STARK DRAPER

COURES NOTES:

OPTIMIZATION THEORY AND ALGORITHMS

COURSE NOTES: VERSION 1.02

October 2019

Contents

Remarks, feedback, and versions

These notes are in development in fall term 2019. These notes are meant to complement and not replace the course text, providing the reader of our specific trajectory through the text and the emphasis of material in our course. Main thanks for this teaching resource are due to Zhipeng Huang and Yanxiao Liu who built up these notes from scratch. Thank you Zhipeng and Yanxiao! As we progress through the semester updated versions with additional chapters and edits will be distributed. The main differences between distributions are noted below. Corrections of typos and errors, and other suggestions are welcome and appreciated. Please email any such comments to `eceCourseProfDraper@gmail.com`. Please include the course number in the subject line of your message, as multiple sets of notes are in parallel development.

Version 1.01: Initial distribution of chapters 1 and 2.
Version 1.02: Initial distribution of chapters 2 and 3.

1
Introduction

This class will introduce you to the fundamental theory and models of optimization as well as the geometry that underlies them. The first portion of the course focuses on geometry: recalling and generalizing linear algebraic concepts you first met in your linear algebra course. The second portion focuses on optimization. Presentation of applications is woven throughout. We will draw examples from diverse areas of the engineering and natural sciences. The material covered in this course will prove of interest to students from all areas of engineering, from the computer sciences and, more generally, from disciplines wherein mathematical structure and the use of numerical data is of central importance.

The main prior courses that we will be building on are vector calculus and linear algebra. No prior exposure to optimization is assumed.

The course text is *Optimization Models*, by G. Calafiore and L. El Ghaoui, Cambridge Univ. Press, 2014. These notes are provided as a supplement to, and not a replacement for, the course text. Many problem set problems will be drawn from the course text.

Notation

We work mainly with finite-dimensional real-valued vectors in the course. Lower-case is used for vectors. A length-n real vector x is an ordered collection of real numbers where the ith coordinate of x is denoted $x_i \in \mathbb{R}$. The default will be column vectors so

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

The length n of the vector is also termed the "dimension" of the vector, which will subsequently be defined formally. Alternately, the

elements of x may be complex, i.e., $x_i \in \mathbb{C}$, or in some other field, $x_i \in \mathbb{F}$. Again, our focus will be in the reals and we compactly denote the space of x as $x \in \mathbb{R}^n$. The transpose of a column vector is a row vector. The transpose x^{T} of x is

$$x^{\mathrm{T}} = [x_1 \; x_2 \; \dots x_n].$$

We often need to work with a set (or a list) of vectors,

$$\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$$

where $x^{(i)} \in \mathbb{R}^n$, $i \in \{1, 2, \dots, m\}$ and $(x^{(i)})^{\mathrm{T}} = [x_1^{(i)} \; x_2^{(i)} \; \dots \; x_n^{(i)}]$. The set $\{1, 2, \dots, m\}$ is the index set of m elements. We often use the shorthand $[m]$ for the index set; in the above we would have written $i \in [m]$. We note that the book is not one hundred percent consistent on this notation. It sometimes reverts to the (simpler) notation $\{x_1, x_2, \dots x_m\}$ where $x_i \in \mathbb{R}^n$ and $i \in [m]$ for sets of vectors. This less burdensome notation is used n settings where sets of vectors are considered, but it is not necessary also to index individual elements of the vectors.

Uppercase is used for matrices. A matrix A consisting of n rows and m columns of real numbers is denoted $A \in \mathbb{R}^{n \times m}$. The element in the ith row and jth column of A is denoted $[A]_{ij}$ (alternately a_{ij}). The transpose of A, A^{T} is the matrix the element in the ith row and jth column of which is $[A]_{ji}$ (alternately a_{ji}).

Sets are denoted using calligraphic font. (I will say "script" in class since "calligraphic" is a mouthful.) For example, the set of vectors described above might be denoted $\mathcal{X} = \{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$. The cardinality of the set $\mathcal{X}$ is denoted $|\mathcal{X}|$; in the above example $|\mathcal{X}| = m$. For some special sets we make an exception. In particular to denote real numbers, complex numbers, and integers we respectively write $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{Z}$. Occassionally we have need to refer to the sets of non-negative and positive real numbers, respectively denoted $\mathbb{R}_+$ and $\mathbb{R}_{++}$.

Functions map elements of one set to another. As with vectors we use lowercase letters to denote functions. While we typically use letters towards the end of the Latin alphabet for vectors (u, v, w, x, y, z), we typically use letters earlier in the alphabet for functions (f, g, h), and letters in the middle for indexing (i, j, k, l, m, n).

We write $f : \mathcal{X} \to \mathcal{Y}$ to denote a function f that maps elements of $\mathcal{X}$ to elements of $\mathcal{Y}$. This notation is akin to strongly-typed programming languages. The function f needs an input in $\mathcal{X}$ to be able to process it. Elements not in $\mathcal{X}$ are not acceptable as inputs. That said, not every element of $\mathcal{X}$ may be acceptable to f. (E.g., if f calculates the average age of students in a class, no age inputted into the function should be negative.) The acceptable subset of $\mathcal{X}$ is the domain

of f, denoted dom$f$. It is often convenient to define $f(x) = \infty$ for all $x \notin \text{dom}f$. In that case $\text{dom}f = \{x \in \mathcal{X} | |f(x)| < \infty\}$. In this course we mostly consider functions of the form $f : \mathbb{R}^n \to \mathbb{R}^m$. Some terminology that you might be aware of concerns the relationship between n and m. If $n \neq m$ then f is a "map". If $n = m$ then f is an "operator". If $m = 1$ then f is a "functional". An example of an $f : \mathbb{R} \to \mathbb{R}$ where dom$f = \mathbb{R}_+$ is plotted in Fig. 1.1.

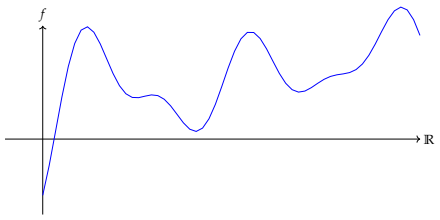


Figure 1.1: A function $f : \mathbb{R} \to \mathbb{R}$.

2
Vectors and functions

① Geometry
  -Vectors and vector spaces
  -Norms
  -Inner product
② Projection
  -Onto subspace
  -Onto affine sets
  -Non-Euclidean
③ Functions
  -Functions and sets
  -Linear and affine
  -Gradients and Taylor approximations

Vector: A collection of numbers.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

where each $x_i \in \mathbb{R}$ or $x_i \in \mathbb{C}$. The length n of the vector is also termed as the "dimension" of the vector, which will subsequently be defined formally.

Our default will be a column vector as we describe above. Transpose x yields a row vector,

$$x^{\mathrm{T}} = [x_1 \ x_2 \ \dots x_n].$$

and occasionally write as a list $(x_1, x_2, \cdots, x_n)$. Note that a vector is not a set of numbers since order matters.

Also, we often need to work with a set(or list) of vectors,

$$\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$$

where $x^{(i)} \in \mathbb{R}^n$, $i \in \{1, 2, \dots, m\}$, $i \in [m] = \{1, 2, \cdots, m\}$ and $(x^{(i)})^{\mathrm{T}} = [x_1^{(i)} \ x_2^{(i)} \ \dots \ x_n^{(i)}]$.

Note: The textbook is not 100% consistent in its use of this notation.

Vector Space

First, we define how to add pairs of vectors and how to scale vectors as follows:

Addition: $u = v^1 + v^2$, means $u_i = v_i^1 + v_i^2$ for all $i \in [n]$.

Scaling: $u = av$, means $u_i = v_i^1 + v_i^2$ for all $i \in [n]$.

Linear combination: $\sum_{i=1}^{m} a_i v^{(i)}$

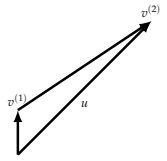


Figure 2.1: Addition

Vector Space: a set of vectors v that is closed under addition and scaling, and satisfy following axioms:

(1) Commutativity: $u + v = v + u$

(2)Associativity: $(u + v) + w = u + (v + w)$

(3) Distributivity: $a(u + v) = au + av$, $(a + b)u = au + bu$

(4) Identity element of addition: $\exists 0 \in \mathcal{V}$ s.t. $u + 0 = u$

(5) Inverse elements of addition: $\exists -u \in \mathcal{V}$ s.t. $u + (-u) = 0$

(6) Identity element of scalar multiplication: $\exists a \in \mathbb{R}$ or $\mathbb{C}$ s.t. $au = u$

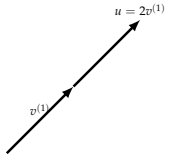


Figure 2.2: Scaling

In this course our focus is on $\mathbb{R}^n$, i.e., finite-length vectors with real elements. It is also useful to note that the geometric ideas could apply to lots of other spaces, such as

① Finite-length complex vector

we need this esp for discussion of eigenvalues and eigenvectors. But it is also important example in quantum computing.

② ∞-length complex sequences

③ Complex functions defined on real line

④ Polynomials of degree at most n-1

$$P_{n-1} = \{P | p(t) = a_{n-1}t^{n-1} + a_{n-2}t^{n-2} + \cdots + a_1 t + a_0\}$$

⑤ Sets of matrices(will discuss later)

Note: Some authors prefer “linear space" rather than "vector space" since elements of space are not always vectors in the sense of a list.

Span and subspace

Let S be a set of vectors in a real vector space V, i.e.,$S = \{v^{(1)}, v^{(2)}, \cdots, v^{(m)}\}$, where each $v^{(i)} \in \mathbb{R}^n$.Then, the span of S, denoted by span(S), is the set consisting of all the vectors that are linear combinations of $\{v^{(1)}, v^{(2)}, \cdots, v^{(m)}\}$, that is,

$$\text{span}(S) = \left\{ \sum_{i=1}^{m} a_i v^{(i)} \middle| a_i \in \mathbb{R}, \forall i \in [m] \right\}$$

This set is also called a **subspace** of V.

Example 1

Let $S = \{v^{(1)}\} = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}$, then

$$\begin{aligned} \text{span}(S) &= \text{span}(v^{(1)}) \\ &= \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \middle| x = y \right\} \\ &= \left\{ a \begin{bmatrix} 1 \\ 1 \end{bmatrix} \middle| a \in \mathbb{R} \right\} \end{aligned}$$

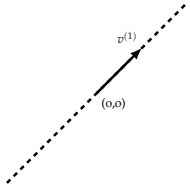


Figure 2.3: Example 1

Example 2

Let $S = \{v^{(1)}, v^{(2)}\} = \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \right\}$, then

$$\begin{aligned} \text{span}(S) &= \{a_1 v^{(1)}, a_2 v^{(2)} | (a_1, a_2) \in \mathbb{R}^2\} \\ &= \left\{ \begin{bmatrix} x \\ y \\ 0 \end{bmatrix} \middle| x \in \mathbb{R}, y \in \mathbb{R} \right\} \\ &= x - y \text{ plane} \end{aligned}$$

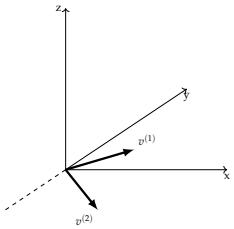


Figure 2.4: Example 2

Note:

(1) $0 \in \mathbb{R}^n$ always included since we can set all coefficients $a_i = 0$ for all i.

(2) Subspace is a "flat" that goes through the origin.

Linear independent set

A set $S = \{v^{(1)}, \cdots, v^{(n)}\}$ is a linearly independent set if there is no element of S can be expressed as a linear combination of the others.

The set S is linearly independent if the only a_i that satisfies

$$\sum_{i=1}^{m} a_i v^{(i)} = 0 \quad \text{is if} \quad a_i = 0 \; \forall i \in [m]$$

Importance of linearly independent

For any $u \in \text{span}(S)$, there is a unique linear combination to express u. That is, only one choice of a_i in the expression.

For example, any 2-d vector u can be uniquely expressed by the following two vectors $v^{(1)}$ and $v^{(2)}$, which form the set S.

$$v^{(1)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, v^{(2)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

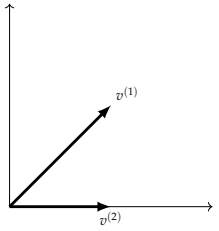


Figure 2.5:

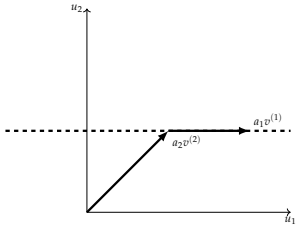


Figure 2.6:

Notice that the two vectors are co-linear so that there is no redundancy in the set S. Now, consider the case there is a redundancy in S, i.e., there is a vector in S can be expressed by the others in S:

$$S = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$$

We can prove that we can always shrink S by removing elements to get a linearly independent set(and also the same subspace before the deletion). Such an irreducible or linearly independent set can serve as a **basis** for span(S).

Any largest linearly independent subset of $S = \{v^{(1)}, \cdots, v^{(m)}\}$, $B = \{v^{(1)}, \cdots, v^{(k)}\}$, $k \leq m$ is a basis for span(S), and the dimension of span(S) is denoted as dim(span(S))=k.

Example 1

The following vectors form an linearly independent spanning set S, and also serve as a basis for the vector space spanned by the set S (i.e., $\mathbb{R}^3$ in this case)

$$v^{(1)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, v^{(2)} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, v^{(3)} = \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix}$$

However, if we redefine v^3, says

$$v^{(3)} = \begin{bmatrix} 3 \\ 4 \\ 2 \end{bmatrix} = 2v^{(1)} + v^{(2)}$$

Then $\left(v^{(1)}, v^{(2)}, v^{(3)}\right)$ is not a basis(since it is linearly dependent now), and it need to be reduced to:

$$\text{span}\left(\{v^{(1)}, v^{(2)}\}\right) = \text{span}\left(\{v^{(1)}, v^{(3)}\}\right) = \text{span}\left(\{v^{(2)}, v^{(3)}\}\right) = \text{span}(S)$$

We can prove that each a basis for span(S) all have same coordinates.

Example 2

The most commonly used basis is the "standard" basis, that is, each vector of the basis has a unit length:

$$v^{(1)} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}, v^{(2)} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}, \cdots, v^{(n)} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

We often use $'e'$ for standard basis, i.e., $e^{(i)} = v^{(i)}$.

Norms: The idea of distance on length on a vector space $\mathcal{V}$

A norm $\|\cdot\|$ is a function such that $\|\cdot\| : \mathcal{V} \mapsto \mathbb{R}$ and satisfies

(a) $\|v\| \geq 0$, $\forall v \in \nu$, and $\|v\| = 0$ iff $v = 0$.

(b) $\|u + v\| \leq \|u\| + \|v\|$, $\forall u, v \in \mathcal{V}$.

(c) $\|au\| = |a|\|u\|$, $\forall a \in \mathbb{R}, u \in \mathcal{V}$

Note that ν can be either $\mathbb{R}$ or $\mathbb{C}$, if $\mathcal{V} \in \mathbb{C}$ we should have $a \in \mathbb{C}$ in (c).

Following is a family of norms that are frequently used:

L_p norm:

$$\|x\|_p = (\sum_{k=1}^{n} |x_k|^p)^{1/p}, 1 \leq p \leq \infty$$

L_2 norm: Euclidean length

$$\|x\|_2 = \sqrt{\sum_{k=1}^{n} |x_k|^2}$$

L_1 norm:

$$\|x\|_1 = \sum_{k=1}^{n} |x_k|$$

L_∞ norm:

$$\|x\|_\infty = \lim_{p \to \infty} \|x_k\|_p = \max_{k \in [n]} |x_k|$$

Length is a notion of "size". A natural notion of its "size" of a set is the number of non zero component, i.e., carnality of non-zero support

$$\text{card}(x) = \sum_{k=1}^{n} \mathbb{1}_{x_k \neq 0}$$

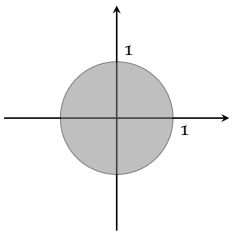
Sometimes it is called "L_0" norm $\|x\|_0$, since $\text{card}(x) = \lim_{p\to 0}(\sum_{k=1}^{n} |x_k|^p)^p$, but it is not a norm (so this terminology is inaccurate). For instance, it doesn't satisfy property (c) of a norm:

$$\text{card}(2x) = \text{card}(x) \neq 2\text{card}(x)$$

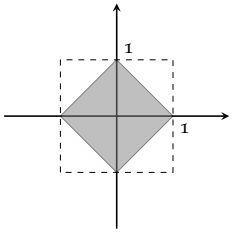
Unit norm-ball

To visualize a norm we often plot the unit norm-ball $\beta_p = \{x | \|x\|_p \leq 1\}$ in $\mathbb{R}^2$. For example,

L_2 norm ball

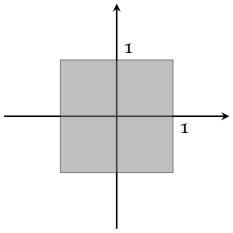


L_1 norm ball : $\{x | |x_1| \leq 1\}$

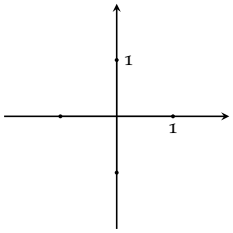


(a) First see inside the box, clearly $|x_1| \leq 1$ and $|x_2| \leq 1$
(b) Look at the position we want, $x_1 + x_2 \leq 1$, i.e., $x_2 \leq 1 - x_1$
(c) Rest by symmetry

L_∞ norm ball: $\{x|\max\{|x_1|,|x_2|\} \leq 1\}$



What about card(x)?



The set $\{x|\text{card}(x) \leq 1\}$ obviously is not much of a "ball".

To visualize a bit more, we look at the **"level sets"** of the norm balls. We define the level set as $\{x||x| = c\}$, and let's see for $c = \frac{1}{2}, 1, 2$. See for the figures on the r.h.s.

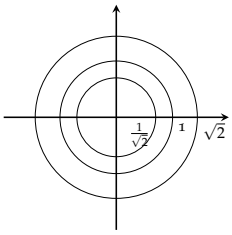


Figure 2.7: L_1 level set

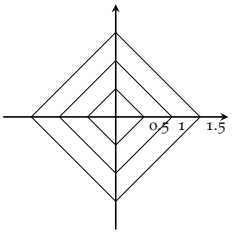


Figure 2.8: L_2 level set

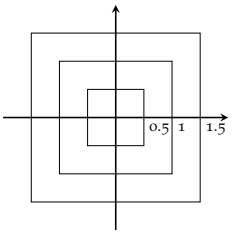


Figure 2.9: L_∞ level set

Why might we be interested in different norms?

Later in the course, we will see applications in optimal control that we want to meet a control objective while minimizing some resources(The objective will be to min a norm of the resources).

Inner Products

Any inner product(aka dot/scalar product) on a (real) vector space Ω maps a pair of elements $x, y \in \Omega$ into the scalar, that is, $\langle \cdot, \cdot \rangle : \Omega \times \Omega \mapsto \mathbb{R}$. For vectors in $\mathbb{R}^n$ the inner product of vectors x and y is given by

$$\langle x, y \rangle = x^{\mathrm{T}} y = \sum_{k=1}^{n} x_k y_k$$

For any $x, y, z \in \Omega$ and $a \in \mathbb{R}$, the following must hold for a inner product:

(1) $\langle x, y \rangle \geq 0$ and $\langle x, y \rangle = 0$ iff $x = 0 \in \Omega$

(2) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$

(3) $\langle ax, y \rangle = a\langle x, y \rangle$

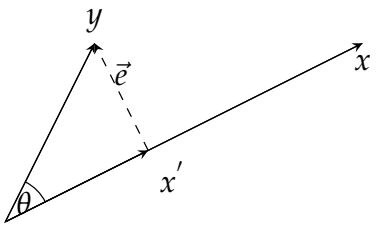
(4) $\langle x, y \rangle = \langle y, x \rangle$

Note:

(a) The above change slightly in complex vector space, e.g., $\langle x, y\rangle = \overline{\langle y, x\rangle}$

(b) The concept we develop apply beyond list vectors in $\mathbb{R}^n$ or $\mathbb{C}_n$, e.g., space of polynomials or of functions, but our focus will be $\mathbb{R}^n$ and $\mathbb{C}_n$.

Let's connect to angle now.



In above picture $x, y \in \mathbb{R}^n$ but since dimspan$(x, y) = 2$ (assuming x and y are not co-linear). The familiar picture in $\mathbb{R}^2$ shall holds.

Since we know that $|\cos\theta| < 1$, rearranging gives

$$|\langle x, y\rangle| = |x^{\mathrm{T}} y| \leq \|x\|_2 \|y\|_2$$

This is the so called Cauchy-Schwartz inequality, and it relates inner product(angle) to norms(length). Such inequality holds for inner product spaces, not just $\mathbb{R}^n$(n-dimensional Euclidean space). Further more, it could relate the inner product to the norms (not only L_2) via a generalization, says, "Holder's inequality":

$$|x^{\mathrm{T}} y| \leq \sum_{k=1}^{n} |x_k y_k| \leq \|x\|_p \|y\|_q$$

for any $p, q \geq 1$ such that $1/p + 1/q = 1$.

(1) If $p = q = 2$, we get the Cauchy-Schwartz inequality.

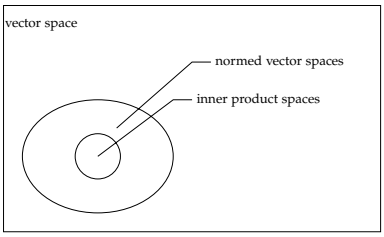
(2) If $p = 1$, $q = \infty$, we get $|x^{\mathrm{T}} y| \leq \|x\|_1 \|y\|_\infty = (\sum_{k=1}^{n} |x_k|)(\max_{k\in[n]} x_k)$.

A second important connection of inner product and norm is that

$$\|x\|_2 = \sqrt{x^{\mathrm{T}} x} = \langle x, x\rangle$$

The L_2 norm is "induced" by the inner product. In fact, any inner product induces a norm (by the properties of inner product) However, there are norms that are not induced by any inner product, e.g., L_1 and L_∞. Inner product space has a more special structure than a "normed" vector space.

Note: There are also spaces with a sense of length (a "metric"). Those are not vector spaces (it can't add and scale elements). Those are "metric" spaces.



Angles between vectors

By Cauchy-Schwartz $\frac{|\langle x,y\rangle|}{\|x\|\|y\|} \leq 1$, hence we have the following cases for the angle θ:

(a) If $|\cos\theta| = +1$, then $\theta = 0^\circ$ or 180°. Vectors x and y are "co-linear", and $|\langle x,y\rangle| = \|x\|\|y\|$

(b) If $|\cos\theta| = 0$, then $\theta = 90^\circ$, and $\frac{|\langle x,y\rangle|}{\|x\|\|y\|} = 0$, or equivalently, $\langle x,y\rangle = 0$ (asumming $x \neq 0$ and $y \neq 0$). In this case, θ is a "right" angle, and x, y are orthogonal vectors.

(c) If $|\theta| < 90^\circ$,then $\cos\theta > 0$, and $\langle x,y\rangle > 0$, and θ is a "acute angle", wheres if $|\theta| > 90^{circ}$, then $\cos\theta < 0$ and $\langle x,y\rangle < 0$, θ is a "obtuse angle".

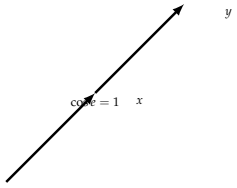


Figure 2.10: (a) $\cos\theta = +1$

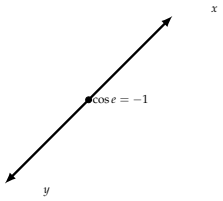


Figure 2.11: (a) $\cos\theta = -1$

Orthogonality

A set of vectors $S = \{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$ is mutually orthogonal if $\langle x^{(i)}, x^{(j)}\rangle = 0$, $\forall i \neq j$

Such sets have nice property that the elements of S are linearly independent and so provide a basis for span(S) and hence dim(span(S))=m.

If, in addition, all elements have unit norm, i.e., $\|x^{(i)}\|_2 = 1$ for all $i \in [m]$ then the set forms an **orthogonal basis**.

Note that we use $\|\cdot\|_2$ to measure length because it is induced by the inner product.

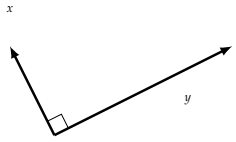


Figure 2.12: (b) $\cos\theta = 0$

Orthogonal complement: Given a subspace $S \in \nu$, a vector $x \in \nu$ is orthogonal to S if $x \perp s \forall s \in S$, i.e., $\perp$ to all vectors in the subspace S. The orthogonal complement to S is defined as a collection of such vectors x, namely

$$S^\perp = \{x \in \nu | x \perp s\}$$

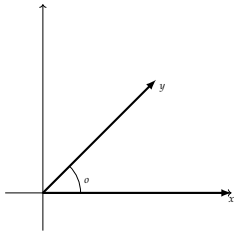


Figure 2.13: (c) $\cos\theta > 0$

Some results of $S^\perp$:

(i) $S^\perp$ is a subspace: clearly it includes $0 \in \gamma$ and is closed under linear combination(all linear combination $\perp S$).

(ii) dim(ν)=dim(S) + dim ($S^\perp$).

(iii) Any $x \in \nu$ can be written in a unique way as $x = x_s + x_{s^\perp}$ for any subspace S.

Note: If $S = \nu$ then $S^\perp = 0$.

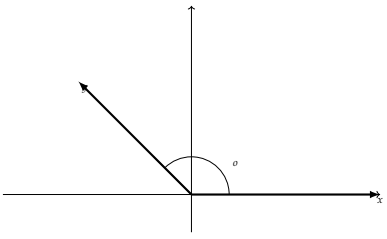


Figure 2.14: (c) $\cos\theta < 0$

Projection

Motivation: Given a point $x \in v$, find the "closest" point in the set S (recall that points $\equiv$ vectors), this point is called the projection of x on the set S. Denote this point as y, formally we have

$$y = \Pi_s(x) = \arg\min_{y \in S} \|y - x\|$$

Let's consider different cases for this optimization question:

(1) S is a subspace of an inner product space associate with L_2 norm

(2) S is an "affine" set(a shifted subspace)

(3) Consider other norms, e.g., L_1, L_∞ for which no inner product(projection in normed vectors space)

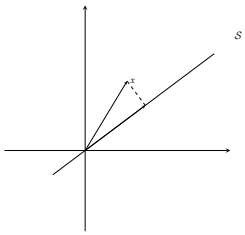


Figure 2.15: S is a subspace of an inner product space

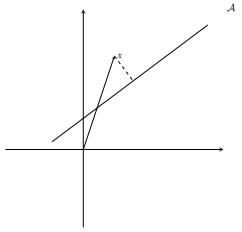


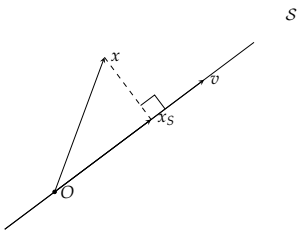
Figure 2.16: S is an "affine" set

Projection onto 1-D subspace

Let's consider the one dimension subspace given by

$$S = span(\{v\}) = \{\lambda v | \lambda \in \mathbb{R}\}$$

The vector x and subspace S are somethings like:



By orthogonal decomposition, $x \in S \oplus S^\perp$, and therefore $\exists x_s \in S, e \in S^\perp$, such that $x = x_s + e$ (an unique expression).

Use this decomposition to solve the optimization problem(find the closest point)

$$\Pi_s(x) = \arg_{y \in S} \min \|y - x\|_2 = \arg_{y \in S} \min \|y - x\|_2^2$$

The objective function can be written as

$$\begin{aligned} \|y - x\|_2^2 &= \langle y - x, y - x \rangle \\ &= \langle (y - x_s) - e, (y - x_s) - e \rangle \\ &= \|y - x_s\|^2 + \|z\|^2 - 2\langle y - x_s, e \rangle \\ &\geq \|e\|_2^2 \end{aligned}$$

where the minimum is attained by setting $y = x_s$. Note that the minimum is unique by uniqueness of orthogonal decomposition and $\|y - x_s\|^2 = 0$ iff $y = x_s$.

To summarize,

$$x_s = \Pi_s(x) = \arg_{y \in S} \min \|y - x\|_2$$

where x_s is in $\perp$-decomposition.

To solve for x_s (the point we want), we use the condition that

$$(x - x_s) \perp S = \{\lambda v | \lambda \in \mathbb{R}\}$$

Since $x \in S$, $\exists a \in \mathbb{R}$ such that $x_s = av$, and now we need to solve for a by

$$0 = \langle x - av, v\rangle = \langle x, v\rangle - \langle av, v\rangle = \langle x, v\rangle - a\langle v, v\rangle$$

Rearranging yields that $a = \frac{\langle x,v\rangle}{a\langle v,v\rangle} = \frac{\langle x,v\rangle}{\|v\|^2}$. Thus, $x_s = av = \frac{\langle x,v\rangle}{\|v\|^2}v$.

Projection onto a general subspace

Observe that all previous steps for 1-D case still hold. Only used S is 1-D when solving for $x^{(s)}$, so we have already done.

Theorem: Let $x \in \Omega$ and $S \in \Omega$, where x is a vector, S is a subspace of Ω and Ω is an inner product space. There exists a unique vector $x^* \in S$ such that

$$x^* = \arg_{y \in S} \min \|x - y\|$$

A necessary and sufficient condition for x^* is
(1). $x^* \in S$
(2). $x - x^* \perp S$

Now let's consider how to Solve for x^* in this general case.

Let $S = \text{span}\left(\{x^{(1)}, x^{(2)}, \cdots x^{(d)}\}\right)$. Notice that $x^* \in S$ can be written as $x^* = \sum_{i=1}^{d} a_i x^{(i)}$ for some a_i, and $(x - x^*) \perp S$, then if $(x - x^*) \perp x^{(k)}\ \forall k \in [d]$, that will be $\perp$ to all linear combination of the spanning set and hence $\perp$ to S.

Accordingly yields d conditions, $\forall k \in [d]$, we have

$$0 = \langle x - x^*, x^{(k)}\rangle = \langle x - \sum_{i=1}^{d} a_i x^{(i)}, x^{(k)}\rangle = \langle x, x^{(k)}\rangle - \sum_{i=1}^{d} a_i \langle x^{(i)}, x^{(k)}\rangle$$

Rearranging yields

$$\sum_{i=1}^{d} a_i \langle x^{(i)}, x^{(k)}\rangle = \langle x, x^{(k)}\rangle,\ \forall k \in [d]$$

Or stacking into a matrix(d equations in d unknowns)

$$\begin{bmatrix} \langle x^{(1)}, x^{(1)}\rangle & \langle x^{(1)}, x^{(2)}\rangle & \cdots & \langle x^{(1)}, x^{(d)}\rangle \\ \vdots & & & \\ \langle x^{(d)}, x^{(1)}\rangle & \cdots & \cdots & \langle x^{(d)}, x^{(d)}\rangle \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_d \end{bmatrix} = \begin{bmatrix} \langle x^{(1)}, x\rangle \\ \vdots \\ \langle x^{(a)}, x\rangle \end{bmatrix}$$

One case where easy to solve the equation is, when the $x^{(k)}$ are all mutually $\perp$ (so the matrix is diagonal), or furthermore, all these

vectors have unit length and mutually orthogonal(so it is an identity matrix).

How do you orthogonalize and normalize a matrix?

Gran-Schmit Procedure:

Let's consider an example: we have already have a basis $x^{(1)}, x^{(2)}$, and we want to find the orthogonal basis $z^{(1)}, z^{(2)}$.

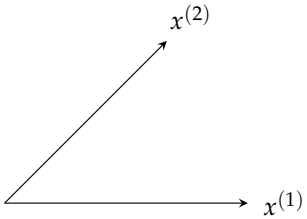


Figure 2.17: Problem setting

Step 1: Normalize $x^{(1)}$

$$z^{(1)} = \frac{x^{(1)}}{\|x^{(1)}\|}$$

Step 2: Orthogonalize $x^{(2)}$

(a). Project $x^{(2)}$ onto $z^{(1)}$

$$\frac{\langle x^{(2)}, z^{(1)} \rangle}{\|z^{(1)}\|} z^{(1)} = \langle x^{(2)}, x^{(1)} \rangle z^{(1)} = u$$

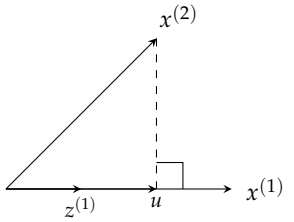


Figure 2.18: Step 1

(b). Normalize to obtain $z^{(2)}$

$$\frac{x^{(2)} - u}{\|x^{(2)} - u\|}$$

The above procedure could be extended to higher dimensions as needed.

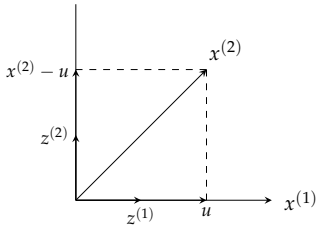


Figure 2.19: Step 2

Stacking up results and yields the **QR decomposition**

$$A = \begin{bmatrix} \vdots & \vdots & \cdots & \vdots \\ x^{(1)} & x^{(2)} & \cdots & x^{(m)} \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix} = \begin{bmatrix} \vdots & \vdots & \cdots & \vdots \\ z^{(1)} & z^{(2)} & \cdots & z^{(m)} \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & \cdots \\ 0 & r_{22} & r_{23} & \cdots \\ 0 & 0 & r_{33} & \cdots \\ \vdots & \ddots & & \end{bmatrix}$$

$$= QR$$

$$= \begin{bmatrix} \vdots & \vdots & \cdots \\ r_{11} z^{(1)} & r_{12} z^{(1)} + r_{22} z^{(1)} & \cdots \\ \vdots & \vdots & \cdots \end{bmatrix}$$

Note that any non singular square matrix A could be decomposed in this way.

Project onto affine set

Recall that all subspace must go through origin, and an "affine" set is a shift/translate of a subspace, and thus it seems that it can't be too difficult to project onto this kind of set. An affine set $\mathcal{A}$ is defined as

$$\mathcal{A} = \{x \in \Omega | x = u + x^{(0)}, u \in U, x^{(0)} \in \mathcal{A}\}$$

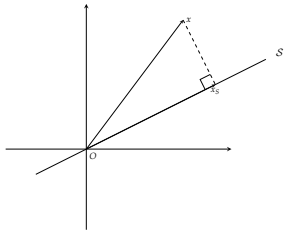


Figure 2.20: Subpace S

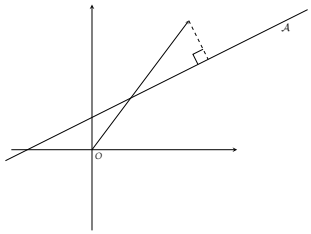


Figure 2.21: Affine set $\mathcal{A}$ as a shifted

Note that we can shift S be any point in $\mathcal{A}$.

The idea of finding projection onto affine set:

Step (0) Goal: to project $x \in \Omega$ onto $\mathcal{A}$.

Step (1) Translate both x and $\mathcal{A}$ by $-x^{(0)}$, and note that the translation of $\mathcal{A}$ is S.

Step (2) Project(translate) $x - x^{(0)}$ onto S(as we did before), and shift result back by $+x^{(0)}$.

Step (3) Get the projection point in $\mathcal{A}$.

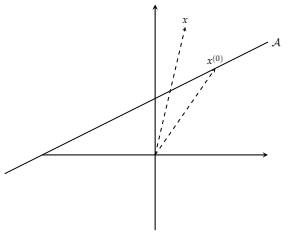


Figure 2.22: Step (0)

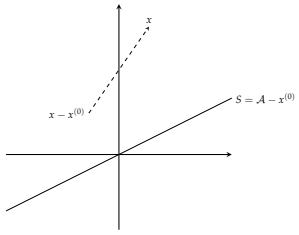


Figure 2.23: Step (1)

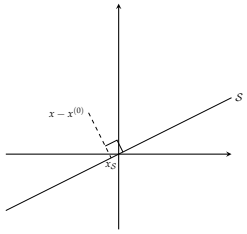


Figure 2.24: Step (2)

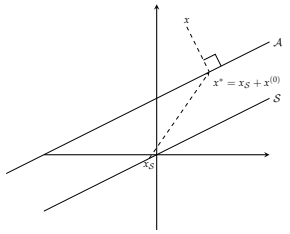


Figure 2.25: Step (3)

Theorem: Projection onto affine set

Let $\mathcal{A} \in \Omega$ be an affine set, where Ω is an inner product space and $\mathcal{A} = S + x^{(c)}$. There is a unique $x^* \in \mathcal{A}$ such that

$$x^* = \arg_{y \in \mathcal{A}} \min \|y - x\|$$

A necessary and sufficient(set of) conditions:

(1). $x^* \in \mathcal{A}$

(2). $(x - x^*) \perp S$

Proof: Any $y \in A$ can be expressed as $y = z + x^{(0)}$ when $z \in S$

$$\begin{aligned} \min_{y \in A} \|y - x\| &= \min_{(z + x^{(0)}) \in A} \|z + x^{(0)} - x\| \\ &= \min_{z \in S} \|z - (x - x^{(0)})\| \end{aligned}$$

Thus $z^* = \arg\min_{z \in S} \|z - (x - x^{(0)})\|$, and translating back, we have

$$x^{(*)} = z^{(*)} + x^{(0)}$$

What are the conditions for optimality? That is,

$z^{(*)} - (x - x^{(0)}) \perp S$, where $z^* \in S$ is obtained by projection onto S.

Thus, in terms of optimal x^*,

(1) $x^{(*)} = z^{(*)} + x^{(0)} \in A$.

(2) $(z^{(*)} + x^{(0)} - x) \perp S \Leftrightarrow (x^{(*)} - x) \perp S$.

Example: Projection onto a hyperplane

A **hyperplane** is an affine set $\mathcal{H}$ specified by the pair $(a, b) \in \mathbb{R}^n \times \mathbb{R}$.

$$\mathcal{H} = \{z \in \mathbb{R}^n | a^T z = b\}$$

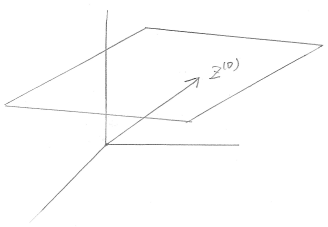
An equivalent definition is

$$\mathcal{H} = \{z \in \mathbb{R}^n | z = u + z^{(0)}, u \in S, z^{(0)} \in \mathcal{H}\}$$

where S is the subspace $S = \{z \in \mathbb{R}^n | a^T z = 0\}$.

Note:

(1) The "equivalent" definition is clearly that of an affine set.



(2) $a \neq 0$ is termed as the "normal" direction.

(3) If $b \neq 0$, then $\mathcal{H} = \mathcal{S}$ and so it is a subspace.

A hyperplane is a special type of affine set. The dimension of the hyperplane is $n-1$ given that the subspace S has a dimension of n. For example,

(1) A (2-D) plane is a hyperplane in $\mathbb{R}^3$.

(2) A line is a hyperplane in $\mathbb{R}^3$.

(3) A line is not hyperplane in $\mathbb{R}_3$.

Exercise: Prove equivalence of above 2 definition

Let's Start with $\mathcal{H} = \{z \in \mathbb{R}^2 | a^T z = b\}$, and let $z^0 \in \mathcal{H}$(any element at $\mathcal{H}$ will do).

Then, since $a^T z^0 = b$ and for any $z \in \mathcal{H}$, we have

$$a^T z - b = 0 \Leftrightarrow a^T - a^T z^{(0)} = 0 \Leftrightarrow a^T (z - z^{(0)}) = 0$$

Thus,

$$\mathcal{H} = \{z | a^T (z - z^{(0)}) = 0\} \qquad (*)$$

Now, since $(\text{span}\{a\}) = \{x | x = \lambda a, \lambda \in \mathcal{R}\}$ is a subspace of dim 1, the set $(\text{span}\{a\})^{\perp}$ is a subspace of dim $n-1$.

Let $\mathcal{S}$ be this subspace.

Observe that by (*), vectors in $\mathcal{H}$ when translate by $-z^{(0)}$ on $\mathcal{S}$.

Therefore $\mathcal{H}$ is $\mathcal{S}$ translated by $+z^{(0)}$ yields $z^{(0)}$.

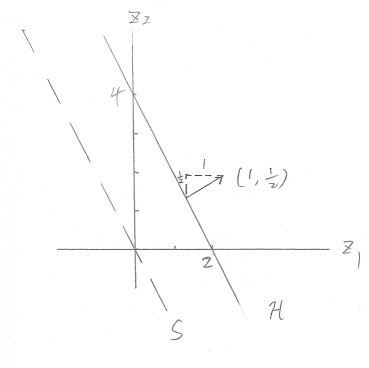
Now, let's start with $\mathcal{H} = \{z \in \mathbb{R}^n | z = u + z^{(0)}, u \in \mathcal{S}, z \in \mathcal{H}\}$.

Let $\{u^{(1)}, \ldots, u^{(n-1)}\}$ be a basis for u, and we choose $a \perp u^{(i)}$, $i \in [n-1]$ and let $b = a^T z^0$. Then for any $z \in \mathcal{H}$, we have $a^T z = 0 + a^T z^{(0)} = b$.

Example: 2-D case Let normal direction be $a = (1, \frac{1}{2})$, offset $b = 2$.

$$\begin{aligned} \mathcal{H} &= \{z \in \mathbb{R}^2 | a^T z = b\} \\ &= \{z \in \mathbb{R}^2 | z_1 + \frac{1}{2} z_2 - 2 = 0\} \end{aligned}$$

$$\mathcal{H} = \{z \in \mathbb{R}^n | z = u + z^{(0)}, u \in \mathcal{S}, z^{(0)} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}\}$$



when

$$\begin{aligned}\mathcal{S} &= \{z \in \mathbb{R}^2 | a^T z = 0\} \\ &= \{z \in \mathbb{R}^2 | z_1 + \frac{1}{2} z_2 = 0\}\end{aligned}$$

Note that recall any value at $z^{(0)} \in \mathcal{H}$ works so alternatively, e.g.,

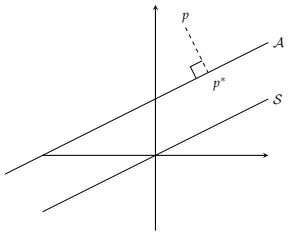
$$\mathcal{H} = \{z \in \mathbb{R}^n | z = u + z^{(0)}, u \in S, z^{(0)} = \begin{bmatrix} 0 \\ 4 \end{bmatrix}\}$$

or

$$\mathcal{H} = \{z \in \mathbb{R}^n | z = u + z^{(0)}, u \in S, z^{(0)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}\} \quad \text{, etc.}$$

Now let's back to the first example, projection onto hyperplane.

$$\mathcal{H} = \{z \in \mathbb{R}^n | a^{\mathrm{T}} z = b\} = \{z \in \mathbb{R}^n | z = x_s + z^{(0)}, x_s \in S, z^{(0)} \in \mathcal{H}\}$$



Recall that $\dim(S) = n - 1$, so $\dim(S^\perp) = 1$, and we want to find the point $p^* = \arg\min_{p \in \mathcal{H}} \|p^* - p\|$.

Observe that $(p - p^*) \perp S$ (by optimal condition), So $(p - p^*) \in S^\perp = \{\lambda a | \lambda \in \mathbb{R}\}$, and $\exists \lambda^*$ such that $p - p^* = \lambda^* a$.

Now, we want to solve for λ^* but notice that there are 2 unknowns (λ^*, p^*), and hence we get rid of p^* dependency by using definition

of $\mathcal{H}$.

$$\begin{aligned} & p - p^* = \lambda^* a \\ \Leftrightarrow\ & a^{\mathrm{T}}(p - p^*) = a^{\mathrm{T}}(\lambda^* a) \\ \Leftrightarrow\ & a^{\mathrm{T}}p - a^{\mathrm{T}}p^*) = \lambda^* a^{\mathrm{T}}a \\ \Leftrightarrow\ & a^{\mathrm{T}}p - b = \lambda^* a^{\mathrm{T}}a \\ \Leftrightarrow\ & \lambda^* = \frac{a^{\mathrm{T}}p - b}{a^{\mathrm{T}}a} \\ \Leftrightarrow\ & \lambda^* = \frac{a^{\mathrm{T}}p - b}{\|a\|^2} \end{aligned}$$

Thus, $p - p^* = \lambda^* a = (\frac{a^{\mathrm{T}}p-b}{\|a\|^2})a$, or $p^* = p - (\frac{a^{\mathrm{T}}p-b}{\|a\|^2})a$.
and

$$\|p - p^*\| = \|\lambda^* a\| = |\lambda^*|\|a\| = \frac{|a^{\mathrm{T}}p - b|}{\|a\|}$$

.

Recall terminology $\|p - p^*\| = \min_{y\in\mathcal{H}} \|y - p\|$, we have

$$p^* = \arg\min_{y\in\mathcal{H}} \|y - p\|$$

Projection w.r.t other norms

Recall that inner product spaces have a notion of angle, have term "orthogonality principle", and the L_2 norm is one such example. In contrast, some norms such as L_1 andL_∞ norms don't come with a sense of angle. However the problem still make senses if $p \neq 2$, e.g., $p = 1$, $p = \infty$, but we cannot apply $\perp$ principle since there is no sense of angle.

In following we will

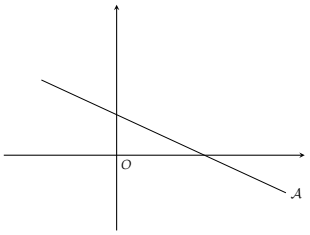
(1). Discuss projection in normed vector spaces, particularly L_1 and L_∞.

(2). Illustrate how the solution differs as you change the norm (change p).

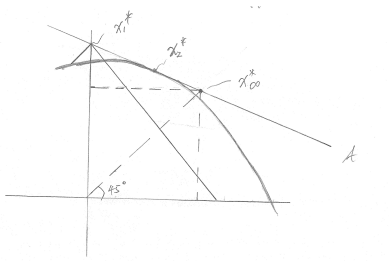
(3). Give you a sense for character of difference such for $p = 1$ and $p = \infty$.

(4). Get a sense of why might pick $p \neq 2$.

Recall norm balls we draw before. Let's project $0 \in \mathbb{R}^2$ onto a line (affine set/hyperplane).



$$x^* = \arg\min_{x \in \mathcal{A}} \|x - 0\|_p = \arg\min_{x \in \mathcal{A}} \|x\|_p$$



Observe that:

x_2^*: Familiar with solution via inner product and $\perp$ theorem, and has a closed form solution.

x_1^*: Solution is "sparse", generally will be the case for affine constraints since vertices of norm-ball are axis-aligned.

x_∞^*: At optimum, $x_{\infty,1}^* = x_{\infty,2}^*$, equal-magnitude coordinate.

Functions

Some terminologies will be used in this material:

"Function": $F : \mathbb{R}^n \mapsto \mathbb{R}$

"Map": $F : \mathbb{R}^n \mapsto \mathbb{R}^m$

However, not all input values may be allowed, input may be a subset of Ω (cf, $\mathbb{R}^n$), this is the "domain" of F.

Aside: Terminology when discussion a pair of vector space $(\mathcal{V}, \mathcal{U})$ over a field $\mathbb{F}$

$F : \mathcal{U} \mapsto \mathcal{V}$, a "map", generally $\dim(\mathcal{U}) \neq \dim(\mathcal{V})$.

$F : \mathcal{U} \mapsto \mathcal{U}$, an "operator", input and output vectors spaces have the same dimension.

$F : \mathcal{U} \mapsto \mathbb{F}$, a "functional", map vector space into a scalar.

In this course, $\mathcal{U}$=$\mathbb{R}^n$, $\mathcal{V}$=$\mathbb{R}^m$, $\mathbb{F}$=$\mathbb{R}$ (or, occasionally $\mathbb{C}$).

Sets related to functions

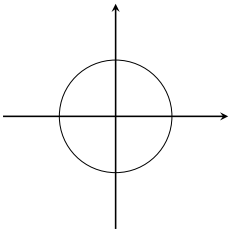


Figure 2.26:

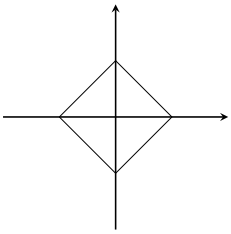


Figure 2.27:

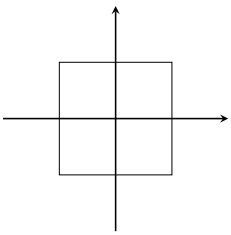


Figure 2.28:

Various sets defined by a function tell us a lot(or sometimes everything) about a function $F : \mathbb{R}^n \mapsto \mathbb{R}$

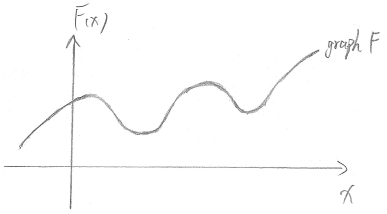
(1) The "graph" (a.k.a, "plot") of F is the set

$$F = \{(x, F(x)) \in \mathbb{R}^{n+1} : x \in \mathbb{R}^n\}$$

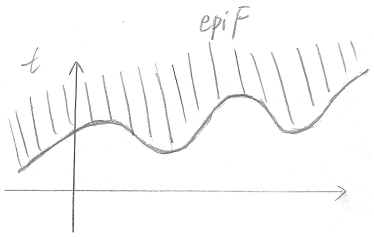
(2) The "epigraph" of F is the set

$$F = \{(x, t) \in \mathbb{R}^{n+1} : x \in \mathbb{R}^n, t \geq F(x)\}$$

Graph



Epigraph



It is also useful to consider points at(or below) a height

(3) The "level" set

$$c_F(t) = \{x \in \mathbb{R}^n : F(x) = t\}$$

(4) The "Sub-level" set

$$L_F(t) = \{x \in \mathbb{R}^n : F(x) \leq t\}$$

Note: graph and epigraph are in $\mathbb{R}^{n+1}$, level and sublevel set are in $\mathbb{R}^n$

Let's sketch these sets for L_2 and L_1 norms in $\mathbb{R}^2$ on the r.h.s.

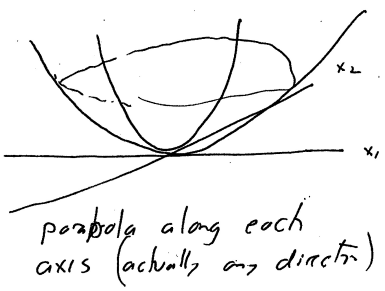


Figure 2.29: Graph 1

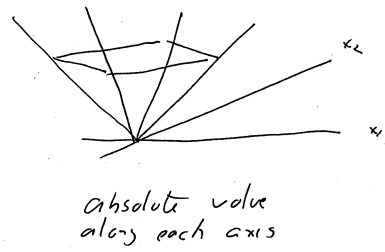


Figure 2.30: Graph 2

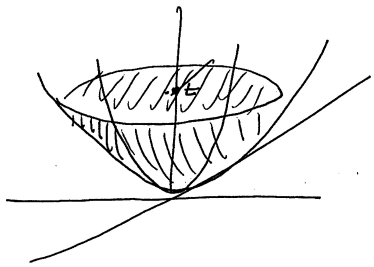


Figure 2.31: Epigraph 1

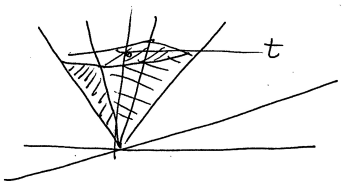


Figure 2.32: Epigraph 2

Linear and affine functions

1. A function $F : \mathbb{R}^n \mapsto \mathbb{R}$ is linear iff following two properties are satisfied

(1) "Homogeneous": $F(ax) = aF(x), \forall x \in \mathbb{R}^n$ and $a \in \mathbb{R}$

(2) "Additivity": $F(x^{(1)} + x^{(2)}) = F(x^{(1)}) + F(x^{(2)})$

Put together to get

$$F(\sum_{i\in[d]} a_i x^{(i)}) = \sum_{i\in[d]} a_i F(x^{(i)})$$

2. A function $F : \mathbb{R}^n \mapsto \mathbb{R}$ is affine iff

Let $\overline{F}$ define pointwise as $\overline{F} = F(x) - F(0), \forall x \in \mathbb{R}^n$ is a linear function. The $F : \mathbb{R}^n \mapsto \mathbb{R}$ is affine iff there is a unique pair $(a, b) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$F(x) = a^{\mathrm{T}}x + b, \ \forall x \in \mathbb{R}^n$$

Since $F(0) = b$, this implies that any linear function can be expressed as $F(x) = a^{\mathrm{T}}x = \langle a, x\rangle$ for some unique $a \in \mathbb{R}^n$.

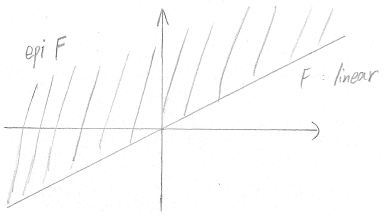
Sets and linear/affine functions

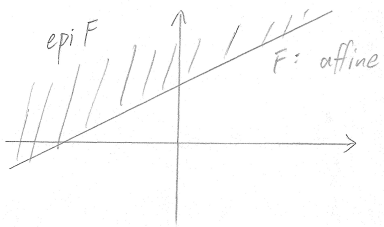
The graph of $F : \mathbb{R}^n \mapsto \mathbb{R}$ is a

- subspace of $\mathbb{R}^{n+1}$ if F is linear.
- hyperplane of $\mathbb{R}^{n+1}$ if F is affine.

The epigraph of $F : \mathbb{R}^n \mapsto \mathbb{R}$ is a

- half-space of $\mathbb{R}^{n+1}$ if F is affine.
- half-space the boarder at which includes $0 \in \mathbb{R}^{n+1}$ if F is linear.





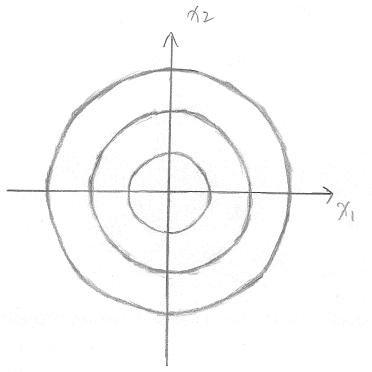


Figure 2.33: Level set 1

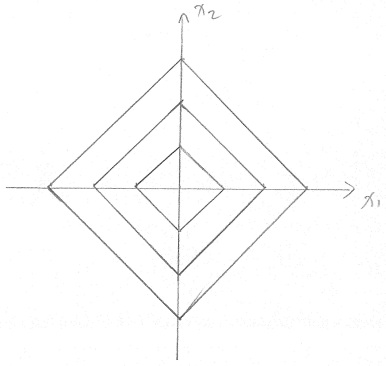


Figure 2.34: Level set 2

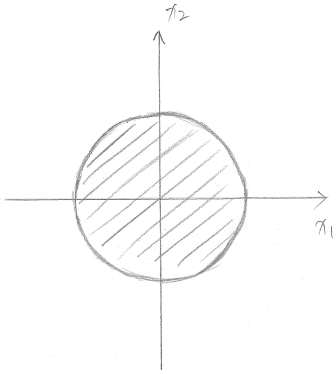


Figure 2.35: Sub-level set 1

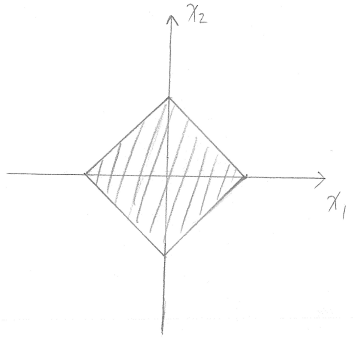
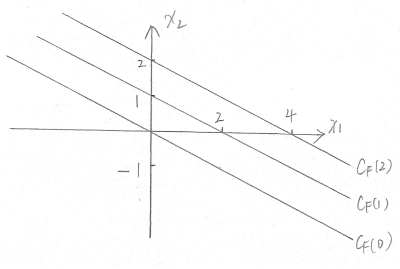


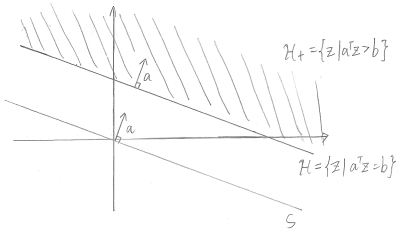
Figure 2.36: Sub-level set 2

Similar statements hold for level sets and sub-level sets in $\mathbb{R}^n$, e.g., level sets of a linear function $F : \mathbb{R}^2 \mapsto \mathbb{R}$ are affine sets in $\mathbb{R}^2$



Definition of a hyperplane:

$$\mathcal{H} = \{z \in \mathbb{R}^n | a^{\mathrm{T}} z = b, a \in \mathbb{R}^n, b \in \mathbb{R}\}$$



Definition of Half-spaces: are on one side or other of a hyperplane(see the r.h.s for a graph of $\mathcal{H}_+$),

$$\mathcal{H}_+ = \{z \in \mathbb{R}^n | a^{\mathrm{T}} z > b\}$$

$$\mathcal{H}_- = \{z \in \mathbb{R}^n | a^{\mathrm{T}} z \leq b\}$$

Gradient

The gradient ∇F of $F : \mathbb{R}^n \mapsto \mathbb{R}$ is the vector of partial derivatives

$$\nabla F = \begin{bmatrix} \frac{\partial F(x)}{\partial x_1} \\ \frac{\partial F(x)}{\partial x_2} \\ \vdots \\ \frac{\partial F(x)}{\partial x_n} \end{bmatrix}, \text{ where } x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Sometime we need to consider compound function, and thus we need chain rule for gradients. Says, $g : \mathbb{R}^n \mapsto \mathbb{R}^m$ and $F : \mathbb{R}^m \mapsto \mathbb{R}$, both F and g are differentiable and we want $\nabla\Phi(x)$, where $\Phi(x) = F(g(x))$. In this case we have

$$\nabla\phi(x) = \begin{bmatrix} \frac{\partial\phi(x)}{\partial x_1} \\ \vdots \\ \vdots \\ \frac{\partial\phi(x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1(x)}{\partial x_1} & \frac{\partial g_2(x)}{\partial x_1} & \cdots & \frac{\partial g_m(x)}{\partial x_1} \\ \frac{\partial g_1(x)}{\partial x_2} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial g_1(x)}{\partial x_n} & \cdots & \cdots & \frac{\partial g_m(x)}{\partial x_n} \end{bmatrix} \nabla F\left(g(x)\right)$$

Example

Let $g : \mathbb{R}^4 \mapsto \mathbb{R}^3$ and $F : \mathbb{R}^3 \mapsto \mathbb{R}$, both F and g are differentiable and we want to find $\nabla\Phi(x)$. To simplify, we let the function g and F takes the form,

$g(x) = Ax + b$, where A is an 3 by 4 matrix, x is a 4 dimensions vector, and b is 3 dimensions vector, so function g maps $\mathbb{R}^4$ to $\mathbb{R}^3$.

$F(x) = Cx$, where C is an 1 by 3 matrix and the input x is a 3 dimensions vector, so function F maps $\mathbb{R}^3$ to $\mathbb{R}$.

Hence, the gradient of $\Phi(x)$ is

$$\begin{aligned}
\nabla\phi(x) &= \begin{bmatrix} \frac{\partial\phi(x)}{\partial x_1} \\ \frac{\partial\phi(x)}{\partial x_2} \\ \frac{\partial\phi(x)}{\partial x_{13}} \\ \frac{\partial\phi(x)}{\partial x_4} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial g_1(x)}{\partial x_1} & \frac{\partial g_2(x)}{\partial x_1} & \frac{\partial g_3(x)}{\partial x_1} \\ \frac{\partial g_1(x)}{\partial x_2} & \frac{\partial g_2(x)}{\partial x_2} & \frac{\partial g_3(x)}{\partial x_2} \\ \frac{\partial g_1(x)}{\partial x_3} & \frac{\partial g_2(x)}{\partial x_3} & \frac{\partial g_3(x)}{\partial x_3} \\ \frac{\partial g_1(x)}{\partial x_4} & \frac{\partial g_2(x)}{\partial x_4} & \frac{\partial g_3(x)}{\partial x_4} \end{bmatrix} \begin{bmatrix} \frac{\partial\phi(x)}{\partial g_1} \\ \frac{\partial\phi(x)}{\partial g_2} \\ \frac{\partial\phi(x)}{\partial g_3} \end{bmatrix} \\
&= \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \\ a_{14} & a_{24} & a_{34} \end{bmatrix} \begin{bmatrix} c_{11} \\ c_{12} \\ c_{13} \end{bmatrix} \\
&= A^{\mathrm{T}} C^{\mathrm{T}}
\end{aligned}$$

Affine approximations

Consider the Taylor series for $F\colon \mathbb{R}^n \to \mathbb{R}$.

$$F(x) = F(x_0) + \nabla F(x_0)^T (x - x_0) + \varepsilon(x)$$

Example 1. $F(x) = 2x_1^2 + x_2^2$

$$F(x) \cong F(x^{(0)}) + \nabla F(x^{(0)})^T (x - x^{(0)})\ \nabla F(x)|_x \qquad (*)$$

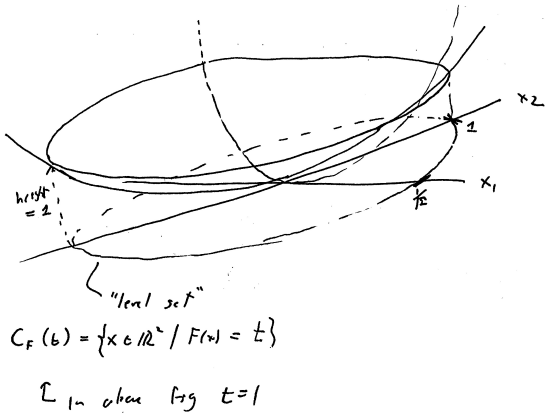
$$\nabla F(x) = \begin{bmatrix} 4x_1 \\ 2x_2 \end{bmatrix}$$

$$\nabla F\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\nabla F\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$$

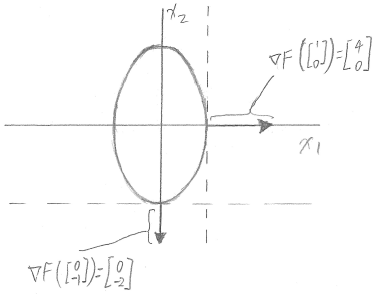
$$\nabla F\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

Let's sketch the "level set" in 2-D,



$C_F(b) = \{x \in \mathbb{R}^2 \mid F(x) = t\}$

In this fig $t = 1$



$$\nabla F\left(\begin{bmatrix}1\\0\end{bmatrix}\right) = \begin{bmatrix}4\\0\end{bmatrix}$$

$$\nabla F\left(\begin{bmatrix}0\\-1\end{bmatrix}\right) = \begin{bmatrix}0\\-2\end{bmatrix}$$

Let's visualize the set such that the increment in $(*)$ is, to first order, constant. That is, which $x \in \mathbb{R}^2$ satisfy the relation

$$\{x | \nabla F(x_0)^T (x - x_0) = c\}$$

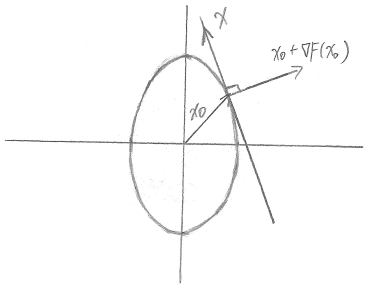
(1) Consider case $c = 0$

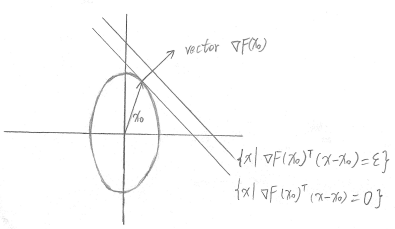
There are points s.t., to approximate $(*)$, has some level as $F(x_0)$ when $c = 0$, we have the set $\{x | \nabla F(x_0)^T (x - x_0) = 0\}$

(2) Consider $c = \varepsilon > 0$, a small positive increment. Then,

$$\{x | \nabla F(x_0)^T (x - x_0) = \varepsilon\}$$

is points that, to first order, have slightly higher cost (value, level). Then $F(x_0)$, the value at $x = x_0$.



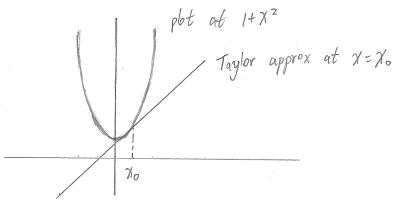


In general the set

$$\begin{aligned}&\{x|\nabla F(x_0)^T(x - x_0) = c\}\\ &= \{x|\nabla F(x_0)^T x = \nabla F(x_0)^T x_0 + c\}\\ &= \{x|a^T x = b\}\end{aligned}$$

which is a "hyperplane", a type of affine set.

Observe that geometry of gradients connects to geometry of level sets. But you might think that is a bit funny. You know a Taylor sense approx is of the function F not the level sets of F. You might also recall there is a tangent approximation involved somewhere, e.g.



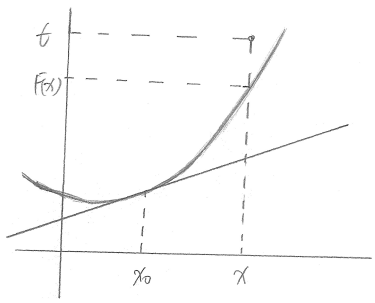
To develop the approximation we need to consider the plot or "graph" at the function F

$$\text{graph } F = \{(x, F(x)|x \in \mathbb{R}^n)\} \subseteq \mathbb{R}^{n+1}$$

e.g., in above example $F(x) = x^2 + 1$, $F\colon \mathbb{R} \to \mathbb{R}$, $s - n = 1$ and plot (graph) is in $\mathbb{R}^2$.

To find the tangent approximation, we will pick a point t "above". The graph, is pick some pair (x, t) s.t. $t \geq F(x)$.

Use Taylor approximation above x_0 to approximate $F(x)$



Recap:

(1) Pick (x, t) s.t. $t \geq F(x)$

(2) Assume x and x_0 are "close" so approximation is accurate.

(3) By Taylor $F(x) = F(x_0) + \nabla F(x_0)^T(x - x_0) + \varepsilon(x)$

(4) By (1), $t \geq F(x) = F(x_0) + \nabla F(x_0)^T(x - x_0) + \varepsilon(x)$

(5) By (2) will drop the $\varepsilon(x)$ term and assume inequality doesn't flip (because x and x_0 are sufficiently close that $\varepsilon(x)$ is sufficient small), and thus yields

$$t \geq F(x_0) + \nabla F(x_0)^T(x - x_0) \qquad (*)$$

Next, we re-arrange

(6)

$$\begin{aligned} 0 &\geq -(t - F(x_0)) + \nabla F(x_0)^T(x - x_0) \\ &= \begin{bmatrix} \nabla F(x_0)^T & -1 \end{bmatrix} \begin{bmatrix} x - x_0 \\ t - F(x_0) \end{bmatrix} \end{aligned}$$

Observe that:

(a) $(x - x_0) \in \mathbb{R}^n$ so vectors are in $\mathbb{R}^{n+1}$.

(b) $t - F(x_0) \in \mathbb{R}$ i.e., in example plot when $n = 1$ in $\mathbb{R}^2$.

Now recall connection between angles and inner products.

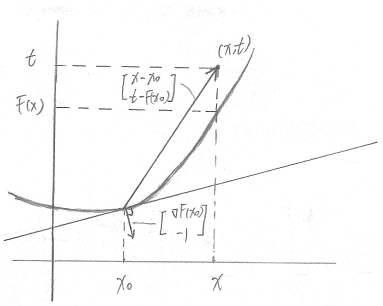
(1). If inner product of 2 vectors is negative, then the angles is obtuse.

(2). Matches picture.

What about vectors when this inner product = 0 ? That is

$$\{u \in \mathbb{R}^{n+1} | \langle u, [\nabla F(x_0), -1]^{\mathrm{T}} \rangle = 0\}$$

writing $u = [x - x_0, t - F(x_0]^{\mathrm{T}}$ and no longer require $t \geq F(x)$.



The condition becomes

$$\begin{bmatrix} x - x_0 \\ t - F(x_0) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \nabla F(x_0) \\ -1 \end{bmatrix} = 0$$

So we recognize the set defines a hyperplane in $\mathbb{R}^{n+1}$,

$$\mathcal{H} = \left\{ \begin{bmatrix} x \\ t \end{bmatrix} \middle| \begin{bmatrix} x^{\mathrm{T}} & t \end{bmatrix} \begin{bmatrix} \nabla F(x_0) \\ -1 \end{bmatrix} = \begin{bmatrix} x_0^{\mathrm{T}} & F(x_0) \end{bmatrix} \begin{bmatrix} \nabla F(x_0) \\ -1 \end{bmatrix} \right\}$$

This is called a "supporting hyperplane" of the epigraph.

Finally, let's look at Taylor approximation one last time (1st order approximation), and recall the geometric interpretation of each of the pieces:

$$\begin{aligned} F(x) &\approx F(x_0) + \nabla F(x_0)^{\mathrm{T}}(x - x_0) \\ &= F(x_0) + \|\nabla F(x_0)\| \|x - x_0\| \left\langle \frac{\nabla F(x_0)}{\|\nabla F(x_0)\|}, \frac{x - x_0}{\|x - x_0\|} \right\rangle \\ &= \text{bias} + \text{steepness} \times \text{distance} \times \text{angle} \end{aligned}$$

3
Matrices and eigen decomposition

3.1 Matrices: array of numbers

Matrices are rectangular arrays of numbers:

$$A = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{m1} & a_{m2} & ... & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

The element in the i^{th} row & j^{th} column: $a_{ij} = [A]_{ij}$(equivalent notation)

The transposition operation works on matrices by exchanging rows and columns:

$$A^T = \begin{bmatrix} a_{11} & a_{21} & ... & a_{1n} \\ a_{12} & a_{22} & ... & a_{m2} \\ ... & ... & ... & ... \\ a_{1n} & a_{2n} & ... & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

So $[A]_{ij} = [A^T]_{ji}$ if $A \in \mathbb{R}^{m \times n}$ and $A^T \in \mathbb{R}^{n \times m}$

Operations of matrices:

1) $A + B = C$, where $[C] = [A]_{ij} + [B]_{ij}$

2) $\alpha A = B$, where $[B]_{ij} = \alpha[A]_{ij}$

Note: the origin is a all-zero matrix.

Definition 3.1. Inner product of matrices

Let $A, B \in \mathbb{R}^{m \times n}$, the inner product of metrics A and B is defined as

$$\langle A, B \rangle = \text{trace}(A^T B) = \text{trace}(B A^T)$$

where $A^T B \in \mathbb{R}^{n \times n}$ and $B^T A \in \mathbb{R}^{m \times m}$, and the trace($X$) for a given matrix X is defined as the sum of the diagonal elements of X.

Length of Matrix: Norm

We introduce the Forbenius norm here but note that there are also other matrix norms(e.g., spectrum norm, nuclear norm)

Forbenius Norm:

$$\|A\|_F = \sqrt{< A, A >} = \sqrt{\text{trace}(A^T A)} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}[A]_{ij}^2}$$

Matrix inverse

An n by n matrix A is called "invertible", if $\exists$ unique A^{-1} s.t. $AA^{-1} = A^{-1}A = I$. The matrix A^{-1} is called the inverse of matrix A

Some properties for invertible matrices:

- $(AB)^{-1} = B^{-1}A^{-1}$, where $A, B \in \mathbb{R}^{n\times n}$
- $(A^{-1})^T = (A^T)^{-1}$
- $\det(A^{-1}) = \frac{1}{\det(A)}$

Now, thinking the matrix $A_{m\times n}$ as a mapping, i.e., $A : \mathbb{R}^n \mapsto \mathbb{R}^m$ (or sometimes denotes as $F : \mathbb{R}^n \mapsto \mathbb{R}^m$, where F is a linear mapping), the inverse of A might thinking as a inverse of the original mapping, that is, map $\mathbb{R}^m$ back to $\mathbb{R}^n$. The question remains is how to achieve this.

Notice that:

(1) If $m < n$, it is impossible to map $\mathbb{R}^m$ back to $\mathbb{R}^n$.

(2) If $m \geq n$, it is possible to achieve this inverse mapping.

Remark: There is a thing called "pseudo inverse" for non square matrix.

Matrices as linear & affine maps

A map $F : \mathcal{V} \mapsto \mathcal{W}$ is "linear if for any two points $x^{(1)}, x^{(2)} \in \mathcal{U}$ and scalar a_1 and a_2 have

$$F(a_1 x^{(1)} + a_2 x^{(2)}) = a_1 F(x^{(1)}) + a_2 F(x^{(2)})$$

It turns out that any linear map is completely specified by a matrix. For example, back to basic linear system, $Ax = y$, the linear mapping maps $x \in \mathbb{R}^n$ to $y \in \mathbb{R}^m$, via the multiplication of matrix A.

Furthermore, affine maps are linear functions plus the offset, i.e., $F(x) = Ax + b$ where $A \in \mathbb{R}^{m,n}$, $b \in \mathbb{R}^m$

Approximations

Recall that previously we have talked about approximation of a function $F : \mathbb{R}^n \mapsto \mathbb{R}$, and now we are going to consider the case $F : \mathbb{R}^n \mapsto \mathbb{R}^m$.

A nonlinear map $F : \mathbb{R}^n \to \mathbb{R}^m$ can be approximated by an affine map(so we are considering a stack of functions now):

$$\begin{aligned}
F(x) &= \begin{bmatrix} F_1(x) \\ F_2(x) \\ \vdots \\ F_m(x) \end{bmatrix} \\
&= \begin{bmatrix} F_1(x_0) \\ F_2(x_0) \\ \vdots \\ F_m(x_0) \end{bmatrix} + \begin{bmatrix} \nabla F_1(x_0)^{\mathrm{T}} \\ \nabla F_2(x_0)^{\mathrm{T}} \\ \vdots \\ \nabla F_m(x_0)^{\mathrm{T}} \end{bmatrix} (x - x_0) + o(\|x - x_0\|) \\
&= \begin{bmatrix} F_1(x_0) \\ F_2(x_0) \\ \vdots \\ F_m(x_0) \end{bmatrix} + \begin{bmatrix} \frac{\partial F_1(x_0)}{\partial x_1} & \cdots & \frac{\partial F_1(x_0)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F_m(x_0)}{\partial x_1} & \cdots & \frac{\partial F_m(x_0)}{\partial x_n} \end{bmatrix} (x - x_0) + o(\|x - x_0\|) \\
&= F(x_0) + J_F(x_0)(x - x_0) + o(\|x - x_0\|)
\end{aligned}$$

where $o(\|x - x_0\|)$ are terms that go to zero faster than 1st order for $x \to x_0$ and $J_F(x_0)$ is the Jacobian of F evaluated at x_0:

$$J_F(x_0) = \begin{bmatrix} \frac{\sigma f_1}{\sigma x_1} & \cdots & \frac{\sigma f_1}{\sigma x_n} \\ \cdots & \cdots & \cdots \\ \frac{\sigma f_m}{\sigma x_1} & \cdots & \frac{\sigma f_m}{\sigma x_n} \end{bmatrix}_{x=x_0}$$

For x 'near' x_0, the variation $\delta_F(x) = F(x) - F(x_0)$ can be approximated described by a linear map:

$$\delta_F(x) = J_F(x_0)\delta_x,\ \delta_x = x - x_0$$

Orthogonal Matrices

Definition 3.2. $U \in \mathbb{R}^{n \times m}$ is **orthogonal** if $U = [U^{(1)} ... U^{(n)}]$ and

$$U^{(i)^T} U^{(j)} = \begin{cases} 0 & \forall i \neq j \\ 1 & \text{if} i = j \end{cases}$$

Then $UU^T = U^T U = I$

Next, let's see how orthogonal transformation do to geometry,i.e., to length and angles between vectors. Let U be an n by n orthogonal matrix which defines a linear map from $x \in \mathbb{R}^n$ to $y \in \mathbb{R}^n$, that is, $y = Ux$.

(a) Length between vectors

$$\|y\|^2 = (Ux)^T(Ux) = x^T U^T U x = x^T x = \|x\|^2$$

(b) Angle between vectors

To compare the angles between vectors, we consider two maps now, i.e., $v = Uw$ and $y = Ux$.

$$\langle y, v \rangle = \langle Ux, Uw \rangle = x^T U^T U w = x^T w = \langle x, w \rangle$$

To summarize, the length of a vector and the angle between two vectors remain the same after an orthogonal transformation.

Range, rank and null space

Let's consider a linear map that $F : x \mapsto Ax$, where $x \in \mathbb{R}^n$ and A is m by n matrix.

Some important terminology regarding this map are illustrated as follows:

- Domain

  $\text{dom}(A) = \mathbb{R}^n$, $A = [a^{(1)} ... a^{(n)}]$
  $\text{dom}(A^T) = \mathbb{R}^m$, $A^T = [a^{(1)} ... a^{(m)}]$

- Range(or, Column space)

  Range of A is the set of vectors y obtained as a linear combination of vectors $x \in \mathbb{R}^n$, and takes the form $y = Ax$.

$$R(A) = \{y \in \mathbb{R}^m | y = Ax = \sum_{i=1}^{n} x_i a^{(i)}\}$$

$$R(A^T) = \{w \in \mathbb{R}^m | w = A^T u = \sum_{i=1}^{m} u_i a^{(i)}\}$$

- Rank

  The dimension of the range of A is called the rank of A:

$$\text{rank}(A) = \dim\{R(A)\} = \dim\{R(A^T)\} = \text{rank}(A^T)$$

- Nullspace (or, Kernel)

  The nullspace of the matrix A is the set of vectors in the input space that are mapped to zero vector:

$$N(A) = \{x \in \mathbb{R}^n | Ax = 0\}$$

- Fundamental Theorem

  We can find that for any $x \in R(A^T)$ and any $z \in N(A)$, it holds that $x^T z = 0$.

  $\mathbb{R}^n = R(A^T) \bigoplus N(A)$: $\forall x \in \mathbb{R}^n$ there is a unique $x = x_{R(A^T)} = x_{N(A)}$.

Theorem 3.3. *Fundamental theorem of linear algebra*

For any given matrix $A \in \mathbb{R}^{m,n}$, it holds that $N(A) \perp R(A^T)$ and $R(A) \perp N(A^T)$, hence

$$\begin{aligned} N(A) \bigoplus R(A^T) &= \mathbb{R}^n \\ R(A) \bigoplus N(A^T) &= \mathbb{R}^m \\ dim N(A) + \text{rank}(A) &= n \\ dim N(A^T) + \text{rank}(A) &= m \end{aligned}$$

Consequently, we can decompose $\forall x \in \mathbb{R}^n$ as the sum of 2 vectors orthogonal to each other, one in the range of A^T and the other one is in the nullspace of A, i.e.,

$$x = A^T \xi + z, \text{ where } z \in \mathrm{N}(A)$$

PageRank

PageRank algorithm: More important web page should be ranked higher

A page's important score can be interpreted as the number of "votes" that a page has received from other pages(or, the sum of importance score received from all its neighbors), and also a page can make a vote to other pages as well. Further more, if one page has made multiple votes to the others, then each vote it made is scaled by its total vote.

Therefore, the importance score of a page(or node) i can be written as:

$$\pi(i) = \sum_{j \to i} \frac{\pi_j}{O_j}$$

where $j \to i$ means the set of all the neighbors(denote each neighbor as j) of i.

Let's consider following example

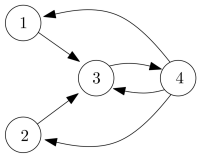


Figure 3.1 A simple example of importance score with 4 webpages and 6 hyperlinks. It is small with much symmetry, leading to a simple calculation of importance scores of the nodes.

Let the score of node 1 be x, and that of node 4 be y. Looking at node 1's incoming links, we see that there is only one such link, coming from node 4 that points to three nodes. So $x = \frac{y}{3}$ and $2x + 2y = 1$. The second equality comes from the normalization of importance scores, and it turns out node 1 and node 2 has the same importance

score, and also node 3 and 4 have the same one as well. So the set of importance scores turns out to be [0.125, 0.125, 0.375, 0.375].

Let's define following terminology:

Matrix H: its (i,j)th entry is $\frac{1}{O_i}$ if there is a hyperlink from webpage i to webpage j, and 0 otherwise.

π: $N \times 1$ column vector denoting the importance scores of the N webpages.

Multiply π^T on the right by matrix H, this is spreading the importance score from the last iteration evenly among the outgoing links, and re-calculating the importance score of each webpage in this iteration by summing up the importance scores from the incoming links. That is

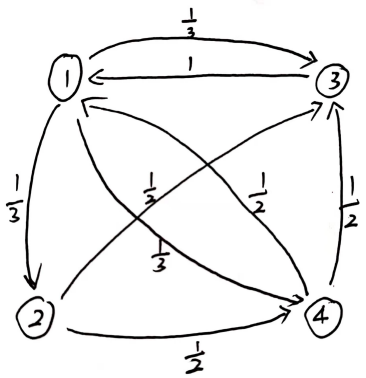
$$\pi^T[k] = \pi^T[k-1]H$$

When k takes a large number(take a large number of iterations), we have the limiting distribution $\pi^{*\mathrm{T}}$ for all the pages,

$$\pi^{*T} = \pi^{*T}H$$

Obviously, π^{*T} is the left eigenvector of H corresponding to the eigenvalue of 1.

Let's take a look in this example. For a graph like this:



$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

The importance scores for each page can be computed as follows

$$\begin{aligned} x &= Ax \\ Ax - x &= 0 \\ (A-I)x &= 0 \\ x &\in N(A-I) \end{aligned}$$

$$x = \frac{1}{s_1}\begin{bmatrix}12\\4\\9\\6\end{bmatrix}$$

Note:

- $x^{(0)}$ is the initial distribution.
- $x_i^{(0)}$ =Pr[start on page i]
- $u^{(i)}, \lambda_i$ is eigen-vector/value pair if $Au^{(i)} = \lambda_i u^{(i)}$, and we arrange this eigenvalues in a decreasing order from $i = 1$, that is, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$

If $Au^{(i)} = \lambda_i u^{(i)}$, A is "diagonalizable" if it has a full set of linear independent eigenvectors. In this case $x^{(0)} = \sum_{i=1}^{n} \alpha_i u^{(i)}$

$$\begin{aligned}
x^{(1)} &= Ax^{(0)} = A[\sum_{i=1}^{n} \alpha_i u^{(1)} = \sum_i \alpha_i (Au^{(i)})] \\
x^{(2)} &= A(Ax^{(0)}) = \sum_{i=1}^{n} \alpha_i (A^2 u^{(i)}) = \sum_{i=1}^{n} \alpha_i (\lambda_i^2 u^{(i)}) \\
&\ldots \\
x^{(k)} &= A^k x(0) = \sum_{i=1}^{n} \alpha_i (\lambda_i)^k u^{(i)} \\
&= \alpha_1 (\lambda_1)^k u^{(1)} + \sum_{i=2}^{n} \alpha_i (\lambda_i)^k u^{(i)} \\
&= \alpha_1 u^{(1)} + \sum_{i=2}^{n} \alpha_i (\lambda_i)^k u^{(i)} \\
&when\, k \to \infty \\
&= \alpha_1 u^{(1)}
\end{aligned}$$

Therefore, we have

$$lim_{k\to\infty} \frac{A^k x^{(0)}}{\|A^k x^{(0)}\|} = u^{(1)}$$

In conclusion, the limiting distribution(the importance score) for each page, is specified by an eigenvector(with the largest eigenvalue).

[Further reading] In the terminology of stochastic process, the matrix H is called the column stochastic matrix (i.e., each column sum up to one and no entries is negative), and important score for each page we want to compute π^* is the stationary distribution (also equal to the limiting distribution when it is an irreducible markov

chain). The $\pi^*(i)$ is interpreted as, the proportion of time you stay in the state i in a long run.

If the matrix A have repeated eigenvalues,

$$Au^{(1)} = \lambda_1 u^{(1)}$$
$$Au^{(2)} = \lambda_2 u^{(2)}$$

clearly:

$$A(\alpha_1 u^{(1)} + \alpha_2 u^{(2)}) = \alpha_1 \lambda_1 u^{(1)} + \alpha_2 \lambda_2 u^{(2)}$$

where $\alpha_1 u^{(1)} + \alpha_2 u^{(2)}$ is this linear combination and $\alpha_1 \lambda_1 u^{(1)} + \alpha_2 \lambda_2 u^{(2)}$ are the maps to a different of same eigenvalues.

1) The "algebraic" multiplicity of an eignevalue λ of a square matrix A is # of eigenvalues $\lambda_i, \lambda_2, ..., \lambda_m$ equal to λ, and we write it as AM(λ).

2) The geometric multiplicity of an eigenvalue λ of a square matrix A is the dimension of $N(A - \lambda I)$, and we write it as GM(λ).

In general, $0 < GM(\lambda) \leq AM(\lambda)$.

If $GM(\lambda_i) = AM(\lambda_i), \forall i$, then A is diagonalizable. If A is diagonalizable, we can write $Au^{(i)} = \lambda_i u^{(i)}$, and now assume all λ_i distinct $GM(\lambda_i) = AM(\lambda_i) = 1, \forall i$

$$\begin{bmatrix} Au^{(1)} & Au^{(2)} & ... & Au^{(n)} \end{bmatrix} = \begin{bmatrix} \lambda_1 u^{(1)} & \lambda_2 u^{(2)} & ... & \lambda_i u^{(i)} \end{bmatrix}$$

$$A \begin{bmatrix} u^{(1)} & u^{(2)} & ... & u^{(n)} \end{bmatrix} = \begin{bmatrix} u^{(1)} & u^{(2)} & ... & u^{(n)} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ ... & ... & ... & \\ 0 & ... & 0 & \lambda_n \end{bmatrix}$$

Or equivalently,

$$AU = U\Lambda$$
$$A = U\Lambda U^{-1}$$
$$\Lambda = U^{-1} A U$$

Recall pagerank:

$$\begin{aligned} A^k x^{(0)} &= (U\Lambda U^{-1})^k x^{(0)} \\ &= U\Lambda^k U^{-1} x^{(0)} \\ &= U \begin{bmatrix} \lambda_1^k & 0 & ... & 0 \\ 0 & \lambda_2^k & ... & 0 \\ ... & ... & ... & \\ 0 & ... & 0 & \lambda_n^k \end{bmatrix} U^{-1} x^{(0)} \end{aligned}$$

It turns out, when a matrix A is diagonalizable, it is easier for us to compute A^k.

Determinant

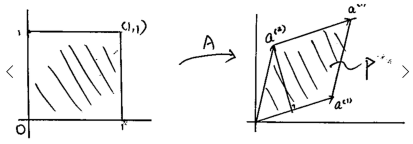
In previous eigenvalue decomposition, we need to solve for λ from $det(A - \lambda I) = 0$ (characteristic function).

The best way the understand determinant net is geometrically as a scaling factor associated with a linear map. Let's take a look at the following example.

Example:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{(1)} & a_{(2)} \end{bmatrix}$$

$$U = \{x \in \mathbb{R}^2 | 0 \leq x_i \leq 1, i \in [2]\}$$
$$P = \{Ax | x \in \mathcal{U}\}$$



In this example, the set $\mathcal{U}$ is mapped to set P, via a linear map(multiply by A). We find that

$$Vol(P) = \det(A) Vol(\mathcal{U})$$

That is, $\det(A)$ plays as scaling factor associate with a linear map.

Recall that if $det(A) = 0$ then A is non-invertible, so if we take a matrix A with $det(A) = 0$ then P will be a line with $Vol(P) = 0$.(Recall that it is impossible to invert the map from a lower dimension space back to a higher dimension space).

Assume A is **diagonalizable**(A is similar with a diagonal matrix):

$$\begin{aligned} A &= U\Lambda U^{-1} \\ |det(A)| &= |det(U\Lambda U^{-1})| \\ &= |det(U)det(\Lambda)det(U^{-1})| \\ &= det(U)det(\Lambda)\frac{1}{det(U)} \\ &= |det(\Lambda)| \\ &= |\prod_{i=1}^{n} \lambda_i| \end{aligned}$$

So the determinant of A is zero if there exists an eigenvalue with zero value, and to summarize, A is not invertible if there is an zero eigenvalue.

4
Symmetric matrices and spectral decomposition

4.1 Symmetric Matrices

The set of n by n square matrix is defined as

$$S^n = \{A \in \mathbb{R}^{n\times n} | A = A^T\}$$

Following are a few examples of symmetric matrix

Example 1: Hessian matrix: A matrix that each element is the 2nd order partial derivative of F

$$[\nabla^2 F]_{ij} = \frac{\sigma}{\sigma x_i}\frac{\sigma}{\sigma x_j}F(x)$$

Example 2: Quadratic Functions: $q : \mathbb{R}^n \to \mathbb{R}$

$$\begin{aligned} q(x) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j + \sum_{i=1}^{n} c_i x_i + d \\ &= x^T A x + c^T x + d \\ &= \frac{1}{2}x^T(A + A^T)x + c^T x + d \\ &= \frac{1}{2}\begin{bmatrix} x^T & 1 \end{bmatrix}\begin{bmatrix} A + A^T & C \\ C^T & 2d \end{bmatrix}\begin{bmatrix} x \\ 1 \end{bmatrix} \end{aligned}$$

1) Let $F(x) = C^T x = \sum_{i=1}^{n} c_i x_i$:

$$\frac{d}{dx_k}F(x) = \frac{d}{dx_k}(\sum_{i=1}^{n} c_i x_i) = c_k$$

$$\nabla F(x) = \begin{bmatrix} \frac{\sigma F(x)}{\sigma x_1} \\ ... \\ \frac{\sigma F(x)}{\sigma x_n} \end{bmatrix} = \begin{bmatrix} c_1 \\ ... \\ c_n \end{bmatrix} = C$$

2) Let

$$\begin{aligned} F(x) &= x^T Ax = \sum_i \sum_j a_{ij} x_i x_j \\ &= a_{11}x_1^2 + a_{12}x_1x_2 + ... + a_{21}x_2x_1 + ... \\ \frac{d}{dx_k}F(x) &= \frac{d}{dx_k}(a_{kk}k_k^2 + \sum_{l\neq k} k_l x_k (a_{kk} + a_{kl})) \\ &= (a_{kk} + a_{kk})x_k + \sum_{l\neq k} x_l(a_{lk} + a_{kl}) \\ &= \sum_{i=1}^{n} (a_{lk} + a_{kl})x_l \\ &= \sum_{i=1}^{n} ([A]_{kl} + [A]_{lk})x_l \end{aligned}$$

Hence,

$$\begin{aligned} \nabla F(x) &= (A + A^T)x \\ [\nabla^2 F(x)]_{kj} &= \frac{d}{dx_j}(\frac{\partial}{\partial x_k}F(x)) \\ &= [A]_{kj} + [A]_{jk} \\ \nabla^2 F(x) &= (A + A^T) \end{aligned}$$

Combine (1) and (2), and because $q(x) = x^T Ax + c^T x + d$, we take Taylor approximation of $q(x)$ up to the second order:

$$\begin{aligned} \tilde{q}(x) &= q(0) + \nabla q(0)^T x + \frac{1}{2}x^T \nabla^2 q(0) x \\ &= d + c^T x + \frac{1}{2}x^T (A + A^T)x \end{aligned}$$

Symmetric Matrices and Eigenvectors

Theorem 4.1. *4.18 & 4.2 in textbook*

For any matrix in $S^n = \{A \in \mathbb{R}^{n\times n} | A = A^T\}$*:*

1) All eigenvalues are purely real(so eigenvectors can be picked purely real).

2) $GM(\lambda_i) = AM(\lambda_i)$*: Symmetric matrix is always diagonalizable.*

3) Eigenvectors of distinct eigenvalues are $\perp$*, i.e.,*

$\xi_{\lambda_i} = N(A - \lambda_i I) \perp \xi_{\lambda_i} = N(A - \lambda_j I)$*, where* ξ_{λ_i} *denotes the eigenspace w.r.t eigenvalue* λ_i *(also, it is the null space of matrix* $(A - \lambda_i I)$*).*

Implication: We can pick the basis for each eigenspace to be an orthogonal basis(e.g, pick the eigenvectors of this symmetric matrix), because we have "Fullset" of eigenvectors(n linearly independent vectors) and we can always write:

"Spectral Decomposition":

$$
\begin{aligned}
A &= U\Lambda U^{-1} \\
&= U\Lambda U^{T} \\
&= \begin{bmatrix} u^{(1)} & u^{(2)} & \cdots & u^{(n)} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_n \end{bmatrix} \begin{bmatrix} u^{(1)^T} \\ \vdots \\ u^{(n)^T} \end{bmatrix} \\
&= \sum_{i=1}^{n} \lambda_i U^{(i)} U^{(i)^T}
\end{aligned}
$$

We summarize our result now: An n by n matrix A is diagonalizable iff there are n linearly independent eigenvectors(subject to scaling factors). In addition, Furthermore, if A is diagonalizable, that is, $A = U\Lambda U^{-1}$, where Λ is diagonal matrix with all its entries are the eigenvalues and U is a collection of all its eigenvectors.

Variational Characterization of eigenvalues of λ_i where $A \in S^n$

We arrange the eigenvalues in a decreasing order, i.e.,

$$\lambda_{max}(A) = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n = \lambda_{min}(A)$$

We define the "Rayleigh quotient" as $\frac{x^T Ax}{x^T x}$ for $x \neq 0$, and we propose a theorem for this ratio as follows

Theorem 4.2. *For $A \in S^n$, we have*

$$\lambda_{min}(A) \leq \frac{x^T Ax}{||x||^2} \leq \lambda_{max}(A), \ \forall x \neq 0$$

Proof.

$$
\begin{aligned}
x^T Ax &= x^T U\Lambda U^T x \\
&= \bar{x}^T \Lambda \bar{x} \\
&= \sum_{i=1}^{n} (\bar{x}_i)^2 \lambda_i \\
&\leq \sum_{i=1}^{n} (\bar{x}_i)^2 \lambda_{max}(A) \\
&= ||\bar{x}||^2 \lambda_{max}(A)
\end{aligned}
$$

where $\bar{x} = U^T x$, and note that $\|\bar{x}\| = \|x\|$ since U is orthogonal. Use similar trick we could obtain the lower bound for $x^T Ax$. By a simple rearrangement we yield the desired result

$$\lambda(A)_{min} ||\bar{x}||^2 \leq x^T Ax \leq ||\bar{x}||^2 \lambda_{max}(A)$$

$$\lambda(A)_{min} \leq \frac{x^T Ax}{||\bar{x}||^2} \leq \lambda_{max}(A)$$

□

Positive (Semi) Definite matrices(PD & PSD)

Definition 4.3. A symmetric matrix $A \in S^n$ is PD (or PSD) if $x^T Ax > 0$, $\forall x \in \mathbb{R}^n$(or $x^T Ax \geq 0$).

Alternatively, we denote the set of PSD matrix and the set of PD matrix as

PSD: $S_+^n = \{A \in S^n | A \succeq 0\}$

PD: $S_{++}^n = \{A \in S^n | A \succ 0\}$

Note: The curled inequality symbol $\succeq$ (and its strict form $\succ$) is used to denote generalized inequality: between vectors, it represents component-wise inequality; between matrices, it represents matrix inequality.

Necessary and sufficient conditions:

(1) A symmetric matrix is PSD iff all its eigenvalues ≥ 0, or, all its **principal minors** are nonnegative.

(2) A symmetric matrix is PD iff all its eigenvalues > 0, or, all its **leading principal minors** are positive (Sylvester's criterion).

Now we prove the argument for PD:

First, assume $A \in S^n$ is PD, we will show that all $\lambda_i > 0$.

$$x^T Ax = x^T U \Lambda U^T x = \bar{x}^T \Lambda \bar{x} = \sum_{i=1}^{n} \lambda_i (\bar{x}_i)^2 > 0$$

Since A is PD and $x \neq 0$, and it is always diagonalizable for a symmetric matrix.

To show this implies $\lambda_j \geq 0, \forall j \in [n]$, we set:

$$\bar{x} = e_j = \begin{bmatrix} 0 \\ 0 \\ ... \\ 1 \\ 0 \\ 0 \\ ... \end{bmatrix}$$

where only the jth entry is 1.

$$0 \leq U^{(i)^T} U \Lambda U^T U^{(1)} = e_j^T \Lambda e_j = \lambda_j$$

And now we Assume all eigenvalues are positive, and we want to show that A is PD:

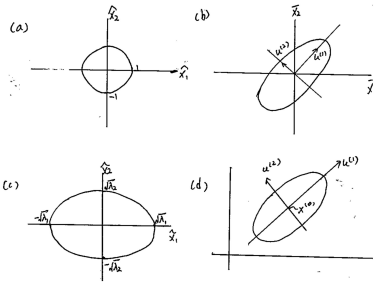
$$x^T Ax = x^T U \Lambda U^T x = \bar{x}^T \Lambda \bar{x} = \sum_{i=1}^{n} (\bar{x}_i)^2 \lambda_i \geq 0$$

The proof is completed.

Recall some previous results and note that:
(1) $\det(A) = \prod_{i=1}^{n} \lambda_i$
(2) $\det(A) = 0$ iff there exist eigenvalue $\lambda_i = 0$:
(3) Combine (1) and (2) and our proof above, we have
→ PD matrices are invertible
→ PSD matrices are invertible only if PD

Ellipses

An ellipse can be defined geometrically as a set or locus of points in the Euclidean plane. Let's consider the following set(an ellipse)



$$\xi = \{x \in \mathbb{R}^n | (x - x^{(0)})^T P (x - x^{(0)}) \leq 1\}$$

where P is PD.

Note that the argument above is a quadratic function:

$$\begin{aligned}(x - x^{(0)})^T P^{-1} (x - x^{(0)}) &= x^T P^{-1} x - 2x^{(0)^T} P^{-1} x + x^{(0)^T} P^{-1} x^{(0)} \\ &= x^T A x + c^T x + d\end{aligned}$$

Let's look at the set ξ, clearly it is centered at $x = x^{(0)}$, and we further simplify it by defining $\bar{x} = x - x^{(0)}$

$$\begin{aligned}1 &\geq (x - x^{(0)})^T P^{-1} (x - x^{(0)}) \\ &= \bar{x}^T P^{-1} \bar{x} \\ &= \bar{x}^T (U \Lambda U^T)^{-1} \bar{x} \\ &= \bar{x}^T (U^T)^{-1} \Lambda^{-1} U^{-1} \bar{x} \\ &= \bar{x}^T U \Lambda^{-1} U^T \bar{x} \\ &= \hat{x}^T \Lambda^{-1} \hat{x} \\ &= \sum_{i=1}^{n} (\frac{\hat{x}_i}{\sqrt{\lambda_i}})^2 \\ &= \sum_{i=1}^{n} (\hat{x}_i)^2 \\ &= ||\hat{x}||^2\end{aligned}$$

where $\hat{x} = U^T \bar{x}$.

Example of where symmetric and PSD matrix are important:
Dataset $x^{(1)}, x^{(2)}, ..., x^{(m)}$ all $x^{(i)} \in \mathbb{R}^n$
Sample mean: $\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$
Sample covariance: $\sum = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu)(x^{(i)} - \mu)^T$, where $(x^{(i)} - \mu)(x^{(i)} - \mu)^T$ is the outer-product of centered data points.

Let's consider an example with $m = 3$:

$$x^{(1)} = \begin{bmatrix}1\\2\end{bmatrix} x^{(2)} = \begin{bmatrix}4\\4\end{bmatrix} x^{(3)} = \begin{bmatrix}4\\0\end{bmatrix} \mu = \begin{bmatrix}3\\2\end{bmatrix} \tilde{x}^{(1)} = \begin{bmatrix}-2\\0\end{bmatrix} \tilde{x}^{(2)} = \begin{bmatrix}1\\2\end{bmatrix} \tilde{x}^{(3)} = \begin{bmatrix}1\\-2\end{bmatrix}$$

where we take $\tilde{x}^{(i)} = x^{(i)} - \mu$, and μ is the sample mean.

So we could compute the covariance matrix by

$$\begin{aligned}\Sigma &= \frac{1}{3}\left(\tilde{x}^{(1)}\tilde{x}^{(1)\mathrm{T}} + \tilde{x}^{(2)}\tilde{x}^{(2)\mathrm{T}} + \tilde{x}^{(3)}\tilde{x}^{(3)\mathrm{T}}\right)\\ &= \begin{bmatrix}2 & 0\\ 0 & \frac{8}{3}\end{bmatrix}\end{aligned}$$

It could be easily verify that the quadratic function $(x-\mu)^T\Sigma^{-1}(x-\mu)$ could be visualize as an ellipses with the choice $\gamma = 2$, i.e., the set $xi_\gamma = \{x|(x-\mu)^T\Sigma^{-1}(x-\mu) \leq \gamma\}$ is an ellipses.

To prove $\Sigma \geq 0$, let's consider sample variance of the scalar product for $i \in [m]$ with choice $\|w\| = 1$

$$S^{(i)} = w^T x^{(i)} = \langle w, x^{(i)}\rangle$$

sample mean:

$$\tilde{S} = \frac{1}{m}\sum_{i=1}^{m} s^{(i)} = \frac{1}{m}\sum_{i=1}^{m} w^T x^{(1)} = w^T\mu$$

sample variance:

$$\begin{aligned}\sigma^2 &= \frac{1}{m}\sum_{i=1}^{m}(s^{(i)} - w^T\mu)^2\\ &= \frac{1}{m}\sum_{i=1}^{m}(w^T(x^{(i)} - \mu))^2\\ &= \frac{1}{m}\sum_{i=1}^{m} w^T(x^{(i)} - \mu)(x^{(i)} - \mu)^T w\\ &= w^T[\frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu)x^{(i)} - \mu)^T]w\\ &= w^T\sum w\end{aligned}$$

Hence it is obviously non negative(so it is PSD) for any choice of w. The proof is completed.

Square-root matrix and Cholesky decomposition

From previous results(spectral decomposition), any PSD(and certainly for any PD) matrix can be written as

$$\begin{aligned}A &= U\Lambda U^{-1}\\ &= U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}U^T\\ &= U\Lambda^{\frac{1}{2}}U^T U\Lambda^{\frac{1}{2}}U^T\end{aligned}$$

where $\Lambda^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\lambda_1} & ... & ... \\ ... & ... & ... \\ ... & ... & \sqrt{\lambda_n} \end{bmatrix}$, and the third equality is obtained since $U^T U = I$ (U is orthogonal and so $U^{-1} = U^T$).

The $A^{\frac{1}{2}} = U\Lambda^{\frac{1}{2}}U^T \rightarrow$ is called the square root matrix of A, and furthermore, A is PSD(PD) iff there exists a unique $A^{\frac{1}{2}}$ is a PSD(PD) matrix.

Now, let's see how to obtain the Cholesky decomposition. We rewrite the matrix decomposition as

$$\begin{aligned} A &= U\Lambda U^{-1} \\ &= U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}U^T \\ &= U\Lambda^{\frac{1}{2}}U^T U\Lambda^{\frac{1}{2}}U^T \\ &= \beta^T\beta \\ &= (QR)^T QR \\ &= R^T Q^T QR \\ &= R^T R \end{aligned}$$

where we let $\beta = \Lambda^{\frac{1}{2}}U^T$, and apply QR decomposition on this square matrix β(recall that it is unique to any square matrix). Finally, we express matrix A as a product of triangular matrices, where R^T is lower triangular and R is upper triangular.

5
Singlar value decomposition

5.1 The Singular Value Decomposition(SVD)

Let's review our previous result before starting the SVD part.

Eigen-decomposition

For any $A \in \mathbb{R}^{n\times n}$ that is diagonalizable, we can express A as

$$A = U\Lambda U^{-1}$$

U: n by n invertiable matrix of linear independent eigenvectors $\in \mathbb{C}^n$, that is, each column of U is an eigenvector of A.

Λ: a diagonal matrix whose diagonal entries are eigenvalues of A, and $\lambda_i \in \mathbb{C}$.

Spectral

For any $A \in S^n$ can express A as

$$A = U\Lambda U^T$$

U: Orthogonal matrix ($\perp$ & normalized) $U^{(i)} \in \Re^n$;

Λ: Diagonal matrix of $\lambda_i \in \Re$.

SVD

Any matrix $A \in \Re^{m\times n}$ can be expressed as

$$A = U\tilde{\Sigma}V^T$$

$U \in \Re^{m\times m}$: orthogonal so $UU^T = U^TU = I_m$

$V \in \Re^{n\times n}$: orthogonal so $VV^T = V^TV = I_n$

$$\tilde{\Sigma} = \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \Re^{m\times n}$$

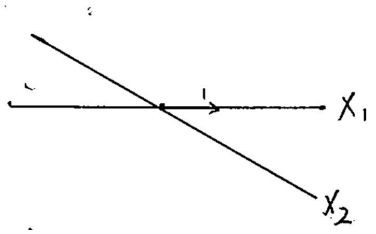
where $\Sigma = diag(\sigma_1, ..., \sigma_r) \geq 0$
Comments:
SVD:

- inherits $\perp$ matrices of spectral decomp and purely real λ_i
- Generalizes to non-square matrices
- Loose property of direction invariance of eigen-decomposition
- Consider $Ax = U\Sigma V^T x$.

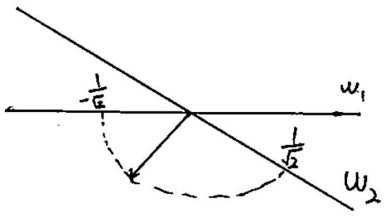
Example 5.1. $y = Ax = U\tilde{\Sigma}V^T x$

$$U = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}} \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} V = \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} A = \begin{bmatrix} -\frac{2}{\sqrt{6}} & \frac{2}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} & \frac{2}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} & \frac{2}{\sqrt{6}} \end{bmatrix}$$

$$a) x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

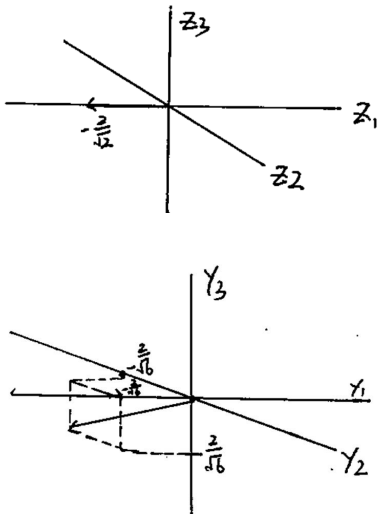


$$b) w = V^T x = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$



$$c) z = \Sigma w = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix}$$

$$d) y = Uz = \begin{bmatrix} -\frac{2}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \end{bmatrix}$$



SVD follows from eigen-decomp of A^TA and $AA^T \rightarrow$ both symmetric, so spectral theorem applies.

Write out $U\&V$ as: (r: rank(A))

$$U = \begin{bmatrix} U^{(1)} & ... & U^{(r)} & U^{(r+1)} & ... & U^{(m)} \end{bmatrix} = \begin{bmatrix} U_r & U_{mr} \end{bmatrix}$$

$$V = \begin{bmatrix} V^{(1)} & ... & V^{(r)} & V^{(r+1)} & ... & V^{(n)} \end{bmatrix} = \begin{bmatrix} V_r & V_{nr} \end{bmatrix}$$

$$AA^T = U\tilde{\Sigma}V^TV\tilde{\Sigma}^TU^T \quad = \begin{bmatrix} \mathcal{U}_r & \mathcal{U}_{mr} \end{bmatrix} \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \Sigma^r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathcal{U}_r^T \\ \mathcal{U}_{mr}^T \end{bmatrix} = \begin{bmatrix} \mathcal{U}_r & \mathcal{U}_{mr} \end{bmatrix} \begin{bmatrix} \Sigma^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathcal{U}_r^T \\ \mathcal{U}_{mr}^T \end{bmatrix} \quad = \sum_{i=1}^{r} (\sigma_i)^2 u^{(i)} (u(i))^T$$

$$= \sum_{i=1}^{m} (\sigma_1)^2 u^{(i)} u^{(i)^T}$$

$(AA^T)u^{(k)}, \rightarrow$ WTS $u^{(k)}$ is k^{th} eigenvector of AA^T.

$(AA^T)u^{(k)} = \sum_{i=1}^{m} \sigma_i^2 u^{(1)} (u^{(1)})^T u^{(k)} = \sum_{i=1}^{n} \sigma_i^2 \prod (i = k) u^{(1)} = \sigma_k^2 u^{(k)}$

$\lambda_k(AA^T) = \sigma_k^2$

same logic applied to A^TA shows $V^{(k)}$ is k^{th} eigenvector of (A^TA)

Computing SVD

1) Singular values: Compute eigenvalues of AA^T or A^TA to find $\lambda_i(A^TA) = \sigma_i^2 \rightarrow \sigma_i = \sqrt{\lambda_i(A^TA)}$

$x^TA^TAx = w^Tw = ||w||^2$

2) Right-Singular vectors: eigenvectors of A^TA;

3) Left-Singular vectors: eigenvectors of AA^T;

4) Because $AA^T\&A^TA$ both symmetric, both have full set of eigenvectors.

2. Consider arbitrary $x \in \Re^n$,

$$\begin{aligned} Ax &= U\tilde{\Sigma}V^T x \\ &= \begin{bmatrix} U_r & U_{mr} \end{bmatrix} \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{mr}^T \end{bmatrix} x \\ &= \begin{bmatrix} U_r & U_{mr} \end{bmatrix} \begin{bmatrix} \Sigma \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} V_r^T \end{bmatrix} x \\ &= \Sigma_{i=1}^r \sigma_i u^{(i)} (V^{(1)})^T x (***) \end{aligned}$$

Eq(*) looses all components of x along $V^{(i)}$ directions when $r+1 \leq i \leq n$:

$\rightarrow$ i.e. columns of V_{nr}

$\rightarrow$ columns of V_{nr} provide basis for $N(A)$.

All direction in output are in span $\{u^{(1)}, ..., u^{(n)}\}$

$\rightarrow$ columns of U_r provide basis for $R(A)$.

Columns of U_{mr} provide basis for $N(A^T)$

Columns of V_r provide basis for $R(A^T)$

2) Consider arbitrary $x \in \Re^n$

$$\begin{aligned} A &= \mathcal{U}\tilde{\Sigma}V^r \\ &= \begin{bmatrix} \mathcal{U}_r & \mathcal{U}_{mr} \end{bmatrix} \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{nr}^T \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{U}_r & \mathcal{U}_{mr} \end{bmatrix} \begin{bmatrix} \Sigma \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} V_r^T \end{bmatrix} \\ &= \mathcal{U}_r \Sigma V_r^T \\ \mathcal{U}_r^T A V_r &= \mathcal{U}_r^T (\mathcal{U}_r \Sigma V_r^T) V_r = \Sigma = \mathcal{U}_r^T A V_r \end{aligned}$$

$$\begin{aligned} A &= \mathcal{U}_r \Sigma V_r^T && (5.1) \\ &= \sum_{i=1}^r \mathcal{U}^{(i)} (V^{(i)})^T \sigma_i && (5.2) \\ &= \sigma_1 \mathcal{U}^{(1)} V^{(1)^T} + \sigma_2 \mathcal{U}^{(2)} V^{(2)^T} && (5.3) \end{aligned}$$

$$\sigma_1 \mathcal{U} V^{(1)^T} = \begin{bmatrix} \sigma_1 V_1^{(1)} \mathcal{U}^{(1)} & \sigma_1 V_2^{(1)} \mathcal{U}^{(1)} & ... & \sigma_1 V_n^{(1)} \mathcal{U}^{(1)} \end{bmatrix}$$

Condition #

Consider $Ax = b$ when A is invertible, solve for x as $x = A^{-1}b$. What if $b = b_r + e$? how much does solution change?

Solution is $\hat{x} = A^{-1}b_T + A^{-1}e$

$$\frac{||\frac{A^{-1}e||}{||A^{-1}b_T||}}{\frac{||e||}{||b||}} = \frac{||A^{-1}e||_2}{||e||_2}\frac{||b||_2}{||A^{-1}b||_2}$$

$$\begin{aligned} max_{e,b\neq 0}\frac{||A^{-1}e||}{||e||}\frac{||b||}{||A^{-1}b||} &= \left[max_{e\neq 0}\frac{||A^{-1}e}{||e||}\right]\left[max_{b\neq 0}\frac{||b||}{||A^{-1}b||}\right] \\ &= \frac{\sigma_{max}(A^{-1})}{\sigma_{min}(A^{-1}} \\ &= \frac{\frac{1}{\sigma_n}}{\frac{1}{\sigma_1}} \\ &= \frac{\sigma_1}{\sigma_n} \\ &= \frac{\sigma_{max}(A)}{\sigma_{min}(A)} \qquad = K(A) \end{aligned}$$

$$A^{-1} = (\mathbb{U}\Sigma V^T)^{-1} \tag{5.4}$$

$$= V\Sigma^{-1}\mathcal{T} \tag{5.5}$$

$$= V\begin{bmatrix} \frac{1}{\sigma_1} & & \\ & \dots & \\ & & \frac{1}{\sigma_n} \end{bmatrix}\mathcal{U}^T \tag{5.6}$$

$$\sigma_{max}(A^{-1}) = \frac{1}{\sigma_n} \tag{5.7}$$

$$\sigma_{min}(A^{-1}) = \frac{1}{\sigma_1} \tag{5.8}$$

Bibliography