p be the vector (p_1, p_2, p_3). Then the update law is

$$p(k+1) = Ap(k) + b,$$

where $b = (0, 0, 1/2)$ and

$$A = \frac{1}{4} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

You're invited to prove that $p(k)$ converges to the vector $(1/6, 1/2, 5/6)$:



Thus the intervals have equal width and the points are their centres, just what you'd like for a quantizer. □

Continuous time

There's a natural continuous-time version of the algorithm.

Example (continued) Think now of p_i, c_i, and V_i as evolving in continuous time (c_i is the centre of V_i):

$$\dot{p}_1 = c_1 - p_1, \quad \dot{p}_2 = c_2 - p_2, \quad \dot{p}_3 = c_3 - p_3.$$

This leads to

$$\dot{p} = Ap + b,$$

where $b = (0, 0, 1/2)$ and

$$A = \frac{1}{4} \begin{bmatrix} -3 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -3 \end{bmatrix}.$$

Again, $p(t)$ converges to the vector $(1/6, 1/2, 5/6)$. □

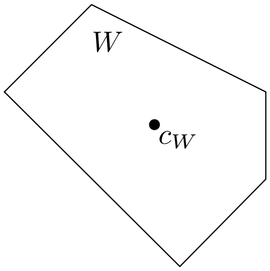
Lloyd's algorithm in 2D

Consider a convex polytope W in $\mathbb{R}^2$ of area A_W. Its centroid (point of balance) c_W satisfies

$$\int_W (q - c_W) dq = 0,$$

and therefore

$$c_W = \frac{1}{A_W} \int_W q dq.$$



The polar moment of inertia of W about a point $p \in W$ is

$$\mathcal{H}(p, W) = \int_W \|q - p\|^2 dq.$$

The parallel axis theorem is a standard result in mechanics:

Lemma 3.3 *For every p in W*

$$\mathcal{H}(p, W) = \mathcal{H}(c_W, W) + A_W \|p - c_W\|^2.$$

Proof The statement is

$$\int_W \|q - p\|^2 dq = \int_W \|q - c_W\|^2 dq + \int_W \|p - c_W\|^2 dq.$$

Now the quantity

$$\left(\int_W \|f(q)\|^2 dq \right)^{1/2}$$

is a norm on the function $f : W \longrightarrow \mathbb{R}^2$, so let's write this quantity as $\|f\|$. Specifically, define $f(q) = q - c_W$ (affine-linear function) and $g(q) = c_W - p$ (constant function). Then we're trying to show

$$\|f + g\|^2 = \|f\|^2 + \|g\|^2,$$

which is an instance of Pythagoras' theorem. So all we have to show is that $f \perp g$:

$$\begin{aligned} \langle f, g \rangle &= \int_W f(q)^T g(q) dq \\ &= \int_W (q - c_W)^T (c_W - p) dq \\ &= \int_W q^T dq (c_W - p) - \int_W dq c_W^T (c_W - p) \\ &= A_W c_W^T (c_W - p) - A_W c_W^T (c_W - p) \\ &= 0. \end{aligned}$$

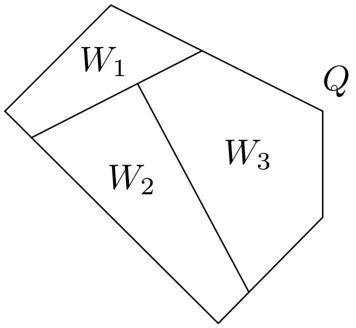
□

Corollary 3.1 *The unique point p that minimizes $\mathcal{H}(p, W)$ is the centroid, $p = c_W$.*

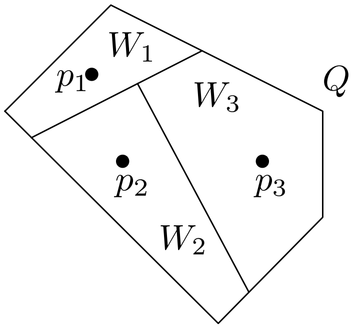
Let's interpret that last result in terms of sensor placement. Suppose we want to place a sensor at a location p in W to optimize coverage. We take $\mathcal{H}(p, W)$ as the cost function, a measure of coverage error—the sum over q of the squares of the distances $\|q - p\|$. Then the optimal location for the sensor is the centroid.

Fixed partition

Let's extend to n sensors. Consider a convex polytope Q in $\mathbb{R}^2$. Suppose $\mathcal{W} = \{W_i\}_{i=1,\ldots,n}$ is a given partition:



Now suppose there are n sensors that are to be placed at locations $\{p_i\}$, one in each cell: $p_i \in W_i$:



The cost function for cell i is $\mathcal{H}(p_i, W_i)$ and the total cost function is

$$\mathcal{H}(p, \mathcal{W}) = \mathcal{H}(p_1, W_1) + \cdots + \mathcal{H}(p_n, W_n),$$

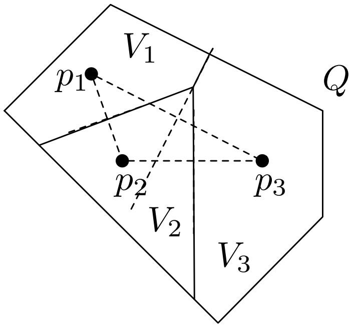
where $p = (p_1, \ldots, p_n)$ denotes the vector of sensor positions. Since $W_1, \ldots, W_n$ are all disjoint,

$$\min_p \mathcal{H}(p, \mathcal{W}) = \min_{p_1} \mathcal{H}(p_1, W_1) + \cdots + \min_{p_n} \mathcal{H}(p_n, W_n).$$

Thus the optimal p_i is the centroid of W_i.

Fixed sensor locations

Now let's suppose the n sensor locations p are fixed but the n cells $\mathcal{W} = \{W_i\}$ are to be designed. The optimal partition turns out to be the Voronoi partition $\mathcal{V}$:



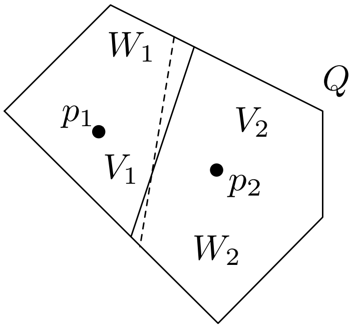
In mathematical terms,

$$V_i = \{q : (\forall j \neq i)\|q - p_i\| \leq \|q - p_j\|\}.$$

Each V_i is the intersection of half planes. The picture just shown is called a Voronoi diagram, and the partition is uniquely determined by p.

Lemma 3.4 *For a given p, the unique partition that minimizes $\mathcal{H}(p, \mathcal{W})$ is the Voronoi partition, $\mathcal{W} = \mathcal{V}$.*

Proof Let's do the case $n = 2$ for simplicity of explanation. Here's the picture:



The solid line through Q defines the Voronoi partition; it bisects the line joining p_1 and p_2. Let $\mathcal{W}$ be any other partition, shown by the dashed line. We'll show that

$$\mathcal{H}(p, \mathcal{V}) \leq \mathcal{H}(p, \mathcal{W}),$$

that is,

$$\int_{V_1} \|q - p_1\|^2 dq + \int_{V_2} \|q - p_2\|^2 dq \leq \int_{W_1} \|q - p_1\|^2 dq + \int_{W_2} \|q - p_2\|^2 dq. \tag{3.2}$$

Let χ_V denote the characteristic function of a set V, that is,

$$\chi_V(q) = 1 \text{ if } q \in V, \quad \chi_V(q) = 0 \text{ if not.}$$

Then (3.2) is equivalent to

$$\int_Q \left[\|q-p_1\|^2\chi_{V_1}(q)+\|q-p_2\|^2\chi_{V_2}(q)\right]dq \le$$

$$\int_Q \left[\|q-p_1\|^2\chi_{W_1}(q)+\|q-p_2\|^2\chi_{W_2}(q)\right]dq.$$

So it suffices to prove that for every q

$$\|q-p_1\|^2\chi_{V_1}(q)+\|q-p_2\|^2\chi_{V_2}(q) \le \|q-p_1\|^2\chi_{W_1}(q)+\|q-p_2\|^2\chi_{W_2}(q). \tag{3.3}$$

Let $q \in V_1$. If $q \in W_1$, then

$$\|q-p_1\|^2\chi_{V_1}(q)=\|q-p_1\|^2\chi_{W_1}(q)$$

and so (3.3) is true with equality; whereas, if $q \in W_2$, then

$$\|q-p_1\|^2\chi_{V_1}(q)\le\|q-p_2\|^2\chi_{W_2}(q)$$

and so (3.3) is true.

Likewise if $q \in V_2$.

□

Fixed number of sensors

Now we turn to the more interesting problem: Both the sensor locations and the cells are designable—only the overall set Q and the number n of sensors are given and fixed. The problem is to minimize $\mathcal{H}(p,\mathcal{W})$ over both p and $\mathcal{W}$.

Lloyd's algorithm is this:

```
Step 0: Start with an arbitrary partition $\{W_i\}$ and arbitrary points $\{p_i\}$, $p_i \in W_i$.

Step 1: Construct the unique Voronoi partition $\{V_i\}$ generated by $\{p_i\}$.

Step 2: Update $p_i$ to be the centroid of $V_i$.

        Return to Step 1.
```

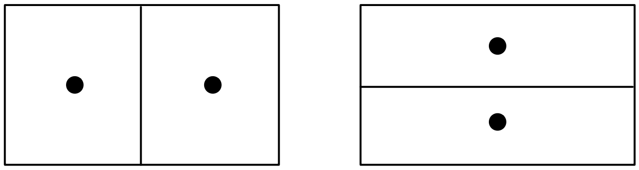
The rationale for the algorithm is this: Regarding Step 1, by Lemma 3.4

$$\mathcal{H}(p,\mathcal{W}) \ge \mathcal{H}(p,\mathcal{V}).$$

Then, regarding Step 2, by Lemma 3.3

$$\begin{aligned}\mathcal{H}(p,\mathcal{V}) &= \sum_i \left[\mathcal{H}(c_{V_i},V_i)+A_{V_i}\|p_i-c_{V_i}\|^2\right]\\ &\ge \sum_i \mathcal{H}(c_{V_i},V_i)\\ &= \mathcal{H}(p_{\text{updated}},\mathcal{V}).\end{aligned}$$

The procedure converges asymptotically to a Voronoi partition with p_i being the centroid of V_i. However, the limit may be only a local optimum for the function $\mathcal{H}$. For example:



If the algorithm is initialized in either of the two ways shown, it terminates immediately. But the right-hand value of $\mathcal{H}$ is larger than the left-hand value.

References

1. J. Cortes, S. Martinez, T. Karatas, and F. Bullo, "Coverage control for mobile sensing networks," *IEEE Transactions on Robotics and Automation*, 20(2): 243–255, 2004.

2. S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982; the material in this paper was presented in part at the Institute of Mathematical Statistics Meeting, Atlantic City, NJ, September, 1957.

3.6 Problems

1. Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be defined by

$$f(x) = \begin{cases} x^2 \sin(1/x), & x \neq 0 \\ 0, & x = 0. \end{cases}$$

Prove that f is differentiable but its derivative is not continuous.

2. Define $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ by

$$f(x_1, x_2) = \begin{cases} x_1 x_2 \dfrac{x_1^2 - x_2^2}{x_1^2 + x_2^2}, & x \neq 0 \\ 0, & x = 0. \end{cases}$$

Find $f_{xx}(0)$.

3. Consider

$$f(x_1, x_2) = (x_2 - x_1^2)^2 + (1 - x_1)^2.$$

Find the global minimizer, if it exists.

4. Consider the problem of maximizing $f(x)$ subject to $g(x) = 0$, where

$$f(x) = \frac{1}{2} x^T A x + b^T x + c, \quad g(x) = Dx.$$

Assume A is symmetric and $A < 0$, and D is surjective. What can you conclude about existence and uniqueness of a solution?

5. Here we apply the theory to a state-space regulation problem. Consider

$$x(k+1) = Ax(k) + Bu(k), \quad A \in \mathbb{R}^{n\times n}, \quad B \in \mathbb{R}^{n\times m}.$$

Suppose, given $x(0)$, we want to drive the state $x(k)$ to the origin. There are at least two ways to do this. The feedback method is to choose F so that $A+BF$ is nilpotent, if possible, and then set $u = Fx$. We'll look at the other method, open-loop control. We have

$$x(1) = Ax(0) + Bu(0)$$

and so on until

$$x(n) = A^n x(0) + W\tilde{u},$$

where

$$W = \left[\begin{array}{cccc} B & AB & \cdots & A^{n-1}B \end{array} \right], \quad \tilde{u} = (u(n-1), \ldots, u(0)).$$

We assume (A, B) is controllable. Therefore W is surjective and there exists $\tilde{u}$ such that $x(n) = 0$. We propose to choose $\tilde{u}$ such that $x(n) = 0$ and $\|\tilde{u}\|^2$ is minimum. The idea is to drive the state to the origin using minimum energy. This is precisely our constrained optimization problem with

$$f(\tilde{u}) = \|\tilde{u}\|^2, \quad g(\tilde{u}) = A^n x(0) + W\tilde{u}.$$

Solve this optimization problem.

6. Consider the nonlinear differential equation

$$\dot{x} = -x(\|x\|^2 - 1),$$

where x is a vector. Find the equilibrium points. Linearize the equation at every equilibrium and see if you can conclude anything about local stability.

7. Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be given by

$$f(x) = \frac{1}{\|x\|} x.$$

The norm is the Euclidean norm, $\|x\| = (x^T x)^{1/2}$. The definition of f makes sense as long as $x \neq 0$; for completeness you can set $f(0) = 0$ (the value is irrelevant). Compute $f_x(x)$, the Jacobian of f at $x \neq 0$.

8. Solve the problem

$$\min_x \|c - Ax\|.$$

Show that there always exists a solution. When is the solution unique?

9. Find the vector in $A\mathrm{Ker}B$ that is closest to c, where

$$A = \begin{bmatrix} 1 & 1 \\ 2 & -2 \\ 3 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad c = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

10. Show that the equation $A^T Ax = A^T b$ is always solvable.

11. This is a problem on sensor placement. Suppose we want to place a sensor to detect, say, temperature.

    (a) Suppose the workspace is the unit interval $[0, 1]$. We want to place a sensor at a location $p \in [0, 1]$ to get "optimal coverage." To make this precise, we have to define a measure of *coverage error*, denoted $H(p)$. Here are some options:

    If q is another point, how well the sensor can measure the temperature at location q depends on the distance $|q - p|$. Suppose the temperature error is in fact proportional to $|q - p|$. Then the average error is proportional to

    $$H_1(p) = \int_0^1 |q - p| dq$$

    while the worst case error is proportional to

    $$H_\infty(p) = \max_{0 \le q \le 1} |q - p|.$$

    Suppose the temperature error is in fact proportional to $|q - p|^2$. Then the average error is proportional to

    $$H_2(p) = \int_0^1 |q - p|^2 dq.$$

    Find the optimal p (the one that minimizes H) in each of the three cases. Is the optimal p the centre of the interval in all three cases?

    (b) Now consider the extension to a convex polygon region in $\mathbb{R}^2$. As in the course notes, for $H_2(p)$, the optimal p is the centroid of the region. Give an example where this isn't true for $H_1(p)$ or $H_\infty(p)$.

Part I

Classical Theories

Chapter 4

Calculus of Variations

References: *Calculus of Variations*, I. M. Gelfand and S. V. Fomin; *Optimization by Vector Space Methods*, D. G. Luenberger.

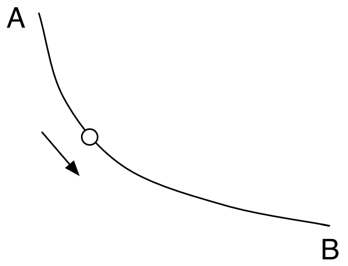
4.1 The Brachistochrone Problem

"Optimal control was born in 1697—300 years ago-in Groningen, a university town in the north of The Netherlands, when Johann Bernoulli, professor of mathematics at the local university from 1695 to 1705, published his solution of the brachystochrone problem." So begins the article

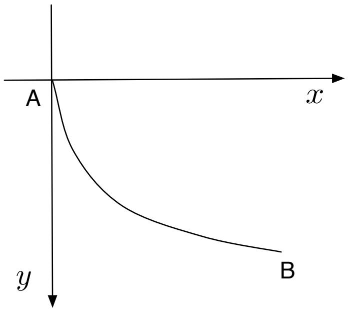
> "300 Years of Optimal Control: From the Brachistochrone to the Maximum Principle,"
> H.J. Sussman and J.C. Willems, IEEE Control Systems Magazine, 1997,

Here we study the brachistochrone problem and in a later chapter, the maximum principle. Brachistochrone means "shortest time."

A tiny spherical wooden bead with a hole drilled through it slides from rest without friction along a rigid wire:



The starting point A is higher in elevation than the end point B, and the curve of the wire lies in a vertical plane like this:



We want the bead to slide under the force of gravity from A to B. The two points A and B are fixed in space, but the wire curve is free for us to design. For what curve does the bead slide from A to B in minimum time? It's not the straight line from A to B.

This is the brachistochrone problem. It was worked on by Newton, the Bernoulli brothers, and other great scientists. The problem is harder than a simple calculus problem because we're looking for an optimal curve instead of an optimal number or vector. The space of curves is infinite dimensional. Let a candidate curve be

$$y(x), \quad \text{at A} : y(0) = 0, \quad \text{at B} : y(x_1) = y_1.$$

So x_1, y_1 are given and the curve $y(x)$ is to be found.

Let t_1 denote the time it takes for the bead to slide from A to B. We have to bring in some physics to specify the time t_1 as a function of the curve. Let the bead start from A at time $t = 0$, let s denote the distance along the curve from A to where the bead is at time t, and let $v = \dot{s}$, where dot denotes derivative with respect to t.

Claim $v^2 = 2gy$
This follows from the conservation of energy:

$$\frac{1}{2}mv^2 = mgy.$$

But here's another derivation:

Proof of claim The force vertically down on the bead is mg and therefore the force tangent to the curve is

$$mg\frac{dy}{ds}.$$

Newton's second law gives

$$mg\frac{dy}{ds} = m\ddot{s},$$

i.e.,

$$g\frac{dy}{ds} = \ddot{s}.$$

Multiply by $2\dot{s}$:

$$2g\dot{y} = 2\dot{s}\ddot{s}.$$

Integrate:

$$\int_0^t 2g\dot{y}d\tau = \int_0^t 2\dot{s}\ddot{s}d\tau.$$

Thus

$$2gy + c = \dot{s}^2 = v^2.$$

At $t = 0$, $y = v = 0$; therefore $c = 0$ and we have proved the claim. □

Next, we have

$$ds^2 = dx^2 + dy^2,$$

and therefore

$$\dot{s}^2 = \dot{x}^2 + \dot{y}^2 = \dot{x}^2 + y'^2\dot{x}^2.$$

where prime denotes derivative with respect to x. On the left replace $\dot{s}^2$ by $v^2 = 2gy$:

$$2gy = (1 + y'^2)\dot{x}^2.$$

Thus

$$dt = \sqrt{\frac{1 + y'^2}{2gy}}dx.$$

Integrate:

$$t_1 = \int_0^{x_1} \sqrt{\frac{1 + y'(x)^2}{2gy(x)}}dx.$$

Changing notation, we have arrived at the problem of finding a curve $x(t)$ to minimize

$$J(x) = \int_{t_1}^{t_2} \left(\frac{1 + \dot{x}^2}{2gx}\right)^{1/2} dt$$

subject to the constraints that $x(t_1), x(t_2)$ are fixed.

4.2 The General Problem

The general problem involves

$$J(x) = \int_{t_1}^{t_2} f[t, x(t), \dot{x}(t)]dt,$$

where $x(t_1) = x_1, x(t_2) = x_2$ are fixed. The function f maps $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}$ and is assumed to be of class C^2.

Let us denote by $\mathcal{X}$ the vector space of C^1 functions $x : \mathbb{R} \longrightarrow \mathbb{R}^n$ and by $\mathcal{X}_a$ the subset such that $x(t_1) = x_1$ and $x(t_2) = x_2$. This latter is the set of admissible curves. The problem is to find $x \in \mathcal{X}_a$ to minimize $J(x)$.

Thus

$$J : \mathcal{X} \longrightarrow \mathbb{R}.$$

A function like this whose domain is a function space and whose co-domain is the reals is usually called a **functional**. Other interesting examples are the length of a curve, the area surrounded by a closed curve, etc.

4.3 The Euler-Lagrange Equation

The Euler-Lagrange equation is a necessary condition for a function to be optimal.

Theorem 4.1 *If $x^o \in \mathcal{X}_a$ minimizes J, then x^o satisfies the equation*

$$f_x = \frac{d}{dt} f_{\dot{x}},$$

that is, for all $t_1 \leq t \leq t_2$

$$f_x[t, x^o(t), \dot{x}^o(t)] = \frac{d}{dt} f_{\dot{x}}[t, x^o(t), \dot{x}^o(t)].$$

Let's finish the brachistochrone problem before the proof.

Example The brachistochrone problem.

We have

$$f(t, x, \dot{x}) = \left(\frac{1 + \dot{x}^2}{2gx} \right)^{1/2}.$$

Thus

$$f_x = -\frac{1}{2x} \left(\frac{1 + \dot{x}^2}{2gx} \right)^{1/2}$$

$$f_{\dot{x}} = \frac{1}{\sqrt{2gx}} \frac{\dot{x}}{(1 + \dot{x}^2)^{1/2}}$$

$$\frac{d}{dt} f_{\dot{x}} = -\frac{\dot{x}}{2x\sqrt{2gx}} \frac{\dot{x}}{(1+\dot{x}^2)^{1/2}} + \frac{1}{\sqrt{2gx}} \frac{\ddot{x}}{(1+\dot{x}^2)^{3/2}}.$$

Thus the Euler-Lagrange equation is

$$-\frac{1}{2x}\left(\frac{1+\dot{x}^2}{2gx}\right)^{1/2} = -\frac{\dot{x}}{2x\sqrt{2gx}} \frac{\dot{x}}{(1+\dot{x}^2)^{1/2}} + \frac{1}{\sqrt{2gx}} \frac{\ddot{x}}{(1+\dot{x}^2)^{3/2}}.$$

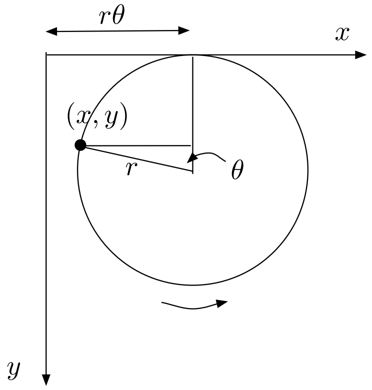
This simplifies to

$$2x\ddot{x} + \dot{x}^2 + 1 = 0.$$

Let's return to the original variables: $y(x)$ instead of $x(t)$:

$$2yy'' + y'^2 + 1 = 0. \tag{4.1}$$

Instead of solving this equation, it's easier to propose a solution and then verify it. The path is a cycloid, that is, a curve generated by a fixed point moving on a rolling wheel:



The wheel rolls along the x-axis as shown. The black dot traces out a cycloid. Thus r is constant, $\theta(t)$ is a function of time, $\theta(0) = 0$, and

$$x = r\theta - r\sin\theta, \quad y = r - r\cos\theta.$$

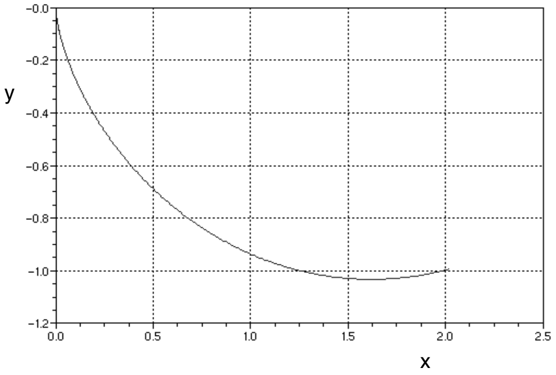
Let us verify that this path satisfies (4.1). We have

$$y' = \frac{\dot{y}}{\dot{x}} = \frac{r\dot{\theta}\sin\theta}{r\dot{\theta}(1-\cos\theta)} = \frac{\sin\theta}{1-\cos\theta}$$

$$y'' = \frac{1}{\dot{x}}\frac{d}{dt}y' = \frac{1}{r\dot{\theta}(1-\cos\theta)}\frac{d}{dt}\frac{\sin\theta}{1-\cos\theta} = -\frac{1}{r(1-\cos\theta)^2}.$$

Substitute these into (4.1).

For the terminal point $(2, -1)$, the graph is this:



It's interesting that the curve goes up near the end. □

The proof of Theorem 4.1 requires a lemma. Recall the spaces $\mathcal{X}$ and $\mathcal{X}_a$. Define $\mathcal{X}_0$ to be the subspace of $\mathcal{X}$ of functions $h(t)$ that equal zero at the two times t_1, t_2.

Lemma 4.1 *Suppose $y(t)$ is a continuous function, $y(t) \in \mathbb{R}^n$, and*

$$(\forall h \in \mathcal{X}_0) \int_{t_1}^{t_2} y(t)^T \dot{h}(t) dt = 0.$$

Then $y(t)$ is a constant vector on $[t_1, t_2]$.

Proof Define the vector c via

$$\int_{t_1}^{t_2} [y(t) - c] dt = 0$$

and let

$$h(t) = \int_{t_1}^{t} [y(\tau) - c] d\tau.$$

Then $h \in \mathcal{X}_0$ and

$$\begin{aligned}
\int_{t_1}^{t_2} \|y(t) - c\|^2 dt &= \int_{t_1}^{t_2} [y(t) - c]^T [y(t) - c] dt \\
&= \int_{t_1}^{t_2} [y(t) - c]^T \dot{h}(t) dt \\
&= \int_{t_1}^{t_2} y(t)^T \dot{h}(t) dt - c^T [h(t_2) - h(t_1)] \\
&= 0.
\end{aligned}$$

Thus $y(t) = c$. □

Proof of Theorem 4.1 Let $h \in \mathcal{X}_0$. Then for every ε, $x^o + \varepsilon h$ is in $\mathcal{X}_a$ and so $J(x^o + \varepsilon h)$ has a minimum at $\varepsilon = 0$. Thus

$$\frac{d}{d\varepsilon} J(x^o + \varepsilon h)\bigg|_{\varepsilon=0} = 0.$$

We have

$$\begin{aligned}\frac{d}{d\varepsilon} J(x^o + \varepsilon h)\bigg|_{\varepsilon=0} &= \int_{t_1}^{t_2} \frac{d}{d\varepsilon} f[t, x^o + \varepsilon h, \dot{x}^o + \varepsilon \dot{h}]\bigg|_{\varepsilon=0} dt \\ &= \int_{t_1}^{t_2} f_x(t, x^o, \dot{x}^o)h + f_{\dot{x}}(t, x^o, \dot{x}^o)\dot{h} dt\end{aligned}$$

Therefore we have

$$\int_{t_1}^{t_2} f_x(t, x^o, \dot{x}^o)h + f_{\dot{x}}(t, x^o, \dot{x}^o)\dot{h} dt = 0.$$

Now if we knew $(d/dt)f_{\dot{x}}$ exists, we could integrate by parts here.

Next, define

$$g(t) = \int_{t_1}^{t} f_x(\tau, x^o(\tau), \dot{x}^o(\tau))d\tau.$$

Then integrate by parts:

$$\int_{t_1}^{t_2} f_x(t, x^o, \dot{x}^o)h dt = -\int_{t_1}^{t_2} g\dot{h} dt.$$

Thus

$$\frac{d}{d\varepsilon} J(x^o + \varepsilon h)\bigg|_{\varepsilon=0} = \int_{t_1}^{t_2} [-g + f_{\dot{x}}(t, x^o, \dot{x}^o)]\dot{h} dt.$$

Then the lemma gives that

$$-g + f_{\dot{x}}(t, x^o, \dot{x}^o) = \text{ constant.}$$

Differentiating with respect to t we get the Euler-Lagrange equation. □

Example

$$\dot{x} = Ax + u$$

Given $x(0)$, find the minimum energy u such that $x(1) = 0$. To set this up, we have

$$\int_0^1 \|u(t)\|^2 dt = \int_0^1 \|\dot{x}(t) - Ax(t)\|^2 dt.$$

So we define

$$f(t, x, \dot{x}) = \|\dot{x} - Ax\|^2.$$

Then

$$f_x = -2\dot{x}^T A + 2x^T A^T A$$

$$f_{\dot{x}} = 2\dot{x}^T - 2x^T A^T.$$

The Euler-Lagrange equation is

$$-2\dot{x}^T A + 2x^T A^T A = 2\ddot{x}^T - 2\dot{x}^T A^T.$$

This reduces to

$$\ddot{x} + (A - A^T)\dot{x} - A^T A x = 0.$$

Thus if there's an optimal state x^o, it satisfies

$$\ddot{x}^o + (A - A^T)\dot{x}^o - A^T A x^o = 0, \quad x^o(0) \text{ given}, \quad x^o(1) = 0.$$

This is a two-point boundary-value problem. (See one of the exercises.) The corresponding optimal control is $u^o = \dot{x}^o - Ax^o$. □

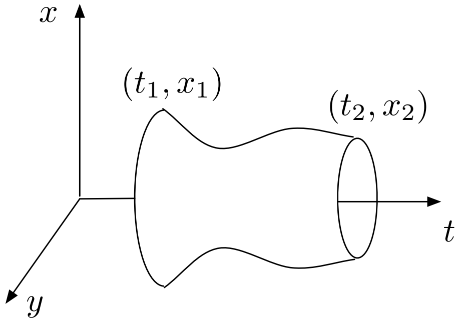
4.4 Problems

1. In the (t, x)-plane, the problem is to find the curve $x(t)$ from (t_1, x_1) to (t_2, x_2) of minimum length (a straight line). Formulate the problem by writing the length of a curve as an integral with respect to t, and solve by the calculus of variations.

2. Consider (x, y, t)-space and consider a curve $x(t)$ in the (t, x)-plane from (t_1, x_1) to (t_2, x_2). Rotate this curve about the t-axis. The surface area is

$$2\pi \int_{t_1}^{t_2} x\sqrt{1 + \dot{x}^2} dt.$$

   Find the curve that minimizes the surface area.



3. Consider the two-point boundary-value problem

$$\ddot{x} = A^T A x, \quad x(0) \text{ given}, \quad x(1) \text{ given}.$$

   Show that it has a unique solution $x(t)$.

4. **Preamble** Consider the problem of Section 3.2, minimizing

$$J(x) = \int_{t_1}^{t_2} f(t, x(t), \dot{x}(t))dt,$$

where $x(t) \in \mathbb{R}$. The values t_1 and t_2 are fixed. Also, $x(t_1)$ is fixed, but (unlike in Section 3.2) $x(t_2)$ is unconstrained. Let $h(t)$ be an arbitrary C^1 function such that $h(t_1) = 0$. It can be derived (don't you do it) that

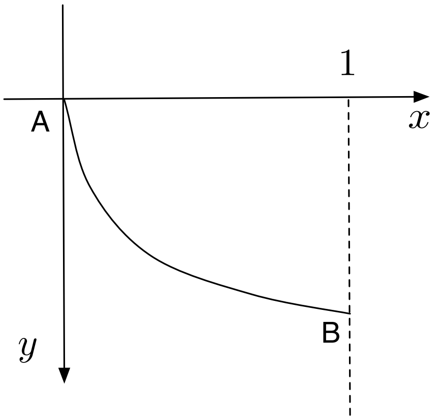
$$\frac{d}{d\varepsilon} J(x + \varepsilon h)\bigg|_{\varepsilon=0} = \int_{t_1}^{t_2} \left(f_x - \frac{d}{dt} f_{\dot{x}} \right) h dt + f_{\dot{x}}(t_2, x(t_2), \dot{x}(t_2))h(t_2).$$

Assume x^o is locally optimal. Then the first term on the right-hand side equals zero from Theorem 3.1. It follows that

$$f_{\dot{x}}(t_2, x^o(t_2), \dot{x}^o(t_2))h(t_2) = 0.$$

This must hold for every $h(t_2)$. It then follows that $f_{\dot{x}}(t_2, x^o(t_2), \dot{x}^o(t_2))$ must equal zero.

The Problem Using the theory just presented, solve the brachistochrone problem (minimum-time path) where A is at the origin but the point B can be any point on the vertical line shown:



5. For the brachistochrone problem, derive the formula $t_1 = \theta_1/\sqrt{g}$. (I think the formula is correct.)

6. Solve the problem of moving a point $p(t)$ in the plane from $p(0) = (1, 1)$ to $p(2) = (0, 0)$ while minimizing the average velocity squared

$$\frac{1}{2} \int_0^2 \|\dot{p}(t)\|^2 dt.$$

Chapter 5

The Maximum Principle

Reference: *Optimal Control*, Athans and Falb, 1966.

This chapter introduces the maximum principle, which was first presented in the famous book

> *The Mathematical Theory of Optimal Processes*, L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze, and E.F. Mischenko, Wiley, New York, 1962

We'll do only a very brief introduction to this approach. You should see Athans and Falb for the wealth of types of problems that can be formulated. Here we do just two special problems to illustrate. Proofs are omitted.

5.1 The Double Integrator

We begin with an example, a kind for which the maximum principle is useful. A cart of mass M moves on wheels in a straight line. The position of the cart is y and a force u is available to control the motion. Thus, if there's no other force, such as friction, then

$$M\ddot{y} = u.$$

We say this system is a double integrator, because y is proportional to the double integral of u. In fact, we might as well take $M = 1$ (by redefining u to be u/M). The problem is to drive the cart from any $(y(0), \dot{y}(0))$ to $(y = 0, \dot{y} = 0)$ in minimum time. This makes sense only if u is bounded, say $|u(t)| \leq 1$.

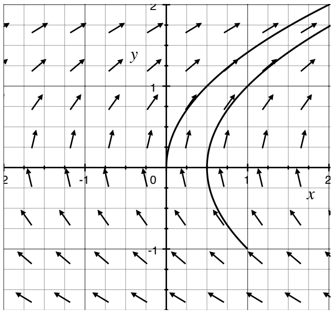
Why can't we use the calculus of variations on this problem?

We shall see that the optimal control signal u^o is always with $+1$ or -1. That is, for every t we have either $u^o(t) = 1$ or $u^o(t) = -1$. Such a control is said to be a **bang-bang** control law—it jumps from one extreme value to the other. Let's use that knowledge and continue.

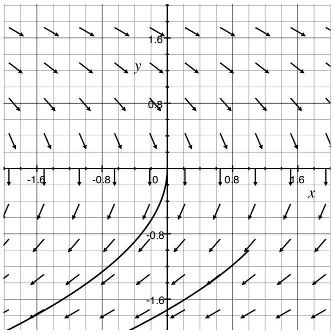
Take the natural state model $x = (y, \dot{y})$:

$$\dot{x} = Ax + Bu, \quad A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

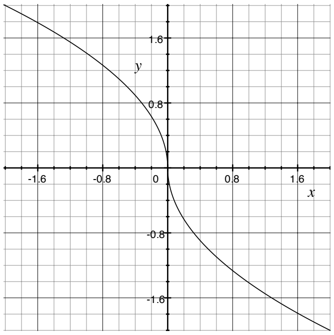
For $u = +1$, $\ddot{y} = 1$ and y is therefore increasing. This graph shows the vector field and two trajectories—one starting at the origin and the other starting at $(1, -1)$:



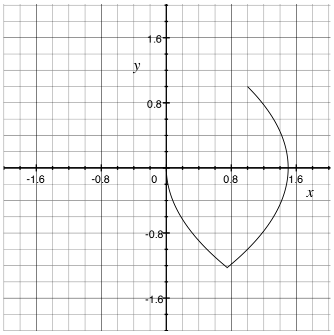
And for $u = -1$ the graph is this:



This leads to the switching curve: The part in the second quadrant is the trajectory backward in time from the origin with $u = -1$; the part in the fourth quadrant is the trajectory backward in time from the origin with $u = 1$:



Every optimal trajectory has at most one switch, to get onto this switching curve. Here's the optimal trajectory starting in the first quadrant:



5.2 Two Special Problems

First Special Problem: Fixed Initial State, Final State, and Final Time

We consider nonlinear state models of the form

$$\dot{x} = f(x, u).$$

It is assumed that f is continuously differentiable in its two arguments.

We assume that t_1 and the boundary states $x(0) = x_0$ and $x(t_1) = v$ are fixed. Furthermore, u is required to be piecewise continuous and to satisfy the constraint $u(t) \in \Omega$, $t \in [0, t_1]$, where the set Ω is fixed. Typically it is a compact set. Finally, the cost to be minimized is

$$J(u) = \int_0^{t_1} L(x(t), u(t))dt,$$

where L is continuously differentiable. Of course J is also a function of x_0, v, t_1, f, but these are not variables in the problem—they are fixed and only u is a variable for design.

Now to the solution in the form of the maximum principle. Define the Hamiltonian, the function

$$H(x, u, \lambda) = L(x, u) + \lambda^T f(x, u), \quad H : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \longrightarrow \mathbb{R}.$$

Theorem 5.1 *Assume a local optimum u^o exists and let x^o be the corresponding state. Then the following conditions hold:*

1. *There exists λ^o such that x^o, λ^o satisfy*

$$\begin{aligned} \dot{x}^o &= H_\lambda(x^o, u^o, \lambda^o)^T = f(x^o, u^o), \quad x^o(0) = x_0, \ x^o(t_1) = v \\ \dot{\lambda}^o &= -H_x(x^o, u^o, \lambda^o)^T. \end{aligned}$$

2. *For each $t \in [0, t_1]$*

$$H[x^o(t), u^o(t), \lambda^o(t)] = \min_{u \in \Omega} H[x^o(t), u, \lambda^o(t)].$$

3. *For each $t \in [0, t_1]$*

$$H[x^o(t), u^o(t), \lambda^o(t)] = 0.$$

The form of the theorem is actually a minimum principle. This is because the problem studied is one of minimization. The name "Maximum Principle" is used nevertheless. Notice that the theorem gives a necessary condition for a local optimum.

Example Let the state space be $\mathbb{R}^2$ and the state equation

$$\dot{x} = u.$$

This represents a point, in the plane, under velocity control. We want to steer the state from $x(0) = 0$ to $x(1) = v$, a given target vector, while minimizing the control energy

$$J(u) = \int_0^1 \|u(t)\|^2 dt,$$

and subject to the constraint $\|u(t)\| \leq 1$. We expect the optimal state trajectory to be a straight line.

The Hamiltonian is

$$H(x, \lambda, u) = \|u\|^2 + \lambda^T u.$$

The state and co-state equations are

$$\begin{aligned} \dot{x}^o &= u^o, \quad x^o(0) = 0,\ x^o(1) = v \\ \dot{\lambda}^o &= 0. \end{aligned}$$

Thus λ^o is a constant. The third condition in the theorem leads to $u^o(t) + \lambda^o(t) = 0$, and hence the optimal u is a constant vector too: $u^o(t) = c_1$. Thus the problem is solvable only if $\|c_1\| \leq 1$. Assuming this inequality, we have from

$$\dot{x}^o = c_1, \quad x^o(0) = 0,\ x^o(1) = v$$

that $x^o(t) = vt$ and $u^o(t) = v$.

To recap, the problem has a solution only if $\|v\| \leq 1$. If an optimal control exists, it equals $u^o(t) = v$, and thus is unique. Finally, it is easily shown that this controller is optimal. □

Second Special Problem: Linear System, Minimum Time

The plant model is linear time-invariant:

$$\dot{x} = Ax + Bu.$$

Controllability of (A, B) is assumed. The goal is to drive the state from $x(0) = x_0$ to $x(t_1) = 0$ in minimum time t_1. The control signal u is required to be piecewise continuous and to satisfy the constraint $u \in \Omega$, defined as the unit cube, i.e.,

$$u \in \Omega \text{ iff } (\forall i)|u_i| \leq 1.$$

For this minimum-time problem the functional to be minimized is

$$\int_0^{t_1} dt = t_1.$$

Thus the Hamiltonian is

$$H(x, u, \lambda) = 1 + \lambda^T(Ax + Bu).$$

Theorem 5.2 *Assume a local optimum u^o and t_1^o exist and let x^o be the corresponding state. Then the following conditions hold:*

1. *There exists λ^o such that x^o, λ^o satisfy*

$$\begin{aligned} \dot{x}^o &= H_\lambda(x^o, u^o, \lambda^o)^T = Ax^o + Bu^o, \quad x^o(0) = x_0,\ x^o(t_1^o) = 0 \\ \dot{\lambda}^o &= -H_x(x^o, u^o, \lambda^o)^T = -A^T\lambda^o. \end{aligned}$$

2. *For each* $t \in [0, t_1^o]$

$$H[x^o(t), u^o(t), \lambda^o(t)] = \min_{u \in \Omega} H[x^o(t), u, \lambda^o(t)].$$

3. *For each* $t \in [0, t_1^o]$

$$H[x^o(t), u^o(t), \lambda^o(t)] = 0.$$

The second condition in the theorem reduces to

$$\lambda^o(t)^T B u^o(t) \leq \lambda^o(t)^T B u, \quad t \in [0, t_1^o], \ u \in \Omega.$$

This implies that $u_i^o(t)$, the ith component of the optimal control, equals -1 if the ith component of $B^T \lambda^o(t)$ is positive and equals $+1$ if the ith component of $B^T \lambda^o(t)$ is negative. Thus the optimal control is confined to the set of vertices of Ω. This is called *bang-bang* control.

Example We return to the double integrator. The co-state equation is

$$\begin{aligned} \dot{\lambda}_1^o &= 0 \\ \dot{\lambda}_2^o &= -\lambda_1^o. \end{aligned}$$

The solution is

$$\lambda_1^o(t) = c_1, \quad \lambda_2^o(t) = -c_1 t + c_2.$$

Since $B^T \lambda = \lambda_2$, we get that $u^o(t)$ equals -1 if $\lambda_2^o(t)$ is positive and equals $+1$ if $\lambda_2^o(t)$ is negative. We conclude that the optimal control signal equals ±1 and switches at most once. That's all the theorem gives us. More details, such as the equation of the switching curve, have to be derived as in the first section of the chapter. □

5.3 Problems

1. Think of the bang-bang optimal control signal for the double integrator. Discuss how it could be implemented. What sensors would be required? Is it a feedback controller? Is this controller robust to sensor noise and modeling errors?

Chapter 6

Dynamic Programming

Dynamic programming (DP) is a clever approach to certain types of optimization problems. It was developed by Richard Bellman and made popular in his book.

6.1 Examples

Example Let $\{x_1, \ldots, x_n\}$ be a finite sequence of real numbers and consider the problem $\min_i x_i$ of finding the minimum. If asked to write a program to solve this, you would undoubtedly write this to compute the minimum, a:

```
$a = x_1$;
for $i = 2 : n$
    $a = \min(a, x_i)$;
end
```

The DP method is exactly the same except in reverse order. Define the **value function**

$$V(i) = \min\ \{x_i, \ldots, x_n\},$$

that is, $V(i)$ is the minimum "cost-to-go" starting at x_i. The value $V(1)$ is what we seek.

Of course, $V(n) = x_n$. Suppose we know $V(i)$ for some i, $1 < i < n$. Then

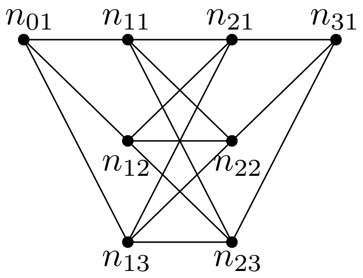
$$\begin{aligned} V(i-1) &= \min\ \{x_{i-1}, \ldots, x_n\} \\ &= \min\ \{x_{i-1}, V(i)\}. \end{aligned}$$

Thus the DP algorithm is

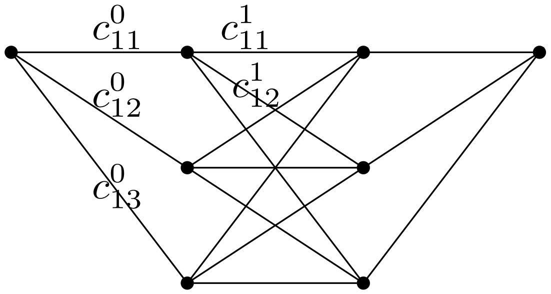
```
$V(n) = x_n$
for $i = n-1, n-2, \ldots, 1$: $V(i-1) = \min\ \{x_{i-1}, V(i)\}$
```

Thus the minimization problem is a recursion of small minimization problems over just pairs of numbers. There are $n-1$ compare operations, and so the complexity is linear. □

Example Another application of DP is to find a minimum-cost path through a graph. Consider this graph:



The nodes are labeled n_{ij}, where i is interpreted as the stage and j as the node number at that stage. Thus there's one node at stage 0, three nodes at stage 1, etc. One wants to travel from the start node n_{01} to the end node n_{31} with minimum cost. Each link has a cost, labeled like this (not all are shown):

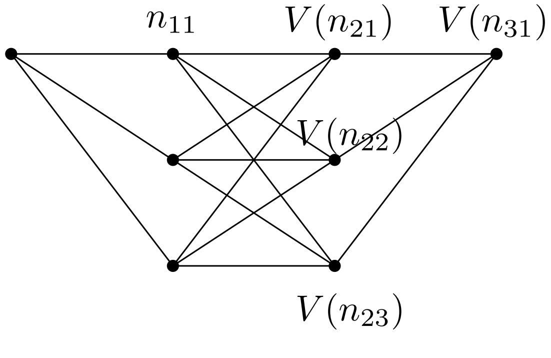


Thus, c^k_{ij} is the cost from node i at stage k to node j at stage $k+1$. The cost of a path is defined to be the sum of the costs of the links.

We define the value function, a real-valued function of the nodes, as follows: $V(n_{ij})$ is the minimum cost to go from node n_{ij} to the end node. The value function at stage 3 is obviously 0. Thus $V(n_{31}) = 0$. The value function at stage 2 is obviously just the cost of the last link:

$$V(n_{21}) = c^2_{11}, \quad V(n_{22}) = c^2_{21}, \quad V(n_{23}) = c^2_{31}.$$

We label these at the nodes:



Now to the value function at stage 1. We will invoke the so-called **principle of optimality**: Consider an optimal path from n_{01} to n_{31}; if this path goes through node n_{1j} at stage 1, then the subpath from node n_{1j} to n_{31} is optimal too. That is, for every optimal path, the cost-to-go is minimum at each point along the path. Note that we're not saying the initial subpath is optimal, but rather the cost-to-go is. Thus at node n_{11}, since there are just three links out, we have

$$V(n_{11}) = \min \{c^1_{11} + V(n_{21}), c^1_{12} + V(n_{22}), c^1_{13} + V(n_{23})\}.$$

After the other values are computed at stage 1, one computes $V(n_{01})$, which equals the minimum cost path from start to end. After the value function is computed at every node, it's easy to find optimal paths by moving left to right. □

6.2 The Hamilton-Jacobi-Bellman Equation

We now illustrate the DP approach by looking at the linear-quadratic regulator (LQR) problem.

Consider the plant equation

$$\dot{x} = Ax + Bu, \quad x(0) = x_0$$

and the cost functional

$$J(x_0, u) = \int_0^\infty x(t)^T Qx(t) + u(t)^T Ru(t)dt.$$

The arguments of J are the initial state x_0 and the input signal u. Implicitly, u is such that J is finite. This problem is a special case of

$$\dot{x} = f(x, u), \quad x(0) = x_0$$

$$J(x_0, u) = \int_0^\infty L[x(t), u(t)]dt.$$

So let's do this more general case and then specialize.

Introduce the **value function**, the optimal cost as a function of arbitrary initial time and state:

$$V(\tau, \xi) = \min_u \left\{ \int_\tau^\infty L(x, u)dt : x(\tau) = \xi \right\}.$$

The argument t in the integrand has been dropped for convenience. Because A, B, Q, R are constant matrices and the upper limit on the integral is ∞, you can check that $V(\tau, \xi)$ is independent of τ, that is, V is a function of only ξ in this instance. Nevertheless, we'll keep the two arguments in order to get the general HJB equation.

For any $\delta\tau > 0$ we have

$$\int_\tau^\infty L(x, u)dt = \int_\tau^{\tau+\delta\tau} L(x, u)dt + \int_{\tau+\delta\tau}^\infty L(x, u)dt.$$

Let u_τ denote the piece of u defined over $(\tau, \tau + \delta\tau)$ and $\overline{u}_\tau$ the piece of u defined over $(\tau + \delta\tau, \infty)$. Then the term $\int_\tau^{\tau+\delta\tau} L(x, u)dt$ is a function of $x(\tau)$ and u_τ, while $\int_{\tau+\delta\tau}^\infty L(x, u)dt$ is a function of

$x(\tau+\delta\tau)$ and $\overline{u}_\tau$; but $x(\tau+\delta\tau)$ is a function of $x(\tau)=\xi$ and u_τ. Minimizing over u we get

$$\begin{aligned}
V(\tau,\xi) &= \min_u \int_\tau^\infty L(x,u)dt \\
&= \min_{u_\tau}\min_{\overline{u}_\tau}\left[\int_\tau^{\tau+\delta\tau} L(x,u)dt + \int_{\tau+\delta\tau}^\infty L(x,u)dt\right] \\
&= \min_{u_\tau}\left[\int_\tau^{\tau+\delta\tau} L(x,u)dt + \min_{\overline{u}_\tau}\int_{\tau+\delta\tau}^\infty L(x,u)dt\right] \\
&= \min_{u_\tau}\left[\int_\tau^{\tau+\delta\tau} L(x,u)dt + V(\tau+\delta\tau, x(\tau+\delta\tau))\right].
\end{aligned}$$

Now we let $\delta\tau$ approach 0. To first order in $\delta\tau$ we have

$$x(\tau+\delta\tau) = \xi + \delta\tau f(\xi, u(\tau))$$

and therefore

$$V(\tau+\delta\tau, x(\tau+\delta\tau)) = V(\tau,\xi) + \delta\tau\frac{\partial V}{\partial\tau}(\tau,\xi) + \delta\tau\frac{\partial V}{\partial x}(\tau,\xi)f(\xi,u(\tau)).$$

Also,

$$\int_\tau^{\tau+\delta\tau} L(x,u)dt = \delta\tau L[\xi,u(\tau)].$$

Thus we have

$$V(\tau,\xi) = \min_{u(\tau)}\left\{\delta\tau L[\xi,u(\tau)] + V(\tau,\xi) + \delta\tau\frac{\partial V}{\partial\tau}(\tau,\xi) + \delta\tau\frac{\partial V}{\partial x}(\tau,\xi)f(\xi,u(\tau))\right\}$$

and hence

$$0 = \min_{u(\tau)}\left\{L[\xi,u(\tau)] + \frac{\partial V}{\partial\tau}(\tau,\xi) + \frac{\partial V}{\partial x}(\tau,\xi)f(\xi,u(\tau))\right\}.$$

In this equation, ξ and $u(\tau)$ are dummy variables. Let's replace them by $x\in\mathbb{R}^n$ and $u\in\mathbb{R}^m$:

$$\min_u\left\{L(x,u) + \frac{\partial V}{\partial\tau}(\tau,x) + \frac{\partial V}{\partial x}(\tau,x)f(x,u)\right\} = 0.$$

We arrive at the Hamilton-Jacobi-Bellman (HJB) equation:

$$\frac{\partial V}{\partial\tau}(\tau,x) + \min_u\left\{L(x,u) + \frac{\partial V}{\partial x}(\tau,x)f(x,u)\right\} = 0.$$

As mentioned at the start of the derivation, $V(\tau,x)$ is a function only of x: $V(x)$. So in this time-invariant case the HJB equation is

$$\min_u\left\{L(x,u) + \frac{dV}{dx}(x)f(x,u)\right\} = 0.$$

The derivation of this equation wasn't rigorous. What we have is an equation satisfied by an optimal control law, u as a function of x, and the value function, $V(x)$, under certain conditions, if an optimal control exists. What to do with the equation? Solve the minimization problem on the left-hand side for u as a function of x and dV/dx; then equate the left-hand side to zero.

Example A very simple example is

$$\dot{x} = x + u, \quad J(x_0) = \int_0^\infty x^2 + u^2 dt.$$

That is,

$$f(x, u) = x + u, \quad L(x, u) = x^2 + u^2.$$

Letting V_x denote dV/dx, we have the HJB equation

$$\min_u \left\{x^2 + u^2 + V_x(x + u)\right\} = 0.$$

To do the minimization, differentiate with respect to u and set the derivative to zero:

$$2u + V_x = 0.$$

Thus $u = -V_x/2$. Substitute this into

$$x^2 + u^2 + V_x(x + u) = 0$$

and solve for V_x:

$$V_x = 2(1 \pm \sqrt{2})x.$$

Thus

$$u = -\frac{1}{2}V_x = -(1 \pm \sqrt{2})x.$$

Only the solution

$$u = -(1 + \sqrt{2})x$$

is valid; the other doesn't yield $J(x_0) < \infty$. □

Now we return to the full LQR problem. We have

$$f(x, u) = Ax + Bu, \quad L(x, u) = x^T Qx + u^T Ru,$$

where Q is positive semidefinite and R positive definite. The HJB equation is

$$\min_u \left\{x^T Qx + u^T Ru + V_x(Ax + Bu)\right\} = 0.$$

The minimizing u satisfies

$$2u^T R + V_x B = 0,$$

and so

$$u = -\frac{1}{2}R^{-1}B^T V_x^T.$$

Substituting into

$$x^T Qx + u^T Ru + V_x(Ax + Bu) = 0$$

gives

$$x^T Qx + V_x Ax - \frac{1}{4}V_x BR^{-1}B^T V_x^T = 0.$$

Now we somehow have to find a $V(x)$ satisfying this equation and such that u makes J finite. It turns out that a quadratic function will work: $V(x) = x^T Px$. Substituting $V_x = 2x^T P$ into the equation gives

$$x^T Qx + 2x^T PAx - x^T PV_x BR^{-1}B^T Px = 0,$$

or equivalently, to get a symmetric matrix,

$$x^T Qx + x^T PAx + x^T A^T Px - x^T PV_x BR^{-1}B^T Px = 0.$$

This can be written as

$$x^T(Q + PA + A^T P - PBR^{-1}B^T P)x = 0$$

and this leads to the Riccati equation:

$$Q + PA + A^T P - PBR^{-1}B^T P = 0.$$

It remains to study when this equation has a solution P such that J is finite for

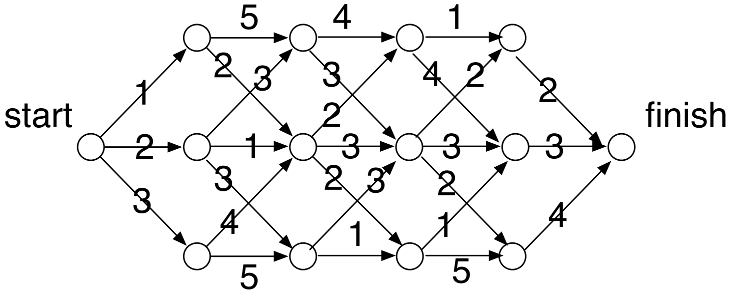
$$u = -R^{-1}B^T Px.$$

We'll do this in a later chapter.

Let's summarize what we've done. We assumed an optimal control law exists and we derived a formula for it, but we don't know when it is valid, that is, we don't know if the Riccati equation has a solution, and, if it does, we don't know if J is finite. This is typical of DP. It provides an existence condition but you still have to do a lot of work.

6.3 Problems

1. Find the minimum cost path.



2. Discrete-time LQR:

$$x(k+1) = Ax(k) + Bu(k), \quad x(0) = x_0$$

$$\sum_{k \geq 0} x(k)^T Q x(k)^T + u(k)^T R u(k)$$

Find the HJB equation.

Part II

More Recent Theories

Preamble to Part II

In the next three chapters we formulate optimal control problems in terms of signal norms. This leads to function spaces and operators on them. This is a branch of mathematics called functional analysis. Here we introduce and motivate the main ideas.

The performance of a system should be measured by norms, for example, how large a tracking error is, or how large a control signal is. So let us begin with some familiar terms for deterministic signals. For a sinusoidal signal $x(t) = A\cos(\omega t + \phi)$, the zero-to-peak value is $|A|$. We write this as $\|x\|_\infty$ and call it the infinity norm: $\|x\|_\infty = \max_t |x(t)|$. In an electric circuit, the power dissipated in a resistor at time t is $i(t)^2 R$ and the energy dissipated is the integral of this over time. Extending this, we shall think of the energy of a signal $x(t)$ as $\int x(t)^2 dt$ and we shall write this as $\|x\|_2^2$, the square of the 2-norm. Thus for a signal $x(t)$, there are two norms: $\|x\|_\infty$, the zero-to-peak value, and $\|x\|_2$, the square-root of the energy. Thus we have two norms to measure signal size for deterministic signals.

Then there are random signals. Let x be a zero-mean random variable. Its root-mean-square (rms) value qualifies as a norm: $\|x\| = (\mathrm{E}\ x^2)^{1/2}$. This extends to a random vector: The norm is $\|x\| = (\mathrm{E}\ x^T x)^{1/2}$, which can also be written $\|x\| = \mathrm{Tr}\ (\mathrm{E}\ xx^T)^{1/2}$, where Tr denotes trace. This extends to zero-mean stationary random signals.

Now we turn to norms of systems. To get a glimpse of this concept, consider the equation $y = Au$, where u, y are vectors and A is a matrix. We shall think of this equation as defining a system—input u, output y. We are going to define two norms for this system: The first is for a specific input, and the second is what is called an induced norm.

For the first system norm, suppose u is a zero-mean white random vector, that is, its covariance matrix equals the identity matrix: $\mathrm{E}\ uu^T = I$. The term "white" refers to the fact that two different components of u are uncorrelated. The covariance matrix of y equals AA^T and therefore the norm of y equals $\mathrm{Tr}\ AA^T$, or equivalently, $\mathrm{Tr}\ A^T A$. This motivates introducing the Frobenius norm of a matrix:

$$\|A\|_F = \left(\mathrm{Tr}\ A^T A\right)^{1/2}.$$

You can check that this equals $\left(\Sigma_{i,j} a_{ij}^2\right)^{1/2}$. Thus, the Frobenius norm of A equals the rms output when the input is the standard white vector.

The second system norm is defined by saying that its square equals the maximum output energy when the input energy equals 1:

$$\sup\{\|Ax\|_2 : \|u\|_2 = 1\}.$$

It is a fact that this induced norm equals $\sigma_{\max}(A)$, the largest singular value of A.

Summary of Spaces

Time domain, discrete time, scalar valued

The domain of all these functions is $\mathbb{Z}$, the integers. The time set could also be the non-negative integers. We could write $\ell(\mathbb{Z})$ etc., but we shall let context determine this.

ℓ The space of all signals $\mathbb{Z} \longrightarrow \mathbb{R}$.
This is a vector space. The time set could also be the non-negative integers.

ℓ^2 The subspace of ℓ of square summable, i.e., finite energy, signals.
A Hilbert space under the inner product $\langle x, y \rangle = \sum_n x[n]y[n]$.
The norm is $\|x\|_2 = \left(\sum_n x[n]^2\right)^{1/2}$.

ℓ^∞ The subspace of ℓ of bounded signals.
A Banach space under the norm $\|x\|_\infty = \sup_n |x[n]|$.

ℓ^1 The subspace of ℓ of absolutely summable signals.
A Banach space under the norm $\|x\|_1 = \sum_n |x[n]|$.

c_{fd} The subspace of ℓ or ℓ^2 or ℓ^∞ of signals that are of finite duration.

Time domain, continuous time, scalar valued

The domain of all these functions is $\mathbb{R}$, the set of real time values. The time set could also be the non-negative reals.

$\mathcal{L}$ The space of all signals $\mathbb{R} \longrightarrow \mathbb{R}$.

$\mathcal{L}^2$ The subspace of $\mathcal{L}$ of square integrable, i.e., finite energy, signals.
A Hilbert space under the inner product $\langle x, y \rangle = \int_t x(t)y(t)dt$.
The norm is $\|x\|_2 = \left(\int_t x(t)^2 dt\right)^{1/2}$.

$\mathcal{L}^\infty$ The subspace of $\mathcal{L}$ of essentially bounded signals,
i.e., bounded except perhaps on a set of measure zero.
A Banach space under the norm $\|x\|_\infty = \text{ess sup } |x(t)|$.

$\mathcal{L}^1$ The subspace of $\mathcal{L}$ of absolutely summable signals.
A Banach space under the norm $\|x\|_1 = \int_t |x(t)|dt$.

Frequency domain, discrete time, scalar valued

These functions are complex-valued and their domain is the unit circle, and sometimes beyond.

$\mathcal{L}^2$ The space of square integrable complex-valued functions $X(\mathrm{e}^{j\omega})$.
A Hilbert space under $\langle X, Y\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\mathrm{e}^{j\omega})\overline{Y(\mathrm{e}^{j\omega})}d\omega$.
The norm is $\|X\|_2 = \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \left|X(\mathrm{e}^{j\omega})\right|^2 d\omega\right)^{1/2}$.
$\mathcal{L}^2$ is the space of DTFTs (discrete-time Fourier transforms) of ℓ^2.

$\mathcal{H}^2$ The space of z-transforms of those functions in ℓ^2 that are 0 for $n < 0$.
This is a closed subspace of $\mathcal{L}^2$, so it has the same inner product.
The functions in $\mathcal{H}^2$ are analytic in the unit disk, $|z| < 1$.

$\mathcal{L}^\infty$ The space of essentially bounded complex-valued functions $X(\mathrm{e}^{j\omega})$.
A Banach space under $\|X\|_\infty = \text{ess sup } |X(\mathrm{e}^{j\omega})|$.

$\mathcal{H}^\infty$ The space of functions that are analytic and bounded in the unit disk.
This is a closed subspace of $\mathcal{L}^\infty$, so it has the same norm.

Frequency domain, continuous time, scalar valued

These functions are complex-valued and their domain is the imaginary axis.

$\mathcal{L}^2$ The space of square integrable complex-valued functions $X(j\omega)$.
A Hilbert space under $\langle X, Y\rangle = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(j\omega)\overline{Y(j\omega)}d\omega$.
The norm is $\|X\|_2 = \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} |X(j\omega)|^2 d\omega\right)^{1/2}$.
$\mathcal{L}^2$ is the space of FTs (Fourier transforms) of $\mathcal{L}^2$.

$\mathcal{H}^2$ The space of Laplace transforms of those functions in $\mathcal{L}^2$ that are 0 for $t < 0$.
This is a closed subspace of $\mathcal{L}^2$, so it has the same inner product.
The functions in $\mathcal{H}^2$ are analytic in the right half-plane, Re $s > 0$.

$\mathcal{L}^\infty$ The space of essentially bounded complex-valued functions $X(j\omega)$.
A Banach space under $\|X\|_\infty = \text{ess sup } |X(j\omega)|$.

$\mathcal{H}^\infty$ The space of functions that are analytic and bounded in the right half-plane.
This is a closed subspace of $\mathcal{L}^\infty$, so it has the same norm.

Chapter 7

Introduction to Function Spaces

7.1 Hilbert Space and Banach Space

Norms

Everyone knows what a vector space is. Unless we say otherwise, the scalars associated with the vector space will be real numbers. Our vector spaces will frequently not be finite dimensional. An example is the space $\mathcal{C}[t_1, t_2]$ of real-valued continuous functions $x(t)$ defined on the time interval $[t_1, t_2]$. It's possible to define a norm on this space, for example,

$$\|x\|_\infty = \max_{t_1 \leq t \leq t_2} |x(t)|.$$

But this isn't the only possibility; another is

$$\|x\|_2 = \left(\int_{t_1}^{t_2} |x(t)|^2 dt \right)^{1/2}.$$

There are three properties a norm must have:

$$\|x\| \geq 0 \text{ and } \|x\| = 0 \Longleftrightarrow x = 0$$

$$\|cx\| = |c| \|x\|$$

$$\|x + y\| \leq \|x\| + \|y\|.$$

A vector space with a norm is a **normed space**.

Inner products

The space $\mathbb{R}^n$ has the inner product $\langle x, y \rangle$, also written as a dot product. If x, y are regarded as column vectors, then the inner product can also be written $x^T y$. Finally, the Euclidean norm can be defined in terms of the inner product:

$$\|x\| = \langle x, x \rangle^{1/2}.$$

In a general vector space $\mathcal{X}$, an inner product $\langle x, y \rangle$ must have three properties: $\langle x, y \rangle = \langle y, x \rangle$; for every y the map $x \mapsto \langle x, y \rangle$ is linear; and $\langle x, x \rangle$ is positive for all nonzero x. A vector space with an inner product is an **inner product space**. The space $\mathcal{C}[t_1, t_2]$ with the norm $\|x\|_2$ is an inner product space, the inner product being

$$\langle x, y \rangle = \int_{t_1}^{t_2} x(t)y(t)dt.$$

Completeness

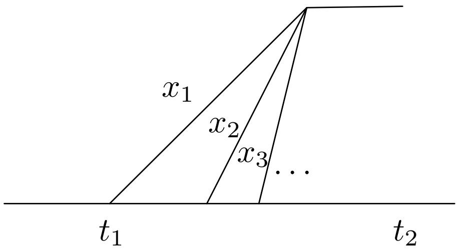
Consider the set $\mathbb{R}$ and a sequence $\{a_n\}$ in it. It is called a **Cauchy sequence** if

$$(\forall \varepsilon)(\exists N)(\forall i, k) i, k > N \Longrightarrow |a_i - a_k| < \varepsilon.$$

Evidently this sequence is "trying to converge" in the sense that the elements in the sequence are getting closer and closer together. That it does converge in $\mathbb{R}$ is a feature of that set, a feature called **completeness**: Every Cauchy sequence in $\mathbb{R}$ has a limit in $\mathbb{R}$. For example, the interval $(0, 1]$ is not complete, because $1/n$ is a Cauchy sequence that converges to $0 \notin (0, 1]$.

In a normed space, every convergent sequence is a Cauchy sequence (a good exercise). If, conversely, every Cauchy sequence converges in the space, the space is said to be **complete**. A complete normed space is called a **Banach space**; a complete inner-product space is called a **Hilbert space**. The advantage of completeness is that in principle one doesn't have to know a limit to test if a sequence converges.

The space $\mathcal{C}[t_1, t_2]$ with the norm $\|x\|_\infty$ is a Banach space, while $\mathcal{C}[t_1, t_2]$ with the norm $\|x\|_2$ is not. To see this latter fact, note that the sequence



is a Cauchy sequence in the norm $\|x\|_2$, but it converges to a step function, which is not continuous and therefore not in $\mathcal{C}[t_1, t_2]$. It is not a Cauchy sequence in the norm $\|x\|_\infty$.

Consider again $\mathcal{C}[t_1, t_2]$ with the norm $\|x\|_\infty$. The set of polynomial functions of t is a subset, $\mathcal{P}[t_1, t_2]$, of $\mathcal{C}[t_1, t_2]$. It is a **subspace**, that is, it is a vector space itself, but it is not complete. For example, the function $x(t) = \sin t$ belongs to $\mathcal{C}[t_1, t_2]$; if $x_n(t)$ denotes the truncation at the nth term of the Taylor series of $x(t)$, then $x_n \in \mathcal{P}[t_1, t_2]$; also, $x_n(t)$ converges to x in the sense that

$$\lim_{n \to \infty} \|x - x_n\|_\infty = 0.$$

The **closure** of $\mathcal{P}[t_1, t_2]$, denoted $\overline{\mathcal{P}[t_1, t_2]}$, is defined to be the limits of all sequences in $\mathcal{P}[t_1, t_2]$ that converge in $\mathcal{C}[t_1, t_2]$. The Weierstrass approximation theorem says that

$$\overline{\mathcal{P}[t_1, t_2]} = \mathcal{C}[t_1, t_2],$$

that is, any continuous function on a closed interval can be approximated uniformly by a polynomial. Now let's enlarge $\mathcal{C}[t_1, t_2]$. There are certainly discontinuous functions that are bounded and therefore for which the norm $\|x\|_\infty$ is finite. We are going to call this class of functions $\mathcal{L}^\infty[t_1, t_2]$. [1] With appropriate consideration, $\mathcal{L}^\infty[t_1, t_2]$ is a Banach space.

On the other hand, consider $\mathcal{C}[t_1, t_2]$ with the norm $\|x\|_2$. As we saw, it's not complete. But it can be embedded in a complete normed space, which is denoted $\mathcal{L}^2[t_1, t_2]$. The construction of this completion is somewhat involved, hence we omit it. For us, it's good enough to accept that $\mathcal{L}^2[t_1, t_2]$ contains functions x for which there is a sequence $\{x_n\}$ of continuous functions, or even polynomials, such that

$$\lim_{n\to\infty} \|x - x_n\|_2 = 0.$$

Thus $\mathcal{L}^2[t_1, t_2]$ is a Hilbert space.

More generally, if the time set is all real t, we write $\mathcal{L}^2(\mathbb{R})$. Indeed, we can view $\mathcal{L}^2[t_1, t_2]$ as the subspace of $\mathcal{L}^2(\mathbb{R})$. of functions zero for t not in $[t_1, t_2]$. Likewise for $\mathcal{L}^2[0, \infty)$. If the time set is known or irrelevant, we may write just $\mathcal{L}^2$.

A different example, from the world of discrete-time signals, is ℓ^2. The elements of this space are square-summable discrete-time signals, $x(k), k \geq 0$, $x(k) \in \mathbb{R}$. The inner product is

$$\langle x, y\rangle = \sum_k x(k)y(k).$$

Consider the subset c_{dn} of signals of duration n, that is, $x(k)$ equals zero for $k > n$. It's routine to prove that c_{dn} is a subspace of the vector space ℓ^2, and that furthermore it is a closed set within ℓ^2. Being closed, c_{dn} is complete, and therefore is a Hilbert space. Now consider the subset c_{fd} of signals of finite duration, that is, $x(k)$ equals zero within a finite time:

$$c_{fd} = \{x : (\exists n)(\forall k > n) x(k) = 0\}.$$

Again, c_{fd} is a subspace of the vector space ℓ^2. However c_{fd} is not a closed set within ℓ^2. To see this, note that the sequence x_i,

$$x_i(k) = \begin{cases} 1/2^k, & k \leq i \\ 0, & k > i \end{cases}$$

belongs to c_{fd}, and it converges in ℓ^2 to the signal

$$x(k) = \frac{1}{2^k},$$

but x doesn't belong to c_{fd}. In fact, the closure of c_{fd} in ℓ^2 is the space ℓ^2 itself.

Vector-valued functions

Now we turn to vector-valued signals. Let $x(t)$ denote a function where t is a real number and $x(t)$ is an n-dimensional real vector. The $\mathcal{L}^2$-norm of x is defined to be

$$\|x\|_2 = \left(\int_{-\infty}^{\infty} \|x(t)\|^2 dt\right)^{1/2}.$$

[1] Actually, we're glossing over a subtle point. Let $x(t)$ be the function defined on the interval $[0, 1]$ by saying that $x(t) = n$ for $t = 1/n$, $n \geq 1$, and $x(t) = 0$ otherwise. Sketch the graph of this function. It is unbounded but is zero except at a countable number of points. We say that x equals zero almost everywhere and we set $\|x\|_\infty = 0$.

The norm $\|x(t)\|$ is the Euclidean norm of the vector $x(t)$. Usually it is irrelevant what the dimension is so we continue to write the space as $\mathcal{L}^2(\mathbb{R})$.

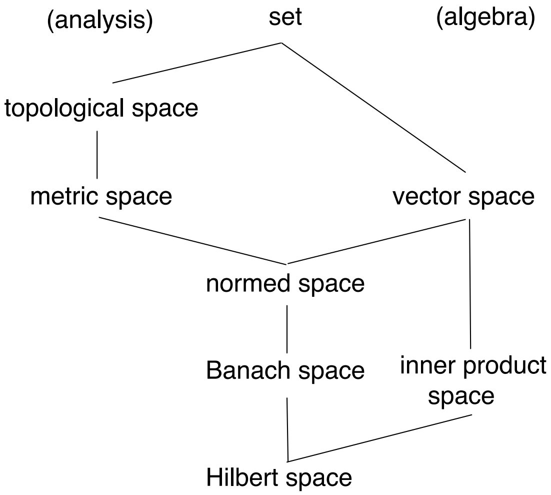
For a bounded signal the norm is

$$\|x\|_\infty = \sup_t \|x(t)\|.$$

Again, the right-hand norm is the Euclidean norm. We write $\mathcal{L}^2(\mathbb{R})$ for the class of such functions.

We have just seen the definition of Hilbert space: a complete inner product space. Thus Hilbert space is an abstract concept for which there are many instances, $\mathcal{L}^2[t_1, t_2]$ being one. This idea of defining an abstract concept is very common in mathematics because one can get a general result that applies in many instances.

In a classification of spaces, Hilbert and Banach spaces sit here:



The notion of orthogonality allows optimization to be done by orthogonal projections. A famous example of this is the Kalman filter.

7.2 The Projection Theorem

Let's begin with some properties of the inner product. The first is the Cauchy-Schwarz inequality.

Lemma 7.1 *In an inner-product space*

$$|\langle x, y\rangle| \leq \|x\|\|y\|.$$

Proof If $y = 0$ then both sides of the inequality equal 0. So now assume $y \neq 0$. Define

$$c = \frac{\langle x, y\rangle}{\|y\|^2}.$$

Then

$$\begin{aligned} 0 &\leq \|x - cy\|^2 \\ &= \langle x - cy, x - cy\rangle \\ &= \|x\|^2 - c\langle y, x\rangle - c\langle x, y\rangle + c^2\|y\|^2. \end{aligned}$$

The second and fourth terms cancel. Thus

$$c\langle x, y\rangle \leq \|x\|^2,$$

and so

$$\langle x, y\rangle^2 \leq \|x\|^2\|y\|^2.$$

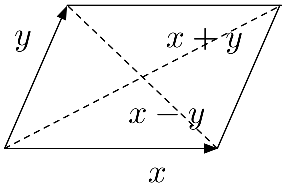
□

The second is the parallelogram equality.

Lemma 7.2 *In an inner-product space*

$$\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2).$$

The proof is left as an exercise. The picture going with the lemma is this:



In an inner product space, we say x, y are **orthogonal** and write $x \perp y$ if $\langle x, y\rangle = 0$. For example, $\mathrm{e}^{jnt} \perp \mathrm{e}^{jmt}$ in $\mathcal{L}^2[0, 2\pi]$. Here, the field of scalars is $\mathbb{C}$ and the inner product is

$$\langle x, y\rangle = \int_0^{2\pi} x(t)\overline{y(t)}dt.$$

A set $\mathcal{V}$ in a vector space is **convex** if, whenever $x, y \in \mathcal{V}$, all points on the line from x to y are in $\mathcal{V}$, i.e.,

$$(\forall x, y)(\forall \lambda)x, y \in \mathcal{V}, 0 \leq \lambda \leq 1 \Longrightarrow \lambda x + (1-\lambda)y \in \mathcal{V}.$$

Theorem 7.1 *Let $\mathcal{X}$ be an inner product space, $\mathcal{V}$ a complete convex subset, and $x \in \mathcal{X}$. There is a unique vector in $\mathcal{V}$ that is closest to x.*

Proof Define

$$\delta = \inf_{v \in \mathcal{V}} \|x - v\|.$$

Then there is a sequence $\{v_n\}$ in $\mathcal{V}$ such that $\|x - v_n\|$ converges to δ. If we can show that $\{v_n\}$ is a Cauchy sequence, then, because $\mathcal{V}$ is complete, $v^o = \lim_n v_n$ exists, belongs to $\mathcal{V}$, and is closest to x.

To prove the sequence is a Cauchy sequence, by the parallelogram equality we have

$$\|(v_n - x) + (x - v_m)\|^2 + \|(v_n - x) - (x - v_m)\|^2 = 2(\|v_n - x\|^2 + \|x - v_m\|^2).$$

Therefore

$$\|v_n - v_m\|^2 = 2\|v_n - x\|^2 + 2\|x - v_m\|^2 - 4\|(v_n + v_m)/2 - x\|^2.$$

By convexity $(v_n + v_m)/2 \in \mathcal{V}$ and so

$$\|(v_n + v_m)/2 - x\| \geq \delta.$$

Therefore

$$\|v_n - v_m\|^2 \leq 2\|v_n - x\|^2 + 2\|x - v_m\|^2 - 4\delta^2.$$

The right-hand side is arbitrarily small for n, m sufficiently large. This proves the Cauchy property.

Finally, uniqueness is proved like this. Suppose

$$\|x - v^o\| = \|x - v\| = \delta.$$

As in the preceding inequality,

$$\|v^o - v\|^2 \leq 2\|v^o - x\|^2 + 2\|x - v\|^2 - 4\delta^2 = 0.$$

Thus $v^o = v$. □

The preceding result is not true in general in a normed space, as shown in an exercise.

A subspace $\mathcal{V}$ in a Hilbert space $\mathcal{X}$ may not be closed, as we saw. Its **orthogonal complement**, denoted $\mathcal{V}^\perp$, is the set of vectors orthogonal to every vector in $\mathcal{V}$. The set $\mathcal{V}^\perp$ is a subspace and it is closed. If $\mathcal{V}$ is closed, then

$$\mathcal{X} = \mathcal{V} \oplus \mathcal{V}^\perp,$$

which means that every vector x can be written uniquely as $v + w$, $v \in \mathcal{V}, w \in \mathcal{V}^\perp$.

Now for the projection theorem. It is trivial to prove that a subspace of a vector space is convex.

Theorem 7.2 *Let $\mathcal{X}$ be a Hilbert space and $\mathcal{V}$ a closed subspace. Let $x \in \mathcal{X}$ and let v^o be the vector in $\mathcal{V}$ that is closest to x. Then $x - v^o \perp \mathcal{V}$.*

Proof Suppose not. Then there is a vector v in $\mathcal{V}$ of unit norm and such that

$$\langle x - v^o, v\rangle = c \neq 0.$$

Then

$$\begin{aligned}\|x - (v^o + cv)\|^2 &= \|(x - v^o) - cv\|^2 \\ &= \|x - v^o\|^2 - 2c\langle x - v^o, v\rangle + c^2 \\ &= \|x - v^o\|^2 - c^2 \\ &< \|x - v^o\|^2.\end{aligned}$$

This contradicts that v^o is closest to x. □

Example Consider a simple first-order model of a motor:

$$\ddot{\theta} + \dot{\theta} = u.$$

The state is $x = (\theta, \dot{\theta})$. Suppose the control objective is to drive the state from $x(0) = (0,0)$ to $x(1) = (1,0)$ using minimum energy, that is,

$$\|u\|_2 = \int_0^1 u(t)^2 dt$$

should be minimum. Thus the optimization space is $\mathcal{U} = \mathcal{L}^2[0,1]$.

Let us turn the objective $x(1) = (1,0)$ into a constraint on u. We have

$$\theta(t) = \int_0^t u(\tau)d\tau - \int_0^t \mathrm{e}^{-(t-\tau)} u(\tau)d\tau$$

$$\dot{\theta}(t) = \int_0^t \mathrm{e}^{-(t-\tau)} u(\tau)d\tau.$$

Define

$$v_1(t) = 1 - \mathrm{e}^{t-1}, \quad v_2(t) = \mathrm{e}^{t-1}.$$

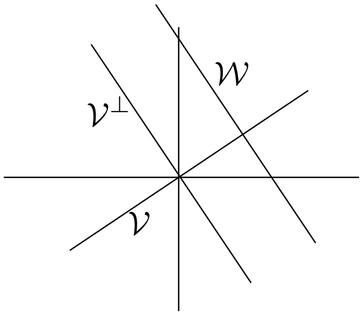
Then the problem is to minimize $\|u\|_2$ subject to

$$\langle v_1, u\rangle = 1, \quad \langle v_2, u\rangle = 0.$$

The vectors v_1, v_2 are linearly independent. Let $\mathcal{V}$ denote their span. Since $\mathcal{V}$ is finite-dimensional, it is closed. Let $\mathcal{W}$ denote the set of control signals driving the state to the desired point:

$$\mathcal{W} = \{u : \langle v_1, u\rangle = 1, \langle v_2, u\rangle = 0\}.$$

If $u \in \mathcal{W}$, then every vector of the form $u + p$, $p \perp \mathcal{V}$, belongs to $\mathcal{W}$. Thus the picture looks like this:



It follows that the optimal u lies at the intersection of $\mathcal{V}$ and $\mathcal{W}$. So write $u = c_1v_1 + c_2v_2$. Substitute this into the constraint equations and solve for c_1, c_2:

$$u^o(t) = \frac{1}{3-\mathrm{e}}(1 + \mathrm{e} - 2\mathrm{e}^t).$$

□

7.3 Operators

Let $\mathcal{X}, \mathcal{Y}$ be normed spaces and let $T : \mathcal{X} \longrightarrow \mathcal{Y}$ be a linear function; function, mapping, transformation are synonymous. We say T is **bounded** if

$$(\exists b)(\forall x)\|Tx\| \leq b\|x\|.$$

The least bound b for which this inequality holds is called the **norm** of T, denoted $\|T\|$. It is a fact that boundedness and continuity are equivalent.

A good example is a BIBO stable system. For example, consider the LTI system with transfer function

$$G(s) = \frac{1}{s+1}$$

and let T denote the time-domain mapping from input to output. If we take $\mathcal{X}, \mathcal{Y}$ both to be the space of bounded continuous functions on the time interval $[0, \infty)$, with norm

$$\|x\|_\infty = \sup_t |x(t)|,$$

then T is bounded. In fact, the norm of T equals

$$\int_0^\infty |g(t)|dt,$$

where $g(t)$ is the inverse Laplace transform of $G(s)$.

A bounded linear map is called an **operator**. The set of operators $T : \mathcal{X} \longrightarrow \mathcal{Y}$ is denoted $B(\mathcal{X}, \mathcal{Y})$. It too is a normed space, and if $\mathcal{X}, \mathcal{Y}$ are Hilbert spaces, $B(\mathcal{X}, \mathcal{Y})$ is a Banach space.

Example Consider the state-space system

$$\dot{x} = Ax + Bu, \quad x(0) = 0.$$

Suppose $\dim x = n, \dim u = m$. Fix a time, say $t = 1$, and consider the map T from u to $x(1)$. Let us take the domain of T, denoted $\mathcal{U}$, to be the Hilbert space $\mathcal{L}^2[0,1]$, that is, m-dimensional vectors $u(t)$ each of whose components lives in $\mathcal{L}^2[0,1]$. The inner product on $\mathcal{U}$ is

$$\langle u, v \rangle = \int_0^1 u(t)^T v(t) dt.$$

The co-domain of T is $\mathbb{R}^n$. It's not hard to show that T is bounded. That it is bounded, even though A may not be stable, is because the time interval is finite. Nothing very bad can happen in finite time. We'll see later what the norm of T is. □

Now restrict $\mathcal{X}, \mathcal{Y}$ to be Hilbert spaces. It is a very important fact that $T : \mathcal{X} \longrightarrow \mathcal{Y}$ has an **adjoint**. This is an operator $T^* : \mathcal{Y} \longrightarrow \mathcal{X}$, in the reverse direction, satisfying the equation

$$\langle Tx, y \rangle = \langle x, T^* y \rangle.$$

Example, continued In the equation

$$\langle Tu, x \rangle = \langle u, T^* x \rangle$$

the left-hand inner product is in $\mathbb{R}^n$ and the right-hand one is in $\mathcal{L}^2[0,1]$. We have

$$\langle Tu, x\rangle = \left(\int_0^1 \mathrm{e}^{A(1-t)}Bu(t)dt\right)^T x = \int_0^1 u(t)^T B^T \mathrm{e}^{A^T(1-t)}xdt.$$

Denoting T^*x by v, we have

$$\langle u, T^*x\rangle = \int_0^1 u(t)^T v(t)dt.$$

Equating the two right-had sides, we have

$$v(t) = B^T \mathrm{e}^{A^T(1-t)}x.$$

Thus T^* is the mapping $x \mapsto v$ given by

$$(T^*x)(t) = v(t) = B^T \mathrm{e}^{A^T(1-t)}x.$$

□

The **image** of T, denoted Im T, is the set of all vectors Tx as x ranges over all $\mathcal{X}$. The image is a subspace of $\mathcal{Y}$, though it may not be closed. The **kernel** of T, denoted Ker T, is the set of all vectors x such that $Tx = 0$. The kernel is a closed subspace of $\mathcal{X}$.

Lemma 7.3 (Im $T)^\perp$ = Ker T^*

Proof

$$\begin{aligned} y \in (\text{Im } T)^\perp &\Leftrightarrow (\forall v \in \text{Im } T)\langle v, y\rangle = 0 \\ &\Leftrightarrow (\forall x)\langle Tx, y\rangle = 0 \\ &\Leftrightarrow (\forall x)\langle x, T^*y\rangle = 0 \\ &\Leftrightarrow T^*y = 0 \\ &\Leftrightarrow y \in \text{Ker } T^* \end{aligned}$$

□

7.4 A Minimization Problem

Let $\mathcal{X}, \mathcal{Y}$ be Hilbert spaces and $T : \mathcal{X} \longrightarrow \mathcal{Y}$ an operator.

Theorem 7.3 *The vector x^o minimizes $\|Tx - y\|$ iff $T^*Tx^o = T^*y$.*

Proof (Necessity) Suppose x^o minimizes $\|Tx - y\|$, that is,

$$(\forall x)\|Tx^o - y\| \leq \|Tx - y\|.$$

Define $y^o = Tx^o \in \text{Im } T$.

It is claimed that $y^o - y \perp \text{Im } T$. To prove this, following the proof of the projection theorem suppose to the contrary that there exists $y_1 \in \text{Im } T$ such that

$$\|y_1\| = 1, \ \ \langle y - y^o, y_1 \rangle = c \neq 0.$$

Then

$$\begin{aligned} \|y - (y^o + cy_1)\|^2 &= \|(y - y^o) - cy_1\|^2 \\ &= \|y - y^o\|^2 - c^2 \\ &< \|y - y^o\|^2. \end{aligned}$$

Since $y^o + cy_1 \in \text{Im } T$, there exists x such that $Tx = y^o + cy_1$. Thus

$$\|y - Tx\| < \|y - Tx^o\|.$$

This contradicts that x^o is optimal, and proves the claim.

Thus

$$y^o - y \in (\text{Im } T)^\perp = \text{Ker } T^*.$$

Hence $T^* y^o = T^* y$, i.e., $T^* T x^o = T^* y$.

(Sufficiency) Assume

$$Tx^o - y \in \text{Ker } T^* = (\text{Im } T)^\perp.$$

Then for any $y_1 \in \text{Im } T$, by Pythagorus

$$\|y_1 - y\|^2 = \|y_1 - Tx^o\|^2 + \|Tx^o - y\|^2 \geq \|Tx^o - y\|^2.$$

Thus x^o is optimal. □

Continuing with the same setup, suppose the equation $Tx = y$ is solvable. If it has more than one solution, then it has infinitely many. Suppose we'd like a solution x of minimum norm.

Theorem 7.4 *Assume the equation $Tx = y$ has a solution and that T^* has closed image. The vector x^o minimizes $\|x\|$ subject to $Tx = y$ iff $x^o = T^* z$ where z is any vector such that $TT^* z = y$.*

Proof Fix one solution $\bar{x}$ of $Tx = y$. Then any other solution has the form $\bar{x} - \tilde{x}$ where $\tilde{x} \in \text{Ker } T$. Thus the problem

$$\min_{Tx=y} \|x\|$$

is equivalent to the problem

$$\min_{\tilde{x} \in \text{Ker } T} \|\bar{x} - \tilde{x}\|.$$

By the projection theorem, and since Ker T is closed, the latter minimum exists and is unique: Let it be achieved by $\tilde{x}^o$. Define $x^o = \bar{x} - \tilde{x}^o$. Also by the projection theorem, x^o belongs to $(\text{Ker } T)^\perp$, and thus to Im T^*, since it is closed. Thus $x^o = T^* z$ for some z. Multiplying this equation by T gives $y = TT^* z$. □

Example, continued Consider the state-space system

$$\dot{x} = Ax + Bu, \quad x(0) = 0.$$

As before, let T denote the mapping from u to $x(1)$. The domain of T, denoted $\mathcal{U}$, is $\mathcal{L}^2[0, 1]$ and the co-domain of T is $\mathbb{R}^n$. The adjoint T^* is the mapping $\mathbb{R}^n \longrightarrow \mathcal{U}$ given by

$$(T^* x)(t) = B^T \mathrm{e}^{A^T(1-t)} x.$$

Let us pose the problem of finding the minimum norm u such that $x(1)$ equals a prescribed target vector. You're asked to solve this in an exercise. □

7.5 Problems

1. Prove that in a normed space every convergent sequence is a Cauchy sequence.

2. Prove the parallelogram equality in an inner product space:

   $$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2).$$

   Show that the norm $\|x\|_\infty$ in $\mathcal{C}[0, 1]$ does not satisfy the parallelogram equality, and therefore $\mathcal{C}[0, 1]$ with this norm is not an inner product space.

3. Prove that in $\mathbb{R}^n$ every subspace is closed.

4. Give an example of a function in $\mathcal{L}^2[0, \infty)$ that does not tend to zero as $t \longrightarrow \infty$.

5. Consider the system with transfer function $G(s) = 1/(s + 1)$. If the input is in $\mathcal{L}^2[0, \infty)$, does that mean the output tends to zero as $t \longrightarrow \infty$?

6. Consider $\mathcal{C}[0, 1]$ with the norm $\|x\|_\infty$. Define $\mathcal{V}$ to be the set of functions satisfying

   $$\int_0^{1/2} x(t)dt - \int_{1/2}^1 x(t)dt = 1.$$

   Show that $\mathcal{V}$ is closed (hence complete) and convex, but it does not have an element of minimum norm, that is, a vector closest to 0.

7. Consider the state-space system

   $$\dot{x} = Ax + Bu, \quad x(0) = 0.$$

   Assume (A, B) is controllable. Find u in $\mathcal{L}^2[0, 1]$ of minimum norm such that $x(1) = v$, a given vector.

8. Let $p(t)$ be a polynomial with real coefficients and of degree at most 1; that is, it has the form $p(t) = c_0 + c_1 t$ with c_0, c_1 real numbers. The task is to approximate the quadratic t^2 over the range $0 \leq t \leq 1$. This is expressed by saying that

   $$\int_0^1 [p(t) - t^2]^2 dt$$

should be minimum. There is also a constraint on $p(t)$, namely,

$$\int_0^1 p(t)dt = 0.$$

This problem can be formulated in $\mathcal{L}^2[0,1]$ as follows. To be minimized is $\|p - t^2\|$; the constraint is $\langle p, 1\rangle = 0$; also p must belong to the span of $\{1, t\}$. Define

$$\mathcal{V} = \{v : v \in \mathcal{L}^2[0,1], v \in \text{Span}\{1, t\}, v \perp 1\}.$$

Then the problem is to find p in $\mathcal{V}$ that is closest to t^2. Solve this problem via the projection theorem.

9. Consider the space $\mathbb{R}^{n\times m}$ of $n \times m$ matrices. There is a natural inner product, namely,

$$\langle X, Y\rangle = \text{trace } X^T Y.$$

The trace of a square matrix is the sum of its diagonal elements. This inner product reduces to the usual one for vectors when $m = 1$. Consider the optimization problem

$$\text{minimize}_X \|A - BXC\|.$$

Here A, B, C are real matrices, A is $n \times m$, B is $n \times p$, and C is $q \times m$. Thus X is $p \times q$. This is a problem of the form

$$\text{minimize}_X \|A - T(X)\|,$$

where T is the linear transformation given by $T(X) = BXC$. Derive the adjoint of T and write down the equation for an optimal X.

10. (a) Let $\mathcal{H}^2$ denote the space of rational transfer functions that have real coefficients and are strictly proper. (It's not a Hilbert space because it is not complete, but it's a perfectly good inner-product space.) Examples are

$$\frac{1}{s+1}, \quad \frac{s^2}{s^3 + 3s^2 + s + 1}$$

but not

$$1, \quad \frac{1}{s-1}, \quad \frac{1}{s+j}, \quad \sin s.$$

Show that this is an inner product space with inner product

$$\langle F, G\rangle = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(-j\omega)G(j\omega)d\omega.$$

(b) Consequently, the $\mathcal{H}^2$-norm is

$$\|F\|_2 = \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} |F(j\omega)|^2 d\omega\right)^{1/2}.$$

What is the corresponding norm of $f(t)$, the inverse Laplace transform of $F(s)$?

(c) If you haven't had a course in complex variables, you'll have to look up the residue theorem. For $F, G \in \mathcal{H}^2$, show that $\langle F, G\rangle$ equals the sum of the residues of $F(-s)G(s)$ at its poles in the left half-plane. For example, if $F(s) = 1/(s+1)$, then $\|F\|_2^2$ equals the residue of

$$F(-s)F(s) = \frac{1}{(-s+1)(s+1)}$$

at the pole $s = -1$, which equals $1/2$.

(d) The function $U(s) = (s-1)/(s+1)$ is called an all-pass function. Draw its magnitude Bode plot to see why.

(e) For the same $U(s)$, show that $\|UF\|_2 = \|F\|_2$ for every F in $\mathcal{H}^2$.

(f) Let

$$G(s) = \frac{1}{s+1}, \quad U(s) = \frac{s-1}{s+1}, \quad V \in \mathcal{H}^2.$$

Prove that G and UV are orthogonal and so

$$\|G + UV\|_2^2 = \|G\|_2^2 + \|V\|_2^2.$$

(g) Let T be the operator $\mathcal{H}^2 \longrightarrow \mathcal{H}^2$ that maps V to UV, U as in the preceding part. Find the adjoint operator T^*. What are TT^* and T^*T? Be careful: The adjoint of T is not the mapping $V(s) \mapsto U(-s)V(s)$; this is because $U(-s)$ has a pole in the right half-plane, and therefore $U(-s)V(s)$ is not in $\mathcal{H}^2$ in general, even though V is in $\mathcal{H}^2$.

11. Let $\mathcal{H}^\infty$ denote the set of functions $G(s)$ that are analytic and bounded in the open right half-plane. For example, if $G(s)$ is rational, then it has no poles in the closed right half-plane and it is proper (its numerator degree is not greater than its denominator degree). On this space define the norm

$$\|G\|_\infty = \sup_\omega |G(j\omega)|.$$

Now let

$$F(s) = \frac{s-1}{s+1}.$$

Prove that $F\mathcal{H}^\infty$, the set of all products FG, where G ranges over $\mathcal{H}^\infty$, is closed in $\mathcal{H}^\infty$.

12. Consider again the problem

$$\text{minimize}_x \|c - Ax\|.$$

Take

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 0 \\ 3 & 1 \end{bmatrix}, \quad c = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}.$$

Solve using the projection theorem.

13. Consider the space $\mathbb{R}^{n\times n}$ with the inner product

$$\langle X, Y\rangle = \text{Trace } X^TY.$$

Let $\mathcal{S}$ denote the subspace of symmetric matrices. Find $\mathcal{S}^\perp$. Given A, find B, C such that

$$A = B + C, \quad B \in \mathcal{S}, \quad C \in \mathcal{S}^\perp.$$

14. Find the distance in the Frobenius norm from a square matrix to the nearest lower triangular matrix.

15. Einstein's summation convention is that an expression like $\sum_k a_{ik}b_{kj}$ is abbreviated to $a_{ik}b_{kj}$. That is, in any product expression, a repeated index (k in this case) means summation. Thus if A and B are matrices such that the product $C = AB$ is defined, then $c_{ij} = a_{ik}b_{kj}$. Using this convention, prove that the trace of AB equals the trace of BA (assuming AB and BA are square).

16. Consider the vector space $\mathbb{R}^{n\times n}$ with the inner product $\langle X, Y\rangle = \text{trace} X^TY$. The set of symmetric matrices is a subspace, $\mathcal{S}$. So the problem

$$\min_{B\in\mathcal{S}} \|A - B\|$$

is that of approximating a matrix by a symmetric matrix. Solve it.

17. Show that every linear transformation $\mathbb{R}^n \longrightarrow \mathbb{R}^m$ is bounded.

18. Consider an imaginary one-dimensional straight road going off to infinity in both directions. Model the road as $\mathbb{R}$. Consider a countably infinite number of cars on the road; assume each car has been labeled with an integer number. At any particular time, the cars could be at any particular points on the road. Let $x_k(t)$ denote the location on the road of car k at time t and let $x(t)$ be the infinite vector

$$x(t) = (\dots, x_{-1}(t), x_0(t), x_1(t), \dots).$$

Let's fix t and drop the argument t in $x(t)$. Thus x is a vector with an infinite number of components, each of which could be any real number. Finally, let ℓ denote the set of all such vectors x.

The set ℓ is a vector space. But it is not a normed space, and hence it cannot be a Hilbert or Banach space. Make ℓ a topological vector space (you'll have to look up the definition of this). Modify the graph on page 86 and show where ℓ fits.

19. Suppose $a_n,\ n \geq 1$ is a non=negative sequence in ℓ^1. Prove that so is $(a_n)^{1/2}/n,\ n \geq 1$.

Chapter 8

$\mathcal{H}^2$ Optimal Control

The symbol $\mathcal{H}^2$ stands for the space of all stable, strictly proper transfer functions, such as

$$\frac{1}{s+1}, \quad \frac{2s-1}{s^2+5s+2}$$

but not

$$\frac{1}{s}, \quad \frac{s}{s+1}.$$

There's a natural inner product (dot product) and norm:

$$< P, Q >= \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{P(j\omega)}Q(j\omega)\ d\omega, \quad \|P\|_2 = \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} |P(j\omega)|^2\ d\omega\right)^{1/2}.$$

Here the bar denotes complex conjugate. The space extends to matrices, as we'll see.

In this chapter we study optimal control design in this space.

8.1 Overview

This section gives an overview of the standard $\mathcal{H}^2$ problem. An example illustrates how to set problems up.

Let $\mathcal{L}(\mathbb{R}, \mathbb{R}^n)$ denote the space of all signals from $\mathbb{R}$ to $\mathbb{R}^n$. We review the $\mathcal{L}^2$-norm of a signal u in $\mathcal{L}(\mathbb{R}, \mathbb{R}^n)$. For each time t, $u(t)$ is a vector in $\mathbb{R}^n$; denote its Euclidean norm by $\|u(t)\|$. The $\mathcal{L}^2$-norm of u is then defined to be

$$\|u\|_2 = \left(\int_{-\infty}^{\infty} \|u(t)\|^2 dt\right)^{1/2}.$$

The space $\mathcal{L}^2(\mathbb{R}, \mathbb{R}^n)$, or just $\mathcal{L}^2(\mathbb{R})$ if convenient, consists of all signals for which this norm is finite. For example, the norm is finite if $u(t)$ converges to 0 exponentially as $t \to \infty$. (Caution: $\|u\|_2 < \infty$ does not imply that $u(t) \to 0$ as $t \to \infty$—think of a counterexample.)

Before defining a norm for a transfer matrix, we have to deal a norm for complex matrices. Let R be a $p \times m$ complex matrix, that is, $R \in \mathbb{C}^{p\times m}$. There are many possible definitions for $\|R\|$; we need one. Let R^* denote the complex-conjugate transpose of R. The matrix R^*R is Hermitian and positive semidefinite. Recall that the *trace* of a square matrix is the sum of the entries on the main diagonal. It is a fact that the trace also equals the sum of the eigenvalues.

The *first definition* for $\|R\|$ (there will be another in the next chapter) is $[\text{trace}(R^*R)]^{1/2}$.

Example

$$\begin{aligned}
R &= \begin{bmatrix} 2+j & j \\ 1-j & 3-2j \end{bmatrix} \\
R^*R &= \begin{bmatrix} 2-j & 1+j \\ -j & 3+2j \end{bmatrix}\begin{bmatrix} 2+j & j \\ 1-j & 3-2j \end{bmatrix} = \begin{bmatrix} 7 & 6+3j \\ 6-3j & 14 \end{bmatrix} \\
\|R\| &= (7+14)^{1/2} = \sqrt{21}
\end{aligned}$$

Observe in this example that if r_{ij} denotes the ijth entry in R, then

$$\|R\| = \left(\sum_i \sum_j |r_{ij}|^2 \right)^{1/2}.$$

This holds in general.

Now we can define the norm of a stable $p \times m$ transfer matrix $G(s)$. Note that for each ω, $G(j\omega)$ is a $p \times m$ complex matrix.

$\mathcal{H}^2$-Norm

$$\|G\|_2 = \left\{ \frac{1}{2'i} \int_{-\infty}^{\infty} \operatorname{trace}\left[G(j\omega)^* G(j\omega)\right] d\omega \right\}^{1/2}$$

Note that the integrand equals the square of the first-definition norm of $G(j\omega)$.

Concerning this definition is an important input-output fact. Let G be a stable, causal, LTI system with input u of dimension m and output y of dimension p. Let e_i, $i = 1, \ldots, m$, denote the standard basis vectors in $\mathbb{R}^m$. Thus, δe_i is an impulse applied to the i^{th} input; $G\delta e_i$ is the corresponding output. Then the $\mathcal{H}^2$-norm of the transfer matrix G is related to the average $\mathcal{L}^2$-norm of the output when impulses are applied at the input channels.

Theorem 8.1 $\|G\|_2^2 = \sum_{i=1}^m \|G\delta e_i\|_2^2$

Thus $\|G_2\|_2$ is an average system gain for known inputs.

It is useful to be able to compute $\|G\|_2$ by state-space methods. Let

$$G(s) = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right],$$

with A stable, that is, all eigenvalues with negative real part. This matrix notation stands for the transfer matrix:

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] := C(sI-A)^{-1}B + D,$$

Then $\|G\|_2 = \infty$ unless $D = 0$, in which case the following procedure does the job:

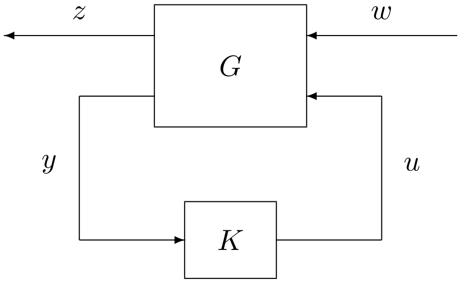
Step 1 Solve for L:

$$AL + AL^T + BB^T = 0.$$

Thus L equals the controllability Gramian.

Step 2 $\|G\|_2^2 = \text{trace } CLC^T$

Consider the standard setup shown here:



We must define the concept of internal stability for this setup. Start with a minimal realization of G:

$$G(s) = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right].$$

The input and output of G are partitioned as

$$\begin{bmatrix} w \\ u \end{bmatrix}, \quad \begin{bmatrix} z \\ y \end{bmatrix}.$$

This induces a corresponding partition of B, C, and D:

$$\begin{bmatrix} B_1 & B_2 \end{bmatrix}, \quad \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}, \quad \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}.$$

We shall *assume* that $D_{22} = 0$, that is, the transfer matrix from u to y is strictly proper. This is a condition to guarantee existence of closed-loop transfer matrices. Thus the realization for G has the form

$$G(s) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & 0 \end{array}\right].$$

Also, bring in a minimal realization of K:

$$K(s) = \left[\begin{array}{c|c} A_K & B_K \\ \hline C_K & D_K \end{array}\right].$$

Now set $w = 0$ and write the state equations describing the controlled system:

$$\begin{aligned} \dot{x} &= Ax + B_2 u \\ y &= C_2 x \\ \dot{x}_K &= A_K x_K + B_K y \\ u &= C_K x_K + D_K y. \end{aligned}$$

Eliminate u and y:

$$\begin{bmatrix} \dot{x} \\ \dot{x}_K \end{bmatrix} = \begin{bmatrix} A + B_2 D_K C_2 & B_2 C_K \\ B_K C_2 & A_K \end{bmatrix} \begin{bmatrix} x \\ x_K \end{bmatrix}.$$

We call this latter matrix the *closed-loop A-matrix*. It can be checked that its eigenvalues do not depend on the particular minimal realizations chosen for G and K. The closed-loop system is said to be *internally stable* if this closed-loop A-matrix is stable, that is, all its eigenvalues have negative real part. It can be proved that, given G, an internally stabilizing K exists iff (A, B_2) is stabilizable and (C_2, A) is detectable.

Let T_{zw} denote the system from w to z, with transfer matrix $T_{zw}(s)$. The $\mathcal{H}^2$-optimal control problem is to compute an internally stabilizing controller K that minimizes $\|T_{zw}\|_2$. The following conditions guarantee the existence of an optimal K:

(A1) (A, B_2) is stabilizable and (C_2, A) is detectable;

(A2) the matrices D_{12} and D_{21} have full column and row rank, respectively;

(A3) the matrices

$$\begin{bmatrix} A - j\omega & B_2 \\ C_1 & D_{12} \end{bmatrix}, \begin{bmatrix} A - j\omega & B_1 \\ C_2 & D_{21} \end{bmatrix}$$

have full column and row rank, respectively, $\forall \omega$;

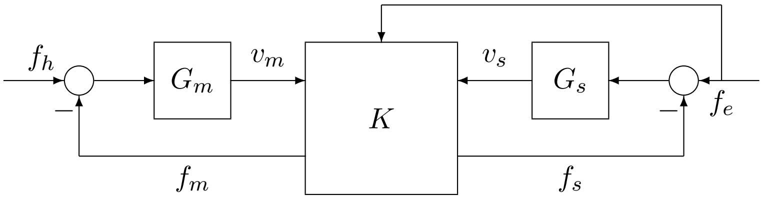
(A4) $D_{11} = 0$.

The first assumption is, as mentioned above, necessary and sufficient for existence of an internally stabilizing controller. In (A2) full column rank of D_{12} means that the control signal u is fully weighted in the output z. This is a sensible assumption, for if, say, some component of u is not weighted, there is no a priori reason for the optimal controller not to try to make this component unbounded. Dually, full row rank of D_{21} means that the exogenous signal w fully corrupts the measured signal y; it's like assuming noise for each sensor. Again, this is sensible, because otherwise the optimal controller may try to differentiate y, that is, the controller may be improper. Assumption (A3) is merely technical—an optimal controller may exist without it. In words, the assumption says there are no imaginary axis zeros in the cross systems from u to z and from w to y. Finally, (A4) guarantees that $\|T_{zw}\|_2$ is finite for every internally stabilizing and strictly proper controller (recall that T_{zw} must be strictly proper).

The problem is said to be *regular* if assumptions (A1) to (A4) are satisfied. Sometimes when we formulate a problem they are not initially satisfied; for example, we may initially not explicitly model sensor noise. Then we must modify the problem so that the assumptions *are* satisfied. This process is called *regularization*.

Under these assumptions, the MATLAB commands *h2syn* and *h2lqg* compute the optimal controller. These functions are part of the Robust Control Toolbox of MATLAB. The following example illustrates the $\mathcal{H}^2$ design technique.

Example Bilateral hybrid telerobot. The setup is shown here:



Two robots, a master, G_m, and a slave, G_s, are controlled by one controller, K. A human provides a force command, f_h, to the master, while the environment applies a force, f_e, to the slave. The controller measures the two velocities, v_m and v_s, together with f_e via a force sensor. In turn it provides two force commands, f_m and f_s, to the master and slave. Ideally, we want motion following ($v_s = v_m$), a desired master compliance (v_m a desired function of f_h), and force reflection ($f_m = f_e$).

For simplicity of computation we shall take G_m and G_s to be SISO with transfer functions

$$G_m(s) = \frac{1}{s}, \quad G_s(s) = \frac{1}{10s}.$$

We shall design K for two test inputs, namely, $f_e(t)$ is the finite-width pulse

$$f_e(t) = \begin{cases} 10, & 0 \le t \le 0.2 \\ 0, & t > 0.2, \end{cases} \tag{8.1}$$

indicating an abrupt encounter between the slave and a stiff environment, and $f_h(t)$ is the triangular pulse

$$f_h(t) = \begin{cases} 2t, & 0 \le t \le 1 \\ -2t+4, & 1 \le t \le 2 \\ 0, & t > 2, \end{cases} \tag{8.2}$$

to mimic a ramp-up, ramp-down command.

The generalized error vector is taken to have four components: the velocity error $v_m - v_s$; the compliance error $f_h - v_m$ (for simplicity, the desired compliance is assumed to be $v_m = f_h$); the force-reflection error $f_m - f_e$; and the slave actuator force. The last component is included as part of regularization, that is, to penalize excessive force applied to the slave. Introducing four weights to be decided later, we arrive at the generalized error vector

$$z = \begin{bmatrix} \alpha_v(v_m - v_s) \\ \alpha_c(f_h - v_m) \\ \alpha_f(f_m - f_e) \\ \alpha_s f_s \end{bmatrix}.$$

The Laplace transforms of f_e and f_h are not rational:

$$F_e(s) = \frac{10}{s}\left(1 - \mathrm{e}^{-0.2s}\right), \quad F_h(s) = \frac{2}{s^2}\left(1 - \mathrm{e}^{-s}\right)^2.$$

To get a tractable problem, we shall use second- and third-order Padé approximations,

$$e^{-Ts} \approx \left[1 - \frac{Ts}{2} + \frac{(Ts)^2}{12}\right] \Bigg/ \left[1 + \frac{Ts}{2} + \frac{(Ts)^2}{12}\right]$$

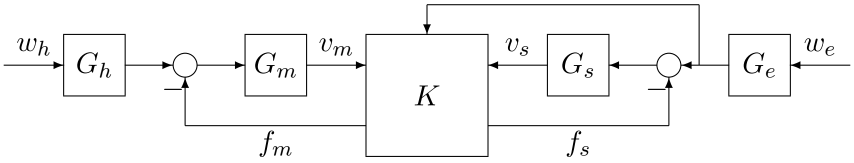
and

$$e^{-Ts} \approx \left[1 - \frac{Ts}{2} + \frac{(Ts)^2}{10} - \frac{(Ts)^3}{120}\right] \Bigg/ \left[1 + \frac{Ts}{2} + \frac{(Ts)^2}{10} + \frac{(Ts)^3}{120}\right].$$

Using the third-order approximation for $F_e(s)$ and the second-order one for $F_h(s)$, we get

$$\begin{aligned}
F_e(s) &\approx 20\left[\frac{0.2}{2} + \frac{0.2^3 s^2}{120}\right] \Bigg/ \left[1 + \frac{0.2s}{2} + \frac{(0.2s)^2}{10} + \frac{(0.2s)^3}{120}\right] \\
&=: G_e(s) \\
F_h(s) &\approx 2 \Bigg/ \left(1 + \frac{s}{2} + \frac{s^2}{12}\right)^2 \\
&=: G_h(s).
\end{aligned}$$

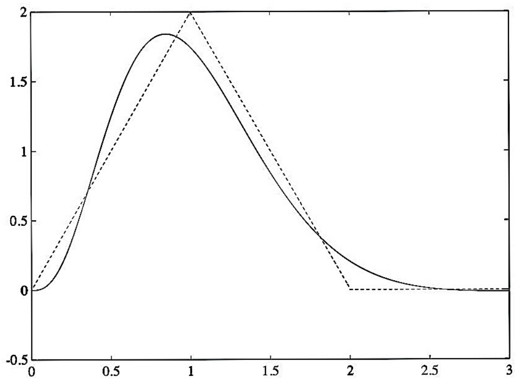
Incorporating these two prefilters into the preceding block diagram leads to this:



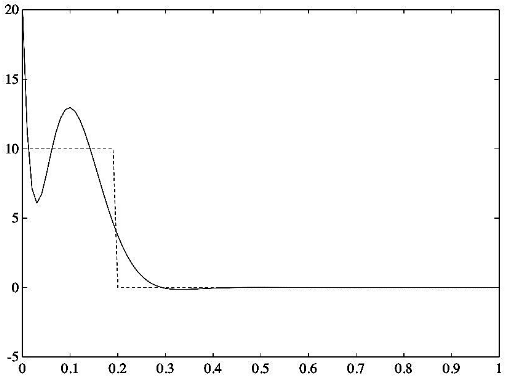
The two exogenous inputs w_h and w_e are unit impulses. The vector of exogenous inputs is therefore

$$w = \begin{bmatrix} w_h \\ w_e \end{bmatrix}.$$

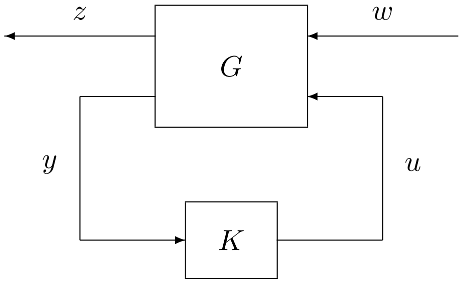
This figure compares $f_h(t)$ with the impulse response of G_h ($f_h(t)$ dash and the impulse response of G_h solid):



And this figure is for $f_e(t)$ ($f_e(t)$ dash and the impulse response of G_e solid):



The error in the second plot is larger because $f_e(t)$ is not continuous. The control system is shown in here



where z and w are as above and

$$y = \begin{bmatrix} f_e \\ v_s \\ v_m \end{bmatrix}, \quad u = \begin{bmatrix} f_m \\ f_s \end{bmatrix}.$$

Beginning with state models for G_h, G_m, G_s, G_e, namely,

$$\left[\begin{array}{c|c} A_h & B_h \\ \hline C_h & 0 \end{array}\right], \quad \left[\begin{array}{c|c} A_m & B_m \\ \hline C_m & 0 \end{array}\right], \quad \left[\begin{array}{c|c} A_s & B_s \\ \hline C_s & 0 \end{array}\right], \quad \left[\begin{array}{c|c} A_e & B_e \\ \hline C_e & 0 \end{array}\right],$$

with corresponding states x_h, x_m, x_s, x_e, and defining the state

$$x = \begin{bmatrix} x_m \\ x_s \\ x_e \\ x_h \end{bmatrix}$$

lead to the following state model for G:

$$\left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0 & D_{12} \\ C_2 & 0 & 0 \end{array}\right] :=$$

$$\left[\begin{array}{cccc|cccc}
A_m & 0 & 0 & B_mC_h & 0 & 0 & -B_m & 0 \\
0 & A_s & B_sC_e & 0 & 0 & 0 & 0 & -B_s \\
0 & 0 & A_e & 0 & 0 & B_e & 0 & 0 \\
0 & 0 & 0 & A_h & B_h & 0 & 0 & 0 \\
\hline
\alpha_vC_m & -\alpha_vC_s & 0 & 0 & 0 & 0 & 0 & 0 \\
-\alpha_cC_m & 0 & 0 & \alpha_cC_h & 0 & 0 & 0 & 0 \\
0 & 0 & -\alpha_fC_e & 0 & 0 & 0 & \alpha_fI & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha_sI \\
 & & & & & & & \\
0 & 0 & C_e & 0 & 0 & 0 & 0 & 0 \\
0 & C_s & 0 & 0 & 0 & 0 & 0 & 0 \\
C_m & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right]. \qquad (8.3)$$

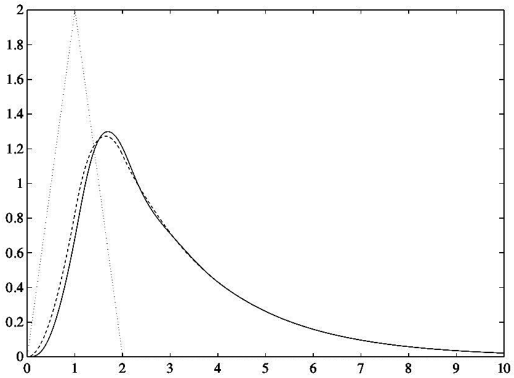
For the data at hand, $D_{21} = 0$, so (A2) fails. Evidently, the condition $D_{21} = 0$ reflects the fact that no sensor noise was modelled, that is, perfect measurements of v_m, v_s, f_e were assumed. Let us add sensor noises, say of magnitude ϵ. Then w is augmented to a 5-vector and the state matrices of G change appropriately so that the realization becomes

$$\left[\begin{array}{c|ccc}
A & 0 & B_1 & B_2 \\
\hline
C_1 & 0 & 0 & D_{12} \\
C_2 & \epsilon I & 0 & 0
\end{array}\right].$$

Some trial-and-error is required to get suitable values for the weights; the following values give reasonable responses:

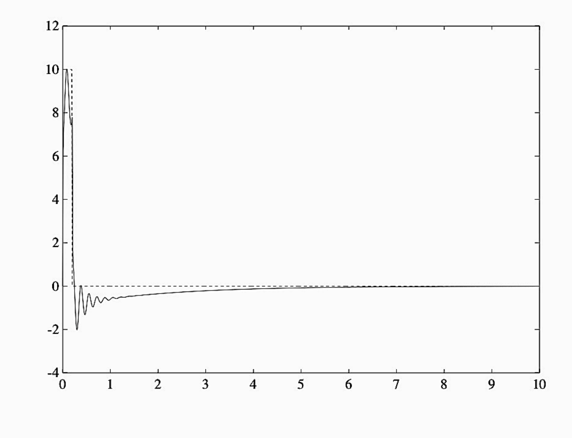
$$\alpha_v = 10, \quad \alpha_c = 5, \quad \alpha_f = 10, \quad \alpha_s = 0.01, \quad \epsilon = 0.1.$$

The MATLAB function *h2syn* can be used to compute the optimal controller. The next figure shows plots of $v_s(t)$ and $v_m(t)$ when the system is commanded by $f_h(t)$ (also shown) (v_s (solid), v_m (dash), and f_h (dot)):



The velocity tracking and compliance are quite good.

The next figure shows the response of $f_m(t)$ commanded by $f_e(t)$ (f_m (solid) and f_e (dash)).



The force reflection is evident, though there is some oscillation in $f_m(t)$. □

Here's the MATLAB code for this example:

```
%
% Program by Dan Davison
%
% Program summary:
%
% (1) - find state-space model of G and regularize it
% (2) - use H2SYN (in mu-tools) to find optimal K
%     - controller is stored in AK,BK,CK,DK
% (3) - simulate response to two types of inputs

clear

%
% (1) - SETUP STATE-SPACE MODEL FOR G
%

numG_m = [1];
denG_m = [1 0];
[A_m,B_m,C_m,D_m] = tf2ss(numG_m,denG_m);

numG_s = [1];
denG_s = [10 0];
[A_s,B_s,C_s,D_s] = tf2ss(numG_s,denG_s);

numG_e = 20*[0.2^3/120  0  0.2/2];
denG_e = [0.2^3/120  0.2^2/10  0.2/2  1];
[A_e,B_e,C_e,D_e] = tf2ss(numG_e,denG_e);

numG_h = [2];
temp = [1/12 1/2 1];
denG_h = conv(temp,temp);
```

```
[A_h,B_h,C_h,D_h] = tf2ss(numG_h,denG_h);

[n_m,m_m]=size(B_m);
[n_s,m_s]=size(B_s);
[n_e,m_e]=size(B_e);
[n_h,m_h]=size(B_h);
[p_m,n_m]=size(C_m);
[p_s,n_s]=size(C_s);
[p_e,n_e]=size(C_e);
[p_h,n_h]=size(C_h);

A = [    A_m       zeros(n_m,n_s) zeros(n_m,n_e)    B_m*C_h
      zeros(n_s,n_m)     A_s         B_s*C_e      zeros(n_s,n_h)
      zeros(n_e,n_m) zeros(n_e,n_s)    A_e        zeros(n_e,n_h)
      zeros(n_h,n_m) zeros(n_h,n_s) zeros(n_h,n_e)     A_h      ];

tmp = [zeros(n_m,m_h) zeros(n_m,m_e)
      zeros(n_s,m_h) zeros(n_s,m_e)
      zeros(n_e,m_h)     B_e
          B_h       zeros(n_h,m_e)];
B1 = [zeros(n_m+n_s+n_e+n_h,3) tmp];

B2 = [ -B_m       zeros(n_m,m_s)
      zeros(n_s,m_m)    -B_s
      zeros(n_e,m_m) zeros(n_e,m_s)
      zeros(n_h,m_m) zeros(n_h,m_s)];

B = [B1 B2];

% weights on, resp, v_m - v_s, f_h - v_m, f_m - f_e, f_s

w_v = 10;
w_z = 5;
w_f = 10;
w_s = .01;
weight = diag([w_v w_z w_f w_s]);


C1 = [    C_m          -C_s       zeros(p_m,n_e) zeros(p_m,n_h)
         -C_m      zeros(p_m,n_s)   zeros(p_m,n_e)     C_h
      zeros(p_e,n_m) zeros(p_e,n_s)    -C_e       zeros(p_e,n_h)
      zeros(p_s,n_m) zeros(p_s,n_s) zeros(p_s,n_e) zeros(p_s,n_h) ];

C1 = weight*C1;

C2 = [ zeros(p_e,n_m) zeros(p_e,n_s)     C_e       zeros(p_e,n_h)
```

```
      zeros(p_s,n_m)     C_s       zeros(p_s,n_e) zeros(p_s,n_h)
          C_m       zeros(p_m,n_s) zeros(p_m,n_e) zeros(p_m,n_h)];

C = [C1;C2];

epsilon = 0.1; % weight on added noise

D11 = zeros(4,5);

D12 = [0 0
       0 0
       1 0
       0 1];
D12 = weight*D12;

D21= [epsilon 0 0 0 0
       0 epsilon 0 0 0
       0 0 epsilon 0 0];

D22= zeros(3,2);

D = [D11 D12;D21 D22];


% run h2syn

plant=pck(A,B,C,D);
[kk,gg,kfi,gfi,hamx,hamy]=h2syn(plant,3,2,2);
[AK,BK,CK,DK]=unpck(kk);


% set up for simulation

[nk,mk]=size(BK);
[pk,nk]=size(CK);
CK1 = CK(1,1:nk);
CK2 = CK(2,1:nk);
BK1 = BK(1:nk,1);
BK2 = BK(1:nk,2);
BK3 = BK(1:nk,3);

if norm(DK) > eps
  error('DK is not zero.  Program needs updating.')
end

%
```

```
% (3) - SIMULATE CLOSED LOOP SYSTEM
%

% Create system M from [f_e,f_h] to [v_m,v_s,f_e,f_m]:

AM = [     A_m      zeros(n_m,n_s)   -B_m*CK1    ;...
      zeros(n_s,n_m)     A_s         -B_s*CK2    ;...
        BK3*C_m       BK2*C_s          AK      ];

BM = [zeros(n_m,m_s)      B_m     ;...
          B_s       zeros(n_s,m_m);...
         BK1       zeros(nk,m_m)];

CM = [    C_m      zeros(1,n_s) zeros(1,nk) ;...
      zeros(1,n_m)    C_s       zeros(1,nk) ;...
      zeros(1,n_m) zeros(1,n_s) zeros(1,nk) ;...
      zeros(1,n_m) zeros(1,n_s)    CK1      ];

DM = [0 0; 0 0; 1 0; 0 0];

% first, simulate with f_e = 0

Tmax = 10;
delT = .01;

T1=0:delT:Tmax;

fe = zeros(length(T1),1);
fh = zeros(length(T1),1);

for i=1:length(T1)*1/Tmax
  t = (i-1)*Tmax/length(T1);
  fh(i) = 2*t;
end

for i=length(T1)*1/Tmax+1:length(T1)*2/Tmax
  t = (i-1)*Tmax/length(T1);
  fh(i) = -2*t+4;
end

Output = lsim(AM,BM,CM,DM,[fe fh],T1);
vm1 = Output(:,1);
vs1 = Output(:,2);
fe1 = Output(:,3);
fm1 = Output(:,4);
fh1=fh;
```

```
% second, simulate with f_h = 0

Tmax = 10;
delT = .01;

T2=0:delT:Tmax;

fe = zeros(length(T2),1);
fh = zeros(length(T2),1);

for i=1:length(T2)*.2/Tmax
  t = (i-1)*Tmax/length(T2);
  fe(i) = 10;
end

Output = lsim(AM,BM,CM,DM,[fe fh],T2);
vm2 = Output(:,1);
vs2 = Output(:,2);
fe2 = Output(:,3);
fm2 = Output(:,4);

%plot(T1, [vs1 vm1 fh1])
plot(T2,[fm2 fe2])
```

After that tutorial on the use of $\mathcal{H}^2$ optimal control, we return to the theory.

8.2 Lyapunov Equation

The equation

$$A^T X + XA + M = 0$$

is called a **Lyapunov equation**. Here A, M, X are all square matrices, say $n \times n$, with M symmetric.

One situation is where A and M are given and the equation is to be solved for X. Existence and uniqueness are easy to establish in principle. Define the linear map

$$\mathbf{L} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}, \quad \mathbf{L}(X) = A^T X + XA.$$

Then the Lyapunov equation has a solution X iff $M \in \operatorname{Im} \mathbf{L}$; if this condition holds, the solution is unique iff $\mathbf{L}$ is one-to-one, hence invertible. Let $\sigma()$ denote the set of eigenvalues—the spectrum—of a matrix or linear transformation. It can be shown that

$$\sigma(\mathbf{L}) = \{\lambda_1 + \lambda_2 : \lambda_1, \lambda_2 \in \sigma(A)\}.$$

So the Lyapunov equation has a unique solution iff A has the property that no two of its eigenvalues add to zero. For example, if A is stable, the unique solution is

$$X = \int_0^\infty \mathrm{e}^{A^T t} M \mathrm{e}^{At} dt.$$

This can be proved as follows. Let $P(t) = \mathrm{e}^{A^T t} M \mathrm{e}^{At}$. Then

$$\dot{P}(t) = A^T P(t) + P(t)A.$$

Integrate from $t = 0$ to ∞.

We'll be more interested in another situation—where we want to infer stability of A.

Theorem 8.2 *Suppose A, M, X satisfy the Lyapunov equation, (M, A) is detectable, and M and X are positive semi-definite. Then A is stable.*

Proof For a proof by contradiction, suppose A has some eigenvalue λ with Re $\lambda \geq 0$. Let x be a corresponding eigenvector. Pre-multiply the Lyapunov equation by x^*, the complex-conjugate transpose, and post-multiply by x to get

$$(2\text{Re } \lambda)x^* X x + x^* M x = 0.$$

Both terms on the left are ≥ 0. Hence $x^* M x = 0$, which implies that $Mx = 0$ since $M \geq 0$. Thus

$$\begin{bmatrix} A - \lambda I \\ M \end{bmatrix} x = 0.$$

By detectability we must have $x = 0$, a contradiction. □

8.3 Spectral Subspaces

Let A be a square matrix. There's a similarity transformation that maps A to a matrix of the form

$$\begin{bmatrix} A^- & 0 \\ 0 & A^+ \end{bmatrix},$$

where the eigenvalues of A^- are all in Re $s < 0$ and those of A^+ in Re $s \geq 0$. That is to say, there are two invariant subspaces of A, say $\mathcal{X}^-(A)$ and $\mathcal{X}^+(A)$, that are independent; furthermore, if $v_1, \ldots, v_k$ is a basis for $\mathcal{X}^-(A)$ and $v_{k+1}, \ldots, v_n$ for $\mathcal{X}^+(A)$, then the matrix V formed from these basis vectors satisfies

$$V^{-1}AV = \begin{bmatrix} A^- & 0 \\ 0 & A^+ \end{bmatrix}.$$

These two subspaces are called the **spectral subspaces** of A.

One can get V from the generalized eigenvectors, although you have to keep the vectors real. A simpler way, at least conceptually, is as follows. Let $p(s)$ denote the characteristic polynomial of A. It can be factored as

$$p(s) = p^-(s)p^+(s),$$

where the roots of p^- are in Re $s < 0$ and those of p^+ in Re $s \geq 0$. Then

$$\mathcal{X}^-(A) = \text{Ker } p^-(A), \quad \mathcal{X}^+(A) = \text{Ker } p^+(A).$$

8.4 Riccati Equation

Let A, P, Q be real $n \times n$ matrices with P and Q symmetric. Define the $2n \times 2n$ matrix

$$H := \begin{bmatrix} A & -P \\ -Q & -A^T \end{bmatrix}.$$

A matrix of this form is called a **Hamiltonian matrix**.

It is claimed that $\sigma(H)$ is symmetric about the imaginary axis. To prove this, introduce the $2n \times 2n$ matrix

$$J := \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}$$

having the property $J^2 = -I$. Then

$$J^{-1}HJ = -JHJ = -H^T$$

so H and $-H^T$ are similar. Thus λ is an eigenvalue iff $-\lambda$ is.

Now assume H has no eigenvalues on the imaginary axis. Then it must have n in Re $s < 0$ and n in Re $s > 0$. Thus the two spectral subspaces $\mathcal{X}^-(H)$ and $\mathcal{X}^+(H)$ both have dimension n. Let's focus on $\mathcal{X}^-(H)$. Finding a basis for it, stacking the basis vectors up to form a matrix, and partitioning the matrix, we get

$$\mathcal{X}^-(H) = \text{Im} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

where $X_1, X_2 \in \mathbb{R}^{n \times n}$. If X_1 is nonsingular, i.e., if the two subspaces

$$\mathcal{X}^-(H), \quad \text{Im} \begin{bmatrix} 0 \\ I \end{bmatrix}$$

are complementary, we can set $X := X_2 X_1^{-1}$ to get

$$\mathcal{X}^-(H) = \text{Im} \begin{bmatrix} I \\ X \end{bmatrix}.$$

Notice that X is then uniquely determined by H, i.e. $H \mapsto X$ is a function. We shall denote this function by *Ric* and write $X = Ric(H)$.

To recap, *Ric* is a (nonlinear) function $\mathbb{R}^{2n \times 2n} \to \mathbb{R}^{n \times n}$ which maps H to X where

$$\mathcal{X}^-(H) = \text{Im} \begin{bmatrix} I \\ X \end{bmatrix}.$$

The domain of *Ric*, denoted *dom Ric*, consists of Hamiltonian matrices H with two properties, namely, H has no eigenvalues on the imaginary axis and the two subspaces

$$\mathcal{X}^-(H), \quad \text{Im} \begin{bmatrix} 0 \\ I \end{bmatrix}$$

are complementary.

Some properties of X are given below.

Lemma 8.1 *Suppose $H \in$ dom Ric and $X = Ric(H)$. Then*

(i) X is symmetric

(ii) X satisfies the algebraic Riccati equation

$$A^T X + XA - XPX + Q = 0$$

(iii) $A - PX$ is stable.

Proof (i) Let X_1, X_2 be as above. It's claimed that

$$X_1^T X_2 \text{ is symmetric.} \tag{8.4}$$

To prove this, note that there exists a stable matrix H^- in $\mathbb{R}^{n \times n}$ such that

$$H \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} H^-.$$

Pre-multiply this equation by

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}^T J$$

to get

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}^T JH \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}^T J \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} H^-. \tag{8.5}$$

Now JH is symmetric; hence so is the left-hand side of (8.5); hence so is the right:

$$\begin{aligned} (-X_1^T X_2 + X_2^T X_1)H^- &= H^{-T}(-X_1^T X_2 + X_2^T X_1)^T \\ &= -H^{-T}(-X_1^T X_2 + X_2^T X_1). \end{aligned}$$

This is a Lyapunov equation. Since H^- is stable, the unique solution is

$$-X_1^T X_2 + X_2^T X_1 = 0.$$

This proves (8.4).

We have $XX_1 = X_2$. Pre-multiply by X_1^T and then use (8.4) to get that $X_1^T X X_1$ is symmetric. Since X_1 is nonsingular, this implies that X is symmetric too.

(ii) Start with the equation

$$H \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} H^-$$

and extract X_1 to get

$$H \begin{bmatrix} I \\ X \end{bmatrix} X_1 = \begin{bmatrix} I \\ X \end{bmatrix} X_1 H^-.$$

Post-multiply by X_1^{-1}:

$$H \left[\begin{array}{c} I \\ X \end{array} \right] = \left[\begin{array}{c} I \\ X \end{array} \right] X_1 H^- X_1^{-1}. \tag{8.6}$$

Now pre-multiply by $\left[\begin{array}{cc} X & -I \end{array} \right]$:

$$\left[\begin{array}{cc} X & -I \end{array} \right] H \left[\begin{array}{c} I \\ X \end{array} \right] = 0.$$

This is precisely the Riccati equation.

(iii) Pre-multiply (8.6) by $\left[\begin{array}{cc} I & 0 \end{array} \right]$ to get

$$A - PX = X_1 H^- X_1^{-1}.$$

Thus $A - PX$ is stable because H^- is. □

The following result gives verifiable conditions under which H belongs to *dom Ric*.

Theorem 8.3 *Suppose H has the form*

$$H = \left[\begin{array}{cc} A & -BR^{-1}B^T \\ -Q & -A^T \end{array} \right]$$

with $Q \geq 0$, $R > 0$, (A, B) stabilizable, and (Q, A) detectable. Then $H \in$ dom Ric. Let $X = Ric(H)$ and $F = -R^{-1}B^T X$. Then $X \geq 0$ and $A + BF$ is stable. Finally, if (Q, A) is observable, then $X > 0$.

Proof We'll first show that H has no imaginary eigenvalues. Suppose, on the contrary, that $j\omega$ is an eigenvalue and $\left[\begin{array}{c} x \\ z \end{array} \right]$ a corresponding eigenvector. Then

$$\begin{aligned} Ax - BR^{-1}B^T z &= j\omega x \qquad (8.7) \\ -Qx - A^T z &= j\omega z. \qquad (8.8) \end{aligned}$$

Re-arrange:

$$\begin{aligned} (A - j\omega I)x &= BR^{-1}B^T z \qquad (8.9) \\ -(A - j\omega I)^* z &= Qx \qquad (8.10) \end{aligned}$$

Thus

$$\begin{aligned} \langle z, (A - j\omega I)x \rangle &= \langle z, BR^{-1}B^T z \rangle = \|R^{-1/2}B^T z\|^2 \\ -\langle x, (A - j\omega I)^* z \rangle &= \langle x, Qx \rangle = \|Q^{1/2}x\|^2 \end{aligned}$$

and hence

$$\begin{aligned} \langle z, (A - j\omega I)x \rangle &= \|R^{-1/2}B^T z\|^2 \\ \langle (A - j\omega I)x, z \rangle &= -\|Q^{1/2}x\|^2. \end{aligned}$$

Thus $\langle (A - j\omega I)x, z\rangle$ is real and

$$-\|Q^{1/2}x\|^2 = \langle (A - j\omega I)x, z\rangle = \|R^{-1/2}B^T z\|^2.$$

Therefore $B^T z = 0$ and $Qx = 0$. So from (8.9) and (8.10)

$$\begin{aligned} (A - j\omega I)x &= 0 \\ (A - j\omega I)^* z &= 0. \end{aligned}$$

Combine the last four equations to get

$$z^* \begin{bmatrix} A - j\omega I & R^{-1}B \end{bmatrix} = 0$$

$$\begin{bmatrix} A - j\omega I \\ Q \end{bmatrix} x = 0.$$

By stabilizability and detectability it follows that $x = z = 0$, a contradiction.

Next, we'll show that

$$\mathcal{X}^-(H), \quad \text{Im} \begin{bmatrix} 0 \\ I \end{bmatrix}$$

are complementary. This requires a preliminary step. As in the proof of Lemma 8.1 bring in X_1, X_2, H^- so that

$$\mathcal{X}^-(H) = \text{Im} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

$$H \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} H^-. \tag{8.11}$$

We want to show that X_1 is nonsingular, i.e. Ker $X_1 = 0$. First, it is claimed that Ker X_1 is H^--invariant. To prove this, let $x \in \text{Ker}\, X_1$. Pre-multiply (8.11) by $\begin{bmatrix} I & 0 \end{bmatrix}$ to get

$$AX_1 - BR^{-1}B^T X_2 = X_1 H^-. \tag{8.12}$$

Pre-multiply by $x^T X_2^T$, post-multiply by x, and use the fact that $X_2^T X_1$ is symmetric (see (8.4)) to get

$$-x^T X_2^T BR^{-1}B^T X_2 x = 0.$$

Thus $B^T X_2 x = 0$. Now post-multiply (8.12) by x to get $X_1 H^- x = 0$, i.e. $H^- x \in \text{Ker}\, X_1$. This proves the claim.

Now to prove that X_1 is nonsingular, suppose on the contrary that Ker $X_1 \neq 0$. Then H restricted to Ker X_1 has an eigenvalue, λ, and a corresponding eigenvector, x:

$$H^- x = \lambda x \tag{8.13}$$

$$\text{Re}\,\lambda < 0, \quad 0 \neq x \in \text{Ker}\, X_1.$$

Pre-multiply (8.11) by $\begin{bmatrix} 0 & I \end{bmatrix}$:

$$-QX_1 - A^T X_2 = X_2 H^-. \tag{8.14}$$

Post-multiply by x and use (8.13):

$$(A^T + \lambda I)X_2 x = 0.$$

Since $B^T X_2 x = 0$ too from above, we have

$$x^* X_2^T \begin{bmatrix} A + \overline{\lambda} I & B \end{bmatrix} = 0.$$

Then stabilizability implies $X_2 x = 0$. But if $X_1 x = 0$ and $X_2 x = 0$, then $x = 0$, a contradiction. This concludes the proof of complementarity, and hence the proof that $H \in domRic$.

That $A + BF$ is stable follows from part (iii) of Lemma 8.1 ($P = BR^{-1}B^T$).

To show that $X \geq 0$, we have the Riccati equation

$$A^T X + XA - XBR^{-1}B^T X + Q = 0,$$

or equivalently

$$(A + BF)^T X + X(A + BF) + XBR^{-1}B^T X + Q = 0.$$

Thus

$$X = \int_0^\infty \mathrm{e}^{(A+BF)^T t}(XBR^{-1}B^T X + Q)\mathrm{e}^{(A+BF)t} dt. \tag{8.15}$$

Since $XBR^{-1}B^T X + Q$ is positive semi-definite, so is X.

Finally, suppose (Q, A) is observable. We'll show that if $x^T X x = 0$, then $x = 0$; thus $X > 0$. Pre-multiply (8.15) by x^T and post-multiply by x:

$$x^T X x = \int_0^\infty \|R^{-1/2} B^T X \mathrm{e}^{(A+BF)t} x\|^2 dt + \int_0^\infty \|Q^{1/2} \mathrm{e}^{(A+BF)t} x\|^2 dt.$$

Thus if $x^T X x = 0$, then $Xx = 0$ and

$$Q\mathrm{e}^{(A+BF)t} x = 0, \quad \forall t \geq 0.$$

But this implies that x belongs to the unobservable subspace of (Q, A) and so $x = 0$. □

8.5 The LQR Problem

The LQR problem is, given a plant model

$$\dot{x} = Ax + Bu, \quad x(0) = x_0,$$

find a control input u that minimizes the cost functional

$$J = \int_0^\infty x(t)^T Q x(t) + u(t)^T R u(t) dt.$$

We have all the machinery now to solve this.

Assume $Q \geq 0$, $R > 0$, (A, B) stabilizable, and (Q, A) detectable. Define

$$H = \begin{bmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{bmatrix}, \quad X = Ric(H), \quad F = -R^{-1}B^T X.$$

By Theorem 8.3, X is well-defined, $X \geq 0$, and $A + BF$ is stable. The associated Riccati equation is

$$A^T X + XA - XBR^{-1}B^T X + Q = 0,$$

Theorem 8.4 *The control signal that minimizes J is $u = Fx$, it is the unique optimal control, and for this control signal $J = x_0^T X x_0$.*

The proof needs a lemma. Let us denote by $\mathcal{L}^2$ the class of signals that are square-integrable on the time interval $[0, \infty)$.

Lemma 8.2 *If J is finite, then $x(t) \to 0$ as $t \to \infty$.*

Proof Assume $J < \infty$. Then $u \in \mathcal{L}^2$ because $R > 0$. Let $C := Q^{1/2}$ and $y := Cx$. Then $y \in \mathcal{L}^2$ too. By detectability, there exists K such that $A + KC$ is stable. A standard observer to estimate x is

$$\begin{aligned} \dot{\hat{x}} &= A\hat{x} + Bu + K(C\hat{x} - y) \\ &= (A + KC)\hat{x} + Bu - Ky. \end{aligned}$$

Since $A + KC$ is stable, $u \in \mathcal{L}^2$, and $y \in \mathcal{L}^2$, so $\hat{x}(t) \to 0$. By observer theory, $\hat{x}(t) - x(t) \to 0$. Thus $x(t) \to 0$. □

Proof of Theorem The proof is a trick using the completion of a square. Let u be an arbitrary control input for which J is finite. We shall differentiate the quadratic form $x(t)^T X x(t)$ along the solution of the plant equation. To simplify notation, we suppress dependence on t. We have

$$\begin{aligned} \frac{d}{dt}(x^T X x) &= \dot{x}^T X x + x^T X \dot{x} \\ &= (Ax + Bu)^T X x + x^T X (Ax + Bu) \\ &= x^T (A^T X + XA) x + 2u^T B^T X x \\ &= x^T (XBR^{-1}B^T X - Q) x + 2u^T B^T X x \text{ from the Riccati equation} \\ &= -x^T Q x + x^T XBR^{-1}B^T X x + 2u^T B^T X x \\ &= -x^T Q x + x^T XBR^{-1}B^T X x + 2u^T B^T X x + (u^T R u - u^T R u) \\ &\quad \text{—this was the completion of squares trick} \\ &= -x^T Q x - u^T R u + \|R^{-1/2} B^T X x + R^{1/2} u\|^2. \end{aligned}$$

Rearranging terms we have

$$x^T Q x + u^T R u = -\frac{d}{dt}(x^T X x) + \|R^{-1/2} B^T X x + R^{1/2} u\|^2.$$

Now integrate from $t = 0$ to $t = \infty$ and use the lemma:

$$J = x_0^T X x_0 + \int_0^\infty \|R^{-1/2} B^T X x + R^{1/2} u\|^2 dt.$$

Thus J is minimum iff $R^{-1/2}B^TXx + R^{1/2}u \equiv 0$, i.e., $u = Fx$. The other conclusion follows. □

The LQR solution provides a very convenient way to stabilize an LTI plant. Given A, B, select Q, R with $Q \geq 0$, (Q, A) detectable, and $R > 0$. Then the optimal F stabilizes $A + BF$. This is the preferred method over pole assignment.

The LQR solution is rarely implementable as it stands, because it requires that there be a sensor for each state variable; that is, x must be fully sensed. We look next at the generalization of the LQR problem to the more general case where x is not fully sensed.

8.6 Solution of the $\mathcal{H}^2$ Problem

In this section we see a derivation of the solution of the $\mathcal{H}^2$ problem in a special case. We take the realization of the transfer matrix G to be of the form

$$G(s) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0 & D_{12} \\ C_2 & D_{21} & 0 \end{array}\right].$$

The following assumptions are made:

(A1) (A, B_1) is stabilizable and (C_1, A) is detectable

(A2) (A, B_2) is stabilizable and (C_2, A) is detectable

(A3) $D_{12}^T \begin{bmatrix} C_1 & D_{12} \end{bmatrix} = \begin{bmatrix} 0 & I \end{bmatrix}$

(A4) $\begin{bmatrix} B_1 \\ D_{21} \end{bmatrix} D_{21}^T = \begin{bmatrix} 0 \\ I \end{bmatrix}$

Assumption (A2) is necessary and sufficient for G to be internally stabilizable. Assumption (A1) is for a technical reason: Together with (A2) it guarantees that two Hamiltonian matrices (H_2 and J_2 below) belong to *dom*(*Ric*).

Assumption (A3) means that C_1x and $D_{12}u$ are orthogonal and that the latter control penalty is nonsingular and normalized. In the conventional LQG setting this means that there is no cross weighting between the state and control input, and that the control weight matrix is the identity . Other nonsingular control weights can easily be converted to this problem with a change of coordinates in u. Relaxing the orthogonality condition introduces a few extra terms in the controller formulas.

Finally, assumption (A4) is dual to (A3) and concerns how the exogenous signal w enters G: The plant disturbance and the sensor noise are orthogonal, and the sensor noise weighting is normalized and nonsingular. Two additional assumptions that are implicit in the assumed realization for G is that $D_{11} = 0$ and $D_{22} = 0$. Relaxing these assumptions complicates the formulas substantially.

By Theorem 8.3 the Hamiltonian matrices

$$H_2 := \begin{bmatrix} A & -B_2B_2^T \\ -C_1^TC_1 & -A^T \end{bmatrix}, \quad J_2 := \begin{bmatrix} A^T & -C_2^TC_2 \\ -B_1B_1^T & -A \end{bmatrix}$$

belong to *dom(Ric)* and, moreover, $X_2 := \mathit{Ric}(H_2)$ and $Y_2 := \mathit{Ric}(J_2)$ are positive semi-definite. Define

$$F_2 := -B_2^TX_2, \quad L_2 := -Y_2C_2^T$$

$$A_{F_2} := A + B_2F_2, \quad C_{1F_2} := C_1 + D_{12}F_2$$

$$A_{L_2} := A + L_2C_2, \quad B_{1L_2} := B_1 + L_2D_{21}$$

$$\hat{A}_2 := A + B_2F_2 + L_2C_2$$

$$G_c(s) := \left[\begin{array}{c|c} A_{F_2} & I \\ \hline C_{1F_2} & 0 \end{array}\right], \qquad G_f(s) := \left[\begin{array}{c|c} A_{L_2} & B_{1L_2} \\ \hline I & 0 \end{array}\right]$$

and let T_{zw} denote the transfer function from w to z.

Theorem 8.5 *The unique optimal controller is*

$$K_{opt}(s) := \left[\begin{array}{c|c} \hat{A}_2 & -L_2 \\ \hline F_2 & 0 \end{array}\right].$$

Moreover,

$$\min \|T_{zw}\|_2^2 = \|G_cB_1\|_2^2 + \|F_2G_f\|_2^2.$$

The first term in the minimum cost, $\|G_cB_1\|_2^2$, is associated with optimal control with state feedback and the second, $\|F_2G_f\|_2^2$, with optimal filtering. These two norms can easily be computed as follows:

$$\|G_cB_1\|_2^2 = \text{trace } (B_1^TX_2B_1)$$

$$A_{F_2}^TX_2 + X_2A_{F_2} + C_{1F_2}^TC_{1F_2} = 0$$

$$\|F_2G_f\|_2^2 = \text{trace } (F_2Y_2F_2^T)$$

$$A_{L_2}Y_2 + Y_2A_{L_2}^T + B_{1L_2}B_{1L_2}^T = 0.$$

The controller K_{opt} has a beautiful separation structure: The controller equations can be written as

$$\dot{\hat{x}} = A\hat{x} + B_2u + L_2(C_2\hat{x} - y)$$

$$u = F_2\hat{x}.$$

The matrix F_2 is the optimal feedback gain were x directly measured; L_2 is the optimal filter gain; $\hat{x}$ is the optimal estimate of x.

The proof involves optimality in an inner-product space; so projection theory applies. Let $\mathcal{H}^2$ denote the (Hardy) space of transfer matrices $P(s)$ that are stable and strictly proper. This has a natural inner product,

$$< P, Q >= \frac{1}{2\pi}\int_{-\infty}^{\infty} \text{trace } [P(j\omega)^*Q(j\omega)] \, d\omega,$$

which is consistent with our norm definition: $\|P\|_2^2 =< P, P >$. Likewise, let $\mathcal{H}^{2^\perp}$ denote the space of transfer matrices that are antistable (all poles in Re $s > 0$) and strictly proper. Same inner product, same norm. Then $\mathcal{H}^2$ and $\mathcal{H}^{2^\perp}$ are orthogonal spaces:

$$P \in \mathcal{H}^2, Q \in \mathcal{H}^{2^\perp} \implies < P, Q >= 0.$$

The sum $\mathcal{H}^2 \oplus \mathcal{H}^{2^\perp}$ consists of all transfer matrices that are strictly proper and have no poles on the imaginary axis.

Finally, let us introduce the notation $P^\sim(s) := P(-s)^T$. A stable matrix $P(s)$ is said to be **allpass** if $P^\sim P = I$. For example,

$$\frac{s-1}{s+1}$$

is an allpass function.

Proof of Theorem 8.5 Let K be any proper, stabilizing controller. Start with the system equations

$$\dot{x} = Ax + B_1 w + B_2 u$$

$$z = C_1 x + D_{12} u$$

and define a new control variable, $v := u - F_2 x$. The equations become

$$\dot{x} = A_{F_2} x + B_1 w + B_2 v$$

$$z = C_{1F_2} x + D_{12} v$$

or in the frequency-domain

$$Z = G_c B_1 W + UV.$$

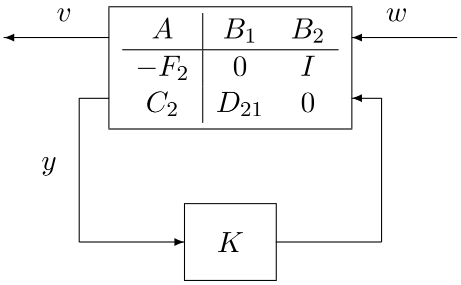
This implies that

$$T_{zw} = G_c B_1 + U T_{vw}$$

You will prove in Problem 5 the following fact: U is allpass and $U^\sim G_c$ belongs to $\mathcal{H}^{2^\perp}$. This implies that $G_c B_1$ and $U T_{vw}$ are orthogonal matrices in $\mathcal{H}^2$ (T_{vw} belongs to $\mathcal{H}^2$ by internal stability). So from the previous equation

$$\|T_{zw}\|_2^2 = \|G_c B_1\|_2^2 + \|T_{vw}\|_2^2.$$

Now look at how v is generated:



Note that K stabilizes G iff K stabilizes the above system (the two closed-loop systems have identical A-matrices). So

$$\min_K \|T_{zw}\|_2^2 = \|G_c B_1\|_2^2 + \min_K \|T_{vw}\|_2^2$$

and therefore the theorem will be proved once we show the following: For the setup in the previous block diagram, the unique optimal controller is

$$\left[\begin{array}{c|c} A + B_2F_2 + L_2C_2 & -L_2 \\ \hline F_2 & 0 \end{array}\right]$$

and the minimum value of $\|T_{vw}\|_2$ equals $\|F_2 G_f\|_2$. Notice in this setup that $A + B_2F_2$ is stable.

By the assignment $C_1 \leftarrow -F_2$, the previous statement becomes this: For

$$G(s) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0 & I \\ C_2 & D_{21} & 0 \end{array}\right]$$

with $A - B_2C_1$ stable, the unique optimal controller is

$$\left[\begin{array}{c|c} A - B_2C_1 + L_2C_2 & L_2 \\ \hline C_1 & 0 \end{array}\right]$$

and the minimum cost is $\|C_1 G_f\|_2$.

The dual of the last statement is this: For

$$G(s) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0 & D_{12} \\ C_2 & I & 0 \end{array}\right]$$

with $A - B_1C_2$ stable, the unique optimal controller is

$$\left[\begin{array}{c|c} A + B_2F_2 - B_1C_2 & B_1 \\ \hline F_2 & 0 \end{array}\right]$$

and the minimum cost is $\|G_c B_1\|_2$.

To prove this, apply the controller and let $\hat{x}$ denote its state. Then the system equations are

$$\dot{x} = Ax + B_1w + B_2u$$

$$z = C_1x + D_{12}u$$

$$y = C_2x + w$$

$$\dot{\hat{x}} = (A + B_2F_2 - B_1C_2)\hat{x} + B_1y$$

$$u = F_2\hat{x},$$

so

$$\dot{\hat{x}} = A\hat{x} + B_2u + B_1(y - C_2\hat{x}).$$

Defining $e := x - \hat{x}$, we get

$$\dot{e} = (A - B_1 C_2)e.$$

It's now easy to infer internal stability from stability of $A + B_2F_2$ and $A - B_1C_2$. For zero initial conditions on x, $\hat{x}$, we have $e(t) \equiv 0$. Hence

$$u = F_2\hat{x} = F_2x. \tag{8.16}$$

For every proper, stabilizing controller the equation

$$\|T_{zw}\|_2^2 = \|G_cB_1\|_2^2 + \|T_{vw}\|_2^2$$

is still valid, showing that

$$\|T_{zw}\|_2 \geq \|G_cB_1\|_2.$$

But for the present controller, (8.16) implies that $v \equiv 0$, i.e., $T_{vw} = 0$. Thus the present controller is optimal and the minimum cost is $\|G_cB_1\|_2$. Finally, for uniqueness it can be shown (an exercise) that the unique solution of $T_{vw} = 0$ is the controller above. □

8.7 Problems

1. Take $G(s) = 1/(s+1)$ and compute the $\mathcal{H}^2$-norm $\|G\|_2$ by the three methods: time-domain, state-space, residue theorem

2. Give an interesting (i.e., nontrivial) example of a 2×1 allpass matrix.

3. Consider

   $$\dot{x} = Ax + Bu, \quad x(0) = x_0$$

   with A stable. True or false: For every u in $\mathcal{L}^2[0, \infty)$, $x(t)$ tends to 0 as t tends to ∞.

4. Suppose u and y are scalar-valued signals and the transfer function from u to y is $1/s^2$. For the standard canonical realization (A, B, C) consider the optimization problem

   $$\min_{u=Fx} \int_0^\infty \rho y(t)^2 + u(t)^2 \, dt,$$

   where ρ is positive. Find the optimal F. Study the eigenvalues of $A + BF$ as $\rho \to 0$ and as $\rho \to \infty$.

5. Prove that U is allpass and $U^\sim G_c \in \mathcal{H}^{2^\perp}$.

6. Prove uniqueness in Theorem 8.5.

7. You know that right half-plane zeros place definite performance limitations on the control of a system. This exercise illustrates this fact in the present context.

   Consider the system

   $$\dot{x} = Ax + Bu, \quad x(0) = x_0$$

$$z = Cx.$$

Then

$$Z(s) = C(sI - A)^{-1}x_0 + C(sI - A)^{-1}BU(s).$$

If A is stable, we might like to see how small we can make $\|Z\|_2$ by suitable choice of stable $U(s)$. In particular, we might like to know if $\|Z\|_2$ can be made arbitrarily small.

Let

$$C(sI - A)^{-1}B = \frac{s-1}{(s+2)(s+3)}.$$

Compute

$$\inf_{U \text{ stable}} \|C(sI - A)^{-1}x_0 + C(sI - A)^{-1}Bu\|_2$$

as a function of x_0. For what values of x_0 is the infimum equal to zero.

Repeat for

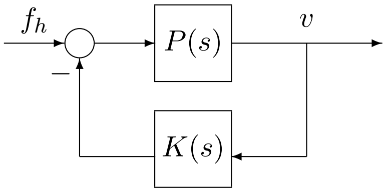
$$C(sI - A)^{-1}B = \frac{s+1}{(s+2)(s+3)}.$$

8. This problem concerns optimization in the space $\mathbb{R}^2$ with respect to three norms:

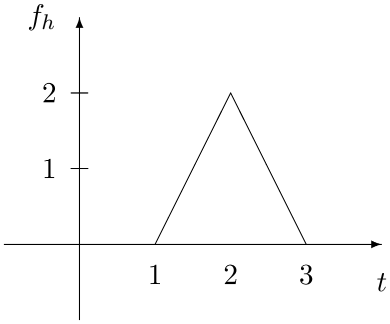
$$\|x\|_1 = |x_1| + |x_2|$$
$$\|x\|_2 = \left(x_1^2 + x_2^2\right)^{1/2}$$
$$\|x\|_\infty = \max\{|x_1|, |x_2|\}$$

Let $\mathcal{V}$ denote the one-dimensional subspace spanned by the vector $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$. For each of the three norms, find the vector in $\mathcal{V}$ that is closest to $\begin{bmatrix} 0 \\ 5 \end{bmatrix}$.

9. Consider the feedback system



Both $P(s)$ and $K(s)$ are SISO transfer functions. The plant is $P(s) = 1/s$ and the human force input f_h is as follows:



The output is a velocity v. It is desired to design a proper transfer function $K(s)$ to achieve internal stability of the feedback system and minimize the compliance error $\|f_h - v\|_2$. Set this up as a problem in $\mathcal{H}^2$ optimal control.

10. This problem relates to the LQR problem and whether or not $J < \infty$ implies $u \in L^2$. Show that it is true if R is positive definite. Hint: You have to show that if $R^{1/2}u$ is in L^2, then u is too.

11. Consider

$$A = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Regard A as the linear map $\mathbb{R}^2 \longrightarrow \mathbb{R}^2$ defined by $x \mapsto Ax$. Let $\mathcal{V}$ be one of the two 1-dimensional invariant subspaces and let V be the linear map $\mathcal{V} \longrightarrow \mathbb{R}^2$ given by $x \mapsto x$. Find the linear map $A_1 : \mathcal{V} \longrightarrow \mathcal{V}$ that satisfies the equation $VA_1 = AV$. This map is called the restriction of A to $\mathcal{V}$.

12. Take the LQR problem with $A = B = Q = R = 1$. Form the Hamiltonian matrix H given in Theorem 6.2. Find its invariant subspaces. Are there any of the form

$$\mathrm{Im} \begin{bmatrix} I \\ P \end{bmatrix}?$$

(We saw that the LQR problem reduces to looking for an invariant subspace of this form.)

13. Consider the LQR problem with

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad Q = \begin{bmatrix} q_1 & 0 \\ 0 & q_2 \end{bmatrix} > 0, \quad R = r > 0.$$

Let λ_1, λ_2 denote the eigenvalues of $A + BF$ for the optimal F. Of course, the two eigenvalues have to be in the left half-plane and be complex conjugates if not real. Are they otherwise freely assignable by choice of q_1, q_2, r?

Chapter 9

$\mathcal{H}^\infty$ Optimal Control

The symbol $\mathcal{H}^\infty$ stands for the space of all stable, proper transfer functions, such as

$$\frac{1}{s+1}, \quad \frac{2s-1}{s^2+5s+2}, \quad \frac{s}{s+1}$$

but not

$$\frac{1}{s}.$$

There's a natural norm, namely,

$$\|G\|_\infty = \sup_\omega |G(j\omega)|,$$

but no inner product. Thus $\mathcal{H}^\infty$ is a Banach space. The space extends to matrices, as we'll see.

In this chapter we study optimal control design in this space. This chapter begins with a tutorial overview, followed by some of the underlying theory.

9.1 Overview

Let R be a complex $p \times m$ matrix. The *singular values* of R are defined as the square roots of the eigenvalues of R^*R. The maximum singular value of R, denoted $\sigma_{\max}(R)$, has the properties required of a norm and is our *second definition* for $\|R\|$.

Example

The singular values of

$$R = \left[\begin{array}{cc} 2+j & j \\ 1-j & 3-2j \end{array} \right]$$

equal 4.2505, 1.7128. These are computed via the function *svd* in MATLAB. Thus $\|R\| = 4.2505$.

The importance of this second definition of matrix norm is derived from the following fact. Let $u \in \mathbb{C}^m$ and let $y = Ru$, so $y \in \mathbb{C}^p$. The fact is that

$$\sigma_{\max}(R) = \max\{\|y\| : \|u\| = 1\}.$$

This has the interpretation that if we think of R as a system with input u and output y, then $\sigma_{\max}(R)$ equals the system's gain, that is, maximum output norm over all inputs of unit norm.

Now we can define the $\mathcal{H}^\infty$ norm of a stable $p \times m$ transfer matrix $G(s)$:

$$\|G\|_\infty = \sup_\omega \sigma_{\max}\left[G(j\omega)\right]$$

So here we used the second-definition norm of $G(j\omega)$. If $G(s)$ is scalar-valued, its norm equals the peak magnitude on the Bode plot.

Concerning this definition is an important input-output fact. Let G be a stable, causal, LTI system with input u of dimension m and output y of dimension p. The norm $\mathcal{H}^\infty$-norm of the transfer matrix G is related to the maximum $\mathcal{L}^2$-norm of the output over all inputs of unit norm.

Theorem 9.1 $\|G\|_\infty = \sup\{\|y\|_2 : \|u\|_2 = 1\}$

Thus the major distinction between $\|G\|_2$ and $\|G\|_\infty$ is that the former is an average system gain for known inputs, while the latter is a worst-case system gain for unknown inputs.

It is useful to be able to compute $\|G\|_\infty$ by state-space methods. Let

$$G(s) = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right],$$

with A stable, that is, all eigenvalues with negative real part. The computation of $\|G\|_\infty$ using state-space methods involves the Hamiltonian matrix

$$H = \left[\begin{array}{cc} A + B(\gamma^2 - D^TD)^{-1}D^TC & \gamma B(\gamma^2 - D^TD)^{-1}D^T \\ -\gamma C^T(\gamma^2 - DD^T)^{-1}C & -[A + B(\gamma^2 - D^TD)^{-1}D^TC]^T \end{array}\right],$$

where γ is a positive number. The matrices $\gamma^2 - DD^T$, $\gamma^2 - D^TD$ are invertible provided they are positive definite, equivalently, γ^2 is greater than the largest eigenvalue of DD^T (or D^TD), equivalently, $\gamma > \sigma_{max}(D)$.

Theorem 9.2 *Let γ_{max} denote the maximum γ such that H has an eigenvalue on the imaginary axis. Then $\|G\|_\infty = \max\{\sigma_{max}(D), \gamma_{max}\}$.*

The theorem suggests the following procedure: Plot, versus γ, the distance from the imaginary axis to the nearest eigenvalue of H; then γ_{max} equals the maximum γ for which the distance equals zero; then $\|G\|_\infty = \max\{\sigma_{max}(D), \gamma_{max}\}$. A more efficient procedure is to compute γ_{max} by a bisection search.

The $\mathcal{H}^\infty$-optimal control problem is to compute an internally stabilizing controller K that minimizes $\|T_{zw}\|_\infty$ for the standard setup. This problem is much harder than the $\mathcal{H}^2$ problem. Instead of seeking a controller that actually minimizes $\|T_{zw}\|_\infty$, a simpler problem is to search for a controller that gives $\|T_{zw}\|_\infty < \gamma$, where γ is a pre-specified parameter. If γ is too small, a controller will not exist, so we need a test for existence. With this, the following procedure leads to a controller that is close to optimal:

1. Start with a large enough γ so that a controller exists.

2. Test existence for smaller and smaller values of γ until eventually γ is close to the minimum γ for existence.

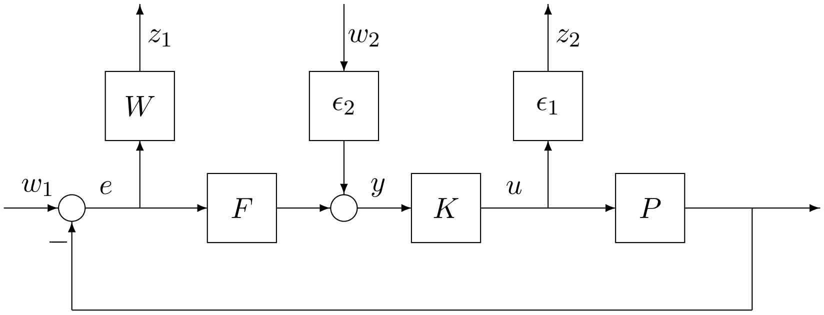
3. Compute a controller so that $\|T_{zw}\|_\infty < \gamma$.

A bisection search can be used.

The MATLAB command *hinfsyn* performs this procedure. The regularity assumptions required are (A1)-(A3), but not (A4), on page 100. The following example illustrates how a typical frequency-domain design problem can be formulated as one of $\mathcal{H}^\infty$-optimization.

Example

The next figure shows a single-loop analog feedback system.



The plant is P and the controller K; F is an antialiasing filter for future digital implementation of the controller (it is a good idea to include F at the start of the analog design so that there are no surprises later due to additional phase lag). The basic control specification is to get good tracking over a certain frequency range, say $[0, \omega_1]$; that is, to make the magnitude of the transfer function from w_1 to e small over this frequency range. The weighted tracking error is z_1 in the figure, where the weight W is selected to be a lowpass filter with bandwidth ω_1. We could attempt to minimize the $\mathcal{H}^\infty$-norm from w_1 to z_1, but this problem is not regular. To regularize it, another input, w_2, is added and another signal, z_2, is penalized. The two weights ϵ_1 and ϵ_2 are small positive scalars. The design problem is to minimize the $\mathcal{H}^\infty$-norm

$$\text{from } w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \text{ to } z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

The preceding figure can then be converted to the standard block diagram by stacking the states of P, F, and W to form the state of G.

The plant transfer function is taken to be

$$P(s) = \frac{20 - s}{(s + 0.01)(20 + s)}.$$

This can be regarded as an approximation of the time-delay system $\frac{1}{s}\mathrm{e}^{-40s}$, an integrator cascaded with a time delay of 40 time units. With a view toward subsequent digital control with sampling period $h = 0.5$, the filter F is taken to have bandwidth $\pi/0.5$, the Nyquist frequency ω_N:

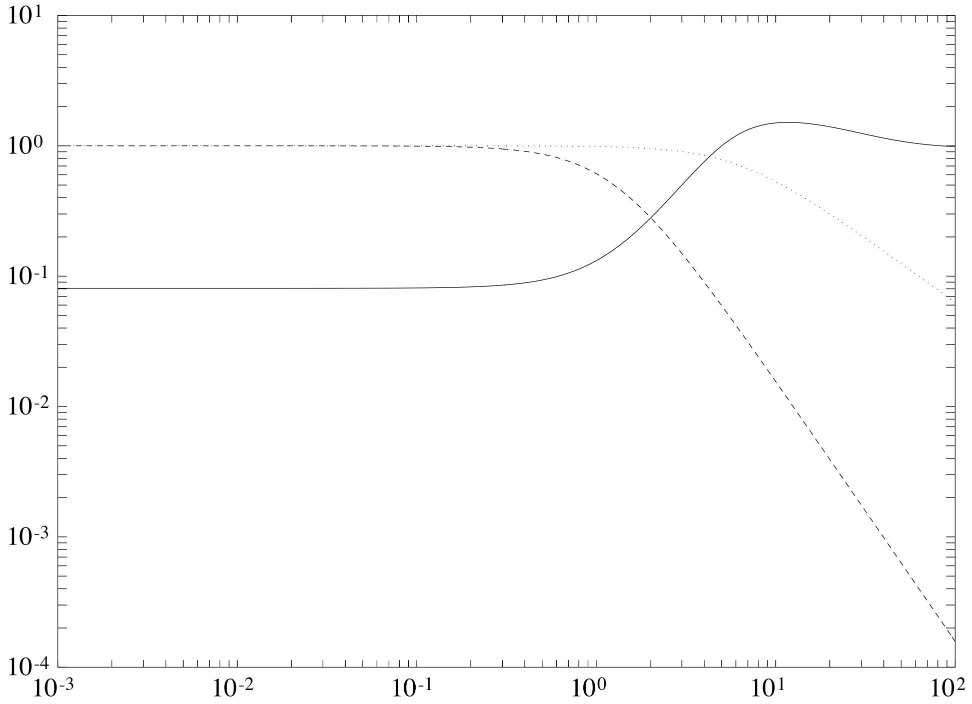
$$F(s) = \frac{1}{(0.5/\pi)s + 1}.$$

The weight W is then taken to have bandwidth one-fifth the Nyquist frequency:

$$W(s) = \left[\frac{1}{(2.5/\pi)s + 1}\right]^2.$$

Finally, ϵ_1 and ϵ_2 are both set to 0.01.

The next figure shows the results of the design using *hinfsyn*:



The solid curve is the Bode magnitude plot of the *sensitivity function*, that is, the transfer function from w_1 to e, namely, $1/(1+PKF)$. Also shown are the magnitude plots for W (dash) and F (dot). Evidently, the design has achieved some tracking error attenuation over the bandwidth of W. A greater degree of attenuation could be achieved by tuning the weights W, ϵ_1, and ϵ_2.

The MATLAB code:

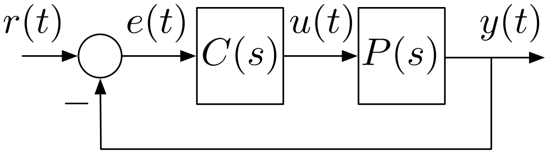
```
% H_inf analog design example with discretization

% input data

clear

% parameters

h=0.021;
z=20;
[AP,BP,CP,DP]=tf2ss([-1 z],conv([1 .01],[1 z]));
[AF,BF,CF,DF]=tf2ss(1,[0.5/pi 1]);
numW=1;
```

```
denW=conv([5/(2*pi) 1],[5/(2*pi) 1]);
[AW,BW,CW,DW]=tf2ss(numW,denW);
[nP,mP]=size(BP);
[nF,mF]=size(BF);
[nW,mW]=size(BW);

eps1=0.01;
eps2=0.01;

% build G

A=[AP zeros(nP,nF) zeros(nP,nW);
-BF*CP AF zeros(nF,nW);
-BW*CP zeros(nW,nF) AW];
B1=[0*BP 0*BP;BF 0*BF;BW 0*BW];
B2=[BP;0*BF;0*BW];
C1=[-DW*CP zeros(1,nF) CW;zeros(1,nP+nF+nW)];
C2=[zeros(1,nP) CF zeros(1,nW)];
D11=[DW 0;0 0];
D12=[0;eps1];
D21=[0 eps2];
D22=0;

% design

p=pck(A,[B1 B2],[C1;C2],[D11 D12;D21 D22]);
[k,g,gfin,ax,ay,hamx,hamy]=hinfsyn(p,1,1,0,.2,.01);
[AK,BK,CK,DK]=unpck(k);


% generate closed-loop analog system S

[nK,mK]=size(BK);
AS=[AP BP*CK zeros(nP,nF);zeros(nK,nP) AK BK*CF;-BF*CP zeros(nF,nK) AF];
BS=[0*BP;0*BK;BF];
CS=[-CP 0*CK 0*CF];
DS=1;


% discretize K

[AK,BK]=c2d(AK,BK,h);

% stability check

Atmp=[AP BP*CF;zeros(nF,nP) AF];
```

```
Btmp=[0*BP;BF];
Ctmp=[0*CP CF];

[Atmp,Btmp]=c2d(Atmp,Btmp,h);
Abar=[Atmp Btmp*CK;-BK*Ctmp AK];
max(abs(eig(Abar)));


% analysis


w=logspace(-3,2,200);
j=sqrt(-1);
p=freqrc(AP,BP,CP,DP,w);
f=freqrc(AF,BF,CF,DF,w);
k=dfreqrc(AK,BK,CK,DK,w,h);
r=(1-exp(-j*h*w))./(j*h*w);
tmp=ones(1,length(w))./(1+p.*r.*k.*f);
magS2=abs(tmp);



[magS1,ph]=bode(AS,BS,CS,DS,1,w);
[magW,ph]=bode(AW,BW,CW,DW,1,w);
[magF,ph]=bode(AF,BF,CF,DF,1,w);
loglog(w,magS1,w,magS2)
%loglog(w,magS1,w,magW,w,magF)
```

9.2 A Simple Feedback Loop

The previous example wasn't so complicated, but we need an even simpler one in order to isolate the Nehari problem that lies at the heart of the solution . Consider the block diagram



with the plant transfer function

$$P(s) = \frac{1}{2s+1}\mathrm{e}^{-s}.$$

It's BIBO stable, but it has a time delay, which makes it hard to control. For simplicity, let us rationalize $P(s)$ via a Padé approximation:

$$P(s) = \frac{1-0.5s}{(2s+1)(1+0.5s)}.$$

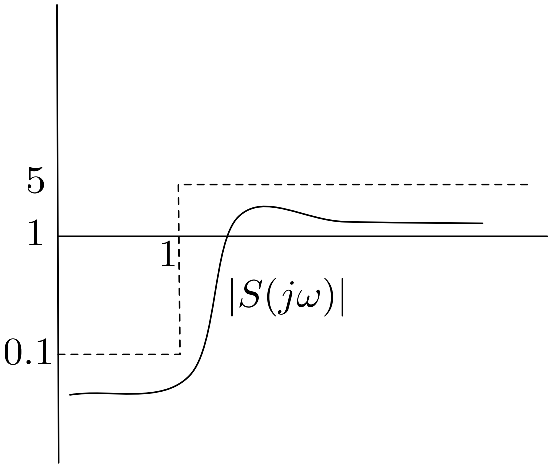
We want to design a controller $C(s)$ so that the feedback loop is stable and also has some measure of stability robustness. A good way to do this is to require the Nyquist plot of PC to stay outside the circle centred at -1 and radius, say, 0.2. Let S denote the transfer function from r to e (known as the sensitivity function). It turns out that $\|S\|_\infty^{-1}$ equals the distance in the complex plane from the critical point -1 to the closest point on the Nyquist plot of PC. (Reference: ECE356 course notes.) Thus stability robustness is equivalent to the inequality

$$|S(j\omega)| \le 5, \quad \forall \omega.$$

Suppose that, in addition to stability robustness, we want to design the controller $C(s)$ so that the system tracks signals $r(t)$ up to, say, 1 rad/s. Thus we want, say,

$$|S(j\omega)| \le 0.1, \quad \forall \omega \le 1.$$

This allows at most 10 percent tracking error for sinusoidal reference signals. Therefore we want the magnitude Bode plot of S to lie under the dashed line:



To handle these two specs together it is convenient to construct a weighting function $W(s)$ such that $|W(j\omega)| \approx 10$ over the frequency range $[0,1]$ and $|W(j\omega)| \approx 0.2$ over the frequency range $[1,\infty)$. Then the two specs become one: $\|WS\|_\infty \le 1$. For computational reasons, we want $W(s)$ to be rational. So its magnitude can't be discontinuous, and we need some transition from magnitude 10 to magnitude 0.2. To keep things simple, let's try the weighting function

$$W(s) = \gamma \frac{\alpha s + 1}{\beta s + 1}.$$

For $|W| = 10$ at $\omega = 0$, we take $\gamma = 10$. Then for $|W| = 0.2$ at $\omega = \infty$, we need

$$10 \frac{\alpha}{\beta} = 0.2.$$

Finally, for $|W| = 1$ at $\omega = 1$, we need

$$10 \left| \frac{\alpha j + 1}{\beta j + 1} \right| = 1.$$

These conditions give

$$\alpha = \sqrt{99/300} = 0.574, \quad \beta = 50\alpha = 28.7.$$

To recap, we have arrived at the problem of designing a controller $C(s)$ that stabilizes $P(s)$ and achieves the inequality $\|WS\|_\infty \le 1$, where

$$W(s) = 10\frac{0.574s+1}{28.7s+1}.$$

The problem may not be solvable. That is, there may not be any stabilizing controller such that $\|WS\|_\infty \le 1$. If so, we have to compromise somehow; relax either the tracking error or the stability margin or both. This is a great feature of this way of doing control design: We can sensibly make tradeoffs.

The obvious problem at hand is to minimize $\|WS\|_\infty$ over all $C(s)$ that stabilize $P(s)$. If this minimum is less than or equal to 1, our specifications are feasible. However,

$$WS = \frac{W}{1+PC}$$

is a nonlinear function of $C(s)$, and moreover $C(s)$ is constrained to stabilize. We need to change the optimization parameter. Notice that the $P(s)$ in our example is strictly proper and belongs to $\mathcal{H}^\infty$.

Lemma 9.1 *A proper rational controller $C(s)$ stabilizes a strictly proper $P \in \mathcal{H}^\infty$ iff it has the form*

$$C = \frac{Q}{1-PQ}, \quad Q \in \mathcal{H}^\infty.$$

Proof (Necessity) Suppose C stabilizes. Let Q equal the transfer function from r to u:

$$Q = \frac{C}{1+PC}.$$

Solve for C to get $C = Q/(1-PQ)$.

(Sufficiency) Suppose C is given by the formula in the lemma. Then all closed-loop transfer functions belong to $\mathcal{H}^\infty$. For example, the transfer function from r to y equals

$$\frac{PC}{1+PC} = PQ.$$

Also, the sensitivity function S equals $1-PQ$. And so on for all other closed-loop transfer functions. □

The lemma changes the problem of minimizing $\|WS\|_\infty$ over C to the problem

$$\min_{Q\in\mathcal{H}^\infty} \|W(1-PQ)\|_\infty.$$

This is rather better because $W(1-PQ)$ is an affine function of Q. Let us write

$$P = P_1P_2, \quad P_1(s) = \frac{1-0.5s}{1+0.5s}, \quad P_2(s) = \frac{1}{2s+1}.$$

Notice that $|P_1(j\omega)| = 1$ for all ω. Therefore for every Q in $\mathcal{H}^\infty$

$$\|W(1-PQ)\|_\infty = \|WP_1(P_1^{-1} - P_2Q)\|_\infty = \|W(P_1^{-1} - P_2Q)\|_\infty = \|WP_1^{-1} - WP_2Q\|_\infty.$$

Let $\mathcal{L}^\infty(j\mathbb{R})$ denote the space of proper transfer functions that have no poles on the imaginary axis. Then $F := WP_1^{-1}$ belongs to $\mathcal{L}^\infty(j\mathbb{R})$, while WP_2Q belongs to $\mathcal{H}^\infty$. Thus the minimum of $\|WS\|_\infty$ over all stabilizing controllers seems to be very close to the distance from F to $\mathcal{H}^\infty$. The gap arises from the fact that the set

$$\{WP_2Q : Q \in \mathcal{H}^\infty\}$$

is a proper subset of $\mathcal{H}^\infty$ because WP_2 is strictly proper. That is, if X is the function in $\mathcal{H}^\infty$ that is closest to F and if we get Q from $WP_2Q = X$, then Q may not be proper. This can be rectified by a high-frequency correction.

9.3 The Nehari Problem

So we are given a function $R(s)$ in $\mathcal{L}^\infty(j\mathbb{R})$ and we want to find $X(s)$ in $\mathcal{H}^\infty$ that minimizes the norm

$$\|R - X\|_\infty := \sup_\omega |R(j\omega) - X(j\omega)|.$$

That is, we want to find an X in $\mathcal{H}^\infty$ that is closest to R in the infinity norm. There are two ways to view this problem: 1) R is unstable and X has to be stable, while both are causal; 2) R is noncausal and X has to be causal, while both are stable.

The second way turns out to be more useful. To view the problem in this way, we have to suppose R and X are two-sided Laplace transforms, e.g.,

$$R(s) = \int_{-\infty}^{\infty} r(t)\mathrm{e}^{-st}dt.$$

The region of convergence is taken to include the imaginary axis, so that the underlying system is stable. Thus for $R(s)$ the ROC must be Re $s < 1$. Therefore the time-domain equation giving rise to $R(s)$ must be

$$y(t) = \int_{-\infty}^{\infty} r(t-\tau)u(\tau)d\tau.$$

The time-domain operator

$$\Lambda_r : \ u \mapsto r * u, \quad \mathcal{L}^2(\mathbb{R}) \longrightarrow \mathcal{L}^2(\mathbb{R}),$$

called the **Laurent operator derived from** r, is equivalent to the frequency-domain operator

$$U \mapsto RU, \ \ \mathcal{L}^2(j\mathbb{R}) \longrightarrow \mathcal{L}^2(j\mathbb{R})$$

in the sense that they have equal induced norms, since the Fourier transform is norm-preserving, by Theorem **??**. The norm of the latter operator equals $\|R\|_\infty = 1$.

Likewise for X: For $X(s)$ the ROC must include the imaginary axis. The time-domain equation giving rise to $X(s)$ must be

$$y(t) = \int_{-\infty}^{\infty} x(t-\tau)u(\tau)d\tau.$$

The Laurent operator

$$\Lambda_x : u \mapsto x * u, \quad \mathcal{L}^2(\mathbb{R}) \longrightarrow \mathcal{L}^2(\mathbb{R})$$

is equivalent to the frequency-domain operator

$$U \mapsto XU : \ \mathcal{L}^2(j\mathbb{R}) \longrightarrow \mathcal{L}^2(j\mathbb{R}).$$

The norm of the latter operator equals $\|X\|_\infty$.

The difference between the two systems is that Λ_x is causal while Λ_r is not. Let us extract the "noncausal" part of Λ_r by taking the input to start at time 0 and looking at the output only before then:

$$y(t) = \int_0^{\infty} r(t-\tau)u(\tau)d\tau, \quad t \leq 0.$$

This operator, $\mathcal{L}^2[0,\infty) \longrightarrow \mathcal{L}^2(-\infty,0]$, is called the **Hankel operator derived from** r, denoted Γ_r. It maps the future into the past. On the other hand, since the system with transfer function X is causal, its Hankel operator Γ_x equals 0.

Notice that a Hankel operator is a piece of a Laurent operator. Thus $\|\Lambda_r\| \geq \|\Gamma_r\|$. Notice also that

$$\|R - X\|_\infty = \|\Lambda_r - \Lambda_x\|.$$

From these two facts, we get

$$\|R - X\|_\infty = \|\Lambda_r - \Lambda_x\| \geq \|\Gamma_r - \Gamma_x\| = \|\Gamma_r\|.$$

Our original problem was to minimize $\|R - X\|_\infty$. We've seen that a lower bounded for this norm is $\|\Gamma_r\|$. However, Nehari's theorem says the lower bound is tight:

Theorem 9.3 *The distance from R in $\mathcal{L}^\infty(j\mathbb{R})$ to $\mathcal{H}_\infty$ equals $\|\Gamma_r\|$. Moreover, the distance is achieved (there is an optimal X).*

So it remains to compute the norm $\|\Gamma_r\|$ and then to compute the optimal X.

9.4 Hankel Operators

We may as well suppose $R(s)$ is strictly proper, rational, with all poles in Re $s < 0$. Then it has a state a state model

$$R(s) = C(sI - A)^{-1}B,$$

where A is antistable (all eigenvalues in Re $s > 0$). Suppose A is $n \times n$. Such R belongs to $\mathcal{L}^\infty(j\mathbb{R})$. The inverse two-sided Laplace transform of $R(s)$ is

$$r(t) = \begin{cases} -Ce^{At}B, & t < 0 \\ 0, & t \geq 0 \end{cases}$$

The Hankel operator Γ_r maps a function u in $\mathcal{L}^2[0,\infty)$ to the function y in $\mathcal{L}^2(-\infty,0]$ defined by

$$y(t) = \int_0^\infty r(t-\tau)u(\tau)d\tau, \quad t < 0,$$

that is,

$$y(t) = -Ce^{At}\int_0^\infty e^{-A\tau}Bu(\tau)d\tau, \quad t < 0.$$

Define two auxiliary operators: the *controllability operator*

$$\Psi_c : \mathcal{L}^2[0,\infty) \to \mathbb{C}^n, \quad \Psi_c u := -\int_0^\infty e^{-A\tau}Bu(\tau)d\tau$$

and the *observability operator*

$$\Psi_o : \mathbb{C}^n \to \mathcal{L}^2(-\infty,0], \quad (\Psi_o x)(t) := Ce^{At}x, \quad t < 0.$$

Then

$$\Gamma_r = \Psi_o\Psi_c.$$

Thus we have the diagram

$$\begin{tikzcd}
\mathcal{L}^2(\mathbb{R}) \arrow[rr, "\overset{\text{Laurent}}{\Lambda_r}"] & & \mathcal{L}^2(\mathbb{R}) \arrow[d] \\
\mathcal{L}^2[0,\infty) \arrow[u] \arrow[rr, "\Gamma_r", "\text{Hankel}"'] \arrow[dr, "\Psi_c"'] & & \mathcal{L}^2(-\infty,0] \\
& \mathbb{C}^n \arrow[ur, "\Psi_o"'] &
\end{tikzcd}$$

Since $\|\Gamma_r\| = \|\Psi_o\Psi_c\|$, it remains to compute the latter norm.

The self-adjoint operators $\Psi_c\Psi_c^*$ and $\Psi_o^*\Psi_o$ map $\mathbb{C}^n$ to itself. Thus they have matrix representations with respect to the standard basis on $\mathbb{C}^n$. Define the *controllability* and *observability gramians*

$$L_c := \int_0^\infty \mathrm{e}^{-At}BB^T\mathrm{e}^{-A^Tt}dt \tag{9.1}$$

$$L_o := \int_0^\infty \mathrm{e}^{-A^Tt}C^TC\mathrm{e}^{-At}dt \tag{9.2}$$

It is routine to show that L_c and L_o are the unique solutions of the Lyapunov equations

$$AL_c + L_cA^T = BB^T \tag{9.3}$$

$$A^TL_o + L_oA = C^TC \tag{9.4}$$

We state without proof this fact: The norm of $\Psi_o\Psi_c$ equals the square root of the norm of $(\Psi_o\Psi_c)^*(\Psi_o\Psi_c)$, and this in turn equals the largest eigenvalue of the matrix L_cL_o.

Example Let's complete the example from the first section. We have

$$F(s) = 10\frac{0.574s+1}{28.7s+1}\frac{1+0.5s}{1-0.5s} = \frac{0.736}{1-0.5s} + (\text{a function in } \mathcal{H}^\infty),$$

and so

$$F(s) = R(s) + (\text{a function in } \mathcal{H}^\infty),$$

where

$$R(s) = -\frac{1.47}{s-2}.$$

A state model for $R(s)$ is

$$A = 2, \quad B = 1, \quad C = -1.47.$$

The Lyapunov equations (9.3), (9.4) yield

$$L_c = 1/4, \quad L_o = 0.541.$$

Thus

$$\|\Gamma_r\| = \sqrt{L_cL_o} = 0.368.$$

Thus the distance from F to $\mathcal{H}^\infty$ equals 0.368. Our design specs are therefore easily feasible.

We omit the construction of X and a controller that meets the specs. MATLAB has tools to design controllers based on the approach in this chapter.

9.5 Problems

1. Let $U(s) = s/(s+1)$. Suppose $G \in \mathcal{H}^\infty$ and we want to approximate it by UV for some $V \in \mathcal{H}^\infty$, that is, we want to minimize $\|G - UV\|_\infty$. In general we can't take $V = G/U$ because G/U has a pole at $s = 0$ unless it's cancelled by a zero of G, so G/U is not in $\mathcal{H}^\infty$. Thus in general the error norm $\|G - UV\|_\infty$ can be made only arbitrarily small, and not zero, by suitable choice of V.

   (a) Write in proper logic notation (using $\forall$ and $\exists$ where appropriate) the mathematical statement of this: "Let G belong to $\mathcal{H}^\infty$. Then the norm $\|G - UV\|_\infty$ can be made arbitrarily small by suitable choice of V in $\mathcal{H}^\infty$." In this statement U is given and fixed and should not be quantified.

   (b) Write in proper logic notation the negation of your logic statement in part (a).

   (c) Convert the preceding logic statement into a natural sounding sentence or sentences in words.

   (d) Write in proper logic notation the mathematical statement of this: "In general, $\|G - UV\|_\infty$ cannot be made equal to zero."

2. In the scalar-valued case prove that $\mathcal{RL}^2$ equals the set of all real-rational functions that are strictly proper and have no poles on the imaginary axis.

3. Show that Ψ_c is surjective if (A, B) is controllable and that Ψ_o is injective if (C, A) is observable.

4. Show that the adjoints of Ψ_c and Ψ_o are as follows:

$$\Psi_c^* : \mathbb{C}^n \to \mathcal{L}^2[0, \infty)$$

$$(\Psi_c^* x)(t) = -B^T \mathrm{e}^{-A^T t} x, \quad t \geq 0$$

$$\Psi_o^* : \mathcal{L}^2(-\infty, 0] \to \mathbb{C}^n$$

$$\Psi_o^* y = \int_{-\infty}^{0} \mathrm{e}^{A^T t} C^T y(t) dt.$$

5. Prove that the matrix representations of $\Psi_c \Psi_c^*$ and $\Psi_o^* \Psi_o$ are L_c and L_o respectively.

Epilogue

So where do we stand now in 2010? Let's review and try to draw some conclusions.

1. The three classical topics, calculus of variations, the maximum principle, dynamic programming, were included for historical interest.

   The brachistochrone problem is beautiful, isn't it? Find a curve that optimizes a scalar quantity, the time to slide down.

   The maximum principle is very general, and can include many kinds of constraints. However, for many problems I don't think a practical solution has been provided by the necessary condition. Try some problem harder than time-optimal control of the double integrator. For example, try a cart-pendulum system with the problem of swinging up the pendulum in minimum time. The state space is $\mathbb{R}^4$ and so the switching set is a 3D hypersurface. It's hard to compute this hypersurface, and then how are you going to implement the controller?

   Dynamic programming is indeed very powerful and the HJB equations are very important and have played, and continue to play, an important role in optimal control.

2. I love the function space method. The reason is that it seems perfectly suited to systems theory. A system is a function that maps an input to an output, that is, a system is a mapping from one set to another. So right from the get-go one is into block diagrams, spaces of signals, and operators. This is, in my view, the cleanest and clearest way to formulate a problem. The subsystems may have differential equation models, but those are just special ways of modeling maps.

3. The $\mathcal{H}^2$ and $\mathcal{H}^\infty$ optimization methods are widely used in control design.

   ... (I haven't finished this.)