Math 554
Linear Analysis
Autumn 2006
Lecture Notes

Ken Bube and James Burke

April 7, 2021

Contents

Linear Algebra and Matrix Analysis

Vector Spaces

Throughout this course, the base field $\mathbb{F}$ of scalars will be $\mathbb{R}$ or $\mathbb{C}$. Recall that a vector space is a nonempty set V on which are defined the operations of addition (for v, $w \in V$, $v+w \in V$) and scalar multiplication (for $\alpha \in \mathbb{F}$ and $v \in V$, $\alpha v \in V$), subject to the following conditions:

1. $x + y = y + x$
2. $(x + y) + z = x + (y + z)$
3. There exists an element $0 \in V$ such that $x + 0 = x$ for all x
4. For each $x \in V$, there is an element of V denoted $-x$ such that $x + (-x) = 0$
5. $\alpha(\beta x) = (\alpha\beta)x$
6. $\alpha(x + y) = \alpha x + \alpha y$
7. $(\alpha + \beta)x = \alpha x + \beta x$
8. $1x = x$

A subset $W \subset V$ is a *subspace* if W is closed under addition and scalar multiplication, so W inherits a vector space structure of its own.

Examples:

(1) $\{0\}$

(2) $\mathbb{F}^n = \left\{ \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} : \text{ each } x_j \in \mathbb{F} \right\}, \quad n \geq 1$

(3) $\mathbb{F}^{m\times n} = \left\{ \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} : \text{ each } a_{ij} \in \mathbb{F} \right\}, \quad m, n \geq 1$

(4) $\mathbb{F}^\infty = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix} : \text{each } x_j \in \mathbb{F} \right\}$

(5) $\ell^1(\mathbb{F}) \subset \mathbb{F}^\infty$, where $\ell^1(\mathbb{F}) = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix} : \sum_{j=1}^\infty |x_j| < \infty \right\}$

$\ell^\infty(\mathbb{F}) \subset \mathbb{F}^\infty$, where $\ell^\infty(\mathbb{F}) = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix} : \sup_j |x_j| < \infty \right\}$

$\ell^1(\mathbb{F})$ and $\ell^\infty(\mathbb{F})$ are clearly subspaces of $\mathbb{F}^\infty$.

Let $0 < p < \infty$, and define $\ell^p(\mathbb{F}) = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix} : \sum_{j=1}^\infty |x_j|^p < \infty \right\}$.

Since

$$\begin{aligned} |x+y|^p &\le (|x|+|y|)^p \le (2\max(|x|,|y|))^p \\ &= 2^p \max(|x|^p, |y|^p) \le 2^p(|x|^p + |y|^p), \end{aligned}$$

it follows that $\ell^p(\mathbb{F})$ is a subspace of $\mathbb{F}^\infty$.

Exercise: Show that $\ell^p(\mathbb{F}) \subsetneqq \ell^q(\mathbb{F})$ if $0 < p < q \le \infty$.

(6) Let X be a nonempty set; then the set of all functions $f : X \to \mathbb{F}$ has a natural structure as a vector space over $\mathbb{F}$: define $f_1 + f_2$ by $(f_1 + f_2)(x) = f_1(x) + f_2(x)$, and define αf by $(\alpha f)(x) = \alpha f(x)$.

(7) For a metric space X, let $C(X, \mathbb{F})$ denote the set of all continuous $\mathbb{F}$-valued functions on X. $C(X, \mathbb{F})$ is a subspace of the vector space defined in (6). Define $C_b(X, \mathbb{F}) \subset C(X, \mathbb{F})$ to be the subspace of all bounded continuous functions $f : X \to \mathbb{F}$.

(8) If $U \subset \mathbb{R}^n$ is a nonempty open set and k is a nonnegative integer, the set $C^k(U, \mathbb{F}) \subset C(U, \mathbb{F})$ of functions all of whose derivatives of order at most k exist and are continuous on U is a subspace of $C(U, \mathbb{F})$. The set $C^\infty(U, \mathbb{F}) = \bigcap_{k=0}^\infty C^k(U, \mathbb{F})$ is a subspace of each of the $C^k(U, \mathbb{F})$.

(9) Define $\mathcal{P}(\mathbb{F}) \subset C^\infty(\mathbb{R}, \mathbb{F})$ to be the space of all $\mathbb{F}$-valued polynomials on $\mathbb{R}$:

$$\mathcal{P}(\mathbb{F}) = \{a_0 + a_1 x + \cdots + a_m x^m : m \ge 0, \text{ each } a_j \in \mathbb{F}\}.$$

Each $p \in \mathcal{P}(\mathbb{F})$ is viewed as a function $p : \mathbb{R} \to \mathbb{F}$ given by $p(x) = a_0 + a_1 x + \cdots + a_m x^m$.

(10) Define $\mathcal{P}_n(\mathbb{F}) \subset \mathcal{P}(\mathbb{F})$ to be the subspace of all polynomials of degree $\le n$.

(11) Let $V = \{u \in C^2(\mathbb{R}, \mathbb{C}) : u'' + u = 0\}$. It is easy to check directly from the definition that V is a subspace of $C^2(\mathbb{R}, \mathbb{C})$. Alternatively, one knows that

$$V = \{a_1 \cos x + a_2 \sin x : a_1, a_2 \in \mathbb{C}\} = \{b_1 e^{ix} + b_2 e^{-ix} : b_1, b_2 \in \mathbb{C}\},$$

from which it is also clear that V is a vector space.

More generally, if $L(u) = u^{(m)} + a_{m-1}u^{(m-1)} + \cdots + a_1 u' + a_0 u$ is an m^{th} order linear constant-coefficient differential operator, then $V = \{u \in C^m(\mathbb{R}, \mathbb{C}) : L(u) = 0\}$ is a vector space. V can be explicitly described as the set of all linear combinations of certain functions of the form $x^j e^{rx}$ where $j \geq 0$ and r is a root of the characteristic polynomial $r^m + a_{m-1}r^{m-1} + \cdots + a_1 r + a_0 = 0$. For details, see Chapter 3 of Birkhoff & Rota.

Convention: Throughout this course, if the field $\mathbb{F}$ is not specified, it is assumed to be $\mathbb{C}$.

Linear Independence, Span, Basis

Let V be a vector space. A *linear combination* of the vectors $v_1, \ldots, v_m \in V$ is a vector $v \in V$ of the form

$$v = \alpha_1 v_1 + \cdots + \alpha_m v_m$$

where each $\alpha_j \in \mathbb{F}$. Let $S \subset V$ be a subset of V. S is called *linearly independent* if for every finite subset $\{v_1, \ldots, v_m\}$ of S, the linear combination $\sum_{i=1}^m \alpha_i v_i = 0$ iff $\alpha_1 = \cdots = \alpha_m = 0$. Otherwise, S is called *linearly dependent.* Define the *span of* S (denoted Span(S)) to be the set of all linear combinations of all finite subsets of S. (Note: a linear combination is by definition a *finite* sum.) If $S = \emptyset$, set $\operatorname{Span}(S) = \{0\}$. S is said to be a *basis* of V if S is linearly independent and $\operatorname{Span}(S) = V$.

Facts: (a) Every vector space has a basis; in fact if S is any linearly independent set in V, then there is a basis of V containing S. The proof of this in infinite dimensions uses Zorn's lemma and is nonconstructive. Such a basis in infinite dimensions is called a *Hamel basis.* Typically it is impossible to identify a Hamel basis explicitly, and they are of little use. There are other sorts of "bases" in infinite dimensions defined using topological considerations which are very useful and which we will consider later.

(b) Any two bases of the same vector space V can be put into $1-1$ correspondence. Define the *dimension* of V (denoted $\dim V$) $\in \{0, 1, 2, \ldots\} \cup \{\infty\}$ to be the number of elements in a basis of V. The vectors $e_1, \ldots, e_n$, where

$$e_j = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow j^{\text{th}} \text{ entry},$$

form the *standard basis* of $\mathbb{F}^n$, and $\dim \mathbb{F}^n = n$.

Remark. Any vector space V over $\mathbb{C}$ may be regarded as a vector space over $\mathbb{R}$ by restriction of the scalar multiplication. It is easily checked that if V is finite-dimensional with basis $\{v_1, \ldots, v_n\}$ over $\mathbb{C}$, then $\{v_1, \ldots, v_n, iv_1, \ldots, iv_n\}$ is a basis for V over $\mathbb{R}$. In particular, $\dim_{\mathbb{R}} V = 2\dim_{\mathbb{C}} V$.

The vectors $e_1, e_2, \ldots \in \mathbb{F}^\infty$ are linearly independent. However, $\text{Span}\{e_1, e_2, \ldots\}$ is the proper subset $\mathbb{F}^\infty_0 \subset \mathbb{F}^\infty$ consisting of all vectors with only finitely many nonzero components. So $\{e_1, e_2, \ldots\}$ is not a basis of $\mathbb{F}^\infty$. But $\{x^m : m \in \{0, 1, 2, \ldots\}\}$ *is* a basis of $\mathcal{P}$.

Now let V be a finite-dimensional vector space, and $\{v_1, \ldots, v_n\}$ be a basis for V. Any $v \in V$ can be written uniquely as $v = \sum\limits_{i=1}^n x_i v_i$ for some $x_i \in \mathbb{F}$. So we can define a map from V into $\mathbb{F}^n$ by $v \mapsto \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$. The x_i's are called the *coordinates* of v with respect to the basis $\{v_1, \ldots, v_n\}$. This coordinate map clearly preserves the vector space operations and is bijective, so it is an isomorphism of V with $\mathbb{F}^n$ in the following sense.

Definition. Let V, W be vector spaces. A map $L : V \to W$ is a *linear transformation* if

$$L(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 L(v_1) + \alpha_2 L(v_2)$$

for all $v_1, v_2 \in V$ and $\alpha_1, \alpha_2 \in \mathbb{F}$. If in addition L is bijective, then L is called a (vector space) *isomorphism.*

Even though every finite-dimensional vector space V is isomorphic to $\mathbb{F}^n$, where $n = \dim V$, the isomorphism depends on the choice of basis. Many properties of V are independent of the basis (e.g. $\dim V$). We could try to avoid bases, but it is very useful to use coordinate systems. So we need to understand how coordinates change when the basis is changed.

Change of Basis

Let V be a finite dimensional vector space. Let $\{v_1, \ldots, v_n\}$ and $\{w_1, \ldots, w_n\}$ be two bases for V. For $v \in V$, let $x = (x_1, \ldots, x_n)^T$ and $y = (y_1, \ldots, y_n)^T$ denote the vectors of coordinates of v with respect to the bases $\mathcal{B}_1 = \{v_1, \ldots, v_n\}$ and $\mathcal{B}_2 = \{w_1, \ldots, w_n\}$, respectively. Here ^{T} denotes the transpose. So $v = \sum_{i=1}^n x_i v_i = \sum_{j=1}^n y_j w_j$. Express each w_j in terms of $\{v_1, \ldots, v_n\} : w_j = \sum_{i=1}^n c_{ij} v_i$ ($c_{ij} \in \mathbb{F}$). Let $C = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & & \\ c_{n1} & \cdots & c_{nn} \end{pmatrix} \in \mathbb{F}^{n \times n}$. Then

$$\sum_{i=1}^n x_i v_i = v = \sum_{j=1}^n y_j w_j = \sum_{i=1}^n \left(\sum_{j=1}^n c_{ij} y_j \right) v_i,$$

so $x_i = \sum_{j=1}^n c_{ij} y_j$, i.e. $x = Cy$. C is called the *change of basis* matrix.

Notation: Horn-Johnson uses $M_{m,n}(\mathbb{F})$ to denote what we denote by $\mathbb{F}^{m \times n}$: the set of $m \times n$ matrices with entries in $\mathbb{F}$. H-J writes $[v]_{\mathcal{B}_1}$ for x, $[v]_{\mathcal{B}_2}$ for y, and ${}_{\mathcal{B}_1}[I]_{\mathcal{B}_2}$ for C, so $x = Cy$

becomes $[v]_{\mathcal{B}_1} = {}_{\mathcal{B}_1}[I]_{\mathcal{B}_2}[v]_{\mathcal{B}_2}$. Similarly, we can express each v_j in terms of $\{w_1, \ldots, w_n\}$: $v_j = \sum_{i=1}^n b_{ij} w_i$ ($b_{ij} \in \mathbb{F}$). Let $B = \begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & & \\ b_{n1} & \cdots & b_{nn} \end{pmatrix} \in \mathbb{F}^{n\times n}$. Then $y = Bx$. We obtain that C and B are invertible and $B = C^{-1}$.

Formal matrix notation: Write the basis vectors $(v_1, \cdots, v_n)$ and $(w_1, \cdots, w_n)$ formally in rows. Then the equations $w_j = \sum_{i=1}^n c_{ij} v_i$ become the formal matrix equation

$$(w_1, \cdots, w_n) = (v_1, \cdots, v_n)C$$

using the usual matrix multiplication rules. In general, $(v_1, \cdots, v_n)$ and $(w_1, \cdots, w_n)$ are not matrices (although in the special case where each v_j and w_j is a column vector in $\mathbb{F}^n$, we have $W = VC$ where $V, W \in \mathbb{F}^{n\times n}$ are the matrices whose columns are the v_j's and the w_j's, respectively). We also have the formal matrix equations $v = (v_1, \cdots, v_n)x$ and $v = (w_1, \cdots, w_n)y$, so

$$(v_1, \cdots, v_n)x = (w_1, \cdots, w_n)y = (v_1, \cdots, v_n)Cy,$$

which gives us $x = Cy$ as before.

Remark. We can read the matrix equation $W = VC$ as saying that the j^{th} column of W is the linear combination of the columns of V whose coefficients are in the j^{th} column of C.

Constructing New Vector Spaces from Given Ones

(1) The intersection of any family of subspaces of V is again a subspace: let $\{W_\gamma : \gamma \in G\}$ be a family of subspaces of V (where G is an index set); then $\bigcap_{\gamma\in G} W_\gamma$ is a subspace of V.

(2) *Sums of subspaces:* If W_1, W_2 are subspaces of V, then

$$W_1 + W_2 = \{w_1 + w_2 : w_1 \in W_1, w_2 \in W_2\}$$

is also a subspace, and $\dim(W_1 + W_2) + \dim(W_1 \cap W_2) = \dim W_1 + \dim W_2$. We say that the sum $W_1 + W_2$ is *direct* if $W_1 \cap W_2 = \{0\}$ (equivalently: for each $v \in W_1 + W_2$, there are unique $w_1 \in W_1$ and $w_2 \in W_2$ for which $v = w_1 + w_2$), and in this case we write $W_1 \oplus W_2$ for $W_1 + W_2$. More generally, if $W_1, \ldots, W_n$ are subspaces of V, then $W_1 + \cdots + W_n = \{w_1 + \cdots + w_n : w_j \in W_j, 1 \le j \le n\}$ is a subspace. We say that the sum is *direct* if whenever $w_j \in W_j$ and $\sum_{j=1}^n w_j = 0$, then each $w_j = 0$, and in this case we write $W_1 \oplus \cdots \oplus W_n$. Even more generally, if $\{W_\gamma : \gamma \in G\}$ is a family of subspaces of V, define $\sum_{\gamma\in G} W_\gamma = \operatorname{span}\left(\bigcup_{\gamma\in G} W_\gamma\right)$. We say that the sum is direct if for each finite subset G' of G, whenever $w_\gamma \in W_\gamma$ for $\gamma \in G'$ and $\sum_{\gamma\in G'} w_\gamma = 0$, then each $w_\gamma = 0$ for $\gamma \in G'$ (equivalently: for each $\beta \in G$, $W_\beta \cap \left(\sum_{\gamma\in G,\gamma\neq\beta} W_\gamma\right) = \{0\}$).

(3) *Direct Products:* Let $\{V_\gamma : \gamma \in G\}$ be a family of vector spaces over $\mathbb{F}$. Define $V = \mathop{\times}\limits_{\gamma\in G} V_\gamma$ to be the set of all functions $v : G \to \bigcup\limits_{\gamma\in G} V_\gamma$ for which $v(\gamma) \in V_\gamma$ for all $\gamma \in G$. We write v_γ for $v(\gamma)$, and we write $v = (v_\gamma)_{\gamma\in G}$, or just $v = (v_\gamma)$. Define $v + w = (v_\gamma + w_\gamma)$ and $\alpha v = (\alpha v_\gamma)$. Then V is a vector space over $\mathbb{F}$. (Example: $G = \mathbb{N} = \{1, 2, \ldots\}$, each $V_n = \mathbb{F}$. Then $\mathop{\times}\limits_{n\geq 1} V_n = \mathbb{F}^\infty$.)

(4) (*External*) *Direct Sums:* Let $\{V_\gamma : \gamma \in G\}$ be a family of vector spaces over $\mathbb{F}$. Define $\bigoplus\limits_{\gamma\in G} V_\gamma$ to be the subspace of $\mathop{\times}\limits_{\gamma\in G} V_\gamma$ consisting of those v for which $v_\gamma = 0$ except for finitely many $\gamma \in G$. (Example: For $n = 0, 1, 2, \ldots$ let $V_n = \operatorname{span}(x^n)$ in $\mathcal{P}$. Then $\mathcal{P}$ can be identified with $\bigoplus\limits_{n\geq 0} V_n$.)

Facts: (a) If G is a finite index set, then $\times V_\gamma$ and $\bigoplus V_\gamma$ are isomorphic.

(b) If each W_γ is a subspace of V and the sum $\sum_{\gamma\in G} W_\gamma$ is direct, then it is naturally isomorphic to the external direct sum $\bigoplus W_\gamma$.

(5) *Quotients:* Let W be a subspace of V. Define on V the equivalence relation $v_1 \sim v_2$ if $v_1 - v_2 \in W$, and define the quotient to be the set V/W of equivalence classes. Let $v + W$ denote the equivalence class of v. Define a vector space structure on V/W by defining $\alpha_1(v_1 + W) + \alpha_2(v_2 + W) = (\alpha_1 v_1 + \alpha_2 v_2) + W$. Define the *codimension* of W in V by $\operatorname{codim}(W) = \dim(V/W)$.

Dual Vector Spaces

Definition. Let V be a vector space. A *linear functional* on V is a function $f : V \to \mathbb{F}$ for which $f(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 f(v_1) + \alpha_2 f(v_2)$ for $v_1, v_2 \in V$, $\alpha_1, \alpha_2 \in \mathbb{F}$. Equivalently, f is a linear transformation from V to the 1-dimensional vector space $\mathbb{F}$.

Examples:

(1) Let $V = \mathbb{F}^n$, and let f be a linear functional on V. Set $f_i = f(e_i)$ for $1 \leq i \leq n$. Then for $x = (x_1, \ldots, x_n)^T = \sum_{i=1}^n x_i e_i \in \mathbb{F}^n$,

$$f(x) = \sum_{i=1}^n x_i f(e_i) = \sum_{i=1}^n f_i x_i$$

So every linear functional on $\mathbb{F}^n$ is a linear combination of the coordinates.

(2) Let $V = \mathbb{F}^\infty$. Given an N and some $f_1, f_2, \ldots, f_N \in \mathbb{F}$, we can define a linear functional $f(x) = \sum_{i=1}^N f_i x_i$ for $x \in \mathbb{F}^\infty$. However, not all linear functionals on $\mathbb{F}^\infty$ are of this form.

(3) Let $V = \ell^1(\mathbb{F})$. If $f \in \ell^\infty(\mathbb{F})$, then for $x \in \ell^1(\mathbb{F})$, $\sum_{i=1}^\infty |f_i x_i| \leq (\sup |f_i|) \sum_{i=1}^\infty |x_i| < \infty$, so the sum $f(x) = \sum_{i=1}^\infty f_i x_i$ converges absolutely, defining a linear functional on $\ell^1(\mathbb{F})$. Similarly, if $V = \ell^\infty(\mathbb{F})$ and $f \in \ell^1(\mathbb{F})$, $f(x) = \sum_{i=1}^\infty f_i x_i$ defines a linear functional on $\ell^\infty(\mathbb{F})$.

(4) Let X be a metric space and $x_0 \in X$. Then $f(u) = u(x_0)$ defines a linear functional on $C(X)$.

(5) If $-\infty < a < b < \infty$, $f(u) = \int_a^b u(x)dx$ defines a linear functional on $C([a,b])$.

Definition. If V is a vector space, the dual space of V is the vector space V' of all linear functionals on V, where the vector space operations on V' are given by $(\alpha_1 f_1 + \alpha_2 f_2)(v) = \alpha_1 f_1(v) + \alpha_2 f_2(v)$.

Remark. When V is infinite dimensional, V' is often called the *algebraic* dual space of V, as it depends only on the algebraic structure of V. We will be more interested in linear functionals related also to a topological structure on V. After introducing norms (which induce metrics on V), we will define V^* to be the vector space of all *continuous* linear functionals on V. (When V is finite dimensional, with any norm on V, every linear functional on V is continuous, so $V^* = V'$.)

Dual Basis in Finite Dimensions

Let V be a finite dimensional vector space and let $\{v_1, \ldots, v_n\}$ be a basis for V. For $1 \leq i \leq n$, define linear functionals $f_i \in V'$ by $f_i(v_j) = \delta_{ij}$, where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

Let $v \in V$, and let $x = (x_1, \ldots, x_n)^T$ be the vector of coordinates of v with respect to the basis $\{v_i, \ldots, v_n\}$, i.e., $v = \sum_{i=1}^n x_i v_i$. Then $f_i(v) = x_i$, i.e., f_i maps v into its coordinate x_i. Now if $f \in V'$, let $a_i = f(v_i)$; then

$$f(v) = f(\sum x_i v_i) = \sum_{i=1}^n a_i x_i = \sum_{i=1}^n a_i f_i(v),$$

so $f = \sum_{i=1}^n a_i f_i$. This representation is unique (exercise), so $\{f_1, \ldots, f_n\}$ is a basis for V', called the *dual basis* to $\{v_1, \ldots, v_n\}$. We get $\dim V' = \dim V$.

If we write the dual basis in a column $\begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}$ and the coordinates $(a_1 \cdots a_n)$ of $f = \sum_{i=1}^n a_i f_i \in V'$ in a row, then $f = (a_1 \cdots a_n) \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}$. The defining equation of the dual basis is (matrix multiply, evaluate)

$$(*) \qquad \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} (v_1 \cdots v_n) = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} = I$$

Change of Basis and Dual Bases: Let $\{w_1, \ldots, w_n\}$ be another basis of V related to the first basis $\{v_1, \ldots, v_n\}$ by the change-of-basis matrix C, i.e., $(w_1 \cdots w_n) = (v_1 \cdots v_n)C$. Left-multiplying $(*)$ by C^{-1} and right-multiplying by C gives

$$C^{-1}\begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}(w_1 \cdots w_n) = I.$$

Therefore

$$\begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix} = C^{-1}\begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} \text{ satisfies } \begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix}(w_1 \cdots w_n) = I$$

and so $\{g_1, \ldots, g_n\}$ is the dual basis to $\{w_1, \ldots, w_n\}$. If $(b_1 \cdots b_n)$ are the coordinates of $f \in V'$ with respect to $\{g_1, \ldots, g_n\}$, then

$$f = (b_1 \cdots b_n)\begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix} = (b_1 \cdots b_n)C^{-1}\begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} = (a_1 \cdots a_n)\begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix},$$

so $(b_1 \cdots b_n)C^{-1} = (a_1 \cdots a_n)$, i.e., $(b_1 \cdots b_n) = (a_1 \cdots a_n)C$ is the transformation law for the coordinates of f with respect to the two dual bases $\{f_1, \ldots, f_n\}$ and $\{g_1, \ldots, g_n\}$.

Linear Transformations

Linear transformations were defined above.

Examples:

(1) Let $T : \mathbb{F}^n \to \mathbb{F}^m$ be a linear transformation. For $1 \leq j \leq n$, write

$$T(e_j) = t_j = \begin{pmatrix} t_{1j} \\ \vdots \\ t_{mj} \end{pmatrix} \in \mathbb{F}^m.$$

If $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{F}^n$, then $T(x) = T(\Sigma x_j e_j) = \Sigma x_j t_j$, which we can write as

$$T(x) = (t_1 \cdots t_n)\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} t_{11} & \cdots & t_{1n} \\ \vdots & & \vdots \\ t_{m1} & \cdots & t_{mn} \end{pmatrix}\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

So every linear transformation from $\mathbb{F}^n$ to $\mathbb{F}^m$ is given by multiplication by a matrix in $\mathbb{F}^{m \times n}$.

(2) One can construct linear transformations $T : \mathbb{F}^\infty \to \mathbb{F}^\infty$ by matrix multiplication. Let

$$T = \left(\begin{array}{ccc} t_{11} & t_{12} & \cdots \\ t_{21} & \ddots & \\ \vdots & & \end{array} \right)$$

be an infinite matrix for which each row has only finitely many nonzero entries. In forming Tx for $x = \left(\begin{array}{c} x_1 \\ \vdots \end{array} \right) \in \mathbb{F}^\infty$, each entry in Tx is given by a finite sum, so Tx makes sense and T clearly defines a linear transformation from $\mathbb{F}^\infty$ to itself. (However, not all linear transformations on $\mathbb{F}^\infty$ are of this form.) The shift operators $(x_1, x_2, \ldots)^T \mapsto (0, x_1, x_2, \ldots)^T$ and $(x_1, x_2, \ldots)^T \mapsto (x_2, x_3, \ldots)^T$ are examples of linear transformations of this form.

(3) If $\sup\limits_{i,j} |t_{ij}| < \infty$ and $x \in \ell^1$, then for each i, $\sum_{j=1}^\infty |t_{ij} x_j| \leq \sup\limits_{i,j} |t_{ij}| \sum_{j=1}^\infty |x_j|$. It follows that matrix multiplication Tx defines a linear transformation $T : \ell^1 \to \ell^\infty$.

(4) There are many ways that linear transformations arise on function spaces, for example:

(a) Let $k \in C([c,d] \times [a,b])$ where $[a,b], [c,d]$ are closed bounded intervals. Define the linear transformation $L : C[a,b] \to C[c,d]$ by $L(u)(x) = \int_a^b k(x,y) u(y) dy$. L is called an *integral operator* and $k(x,y)$ is called its *kernel*.

(b) Let X be a metric space and let $m \in C(X)$. Then $L(u)(x) = m(x)u(x)$ defines a *multiplier operator* L on $C(X)$.

(c) Let X and Y be metric spaces and let $g : X \to Y$ be continuous. Then $L(u)(x) = u(g(x))$ defines a *composition operator* $L : C(Y) \to C(X)$.

(d) $u \mapsto u'$ defines a *differential operator* $L : C^1[a,b] \to C[a,b]$.

Suppose V, W are finite-dimensional with bases $\{v_1, \ldots, v_n\}$, $\{w_1, \ldots, w_m\}$, respectively and suppose $L : V \to W$ is linear. For $1 \leq j \leq n$, we can write $Lv_j = \sum_{i=1}^m t_{ij} w_i$. The matrix

$$T = \left(\begin{array}{ccc} t_{11} & \cdots & t_{1n} \\ \vdots & & \vdots \\ t_{m1} & \cdots & t_{mn} \end{array} \right) \in \mathbb{F}^{m \times n}$$

is called the matrix of L with respect to the bases $\mathcal{B}_1 = \{v_1, \ldots, v_n\}$, $\mathcal{B}_2 = \{w_1, \ldots, w_m\}$ (H-J writes $T = {}_{\mathcal{B}_2}[L]_{\mathcal{B}_1}$.) Let $v \in V$ and let $\left(\begin{array}{c} x_1 \\ \vdots \\ x_n \end{array} \right)$ be the coordinates of v with respect to $\mathcal{B}_1$ and $\left(\begin{array}{c} y_1 \\ \vdots \\ y_m \end{array} \right)$ the coordinates of Lv with respect to $\mathcal{B}_2$. Then

$$\sum_{i=1}^m y_i w_i = Lv = L\left(\sum_{j=1}^n x_j v_j \right) = \sum_{i=1}^m \left(\sum_{j=1}^n t_{ij} x_j \right) w_i,$$

so for $1 \leq i \leq m$, we have $y_i = \sum_{j=1}^n t_{ij} x_j$, i.e. $y = Tx$. Thus, in terms of the coordinates relative to these bases, L is represented by matrix multiplication by T.

Note that the relations defining T can be rewritten as $L(v_1 \cdots v_n) = (w_1 \cdots w_n)T$. Suppose now that we choose different bases $\mathcal{B}_1' = \{v_1', \ldots, v_n'\}$ and $\mathcal{B}_2' = \{w_1', \ldots, w_m'\}$ for V and W, respectively, with change-of-bases matrices $C \in \mathbb{F}^{n\times n}$, $D \in \mathbb{F}^{m\times m}$:

$$(v_1' \cdots v_n') = (v_1 \cdots v_n)C \quad \text{and} \quad (w_1' \cdots w_m') = (w_1 \cdots w_m)D.$$

Then

$$L(v_1' \cdots v_n') = (w_1 \cdots w_n)TC = (w_1' \cdots w_m')D^{-1}TC,$$

so the matrix of L in the new bases is $D^{-1}TC$. In particular, if $W = V$ and we choose $\mathcal{B}_2 = \mathcal{B}_1$ and $\mathcal{B}_2' = \mathcal{B}_1'$, then $D = C$, so the matrix of L in the new basis is $C^{-1}TC$. A matrix of the form $C^{-1}TC$ is said to be *similar* to T. Therefore similar matrices can be viewed as representations of the same linear transformation with respect to different bases.

Linear transformations can be studied abstractly or in terms of matrix representations. For $L : V \to W$, the range $\mathcal{R}(L)$, null space $\mathcal{N}(L)$ (or kernel $\ker(L)$), $\operatorname{rank}(L) = \dim(\mathcal{R}(L))$, etc., can be defined directly in terms of L, or in terms of matrix representations. If $T \in \mathbb{F}^{n\times n}$ is the matrix of $L : V \to V$ in some basis, it is easiest to define $\det L = \det T$ and $\operatorname{tr} L = \operatorname{tr} T$. Since $\det(C^{-1}TC) = \det T$ and $\operatorname{tr}(C^{-1}TC) = \operatorname{tr} T$, these are independent of the basis.

Vector Spaces of Linear Transformations

Let V, W be vector spaces. We denote by $\mathcal{L}(V, W)$ the set of all linear transformations from V to W. The set $\mathcal{L}(V, W)$ has a natural vector space structure: if L_1, $L_2 \in \mathcal{L}$ and $\alpha_1, \alpha_2 \in \mathbb{F}$, define $\alpha_1 L_1 + \alpha_2 L_2 \in \mathcal{L}(V, W)$ by $(\alpha_1 L_1 + \alpha_2 L_2)(v) = \alpha_1 L_1(v) + \alpha_2 L_2(v)$. In the infinite-dimensional case, we will be more interested in the vector space $\mathcal{B}(V, W)$ of all bounded linear transformations (to be defined) from V to W with respect to norms on V and W. When V and W are finite-dimensional, it will turn out that $\mathcal{B}(V, W) = \mathcal{L}(V, W)$.

If V, W have dimensions n, m, respectively, then the matrix representation above shows that $\mathcal{L}(V, W)$ is isomorphic to $\mathbb{F}^{m\times n}$, so it has dimension nm. When $V = W$, we denote $\mathcal{L}(V, V)$ by $\mathcal{L}(V)$. Since the composition $M \circ L : V \to U$ of linear transformations $L : V \to W$ and $M : W \to U$ is also linear, $\mathcal{L}(V)$ is naturally an algebra with composition as the multiplication operation.

Projections

Suppose W_1, W_2 are subspaces of V and $V = W_1 \oplus W_2$. Then we say W_1 and W_2 are *complementary subspaces*. Any $v \in V$ can be written uniquely as $v = w_1 + w_2$ with $w_1 \in W_1$, $w_2 \in W_2$. So we can define maps $P_1 : V \to W_1$, $P_2 : V \to W_2$ by $P_1 v = w_1$, $P_2 v = w_2$. It is easy to check that P_1, P_2 are linear. We usually regard P_1, P_2 as mapping V into itself (as $W_1 \subset V$, $W_2 \subset V$). P_1 is called the *projection onto* W_1 *along* W_2 (and P_2 the projection of W_2 along W_1). It is important to note that P_1 is not determined solely by the subspace $W_1 \subset V$, but also depends on the choice of the complementary subspace W_2. Since a linear transformation is determined by its restrictions to direct summands of its domains, P_1 is uniquely characterized as that linear transformation on V which satisfies

$$P_1\Big|_{W_1} = I\Big|_{W_1} \qquad \text{and} \qquad P_1\Big|_{W_2} = 0.$$

It follows easily that

$$P_1^2 = P_1, \quad P_2^2 = P_2, \quad P_1 + P_2 = I, \quad P_1P_2 = P_2P_1 = 0.$$

In general, an element q of an algebra is called *idempotent* if $q^2 = q$. If $P : V \to V$ is a linear transformation and P is idempotent, then P is a projection in the above sense: it is the projection onto $\mathcal{R}(P)$ along $\mathcal{N}(P)$.

This discussion extends to the case in which $V = W_1 \oplus \cdots \oplus W_m$ for subspaces W_i. We can define projections $P_i : V \to W_i$ in the obvious way: P_i is the projection onto W_i along $W_1 \oplus \cdots \oplus W_{i-1} \oplus W_{i+1} \oplus \cdots \oplus W_m$. Then

$$P_i^2 = P_i \text{ for } 1 \le i \le m, \quad P_1 + \cdots + P_m = I, \quad \text{and} \quad P_iP_j = P_jP_i = 0 \text{ for } i \ne j.$$

If V is finite dimensional, we say that a basis $\{w_1, \ldots, w_p, u_1, \ldots, u_q\}$ for $V = W_1 \oplus W_2$ is *adapted to the decomposition* $W_1 \oplus W_2$ if $\{w_1, \ldots, w_p\}$ is a basis for W_1 and $\{u_1, \ldots, u_q\}$ is a basis for W_2. With respect to such a basis, the matrix representations of P_1 and P_2 are

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}, \text{ where the block structure is } \begin{bmatrix} p\times p & p\times q \\ q\times p & q\times q \end{bmatrix},$$

abbreviated as: $$\begin{array}{cc} & \begin{array}{cc} p & q \end{array} \\ \begin{array}{c} p \\ q \end{array} & \begin{bmatrix} * & * \\ * & * \end{bmatrix} \end{array}.$$

Invariant Subspaces

We say that a subspace $W \subset V$ is invariant under a linear transformation $L : V \to V$ if $L(W) \subset W$. If V is finite dimensional and $\{w_1, \ldots, w_p\}$ is a basis for W which we complete to some basis $\{w_1, \ldots, w_p, u_1, \ldots, u_q\}$ of V, then W is invariant under L iff the matrix of L in this basis is of the form

$$\begin{array}{cc} & \begin{array}{cc} p & q \end{array} \\ \begin{array}{c} p \\ q \end{array} & \begin{bmatrix} * & * \\ 0 & * \end{bmatrix} \end{array},$$

i.e., block upper-triangular.

We say that $L : V \to V$ preserves the decomposition $W_1 \oplus \cdots \oplus W_m = V$ if each W_i is invariant under L. In this case, L defines linear transformations $L_i : W_i \to W_i$, $1 \le i \le m$, and we write $L = L_1 \oplus \cdots \oplus L_m$. Clearly L preserves the decomposition iff the matrix T of L with respect to an adapted basis is of block diagonal form

$$T = \begin{bmatrix} T_1 & & & 0 \\ & T_2 & & \\ & & \ddots & \\ 0 & & & T_m \end{bmatrix},$$

where the T_i's are the matrices of the L_i's in the bases of the W_i's.

Nilpotents

A linear transformation $L : V \to V$ is called *nilpotent* if $L^r = 0$ for some $r > 0$. A basic example is a shift operator on $\mathbb{F}^n$: define $Se_1 = 0$, and $Se_i = e_{i-1}$ for $2 \le i \le n$. The matrix of S is denoted S_n:

$$S_n = \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & & 1 \\ 0 & & & 0 \end{bmatrix} \in \mathbb{F}^{n\times n}.$$

Note that S^m shifts by $m : S^m e_i = 0$ for $1 \le i \le m$, and $S^m e_i = e_{i-m}$ for $m+1 \le i \le n$. Thus $S^n = 0$. For $1 \le m \le n-1$, the matrix $(S_n)^m$ of S^m is zero except for 1's on the m^{th} super diagonal (i.e., the ij elements for $j = i + m$ $(1 \le i \le n-m)$ are 1's):

$$(S_n)^m = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 \\ & \ddots & \ddots & \ddots & 1 \\ & & \ddots & \ddots & 0 \\ 0 & & & \ddots & 0 \end{bmatrix} \begin{array}{l} \longleftarrow (1, m+1) \text{ element} \\ \longleftarrow (n-m, n) \text{ element.} \\ \\ \\ \end{array}$$

Note, however that the analogous shift operator on $\mathbb{F}^\infty$ defined by: $Se_1 = 0$, $Se_i = e_{i-1}$ for $i \ge 2$, is *not* nilpotent.

Structure of Nilpotent Operators in Finite Dimensions

We next prove a theorem which describes the structure of all nilpotent operators in finite dimensions. This is an important result in its own right and will be a key step in showing that every matrix is similar to a matrix in Jordan form.

Theorem. Let V be finite dimensional and $L : V \to V$ be nilpotent. There is a basis for V in which L is a direct sum of shift operators.

Proof. Since L is nilpotent, there is an integer r so that $L^r = 0$ but $L^{r-1} \ne 0$. Let $v_1, \ldots, v_{l_1}$ be a basis for $\mathcal{R}(L^{r-1})$, and for $1 \le i \le l_1$, choose $w_i \in V$ for which $v_i = L^{r-1} w_i$. (As an aside, observe that

$$V = \mathcal{N}(L^r) = \mathcal{N}(L^{r-1}) \oplus \operatorname{span}\{w_1, \ldots, w_{l_1}\}.)$$

We claim that the set

$$\mathcal{S}_1 = \{L^k w_i : 0 \le k \le r-1, 1 \le i \le l_1\}$$

is linearly independent. Suppose

$$\sum_{i=1}^{l_1} \sum_{k=0}^{r-1} c_{ik} L^k w_i = 0.$$

Apply L^{r-1} to obtain

$$\sum_{i=1}^{l_1} c_{i0} L^{r-1} w_i = 0.$$

Hence $\sum_{i=1}^{l_1} c_{i0} v_i = 0$, so $c_{i0} = 0$ for $1 \le i \le l_1$. Now apply L^{r-2} to the double sum to obtain

$$0 = \sum_{i=1}^{l_1} c_{i1} L^{r-1} w_i = \sum_{i=1}^{l_1} c_{i1} v_i,$$

so $c_{i1} = 0$ for $1 \le i \le l_1$. Successively applying lower powers of L shows that all $c_{ik} = 0$.

Observe that for $1 \le i \le l_1$, $\operatorname{span}\{L^{r-1}w_i, L^{r-2}w_i, \ldots, w_i\}$ is invariant under L, and L acts by shifting these vectors. It follows that on $\operatorname{span}(\mathcal{S}_1)$, L is the direct sum of l_1 copies of the $(r \times r)$ shift S_r, and in the basis

$$\{L^{r-1}w_1, L^{r-2}w_1, \ldots, w_1, L^{r-1}w_2, \ldots, w_2, \ldots, L^{r-1}w_{l_1}, \ldots, w_{l_1}\}$$

for $\operatorname{span}(\mathcal{S}_1)$, L has the matrix $\begin{bmatrix} S_r & & 0 \\ & \ddots & \\ 0 & & S_r \end{bmatrix}$. In general, $\operatorname{span}(\mathcal{S}_1)$ need not be all of V, so we aren't done.

We know that $\{L^{r-1}w_1, \ldots, L^{r-1}w_{l_1}\}$ is a basis for $\mathcal{R}(L^{r-1})$, and that

$$\{L^{r-1}w_1, \ldots, L^{r-1}w_{l_1}, L^{r-2}w_1, \ldots, L^{r-2}w_{l_1}\}$$

are linearly independent vectors in $\mathcal{R}(L^{r-2})$. Complete the latter to a basis of $\mathcal{R}(L^{r-2})$ by appending, if necessary, vectors $\widetilde{u}_1, \ldots, \widetilde{u}_{l_2}$. As before, choose $\widetilde{w}_{l_1+j}$ for which

$$L^{r-2}\widetilde{w}_{l_1+j} = \widetilde{u}_j, \quad 1 \le j \le l_2.$$

We will replace $\widetilde{w}_{l_1+j}$ $(1 \le j \le l_2)$ by vectors in $\mathcal{N}(L^{r-1})$. Note that

$$L\widetilde{u}_j = L^{r-1}\widetilde{w}_{l_1+j} \in \mathcal{R}(L^{r-1}),$$

so we may write

$$L^{r-1}\widetilde{w}_{l_1+j} = \sum_{i=1}^{l_1} a_{ij} L^{r-1} w_i$$

for some $a_{ij} \in \mathbb{F}$. For $1 \le j \le l_2$, set

$$w_{l_1+j} = \widetilde{w}_{l_1+j} - \sum_{i=1}^{l_1} a_{ij} w_i \quad \text{and} \quad u_j = L^{r-2} w_{l_1+j}.$$

Replacing the $\widetilde{u}_j$'s by the u_j's still gives a basis of $\mathcal{R}(L^{r-2})$ as above (exercise). Clearly $L^{r-1}w_{l_1+j} = 0$ for $1 \le j \le l_2$. (Again as an aside, observe that we now have the direct sum decomposition

$$\mathcal{N}(L^{r-1}) = \mathcal{N}(L^{r-2}) \oplus \operatorname{span}\{Lw_1, \ldots, Lw_{l_1}, w_{l_1+1}, \ldots, w_{l_1+l_2}\}.)$$

So we now have a basis for $\mathcal{R}(L^{r-2})$ of the form

$$\{L^{r-1}w_1,\ldots,L^{r-1}w_{l_1},L^{r-2}w_1,\ldots,L^{r-2}w_{l_1},L^{r-2}w_{l_1+1},\ldots,L^{r-2}w_{l_1+l_2}\}$$

for which $L^{r-1}w_{l_1+j}=0$ for $1\le j\le l_2$. By the same argument as above, upon setting

$$\mathcal{S}_2=\{L^k w_{l_1+j}:0\le k\le r-2,1\le j\le l_2\},$$

we conclude that $\mathcal{S}_1\cup\mathcal{S}_2$ is linearly independent, and L acts on $\operatorname{span}(\mathcal{S}_2)$ as a direct sum of l_2 copies of the $(r-1)\times(r-1)$ shift S_{r-1}. We can continue this argument, decreasing r one at a time and end up with a basis of $\mathcal{R}(L^0)=V$ in which L acts as a direct sum of shift operators:

$$L=\overbrace{S_r\oplus\cdots\oplus S_r}^{l_1}\oplus\overbrace{S_{r-1}\oplus\cdots\oplus S_{r-1}}^{l_2}\oplus\cdots\oplus\overbrace{S_1\oplus\cdots\oplus S_1}^{l_r}\qquad\text{(Note: }S_1=0\in\mathbb{F}^{1\times1})$$

□

Remarks:

(1) For $1\le j$, let $k_j=\dim(\mathcal{N}(L^j))$. It follows easily from the above that $0<k_1<k_2<\cdots<k_r=k_{r+1}=k_{r+2}=\cdots=n$, and thus $r\le n$.

(2) The structure of L is determined by knowing r and $l_1,\ldots,l_r$. These, in turn, are determined by knowing $k_1,\ldots,k_n$ (see the homework).

(3) General facts about nilpotent transformations follow from this normal form. For example, if $\dim V=n$ and $L:V\to V$ is nilpotent, then

(i) $L^n=0$

(ii) $\operatorname{tr}L=0$

(iii) $\det L=0$

(iv) $\det(I+L)=1$

(v) for any $\lambda\in\mathbb{F}$, $\det(\lambda I-L)=\lambda^n$

Dual Transformations

Recall that if V and W are vector spaces, we denote by V' and $\mathcal{L}(V,W)$ the dual space of V and the space of linear transformations from V to W, respectively.

Let $L\in\mathcal{L}(V,W)$. We define the *dual*, or *adjoint* transformation $L':W'\to V'$ by $(L'g)(v)=g(Lv)$ for $g\in W'$, $v\in V$. Clearly $L\mapsto L'$ is a linear transformation from $\mathcal{L}(V,W)$ to $\mathcal{L}(W',V')$ and $(L\circ M)'=M'\circ L'$ if $M\in\mathcal{L}(U,V)$.

When V, W are finite dimensional and we choose bases for V and W, we get corresponding dual bases, and we can represent vectors in V, W, V', W' by their coordinate vectors

$$x=\begin{pmatrix}x_1\\\vdots\\x_n\end{pmatrix}\qquad y=\begin{pmatrix}y_1\\\vdots\\y_m\end{pmatrix}\qquad a=(a_1\cdots a_n)\qquad b=(b_1\cdots b_m).$$

Also, L is represented by a matrix $T \in \mathbb{F}^{m\times n}$ for which $y = Tx$. Now if $g \in W'$ has coordinates $b = (b_1 \cdots b_m)$, we have

$$g(Lv) = (b_1 \cdots b_m)T \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

so $L'g$ has coordinates $(a_1 \cdots a_n) = (b_1 \cdots b_m)T$. Thus L is represented by left-multiplication by T on column vectors, and L' is represented by right-multiplication by T on row vectors. Another common convention is to represent the dual coordinate vectors also as columns; taking the transpose in the above gives

$$\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = T^T \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix},$$

so L' is represented by left-multiplication by T^T on column vectors. (T^T is the transpose of T: $(T^T)_{ij} = t_{ji}$.)

We can take the dual of V' to obtain V''. There is a natural inclusion $V \to V''$: if $v \in V$, then $f \mapsto f(v)$ defines a linear functional on V'. This map is injective since if $v \neq 0$, there is an $f \in V'$ for which $f(v) \neq 0$. (Proof: Complete $\{v\}$ to a basis for V and take f to be the first vector in the dual basis.)

We identify V with its image, so we can regard $V \subset V''$. If V is finite dimensional, then $V = V''$ since $\dim V = \dim V' = \dim V''$. If V is infinite dimensional, however, then there are elements of V'' which are not in V.

If $S \subset V$ is a subset, we define the annihilator $S^a \subset V'$ by

$$S^a = \{f \in V' : f(v) = 0 \text{ for all } v \in S\}.$$

Clearly $S^a = (\mathrm{span}(S))^a$. Now $(S^a)^a \subset V''$, and if $\dim V < \infty$, we can identify $V'' = V$ as above.

Proposition. If $\dim V < \infty$, then $(S^a)^a = \mathrm{span}(S)$.

Proof. It follows immediately from the definition that $\mathrm{span}(S) \subset (S^a)^a$. To show $(S^a)^a \subset \mathrm{span}(S)$, assume without loss of generality that S is a subspace. We claim that if W is an m-dimensional subspace of V and $\dim V = n$, then $\dim W^a = \mathrm{codim} W = n-m$. To see this, choose a basis $\{w_1, \ldots, w_m\}$ for W and complete it to a basis $\{w_1, \ldots, w_{m+1}, \ldots, w_n\}$ for V. Then clearly the dual basis vectors $\{f_{m+1}, \ldots, f_n\}$ are a basis for W^a, so $\dim W^a = n - m$. Hence $\dim(S^a)^a = n - \dim S^a = n - (n - \dim S) = \dim S$. Since we know $S \subset (S^a)^a$, the result follows. □

In complete generality, we have

Proposition. Suppose $L \in \mathcal{L}(V, W)$. Then $\mathcal{N}(L') = \mathcal{R}(L)^a$.

Proof. Clearly both are subspaces of W'. Let $g \in W'$. Then $g \in \mathcal{N}(L') \iff L'g = 0 \iff (\forall v \in V)\,(L'g)(v) = 0 \iff (\forall\, v \in V)\, g(Lv) = 0 \iff g \in \mathcal{R}(L)^a$. □

As an immediate consequence of the two Propositions above we conclude:

Corollary. Suppose $L \in \mathcal{L}(V, W)$ and $\dim W < \infty$. Then $\mathcal{R}(L) = \mathcal{N}(L')^a$.

We are often interested in identifying $\mathcal{R}(L)$ for some $L \in \mathcal{L}(V, W)$, or equivalently in determining those $w \in W$ for which there exists $v \in V$ satisfying $Lv = w$. If V and W are finite-dimensional, choose bases of V, W, thereby obtaining coordinate vectors $x \in \mathbb{F}^n$, $y \in \mathbb{F}^m$ for v, w and a matrix T representing L. This question then amounts to determining those $y \in \mathbb{F}^m$ for which the linear system $Tx = y$ can be solved. According to the Corollary above, we have $\mathcal{R}(L) = \mathcal{N}(L')^a$. Thus there exists $v \in V$ satisfying $Lv = w$ iff $g(w) = 0$ for all $g \in W'$ for which $L'g = 0$. In terms of matrices, $Tx = y$ is solvable iff $(b_1 \cdots b_m) \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = 0$ for all $(b_1 \cdots b_m)$ for which $(b_1 \cdots b_m)T = 0$, or equivalently, $T^T \begin{pmatrix} b_1 \\ \cdots \\ b_m \end{pmatrix} = 0$. These are often called the *compatibility conditions* for solving the linear system $Tx = y$.

Bilinear Forms

A function $\varphi : V \times V \to \mathbb{F}$ is called a *bilinear form* if it is linear in each variable separately.

Examples:

(1) Let $V = \mathbb{F}^n$. For any matrix $A \in \mathbb{F}^{n \times n}$,

$$\varphi(y, x) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} y_i x_j = y^T A x$$

is a bilinear form. In fact, all bilinear forms on $\mathbb{F}^n$ are of this form: since

$$\varphi\left(\sum y_i e_i, \sum x_j e_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} y_i x_j \varphi(e_i, e_j),$$

we can just set $a_{ij} = \varphi(e_i, e_j)$. Similarly, for any finite-dimensional V, we can choose a basis $\{v_1, \ldots, v_n\}$; if φ is a bilinear form on V and $v = \sum x_j v_j$, $w = \sum y_i v_i$, then

$$\varphi(w, v) = \sum_{i=1}^{n} \sum_{j=1}^{n} y_i x_j \varphi(v_i, v_j) = y^T A x$$

where $A \in \mathbb{F}^{n \times n}$ is given by $a_{ij} = \varphi(v_i, v_j)$. A is called the *matrix of* φ with respect to the basis $\{v_1, \ldots, v_n\}$.

(2) One can also use infinite matrices $(a_{ij})_{i,j \geq 1}$ for $V = \mathbb{F}^\infty$ as long as convergence conditions are imposed. For example, if all $|a_{ij}| \leq M$, then $\varphi(y, x) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} y_i x_j$ defines a bilinear form on ℓ^1 since

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |a_{ij} y_i x_j| \leq M \left(\sum_{i=1}^{\infty} |y_i|\right) \left(\sum_{j=1}^{\infty} |x_j|\right).$$

Similarly if $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|a_{ij}|<\infty$, then we get a bilinear form on ℓ^∞.

(3) If V is a vector space and $f,g\in V'$, then $\varphi(w,v)=g(w)f(v)$ is a bilinear form.

(4) If $V=C[a,b]$, then the following are all examples of bilinear forms:

(i) $\varphi(v,u)=\int_a^b\int_a^b k(x,y)v(x)u(y)dxdy$, for $k\in C([a,b]\times[a,b])$

(ii) $\varphi(v,u)=\int_a^b h(x)v(x)u(x)dx$, for $h\in C([a,b])$

(iii) $\varphi(v,u)=v(x_0)\int_a^b u(x)dx$, for $x_0\in[a,b]$

We say that a bilinear form is *symmetric* if $(\forall\, v,w\in V)\varphi(v,w)=\varphi(w,v)$. In the finite-dimensional case, this corresponds to the condition that the matrix A be symmetric, i.e., $A=A^T$, or $(\forall\, i,j)\, a_{ij}=a_{ji}$.

Returning to Example (1) above, let V be finite-dimensional and consider how the matrix of the bilinear form φ changes when the basis of V is changed. Let $(v_1',\ldots,v_n')$ be another basis for V related to the original basis $(v_1,\ldots,v_n)$ by change of basis matrix $C\in\mathbb{F}^{n\times n}$. We have seen that the coordinates x' for v relative to $(v_1',\ldots,v_n')$ are related to the coordinates x relative to $(v_1,\ldots,v_n)$ by $x=Cx'$. If y' and y denote the coordinates of w relative to the two bases, we have $y=Cy'$ and therefore

$$\varphi(w,v)=y^TAx=y'^TC^TACx'.$$

It follows that the matrix of φ in the basis $(v_1',\ldots,v_n')$ is C^TAC. Compare this with the way the matrix representing a linear transformation L changed under change of basis: if T was the matrix of L in the basis $(v_1,\ldots,v_n)$, then the matrix in the basis $(v_1',\ldots,v_n')$ was $C^{-1}TC$. Hence the way a matrix changes under change of basis depends on whether the matrix represents a linear transformation or a bilinear form.

Sesquilinear Forms

When $\mathbb{F}=\mathbb{C}$, we will more often use sesquilinear forms: $\varphi: V\times V\to\mathbb{C}$ is called *sesquilinear* if φ is linear in the second variable and conjugate-linear in the first variable, i.e.,

$$\varphi(\alpha_1w_1+\alpha_2w_2,v)=\bar{\alpha}_1\varphi(w_1,v)+\bar{\alpha}_2\varphi(w_2,v).$$

(Sometimes the convention is reversed and φ is conjugate-linear in the second variable. The two possibilities are equivalent upon interchanging the variables.) For example, on $\mathbb{C}^n$ all sesquilinear forms are of the form $\varphi(w,z)=\sum_{i=1}^n\sum_{j=1}^n a_{ij}\bar{w}_iz_j$ for some $A\in\mathbb{C}^{n\times n}$. To be able to discuss bilinear forms over $\mathbb{R}$ and sesquilinear forms over $\mathbb{C}$ at the same time, we will speak of a sesquilinear form over $\mathbb{R}$ and mean just a bilinear form over $\mathbb{R}$. A sesquilinear form is said to be Hermitian-symmetric (or sometimes just Hermitian) if

$$(\forall\, v,w\in V)\,\varphi(w,v)=\overline{\varphi(v,w)}$$

(when $\mathbb{F}=\mathbb{R}$, we say the form is symmetric). When $\mathbb{F}=\mathbb{C}$, this corresponds to the condition that $A=A^H$, where $A^H=\bar{A}^T$ (i.e., $(A^H)_{ij}=\overline{a_{ji}}$) is the Hermitian transpose (or conjugate transpose) of A, and a matrix $A \in \mathbb{C}^{n\times n}$ satisfying $A = A^H$ is called *Hermitian.* When $\mathbb{F} = \mathbb{R}$, this corresponds to the condition $A = A^T$ (i.e., A is symmetric).

To a sesquilinear form, we can associate the quadratic form $\varphi(v, v)$. We say that φ is nonnegative (or positive semi-definite) if $(\forall\, v \in V)\, \varphi(v, v) \geq 0$, and that φ is positive (or positive definite) if $\varphi(v, v) > 0$ for all $v \neq 0$ in V. By an *inner product* on V, we will mean a positive-definite Hermitian-symmetric sesquilinear form.

Examples:

(1) $\mathbb{F}^n$ with the Euclidean inner product $\langle y, x\rangle = \sum_{i=1}^n \overline{y_i} x_i$.

(2) Let $V = \mathbb{F}^n$, and let $A \in \mathbb{F}^{n\times n}$ be Hermitian-symmetric. Define

$$\langle y, x\rangle_A = \sum_{i=1}^n \sum_{j=1}^n a_{ij} \overline{y_i} x_j = \overline{y}^T A x$$

The requirement that $\langle x, x\rangle_A > 0$ for $x \neq 0$ for $\langle\cdot, \cdot\rangle_A$ to be an inner product serves to define positive-definite matrices.

(3) If V is any finite-dimensional vector space, we can choose a basis and thus identify $V \cong \mathbb{F}^n$, and then transfer the Euclidean inner product to V in the coordinates of this basis. The resulting inner product depends on the choice of basis — there is no canonical inner product on a general vector space. With respect to the coordinates induced by a basis, any inner product on a finite-dimensional vector space V is of the form (2).

(4) One can define an inner product on ℓ^2 by $\langle y, x\rangle = \sum_{i=1}^\infty \overline{y_i} x_i$. To see (from first principles) that this sum converges absolutely, apply the finite-dimensional Cauchy-Schwarz inequality to obtain

$$\sum_{i=1}^n |\overline{y_i} x_i| \leq \left(\sum_{i=1}^n |x_i|^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^n |y_i|^2\right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^\infty |x_i|^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^\infty |y_i|^2\right)^{\frac{1}{2}}.$$

Now let $n \to \infty$ to deduce that the series $\sum_{i=1}^\infty \overline{y_i} x_i$ converges absolutely.

(5) The L^2-inner product on $C([a, b])$ is given by $\langle v, u\rangle = \int_a^b \overline{v(x)} u(x) dx$. (Exercise: show that this is indeed positive definite on $C([a, b])$.)

An inner product on V determines an injection $V \to V'$: if $w \in V$, define $w^* \in V'$ by $w^*(v) = \langle w, v\rangle$. Since $w^*(w) = \langle w, w\rangle$ it follows that $w^* = 0 \Rightarrow w = 0$, so the map $w \mapsto w^*$ is injective. The map $w \mapsto w^*$ is *conjugate-linear* (rather than linear, unless $\mathbb{F} = \mathbb{R}$) since $(\alpha w)^* = \bar{\alpha} w^*$. The image of this map is a subspace of V'. If $\dim V < \infty$, then this map is surjective too since $\dim V = \dim V'$. In general, it is not surjective.

Let $\dim V < \infty$, and represent vectors in V as elements of $\mathbb{F}^n$ by choosing a basis. If v, w have coordinates $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, respectively, and the inner product has matrix

$A \in \mathbb{F}^{n\times n}$ in this basis, then

$$w^*(v) = \langle w, v\rangle = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} a_{ij}\overline{y_i}\right) x_j.$$

It follows that w^* has components $b_j = \sum_{j=1}^{n} a_{ij}\overline{y_i}$ with respect to the dual basis. Recalling that the components of dual vectors are written in a row, we can write this as $b = \overline{y}^T A$.

An inner product on V allows a reinterpretation of annihilators. If $W \subset V$ is a subspace, define the orthogonal complement $W^\perp = \{v \in V : \langle w, v\rangle = 0 \text{ for all } w \in W\}$. Clearly $W^\perp$ is a subspace of V and $W \cap W^\perp = \{0\}$. The orthogonal complement $W^\perp$ is closely related to the annihilator W^a: it is evident that for $v \in V$, we have $v \in W^\perp$ if and only if $v^* \in W^a$. If $\dim V < \infty$, we saw above that every element of V' is of the form v^* for some $v \in V$. So we conclude in the finite-dimensional case that

$$W^a = \{v^* : v \in W^\perp\}.$$

It follows that $\dim W^\perp = \dim W^a = \text{codim} W$. From this and $W \cap W^\perp = \{0\}$, we deduce that $V = W \oplus W^\perp$. So in a finite dimensional inner product space, a subspace W has a natural complementary subspace, namely $W^\perp$. The induced projection onto W along $W^\perp$ is called the *orthogonal projection* onto W.

Norms

A norm is a way of measuring the length of a vector. Let V be a vector space. A *norm* on V is a function $\|\cdot\| : V \to [0, \infty)$ satisfying

(i) $\|v\| = 0$ iff $v = 0$

(ii) $\|\alpha v\| = |\alpha| \cdot \|v\|$ for $\alpha \in \mathbb{F}$, $v \in V$

(iii) (triangle inequality) $\|v + w\| \leq \|v\| + \|w\|$ for $v, w \in V$.

The pair $(V, \|\cdot\|)$ is called a *normed linear space* (or normed vector space).

Fact: A norm $\|\cdot\|$ on a vector space V induces a metric d on V by

$$d(v, w) = \|v - w\|.$$

(Exercise: Show d is a metric on V.) All topological properties (e.g. open sets, closed sets, convergence of sequences, continuity of functions, compactness, etc.) will refer to those of the metric space (V, d).

Examples:

(1) ℓ^p norm on $\mathbb{F}^n$ $(1 \leq p \leq \infty)$

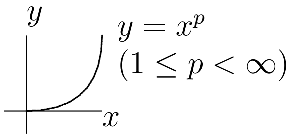
(a) $p = \infty$ $\qquad \|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$ $\qquad$ is a norm on $\mathbb{F}^n$.

(b) $1 \leq p < \infty$ $\qquad \|x\|_p = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$ $\qquad$ is a norm on $\mathbb{F}^n$.

The triangle inequality

$$\left(\sum_{i=1}^n |x_i + y_i|^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^n |y_i|^p\right)^{\frac{1}{p}}$$

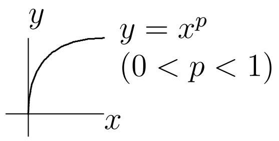
is known as "Minkowski's inequality." It is a consequence of Hölder's inequality. Integral versions of these inequalities are proved in real analysis texts, e.g., Folland or Royden. The proofs for vectors in $\mathbb{F}^n$ are analogous to the proofs for integrals. The fact that the triangle inequality holds is related to the observation that for $1 \leq p < \infty$, the function $x \mapsto x^p$ for $x \geq 0$ is convex:



(c) $0 < p < 1$ $\qquad (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$ is *not* a norm on $\mathbb{F}^n$ $(n > 1)$.

The triangle inequality does not hold. For a counterexample, let $x = e_1$ and $y = e_2$. Then $(\sum_{i=1}^n |x_i + y_i|^p)^{\frac{1}{p}} = 2^{\frac{1}{p}} > 2 = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}} + (\sum_{i=1}^n |y_i|^p)^{\frac{1}{p}}$.

The failure of the triangle inequality is related to the observation that for $0 < p < 1$, the map $x \mapsto x^p$ for $x \geq 0$ is *not* convex:



(2) ℓ^p norm on ℓ^p (subspace of $\mathbb{F}^\infty$) $(1 \leq p \leq \infty)$

(a) $p = \infty$ Recall $\ell^\infty = \{x \in \mathbb{F}^\infty : \sup_{i \geq 1} |x_i| < \infty\}$.

$\|x\|_\infty = \sup_{i\geq 1} |x_i|$ is a norm on ℓ^∞.

(b) $1 \leq p < \infty$ Recall $\ell^p = \left\{x \in \mathbb{F}^\infty : (\sum_{i=1}^\infty |x_i|^p)^{\frac{1}{p}} < \infty\right\}$.

$\|x\|_p = (\sum_{i=1}^\infty |x|^p)^{\frac{1}{p}}$ is a norm on ℓ^p.

The triangle inequality follows from the finite-dimensional case: exercise.

(3) L^p norm on $C([a,b])$ $(1 \leq p \leq \infty)$ $(-\infty < a < b < \infty)$

(a) $p = \infty$ $\|f\|_\infty = \sup_{a \leq x \leq b} |f(x)|$ is a norm on $C([a,b])$.

Since $|f(x)|$ is a continuous, real-valued function on the compact set $[a,b]$, it takes on its maximum, so the "sup" is actually a "max" here:

$$\|f\|_\infty = \max_{a \leq x \leq b} |f(x)|.$$

(b) $1 \leq p < \infty$ $\|f\|_p = \left(\int_a^b |f(x)|^p dx\right)^{\frac{1}{p}}$ is a norm on $C([a,b])$.

Use continuity of f to show that $\|f\|_p = 0 \Rightarrow f(x) \equiv 0$ on $[a,b]$.

The triangle inequality

$$\left(\int_a^b |f(x)+g(x)|^p dx\right)^{\frac{1}{p}} \leq \left(\int_a^b |f(x)|^p dx\right)^{\frac{1}{p}} + \left(\int_a^b |g(x)|^p dx\right)^{\frac{1}{p}}$$

is Minkowski's inequality, a consequence of Hölder's inequality.

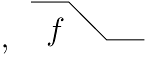
(c) $0 < p < 1$ $\left(\int_a^b |f(x)|^p dx\right)^{\frac{1}{p}}$ is *not* a norm on $C([a,b])$.

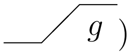
"Pseudo-example": Let $a = 0$, $b = 1$,

$$f(x) = \left\{ \begin{array}{ll} 1 & 0 \leq x < \frac{1}{2} \\ 0 & \frac{1}{2} < x \leq 1 \end{array} \right. , \qquad g(x) = \left\{ \begin{array}{ll} 0 & 0 \leq x < \frac{1}{2} \\ 1 & \frac{1}{2} < x \leq 1 \end{array} \right. .$$

Then

$$\left(\int_0^1 |f(x)|^p dx\right)^{\frac{1}{p}} + \left(\int_0^1 |g(x)|^p dx\right)^{\frac{1}{p}}$$
$$= \left(\frac{1}{2}\right)^{\frac{1}{p}} + \left(\frac{1}{2}\right)^{\frac{1}{p}} = 2^{1-\frac{1}{p}} < 1 = \left(\int_0^1 |f(x)+g(x)|^p dx\right)^{\frac{1}{p}},$$

so the triangle inequality fails. This is only a pseudo-example because f and g are not continuous. Exercise: Adjust these f and g to be continuous (e.g., 



) to construct a legitimate counterexample to the triangle inequality.

Remark: There is also a Minkowski inequality for integrals: if $1 \leq p < \infty$ and $u \in C([a,b] \times [c,d])$, then

$$\left(\int_a^b \left|\int_c^d u(x,y)dy\right|^p dx\right)^{\frac{1}{p}} \leq \int_c^d \left(\int_a^b |u(x,y)|^p \, dx\right)^{\frac{1}{p}} dy.$$

Equivalence of Norms

Lemma. If $(V, \|\cdot\|)$ is a normed linear space, then $\|\cdot\| : (V, \|\cdot\|) \to \mathbb{R}$ is continuous.

Proof. For $v_1, v_2 \in V$, $\|v_1\| = \|v_1 - v_2 + v_2\| \leq \|v_1 - v_2\| + \|v_2\|$, and thus $\|v_1\| - \|v_2\| \leq \|v_1 - v_2\|$. Similarly, $\|v_2\| - \|v_1\| \leq \|v_2 - v_1\| = \|v_1 - v_2\|$. So $|\|v_1\| - \|v_2\|| \leq \|v_1 - v_2\|$. Given $\epsilon > 0$, let $\delta = \epsilon$, etc. □

Definition. Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$, both on the same vector space V, are called *equivalent norms* on V if there are constants $C_1, C_2 > 0$ such that for all $v \in V$,

$$\frac{1}{C_1}\|v\|_2 \leq \|v\|_1 \leq C_2\|v\|_2.$$

Remarks.

(1) If $\|v\|_1 \leq C_2\|v\|_2$, then $\|v_k - v\|_2 \to 0 \Rightarrow \|v_k - v\|_1 \to 0$, so the identity map $I : (V, \|\cdot\|_2) \to (V, \|\cdot\|_1)$ is continuous. Likewise if $\|v\|_2 \leq C_1\|v\|_1$, then the identity map $I : (V, \|\cdot\|_1) \to (V, \|\cdot\|_2)$ is continuous. So if two norms are equivalent, then the identity map is bicontinuous. It is not hard to show (and it follows from a Proposition in the next chapter) that the converse is true as well: if the identity map is bicontinuous, then the two norms are equivalent. Thus two norms are equivalent if and only if they induce the same topologies on V, that is, the associated metrics have the same open sets.

(2) Equivalence of norms is easily checked to be an equivalence relation on the set of all norms on a fixed vector space V.

The next theorem establishes the fundamental fact that on a finite dimensional vector space, all norms are equivalent.

Norm Equivalence Theorem If V is a finite dimensional vector space, then any two norms on V are equivalent.

Proof. Fix a basis $\{v_1, \ldots, v_n\}$ for V, and identify V with $\mathbb{F}^n$ ($v \in V \leftrightarrow x \in \mathbb{F}^n$, where $v = x_1v_1 + \cdots + x_nv_n$). Using this identification, it suffices to prove the result for $\mathbb{F}^n$. Let

$|x| = (\sum_{i=1}^n |x_i|^2)^{\frac{1}{2}}$ denote the Euclidean norm (i.e., ℓ^2 norm) on $\mathbb{F}^n$. Because equivalence of norms is an equivalence relation, it suffices to show that any given norm $\|\cdot\|$ on $\mathbb{F}^n$ is equivalent to the Euclidean norm $|\cdot|$. For $x \in \mathbb{F}^n$,

$$\|x\| = \left\|\sum_{i=1}^n x_i e_i\right\| \leq \sum_{i=1}^n |x_i| \cdot \|e_i\| \leq \left(\sum_{i=1}^n |x_i|^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^n \|e_i\|^2\right)^{\frac{1}{2}}$$

by the Schwarz inequality in $\mathbb{R}^n$. Thus $\|x\| \leq M|x|$, where $M = (\sum_{i=1}^n \|e_i\|^2)^{\frac{1}{2}}$. Thus the identity map $I : (\mathbb{F}^n, |\cdot|) \to (\mathbb{F}^n, \|\cdot\|)$ is continuous, which is half of what we have to show.

Composing the map with $\|\cdot\| : (\mathbb{F}^n, \|\cdot\|) \to \mathbb{R}$ (which is continuous by the preceding Lemma), we conclude that $\|\cdot\| : (\mathbb{F}^n, |\cdot|) \to \mathbb{R}$ is continuous. Let $S = \{x \in \mathbb{F}^n : |x| = 1\}$. Then S is compact in $(\mathbb{F}^n, |\cdot|)$, and thus $\|\cdot\|$ takes on its minimum on S, which must be > 0 since $0 \notin S$. Let $m = \min_{|x|=1} \|x\| > 0$. So if $|x| = 1$, then $\|x\| \geq m$. For any $x \in \mathbb{F}^n$ with $x \neq 0$, $\left|\frac{x}{|x|}\right| = 1$, so $\left\|\frac{x}{|x|}\right\| \geq m$, i.e. $|x| \leq \frac{1}{m}\|x\|$. Thus $\|\cdot\|$ and $|\cdot|$ are equivalent. □

Remarks.

(1) So all norms on a fixed finite dimensional vector space are equivalent. Be careful, though, when studying problems (e.g. in numerical PDE) where there is a sequence of finite dimensional spaces of increasing dimensions: the constants C_1 and C_2 in the equivalence can depend on the dimension (e.g. $\|x\|_2 \leq \sqrt{n}\|x\|_\infty$ in $\mathbb{F}^n$).

(2) The Norm Equivalence Theorem is *not* true in infinite dimensional vector spaces, as the following examples show.

Example. Recall that $\mathbb{F}_0^\infty = \{x \in \mathbb{F}^\infty : (\exists\, N)(\forall\, n \geq N) \quad x_n = 0\}$. On $\mathbb{F}_0^\infty$, the ℓ^p norm and ℓ^q norm are *not* equivalent for $1 \leq p < q \leq \infty$. We will show the case $p = 1$, $q = \infty$ here. Note that $\|x\|_\infty \leq \sum_{i=1}^\infty |x_i| = \|x\|_1$, so $I : (\mathbb{F}_0^\infty, \|\cdot\|_1) \to (\mathbb{F}_0^\infty, \|\cdot\|_\infty)$ *is* continuous. But if

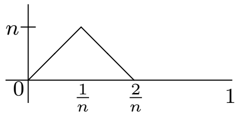
$$y_1 = (1, 0, 0, \cdots), \qquad y_2 = (1, 1, 0, \cdots), \qquad y_3 = (1, 1, 1, 0, \cdots), \qquad \cdots$$

then $\|y_n\|_\infty = 1\, \forall\, n$, but $\|y_n\|_1 = n$. So there does *not* exist a constant C for which $(\forall\, x \in \mathbb{F}_0^\infty)\|x\|_1 \leq C\|x\|_\infty$.

Example. On $C([a, b])$, for $1 \leq p < q \leq \infty$, the L^p and L^q norms are *not* equivalent. We will show the case $p = 1$, $q = \infty$ here. We have

$$\|u\|_1 = \int_a^b |u(x)|dx \leq \int_a^b \|u\|_\infty dx = (b - a)\|u\|_\infty,$$

so $I : (C([a, b]), \|\cdot\|_\infty) \to (C([a, b]), \|\cdot\|_1)$ *is* continuous.
(Remark: The integral $\mathcal{I}(u) = \int_a^b u(x)dx$ is continuous on $(C([a, b]), \|\cdot\|_1)$ since $|\mathcal{I}(u_1) - \mathcal{I}(u_2)| \leq \int_a^b |u_1(x) - u_2(x)|dx = \|u_1 - u_2\|_1$. So composing these two continuous operators implies the standard result that if $u_n \to u$ uniformly on $[a, b]$, then $\int_a^b u_n(x)dx \to \int_a^b u(x)dx$.)
To see that the inequality the other direction fails, WLOG assume $a = 0$, $b = 1$. Let u_n have graph:



Then $\|u_n\|_1 = 1$, but $\|u_n\|_\infty = n$. So there does not exist a constant C for which $(\forall\, u \in C([a,b]))$ $\|u\|_\infty \leq C\|u\|_1$.

(3) It can be shown that, for a normed linear space V, the closed unit ball $\{v \in V : \|v\| \leq 1\}$ is compact iff $\dim V < \infty$.

Example. In ℓ^2 (subspace of $\mathbb{F}^\infty$) with ℓ^2 norm $\|x\|_2 = \sqrt{\sum_{i=1}^\infty |x_i|^2}$ considered above, the closed unit ball $\{x \in \ell^2 : \|x\|_2 \leq 1\}$ is *not* compact. The sequence $e_1, e_2, e_3, \ldots$ is in the closed unit ball, but no subsequence converges because $\|e_i - e_j\|_2 = \sqrt{2}$ for $i \neq j$. Note that this shows that in an infinite dimensional normed linear space, a closed bounded set need not be compact.

Norms induced by inner products

Let V be a vector space and $\langle \cdot, \cdot \rangle$ be an inner product on V. Define $\|v\| = \sqrt{\langle v, v\rangle}$. By the properties of an inner product,

$$\|v\| \geq 0 \text{ with } \|v\| = 0 \text{ iff } v = 0, \text{ and}$$
$$(\forall\, \alpha \in \mathbb{F})(\forall\, v \in V) \quad \|\alpha v\| = |\alpha| \cdot \|v\|.$$

To show that $\|\cdot\|$ is actually a norm on V we need the triangle inequality. For this, it is helpful to observe that for any two vectors $u, v \in V$ we have

$$\begin{aligned} \|u+v\|^2 &= \langle u+v,\ u+v\rangle \\ &= \langle u,\ u\rangle + \langle u,\ v\rangle + \langle v,\ u\rangle + \langle v,\ v\rangle \\ &= \langle u,\ u\rangle + \langle u,\ v\rangle + \overline{\langle u,\ v\rangle} + \langle v,\ v\rangle \\ &= \|u\|^2 + 2\mathcal{R}e\,\langle u,\ v\rangle + \|v\|^2\,. \end{aligned}$$

Consequently, if u and v are orthogonal ($\langle u,\ v\rangle = 0$), then

$$\|u+v\|^2 = \|u\|^2 + \|v\|^2\,.$$

Cauchy-Schwarz inequality: For all v,$w \in V$:

$$|\langle v, w\rangle| \leq \|v\| \cdot \|w\|$$

Moreover, equality holds if and only if v and w are linearly dependent. (This latter statement is sometimes called the "converse of Cauchy-Schwarz.")

Proof. Case (i) If $v = 0$, we are done.

<u>Case (ii)</u> Assume $v \neq 0$, and set

$$u := w - \frac{\langle w,\ v\rangle}{\|v\|^2} v,$$

so that $\langle u,\ v\rangle = 0$, i.e. $u \perp v$. Then, by orthogonality,

$$\begin{aligned} \|w\|^2 &= \left\| u + \frac{\langle w,\ v\rangle}{\|v\|^2} v \right\|^2 \\ &= \|u\|^2 + \left\| \frac{\langle w,\ v\rangle}{\|v\|^2} v \right\|^2 \\ &= \|u\|^2 + \frac{|\langle w,\ v\rangle|^2}{\|v\|^2} \\ &\geq \frac{|\langle w,\ v\rangle|^2}{\|v\|^2}, \end{aligned}$$

with equality if and only if $u = 0$.

Now the triangle inequality follows:

$$\begin{aligned} \|v+w\|^2 &= \langle v+w, v+w\rangle = \langle v, v\rangle + 2\mathcal{R}e\langle v, w\rangle + \langle w, w\rangle \\ &\leq \|v\|^2 + 2|\langle v, w\rangle| + \|w\|^2 \leq \|v\|^2 + 2\|v\| \cdot \|w\| + \|w\|^2 = (\|v\| + \|w\|)^2. \end{aligned}$$

So $\|v\| = \sqrt{\langle v, v\rangle}$ is a norm on V, called the norm induced by the inner product $\langle \cdot, \cdot \rangle$. An inner product induces a norm, which induces a metric $(V, \langle \cdot, \cdot \rangle) \to (V, \|\cdot\|) \to (V, d)$.

Examples.

(1) The Euclidean norm (i.e. ℓ^2 norm) on $\mathbb{F}^n$ is induced by the standard inner product $\langle y, x\rangle = \sum_{i=1}^n \overline{y_i} x_i$: $\|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$.

(2) Let $A \in \mathbb{F}^{n\times n}$ be Hermitian symmetric and positive definite, and let

$$\langle y, x\rangle_A = \sum_{i=1}^n \sum_{j=1}^n \overline{y_i} a_{ij} x_j$$

for $x, y \in \mathbb{F}^n$. Then $\langle \cdot, \cdot \rangle_A$ is an inner product on $\mathbb{F}^n$, which induces the norm $\|x\|_A$ given by

$$\|x\|_A^2 = \langle x, x\rangle_A = \sum_{i=1}^n \sum_{j=1}^n \overline{x_i} a_{ij} x_j = \overline{x}^T A x = x^H A x,$$

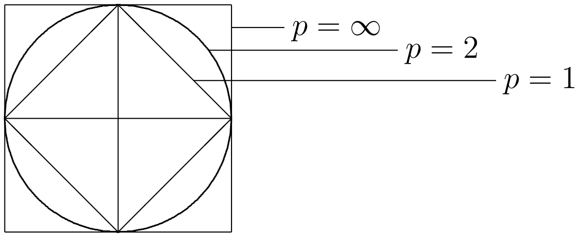
where $x^H = \overline{x}^T$.

(3) The ℓ^2 norm on ℓ^2 (subspace of $\mathbb{F}^\infty$) is induced by the inner product $\langle y, x\rangle = \sum_{i=1}^\infty \overline{y_i} x_i$: $\|x\|_2^2 = \sum_{i=1}^\infty \overline{x_i} x_i = \sum_{i=1}^\infty |x_i|^2$.

(4) The L^2 norm $\|u\|_2 = \left(\int_a^b |u(x)|^2 dx\right)^{\frac{1}{2}}$ on $C([a, b])$ is induced by the inner product $\langle v, u\rangle = \int_a^b \overline{v(x)} u(x) dx$.

Closed unit balls in finite dimensional normed linear spaces

Example. The unit balls for the ℓ^p norms in $\mathbb{R}^2$ look like:

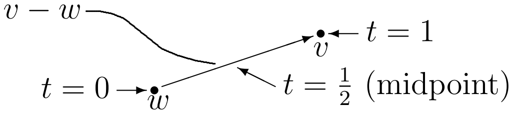


Definition. A subset C of a vector space V is called *convex* if

$$(\forall\, v, w \in C)(\forall\, t \in [0,1]) \qquad tv + (1-t)w \in C.$$

Remarks.

(1) This means that the line segment joining v and w is in C:



$tv + (1-t)w = w + t(v - w)$ is on this line segment.

(2) The linear combination $tv + (1-t)w$ for $t \in [0,1]$ is often called a *convex combination* of v and w.

It is easily seen that the closed unit ball $B = \{v \in V : \|v\| \leq 1\}$ in a *finite dimensional* normed linear space $(V, \|\cdot\|)$ satisfies:

1. B is convex.

2. B is compact.

3. B is symmetric. (This means that if $v \in B$ and $\alpha \in \mathbb{F}$ with $|\alpha| = 1$, then $\alpha v \in B$.)

4. The origin is in the interior of B. (Remark: The condition that 0 be in the interior of a set is independent of the norm: by the norm equivalence theorem, all norms induce the same topology on V, i.e. have the same collection of open sets.)

Conversely, if $\dim V < \infty$ and $B \subset V$ satisfies these four conditions, then there is a unique norm on V for which B is the closed unit ball. In fact, the norm can be obtained from the set B by:

$$\|v\| = \inf\{c > 0 : \frac{v}{c} \in B\}.$$

Exercise: Show that this defines a norm, and that B is its closed unit ball. Uniqueness follows from the fact that in any normed linear space, $\|v\| = \inf\{c > 0 : \frac{v}{c} \in B\}$ where B is the closed unit ball $B = \{v : \|v\| \le 1\}$.

Hence there is a one-to-one correspondence between norms on a finite dimensional vector space and subsets B satisfying these four conditions.

Completeness

Completeness in a normed linear space $(V, \|\cdot\|)$ means completeness in the metric space (V, d), where $d(v, w) = \|v - w\|$: every Cauchy sequence $\{v_n\}$ in V has a limit in V. (The sequence v_n is Cauchy if $(\forall\, \epsilon > 0)(\exists\, N)(\forall\, n, m \ge N)\|v_n - v_m\| < \epsilon$, and the sequence has a limit in V if $(\exists\, v \in V)\|v_n - v\| \to 0$ as $n \to \infty$.)

Example. $\mathbb{F}^n$ in the Euclidean norm $\|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$ is complete.

It is immediate from the definition of a Cauchy sequence that if two norms on V are equivalent, then a sequence is Cauchy with respect to one of the norms if and only if it is Cauchy with respect to the other. Therefore, if two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on a vector space V are equivalent, then $(V, \|\cdot\|_1)$ is complete iff $(V, \|\cdot\|_2)$ is complete.

Every finite-dimensional normed linear space is complete. In fact, we can choose a basis and use it to identify V with $\mathbb{F}^n$. Since $\mathbb{F}^n$ is complete in the Euclidean norm, it is complete in any norm by the norm equivalence theorem.

But not every infinite dimensional normed linear space is complete.

Definition. A complete normed linear space is called a *Banach space*. An inner product space for which the induced norm is complete is called a *Hilbert space*.

Examples. To show that a normed linear space is complete, we must show that every Cauchy sequence converges in that space. Given a Cauchy sequence,

(1) construct what you think is its limit;

(2) show the limit is in the space V;

(3) show the sequence converges to the limit in the norm of V.

(1) Let X be a metric space. Let $C(X)$ denote the vector space of continuous functions $u : X \to \mathbb{F}$. Let $C_b(X)$ denote the subspace of $C(X)$ consisting of all bounded continuous functions $C_b(X) = \{u \in C(X) : (\exists\, K)(\forall\, x \in X)|u(x)| \le K\}$. On $C_b(X)$, define the sup norm $\|u\| = \sup_{x \in X} |u(x)|$.

Fact. $(C_b(X)), \|\cdot\|)$ is complete.

Proof. Let $\{u_n\} \subset C_b(X)$ be Cauchy in $\|\cdot\|$. For each $x \in X$, $|u_n(x) - u_m(x)| \leq \|u_n - u_m\|$, so for each $x \in X$, $\{u_n(x)\}$ is a Cauchy sequence in $\mathbb{F}$. Since $\mathbb{F}$ is complete, this sequence has a limit in $\mathbb{F}$ which we will call $u(x)$: $u(x) = \lim_{n\to\infty} u_n(x)$. This defines a function on X. The fact that $\{u_n\}$ is Cauchy in $C_b(X)$ says that for each $\epsilon > 0$, $(\exists\, N)(\forall\, n, m \geq N)(\forall\, x \in X)|u_n(x) - u_m(x)| < \epsilon$. Taking the limit (for each fixed x) as $m \to \infty$, we get $(\forall\, n \geq N)(\forall\, x \in X)|u_n(x) - u(x)| \leq \epsilon$. Thus $u_n \to u$ uniformly, so u is continuous (since the uniform limit of continuous functions is continuous). Clearly u is bounded (choose N for $\epsilon = 1$; then $(\forall\, x \in X)|u(x)| \leq \|u_N\| + 1$), so $u \in C_b(X)$. And now we have $\|u_n - u\| \to 0$ as $n \to \infty$, i.e., $u_n \to u$ in $(C_b(X), \|\cdot\|)$. □

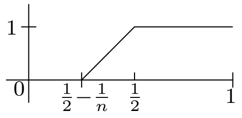
(2) ℓ^p is complete for $1 \leq p \leq \infty$.

$p = \infty$. This is a special case of (1) where $X = \mathbb{N} = \{1, 2, 3, \ldots\}$.

$1 \leq p < \infty$. Let $\{x^n\}$ be a Cauchy sequence in ℓ^p; write $x^n = (x_1^n, x_2^n, \ldots)$. Given $\epsilon > 0$, $(\exists\, N)(\forall\, n, m \geq N)\|x^n - x^m\|_p < \epsilon$. For each $k \in \mathbb{N}$, $|x_k^n - x_k^m| \leq (\sum_{k=1}^\infty |x_k^n - x_k^m|^p)^{\frac{1}{p}} = \|x^n - x^m\|$, so for each $k \in \mathbb{N}$, $\{x_k^n\}_{n=1}^\infty$ is a Cauchy sequence in $\mathbb{F}$, which has a limit: let $x_k = \lim_{n\to\infty} x_k^n$. Let x be the sequence $x = (x_1, x_2, x_3, \ldots)$; so far, we just know that $x \in \mathbb{F}^\infty$. Given $\epsilon > 0$, $(\exists\, N)(\forall\, n, m \geq N)\|x^n - x^m\| < \epsilon$. Then for any K and for $n, m \geq N$, $\left(\sum_{k=1}^K |x_k^n - x_k^m|^p\right)^{\frac{1}{p}} < \epsilon$; taking the limit as $m \to \infty$, $\left(\sum_{k=1}^K |x_k^n - x_k|^p\right)^{\frac{1}{p}} \leq \epsilon$; then taking the limit as $K \to \infty$, $(\sum_{k=1}^\infty |x_k^n - x_k|^p)^{\frac{1}{p}} \leq \epsilon$. Thus $x^N - x \in \ell^p$, so also $x = x^N - (x^N - x) \in \ell^p$, and we have $(\forall\, n \geq N)\|x^n - x\|_p \leq \epsilon$. Thus $\|x^n - x\|_p \to 0$ as $n \to \infty$, i.e., $x_n \to x$ in ℓ^p. □

(3) If X is a compact metric space, then every continuous function $u : X \to \mathbb{F}$ is bounded, so $C(X) = C_b(X)$. In particular, $C(X)$ is complete in the sup norm $\|u\| = \sup_{x\in X} |u(x)|$ (special case of (1)). For example, $C([a, b])$ is complete in the L^∞ norm.

(4) For $1 \leq p < \infty$, $C([a, b])$ is *not* complete in the L^p norm.

Example. On $[0, 1]$, let u_n be: 

Then $u_n \in C([0, 1])$.

Exercise: Show that $\{u_n\}$ is Cauchy in $\|\cdot\|_p$.

We must show that there does *not* exist a $u \in C([0, 1])$ for which $\|u_n - u\|_p \to 0$.

Exercise: Show that if $u \in C([0, 1])$ and $\|u_n - u\|_p \to 0$, then $u(x) \equiv 0$ for $0 \leq x < \frac{1}{2}$ and $u(x) \equiv 1$ for $\frac{1}{2} < x \leq 1$, contradicting the continuity of u at $x = \frac{1}{2}$.

(5) $\mathbb{F}_0^\infty = \{x \in \mathbb{F}^\infty : (\exists\, N)(\forall\, i \geq N)x_i = 0\}$ is *not* complete in any ℓ^p norm $(1 \leq p \leq \infty)$.

$1 \leq p < \infty$. Choose any $x \in \ell^p \setminus \mathbb{F}_0^\infty$, and consider the truncated sequences $y_1 = (x_1, 0, \ldots)$; $y_2 = (x_1, x_2, 0, \ldots)$; $y_3 = (x_1, x_2, x_3, 0, \ldots)$, etc.

Exercise: Show that $\{y_n\}$ is Cauchy in $(\mathbb{F}_0^\infty, \|\cdot\|_p)$, but that there is no $y \in \mathbb{F}_0^\infty$ for which $\|y_n - y\|_p \to 0$.

$p = \infty$. Same idea: choose any $x \in \ell^\infty \setminus \mathbb{F}_0^\infty$ for which $\lim_{i\to\infty} x_i = 0$, and consider the sequence of truncated sequences.

Completion of a Metric Space

Fact. Let (X, d) be a metric space. Then there exists a complete metric space $(\bar{X}, \bar{d})$ and an "inclusion map" $i : X \to \bar{X}$ for which i is injective, i is an isometry from X to $i(X)$ (i.e. $(\forall\, x, y \in X) d(x, y) = \bar{d}(i(x), i(y)))$, and $i(X)$ is dense in $\bar{X}$. Moreover, all such $(\bar{X}, \bar{d})$ are isometrically isomorphic. The metric space $(\bar{X}, \bar{d})$ is called the *completion* of (X, d). (One way to construct such an $\bar{X}$ is to take equivalence classes of Cauchy sequences in X to be elements of $\bar{X}$.)

Representations of Completions

In some situations, the completion of a metric space can be identified with a larger vector space which actually includes X, and whose elements are objects of a similar nature to the elements of X. One example is $\mathbb{R}$ = completion of the rationals $\mathbb{Q}$. The completion of $C([a, b])$ in the L^p norm (for $1 \leq p < \infty$) can be represented as $L^p([a, b])$, the vector space of [equivalence classes of] Lebesgue measurable functions $u : [a, b] \to \mathbb{F}$ for which $\int_a^b |u(x)|^p dx < \infty$, with norm $\|u\|_p = \left(\int_a^b |u(x)|^p dx\right)^{\frac{1}{p}}$. We will discuss this example in more detail next quarter. Recall the fact from metric space theory that a subset of a complete metric space is complete in the restricted metric iff it is closed. This implies

Proposition. Let V be a Banach space, and $W \subset V$ be a subspace. The norm on V restricts to a norm on W. We have:

$$W \text{ is complete} \qquad \text{iff} \qquad W \text{ is closed.}$$

(If you're not familiar with this fact, it is extremely easy to prove. Just follow your nose.)

Further Examples. Define

$C_0(\mathbb{R}^n) = \{u \in C_b(\mathbb{R}^n) : \lim_{|x|\to\infty} u(x) = 0\}$

$C_c(\mathbb{R}^n) = \{u \in C_b(\mathbb{R}^n) : (\exists\, K > 0) \text{ such that } u(x) = 0 \text{ for } |x| \geq K\}.$

We remark that if X is a metric space and $u : X \to \mathbb{F}$ is a function, the *support* of u is defined to be the *closure* of $\{x \in X : u(x) \neq 0\}$. $C_c(\mathbb{R}^n)$ is thus the space of continuous functions on $\mathbb{R}^n$ with *compact support.*

(6) $C_0(\mathbb{R}^n)$ is complete in the sup norm (exercise). This can either be shown directly, or by showing that $C_0(\mathbb{R}^n)$ is a closed subspace of $C_b(\mathbb{R}^n)$.

(7) $C_c(\mathbb{R}^n)$ is *not* complete in the sup norm. In fact, $C_c(\mathbb{R}^n)$ is dense in $C_0(\mathbb{R}^n)$. So $C_0(\mathbb{R}^n)$ is a representation of the completion of $C_c(\mathbb{R}^n)$ in the sup norm.

Series in normed linear spaces

Let $(V, \|\cdot\|)$ be a normed linear space. Consider a series $\sum_{n=1}^{\infty} v_n$ in V.

Definition. We say the series *converges in* V if $\exists\, v \in V$ such that $\lim_{N\to\infty} \|S_N - v\| = 0$, where $S_N = \sum_{n=1}^{N} v_n$ is the N^{th} partial sum. We say this series *converges absolutely* if $\sum_{n=1}^{\infty} \|v_n\| < \infty$.

Caution: If a series converges absolutely in a normed linear space, it does not have to converge in that space.

Example. The series $(1, 0\cdots) + \left(0, \frac{1}{2}, 0\cdots\right) + \left(0, 0, \frac{1}{4}, 0\cdots\right)$ converges absolutely in $\mathbb{F}_0^{\infty}$ with the ℓ^{∞} norm, but it doesn't converge in this space.

Proposition. A normed linear space $(V, \|\cdot\|)$ is complete iff every absolutely convergent series actually converges in $(V, \|\cdot\|)$.

Sketch of proof $(\Rightarrow)$: Given an absolutely convergent series, show that the sequence of partial sums is Cauchy: for $m > n$ $\|S_m - S_n\| \leq \sum_{j=n+1}^{m} \|v_j\|$.
$(\Leftarrow)$: Given a Cauchy sequence $\{v_n\}$, choose $n_1 < n_2 < \cdots$ inductively so that for $k = 1, 2, \ldots$, $(\forall\, n, m \geq n_k)\|v_n - v_m\| \leq 2^{-k}$. Then in particular $\|v_{n_k} - v_{n_{k+1}}\| \leq 2^{-k}$. Show that the series $v_{n_1} + \sum_{k=2}^{\infty}(v_{n_k} - v_{n_{k-1}})$ is absolutely convergent. Let v be its limit. Show that $v_n \to v$. □

A special case encountered in the classical literature is the Weierstrass M-test, obtained by taking $V = C_b(X)$ with the sup norm on a metric space X. The Weierstrass M-test states that if $u_j \in C_b(X)$ and $\sum_{j=1}^{\infty} \sup|u_j| < \infty$, then the series $\sum_{j=1}^{\infty} u_j$ converges uniformly to a continuous function. This is traditionally called the M-test because one sets $M_j = \sup|u_j|$ so that the hypothesis is $\sum_{j=1}^{\infty} M_j < \infty$.

Norms on Operators

If V, W are vector spaces then so is $\mathcal{L}(V,W)$, the space of linear transformations from V to W. We now consider norms on subspaces of $\mathcal{L}(V,W)$.

Bounded Linear Operators and Operator Norms

Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be normed linear spaces. An operator $L \in \mathcal{L}(V,W)$ is called a *bounded linear operator* if

$$\sup_{\|v\|_V=1} \|Lv\|_W < \infty.$$

Let $\mathcal{B}(V,W)$ denote the set of all bounded linear operators from V to W. In the special case $W = \mathbb{F}$ we have *bounded linear functionals*, and we set $V^* = \mathcal{B}(V, \mathbb{F})$.

Proposition. If $\dim V < \infty$, then every linear operator is bounded, so $\mathcal{L}(V,W) = \mathcal{B}(V,W)$ and $V^* = V'$.

Proof. Choose a basis $\{v_1, \ldots, v_n\}$ for V with corresponding coordinates $x_1, \ldots, x_n$. Then $\sum_{i=1}^n |x_i|$ is a norm on V, so by the Norm Equivalence Theorem, there exists M so that $\sum_{i=1}^n |x_i| \leq M\|v\|$ for $v = \sum x_i v_i$. Then

$$\begin{aligned}
\|Lv\|_W &= \left\| L\left(\sum_{i=1}^n x_i v_i\right)\right\|_W \\
&\leq \sum_{i=1}^n |x_i| \cdot \|Lv_i\|_W \\
&\leq \left(\max_{1\leq i\leq n} \|Lv_i\|_W\right) \sum_{i=1}^n |x_i| \\
&\leq \left(\max_{1\leq i\leq n} \|Lv_i\|_W\right) M\|v\|_V,
\end{aligned}$$

so

$$\sup_{\|v\|_V=1} \|Lv\|_W \leq \left(\max_{1\leq i\leq n} \|Lv_i\|_W\right) \cdot M < \infty.$$

□

Caution. If L is a bounded linear operator, it is not necessarily the case that $\{\|Lv\|_W : v \in V\}$ is a bounded set of $\mathbb{R}$. In fact, if it is, then $L \equiv 0$ (exercise). Similarly, if a linear functional is a bounded linear functional, it does *not* mean that there is an M for which $|f(v)| \leq M$ for all $v \in V$. The word "bounded" is used in different ways in different contexts.

Remark: It is easy to see that if $L \in \mathcal{L}(V,W)$, then

$$\sup_{\|v\|_V=1} \|Lv\|_W = \sup_{\|v\|_V\leq 1} \|Lv\|_W = \sup_{v\neq 0}(\|Lv\|_W/\|v\|_V).$$

Therefore L is bounded if and only if there is a constant M so that $\|Lv\|_W \leq M\|v\|_V$ for all $v \in V$.

Examples:

(1) Let $V = \mathcal{P}$ be the space of polynomials with norm $\|p\| = \sup_{0\leq x\leq 1} |p(x)|$. The differentiation operator $\frac{d}{dx} : \mathcal{P} \to \mathcal{P}$ is not a bounded linear operator: $\|x^n\| = 1$ for all $n \geq 1$, but $\left\|\frac{d}{dx}x^n\right\| = \|nx^{n-1}\| = n$.

(2) Let $V = \mathbb{F}_0^\infty$ with ℓ^p-norm for some p, $1 \leq p \leq \infty$. Let L be diagonal, so $Lx = (\lambda_1 x_1, \lambda_2 x_2, \lambda_3 x_3, \ldots)^T$ for $x \in \mathbb{F}_0^\infty$, where $\lambda_i \in \mathbb{C}$, $i \geq 1$. Then L is a bounded linear operator iff $\sup_i |\lambda_i| < \infty$.

One of the reasons that boundedness of a linear transformation is an important concept is the following result.

Proposition. Let $L : V \to W$ be a linear transformation between normed vector spaces. Then L is bounded iff L is continuous.

Proof. First suppose L is bounded. Thus there is C so that $\|Lv\|_W \leq C\|v\|_V$ for all $v \in V$. Then for all v_1, $v_2 \in V$,

$$\|Lv_1 - Lv_2\|_W = \|L(v_1 - v_2)\|_W \leq C\|v_1 - v_2\|_V.$$

Hence L is continuous (given ϵ, let $\delta = C^{-1}\epsilon$, etc.). In fact, L is Lipschitz continuous with Lipschitz constant C (and in particular is uniformly continuous).

Conversely, suppose L is continuous. Then L is continuous at $v = 0$. Let $\epsilon = 1$. Then $\exists\,\delta > 0$ so that if $\|v\|_V \leq \delta$, then $\|Lv\|_W \leq 1$. For any $v \neq 0$, $\left\|\frac{\delta}{\|v\|_V}v\right\|_V = \delta$, so $\left\|L\left(\frac{\delta}{\|v\|_V}v\right)\right\|_W \leq 1$, i.e., $\|Lv\|_W \leq \frac{1}{\delta}\|v\|_V$. Let $C = \frac{1}{\delta}$. □

Definition. Let $L : V \to W$ be a bounded linear operator between normed linear spaces, i.e., $L \in \mathcal{B}(V, W)$. Define the operator norm of L to be

$$\|L\| = \sup_{\|v\|_V\leq 1} \|Lv\|_W \left(= \sup_{\|v\|_V=1} \|Lv\|_W = \sup_{v\neq 0} \left(\|Lv\|_W / \|v\|_V\right)\right).$$

Remark. We have $\|Lv\|_W \leq \|L\| \cdot \|v\|_V$ for all $v \in V$. In fact, $\|L\|$ is the smallest constant with this property: $\|L\| = \min\{C \geq 0 : (\forall\, v \in V)\ \|Lv\|_W \leq C\|v\|_V\}$.

We now show that $\mathcal{B}(V, W)$ is a vector space (a subspace of $\mathcal{L}(V, W)$). If $L \in \mathcal{B}(V, W)$ and $\alpha \in \mathbb{F}$, clearly $\alpha L \in \mathcal{B}(V, W)$ and $\|\alpha L\| = |\alpha| \cdot \|L\|$. If $L_1, L_2 \in \mathcal{B}(V, W)$, then

$$\|(L_1 + L_2)v\|_W \leq \|L_1 v\|_W + \|L_2 v\|_W \leq (\|L_1\| + \|L_2\|)\|v\|_V,$$

so $L_1 + L_2 \in \mathcal{B}(V, W)$, and $\|L_1 + L_2\| \leq \|L_1\| + \|L_2\|$. It follows that the operator norm is indeed a norm on $\mathcal{B}(V, W)$. $\|\cdot\|$ is sometimes called the operator norm on $\mathcal{B}(V, W)$ induced by the norms $\|\cdot\|_V$ and $\|\cdot\|_W$ (as it clearly depends on both $\|\cdot\|_V$ and $\|\cdot\|_W$).

Completeness of $\mathcal{B}(V,W)$ when W is complete

Proposition. If W is complete, then $\mathcal{B}(V,W)$ is complete. In particular, V^* is always complete (since $\mathbb{F}$ is), whether or not V is.

Proof. If $\{L_n\}$ is Cauchy in $\mathcal{B}(V,W)$, then $(\forall\, v \in V)\{L_n v\}$ is Cauchy in W, so the limit $\lim_{n\to\infty} L_n v \equiv Lv$ exists in W. Clearly the map $L : V \to W$ so defined is linear, and it is easy to see that $L \in \mathcal{B}(V,W)$ and $\|L_n - L\| \to 0$. □

Dual norms

In the special case $W = \mathbb{F}$, the norm $\|f\|^* = \sup_{\|v\|\leq 1} |f(v)|$ on V^* is called the *dual norm* to that on V. If $\dim V < \infty$, then we can choose bases and identify V and V^* with $\mathbb{F}^n$. Thus every norm on $\mathbb{F}^n$ has a dual norm on $\mathbb{F}^n$. We sometimes write $\mathbb{F}^{n*}$ for $\mathbb{F}^n$ when it is being identified with V^*. Consider some examples.

(1) If $\mathbb{F}^n$ is given the ℓ^1-norm, then the dual norm is

$$\|f\|^* = \max_{\|x\|_1\leq 1} \left|\sum_{i=1}^n f_i x_i\right| \qquad \text{for} \qquad f = (f_1, \ldots, f_n) \in \mathbb{F}^{n*},$$

which is easily seen to be the ℓ^∞-norm $\|f\|_\infty$ (exercise).

(2) If $\mathbb{F}^n$ is given the ℓ^∞-norm, then the dual norm is

$$\|f\|^* = \max_{\|x\|_\infty\leq 1} \left|\sum_{i=1}^n f_i x_i\right| \qquad \text{for} \qquad f = (f_1, \ldots, f_n) \in \mathbb{F}^{n*},$$

which is easily seen to be the ℓ^1-norm $\|f\|_1$ (exercise).

(3) The dual norm to the ℓ^2-norm on $\mathbb{F}^n$ is again the ℓ^2-norm; this follows easily from the Schwarz inequality (exercise). The ℓ^2-norm is the only norm on $\mathbb{F}^n$ which equals its own dual norm; see the homework.

(4) Let $1 < p < \infty$. The dual norm to the ℓ^p-norm on $\mathbb{F}^n$ is the ℓ^q-norm, where $\frac{1}{p} + \frac{1}{q} = 1$. The key inequality is Hölder's inequality: $|\sum_{i=1}^n f_i x_i| \leq \|f\|_q \cdot \|x\|_p$. We will be primarily interested in the cases $p = 1, 2, \infty$. (Note: $\frac{1}{p} + \frac{1}{q} = 1$ in an extended sense when $p = 1$ and $q = \infty$, or when $p = \infty$ and $q = 1$; Hölder's inequality is trivial in these cases.)

It is instructive to consider linear functionals and the dual norm geometrically. Recall that a norm on $\mathbb{F}^n$ can be described geometrically by its closed unit ball B, a compact convex set. The geometric realization of a linear functional (excluding the zero functional) is a hyperplane. A hyperplane in $\mathbb{F}^n$ is a set of the form $\{x \in \mathbb{F}^n : \sum_{i=1}^n f_i x_i = c\}$, where f_i, $c \in \mathbb{F}$ and not all $f_i = 0$ (sets of this form are sometimes called *affine* hyperplanes if the term "hyperplane" is being reserved for a subspace of $\mathbb{F}^n$ of dimension $n-1$). In fact, there is a natural $1-1$ correspondence between $\mathbb{F}^{n*}\backslash\{0\}$ and the set of hyperplanes in $\mathbb{F}^n$ which do not contain the origin: to $f=(f_1,\ldots,f_n)\in\mathbb{F}^{n*}$, associate the hyperplane $\{x\in\mathbb{F}^n : f(x)=f_1x_1+\cdots+f_nx_n=1\}$; since every hyperplane not containing 0 has a unique equation of this form, this is a $1-1$ correspondence as claimed.

If $\mathbb{F}=\mathbb{C}$ it is often more appropriate to use real hyperplanes in $\mathbb{C}^n=\mathbb{R}^{2n}$; if $z\in\mathbb{C}^n$ and we write $z_j=x_j+iy_j$, then a real hyperplane not containing $\{0\}$ has a unique equation of the form $\sum_{j=1}^n(a_jx_j+b_jy_j)=1$ where $a_j,b_j\in\mathbb{R}$, and not all of the a_j's and b_j's vanish. Observe that this equation is of the form $\mathcal{R}e\left(\sum_{j=1}^n f_jz_j\right)=1$ where $f_j=a_j-ib_j$ is uniquely determined. Thus the real hyperplanes in $\mathbb{C}^n$ not containing $\{0\}$ are all of the form $\mathcal{R}ef(z)=1$ for a unique $f\in\mathbb{C}^{n*}\backslash\{0\}$. For $\mathbb{F}=\mathbb{R}$ or $\mathbb{C}$ and $f\in\mathbb{F}^{n*}\backslash\{0\}$, we denote by H_f the real hyperplane $H_f=\{v\in\mathbb{F}^n : \mathcal{R}ef(v)=1\}$.

Lemma. If $(V,\|\cdot\|)$ is a normed linear space and $f\in V^*$, then the dual norm of f satisfies $\|f\|^*=\sup_{\|v\|\leq 1}\mathcal{R}ef(v)$.

Proof. Since $\mathcal{R}ef(v)\leq|f(v)|$,

$$\sup_{\|v\|\leq 1}\mathcal{R}ef(v)\leq\sup_{\|v\|\leq 1}|f(v)|=\|f\|^*.$$

For the other direction, choose a sequence $\{v_j\}$ from V with $\|v_j\|=1$ and $|f(v_j)|\to\|f\|^*$. Taking $\theta_j=-\arg f(v_j)$ and setting $w_j=e^{i\theta_j}v_j$, we have $\|w_j\|=1$ and $f(w_j)=|f(v_j)|\to\|f\|^*$, so $\sup_{\|v\|\leq 1}\mathcal{R}ef(v)\geq\|f\|^*$. □

We can reformulate the Lemma geometrically in terms of real hyperplanes and the unit ball in the original norm. Note that any real hyperplane H_f in $\mathbb{F}^n$ not containing the origin divides $\mathbb{F}^n$ into two closed half-spaces whose intersection is H_f. The one of these half-spaces which contains the origin is $\{v\in\mathbb{F}^n : \mathcal{R}ef(v)\leq 1\}$.

Proposition. Let $\|\cdot\|$ be a norm on $\mathbb{F}^n$ with closed unit ball B. A linear functional $f\in\mathbb{F}^{n*}\backslash\{0\}$ satisfies $\|f\|^*\leq 1$ if and only if B lies completely in the half space $\{v\in\mathbb{F}^n : \mathcal{R}ef(v)\leq 1\}$ determined by f containing the origin.

Proof. This is just a geometric restatement of the Lemma above. The Lemma shows that $f\in\mathbb{F}^{n*}$ satisfies $\|f\|^*\leq 1$ iff $\sup_{\|v\|\leq 1}\mathcal{R}ef(v)\leq 1$, which states geometrically that B is contained in the closed half-space $\mathcal{R}ef(v)\leq 1$. □

It is interesting to translate this criterion into a concrete geometric realization of the dual unit ball in specific examples; see the homework.

The following Theorem is a fundamental result concerning dual norms.

Theorem. Let $(V,\|\cdot\|)$ be a normed linear space and $v\in V$. Then there exists $f\in V^*$ such that $\|f\|^*=1$ and $f(v)=\|v\|$.

This is an easy consequence of the following result.

Theorem. Let $(V,\|\cdot\|)$ be a normed linear space and let $W\subset V$ be a subspace. Give W the norm which is the restriction of the norm on V. If $g\in W^*$, there exists $f\in V^*$ satisfying $f|_W=g$ and $\|f\|_{V^*}=\|g\|_{W^*}$.

The first Theorem follows from the second as follows. If $v \neq 0$, take $W = \operatorname{span}\{v\}$ and define $g \in W^*$ by $g(v) = \|v\|$, extended by linearity. Then $\|g\|_{W^*} = 1$, so the f produced by the second Theorem does the job in the first theorem. If $v = 0$, choose any nonzero vector in V and apply the previous argument to it to get f.

The second Theorem above is a version of the Hahn-Banach Theorem. Its proof in infinite dimensions is non-constructive, uses Zorn's Lemma, and will not be given here. We refer to a real analysis or functional analysis text for the proof (e.g., Royden *Real Analysis* or Folland *Real Analysis*). In this course we will not have further occasion to consider the second Theorem. For convenience we will refer to the first Theorem as the Hahn-Banach Theorem, even though this is not usual terminology.

We will give a direct proof of the first Theorem in the case when V is finite-dimensional. The proof uses the Proposition above and some basic properties of closed convex sets in $\mathbb{R}^n$. When we get to the proof of the first Theorem below, the closed convex set will be the closed unit ball in the given norm.

Let K be a closed convex set in $\mathbb{R}^n$ such that $K \neq \emptyset$, $K \neq \mathbb{R}^n$. A hyperplane $H \subset \mathbb{R}^n$ is said to be a *supporting hyperplane* for K if the following two conditions hold:

1. K lies completely in (at least) one of the two closed half-spaces determined by H
2. $H \cap K \neq \emptyset$.

We associate to K the family $\mathcal{H}_K$ of all closed half-spaces S containing K and bounded by a supporting hyperplane for K.

Fact 1. $K = \cap_{S \in \mathcal{H}_K} S$.

Fact 2. If $x \in \partial K$, then there is a supporting hyperplane for K containing x.

In order to prove these facts, it is useful to use the constructions of Euclidean geometry in $\mathbb{R}^n$ (even though the above statements just concern the vector space structure and are independent of the inner product).

Proof of Fact 1. Clearly $K \subset \cap_{\mathcal{H}_K} S$. For the other inclusion, suppose $y \notin K$. Let $x \in K$ be a closest point in K to y. Let H be the hyperplane through x which is normal to $y - x$. We claim that K is contained in the half-space S determined by H which does not contain y. If so, then $S \in \mathcal{H}_K$ and $y \notin S$, so $y \notin \cap_{\mathcal{H}_K} S$ and we are done. To prove the claim, if there were a point z in K on the same side of H as y, then by convexity of K, the whole line segment between z and x would be in K. But then an easy computation shows that points on this line segment near x would be closer than x to y, a contradiction. Note as a consequence of this argument that a closed convex set K has a unique closest point to a point in $\mathbb{R}^n \setminus K$.

Proof of Fact 2. Let $x \in \partial K$ and choose points $y_j \in \mathbb{R}^n \setminus K$ so that $y_j \to x$. For each such y_j, let $x_j \in K$ be the closest point in K to y_j. By the argument in Fact 1, the hyperplane through x_j and normal to $y_j - x_j$ is supporting for K. Since $y_j \to x$, we must have also $x_j \to x$. If we let $u_j = \frac{y_j - x_j}{|y_j - x_j|}$, then u_j is a sequence on the Euclidean unit sphere, so a subsequence converges to a unit vector u. By taking limits, it follows that the hyperplane through x and normal to u is supporting for K.

Proof of first Theorem if $\dim(V) < \infty$. By choosing a basis we may assume that $V = \mathbb{F}^n$. Clearly it suffices to assume that $\|v\| = 1$. The closed unit ball B is a closed convex set in $\mathbb{F}^n$ ($= \mathbb{R}^n$ or $\mathbb{R}^{2n}$). By Fact 2 above, there is a supporting (real) hyperplane H for B with $v \in H$. Since the origin is in the interior of B, the origin is not in H. Thus this supporting hyperplane is of the form H_f for some uniquely determined $f \in \mathbb{F}^{n*}$. Since H is supporting, B lies completely in the half-space determined by f containing the origin, so the Proposition above shows that $\|f\|^* \leq 1$. Then we have

$$1 = \mathcal{R}e f(v) \leq |f(v)| \leq \|f\|^* \|v\| = \|f\|^* \leq 1.$$

It follows that all terms in the above string of inequalities must be 1, from which we deduce that $\|f\|^* = 1$ and $f(v) = 1$ as desired. □

The Second Dual

Let $(V, \|\cdot\|)$ be a normed linear space, V^* be its dual equipped with the dual norm, and V^{**} be the dual of V^* with the norm dual to that on V^*. Given $v \in V$, define $v^{**} \in V^{**}$ by $v^{**}(f) = f(v)$; since $|v^{**}(f)| \leq \|f\|^* \cdot \|v\|$, $v^{**} \in V^{**}$ and $\|v^{**}\| \leq \|v\|$. By the Hahn-Banach theorem, $\exists\, f \in V^*$ with $\|f\|^* = 1$ and $f(v) = \|v\|$. Thus $v^{**}(f) = \|v\|$, so $\|v^{**}\| = \sup_{\|f\|^*=1} |v^{**}(f)| \geq \|v\|$. Hence $\|v^{**}\| = \|v\|$, so the mapping $v \mapsto v^{**}$ from V into V^{**} is an isometry of V onto the range of this map. In general, this embedding is not surjective; if it is, then $(V, \|\cdot\|)$ is called *reflexive*.

In finite dimensions, dimension arguments imply this map is surjective. Thus we can identify V with V^{**}, and under this identification the dual norm to the dual norm is just the original norm on V.

The second dual provides one way to realize the completion of a normed linear space $(V, \|\cdot\|)$. Let us identify V with its image under the map $v \mapsto v^{**}$, so that we view $V \subset V^{**}$. Since V^{**} is always complete whether V is or not, the closure $\overline{V} \subset V^{**}$ is a closed subspace of a complete normed linear space, and hence is complete when given the norm which is the restriction of the norm on V^{**}. Since V is dense in $\overline{V}$, $\overline{V}$ is a realization of the completion of V.

Submultiplicative Norms

If V is a vector space, $\mathcal{L}(V, V) = \mathcal{L}(V)$ is an algebra with composition as multiplication. A norm on a subalgebra of $\mathcal{L}(V)$ is said to be *submultiplicative* if $\|A \circ B\| \leq \|A\| \cdot \|B\|$. (Horn-Johnson calls this a matrix norm in finite dimensions.)

Example: For $A \in \mathbb{C}^{n \times n}$, define $\|A\| = \sup_{1 \leq i,j \leq n} |a_{ij}|$. This norm is not submultiplicative: if $A = B = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix}$, then $\|A\| = \|B\| = 1$, but $AB = A^2 = nA$ so $\|AB\| = n$. However, the norm $A \mapsto n \sup_{1 \leq i,j \leq n} |a_{ij}|$ is submultiplicative (exercise).

Fact: Operator norms are always submultiplicative.

In fact, if U, V, W are normed linear spaces and $L \in \mathcal{B}(U, V)$ and $M \in \mathcal{B}(V, W)$, then for $u \in U$,

$$\|(M \circ L)(u)\|_W = \|M(Lu)\|_W \leq \|M\| \cdot \|Lu\|_V \leq \|M\| \cdot \|L\| \cdot \|u\|_U,$$

so $M \circ L \in \mathcal{B}(U, W)$ and $\|M \circ L\| \leq \|M\| \cdot \|L\|$. The special case $U = V = W$ shows that the operator norm on $\mathcal{B}(V)$ is submultiplicative (and $L, M \in \mathcal{B}(V) \Rightarrow M \circ L \in \mathcal{B}(V)$).

Norms on Matrices

If $A \in \mathbb{C}^{m\times n}$, we denote by $A^T \in \mathbb{C}^{n\times m}$ the transpose of A, and by $A^* = A^H = \bar{A}^T$ the conjugate-transpose (or Hermitian transpose) of A. Commonly used norms on $\mathbb{C}^{m\times n}$ are the following. (We use the notation of H-J.)

$$\begin{aligned}
\|A\|_1 &= \sum_{i=1}^m \sum_{j=1}^n |a_{ij}| && \text{(the } \ell^1\text{-norm on } A \text{ as if it were in } \mathbb{C}^{mn}) \\
\|A\|_\infty &= \max\nolimits_{i,j} |a_{ij}| && \text{(the } \ell^\infty\text{-norm on } A \text{ as if it were in } \mathbb{C}^{mn}) \\
\|A\|_2 &= \left(\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2\right)^{\frac{1}{2}} = (\operatorname{tr} A^*A)^{\frac{1}{2}} && \text{(the } \ell^2\text{-norm on } A \text{ as if it were in } \mathbb{C}^{mn})
\end{aligned}$$

The norm $\|A\|_2$ is called the Hilbert-Schmidt norm of A, or the Frobenius norm of A, and is often denoted $\|A\|_{HS}$ or $\|A\|_F$. It is sometimes called the Euclidean norm of A. This norm comes from the inner product $\langle B, A\rangle = \sum_{i=1}^m \sum_{j=1}^n \overline{b_{ij}} a_{ij} = \operatorname{tr}(B^*A)$.

We also have the following p-norms for matrices: let $1 \leq p \leq \infty$ and $A \in \mathbb{C}^{m\times n}$, then

$$|||A|||_p = \max_{\|x\|_p=1} \|Ax\|_p.$$

This is the operator norm induced by the ℓ^p norm $\|\cdot\|_p$ on $\mathbb{C}^m$ and $\mathbb{C}^n$.

Caution: $|||A|||_p$ is a quite non-standard notation; the standard notation is $\|A\|_p$. In numerical analysis, the Frobenius norm is typically denoted $\|A\|_F$. We will, however, go ahead and use the notation of H-J.

Using arguments similar to those identifying the dual norms to the ℓ^1- and ℓ^∞-norms on $\mathbb{C}^n$, it can be easily shown that

$$\begin{aligned}
|||A|||_1 &= \max\nolimits_{1\leq j\leq n} \sum\nolimits_{i=1}^m |a_{ij}| && \text{(maximum (absolute) column sum)} \\
|||A|||_\infty &= \max\nolimits_{1\leq i\leq m} \sum\nolimits_{j=1}^n |a_{ij}| && \text{(maximum (absolute) row sum)}
\end{aligned}$$

The norm $|||A|||_2$ is often called the spectral norm (we will show later that it equals the square root of the largest eigenvalue of A^*A).

Except for $\|\cdot\|_\infty$, all the above norms are submultiplicative on $\mathbb{C}^{n\times n}$. In fact, they satisfy a stronger submultiplicativity property even for non-square matrices known as consistency, which we discuss next.

Consistent Matrix Norms

The concept of submultiplicativity can be extended to non-square matrices.

Definition. Let μ, ν, ρ be norms on $\mathbb{C}^{m\times n}$, $\mathbb{C}^{n\times k}$, $\mathbb{C}^{m\times k}$, respectively. We say that μ, ν, ρ are *consistent* if $\forall\, A \in \mathbb{C}^{m\times n}$ and $\forall\, B \in \mathbb{C}^{n\times k}$,

$$\rho(AB) \leq \mu(A)\nu(B).$$

Observe that in the case $m = n = k$ and $\rho = \mu = \nu$, this definition is equivalent to stating that the norm on $\mathbb{C}^{n\times n}$ is submultiplicative.

The next proposition shows that with the exception of $\|\cdot\|_\infty$, each of the norms defined above constitutes a consistent family of matrix norms.

Proposition. Let $\|\cdot\|$ be any one of the norms $|||\cdot|||_p$, $1 \leq p \leq \infty$, or $\|\cdot\|_1$, or $\|\cdot\|_2$. If $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{n\times k}$, then we have

$$\|AB\| \leq \|A\|\|B\|.$$

Proof. We saw in the previous section that operator norms always satisfy this consistency property. This establishes the result for the norms $|||\cdot|||_p$, $1 \leq p \leq \infty$. We reproduce the argument:

$$|||AB|||_p = \max_{\|x\|_p=1} \|ABx\|_p \leq \max_{\|x\|_p=1} |||A|||_p\|Bx\|_p \leq \max_{\|x\|_p=1} |||A|||_p|||B|||_p\|x\|_p = |||A|||_p|||B|||_p.$$

For $\|\cdot\|_1$ and $\|\cdot\|_2$, the proposition follows by direct estimation:

$$\|AB\|_1 = \sum_{i=1}^{m}\sum_{l=1}^{k}\left|\sum_{j=1}^{n} a_{ij}b_{jl}\right| \leq \sum_{i=1}^{m}\sum_{l=1}^{k}\sum_{j=1}^{n}\sum_{r=1}^{n} |a_{ij}||b_{rl}| = \|A\|_1\|B\|_1$$

$$\|AB\|_2^2 = \sum_{i=1}^{m}\sum_{l=1}^{k}\left|\sum_{j=1}^{n} a_{ij}b_{jl}\right|^2 \leq \sum_{i=1}^{m}\sum_{l=1}^{k}\left(\sum_{j=1}^{n}|a_{ij}|^2\right)\left(\sum_{j=1}^{n}|b_{jl}|^2\right) = \|A\|_2^2\|B\|_2^2$$

□

Note that the proposition fails for $\|\cdot\|_\infty$: we have already seen that submultiplicativity fails for this norm on $\mathbb{C}^{n\times n}$. Note also that $\|\cdot\|_1$ and $\|\cdot\|_2$ on $\mathbb{C}^{n\times n}$ are definitely not the operator norm for any norm on $\mathbb{C}^n$, since $\|I\| = 1$ for any operator norm.

Proposition. For $A \in \mathbb{C}^{m\times n}$, we have

$$|||A|||_1 \leq \|A\|_1 \qquad \text{and} \qquad |||A|||_2 \leq \|A\|_2$$

Proof. The first of these follows immediately from the explicit description of these norms. However, both of these also follow from the general fact that the operator norm on $\mathbb{C}^{m\times n}$

determined by norms on $\mathbb{C}^m$ and $\mathbb{C}^n$ is the smallest norm consistent with these two norms. Specifically, using the previous proposition with $k = 1$, we have

$$|||A|||_2 = \max_{\|x\|_2=1} \|Ax\|_2 \leq \max_{\|x\|_2=1} \|A\|_2 \|x\|_2 = \|A\|_2$$

and similarly for the 1-norms. □

The bound $|||A|||_2 \leq \|A\|_2$ is useful for estimating $|||A|||_2$ since this is not explicit, but $\|A\|_2$ is explicit.

Analysis with Operators

Throughout this discussion, let V be a Banach space. Since V is complete, $\mathcal{B}(V) = \mathcal{B}(V, V)$ is also complete (in the operator norm). We want to define functions of an operator $L \in \mathcal{B}(V)$. We can compose L with itself, so we can form powers $L^k = L \circ \cdots \circ L$, and thus we can define polynomials in L: if $p(z) = a_0 + a_1 z + \cdots + a_n z^n$, then $p(L) \equiv a_0 I + a_1 L + \cdots + a_n L^n$. By taking limits, we can form power series, and thus analytic functions of L. For example, consider the series

$$e^L = \sum_{k=0}^{\infty} \frac{1}{k!} L^k = I + L + \frac{1}{2} L^2 + \cdots$$

(note L^0 is the identity I by definition). This series converges in the operator norm on $\mathcal{B}(V)$: by submultiplicativity, $\|L^k\| \leq \|L\|^k$, so

$$\sum_{k=0}^{\infty} \frac{1}{k!} \|L^k\| \leq \sum_{k=0}^{\infty} \frac{1}{k!} \|L\|^k = e^{\|L\|} < \infty;$$

since the series converges absolutely and $\mathcal{B}(V)$ is complete (recall V is a Banach space), it converges in the operator norm to an operator in $\mathcal{B}(V)$ which we call e^L (note that $\|e^L\| \leq e^{\|L\|}$). In the finite dimensional case, this says that for $A \in \mathbb{F}^{n \times n}$, each component of the partial sum $\sum_{k=0}^{N} \frac{1}{k!} A^k$ converges as $N \to \infty$; the limiting matrix is e^A.

Another fundamental example is the Neumann series. We will say that an operator in $\mathcal{B}(V)$ is invertible if it is bijective (i.e., invertible as a point-set mapping from V onto V, which implies that the inverse map is well-defined and linear) *and* that its inverse is also in $\mathcal{B}(V)$. It is a consequence of the closed graph theorem (see Royden or Folland) that if $L \in \mathcal{B}(V)$ is bijective (and V is a Banach space), then its inverse map L^{-1} is also in $\mathcal{B}(V)$.

Note: $\mathcal{B}(V)$ has a ring structure using the addition of operators and composition of operators as the multiplication; the identity of multiplication is just the identity operator I. Our concept of invertibility is equivalent to invertibility in this ring: if $L \in \mathcal{B}(V)$ and $\exists\, M \in \mathcal{B}(V) \ni LM = ML = I$, then $ML = I \Rightarrow L$ injective and $LM = I \Rightarrow L$ surjective. Note that this ring in general is *not* commutative.

Proposition. If $L \in \mathcal{B}(V)$ and $\|L\| < 1$, then $I - L$ is invertible, and the Neumann series $\sum_{k=0}^{\infty} L^k$ converges in the operator norm to $(I - L)^{-1}$.

Remark. Formally we can guess this result since the power series of $\frac{1}{1-z}$ centered at $z = 0$ is $\sum_{k=0}^{\infty} z^k$ with radius of convergence 1.

Proof. If $\|L\| < 1$, then

$$\sum_{k=0}^{\infty} \|L^k\| \le \sum_{k=0}^{\infty} \|L\|^k = \frac{1}{1-\|L\|} < \infty,$$

so the Neumann series $\sum_{k=0}^{\infty} L^k$ converges to an operator in $\mathcal{B}(V)$. Now if $S_j, S, T \in \mathcal{B}(V)$ and $S_j \to S$ in $\mathcal{B}(V)$, then $\|S_j - S\| \to 0$, so $\|S_jT - ST\| \le \|S_j - S\| \cdot \|T\| \to 0$ and $\|TS_j - TS\| \le \|T\| \cdot \|S_j - S\| \to 0$, and thus $S_jT \to ST$ and $TS_j \to TS$ in $\mathcal{B}(V)$. Thus

$$(I-L)\left(\sum_{k=0}^{\infty} L^k\right) = \lim_{N\to\infty}(I-L)\sum_{k=0}^{N} L^k = \lim_{N\to\infty}(I - L^{N+1}) = I$$

(using $\|L^{N+1}\| \le \|L\|^{N+1} \to 0$), and similarly $\left(\sum_{k=0}^{\infty} L^k\right)(I-L) = I$. So $I-L$ is invertible and $(I-L)^{-1} = \sum_{k=0}^{\infty} L^k$. □

This is a very useful fact: a perturbation of I by an operator of norm < 1 is invertible. This implies, among other things, that the set of invertible operators in $\mathcal{B}(V)$ is an open subset of $\mathcal{B}(V)$ (in the operator norm); see the homework.

Clearly the power series arguments used above can be generalized. Let $f(z)$ be analytic on the disk $\{|z| < R\} \subset \mathbb{C}$, with power series $f(z) = \sum_{k=0}^{\infty} a_k z^k$ (which has radius of convergence at least R). If $L \in \mathcal{B}(V)$ and $\|L\| < R$, then the series $\sum_{k=0}^{\infty} a_k L^k$ converges absolutely, and thus converges to an element of $\mathcal{B}(V)$ which we call $f(L)$ (recall V is a Banach space). It is easy to check that usual operational properties hold, for example $(fg)(L) = f(L)g(L) = g(L)f(L)$. However, one must be careful to remember that operators do not commute in general. So, for example, $e^{L+M} \neq e^L e^M$ in general, although if L and M commute (i.e. $LM = ML$), then $e^{L+M} = e^L e^M$.

Let $L(t)$ be a 1-parameter family of operators in $\mathcal{B}(V)$, where $t \in (a,b)$. Since $\mathcal{B}(V)$ is a metric space, we know what it means for $L(t)$ to be a continuous function of t. We can define differentiability as well: $L(t)$ is differentiable at $t = t_0 \in (a,b)$ if $L'(t_0) = \lim_{t\to t_0} \frac{L(t)-L(t_0)}{t-t_0}$ exists in the norm on $\mathcal{B}(V)$. For example, it is easily checked that for $L \in \mathcal{B}(V)$, e^{tL} is differentiable in t for all $t \in \mathbb{R}$, and $\frac{d}{dt}e^{tL} = Le^{tL} = e^{tL}L$.

We can similarly consider families of operators in $\mathcal{B}(V)$ depending on several real parameters or on complex parameter(s). A family $L(z)$ where $z = x + iy \in \Omega^{\text{open}} \subset \mathbb{C}$ $(x, y \in \mathbb{R})$ is said to be holomorphic in Ω if the partial derivatives $\frac{\partial}{\partial x}L(z)$, $\frac{\partial}{\partial y}L(z)$ exist and are continuous in Ω, and $L(z)$ satisfies the Cauchy-Riemann equation $\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right)L(z) = 0$ in Ω. As in complex analysis, this is equivalent to the assumption that in a neighborhood of each point $z_0 \in \Omega$, $L(z)$ is given by the $\mathcal{B}(V)$-norm convergent power series $L(z) = \sum_{k=0}^{\infty} \frac{1}{k!}(z-z_0)^k \left(\frac{d}{dz}\right)^k L(z_0)$.

One can also integrate families of operators. If $L(t)$ depends continuously on $t \in [a,b]$, then it can be shown using the same estimates as for $\mathbb{F}$-valued functions (and the uniform continuity of $L(t)$ since $[a,b]$ is compact) that the Riemann sums

$$\frac{b-a}{N}\sum_{k=0}^{n-1} L\left(a + \frac{k}{N}(b-a)\right)$$

converge in $\mathcal{B}(V)$-norm (recall V is a Banach space) as $n \to \infty$ to an operator in $\mathcal{B}(V)$, denoted $\int_a^b L(t)dt$. (More general Riemann sums than just the left-hand "rectangular rule" with equally spaced points can be used.) Many results from standard calculus carry over, including

$$\left\| \int_a^b L(t)dt \right\| \leq \int_a^b \|L(t)\|dt$$

which follows directly from

$$\left\| \frac{b-a}{N} \sum_{k=0}^{N-1} L\left(a + \frac{k}{N}(b-a)\right) \right\| \leq \frac{b-a}{N} \sum_{k=0}^{N-1} \left\| L\left(a + \frac{k}{N}(b-a)\right) \right\|.$$

By parameterizing paths in $\mathbb{C}$, one can define line integrals of holomorphic families of operators. We will discuss such constructions further as we need them.

Adjoint Transformations

Recall that if $L \in \mathcal{L}(V,W)$, the adjoint transformation $L' : W' \to V'$ is given by $(L'g)(v) = g(Lv)$.

Proposition. Let V, W be normed linear spaces. If $L \in \mathcal{B}(V,W)$, then $L'(W^*) \subset V^*$. Moreover, $L' \in \mathcal{B}(W^*,V^*)$ and $\|L'\| = \|L\|$.

Proof. For $g \in W^*$,

$$|(L'g)(v)| = |g(Lv)| \leq \|g\| \cdot \|Lv\| \leq \|g\| \cdot \|L\| \cdot \|v\|,$$

so $L'g \in V^*$, and $\|L'g\| \leq \|g\| \cdot \|L\|$. Thus $L' \in \mathcal{B}(W^*,V^*)$ and $\|L'\| \leq \|L\|$.

Given $v \in V$, apply the Hahn-Banach theorem to Lv to conclude that $\exists\, g \in W^*$ with $\|g\| = 1$ and $(L'g)(v) = g(Lv) = \|Lv\|$. So

$$\|L'\| = \sup_{\|g\|\leq 1} \|L'g\| = \sup_{\|g\|\leq 1} \sup_{\|v\|\leq 1} |(L'g)(v)| \geq \sup_{\|v\|\leq 1} \|Lv\| = \|L\|.$$

Hence $\|L'\| = \|L\|$. □

Transposes and Adjoints

Recall that if $A \in \mathbb{C}^{m\times n}$, we denote by $A^T \in \mathbb{C}^{n\times m}$ the transpose of A, and by $A^* = A^H = \bar{A}^T$ the conjugate-transpose (or Hermitian transpose) of A. If $u, v \in \mathbb{C}^l$ are represented as column vectors, then the Euclidean inner product can be represented in terms of matrix multiplication as $\langle v, u\rangle = v^*u$. For $A \in \mathbb{C}^{m\times n}$, $x \in \mathbb{C}^n$, $y \in \mathbb{C}^m$, we then have $\langle y, Ax\rangle = \langle A^*y, x\rangle$ since $y^*Ax = (A^*y)^*x$.

If $(V, \langle\cdot,\cdot\rangle_V)$ and $(W, \langle\cdot,\cdot\rangle_W)$ are finite dimensional inner product spaces and $L \in \mathcal{L}(V,W)$, then the adjoint operator $L^* \in \mathcal{L}(W,V)$ is defined as follows. For $w \in W$, the assignment $v \to \langle w, Lv\rangle_W$ defines a linear functional on V. We have seen that every linear functional on a finite-dimensional inner product space arises by taking the inner product with a uniquely determined vector in that space. So there exists a unique vector in V which we denote L^*w with the property that

$$\langle L^*w, v\rangle_V = \langle w, Lv\rangle_W$$

for all $v \in V$. The map $w \to L^*w$ defined in this way is a linear transformation $L^* \in \mathcal{L}(W, V)$. However, L^* depends conjugate-linearly on L: $(\alpha L)^* = \bar{\alpha}L^*$.

In the special case in which $V = \mathbb{C}^n$, $W = \mathbb{C}^m$, the inner products are Euclidean, and L is multiplication by $A \in \mathbb{C}^{m\times n}$, it follows that L^* is multiplication by A^*. Compare this with the dual operator L'. Recall that $L' \in \mathcal{L}(W', V')$ is given by right multiplication by A on row vectors, or equivalently by left multiplication by A^T on column vectors. Thus the matrices representing L^* and L' differ by a complex conjugation. In the abstract setting, the operators L^* and L' are related by the conjugate linear isomorphisms $V \cong V'$, $W \cong W'$ induced by the inner products.

Condition Number and Error Sensitivity

In this section we apply our material on norms to the analysis of error sensitivity when solving inhomogeneous systems of linear equations.

Throughout this discussion $A \in \mathbb{C}^{n\times n}$ will be assumed to be invertible. We are interested in determining the sensitivity of the solution of the linear system $Ax = b$ (for a given $b \in \mathbb{C}^n$) to perturbations in the right-hand-side (RHS) vector b or to perturbations in A. One can think of such perturbations as arising from errors in measured data in computational problems, as often occurs when the entries in A and/or b are measured. As we will see, the fundamental quantity is the *condition number*

$$\kappa(A) = \|A\| \cdot \|A^{-1}\|$$

of A, relative to a submultiplicative norm $\|\cdot\|$ on $\mathbb{C}^{n\times n}$. Since $\|I\| \geq 1$ in any submultiplicative norm ($\|I\| = \|I^2\| \leq \|I\|^2 \Rightarrow \|I\| \geq 1$), we have $\kappa(A) = \|A\| \cdot \|A^{-1}\| \geq \|A \cdot A^{-1}\| = \|I\| \geq 1$.

Suppose $\|\cdot\|$ is a norm on $\mathbb{C}^{n\times n}$ consistent with a norm $\|\cdot\|$ on $\mathbb{C}^n$ (i.e. $\|Ax\| \leq \|A\| \cdot \|x\|$ as defined previously). Suppose first that the RHS vector b is subject to error, but the matrix A is not. Then one actually solves the system $A\widehat{x} = \widehat{b}$ for $\widehat{x}$, where $\widehat{b}$ is presumably close to b, instead of the system $Ax = b$ for x. Let x, $\widehat{x}$ be the solutions of $Ax = b$, $A\widehat{x} = \widehat{b}$, respectively. Define the *error vector* $e = x - \widehat{x}$, and the *residual vector* $r = b - \widehat{b} = b - A\widehat{x}$ (the amount by which $A\widehat{x}$ fails to match b). Then $Ae = A(x - \widehat{x}) = b - \widehat{b} = r$, so $e = A^{-1}r$. Thus $\|e\| \leq \|A^{-1}\| \cdot \|r\|$. Since $Ax = b$, $\|b\| \leq \|A\| \cdot \|x\|$. Multiplying these two inequalities gives $\|e\| \cdot \|b\| \leq \|A\| \cdot \|A^{-1}\| \cdot \|x\| \cdot \|r\|$, i.e.

$$\frac{\|e\|}{\|x\|} \leq \kappa(A)\frac{\|r\|}{\|b\|}.$$

So the *relative error* $\frac{\|e\|}{\|x\|}$ is bounded by the condition number $\kappa(A)$ times the *relative residual* $\frac{\|r\|}{\|b\|}$.

Exercise: Show that also $\frac{\|e\|}{\|x\|} \geq \frac{1}{\kappa(A)}\frac{\|r\|}{\|b\|}$.

Matrices for which $\kappa(A)$ is large are called *ill-conditioned* (relative to the norm $\|\cdot\|$); those for which $\kappa(A)$ is close to $\|I\|$ (which is 1 if $\|\cdot\|$ is the operator norm) are called *well-conditioned* (and perfectly conditioned if $\kappa(A) = \|I\|$). If A is ill-conditioned, small relative errors in the data (RHS vector b) *can* result in large relative errors in the solution.

If $\widehat{x}$ is the result of a numerical algorithm (with round-off error) for solving $Ax = b$, then the error $e = x - \widehat{x}$ is not computable, but the residual $r = b - A\widehat{x}$ is computable, so we obtain an upper bound on the relative error $\frac{\|e\|}{\|x\|} \leq \kappa(A)\frac{\|r\|}{\|b\|}$. In practice, we don't know $\kappa(A)$ (although we may be able to estimate it), and this upper bound may be much larger than the actual relative error.

Suppose now that A is subject to error, but b is not. Then $\widehat{x}$ is now the solution of $(A+E)\widehat{x} = b$, where we assume that the error $E \in \mathbb{C}^{n\times n}$ in the matrix is small enough that $\|A^{-1}E\| < 1$, so $(I + A^{-1}E)^{-1}$ exists and can be computed by a Neumann series. Then $A + E$ is invertible and $(A+E)^{-1} = (I + A^{-1}E)^{-1}A^{-1}$. The simplest inequality bounds $\frac{\|e\|}{\|\widehat{x}\|}$, the error relative to $\widehat{x}$, in terms of the relative error $\frac{\|E\|}{\|A\|}$ in A: the equations $Ax = b$ and $(A+E)\widehat{x} = b$ imply $A(x - \widehat{x}) = E\widehat{x}$, $x - \widehat{x} = A^{-1}E\widehat{x}$, and thus $\|x - \widehat{x}\| \leq \|A^{-1}\| \cdot \|E\| \cdot \|\widehat{x}\|$, so that

$$\frac{\|e\|}{\|\widehat{x}\|} \leq \kappa(A)\frac{\|E\|}{\|A\|}.$$

To estimate the error relative to x is more involved; see below.

Consider next the problem of estimating the change in A^{-1} due to a perturbation in A. Suppose $\|A^{-1}\| \cdot \|E\| < 1$. Then as above $A + E$ is invertible, and

$$\begin{aligned} A^{-1} - (A+E)^{-1} &= A^{-1} - \sum_{k=0}^{\infty}(-1)^k (A^{-1}E)^k A^{-1} \\ &= \sum_{k=1}^{\infty}(-1)^{k+1}(A^{-1}E)^k A^{-1}, \end{aligned}$$

so

$$\begin{aligned} \|A^{-1} - (A+E)^{-1}\| &\leq \sum_{k=1}^{\infty} \|A^{-1}E\|^k \cdot \|A^{-1}\| \\ &= \frac{\|A^{-1}E\|}{1 - \|A^{-1}E\|}\|A^{-1}\| \\ &\leq \frac{\|A^{-1}\| \cdot \|E\|}{1 - \|A^{-1}\| \cdot \|E\|}\|A^{-1}\| \\ &= \frac{\kappa(A)}{1 - \kappa(A)\|E\|/\|A\|}\frac{\|E\|}{\|A\|}\|A^{-1}\|. \end{aligned}$$

So the relative error in the inverse satisfies

$$\frac{\|A^{-1} - (A+E)^{-1}\|}{\|A^{-1}\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\|E\|/\|A\|}\frac{\|E\|}{\|A\|}.$$

Note that if $\kappa(A)\frac{\|E\|}{\|A\|} = \|A^{-1}\| \cdot \|E\|$ is small, then $\frac{\kappa(A)}{1-\kappa(A)\|E\|/\|A\|} \approx \kappa(A)$. So the relative error in the inverse is bounded (approximately) by the condition number $\kappa(A)$ of A times the relative error $\frac{\|E\|}{\|A\|}$ in the matrix A.

Using a similar argument, one can derive an estimate for the error relative to x for the computed solution to $Ax = b$ when A and b are simultaneously subject to error. One can show that if $\widehat{x}$ is the solution of $(A+E)\widehat{x} = \widehat{b}$ with both A and b perturbed, then

$$\frac{\|e\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\|E\|/\|A\|}\left(\frac{\|E\|}{\|A\|} + \frac{\|r\|}{\|b\|}\right).$$

(Use $(A+E)x = b + Ex$ and $(A+E)\widehat{x} = \widehat{b}$ to show $x - \widehat{x} = (A+E)^{-1}(Ex + r)$, and also use $\|r\| \leq \frac{\|r\|}{\|b\|}\|A\| \cdot \|x\|$.)

Finite Dimensional Spectral Theory

We begin with a brief review; see Chapter 1 of H-J for more details. Let V be finite dimensional and let $L \in \mathcal{L}(V)$. Unless stated otherwise, $\mathbb{F} = \mathbb{C}$.

Definition. $\lambda \in \mathbb{C}$ is an *eigenvalue* of L if $\exists\, v \in V$, $v \neq 0$ such that $Lv = \lambda v$. The vector v is called an eigenvector associated with the eigenvalue λ.

Thus, if (λ, v) is a eigenvalue-eigenvector pair for L, then $\operatorname{span}\{v\}$ is a one-dimensional invariant subspace under L, and L acts on $\operatorname{span}\{v\}$ by scalar multiplication by λ. Denote by $E_\lambda = \mathcal{N}(\lambda I - L)$ the λ-eigenspace of L. Every nonzero vector in E_λ is an *eigenvector* of L associated with the eigenvalue λ. Define the *geometric multiplicity* of λ to be $m_G(\lambda) = \dim E_\lambda$, i.e., the maximum number of linearly independent eigenvectors associated with λ. The *spectrum* $\sigma(L)$ of L is the set of its eigenvalues, and the *spectral radius* of L is

$$\rho(L) = \max\{|\lambda| : \lambda \in \sigma(L)\}.$$

Clearly $\lambda \in \sigma(L) \Leftrightarrow \lambda I - L$ is singular $\Leftrightarrow \det(\lambda I - L) = 0 \Leftrightarrow p_L(\lambda) = 0$, where $p_L(t) = \det(tI - L)$ is the *characteristic polynomial* of L; p_L is a monic polynomial of degree $n = \dim V$ whose roots are exactly the eigenvalues of L. By the fundamental theorem of algebra, p_L has n roots counting multiplicity; we define the *algebraic multiplicity* $m_A(\lambda)$ of an eigenvalue λ of L to be its multiplicity as a root of p_L.

Facts:

(1) $m_G(\lambda) \leq m_A(\lambda)$ for $\lambda \in \sigma(L)$

(2) Eigenvectors corresponding to different eigenvalues are linearly independent; i.e., if $v_i \in E_{\lambda_i}\backslash\{0\}$ for $1 \leq i \leq k$ and $\lambda_i \neq \lambda_j$ for $i \neq j$, then $\{v_1, \ldots, v_k\}$ is linearly independent. Moreover, if $\{v_1, \ldots, v_k\}$ is a set of eigenvectors with the property that for each $\lambda \in \sigma(L)$, the subset of $\{v_1, \ldots, v_k\}$ corresponding to λ (if nonempty) is linearly independent, then $\{v_1, \ldots, v_k\}$ is linearly independent.

Definition. $L \in \mathcal{L}(V)$ is *diagonalizable* if there is a basis $\mathcal{B} = \{v_1, \ldots, v_n\}$ of V consisting of eigenvectors of L. This definition is clearly equivalent to the alternate definition: L is diagonalizable if there is a basis $\mathcal{B} = \{v_1, \ldots, v_n\}$ of V for which the matrix of L with respect to $\mathcal{B}$ is a diagonal matrix in $\mathbb{C}^{n\times n}$.

Since $\sum_{\lambda\in\sigma(L)} m_A(\lambda) = n = \dim V$ and $m_G(\lambda) \leq m_A(\lambda)$, it follows that $\sum_{\lambda\in\sigma(L)} m_G(\lambda) \leq n$ with equality iff $m_G(\lambda) = m_A(\lambda)$ for all $\lambda \in \sigma(L)$. By Fact 2, $\sum_{\lambda\in\sigma(L)} m_G(\lambda)$ is the maximum number of linearly independent eigenvectors of L. Thus L is diagonalizable $\Leftrightarrow$ $m_G(\lambda) = m_A(\lambda)$ for all $\lambda \in \sigma(L)$. In particular, since $m_G(\lambda) \geq 1$ for all $\lambda \in \sigma(L)$), it follows that if L has n distinct eigenvalues, then L is diagonalizable.

We say that a matrix $A \in \mathbb{C}^{n\times n}$ is diagonalizable iff A is similar to a diagonal matrix, i.e., there exists an invertible $S \in \mathbb{C}^{n\times n}$ for which $S^{-1}AS = D$ is diagonal. Consider the linear transformation $L : \mathbb{C}^n \to \mathbb{C}^n$ given by $L : x \mapsto Ax$. Since matrices similar to A correspond to the matrices of L with respect to different bases, clearly the matrix A is diagonalizable iff L is diagonalizable. Since S is the change of basis matrix and $e_1, \ldots, e_n$ are eigenvectors of

D, it follows that the columns of S are linearly independent eigenvectors of A. This is also clear by computing the matrix equality $AS = SD$ column by column.

We will restrict our attention to $\mathbb{C}^n$ with the Euclidean inner product $\langle \cdot, \cdot \rangle$; here $\| \cdot \|$ will denote the norm induced by $\langle \cdot, \cdot \rangle$ (i.e., the ℓ^2-norm on $\mathbb{C}^n$), and we will denote by $\|A\|$ the operator norm induced on $\mathbb{C}^{n\times n}$ (previously denoted $|||A|||_2$). Virtually all the classes of matrices we are about to define generalize to any Hilbert space V, but we must first know that for $y \in V$ and $A \in \mathcal{B}(V)$, $\exists\, A^*y \in V \ni \langle y, Ax \rangle = \langle A^*y, x \rangle$; we will prove this next quarter. So far, we know that we can define the transpose operator $A' \in \mathcal{B}(V^*)$, so we need to know that we can identify $V^* \cong V$ as we can do in finite dimensions in order to obtain A^*. For now we restrict to $\mathbb{C}^n$.

One can think of many of our operations and sets of matrices in $\mathbb{C}^{n\times n}$ as analogous to corresponding objects in $\mathbb{C}$. For example, the operation $A \mapsto A^*$ is thought of as analogous to conjugation $z \mapsto \bar{z}$ in $\mathbb{C}$. The analogue of a real number is a Hermitian matrix.

Definition. $A \in \mathbb{C}^{n\times n}$ is said to be Hermitian symmetric (or self-adjoint or just Hermitian) if $A = A^*$. $A \in \mathbb{C}^{n\times n}$ is said to be *skew-Hermitian* if $A^* = -A$.

Recall that we have already given a definition of what it means for a sesquilinear form to be Hermitian symmetric. Recall also that there is a $1-1$ correspondence between sesquilinear forms and matrices $A \in \mathbb{C}^{n\times n}$: A corresponds to the form $\langle y, x \rangle_A = \langle y, Ax \rangle$. It is easy to check that A is Hermitian iff the sesquilinear form $\langle \cdot, \cdot \rangle_A$ is Hermitian-symmetric.

Fact: A is Hermitian iff iA is skew-Hermitian (exercise).

The analogue of the imaginary numbers in $\mathbb{C}$ are the skew-Hermitian matrices. Also, any $A \in \mathbb{C}^{n\times n}$ can be written uniquely as $A = B + iC$ where B and C are Hermitian: $B = \frac{1}{2}(A + A^*)$, $C = \frac{1}{2i}(A - A^*)$. Then $A^* = B - iC$. Almost analogous to the $\mathcal{R}e$ and $\mathcal{I}m$ part of a complex number, B is called the *Hermitian part* of A, and iC (not C) is called the *skew-Hermitian part* of A.

Proposition. $A \in \mathbb{C}^{n\times n}$ is Hermitian iff $(\forall\, x \in \mathbb{C}^n)\langle x, Ax \rangle \in \mathbb{R}$.

Proof. If A is Hermitian, $\langle x, Ax \rangle = \frac{1}{2}(\langle x, Ax \rangle + \langle Ax, x \rangle) = \mathcal{R}e\langle x, Ax \rangle \in \mathbb{R}$. Conversely, suppose $(\forall\, x \in \mathbb{C}^n)\langle x, Ax \rangle \in \mathbb{R}$. Write $A = B + iC$ where B, C are Hermitian. Then $\langle x, Bx \rangle \in \mathbb{R}$ and $\langle x, Cx \rangle \in \mathbb{R}$, so $\langle x, Ax \rangle \in \mathbb{R} \Rightarrow \langle x, Cx \rangle = 0$. Since any sesquilinear form $\{y, x\}$ over $\mathbb{C}$ can be recovered from the associated quadratic form $\{x, x\}$ by polarization:

$$4\{y, x\} = \{y + x, y + x\} - \{y - x, y - x\} - i\{y + ix, y + ix\} + i\{y - ix, y - ix\},$$

we conclude that $\langle y, Cx \rangle = 0\ \forall\, x, y \in \mathbb{C}^n$, and thus $C = 0$, so $A = B$ is Hermitian. □

The analogue of the nonnegative reals are the positive semi-definite matrices.

Definition. $A \in \mathbb{C}^{n\times n}$ is called *positive semi-definite* (or nonnegative) if $\langle x, Ax \rangle \geq 0$ for all $x \in \mathbb{C}^n$. By the previous proposition, a positive semi-definite $A \in \mathbb{C}^{n\times n}$ is automatically Hermitian, but one often says Hermitian positive semi-definite anyway.

Caution: If $A \in \mathbb{R}^{n\times n}$ and $(\forall\, x \in \mathbb{R}^n)$ $\langle x, Ax \rangle \geq 0$, A need not be symmetric. For example, if $A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$, then $\langle Ax, x \rangle = \langle x, x \rangle\ \forall\, x \in \mathbb{R}^n$.

If $A \in \mathbb{C}^{n\times n}$ then we can think of A^*A as the analogue of $|z|^2$ for $z \in \mathbb{C}$. Observe that A^*A is positive semi-definite: $\langle A^*Ax, x\rangle = \langle Ax, Ax\rangle = \|Ax\|^2 \geq 0$. It is also the case that $\|A^*A\| = \|A\|^2$. To see this, we first show that $\|A^*\| = \|A\|$. In fact, we previously proved that if $L \in \mathcal{B}(V, W)$ for normed vector spaces V, W, then $\|L'\| = \|L\|$, and we also showed that for $A \in \mathbb{C}^{n\times n}$, A' can be identified with $A^T = \bar{A}^*$. Since it is clear from the definition that $\|\bar{A}\| = \|A\|$, we deduce that $\|A^*\| = \|\bar{A}^*\| = \|A'\| = \|A\|$. Using this, it follows on the one hand that

$$\|A^*A\| \leq \|A^*\| \cdot \|A\| = \|A\|^2$$

and on the other hand that

$$\begin{aligned}
\|A^*A\| &= \sup_{\|x\|=1} \|A^*Ax\| \\
&= \sup_{\|x\|=1} \sup_{\|y\|=1} |\langle y, A^*Ax\rangle| \\
&\geq \sup_{\|x\|=1} \langle x, A^*Ax\rangle \\
&= \sup_{\|x\|=1} \|Ax\|^2 \ = \ \|A\|^2 \ .
\end{aligned}$$

Together these imply $\|A^*A\| = \|A\|^2$.

The analogue of complex numbers of modulus 1 are the unitary matrices.

Definition. $A \in \mathbb{C}^{n\times n}$ is *unitary* if $A^*A = I$.

Since injectivity is equivalent to surjectivity for $A \in \mathbb{C}^{n\times n}$, it follows that $A^* = A^{-1}$ and $AA^* = I$ (each of these is equivalent to $A^*A = I$).

Proposition. For $A \in \mathbb{C}^{n\times n}$, the following conditions are all equivalent:

(1) A is unitary.

(2) The columns of A form an orthonormal basis of $\mathbb{C}^n$.

(3) The rows of A form an orthonormal basis of $\mathbb{C}^n$.

(4) A preserves the Euclidean norm: $(\forall\, x \in \mathbb{C}^n)$ $\|Ax\| = \|x\|$.

(5) A preserves the Euclidean inner product: $(\forall\, x, y \in \mathbb{C}^n)$ $\langle Ay, Ax\rangle = \langle y, x\rangle$.

Proof Sketch. Let $a_1, \ldots, a_n$ be the columns of A. Clearly $A^*A = I \Leftrightarrow a_i^*a_j = \delta_{ij}$. So (1) $\Leftrightarrow$ (2). Similarly (1) $\Leftrightarrow$ $AA^* = I \Leftrightarrow$ (3). Since $\|Ax\|^2 = \langle Ax, Ax\rangle = \langle x, A^*Ax\rangle$, (4) $\Leftrightarrow$ $\langle x, (A^*A - I)x\rangle = 0\,\forall\, x \in \mathbb{C}^n \Leftrightarrow A^*A = I \Leftrightarrow$ (1). Finally, clearly (5) $\Rightarrow$ (4), and (4) $\Rightarrow$ (5) by polarization. □

Normal matrices don't really have an analogue in $\mathbb{C}$.

Definition. $A \in \mathbb{C}^{n\times n}$ is *normal* if $AA^* = A^*A$.

Proposition. For $A \in \mathbb{C}^{n\times n}$, the following conditions are equivalent:

(1) A is normal.

(2) The Hermitian and skew-Hermitian parts of A commute, i.e., if $A = B + iC$ with B, C Hermitian, $BC = CB$.

(3) $(\forall\, x \in \mathbb{C}^n)\ \|Ax\| = \|A^*x\|$.

Proof sketch. (1) $\Leftrightarrow$ (2) is an easy exercise. Since $\|Ax\|^2 = \langle x, A^*Ax\rangle$ and $\|A^*x\|^2 = \langle x, AA^*x\rangle$, we get

$$(3) \quad \Leftrightarrow \quad (\forall\, x \in \mathbb{C}^n)\ \langle x, (A^*A - AA^*)x\rangle = 0 \quad \Leftrightarrow \quad (1).$$

□

Observe that Hermitian, skew-Hermitian, and unitary matrices are all normal.

The above definitions can all be specialized to the real case. Real Hermitian matrices are (real) *symmetric* matrices: $A^T = A$. Every $A \in \mathbb{R}^{n\times n}$ can be written uniquely as $A = B + C$ where $B = B^T$ is symmetric and $C = -C^T$ is skew-symmetric: $B = \frac{1}{2}(A + A^T)$ is called the *symmetric part* of A; $C = \frac{1}{2}(A - A^T)$ is the *skew-symmetric part*. Real unitary matrices are called *orthogonal matrices*, characterized by $A^TA = I$ or $A^T = A^{-1}$. Since $(\forall\, A \in \mathbb{R}^{n\times n})(\forall\, x \in \mathbb{R}^n)\langle x, Ax\rangle \in \mathbb{R}$, there is no characterization of symmetric matrices analogous to that given above for Hermitian matrices. Also unlike the complex case, the values of the quadratic form $\langle x, Ax\rangle$ for $x \in \mathbb{R}^n$ only determine the symmetric part of A, not A itself (the real polarization identity $4\{y, x\} = \{y + x, y + x\} - \{y - x, y - x\}$ is valid only for *symmetric* bilinear forms $\{y, x\}$ over $\mathbb{R}^n$). Consequently, the definition of real positive semi-definite matrices includes symmetry in the definition, together with $(\forall\, x \in \mathbb{R}^n)\ \langle x, Ax\rangle \geq 0$. (This is standard, but not universal. In some mathematical settings, symmetry is not assumed automatically. This is particularly the case in monotone operator theory and optimization theory where it is essential to the theory and the applications that positive definite matrices and operators are **not** assumed to be symmetric.)

The analogy with the complex numbers is particularly clear when considering the eigenvalues of matrices in various classes. For example, consider the characteristic polynomial of a matrix $A \in \mathbb{C}^{n\times n}$, $p_A(t)$. Since $\overline{p_A(t)} = p_{A^*}(\bar{t})$, we have $\lambda \in \sigma(A) \Leftrightarrow \bar{\lambda} \in \sigma(A^*)$. If A is Hermitian, then all eigenvalues of A are real: if x is an eigenvector associated with λ, then

$$\lambda\langle x, x\rangle = \langle x, Ax\rangle = \langle Ax, x\rangle = \bar{\lambda}\langle x, x\rangle,$$

so $\lambda = \bar{\lambda}$. Also eigenvectors corresponding to different eigenvalues are orthogonal: if $Ax = \lambda x$ and $Ay = \mu y$, then

$$\lambda\langle y, x\rangle = \langle y, Ax\rangle = \langle Ay, x\rangle = \mu\langle y, x\rangle,$$

so $\langle y, x\rangle = 0$ if $\lambda \neq \mu$. Any eigenvalue λ of a unitary matrix satisfies $|\lambda| = 1$ since $|\lambda| \cdot \|x\| = \|Ax\| = \|x\|$. Again, eigenvectors corresponding to different eigenvalues of a unitary matrix are orthogonal: if $Ax = \lambda x$ and $Ay = \mu y$, then

$$\lambda\langle y, x\rangle = \langle y, Ax\rangle = \langle A^{-1}y, x\rangle = \langle \mu^{-1}y, x\rangle = \bar{\mu}^{-1}\langle y, x\rangle = \mu\langle y, x\rangle.$$

Matrices which are both Hermitian and unitary, i.e., $A = A^* = A^{-1}$, satisfy $A^2 = I$. The linear transformations determined by such matrices can be thought of as generalizations of reflections: one example is $A = -I$, corresponding to reflections about the origin. Householder transformations are also examples of matrics which are both Hermitian and unitary: by definition, they are of the form

$$I - \frac{2}{\langle y, y\rangle} yy^*$$

where $y \in \mathbb{C}^n \backslash \{0\}$. Such a transformation corresponds to reflection about the hyperplane orthogonal to y. This follows since

$$x \mapsto x - 2\frac{\langle y, x\rangle}{\langle y, y\rangle} y;$$

$\frac{\langle y,x\rangle}{\langle y,y\rangle} y$ is the orthogonal projection onto span$\{y\}$ and $x - \frac{\langle y,x\rangle}{\langle y,y\rangle} y$ is the projection onto $\{y\}^\perp$. (Or simply note that $y \mapsto -y$ and $\langle y, x\rangle = 0 \Rightarrow x \mapsto x$.)

Unitary Equivalence

Similar matrices represent the same linear transformation. There is a special case of similarity which is of particular importance.

Definition. We say that $A, B \in \mathbb{C}^{n\times n}$ are *unitarily equivalent* (or unitarily similar) if there is a unitary matrix $U \in \mathbb{C}^{n\times n}$ such that $B = U^*AU$, i.e., A and B are similar via a unitary similarity transformation (recall: $U^* = U^{-1}$).

Unitary equivalence is important for several reasons. One is that the Hermitian transpose of a matrix is much easier to compute than its inverse, so unitary similarity is computationally advantageous. Another is that, with respect to the operator norm $\|\cdot\|$ on $\mathbb{C}^{n\times n}$ induced by the Euclidean norm on $\mathbb{C}^n$, a unitary matrix U is perfectly conditioned: $(\forall\, x \in \mathbb{C}^n)$ $\|Ux\| = \|x\| = \|U^*x\|$ implies $\|U\| = \|U^*\| = 1$, so $\kappa(U) = \|U\| \cdot \|U^{-1}\| = \|U\| \cdot \|U^*\| = 1$. Moreover, unitary similarity preserves the condition number of a matrix relative to $\|\cdot\|$: $\kappa(U^*AU) = \|U^*AU\| \cdot \|U^*A^{-1}U\| \leq \kappa(A)$ and likewise $\kappa(A) \leq \kappa(U^*AU)$. (In general, for any submultiplicative norm on $\mathbb{C}^{n\times n}$, we obtain the often crude estimate $\kappa(S^{-1}AS) = \|S^{-1}AS\| \cdot \|S^{-1}A^{-1}S\| \leq \|S^{-1}\|^2\|A\| \cdot \|A^{-1}\| \cdot \|S\|^2 = \kappa(S)^2\kappa(A)$, indicating that similarity transformations can drastically change the condition number of A if the transition matrix S is poorly conditioned; note also that $\kappa(A) \leq \kappa(S)^2\kappa(S^{-1}AS)$.) Another basic reason is that unitary similarity preserves the Euclidean operator norm $\|\cdot\|$ and the Frobenius norm $\|\cdot\|_F$ of a matrix.

Proposition. Let $U \in \mathbb{C}^{n\times n}$ be unitary, and $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{n\times k}$. Then

(1) In the operator norms induced by the Euclidean norms, $\|AU\| = \|A\|$ and $\|UB\| = \|B\|$.

(2) In the Frobenius norms, $\|AU\|_F = \|A\|_F$ and $\|UB\|_F = \|B\|_F$.

So multiplication by a unitary matrix on either side preserves $\|\cdot\|$ and $\|\cdot\|_F$.

Proof sketch.

(1) $(\forall x \in \mathbb{C}^k)$ $\|UBx\| = \|Bx\|$, so $\|UB\| = \|B\|$. Likewise, since U^* is also unitary, $\|AU\| = \|(AU)^*\| = \|U^*A^*\| = \|A^*\| = \|A\|$.

(2) Let $b_1, \ldots, b_k$ be the columns of B. Then $\|UB\|_F^2 = \sum_{j=1}^k \|Ub_j\|_2^2 = \sum_{j=1}^k \|b_j\|_2^2 = \|B\|_F^2$. Likewise, since U^* is also unitary, $\|AU\|_F = \|U^*A^*\|_F = \|A^*\|_F = \|A\|_F$.

□

Observe that $\|U\|_F = \sqrt{n}$.

Schur Unitary Triangularization Theorem

Theorem. Any matrix $A \in \mathbb{C}^{n\times n}$ is unitarily equivalent to an upper triangular matrix T. If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A in any prescribed order, then one can choose a unitary similarity transformation so that the diagonal entries of T are $\lambda_1, \ldots, \lambda_n$ in that order.

Proof sketch (see also pp. 79–80 of H-J). By induction on n. Obvious for $n = 1$. Assume true for $n-1$. Given $A \in \mathbb{C}^{n\times n}$ and an ordering $\lambda_1, \ldots, \lambda_n$ of its eigenvalues, choose an eigenvector x for λ_1 with Euclidean norm $\|x\| = 1$. Extend $\{x\}$ to a basis of $\mathbb{C}^n$ and apply the Gram-Schmidt procedure to obtain an orthonormal basis $\{x, u_2, \ldots, u_n\}$ of $\mathbb{C}^n$. Let $U_1 = [xu_2 \cdots u_n] \in \mathbb{C}^{n\times n}$ be the unitary matrix whose columns are $x, u_2, \ldots, u_n$. Since $Ax = \lambda_1 x$, $U_1^* A U_1 = \begin{bmatrix} \lambda_1 & y_1^* \\ 0 & B \end{bmatrix}$ for some $y_1 \in \mathbb{C}^{n-1}$, $B \in \mathbb{C}^{(n-1)\times(n-1)}$. Since similar matrices have the same characteristic polynomial,

$$\begin{aligned} p_A(t) &= \det\left(tI - \begin{bmatrix} \lambda_1 & y_1^* \\ 0 & B \end{bmatrix}\right) \\ &= (t-\lambda_1)\det(tI - B) \\ &= (t-\lambda_1)p_B(t), \end{aligned}$$

so the eigenvalues of B are $\lambda_2, \ldots, \lambda_n$. By the induction hypothesis, $\exists$ a unitary $\widetilde{U} \in \mathbb{C}^{(n-1)\times(n-1)}$ and upper triangular $\widetilde{T} \in \mathbb{C}^{(n-1)\times(n-1)} \ni \widetilde{U}^* B \widetilde{U} = \widetilde{T}$ and the diagonal entries of $\widetilde{T}$ are $\lambda_2, \ldots, \lambda_n$ in that order. Let $U_2 = \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{U} \end{bmatrix} \in \mathbb{C}^{n\times n}$. Then U_2 is unitary, and

$$U_2^* U_1^* A U_1 U_2 = \begin{bmatrix} \lambda_1 & y_1^* \widetilde{U} \\ 0 & \widetilde{U}^* B \widetilde{U} \end{bmatrix} = \begin{bmatrix} \lambda_1 & y_1^* \widetilde{U} \\ 0 & \widetilde{T} \end{bmatrix} \equiv T.$$

Since $U \equiv U_1 U_2$ is unitary and $U^*AU = T$, the statement is true for n as well. □

Note: The basic iterative step that reduces the dimension by 1 is called a deflation. The deflation trick is used to derive a number of important matrix factorizations.

Fact: Unitary equivalence preserves the classes of Hermitian, skew-Hermitian, and normal matrices: e.g., if $A^* = A$, then $(U^*AU)^* = U^*A^*U = U^*AU$ is also Hermitian; if $A^*A = AA^*$, then $(U^*AU)^*(U^*AU) = U^*A^*AU = U^*AA^*U = (U^*AU)(U^*AU)^*$ is normal.

Spectral Theorem. *Let $A \in \mathbb{C}^{n\times n}$ be normal. Then A is unitarily diagonalizable, i.e., A is unitarily similar to a diagonal matrix; so there is an orthonormal basis of eigenvectors.*

Proof sketch. By the Schur Triangularization Theorem, $\exists$ unitary $U \ni U^*AU = T$ is upper triangular. Since A is normal, T is normal: $T^*T = TT^*$. By equating the diagonal entries of T^*T and TT^*, we show T is diagonal. The $(1,1)$ entry of T^*T is $|t_{11}|^2$; that of TT^* is $\sum_{j=1}^n |t_{1j}|^2$. Since $|t_{11}|^2 = \sum_{j=1}^n |t_{1j}|^2$, it must be the case that $t_{1j} = 0$ for $j \geq 2$. Now the $(2,2)$ entry of T^*T is $|t_{22}|^2$; that of TT^* is $\sum_{j=2}^n |t_{2j}|^2$; so again it must be the case that $t_{2j} = 0$ for $j \geq 3$. Continuing with the remaining rows yields the result. □

Cayley-Hamilton Theorem

The Schur Triangularization Theorem gives a quick proof of:

Theorem. (Cayley-Hamilton) *Every matrix $A \in \mathbb{C}^{n\times n}$ satisfies its characteristic polynomial: $p_A(A) = 0$.*

Proof. By Schur, $\exists$ unitary $U \in \mathbb{C}^{n\times n}$ and upper triangular $T \in \mathbb{C}^{n\times n} \ni U^*AU = T$, where the diagonal entries of T are the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A (in some order). Since $A = UTU^*$, $A^k = UT^kU^*$, so $p_A(A) = Up_A(T)U^*$. Writing $p_A(t)$ as

$$p_A(t) = (t - \lambda_1)(t - \lambda_2)\cdots(t - \lambda_n)$$

gives

$$p_A(T) = (T - \lambda_1 I)(T - \lambda_2 I)\cdots(T - \lambda_n I).$$

Since $T - \lambda_j I$ is upper triangular with its jj entry being zero, it follows easily that $p_A(T) = 0$ (accumulate the product from the left, in which case one shows by induction on k that the first k columns of $(T - \lambda_1 I)\cdots(T - \lambda_k I)$ are zero). □

Rayleigh Quotients and the Courant-Fischer Minimax Theorem

There is a very useful variational characterization of the eigenvalues of a Hermitian matrix. We will use this approach later in the course to prove the existence of eigenvalues of compact Hermitian operators in infinite dimensions.

For $A \in \mathbb{C}^{n\times n}$ and $x \in \mathbb{C}^n\backslash\{0\}$, define the *Rayleigh quotient of x* (for A) to be

$$r_A(x) = \frac{\langle x, Ax\rangle}{\langle x, x\rangle}$$

(This is not a standard notation; often $\rho(x)$ is used, but we avoid this notation to prevent possible confusion with the spectral radius $\rho(A)$.) Rayleigh quotients are most useful for Hermitian matrices $A \in \mathbb{C}^{n\times n}$ because $r_A(x) \in \mathbb{R}$ if A is Hermitian. Note that $r_A(x) = \langle x, Ax\rangle$ if $\|x\| = 1$, and in general $r_A(x) = r_A\left(\frac{x}{\|x\|}\right)$, so consideration of $r_A(x)$ for general $x \neq 0$ is equivalent to consideration of $\langle x, Ax\rangle$ for $\|x\| = 1$.

Proposition. Let $A \in \mathbb{C}^{n\times n}$ be Hermitian with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$. For $x \neq 0$, $\lambda_1 \leq r_A(x) \leq \lambda_n$. Moreover, $\min_{x\neq 0} r_A(x) = \lambda_1$ and $\max_{x\neq 0} r_A(x) = \lambda_n$.

The Proposition is a consequence of an explicit formula for $\langle x, Ax\rangle$ in terms of the coordinates relative to an orthonormal basis of eigenvectors for A. By the spectral theorem, there is an orthonormal basis $\{u_1, \ldots, u_n\}$ of $\mathbb{C}^n$ consisting of eigenvectors of A corresponding to $\lambda_1, \ldots, \lambda_n$. Let $U = [u_1 \cdots u_n] \in \mathbb{C}^{n\times n}$, so U is unitary and

$$U^*AU = \Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n).$$

Given $x \in \mathbb{C}^n$, let $y = U^*x$, so $x = Uy = y_1u_1 + \cdots + y_nu_n$. Then

$$\langle x, Ax\rangle = x^*Ax = y^*U^*AUy = y^*\Lambda y = \sum_{i=1}^n \lambda_i|y_i|^2,$$

and of course $\|x\|^2 = \|y\|^2$. It is clear that if $\|y\|^2 = 1$, then $\lambda_1 \le \sum_{i=1}^n \lambda_i|y_i|^2 \le \lambda_n$, with equality for $y = e_1$ and $y = e_n$, corresponding to $x = u_1$ and $x = u_n$. This proves the Proposition.

Analogous reasoning identifies the Euclidean operator norm of a normal matrix.

Proposition. If $A \in \mathbb{C}^{n\times n}$ is normal, then the Euclidean operator norm of A satisfies $\|A\| = \rho(A)$, the spectral radius of A.

Proof. If U,λ_i and y_i are as above, then

$$\|Ax\|^2 = \|U^*Ax\|^2 = \|U^*AUy\|^2 = \|\Lambda y\|^2 = \sum_{i=1}^n |\lambda_i|^2|y_i|^2.$$

Just as above, it follows that

$$\sup_{\|x\|=1} \|Ax\|^2 = \sup_{\|y\|=1} \sum_{i=1}^n |\lambda_i|^2|y_i|^2 = \max|\lambda_i|^2 = \rho(A)^2.$$

□

Caution. It is not true that $\|A\| = \rho(A)$ for all $A \in \mathbb{C}^{n\times n}$. For example, take $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$.

There follows an identification of the Euclidean operator norm of any, possibly nonsquare, matrix.

Corollary. If $A \in \mathbb{C}^{m\times n}$, then $\|A\| = \sqrt{\rho(A^*A)}$, where $\|A\|$ is the operator norm induced by the Euclidean norms on $\mathbb{C}^m$ and $\mathbb{C}^n$.

Proof. Note that $A^*A \in \mathbb{C}^{n\times n}$ is positive semidefinite Hermitian, so the first proposition of this section shows that $\max_{\|x\|=1}\langle x, A^*Ax\rangle = \max_{\|x\|=1} r_{A^*A}(x) = \rho(A^*A)$. But $\|Ax\|^2 = \langle x, A^*Ax\rangle$, giving

$$\|A\|^2 = \max_{\|x\|=1} \|Ax\|^2 = \max_{\|x\|=1} \langle x, A^*Ax\rangle = \rho(A^*A).$$

□

The first proposition of this section gives a variational characterization of the largest and smallest eigenvalues of an Hermitian matrix. The next Theorem extends this to provide a variational characterization of all the eigenvalues.

Courant-Fischer Minimax Theorem. *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian. In what follows, S_k will denote an arbitrary subspace of $\mathbb{C}^n$ of dimension k, and* $\min_{S_k}$ *and* $\max_{S_k}$ *denote taking the min or max over all subspaces of $\mathbb{C}^n$ of dimension k.*

(1) For $1 \leq k \leq n$, $\min_{S_k} \max_{x\neq 0, x\in S_k} r_A(x) = \lambda_k$ *(minimax)*

(2) For $1 \leq k \leq n$, $\max_{S_{n-k+1}} \min_{x\neq 0, x\in S_{n-k+1}} r_A(x) = \lambda_k$ *(maximin)*

where $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of A.

Proof. Let $u_1, \ldots, u_n$ be orthonormal eigenvectors of A corresponding to $\lambda_1, \ldots, \lambda_n$. Let $U = [u_1 \cdots u_n] \in \mathbb{C}^{n\times n}$, so $U^*AU = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, and for $x \in \mathbb{C}^n$, let $y = U^*x$, so $x = Uy = y_1u_1 + \cdots + y_nu_n$. To prove (1), let $W = \operatorname{span}\{u_k, \ldots, u_n\}$, so $\dim W = n-k+1$. If $\dim S_k = k$, then by dimension arguments $\exists\, x \in S_k \cap W\backslash\{0\}$, which we can assume to satisfy $\|x\| = 1$. Then $r_A(x) = \sum_{i=k}^n \lambda_i |y_i|^2 \geq \lambda_k$. Thus $\forall\, S_k$, $\max_{\|x\|=1, x\in S_k} r_A(x) \geq \lambda_k$. But for $x \in \operatorname{span}\{u_1, \ldots, u_k\}\backslash\{0\}$, $r_A(x) = \sum_{i=1}^k \lambda_i|y_i|^2 \leq \lambda_k$. So for $S_k = \operatorname{span}\{u_1, \ldots, u_k\}$, the $\max = \lambda_k$. Thus $\min\max = \lambda_k$.

The proof of (2) is similar. Let $W = \operatorname{span}\{u_1, \ldots, u_k\}$, so $\dim W = k$. If $\dim S_{n-k+1} = n-k+1$, then $\exists\, x \in S_{n-k+1} \cap W\backslash\{0\}$ with $\|x\| = 1$, and $r_A(x) \leq \lambda_k$. Thus $\forall\, S_{n-k+1}$, $\min_{\|x\|=1, x\in S_{n-k+1}} r_A(x) \leq \lambda_k$. But if we take $S_{n-k+1} = \operatorname{span}\{u_k, \ldots, u_n\}$, the min is λ_k, so $\max\min = \lambda_k$. □

Remark. (1) for $k = 1$ and (2) for $k = n$ give the previous result for the largest and smallest eigenvalues.

Non-Unitary Similarity Transformations

Despite the advantages of unitary equivalence, there are limitations. Not every diagonalizable matrix is unitarily diagonalizable. For example, consider an upper-triangular matrix T with distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. We know that T is diagonalizable. However, T is not unitarily similar to a diagonal matrix unless it is already diagonal. This is because unitary equivalence preserves the Frobenius norm: $\|T\|_F^2 = \sum_{i=1}^n |\lambda_i|^2 + \sum_{i<j} |t_{ij}|^2$, but any diagonal matrix similar to T has Frobenius norm $\sum_{i=1}^n |\lambda_i|^2$. In order to diagonalize T it is necessary to use non-unitary similarity transformations.

Proposition. Let $A \in \mathbb{C}^{n\times n}$ and let $\lambda_1, \ldots, \lambda_k$ be the *distinct* eigenvalues of A, with multiplicities $m_1, \ldots, m_k$, respectively (so $m_1 + \cdots + m_k = n$). Then A is similar to a block diagonal matrix of the form

$$\begin{bmatrix} T_1 & & & 0 \\ & T_2 & & \\ & & \ddots & \\ 0 & & & T_k \end{bmatrix},$$

where each $T_i \in \mathbb{C}^{m_i\times m_i}$ is upper triangular with λ_i as each of its diagonal entries.

Proof. By Schur, A is similar to an upper triangular T with diagonal entries ordered $\overbrace{\lambda_1, \ldots, \lambda_1}^{m_1}, \overbrace{\lambda_2, \ldots, \lambda_2}^{m_2}, \cdots, \overbrace{\lambda_k, \ldots, \lambda_k}^{m_k}$. We use a strategy as in Gaussian elimination (but must be sure to do similarity transformations). Let $E_{rs} \in \mathbb{C}^{n\times n}$ have 1 in the (r,s)-entry and 0 elsewhere. Left multiplication of T by E_{rs} moves the sth row of T to the rth row and zeros out all other elements, that is, the elements of the matrix $E_{rs}T$ are all zero except for those in the rth row which is just the sth row of T. Therefore, left multiplication of T by the matrix $(I - \alpha E_{rs})$ corresponds to subtracting α times the sth row of T from the rth row of T. This is one of the elementary row operations used in Gaussian elimination. Note that if $r < s$, then this operation introduces no new non-zero entries below the main diagonal of T, that is, $E_{rs}T$ is still upper triangular.

Similarly, right multiplication of T by E_{rs} moves the rth column of T to the sth column and zeros out all other entries in the matrix, that is, the elements of the matrix TE_{rs} are all zero except for those in the sth column which is just the rth column of T. Therefore, right multiplication of T by $(I + \alpha E_{rs})$ corresponds to adding α times the rth column of T to the sth column of T. If $r < s$, then this operation introduces no new non-zero entries below the main diagonal of T, that is, TE_{rs} is still upper triangular.

Because of the properties described above, the matrices $(I \pm \alpha E_{rs})$ are sometimes referred to as *Gaussian elimination matrices*. Note that $E_{rs}^2 = 0$ whenever $r \neq s$, and so

$$(I - \alpha E_{rs})(I + \alpha E_{rs}) = I - \alpha E_{rs} + \alpha E_{rs} - \alpha^2 E_{rs}^2 = I.$$

Thus $(I + \alpha E_{rs})^{-1} = (I - \alpha E_{rs})$.

Now consider the similarity transformation

$$T \mapsto (I + \alpha E_{rs})^{-1} T (I + \alpha E_{rs}) = (I - \alpha E_{rs}) T (I + \alpha E_{rs})$$

with $\alpha \in \mathbb{C}$ and $r < s$. Since T is upper triangular (as are $I \pm \alpha E_{rs}$ for $r < s$), it follows that this similarity transformation only changes T in the s^{th} column above (and including) the r^{th} row, and in the r^{th} row to the right of (and including) the s^{th} column, and that t_{rs} gets replaced by $t_{rs} + \alpha(t_{rr} - t_{ss})$. So if $t_{rr} \neq t_{ss}$ it is possible to choose α to make $t_{rs} = 0$. Using these observations, it is easy to see that such transformations can be performed successively without destroying previously created zeroes to zero out all entries except those in the desired block diagonal form (work backwards row by row starting with row $n - m_k$; in each row, zero out entries from left to right). □

Jordan Form

Let $T \in \mathbb{C}^{n \times n}$ be an upper triangular matrix in block diagonal form

$$T = \begin{bmatrix} T_1 & & 0 \\ & \ddots & \\ 0 & & T_k \end{bmatrix}$$

as in the previous Proposition, i.e., $T_i \in \mathbb{C}^{m_i \times m_i}$ satisfies $T_i = \lambda_i I + N_i$ where $N_i \in \mathbb{C}^{m_i \times m_i}$ is strictly upper triangular, and $\lambda_1, \ldots, \lambda_k$ are distinct. Then for $1 \leq i \leq k$, $N_i^{m_i} = 0$, so N is nilpotent. Recall that any nilpotent operator is a direct sum of shift operators in an appropriate basis, so the matrix N_i is similar to a direct sum of shift matrices

$$S_l = \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & 1 \\ 0 & & & 0 \end{bmatrix} \in \mathbb{C}^{l \times l}$$

of varying sizes l. Thus each T_i is similar to a direct sum of *Jordan blocks*

$$J_l(\lambda) = \lambda I_l + S_l = \begin{bmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & 1 \\ 0 & & & \lambda \end{bmatrix} \in \mathbb{C}^{l\times l}$$

of varying sizes l (with $\lambda = \lambda_i$).

Definition. A matrix J is in *Jordan normal form* if it is the direct sum of finitely many Jordan blocks (with, of course, possibly different values of λ and l).

The previous Proposition, together with our results on the structure of nilpotent operators as discussed above, establishes the following Theorem.

Theorem. *Every matrix $A \in \mathbb{C}^{n\times n}$ is similar to a matrix in Jordan normal form.*

Remarks:

(1) The Jordan form of A is not quite unique since the blocks may be arbitrarily reordered by a similarity transformation. As we will see, this is the only nonuniqueness: the number of blocks of each size for each eigenvalue λ is determined by A.

(2) Pick a Jordan matrix similar to A. For $\lambda \in \sigma(A)$ and $j \geq 1$, let $b_j(\lambda)$ denote the number of $j \times j$ blocks associated with λ, and let $r(\lambda) = \max\{j : b_j(\lambda) > 0\}$ be the size of the largest block associated with λ. Let $k_j(\lambda) = \dim(\mathcal{N}(A - \lambda I)^j)$. Then from our remarks on nilpotent operators,

$$0 < k_1(\lambda) < k_2(\lambda) < \cdots < k_{r(\lambda)}(\lambda) = k_{r(\lambda)+1}(\lambda) = \cdots = m(\lambda),$$

where $m(\lambda)$ is the algebraic multiplicity of λ. By considering the form of powers of shift matrices, one can easily show that

$$b_j(\lambda) + b_{j+1}(\lambda) + \cdots + b_{r(\lambda)}(\lambda) = k_j(\lambda) - k_{j-1}(\lambda),$$

i.e., the number of blocks of size $\geq j$ associated with λ is $k_j(\lambda) - k_{j-1}(\lambda)$. (In particular, for $j = 1$, the number of Jordan blocks associated with λ is $k_1(\lambda)$ = the geometric multiplicity of λ.) Thus,

$$b_j(\lambda) = -k_{j+1}(\lambda) + 2k_j(\lambda) - k_{j-1}(\lambda),$$

which is completely determined by A. (Compare with problem (6) on HW # 2.) Since $k_j(\lambda)$ is invariant under similarity transformations, we conclude:

Proposition. (a) The Jordan form of A is unique up to reordering of the Jordan blocks. (b) Two matrices in Jordan form are similar iff they can be obtained from each other by reordering the blocks.

Remarks continued:

(3) Knowing the algebraic and geometric multiplicities of each eigenvalue of A is not sufficient to determine the Jordan form (unless the algebraic multiplicity of each eigenvalue is at most one greater than its geometric multiplicity).

Exercise: Why is it determined in this case?

For example,

$$N_1 = \begin{bmatrix} 0 & 1 & & \\ & 0 & 1 & \\ & & 0 & \\ & & & 0 \end{bmatrix} \quad \text{and} \quad N_2 = \begin{bmatrix} 0 & 1 & & \\ 0 & 0 & & \\ & & 0 & 1 \\ & & 0 & 0 \end{bmatrix}$$

are not similar as $N_1^2 \neq 0 = N_2^2$, but both have 0 as the only eigenvalue with algebraic multiplicity 4 and geometric multiplicity 2.

(4) The expression for $b_j(\lambda)$ in remark (2) above can also be given in terms of $r_j(\lambda) \equiv \operatorname{rank}((A-\lambda I)^j) = \dim(\mathcal{R}(A-\lambda I)^j) = m(\lambda) - k_j(\lambda)$: $b_j = r_{j+1} - 2r_j + r_{j-1}$.

(5) A necessary and sufficient condition for two matrices in $\mathbb{C}^{n\times n}$ to be similar is that they are both similar to the same Jordan normal form matrix.

Spectral Decomposition

There is a useful invariant (i.e., basis-free) formulation of some of the above. Let $L \in \mathcal{L}(V)$ where $\dim V = n < \infty$ (and $\mathbb{F} = \mathbb{C}$). Let $\lambda_1, \ldots, \lambda_k$ be the distinct eigenvalues of L, with algebraic multiplicities $m_1, \ldots, m_k$. Define the *generalized eigenspaces* $\widetilde{E}_i$ of L to be $\widetilde{E}_i = \mathcal{N}((L-\lambda_i I)^{m_i})$. (The eigenspaces are $E_{\lambda_i} = \mathcal{N}(L-\lambda_i I)$. Vectors in $\widetilde{E}_i \backslash E_{\lambda_i}$ are sometimes called *generalized eigenvectors.*) Upon choosing a basis for V in which L is represented as a block-diagonal upper triangular matrix as above, one sees that

$$\dim \widetilde{E}_i = m_i \ (1 \leq i \leq k) \quad \text{and} \quad V = \bigoplus_{i=1}^{k} \widetilde{E}_i.$$

Let P_i $(1 \leq i \leq k)$ be the projections associated with this decomposition of V, and define $D = \sum_{i=1}^k \lambda_i P_i$. Then D is a diagonalizable transformation since it is diagonal in the basis in which L is block diagonal upper triangular. Using the same basis, one sees that the matrix of $N \equiv L - D$ is strictly upper triangular, and thus N is nilpotent (in fact $N^m = 0$ where $m = \max m_i$); moreover, $N = \sum_{i=1}^k N_i$ where $N_i = P_i N P_i$. Also $L\widetilde{E}_i \subset \widetilde{E}_i$, and $LD = DL$ since D is a multiple of the identity on each of the L-invariant subspaces $\widetilde{E}_i$, and thus also $ND = DN$. We have proved:

Theorem. *Any $L \in \mathcal{L}(V)$ can be written as $L = D + N$ where D is diagonalizable, N is nilpotent, and $DN = ND$. If P_i is the projection onto the λ_i-generalized eigenspace and $N_i = P_i N P_i$, then $D = \sum_{i=1}^k \lambda_i P_i$ and $N = \sum_{i=1}^k N_i$. Moreover,*

$$LP_i = P_i L = P_i L P_i = \lambda_i P_i + N_i \ (1 \leq i \leq k),$$

$$P_iP_j = \delta_{ij}P_i \text{ and } P_iN_j = N_jP_i = \delta_{ij}N_j \ (1 \le i \le k)(1 \le j \le k),$$

and

$$N_iN_j = N_jN_i = 0 \quad (1 \le i < j \le k),$$

where $\delta_{ij} = 0$ *if* $i \neq j$ *and* $\delta_{ij} = 1$ *if* $i = j$.

Note: D and N are uniquely determined by L, but we will not prove this here.

If V has an inner product $\langle \cdot, \cdot \rangle$, and L is normal, then we know that L is diagonalizable, so $N = 0$. In this case we know that eigenvectors corresponding to different eigenvalues are orthogonal, so the subspaces $\widetilde{E}_i$ ($= E_{\lambda_i}$ in this situation) are mutually orthogonal in V. The associated projections P_i are orthogonal projections (since $E_{\lambda_i}^{\perp} = E_{\lambda_1} \oplus \cdots \oplus E_{\lambda_{i-1}} \oplus E_{\lambda_{i+1}} \oplus \cdots \oplus E_{\lambda_k}$).

There is a useful characterization of when a projection is orthogonal. Recall that any $P \in \mathcal{L}(V)$ satisfying $P^2 = P$ is a projection: one has $V = \mathcal{R}(P) \oplus \mathcal{N}(P)$, and P is the projection of V onto $\mathcal{R}(P)$ along $\mathcal{N}(P)$. Recall also that P is called an *orthogonal projection* relative to an inner product on V if $\mathcal{R}(P) \perp \mathcal{N}(P)$.

Proposition. A projection P on an inner product space is orthogonal iff it is self-adjoint (i.e., P is Hermitian: $P^* = P$, where P^* is the adjoint of P with respect to the inner product).

Proof. Let $P \in \mathcal{L}(V)$ be a projection. If $P^* = P$, then $\langle y, Px \rangle = \langle Py, x \rangle \ \forall\, x, y \in V$. So $y \in \mathcal{N}(P)$ iff $(\forall\, x \in V)\ \langle y, Px \rangle = \langle Py, x \rangle = 0$ iff $y \in \mathcal{R}(P)^{\perp}$, so P is an orthogonal projection. Conversely, suppose $\mathcal{R}(P) \perp \mathcal{N}(P)$. We must show that $\langle y, Px \rangle = \langle Py, x \rangle$ for all $x, y \in V$. Since $V = \mathcal{R}(P) \oplus \mathcal{N}(P)$, it suffices to check this separately in the four cases $x, y \in \mathcal{R}(P)$, $\mathcal{N}(P)$. Each of these cases is straightforward since $Pv = v$ for $v \in \mathcal{R}(P)$ and $Pv = 0$ for $v \in \mathcal{N}(P)$. □

Jordan form depends discontinuously on A

Ignoring the reordering question, the Jordan form of A is discontinuous at every matrix A except those with distinct eigenvalues. For example, when $\epsilon = 0$, the Jordan form of $\begin{pmatrix} \epsilon & 1 \\ 0 & 0 \end{pmatrix}$ is $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, but for $\epsilon \neq 0$, the Jordan form is $\begin{pmatrix} \epsilon & 0 \\ 0 & 0 \end{pmatrix}$. So small perturbations in A can significantly change the Jordan form. For this reason, the Jordan form is almost never used for numerical computation.

Jordan Form over $\mathbb{R}$

The previous results do not hold for real matrices: for example, in general a real matrix is not similar to a real upper-triangular matrix via a real similarity transformation. If it were, then its eigenvalues would be the real diagonal entries, but a real matrix need not have only real eigenvalues. However, nonreal eigenvalues are the only obstruction to carrying out our previous arguments. Precisely, if $A \in \mathbb{R}^{n \times n}$ has real eigenvalues, then A is orthogonally similar to a real upper triangular matrix, and A can be put into block diagonal and Jordan form using real similarity transformations, by following the same arguments as before. If A

does have some nonreal eigenvalues, then there are substitute normal forms which can be obtained via real similarity transformations.

Recall that nonreal eigenvalues of a real matrix $A \in \mathbb{R}^{n\times n}$ came in complex conjugate pairs: if $\lambda = a + ib$ (with $a, b \in \mathbb{R}$, $b \neq 0$) is an eigenvalue of A, then since $p_A(t)$ has real coefficients, $0 = \overline{p_A(\lambda)} = p_A(\bar{\lambda})$, so $\bar{\lambda} = a - ib$ is also an eigenvalue. If $u + iv$ (with $u, v \in \mathbb{R}^n$) is an eigenvector of A for λ, then

$$A(u - iv) = A\overline{(u + iv)} = \overline{A(u + iv)} = \overline{\lambda(u + iv)} = \bar{\lambda}(u - iv),$$

so $u - iv$ is an eigenvector of A for $\bar{\lambda}$. It follows that $u + iv$ and $u - iv$ (being eigenvectors for different eigenvalues) are linearly independent over $\mathbb{C}$, and thus $u = \frac{1}{2}(u + iv) + \frac{1}{2}(u - iv)$ and $v = \frac{1}{2i}(u + iv) - \frac{1}{2i}(u - iv)$ are linearly independent over $\mathbb{C}$, and consequently also over $\mathbb{R}$. Since $A(u + iv) = (a + ib)(u + iv) = (au - bv) + i(bu + av)$, it follows that $Au = au - bv$ and $Av = bu + av$. Thus $\operatorname{span}\{u, v\}$ is a 2-dimensional real invariant subspace of $\mathbb{R}^n$ for A, and the matrix of A restricted to the subspace $\operatorname{span}\{u, v\}$ with respect to the basis $\{u, v\}$ is $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$ (observe that this 2×2 matrix has eigenvalues $\lambda, \bar{\lambda}$).

Over $\mathbb{R}$, the best one can generally do is to have such 2×2 diagonal blocks instead of upper triangular matrices with $\lambda, \bar{\lambda}$ on the diagonal. For example, the real Jordan blocks for $\lambda, \bar{\lambda}$ are

$$J_l(\lambda, \bar{\lambda}) = \begin{bmatrix} \begin{bmatrix} a & b \\ -b & a \end{bmatrix} & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} & & 0 \\ & \ddots & \ddots & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ 0 & & & \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \end{bmatrix} \in \mathbb{R}^{2l\times 2l}.$$

The real Jordan form of $A \in \mathbb{R}^{n\times n}$ is a direct sum of such blocks, with the usual Jordan blocks for the real eigenvalues. See H-J for details.

Non-Square Matrices

There is a useful variation on the concept of eigenvalues and eigenvectors which is defined for both square and non-square matrices. Throughout this discussion, for $A \in \mathbb{C}^{m\times n}$, let $\|A\|$ denote the operator norm induced by the Euclidean norms on $\mathbb{C}^n$ and $\mathbb{C}^m$ (which we denote by $\|\cdot\|$), and let $\|A\|_F$ denote the Frobenius norm of A. Note that we still have

$$\langle y, Ax\rangle_{\mathbb{C}^m} = y^*Ax = \langle A^*y, x\rangle_{\mathbb{C}^n} \quad \text{for} \quad x \in \mathbb{C}^n,\ y \in \mathbb{C}^m.$$

From $A \in \mathbb{C}^{m\times n}$ one can construct the square matrices $A^*A \in \mathbb{C}^{n\times n}$ and $AA^* \in \mathbb{C}^{m\times m}$. Both of these are Hermitian positive semi-definite. In particular A^*A and AA^* are diagonalizable with real non-negative eigenvalues. Except for the multiplicities of the zero eigenvalue, these matrices have the same eigenvalues; in fact, we have:

Proposition. Let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{n\times m}$ with $m \le n$. Then the eigenvalues of BA (counting multiplicity) are the eigenvalues of AB, together with $n-m$ zeroes. (Remark: For $n = m$, this was Problem 4 on Problem Set 5.)

Proof. Consider the $(n+m)\times(n+m)$ matrices

$$C_1 = \begin{bmatrix} AB & 0 \\ B & 0 \end{bmatrix} \quad \text{and} \quad C_2 = \begin{bmatrix} 0 & 0 \\ B & BA \end{bmatrix}.$$

These are similar since $S^{-1}C_1S = C_2$ where

$$S = \begin{bmatrix} I & A \\ 0 & I \end{bmatrix} \quad \text{and} \quad S^{-1} = \begin{bmatrix} I & -A \\ 0 & I \end{bmatrix}.$$

But the eigenvalues of C_1 are those of AB along with n zeroes, and the eigenvalues of C_2 are those of BA along with m zeroes. The result follows. □

So for any m, n, the eigenvalues of A^*A and AA^* differ by $|n-m|$ zeroes. Let $p = \min(m,n)$ and let $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p$ (≥ 0) be the joint eigenvalues of A^*A and AA^*.

Definition. The *singular values* of A are the numbers

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_p \ge 0,$$

where $\sigma_i = \sqrt{\lambda_i}$. (When $n > m$, one often also defines singular values $\sigma_{m+1} = \cdots = \sigma_n = 0$.)

It is a fundamental result that one can choose orthonormal bases for $\mathbb{C}^n$ and $\mathbb{C}^m$ so that A maps one basis into the other, scaled by the singular values. Let $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{C}^{m\times n}$ be the "diagonal" matrix whose ii entry is σ_i $(1 \le i \le p)$.

Singular Value Decomposition (SVD)

If $A \in \mathbb{C}^{m\times n}$, then there exist unitary matrices $U \in \mathbb{C}^{m\times m}$, $V \in \mathbb{C}^{n\times n}$ such that $A = U\Sigma V^*$, where $\Sigma \in \mathbb{C}^{m\times n}$ is the diagonal matrix of singular values.

Proof. By the same argument as in the square case, $\|A\|^2 = \|A^*A\|$. But

$$\|A^*A\| = \lambda_1 = \sigma_1^2, \quad \text{so} \quad \|A\| = \sigma_1.$$

So we can choose $x \in \mathbb{C}^n$ with $\|x\| = 1$ and $\|Ax\| = \sigma_1$. Write $Ax = \sigma_1 y$ where $\|y\| = 1$. Complete x and y to unitary matrices

$$V_1 = [x, \widetilde{v}_2, \cdots, \widetilde{v}_n] \in \mathbb{C}^{n\times n} \quad \text{and} \quad U_1 = [y, \widetilde{u}_2, \cdots, \widetilde{u}_m] \in \mathbb{C}^{m\times m}.$$

Since $U_1^*AV_1 \equiv A_1$ is the matrix of A in these bases it follows that

$$A_1 = \begin{bmatrix} \sigma_1 & w^* \\ 0 & B \end{bmatrix}$$

for some $w \in \mathbb{C}^{n-1}$ and $B \in \mathbb{C}^{(m-1)\times(n-1)}$. Now observe that

$$\begin{aligned} \sigma_1^2 + w^*w &\leq \left\| \begin{bmatrix} \sigma_1^2 + w^*w \\ Bw \end{bmatrix} \right\| \\ &= \left\| A_1 \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\| \\ &\leq \|A_1\| \cdot \left\| \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\| \\ &= \sigma_1(\sigma_1^2 + w^*w)^{\frac{1}{2}} \end{aligned}$$

since $\|A_1\| = \|A\| = \sigma_1$ by the invariance of $\|\cdot\|$ under unitary multiplication.

It follows that $(\sigma_1^2 + w^*w)^{\frac{1}{2}} \leq \sigma_1$, so $w = 0$, and thus

$$A_1 = \begin{bmatrix} \sigma_1 & 0 \\ 0 & B \end{bmatrix}.$$

Now apply the same argument to B and repeat to get the result. For this, observe that

$$\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & B^*B \end{bmatrix} = A_1^*A_1 = V_1^*A^*AV_1$$

is unitarily similar to A^*A, so the eigenvalues of B^*B are $\lambda_2 \geq \cdots \geq \lambda_n$ (≥ 0). Observe also that the same argument shows that if $A \in \mathbb{R}^{m\times n}$, then U and V can be taken to be real orthogonal matrices. □

This proof given above is direct, but it masks some of the key ideas. We now sketch an alternative proof that reveals more of the underlying structure of the SVD decomposition.

Alternative Proof of SVD: Let $\{v_1, \ldots, v_n\}$ be an orthonormal basis of $\mathbb{C}^n$ consisting of eigenvectors of A^*A associated with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ (≥ 0), respectively, and let $V = [v_1 \cdots v_n] \in \mathbb{C}^{n\times n}$. Then V is unitary, and

$$V^*A^*AV = \Lambda \equiv \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}.$$

For $1 \leq i \leq n$,

$$\|Av_i\|^2 = e_i^* V^* A^* A V e_i = \lambda_i = \sigma_i^2 \ .$$

Choose the integer r such that

$$\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$$

(r turns out to be the rank of A). Then for $1 \leq i \leq r$, $Av_i = \sigma_i u_i$ for a unique $u_i \in \mathbb{C}^m$ with $\|u_i\| = 1$. Moreover, for $1 \leq i, j \leq r$,

$$u_i^* u_j = \frac{1}{\sigma_i \sigma_j} v_i^* A^* A v_j = \frac{1}{\sigma_i \sigma_j} e_i^* \Lambda e_j = \delta_{ij}.$$

So we can append vectors $u_{r+1}, \ldots, u_m \in \mathbb{C}^m$ (if necessary) so that $U = [u_1 \cdots u_m] \in \mathbb{C}^{m \times m}$ is unitary. It follows easily that $AV = U\Sigma$, so $A = U\Sigma V^*$. □

The ideas in this second proof are derivable from the equality $A = U\Sigma V^*$ expressing the SVD of A (no matter how it is constructed). The SVD equality is equivalent to $AV = U\Sigma$. Interpreting this equation columnwise gives

$$Av_i = \sigma_i u_i \quad (1 \leq i \leq p),$$

and

$$Av_i = 0 \quad \text{for } i > m \text{ if } n > m,$$

where $\{v_1, \ldots, v_n\}$ are the columns of V and $\{u_1, \ldots, u_m\}$ are the columns of U. So A maps the orthonormal vectors $\{v_1, \ldots, v_p\}$ into the orthogonal directions $\{u_1, \ldots, u_p\}$ with the singular values $\sigma_1 \geq \cdots \geq \sigma_p$ as scale factors. (Of course if $\sigma_i = 0$ for an $i \leq p$, then $Av_i = 0$, and the direction of u_i is not represented in the range of A.)

The vectors $v_1, \ldots, v_n$ are called the *right singular vectors* of A, and $u_1, \ldots, u_m$ are called the *left singular vectors* of A. Observe that

$$A^*A = V\Sigma^*\Sigma V^* \quad \text{and} \quad \Sigma^*\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2) \in \mathbb{R}^{n \times n}$$

even if $m < n$. So

$$V^*A^*AV = \Lambda = \operatorname{diag}(\lambda_1, \ldots \lambda_n),$$

and thus the columns of V form an orthonormal basis consisting of eigenvectors of $A^*A \in \mathbb{C}^{n \times n}$. Similarly $AA^* = U\Sigma\Sigma^*U^*$, so

$$U^*AA^*U = \Sigma\Sigma^* = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_p^2, \overbrace{0, \ldots, 0}^{m-n \text{ zeroes if } m>n}) \in \mathbb{R}^{m \times m},$$

and thus the columns of U form an orthonormal basis of $\mathbb{C}^m$ consisting of eigenvectors of $AA^* \in \mathbb{C}^{m \times m}$.

Caution. We cannot choose the bases of eigenvectors $\{v_1, \ldots, v_n\}$ of A^*A (corresponding to $\lambda_1, \ldots, \lambda_n$) and $\{u_1, \ldots, u_m\}$ of AA^* (corresponding to $\lambda_1, \ldots, \lambda_p, [0, \ldots, 0]$) independently: we must have $Av_i = \sigma_i u_i$ for $\sigma_i > 0$.

In general, the SVD is not unique. Σ is uniquely determined but if A^*A has multiple eigenvalues, then one has freedom in the choice of bases in the corresponding eigenspace, so V (and thus U) is not uniquely determined. One has complete freedom of choice of orthonormal bases of $\mathcal{N}(A^*A)$ and $\mathcal{N}(AA^*)$: these form the right-most columns of V and U, respectively. For a nonzero multiple singular value, one can choose the basis of the eigenspace of A^*A (choosing columns of V), but then the corresponding columns of U are determined; or, one can choose the basis of the eigenspace of AA^* (choosing columns of U), but then the corresponding columns of V are determined. If all the singular values $\sigma_1, \ldots, \sigma_n$ of A are distinct, then each column of V is uniquely determined up to a factor of modulus 1, i.e., V is determined up to right multiplication by a diagonal matrix

$$D = \operatorname{diag}(e^{i\theta_1}, \ldots, e^{i\theta_n}).$$

Such a change in V must be compensated for by multiplying the first n columns of U by D (the first $n-1$ cols. of U by $\operatorname{diag}(e^{i\theta_1}, \ldots, e^{i\theta_{n-1}})$ if $\sigma_n = 0$); of course if $m > n$, then the last $m-n$ columns of U have further freedom (they are in $\mathcal{N}(AA^*)$).

There is an abbreviated form of SVD useful in computation. Since rank is preserved under unitary multiplication, $\operatorname{rank}(A) = r$ iff $\sigma_1 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1} = \cdots$. Let $U_r \in \mathbb{C}^{m\times r}$, $V_r \in \mathbb{C}^{n\times r}$ be the first r columns of U, V, respectively, and let $\Sigma_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$. Then $A = U_r\Sigma_r V_r^*$ (exercise).

Applications of SVD

If $m = n$, then $A \in \mathbb{C}^{n\times n}$ has eigenvalues as well as singular values. These can differ significantly. For example, if A is nilpotent, then all of its eigenvalues are 0. But all of the singular values of A vanish iff $A = 0$. However, for A normal, we have:

Proposition. Let $A \in \mathbb{C}^{n\times n}$ be normal, and order the eigenvalues of A as

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

Then the singular values of A are $\sigma_i = |\lambda_i|$, $1 \leq i \leq n$.

Proof. By the Spectral Theorem for normal operators, there is a unitary $V \in \mathbb{C}^{n\times n}$ for which $A = V\Lambda V^*$, where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. For $1 \leq i \leq n$, choose $d_i \in \mathbb{C}$ for which $\bar{d}_i\lambda_i = |\lambda_i|$ and $|d_i| = 1$, and let $D = \operatorname{diag}(d_1, \ldots, d_n)$. Then D is unitary, and

$$A = (VD)(D^*\Lambda)V^* \equiv U\Sigma V^*,$$

where $U = VD$ is unitary and $\Sigma = D^*\Lambda = \operatorname{diag}(|\lambda_1|, \ldots, |\lambda_n|)$ is diagonal with decreasing nonnegative diagonal entries. □

Note that both the right and left singular vectors (columns of V, U) are eigenvectors of A; the columns of U have been scaled by the complex numbers d_i of modulus 1.

The Frobenius and Euclidean operator norms of $A \in \mathbb{C}^{m\times n}$ are easily expressed in terms of the singular values of A:

$$\|A\|_F = \left(\sum_{i=1}^{n} \sigma_i^2\right)^{\frac{1}{2}} \quad \text{and} \quad \|A\| = \sigma_1 = \sqrt{\rho(A^*A)},$$

as follows from the unitary invariance of these norms. There are no such simple expressions (in general) for these norms in terms of the eigenvalues of A if A is square (but not normal). Also, one cannot use the spectral radius $\rho(A)$ as a norm on $\mathbb{C}^{n\times n}$ because it is possible for $\rho(A) = 0$ and $A \neq 0$; however, on the *subspace* of $\mathbb{C}^{n\times n}$ consisting of the normal matrices, $\rho(A)$ is a norm since it agrees with the Euclidean operator norm for normal matrices.

The SVD is useful computationally for questions involving rank. The rank of $A \in \mathbb{C}^{m\times n}$ is the number of nonzero singular values of A since rank is invariant under pre- and post-multiplication by invertible matrices. There are stable numerical algorithms for computing SVD (try on `matlab`). In the presence of round-off error, row-reduction to echelon form usually fails to find the rank of A when its rank is $< \min(m, n)$; for such a matrix, the computed SVD has the zero singular values computed to be on the order of machine ϵ, and these are often identifiable as "numerical zeroes." For example, if the computed singular values of A are $10^2, 10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-15}, 10^{-15}, 10^{-16}$ with machine $\epsilon \approx 10^{-16}$, one can safely expect $\operatorname{rank}(A) = 7$.

Another application of the SVD is to derive the polar form of a matrix. This is the analogue of the polar form $z = re^{i\theta}$ in $\mathbb{C}$. (Note from problem 1 on Prob. Set 6, $U \in \mathbb{C}^{n\times n}$ is unitary iff $U = e^{iH}$ for some Hermitian $H \in \mathbb{C}^{n\times n}$).

Polar Form

Every $A \in \mathbb{C}^{n\times n}$ may be written as $A = PU$, where P is positive semi-definite Hermitian and U is unitary.

Proof. Let $A = U\Sigma V^*$ be a SVD for A, and write

$$A = (U\Sigma U^*)(UV^*).$$

Then $U\Sigma U^*$ is positive semi-definite Hermitian and UV^* is unitary. □

Observe in the proof that the eigenvalues of P are the singular values of A; this is true for any polar decomposition of A (exercise). We note that in the polar form $A = PU$, P is always uniquely determined and U is uniquely determined if A is invertible (as in $z = re^{i\theta}$). The uniqueness of P follows from the following two facts:

(i) $AA^* = PUU^*P^* = P^2$ and

(ii) every positive semi-definite Hermitian matrix has a unique positive semi-definite Hermitian square root (see H-J, Theorem 7.2.6).

If A is invertible, then so is P, so $U = P^{-1}A$ is also uniquely determined. There is also a version of the polar form for non-square matrices; see H-J for details.

Linear Least Squares Problems

If $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$, the linear system $Ax = b$ might not be solvable. Instead, we can solve the minimization problem: find $x \in \mathbb{C}^n$ to attain $\inf_{x\in\mathbb{C}^n} \|Ax - b\|^2$ (Euclidean norm).

This is called a least-squares problem since the square of the Euclidean norm is a sum of squares. At a minimum of $\varphi(x) = \|Ax - b\|^2$ we must have $\nabla\varphi(x) = 0$, or equivalently

$$\varphi'(x; v) = 0 \quad \forall\, v \in \mathbb{C}^n,$$

where

$$\varphi'(x; v) = \frac{d}{dt}\varphi(x + tv)\bigg|_{t=0}$$

is the directional derivative. If $y(t)$ is a differentiable curve in $\mathbb{C}^m$, then

$$\frac{d}{dt}\|y(t)\|^2 = \langle y'(t), y(t)\rangle + \langle y(t), y'(t)\rangle = 2\mathcal{R}e\langle y(t), y'(t)\rangle.$$

Taking $y(t) = A(x + tv) - b$, we obtain that

$$\nabla\varphi(x) = 0 \Leftrightarrow (\forall\, v \in \mathbb{C}^n)\ 2\mathcal{R}e\langle Ax - b, Av\rangle = 0 \Leftrightarrow A^*(Ax - b) = 0,$$

i.e.,

$$A^*Ax \;=\; A^*b\ .$$

These are called the *normal equations* (they say $(Ax - b) \perp \mathcal{R}(A)$).

Linear Least Squares, SVD, and Moore-Penrose Pseudoinverse

The Projection Theorem (for finite dimensional S)

Let V be an inner product space and let $S \subset V$ be a finite dimensional subspace. Then

(1) $V = S \oplus S^\perp$, i.e., given $v \in V$, $\exists$ unique $y \in S$ and $z \in S^\perp$ for which

$$v = y + z$$

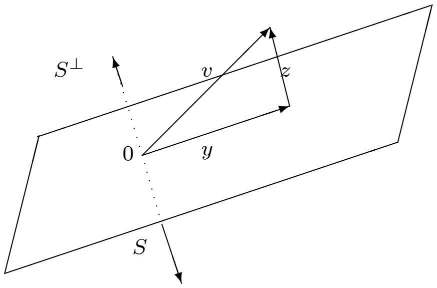
(so $y = Pv$, where P is the orthogonal projection of V onto S; also $z = (I - P)v$, and $I - P$ is the orthogonal projection of V onto $S^\perp$).

(2) Given $v \in V$, the y in (1) is the unique element of S which satisfies $v - y \in S^\perp$.

(3) Given $v \in V$, the y in (1) is the unique element of S realizing the minimum

$$\min_{s \in S}\ \|v - s\|^2\ .$$

Remark. The content of the Projection Theorem is contained in the following picture:



Proof. (1) Let $\{\psi_1, \ldots, \psi_r\}$ be an orthonormal basis of S. Given $v \in V$, let

$$y = \sum_{j=1}^{r} \langle \psi_j, v \rangle \psi_j \quad \text{and} \quad z = v - y.$$

Then $v = y + z$ and $y \in S$. For $1 \le k \le r$,

$$\langle \psi_k, z \rangle = \langle \psi_k, v \rangle - \langle \psi_k, y \rangle = \langle \psi_k, v \rangle - \langle \psi_k, v \rangle = 0,$$

so $z \in S^\perp$. Uniqueness follows from the fact that $S \cap S^\perp = \{0\}$.

(2) Since $z = v - y$, this is just a restatement of $z \in S^\perp$.

(3) For any $s \in S$,

$$v - s = \underbrace{y - s}_{\in S} + \underbrace{z}_{\in S^\perp},$$

so by the Pythagorean Theorem ($p \perp q \Rightarrow \|p \pm q\|^2 = \|p\|^2 + \|q\|^2$),

$$\|v - s\|^2 = \|y - s\|^2 + \|z\|^2.$$

Therefore, $\|v - s\|^2$ is minimized iff $s = y$, and then $\|v - y\|^2 = \|z\|^2$. □

Theorem: [Normal Equations for Linear Least Squares]

Let $A \in \mathbb{C}^{m \times n}$, $b \in \mathbb{C}^m$ and $\|\cdot\|$ be the Euclidean norm. Then $x \in \mathbb{C}^n$ realizes the minimum:

$$\min_{x \in \mathbb{C}^n} \|b - Ax\|^2$$

if and only if x is a solution to the normal equations $A^*Ax = A^*b$.

Proof. Recall from early in the course that we showed that $\mathcal{R}(L)^a = \mathcal{N}(L')$ for any linear transformation L. If we identify $\mathbb{C}^n$ and $\mathbb{C}^m$ with their duals using the Euclidean inner product and take L to be multiplication by A, this can be rewritten as $\mathcal{R}(A)^\perp = \mathcal{N}(A^*)$.

Now apply the Projection Theorem, taking $S = \mathcal{R}(A)$ and $v = b$. Any $s \in S$ can be represented as Ax for some $x \in \mathbb{C}^n$ (not necessarily unique if $\operatorname{rank}(A) < n$). We conclude that $y = Ax$ realizes the minimum iff

$$b - Ax \in \mathcal{R}(A)^\perp = \mathcal{N}(A^*),$$

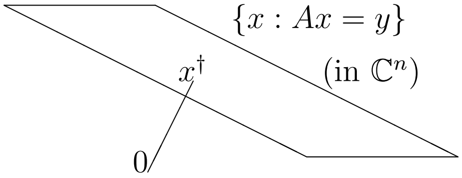
or equivalently $A^*Ax = A^*b$. □

The minimizing element $s = y \in S$ is unique. Since $y \in \mathcal{R}(A)$, there exists $x \in \mathbb{C}^n$ for which $Ax = y$, or equivalently, there exists $x \in \mathbb{C}^n$ minimizing $\|b - Ax\|^2$. Consequently, there is an $x \in \mathbb{C}^n$ for which $A^*Ax = A^*b$, that is, the normal equations are consistent.

If $\operatorname{rank}(A) = n$, then there is a unique $x \in \mathbb{C}^n$ for which $Ax = y$. This x is the unique minimizer of $\|b - Ax\|^2$ as well as the unique solution of the normal equations $A^*Ax = A^*b$. However, if $\operatorname{rank}(A) = r < n$, then the minimizing vector x is not unique; x can be modified by adding any element of $\mathcal{N}(A)$. (Exercise. Show $\mathcal{N}(A) = \mathcal{N}(A^*A)$.) However, there is a unique

$$x^\dagger \in \{x \in \mathbb{C}^n : x \text{ minimizes } \|b - Ax\|^2\} = \{x \in \mathbb{C}^n \,:\, Ax = y\}$$

of minimum norm ($x^\dagger$ is read "x dagger").



To see this, note that since $\{x \in \mathbb{C}^n \,:\, Ax = y\}$ is an affine translate of the subspace $\mathcal{N}(A)$, a translated version of the Projection Theorem shows that there is a unique $x^\dagger \in \{x \in \mathbb{C}^n \,:\, Ax = y\}$ for which $x^\dagger \perp \mathcal{N}(A)$ and this $x^\dagger$ is the unique element of $\{x \in \mathbb{C}^n \,:\, Ax = y\}$ of minimum norm. Notice also that $\{x : Ax = y\} = \{x : A^*Ax = A^*b\}$.

In summary: given $b \in \mathbb{C}^m$, then $x \in \mathbb{C}^n$ minimizes $\|b - Ax\|^2$ over $x \in \mathbb{C}^n$ iff Ax is the orthogonal projection of b onto $\mathcal{R}(A)$, and among this set of solutions there is a unique $x^\dagger$ of minimum norm. Alternatively, $x^\dagger$ is the unique solution of the normal equations $A^*Ax = A^*b$ which also satisfies $x^\dagger \in \mathcal{N}(A)^\perp$.

The map $A^\dagger : \mathbb{C}^m \to \mathbb{C}^n$ which maps $b \in \mathbb{C}^m$ into the unique minimizer $x^\dagger$ of $\|b - Ax\|^2$ of minimum norm is called the Moore-Penrose pseudo-inverse of A. We will see momentarily that $A^\dagger$ is linear, so it is represented by an $n \times m$ matrix which we also denote by $A^\dagger$ (and we also call this matrix the Moore-Penrose pseudo-inverse of A). If $m = n$ and A is invertible, then every $b \in \mathbb{C}^n$ is in $\mathcal{R}(A)$, so $y = b$, and the solution of $Ax = b$ is unique, given by $x = A^{-1}b$. In this case $A^\dagger = A^{-1}$. So the pseudo-inverse is a generalization of the inverse to possibly non-square, non-invertible matrices.

The linearity of the map $A^\dagger$ can be seen as follows. For $A \in \mathbb{C}^{m\times n}$, the above considerations show that $A|_{\mathcal{N}(A)^\perp}$ is injective and maps onto $\mathcal{R}(A)$. Thus $A|_{\mathcal{N}(A)^\perp} : \mathcal{N}(A)^\perp \to \mathcal{R}(A)$ is an isomorphism. The definition of $A^\dagger$ amounts to the formula

$$A^\dagger = (A|_{\mathcal{N}(A)^\perp})^{-1} \circ P_1,$$

where $P_1 : \mathbb{C}^m \to \mathcal{R}(A)$ is the orthogonal projection onto $\mathcal{R}(A)$. Since P_1 and $(A|_{\mathcal{N}(A)^\perp})^{-1}$ are linear transformations, so is $A^\dagger$.

The pseudo-inverse of A can be written nicely in terms of the SVD of A. Let $A = U\Sigma V^*$ be an SVD of A, and let $r = \text{rank}\,(A)$ (so $\sigma_1 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1} = \cdots$). Define

$$\Sigma^\dagger = \text{diag}\,(\sigma_1^{-1}, \ldots, \sigma_r^{-1}, 0, \ldots, 0) \in \mathbb{C}^{n\times m}.$$

(Note: It is appropriate to call this matrix $\Sigma^\dagger$ as it is easily shown that the pseudo-inverse of $\Sigma \in \mathbb{C}^{m\times n}$ is this matrix (exercise).)

Proposition. If $A = U\Sigma V^*$ is an SVD of A, then $A^\dagger = V\Sigma^\dagger U^*$ is an SVD of $A^\dagger$.

Proof. Denote by $u_1, \cdots u_m$ the columns of U and by $v_1, \cdots, v_n$ the columns of V. The statement that $A = U\Sigma V^*$ is an SVD of A is equivalent to the three conditions:

1. $\{u_1, \cdots u_m\}$ is an orthonormal basis of $\mathbb{C}^m$ such that $\text{span}\{u_1, \cdots u_r\} = \mathcal{R}(A)$

2. $\{v_1, \cdots, v_n\}$ is an orthonormal basis for $\mathbb{C}^n$ such that $\text{span}\{v_{r+1}, \cdots, v_n\} = \mathcal{N}(A)$

3. $Av_i = \sigma_i u_i$ for $1 \leq i \leq r$.

The conditions on the spans in 1. and 2. are equivalent to $\text{span}\{u_{r+1}, \cdots u_m\} = \mathcal{R}(A)^\perp$ and $\text{span}\{v_1, \cdots, v_r\} = \mathcal{N}(A)^\perp$. The formula

$$A^\dagger = (A|_{\mathcal{N}(A)^\perp})^{-1} \circ P_1$$

shows that $\text{span}\{u_{r+1}, \cdots u_m\} = \mathcal{N}(A^\dagger)$, $\text{span}\{v_1, \cdots, v_r\} = \mathcal{R}(A^\dagger)$, and that $A^\dagger u_i = \sigma_i^{-1} v_i$ for $1 \leq i \leq r$. Thus the conditions 1.-3. hold for $A^\dagger$ with U and V interchanged and with σ_i replaced by σ_i^{-1}. Hence $A^\dagger = V\Sigma^\dagger U^*$ is an SVD for $A^\dagger$. □

A similar formula can be written for the abbreviated form of the SVD. If $U_r \in \mathbb{C}^{m\times r}$ and $V_r \in \mathbb{C}^{n\times r}$ are the first r columns of U, V, respectively, and

$$\Sigma_r = \text{diag}\,(\sigma_1, \ldots, \sigma_r) \in \mathbb{C}^{r\times r},$$

then the abbreviated form of the SVD of A is $A = U_r\Sigma_r V_r^*$. The above Proposition shows that $A^\dagger = V_r\Sigma_r^{-1}U_r^*$.

One rarely actually computes $A^\dagger$. Instead, to minimize $\|b - Ax\|^2$ using the SVD of A one computes

$$x^\dagger \;=\; V_r(\Sigma_r^{-1}(U_r^* b)) \;.$$

For $b \in \mathbb{C}^m$, we saw above that if $x^\dagger = A^\dagger b$, then $Ax^\dagger = y$ is the orthogonal projection of b onto $\mathcal{R}(A)$. Thus $AA^\dagger$ is the orthogonal projection of $\mathbb{C}^m$ onto $\mathcal{R}(A)$. This is also clear directly from the SVD:

$$AA^\dagger = U_r\Sigma_r V_r^* V_r \Sigma_r^{-1} U_r^* = U_r\Sigma_r\Sigma_r^{-1}U_r^* = U_rU_r^* = \Sigma_{j=1}^r u_j u_j^*$$

which is clearly the orthogonal projection onto $\mathcal{R}(A)$. (Note that $V_r^*V_r = I_r$ since the columns of V are orthonormal.) Similarly, since $w = A^\dagger(Ax)$ is the vector of least length satisfying $Aw = Ax$, $A^\dagger A$ is the orthogonal projection of $\mathbb{C}^n$ onto $\mathcal{N}(A)^\perp$. Again, this also is clear directly from the SVD:

$$A^\dagger A = V_r \Sigma_r^{-1} U_r^* U_r \Sigma_r V_r^* = V_r V_r^* = \Sigma_{j=1}^r v_j v_j^*$$

is the orthogonal projection onto $\mathcal{R}(V_r) = \mathcal{N}(A)^\perp$. These relationships are substitutes for $AA^{-1} = A^{-1}A = I$ for invertible $A \in \mathbb{C}^{n\times n}$. Similarly, one sees that

(i) $AXA = A$,

(ii) $XAX = X$,

(iii) $(AX)^* = AX$,

(iv) $(XA)^* = XA$,

where $X = A^\dagger$. In fact, one can show that $X \in \mathbb{C}^{n\times m}$ is $A^\dagger$ if and only if X satisfies (i), (ii), (iii), (iv). (Exercise — see section 5.54 in Golub and Van Loan.)

The pseudo inverse can be used to extend the (Euclidean operator norm) condition number to general matrices: $\kappa(A) = \|A\| \cdot \|A^\dagger\| = \sigma_1/\sigma_r$ (where $r = \operatorname{rank} A$).

LU Factorization

All of the matrix factorizations we have studied so far are spectral factorizations in the sense that in obtaining these factorizations, one is obtaining the eigenvalues and eigenvectors of A (or matrices related to A, like A^*A and AA^* for SVD). We end our discussion of matrix factorizations by mentioning two non-spectral factorizations. These non-spectral factorizations can be determined directly from the entries of the matrix, and are computationally less expensive than spectral factorizations. Each of these factorizations amounts to a reformulation of a procedure you are already familiar with. The LU factorization is a reformulation of Gaussian Elimination, and the QR factorization is a reformulation of Gram-Schmidt orthogonalization.

Recall the method of Gaussian Elimination for solving a system $Ax = b$ of linear equations, where $A \in \mathbb{C}^{n\times n}$ is invertible and $b \in \mathbb{C}^n$. If the coefficient of x_1 in the first equation is nonzero, one eliminates all occurrences of x_1 from all the other equations by adding appropriate multiples of the first equation. These operations do not change the set of solutions to the equation. Now if the coefficient of x_2 in the new second equation is nonzero, it can be used to eliminate x_2 from the further equations, etc. In matrix terms, if

$$A = \begin{bmatrix} a & v^T \\ u & \widetilde{A} \end{bmatrix} \in \mathbb{C}^{n\times n}$$

with $a \neq 0$, $a \in \mathbb{C}$, $u, v \in \mathbb{C}^{n-1}$, and $\widetilde{A} \in \mathbb{C}^{(n-1)\times(n-1)}$, then using the first row to zero out u amounts to left multiplication of the matrix A by the matrix

$$\begin{bmatrix} 1 & 0 \\ -\frac{u}{a} & I \end{bmatrix}$$

to get

$$(*) \qquad \begin{bmatrix} 1 & 0 \\ -\frac{u}{a} & I \end{bmatrix} \begin{bmatrix} a & v^T \\ u & \widetilde{A} \end{bmatrix} = \begin{bmatrix} a & v^T \\ 0 & A_1 \end{bmatrix} .$$

Define

$$L_1 = \begin{bmatrix} 1 & 0 \\ \frac{u}{a} & I \end{bmatrix} \in \mathbb{C}^{n\times n} \quad \text{and} \quad U_1 = \begin{bmatrix} a & v^T \\ 0 & A_1 \end{bmatrix}$$

and observe that

$$L_1^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{u}{a} & I \end{bmatrix} .$$

Hence (*) becomes

$$L_1^{-1}A = U_1, \text{ or equivalently,} \quad A = L_1U_1 .$$

Note that L_1 is lower triangular and U_1 is block upper-triangular with one 1×1 block and one $(n-1) \times (n-1)$ block on the block diagonal. The components of $\frac{u}{a} \in \mathbb{C}^{n-1}$ are called *multipliers*, they are the multiples of the first row subtracted from subsequent rows, and they are computed in the Gaussian Elimination algorithm. The multipliers are usually denoted

$$u/a = \begin{bmatrix} m_{21} \\ m_{31} \\ \vdots \\ m_{n1} \end{bmatrix} .$$

Now, if the $(1,1)$ entry of A_1 is not 0, we can apply the same procedure to A_1: if

$$A_1 = \begin{bmatrix} a_1 & v_1^T \\ u_1 & \widetilde{A}_1 \end{bmatrix} \in \mathbb{C}^{(n-1)\times(n-1)}$$

with $a_1 \neq 0$, letting

$$\widetilde{L}_2 = \begin{bmatrix} 1 & 0 \\ \frac{u_1}{a_1} & I \end{bmatrix} \in \mathbb{C}^{(n-1)\times(n-1)}$$

and forming

$$\widetilde{L}_2^{-1}A_1 = \begin{bmatrix} 1 & 0 \\ -\frac{u_1}{a_1} & I \end{bmatrix} \begin{bmatrix} a_1 & v_1^T \\ u_1 & \widetilde{A}_1 \end{bmatrix} = \begin{bmatrix} a_1 & v_1^T \\ 0 & A_2 \end{bmatrix} \equiv \widetilde{U}_2 \in \mathbb{C}^{(n-1)\times(n-1)}$$

(where $A_2 \in \mathbb{C}^{(n-2)\times(n-2)}$) amounts to using the second row to zero out elements of the second column below the diagonal. Setting $L_2 = \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{L}_2 \end{bmatrix}$ and $U_2 = \begin{bmatrix} a & v^T \\ 0 & \widetilde{U}_2 \end{bmatrix}$, we have

$$L_2^{-1}L_1^{-1}A = \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{L}_2^{-1} \end{bmatrix} \begin{bmatrix} a & v^T \\ 0 & A_1 \end{bmatrix} = U_2,$$

which is block upper triangular with two 1×1 blocks and one $(n-2) \times (n-2)$ block on the block diagonal. The components of $\frac{u_1}{a_1}$ are multipliers, usually denoted

$$\frac{u_1}{a_1} = \begin{bmatrix} m_{32} \\ m_{42} \\ \vdots \\ m_{n2} \end{bmatrix}.$$

Notice that these multipliers appear in L_2 in the *second* column, below the diagonal. Continuing in a similar fashion,

$$L_{n-1}^{-1} \cdots L_2^{-1} L_1^{-1} A = U_{n-1} \equiv U$$

is upper triangular (provided along the way that the $(1,1)$ entries of $A, A_1, A_2, \ldots, A_{n-2}$ are nonzero so the process can continue). Define $L = (L_{n-1}^{-1} \cdots L_1^{-1})^{-1} = L_1 L_2 \cdots L_{n-1}$. Then $A = LU$. (Remark: A lower triangular matrix with 1's on the diagonal is called a *unit* lower triangular matrix, so $L_j, L_j^{-1}, L_{j-1}^{-1} \cdots L_1^{-1}, L_1 \cdots L_j, L^{-1}, L$ are all unit lower triangular.) For an invertible $A \in \mathbb{C}^{n \times n}$, writing $A = LU$ as a product of a unit lower triangular matrix L and a (necessarily invertible) upper triangular matrix U (both in $\mathbb{C}^{n \times n}$) is called the *LU factorization* of A.

Remarks:

(1) If $A \in \mathbb{C}^{n \times n}$ is invertible and has an LU factorization, it is unique (exercise).

(2) One can show that $A \in \mathbb{C}^{n \times n}$ has an LU factorization iff for $1 \leq j \leq n$, the upper left $j \times j$ principal submatrix $\begin{bmatrix} a_{11} & \cdots & a_{1j} \\ \vdots & & \\ a_{j1} & \cdots & a_{jj} \end{bmatrix}$ is invertible.

(3) Not every invertible $A \in \mathbb{C}^{n \times n}$ has an LU-factorization. (Example: $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ doesn't.) Typically, one must permute the rows of A to move nonzero entries to the appropriate spot for the elimination to proceed. Recall that a permutation matrix $P \in \mathbb{C}^{n \times n}$ is the identity I with its rows (or columns) permuted. Any such $P \in \mathbb{R}^{n \times n}$ is orthogonal, so $P^{-1} = P^T$. Permuting the rows of A amounts to left multiplication by a permutation matrix P^T; then $P^T A$ has an LU factorization, so $A = PLU$ (called the PLU factorization of A).

(4) Fact: Every invertible $A \in \mathbb{C}^{n \times n}$ has a (not necessarily unique) PLU factorization.

(5) It turns out that $L = L_1 \cdots L_{n-1} = \begin{bmatrix} 1 & & 0 \\ m_{21} & & \\ \vdots & \ddots & \\ m_{n-1} & \cdots & 1 \end{bmatrix}$ has the multipliers m_{ij} below the diagonal.

(6) The LU factorization can be used to solve linear systems $Ax = b$ (where $A = LU \in \mathbb{C}^{n\times n}$ is invertible). The system $Ly = b$ can be solved by forward substitution (first equation gives x_1, etc.), and $Ux = y$ can be solved by back-substitution (n^{th} equation gives x_n, etc.), giving the solution of $Ax = LUx = b$. See section 3.5 of H-J.

QR Factorization

Recall first the Gram-Schmidt orthogonalization process. Let V be an inner product space, and suppose $a_1, \ldots, a_n \in V$ are linearly independent. Define $q_1, \ldots, q_n$ inductively, as follows: let $p_1 = a_1$ and $q_1 = p_1/\|p_1\|$; then for $2 \leq j \leq n$, let

$$p_j = a_j - \sum_{i=1}^{j-1} \langle q_i, a_j \rangle q_i \qquad \text{and} \qquad q_j = p_j/\|p_j\|.$$

Since clearly for $1 \leq k \leq n$ we have $q_k \in \operatorname{span}\{a_1, \ldots, a_k\}$, each p_j is nonzero by the linear independence of $\{a_1, \ldots, a_n\}$, so each q_j is well-defined. It is easily seen that $\{q_1, \ldots, q_n\}$ is an orthonormal basis for $\operatorname{span}\{a_1, \ldots, a_n\}$. The relations above can be solved for a_j to express a_j in terms of the q_i with $i \leq j$. Defining $r_{jj} = \|p_j\|$ (so $p_j = r_{jj} q_j$) and $r_{ij} = \langle q_i, a_j \rangle$ for $1 \leq i < j \leq n$, we have: $a_1 = r_{11} q_1$, $a_2 = r_{12} q_1 + r_{22} q_2$, and in general $a_j = \sum_{i=1}^{j} r_{ij} q_i$.

Remarks:

(1) If $a_1, a_2, \cdots$ is a linearly independent sequence in V, we can apply the Gram-Schmidt process to obtain an orthonormal sequence $q_1, q_2, \ldots$ with the property that for $k \geq 1$, $\{q_1, \ldots, q_k\}$ is an orthonormal basis for $\operatorname{span}\{a_1, \ldots, a_k\}$.

(2) If the a_j's are linearly dependent, then for some value(s) of k, $a_k \in \operatorname{span}\{a_1, \ldots, a_{k-1}\}$, and then $p_k = 0$. The process can be modified by setting $q_k = 0$ and proceeding. We end up with orthogonal q_j's, some of which have $\|q_j\| = 1$ and some have $\|q_j\| = 0$. Then for $k \geq 1$, the nonzero vectors in the set $\{q_1, \ldots, q_k\}$ form an orthonormal basis for $\operatorname{span}\{a_1, \ldots, a_k\}$.

(3) The classical Gram-Schmidt algorithm described above applied to n linearly independent vectors $a_1, \ldots, a_n \in \mathbb{C}^m$ (where of course $m \geq n$) does not behave well computationally. Due to the accumulation of round-off error, the computed q_j's are not as orthogonal as one would want (or need in applications): $\langle q_j, q_k \rangle$ is small for $j \neq k$ with j near k, but not so small for $j \ll k$ or $j \gg k$. An alternate version, "Modified Gram-Schmidt," is equivalent in exact arithmetic, but behaves better numerically. In the following "pseudo-codes," p denotes a temporary storage vector used to accumulate the sums defining the p_j's.

```
Classic Gram-Schmidt
For  $j = 1, \cdots, n$ do
    $p := a_j$
    For $i = 1, \ldots, j-1$ do
        $r_{ij} = \langle q_i, a_j \rangle$
        $p := p - r_{ij} q_i$
    $r_{jj} := \|p\|$
    $q_j := p/r_{jj}$
```

```
Modified Gram-Schmidt
For  $j = 1, \ldots, n$ do
    $p := a_j$
    For $i = 1, \ldots, j-1$ do
        $r_{ij} = \langle q_i, p \rangle$
        $p := p - r_{ij} q_i$
    $r_{jj} = \|p\|$
    $q_j := p/r_{jj}$
```

The only difference is in the computation of r_{ij}: in Modified Gram-Schmidt, we orthogonalize the accumulated partial sum for p_j against each q_i successively.

Proposition. Suppose $A \in \mathbb{C}^{m\times n}$ with $m \geq n$. Then $\exists\, Q \in \mathbb{C}^{m\times m}$ which is unitary and an upper triangular $R \in \mathbb{C}^{m\times n}$ (i.e. $r_{ij} = 0$ for $i > j$) for which $A = QR$. If $\widetilde{Q} \in \mathbb{C}^{m\times n}$ denotes the first n columns of Q and $\widetilde{R} \in \mathbb{C}^{n\times n}$ denotes the first n rows of R, then clearly also $A = QR = [\widetilde{Q}\ *]\begin{bmatrix} \widetilde{R} \\ 0 \end{bmatrix} = \widetilde{Q}\widetilde{R}$. Moreover

(a) We may choose an R with nonnegative diagonal entries.

(b) If A is of full rank (i.e. $\operatorname{rank}(A) = n$, or equivalently the columns of A are linearly independent), then we may choose an R with positive diagonal entries, in which case the condensed factorization $A = \widetilde{Q}\widetilde{R}$ is unique (and thus in this case if $m = n$, the factorization $A = QR$ is unique since then $Q = \widetilde{Q}$ and $R = \widetilde{R}$).

(c) If A is of full rank, the condensed factorization $A = \widetilde{Q}\widetilde{R}$ is essentially unique: if $A = \widetilde{Q}_1\widetilde{R}_1 = \widetilde{Q}_2\widetilde{R}_2$, then $\exists$ a unitary diagonal matrix $D \in \mathbb{C}^{n\times n}$ for which $\widetilde{Q}_2 = \widetilde{Q}_1 D^*$ (rescaling the columns of $\widetilde{Q}_1$) and $\widetilde{R}_2 = D\widetilde{R}_1$ (rescaling the rows of $\widetilde{R}_1$).

Proof. If the columns of A are linearly independent, we can apply the Gram-Schmidt process described above. Let $\widetilde{Q} = [q_1, \ldots, q_n] \in \mathbb{C}^{m\times n}$, and define $\widetilde{R} \in \mathbb{C}^{n\times n}$ by setting $r_{ij} = 0$ for $i > j$, and r_{ij} to be the value computed in Gram-Schmidt for $i \leq j$. Then $A = \widetilde{Q}\widetilde{R}$. Extending $\{q_1, \ldots, q_n\}$ to an orthonormal basis $\{q_1, \ldots, q_m\}$ of $\mathbb{C}^m$, and setting $Q = [q_1, \ldots, q_m]$ and $R = \begin{bmatrix} \widetilde{R} \\ 0 \end{bmatrix} \in \mathbb{C}^{m\times n}$, we have $A = QR$. Since $r_{jj} > 0$ in G-S, we have the existence part of (b). Uniqueness follows by induction passing through the G-S process again, noting that at each step we have no choice. (c) follows easily from (b) since if $\operatorname{rank}(A) = n$, then $\operatorname{rank}(\widetilde{R}) = n$ in any $\widetilde{Q}\widetilde{R}$ factorization of A.

If the columns of A are linearly dependent, we alter the Gram-Schmidt algorithm as in Remark (2) above. Notice that $q_k = 0$ iff $r_{kj} = 0\,\forall\, j$, so if $\{q_{k_1}, \ldots, q_{k_r}\}$ are the nonzero vectors in $\{q_1, \ldots, q_n\}$ (where of course $r = \operatorname{rank}(A)$), then the nonzero rows in R are precisely rows $k_1, \ldots, k_r$. So if we define $\widehat{Q} = [q_{k_1} \cdots q_{k_r}] \in \mathbb{C}^{m\times r}$ and $\widehat{R} \in \mathbb{C}^{r\times n}$ to be these nonzero rows, then $\widehat{Q}\widehat{R} = A$ where $\widehat{Q}$ has orthonormal columns and $\widehat{R}$ is upper triangular. Let Q be a unitary matrix whose first r columns are $\widehat{Q}$, and let $R = \begin{bmatrix} \widehat{R} \\ 0 \end{bmatrix} \in \mathbb{C}^{m\times n}$. Then $A = QR$. (Notice that in addition to (a), we actually have constructed an R for which, in each nonzero row, the first nonzero element is positive.) □

Remarks (continued):

(4) If $A \in \mathbb{R}^{m\times n}$, everything can be done in real arithmetic, so, e.g., $Q \in \mathbb{R}^{m\times m}$ is orthogonal and $R \in \mathbb{R}^{m\times n}$ is real, upper triangular.

(5) In practice, there are more efficient and better computationally behaved ways of calculating the Q and R factors. The idea is to create zeros below the diagonal (successively in columns $1, 2, \ldots$) as in Gaussian Elimination, except we now use Householder transformations (which are unitary) instead of the unit lower triangular matrices L_j. Details will be described in an upcoming problem set.

Using QR Factorization to Solve Least Squares Problems

Suppose $A \in \mathbb{C}^{m\times n}$, $b \in \mathbb{C}^m$, and $m \geq n$. Assume A has full rank ($\operatorname{rank}(A) = n$). The QR factorization can be used to solve the least squares problem of minimizing $\|b - Ax\|^2$ (which has a unique solution in this case). Let $A = QR$ be a QR factorization of A, with condensed form $\widetilde{Q}\widetilde{R}$, and write $Q = [\widetilde{Q}\ Q']$ where $Q' \in \mathbb{C}^{m\times(m-n)}$. Then

$$\|b - Ax\|^2 = \|b - QRx\|^2 = \|Q^*b - Rx\|^2 = \left\| \begin{bmatrix} \widetilde{Q}^* \\ Q'^* \end{bmatrix} b - \begin{bmatrix} \widetilde{R} \\ 0 \end{bmatrix} x \right\|^2$$

$$= \left\| \begin{bmatrix} \widetilde{Q}^*b - \widetilde{R}x \\ Q'^*b \end{bmatrix} \right\|^2 = \|\widetilde{Q}^*b - \widetilde{R}x\|^2 + \|Q'^*b\|^2.$$

Here $\widetilde{R} \in \mathbb{C}^{n\times n}$ is an invertible upper triangle matrix, so that x minimizes $\|b - Ax\|^2$ iff $\widetilde{R}x = \widetilde{Q}^*b$. This invertible upper triangular $n \times n$ system for x can be solved by back-substitution. Note that we only need $\widetilde{Q}$ and $\widetilde{R}$ to solve for x.

The QR Algorithm

The QR algorithm is used to compute a specific Schur unitary triangularization of a matrix $A \in \mathbb{C}^{n\times n}$. The algorithm is *iterative*: We generate a sequence $A = A_0, A_1, A_2, \ldots$ of matrices which are unitarily similar to A; the goal is to get the subdiagonal elements to converge to zero, as then the eigenvalues will appear on the diagonal. If A is Hermitian, then so also are $A_1, A_2, \ldots$, so if the subdiagonal elements converge to 0, also the superdiagonal elements converge to 0, and (in the limit) we have diagonalized A. The QR algorithm is the most commonly used method for computing all the eigenvalues (and eigenvectors if wanted) of a matrix. It behaves well numerically since all the similarity transformations are unitary.

When used in practice, a matrix is first reduced to *upper-Hessenberg form* ($h_{ij} = 0$ for $i > j + 1$) using unitary similarity transformations built from Householder reflections (or Givens rotations), quite analogous to computing a QR factorization. Here, however, similarity transformations are being performed, so they require left and right multiplication by the Householder transformations — leading to an inability to zero out the first subdiagonal ($i = j + 1$) in the process. If A is Hermitian and upper-Hessenberg, A is tridiagonal. This initial reduction is to decrease the computational cost of the iterations in the QR algorithm. It is successful because upper-Hessenberg form is preserved by the iterations: if A_k is upper Hessenberg, so is A_{k+1}.

There are many sophisticated variants of the QR algorithm (shifts to speed up convergence, implicit shifts to allow computing a real quasi-upper triangular matrix similar to a real matrix using only real arithmetic, etc.). We consider the basic algorithm over $\mathbb{C}$.

The (Basic) QR Algorithm

Given $A \in \mathbb{C}^{n\times n}$, let $A_0 = A$. For $k = 0, 1, 2, \ldots$, starting with A_k, do a QR factorization of A_k: $A_k = Q_k R_k$, and then define $A_{k+1} = R_k Q_k$.

Remark. $R_k = Q_k^* A_k$, so $A_{k+1} = Q_k^* A_k Q_k$ is unitarily similar to A_k. The algorithm uses the Q of the QR factorization of A_k to perform the next unitary similarity transformation.

Convergence of the QR Algorithm

We will show under mild hypotheses that all of the subdiagonal elements of A_k converge to 0 as $k \to \infty$. See section 2.6 in H-J for examples where the QR algorithm does not converge. See also sections 7.5, 7.6, 8.2 in Golub and Van Loan for more discussion.

Lemma. Let Q_j ($j = 1, 2, \ldots$) be a sequence of unitary matrices in $\mathbb{C}^{n\times n}$ and R_j ($j = 1, 2, \ldots$) be a sequence of upper triangular matrices in $\mathbb{C}^{n\times n}$ with positive diagonal entries. Suppose $Q_j R_j \to I$ as $j \to \infty$. Then $Q_j \to I$ and $R_j \to I$.

Proof Sketch. Let Q_{j_k} be any subsequence of Q_j. Since the set of unitary matrices in $\mathbb{C}^{n\times n}$ is compact, $\exists$ a sub-subsequence $Q_{j_{k_l}}$ and a unitary $Q \ni Q_{j_{k_l}} \to Q$. So $R_{j_{k_l}} = Q_{j_{k_l}}^* Q_{j_{k_l}} R_{j_{k_l}} \to Q^* \cdot I = Q^*$. So Q^* is unitary, upper triangular, with nonnegative diagonal elements, which implies easily that $Q^* = I$. Thus every subsequence of Q_j has in turn a sub-subsequence converging to I. By standard metric space theory, $Q_j \to I$, and thus $R_j = Q_j^* Q_j R_j \to I \cdot I = I$. □

Theorem. *Suppose $A \in \mathbb{C}^{n\times n}$ has eigenvalues $\lambda_1, \ldots, \lambda_n$ with $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$. Choose $X \in \mathbb{C}^{n\times n} \ni X^{-1}AX = \Lambda \equiv \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, and suppose X^{-1} has an LU decomposition. Generate the sequence $A_0 = A, A_1, A_2, \ldots$ using the QR algorithm. Then the subdiagonal entries of $A_k \to 0$ as $k \to \infty$, and for $1 \le j \le n$, the j^{th} diagonal entry $\to \lambda_j$.*

Proof. Define $\widetilde{Q}_k = Q_0 Q_1 \cdots Q_k$ and $\widetilde{R}_k = R_k \cdots R_0$. Then $A_{k+1} = \widetilde{Q}_k^* A \widetilde{Q}_k$.

Claim: $\widetilde{Q}_k \widetilde{R}_k = A^{k+1}$.

Proof: Clear for $k = 0$. Suppose $\widetilde{Q}_{k-1} \widetilde{R}_{k-1} = A^k$. Then

$$R_k = A_{k+1} Q_k^* = \widetilde{Q}_k^* A \widetilde{Q}_k Q_k^* = \widetilde{Q}_k^* A \widetilde{Q}_{k-1},$$

so

$$\widetilde{R}_k = R_k\widetilde{R}_{k-1} = \widetilde{Q}_k^* A\widetilde{Q}_{k-1}\widetilde{R}_{k-1} = \widetilde{Q}_k^* A^{k+1},$$

so $\widetilde{Q}_k\widetilde{R}_k = A^{k+1}$.

Now, choose a QR factorization of X and an LU factorization of X^{-1}: $X = QR$, $X^{-1} = LU$ (where Q is unitary, L is unit lower triangular, R and U are upper triangular with nonzero diagonal entries). Then

$$A^{k+1} = X\Lambda^{k+1}X^{-1} = QR\Lambda^{k+1}LU = QR(\Lambda^{k+1}L\Lambda^{-(k+1)})\Lambda^{k+1}U.$$

Let $E_{k+1} = \Lambda^{k+1}L\Lambda^{-(k+1)} - I$ and $F_{k+1} = RE_{k+1}R^{-1}$.

Claim: $E_{k+1} \to 0$ (and thus $F_{k+1} \to 0$) as $k \to \infty$.

Proof: Let l_{ij} denote the elements of L. E_{k+1} is strictly lower triangular, and for $i > j$ its ij element is $\left(\frac{\lambda_i}{\lambda_j}\right)^{k+1} l_{ij} \to 0$ as $k \to \infty$ since $|\lambda_i| < |\lambda_j|$.

Now $A^{k+1} = QR(I + E_{k+1})\Lambda^{k+1}U$, so $A^{k+1} = Q(I + F_{k+1})R\Lambda^{k+1}U$. Choose a QR factorization of $I+F_{k+1}$ (which is invertible) $I+F_{k+1} = \widehat{Q}_{k+1}\widehat{R}_{k+1}$ where $\widehat{R}_{k+1}$ has positive diagonal entries. By the Lemma, $\widehat{Q}_{k+1} \to I$ and $\widehat{R}_{k+1} \to I$. Since $A^{k+1} = (Q\widehat{Q}_{k+1})(\widehat{R}_{k+1}R\Lambda^{k+1}U)$ and $A^{k+1} = \widetilde{Q}_k\widetilde{R}_k$, the essential uniqueness of QR factorizations of invertible matrices implies $\exists$ a unitary diagonal matrix D_k for which $Q\widehat{Q}_{k+1}D_k^* = \widetilde{Q}_k$ and $D_k\widehat{R}_{k+1}\Lambda^{k+1}U = \widetilde{R}_k$. So $\widetilde{Q}_kD_k = Q\widehat{Q}_{k+1} \to Q$, and thus

$$D_k^*A_{k+1}D_k = D_k^*\widetilde{Q}_k^*A\widetilde{Q}_kD_k \to Q^*AQ \quad \text{as} \;\; k \to \infty.$$

But

$$Q^*AQ = Q^*(QR\Lambda X^{-1})QRR^{-1} = R\Lambda R^{-1}$$

is upper triangular with diagonal entries $\lambda_1, \ldots, \lambda_n$ in that order. Since D_k is unitary and diagonal, the lower triangular part of $R\Lambda R^{-1}$ and of $D_kR\Lambda R^{-1}D_k^*$ are the same, namely

$$\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}.$$

Thus

$$\|A_{k+1} - D_kR\Lambda R^{-1}D_k^*\| = \|D_k^*A_{k+1}D_k - R\Lambda R^{-1}\| \to 0,$$

and the Theorem follows. □

Note that the proof shows that $\exists$ a sequence $\{D_k\}$ of unitary diagonal matrices for which $D_k^*A_{k+1}D_k \to R\Lambda R^{-1}$. So although the superdiagonal ($i < j$) elements of A_{k+1} may not converge, the magnitude of each superdiagonal element converges.

As a partial explanation for why the QR algorithm works, we show how the convergence of the first column of A_k to $\begin{bmatrix} \lambda_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ follows from the power method. (See Problem 6 on

Homework # 5.) Suppose $A \in \mathbb{C}^{n\times n}$ is diagonalizable and has a unique eigenvalue λ_1 of maximum modulus, and suppose for simplicity that $\lambda_1 > 0$. Then if $x \in \mathbb{C}^n$ has nonzero component in the direction of the eigenvector corresponding to λ_1 when expanded in terms of the eigenvectors of A, it follows that the sequence $A^k x/\|A^k x\|$ converges to a unit eigenvector corresponding to λ_1. The condition in the Theorem above that X^{-1} has an LU factorization implies that the $(1,1)$ entry of X^{-1} is nonzero, so when e_1 is expanded in terms of the eigenvectors $x_1, \ldots, x_n$ (the columns of X), the x_1-coefficient is nonzero. So $A^{k+1}e_1/\|A^{k+1}e_1\|$ converges to αx_1 for some $\alpha \in \mathbb{C}$ with $|\alpha| = 1$. Let $(\widetilde{q}_k)_1$ denote the first column of $\widetilde{Q}_k$ and $(\widetilde{r}_k)_{11}$ denote the $(1,1)$-entry of $\widetilde{R}_k$; then

$$A^{k+1}e_1 = \widetilde{Q}_k\widetilde{R}_k e_1 = (\widetilde{r}_k)_{11}\widetilde{Q}_k e_1 = (\widetilde{r}_k)_{11}(\widetilde{q}_k)_1,$$

so $(\widetilde{q}_k)_1 \to \alpha x_1$. Since $A_{k+1} = \widetilde{Q}_k^* A \widetilde{Q}_k$, the first column of A_{k+1} converges to $\begin{bmatrix} \lambda_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$.

Further insight into the relationship between the QR algorithm and the power method, inverse power method, and subspace iteration, can be found in this delightful paper: “Understanding the QR Algorithm” by D. S. Watkins (SIAM Review, vol. 24, 1982, pp. 427–440).

Resolvent

Let V be a finite-dimensional vector space and $L \in \mathcal{L}(V)$. If $\zeta \notin \sigma(L)$, then the operator $L - \zeta I$ is invertible, so we can form

$$R(\zeta) = (L - \zeta I)^{-1}$$

(which we sometimes denote by $R(\zeta, L)$). The function $R : \mathbb{C}\backslash\sigma(L) \to \mathcal{L}(V)$ is called the *resolvent* of L. It provides an analytic approach to questions about the spectral theory of L. The set $\mathbb{C}\backslash\sigma(L)$ is called the *resolvent set* of L. Since the inverses of commuting invertible linear transformations also commute, $R(\zeta_1)$ and $R(\zeta_2)$ commute for $\zeta_1, \zeta_2 \in \mathbb{C}\backslash\sigma(L)$. Since a linear transformation commutes with its inverse, it also follows that L commutes with all $R(\zeta)$.

We first want to show that $R(\zeta)$ is a holomorphic function of $\zeta \in \mathbb{C}\backslash\sigma(L)$ with values in $\mathcal{L}(V)$. Recall our earlier discussion of holomorphic functions with values in a Banach space; one of the equivalent definitions was that the function is given by a norm-convergent power series in a neighborhood of each point in the domain. Observe that

$$\begin{aligned} R(\zeta) &= (L - \zeta I)^{-1} \\ &= (L - \zeta_0 I - (\zeta - \zeta_0)I)^{-1} \\ &= (L - \zeta_0 I)^{-1}[I - (\zeta - \zeta_0)R(\zeta_0)]^{-1}. \end{aligned}$$

Let $\|\cdot\|$ be a norm on V, and $\|\cdot\|$ denote the operator norm on $\mathcal{L}(V)$ induced by this norm. If

$$|\zeta - \zeta_0| < \frac{1}{\|R(\zeta_0)\|},$$

then the second inverse above is given by a convergent Neumann series:

$$\begin{aligned} R(\zeta) &= R(\zeta_0)\sum_{k=0}^{\infty} R(\zeta_0)^k(\zeta - \zeta_0)^k \\ &= \sum_{k=0}^{\infty} R(\zeta_0)^{k+1}(\zeta - \zeta_0)^k. \end{aligned}$$

Thus $R(\zeta)$ is given by a convergent power series about any point $\zeta_0 \in \mathbb{C}\backslash\sigma(L)$ (and of course the resolvent set $\mathbb{C}\backslash\sigma(L)$ is open), so $R(\zeta)$ defines an $\mathcal{L}(V)$-valued holomorphic function on the resolvent set $\mathbb{C}\backslash\sigma(L)$ of L. Note that from the series one obtains that

$$\left(\frac{d}{d\zeta}\right)^k R(\zeta)\bigg|_{\zeta_0} = k!R(\zeta_0)^{k+1}.$$

Hence for any $\zeta \in \mathbb{C}\backslash\sigma(L)$,

$$\left(\frac{d}{d\zeta}\right)^k R(\zeta) = k!R(\zeta)^{k+1}.$$

This can be remembered easily by noting that it follows formally by differentiating $R(\zeta) = (L - \zeta)^{-1}$ with respect to ζ, treating L as a parameter.

The argument above showing that $R(\zeta)$ is holomorphic has the advantage that it generalizes to infinite dimensions. Although the following alternate argument only applies in finite dimensions, it gives stronger results in that case. Let $n = \dim V$, choose a basis for V, and represent L by a matrix in $\mathbb{C}^{n\times n}$, which for simplicity we will also call L. Then the matrix of $(L-\zeta I)^{-1}$ can be calculated using Cramer's rule. First observe that

$$\det(L-\zeta I) = (-1)^n p_L(\zeta).$$

Also each of the components of the classical adjoint matrix of $L-\zeta I$ is a polynomial in ζ of degree at most $n-1$. It follows that each component of $(L-\zeta I)^{-1}$ is a rational function of ζ (which vanishes at ∞), so in that sense $R(\zeta)$ is a rational $\mathcal{L}(V)$-valued function. Also each eigenvalue λ_i of L is a pole of $R(\zeta)$ of order at most m_i, the algebraic multiplicity of λ_i. Of course $R(\zeta)$ cannot have a removable singularity at $\zeta = \lambda_i$, for otherwise letting $\zeta \to \lambda_i$ in the equation $(L-\zeta I)R(\zeta) = I$ would show that $L - \lambda_i I$ is invertible, which it is not.

We calculated above the Taylor expansion of $R(\zeta)$ about any point $\zeta_0 \in \mathbb{C}\backslash\sigma(L)$. It is also useful to calculate the Laurent expansion about the poles. Recall the spectral decomposition of L: if $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of L with algebraic multiplicities $m_1, \ldots, m_k$, and

$$\widetilde{E}_i = \mathcal{N}((L-\lambda_i I)^{m_i})$$

are the generalized eigenspaces, then

$$V = \bigoplus_{i=1}^{k} \widetilde{E}_i,$$

and each $\widetilde{E}_i$ is invariant under L. Let $P_1, \ldots, P_k$ be the associated projections, so that

$$I = \sum_{i=1}^{k} P_i.$$

Let $N_1, \ldots, N_k$ be the associated nilpotent transformations. We may regard each N_i as an element of $\mathcal{L}(V)$ (in which case

$$N_i = P_i N P_i$$

where

$$N = N_1 + \cdots + N_k,$$

so

$$N_i[\widetilde{E}_i] \subset \widetilde{E}_i \text{ and } N_i[\widetilde{E}_j] = 0 \text{ for } j \neq i),$$

or we may regard N_i as its restriction to $\widetilde{E}_i$ with

$$N_i : \widetilde{E}_i \to \widetilde{E}_i.$$

Now

$$L = \sum_{i=1}^{k} (\lambda_i P_i + N_i),$$

so

$$L - \zeta I = \sum_{i=1}^{k} [(\lambda_i - \zeta)P_i + N_i].$$

Clearly to invert $L - \zeta I$, it suffices to invert each $(\lambda_i - \zeta)P_i + N_i$ on $\widetilde{E}_i$. But on $\widetilde{E}_i$,

$$(\lambda_i - \zeta)P_i + N_i = (\lambda_i - \zeta)[I - (\zeta - \lambda_i)^{-1} N_i].$$

For a nilpotent operator N with $N^m = 0$,

$$(I - N)^{-1} = I + N + N^2 + \cdots + N^{m-1}.$$

This is a special case of a Neumann series which converges since it terminates. Thus

$$\left([(\lambda_i - \zeta)P_i + N_i] \Big|_{\widetilde{E}_i} \right)^{-1} = (\lambda_i - \zeta)^{-1} \sum_{\ell=0}^{m_i - 1} (\zeta - \lambda_i)^{-\ell} N_i^{\ell} = - \sum_{\ell=0}^{m_i - 1} (\zeta - \lambda_i)^{-\ell-1} N_i^{\ell}.$$

The direct sum of these operators gives $(L - \zeta I)^{-1}$, so we obtain

$$R(\zeta) = - \sum_{i=1}^{k} \left[(\zeta - \lambda_i)^{-1} P_i + \sum_{\ell=1}^{m_i - 1} (\zeta - \lambda_i)^{-\ell-1} N_i^{\ell} \right].$$

This result is called the *partial fractions decomposition* of the resolvent. Recall that any rational function $q(\zeta)/p(\zeta)$ with $\deg q < \deg p$ has a unique partial fractions decomposition of the form

$$\sum_{i=1}^{k} \left[\sum_{j=1}^{m_i} \frac{a_{ij}}{(\zeta - r_i)^j} \right]$$

where $a_{ij} \in \mathbb{C}$ and

$$p(\zeta) = \prod_{i=1}^{k} (\zeta - r_i)^{m_i}$$

is the factorization of p (normalized to be monic, r_i distinct). The above is such a decomposition for $R(\zeta)$.

Observe that the partial fractions decomposition gives the Laurent expansion of $R(\zeta)$ about all of its poles all at once: about $\zeta = \lambda_i$ the holomorphic part of $R(\zeta)$ is the sum over all other eigenvalues. For the coefficients of $(\zeta - \lambda_i)^{-1}$ and $(\zeta - \lambda_i)^{-2}$ we have

$$\underset{\zeta=\lambda_i}{\mathcal{R}es}[R(\zeta)] = -P_i \qquad \text{and} \qquad \underset{\zeta=\lambda_i}{\mathcal{R}es}[(\zeta - \lambda_i)R(\zeta)] = -N_i.$$

So the full spectral decomposition of L is encoded in $R(\zeta)$. It is in fact possible to give a complete treatment of the spectral problem — including a proof of the spectral decomposition — based purely on a study of the resolvent and its properties. Beginning with the fact that $R(\zeta)$ has poles at the λ_i's, one can show that for each i, $-\underset{\zeta=\lambda_i}{\mathcal{R}es}[R(\zeta)]$ is a projection and $-\underset{\zeta=\lambda_i}{\mathcal{R}es}[(\zeta - \lambda_i)R(\zeta)]$ is nilpotent, that the sum of the projections is the identity, etc. See Kato for such a treatment.

The special case of the partial fractions decomposition in which L is diagonalizable is particularly easy to derive and remember. If L is diagonalizable then each $\widetilde{E}_i = E_{\lambda_i}$ is the eigenspace and each $N_i = 0$. If $v \in V$, we may uniquely decompose

$$v = \sum_{i=1}^{k} v_i \qquad \text{where} \qquad v_i = P_i v \in E_{\lambda_i}.$$

Then $Lv = \sum_{i=1}^{k} \lambda_i v_i$, so

$$(L - \zeta I)v = \sum_{i=1}^{k} (\lambda_i - \zeta) v_i,$$

so clearly

$$R(\zeta)v = \sum_{i=1}^{k} (\lambda_i - \zeta)^{-1} P_i v,$$

and thus

$$R(\zeta) = \sum_{i=1}^{k} (\lambda_i - \zeta)^{-1} P_i.$$

The powers $(\lambda_i - \zeta)^{-1}$ arise from inverting $(\lambda_i - \zeta)I$ on each E_{λ_i}.

We discuss briefly two applications of the resolvent — each of these has many ramifications which we do not have time to investigate fully. Both applications involve contour integration of operator-valued functions. If $M(\zeta)$ is a continuous function of ζ with values in $\mathcal{L}(V)$ and γ is a C^1 contour in $\mathbb{C}$, we may form $\int_\gamma M(\zeta) d\zeta \in \mathcal{L}(V)$. This can be defined by choosing a fixed basis for V, representing $M(\zeta)$ as matrices, and integrating componentwise, or as a norm-convergent limit of Riemann sums of the parameterized integrals. By considering the componentwise definition it is clear that the usual results in complex analysis automatically extend to the operator-valued case, for example if $M(\zeta)$ is holomorphic in a neighborhood of the closure of a region bounded by a closed curve γ except for poles $\zeta_1, \ldots, \zeta_k$, then $\frac{1}{2\pi i} \int_\gamma M(\zeta) d\zeta = \sum_{i=1}^{k} \mathcal{R}es(M, \zeta_i)$.

Perturbation of Eigenvalues and Eigenvectors

One major application of resolvents is the study of perturbation theory of eigenvalues and eigenvectors. We sketch how resolvents can be used to study continuity properties of eigenvectors. Suppose $A_t \in \mathbb{C}^{n \times n}$ is a family of matrices depending continuously on a parameter t. (In our examples, the domain of t will be a subset of $\mathbb{R}$, but in general the domain of t could be any metric space.) It is a fact that the eigenvalues of A_t depend continuously on t, but this statement must be properly formulated since the eigenvalues are only determined up to order. Since the eigenvalues are the roots of the characteristic polynomial of A_t, and the coefficients of the characteristic polynomial depend continuously on t, (since, by norm equivalence, the entries of A_t depend continuously on t), it suffices to see that the roots of a monic polynomial (of fixed degree) depend continuously on its coefficients. Consider first the case of a simple root: suppose $z_0 \in \mathbb{C}$ is a simple root of a polynomial p_0. We may choose a closed disk about z_0 containing no other zero of p_0; on the boundary γ of this disk p_0 does not vanish, so all polynomials with coefficients sufficiently close to those of p_0 also do not vanish on γ. So for such p, $p'(z)/p(z)$ is continuous on γ, and by the argument principle,

$$\frac{1}{2\pi i}\int_\gamma \frac{p'(z)}{p(z)}dz$$

is the number of zeroes of p (including multiplicities) in the disk. For p_0, we get 1. Since $p \neq 0$ on γ, $\frac{p'}{p}$ varies continuously with the coefficients of p, so

$$\frac{1}{2\pi i}\int_\gamma \frac{p'(z)}{p(z)}dz$$

also varies continuously with the coefficients of P. As it is integer-valued we conclude that it must be the constant 1, so all nearby polynomials have exactly one zero in the disk. Now the residue theorem gives that

$$\frac{1}{2\pi i}\int_\gamma \frac{p'(z)}{p(z)}z\,dz = z_p$$

is the unique root of p in the disk. As the left hand side varies continuously with p, it follows that its simple root z_p does too.

One can also obtain information near multiple zeroes using such arguments. If z_0 is a root of p_0 of multiplicity $m > 1$, then it follows as above that in *any* sufficiently small disk about z_0, any polynomial p sufficiently close to p_0 (where "sufficiently close" depends on the radius of the disk) will have exactly m zeroes in that disk (counting multiplicities). This is one sense in which it can be said that the eigenvalues depend continuously on the coefficients. There are stronger senses as well.

However, eigenvectors do not generally depend continuously on parameters. Consider for example the family given by

$$A_t = \begin{bmatrix} t & 0 \\ 0 & -t \end{bmatrix} \text{ for } t \geq 0 \text{ and } A_t = \begin{bmatrix} 0 & t \\ t & 0 \end{bmatrix} \text{ for } t \leq 0.$$

For each t, the eigenvalues of A_t are t, $-t$. Clearly A_t is diagonalizable for all t. But it is impossible to find a continuous function $v : \mathbb{R} \to \mathbb{R}^2$ such that $v(t)$ is an eigenvector of A_t for each t. For $t > 0$, the eigenvectors of A_t are multiples of

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

while for $t < 0$ they are multiples of

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ and } \begin{bmatrix} -1 \\ 1 \end{bmatrix};$$

clearly it is impossible to join up such multiples continuously by a vector $v(t)$ which doesn't vanish at $t = 0$. (Note that a similar C^∞ example can be constructed: let

$$A_t = \begin{bmatrix} \varphi(t) & 0 \\ 0 & -\varphi(t) \end{bmatrix} \text{ for } t \geq 0, \text{ and } A_t = \begin{bmatrix} 0 & \varphi(t) \\ \varphi(t) & 0 \end{bmatrix} \text{ for } t \leq 0,$$

where

$$\varphi(t) = e^{-1/|t|} \text{ for } t \neq 0 \text{ and } \varphi(0) = 0\ .)$$

In the example above, A_0 has an eigenvalue of multiplicity 2. We show, using the resolvent, that if an eigenvalue of A_0 has algebraic multiplicity 1, then the corresponding eigenvector can be chosen to depend continuously on t, at least in a neighborhood of $t = 0$. Suppose λ_0 is an eigenvalue of A_0 of multiplicity 1. We know from the above that A_t has a unique eigenvalue λ_t near λ_0 for t near 0; moreover λ_t is simple and depends continuously on t for t near 0. If γ is a circle about λ_0 as above, and we set

$$R_t(\zeta) = (A_t - \zeta I)^{-1},$$

then

$$-\frac{1}{2\pi i}\int_\gamma R_t(\zeta)d\zeta = -\mathcal{R}es_{\zeta=\lambda_t}R_t(\zeta) = P_t,$$

where P_t is the spectral projection onto the 1-dimensional eigenspace of A_t corresponding to λ_t. Observe that for t near 0 and $\zeta \in \gamma$, $A_t - \zeta I$ is invertible, and it is clear that $R_t(\zeta)$ depends continuously on t (actually, uniformly in $\zeta \in \gamma$). So P_t depends continuously on t for t near 0. We can obtain a continuously-varying eigenvector by projecting a fixed vector: let v_0 be a unit eigenvector for A_0 corresponding to λ_0, and set

$$v_t = P_t v_0 = -\frac{1}{2\pi i}\int_\gamma R_t(\zeta)v_0 d\zeta\ .$$

The right hand side varies continuously with t, so v_t does too and

$$v_t\Big|_{t=0} = v_0.$$

Hence $v_t \neq 0$ for t near 0, and since v_t is in the range of P_t, v_t is an eigenvector of A_t corresponding to λ_t, as desired.

Remark. These ideas can show that if A_t is a C^k function of t, i.e. each $a_{ij}(t)$ has k continuous derivatives, then also λ_t and v_t are C^k functions of t.

This approach using the resolvent indicates that it is possible to obtain something continuous even when there are multiple eigenvalues. As long as no eigenvalues of A_t hit γ, the expression $-\frac{1}{2\pi i}\int_\gamma R_t(\zeta)d\zeta$ depends continuously on t. By the Residue Theorem, for each t this is the sum of the projections onto all generalized eigenspaces corresponding to eigenvalues in the disk enclosed by γ, so this sum of projections is always continuous.

Spectral Radius

We now show how the resolvent can be used to give a formula for the spectral radius of an operator which does not require knowing the spectrum explicitly; this is sometimes useful. As before, let $L \in \mathcal{L}(V)$ where V is finite dimensional. Then

$$R(\zeta) = (L - \zeta I)^{-1}$$

is a rational $\mathcal{L}(V)$ - valued function of ζ with poles at the eigenvalues of L. In fact, from Cramers rule we saw that

$$R(\zeta) = \frac{Q(\zeta)}{p_L(\zeta)},$$

where $Q(\zeta)$ is an $\mathcal{L}(V)$ - valued polynomial in ζ of degree $\leq n-1$ and p_L is the characteristic polynomial of L. Since $\deg p_L > \deg Q$, it follows that $R(\zeta)$ is holomorphic at ∞ and vanishes there; i.e., for large $|\zeta|$, $R(\zeta)$ is given by a convergent power series in $\frac{1}{\zeta}$ with zero constant term. We can identify the coefficients in this series (which are in $\mathcal{L}(V)$) using Neumann series: for $|\zeta|$ sufficiently large,

$$R(\zeta) = -\zeta^{-1}(I - \zeta^{-1}L)^{-1} = -\zeta^{-1}\sum_{k=0}^{\infty}\zeta^{-k}L^k = -\sum_{k=0}^{\infty}L^k\zeta^{-k-1}\ .$$

The coefficients in the expansion are (minus) the powers of L. For any submultiplicative norm on $\mathcal{L}(V)$, this series converges for $\|\zeta^{-1}L\| < 1$, i.e., for $|\zeta| > \|L\|$.

Recall from complex analysis that the radius of convergence r of a power series

$$\sum_{k=0}^{\infty} a_k z^k$$

can be characterized in two ways: first, as the radius of the largest open disk about the origin in which the function defined by the series has a holomorphic extension, and second directly in terms of the coefficients by the formula

$$\frac{1}{r} = \overline{\lim}_{k\to\infty}|a_k|^{\frac{1}{k}}.$$

These characterizations also carry over to operator-valued series

$$\sum_{k=0}^{\infty} A_k z^k \qquad \text{(where } A_k \in \mathcal{L}(V)\text{)}.$$

Such a series also has a radius of convergence r, and both characterizations generalize: the first is unchanged; the second becomes

$$\frac{1}{r} = \overline{\lim}_{k\to\infty}\|A_k\|^{\frac{1}{k}}.$$

Note that the expression

$$\overline{\lim}_{k\to\infty}\|A_k\|^{\frac{1}{k}}$$

is independent of the norm on $\mathcal{L}(V)$ by the Norm Equivalence Theorem since $\mathcal{L}(V)$ is finite-dimensional. These characterizations in the operator-valued case can be obtained by considering the series for each component in any matrix realization.

Apply these two characterizations to the power series

$$\sum_{k=0}^{\infty} L^k\zeta^{-k} \quad \text{in} \quad \zeta^{-1}$$

for $-\zeta R(\zeta)$. We know that $R(\zeta)$ is holomorphic in $|\zeta| > \rho(L)$ (including at ∞) and that $R(\zeta)$ has poles at each eigenvalue of L, so the series converges for $|\zeta| > \rho(L)$, but in no larger disk about ∞. The second formula gives

$$\{\zeta : |\zeta| > \overline{\lim}_{k\to\infty}\|L^k\|^{\frac{1}{k}}\}$$

as the largest disk of convergence, and thus

$$\rho(L) = \overline{\lim}_{k\to\infty}\|L^k\|^{\frac{1}{k}}.$$

Lemma. If $L \in \mathcal{L}(V)$ has eigenvalues $\lambda_1, \ldots, \lambda_n$, repeated with multiplicities, then the eigenvalues of L^ℓ are $\lambda_1^\ell, \ldots, \lambda_n^\ell$.

Remark. This is a special case of the Spectral Mapping Theorem which we will study soon.

Proof. If L has spectral decomposition

$$L = \sum_{i=1}^{k}(\mu_i P_i + N_i)$$

where $\mu_1, \cdots, \mu_k$ are the distinct eigenvalues of L, then

$$L^\ell = \sum_{i=1}^{k}(\mu_i^\ell P_i + N_i'),$$

where

$$N_i' = \sum_{j=1}^{\ell}\left(\begin{array}{c}\ell\\ j\end{array}\right)\mu_i^{\ell-j}N_i^j$$

is nilpotent. The result follows from the uniqueness of the spectral decomposition. □

Remark. An alternate formulation of the proof goes as follows. By the Schur Triangularization Theorem, or by Jordan form, there is a basis of V for which the matrix of L is upper triangular. The diagonal elements of a power of a triangular matrix are that power of the diagonal elements of the matrix. The result follows.

Proposition. If $\dim V < \infty$, $L \in \mathcal{L}(V)$, and $\|\cdot\|$ is any norm on $\mathcal{L}(V)$, then

$$\rho(L) = \lim_{k\to\infty}\|L^k\|^{\frac{1}{k}}.$$

Proof. We have already shown that

$$\rho(L) = \overline{\lim}_{k\to\infty}\|L^k\|^{\frac{1}{k}},$$

so we just have to show the limit exists. By norm equivalence, the limit exists in one norm iff it exists in every norm, so it suffices to show the limit exists if $\|\cdot\|$ is submultiplicative. Let $\|\cdot\|$ be submultiplicative. Then $\rho(L) \leq \|L\|$. By the lemma,

$$\rho(L^k) = \rho(L)^k \quad \text{so} \quad \rho(L)^k = \rho(L^k) \leq \|L^k\|.$$

Thus

$$\rho(L) \leq \varliminf_{k\to\infty} \|L^k\|^{\frac{1}{k}} \leq \overline{\lim}_{k\to\infty}\|L^k\|^{\frac{1}{k}} = \rho(L),$$

so the limit exists and is $\rho(L)$. □

This formula for the spectral radius $\rho(L)$ of L allows us to extend the class of operators in $\mathcal{L}(V)$ for which we can guarantee that certain series converge. Recall that if $\varphi(z) = \sum_{k=0}^{\infty} a_k z^k$ is holomorphic for $|z| < r$ and $L \in \mathcal{L}(V)$ satisfies $\|L\| < r$ for some submultiplicative norm, then $\varphi(L)$ can be defined as the limit of the norm-convergent series $\sum_{k=0}^{\infty} a_k L^k$. In fact, this series converges under the (apparently weaker) assumption that $\rho(L) < r$: choose $\epsilon > 0$ so that $\rho(L) + \epsilon < r$; for k sufficiently large, $\|L^k\|^{\frac{1}{k}} \leq \rho(L) + \epsilon$, so

$$\sum_{k \text{ large}} \|a_k L^k\| \leq \sum |a_k|(\rho(L)+\epsilon)^k < \infty.$$

For example, the Neumann series

$$(I-L)^{-1} = \sum_{k=0}^{\infty} L^k$$

converges whenever $\rho(L) < 1$. It may happen that $\rho(L) < 1$ and yet $\|L\| > 1$ for certain natural norms (like the operator norms induced by the ℓ^p norms on $\mathbb{C}^n$, $1 \leq p \leq \infty$). An extreme case occurs when L is nilpotent, so $\rho(L) = 0$, but $\|L\|$ can be large (e.g. the matrix

$$\left[\begin{array}{cc} 0 & 10^{10} \\ 0 & 0 \end{array}\right]\Bigg);$$

in this case, of course, any series $\sum_{k=0}^{\infty} a_k L^k$ converges since it terminates.

The following question has arisen a couple of times in the discussion of the spectral radius: given a *fixed* $L \in \mathcal{L}(V)$, what is the infimum of $\|L\|$ as $\|\cdot\|$ ranges over all submultiplicative norms on $\mathcal{L}(V)$? What if we only consider operator norms on $\mathcal{L}(V)$ induced by norms on V? How about restricting further to operator norms on $\mathcal{L}(V)$ induced by inner products on V? We know that $\rho(L) \leq \|L\|$ in these situations. It turns out that the infimum in each of these situations is actually $\rho(L)$.

Proposition. Given $A \in \mathbb{C}^{n\times n}$ and $\epsilon > 0$, there exists a norm $\|\cdot\|$ on $\mathbb{C}^n$ for which, in the operator norm induced by $\|\cdot\|$, we have $\|A\| \leq \rho(A) + \epsilon$.

Caution: The norm depends on A and ϵ.

Proof. Choose an invertible matrix $S \in \mathbb{C}^{n\times n}$ for which

$$J = S^{-1}AS$$

is in Jordan form. Write $J = \Lambda + Z$, where

$$\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)$$

is the diagonal part of J and Z is a matrix with only zero entries except possibly for some one(s) on the first superdiagonal ($i = j + 1$). Let

$$D = \mathrm{diag}\,(1, \epsilon, \epsilon^2, \ldots, \epsilon^{n-1}).$$

Then

$$D^{-1}JD = \Lambda + \epsilon Z.$$

Fix any p with $1 \leq p \leq \infty$. Then in the operator norm $|||\cdot|||_p$ on $\mathbb{C}^{n\times n}$ induced by the ℓ^p-norm $\|\cdot\|_p$ on $\mathbb{C}^n$,

$$|||\Lambda|||_p = \max\{|\lambda_j| : 1 \leq j \leq n\} = \rho(A)$$

and $|||Z|||_p \leq 1$, so

$$|||\Lambda + \epsilon Z|||_p \leq \rho(A) + \epsilon.$$

Define $\|\cdot\|$ on $\mathbb{C}^n$ by

$$\|x\| = \|D^{-1}S^{-1}x\|_p.$$

Then

$$\begin{aligned} \|A\| &= \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{y\neq 0} \frac{\|ASDy\|}{\|SDy\|} = \sup_{y\neq 0} \frac{\|D^{-1}S^{-1}ASDy\|_p}{\|y\|_p} \\ &= |||\Lambda + \epsilon Z|||_p \leq \rho(A) + \epsilon \ . \end{aligned}$$

□

Exercise: Show that we can choose an inner product on $\mathbb{C}^n$ which induces such a norm.

Remarks:

(1) This proposition is easily extended to $L \in \mathcal{L}(V)$ for $\dim V < \infty$.

(2) This proposition gives another proof that if $\varphi(z) = \sum_{k=0}^{\infty} a_k z^k$ is holomorphic for $|z| < r$ and $L \in \mathcal{L}(V)$ satisfies $\rho(L) < r$, then the series $\sum_{k=0}^{\infty} a_k L^k$ converges: choose $\epsilon > 0$ so that $\rho(L) + \epsilon < r$, and then choose a submultiplicative norm on $\mathcal{L}(V)$ for which $\|L\| \leq \rho(L) + \epsilon$; then $\|L\| < r$ and the series converges.

(3) One can use the Schur Triangularization Theorem instead of Jordan form in the proof; see Lemma 5.6.10 in H-J.

We conclude this discussion of the spectral radius with two corollaries of the formula

$$\rho(L) = \lim_{k\to\infty} \|L^k\|^{\frac{1}{k}}$$

for $L \in \mathcal{L}(V)$ with $\dim V < \infty$.

Corollary. $\rho(L) < 1$ iff $L^k \to 0$.

Proof. By norm equivalence, we may use a submultiplicative norm on $\mathcal{L}(V)$. If $\rho(L) < 1$, choose $\epsilon > 0$ with $\rho(L) + \epsilon < 1$. For large k, $\|L^k\| \leq (\rho(L) + \epsilon)^k \to 0$ as $k \to \infty$. Conversely, if $L^k \to 0$, then $\exists\, k \geq 1$ with $\|L^k\| < 1$, so $\rho(L^k) < 1$, so by the lemma, the eigenvalues $\lambda_1, \ldots, \lambda_n$ of L all satisfy $|\lambda_j^k| < 1$ and thus $\rho(L) < 1$. □

Corollary. $\rho(L) < 1$ iff there is a submultiplicative norm on $\mathcal{L}(V)$ and an integer $k \geq 1$ such that $\|L^k\| < 1$.

Functional Calculus

Our last application of resolvents is to define functions of an operator. We do this using a method providing good operational properties, so this is called a functional "calculus."

Let $L \in \mathcal{L}(V)$ and suppose that φ is holomorphic in a neighborhood of the closure of a bounded open set $\Delta \subset \mathbb{C}$ with C^1 boundary satisfying $\sigma(L) \subset \Delta$. For example, Δ could be a large disk containing all the eigenvalues of L, or the union of small disks about each eigenvalue, or an appropriate annulus centered at $\{0\}$ if L is invertible. Give the curve $\partial\Delta$ the orientation induced by the boundary of Δ (i.e. the winding number $n(\partial\Delta, z) = 1$ for $z \in \Delta$ and $= 0$ for $z \in \mathbb{C}\backslash\bar{\Delta}$.) We define $\varphi(L)$ by requiring that the Cauchy integral formula for φ should hold.

Definition.

$$\varphi(L) = -\frac{1}{2\pi i}\int_{\partial\Delta} \varphi(\zeta)R(\zeta)d\zeta = \frac{1}{2\pi i}\int_{\partial\Delta} \varphi(\zeta)(\zeta I - L)^{-1}d\zeta.$$

We first observe that the definition of $\varphi(L)$ is independent of the choice of Δ. In fact, since $\varphi(\zeta)R(\zeta)$ is holomorphic except for poles at the eigenvalues of L, we have by the residue theorem that

$$\varphi(L) = -\sum_{i=1}^{k} \mathcal{R}es_{\zeta=\lambda_i}[\varphi(\zeta)R(\zeta)],$$

which is clearly independent of the choice of Δ. In the special case where $\Delta_1 \subset \Delta_2$, it follows from Cauchy's theorem that

$$\int_{\partial\Delta_2} \varphi(\zeta)R(\zeta)d\zeta - \int_{\partial\Delta_1} \varphi(\zeta)R(\zeta)d\zeta = \int_{\partial(\Delta_2\backslash\bar{\Delta}_1)} \varphi(\zeta)R(\zeta)d\zeta = 0$$

since $\varphi(\zeta)R(\zeta)$ is holomorphic in $\Delta_2\backslash\bar{\Delta}_1$. This argument can be generalized as well.

Next, we show that this definition of $\varphi(L)$ agrees with previous definitions. For example, suppose φ is the constant function 1. Then the residue theorem gives

$$\varphi(L) = -\sum_{i=1}^{k} \mathcal{R}es_{\zeta=\lambda_i}R(\zeta) = \sum_{i=1}^{k} P_i = I.$$

If $\varphi(\zeta) = \zeta^n$ for an integer $n > 0$, then take Δ to be a large disk containing $\sigma(L)$ with boundary γ, so

$$\begin{aligned}
\varphi(L) &= \frac{1}{2\pi i}\int_\gamma \zeta^n(\zeta I - L)^{-1}d\zeta \\
&= \frac{1}{2\pi i}\int_\gamma [(\zeta I - L) + L]^n(\zeta I - L)^{-1}d\zeta \\
&= \frac{1}{2\pi i}\int_\gamma \sum_{j=0}^{n} \begin{pmatrix} n \\ j \end{pmatrix} L^j(\zeta I - L)^{n-j-1}d\zeta \\
&= \frac{1}{2\pi i}\sum_{j=0}^{n} \begin{pmatrix} n \\ j \end{pmatrix} L^j \int_\gamma (\zeta I - L)^{n-j-1}d\zeta\ .
\end{aligned}$$

For $j < n$, the integrand is holomorphic in Δ, so by the Cauchy theorem

$$\int_\gamma (\zeta I - L)^{n-j-1}d\zeta = 0.$$

For $j = n$, we obtain

$$\frac{1}{2\pi i}\int_\gamma (\zeta I - L)^{-1}d\zeta = I$$

as above, so $\varphi(L) = L^n$ as desired. It follows that this new definition of $\varphi(L)$ agrees with the usual definition of $\varphi(L)$ when φ is a polynomial.

Consider next the case in which

$$\varphi(\zeta) = \sum_{k=0}^{\infty} a_k\zeta^k,$$

where the series converges for $|\zeta| < r$. We have seen that if $\rho(L) < r$, then the series $\sum_{k=0}^{\infty} a_k L^k$ converges in norm. We will show that this definition of $\varphi(L)$ (via the series) agrees with our new definition of $\varphi(L)$ (via contour integration). Choose

$$\Delta \subset \{\zeta : |\zeta| < r\} \text{ with } \sigma(L) \subset \Delta \text{ and } \gamma \equiv \partial\Delta \subset \{\zeta : |\zeta| < r\}.$$

We want to show that

$$\frac{-1}{2\pi i}\int_\gamma \varphi(\zeta)R(\zeta)d\zeta = \sum_{k=0}^{\infty} a_k L^k.$$

Set

$$\varphi_N(\zeta) = \sum_{k=0}^{N} a_k\zeta^k.$$

Then $\varphi_N \to \varphi$ uniformly on compact subsets of $\{\zeta : |\zeta| < r\}$; in particular $\varphi_N \to \varphi$ uniformly on γ. If $A(t)$ is a continuous $\mathcal{L}(V)$-valued function of $t \in [a, b]$, then for any norm on $\mathcal{L}(V)$,

$$\left\| \int_a^b A(t)dt \right\| \le \int_a^b \|A(t)\|dt$$

(this follows from the triangle inequality applied to the Riemann sums approximating the integrals upon taking limits). So

$$\left\| \int_\gamma (\varphi(\zeta) - \varphi_N(\zeta)) R(\zeta) d\zeta \right\| \leq \int_\gamma |\varphi(\zeta) - \varphi_N(\zeta)| \cdot \|R(\zeta)\| d|\zeta|.$$

Since $\|R(\zeta)\|$ is bounded on γ, it follows that

$$\lim_{N\to\infty} \int_\gamma \varphi_N(\zeta) R(\zeta) d\zeta = \int_\gamma \varphi(\zeta) R(\zeta) d\zeta$$

in norm. But φ_N is a polynomial, so

$$-\frac{1}{2\pi i} \int_\gamma \varphi_N(\zeta) R(\zeta) d\zeta = \varphi_N(L)$$

as above. Thus

$$-\frac{1}{2\pi i} \int_\gamma \varphi(\zeta) R(\zeta) d\zeta = \lim_{N\to\infty} \left(-\frac{1}{2\pi i} \int_\gamma \varphi_N(\zeta) R(\zeta) d\zeta \right) = \lim_{N\to\infty} \varphi_N(L) = \sum_{k=0}^\infty a_k L^k,$$

and the two definitions of $\varphi(L)$ agree.

Operational Properties

Lemma. (The First Resolvent Equation)

If $L \in \mathcal{L}(V)$, $\zeta_1, \zeta_2 \notin \sigma(L)$, and $\zeta_1 \neq \zeta_2$, then

$$R(\zeta_1) \circ R(\zeta_2) = \frac{R(\zeta_1) - R(\zeta_2)}{\zeta_1 - \zeta_2}.$$

Proof.

$$R(\zeta_1) - R(\zeta_2) = R(\zeta_1)(L - \zeta_2 I) R(\zeta_2) - R(\zeta_1)(L - \zeta_1 I) R(\zeta_2) = (\zeta_1 - \zeta_2) R(\zeta_1) R(\zeta_2).$$

□

Proposition. Suppose $L \in \mathcal{L}(V)$ and φ_1 and φ_2 are both holomorphic in a neighborhood of $\sigma(L)$. Then

(a) $(a_1\varphi_1 + a_2\varphi_2)(L) = a_1\varphi_1(L) + a_2\varphi_2(L)$, and

(b) $(\varphi_1\varphi_2)(L) = \varphi_1(L) \circ \varphi_2(L)$.

Proof. (a) follows immediately from the linearity of contour integration. By the lemma,

$$\begin{aligned}\varphi_1(L)\circ\varphi_2(L) &= \frac{1}{(2\pi i)^2}\int_{\gamma_1}\varphi_1(\zeta_1)R(\zeta_1)d\zeta_1\circ\int_{\gamma_2}\varphi_2(\zeta_2)R(\zeta_2)d\zeta_2\\ &= \frac{1}{(2\pi i)^2}\int_{\gamma_1}\int_{\gamma_2}\varphi_1(\zeta_1)\varphi_2(\zeta_2)R(\zeta_1)\circ R(\zeta_2)d\zeta_2 d\zeta_1\\ &= \frac{1}{(2\pi i)^2}\int_{\gamma_1}\int_{\gamma_2}\varphi_1(\zeta_1)\varphi_2(\zeta_2)\frac{R(\zeta_1)-R(\zeta_2)}{\zeta_1-\zeta_2}d\zeta_2 d\zeta_1.\end{aligned}$$

Thus far, γ_1 and γ_2 could be any curves encircling $\sigma(L)$; the curves could cross and there is no problem since

$$\frac{R(\zeta_1)-R(\zeta_2)}{\zeta_1-\zeta_2}$$

extends to $\zeta_1=\zeta_2$. However, we want to split up the $R(\zeta_1)$ and $R(\zeta_2)$ pieces, so we need to make sure the curves don't cross. For definiteness, let γ_1 be the union of small circles around each eigenvalue of L, and let γ_2 be the union of slightly larger circles. Then

$$\begin{aligned}\varphi_1(L)\circ\varphi_2(L) = \frac{1}{(2\pi i)^2}\Bigg[&\int_{\gamma_1}\varphi_1(\zeta_1)R(\zeta_1)\int_{\gamma_2}\frac{\varphi_2(\zeta_2)}{\zeta_1-\zeta_2}d\zeta_2 d\zeta_1\\ &-\int_{\gamma_2}\varphi_2(\zeta_2)R(\zeta_2)\int_{\gamma_1}\frac{\varphi_1(\zeta_1)}{\zeta_1-\zeta_2}d\zeta_1 d\zeta_2\Bigg].\end{aligned}$$

Since ζ_1 is inside γ_2 but ζ_2 is outside γ_1,

$$\frac{1}{2\pi i}\int_{\gamma_2}\frac{\varphi_2(\zeta_2)}{\zeta_2-\zeta_1}d\zeta_2=\varphi_2(\zeta_1)\quad\text{and}\quad\frac{1}{2\pi i}\int_{\gamma_1}\frac{\varphi_1(\zeta_1)}{\zeta_1-\zeta_2}d\zeta_1=0,$$

so

$$\varphi_1(L)\circ\varphi_2(L)=-\frac{1}{2\pi i}\int_{\gamma_1}\varphi_1(\zeta_1)\varphi_2(\zeta_1)R(\zeta_1)d\zeta_1=(\varphi_1\varphi_2)(L),$$

as desired. □

Remark. Since $(\varphi_1\varphi_2)(\zeta)=(\varphi_2\varphi_1)(\zeta)$, (b) implies that $\varphi_1(L)$ and $\varphi_2(L)$ always commute.

Example. Suppose $L\in\mathcal{L}(V)$ is invertible and $\varphi(\zeta)=\frac{1}{\zeta}$. Since $\sigma(L)\subset\mathbb{C}\backslash\{0\}$ and φ is holomorphic on $\mathbb{C}\backslash\{0\}$, $\varphi(L)$ is defined. Since $\zeta\cdot\frac{1}{\zeta}=\frac{1}{\zeta}\cdot\zeta=1$, $L\varphi(L)=\varphi(L)L=I$. Thus $\varphi(L)=L^{-1}$, as expected.

Similarly, one can show that if

$$\varphi(\zeta)=\frac{p(\zeta)}{q(\zeta)}$$

is a rational function (p,q are polynomials) and $\sigma(L)\subset\{\zeta:q(\zeta)\neq 0\}$, then

$$\varphi(L)=p(L)q(L)^{-1},$$

as expected.

To study our last operational property (composition), we need to identify $\sigma(\varphi(L))$.

The Spectral Mapping Theorem

Suppose $L \in \mathcal{L}(V)$ and φ is holomorphic in a neighborhood of $\sigma(L)$ (so $\varphi(L)$ is well-defined). Then

$$\sigma(\varphi(L)) = \varphi(\sigma(L)) \quad \text{including multiplicities,}$$

i.e., if $\mu_1, \ldots, \mu_n$ are the eigenvalues of L counting multiplicities, then $\varphi(\mu_1), \ldots, \varphi(\mu_n)$ are the eigenvalues of $\varphi(L)$ counting multiplicities.

Proof. Let $\lambda_1, \ldots, \lambda_k$ be the distinct eigenvalues of L, with algebraic multiplicities

$$m_1, \ldots, m_k,$$

respectively. By the residue theorem,

$$\begin{aligned} \varphi(L) &= -\frac{1}{2\pi i}\int_{\partial\Delta} \varphi(\zeta)R(\zeta)d\zeta \\ &= -\sum_{i=1}^{k} \mathcal{R}es_{\zeta=\lambda_i}[\varphi(\zeta)R(\zeta)]. \end{aligned}$$

By the partial fractions decomposition of the resolvent,

$$-R(\zeta) = \sum_{i=1}^{k}\left(\frac{P_i}{\zeta-\lambda_i} + \sum_{\ell=1}^{m_i-1}(\zeta-\lambda_i)^{-\ell-1}N_i^{\ell}\right).$$

It follows that

$$\begin{aligned} -\mathcal{R}es_{\zeta=\lambda_i}\varphi(\zeta)R(\zeta) &= \varphi(\lambda_i)P_i + \sum_{\ell=1}^{m_i-1}\mathcal{R}es_{\zeta=\lambda_i}[\varphi(\zeta)(\zeta-\lambda_i)^{-\ell-1}]N_i^{\ell} \\ &= \varphi(\lambda_i)P_i + \sum_{\ell=1}^{m_i-1}\frac{1}{\ell!}\varphi^{(\ell)}(\lambda_i)N_i^{\ell}. \end{aligned}$$

Thus

$$(*) \qquad \varphi(L) = \sum_{i=1}^{k}[\varphi(\lambda_i)P_i + \sum_{\ell=1}^{m_i-1}\frac{1}{\ell!}\varphi^{(\ell)}(\lambda_i)N_i^{\ell}]$$

This is an explicit formula for $\varphi(L)$ in terms of the spectral decomposition of L and the values of φ and its derivatives at the eigenvalues of L. (In fact, this could have been used to define $\varphi(L)$, but our definition in terms of contour integration has the advantage that it generalizes to the infinite-dimensional case.) Since

$$\sum_{\ell=1}^{m_i-1}\frac{1}{\ell!}\varphi^{(\ell)}(\lambda_i)N_i^{\ell}$$

is nilpotent for each i, it follows that $(*)$ is the (unique) spectral decomposition of $\varphi(L)$. Thus

$$\sigma(\varphi(L)) = \{\varphi(\lambda_1), \ldots, \varphi(\lambda_k)\} = \varphi(\sigma(L)).$$

Moreover, if $\{\varphi(\lambda_1), \ldots, \varphi(\lambda_k)\}$ are distinct, then the algebraic multiplicity of $\varphi(\lambda_i)$ as an eigenvalue of $\varphi(L)$ is the same as that of λ_i for L, and they have the same eigenprojection P_i. In general, one must add the algebraic multiplicities and eigenprojections over all those i with the same $\varphi(\lambda_i)$. □

Remarks:

(1) The special case in which L is diagonalizable is easy to remember:

$$\text{if} \quad L = \sum_{i=1}^{k} \lambda_i P_i, \quad \text{then} \quad \varphi(L) = \sum_{i=1}^{k} \varphi(\lambda_i) P_i.$$

(2) Other consequences of $(*)$ for general L are

$$\operatorname{tr} \varphi(L) = \sum_{i=1}^{k} m_i \varphi(\lambda_i) \quad \text{and} \quad \det \varphi(L) = \prod_{i=1}^{k} \varphi(\lambda_i)^{m_i}.$$

We now study composition.

Proposition. Suppose $L \in \mathcal{L}(V)$, φ_1 is holomorphic in a neighborhood of $\sigma(L)$, and φ_2 is holomorphic in a neighborhood of $\sigma(\varphi_1(L)) = \varphi_1(\sigma(L))$ (so $\varphi_2 \circ \varphi_1$ is holomorphic in a neighborhood of $\sigma(L)$). Then

$$(\varphi_2 \circ \varphi_1)(L) = \varphi_2(\varphi_1(L)).$$

Proof. Let Δ_2 contain $\sigma(\varphi_1(L))$ and let $\gamma_2 = \partial \Delta_2$. Then

$$\varphi_2(\varphi_1(L)) = \frac{1}{2\pi i} \int_{\gamma_2} \varphi_2(\zeta_2)(\zeta_2 I - \varphi_1(L))^{-1} d\zeta_2.$$

Here, $(\zeta_2 I - \varphi_1(L))^{-1}$ means of course the inverse of $\zeta_2 I - \varphi_1(L)$. For fixed $\zeta_2 \in \gamma_2$, we can also apply the functional calculus to the function $(\zeta_2 - \varphi_1(\zeta_1))^{-1}$ of ζ_1 to define this function of L: let Δ_1 contain $\sigma(L)$ and suppose that $\varphi_1(\bar{\Delta}_1) \subset \Delta_2$; then since $\zeta_2 \in \gamma_2$ is outside $\varphi_1(\bar{\Delta}_1)$, the map

$$\zeta_1 \mapsto (\zeta_2 - \varphi_1(\zeta_1))^{-1}$$

is holomorphic in a neighborhood of $\bar{\Delta}_1$, so we can evaluate this function of L; just as for

$$\zeta \mapsto \frac{1}{\zeta}$$

in the example above, we obtain the usual inverse of $\zeta_2 - \varphi_1(L)$. So

$$(\zeta_2 - \varphi_1(L))^{-1} = -\frac{1}{2\pi i} \int_{\gamma_1} (\zeta_2 - \varphi_1(\zeta_1))^{-1} R(\zeta_1) d\zeta_1.$$

Hence

$$\begin{aligned}
\varphi_2(\varphi_1(L)) &= -\frac{1}{(2\pi i)^2}\int_{\gamma_2}\varphi_2(\zeta_2)\int_{\gamma_1}(\zeta_2-\varphi_1(\zeta_1))^{-1}R(\zeta_1)d\zeta_1 d\zeta_2 \\
&= -\frac{1}{(2\pi i)^2}\int_{\gamma_1}R(\zeta_1)\int_{\gamma_2}\frac{\varphi_2(\zeta_2)}{\zeta_2-\varphi_1(\zeta_1)}d\zeta_2 d\zeta_1 \\
&= -\frac{1}{2\pi i}\int_{\gamma_1}R(\zeta_1)\varphi_2(\varphi_1(\zeta_1))d\zeta_1 \qquad\qquad (\text{as } n(\gamma_2,\varphi_1(\zeta_1))=1) \\
&= (\varphi_2\circ\varphi_1)(L).
\end{aligned}$$

□

Logarithms of Invertible Matrices

As an application, let $L \in \mathcal{L}(V)$ be invertible. We can choose a branch of $\log\zeta$ which is holomorphic in a neighborhood of $\sigma(L)$ and we can choose an appropriate Δ in which $\log\zeta$ is defined, so we can form

$$\log L = -\frac{1}{2\pi i}\int_\gamma \log\zeta R(\zeta)d\zeta \quad (\text{where } \gamma=\partial\Delta).$$

This definition will of course depend on the particular branch chosen, but since $e^{\log\zeta}=\zeta$ for any such branch, it follows that for any such choice,

$$e^{\log L} = L.$$

In particular, *every* invertible matrix is in the range of the exponential. This definition of the logarithm of an operator is much better than one can do with series: one could define

$$\log(I+A) = \sum_{\ell=1}^{\infty}(-1)^{\ell+1}\frac{A^\ell}{\ell},$$

but the series only converges absolutely in norm for a restricted class of A, namely $\{A : \rho(A)<1\}$.

Ordinary Differential Equations

Existence and Uniqueness Theory

Let $\mathbb{F}$ be $\mathbb{R}$ or $\mathbb{C}$. Throughout this discussion, $|\cdot|$ will denote the Euclidean norm (i.e ℓ^2-norm) on $\mathbb{F}^n$ (so $\|\cdot\|$ is free to be used for norms on function spaces). An ordinary differential equation (ODE) is an equation of the form

$$g(t, x, x', \ldots, x^{(m)}) = 0$$

where g maps a subset of $\mathbb{R} \times (\mathbb{F}^n)^{m+1}$ into $\mathbb{F}^n$. A *solution* of this ODE on an interval $I \subset \mathbb{R}$ is a function $x : I \to \mathbb{F}^n$ for which $x', x'', \ldots, x^{(m)}$ exist at each $t \in I$, and

$$(\forall\, t \in I) \qquad g(t, x(t), x'(t), \ldots, x^{(m)}(t)) = 0\ .$$

We will focus on the case where $x^{(m)}$ can be solved for explicitly, i.e., the equation takes the form

$$x^{(m)} = f(t, x, x', \ldots, x^{(m-1)}),$$

and where the function f mapping a subset of $\mathbb{R} \times (\mathbb{F}^n)^m$ into $\mathbb{F}^n$ is continuous. This equation is called an m^{th}-*order* $n \times n$ system of ODE's. Note that if x is a solution defined on an interval $I \subset \mathbb{R}$ then the existence of $x^{(m)}$ on I (including one-sided limits at the endpoints of I) implies that $x \in C^{m-1}(I)$, and then the equation implies $x^{(m)} \in C(I)$, so $x \in C^m(I)$.

Reduction to First-Order Systems

Every m^{th}-order $n \times n$ system of ODE's is equivalent to a first-order $mn \times mn$ system of ODE's. Defining

$$y_j(t) = x^{(j-1)}(t) \quad \in \mathbb{F}^n \quad \text{for} \quad 1 \le j \le m$$

and

$$y(t) = \begin{bmatrix} y_1(t) \\ \vdots \\ y_m(t) \end{bmatrix} \in \mathbb{F}^{mn},$$

the system

$$x^{(m)} = f(t, x, \ldots, x^{(m-1)})$$

is equivalent to the first-order $mn \times mn$ system

$$y' = \begin{bmatrix} y_2 \\ y_3 \\ \vdots \\ y_m \\ f(t, y_1, \ldots, y_m) \end{bmatrix}$$

(see problem 1 on Problem Set 9).

Relabeling if necessary, we will focus on first-order $n \times n$ systems of the form $x' = f(t,x)$, where f maps a subset of $\mathbb{R} \times \mathbb{F}^n$ into $\mathbb{F}^n$ and f is continuous.

Example: Consider the $n \times n$ system $x'(t) = f(t)$ where $f : I \to \mathbb{F}^n$ is continuous on an interval $I \subset \mathbb{R}$. (Here f is independent of x.) Then calculus shows that for a fixed $t_0 \in I$, the general solution of the ODE (i.e., a form representing all possible solutions) is

$$x(t) = c + \int_{t_0}^{t} f(s)ds,$$

where $c \in \mathbb{F}^n$ is an arbitrary constant vector (i.e., $c_1, \ldots, c_n$ are n arbitrary constants in $\mathbb{F}$).

Provided f satisfies a Lipschitz condition (to be discussed soon), the general solution of a first-order system $x' = f(t,x)$ involves n arbitrary constants in $\mathbb{F}$ [or an arbitrary vector in $\mathbb{F}^n$] (whether or not we can express the general solution explicitly), so n scalar conditions [or one vector condition] must be given to specify a particular solution. For the example above, clearly giving $x(t_0) = x_0$ (for a known constant vector x_0) determines c, namely, $c = x_0$. In general, specifying $x(t_0) = x_0$ (these are called *initial conditions* (IC), even if t_0 is not the left endpoint of the t-interval I) determines a particular solution of the ODE.

Initial-Value Problems for First-order Systems

An initial value problem (IVP) for the first-order system is the differential equation

$$DE: \qquad x' = f(t,x),$$

together with initial conditions

$$IC: \qquad x(t_0) = x_0\ .$$

A solution to the IVP is a solution $x(t)$ of the DE defined on an interval I containing t_0, which also satisfies the IC, i.e., for which $x(t_0) = x_0$.

Examples:

(1) Let $n = 1$. The solution of the IVP:

$$\begin{aligned} DE: &\qquad x' = x^2 \\ IC: &\qquad x(1) = 1 \end{aligned}$$

is $x(t) = \frac{1}{2-t}$, which blows up as $t \to 2$. So even if f is C^∞ on all of $\mathbb{R} \times \mathbb{F}^n$, solutions of an IVP do not necessarily exist for all time t.

(2) Let $n = 1$. Consider the IVP:

$$\begin{array}{rl} DE: & x' = 2\sqrt{|x|} \\ IC: & x(0) = 0\ . \end{array}$$

For any $c \geq 0$, define $x_c(t) = 0$ for $t \leq c$ and $x_c(t) = (t-c)^2$ for $t \geq c$. Then every $x_c(t)$ for $c \geq 0$ is a solution of this IVP. So in general for continuous $f(t,x)$, the solution of an IVP might not be unique. (The difficulty here is that $f(t,x) = 2\sqrt{|x|}$ is not Lipschitz continuous near $x = 0$.)

An Integral Equation Equivalent to an IVP

Suppose $x(t) \in C^1(I)$ is a solution of the IVP:

$$\begin{array}{rl} DE: & x' = f(t,x) \\ IC: & x(t_0) = x_0 \end{array}$$

defined on an interval $I \subset \mathbb{R}$ with $t_0 \in I$. Then for all $t \in I$,

$$\begin{aligned} x(t) &= x(t_0) + \int_{t_0}^{t} x'(s)ds \\ &= x_0 + \int_{t_0}^{t} f(s, x(s))ds, \end{aligned}$$

so $x(t)$ is also a solution of the *integral equation*

(IE) $$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))ds \qquad (t \in I).$$

Conversely, suppose $x(t) \in C(I)$ is a solution of the integral equation (IE). Then $f(t, x(t)) \in C(I)$, so

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))ds \in C^1(I)$$

and $x'(t) = f(t, x(t))$ by the Fundamental Theorem of Calculus. So x is a C^1 solution of the DE on I, and clearly $x(t_0) = x_0$, so x is a solution of the IVP. We have shown:

Proposition. On an interval I containing t_0, x is a solution of the IVP: $DE: x' = f(t,x)$; $IC: x(t_0) = x_0$ (where f is continuous) with $x \in C^1(I)$ if and only if x is a solution of the integral equation (IE) on I with $x \in C(I)$.

The integral equation (IE) is a useful way to study the IVP. We can deal with the function space of continuous functions on I without having to be concerned about differentiability: continuous solutions of (IE) are automatically C^1. Moreover, the initial condition is built into the integral equation.

We will solve (IE) using a fixed-point formulation.

Definition. Let (X, d) be a metric space, and suppose $F : X \to X$. We say that F is a *contraction* [on X] if there exists $c < 1$ such that

$$(\forall\, x, y \in X) \qquad d(F(x), F(y)) \leq cd(x, y)$$

(c is sometimes called the contraction constant). A point $x_* \in X$ for which

$$F(x_*) = x_*$$

is called a *fixed point* of F.

Theorem (Contraction Mapping Fixed-Point Theorem).
Let (X, d) be a complete *metric space and $F : X \to X$ be a contraction (with contraction constant $c < 1$). Then F has a unique fixed point $x_* \in X$. Moreover, for any $x_0 \in X$, if we generate the sequence $\{x_k\}$ iteratively by* functional iteration

$$x_{k+1} = F(x_k) \quad \textit{for} \quad k \geq 0$$

(sometimes called fixed-point iteration*), then $x_k \to x_*$.*

Proof. Fix $x_0 \in X$, and generate $\{x_k\}$ by $x_{k+1} = F(x_k)$. Then for $k \geq 1$,

$$d(x_{k+1}, x_k) = d(F(x_k), F(x_{k-1})) \leq cd(x_k, x_{k-1}).$$

By induction

$$d(x_{k+1}, x_k) \leq c^k d(x_1, x_0).$$

So for $n < m$,

$$\begin{aligned} d(x_m, x_n) &\leq \sum_{j=n}^{m-1} d(x_{j+1}, x_j) \leq \left(\sum_{j=n}^{m-1} c^j \right) d(x_1, x_0) \\ &\leq \left(\sum_{j=n}^{\infty} c^j \right) d(x_1, x_0) = \frac{c^n}{1-c} d(x_1, x_0). \end{aligned}$$

Since $c^n \to 0$ as $n \to \infty$, $\{x_k\}$ is Cauchy. Since X is complete, $x_k \to x_*$ for some $x_* \in X$. Since F is a contraction, clearly F is continuous, so

$$F(x_*) = F(\lim x_k) = \lim F(x_k) = \lim x_{k+1} = x_*,$$

so x_* is a fixed point. If x and y are two fixed points of F in X, then

$$d(x, y) = d(F(x), F(y)) \leq cd(x, y),$$

so $(1 - c)d(x, y) \leq 0$, and thus $d(x, y) = 0$ and $x = y$. So F has a unique fixed point. $\square$

Applications.

(1) *Iterative methods for linear systems* (see problem 3 on Problem Set 9).

(2) *The Inverse Function Theorem* (see problem 4 on Problem Set 9). If $\Phi : U \to \mathbb{R}^n$ is a C^1 mapping on a neighborhood $U \subset \mathbb{R}^n$ of $x_0 \in \mathbb{R}^n$ satisfying $\Phi(x_0) = y_0$ and $\Phi'(x_0) \in \mathbb{R}^{n\times n}$ is invertible, then there exist neighborhoods $U_0 \subset U$ of x_0 and V_0 of y_0 and a C^1 mapping $\Psi : V_0 \to U_0$ for which $\Phi[U_0] = V_0$ and $\Phi \circ \Psi$ and $\Psi \circ \Phi$ are the identity mappings on V_0 and U_0, respectively.

(In problem 4 of Problem Set 9, you will show that Φ has a continuous right inverse defined on some neighborhood of y_0. Other arguments are required to show that $\Psi \in C^1$ and that Ψ is a two-sided inverse; these are not discussed here.)

Remark. Applying the Contraction Mapping Fixed-Point Theorem (C.M.F.-P.T.) to a mapping F usually requires two steps:

(1) Construct a complete metric space X and a closed subset $S \subset X$ for which $F(S) \subset S$.

(2) Show that F is a contraction on S.

To apply the C.M.F.-P.T. to the integral equation (IE), we need a further condition on the function $f(t, x)$.

Definition. Let $I \subset \mathbb{R}$ be an interval and $\Omega \subset \mathbb{F}^n$. We say that $f(t, x)$ mapping $I \times \Omega$ into $\mathbb{F}^n$ is *uniformly Lipschitz continuous with respect to* x if there is a constant L (called the *Lipschitz constant*) for which

$$(\forall\, t \in I)(\forall\, x, y \in \Omega) \qquad |f(t, x) - f(t, y)| \le L|x - y| \ .$$

We say that f is in (C, Lip) on $I \times \Omega$ if f is continuous on $I \times \Omega$ and f is uniformly Lipschitz continuous with respect to x on $I \times \Omega$.

For simplicity, we will consider intervals $I \subset \mathbb{R}$ for which t_0 is the left endpoint. Virtually identical arguments hold if t_0 is the right endpoint of I, or if t_0 is in the interior of I (see Coddington & Levinson).

Theorem (Local Existence and Uniqueness for (IE) for Lipschitz f)
Let $I = [t_0, t_0 + \beta]$ and $\Omega = \overline{B_r(x_0)} = \{x \in \mathbb{F}^n : |x - x_0| \le r\}$, and suppose $f(t, x)$ is in (C, Lip) on $I \times \Omega$. Then there exisits $\alpha \in (0, \beta]$ for which there is a unique solution of the integral equation

$$\text{(IE)} \qquad x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))ds$$

in $C(I_\alpha)$, where $I_\alpha = [t_0, t_0 + \alpha]$. Moreover, we can choose α to be any positive number satisfying

$$\alpha \le \beta,\ \alpha \le \frac{r}{M}, \quad \textit{and} \quad \alpha < \frac{1}{L}, \ \textit{where} \quad M = \max_{(t,x)\in I\times\Omega} |f(t, x)|$$

and L is the Lipschitz constant for f in $I \times \Omega$.

Proof. For any $\alpha \in (0, \beta]$, let $\|\cdot\|_\infty$ denote the max-norm on $C(I_\alpha)$:

$$\text{for} \quad x \in C(I_\alpha), \quad \|x\|_\infty = \max_{t_0 \le t \le t_0+\alpha} |x(t)| \ .$$

Although this norm clearly depends on α, we do not include α in the notation. Let $x_0 \in C(I_\alpha)$ denote the constant function $x_0(t) \equiv x_0$. For $\rho > 0$ let

$$X_{\alpha,\rho} = \{x \in C(I_\alpha) : \|x - x_0\|_\infty \leq \rho\}.$$

Then $X_{\alpha,\rho}$ is a complete metric space since it is a closed subset of the Banach space $(C(I_\alpha), \| \cdot \|_\infty)$. For any $\alpha \in (0, \beta]$, define $F : X_{\alpha,r} \to C(I_\alpha)$ by

$$(F(x))(t) = x_0 + \int_{t_0}^{t} f(s, x(s))ds.$$

Note that F is well-defined on $X_{\alpha,r}$ and $F(x) \in C(I_\alpha)$ for $x \in X_{\alpha,r}$ since f is continuous on $I \times \overline{B_r(x_0)}$. Fixed points of F are solutions of the integral equation (IE).

Claim. Suppose $\alpha \in (0, \beta]$, $\alpha \leq \frac{r}{M}$, and $\alpha < \frac{1}{L}$. Then F maps $X_{\alpha,r}$ into itself and F is a contraction on $X_{\alpha,r}$.

Proof of Claim: If $x \in X_{\alpha,r}$, then for $t \in I_\alpha$,

$$|(F(x))(t) - x_0| \leq \int_{t_0}^{t} |f(s, x(s))|ds \leq M\alpha \leq r,$$

so $F : X_{\alpha,r} \to X_{\alpha,r}$. If $x, y \in X_{\alpha,r}$, then for $t \in I_\alpha$,

$$\begin{aligned} |(F(x))(t) - (F(y))(t)| &\leq \int_{t_0}^{t} |f(s, x(s)) - f(s, y(s))|ds \\ &\leq \int_{t_0}^{t} L|x(s) - y(s)|ds \\ &\leq L\alpha\|x - y\|_\infty, \end{aligned}$$

so

$$\|F(x) - F(y)\|_\infty \leq L\alpha\|x - y\|_\infty, \quad \text{and} \quad L\alpha < 1.$$

So by the C.M.F.-P.T., for α satisfying $0 < \alpha \leq \beta$, $\alpha \leq \frac{r}{M}$, and $\alpha < \frac{1}{L}$, F has a unique fixed point in $X_{\alpha,r}$, and thus the integral equation (IE) has a unique solution $x_*(t)$ in $X_{\alpha,r} = \{x \in C(I_\alpha) : \|x - x_0\|_\infty \leq r\}$. This is *almost* the conclusion of the Theorem, except we haven't shown x_* is the only solution in all of $C(I_\alpha)$. This uniqueness is better handled by techniques we will study soon, but we can still eke out a proof here. (Since f is only defined on $I \times \overline{B_r(x_0)}$, technically $f(t, x(t))$ does not make sense if $x \in C(I_\alpha)$ but $x \notin X_{\alpha,r}$. To make sense of the uniqueness statement for general $x \in C(I_\alpha)$, we choose some continuous extension of f to $I \times \mathbb{F}^n$.) Fix α as above. Then clearly for $0 < \gamma \leq \alpha$, $x_*|_{I_\gamma}$ is the unique fixed point of F on $X_{\gamma,r}$. Suppose $y \in C(I_\alpha)$ is a solution of (IE) on I_α (using perhaps an extension of f) with $y \not\equiv x_*$ on I_α. Let

$$\gamma_1 = \inf\{\gamma \in (0, \alpha] : y(t_0 + \gamma) \neq x_*(t_0 + \gamma)\}.$$

By continuity, $\gamma_1 < \alpha$. Since $y(t_0) = x_0$, continuity implies

$$\exists\, \gamma_0 \in (0, \alpha] \ni y|_{I_{\gamma_0}} \in X_{\gamma_0, r},$$

and thus $y(t) \equiv x_*(t)$ on I_{γ_0}. So $0 < \gamma_1 < \alpha$. Since $y(t) \equiv x_*(t)$ on I_{γ_1}, $y|_{I_{\gamma_1}} \in X_{\gamma_1, r}$. Let $\rho = M\gamma_1$; then $\rho < M\alpha \le r$. For $t \in I_{\gamma_1}$,

$$|y(t) - x_0| = |(F(y))(t) - x_0| \le \int_{t_0}^{t} |f(s, y(s))| ds \le M\gamma_1 = \rho,$$

so $y|_{I_{\gamma_1}} \in X_{\gamma_1, \rho}$. By continuity, there exists $\gamma_2 \in (\gamma_1, \alpha] \ni y|_{I_{\gamma_2}} \in X_{\gamma_1, r}$. But then $y(t) \equiv x_*(t)$ on I_{γ_2}, contradicting the definition of γ_1. □

The Picard Iteration

Although hidden in a few too many details, the main idea of the proof above is to study the convergence of functional iterates of F. If we choose the initial iterate to be $x_0(t) \equiv x_0$, we obtain the classical Picard Iteration:

$$\left\{ \begin{array}{rcl} x_0(t) & \equiv & x_0 \\ x_{k+1}(t) & = & x_0 + \int_{t_0}^{t} f(s, x_k(s)) ds \quad \text{for} \quad k \ge 0 \end{array} \right.$$

The argument in the proof of the C.M.F.-P.T. gives only *uniform* estimates of, e.g., $x_{k+1} - x_k$: $\|x_{k+1} - x_k\|_\infty \le L\alpha \|x_k - x_{k+1}\|_\infty$, leading to the condition $\alpha < \frac{1}{L}$. For the Picard iteration (and other iterations of similar nature, e.g., for Volterra integral equations of the second kind), we can get better results using *pointwise* estimates of $x_{k+1} - x_k$. The condition $\alpha < \frac{1}{L}$ turns out to be unnecessary (we will see another way to eliminate this assumption when we study continuation of solutions). For the moment, we will set aside the uniqueness question and focus on existence.

Theorem (Picard Global Existence for (IE) for Lipschitz f). *Let $I = [t_0, t_0 + \beta]$, and suppose $f(t, x)$ is in (C, Lip) on $I \times \mathbb{F}^n$. Then there exists a solution $x_*(t)$ of the integral equation (IE) in $C(I)$.*

Theorem (Picard Local Existence for (IE) for Lipschitz f). *Let $I = [t_0, t_0 + \beta]$ and $\Omega = \overline{B_r(x_0)} = \{x \in \mathbb{F}^n : |x - x_0| \le r\}$, and suppose $f(t, x)$ is in (C, Lip) on $I \times \Omega$. Then there exists a solution $x_*(t)$ of the integral equation (IE) in $C(I_\alpha)$, where $I_\alpha = [t_0, t_0 + \alpha]$, $\alpha = \min\left(\beta, \frac{r}{M}\right)$, and $M = \max_{(t,x) \in I \times \Omega} |f(t, x)|$.*

Proofs. We prove the two theorems together. For the global theorem, let $X = C(I)$ (i.e., $C(I, \mathbb{F}^n)$), and for the local theorem, let

$$X = X_{\alpha, r} \equiv \{x \in C(I_\alpha) : \|x - x_0\|_\infty \le r\}$$

as before (where $x_0(t) \equiv x_0$). Then the map

$$(F(x))(t) = x_0 + \int_{t_0}^{t} f(s, x(s)) ds$$

maps X into X in both cases, and X is complete. Let

$$x_0(t) \equiv x_0, \quad \text{and} \quad x_{k+1} = F(x_k) \quad \text{for} \quad k \geq 0.$$

Let

$$\begin{aligned} M_0 &= \max_{t \in I} |f(t, x_0)| \qquad \text{(global theorem)}, \\ M_0 &= \max_{t \in I_\alpha} |f(t, x_0)| \qquad \text{(local theorem)}. \end{aligned}$$

Then for $t \in I$ (global) or $t \in I_\alpha$ (local),

$$\begin{aligned} |x_1(t) - x_0| &\leq \int_{t_0}^{t} |f(s, x_0)| ds \leq M_0(t - t_0) \\ |x_2(t) - x_1(t)| &\leq \int_{t_0}^{t} |f(s, x_1(s)) - f(s, x_0(s))| ds \\ &\leq L \int_{t_0}^{t} |x_1(s) - x_0(s)| ds \\ &\leq M_0 L \int_{t_0}^{t} (s - t_0) ds = \frac{M_0 L (t - t_0)^2}{2!} \end{aligned}$$

By induction, suppose $|x_k(t) - x_{k-1}(t)| \leq M_0 L^{k-1} \frac{(t-t_0)^k}{k!}$. Then

$$\begin{aligned} |x_{k+1}(t) - x_k(t)| &\leq \int_{t_0}^{t} |f(s, x_k(s)) - f(s, x_{k-1}(s))| ds \\ &\leq L \int_{t_0}^{t} |x_k(s) - x_{k-1}(s)| ds \\ &\leq M_0 L^k \int_{t_0}^{t} \frac{(s - t_0)^k}{k!} ds = M_0 L^k \frac{(t - t_0)^{k+1}}{(k+1)!}. \end{aligned}$$

So $\sup_t |x_{k+1}(t) - x_k(t)| \leq M_0 L^k \frac{\gamma^{k+1}}{(k+1)!}$, where $\gamma = \beta$ (global) or $\gamma = \alpha$ (local). Hence

$$\begin{aligned} \sum_{k=0}^{\infty} \sup_t |x_{k+1}(t) - x_k(t)| &\leq \frac{M_0}{L} \sum_{k=0}^{\infty} \frac{(L\gamma)^{k+1}}{(k+1)!} \\ &= \frac{M_0}{L}(e^{L\gamma} - 1). \end{aligned}$$

It follows that the series $x_0 + \sum_{k=0}^{\infty}(x_{k+1}(t) - x_k(t))$, which has x_{N+1} as its N^{th} partial sum, converges absolutely and uniformly on I (global) or I_α (local) by the Weierstrass M-test. Let $x_*(t) \in C(I)$ (global) or $\in C(I_\alpha)$ (local) be the limit function. Since

$$|f(t, x_k(t)) - f(t, x_*(t))| \leq L|x_k(t) - x_*(t)|,$$

$f(t, x_k(t))$ converges uniformly to $f(t, x_*(t))$ on I (global) or I_α (local), and thus

$$\begin{aligned} F(x_*)(t) &= x_0 + \int_{t_0}^t f(s, x_*(s))ds \\ &= \lim_{k\to\infty} (x_0 + \int_{t_0}^t f(s, x_k(x))ds) \\ &= \lim_{k\to\infty} x_{k+1}(t) = x_*(t), \end{aligned}$$

for all $t \in I$ (global) or I_α (local). Hence $x_*(t)$ is a fixed point of F in X, and thus also a solution of the integral equation (IE) in $C(I)$ (global) or $C(I_\alpha)$ (local). □

Corollary. The solution $x_*(t)$ of (IE) satisfies

$$|x_*(t) - x_0| \leq \frac{M_0}{L}(e^{L(t-t_0)} - 1)$$

for $t \in I$ (global) or $t \in I_\alpha$ (local), where $M_0 = \max_{t\in I} |f(t, x_0)|$ (global), or $M_0 = \max_{t\in I_\alpha} |f(t, x_0)|$ (local).

Proof. This is established in the proof above. □

Remark. In each of the statements of the last three Theorems, we could replace "solution of the integral equation (IE)" with "solution of the IVP: $DE : x' = f(t, x)$; $IC : x(t_0) = x_0$" because of the equivalence of these two problems.

Examples.

(1) Consider a *linear* system $x' = A(t)x + b(t)$, where $A(t) \in \mathbb{C}^{n\times n}$ and $b(t) \in \mathbb{C}^n$ are in $C(I)$ (where $I = [t_0, t_0 + \beta]$). Then f is in (C, Lip) on $I \times \mathbb{F}^n$:

$$|f(t, x) - f(t, y)| \leq |A(t)x - A(t)y| \leq \left(\max_{t\in I} \|A(t)\|\right) |x - y|.$$

Hence there is a solution of the IVP: $x' = A(t)x + b(t)$, $x(t_0) = x_0$ in $C^1(I)$.

(2) $(n = 1)$ Consider the IVP: $x' = x^2$, $x(0) = x_0 > 0$. Then $f(t, x) = x^2$ is not in (C, Lip) on $I \times \mathbb{R}$. It is, however, in (C, Lip) on $I \times \Omega$ where $\Omega = \overline{B_r(x_0)} = [x_0 - r, x_0 + r]$ for each fixed r. For a given $r > 0$, $M = (x_0 + r)^2$, and $\alpha = \frac{r}{M} = \frac{r}{(x_0+r)^2}$ in the local theorem is maximized for $r = x_0$, for which $\alpha = (4x_0)^{-1}$. So the local theorem guarantees a solution in $[0, (4x_0)^{-1}]$. The actual solution $x_*(t) = (x_0^{-1} - t)^{-1}$ exists in $[0, (x_0)^{-1})$.

Local Existence for Continuous f

Some condition similar to the Lipschitz condition is needed to guarantee that the Picard iterates converge; it is also needed for uniqueness, which we will return to shortly. It is, however, still possible to prove a local existence theorem assuming only that f is continuous, without assuming the Lipschitz condition. We will need the following form of Ascoli's Theorem:

Theorem (Ascoli). *Let X and Y be metric spaces with X compact. Let $\{f_k\}$ be an* equicontinuous *sequence of functions $f_k : X \to Y$, i.e.,*

$$\begin{gathered}(\forall\, \epsilon > 0)(\exists\, \delta > 0) \quad \textit{such that} \quad (\forall\, k \geq 1)(\forall\, x_1, x_2 \in X) \\ d_X(x_1, x_2) < \delta \Rightarrow d_Y(f_k(x_1), f_k(x_2)) < \epsilon\end{gathered}$$

(in particular, each f_k is continuous), and suppose for each $x \in X$, $\overline{\{f_k(x) : k \geq 1\}}$ is a compact subset of Y. Then there is a subsequence $\{f_{k_j}\}_{j=1}^{\infty}$ and a continuous $f : X \to Y$ such that

$$f_{k_j} \to f \quad \textit{uniformly on } X.$$

Remark. If $Y = \mathbb{F}^n$, the condition $(\forall\, x \in X)$ $\overline{\{f_k(x) : k \geq 1\}}$ is compact is equivalent to the sequence $\{f_k\}$ being *pointwise bounded*, i.e.,

$$(\forall\, x \in X)(\exists\, M_x) \quad \text{such that} \quad (\forall\, k \geq 1) \quad |f_k(x)| \leq M_x.$$

Example. Suppose $f_k : [a, b] \to \mathbb{R}$ is a sequence of C^1 functions, and suppose there exists $M > 0$ such that

$$(\forall\, k \geq 1) \quad \|f_k\|_\infty + \|f_k'\|_\infty \leq M$$

(where $\|f\|_\infty = \max_{a \leq x \leq b} |f(x)|$). Then for $a \leq x_1 < x_2 \leq b$,

$$|f_k(x_2) - f_k(x_1)| \leq \int_{x_1}^{x_2} |f_k'(x)| dx \leq M|x_2 - x_1|,$$

so $\{f_k\}$ is equicontinuous (take $\delta = \frac{\epsilon}{M}$), and $\|f_k\|_\infty \leq M$ certainly implies $\{f_k\}$ is pointwise bounded. So by Ascoli's Theorem, some subsequence of $\{f_k\}$ converges uniformly to a continuous function $f : [a, b] \to \mathbb{R}$.

Theorem (Cauchy-Peano Existence Theorem).
Let $I = [t_0, t_0 + \beta]$ and $\Omega = \overline{B_r(x_0)} = \{x \in \mathbb{F}^n : |x - x_0| \leq r\}$, and suppose $f(t, x)$ is continuous on $I \times \Omega$. Then there exists a solution $x_*(t)$ of the integral equation

$$\text{(IE)} \qquad x(t) = x_0 + \int_{t_0}^{t} f(s, x(s)) ds$$

in $C(I_\alpha)$ where $I_\alpha = [t_0, t_0 + \alpha]$, $\alpha = \min\left(\beta, \frac{r}{M}\right)$, and $M = \max_{(t,x) \in I \times \Omega} |f(t, x)|$ (and thus $x_*(t)$ is a C^1 solution on I_α of the IVP: $x' = f(t, x)$; $x(t_0) = x_0$).

Proof. The idea of the proof is to construct continuous approximate solutions explicitly (we will use the piecewise linear interpolants of grid functions generated by Euler's method), and use Ascoli's Theorem to take the uniform limit of some subsequence. For each integer $k \geq 1$, define $x_k(t) \in C(I_\alpha)$ as follows: partition $[t_0, t_0+\alpha]$ into k equal subintervals (for $0 \leq \ell \leq k$, let $t_\ell = t_0 + \ell\frac{\alpha}{k}$ (note: t_ℓ depends on k too)), set $x_k(t_0) = x_0$, and for $\ell = 1, 2, \ldots, k$ define $x_k(t)$ in $(t_{\ell-1}, t_\ell]$ inductively by $x_k(t) = x_k(t_{\ell-1}) + f(t_{\ell-1}, x_k(t_{\ell-1}))(t - t_{\ell-1})$. For this to be well-defined we must check that $|x_k(t_{\ell-1}) - x_0| \leq r$ for $2 \leq \ell \leq k$ (it is obvious for $\ell = 1$); inductively, we have

$$\begin{aligned}
|x_k(t_{\ell-1}) - x_0| &\leq \sum_{i=1}^{\ell-1} |x_k(t_i) - x_k(t_{i-1})| \\
&= \sum_{i=1}^{\ell-1} |f(t_{i-1}, x_k(t_{i-1}))| \cdot |t_i - t_{i-1}| \\
&\leq M \sum_{i=1}^{\ell-1} (t_i - t_{i-1}) \\
&= M(t_{\ell-1} - t_0) \leq M\alpha \leq r
\end{aligned}$$

by the choice of α. So $x_k(t) \in C(I_\alpha)$ is well defined. A similar estimate shows that for $t, \tau \in [t_0, t_0+\alpha]$,

$$|x_k(t) - x_k(\tau)| \leq M|t - \tau|.$$

This implies that $\{x_k\}$ is equicontinuous; it also implies that

$$(\forall\, k \geq 1)(\forall\, t \in I_\alpha) \quad |x_k(t) - x_0| \leq M\alpha \leq r,$$

so $\{x_k\}$ is pointwise bounded (in fact, uniformly bounded). So by Ascoli's Theorem, there exists $x_*(t) \in C(I_\alpha)$ and a subsequence $\{x_{k_j}\}_{j=1}^\infty$ converging uniformly to $x_*(t)$. It remains to show that $x_*(t)$ is a solution of (IE) on I_α. Since each $x_k(t)$ is continuous and piecewise linear on I_α,

$$x_k(t) = x_0 + \int_{t_0}^t x_k'(s)ds$$

(where $x_k'(t)$ is piecewise constant on I_α and is defined for all t except t_ℓ $(1 \leq \ell \leq k-1)$, where we define it to be $x_k'(t_\ell^+)$). Define

$$\Delta_k(t) = x_k'(t) - f(t, x_k(t)) \quad \text{on} \quad I_\alpha$$

(note that $\Delta_k(t_\ell) = 0$ for $0 \leq \ell \leq k-1$ by definition). We claim that $\Delta_k(t) \to 0$ uniformly on I_α as $k \to \infty$. Indeed, given k, we have for $1 \leq \ell \leq k$ and $t \in (t_{\ell-1}, t_\ell)$ (including t_k if $\ell = k$), that

$$|x_k'(t) - f(t, x_k(t))| = |f(t_{\ell-1}, x_k(t_{\ell-1})) - f(t, x_k(t))|.$$

Noting that $|t - t_{\ell-1}| \leq \frac{\alpha}{k}$ and

$$|x_k(t) - x_k(t_{\ell-1})| \leq M|t - t_{\ell-1}| \leq M\frac{\alpha}{k},$$

the uniform continuity of f (being continuous on the compact set $I \times \Omega$) implies that

$$\max_{t \in I_\alpha} |\Delta_k(t)| \to 0 \quad \text{as} \quad k \to \infty.$$

Thus, in particular, $\Delta_{k_j}(t) \to 0$ uniformly on I_α. Now

$$\begin{aligned} x_{k_j}(t) &= x_0 + \int_{t_0}^t x'_{k_j}(s)ds \\ &= x_0 + \int_{t_0}^t f(s, x_{k_j}(s))ds + \int_{t_0}^t \Delta_{k_j}(s)ds. \end{aligned}$$

Since $x_{k_j} \to x_*$ uniformly on I_α, the uniform continuity of f on $I \times \Omega$ now implies that $f(t, x_{k_j}(t)) \to f(t, x_*(t))$ uniformly on I_α, so taking the limit as $j \to \infty$ on both sides of this equation for each $t \in I_\alpha$, we obtain that x_* satisfies (IE) on I_α. □

Remark. In general, the choice of a subsequence of $\{x_k\}$ is necessary: there are examples where the sequence $\{x_k\}$ does not converge. (See Problem 12, Chapter 1 of Coddington & Levinson.)

Uniqueness

Uniqueness theorems are typically proved by comparison theorems for solutions of scalar differential equations, or by inequalities. The most fundamental of these inequalities is Gronwall's inequality, which applies to real first-order linear scalar equations.

Recall that a first-order linear scalar initial value problem

$$u' = a(t)u + b(t), \quad u(t_0) = u_0$$

can be solved by multiplying by the integrating factor $e^{-\int_{t_0}^t a}$ (i.e., $e^{-\int_{t_0}^t a(s)ds}$), and then integrating from t_0 to t. That is,

$$\frac{d}{dt}\left(e^{-\int_{t_0}^t a} u(t)\right) = e^{-\int_{t_0}^t a} b(t),$$

which implies that

$$\begin{aligned} e^{-\int_{t_0}^t a} u(t) - u_0 &= \int_{t_0}^t \frac{d}{ds}\left(e^{-\int_{t_0}^s a} u(s)\right) ds \\ &= \int_{t_0}^t e^{-\int_{t_0}^s a} b(s)ds \end{aligned}$$

which in turn implies that

$$u(t) = u_0 e^{\int_{t_0}^t a} + \int_{t_0}^t e^{\int_s^t a} b(s)ds.$$

Since $f(t) \le g(t)$ on $[c, d]$ implies $\int_c^d f(t)dt \le \int_c^d g(t)dt$, the identical argument with "=" replaced by "$\le$" gives

Theorem (Gronwall's Inequality - differential form). *Let $I = [t_0, t_1]$. Suppose $a : I \to \mathbb{R}$ and $b : I \to \mathbb{R}$ are continuous, and suppose $u : I \to \mathbb{R}$ is in $C^1(I)$ and satisfies*

$$u'(t) \le a(t)u(t) + b(t) \quad \textit{for} \quad t \in I, \quad \textit{and} \quad u(t_0) = u_0.$$

Then

$$u(t) \leq u_0 e^{\int_{t_0}^t a} + \int_{t_0}^t e^{\int_s^t a} b(s)ds.$$

Remarks:

(1) Thus a solution of the differential inequality is bounded above by the solution of the equality (i.e., the differential equation $u' = au + b$).

(2) The result clearly still holds if u is only continuous and piecewise C^1, and $a(t)$ and $b(t)$ are only piecewise continuous.

(3) There is also an integral form of Gronwall's inequality (i.e., the hypothesis is an integral inequality): if $\varphi, \psi, \alpha \in C(I)$ are real-valued with $\alpha \geq 0$ on I, and

$$\varphi(t) \leq \psi(t) + \int_{t_0}^t \alpha(s)\varphi(s)ds \quad \text{for} \quad t \in I,$$

then

$$\varphi(t) \leq \psi(t) + \int_{t_0}^t e^{\int_s^t \alpha} \alpha(s)\psi(s)ds.$$

In particular, if $\psi(t) \equiv c$ (a constant), then $\varphi(t) \leq c e^{\int_{t_0}^t \alpha}$. (The differential form is applied to the C^1 function $u(t) = \int_{t_0}^t \alpha(s)\varphi(s)ds$ in the proof.)

(4) For $a(t) \geq 0$, the differential form is also a consequence of the integral form: integrating

$$u' \leq a(t)u + b(t)$$

from t_0 to t gives

$$u(t) \leq \psi(t) + \int_{t_0}^t a(s)u(s)ds,$$

where

$$\psi(t) = u_0 + \int_{t_0}^t b(s)ds,$$

so the integral form and then integration by parts give

$$\begin{aligned} u(t) &\leq \psi(t) + \int_{t_0}^t e^{\int_s^t a} a(s)\psi(s)ds \\ &= \cdots = u_0 e^{\int_{t_0}^t a} + \int_{t_0}^t e^{\int_s^t a} b(s)ds. \end{aligned}$$

(5) Caution: a differential inequality implies an integral inequality, but *not* vice versa: $f \leq g \not\Rightarrow f' \leq g'$.

(6) The integral form doesn't require $\varphi \in C^1$ (just $\varphi \in C(I)$), but is restricted to $\alpha \geq 0$. The differential form has no sign restriction on $a(t)$, but it requires a stronger hypothesis (in view of (5) and the requirement that u be continuous and piecewise C^1).

Uniqueness for Locally Lipschitz f

We start with a one-sided local uniqueness theorem for the initial value problem

$$IVP: \qquad x' = f(t,x); \quad x(t_0) = x_0.$$

Theorem. *Suppose for some $\alpha > 0$ and some $r > 0$, $f(t,x)$ is in (C,Lip) on $I_\alpha \times \overline{B_r(x_0)}$, and suppose $x(t)$ and $y(t)$ both map I_α into $\overline{B_r(x_0)}$ and both are C^1 solutions of (IVP) on I_α, where $I_\alpha = [t_0, t_0+\alpha]$. Then $x(t) = y(t)$ for $t \in I_\alpha$.*

Proof. Set

$$u(t) = |x(t) - y(t)|^2 = \langle x(t) - y(t), x(t) - y(t)\rangle$$

(in the Euclidean inner product on $\mathbb{F}^n$). Then $u : I_\alpha \to [0,\infty)$ and $u \in C^1(I_\alpha)$ and for $t \in I_\alpha$,

$$\begin{aligned} u' &= \langle x-y, x'-y'\rangle + \langle x'-y', x-y\rangle \\ &= 2\mathcal{R}e\langle x-y, x'-y'\rangle \le 2|\langle x-y, x'-y'\rangle| \\ &= 2|\langle x-y, (f(t,x) - f(t,y))\rangle| \\ &\le 2|x-y| \cdot |f(t,x) - f(t,y)| \\ &\le 2L|x-y|^2 = 2Lu\ . \end{aligned}$$

Thus $u' \le 2Lu$ on I_α and $u(t_0) = x(t_0) - y(t_0) = x_0 - x_0 = 0$. By Gronwall's inequality, $u(t) \le u_0 e^{2Lt} = 0$ on I_α, so since $u(t) \ge 0$, $u(t) \equiv 0$ on I_α. □

Corollary.

(i) The same result holds if $I_\alpha = [t_0 - \alpha, t_0]$.

(ii) The same result holds if $I_\alpha = [t_0 - \alpha, t_0 + \alpha]$.

Proof. For (i), let $\widetilde{x}(t) = x(2t_0 - t)$, $\widetilde{y}(t) = y(2t_0 - t)$, and $\widetilde{f}(t,x) = -f(2t_0 - t, x)$. Then $\widetilde{f}$ is in (C,Lip) on $[t_0, t_0+\alpha] \times \overline{B_r(x_0)}$, and $\widetilde{x}$ and $\widetilde{y}$ both satisfy the IVP

$$x' = \widetilde{f}(t,x); \quad x'(t_0) = x_0 \quad \text{on} \quad [t_0, t_0+\alpha].$$

So by the Theorem, $\widetilde{x}(t) = \widetilde{y}(t)$ for $t \in [t_0, t_0+\alpha]$, i.e., $x(t) = y(t)$ for $t \in [t_0 - \alpha, t_0]$. Now (ii) follows immediately by applying the Theorem in $[t_0, t_0+\alpha]$ and applying (i) in $[t_0 - \alpha, t_0]$. □

Remark. The idea used in the proof of (i) is often called "time-reversal." The important part is that $\widetilde{x}(t) = x(c - t)$, etc., for some constant c, so that $\widetilde{x}'(t) = -x'(c-t)$, etc. The choice of $c = 2t_0$ is convenient but not essential.

The main uniqueness theorem is easiest to formulate in the case when the initial point (t_0, x_0) is in the interior of the domain of definition of f. There are analogous results with essentially the same proof when (t_0, x_0) is on the boundary of the domain of definition of f.

(Exercise: State precisely a theorem corresponding to the upcoming theorem which applies in such a situation.)

Definition. Let $\mathcal{D}$ be an open set in $\mathbb{R} \times \mathbb{F}^n$. We say that $f(t, x)$ mapping $\mathcal{D}$ into $\mathbb{F}^n$ is *locally Lipschitz continuous with respect to* x if for each $(t_1, x_1) \in \mathcal{D}$ there exists

$$\alpha > 0, \quad r > 0, \quad \text{and} \quad L > 0$$

for which $[t_1 - \alpha, t_1 + \alpha] \times \overline{B_r(x_1)} \subset \mathcal{D}$ and

$$(\forall\, t \in [t_1 - \alpha, t_1 + \alpha])(\forall\, x, y \in \overline{B_r(x_1)}) \quad |f(t, x) - f(t, y)| \le L|x - y|$$

(i.e., f is uniformly Lipschitz continuous with respect to x in $[t_1 - \alpha, t_1 + \alpha] \times \overline{B_r(x_1)}$). We will say $f \in (C, \mathrm{Lip}_{\mathrm{loc}})$ (not a standard notation) on $\mathcal{D}$ if f is continuous on $\mathcal{D}$ and locally Lipschitz continuous with respect to x on $\mathcal{D}$.

Example. Let $\mathcal{D}$ be an open set of $\mathbb{R} \times \mathbb{F}^n$. Suppose $f(t, x)$ maps $\mathcal{D}$ into $\mathbb{F}^n$, f is continuous on $\mathcal{D}$, and for $1 \le i, j \le n$, $\frac{\partial f_i}{\partial x_j}$ exists and is continuous in $\mathcal{D}$. (Briefly, we say f is continuous on $\mathcal{D}$ and C^1 with respect to x on $\mathcal{D}$.) Then $f \in (C, \mathrm{Lip}_{\mathrm{loc}})$ on $\mathcal{D}$. (Exercise.)

Main Uniqueness Theorem. *Let $\mathcal{D}$ be an open set in $\mathbb{R} \times \mathbb{F}^n$, and suppose $f \in (C, \mathrm{Lip}_{\mathrm{loc}})$ on $\mathcal{D}$. Suppose $(t_0, x_0) \in \mathcal{D}$, $I \subset \mathbb{R}$ is some interval containing t_0 (which may be open or closed at either end), and suppose $x(t)$ and $y(t)$ are both solutions of the initial value problem*

$$IVP: \qquad x' = f(t, x); \quad x(t_0) = x_0$$

in $C^1(I)$. (Included in this hypothesis is the assumption that $(t, x(t)) \in \mathcal{D}$ and $(t, y(t)) \in \mathcal{D}$ for $t \in I$.) Then $x(t) \equiv y(t)$ on I.

Proof. Let $A = \{t \in I : x(t) = y(t)\}$. A is clearly a nonempty relatively closed subset of I. We show that A is open in I, from which it follows that $A = I$ as desired.

Suppose $t_1 \in A$. Set $x_1 = x(t_1) = y(t_1)$. By continuity and the openness of $\mathcal{D}$ (as $(t_1, x_1) \in \mathcal{D}$), there exist $\alpha > 0$ and $r > 0$ such that $[t_1 - \alpha, t_1 + \alpha] \times \overline{B_r(x_1)} \subset \mathcal{D}$, f is uniformly Lipschitz continuous with respect to x in $[t_1 - \alpha, t_1 + \alpha] \times \overline{B_r(x_1)}$, and $x(t) \in \overline{B_r(x_1)}$ and $y(t) \in \overline{B_r(x_1)}$ for all t in $I \cap [t_1 - \alpha, t_1 + \alpha]$. By the previous theorem, $x(t) \equiv y(t)$ in $I \cap [t_1 - \alpha, t_1 + \alpha]$. Hence A is open in I. □

Remark. t_0 is allowed to be the left or right endpoint of I.

Comparison Theorem for Nonlinear Real Scalar Equations

We conclude this section with a version of Gronwall's inequality for nonlinear equations.

Theorem. *Let $n = 1$, $\mathbb{F} = \mathbb{R}$. Suppose $f(t, u)$ is continuous in t and u and Lipschitz continuous in u. Suppose $u(t)$, $v(t)$ are C^1 for $t \ge t_0$ (or some interval $[t_0, b)$ or $[t_0, b]$) and satisfy*

$$u'(t) \le f(t, u(t)), \qquad v'(t) = f(t, v(t))$$

and $u(t_0) \le v(t_0)$. Then $u(t) \le v(t)$ for $t \ge t_0$.

Proof. By contradiction. If $u(T) > v(T)$ for some $T > t_0$, then set

$$t_1 = \sup\{t : t_0 \leq t < T \quad \text{and} \quad u(t) \leq v(t)\}.$$

Then $t_0 \leq t_1 < T$, $u(t_1) = v(t_1)$, and $u(t) > v(t)$ for $t > t_1$ (using continuity of $u - v$). For $t_1 \leq t \leq T$, $|u(t) - v(t)| = u(t) - v(t)$, so we have

$$(u - v)' \leq f(t, u) - f(t, v) \leq L|u - v| = L(u - v).$$

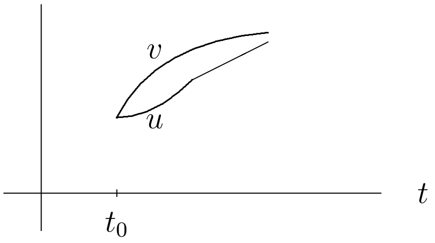
By Gronwall's inequality (applied to $u - v$ on $[t_1, T]$, with $(u - v)(t_1) = 0$, $a(t) \equiv L$, $b(t) \equiv 0$), $(u - v)(t) \leq 0$ on $[t_1, T]$, a contradiction. □

Remarks.

1. As with the differential form of Gronwall's inequality, a solution of the differential inequality $u' \leq f(t, u)$ is bounded above by the solution of the equality (i.e., the DE $v' = f(t, v)$).

2. It can be shown under the same hypotheses that if $u(t_0) < v(t_0)$, then $u(t) < v(t)$ for $t \geq t_0$ (problem 4 on Problem Set 1).

3. Caution: It may happen that $u'(t) > v'(t)$ for some $t \geq t_0$. It is not true that $u(t) \leq v(t) \Rightarrow u'(t) \leq v'(t)$, as illustrated in the picture below.



Corollary. Let $n = 1$, $\mathbb{F} = \mathbb{R}$. Suppose $f(t, u) \leq g(t, u)$ are continuous in t and u, and one of them is Lipschitz continuous in u. Suppose also that $u(t)$, $v(t)$ are C^1 for $t \geq t_0$ (or some interval $[t_0, b)$ or $[t_0, b]$) and satisfy $u' = f(t, u)$, $v' = g(t, v)$, and $u(t_0) \leq v(t_0)$. Then $u(t) \leq v(t)$ for $t \geq t_0$.

Proof. Suppose first that g satisfies the Lipschitz condition. Then $u' = f(t, u) \leq g(t, u)$. Now apply the theorem. If f satisfies the Lipschitz condition, apply the first part of this proof to $\widetilde{u}(t) \equiv -v(t)$, $\widetilde{v}(t) \equiv -u(t)$, $\widetilde{f}(t, u) = -g(t, -u)$, $\widetilde{g}(t, u) = -f(t, -u)$. □

Remark. Again, if $u(t_0) < v(t_0)$, then $u(t) < v(t)$ for $t \geq t_0$.