

STARK DRAPER

# COURSES NOTES: OPTIMIZATION THEORY AND ALGORITHMS

COURSE NOTES: VERSION 1.01

Copyright © 2019 Stark Draper

*September 2019*

## *Contents*

1	<i>Introduction</i>	5
2	<i>Vectors and functions</i>	9



# 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  $i$ th 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

## 6 COURSES NOTES: OPTIMIZATION THEORY AND ALGORITHMS

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

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

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  $(x^{(i)})^T = [x_1^{(i)} \ x_2^{(i)} \ \dots \ x_n^{(i)}]$ .

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  $i$ th row and  $j$ th column of  $A$  is denoted  $[A]_{ij}$  (alternately  $a_{ij}$ ). The transpose of  $A$ ,  $A^T$  is the matrix the element in the  $i$ th row and  $j$ th 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  $f : \mathcal{X} \rightarrow \mathcal{Y}$  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  $\text{dom}f$ . It is often convenient to define  $f(x) = \infty$  for all  $x \notin \text{dom}f$ . In that case  $\text{dom}f = \{x \in \mathcal{X} \mid |f(x)| < \infty\}$ . In this course we mostly consider functions of the form  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ . 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  $f : \mathbb{R} \rightarrow \mathbb{R}$  where  $\text{dom}f = \mathbb{R}_+$  is plotted in Fig. 1.1.

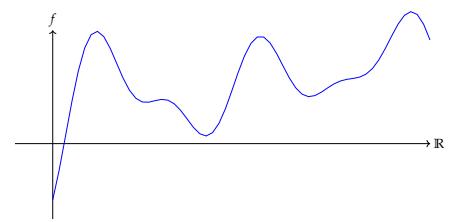


Figure 1.1: A function  $f : \mathbb{R} \rightarrow \mathbb{R}$ .



## 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.

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

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  $\mathbf{x}$  yields a row vector,

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

and occasionally write as a list  $(x_1, x_2, \dots, x_n)$ . 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  $x^{(i)} \in \mathbb{R}^n, i \in \{1, 2, \dots, m\}, i \in [m] = \{1, 2, \dots, m\}$  and  $(x^{(i)})^T = [x_1^{(i)} \ x_2^{(i)} \ \dots \ x_n^{(i)}]$ .

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

**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.

$$au = u$$

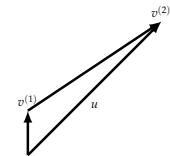


Figure 2.1: Addition

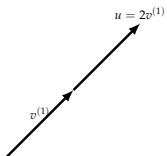


Figure 2.2: Scaling

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  $\infty$ -length complex sequences

③ Complex functions defined on real line

④ Polynomials of degree at most  $n-1$

$$P_{n-1} = \{P \mid p(t) = a_{n-1}t^{n-1} + a_{n-2}t^{n-2} + \cdots + a_1t + a_0\}$$

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.,  $S = \{v^{(1)}, v^{(2)}, \dots, v^{(m)}\}$ , where each  $v^{(i)} \in \mathbb{R}^n$ . Then, the span of  $S$ , denoted by  $\text{span}(S)$ , is the set consisting of all the vectors that are linear combinations of  $\{v^{(1)}, v^{(2)}, \dots, v^{(m)}\}$ , that is,

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

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

#### Example 1

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

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

#### 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

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

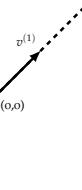


Figure 2.3: Example 1

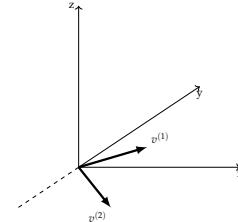


Figure 2.4: Example 2

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  $S = \{v^{(1)}, \dots, v^{(n)}\}$  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

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

### 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  $S$ .

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

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  $\text{span}(S)$ .

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

### 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)

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

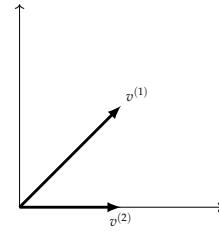


Figure 2.5:

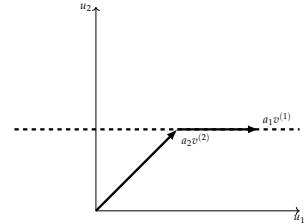


Figure 2.6:

However, if we redefine  $v^3$ , says

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

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

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

We can prove that each a basis for  $\text{span}(S)$  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:

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

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)  $\|v\| \geq 0, \forall v \in \mathcal{V}$ , 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  $v$  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:

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

$L_2$  norm: Euclidean length

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

$L_1$  norm:

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

$L_\infty$  norm:

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

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

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

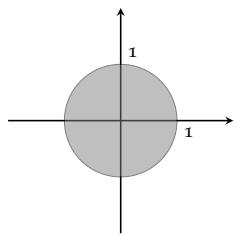
Sometimes it is called " $L_0$ " norm  $\|x\|_0$ , since  $\text{card}(x) = \lim_{p \rightarrow 0} (\sum_{k=1}^n |x_k|^p)^p$ , but it is not a norm (so this terminology is inaccurate). For instance, it doesn't satisfy property (c) of a norm:

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

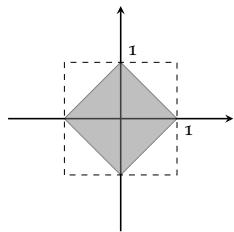
### Unit norm-ball

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

$L_2$  norm ball

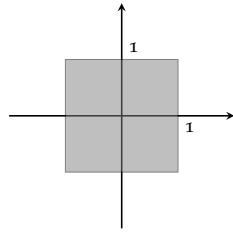


$L_1$  norm ball :  $\{x \mid |x_1| \leq 1\}$

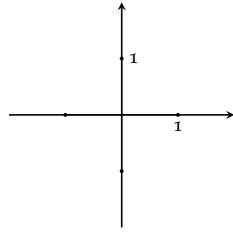


- (a) First see inside the box, clearly  $|x_1| \leq 1$  and  $|x_2| \leq 1$
- (b) Look at the position we want,  $x_1 + x_2 \leq 1$ , i.e.,  $x_2 \leq 1 - x_1$
- (c) Rest by symmetry

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



What about  $\text{card}(x)$ ?



The set  $\{x \mid \text{card}(x) \leq 1\}$  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  $\{x \mid \|x\| = c\}$ , and let's see for  $c = \frac{1}{2}, 1, 2$ . See for the figures on the r.h.s.

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  $\Omega$  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

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

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:

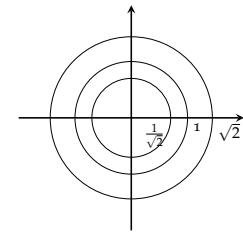


Figure 2.7:  $L_1$  level set

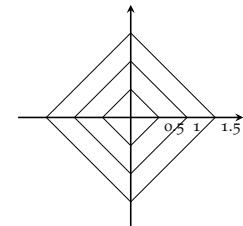


Figure 2.8:  $L_2$  level set

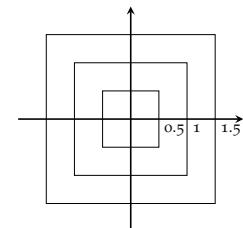
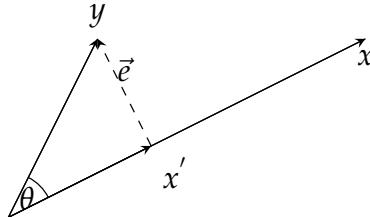


Figure 2.9:  $L_\infty$  level set

- (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  $\text{dimspan}(x, y) = 2$  (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

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

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":

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

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

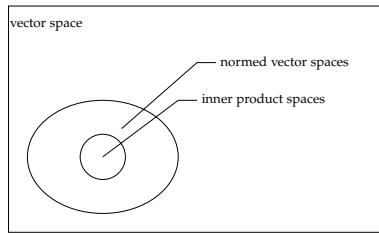
- (1) If  $p = q = 2$ , we get the Cauchy-Schwartz inequality.
- (2) If  $p = 1, q = \infty$ , we get  $|x^T y| \leq \|x\|_1 \|y\|_\infty = (\sum_{k=1}^n |x_k|)(\max_{k \in [n]} |x_k|)$ .

A second important connection of inner product and norm is that

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

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  $L_\infty$ . 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  $\theta$ :

(a) If  $|\cos \theta| = +1$ , then  $\theta = 0^\circ$  or  $180^\circ$ . 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,  $\langle x, y \rangle = 0$  (assuming  $x \neq 0$  and  $y \neq 0$ ). In this case,  $\theta$  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  $\theta$  is a "acute angle", whereas if  $|\theta| > 90^\circ$ , then  $\cos \theta < 0$  and  $\langle x, y \rangle < 0$ ,  $\theta$  is a "obtuse angle".

### Orthogonality

A set of vectors  $S = \{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$  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  $\text{span}(S)$  and hence  $\dim(\text{span}(S))=m$ .

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.

**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

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

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)  $\dim(\nu) = \dim(S) + \dim(S^\perp)$ .

(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  $S^\perp = 0$ .

### Projection

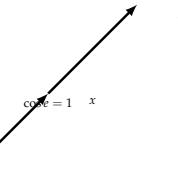


Figure 2.10: (a)  $\cos \theta = +1$

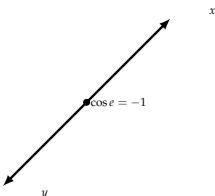


Figure 2.11: (a)  $\cos \theta = -1$

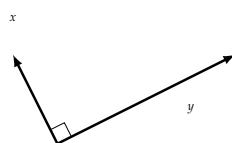


Figure 2.12: (b)  $\cos \theta = 0$

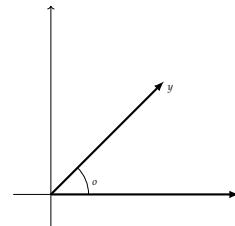


Figure 2.13: (c)  $\cos \theta > 0$

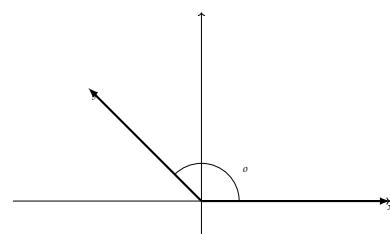


Figure 2.14: (c)  $\cos \theta < 0$

Motivation: Given a point  $x \in \nu$ , 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:

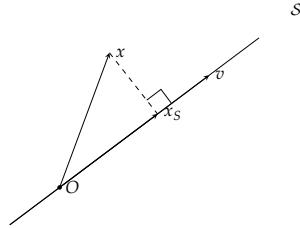
- (1)  $S$  is a subspace of an inner product space associate with  $L_2$  norm
- (2)  $S$  is an "affine" set(a shifted subspace)
- (3) Consider other norms, e.g.,  $L_1$ ,  $L_\infty$  for which no inner product(projection in normed vectors space)

### Projection onto 1-D subspace

Let's consider the one dimension subspace given by

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

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)

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

The objective function can be written as

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

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,

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

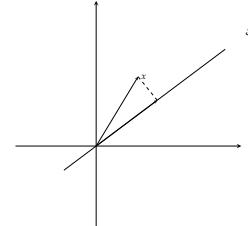


Figure 2.15:  $S$  is a subspace of an inner product space

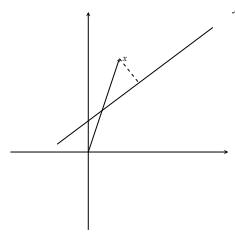


Figure 2.16:  $S$  is an "affine" set

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  $a = \frac{\langle x, v \rangle}{\langle v, v \rangle} = \frac{\langle x, v \rangle}{\|v\|^2}$ . 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  $\Omega$  and  $\Omega$  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  $S = \text{span}(\{x^{(1)}, x^{(2)}, \dots, x^{(d)}\})$ . 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)

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

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?

### Gram-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  $z^{(1)}, z^{(2)}$ .

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

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

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

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

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

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

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

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

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

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

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

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

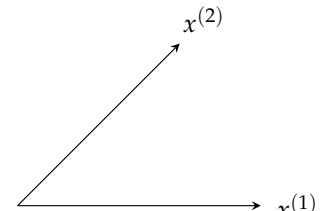


Figure 2.17: Problem setting

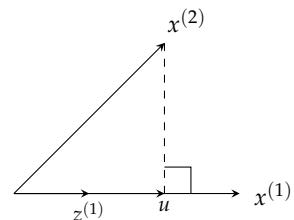


Figure 2.18: Step 1

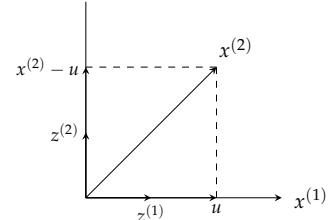


Figure 2.19: Step 2

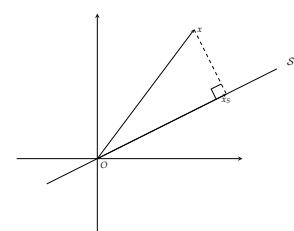


Figure 2.20: Subspace  $S$

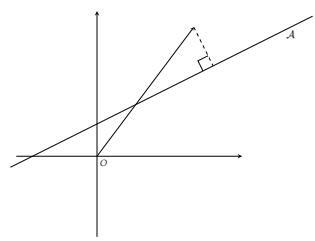


Figure 2.21: Affine set  $\mathcal{A}$  as a shifted subspace

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

The idea of finding projection onto affine set:

Step (o) 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  $\mathcal{A}$ .

### Theorem: Projection onto affine set

Let  $\mathcal{A} \in \Omega$  be an affine set, where  $\Omega$  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  $y \in \mathcal{A}$  can be expressed as  $y = z + x^{(0)}$  when  $z \in S$

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

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)  $x^{(*)} = z^{(*)} + x^{(0)} \in \mathcal{A}$ .
- (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  $(a, b) \in \mathbb{R}^n \times \mathbb{R}$ .

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

An equivalent definition is

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

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

Note:

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

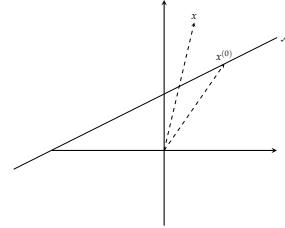


Figure 2.22: Step (o)

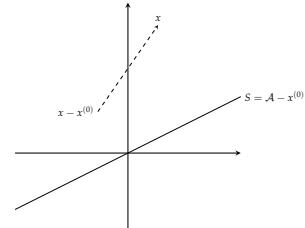


Figure 2.23: Step (1)

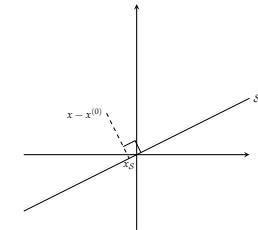


Figure 2.24: Step (2)

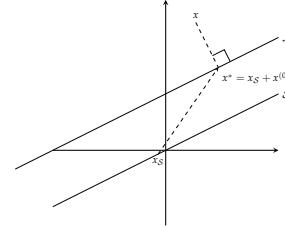
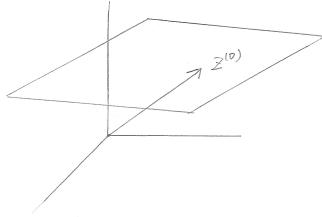


Figure 2.25: Step (3)



(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  $\mathcal{S}$  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,

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

Now, since  $(\text{span}\{a\}) = \{x | x = \lambda a, \lambda \in \mathbb{R}\}$  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  $z^{(0)}$ .

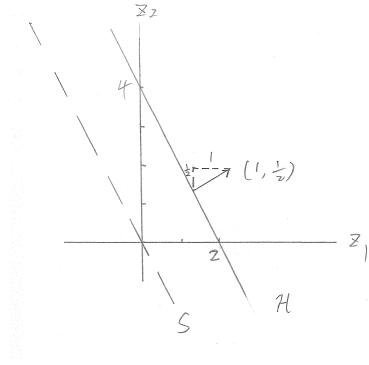
Now, let's start with  $\mathcal{H} = \{z \in \mathbb{R}^n | z = u + z^{(0)}, u \in \mathcal{S}, z \in \mathcal{H}\}$ .

Let  $\{u^{(1)}, \dots, u^{(n-1)}\}$  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.,

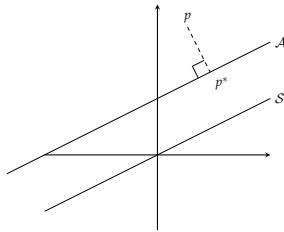
$$\mathcal{H} = \{z \in \mathbb{R}^n | z = u + z^{(0)}, u \in S, z^{(0)} = \begin{bmatrix} 0 \\ 4 \end{bmatrix}\}$$

or

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

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

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



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  $\lambda^*$  but notice that there are 2 unknowns  $(\lambda^*, p^*)$ , and hence we get rid of  $p^*$  dependency by using definition

of  $\mathcal{H}$ .

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

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

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

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

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

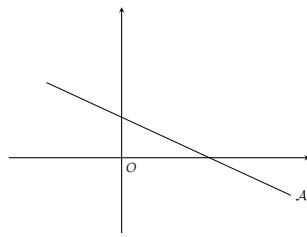
### Projection w.r.t other norms

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  $L_1$  and  $L_\infty$  norms don't come with a sense of angle. However the problem still make sense 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  $L_\infty$ .
- (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).



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

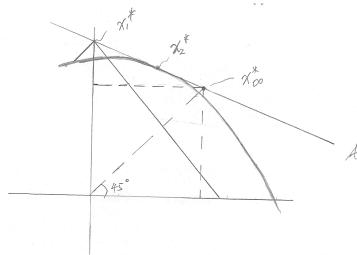


Figure 2.26:

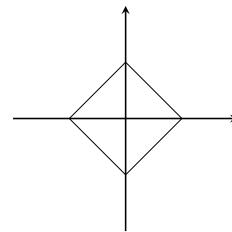


Figure 2.27:

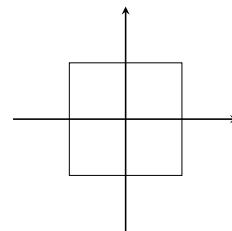


Figure 2.28:

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.

$x_\infty^*$ : 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  $\Omega$  (cf,  $\mathbb{R}^n$ ), this is the "domain" of  $F$ .

Aside: Terminology when discussing a pair of vector spaces  $(\mathcal{V}, \mathcal{U})$  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 vector 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

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

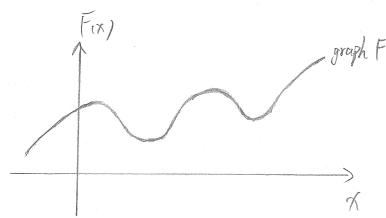
- (1) The "graph" (a.k.a, "plot") of  $F$  is the set

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

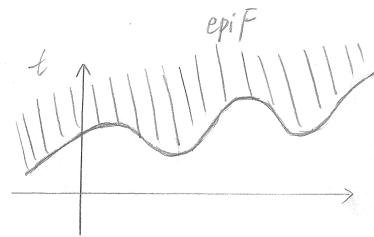
- (2) The "epigraph" of  $F$  is the set

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

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.

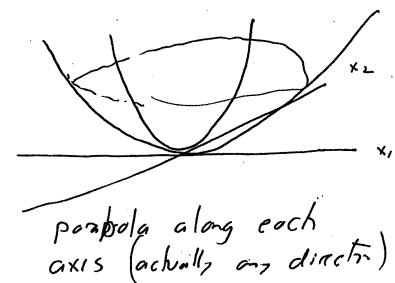


Figure 2.29: Graph 1

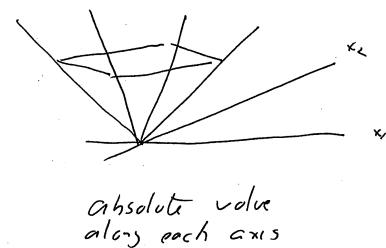


Figure 2.30: Graph 2

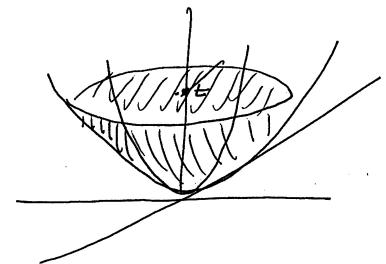


Figure 2.31: Epigraph 1

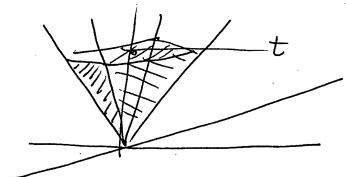


Figure 2.32: Epigraph 2

### 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

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

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

Let  $\bar{F}$  define pointwise as  $\bar{F} = F(x) - F(0), \forall x \in \mathbb{R}^n$  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

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

Since  $F(0) = b$ , this implies that any linear function can be expressed as  $F(x) = a^T x = \langle a, x \rangle$  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.

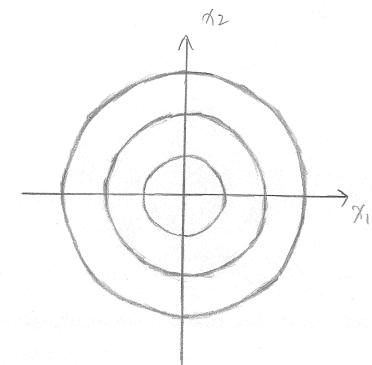
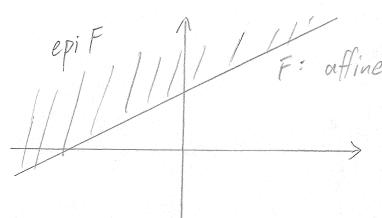
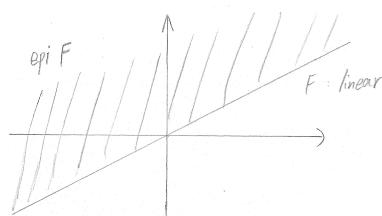


Figure 2.33: Level set 1

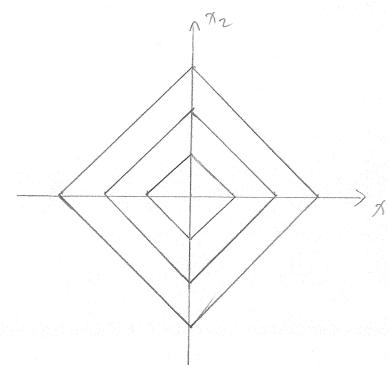


Figure 2.34: Level set 2

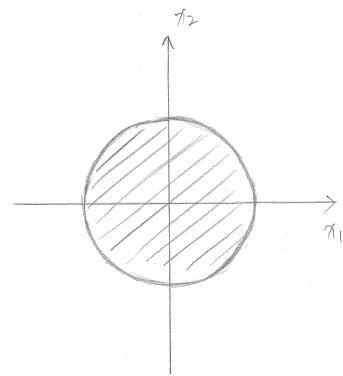


Figure 2.35: Sub-level set 1

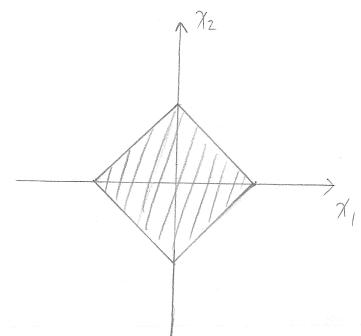
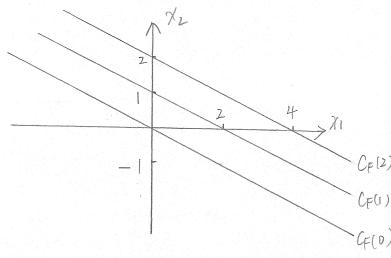


Figure 2.36: Sub-level set 2

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  $\mathbb{R}^2$



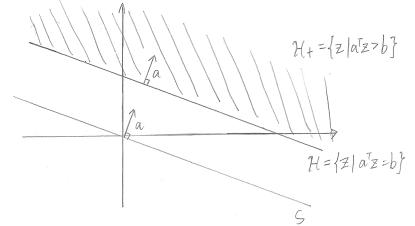
Definition of a hyperplane:

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

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

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

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



### Gradient

The gradient  $\nabla F$  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

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

### 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

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

### Affine approximations

Consider the Taylor series for  $F: \mathbb{R}^n \rightarrow \mathbb{R}$ .

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

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

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

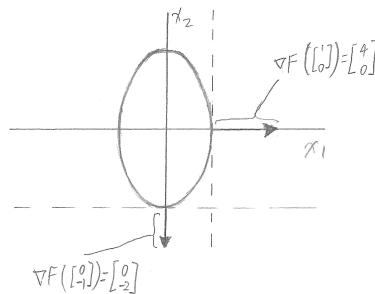
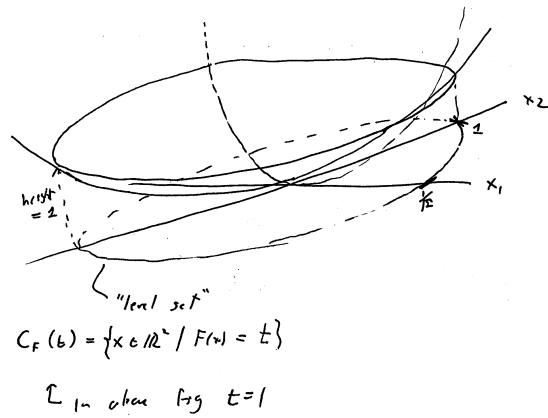
$$\nabla F(x) = \begin{bmatrix} 4x_1 \\ 2x_2 \end{bmatrix}$$

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

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

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

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



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

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

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

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

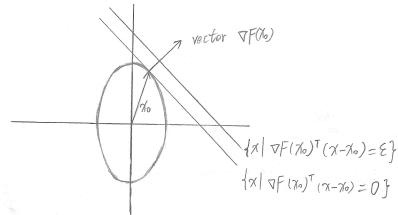
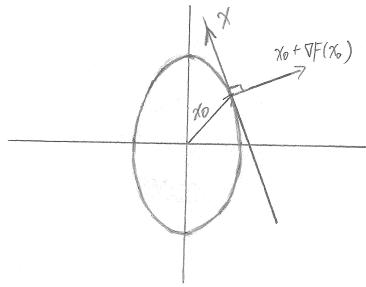
(1) Consider case  $c = 0$

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  $x = x_0$ .

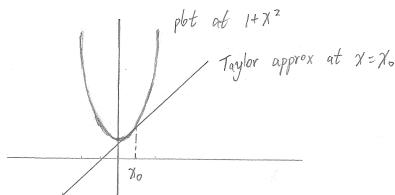


In general the set

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

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.

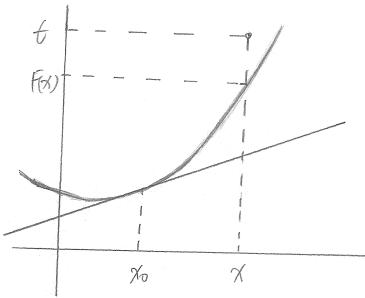


To develop the approximation we need to consider the plot or "graph" at the function  $F$

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

e.g., in above example  $F(x) = x^2 + 1$ ,  $F: \mathbb{R} \rightarrow \mathbb{R}$ ,  $s - n = 1$  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

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

Next, we re-arrange

(6)

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

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.

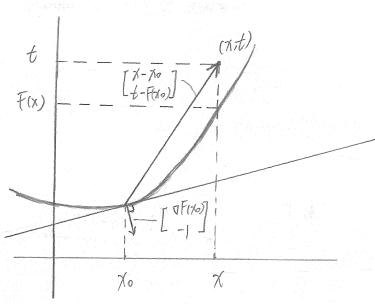
(1). If inner product of 2 vectors is negative, then the angle is obtuse.

(2). Matches picture.

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

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

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



The condition becomes

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

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

$$\mathcal{H} = \left\{ \begin{bmatrix} x \\ t \end{bmatrix} \middle| \begin{bmatrix} x^T & t \end{bmatrix} \begin{bmatrix} \nabla F(x_0) \\ -1 \end{bmatrix} = 0 \right\} = \left\{ \begin{bmatrix} x \\ t \end{bmatrix} \middle| \begin{bmatrix} x_0^T & F(x_0) \end{bmatrix} \begin{bmatrix} \nabla F(x_0) \\ -1 \end{bmatrix} = 0 \right\}$$

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:

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

