ICME Linear Algebra Refresher Course
Lecture 1: Preliminaries

Ron Estrin

September 20, 2016

Introduction

This course: A short refresher on linear algebra, meant to prepare you for *CME 302*, *CME 200*, or other courses involving linear algebra.

Prerequisites: Some level of exposure to linear algebra in your undergrad career.

Hopefully most of what you'll see is review, but if we're ever going too fast (or slow), **ask a question!**

When: Tues/Wed/Thurs, 10:30am - 11:45am.

Slides and material accessible at:
`http://stanford.edu/~tym1/refresher/index.html`.

Much of the material is shamelessly re-used from offerings of previous years (in particular, Victor Minden's slides 2014 slides).

Useful Resources

- Matrix Computations 3ed by Golub and Van Loan.
  There's also a 4th ed available. An encyclopedia of nearly everything you need to know, but not particularly light-reading material.
- Numerical Linear Algebra by Trefethen and Bau.
  Easier to read book with many useful exercises and a more 'conversational' tone.
- A First Course in Numerical Methods by Chen Greif and Uri Ascher.
  Broader focus than just numerical linear algebra, but good for first-time exposure to computational aspects of linear algebra.

Introduction

A little about me:
I'm a third-year ICME PhD student working in linear algebra and optimization.

Email: `restrin@stanford.edu`.
Webpage: `http://stanford.edu/~restrin`.

Let's begin!

Vector Spaces

Definition

A **vector space** is a set V and field $\mathbb{F}$ with a binary operation addition ($u + v = w \in V$ for all $u, v \in V$), and scalar multiplication ($\alpha u = v \in V$ for all $u \in V$, $\alpha \in \mathbb{F}$) such that the following axioms hold:

- Commutativity: $u + v = v + u$
- Associativity: $(u + v) + w = u + (v + w)$
- Additive identity: There exists $0 \in V$ s.t. $v + 0 = v$, $\forall v \in V$
- Additive inverse: $\forall v \in V$ there exists $w \in V$ s.t. $v + w = 0$
- Multiplicative identity: There exists $1 \in \mathbb{F}$ s.t. $1v = v$, $\forall v \in V$
- Distributativity: $\alpha(u + v) = \alpha u + \alpha v$ and $(\alpha + \beta)v = \alpha v + \beta v$

Subspaces

Definition

A **subspace** U of a vector space V (with the field $\mathbb{F}$) is a subset $U \subseteq V$ such that $0 \in U$ and

- Closed under addition: $u + v \in U$ for all $u, v \in U$
- Closed under scalar multiplication: $\alpha v \in U$ for all $v \in U$

Important: A subspace is itself a vector space.

Examples of vector spaces

- **Euclidean space**: (Everyone's favourite) $\mathbb{R}^n$ or $\mathbb{C}^n$ (columns of numbers).
  Example subspace: Choose set $\alpha_i \in \mathbb{R}$, then
  $U = \{x \in \mathbb{R}^n \mid \sum_{i=n}^{n} \alpha_i x_i = 0\}$ is a subspace.
  Generally we'll discuss Euclidean space with either $\mathbb{R}^n$ or $\mathbb{C}^n$.

- **Continuous real-valued functions on [0,1]**.
  Example subspaces: Polynomials of degree $\leq n$ ($\mathbb{P}_n(x)$),
  $U = \{f \in \mathcal{C}[0,1] \mid f(0) = f(1) = 0\}$.

Span of vectors

Definition

The **span** of a set of vectors is the subspace of all linear linear combinations of those vectors

$$\text{span}\{v_1, \ldots, v_k\} = \left\{ w \mid w = \sum_{i=1}^{n} \alpha_i v_i \right\}.$$

Examples:

- $$\text{span}\left\{ \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \right\} = \left\{ \begin{pmatrix} \alpha_1 \\ 0 \\ \alpha_2 \end{pmatrix} \mid \alpha_1, \alpha_2 \in \mathbb{R} \right\}.$$

- $$\text{span}\left\{ \{x^{2k} \mid k \in \mathbb{N}\} \right\} = \{\text{Polynomials with even degree terms}\}$$

Linear independence/dependence

Definition

A set of vectors $\{v_i\}_{i=1}^n$ is **linearly independent** if

$$\sum_{i=1}^{n} \alpha_i v_i = 0 \implies \alpha_i = 0, i = 1, \ldots, n.$$

Otherwise, the set is **linearly dependent**.

Linearly dependent sets are *redundant*, since we can represent any vector (if $\alpha_j \neq 0$) as

$$v_j = \frac{1}{\alpha_j} \sum_{i \neq j} \alpha_i v_i.$$

Linear independence/dependence

Examples:

- The set $\{v_1, v_2, v_3\} = \left\{ \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} \right\}$ is linearly dependent since $v_1 + v_2 - v_3 = 0$.
- The set $\{v_1, v_2, v_3\} = \left\{ \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} \right\}$ is linearly independent.

Bases

Definition

A set of vectors $\{v_i\}_{i=1}^n$ **generates** a vector space U if $\operatorname{span}\{v_i\} = U$.

Definition

A set of vectors $\{v_i\}_{i=1}^n$ is a **basis** for a vector space U if $\operatorname{span}\{v_i\} = U$ and the set $\{v_i\}_{i=1}^n$ is linearly independent.

With a basis, we can express any $u \in U$ in the basis $\{v_i\}_{i=1}^n$ as

$$u = \sum_{i=1}^{n} \alpha_i v_i,$$

for some coefficients α_i.

Dimension of a vector space

Definition

The **dimension** of a vector space V is the number of vectors in any fixed basis of V,

$$\dim(V) = |\text{vectors in basis of } V| .$$

Remember: The dimension depends only on the vector space, not on the basis!

Not all vector spaces are finite dimensional (e.g. space of continuous functions), but for numerical linear algebra, we'll generally only care about the finite-dimensional ones.

Example bases

One basis for $\mathbb{P}_n(x)$ is the set of monomials $\{1, x, x^2, \dots, x^n\}$.

Another basis for the same space is the set of Chebyshev polynomials of the first kind, $\{P_0, P_1, \dots, P_n\}$:

$$
\begin{aligned}
P_0(x) &= 1 \\
P_1(x) &= x \\
P_n(x) &= 2xP_{n-1}(x) - P_{n-2}(x).
\end{aligned}
$$

In both cases, the cardinality of the basis sets is $n+1$, so the dimension of the space is $n+1$.

Although they span the same space, these bases have very different properties!

Definition

An **inner product space** is a vector space V with a defined inner product

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F},$$

such that the following properties hold:

- Conjugate symmetry: $\langle u, v \rangle = \overline{\langle v, u \rangle}$
- Linearity in first argument: $\langle \alpha u + v, w \rangle = \alpha \langle u, w \rangle + \langle v, w \rangle$.
- Positive-Definiteness: $\langle u, u \rangle \geq 0$ with equality iff $u = 0$.

Formal definition for 'products of vectors'.

Examples:

- **Dot-product for $\mathbb{C}^n$:** (Everyone's favourite) Defined as

$$\langle u, v \rangle = \sum_{i=1}^{n} \bar{u}_i v_i = v^* u.$$

  Also known as the ℓ_2 inner product.
- **L^2 inner product for functions on $[0,1]$.** Defined as

$$\langle f, g \rangle_{L^2} = \int_0^1 f(x)\overline{g(x)}dx.$$

Norms

Definition

A **norm** on a vector space V is a function $\|\cdot\| : V \to \mathbb{R}_+ \cup 0$ such that the following properties hold:

- Absolute homogeneity: $\|\alpha v\| = |\alpha| \, \|v\|$ for all $\alpha \in \mathbb{F}$ and $v \in V$
- Sub-additivity (triangle inequality): $\|u + v\| \leq \|u\| + \|v\|$
- Nondegeneracy: $\|v\| = 0$ iff $v = 0$

Norms generalize the idea "length" of vectors. All norms are convex functions.

Examples:

- **Euclidean norm**: Defined as

$$\|u\|_2 = \left(\sum_{i=1}^{n} |u_i|^2\right)^{\frac{1}{2}} = u^* u.$$

  Also called the ℓ_2-norm. An example of a norm defined by an inner product.

- ℓ_p**-norm**: Defined as

$$\|u\|_p = \left(\sum_{i=1}^{n} |u_i|^p\right)^{\frac{1}{p}}.$$

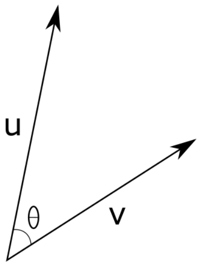
- ℓ_∞**-norm**: Define as

$$\|u\|_\infty = \max_{1\le i\le n} |u_i|.$$

Euclidean inner product and geometry

Let u and v be two vectors, with angle θ in between. The Euclidean norm is exactly the usual notion of 'length' of a vector, and the inner product satisfies

$$\langle u, v \rangle_2 = \|u\|_2 \, \|v\|_2 \cos\theta.$$



Exercises

1. Prove that every inner product defines a norm. That is, show that
$$\|u\| = (\langle u, u \rangle)^{\frac{1}{2}},$$
is a norm.
2. Prove the cosine law. If a, b, c are the sides of the triangle, and θ is the angle between a and b, then
$$|c|^2 = |a|^2 + |b|^2 - 2|a||b|\cos\theta.$$

Important inequalities

Triangle Inequality:

$$\|u+v\| \leq \|u\| + \|v\| .$$

Reverse Triangle Inequality:

$$\|u-v\| \geq |\|u\| - \|v\|| .$$

Cauchy-Schwarz Inequality: Let the norm $\|\cdot\|$ be induced by the inner product $\langle \cdot, \cdot \rangle$. Then

$$|\langle u, v \rangle| \leq \|u\| \, \|v\| \, .$$

It basically says that the size of the inner product is bounded by the product of the size of the vectors themselves.

Important inequalities

Cauchy-Schwarz Inequality: Let the norm $\|\cdot\|$ be induced by the inner product $\langle \cdot, \cdot \rangle$. Then

$$|\langle u, v \rangle| \leq \|u\| \, \|v\| \, .$$

It basically says that the size of the inner product is bounded by the product of the size of the vectors themselves.

Recall that for the Euclidean inner product,

$$\langle u, v \rangle_2 = \|u\|_2 \, \|v\|_2 \cos \theta.$$

Since $0 \leq |\cos \theta| \leq 1$, Cauchy-Schwarz clearly holds, and we can observe when sharpness occurs: when $\theta = 0$.

Definition

Vectors u, v are **orthogonal** with respect to an inner product if

$$\langle u, v \rangle = 0.$$

For the Euclidean inner product, this is the usual notion of orthogonality, i.e. two vectors are orthogonal if $\theta = \pi/2$ since

$$\langle u, v \rangle_2 = \|u\|_2 \|v\|_2 \cos\theta.$$

Definition

An **orthogonal basis** is a basis $\{v_i\}_{i=1}^n$ such that $\langle v_i, v_j \rangle = 0$ for $i \neq j$. A basis is **orthonormal** if it is orthogonal and additionally, $\langle v_i, v_i \rangle = 1$ for all i.

Definition

Two vector spaces U and V are orthogonal if $\langle u, v \rangle = 0$ for all $u \in U$, $v \in V$.

Orthonormal bases

Orthonormal bases $\{q_i\}_{i=1}^n$ are really nice for several reasons. Consider computing the inner product of $u = \sum_{i=1}^n \alpha_i q_i$ and $v = \sum_{j=1}^n \beta_j q_j$.

$$
\begin{aligned}
\langle u, v \rangle &= \left\langle \sum_{i=1}^n \alpha_i q_i, \sum_{j=1}^n \beta_j q_j \right\rangle \\
&= \sum_{i=1}^n \alpha_i \left\langle q_i, \sum_{j=1}^n \beta_j q_j \right\rangle \\
&= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \bar{\beta}_j \langle q_i, q_j \rangle \\
&= \sum_{i=1}^n \alpha_i \bar{\beta}_i.
\end{aligned}
$$

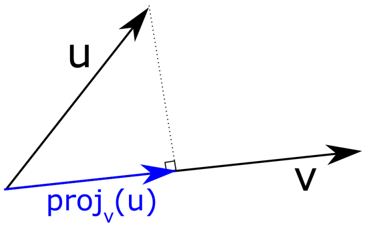
We can compute the norm $\|u\|^2 = \sum_{i=1}^n |\alpha_i|^2$.

Vector projection

The **vector projection** of a vector u onto a vector v is

$$\mathrm{proj}_v(u) = \frac{\langle u, v\rangle}{\langle v, v\rangle} v = (\|u\| \cos\theta)\, \hat{v},$$

where $\hat{v} = v / \|v\|$.
The projection points along v, with magnitude equal to the inner product between u and v.



Gram-Schmidt Process

The **Gram-Schmidt Process** is a way to form an orthonormal basis $\{v_i\}_{i=1}^n$ from a set of vectors $\{u_i\}_{i=1}^n$, so that

$$\operatorname{span}\{v_1, \ldots, v_k\} = \operatorname{span}\{u_1, \ldots, u_k\}$$

for all k. The main idea is that given an orthonormal basis $\{v_1, \ldots, v_{k-1}\}$, then

$$u_k - \operatorname{proj}_{v_1}(u_k) - \cdots - \operatorname{proj}_{v_{k-1}}(u_k) \perp \operatorname{span}\{v_1, \ldots, v_{k-1}\}.$$

Pretty much everything we've talked about for vectors so far applies to matrices:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \in \mathbb{F}^{m \times n}$$

Operations on Matrices

As long as your dimensions make sense, you can:

- Add/Scalar multiply: $C = \alpha A + B \iff c_{ij} = \alpha a_{ij} + b_{ij}$
- Transpose: $A^T \in \mathbb{F}^{n\times m}$ where $(A^T)_{ij} = a_{ji}$
- (Complex) Adjoint: $A^* \in \mathbb{C}^{n\times m}$ where $(A^*)_{ij} = \bar{a}_{ji}$
- Multiply vector by matrix:

$$(Ax)_i = \sum_{j=1}^{n} a_{ij} x_j$$

- Multiply matrix by matrix:

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Matrix-Vector multiplication

Few ways to think about it:

- As dot-products with rows,

$$Ax = \begin{pmatrix} -r_1^T- \\ -r_2^T- \\ \vdots \\ -r_m^T- \end{pmatrix} x = \begin{pmatrix} r_1^T x \\ r_2^T x \\ \vdots \\ r_m^T x \end{pmatrix}$$

- As linear combination of columns,

$$Ax = \begin{pmatrix} | & | & & | \\ c_1 & c_2 & \dots & c_n \\ | & | & & | \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$= x_1 \begin{pmatrix} | \\ c_1 \\ | \end{pmatrix} + x_2 \begin{pmatrix} | \\ c_2 \\ | \end{pmatrix} + \cdots + x_n \begin{pmatrix} | \\ c_n \\ | \end{pmatrix}$$

Matrix-Vector multiplication

Row vector multiplication is similar but reversed:

- As linear combination of rows,

$$x^T A = \begin{pmatrix} x_1 & x_2 & \cdots & x_m \end{pmatrix} \begin{pmatrix} -r_1^T- \\ -r_2^T- \\ \vdots \\ -r_m^T- \end{pmatrix} = \sum_{i=1}^{m} x_i r_i^T$$

- As dot-products with columns,

$$\begin{aligned} x^T A &= \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} | & | & & | \\ c_1 & c_2 & \dots & c_n \\ | & | & & | \end{pmatrix} \\ &= \begin{pmatrix} x^T c_1 & x^T c_2 & \cdots & x^T c_n \end{pmatrix} \end{aligned}$$

Matrix-Matrix products

Approximately 171985318 different ways to think about AB, so pick whatever is most convenient:

- The usual entry-wise dot-product approach
- A applied to columns of A
- B applied to rows of A
- As a sum of **outer products**:

$$AB = \begin{pmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{pmatrix} \begin{pmatrix} -b_1^T- \\ -b_2^T- \\ \vdots \\ -b_n^T- \end{pmatrix} = a_1 b_1^T + \ldots a_n b_n^T$$

- Blocking, ...

Special Matrices

- **Diagonal**, **Triangular**
- **Orthogonal (Real)** and **Unitary (Complex)**: $Q^*Q = I$
- **Symmetric (Real)** and **Hermitian (Complex)**: $A^* = A$
- **Normal**: $AA^* = A^*A$
- **Symmetric (or Hemitian) Positive Definite**: $A = A^*$ and $x^*Ax > 0$ for $x \neq 0$. Also called SPD or HPD.
- **Projection**: $P^2 = P$
- **Rotation**:
$$R(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$
Rotates a point by an angle θ. Note: Also orthogonal.

SPD Matrices

SPD matrices induce an inner product and therefore a norm as well (often called the energy-norm):

$$\langle u, v \rangle_A = u^T A v$$
$$\|u\|_A = \sqrt{\langle u, u \rangle_A}$$

These kinds of inner products come up quite often in applications, as well as in some methods for solving linear systems (e.g. the conjugate gradient algorithm). Keep it in mind!

Matrix norms

Same definition as vector norms (homogeneity, sub-additivity, nondegeneracy). Additional common property is **sub-multiplicativity** (but not required): $\|AB\| \leq \|A\| \|B\|$.

- **Induced norm**: Given norm $\|\cdot\|$ on vector, can define matrix norm as

$$\|A\| = \sup_{\|x\|=1} \|Ax\| .$$

  Can define ℓ_p-norm on matrices this way.
- **Frobenius norm**:

$$\|A\|_F = \sqrt{\sum_{i,j} |A_{ij}|^2} = \operatorname{tr}(A^* A)$$

- **Max norm**:

$$\|A\|_{\max} = \max_{ij} |A_{ij}|$$

  This norm is **not** sub-multiplicative.

Matrix norms

Some induced norms have simpler expressions.

- ℓ_1**-norm**
$$\|A\|_1 = \text{max absolute column-sum.}$$
- ℓ_∞**-norm**
$$\|A\|_\infty = \text{max absolute row-sum.}$$

Linear Transformations

Just like vector spaces aren't just columns of numbers, linear transformations are more than just matrices.

Definition

A **linear transformation** from a vector space V to vector space U is a map $T : V \to U$ such that for all $v_1, v_2 \in V$

$$T(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 T(v_1) + \alpha_2 T(v_2).$$

Note that this implies that $T(0) = 0$ for any linear transformation.

Examples:

- Matrix-vector multiplication
- Differentiation of differentiable functions.

Linear Transformations

Any linear transformation $T : V \to U$ on a finite-dimensional vector space can be expressed as a matrix once a basis for V and U is decided upon.

So if transformation T_i is expressed by the matrix A_i, function composition $T_2(T_1(\cdot))$ is just matrix multiplication $A_2 \cdot A_1$!

Exercise

1. Prove the sin addition identity:

$$\sin(\theta + \phi) = \sin\theta\cos\phi + \cos\theta\sin\phi.$$

   Recall that a rotation matrix is

$$R(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

2. Verify that if A is SPD, then $\langle u, v\rangle_A = u^T A v$ is a valid inner product.

The range of a matrix

Definition

Let $A \in \mathbb{R}^{m \times n}$. The **range** (or "column-space") is a subspace of $\mathbb{R}^m$ given by

$$\begin{aligned} \mathcal{R}(A) &= \{Ax \mid x \in \mathbb{R}^n\} \\ &= \operatorname{span}\{\text{columns of } A\} \end{aligned}$$

This is indeed a subspace. Note that if $b \notin \mathcal{R}(A)$, then no x exists such that $Ax = b$.

The rank of a matrix

Definition

The **rank** of a matrix A is the dimension of its range: $\operatorname{rank}(A) = \dim(\mathcal{R}(A))$.

Theorem:
The dimension of the column space of A is the same as the dimension of the column space of A^T (the row-space),

$$\operatorname{rank}(A) = \operatorname{rank}\left(A^T\right).$$

The null-space of a matrix

Definition

Let $A \in \mathbb{R}^{m \times n}$. The **null-space** (or "kernel") is a subspace of $\mathbb{R}^n$ given by

$$\ker(A) = \{x \in \mathbb{R}^n \mid Ax = 0\}.$$

The dimension of the null-space, $\dim(\ker(A))$ is called the **nullity** of A.

The rank-nullity theorem

Theorem

Let $A \in \mathbb{R}^{m \times n}$. Then

$$\operatorname{rank}(A) + \dim(\ker(A)) = \dim(\mathbb{R}^n) = n.$$

This is the **rank-nullity theorem**.

The four fundamental subspaces

Definition

Let $A \in \mathbb{R}^{m \times n}$. The **four fundamental subspaces** of A are the range and null-spaces of A and A^T:

- $\mathcal{R}(A) \subseteq \mathbb{R}^m$
- $\ker(A) \subseteq \mathbb{R}^n$
- $\mathcal{R}(A^T) \subseteq \mathbb{R}^n$
- $\ker(A^T) \subseteq \mathbb{R}^m$

Theorem: $\mathcal{R}(A)$ and $\ker(A^T)$ are orthogonal ($\mathcal{R}(A) \perp \ker(A^T)$) with respect to the ℓ_2 inner product.

Proof: Let $v \in \mathcal{R}(A)$ and $u \in \ker(A^T)$. Then $v = Aw$ for some w, and

$$v^T u = w^T A^T u = w^T (A^T u) = 0.$$

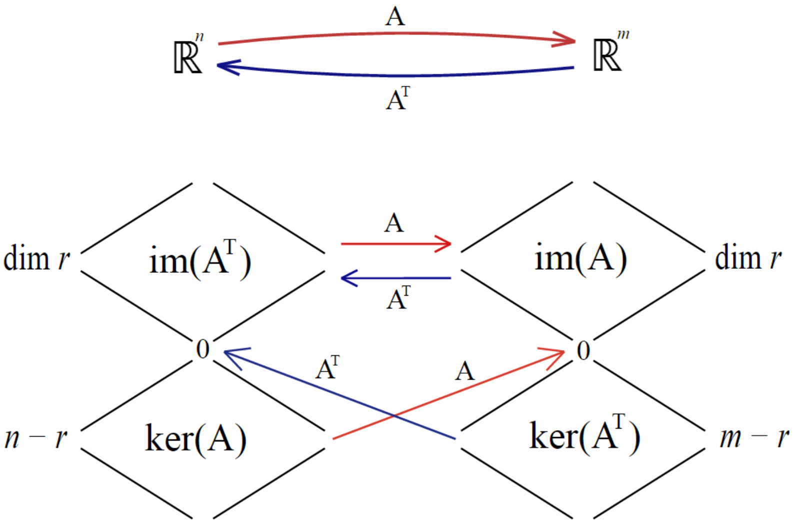
The Fundamental Theorem of Linear Algebra

Theorem

Let $A \in \mathbb{R}^{m \times n}$. The four fundamental subspaces satisfy

$$\mathcal{R}(A) \perp \ker(A^T) \text{ and } \mathcal{R}(A) \cup \ker(A^T) = \mathbb{R}^m$$
$$\mathcal{R}(A^T) \perp \ker(A) \text{ and } \mathcal{R}(A^T) \cup \ker(A) = \mathbb{R}^n$$

Figure next page: The four subspaces by Cronholm144 - Own work. Licensed under Creative Commons Attribution-Share Alike 3.0 via Wikimedia Commons.


A
$\mathbb{R}^n$
$\mathbb{R}^m$
A^T
dim r
im(A^T)
A
A^T
im(A)
dim r
0
A^T
A
0
$n - r$
ker(A)
ker(A^T)
$m - r$

The determinant

The **determinant** is a function of square matrices with a gross entry-wise formula (best to look it up).

Properties:

- $\det(A) = \det(A^T)$
- $\det(I) = 1$, and if Q is orthogonal, then $\det(Q) = \pm 1$
- $\det(AB) = \det(A)\det(B)$
- $\det(A) = 0 \implies \dim(\ker(A)) \geq 1$
- $\det(c \cdot A) = c^n \cdot \det(A)$ for $n \times n$ matrices
- $\det(L) = \prod_{i=1}^{n} l_{ii}$ if L is triangular
- Intuition: $\det(A)$ is the volume of the parallelepiped formed by the columns (or rows) of A.

The trace

The **trace** is a function of square matrices defined as

$$\mathrm{tr}\,(A) = \sum_{i=1}^{n} a_{ii}.$$

Properties:

- $\mathrm{tr}\,(A + B) = \mathrm{tr}\,(A) + \mathrm{tr}\,(B)$
- $\mathrm{tr}\,(c \cdot A) = c \cdot \mathrm{tr}\,(A)$
- $\mathrm{tr}\,(ABC) = \mathrm{tr}\,(CAB) = \mathrm{tr}\,(BCA)$

ICME Linear Algebra Refresher Course
Lecture 2: Solving Linear Systems

Ron Estrin

September 22, 2016

Focus of this lecture: Given matrix $A \in \mathbb{R}^{m \times n}$ and vector $b \in \mathbb{R}^m$, want to find $x \in \mathbb{R}^n$ such that

$$Ax = b, (\text{ or } Ax \approx b).$$

Focus of this lecture: Given matrix $A \in \mathbb{R}^{m \times n}$ and vector $b \in \mathbb{R}^m$, want to find $x \in \mathbb{R}^n$ such that

$$Ax = b, (\text{ or } Ax \approx b).$$

We have 3 cases:

- No solution: $b \notin \mathcal{R}(A)$. The system is **inconsistent**.
- Infinitely many solutions: $\ker(A)$ is nontrivial. The system is **ill-posed.**
- Exactly one solution: Everything else.

Applications

- Solving PDEs via finite difference or finite elements.

$$\text{e.g. 1D: } -\nabla^2 u = f \implies -\frac{u_{i-1} - 2u_i + u_{i+1}}{2} = f_i.$$

- Least-squares fitting: Fitting n parameters of linear model to $m \gg n$ datapoints.
- Computational kernel for solving optimization problems.

$$\begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}.$$

The matrix inverse

Definition

Let A be an $n \times n$ square matrix. A matrix A is **invertible** if there exists a matrix A^{-1} such that

$$A^{-1}A = AA^{-1} = I.$$

The matrix inverse

Definition

Let A be an $n \times n$ square matrix. A matrix A is **invertible** if there exists a matrix A^{-1} such that

$$A^{-1}A = AA^{-1} = I.$$

Some comments:

- The inverse is **unique**.
- The inverse doesn't always exist. Matrices without inverses are called **singular**.
- Non-square matrices do not have inverses, but there are suitable generalizations.

The Invertible Matrix Theorem

The following statements are equivalent:

- A is invertible (has an inverse).
- $\det(A) \neq 0$.
- A has full rank, $\text{rank}(A) = n$.
- $Ax = 0$ has only the solution $x = 0$.
- $\ker(A) = \{0\}$.
- $Ax = b$ has exactly one solution for each b.
- The columns/rows of A are linearly independent.
- 0 is not an eigenvalue of A.
- The columns/rows of A form a basis for $\mathbb{R}^n$.

... and more.

Exercises

Let A be a square matrix.

1. Prove that the left- and right-inverses of A are the same (if $AB = I$ and $CA = I$, then $B = C$). Then prove that the inverse is unique.

2. Prove that if A has full-rank, then an inverse exists.

Solving Nonsingular Matrices

Let's focus on square nonsingular matrices first: $Ax = b$.
Since A is nonsingular, it has an inverse and so

$$x = A^{-1}b.$$

Theoretically, we can compute A^{-1} and apply it to b. This is in general a bad idea:

- Computing A^{-1} in finite precision can incur a lot of numerical error (too **inaccurate**).
- If we only want to solve one rhs, then computing A^{-1} may result in unnecessary extra work (too **too slow**).

Direct Solvers for $Ax = b$

Instead of inverting A and multiplying b, **direct solvers** use the following strategy:

1. Factor $A = A_1 A_2 \dots A_k$ into a product of easy to solve matrices A_i.
2. Set $x^{(1)} = b$, and solve $A_i x^{(i+1)} = x^{(i)}$, until we get $x = x^{(k)}$.

Classically we have $k = 2, 3$ factors (although some of the modern approaches can have k very large).

Easy to Solve Matrices

Some easy to solve matrices:

- Diagonal matrices:

$$Dx = b \implies x_i = b_i / D_{ii}.$$

- Unitary matrices:

$$Qx = b \implies x = Q^* b.$$

- Permutation matrices ($Pe_i = e_{\pi(i)}$):

$$Px = b \implies x_i = b_{\pi^{-1}(i)}.$$

- Lower (or upper) triangular matrices.

Triangular Matrices

Want to solve

$$\begin{pmatrix} l_{11} & & & & \\ l_{21} & l_{22} & & & \\ l_{31} & l_{32} & l_{33} & & \\ \vdots & & & \ddots & \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix}$$

Can solve this via **forward-substitution**:

$$\begin{aligned} x_1 &= b_1 / l_{11}, \\ x_2 &= \left(b_1 - l_{21} x_1\right) / l_{22}, \\ &\vdots \\ x_n &= \left(b_n - l_{n1} x_1 - l_{n2} x_2 - \cdots - l_{n,n-1} x_{n-1}\right) / l_{nn}. \end{aligned}$$

Triangular Matrices

- L is nonsingular as long as all diagonal entries are nonzero, so the process will not fail.
- The process is analogous for upper-triangular matrices. It is called **backward-substitution**, which starts from the bottom and works its way up.
- Computationally cheap: $O(n^2)$ flops to solve.

Common factorizations useful for solving linear systems:
LT: lower-triangular, UT: upper-triangular, Prm: permutation, Orth: Orthogonal, Diag: Diagonal.

- **LU**: $A = LU$
  L is LT, U is UT.
- **Partial-pivoted LU**: $P_1 A = LU$.
  P_1 is Prm, L is LT, U is UT.
- **Complete-pivoted LU**: $P_1 A P_2 = LU$.
  P_1, P_2 are Prm, L is LT, U is UT.
- **QR**: $A = QR$.
  Q is Orth, R is UT.
- **SVD**: $A = U\Sigma V^*$.
  U, V are Orth, Σ is Diag.

Gaussian Elimination for solving $Ax = b$:

- Perform elementary row operations to turn $[A|b] \to [U|y]$ where U is upper triangular.
- Perform backward substitution on $Ux = y$.

Elementary row operations:

- Scale a row.
- Add a multiple of one row to another.
- Permute rows.

Gaussian Elimination on 3×3 Matrix

1. Form augmented system $[A|b]$.

$$\left[\begin{array}{ccc|c} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{array}\right]$$

2. Perform elementary row operation to introduce zeros below the diagonal.

$$\left[\begin{array}{ccc|c} \boxed{\times} & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{array}\right] \xrightarrow{L_1} \left[\begin{array}{ccc|c} \times & \times & \times & \times \\ 0 & + & + & + \\ 0 & + & + & + \end{array}\right]$$

$$\left[\begin{array}{ccc|c} \times & \times & \times & \times \\ 0 & \boxed{\times} & \times & \times \\ 0 & \times & \times & \times \end{array}\right] \xrightarrow{L_2} \left[\begin{array}{ccc|c} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & + & + \end{array}\right] = [U|y]$$

3. Solve the system $Ux = y$ by back-substitution.

Gaussian Elimination

Note that we didn't pivot in the previous example, which may be necessary if a zero-pivot (or a really small one) occurs.

This is in theory how to compute the LU factorization.

- U is the resulting upper triangular matrix.
- If we keep track of our row-operations, this would form the L factor.
- You'll be going through this in gory detail in CME 302...

Non-square or Singular Systems

We have two cases for $Ax = b$:

- $b \notin \mathcal{R}(A)$ and there is no solution.
  - Perhaps $A \in \mathbb{R}^{m \times n}$, $m > n$ (tall-skinny or overdetermined)
  - Perhaps A is singular
- $\ker(A)$ is nontrivial, and there are infinitely many solutions.
  - Perhaps $A \in \mathbb{R}^{m \times n}$, $m < n$ (short-fat or underdetermined)

Overdetermined Systems

We'll set aside the case where A is rank-deficient for now.

Suppose that $m > n$ and $b \notin \mathcal{R}(A)$. Need a sense of what a "good" solution is.

Overdetermined Systems

We'll set aside the case where A is rank-deficient for now.

Suppose that $m > n$ and $b \notin \mathcal{R}(A)$. Need a sense of what a "good" solution is.

We can solve a **Least-squares problem**:

$$\min_x \|Ax - b\|_2.$$

Define $r = b - Ax$ as the **residual**.

Least-Squares problem

Suppose we have m data points (x_i, y_i), and we want to fit an $n - 1 < m$ degree polynomial

$$f(x) = a_1 + a_2 x + a_3 x^3 + \cdots + a_n x^{n-1}$$

to the data. This results in the **Vandemonde** matrix

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^{n-1} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

Solving Least-Squares problems

If $A \in \mathbb{R}^{m\times n}$ is tall-skinny but full rank ($\text{rank}(A) = n$), we can find least-squares solution explicitly.

$$\begin{bmatrix} \\ A \\ \\ \end{bmatrix} x = \begin{bmatrix} \\ b \\ \\ \end{bmatrix} \implies \begin{bmatrix} & A^T & \end{bmatrix} \begin{bmatrix} \\ A \\ \\ \end{bmatrix} x = A^T A x = \begin{bmatrix} & A^T & \end{bmatrix} \begin{bmatrix} \\ b \\ \\ \end{bmatrix}.$$

Notice that $A^T A$ is square and nonsingular. The system on the right called the **Normal equations**. The least-squares solution is:

$$x_{LS} = (A^T A)^{-1} A^T b.$$

IMPORTANT: Forming the normal equations and solving them is usually a very bad idea due to numerical errors. You'll see proper ways of solving LS problems your classes.

Exercises

1. Show that $P = A(A^T A)^{-1} A^T$ is an (orthogonal) projector. Recall that this means that $P^2 = P$ and $P = P^T$. What space does this operator project onto?

2. The least-squares solution is $x_{LS} = (A^T A)^{-1} A^T b$, and the residual is $r = b - Ax_{LS}$. How does the residual relate to $\mathcal{R}(A)$?

Alternative approaches for over-determined systems

Other common approaches include minimizing residual in the ℓ_1 or ℓ_∞ norm. These don't have closed-form solutions; they are linear programs.

Minimizing in the ℓ_1 norm promotes sparsity in the residual (few non-zero entries).

Minimizing in the ℓ_∞ norm promotes all of the residual entries to be roughly the same size (but small).

Moore-Penrose Pseudo-inverse

Definition

Let A be an $m \times n$ matrix. The **Moore-Penrose pseudoinverse** is an $n \times m$ matrix $A^\dagger$ which satisfies

$$AA^\dagger A = A \qquad (AA^\dagger)^T = AA^\dagger$$
$$A^\dagger AA^\dagger = A^\dagger \qquad (A^\dagger A)^T = A^\dagger A.$$

When A is tall and skinny, $A^\dagger = (A^T A)^{-1} A^T$.
This means that $x_{LS} = A^\dagger b$.

When A is short and fat, $A^\dagger = (AA^T)^{-1} A$. This version will also play a role soon.

Underdetermined systems

Suppose instead we have $m < n$, $b \in \mathcal{R}(A)$, so that

$$\begin{bmatrix} & A & \end{bmatrix} x = b.$$

If $A\hat{x} = b$, and $z \in \ker(A)$, then $A(\hat{x} + z) = b$! We have infinitely many solutions so how can we choose?

Underdetermined systems

Suppose instead we have $m < n$, $b \in \mathcal{R}(A)$, so that

$$\begin{bmatrix} & A & \end{bmatrix} x = b.$$

If $A\hat{x} = b$, and $z \in \ker(A)$, then $A(\hat{x} + z) = b$! We have infinitely many solutions so how can we choose?

We can solve a **minimum norm problem**:

$$\min_x \|x\|_2 \text{ s.t. } Ax = b.$$

Underdetermined systems

Suppose instead we have $m < n$, $b \in \mathcal{R}(A)$, so that

$$\begin{bmatrix} & A & \end{bmatrix} x = b.$$

If $A\hat{x} = b$, and $z \in \ker(A)$, then $A(\hat{x} + z) = b$! We have infinitely many solutions so how can we choose?

We can solve a **minimum norm problem**:

$$\min_x \|x\|_2 \text{ s.t. } Ax = b.$$

The solution is $x = A^\dagger b$ again! (But don't form the normal equations!)

Inconsistent and Singular systems

What if $Ax = b$ is both a singular and inconsistent system (or A has bad numerical properties)?

Typical approaches blend the two ideas we've covered via **regularization**:

$$\min_x \|Ax - b\|_2^2 + \lambda^2 \|x\|_2^2. \iff \min_x \left\| \begin{pmatrix} A \\ \lambda I \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2^2.$$

λ is a parameter which controls the trade-off between agreement with the data and numerical stability.

High Level View to Solving Linear Systems

Two approaches for $Ax = b$: **Direct** and **Iterative** methods.

Direct Methods:

- Factor A into easy to solve matrices and solve against each one.
- e.g. LU, QR, SVD ...
- Good for solving many right-hand sides efficiently (factor once, solve many times).
- Need matrix explicitly.

High Level View to Solving Linear Systems

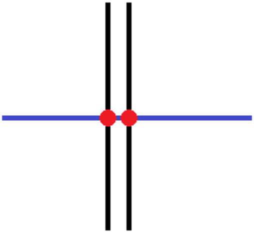
Two approaches for $Ax = b$: **Direct** and **Iterative** methods.

Iterative Methods:

- Stationary Methods:
  - Update process $x_{k+1} \leftarrow G(x_k)$ to successively approximate solution. G is some function satisfying certain conditions.
  - e.g. Jacobi, Gauss-Seidel, Successive Over-Relaxation
- Search Methods:
  - Generate search space for solution, then approximate solution within the search space by solving minimization problem.
  - Example minimization problem: minimize residual
  - e.g. Krylov subspace methods: CG, MINRES, GMRES ...
  - Requires only matrix-vector products with A.

Conditioning and Stability

Consider a 2×2 system $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}$. This is the intersection of two lines (blue and black), solution is red dot. Suppose we perturb the black line $c \to c + \Delta c$, $d \to d + \Delta d$, and $g \to g + \Delta g$.



(a) **Small** perturbation to problem, **small** perturbation to solution.

(b) **Small** perturbation to problem, **large** perturbation to solution.

Conditioning and Stability

Suppose we want to solve system $Ax = b$, and we end up solving $A\hat{x} = \hat{b}$. Recall the residual is $r = b - A\hat{x} = A(x - \hat{x})$.
Then:

$$\begin{aligned} \|b\|_2 &= \|Ax\|_2 && \implies \|b\|_2 \leq \|A\|_2\|x\|_2 \\ \|x - \hat{x}\|_2 &= \|A^{-1}r\|_2 && \implies \|x - \hat{x}\|_2 \leq \|A^{-1}\|_2\|r\|_2 \end{aligned}$$

This implies:

$$\frac{\|x - \hat{x}\|_2}{\|x\|_2} \leq \|A\|_2\|A^{-1}\|_2\frac{\|r\|_2}{\|b\|_2}.$$

Thus we obtain an upper bound on the **forward error** (which we can't compute) using the residual (which we can compute).

The Condition Number

Definition

Given a nonsingular square matrix A, the quantity $\kappa(A) = \|A\|_2 \|A^{-1}\|_2$ is know as the **condition number** of A.

The condition number is a measure of how well-conditioned the matrix A (i.e. how much perturbations in the data may perturb the solution).

Rule of thumb: If the condition number is $\kappa(A) \approx 10^p$, then your computed solution loses p digits of accuracy when using direct methods. **Example:** If you have 16 digits of precision (e.g. double type), and $\kappa(A) \approx 10^6$, you typically expect 10 correct digits.

ICME Linear Algebra Refresher Course
Lecture 3: Spectral Theory of Matrices

Ron Estrin

September 23, 2016

Definition

The **resolvent** of a square matrix A is the matrix-valued mapping

$$R(z) = (A - zI)^{-1}.$$

The entries of the resolvent are rational functions of the scalar z. The resolvent fails to exist if z is a pole of any of these rational functions (i.e. if $A - zI$ becomes singular).

Definition

The **resolvent** of a square matrix A is the matrix-valued mapping

$$R(z) = (A - zI)^{-1}.$$

The entries of the resolvent are rational functions of the scalar z. The resolvent fails to exist if z is a pole of any of these rational functions (i.e. if $A - zI$ becomes singular).

Definition

A scalar λ is an **eigenvalue** of the square matrix A if $R(\lambda)$ does not exist (i.e. $A - \lambda I$ is singular). A nonzero vector v is an **eigenvector** of A associated with eigenvalue λ if $v \in \ker(A - \lambda I)$ or equivalently

$$Av = \lambda v.$$

The pair (λ, v) satisfying $Av = \lambda v$ form an **eigenpair**. The space $\ker(A - \lambda)$ is called the **eigenspace** of A associated with eigenvalue λ.

The set of eigenvalues

$$\sigma(A) = \{\lambda \in \mathbb{C} | A - \lambda I \text{ is singular. } \}$$

is called the **spectrum** of A.
The **spectral radius** is the magnitude of the largest eigenvalue (in magnitude)

$$\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$$

How big can $\rho(A)$ be?

Neumann Series

Theorem

Let A be a square matrix with $\|A\| < 1$. Then $I - A$ is nonsingular and

$$(I - A)^{-1} = \sum_{i=0}^{\infty} A^i,$$

$$\|(I - A)^{-1}\| \leq \frac{1}{1 - \|A\|}.$$

Neumann Series

Theorem

Let A be a square matrix with $\|A\| < 1$. Then $I - A$ is nonsingular and

$$(I - A)^{-1} = \sum_{i=0}^{\infty} A^i,$$

$$\|(I - A)^{-1}\| \leq \frac{1}{1 - \|A\|}.$$

If $|\lambda| > \|A\|$, then $\|A/\lambda\| < 1$,

$$\|R(\lambda)\| = \|(\lambda I - A)^{-1}\| \leq \frac{1}{|\lambda| - \|A\|} < \infty.$$

Thus $\rho(A) \leq \|A\|$!

Applications

- PageRank
- Graph Clustering
- Schrödinger's equation

Similarity Transform

Two matrices are called **similar** if

$$B = X^{-1}AX$$

where X is a nonsingular matrix.

A and B can be viewed as 'same' linear transformation under different bases. A and B have the **same** eigenvalues. If v is an eigenvector of A, then $X^{-1}v$ is an eigenvector of B.

Two matrices are **conjugate** if

$$B = X^T A X$$

where X is nonsingular (compare this to similarity).
The triple (n_+, n_-, n_0) denoting the number of **positive**, **negative** and **zero** eigenvalues respectively is called the **inertia** of A.

Sylverster's law of Inertia says that the inertia of a matrix A is preserved under conjugation.

Computing Eigenvalues (not in practice!)

Definition

The **characteristic polynomial** is

$$\chi(A; z) = det(A - zI) = \prod_{i=1}^{n} (\lambda_i - z)$$

where λ_i are the (not necessarily distinct) eigenvalues of A.

Example:

$$A = \begin{pmatrix} 9 & 0 & -6 \\ -1 & 4 & 2 \\ 2 & 1 & 2 \end{pmatrix}, \qquad A - zI = \begin{pmatrix} 9 - z & 0 & -6 \\ -1 & 4 - z & 2 \\ 2 & 1 & 2 - z \end{pmatrix},$$

$$\chi(A; z) = -z^3 + 15z^2 - 72z + 108.$$

Aside: Computing roots of polynomials

Suppose we want to compute the roots of the polynomial

$$p(t) = a_0 + a_1 t + \cdots + a_{n-1} t^{n-1} + t^n.$$

The **companion matrix** is

$$C(p) = \begin{pmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \cdots & 0 & -a_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -a_{n-1} \end{pmatrix}.$$

Check that $\chi(A; z) = p(z)$, so that the eigenvalues of the companion matrix are the roots of the polynomial.

Computing Eigenvalues (not in practice!)

Once a root λ is found of $\chi(A; z)$, the corresponding eigenvector satisfies $v \in \ker(A - \lambda I)$.

Example: The roots of $\chi(A)$ are $\lambda = 3, 6, 6$. For $\lambda = 3$,

$$A - 3I = \begin{pmatrix} 6 & 0 & -6 \\ -1 & 1 & 2 \\ 2 & 1 & -1 \end{pmatrix} \implies v_1 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

Computing Eigenvalues (not in practice!)

$$A - 6I = \begin{pmatrix} 3 & 0 & -6 \\ -1 & -2 & 2 \\ 2 & 1 & -4 \end{pmatrix}$$
$$\sim \begin{pmatrix} 3 & 0 & -6 \\ 0 & -2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \implies v_2 = \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}$$

Notice $\dim \{\ker(A - 6I)\} = 1$, even though $\lambda = 6$ is a double root in the characteristic polynomial.

Diagonalizable matrices

Definition

An $n \times n$ matrix A is **diagonalizable** if it has n linearly independent eigenvectors.

If A has n eigenvectors that are also mutually orthogonal, we call A **unitarily diagonalizable**.

Most square matrices (in a mathematically rigorous sense) are diagonalizable. Important examples:

- Symmetric matrices: $A = A^T$
- Normal matrices: $AA^T = A^T A$
- Matrices with n distinct eigenvalues

Non-diagonalizable matrices are **defective**.

Geometric vs. Algebraic multiplicity

Definition

The **algebraic multiplicity** of the eigenvalue λ_i is the multiplicity of the root in the characteristic polynomial.
The **geometric multiplicity** of the eigenvalue λ_i is the dimension of the associated eigenspace $\dim \{\ker(A - \lambda_i I)\}$.

Notice that always the geometric multiplicity is at most the algebraic multiplicity.

Example: The algebraic multiplicity of $\lambda = 6$ is 2, but the geometric multiplicity is 1.
This means that the matrix A is **defective**.

Matrices for which it's easy to find eigenvalues and eigenvectors:

- Diagonal matrices (it's already diagonalized!)
- Triangular matrices

Gershgorin's Disc Theorem

The ith Gershgorin disc is a ball of radius $r_i = \sum_{j \neq i} |a_{ij}|$ centered at a_{ii} in the complex plane,

$$\mathcal{D}_i = \left\{ z \in \mathbb{C} \mid |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\}.$$

Theorem

Every eigenvalue of A sits in a Gershgorin disc.

Schur Decomposition

Definition

For all square matrices A, there exist unitary Q and upper-triangular T such that

$$A = QTQ^*.$$

This is the **Schur Decomposition**.

Properties:

- Since T is triangular its eigenvalues are on the diagonal and T and A are similar, the eigenvalues of A are on the diagonal of T.
- This decomposition exists for all square matrices
- Not unique

Determinant and Trace Revisited

Exercise: Using the Schur Decomposition, find expressions for the determinant and trace of a matrix in terms of its eigenvalues.

Exercise: Using the Schur Decomposition, find expressions for the determinant and trace of a matrix in terms of its eigenvalues.

$$
\begin{aligned}
\det(A) &= \det(QTQ^*) \\
&= \det(T) = \prod_{\lambda \in \sigma(A)} \lambda, \\
\operatorname{trace}(A) &= \operatorname{trace}(QTQ^*) \\
&= \operatorname{trace}(Q^*QT) = \sum_{\lambda \in \sigma(A)} \lambda
\end{aligned}
$$

Eigenvalue Decomposition

Definition

If A is diagonalizable, then there exists an invertible matrix X and diagonal matrix Λ such that

$$A = X \Lambda X^{-1}.$$

The eigenvalues of A are on the diagonal of Λ.

Properties:

- If A is symmetric, then A is diagonalizable and X is orthogonal. Furthermore the eigenvalues are necessarily real.
- **Exercise:** Prove that the eigenvalue decomposition exists for symmetric matrices.
  Prove that the eigenvalues of a Hermitian matrix are real.
- Unique up to ordering, but does not always exist!
- For A SPD: $x^T A x \geq 0, \forall x \neq 0 \iff \lambda \geq 0, \forall \lambda \in \sigma(A)$

Exercises

1. Given an eigenvalue decomposition of $A = X\Lambda X^{-1}$, how can you compute A^n quickly?
2. The fibonacci sequence is defined by $F_0 = 0$, $F_1 = 1$ and $F_n = F_{n-1} + F_{n-2}$. Prove that

$$F_n = \frac{\phi^n + \psi^n}{\sqrt{5}}, \qquad \phi = \frac{1+\sqrt{5}}{2}, \psi = \frac{1-\sqrt{5}}{2},$$

by using the matrix

$$\begin{pmatrix} F_n \\ F_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} F_{n-1} \\ F_{n-2} \end{pmatrix}.$$

The Spectral Theorem

Theorem

A matrix A is unitarily diagonalizable if and only if it is normal, that is it satisfies

$$AA^* = A^*A.$$

This exactly classifies when there exists a unitary Q and diagonal Λ such that

$$A = Q\Lambda Q^*.$$

Jordan Canonical form

Definition

Let A be a square matrix with distinct eigenvalues λ_i with algebraic multiplicity a_i and geometric multiplicity g_i. Define a **Jordan block** as

$$J_i = \left(\begin{array}{ccccc} \lambda_i I_{g_i-1} & & & & \\ \hline & \lambda_i & 1 & & \\ & & \lambda_i & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda_i \end{array}\right).$$

Then there exists a nonsingular matrix X such that $A = XJX^{-1}$ where $J = \mathrm{diag}(J_i)$.

For when the eigenvalue decomposition doesn't exist.

For symmetric (Hermitian) matrices, the eigenvalue decomposition are extremely useful:

- It always exists
- Eigenvalues form an orthogonal basis of $\mathbb{C}^n$
- The eigenvalues are real
- Eigenvalues give us the norm of A: $\|A\|_2 = \max \lambda$, $\|A\|_F = \sum \lambda$.

For general matrices:

- Eigenvalue decomposition doesn't necessarily exist. Schur and Jordan form exist only for square matrices.
- Defective matrices don't have eigenvalues which span all of $\mathbb{C}^n$
- Eigenvalues may be complex
- Eigenvalues no longer characterize A:

$$A = \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix}$$

has $\|A\|_2 = O(\alpha)$ but all eigenvalues are 1.

Need a better decomposition for general matrices...

Singular Value Decomposition

Definition

Let A be an $n \times m$ matrix (assume $n \geq m$). There exist unitary matrices $U \in \mathbb{C}^{n \times n}$ and $V \in \mathbb{C}^{m \times m}$, and diagonal matrix $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_m)$, with $\sigma_1 \geq \ldots \sigma_m \geq 0$, such that

$$A = U \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^*.$$

This is called the **Singular Value Decomposition**.

- σ_i are the **singular values**
- $U = (u_1, \ldots, u_n)$ are the **left singular vectors**
- $V = (v_1, \ldots, v_m)$ are the **right singular vectors**

Singular Value Decomposition

Notice that we can then write A as a sum of outer products

$$A = \sigma_1 u_1 v_1^* + \sigma_2 u_2 v_2^* + \cdots + \sigma_m u_m v_m^*$$

Suppose that $\sigma_{k+1} = \sigma_{k+2} = \cdots = \sigma_m = 0$ for some k.
We can make the **economy-size SVD** with $\sigma_1, \ldots, \sigma_k > 0$, and we can split the singular vectors $U = (U_1,\ U_2)$, $V = (V_1,\ V_2)$ with

- $U_1 = (u_1, \ldots, u_k)$ and $U_2 = (u_{k+1}, \ldots, u_n)$,
- $V_1 = (v_1, \ldots, v_k)$ and $V_2 = (v_{k+1}, \ldots, v_m)$

so that

$$A = U_1 \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{pmatrix} V_1^*.$$

(Some) Properties:

- It is unique (up to singular vectors with same singular value)
- If $\text{rank}(A) = k$ then $\sigma_{k+1} = \cdots = \sigma_m = 0$. Similarly, if $\text{null}(A) = n - k$ then $n - k$ of the singular values are zero.
- $\{u_1, \ldots, u_k\}$ form an orthogonal basis for $\text{range}(A)$.
- $\{u_{k+1}, \ldots, u_n\}$ form an orthogonal basis for $\ker(A^*)$
- $\{v_1, \ldots, v_k\}$ form an orthogonal basis for $\text{range}(A^*)$.
- $\{v_{k+1}, \ldots, v_m\}$ form an orthogonal basis for $\ker(A)$
- $\|A\|_2 = \sigma_1$, $\|A^{-1}\|_2 = \frac{1}{\sigma_n}$ and $\|A\|_F = \left(\sum_{i=1}^k \sigma_i^2\right)^{\frac{1}{2}}$.

More properties:

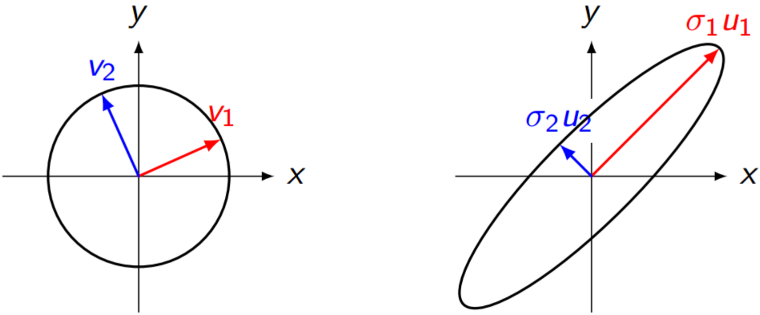
- Condition number for square matrices: $\kappa(A) = \sigma_1/\sigma_n$. In general if A has rank k, then $\kappa(A) = \sigma_1/\sigma_k$ (or ∞ depending on what you're trying to do).
- Eigenvalue decompositions:

$$A^*A = V\Sigma^2V^*, \qquad AA^* = U\begin{pmatrix} \Sigma^2 & \\ & 0 \end{pmatrix}U^*.$$

- Pseudo-inverse:

$$A^\dagger = U_1\Sigma^{-1}V_1^*.$$

Geometric interpretation of SVD



Exercises

1. How do the eigenvalues and singular values of A^{-1} relate to A (for A invertible)?

2. Let A be a SPD matrix. What does this say about the eigenvalues of A?

Low-Rank Matrix Approximations

Theorem

Let A be an $n \times m$ matrix, and $k < \min(m, n)$, then

$$\min_{\text{rank}(B)=k} \|A - B\|_2 = \sigma_{k+1},$$

and the minimum is attained by $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^*$.

Theorem

Let A be an $n \times m$ matrix, and $k < \min(n, m)$, then

$$\min_{\text{rank}(B)=k} \|A - B\|_F = \sqrt{\sum_{i=k+1}^{\min(m,n)} \sigma_i^2},$$

and the minimum is attained by $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^*$.

Problem:
Consider a $n \times d$ matrix D of data (n datapoints, d variables). Assume that each column has mean 0. We want to find $k \leq d$ vectors which best capture the variance in the data.

Solution:
Compute the SVD, $D = U\Sigma V^T$, and take the first k singular vectors (and the singular values are related to the variance of the data along the principal directions).

Application: Principal Component Analysis

Problem:

Consider a $n \times d$ matrix D of data (n datapoints, d variables). Assume that each column has mean 0. We want to find $k \leq d$ vectors which best capture the variance in the data.