Additions to Linear Algebra

Peter Petersen

September 26, 2012

Abstract

In this document we've added corrections as well as included several sections that expand upon the material in the text.

1 Corrections

This is were typos will be listed.

59_{10} Should read $M = \left\{ (\alpha_1, ..., \alpha_n) \in \mathbb{F}^n : \alpha_{j_1} = \cdots = \alpha_{j_{n-k}} = 0 \right\}$

103_2 Should read $\ker(L) = \operatorname{im}(L')^o$

Hint for Exercise 2.6.12.b. Many people seem to think that this problem can only be done using quotient spaces. Here are a few hints towards a solution that does not use quotient spaces. First observe that $\chi_L = \mu_L$, see also Exercise 2.6.7. Let $M \subset V$ be an L-invariant subspace. Let $p = \mu_{L|_M}$ and factor $\chi_L = \mu_L = p \cdot q$. Show that $M \subset \ker(p(L))$. If $M \neq \ker(p(L))$ select a complement $V = \ker(p(L)) \oplus N$ and consider the corresponding block decomposition

$$L = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}$$

where A corresponds to the restriction of L to $\ker(p(L))$. Let r be the characteristic polynomial for C. Show that L is a root of $p \cdot r$ by showing that $r(L)() \subset \ker(p(L))$. Show that $\mu_L = p \cdot r$ and reach a contradiction.

Ignore Exercise 3.3.14

2 Additional Exercises

Exercise 23 gives a beautiful effective algorithm for the Jordan-Chevalley decomposition for linear operators over any field of characteristic 0.

1. Show directly that an upper triangular matrix

$$A = \begin{bmatrix} \alpha_{11} & * & & * \\ 0 & \alpha_{22} & * & \\ \vdots & & \ddots & * \\ 0 & 0 & \cdots & \alpha_{nn} \end{bmatrix}$$

is a root of its characteristic polynomial.

2. Show that a linear operator on a finite dimensional complex vector space admits a basis so that it's matrix representation is upper triangular. Hint: Decompose the vector space in to a direction sum of an eigenspace and a complement and use induction on dimension.

3. Let $L : V \to V$ be a linear operator, where V is not necessarily finite dimensional. If $p \in \mathbb{F}[t]$ has a factorization $p = p_1 \cdots p_k$ where the factors p_i are pairwise relative prime, then

$$\ker(p(L)) = \ker(p_1(L)) \oplus \cdots \oplus \ker(p_k(L))$$

4. Hint: Start with $k = 2$. The use induction on k and that p_k is relatively prime to $p_1 \cdots p_{k-1}$.

5. Show that if a linear operator on a finite dimensional vector space is irreducible, i.e., it has no nontrivial invariant subspaces, then its minimal polynomial is irreducible.

6. Show that if a linear operator on a finite dimensional vector space is indecomposable, i.e., the vector space cannot be written as a direct sum of nontrivial subspaces, then the minimal polynomial is a power of an irreducible polynomial.

7. Assume that $L : V \to V$ has minimal polynomial $m_L(t) = (t-1)^3(t-2)$ and $\chi_L(t) = (t-1)^3(t-2)^3$. Find the Jordan canonical form for L.

8. Assume that $L : V \to V$ has minimal polynomial $m_L(t) = (t-1)^3(t-2)$ and $\chi_L(t) = (t-1)^4(t-2)^3$. Find the Jordan canonical form for L.

9. Find the Jordan canonical form for the following matrices

   (a) $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 8 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -16 & 0 & 0 & 0 \end{bmatrix}$

   (b) $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & -2 & 0 & 2 \end{bmatrix}$

   (c) $\begin{bmatrix} 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 0 \end{bmatrix}$

10. Find the Jordan canonical form for the following matrices

(a) $\begin{bmatrix} 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

(b) $\begin{bmatrix} 0 & \frac{1}{2} & -1 & -3 \\ 0 & 0 & 0 & -2 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

(c) $\begin{bmatrix} -1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2 & 2 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$

11. Find the Jordan canonical form and also a Jordan basis for $D = \frac{d}{dt}$ on each of the following subspaces defined as kernels.

    (a) $\ker\left((D-1)^2(D+1)^2\right)$.

    (b) $\ker\left((D-1)^3(D+1)\right)$.

    (c) $\ker\left(D^2+2D+1\right)$.

12. Find the Jordan canonical form on P_3 for each of the following operators.

    (a) $L = T \circ D$, where $T(f)(t) = tf(t)$.

    (b) $L = D \circ T$.

    (c) $L = T \circ D^2 + 3D + 1$.

13. For $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{C}$ decide which of the matrices are similar (the answer depends on how the λs are related to each other)

$$\begin{bmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_2 & 1 \\ 0 & 0 & \lambda_3 \end{bmatrix}, \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}, \begin{bmatrix} \lambda_1 & 1 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix},$$
$$\begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 1 \\ 0 & 0 & \lambda_3 \end{bmatrix}, \begin{bmatrix} \lambda_1 & 0 & 1 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

14. For each n give examples of $n \times n$ matrices that are similar but not unitarily equivalent.

15. Let $L : V \to V$ be a linear operator with

$$\begin{aligned} \chi_L(t) &= (t-\lambda_1)^{n_1} \cdots (t-\lambda_k)^{n_k}, \\ m_L(t) &= (t-\lambda_1)^{m_1} \cdots (t-\lambda_k)^{m_k}. \end{aligned}$$

If $m_i = 1$ or $n_i - m_i \leq 1$ for each $i = 1, ..., k$, then the Jordan canonical form is completely determined by χ_L and m_L. (Note that for some i we might have $m_i = 1$, while for other j the second condition $n_j - m_j \leq 1$ will hold.)

16. Let $L : \mathbb{R}^2 \to \mathbb{R}^2$ be given by $\begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix}$ with respect to the standard basis. Find the rational canonical form and the basis that yields that form.

17. Let $A \in \text{Mat}_{n \times n}(\mathbb{R})$ satisfy $A^2 = -1_{\mathbb{R}^n}$. Find the rational canonical form for A.

18. Find the real rational canonical forms for the differentiation operator

$$D : C^{\infty}(\mathbb{R}, \mathbb{R}) \to C^{\infty}(\mathbb{R}, \mathbb{R})$$

on each of the following kernels of real functions.

(a) $\ker\left(\left(D^2 + 1\right)^2\right)$.

(b) $\ker\left(\left(D^2 + D + 1\right)^2\right)$.

19. Let $L : V \to V$ be a linear operator.

(a) If $m_L(t) = p(t)$ and p is irreducible, then L is semi-simple, i.e., completely reducible, i.e., every invaraint subspace has an invariant complement. Hint: Use that

$$\begin{aligned} V &= C_{x_1} \oplus \cdots \oplus C_{x_k}, \\ \chi_{L|_{C_{x_i}}}(t) &= m_{L|_{C_{x_i}}}(t) = p(t) \end{aligned}$$

where C_{x_i} has no nontrivial invariant subspaces.

(b) If $m_L(t) = p_1(t) \cdots p_k(t)$, where p_1, ..., p_k are distinct irreducible polynomials, then L is semi-simple. Hint: Show that if $M \subset V$ is L invariant then

$$M = (M \cap \ker(p_1(L))) \oplus \cdots \oplus (M \cap \ker(p_k(L))).$$

20. Assume that $\mathbb{F} \subset \mathbb{L}$, e.g., $\mathbb{R} \subset \mathbb{C}$. Let $A \in \text{Mat}_{n \times n}(\mathbb{F})$. Show that $A : \mathbb{F}^n \to \mathbb{F}^n$ is semi-simple if and only if $A : \mathbb{L}^n \to \mathbb{L}^n$ is semi-simple.

21. *(The generalized Jordan Canonical Form)* Let $L : V \to V$ be a linear operator on a finite dimensional vector space V.

(a) Assume that

$$m_L(t) = (p(t))^m = \chi_L(t),$$

where $p(t)$ is irreducible in $\mathbb{F}[t]$. Show that if $V = C_x$, then

$$e_{ij} = (p(L))^{i-1} (L)^{j-1} (x),$$

where $i = 1, ..., m$ and $j = 1, ..., \deg(p)$ form a basis for V. Hint: It suffices to show that they span V.

(b) With the assumptions as in a. and $k = \deg(p)$ show that if we order the basis as follows

$$e_{m1}, ..., e_{mk}, e_{m-1,1}, ..., e_{m-1,k}, ..., e_{11}, ..., e_{1k}$$

then the matrix representation looks like

$$\begin{bmatrix} C_p & E & \cdots & 0 \\ 0 & C_p & \ddots & \vdots \\ \vdots & & \ddots & E \\ 0 & \cdots & 0 & C_p \end{bmatrix},$$

$$E = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix}$$

where the companion matrix C_p appears on the diagonal, the E matrices right above the diagonal and all other entries are zero.

(c) Explain how a. and b. lead to a generalized Jordan canonical form for any $L : V \to V$.

(d) *(The Jordan-Chevalley decomposition)* Let

$$m_L(t) = (p_1(t))^{m_1} \cdots (p_k(t))^{m_k}$$

be the factorization of the minimal polynomial into distinct irreducible factors. Using the previous exercises show that $L = S + N$, where S is semi-simple with $m_S(t) = p_1(t) \cdots p_k(t)$, N nilpotent, $S = p(L)$, and $N = q(L)$ for suitable polynomials p and q. For a different proof that creates an effective algorithm see the next couple of exercises.

22. Let $p \in \mathbb{F}[t]$. We show how to construct a separable polynomial that has the same roots as p in the algebraic closure, i.e., a polynomial without repeated roots in the algebraic closure.

(a) Show that $\left\{ q \in \mathbb{F}[t] : p \mid q^k \text{ for some } k \geq 1 \right\}$ is is an ideal and therefore generated by a unique monic polynomial s_p.

(b) Show that $s_p \mid p$.

(c) Show that if $q^2 \mid s_p$ then q is a constant.

(d) Show that if $\mathbb{F}$ has characteristic 0, then

$$s_p = \frac{p}{\gcd\{p, Dp\}}.$$

23. Let $L : V \to V$ be a linear operator on a finite dimensional vector space. Let μ be its minimal polynomial and $s = s_\mu$ the corresponding separable polynomial, and s' its derivative. The goal is to show that the Jordan-Chevalley decomposition $L = S + N$ can be computed via an effective algorithm. We know that S has to be semi-simple so it is natural to look for solutions to $s(S) = 0$. This suggests that we seek S via Newton's method

$$\begin{aligned} L_{k+1} &= L_k - (s'(L_k))^{-1} s(L_k) \\ L_0 &= L \end{aligned}$$

where $(s')^{-1}(t) = q(t)$ is interpreted as a polynomial we get from $qs' + ps = 1$, i.e., the inverse modulo s.

(a) Show that such a q exists and can be computed. Hint: use the previous exercise.

(b) Show that

$$L - L_{k+1} = \sum_{i=0}^{k} q(L_i)\, s(L_i)$$

(c) Show that $L - L_k$ is nilpotent for all k.

(d) Use Taylor's formula for polynomials

$$f(t+h) = f(t) + f'(t)\, h + h^2 g(t,h)$$

to conclude that there is a polynomial g such that

$$s(L_{k+1}) = (s(L_k))^2\, g(L_k).$$

(e) Finally let m be the smallest integer so that $\mu \mid s^m$ and show that L_k is semi-simple provided $2^k \geq m$.

(f) Conclude that with these choices we obtain a Jordan-Chevalley decomposition

$$L = L_k + L - L_k = S + N$$

where there are suitable polynomial $p, r \in \mathbb{F}[t]$ such that $S = p(L)$ and $N = r(L)$.

24. Use the previous exercise to show that any invertible $L : V \to V$, where V is finite dimensional can be written as

$$L = SU$$

where S is the same semi-simple operator as in the Jordan-Chevalley decomposition, and U is unipotent, i.e., $U - 1_V$ is nilpotent. Show that $U = q(L)$ for some polynomial q.

3 Linear Algebra in Multivariable Calculus

Linear maps play a big role in multivariable calculus and are used in a number of ways to clarify and understand certain constructions. The fact that linear algebra is the basis for multivariable calculus should not be surprising as linear algebra is merely a generalization of vector algebra.

Let $F : \Omega \to \mathbb{R}^n$ be a differentiable function defined on some open domain $\Omega \subset \mathbb{R}^m$. The differential of F at $x_0 \in \Omega$ is a linear map $DF_{x_0} : \mathbb{R}^m \to \mathbb{R}^n$ that can be defined via the limiting process

$$DF_{x_0}(h) = \lim_{t\to 0} \frac{F(x_0 + th) - F(x_0)}{t}.$$

Note that $x_0 + th$ describes a line parametrized by t passing through x_0 and points in the direction of h. This definition tells us that DF_{x_0} preserves scalar multiplication as

$$\begin{aligned} DF_{x_0}(\alpha h) &= \lim_{t\to 0} \frac{F(x_0 + t\alpha h) - F(x_0)}{t} \\ &= \alpha \lim_{t\to 0} \frac{F(x_0 + t\alpha h) - F(x_0)}{t\alpha} \\ &= \alpha \lim_{t\alpha\to 0} \frac{F(x_0 + t\alpha h) - F(x_0)}{t\alpha} \\ &= \alpha \lim_{s\to 0} \frac{F(x_0 + sh) - F(x_0)}{s} \\ &= \alpha DF_{x_0}(h). \end{aligned}$$

Additivity is another matter however. Thus one usually defines F to be differentiable at x_0 provided we can find a linear map $L : \mathbb{R}^m \to \mathbb{R}^n$ satisfying

$$\lim_{|h|\to 0} \frac{|F(x_0 + h) - F(x_0) - L(h)|}{|h|} = 0$$

One then proves that such a linear map must be unique and then renames it $L = DF_{x_0}$. If F is continuously differentiable, i.e. all of its partial derivatives exist and are continuous, then DF_{x_0} is also given by the $n \times m$ matrix of partial derivatives

$$\begin{aligned} DF_{x_0}(h) &= DF_{x_0}\left(\begin{bmatrix} h_1 \\ \vdots \\ h_m \end{bmatrix}\right) \\ &= \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_n}{\partial x_1} & \cdots & \frac{\partial F_n}{\partial x_m} \end{bmatrix} \begin{bmatrix} h_1 \\ \vdots \\ h_m \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial F_1}{\partial x_1} h_1 + \cdots + \frac{\partial F_1}{\partial x_m} h_m \\ \vdots \\ \frac{\partial F_n}{\partial x_1} h_1 + \cdots + \frac{\partial F_n}{\partial x_m} h_m \end{bmatrix} \end{aligned}$$

One of the main ideas in differential calculus (of several variables) is that linear maps are simpler to work with and that they give good local approximations to differentiable maps. This can be made more precise by observing that we have the *first order approximation*

$$\begin{aligned} F\left(x_0+h\right) &= F\left(x_0\right)+DF_{x_0}\left(h\right)+o\left(h\right), \\ \lim_{\left|h\right|\to 0}\frac{\left|o\left(h\right)\right|}{\left|h\right|} &= 0 \end{aligned}$$

One of the goals of differential calculus is to exploit knowledge of the linear map DF_{x_0} and then use this first order approximation to get a better understanding of the map F itself.

In case $f:\Omega\to\mathbb{R}$ is a function one often sees the differential of f defined as the expression

$$df=\frac{\partial f}{\partial x_1}dx_1+\cdots+\frac{\partial f}{\partial x_m}dx_m.$$

Having now interpreted dx_i as a linear function we then observe that df itself is a linear function whose matrix description is given by

$$\begin{aligned} df\left(h\right) &= \frac{\partial f}{\partial x_1}dx_1\left(h\right)+\cdots+\frac{\partial f}{\partial x_m}dx_m\left(h\right) \\ &= \frac{\partial f}{\partial x_1}h_1+\cdots+\frac{\partial f}{\partial x_m}h_m \\ &= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_m} \end{bmatrix}\begin{bmatrix} h_1 \\ \vdots \\ h_m \end{bmatrix}. \end{aligned}$$

More generally, if we write

$$F=\begin{bmatrix} F_1 \\ \vdots \\ F_n \end{bmatrix},$$

then

$$DF_{x_0}=\begin{bmatrix} dF_1 \\ \vdots \\ dF_n \end{bmatrix}$$

with the understanding that

$$DF_{x_0}\left(h\right)=\begin{bmatrix} dF_1\left(h\right) \\ \vdots \\ dF_n\left(h\right) \end{bmatrix}.$$

Note how this conforms nicely with the above matrix representation of the differential.

As we shall see in this section many of the things we have learned about linear algebra can be used to great effect in multivariable calculus. We are going to study the behavior of smooth vector functions $F : \Omega \to \mathbb{R}^n$, where $\Omega \subset \mathbb{R}^m$ is an open domain. The word smooth is somewhat vague but means that functions will always be at least continuously differentiable, i.e., $(x_0, h) \to DF_{x_0}(h)$ is continuous. The main idea is simply that a smooth function F is approximated via the differential near any point x_0 in the following way

$$F(x_0 + h) \simeq F(z_0) + DF_{x_0}(h).$$

Since the problem of understanding the linear map $h \to DF_{x_0}(h)$ is much simpler and this map also approximates F for small h; the hope is that we can get some information about F in a neighborhood of x_0 through such an investigation.

The graph of $G : \Omega \to \mathbb{R}^n$ is defined as the set

$$\operatorname{Graph}(G) = \{(x, G(x)) \in \mathbb{R}^m \times \mathbb{R}^n : x \in \Omega\}.$$

We picture it as an m-dimensional curved object. Note that the projection $P : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ when restricted to $\operatorname{Graph}(G)$ is one-to-one. This is the key to the fact that the subset $\operatorname{Graph}(G) \subset \mathbb{R}^m \times \mathbb{R}^n$ is the graph of a function from some subset of $\mathbb{R}^m$.

More generally suppose we have some curved set $S \subset \mathbb{R}^{m+n}$ (S stands for surface). Loosely speaking, such a set is has dimension m if near every point $z \in S$ we can decompose the ambient space $\mathbb{R}^{m+n} = \mathbb{R}^m \times \mathbb{R}^n$ in such a way that the projection $P : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m$ when restricted to S, i.e., $P|_S : S \to \mathbb{R}^m$ is one-to-one near z. Thus S can near z be viewed as a graph by considering the function $G : U \to \mathbb{R}^n$, defined via $P(x, G(x)) = x$. The set $U \subset \mathbb{R}^m$ is some small open set where the inverse to $P|_S$ exists. Note that, unlike the case of a graph, the $\mathbb{R}^m$ factor of $\mathbb{R}^{m+n}$ does not have to consist of the first m coordinates in $\mathbb{R}^{m+n}$, nor does it always have to be the same coordinates for all z. We say that S is a *smooth m-dimensional surface* if near every z we can choose the decomposition $\mathbb{R}^{m+n} = \mathbb{R}^m \times \mathbb{R}^n$ so that the graph functions G are smooth.

Example 3.1. Let $S = \left\{z \in \mathbb{R}^{m+1} : |z| = 1\right\}$ be the unit sphere. This is an m-dimensional smooth surface. To see this fix $z_0 \in S$. Since $z_0 = (\alpha_1, ..., \alpha_{n+1}) \neq 0$, there will be some i so that $\alpha_i \neq 0$ for all z near z_0. Then we decompose $\mathbb{R}^{m+1} = \mathbb{R}^m \times \mathbb{R}$ so that $\mathbb{R}$ records the i^{th} coordinate and $\mathbb{R}^m$ the rest. Now consider the equation for S written out in coordinates $z = (\xi_1, ..., \xi_{n+1})$

$$\xi_1^2 + \cdots + \xi_i^2 + \cdots + \xi_{n+1}^2 = 1,$$

and solve it for ξ_i in terms of the rest of the coordinates

$$\xi_i = \pm\sqrt{1 - \left(\xi_1^2 + \cdots + \widehat{\xi_i^2} + \cdots + \xi_{n+1}^2\right)}.$$

Depending on the sign of α_i we can choose the sign in the formula to write S near z_0 as a graph over some small subset in $\mathbb{R}^m$. What is more, since $\alpha_i \neq 0$

we have that $\xi_1^2+\cdots+\widehat{\xi_i^2}+\cdots+\xi_{n+1}^2<1$ for all $z=(\xi_1,...,\xi_{n+1})$ near z_0. Thus the function is smooth near $(\alpha_1,...,\widehat{\alpha_i},...,\alpha_{n+1})$.

The Implicit Function Theorem gives us a more general approach to decide when surfaces defined using equations are smooth.

Theorem 3.2. (The Implicit Function Theorem) *Let $F:\mathbb{R}^{m+n}\to\mathbb{R}^n$ be smooth. If $F(z_0)=c\in\mathbb{R}^n$ and* $\operatorname{rank}(DF_{z_0})=n$, *then we can find a coordinate decomposition $\mathbb{R}^{m+n}=\mathbb{R}^m\times\mathbb{R}^n$ near z_0 such that the set $S=\{z\in\mathbb{R}^{m+n}:F(z)=c\}$ is a smooth graph over some open set $U\subset\mathbb{R}^m$.*

Proof. We are not going to give a complete proof this theorem here, but we can say a few things that might elucidate matters a little. It is convenient to assume $c=0$, this can always be achieved by changing F to $F-c$ if necessary. Note that this doesn't change the differential.

First let us consider the simple situation where F is linear. Then $DF=F$ and so we are simply stating that F has rank n. This means that $\ker(F)$ is m-dimensional. Thus we can find a coordinate decomposition $\mathbb{R}^{m+n}=\mathbb{R}^m\times\mathbb{R}^n$ such that the projection $P:\mathbb{R}^{m+n}=\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}^m$ is an isomorphism when restricted to $\ker(F)$. Therefore, we have an inverse L to $P|_{\ker(F)}$ that maps $L:\mathbb{R}^m\to\ker(F)\subset\mathbb{R}^{m+n}$. In this way we have exhibited $\ker(F)$ as a graph over $\mathbb{R}^m$. Since $\ker(F)$ is precisely the set where $F=0$ we have therefore solved our problem.

In the general situation we use that $F(z_0+h)\simeq DF_{z_0}(h)$ for small h. This indicates that it is natural to suppose that near z_0 the sets S and $\{z_0+h:h\in\ker(DF_{z_0})\}$ are very good approximations to each other. In fact the picture we have in mind is that $\{z_0+h:h\in\ker(DF_{z_0})\}$ is the *tangent space* to S at z_0. The linear map $DF_{z_0}:\mathbb{R}^{m+n}\to\mathbb{R}^n$ evidently is assumed to have rank n and hence nullity m. We can therefore find a decomposition $\mathbb{R}^{m+n}=\mathbb{R}^m\times\mathbb{R}^n$ such that the projection $P:\mathbb{R}^{m+n}\to\mathbb{R}^m$ is an isomorphism when restricted to $\ker(DF_{z_0})$. This means that the tangent space to S at z_0 is m-dimensional and a graph.

It is not hard to believe that a similar result should be true for S itself near z_0. The actual proof can be given using a Newton iteration. In fact if $z_0=(x_0,y_0)\in\mathbb{R}^m\times\mathbb{R}^n$ and $x\in\mathbb{R}^m$ is near x_0, then we find $y=y(x)\in\mathbb{R}^n$ as a solution to $F(x,y)=0$. This is done iteratively by successively solving infinitely many linear systems. We start by using the approximate guess that y is y_0. In order to correct this guess we find the vector $y_1\in\mathbb{R}^n$ that solves the linear equation that best approximates the equation $F(x,y_1)=0$ near (x,y_0), i.e.,

$$F(x,y_1)\simeq F(x,y_0)+DF_{(x,y_0)}(y_1-y_0)=0.$$

The assumption guarantees that $DF_{(x_0,y_0)}|_{\mathbb{R}^n}:\mathbb{R}^n\to\mathbb{R}^n$ is invertible. Since we also assumed that $(x,y)\to DF_{(x,y)}$ is continuous this means that $DF_{(x,y_0)}|_{\mathbb{R}^n}$ will also be invertible as long as x is close to x_0. With this we get the formula:

$$y_1=y_0-\left(DF_{(x,y_0)}|_{\mathbb{R}^n}\right)^{-1}(F(x,y_0)).$$

Repeating this procedure gives us an iteration

$$y_{n+1} = y_n - \left(DF_{(x,y_n)}|_{\mathbb{R}^n}\right)^{-1}\left(F\left(x, y_n\right)\right),$$

that starts at y_0.

It is slightly nasty that we have to keep inverting the map $DF_{(x,y_n)}|_{\mathbb{R}^n}$ as y_n changes. It turns out that one is allowed to always use the approximate differential $DF_{(x_0,y_0)}|_{\mathbb{R}^n}$. This gives us the much simpler iteration

$$y_{n+1} = y_n - \left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(F\left(x, y_n\right)\right).$$

It remains to show that the sequence $\left(y_n\right)_{n\in\mathbb{N}_0}$ converges and that the correspondence $x \to y\left(x\right)$ thus defined, gives a smooth function that solves $F\left(x, y\left(x\right)\right) = 0$. Note, however, that if $y_n \to y\left(x\right)$, then we have

$$\begin{aligned}
y\left(x\right) &= \lim_{n\to\infty} y_{n+1} \\
&= \lim_{n\to\infty}\left(y_n - \left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(F\left(x, y_n\right)\right)\right) \\
&= \lim_{n\to\infty} y_n - \lim_{n\to\infty}\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(F\left(x, y_n\right)\right) \\
&= y\left(x\right) - \left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(F\left(x, \lim_{n\to\infty} y_n\right)\right) \\
&= y\left(x\right) - \left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(F\left(x, y\left(x\right)\right)\right).
\end{aligned}$$

Thus $\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(F\left(x, y\left(x\right)\right)\right) = 0$ and hence $F\left(x, y\left(x\right)\right) = 0$ as desired. The convergence of $\left(y_n\right)_{n\in\mathbb{N}_0}$ hinges on the completeness of real numbers but can otherwise be handled when we have introduced norms. Continuity requires some knowledge of uniform convergence of functions. Smoothness can be checked using continuity of $x \to y\left(x\right)$ and smoothness of F. □

The Implicit Function Theorem gives us the perfect criterion for deciding when solutions to equations give us nice surfaces.

Corollary 3.3. *Let $F : \mathbb{R}^{m+n} \to \mathbb{R}^n$ be smooth and define*

$$S_c = \left\{z \in \mathbb{R}^{m+n} : F\left(z\right) = c\right\}.$$

If rank $\left(DF_z\right) = n$ *for all $z \in S$, then S is a smooth m-dimensional surface.*

Note that $F : \mathbb{R}^{m+n} \to \mathbb{R}^n$ is a collection of n functions $F_1, ..., F_n$. If we write $c = \left(c_1, ..., c_n\right)$ we see that the set S_c is the intersection of the sets $S_{c_i} = \left\{z \in \mathbb{R}^{m+n} : F_i\left(z\right) = c_i\right\}$. We can apply the above corollary to each of these sets and see that they form $m+n-1$ dimensional surfaces provided $DF_i = dF_i$ always has rank 1 on S_{c_i}. This is quite easy to check since this simply means that dF_i is never zero. Each of the linear functions dF_i at some specified point $z \in \mathbb{R}^{m+n}$ can be represented as $1 \times \left(m+n\right)$ row matrices via the partial derivatives for F_i. Thus they lie in a natural vector space and when stacked on top of each other yield the matrix for DF. The rank condition on DF for ensuring that

S_c is a smooth m-dimensional surface on the other hand is a condition on the columns of DF. Now matrices do satisfy the magical condition of having equal row and column rank. Thus DF has rank n if and only if it has row rank n. The latter statement is in turn equivalent to saying that $dF_1, ..., dF_n$ are linearly independent or equivalently span an n-dimensional subspace of $\mathrm{Mat}_{1\times(n+m)}$.

Recall that we say that a function $f : \mathbb{R}^m \to \mathbb{R}$, has a *critical point* at $x_0 \in \mathbb{R}^m$ if $df_{x_0} = 0$. One reason why these points are important lies in the fact that extrema, i.e., local maxima and minima, are critical points. To see this note that if x_0 is a local maximum for f, then

$$f(x_0 + h) \leq f(x_0),$$

for small h. Since

$$df_{x_0}(h) = \lim_{t\to 0} \frac{f(x_0 + th) - f(x_0)}{t},$$

we have that

$$df_{x_0}(h) \leq 0,$$

for all h! This is not possible unless $df_{x_0} = 0$. Note that the level sets $S_c = \{x : f(x) = c\}$ must have the property that either they contain a critical point or they are $(n-1)$-dimensional smooth surfaces.

To make things more interesting let us see what happens when we restrict or constrain a function $f : \mathbb{R}^{m+n} \to \mathbb{R}$ to a smooth surface $S_c = \{z : F(z) = c\}$. Having extrema certainly makes sense so let us see what happens if we assume that $f(z) \leq f(z_0)$ for all $z \in S_c$ near z_0. Note that this is not as simple as the unconstrained situation. To simplify the situation let us assume that we have decomposed $\mathbb{R}^{m+n} = \mathbb{R}^m \times \mathbb{R}^n$ (and coordinates are written $z = (x, y) \in \mathbb{R}^m \times \mathbb{R}^n$) near z_0 and written S_c as a graph of $G : U \to \mathbb{R}^n$, where $U \subset \mathbb{R}^m$. Then $f : S_c \to \mathbb{R}$ can near z_0 be thought of as simply $g(x) = f(x, G(x)) : U \to \mathbb{R}$. So if $f|_{S_c}$ has a local maximum at z_0, then g will have a local maximum at x_0. Since the maximum for g is unconstrained we then conclude $dg_{x_0} = 0$. Using the chain rule on $g(x) = f(x, G(x))$, this leads us to

$$\begin{aligned} 0 &= dg_{x_0}(h) \\ &= df_{z_0}(h, DG_{x_0}(h)). \end{aligned}$$

Note that the vectors $(h, DG_{x_0}(h))$ are precisely the tangent vectors to the graph of G at $(x_0, y_0) = z_0$. We see that the relationship $F(x, G(x)) = 0$ when differentiated gives $DF_{z_0}(h, DG(h)) = 0$. Thus $\ker(DF_{z_0}) = \{(h, DG_{x_0}(h)), h \in \mathbb{R}^n\}$. This means that if we define $z_0 \in S_c$ to be critical for $f|_{S_c}$ when df_{z_0} vanishes on $\ker(DF_{z_0})$, then we have a definition which again guarantees that local extrema are critical. Since it can be nasty to calculate $\ker(DF_{z_0})$ and check that df_{z_0} vanishes on the kernel we seek a different condition for when this happens. Recall that each of $dF_1, ..., dF_n$ vanish on $\ker(DF_{z_0})$, moreover as we saw these linear maps are linearly independent. We also know that the dimension of the space of linear maps $\mathbb{R}^{m+n} \to \mathbb{R}$ that vanish on the m-dimensional space $\ker(DF_{z_0})$ must have dimension n. Thus $dF_1, ..., dF_n$ form a basis for this space. This means that df_{z_0} vanishes on $\ker(DF_{z_0})$ if and only if we can find $\lambda_1, ..., \lambda_n \in \mathbb{R}$ such that

$$df_{z_0} = \lambda_1 dF_1|_{z_0} + \cdots + \lambda_n dF_n|_{z_0}.$$

Using λs for the numbers $\lambda_1, ..., \lambda_n$ is traditional, they are called *Lagrange multipliers.*

Note that we have completely ignored the boundary of the domain Ω and also boundaries of the smooth surfaces. This is mostly so as not to complicate matters more than necessary. While it is not possible to ignore the boundary of domains when discussing optimization, it is possible to do so when dealing with smooth surfaces. Look, e.g., at the sphere as a smooth surface. The crucial fact that the sphere shares with other "closed" smooth surfaces is that it is compact without having boundary. What we are interested in gaining in the use of such surfaces is the guarantee that continuous functions must have a maximum and a minimum.

Another important question in multivariable calculus is when a smooth function can be inverted and still remain smooth. An obvious condition is that it be bijective, but a quick look at $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$ shows that this isn't enough. Assume for a minute that $F : \Omega \to \mathbb{R}^n$ has an inverse $G : F(\Omega) \to \Omega \subset \mathbb{R}^m$ that is also smooth. Then we have $G \circ F(x) = x$ and $F \circ G(y) = y$. Taking derivatives and using the chain rule tells us

$$\begin{aligned} DG_{F(x)} \circ DF_x &= 1_{\mathbb{R}^m}, \\ DF_{G(y)} \circ DF_x &= 1_{\mathbb{R}^n}. \end{aligned}$$

This means that the differentials themselves are isomorphisms and that $n = m$. It turns us that this is precisely the correct condition for ensuring smoothness of the inverse.

Theorem 3.4. (The Inverse Function Theorem) *Let $F : \Omega \to \mathbb{R}^m$ be smooth and assume that we have $x_0 \in \Omega$ where DF_{x_0} is an isomorphism. Then we can find neighborhoods U of x_0 and V of $F(x_0)$ such that $F : U \to V$ is a bijection, that has a smooth inverse $G : V \to U$.*

Corollary 3.5. *Let $F : \Omega \to \mathbb{R}^m$ be smooth and assume that F is one-to-one and that DF_x is an isomorphism for all $x \in \Omega$, then $F(\Omega) \subset \mathbb{R}^m$ is an open domain and there is a smooth inverse $G : F(\Omega) \to \Omega$.*

It is not hard to see that the Inverse Function Theorem follows from the Implicit Function Theorem and vice versa. Note that, when $m = 1$, having nonzero derivative is enough to ensure that the function is bijective as it must be strictly monotone. When $m \geq 2$, this is no longer true as can be seen from $F : \mathbb{C} \to \mathbb{C} - \{0\}$ defined by $F(z) = e^z$. As a two variable function it can also be represented by $F(\alpha, \beta) = e^{\alpha}(\cos\beta, \sin\beta)$. This function maps onto the punctured plane, but all choices $\beta \pm n2\pi, n \in \mathbb{N}_0$ yield the same values for F. The differential is represented by the matrix

$$DF = e^{\alpha} \begin{bmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{bmatrix},$$

that has an inverse given by

$$e^{-\alpha}\left[\begin{array}{cc}\cos\beta & \sin\beta\\ -\sin\beta & \cos\beta\end{array}\right].$$

So the map is locally, but not globally invertible.

Linearization procedures can be invoked in trying to understand several other nonlinear problems. As an example one can analyze the behavior of a fixed point x_0 for $F:\mathbb{R}^n\to\mathbb{R}^n$, i.e., $F(x_0)=x_0$, using the differential DF_{x_0} since we know that $F(x_0+h)\simeq x_0+DF_{x_0}(h)$.

3.1 Exercises

1. We say that $F:\Omega\to\mathbb{R}$ depends functionally on a collection of functions $F_1,...,F_m:\Omega\to\mathbb{R}$ near $x_0\in\Omega$ if $F=\Theta(F_1,...,F_m)$ near x_0 for some function Θ. We say that $F_1,...,F_m:\Omega\to\mathbb{R}$ near $x_0\in\Omega$ are functionally independent if none of the functions are functionally dependent on the rest near x_0.

   (a) Show that if $dF_1|_{x_0},...,dF_m|_{x_0}$ are linearly independent as linear functionals, then $F_1,...,F_m$ are also functionally independent near x_0.

   (b) Assume that $\Omega\subset\mathbb{R}^n$ and $m>n$. Show that, if $\operatorname{span}\{dF_1|_{x_0},...,dF_m|_{x_0}\}$ has dimension n, then we can find $F_{i_1},...,F_{i_n}$ such that all the other functions $F_{j_1},...,F_{j_{m-n}}$ depend functionally on $F_{i_1},...,F_{i_n}$ near x_0.

4 Norms

Before embarking on the richer theory of inner products we wish to cover the more general notion of a *norm*. A norm on a vector space is simply a way of assigning a length or size to each vector. We are going to confine ourselves to the study of vector spaces where the scalars are either real or complex. If V is a vector space, then a norm is a function $\|\cdot\|:V\to[0,\infty)$ that satisfies

1. If $\|x\|=0$, then $x=0$.
2. The scaling condition: $\|\alpha x\|=|\alpha|\,\|x\|$, where α is either a real or complex scalar.
3. The Triangle Inequality: $\|x+y\|\leq\|x\|+\|y\|$.

The first condition just says that the only vector of norm zero is the zero vector. The second condition on scaling conforms to our picture of how the length of a vector changes as we scale it. When we allow complex scalars we note that multiplication by i does not change the size of the vector. Finally the third and truly crucial condition states the fact that in any triangle the sum of two sides is always longer than the third. We can see this by letting three vectors x,y,z be the vertices of the triangle and agreeing that the three numbers

$\|x-z\|$, $\|x-y\|$, $\|y-z\|$ measure the distance between the vertices, i.e., the side lengths. The triangle inequality now says

$$\|x-z\| \leq \|x-y\| + \|y-z\|.$$

An important alternative version of the triangle inequality is the inequality

$$|\|x\| - \|y\|| \leq \|x-y\|.$$

This is obtained by noting that $\|x-y\| = \|y-x\|$ and

$$\begin{aligned} \|x\| &\leq \|y\| + \|x-y\|, \\ \|y\| &\leq \|x\| + \|y-x\|. \end{aligned}$$

There are a plethora of interesting norms on the vector spaces we have considered so far. We shall not establish the three axioms for the norms defined. It is, however, worth pointing out that while the first two properties are usually easy to establish, the triangle inequality can be very tricky to prove.

Example 4.1. The most basic example is $\mathbb{R}^n$ or $\mathbb{C}^n$ with the euclidean norm

$$\|x\|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2}.$$

This norm evidently comes from the inner product via $||x||_2^2 = (x|x)$. The subscript will be explained in the next example.

We stick to $\mathbb{R}^n$ or $\mathbb{C}^n$ and define two new norms

$$\begin{aligned} \|x\|_1 &= |x_1| + \cdots + |x_n|, \\ \|x\|_\infty &= \max\{|x_1|, ..., |x_n|\}. \end{aligned}$$

Note that

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq n\|x\|_\infty.$$

More generally for $p \geq 1$ we have the p-norm

$$\|x\|_p = \sqrt[p]{|x_1|^p + \cdots + |x_n|^p}.$$

If $p \leq q$ we have

$$\|x\|_\infty \leq \|x\|_q \leq \|x\|_p \leq \sqrt[p]{n}\,\|x\|_\infty.$$

The trick that allows us to conclude that $\|x\|_q \leq \|x\|_p$ is by first noting that both norms have the scaling property. Thus it suffices to show the inequality when $\|x\|_q = 1$. This means that we need to show that

$$|x_1|^p + \cdots + |x_n|^p \geq 1$$

when

$$|x_1|^q + \cdots + |x_n|^q = 1.$$

In this case we know that $|x_i| \leq 1$. Thus

$$|x_i|^q \leq |x_i|^p$$

as $q > p$. This implies the inequality.

In addition,

$$\|x\|_p \leq \sqrt[p]{n}\,\|x\|_\infty$$

so

$$\lim_{p\to\infty} \|x\|_p = \|x\|_\infty .$$

This explains all of the subscripts for these norms and also how they relate to each other.

Of all these norms only the 2-norm comes from an inner product. The other norms can be quite convenient at times when one is studying analysis. The 2-norm and the ∞-norm will be used below to justify certain claims we made in the first and second chapter regarding differential equations and multivariable calculus. We shall also see that for linear operators there are two equally natural norm concepts, were only one comes from an inner product.

Example 4.2. The p-norm can be generalized to functions using integration rather than summation. We let $V = C^0\left([a,b],\mathbb{C}\right)$ and define

$$\|f\|_p = \left(\int_a^b |f(t)|^p\,dt\right)^{\frac{1}{p}}.$$

This time the relation between the norms is quite different. If $p \leq q$, then

$$\|f\|_p \leq (b-a)^{\frac{1}{p}-\frac{1}{q}}\,\|f\|_q,$$

or in a more memorable form using normalized integrals:

$$\begin{aligned}
(b-a)^{-\frac{1}{p}}\|f\|_p &= \left(\frac{1}{b-a}\int_a^b |f(t)|^p\,dt\right)^{\frac{1}{p}} \\
&\leq \left(\frac{1}{b-a}\int_a^b |f(t)|^q\,dt\right)^{\frac{1}{q}} \\
&= (b-a)^{-\frac{1}{q}}\|f\|_q .
\end{aligned}$$

Moreover,

$$\|f\|_\infty = \lim_{p\to\infty}\left(\frac{1}{b-a}\int_a^b |f(t)|^p\,dt\right)^{\frac{1}{p}}.$$

Here the ∞-norm is defined as

$$\|f\|_\infty = \sup_{t\in[a,b]} |f(t)| .$$

Assuming that f is continuous this supremum is a maximum, i.e., $|f(t)|$ has a maximum value that we define to be $\|f\|_{\infty}$. See also the next section for more on this ∞-norm.

Aside from measuring the size of vectors the norm is used to define *convergence* on vector spaces. We say that a sequence $x_n \in V$ converges to $x \in V$ with respect to the norm $\|\cdot\|$ if $\|x_n - x\| \to 0$ as $n \to \infty$. Clearly this concept depends on having a norm and might even take on different meanings depending on what norm we use. Note, however, that the norms we defined on $\mathbb{R}^n$ and $\mathbb{C}^n$ are related to each other via

$$\|\cdot\|_{\infty} \leq \|\cdot\|_p \leq \sqrt[p]{n}\, \|\cdot\|_{\infty} .$$

Thus convergence in the p-norm and convergence in the ∞-norm means the same thing. Hence all of these norms yield the same convergence concept.

For the norms on $C^0([a,b],\mathbb{C})$ a very different picture emerges. We know that

$$(b-a)^{-\frac{1}{p}} \|f\|_p \leq (b-a)^{-\frac{1}{q}} \|f\|_q \leq (b-a)^{-1} \|f\|_{\infty} .$$

Thus convergence in the ∞-norm or in the q-norm implies convergence in the p-norm for $p \leq q$. The converse is, however, not at all true.

Example 4.3. Let $[a,b] = [0,1]$ and define $f_n(t) = t^n$. We note that

$$\|f_n\|_p = \sqrt[p]{\frac{1}{np+1}} \to 0 \text{ as } n \to \infty.$$

Thus f_n converges to the zero function in all of the p-norms when $p < \infty$. On the other hand

$$\|f\|_{\infty} = 1$$

so f_n does not converge to the zero function, or indeed any continuous function, in the ∞-norm.

If V and W both have norms then we can also define a norm on $\operatorname{Hom}(V,W)$. This norm, known as the *operator norm*, is defined so that for $L : V \to W$ we have

$$\|L(x)\| \leq \|L\| \|x\| .$$

Using the scaling properties of the norm and linearity of L this is the same as saying

$$\left\| L\left(\frac{x}{\|x\|}\right) \right\| \leq \|L\| , \text{ for } x \neq 0.$$

Since $\left\| \frac{x}{\|x\|} \right\| = 1$, we can then define the operator norm as

$$\|L\| = \sup_{\|x\|=1} \|L(x)\| .$$

It might happen that this norm is infinite. We say that L is *bounded* if $\|L\| < \infty$ and *unbounded* if $\|L\| = \infty$. Note that bounded operators are continuous and that they form a subspace $\mathcal{B}(V,W) \subset \operatorname{Hom}(V,W)$ (see also exercises to this section). In the optional section "Completeness and Compactness" we shall show that linear maps on finite dimensional spaces are always bounded. In case the linear map is defined on a finite dimensional inner product space we give a completely elementary proof of this result in "Orthonormal Bases".

Example 4.4. Let $V = C^\infty([0,1],\mathbb{C})$. Differentiation $D : V \to V$ is unbounded if we use $\|\cdot\|_\infty$ on both spaces. This is because $x_n = t^n$ has norm 1, while $D(x_n) = nx_{n-1}$ has norm $n \to \infty$. If we used $\|\cdot\|_2$, things wouldn't be much better as

$$\begin{aligned} \|x_n\|_2 &= \sqrt{\frac{1}{2n+1}} \to 0, \\ \|Dx_n\|_2 &= n\|x_{n-1}\|_2 = n\sqrt{\frac{1}{2n-1}} \to \infty. \end{aligned}$$

If we try

$$\begin{aligned} M &: C^0([0,1],\mathbb{C}) \to C^0([0,1],\mathbb{C}), \\ S &: C^0([0,1],\mathbb{C}) \to C^0([0,1],\mathbb{C}), \end{aligned}$$

then things are much better as

$$\begin{aligned} \|M(x)\|_\infty &= \sup_{t\in[0,1]} t|x(t)| \\ &\leq \sup_{t\in[0,1]} |x(t)| \\ &= \|x\|_\infty, \\ \|S(x)\|_\infty &= \left\| \int_0^t x(s)\,ds \right\|_\infty \\ &\leq \|x\|_\infty. \end{aligned}$$

Thus both of these operators are bounded in the ∞-norm. It is equally easy to show that they are bounded with respect to all of the p-norms for $1 \leq p \leq \infty$.

4.1 Exercises

1. Let $\mathcal{B}(V,W) \subset \operatorname{Hom}(V,W)$ be the subset of bounded operators.
   (a) Show that $\mathcal{B}(V,W)$ is subspace of $\operatorname{Hom}(V,W)$.
   (b) Show that the operator norm defines a norm on $\mathcal{B}(V,W)$.
2. Show that a bounded linear map is continuous.

5 Completeness and Compactness

In this section we wish to discuss some further properties of norms and how they relate to convergence. This will primarily allow us to show that in the finite dimensional setting nothing nasty or new happens. However, it will also attempt to make the reader aware of certain problems in the infinite dimensional setting. Another goal is to reinforce the importance of the fundamental analysis concepts of compactness and completeness. Finally we shall show in one of the final sections of this chapter how these investigations can help us in solving some of the issues that came up in our earlier sections on differential equations and multivariable calculus.

A vector space with a norm is called a *normed vector space*. It often happens that the norm is not explicitly stated and we shall often just use the same generic symbol $\|\cdot\|$ for several different norms on different vector spaces.

Using norms we can define *continuity* for functions $f : V \to \mathbb{F}$ and more generally for maps $F : V \to W$ between normed vector spaces. The condition is that if $x_n \to x$ in V, then $F(x_n) \to F(x)$ in W.

Another important concept is that of *compactness*. A set $C \subset V$ in a normed vector space is said to be (sequentially) compact if every sequence $x_n \in C$ has a convergent subsequence x_{n_k} whose limit point is in C. It is a crucial property of $\mathbb{R}$ that all closed intervals $[a, b]$ are compact. In $\mathbb{C}$ the unit disc $\Delta = \{\zeta \in \mathbb{C} : |\zeta| \leq 1\}$ is compact. More generally products of these sets $[a, b]^n \subset \mathbb{R}^n$, $\Delta^n \subset \mathbb{C}^n$ are also compact if we use any of the equivalent p-norms. The boundaries of these sets are evidently also compact.

To see why $[0, 1]$ is compact select a sequence $x_n \in [0, 1]$. If we divide $[0, 1]$ into two equal parts $\left[0, \frac{1}{2}\right]$ and $\left[\frac{1}{2}, 1\right]$, then one of these intervals contains infinitely many elements from the sequence. Call this chosen interval I_1 and select an element $x_{n_1} \in I_1$ from the sequence. Next we divide I_1 in half and select a interval I_2 that contains infinitely many elements from the sequence. In this way we obtain a subsequence (x_{n_k}) such that all of the elements x_{n_k} belong to an interval I_k of length 2^{-k}, where $I_{k+1} \subset I_k$. The intersection $\cap_{k=1}^{\infty} I_k$ consists of a point. This is quite plausible if we think of real numbers as represented in binary notation, for then $\cap_{k=1}^{\infty} I_k$ indicates a binary number from the way we chose the intervals. Certainly $\cap_{k=1}^{\infty} I_k$ can't contain more than one point, because if $\alpha, \beta \in \cap_{k=1}^{\infty} I_k$, then also all numbers that lie between α and β lie in $\cap_{k=1}^{\infty} I_k$ as each I_k is an interval. The fact that the intersection is nonempty is a fundamental property of the real numbers. Had we restricted attention to rational numbers the intersection is quite likely to be empty. Clearly the element in $\cap_{k=1}^{\infty} I_k$ is the limit point for (x_{n_k}) and indeed for any sequence (x_k) that satisfies $x_k \in I_{k.}$

The proof of compactness of closed intervals leads us to another fundamental concept. A normed vector space is said to be *complete* if Cauchy's convergence criterion holds true: x_n is convergent if and only if $\|x_n - x_m\| \to 0$ as $m, n \to \infty$. Note that we assert that a sequence is convergent without specifying the limit. This is quite important in many contexts. It is a fundamental property of the real numbers that they are complete. Note that completeness could have been used to establish the convergence of the sequence (x_{n_k}) in the proof of compactness of $[0,1]$. From completeness of $\mathbb{R}$ ones sees that $\mathbb{C}$ and $\mathbb{R}^n$, $\mathbb{C}^n$ are complete since convergence is the same as coordinate convergence. From that we will in a minute be able to conclude that all finite dimensional vector spaces are complete. Note that the rationals $\mathbb{Q}$ are not complete as we can find sequences of rational numbers converging to any real number. These sequences do satisfy $\|x_n - x_m\| \to 0$ as $m, n \to \infty$, but they don't necessarily converge to a rational number. This is why we insist on only using real or complex scalars in connections with norms and inner products.

A crucial result connects continuous functions to compactness.

Theorem 5.1. *Let $f : V \to \mathbb{R}$ be a continuous function on a normed vector space. If $C \subset V$ is compact, then we can find $x_{\min}, x_{\max} \in C$ so that $f(x_{\min}) \leq f(x) \leq f(x_{\max})$ for all $x \in C$.*

Proof. Let us show how to find $x_{\max}$. The other point is found in a similar fashion. We consider the image $f(C) \subset \mathbb{R}$ and compute the smallest upper bound $y_0 = \sup f(C)$. That this number exists is one of the crucial properties of real numbers related to completeness. Now select a sequence $x_n \in C$ such that $f(x_n) \to y_0$. Since C is compact we can select a convergent subsequence $x_{n_k} \to x \in C$. This means that $f(x_{n_k}) \to f(x) = y_0$. In particular, y_0 is not infinite and the limit point x must be the desired $x_{\max}$. □

Example 5.2. The space $C^0([a,b], \mathbb{C})$ may or may not be complete depending on what norm we use. First we show that it is not complete with respect to any of the p-norms for $p < \infty$. To see this observe that we can find a sequence of continuous functions f_n on $[0,2]$ defined by

$$f_n(t) = \begin{cases} 1 & \text{for } t \geq 1 \\ t^n & \text{for } t < 1 \end{cases}$$

whose graphs converge to a step function

$$f(t) = \begin{cases} 1 & \text{for } t \geq 1 \\ 0 & \text{for } t < 1 \end{cases}.$$

We see that

$$\begin{aligned} \|f - f_n\|_p &\to 0, \\ \|f_m - f_n\|_p &\to 0 \end{aligned}$$

for all $p < \infty$. However, the limit function is not continuous and so the p-norm is not complete.

On the other hand the ∞-norm is complete. To see this suppose we have a sequence $f_n \in C^0([a,b], \mathbb{C})$ such that $\|f_n - f_m\|_\infty \to 0$. For each fixed t we have

$$|f_n(t) - f_m(t)| \leq \|f_n - f_m\|_\infty \to 0$$

as $n, m \to \infty$. Since $f_n(t) \in \mathbb{C}$ we can find $f(t) \in \mathbb{C}$ so that $f_n(t) \to f(t)$. To show that $\|f_n - f\|_\infty \to 0$ and $f \in C^0([a,b], \mathbb{C})$ fix $\varepsilon > 0$ and N so that

$$\|f_n - f_m\|_\infty \leq \varepsilon \text{ for all } n, m \geq N.$$

This implies that

$$|f_n(t) - f_m(t)| \leq \varepsilon, \text{ for all } t.$$

If we let $m \to \infty$ in this inequality we obtain

$$|f_n(t) - f(t)| \leq \varepsilon \text{ for all } n \geq N.$$

In particular

$$\|f_n - f\|_\infty \leq \varepsilon \text{ for all } n \geq N.$$

This implies that $f_n \to f$. Having proved this we next see that

$$\begin{aligned} |f(t) - f(t_0)| &\leq |f(t) - f_n(t)| + |f_n(t) - f_n(t_0)| + |f_n(t_0) - f(t_0)| \\ &\leq \|f_n - f\|_\infty + |f_n(t) - f_n(t_0)| + \|f_n - f\|_\infty \\ &= 2\|f_n - f\|_\infty + |f_n(t) - f_n(t_0)| \end{aligned}$$

Since f_n is continuous and $\|f_n - f\|_\infty \to 0$ as $n \to \infty$ we can easily see that f is also continuous.

Convergence with respect to the ∞-norm is also often referred to as *uniform convergence.*

Our first crucial property for finite dimensional vector spaces is that convergence is independent of the norm.

Theorem 5.3. *Let V be a finite dimensional vector space with a norm $\|\cdot\|$ and $e_1, ..., e_m$ a basis for V. Then (x_n) is convergent if and only if all of the coordinates $(\alpha_{1n}), ..., (\alpha_{mn})$ from the expansion*

$$x_n = \begin{bmatrix} e_1 & \cdots & e_m \end{bmatrix} \begin{bmatrix} \alpha_{1n} \\ \vdots \\ \alpha_{mn} \end{bmatrix}$$

are convergent.

Proof. We define a new ∞-norm on V by

$$\begin{aligned} \|x\|_\infty &= \max\{|\alpha_1|, ..., |\alpha_m|\}, \\ x &= e_1\alpha_1 + \cdots + e_m\alpha_m. \end{aligned}$$

That this defines a norm follows from the fact that it is a norm on $\mathbb{F}^n$. Note that coordinate convergence is the same as convergence with respect to this ∞-norm.

Now observe that

$$\begin{aligned} |\|x\| - \|y\|| &\leq \|x - y\| \\ &\leq \|e_1(\alpha_1 - \beta_1) + \cdots + e_m(\alpha_m - \beta_m)\| \\ &\leq |\alpha_1 - \beta_1| \|e_1\| + \cdots + |\alpha_m - \beta_m| \|e_m\| \\ &\leq ||x - y||_\infty \max\{\|e_1\|, ..., \|e_m\|\}. \end{aligned}$$

In other words

$$\|\cdot\| : V \to \mathbb{F}$$

is continuous if we use the norm $\|\cdot\|_\infty$ on V. Now consider the set

$$S = \{x \in V : \|x\|_\infty = 1\}.$$

This is the boundary of the compact set $B = \{x \in V : \|x\|_\infty \leq 1\}$. Thus any continuous function on S must have a maximum and a minimum. Since $\|x\| \neq 0$ on S we can find $C > c > 0$ so that

$$c \leq \|x\| \leq C \text{ for } \|x\|_\infty = 1.$$

Using the scaling properties of the norm this implies

$$c\|x\|_\infty \leq \|x\| \leq C\|x\|_\infty.$$

Thus convergence with respect to either of the norms imply convergence with respect to the other of these norms. □

All of this shows that in finite dimensional vector spaces the only way of defining convergence is that borrowed from $\mathbb{F}^n$. Next we show that all linear maps on finite dimensional normed vector spaces are bounded and hence continuous.

Theorem 5.4. *Let $L : V \to W$ be a linear map between normed vector spaces. If V is finite dimensional, then L is bounded.*

Proof. Let us fix a basis $e_1, ..., e_m$ for V and use the notation from the proof just completed.

Using

$$L(x) = \begin{bmatrix} L(e_1) & \cdots & L(e_m) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}.$$

We see that

$$\begin{aligned} \|L(x)\| &\leq m\|x\|_\infty \max\{\|L(e_1)\|, ..., \|L(e_m)\|\} \\ &\leq mc^{-1}\|x\| \max\{\|L(e_1)\|, ..., \|L(e_m)\|\}, \end{aligned}$$

which implies that L is bounded. □

In infinite dimensions things are much trickier as there are many different ways in which one can define convergence Moreover, a natural operator such as the one defined by differentiation is not bounded or even continuous.

One can prove that if W (but not necessarily V) is complete, then the space of bounded linear maps $\mathcal{B}(V, W)$ is also complete. The situations we are mostly interested in are when both V and W are finite dimensional. From what we have just proven this means that $\mathcal{B}(V, W) = \operatorname{Hom}(V, W)$ and since $\operatorname{Hom}(V, W)$ is finite dimensional completeness also becomes automatic.

We have a very good example of an infinite dimensional complete inner product space.

Example 5.5. The space ℓ^2 with the norm $\|x\|_2 = \sqrt{(x|x)}$ is, unlike $C^0([a,b],\mathbb{C})$, a complete infinite dimensional inner product space.

To prove this we take a sequence $x_k = (\alpha_{n,k}) \in \ell^2$ such that $\|x_k - x_m\|_2 \to 0$ as $k, m \to \infty$. If we fix a coordinate entry n we have that

$$|\alpha_{n,k} - \alpha_{n,m}| \leq \|x_k - x_m\|_2 .$$

So for fixed n we have a sequence $(\alpha_{n,k})$ of complex numbers that must be convergent; $\lim_{k\to\infty} \alpha_{n,k} = \alpha_n$. This gives us a potential limit point $x = (\alpha_n)$ for x_n. For simplicity let us assume that the index set for the coordinates is $\mathbb{N}$. If we assume that

$$\|x_k - x_m\|_2 \leq \varepsilon$$

for all $k, m \geq N$, then

$$\sum_{i=1}^{n} |\alpha_{i,k} - \alpha_{i,m}|^2 \leq \varepsilon^2.$$

If we let $m \to \infty$ in this sum, then we obtain

$$\sum_{i=1}^{n} |\alpha_{i,k} - \alpha_i|^2 \leq \varepsilon^2.$$

Since this holds for all n we can also let $n \to \infty$ in order to get

$$\|x_k - x\|_2 = \sqrt{\sum_{i=1}^{\infty} |\alpha_{i,k} - \alpha_i|^2} \leq \varepsilon \text{ for all } k \geq N.$$

This tells us that $x_k \to x$ as $k \to \infty$. To see that $x \in \ell^2$ just use that $x = x_k + (x - x_k)$ and that we have just shown $(x - x_k) \in \ell^2$.

With this in mind we can now prove the result that connects our two different concepts of completeness.

Theorem 5.6. *Let V be a complete inner product space with a complete basis $e_1, e_2, ..., e_n, ...$ If V is finite dimensional then it is isometric to $\mathbb{F}^n$ and if $e_1, e_2, ..., e_n, ...$ is infinite, then V is isometric to ℓ^2, where we use real or complex sequences in ℓ^2 according to the fields we have used for V.*

Proof. All we need to prove is that the map $V \to \ell^2$ is onto in the case $e_1, e_2, ..., e_n, ...$ is infinite. To see this let $(\alpha_i) \in \ell^2$. We claim that the series $\sum_i \alpha_i e_i$ is convergent. The series $\sum_i \|\alpha_i e_i\|^2 = \sum_i |\alpha_i|^2$ is assumed to be convergent. Using Pythagoras we obtain

$$\begin{aligned} \left\| \sum_{i=m}^{n} \alpha_i e_i \right\|^2 &= \sum_{i=m}^{n} \|\alpha_i e_i\|^2 \\ &= \sum_{i=m}^{n} |\alpha_i|^2 \to 0 \text{ as } n, m \to \infty. \end{aligned}$$

This implies that the sequence $x_n = \sum_{i=1}^{n} \alpha_i e_i$ of partial sums satisfies

$$\|x_n - x_m\| \to 0 \text{ as } n, m \to \infty.$$

Cauchy's convergence criterion can then be applied to show convergence as we assumed that V is complete. □

A complete inner product space is usually referred to as a *Hilbert space*. Hilbert introduced the complete space ℓ^2, but did not study more abstract infinite dimensional spaces. It was left to von Neumann to do that and also coin the term Hilbert space. We just saw that ℓ^2 is in a sense universal provided one can find suitable orthonormal collections of vectors. The goal of the next section is to attempt to do this for the space of periodic functions $C^0_{2\pi}(\mathbb{R}, \mathbb{C})$.

In normed vector spaces completeness implies the important *absolute convergence criterion* for series. Recall that a series $\sum_{n=1}^{\infty} x_n$ is convergent if the partial sums $z_m = \sum_{n=1}^{m} x_n = x_1 + \cdots + x_m$ form a convergent series. The limit is denoted by $\sum_{n=1}^{\infty} x_n$. The absolute convergence criterion states that $\sum_{n=1}^{\infty} x_n$ is convergent if it is absolutely convergent, i.e., $\sum_{n=1}^{\infty} \|x_n\|$ is convergent. It is known from calculus that a series of numbers, such as $\sum_{n=1}^{\infty} \frac{(-1)^n}{n}$, can be convergent without being absolutely convergent. Using the principle of absolute convergence it is sometimes possible to reduce convergence of series to the simpler question of convergence of series with nonnegative numbers, a subject studied extensively in calculus. To justify our claim note that

$$\|z_m - z_k\| = \|x_{k+1} + \cdots + x_m\| \leq \|x_{k+1}\| + \cdots + \|x_m\| \to 0$$

as $k, m \to \infty$ since $\sum_{n=1}^{\infty} \|x_n\|$ is convergent.

6 Orthonormal Bases in Infinite Dimensions

The goal of this section is to find complete orthonormal sets for 2π-periodic functions on $\mathbb{R}$. Recall that this space is denoted $C^0_{2\pi}(\mathbb{R}, \mathbb{R})$ if they are real valued and $C^0_{2\pi}(\mathbb{R}, \mathbb{C})$ if complex valued. For simplicity we shall concentrate on the later space. The inner product we use is given by

$$(f|g) = \frac{1}{2\pi} \int_0^{2\pi} f(t)\, \overline{g(t)} dt.$$

First we recall that $C^0_{2\pi}(\mathbb{R}, \mathbb{C})$ is not complete with this inner product. We can therefore not expect this space to be isometric to ℓ^2. Next recall that this space is complete if we use the stronger norm

$$\|f\|_{\infty} = \max_{t \in \mathbb{R}} |f(t)|.$$

We have a natural candidate for a complete orthonormal basis by using the functions $e_n = \exp(int)$ for $n \in \mathbb{Z}$. It is instructive to check that this is an orthonormal collection of functions. First we see that they are of unit length

$$\begin{aligned}\|e_n\|^2 &= \frac{1}{2\pi}\int_0^{2\pi} |\exp(int)|\, dt \\ &= \frac{1}{2\pi}\int_0^{2\pi} 1dt \\ &= 1.\end{aligned}$$

Next for $n \neq m$ we compute the inner product

$$\begin{aligned}(e_n|e_m) &= \frac{1}{2\pi}\int_0^{2\pi} \exp(int)\exp(-imt)\, dt \\ &= \frac{1}{2\pi}\int_0^{2\pi} \exp(i(n-m)t)\, dt \\ &= \frac{1}{2\pi}\left(\frac{\exp(i(n-m)t)}{i(n-m)}\right)\Bigg|_0^{2\pi} \\ &= 0\end{aligned}$$

since $\exp(i(n-m)t)$ is 2π-periodic.

We use a special notation for the Fourier coefficients $f_k = (f|e_k)$ of f indicating that they depend on f and k. One also often sees the notation

$$\hat{f}_k = (f|e_k).$$

The *Fourier expansion* for f is denoted

$$\sum_{k=-\infty}^{\infty} f_k \exp(ikt).$$

We also write

$$f \sim \sum_{k=-\infty}^{\infty} f_k \exp(ikt).$$

The $\sim$ indicates that the two expressions may not be equal. In fact as things stand there is no guarantee that the Fourier expansion represents a function and even less that it should represent f. We wish to show that

$$\left\| f - \sum_{k=-n}^{n} f_k \exp(ikt) \right\| \to 0$$

as $n \to \infty$, thus showing that we have a complete orthonormal basis. Even this, however, still does not tell us anything about pointwise or uniform convergence of the Fourier expansion.

From Bessel's inequality we derive a very useful result which is worthwhile stating separately.

Proposition 6.1. *Given a function $f \in C^0_{2\pi}(\mathbb{R}, \mathbb{C})$, then the Fourier coefficients satisfy:*

$$\begin{aligned} f_n &\to 0 \text{ as } n \to \infty \\ f_{-n} &\to 0 \text{ as } n \to \infty \end{aligned}$$

Proof. We have that

$$\begin{aligned} \sum_{n=-\infty}^{\infty} |f_n|^2 &\leq \|f\|^2 \\ &= \frac{1}{2\pi} \int_0^{2\pi} |f(t)|^2 \, dt \\ &< \infty \end{aligned}$$

Thus both of the series $\sum_{n=0}^{\infty} |c_n|^2$ and $\sum_{n=0}^{\infty} |c_{-n}|^2$ are convergent. Hence the terms go to zero as $n \to \pm\infty$. □

By looking at the proof we note that it wasn't really necessary for f to be continuous only that we know how to integrate $|f(t)|^2$ and $f(t)\exp(int)$. This means that the result still holds if f is piecewise continuous. This will come in handy below.

Before explaining the first result on convergence of the Fourier expansion we need to introduce the Dirichlet kernel.

Define

$$\begin{aligned} D_n(t_0 - t) &= \sum_{k=-n}^{n} \exp(ik(t_0 - t)) \\ &= \frac{\exp(i(n+1)(t_0 - t)) - \exp(-in(t_0 - t))}{\exp(i(t_0 - t)) - 1}. \end{aligned}$$

This formula follows from the formula for the sum of a finite geometric progression

$$\sum_{k=0}^{n} z^k = \frac{z^{n+1} - 1}{z - 1}$$

Specifically we have

$$\begin{aligned} \sum_{k=-n}^{n} \exp(ik(t_0 - t)) &= \sum_{l=0}^{2n} \exp(i(l-n)(t_0 - t)) \\ &= \exp(-in(t_0 - t)) \sum_{l=0}^{2n} \exp(il(t_0 - t)) \\ &= \exp(-in(t_0 - t)) \frac{\exp(i(2n+1)(t_0 - t)) - 1}{\exp(i(t_0 - t)) - 1} \\ &= \frac{\exp(i(n+1)(t_0 - t)) - \exp(-in(t_0 - t))}{\exp(i(t_0 - t)) - 1}. \end{aligned}$$

Note that

$$\frac{1}{2\pi}\int_0^{2\pi} D_n\left(t_0-t\right)dt = 1,$$

since the only term in the formula $D_n\left(t_0-t\right)=\sum_{k=-n}^{n}\exp\left(ik\left(t_0-t\right)\right)$ that has nontrivial integral is $\exp\left(i0\left(t_0-t\right)\right)=1$.

The importance of the Dirichlet kernel lies in the fact that the partial sums

$$s_n\left(t\right)=\sum_{k=-n}^{n} f_k\exp\left(ikt\right)$$

can be written in the condensed form

$$\begin{aligned}
s_n\left(t_0\right) &= \sum_{k=-n}^{n} f_k\exp\left(ikt_0\right)\\
&= \sum_{k=-n}^{n}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} f\left(t\right)\exp\left(-ikt\right)dt\right)\exp\left(ikt_0\right)\\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(f\left(t\right)\sum_{k=-n}^{n}\exp\left(ik\left(t_0-t\right)\right)\right)dt\\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(f\left(t\right)\frac{\exp\left(i\left(n+1\right)\left(t_0-t\right)\right)-\exp\left(-in\left(t_0-t\right)\right)}{\exp\left(i\left(t_0-t\right)\right)-1}\right)dt\\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} f\left(t\right)D_n\left(t_0-t\right)dt.
\end{aligned}$$

The partial sums of the Fourier expansion can therefore be computed without calculating the Fourier coefficients. This is often very useful both in applications and for mathematical purposes. Note also that the partial sum of f represents the orthogonal projection of f onto $\operatorname{span}\left\{1,\exp\left(\pm t\right),...,\exp\left(\pm nt\right)\right\}$ and is therefore the element in $\operatorname{span}\left\{1,\exp\left(\pm t\right),...,\exp\left(\pm nt\right)\right\}$ that is closest to f.

We can now prove a result on pointwise convergence of Fourier series.

Theorem 6.2. *Let $f\left(t\right)\in C^0_{2\pi}\left(\mathbb{R},\mathbb{C}\right)$. If f is continuous and differentiable at t_0, then the Fourier series for f converges to $f\left(t_0\right)$ at t_0.*

Proof. We must show that $s_n\left(t_0\right)\to f\left(t_0\right)$. The proof proceeds by a direct and fairly simple calculation of the partial sum of the Fourier series for f.

$$\begin{aligned}
& s_n(t_0) \\
= {} & \frac{1}{2\pi}\int_0^{2\pi} f(t)\, D_n(t_0-t)\, dt \\
= {} & \frac{1}{2\pi}\int_0^{2\pi} f(t_0)\, D_n(t_0-t)\, dt + \frac{1}{2\pi}\int_0^{2\pi} (f(t)-f(t_0))\, D_n(t_0-t)\, dt \\
= {} & f(t_0)\frac{1}{2\pi}\int_0^{2\pi} D_n(t_0-t)\, dt \\
& + \frac{1}{2\pi}\int_0^{2\pi} \frac{f(t)-f(t_0)}{\exp(i(t_0-t))-1}\left(\exp(i(n+1)(t_0-t))-\exp(-in(t_0-t))\right) dt \\
= {} & f(t_0) + \frac{1}{2\pi}\int_0^{2\pi} g(t)\left(\exp(i(n+1)(t_0-t))-\exp(-in(t_0-t))\right) dt \\
= {} & f(t_0) + \exp(i(n+1)t_0)\frac{1}{2\pi}\int_0^{2\pi} g(t)\exp(-i(n+1)t)\, dt \\
& - \exp(-int_0)\frac{1}{2\pi}\int_0^{2\pi} g(t)\exp(int)\, dt \\
= {} & f(t_0) + \exp(i(n+1)t_0)\, g_{n+1} - \exp(-int_0)\, g_{-n},
\end{aligned}$$

where

$$g(t) = \frac{f(t)-f(t_0)}{\exp(i(t_0-t))-1}.$$

Since $g(t)$ is nicely defined everywhere except at $t = t_0$ and f is continuous it must follow that g is continuous except possibly at t_0. At t_0 we can use L'Hospital's rule to see that g can be defined at t_0 so as to be a continuous function:

$$\begin{aligned}
\lim_{t\to t_0} g(t) &= \lim_{t\to t_0}\frac{f(t)-f(t_0)}{\exp(i(t_0-t))-1} \\
&= \frac{\left(\frac{d}{dt}(f(t)-f(t_0))\right)(\text{at } t=t_0)}{\left(\frac{d}{dt}(\exp(i(t_0-t))-1)\right)(\text{at } t=t_0)} \\
&= \frac{(f'(t))(\text{at } t=t_0)}{(-\exp(i(t_0-t)))(\text{at } t=t_0)} \\
&= \frac{f'(t_0)}{(-\exp(i(t_0-t_0)))} \\
&= -f'(t_0).
\end{aligned}$$

Having now established that $g \in C^0_{2\pi}(\mathbb{R},\mathbb{C})$ it follows that the Fourier coefficients g_{n+1} and g_{-n} go to zero as $n \to \infty$. Thus the partial sum converges to $f(t_0)$. □

If we make some further assumptions about the differentiability of f then we can use this pointwise convergence result to show convergence of the Fourier expansion of f.

Proposition 6.3. *If $f \in C^0_{2\pi}(\mathbb{R},\mathbb{C})$, and f' is piecewise continuous, then the Fourier coefficients for f and f' are related by*

$$f'_k = (ik) \cdot f_k$$

Proof. First we treat the case when $k = 0$

$$\begin{aligned} f'_0 &= \frac{1}{2\pi}\int_0^{2\pi} f'(t)\,dt \\ &= \frac{1}{2\pi} f(t)|_0^{2\pi} \\ &= 0, \end{aligned}$$

since $f(0) = f(2\pi)$. The general case follows from integration by parts

$$\begin{aligned} f'_k &= \frac{1}{2\pi}\int_0^{2\pi} f'(t)\exp(-ikt)\,dt \\ &= \frac{1}{2\pi} f(t)\exp(-ikt)\Big|_0^{2\pi} - \frac{1}{2\pi}\int_0^{2\pi} f(t)(-ik)\exp(-ikt)\,dt \\ &= \frac{1}{2\pi}(ik)\int_0^{2\pi} f(t)\exp(-ikt)\,dt \\ &= (ik) f_k \end{aligned}$$

□

We can now prove the first good convergence result for Fourier expansions

Theorem 6.4. *Let $f \in C^0_{2\pi}(\mathbb{R},\mathbb{C})$, and assume in addition that f' is piecewise continuous, then the Fourier expansion for f converges uniformly to f.*

Proof. It follows from the above result that the Fourier expansion converges pointwise to f except possibly at a finite number of points were f' is not defined. Therefore, if we can show that the Fourier expansion is uniformly convergent it must converge to a continuous function that agrees with f except possibly at the points where f' is not defined. However, if two continuous functions agree except at a finite number of points then they must be equal.

We evidently have that

$$f'_k = (ik) f_k.$$

Thus

$$|f_k| \leq \frac{1}{k}|f'_k|.$$

Now we know that both of the sequences $\left(\frac{1}{k}\right)_{k\in\mathbb{Z}-\{0\}}$ and $(|f'_k|)_{k\in\mathbb{Z}}$ lie in $\ell^2(\mathbb{Z})$. Thus the inner product of these two sequences

$$\sum_{k\neq 0}\frac{1}{k}|f'_k|$$

is well defined and represents a convergent series. This implies that

$$\sum_{k=-\infty}^{\infty} f_k$$

is absolutely convergent. Recall that $C^0_{2\pi}(\mathbb{R},\mathbb{C})$ is complete when we use the norm $||\cdot||_\infty$. Since

$$\|f_k \exp(ikt)\|_\infty = |f_k|$$

we get that

$$\sum_{k=-\infty}^{\infty} f_k \exp(ikt)$$

is uniformly convergent. □

The above result can be illustrated rather nicely.

Example 6.5. Consider the function given by $f(x) = |x|$ on $[-\pi,\pi]$. The Fourier coefficients are

$$\begin{aligned} f_0 &= \frac{\pi}{2}, \\ f_k &= \frac{1}{ik} f'_k \\ &= \frac{1}{ik}\frac{1}{2\pi}\left(\int_{-\pi}^{0} -\exp(-ikt)\,dt + \int_0^\pi \exp(-ikt)\,dt\right) \\ &= \frac{1}{ik}\frac{1}{2\pi}\left(2i\frac{-1+\cos\pi k}{k}\right) \\ &= \frac{1}{k^2}\frac{1}{\pi}\left(-1+(-1)^k\right) \end{aligned}$$

Thus we see that

$$\left|f_k e^{ikt}\right| \leq \frac{2}{\pi}\frac{1}{k^2}.$$

Hence we are in the situation where we have uniform convergence of the Fourier expansion. We can even sketch s_8 and compare it to f to convince ourselves that the convergence is uniform.

If we calculate the function and the Fourier series at $t = \pi$ we get

$$\pi = \frac{\pi}{2} + \sum_{k\neq 0} \frac{1}{\pi}\frac{-1+(-1)^k}{k^2}\exp(ik\pi).$$

This means that

$$\begin{aligned} \frac{\pi^2}{2} &= 2\sum_{k=1}^{\infty} \frac{-1+(-1)^k}{k^2}(-1)^k \\ &= 4\sum_{l=0}^{\infty} \frac{1}{(2l+1)^2} \end{aligned}$$

Thus yielding the formula

$$\frac{\pi^2}{8} = 1 + \frac{1}{9} + \frac{1}{25} + \cdots$$

In case f is not continuous there is, however, no hope that we could have uniform convergence. This is evident from our theory as the partial sums of the Fourier series always represent continuous functions. If the Fourier series converges uniformly, it must therefore converge to a continuous function. Perhaps the following example will be even more convincing.

Example 6.6. If $f(x) = x$ on $[-\pi, \pi]$, then $f(x)$ is not continuous when thought of as a 2π-periodic function. In this case the Fourier coefficients are

$$\begin{aligned} f_0 &= 0, \\ f_k &= \frac{i(-1)^k}{k}. \end{aligned}$$

Thus

$$\left|f_k e^{ikx}\right| = \frac{1}{k}$$

and we clearly can't guarantee uniform convergence. This time the partial sum looks like.

This clearly approximates f, but not uniformly due to the jump discontinuities.

The last result shows that we nevertheless do have convergence in the norm that comes from the inner product on $C^0_{2\pi}(\mathbb{R}, \mathbb{C})$.

Theorem 6.7. *Let $f \in C^0_{2\pi}(\mathbb{R}, \mathbb{C})$, then the Fourier series converges to f in the sense that*

$$\|f - s_n\| \to 0 \text{ as } n \to \infty.$$

Proof. First suppose in addition that f' exists and is piecewise continuous. Then we have from the previous result that $|f(t) - s_n(t)|$ and consequently also $|f(t) - s_n(t)|^2$ converge uniformly to zero. Hence

$$\begin{aligned} \|f - s_n\|_2^2 &= \frac{1}{2\pi}\int_0^{2\pi} |f(t) - s_n(t)|^2 \, dx \\ &\leq \|f - s_n\|_\infty \to 0. \end{aligned}$$

In the more general situation we must use that for each small number $\varepsilon > 0$ the function f can be approximated by functions $f_\varepsilon \in C^0_{2\pi}(\mathbb{R}, \mathbb{C})$ with piecewise continuous f' such that

$$\|f - f_\varepsilon\| < \varepsilon.$$

Supposing that we can find such f_ε we can show that $\|f - s_n\|_2$ can be made as small as we like. Denote by $s_n^\varepsilon(t)$ the n-th partial sum in the Fourier expansion for f_ε. Since $s_n^\varepsilon(t)$ and $s_n(t)$ are linear combinations of the same functions

$\exp(ikt), k = 0, \pm 1, ..., \pm n$ and $s_n(t)$ is the best approximation of f we must have

$$\|f - s_n\|_2 \leq \|f - s_n^{\varepsilon}\|_2 .$$

We can now apply the triangle inequality to obtain

$$\begin{aligned} \|f - s_n\|_2 &\leq \|f - s_n^{\varepsilon}\|_2 \\ &\leq \|f - f_{\varepsilon}\|_2 + \|f_{\varepsilon} - s_n^{\varepsilon}\|_2 \\ &\leq \varepsilon + \|f_{\varepsilon} - s_n^{\varepsilon}\|_2 . \end{aligned}$$

Using that $\|f_{\varepsilon} - s_n^{\varepsilon}\|_2 \to 0$ as $n \to \infty$, we can choose $N > 0$ so that $\|f_{\varepsilon} - s_n^{\varepsilon}\|_2 \leq \varepsilon$ for all $n \geq N$. This implies that

$$\begin{aligned} \|f - s_n\|_2 &\leq \varepsilon + \|f_{\varepsilon} - s_n^{\varepsilon}\|_2 \\ &= 2\varepsilon. \end{aligned}$$

as long as $n \geq N$. As we can pick $\varepsilon > 0$ as we please, it must follow that

$$\lim_{n \to \infty} \|f - s_n\|_2 = 0.$$

It now remains to establish that we can approximate f by the appropriate functions. Clearly this amounts to showing that we can find nice functions f_{ε} such that the area under the graph of $|f(t) - f_{\varepsilon}(t)|^2$ is small for small ε. The way to see that this can be done is to approximate f by a spline or piecewise linear function g_{ε}. For that construction we simply subdivide $[0, 2\pi]$ into intervals whose endpoints are given by $0 = t_0 < t_1 < \cdots < t_N = 2\pi$. Then we define

$$g(t_k) = f(t_k)$$

and

$$g(st_k + (1 - s)t_{k-1}) = sf(t_k) + (1 - s) f(t_{k-1})$$

for $0 < s < 1$. This defines a function $g \in C^0_{2\pi}(\mathbb{R}, \mathbb{C})$ that is glued together by line segments. Using that f is uniformly continuous on $[0, 2\pi]$ we can make $|f(t) - g(t)|^2$ as small as we like by choosing the partition sufficiently fine. Thus also $\|f - g\|_2 \leq \|f - g\|_{\infty}$ is small. □

6.1 Exercises

1. Show that
$$1, \sqrt{2}\cos(t), \sqrt{2}\sin(t), \sqrt{2}\cos(2t), \sqrt{2}\sin(2t), ...$$
forms a complete orthonormal set for $C^0_{2\pi}(\mathbb{R}, \mathbb{C})$. Use this to conclude that it is also a complete orthonormal set for $C^0_{2\pi}(\mathbb{R}, \mathbb{R})$.

2. Show that $1, \sqrt{2}\cos(t), \sqrt{2}\cos(2t), ..$ respectively $\sqrt{2}\sin(t), \sqrt{2}\sin(2t), ...$ form complete orthonormal sets for the even respectively odd functions in $C^0_{2\pi}(\mathbb{R}, \mathbb{R})$.

3. Show that for any piecewise continuous function f on $[0, 2\pi]$, one can for each $\varepsilon > 0$ find $f_{\varepsilon} \in C^0_{2\pi}(\mathbb{R}, \mathbb{C})$ such that $\|f - f_{\varepsilon}\|_2 \leq \varepsilon$. Conclude that the Fourier expansion converges to f for such functions.

7 Applications of Norms

In this section we complete some unfinished business on existence and uniqueness of solutions to linear differential equations and the proof of the implicit function theorem. Both of these investigations use completeness and operator norms rather heavily and are therefore perfect candidates for justifying all of the notions relating to normed vector spaces introduced earlier in this chapter.

We are now also ready to complete the proof of the implicit function theorem. Let us recall the theorem and the set-up for the proof as far as it went.

Theorem 7.1. (The Implicit Function Theorem) *Let $F : \mathbb{R}^{m+n} \to \mathbb{R}^n$ be smooth. If $F(z_0) = c \in \mathbb{R}^n$ and $\operatorname{rank}(DF_{z_0}) = n$, then we can find a coordinate decomposition $\mathbb{R}^{m+n} = \mathbb{R}^m \times \mathbb{R}^n$ near z_0 such that the set $S = \{z \in \mathbb{R}^{m+n} : F(z) = c\}$ is a smooth graph over some open set $U \subset \mathbb{R}^m$.*

Proof. We assume that $c = 0$ and split $\mathbb{R}^{m+n} = \mathbb{R}^m \times \mathbb{R}^n$ so that the projection $P : \mathbb{R}^{m+n} \to \mathbb{R}^m$ is an isomorphism when restricted to $\ker(DF_{z_0})$. Then $DF_{z_0}|_{\mathbb{R}^n} : \mathbb{R}^n \to \mathbb{R}^n$ is an isomorphism. Note that the version of $\mathbb{R}^n$ that appears in the domain for DF might have coordinates that are differently indexed than the usual indexing used in the image version of $\mathbb{R}^n$. Next rename the coordinates $z = (x, y) \in \mathbb{R}^m \times \mathbb{R}^n$ and set $z_0 = (x_0, y_0)$. The goal is to find $y = y(x) \in \mathbb{R}^n$ as a solution to $F(x, y) = 0$. To make things more rigorous we choose norms on all of the vector spaces. Then we can consider the closed balls $\bar{B}_\varepsilon = \{x \in \mathbb{R}^m : \|x - x_0\| \leq \varepsilon\}$, which are compact subsets of $\mathbb{R}^m$ and where ε is to be determined in the course of the proof. The appropriate vector space where the function $x \to y(x)$ lives is the space of continuous functions $V = C^0\left(\bar{B}_\varepsilon, \mathbb{R}^n\right)$ where we use the norm

$$\|y\|_\infty = \max_{x \in \bar{B}_\varepsilon} \|y(x)\|.$$

With this norm the space is a complete normed vector space just like $C^0([a, b], \mathbb{C})$.

The iteration for constructing $y(x)$ is

$$y_{n+1} = y_n - \left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}(F(x, y_n))$$

and starts with $y_0(x) = y_0$. First we show that $y_n(x)$ is never far away from y_0.

This is done as follows

$$\begin{aligned}
& y_{n+1}-y_0 \\
= \quad & y_n-y_0-\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(F\left(x,y_n\right)\right) \\
= \quad & y_n-y_0 \\
& -\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(DF_{(x_0,y_0)}|_{\mathbb{R}^m}\left(x-x_0\right)+DF_{(x_0,y_0)}|_{\mathbb{R}^n}\left(y_n-y_0\right)+R\right) \\
= \quad & y_n-y_0 \\
& -\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}DF_{(x_0,y_0)}|_{\mathbb{R}^n}\left(y_n-y_0\right) \\
& -\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(DF_{(x_0,y_0)}|_{\mathbb{R}^m}\left(x-x_0\right)+R\right) \\
= \quad & y_n-y_0 \\
& -\left(y_n-y_0\right) \\
& -\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(DF_{(x_0,y_0)}|_{\mathbb{R}^m}\left(x-x_0\right)+R\right) \\
= \quad & -\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(DF_{(x_0,y_0)}|_{\mathbb{R}^m}\left(x-x_0\right)+R\right)
\end{aligned}$$

where the remainder is

$$R=F\left(x,y_n\right)-F\left(x_0,y_0\right)-DF_{(x_0,y_0)}|_{\mathbb{R}^m}\left(x-x_0\right)-DF_{(x_0,y_0)}|_{\mathbb{R}^n}\left(y_n-x_0\right)$$

and has the property that

$$\frac{\|R\|}{\|y_n-y_0\|+\|x-x_0\|}\rightarrow 0 \text{ as } \|y_n-y_0\|+\|x-x_0\|\rightarrow 0.$$

Thus we have

$$\|y_{n+1}-y_0\|\leq\left\|\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\right\|\left(\left\|DF_{(x_0,y_0)}|_{\mathbb{R}^m}\right\|\|x-x_0\|+\|R\|\right).$$

Here $\left\|\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\right\|$ and $\left\|DF_{(x_0,y_0)}|_{\mathbb{R}^m}\right\|$ are fixed quantities, while $\|x-x_0\|\leq\varepsilon$ and we can also assume

$$\begin{aligned}
\|R\| & \leq \frac{1}{4\left\|\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\right\|}\left(\|y_n-y_0\|+\|x-x_0\|\right) \\
& \leq \frac{1}{4\left\|\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\right\|}\left(\|y_n-y_0\|+\varepsilon\right)
\end{aligned}$$

provided $\|y_n-y_0\|,\|x-x_0\|$ are small. This means that

$$\begin{aligned}
\|y_{n+1}-y_0\| \leq & \left\|\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\right\|\left\|DF_{(x_0,y_0)}|_{\mathbb{R}^m}\right\|\varepsilon \\
& +\frac{1}{4}\left(\|y_n-y_0\|+\varepsilon\right).
\end{aligned}$$

This means that we can control the distance $\|y_{n+1}-y_0\|$ in terms of $\|y_n-y_0\|$ and ε. In particular we can for any $\delta>0$ find $\varepsilon=\varepsilon\left(\delta\right)>0$ so that $\|y_{n+1}-y_0\|$

$\leq \delta$ for all n. This means that the y_n functions stay close to y_0. This will be important in the next part of the proof.

Next let us see how far successive functions are from each other

$$\begin{aligned}
y_{n+1} - y_n &= -\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(F\left(x, y_n\right)\right) \\
&= -\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(F\left(x, y_{n-1}\right) + DF_{(x,y_{n-1})}\left(y_n - y_{n-1}\right) + R\right),
\end{aligned}$$

where

$$R = F\left(x, y_n\right) - F\left(x, y_{n-1}\right) - DF_{(x,y_{n-1})}\left(y_n - y_{n-1}\right)$$

and has the property that

$$\frac{\|R\|}{\|y_n - y_{n-1}\|} \to 0 \text{ as } \|y_n - y_{n-1}\| \to 0.$$

This implies

$$\begin{aligned}
y_{n+1} - y_n &= -\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(F\left(x, y_{n-1}\right)\right) \\
&\quad -\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(DF_{(x,y_{n-1})}\left(y_n - y_{n-1}\right)\right) \\
&\quad -\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}(R) \\
&= \left(y_n - y_{n-1}\right) \\
&\quad -\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(DF_{(x_0,y_0)}\left(y_n - y_{n-1}\right)\right) \\
&\quad +\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(\left(DF_{(x_0,y_0)} - DF_{(x,y_{n-1})}\right)\left(y_n - y_{n-1}\right)\right) \\
&\quad -\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}(R) \\
&= \left(y_n - y_{n-1}\right) - \left(y_n - y_{n-1}\right) \\
&\quad +\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(\left(DF_{(x_0,y_0)} - DF_{(x,y_{n-1})}\right)\left(y_n - y_{n-1}\right)\right) \\
&\quad -\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}(R) \\
&= \left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\left(\left(DF_{(x_0,y_0)} - DF_{(x,y_{n-1})}\right)\left(y_n - y_{n-1}\right)\right) \\
&\quad -\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}(R).
\end{aligned}$$

Thus

$$\begin{aligned}
\|y_{n+1} - y_n\| &\leq \left\|\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\right\| \left\|\left(DF_{(x_0,y_0)} - DF_{(x,y_{n-1})}\right)\right\| \|y_n - y_{n-1}\| \\
&\quad + \left\|\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\right\| \|R\|.
\end{aligned}$$

The fact that (x, y_{n-1}) is always close to (x_0, y_0) together with the assumption that $DF_{(x,y)}$ is continuous shows us that we can assume

$$\left\|\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\right\| \left\|\left(DF_{(x_0,y_0)} - DF_{(x,y_{n-1})}\right)\right\| \leq \frac{1}{4}$$

provided ε and δ are sufficiently small. The same is evidently true for

$$\left\|\left(DF_{(x_0,y_0)}|_{\mathbb{R}^n}\right)^{-1}\right\| \|R\|$$

and so we have

$$\|y_{n+1} - y_n\| \leq \frac{1}{2} \|y_n - y_{n-1}\| .$$

Iterating this we obtain

$$\begin{aligned} \|y_{n+1} - y_n\| &\leq \frac{1}{2} \|y_n - y_{n-1}\| \\ &\leq \frac{1}{2}\frac{1}{2} \|y_{n-1} - y_{n-2}\| \\ &\leq \left(\frac{1}{2}\right)^n \|y_1 - y_0\| . \end{aligned}$$

Now consider the telescopic series

$$\sum_{n=0}^{\infty} (y_{n+1} - y_n) .$$

This series is absolutely convergent as $\|y_{n+1} - y_n\| \leq \left(\frac{1}{2}\right)^n \|y_1 - y_0\|$ and the series

$$\|y_1 - y_0\| \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^n = 2 \|y_1 - y_0\|$$

is convergent. Since it is telescopic it converges to

$$\left(\lim_{n\to\infty} y_n\right) - y_0.$$

Thus we have shown that y_n converges in $V = C^0 \left(\bar{B}_\varepsilon, \mathbb{R}^n\right)$ to a function $y(x)$ that must solve $F(x, y(x)) = 0$. It remains to show that y is differentiable and compute its differential.

Using

$$\begin{aligned} 0 &= F(x+h, y(x+h)) - F(x, y(x)) \\ &= DF_{(x,y(x))}|_{\mathbb{R}^m}(h) + DF_{(x,y(x))}|_{\mathbb{R}^n}(y(x+h) - y(x)) + R \end{aligned}$$

and that $DF_{(x,y(x))}|_{\mathbb{R}^n}$ is invertible (an unjustified fact that follows from the fact that it is close to $DF_{(x_0,y_0)}|_{\mathbb{R}^n}$, see also exercises) we see that

$$y(x+h) - y(x) + \left(DF_{(x,y(x))}|_{\mathbb{R}^n}\right)^{-1} DF_{(x,y(x))}|_{\mathbb{R}^m}(h) = \left(DF_{(x,y(x))}|_{\mathbb{R}^n}\right)^{-1}(-R) .$$

This certainly indicates that y should be differentiable with derivative

$$-\left(DF_{(x,y(x))}|_{\mathbb{R}^n}\right)^{-1} DF_{(x,y(x))}|_{\mathbb{R}^m} .$$

This derivative varies continuously so y is continuously differentiable. To establish rigorously that the derivative is indeed correct we need only justify that

$$\lim_{\|h\|\to 0} \frac{\|-R\|}{\|h\|} = 0.$$

This follows from the definition of R and continuity of y. □

7.1 Exercises

1. Let $C \subset V$ be a closed subset of a real vector space. Assume that if $x, y \in C$, then $x + y \in C$ and $\frac{1}{2}x \in C$. Show that C is a real subspace.

2. Let $L : V \to W$ be a continuous additive map between normed vector spaces over $\mathbb{R}$. Show that L is linear. Hint: Use that it is linear with respect to $\mathbb{Q}$.

3. Let $f(z) = \sum_{n=0}^{\infty} a_n z^n$ define a power series. Let $A \in \mathrm{Mat}_{n\times n}(\mathbb{F})$. Show that one can define $f(A)$ as long as $\|A\| <$ radius of convergence.

4. Let $L : V \to V$ be a bounded operator on a normed vector space.

   (a) If $\|L\| < 1$, then $1_V + L$ has an inverse. Hint: $(1_V + L)^{-1} = \sum_{n=1}^{\infty} (-1)^n C^n$.

   (b) With L as above show

   $$\begin{aligned} \left\|L^{-1}\right\| &\leq \frac{1}{1 - \|L\|}, \\ \left\|(1_V + L)^{-1} - 1_V\right\| &\leq \frac{\|L\|}{1 - \|L\|}. \end{aligned}$$

   (c) If $\left\|L^{-1}\right\| \leq \varepsilon^{-1}$ and $\|L - K\| < \varepsilon$, then K is invertible and

   $$\begin{aligned} \left\|K^{-1}\right\| &\leq \frac{\left\|L^{-1}\right\|}{1 - \left\|L^{-1}(K - L)\right\|}, \\ \left\|L^{-1} - K^{-1}\right\| &\leq \frac{\left\|L^{-1}\right\|^2}{\left(1 - \left\|L^{-1}\right\| \|L - K\|\right)^2} \|L - K\|. \end{aligned}$$

5. Let $L : V \to V$ be a bounded operator on a normed vector space.

   (a) If λ is an eigenvalue for L, then

   $$|\lambda| \leq \|L\|.$$

   (b) Given examples of 2×2 matrices where strict inequality always holds.

6. Show that

   $$x(t) = \left(\exp(A(t - t_0)) \int_{t_0}^{t} \exp(-A(s - t_0)) f(s)\, ds\right) x_0$$

   solves the initial value problem $\dot{x} = Ax + f$, $x(t_0) = x_0$.

7. Let $A = B + C \in \mathrm{Mat}_{n\times n}(\mathbb{R})$ where B is invertible and $\|C\|$ is very small compared to $\|B\|$.

(a) Show that $B^{-1}-B^{-1}CB^{-1}$ is a good approximation to A^{-1}.

(b) Use this to approximate the inverse to $\begin{bmatrix} 1 & 0 & 1000 & 1 \\ 0 & -1 & 1 & 1000 \\ 2 & 1000 & -1 & 0 \\ 1000 & 3 & 2 & 0 \end{bmatrix}$.

8 Infinite Dimensional Extensions

Recall that our definition of adjoints rested on knowing that all linear functionals where of the form $x \to (x|y)$. This fact does not hold in infinite dimensional spaces unless we assume that they are complete. Even in that case we need to assume that the functionals are continuous for this result to hold.

Instead of trying to generalize the entire theory to infinite dimensions we are going to discuss a very important special case. Let $V = C_{2\pi}^{\infty}(\mathbb{R},\mathbb{C})$ be the space of of smooth 2π periodic functions with the inner product

$$(f|g) = \frac{1}{2\pi}\int_0^{2\pi} f(t)\overline{g(t)}dt.$$

The evaluation functional $L(f) = f(t_0)$ that evaluates a function in V at t_0 is not continuous nor is it of the form

$$L(f) = \frac{1}{2\pi}\int_0^{2\pi} f(t)\overline{g(t)}dt$$

no matter what class of functions g belongs to. Next consider

$$\begin{aligned} L(f) &= \frac{1}{2\pi}\int_0^{2\pi} f(t)\overline{g(t)}dt \\ &= \frac{1}{2\pi}\int_0^{\pi} f(t)\,dt \end{aligned}$$

where

$$g = \begin{cases} 1 & t \in [0,\pi] \\ 0 & t \in (\pi, 2\pi) \end{cases}$$

This functional is continuous but cannot be represented in the desired form using $g \in C_{2\pi}^{\infty}(\mathbb{R},\mathbb{C})$.

While there are very good ways of dealing with these problems in general we are only going to study operators where we can easily guess the adjoint. The basic operator we wish to study is the differentiation operator $D : C_{2\pi}^{\infty}(\mathbb{R},\mathbb{C}) \to C_{2\pi}^{\infty}(\mathbb{R},\mathbb{C})$. We have already shown that this map is skew-adjoint

$$(Df|g) = -(f|Dg).$$

This map yields an operator $D : V_0 \to V_0$, where $V_0 = \left\{ f \in V : \int_0^{2\pi} f(t)\,dt = 0 \right\}$. Clearly we can define D on V_0, the important observation is that

$$\int_0^{2\pi} (Df)(t)\,dt = f(t)|_0^{2\pi} = 0.$$

Thus $Df \in V_0$ for all $f \in V$. Apparently the function $f(t) \equiv 1$ does not belong to V_0. In fact V_0 is by definition the subspace of all functions that are perpendicular to 1. Since $\ker(D) = \operatorname{span}\{1\}$, we have that $V_0 = (\ker(D))^{\perp}$. The Fredholm alternative then indicates that we might expect $\operatorname{im}(D) = V_0$. This is not hard to verify directly. Let $g \in V_0$ and define

$$f(t) = \int_0^t g(s)\, ds.$$

Clearly g is smooth since f is smooth. Moreover since $f(2\pi) = \int_0^{2\pi} g(s)\, ds = 0 = f(0)$ it is also 2π periodic. Thus $f \in V$ and $Df = g$.

Our next important observation about D is that it is diagonalized by the complete orthonormal set $\exp(int), n \in \mathbb{Z}$ of vectors as

$$D(\exp(int)) = in \exp(int).$$

This is one reason why it is more convenient to work with complex valued functions as D does not have any eigenvalues aside from 0 on $C^{\infty}_{2\pi}(\mathbb{R}, \mathbb{R})$. Note that this also implies that D is unbounded since $\|D(\exp(int))\|_2 = |n| \to \infty$, while $\|\exp(int)\|_2 = 1$.

If we expand the function $f(t) \in V$ according to its Fourier expansion $f = \sum f_n \exp(int)$, then we see that the Fourier expansion for Df is

$$Df = \sum (in) f_n \exp(int).$$

This tells us that we cannot extend D to be defined on the Hilbert space $\ell^2(\mathbb{Z})$ as $((in) f_n)_{n\in\mathbb{Z}}$ doesn't necessarily lie in this space as long as we only assume $(f_n)_{n\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$. A good example of this is $f_n = 1/n$ for $n \neq 0$.

The expression for Df together with Parseval's formula tells us something quite interesting about the operator D, namely, we have *Wirtinger's inequality* for $f \in V_0$

$$\begin{aligned} \|f\|_2^2 &= \sum_{n\neq 0} |f_n|^2 \\ &\leq \sum_{n\neq 0} |in|^2 |f_n|^2 \\ &= \|Df\|_2^2. \end{aligned}$$

Thus the inverse $D^{-1} : V_0 \to V_0$ must be a bounded operator. At the level of Fourier series this map is evidently given by

$$D^{-1}\left(\sum_{n\neq 0} g_n \exp(int)\right) = \sum_{n\neq 0} \frac{g_n}{in} \exp(int).$$

In contrast to D we therefore have that D^{-1} does define a map $\ell^2(\mathbb{Z} - \{0\}) \to \ell^2(\mathbb{Z} - \{0\})$.

With all of this information about D we can now attempt to generalize to the situation to the operator $p(D) : C^{\infty}_{2\pi}(\mathbb{R},\mathbb{C}) \to C^{\infty}_{2\pi}(\mathbb{R},\mathbb{C})$, where $p(t) \in \mathbb{C}[t]$ is a complex polynomial. Having already seen that $D^* = -D$ we can define the adjoint $(p(D))^*$ by

$$\begin{aligned}
(p(D))^* &= \left(a_n D^n + a_{n-1} D^{n-1} + \cdots + a_1 D + a_0\right)^* \\
&= \bar{a}_n (-1)^n D^n + \bar{a}_{n-1} (-1)^{n-1} D^{n-1} + \cdots + \bar{a}_1 (-1) D + \bar{a}_0 \\
&= p^*(D).
\end{aligned}$$

Note that the "adjoint" polynomial $p^*(t)$ satisfies

$$\begin{aligned}
p^*(t) &= \overline{p(-t)}, \\
p^*(it) &= \overline{p(it)}
\end{aligned}$$

for all $t \in \mathbb{R}$. It is easy to check that $p^*(D)$ satisfies the usual adjoint property

$$(p(D) f, g) = (f, p^*(D) g).$$

We would expect $p(D)$ to be diagonalizable as it is certainly a normal operator. In fact we have

$$p(D)(\exp(int)) = p(in)\exp(int).$$

Thus we have the same eigenvectors as for D and the eigenvalues are simply $p(in)$. The adjoint then also has the same eigenvectors, but with conjugate eigenvalues as one would expect:

$$\begin{aligned}
p^*(D)(\exp(int)) &= p^*(in)\exp(int) \\
&= \overline{p(in)}\exp(int).
\end{aligned}$$

This immediately tells us that each eigenvalue can have at most $\deg(p)$ eigenvectors in the set $\{\exp(int) : n \in \mathbb{Z}\}$. In particular,

$$\begin{aligned}
\ker(p(D)) &= \ker(p^*(D)) \\
&= \operatorname{span}\{\exp(int) : p(in) = 0\}
\end{aligned}$$

and

$$\dim(\ker(p(D))) \leq \deg(p).$$

Since $\ker(p(D))$ is finite dimensional we have an orthogonal projection onto $\ker(p(D))$. Hence the orthogonal complement is well-defined and we have that

$$C^{\infty}_{2\pi}(\mathbb{R},\mathbb{C}) = \ker(p(D)) \oplus (\ker(p(D)))^{\perp}.$$

What is more, the Fredholm alternative also suggests that

$$\operatorname{im}(p(D)) = \operatorname{im}(p^*(D)) = (\ker(p(D)))^{\perp}.$$

Our eigenvalue expansion shows that

$$p(D)(f), p^*(D)(f) \in (\ker(p(D)))^{\perp}.$$

Moreover for each n where $p(in) \neq 0$ we have

$$\begin{aligned}
\exp(int) &= p(D)\left(\frac{1}{p(in)}\exp(int)\right), \\
\exp(int) &= p^{*}(D)\left(\frac{1}{\overline{p(in)}}\exp(int)\right).
\end{aligned}$$

Hence

$$\operatorname{im}(p(D)) = \operatorname{im}(p^{*}(D)) = (\ker(p(D)))^{\perp}.$$

Finally we can also generalize Wirtinger's inequality to the effect that we can find some $C > 0$ depending on $p(t)$ such that for all $f \in \operatorname{im}(p(D))$ we have

$$\|f\|_2^2 \leq C\,\|p(D)(f)\|_2^2.$$

To find C we must show that

$$C^{-1} = \inf\{|p(in)| : p(in) \neq 0\} > 0.$$

This follows from the fact that unless $\deg(p) = 0$ we have $|(p(z_n))| \to \infty$ for any sequence (z_n) of complex numbers such that $|z_n| \to \infty$ as $n \to \infty$. Thus $\inf\{|p(in)| : p(in) \neq 0\}$ is obtained for some value of n. In concrete situations it is quite easy to identify both the n such that $p(in) = 0$ and also the n that minimizes $|p(in)|$. The generalized Wirtinger inequality tells us that we have a bounded operator

$$(p(D))^{-1} : \operatorname{im}(p(D)) \to \operatorname{im}(p(D))$$

that extends to $\ell^2(\{n \in \mathbb{Z} : p(in) \neq 0\})$.

Let us collect some of these results in a theorem.

Theorem 8.1. *Consider* $p(D) : C^{\infty}_{2\pi}(\mathbb{R},\mathbb{C}) \to C^{\infty}_{2\pi}(\mathbb{R},\mathbb{C})$, *where* $p(t) \in \mathbb{C}[t]$. *Then*

$p(D)(\exp(int)) = p(in)\exp(int)$.

$\dim(\ker(p(D))) \leq \deg(p)$

$C^{\infty}_{2\pi}(\mathbb{R},\mathbb{C}) = \ker(p(D)) \oplus (\ker(p(D)))^{\perp} = \ker(p(D)) \oplus \operatorname{im}(p(D))$

$p(D) : \operatorname{im}(p(D)) \to \operatorname{im}(p(D))$ *is one-to-one and onto with bounded inverse.*

If $g \in \operatorname{im}(p(D))$, *then* $p(D)(x) = g$ *has a unique solution* $x \in \operatorname{im}(p(D))$.

This theorem comes in quite handy when trying to find periodic solutions to differential equations. We can illustrate this through a few examples.

Example 8.2. Consider $p(D) = D^2 - 1$. Then $p(t) = t^2 - 1$ and we see that $p(in) = -n^2 - 1 \leq -1$. Thus $\ker(p(D)) = \{0\}$. This should not come as a surprise as $p(D) = 0$ has two linearly independent solutions $\exp(\pm t)$ that are not periodic. We then conclude that $p(D) : C^{\infty}_{2\pi}(\mathbb{R},\mathbb{C}) \to C^{\infty}_{2\pi}(\mathbb{R},\mathbb{C})$ is an isomorphism with

$$\|f\|_2 \leq \|p(D)(f)\|_2,$$

and the equation $p(D)(x) = g \in C^{\infty}_{2\pi}(\mathbb{R},\mathbb{C})$ has unique solution $x \in C^{\infty}_{2\pi}(\mathbb{R},\mathbb{C})$. This solution can be found directly from the Fourier expansion of $g = \sum_{n\in\mathbb{Z}} g_n \exp(int)$:

$$x = \sum_{n\in\mathbb{Z}} \frac{g_n}{-n^2-1} \exp(int)$$

Consider $p(D) = D^2 + 1$. Then $p(t) = t^2 + 1$ and we have $p(\pm i) = 0$. Consequently

$$\begin{aligned} \ker(p(D)) &= \operatorname{span}\{\exp(it), \exp(-it)\} \\ &= \operatorname{span}\{\cos(t), \sin(t)\}. \end{aligned}$$

The orthogonal complement has the property that the ±1 term in the Fourier expansion is 0. So if

$$g = \sum_{n\neq\pm1} g_n \exp(int)$$

then the solution to

$$p(D)(x) = g$$

that lies in $\operatorname{im}(p(D))$ is given by

$$x = \sum_{n\neq\pm1} \frac{g_n}{-n^2+1} \exp(int).$$

We are going to have problems solving

$$D_t^2 x + x = \exp(\pm it)$$

even if we don't just look for periodic solutions. Usually one looks for solutions that look like the forcing terms g unless g is itself a solution to the homogeneous equation. Otherwise we have to multiply the forcing term by a polynomial of the appropriate degree. In this case we see that

$$x(t) = \frac{\mp it}{2} \exp(\pm it)$$

is a solution to the inhomogeneous equation. This is clearly not periodic, but it does yield a discontinuous 2π periodic solution if we declare that it is given by $x(t) = \frac{\mp it}{2}\exp(\pm it)$ on $[-\pi,\pi]$.

To end this section let us give a more geometric application of what has be developed so far. The classical isoperimetric problem asks, if among all domains in the plane with fixed perimeter $2\pi R$ the circle has the largest area πR^2? Thus the problem is to show that for a plane region $\Omega \subset \mathbb{C}$ we have that $\operatorname{area}(\Omega) \leq \pi R^2$ if the perimeter of $\partial\Omega$ is $2\pi R$. This is were the functions from the space $C^{\infty}_{2\pi}(\mathbb{R},\mathbb{C})$ come in handy in a different way. Assume that the perimeter is 2π and then parametrize it by arclength via a function $f(t) \in C^{\infty}_{2\pi}(\mathbb{R},\mathbb{C})$. The length of the perimeter is then calculated by

$$\int_0^{2\pi} |(Df)(t)|\, dt = 2\pi$$

Note that multiplication by $\pm i$ rotates a vector by 90° so $\pm i\,(Df)\,(t)$ represent the unit normal vectors to the domain at $f\,(t)$ since $Df\,(t)$ is a unit vector.

To find a formula for the area we use Green's theorem in the plane

$$\begin{aligned}\operatorname{area}(\Omega) &= \int\int_{\Omega} 1dxdy \\ &= \frac{1}{2}\left|\int_0^{2\pi} \operatorname{Re}\left(f\,(t)\,|i\,(Df)\,(t)\right)dt\right| \\ &= \frac{1}{2}2\pi\left|\frac{1}{2\pi}\int_0^{2\pi} \operatorname{Re}\left(f\,(t)\,|i\,(Df)\,(t)\right)dt\right| \\ &= \pi\left|\operatorname{Re}\left(f|iDf\right)\right|.\end{aligned}$$

Cauchy-Schwarz then implies that

$$\begin{aligned}\operatorname{area}(\Omega) &= \pi\left|\operatorname{Re}\left(f|iDf\right)\right| \\ &\leq \pi\left\|f\right\|_2\left\|iDf\right\|_2 \\ &= \pi\left\|f\right\|_2\left\|Df\right\|_2 \\ &= \pi\left\|f\right\|_2\end{aligned}$$

Now translate the region, so that $\int_0^{2\pi} f\,(t)\,dt = 0$. This can be done without affecting the area and differential so the above formula for the area still holds. Wirtinger's inequality then implies that

$$\begin{aligned}\operatorname{area}(\Omega) &\leq \pi\left\|f\right\|_2 \\ &\leq \pi\left\|Df\right\|_2 \\ &= \pi,\end{aligned}$$

which is what we wanted to prove. In case the length of the perimeter is $2\pi R$ we need to scale the parameter so that the function remains 2π periodic. This means that f looks like $f\,(t\cdot R)$ and $|Df| = R$. With this change the argument is easily repeated.

This proof also yields the rigidity statement that only the circle has maximal area with fixed circumference. To investigate that we observe that equality in Wirtinger's inequality occurs only when $f\,(t) = f_1\exp\left(it\right) + f_{-1}\exp\left(-it\right)$. The condition that the curve was parametrized by arclength then implies

$$\begin{aligned}1 &= \left|Df\,(t)\right|^2 \\ &= \left|if_1\exp\left(it\right) - if_{-1}\exp\left(-it\right)\right|^2 \\ &= \left|f_1\right|^2 + \left|f_2\right|^2 - 2\operatorname{Re}\left(f_1\overline{f_{-1}}\exp\left(2it\right)\right)\end{aligned}$$

Since $\operatorname{Re}\left(\exp\left(2it\right)\right)$ is not constant in t we conclude that either $f_1 = 0$ or $f_{-1} = 0$. Thus $f\,(t) = f_{\pm 1}\exp\left(\pm it\right)$ parametrizes a circle.

8.1 Exercises

1. Study the differential equation $p\,(D)\,(x) = (D-i)\,(D+2i)\,(x) = g\,(t)$. Find the kernel, image, the constant in Wirtinger's inequality etc.

2. Consider a differential equation $p(D)(x) = g(t)$ such that the homogeneous equation $p(D)(x) = 0$ has a solution in $C_{2\pi}^{\infty}(\mathbb{R}, \mathbb{C})$. If $g(t) \in C_{2\pi}^{\infty}(\mathbb{R}, \mathbb{C})$ show that the inhomogeneous equation has either infinitely many or no solutions in $C_{2\pi}^{\infty}(\mathbb{R}, \mathbb{C})$.

9 Calculating the Jordan Canonical Form

The purpose of this section is to elaborate on the proof of the Jordan canonical form. The goal is to give an algorithm that for each eigenvalue computes the number of blocks of a given size in the Jordan Canonical form. Recall that we explained what happens in dimension 2 and 3 in chapter 2 "The Jordan Canonical From" so we are mostly concerned with the higher dimensional cases here. Initially we shall simply consider a *nilpotent* operator $N : V \to V$, i.e., $N^k = 0$ for some k. This implies that $\lambda = 0$ is the only possible eigenvalue and hence that $\chi_N(t) = t^n$. The cyclic subspace decomposition gives us a direct sum decomposition $V = C_{x_1} \oplus \cdots \oplus C_{x_s}$, where $\dim C_{x_1} \geq \cdots \geq \dim C_{x_s}$ and each C_{x_i} corresponds to a Jordan block of size $\dim C_{x_i}$. This gives us a *partition* of n

$$\begin{aligned} n &= \dim C_{x_1} + \cdots + \dim C_{x_s}, \text{ where} \\ \dim C_{x_1} &\geq \cdots \geq \dim C_{x_s}. \end{aligned}$$

The goal of this section is to find this partition from a different set of numbers that are simpler to calculate. The numbers are the following dimensions.

$$\begin{aligned} k_1 &= \dim(\ker(N)) \\ k_2 &= \dim\left(\ker\left(N^2\right)\right) \\ &\vdots \\ k_n &= \dim(\ker(N^n)) \end{aligned}$$

We already know that $C_{x_i} \cap \dim(\ker(N))$ is one dimensional as each cyclic subspace has one dimensional eigenspace for $\lambda = 0$. Thus $k_1 = s$ gives the number of Jordan blocks. In a similar way we observe that $k_2 - k_1$ must give us the number of Jordan blocks of size ≥ 2. And more generally $k_i - k_{i-1}$ is the number of Jordan blocks of size $\geq i$. Thus we see that the number of Jordan blocks of size i must be given by the number $(k_i - k_{i-1}) - (k_{i+1} - k_i)$. These are precisely the numbers we wish to find and they can clearly be found from just knowing $k_1, ..., k_n$.

This information gives a different partition of n. First find the smallest m such that $k_{m+1} = k_m = n$. Note that this means $m_N(t) = t^m$. We then have a partition

$$\begin{aligned} n &= k_1 + (k_2 - k_1) + \cdots + (k_m - k_{m-1}), \text{ where} \\ k_1 &\geq (k_2 - k_1) \geq \cdots \geq (k_m - k_{m-1}) > 0. \end{aligned}$$

Moreover, this partition tells us the number of blocks of a given size. The information can be encoded in a so called *Young diagram* or *Young tableau* (table)

N	n	k_1	k_2-k_1	$\cdots$	$k_{m-1}-k_{m-2}$	k_m-k_{m-1}
C_{x_1}	m	$\bullet$	$\bullet$	$\cdots$	$\bullet$	$\bullet$
C_{x_2}	m	$\bullet$	$\bullet$	$\cdots$	$\bullet$	$\bullet$
C_{x_3}	$m-1$	$\bullet$	$\bullet$	$\cdots$	$\bullet$	
	$\vdots$	$\vdots$	$\vdots$			
	?	$\bullet$	$\bullet$	$\cdots$		
$C_{x_{s-1}}$	1	$\bullet$				
C_{x_s}	1	$\bullet$				

The first column first records the linear map N and then the decomposition of N into subspaces corresponding to Jordan blocks. The second records the dimensions of these subspaces, thus giving us the first decomposition of n. The second decomposition of n is the first row starting with k_1. Finally the columns headed by $k_i - k_{i-1}$ has $(k_i - k_{i-1})$ dots starting from the top. To find the number of blocks of size i we simply go to the column headed by $k_i - k_{i-1}$ and check how many dots we have at the bottom of that column which do not have a dots to the immediate right. In this way we can by starting from the right hand column find the size of each of the blocks and then record that in the second column.

For $n = 2, 3, 4$ all Young diagrams look like

N	2	2
C_{x_1}	1	$\bullet$
C_{x_2}	1	$\bullet$

,

N	2	1	1
C_{x_1}	2	$\bullet$	$\bullet$

N	3	3
C_{x_1}	1	$\bullet$
C_{x_2}	1	$\bullet$
C_{x_3}	1	$\bullet$

,

N	3	2	1
C_{x_1}	2	$\bullet$	$\bullet$
C_{x_2}	1	$\bullet$	

,

N	3	1	1	1
C_{x_1}	3	$\bullet$	$\bullet$	$\bullet$

N	4	4
C_{x_1}	1	$\bullet$
C_{x_2}	1	$\bullet$
C_{x_3}	1	$\bullet$
C_{x_4}	1	$\bullet$

,

N	4	3	1
C_{x_1}	2	$\bullet$	$\bullet$
C_{x_2}	1	$\bullet$	
C_{x_3}	1	$\bullet$	

,

N	4	2	2
C_{x_1}	2	$\bullet$	$\bullet$
C_{x_2}	2	$\bullet$	$\bullet$

,

N	4	2	1	1
C_{x_1}	3	$\bullet$	$\bullet$	$\bullet$
C_{x_2}	1	$\bullet$		

,

N	4	1	1	1	1
C_{x_1}	4	$\bullet$	$\bullet$	$\bullet$	$\bullet$

The first observation we make is that n and k_1 determine the entire Young diagram in all cases except for the following two situations

N	4	2	2
C_{x_1}	2	●	●
C_{x_2}	2	●	●

,

N	4	2	1	1
C_{x_1}	3	●	●	●
C_{x_2}	1	●		

where $n = 4$ and $k_1 = 2$. In this case it is necessary to compute k_2 in order to find the Jordan block structure. The prototypical example of two 4×4 matrices that conform to those two Young diagrams are

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

This is enough information to compute virtually every Jordan canonical form that can be done by hand. If we start with a linear transformation L such that $\chi_L(t) = (t - \lambda_1)^{n_1} \cdots (t - \lambda_k)^{n_k}$ and we know that $n_1, ..., n_k \leq 3$, then we can find the Jordan canonical form by simply calculating $k_1(\lambda_i) = \dim(\ker(L - \lambda_i 1_V))$. The Jordan block structure corresponding to the eigenvalue λ_i is then determined by the two numbers n_i and $k_1(\lambda_i)$. In the more general case were we allow $n_i = 4$ and $k_1(\lambda_i) = 2$, we need to also compute $k_2(\lambda_i) = \dim\left(\ker\left((L - \lambda_i 1_V)^2\right)\right)$ to decide the Jordan block structure.

All of these investigations also lead us to a procedure for deciding when two matrices are similar. We say that two linear maps $L_1 : V_1 \to V_1$ and $L_2 : V_2 \to V_2$ are *similar* if there is an isomorphism $K : V_1 \to V_2$ such that $L_2 = K \circ L_1 \circ K^{-1}$. Clearly this is equivalent to finding bases for V_1 and V_2 such that the matrix representations are the same

$$\begin{tikzcd}
V_1 & \mathbb{F}^n \arrow[l] \arrow[r] & V_2 \\
V_1 \arrow[u, "L_1"'] & \mathbb{F}^n \arrow[l] \arrow[r] \arrow[u, "{[L_1]}", "{[L_2]}"'] & V_2 \arrow[u, "L_2"']
\end{tikzcd}$$

Note that unitary equivalence is a much stronger condition, which in any case only makes sense when we have inner products. Similarity is the correct concept when we are in abstract vector spaces. With this definition it is clear that two linear maps with characteristic polynomials that split are similar if and only if they have the same Jordan canonical form. Thus the following theorem completely determines the similarity type of a linear operator.

Theorem 9.1. (Construction of the Jordan Canonical Form) *Let $L : V \to V$ be a linear operator on an n dimensional vector space. If*

$$\chi_L(t) = (t - \lambda_1)^{n_1} \cdots (t - \lambda_k)^{n_k}$$

splits, then the Jordan canonical form is completely determined by the numbers

$$\begin{aligned} k_1(\lambda_i) &= \dim(\ker(L-\lambda_i 1_V)), \\ k_2(\lambda_i) &= \dim\left(\ker\left((L-\lambda_i 1_V)^2\right)\right), \\ &\vdots \\ k_{n_i}(\lambda_i) &= \dim(\ker((L-\lambda_i 1_V)^{n_i})), \end{aligned}$$

where $i = 1, ..., k$. Moreover, the Jordan block structure for the eigenvalue λ_i can be found from the Young diagram associated with the decomposition

$$n_i = k_1(\lambda_i) + (k_2(\lambda_i) - k_1(\lambda_i)) + \cdots + (k_{m_i}(\lambda_i) - k_{m_i-1}(\lambda_i)),$$

where m_i is the smallest integer so that

$$k_{m_i}(\lambda_i) = \dim(\ker((L-\lambda_i 1_V)^{m_i})) = n_i.$$

In the "The Smith Normal Form" we give a different procedure for determining the similarity type of a linear operator. This procedure does not depend on assuming that the characteristic polynomial splits. However, in case $\mathbb{F} \subset \mathbb{C}$ it will turn out that two linear operators are similar if and only if their matrix representations are similar as complex matrices. Thus the above result is far more general than it appears.

First let us see that the minimal and characteristic polynomials are not sufficient information if we seek to find the similarity type of a linear operator. On $\mathbb{F}^4$ the two matrices

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

have $\chi(t) = t^4$ and $m(t) = t^2$, but they are not similar as $\dim(\ker(A)) = 3$, while $\dim(\ker(B)) = 2$. The Young diagrams for these two matrices look like

A	4	3	1
C_{x_1}	2	●	●
C_{x_2}	1	●	
C_{x_3}	1	●	

B	4	2	2
C_{x_1}	2	●	●
C_{x_2}	2	●	●

Finally let us give a more computationally challenging example.

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

We see immediately that $\chi(t)=t^4(t-1)^2(t-2)^2$. We then seek to find the dimensions of the kernels $\ker(A), \ker\left(A^2\right), \ker\left(A-1_{\mathbb{F}^8}\right), \ker\left(A-2\cdot 1_{\mathbb{F}^8}\right)$. From A it self we see that $\dim(\ker(A))=2$. We then calculate (in fact since A is in block form we only need to worry about the upper left hand 4×4 block.)

$$A^2=\begin{bmatrix} 0 & 0 & 1 & -1 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -4 & 3 \\ 0 & 0 & 0 & 0 & 1 & -2 & -3 & 6 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & -4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}$$

which clearly has $\dim\left(\ker\left(A^2\right)\right)=3$. Thus the Jordan blocks for 0 are

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, [0].$$

Next we calculate $A-1_{\mathbb{F}^8}$ and $A-2\cdot 1_{\mathbb{F}^8}$

$$A-1_{\mathbb{F}^8}=\begin{bmatrix} -1 & -1 & 1 & -1 & 0 & 1 & -1 & -1 \\ 0 & -1 & -1 & 1 & 0 & -1 & 1 & 1 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A-2\cdot 1_{\mathbb{F}^8}=\begin{bmatrix} -2 & -1 & 1 & -1 & 0 & 1 & -1 & -1 \\ 0 & -2 & -1 & 1 & 0 & -1 & 1 & 1 \\ 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 & 2 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

each of which has 1 dimensional kernel. Thus the Jordan blocks are

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

This means that the Jordan canonical form for A is

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

10 The Rational Canonical Form

The purpose of this section is to explain what can be said about linear maps when the characteristic polynomial doesn't split. From the minimal polynomial section we know that one can expect to decompose any $L : V \to V$ into L invariant subspaces V_i such that $V = \oplus_i V_i$ and $L|_{V_i}$ has the property that $\chi_{L|_{V_i}}(t) = m_{L|_{V_i}}(t)$. Having achieved this we are then left with the problem of discovering the simplest matrix representation for a linear operator $L : V \to V$ where $\chi_L(t) = m_L(t)$. At that level of generality it would take some work to answer this question (see exercises). Instead we are going to restrict ourselves to a simpler case as in the case of the Jordan blocks where $\chi_L(t) = m_L(t) = (t-\lambda)^n$. In the more general situation at hand where we can't necessarily guarantee roots of polynomials we shall decompose V into subspaces where $\chi_{L|_{V_i}}(t) = m_{L|_{V_i}}(t) = (p_i(t))^{m_i}$ and $p_i(t)$ cannot be written as a product of polynomials in $\mathbb{F}[t]$. To clarify the importance of these polynomials we introduce some notation. The proofs of the facts we use are covered in "polynomials" from chapter 2.

We say that a monic polynomial $p \in \mathbb{F}[t]$ is *irreducible* if p cannot be written as a product of two monic polynomials of degree ≥ 1 in $\mathbb{F}[t]$. Thus the irreducible polynomials in $\mathbb{C}[t]$ are precisely the linear ones $t-\lambda$. While the irreducible polynomials in $\mathbb{R}[t]$ look like $(t-\lambda)$ or $t^2+\alpha t+\beta$, where $\alpha^2 - 4\beta < 0$. There is also a relatively simple way of checking whether a polynomial in $\mathbb{Q}[t]$ is irreducible. This is the so called *Eisenstein criterion.*

Lemma 10.1. *If we take a monic polynomial $q(t) \in \mathbb{Q}[t]$ and multiply it by an integer so that all coefficients become integers*

$$kq(t) = a_n t^n + \cdots a_1 t + a_0,$$

then q is irreducible provided we can find a prime number p so that p does not divide a_n, p divides all of the other coefficients $a_{n-1}, ..., a_1, a_0$, and p^2 does not divide a_0.

Just as one can always factor an integer into prime factors one can also factor a monic polynomial $p(t) \in \mathbb{F}[t]$ into irreducible factors

$$p(t) = (p_1(t))^{n_1} \cdots (p_k(t))^{n_k}.$$

If we fix a linear operator $L : V \to V$ and factor its characteristic polynomial

$$\chi_L(t) = (p_1(t))^{n_1} \cdots (p_k(t))^{n_k},$$

then it is natural to suppose that the L invariant subspaces $\ker(p(L))$ and $\ker\left((p(L))^k\right)$, where $p(t)$ is some irreducible factor of $\chi_L(t)$, in a natural way replace $\ker(L - \lambda 1_V)$ and $\ker\left((L - \lambda 1_V)^k\right)$.

The following proposition shows to what extend irreducible factors of the characteristic polynomial mimic eigenvalues. The keen reader will, however, observe that the proof of the rational canonical form below does not depend on this nice characterization.

Proposition 10.2. *Let $p(t) \in \mathbb{F}[t]$ be an irreducible polynomial and $L : V \to V$ a linear operator on a finite dimensional vector space. Then the following conditions are equivalent.*

1. $\ker(p(L)) \neq \{0\}$.

2. $p(t)$ *divides* $m_L(t)$.

3. $p(t)$ *divides* $\chi_L(t)$.

Proof. Note that if 1 holds then the minimal polynomial for $L|_{\ker(p(L))}$ must divide $p(t)$ and therefore be $p(t)$ as $p(t)$ was assumed to be irreducible. Conversely if 2 holds then we have

$$0 = m_L(L) = p(L)\, q(L)$$

for some $q(t) \in \mathbb{F}[t]$. If $\ker(p(L)) = \{0\}$, then $p(L)$ must be an isomorphism and hence it must follow that $q(L) = 0$. But this means that $m_L(t)$ divides $q(t)$ which contradicts that $\deg(p(t)) \geq 1$. Thus 1. and 2. are equivalent. Moreover since $m_L(t)$ divides $\chi_L(t)$ it will also follow that $p(t)$ divides $\chi_L(t)$ provided it divides $m_L(t)$.

The final step that 3. implies 2. is a little more involved. First pick a cyclic subspace decomposition of V. Thus L has a matrix representation

$$[L] = \begin{bmatrix} C_{p_1} & 0 & & 0 \\ 0 & C_{p_2} & & \\ & & \ddots & \\ 0 & & & C_{p_k} \end{bmatrix}$$

where

$$\chi_L(t) = \chi_{[L]}(t) = p_1(t) \cdots p_k(t).$$

Since $p(t)$ is irreducible and divides $\chi_L(t)$ it must also divide one of the polynomials $p_1(t), ..., p_k(t)$. If we assume that $p(t)$ divides $p_i(t)$, then we have that $p(t)$ divides $m_{C_{p_i}}(t) = p_i(t)$. Since $m_{C_{p_i}}(t)$ divides $m_L(t)$, this shows that $p(t)$ divides $m_L(t)$. □

Corollary 10.3. *Let $L : V \to V$ be a linear operator on an n-dimensional space, then*

$$\begin{aligned} \chi_L(t) &= (p_1(t))^{n_1} \cdots (p_k(t))^{n_k}, \\ m_L(t) &= (p_1(t))^{m_1} \cdots (p_k(t))^{m_k}, \end{aligned}$$

where $p_i(t)$ are irreducible, $1 \leq m_i \leq n_i \leq n$, and

$$n_1 \deg(p_1(t)) + \cdots + n_k \deg(p_k(t)) = n.$$

The proof of how a linear transformation can be reduced and given a canonical form at this general level now follows the outline that was used for the Jordan canonical form. Namely, we decompose V into cyclic subspaces with the property that no further decompositions are possible. We then show that these indecomposable blocks give a simple matrix structure for L.

The reduction process works by induction on $\dim V$. Thus we fix a linear operator $L : V \to V$. Let

$$m_L(t) = (p_1(t))^{m_1} \cdots (p_k(t))^{m_k}$$

be the factorization of $m_L(t)$ into irreducible polynomials $p_i(t)$. Then we have an L invariant decomposition

$$V = \ker((p_1(L))^{m_1}) \oplus \cdots \oplus \ker((p_k(L))^{m_k}).$$

Therefore we can restrict our efforts to the situation where $L : V \to V$ is a linear operator with $m_L(t) = (p(t))^m$ and $p(t)$ is irreducible in $\mathbb{F}[t]$. Next we use the cyclic subspace decomposition to decompose V further. This reduces us to the situation where $L : V \to V$ is a linear operator with $m_L(t) = (p(t))^m = \chi_L(t)$. For such operators we claim that no further decomposition of V is possible:

Lemma 10.4. *Let $L : V \to V$ be a linear operator with $m_L(t) = (p(t))^m = \chi_L(t)$, where $p(t)$ is irreducible in $\mathbb{F}[t]$. Then there are no nontrivial L invariant decompositions $V = M \oplus N$. Moreover V has a cyclic basis which makes $[L]$ into a companion matrix.*

Proof. First suppose that we have an L invariant decomposition $V = M \oplus N$. Because $p(t)$ is irreducible it follows that

$$\begin{aligned} \chi_{L|_M}(t) &= (p(t))^k, \\ \chi_{L|_N}(t) &= (p(t))^l, \end{aligned}$$

where $l + k = m$. Then

$$\begin{aligned} m_{L|_M}(t) &= (p(t))^{k'}, \\ m_{L|_N}(t) &= (p(t))^{l'}, \end{aligned}$$

where $k' \leq k$ and $l' \leq l$. If $z = x + y$ where $x \in M$ and $y \in N$, then

$$\begin{aligned}
(p(L))^{\max\left\{k',l'\right\}}(z) &= (p(L))^{\max\left\{k',l'\right\}}(x) + (p(L))^{\max\left\{k',l'\right\}}(y) \\
&= (p(L|_M))^{\max\left\{k',l'\right\}}(x) + (p(L|_N))^{\max\left\{k',l'\right\}}(y) \\
&= 0 + 0.
\end{aligned}$$

But then it must follow that $(p(L))^{\max\left\{k',l'\right\}} = 0$. Hence $\max\{k', l'\} = m$ and thus $M = V$ or $N = V$ as desired.

This clearly implies that the only cyclic subspace decomposition of V is the trivial one $V = C_x$. This finishes the proof. □

Theorem 10.5. (The Rational Canonical Form) *Let $L : V \to V$ be a linear operator, Then we can find an L invariant decomposition*

$$V = C_{x_1} \oplus \cdots \oplus C_{x_s}$$

where

$$\chi_{L|_{C_{x_i}}}(t) = m_{L|_{C_{x_i}}}(t) = (p_i(t))^{k_i}$$

and $p_i(t) \in \mathbb{F}[t]$ are monic irreducible polynomials.

To see how such canonical forms work in practice let us consider linear operators of finite order as in "Diagonalizability Redux". Thus we have a linear operator $L : V \to V$ such that $L^k = 1_V$. We start by assuming that $\mathbb{F} = \mathbb{R}$. The minimal polynomial divides $t^k - 1$. So to find the potential irreducible factors of m_L we seek the irreducible factors of $t^k - 1$. We always have that $t - 1$ divides $t^k - 1$, and when k is even also $t + 1$ divides $t^k - 1$. Otherwise we know that the complex polynomials come in conjugate pairs that look like $e^{i2\pi\frac{l}{k}}, e^{-i2\pi\frac{l}{k}}$ where $0 < l < k/2$. Thus

$$\left(t - e^{i2\pi\frac{l}{k}}\right)\left(t - e^{-i2\pi\frac{l}{k}}\right) = t^2 - 2\cos\left(2\pi\frac{l}{k}\right)t + 1$$

is an irreducible factor. We also see that each of these irreducible factors only occurs once. Thus L has a rational canonical form where the blocks look like

$$[1], [-1], \begin{bmatrix} 0 & -1 \\ 1 & 2\cos\left(2\pi\frac{l}{k}\right) \end{bmatrix}, \text{ where } 0 < l < k/2.$$

In case $\mathbb{F} = \mathbb{Q}$ things are a good deal more complicated. The irreducible factorization of $t^k - 1$ is something that is discussed in more advanced algebra courses covering Galois Theory. We are going to consider the cases where $k = 2, 3, 4, 5$ as examples of what might happen.

When $k = 2$ we have $t^2 - 1 = (t - 1)(t + 1)$. Thus L is diagonalizable with eigenvalues 1 and/or -1.

When $k = 3$ we have $t^3 - 1 = (t - 1)\left(t^2 + t + 1\right)$. Here $t^2 + t + 1$ does not have rational roots and is therefore irreducible. Thus the blocks in the rational form look like

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

When $k = 4$ we have $t^4 - 1 = (t-1)(t+1)\left(t^2+1\right)$ and $t^2 + 1$ is irreducible. Thus the blocks in the rational form look like

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

Finally when $k = 5$ we have $t^5 - 1 = (t-1)\left(t^4 + t^3 + t^2 + t + 1\right)$. Here we can with a little work show that $p(t) = t^4 + t^3 + t^2 + t + 1$ is irreducible. As it stands we cannot apply Eisenstein's criterion. However we have that

$$\begin{aligned} p(t+1) &= (t+1)^4 + (t+1)^3 + (t+1)^2 + (t+1) + 1 \\ &= t^4 + 5t^3 + 10t^2 + 10t + 5. \end{aligned}$$

Thus 5 doesn't divide the coefficient in front of t^4, it does divide all the other coefficients, and finally 5^2 does not divide the constant term. This implies that $p(t+1)$ is irreducible. This shows that $p(t)$ must also be irreducible. Thus the blocks in the rational form look like

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

11 Control Theory

While this chapter has been quite abstract and theoretical it is in fact no less useful for applications. As an example where most of the notions introduced both in this chapter and in the chapters on inner products can be used we have chosen control theory. The account is just very brief overview.

The idea of control theory is to start with a state space V and an operator $L : V \to V$. The state space V can be the three dimensional space we live in, or the number of animals of different species being studied, or a collection of relevant economic indicators. The operator L then dictates how the system changes with time if we make no interventions. Here we will only consider linear operators L.

In this simplistic set-up time is discrete and the iterates of an initial state $x_0 \in V$:

$$\begin{aligned} x_1 &= L(x_0), \\ x_2 &= L^2(x_0), .. \end{aligned}$$

then describes how this states evolves in time without external influence. We are interested in forcing the system to behave in a predictable or controlled manner. This means that we need at each time to change how the system evolves in a way that depends on the state of the system. By adding a forcing term $f_n \in V$ to each iteration we see that the system evolves as follows

$$\begin{aligned} x_1 &= L(x_0) + f_1, \\ x_2 &= L(L(x_0) + f_1) + f_2, ... \end{aligned}$$

or

$$\begin{aligned} x_1 &= L(x_0) + f_1, \\ x_2 &= L^2(x_0) + L(f_1) + f_2, \ldots \end{aligned}$$

Note how each forcing term has an effect on the next iteration and consequently on what happens for all future time. Thus applying a forcing term at time 1 will affect the system forever. In most realistic problems there will be some natural constraints on what types of forcings can be allowed. A fairly realistic assumption is that

$$\begin{aligned} f_k &= \begin{bmatrix} b_1 & \cdots & b_m \end{bmatrix} u_k \\ &= \begin{bmatrix} b_1 & \cdots & b_m \end{bmatrix} \begin{bmatrix} u_{1k} \\ \vdots \\ u_{mk} \end{bmatrix} \end{aligned}$$

where the scalars $u_{1k}, ..., u_{mk} \in \mathbb{F}$ can be chosen freely at each time and the vectors $b_1, ..., b_m \in V$ remain fixed. We are therefore forced to exert control only in the subspace generated by $b_1, ..., b_m$. If we denote vectors in $\mathbb{F}^m$ by u and the linear map $\begin{bmatrix} b_1 & \cdots & b_m \end{bmatrix} : \mathbb{F}^m \to V$ by B, then we see that the iterations look like

$$\begin{aligned} x_0 &= x_0 \\ x_1 &= L(x_0) + Bu_1, \\ x_2 &= L^2(x_0) + LBu_1 + Bu_2, \\ &\vdots \\ x_n &= L^n(x_0) + L^{n-1}Bu_1 + \cdots + LBu_{n-1} + Bu_n \\ &= L^n(x_0) + \sum_{k=1}^{n} L^{n-k}Bu_k. \end{aligned}$$

Given L and B our problem is to determine whether it is possible to choose $u_1, ..., u_n \in \mathbb{F}^m$ so that starting at x_0 we can get to a desired state x_n at time n. We also want to decide how small we can choose n when going from x_0 to x_n.

The first observation we make here is that $x_k \in \operatorname{im}(L) + \operatorname{im}(B)$, for $k = 1, ..., n$. Thus we must assume that these two images are transversal, i.e., $\operatorname{im}(L) + \operatorname{im}(B) = V$, in order to get to an arbitrarily chosen point $x_n \in V$. More generally we have.

Lemma 11.1. *Given $L : V \to V$ and a control $B : \mathbb{F}^m \to V$, we can go from any $x_0 \in V$ to any other $x_n \in V$ in time n if and only if*

$$V = \operatorname{im}\left(L^{n-1}B\right) + \cdots + \operatorname{im}(LB) + \operatorname{im}(B).$$

Proof. Assuming that we can get from $x_0 = 0$ to any $x_n \in V$ in time n, it must follow that

$$V = \operatorname{im}\left(L^{n-1}B\right) + \cdots + \operatorname{im}(LB) + \operatorname{im}(B)$$

as

$$x_n = \sum_{k=1}^{n} L^{n-k} B u_k.$$

Conversely if

$$V = \operatorname{im}\left(L^{n-1}B\right) + \cdots + \operatorname{im}(LB) + \operatorname{im}(B),$$

then we can, given $x_0, x_n \in V$, choose $u_1, ..., u_n$ so that

$$x_n - L^n(x_0) = \sum_{k=1}^{n} L^{n-k} B u_n.$$

□

With this in mind it is perhaps becoming clear that companion matrices, cyclic subspaces, etc. can be quite helpful in investigating this problem. If we use $B = \begin{bmatrix} b_1 & \cdots & b_m \end{bmatrix}$, then we see that

$$\begin{aligned} \operatorname{im}\left(L^{n-1}B\right) + \cdots + \operatorname{im}(LB) + \operatorname{im}(B) &= \operatorname{span}\left\{b_1, L(b_1), ..., L^{n-1}(b_1)\right\} \\ &\quad + \cdots + \operatorname{span}\left\{b_m, L(b_m), ..., L^{n-1}(b_m)\right\}. \end{aligned}$$

This means that as long as $n \geq \deg(m_L)$ then

$$\operatorname{im}\left(L^{n-1}B\right) + \cdots + \operatorname{im}(LB) + \operatorname{im}(B) = C_{b_1} + \cdots + C_{b_m}.$$

Therefore, if we can get from any x_0 to any x_n in n steps, then we can also do it in $\deg(m_L)$ steps. Thus the degree of the minimal polynomial tells us the minimum time or number of steps we should expect to take. For some initial conditions x_0 we might of course be able to do this faster.

Next we see that m must be at least the same as the smallest number of subspaces in a cyclic subspace decomposition for L. Since companion matrices always have one dimensional eigenspaces we see that

$$m \geq \max\left\{\dim\left(\ker\left(L - \lambda 1_V\right)\right) : \lambda \subset \mathbb{F}\right\}.$$

In case L doesn't have any eigenvalues or m_L doesn't split we get the refined formula

$$m \geq \max\left\{\dim\left(\ker\left(p(L)\right)\right) : p(t) \in \mathbb{F}[t] \text{ is irreducible}\right\}.$$

Note that if $p(t)$ is irreducible and doesn't divide $m_L(t)$, then $\dim(\ker(p(L))) = 0$. So only the irreducible factors of m_L are relevant when computing the maximum.

All of this means that we can use the minimal polynomial, its factorization, and corresponding eigenspaces to decide how to choose $b_1, ..., b_m$ so as to solve our problem of moving from x_0 to x_m.

When V is an inner product space we can also find a minimum energy control using the Moore-Penrose inverse.

If

$$x_n - L^n\left(x_0\right) = \sum_{k=1}^{n} L^{n-k} B u_k$$

then the *energy* that it costs to go from x_0 to x_n using the controls $u_1, ..., u_n \in \mathbb{F}^m$ is simply

$$\sum_{k=1}^{n} \left\| u_k \right\|^2.$$

We wish to select the controls so that this energy is as small as possible. To this end define

$$\begin{aligned} Y_n &: \mathbb{F}^{m\cdot n} \to V, \\ Y_n\left(u_1, ..., u_n\right) &= \sum_{k=1}^{n} L^{n-k} B u_k \end{aligned}$$

where we think of elements in $\mathbb{F}^{m\cdot n}$ as n-tuples $U_n = \left(u_1, ..., u_n\right)$ of vectors in $\mathbb{F}^m$, or in other words $m \times n$ matrices. The inner product on $\mathbb{F}^{m\cdot n}$ is given by

$$\left(\left(u_1, ..., u_n\right) \middle| \left(v_1, ..., v_n\right)\right) = \sum_{k=1}^{n} \left(u_k | v_k\right).$$

We are therefore trying to find the solution to

$$Y_n U_n = x_n - L^n\left(x_0\right)$$

where $U_n \in \mathbb{F}^{m\cdot n}$ is smallest possible. We know that the Moore-Penrose inverse gives the answer to this question. Thus

$$U_n = Y_n^{\dagger}\left(x_n - L^n\left(x_0\right)\right).$$

This solution can alternatively be described as the unique solution that lies in $\left(\ker\left(Y_n\right)\right)^{\perp} = \operatorname{im}\left(Y_n^*\right)$. It is therefore natural to seek a solution of the form

$$U_n = Y_n^* x$$

where $x \in V$. Note that

$$\begin{aligned} Y_n^* x &= \left(\left(L^{n-1}B\right)^* x, ..., B^* x\right) \\ &= \left(B^*\left(L^{n-1}\right)^* x, ..., B^* x\right) \end{aligned}$$

and

$$Y_n Y_n^* x = \sum_{k=1}^{n} L^{n-k} B B^* \left(L^{n-k}\right)^* x.$$

We are therefore looking for solutions $x \in V$ to the equation

$$Y_n Y_n^* x = x_n - L^n\left(x_0\right).$$

Since

$$(Y_n Y_n^* x, x) = \|Y_n^* x\|^2$$

we have that

$$\ker (Y_n Y_n^*) = \ker (Y_n^*) = \operatorname{im} (Y_n)^{\perp} .$$

Thus $Y_n Y_n^* : V \to V$ is an isomorphism provided Y_n is onto and the equation

$$Y_n Y_n^* x = x_n - L^n (x_0)$$

will have a unique solution, which has the property that $Y_n^* x$ consists of the minimal energy controls that take us from x_0 to x_n in n steps.