STARK DRAPER

COURES NOTES:

OPTIMIZATION THEORY AND ALGORITHMS

COURSE NOTES: VERSION 1.01

September 2019

Contents

List of Figures

List of Tables

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 $\mathrm{dom} f$. It is often convenient to define $f(x) = \infty$ for all $x \notin \mathrm{dom} f$. In that case $\mathrm{dom} f = \{x \in \mathcal{X} | |f(x)| < \infty\}$. 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 $\mathrm{dom} f = \mathbb{R}_+$ is plotted in Fig. 1.1.

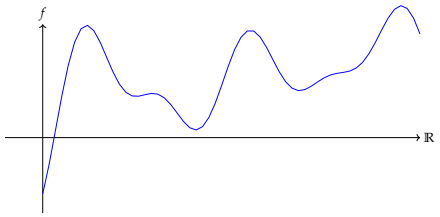


Figure 1.1: A function $f : \mathbb{R} \to \mathbb{R}$.

2
Vectors and functions

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

As mentioned in introduction, the 1st part of this course will focus on geometry. Linear algebra is the mathematical study of geometry in (arbitrary large) dimensions.

It turns out that your geometry xxx from $\mathbb{R}^2$ to $\mathbb{R}^3$(planes and space) is extremely helpful to conceive of large dimensional sets and understand operations on them.

Lots(but not all) of what we cover in the first few topics will repeat what you saw in your linear algebra course(Math 188/185).

Why repeat?

-Linear algebra in year 1, perhaps semester 1.

-It takes time xxx, to "get" linear algebra. If you are in this course, you are likely to come up a lot going forward.

-In your linear algebra course may have concentrate more on the "algebra" side of linear algebra rather than the geometry side. Our focus will be on the latter. E.g., $y - x^2 = 0$ is an algebraic relation but defines a geometric object(a parabola).

-If you took ECE216 with me you will see some familiar examples, and I encourage you to connecting questions (perhaps after class as xxx all students took ECE216).

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 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: book not 100% consistent, XXX

Vector Space

To this point a vector is just a lost of numbers.

To get to geometry, we need to define how to add pairs of vectors and how to scale vectors.

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

Note that If $a = 0$, then $u = av^{(1)} = 0$.

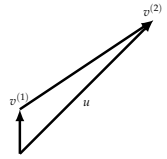


Figure 2.1: Add

With respect to the XXX at the origin can think of as xxx as a displacement(a move through the space) from the origin.

For any vector $v \in \mathbb{R}$, $v = v - 0$.

Note: XXX

Vector Space: a set of vectors that is closed under addition and scaling.

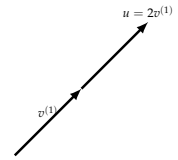


Figure 2.2: Scale

Formally we need following axioms:

Commutativity: $u + v = v + u$

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

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

Identity element of addition: $\exists 0 \in \gamma$ s.t. $u + 0 = u$

Inverse elements of addition: $\exists -u \in \gamma$ s.t. $u + (-u) = 0$

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

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.

① Finite-length complex vector

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

② ∞-length complex sequences(DT signals&system)

③ Complex functions defined on real line(CT signals&system)

④ Polynomials of degree at most n-1

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

It can xxx linear combinations of polynomials and doesn't increase degree, so closed.

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

If I give you a set of vectors, thinking of each as a displacement, anywhere you can get to via linear combinations is the "span" at the set.

If $S = \{v^{(1)}, v^{(2)}, \cdots, v^{(m)}\}$, where each $v^{(i)} \in \mathbb{R}^n$

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

Example 1

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

then

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

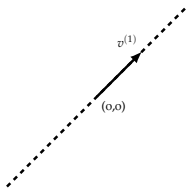


Figure 2.3:

Example 2

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

$$span(S) = \{a_1 v^{(1)}, a_2 v^{(2)} | (a_1, a_2) \in \mathbb{R}_2\}$$
$$= \left\{ \begin{bmatrix} x \\ y \\ 0 \end{bmatrix} | x \in \mathbb{R}, y \in \mathbb{R} \right\}$$
$$= \text{x-y plane}$$

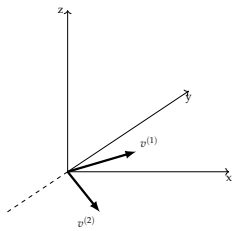


Figure 2.4:

In fact, the span of a set of vectors is a "subspace" is a subset of the XXX vector space($\mathbb{R}^n$ or $\mathbb{C}^n$). XXXX properties of a vector space

Note: $0 \in \mathbb{R}_n$ always included since we can set all coefficients $a_i = 0$ for all i.

Subspace is a "flat" that goes through the origin.

Linear independent set

$S = \{v^{(1)}, \cdots, v^{(n)}\}$ is a linear, independent set if 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 $\sum_{i=1}^{m} a_i v^{(i)} = 0$ is if $a_i = 0 \; \forall i \in [m]$

If this were not true, letting $l \in [m]$ be s.t. $a_l \neq 0$ would have

$$a_l v^l + \sum_{i \neq l} a_i v^{(i)} = 0$$

$$v^{(l)} = \sum_{i \neq l} (\frac{-a_i}{a_l}) v^{(i)}$$

Note that $a_l \neq 0$. So it is not linearly independent.

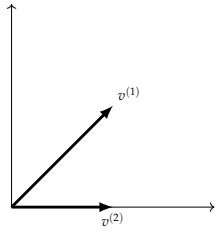


Figure 2.5:

Importance of linearly independent

For any $u \in span(S)$ that is a unique choice of the a_i in the expression $u = \sum_{i=1}^{m} a_i v^{(i)}$, i.e., only one way to express.

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

No redundancy is representation.

observe: it did have redundancy in S. For example,

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

can always shrink S by removing elements to get a linearly independent set. Such an irreducible or linearly independent set can serve as a basis for span(S).

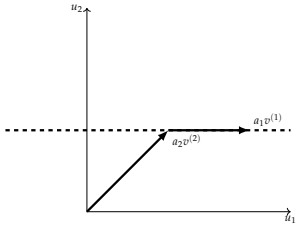


Figure 2.6:

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), dim(span(S))=k.

Example

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

an linearly independent spanning set form a basis for

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

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

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

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

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

Example

Perhaps most familiar basis is the "standard" basis

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

often see "e" for standard basis, i.e., $e^{(i)} = v^{(i)}$.

Note that our book uses e_i for $e^{(i)}$.

Norms: Another important property, idea of distance on length

Familiar: Euclidean distance

But not only notion of distance. E.g. walking through downtown Toronto or NYC. Blocks you walk along the shortest park.

figure here

So, multiple sense of distance, a sense of distance for a vector is a "norm", some properties on norm must satisfy.

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

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

(b)$\|u + v\| \leq \|u\| + \|v\|, \forall u, v \in \gamma$.

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

Note:γ can be either $\mathbb{R}$ or $\mathbb{C}$, if $\gamma \in \mathbb{C}$ we should have $a \in \mathbb{C}$ in (c).

a family of norms but will come up after are:

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
XXX

i.e., carnality of non-zero support

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

Sometimes it is called "L_0" $norm \|x\|_0$ since

$$card(x) = \lim_{p\to 0} (\sum_{k=1}^{n} |x_k|^p)^p$$

But not a norm (so this terminology is inaccurate). E.g., it doesn't satisfy property (c),

$$card(2x) = card(x) \neq 2card(x)$$

To visualize a norm we often plot it unit norm-ball

$$\beta_p = \{x | \|x\|_p \leq 1\}$$

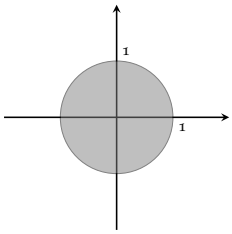


Figure 2.7:

L_2

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

(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

$L_\infty : \{x | \max\{|x_1|, |x_2|\} \leq 1\}$

what about "L_0"? $\{x | card(x) \leq 1\}$

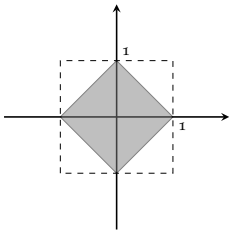
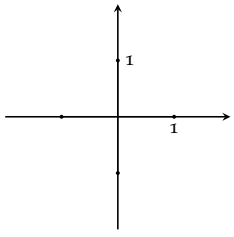


Figure 2.8:



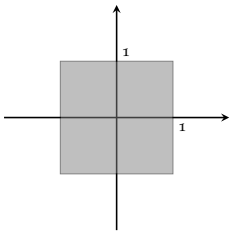


Figure 2.9:

Not much of a "ball"

To visualize a bit more look at "level sets" of the norm balls.

$\{x||x| = c\}$, see for $c = \frac{1}{2}, 1, 2$

L_1

L_2

L_∞

Why might we be interested in different norms?

Ex: Later see applications in optimal control XXX want to meet a XXX objective (move a XXX from point a to XXX point b) while minimizing some resources $\rightarrow$ XXX will be XXX min a norm XXX XXX XXX

$L_2 \rightarrow$ get min energy solution

$L_1 \rightarrow$ get a sparse solution not many forms of jet, useful when "XXX up" overhead

L_∞ all uses of resource will be equal in XXX "XXX -XXX " XXX in XXX XXX

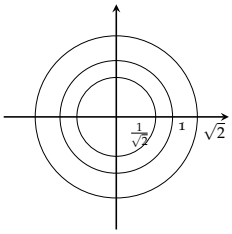


Figure 2.10:

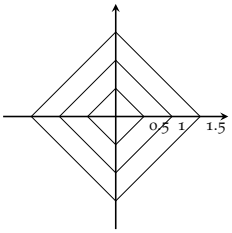


Figure 2.11:

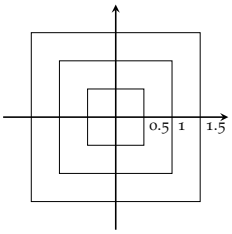


Figure 2.12:

Inner Products

Final concept in geometry is angles which XXX to concept of inner product between elements of a vector space.

The "standard" inner product in $\mathbb{R}_n$ (aka dot/scalar product)

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

More general denote an inner product as $\langle x, y \rangle$

Definition: Any inner 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 any $x, y, z \in \Omega$ and $a \in \mathbb{R}$, the following holds:

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

$\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$

$\langle ax, y \rangle = a \langle x, y \rangle$

$\langle x, y \rangle = \langle y, x \rangle$

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

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

Let's connect to angle now.

Note: 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.

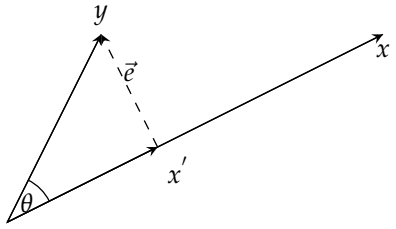


Figure 2.13:

Since we know that $|\cos\theta| < 1$ XXX give

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

Cauchy-Schwartz relates inner product(angle) to norms(length)

Holds for the inner products, not just in $\mathbb{R}_n$

But can also related $|\langle x, y\rangle|$ to the norms (not L_2) via a generalization, "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 $1/p + 1/q = 1$

If $p = q = 2$, get c-s

If $p = 1$, $q = \infty$, 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 XXX 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 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.

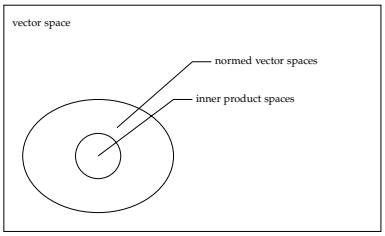


Figure 2.14:

Angles between vectors

Important XXX: since by Cauchy-Schwartz $\frac{|\langle x,y\rangle|}{\|x\|\|y\|} \leq 1$

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

(b) Perhaps more important if $|\cos\theta| = 0$, then $\theta 90^\circ$, and $\frac{|\langle x,y\rangle|}{\|x\|\|y\|} = 0$, or equivalently, $\langle x, y\langle = 0$ asumming $x \neq 0$ and $y \neq 0$

θ is a "right" angle, and x, y are orthogonal vectors.

(c) If $|\theta| < 90^{circ}$,then $\cos\theta > 0$. So $\langle x, y\rangle > 0$ on "acute angle".

wheres if $|\theta| > 90^{circ}$, so $\langle x, y\rangle < 0$ on "abluse angle".

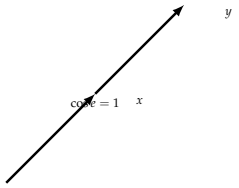


Figure 2.15:

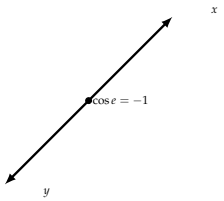


Figure 2.16:

Orthogonality

A set of vectors $S = \{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\}$ 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) & span(S)=m.

If, in addition, all elements have unit norm, i.e., $\|x^i\|_2 = 1$ for all $i \in [m]$ that the set forms an "orthogonal" basis.

Note that $\|\cdot\|_2$ XXX measure length XXX if is induced by the inner product.

When (shortly) get to projection will see orthogonal basis are easy to XXX XXX

Orthogonal complement: Given $S \in \gamma$, a subspace of γ, a vector $x \in \gamma$ is orthogonal to S if $x \perp s \forall s \in S$, i.e., $\perp$ to all vectors in S

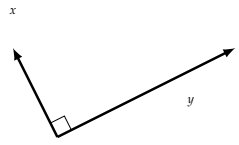


Figure 2.17:

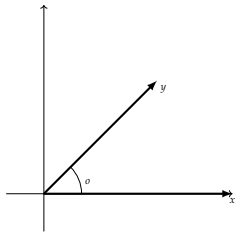
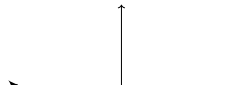


Figure 2.18:



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

figure here

Some results:

(i) $S^{\perp}$ is a subspace: clearly include $0 \in \gamma$ and is closed under linear combination(all linear combination $\perp S$)

(ii) dim(γ)=dim(S) + dim ($S^{\perp}$)

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

Note: If $S = \gamma$ then $S^{\perp} = 0$

Projection

Basic problem: Given a point $x \in \gamma$, find the "closeest" point in the set S (recall that points $\equiv$ vectors)

$$\Pi_s(x) = \arg\min \|y - x\|$$

(1) First for S a subspace of an inner product space, L_2

(2) Second for S, an "affine" set.

Basically a shift of a subspace

Work XXX of geometry

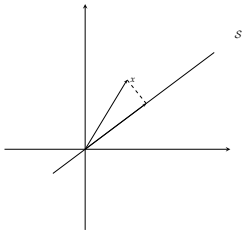


Figure 2.22:

(3) Third will realize can do even for other norms, L_1, L_∞ for which XXX inner product(projection in normed vectors space)

Projection and 1-D subspace

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

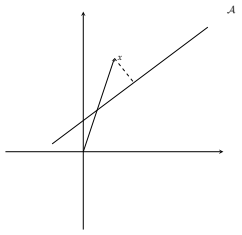


Figure 2.23:

Figure here

Be orthogonal decomposition $x \in S \oplus S^{\perp}$

So $\exists x_s \in S, e \in S^{\perp}$

s.t. $x = x_s + e$, unique expression

$(x - x_s) = e \in S^{\perp}$

Use this decomposition to solve optimization problem

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

Look at the objective function

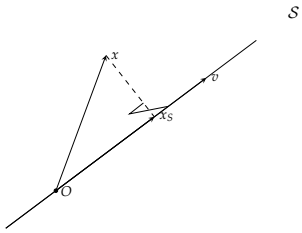


Figure 2.24:

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

where minimum 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: use condition that

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

But $x \in S$, so $\exists a \in \mathbb{R}$ s.t. $x_s = av$, need to solve for a.

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

so $a = \frac{\langle x,v\rangle}{a\langle v,v\rangle} = \frac{\langle x,v\rangle}{\|v\|^2}$

Thus, $x_s = av = \frac{\langle x,v\rangle}{\|v\|^2} v$

Notes: Recall earlier left XXX derivative that concept $\cos\theta$ to inner product

(1) $\cos\theta = \frac{x_s}{\|x\|} = \frac{1}{\|x\|} \frac{|\rangle x,v\rangle|}{\|v\|} \|v\|$

$\cos\theta = \frac{|\langle x,v\rangle|}{\|v\|\|x\|}$

(2) Nice way to remember

$x_s = \langle x, \frac{x}{\|v\|}\rangle \frac{x}{\|v\|}$

Now consider projection onto a subspace in general

figure here

Observe: 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 Ω be an inner product space. Let $x \in \Omega$ and let $S \in \Omega$ be a subspace. There exists a unique vector $x^* \in S$

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

A necessary and sufficient condition for x^* is

1. $x^* \in S$
2. $x - x^* \perp S$

Solving for x^*(general case)

Let $S = span\left(\{x^{(1)}, x^{(2)}, \cdots x^{(n)}\}\right)$

Since $x^* \in S$ can be written as

$x^* = \sum_{i=1}^d a_i x^{(i)}$ for some(as yet unknown) a_i

since $(x - x^*) \perp S$

If $(x - x^*) \perp x^{(k)}$ $\forall k \in [d]$ that will be $\perp$ to all linear combination and hence $\perp$ *toS*

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

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

Re-arranging 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 equations:

when the $x^{(k)}$ are all mutually $\perp$ matrix is diagonal.

If orthogonal and normalized matrix is identity matrix (even easier)

How do you orthogonal and normalize a matrix? Gran-Schmit Procedure

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

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

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

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

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

b. Normalize

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

XXX to higher dimensions as needed

Stacking up results to XXX yields "QR" decomposition

matrix here

To date what have seen in linear algebra class

Next something(perhaps) new: project onto affine set

all subspace go through origin

can't be too difficult to project onto XXX line(subspace) that doesn't include origin

Seems there must be some way to XXX our results XXX to this slightly modified geometry.

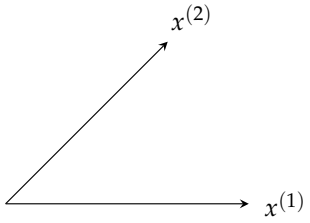


Figure 2.25:

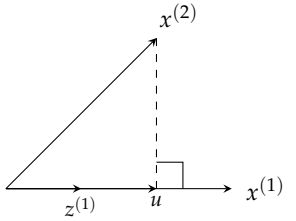


Figure 2.26:

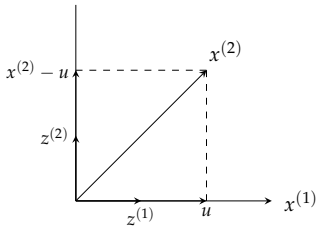


Figure 2.27:

Definition: an "affine" set is a shift/translate of a subspace

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

figure here

figure here

Note: can shift S be any point in A(since any other point in A is XXX point + a vector in S)

Idea of projection onto affine set

⓪ To project $x \in \Omega$ onto A

① First translate both x and A by $-x^{(0)}$

translation of A is S

② Project(translate) $x - x^{(0)}$ onto S(as before)

shift result back by $+x^{(0)}$

③ Get point in A

Theorem: Projection onto affine set

Let Ω be an inner product space

Let $A \in \Omega$ be an affine set where $A = S + x^{(c)}$

There is a unique $x^* \in A$ such that

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

A necessary and sufficient(set of) conditions:

1. $x^* \in A$

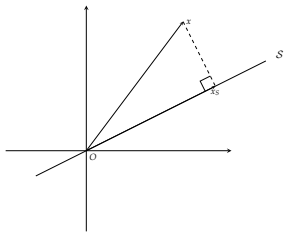


Figure 2.28:

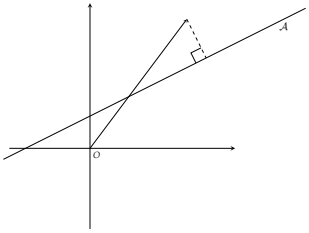


Figure 2.29:

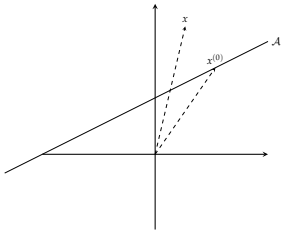


Figure 2.30:

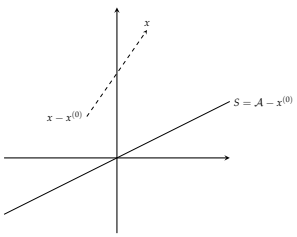


Figure 2.31:

2. $(x - x^*) \perp S$
Before XXX note $(x - x^*) \perp S$ not that
If $(x - x^*) \perp A$, then $(x - x^*) \perp allvectorsinA$
But can see not case
The book writes $(p - p^{(*)}) \perp H$
Proof: any $y \in A$ can be expressed as $y = z + x^{(0)}$ when $z \in S$

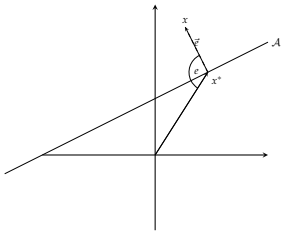


Figure 2.34:

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

Thus $z^* = \arg\min_{z \in S} \|z - (x - x^{(0)})\|$
and translating back,
$x^{(*)} = z^{(*)} + x^{(0)}$
What are the conditons for optimality?
$z^{(*)} - (x - x^{(0)}) \perp S$ and z^* by projection XXX
Thus, in terms of optimal x^*
$x^{(*)} = z^{(*)} + x^{(0)} \in A$
$z^{(*)} + x^{(0)} - x \perp S \equiv x^{(*)} - x \perp S$
Example: Projection onto a hyperplane
example here
Exercise: Prove equivalence of 2 definition
Example: 2-D case
Back to projection onto hyperplane
$H = \{z \in \mathbb{R}_n | a^{\mathrm{T}} z = b\} = \{z \in \mathbb{R}_n | z = x_s + z^{(0)}, x_s \in S, z^{(0)} \in H\}$
Recall $dim(S) = n - 1$, so $dim(S^{\perp}) = 1$
want $p^{(*)} = \arg\min_{p \in H} \|p^* - p\|$
Observe $p - p^{9*} \perp S$ (optimal condition)
So $p - p^* \in S^{(\perp)} = \{\lambda a | \lambda \in \mathbb{R}\}$
Therefore, $\exists \lambda^*$ s.t. $p - p^* = \lambda^* a$
want to solve for λ^* but have 2 unknowns (λ^*, p^*)
will get rid of p^* dependency by using definition of H

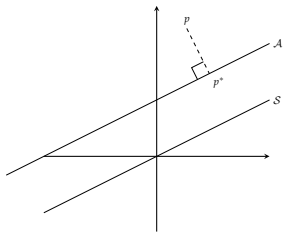


Figure 2.35:

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

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

$p^* = p - (\frac{a^T p - b}{\|a\|^2})a$

and

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

Recall terminology

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

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

Projection w.r.t other norms

So far looked at projection in inner product space.

Recall inner product spaces have a notion of angle, have term "orthogonality principle", L_2 norm is one such example

In contrast, XXX L_1 andL_∞ norms don't come with a sense of angle. However the problem

still makes sens if $p \neq 2$, e.g., $p = 1$, $p = \infty$, but cannot apply $\perp$ principle since 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 norm (change p)

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

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

Recall norm balls

Let's project $x = 0 \in \mathbb{R}^2$ onto a line (affine set/hyperplane)

figure here

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

figure here

Observe:

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

x_1^*: Solution is "sparse", generally will be cost for other constraints since norm-ball XXX + XXX axis+aligned

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

Applications(later in course in XXX x will be control vector over time)

L_2: energy control

L_2^*: sparse solution:

L_∞:equal-effect: useful in "bang-bang" control, XXX XXX XXX XXX, e.g., in rockets

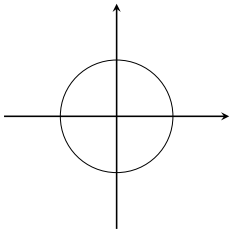


Figure 2.36:

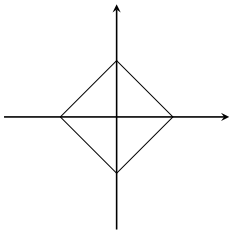


Figure 2.37:

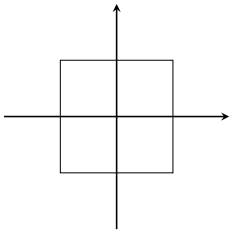


Figure 2.38:

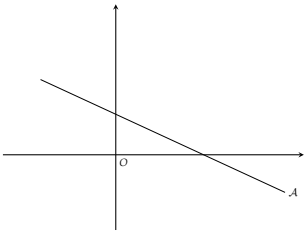


Figure 2.39:

Functions

To date our focus has been on vectors

Now will discuss functions that map vectors inputs to real number. This is importance when come to optimization, e.g., in last example

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

$x \in \mathbb{R}_n$ but the x is chosen to minimize the function $\|\cdot\| : \mathbb{R}_n \mapsto \mathbb{R}$

Some terminology in this book:

"Function": $F : \mathbb{R}_n \mapsto \mathbb{R}$

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

Note: above is like a XXX, XXX programming language

$$F : x \mapsto y$$

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

dom F= $\{x \in \mathbb{R}^n | |F(x)| \leq \infty\}$

example here

figure here

figure here

not the same functions since domains differ.

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

F:$u \mapsto \gamma$, a "map", generally $dim(u) \neq dim(\gamma)$

F:$u \mapsto u$, an " ", input and output vectors spaces have the same dimension

F:$u \mapsto \mathbb{F}$, a "function", map vector space into a scalar

in this course, $u = \mathbb{R}_n$, $\gamma = \mathbb{R}_m$, $\mathbb{F} = \mathbb{R}$(or, occasionally $\mathbb{C}$)

Sets related to functions

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

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

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

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

figure here

figure here

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 are in $\mathbb{R}_n$

Let's sketch these sets for L_2 and L_1 norms in $\mathbb{R}_2$

Graph:

figure here

Epigraph:

figure here

Level sets:

figure here

Sub-level sets:

figure here

Linear and affine functions: important classes

$F : \mathbb{R}_n \mapsto \mathbb{R}$ is linear iff

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

(2)"Additivity": $F(x^{(1)} + x^{92}) = F(x^{(1)}) + F(x^{(2)})$

Put together and recurse to get

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

$F : \mathbb{R}_n \mapsto \mathbb{R}$ is affine iff

$\overline{F}$ define pointwise as $\overline{F} = F(x) - F(0)$, $\forall x \in \mathbb{R}_n$ is a linear function.

Turns out(wasn't prove-see "Linear Algebra Done right")

The $F : \mathbb{R}_n \mapsto \mathbb{R}$ is affine iff there is a unique pair $(a, b) \in \mathbb{R}_n \times \mathbb{R}$ s.t.

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

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 here

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 $\mathbb{R}_2$

figure here

Exercise: Use definition of graph and epigraph to prove:

XXX round a best way to describe hyperplane and half-spaces, direct XXX next:

proof here

Proof: Graph F is a hyperplane when F is affine

proof here

Recalling definition of a hyperplane

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

Half-spaces are on one side or other of hyperplane

$$H_+ = \{z \in \mathbb{R}_n | a^{\mathrm{T}} z > b\}$$

$$H_- = \{z \in \mathbb{R}_n | a^{\mathrm{T}} z \leq b\}$$

Recall that a is $\perp$ to the subspace but defines H, euqivalently the boarder of H_+ or H_-

figure here

Why is H_+(rather than H_-) the side of H XXX which the normal direction is painting

figure here

Note a and $(z - z^{(c)})$ label the vectors, is the displacements, not the end-points. The end-points are labeled as z, $z^{(c)}$, $z^{(c)} + a$

Let's consider the inner product

$$\begin{aligned} \langle z - z^c, (z^c + a) - a \rangle &= \langle z - z^c, a \rangle \\ &= z^{\mathrm{T}} a - (z^{(c)})^{\mathrm{T}} a \\ &= z^{\mathrm{T}} a - b \end{aligned}$$

To date

(1)Geometry

(2)Functions

In optimization you have some parametric vectors $x \in \mathbb{R}_n$. You want to search all allowable x by moving around $\mathbb{R}_n$(geometry!) to minimize some cost function $F : \mathbb{R}_n \mapsto \mathbb{R}$. Each parameter vector $x \in \mathbb{R}_n$ will be associated with a cost $F(x)$.

How might you solve such problem?

Classic example in 2-D: Finding your way off a mountain and XXX of a faster to a town.

Anyone know the classic method? Follow a stream downhill.

Why is that a good method?

Eventually, would get to ocean, towns built by streams

In mountains no towns upstream.

But equally important, don't walk in circles since water doesn't flow uphill

To go downstream XXX easy to figure out when to get to stream - see which XXX water is flowing - Follow the negative gradient

In fact do some thing in $\mathbb{R}_n$ not just $\mathbb{R}_2$ "gradient descent".

So, gradient is importance, let's remind ourselves what it is.

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 consider compound function, need chain rule for gradients

Say,

$g : \mathbb{R}^n \mapsto \mathbb{R}_m$

$F : \mathbb{R}^m \mapsto \mathbb{R}$

Both F and g are differentiable and we want $\nabla\Phi(x)$, where $\Phi(x) = F(g(x))$.

Before give formula, note that

1. $\nabla\Phi(x) \in \mathbb{R}_n$ XXX they are n "lengths" XXX XXX in x.
2. Helps to consider an intermediate value z
3. First consider the k element of $\nabla\Phi$, i.e., $[\nabla\Phi(x)]_k$

$$\begin{aligned} [\nabla\Phi(x)]_k &= \frac{\partial\Phi(x)}{\partial x_k} \\ &= [\frac{\partial g_1(x)}{\partial x_k}\frac{\partial g_2(x)}{\partial x_k}\cdots\frac{\partial g_m(x)}{\partial x_k}]\nabla F(z)|_{z=g(x)} \\ &= \sum_{i=1}^{m}\frac{\partial g_i(x)}{\partial x_k}\frac{\partial F(z)}{\partial z_i}|_{z=g(x)} \end{aligned}$$

Stacking up

matrix here

Example: $\Phi(x) = F(g(x))$, where $g : \mathbb{R}_n \mapsto \mathbb{R}_m$ is affine.

In particular, $g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{bmatrix}$

and $g_i(x) = a_i^{\mathrm{T}}x + b_i, i \in [m]$, $a_i \in \mathbb{R}_n$, $b_i \in \mathbb{R}$

$$\begin{aligned} \frac{\partial g_i(x)}{\partial x_k} = \frac{\partial}{\partial x_k}(a_i^{\mathrm{T}}x + b) &= \frac{\partial}{\partial x_k}(\sum_{j=1}^{m} a_{ij}x_j) \\ &= a_{ik} \end{aligned}$$

Thus, $[\nabla\Phi(x)]_k = [a_{1k}a_{2k}\cdots a_{mk}]\nabla F(z)|_{z=g(x)}$

Stacking up we have

matrix here

Affine approximation

Consider the Taylor series for $F : \mathbb{R}_n \mapsto \mathbb{R}$

$$F(x) = F(x_0) + \nabla F(x_0)^{\mathrm{T}}(x - x_0) + \epsilon(x)$$

Example here

3
Matrices and eigen decomposition

s

4
Symmetric matrices and spectral decomposition

5
Singlar value decomposition

6
Linear equations and least squares

7
Linear, quadratic, and geometric models

8
Convexity

9
First and second order methods

Bibliography