STARK DRAPER

COURSE NOTES: OPTIMIZATION THEORY AND ALGORITHMS

COURSE NOTES: VERSION 1.11

December 2019

Contents

1 Introduction 5

2 Vectors and functions 9

3 Matrices and eigen decomposition 35

4 Symmetric matrices and spectral decomposition 45

5 Singular value decomposition 53

6 Linear equations and least squares 59

7 Linear, quadratic, and quadratically-constrained quadratic programs 71

8 Convex sets and functions 97

9 Convex optimization 117

10 Duality 131

Bibliography 153

Remarks, feedback, and versions

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

Version 1.01: Initial distribution of chapters 1 and 2.
Version 1.02: Initial distribution of chapters 3 and 4.
Version 1.03: Initial distribution of chapter 5.
Version 1.04: Initial distribution of chapter 6.
Version 1.05: Initial distribution of chapter 7, including linear programs.
Version 1.06: Distribution of remainder of chapter 7, including quadratic programs and quadratically-constrained quadratic programs.
Version 1.07: Initial distribution of rough notes for chapter 8; to be revised.
Version 1.08: Revision of chapter 8; initial bit of chapter 9 included.
Version 1.09: Revision of chapter 9; initial bit of chapter 10 included.
Version 1.10: Revision of chapter 10.
Version 1.11: Revision of chapter 10, added material on KKT conditions.

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

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

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

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

where $x^{(i)} \in \mathbb{R}^n$, $i \in \{1, 2, \ldots, m\}$ and $(x^{(i)})^{\mathrm{T}} = [x_1^{(i)} \; x_2^{(i)} \; \ldots \; x_n^{(i)}]$. The set $\{1, 2, \ldots, m\}$ 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 $\{x_1, x_2, \ldots x_m\}$ 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 $\mathcal{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$. 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 $\mathbb{Z}$. Occasionally we have need to refer to the sets of non-negative and positive real numbers, respectively denoted $\mathbb{R}_+$ and $\mathbb{R}_{++}$.

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

We write $f : \mathcal{X} \to \mathcal{Y}$ to denote a function f that maps elements of $\mathcal{X}$ to elements of $\mathcal{Y}$. This notation is akin to strongly-typed programming languages. The function f needs an input in $\mathcal{X}$ to be able to process it. Elements not in $\mathcal{X}$ are not acceptable as inputs. That said, not every element of $\mathcal{X}$ may be acceptable to f. (E.g., if f calculates the average age of students in a class, no age inputted into the function should be negative.) The acceptable subset of $\mathcal{X}$ is the domain of f, denoted dom$f$. It is often convenient to define $f(x) = \infty$ for all $x \notin$ dom$f$. In that case dom$f = \{x \in \mathcal{X} | |f(x)| < \infty\}$. 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 dom$f = \mathbb{R}_+$ is plotted in Fig. 1.1.

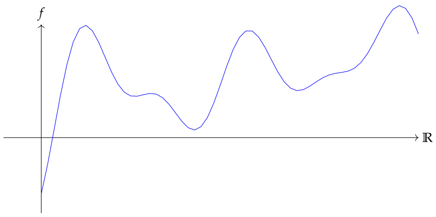


Figure 1.1: A function $f : \mathbb{R} \rightarrow \mathbb{R}$.

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.

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

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

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

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

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,

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

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

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

Vector Space

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

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

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

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

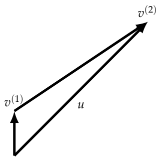


Figure 2.1: Addition

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

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

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

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

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

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

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

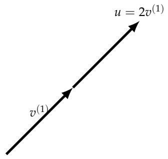


Figure 2.2: Scaling

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

① Finite-length complex vector

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

② ∞-length complex sequences

③ Complex functions defined on real line

④ Polynomials of degree at most n-1

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

⑤ Sets of matrices(will discuss later)

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

Span and subspace

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

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

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

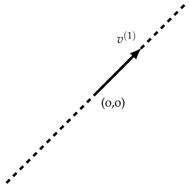


Figure 2.3: Example 1

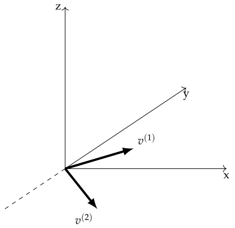


Figure 2.4: Example 2

Example 1

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

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

Example 2

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

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

Note:

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

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

Linear independent set

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

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

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

Importance of linearly independent

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

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

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

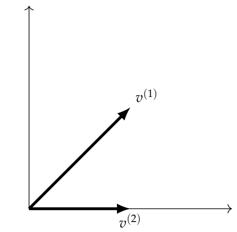


Figure 2.5:

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:

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

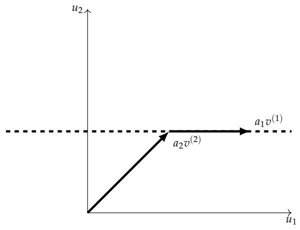


Figure 2.6:

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

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

Example 1

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

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

However, if we redefine v^3, says

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

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

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

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

Example 2

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

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

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

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

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

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

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

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

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

Following is a family of norms that are frequently used:

L_p norm:

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

L_2 norm: Euclidean length

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

L_1 norm:

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

L_∞ norm:

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

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

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

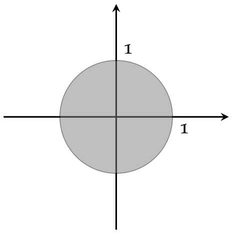
Sometimes it is called "L_0" norm $\|x\|_0$, since $\operatorname{card}(x) = \lim_{p \to 0} (\sum_{k=1}^{n} |x_k|^p)^p$, but it is not a norm (so this terminology is inaccurate). For instance, it doesn't satisfy property (c) of a norm:

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

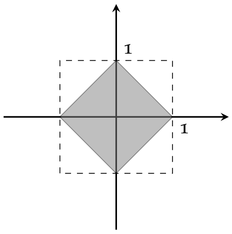
Unit norm-ball

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

L_2 norm ball

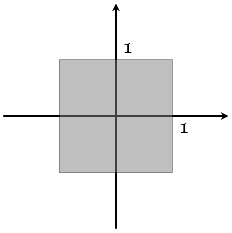


L_1 norm ball : $\{x||x_1| \leq 1\}$

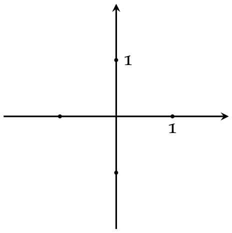


(a) First see inside the box, clearly $|x_1| \leq 1$ and $|x_2| \leq 1$
(b) Look at the position we want, $x_1 + x_2 \leq 1$, i.e., $x_2 \leq 1 - x_1$
(c) Rest by symmetry

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



What about card(x)?



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

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

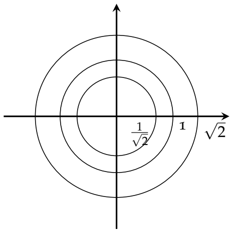


Figure 2.7: L_1 level set

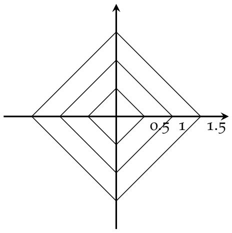


Figure 2.8: L_2 level set

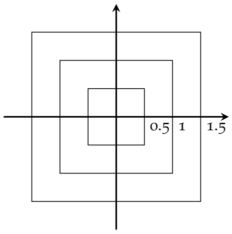


Figure 2.9: L_∞ level set

Why might we be interested in different norms?

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

Inner Products

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

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

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

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

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

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

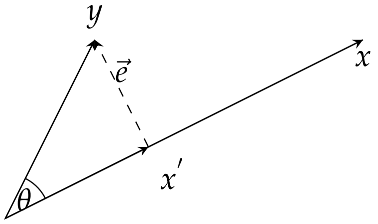
(4) $\langle x, y \rangle = \langle y, x \rangle$

Note:

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

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

Let's connect to angle now.



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

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

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

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

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

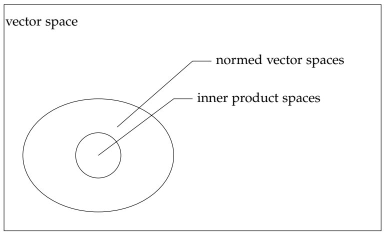
for any $p, q \geq 1$ such that $1/p + 1/q = 1$.
(1) If $p = q = 2$, we get the Cauchy-Schwartz inequality.
(2) If $p = 1$, $q = \infty$, we get $|x^{\mathrm{T}}y| \leq \|x\|_1\|y\|_\infty = (\sum_{k=1}^{n} |x_k|)(\max_{k\in[n]} x_k)$.

A second important connection of inner product and norm is that

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

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

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



Angles between vectors

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

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

(b) If $|\cos\theta| = 0$, then $\theta = 90°$, 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, θ is a "right" angle, and x, y are orthogonal vectors.

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

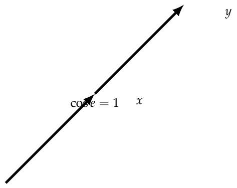


Figure 2.10: (a) $\cos\theta = +1$

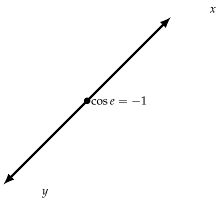


Figure 2.11: (a) $\cos\theta = -1$

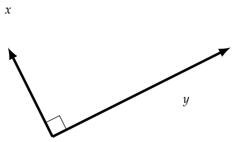


Figure 2.12: (b) $\cos\theta = 0$

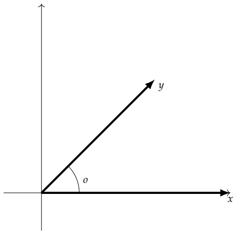


Figure 2.13: (c) $\cos\theta > 0$

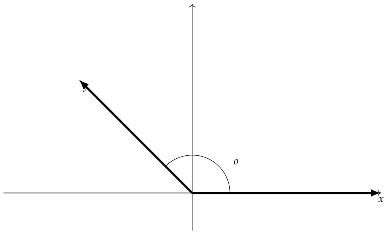


Figure 2.14: (c) $\cos\theta < 0$

Orthogonality

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

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

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

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

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(ν)=dim(S) + dim ($S^{\perp}$).

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

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

Projection

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_∞ for which no inner product(projection in normed vectors space)

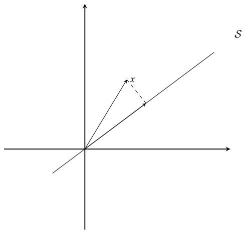


Figure 2.15: S is a subspace of an inner product space

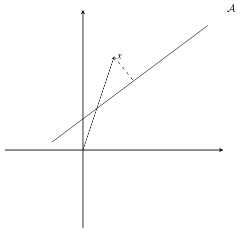


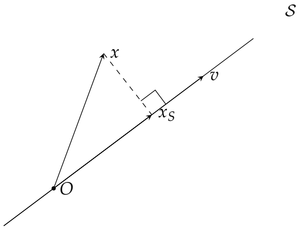
Figure 2.16: S is an "affine" set

Projection onto 1-D subspace

Let's consider the one dimension subspace given by

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

The vector x and subspace S are something like:



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

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

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

The objective function can be written as

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

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

To summarize,

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

where x_s is in $\perp$-decomposition.

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

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

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

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

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

Projection onto a general subspace

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

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

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

A necessary and sufficient condition for x^* is

(1). $x^* \in S$

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

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

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

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

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

Rearranging yields

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

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

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

One case where easy to solve the equation is, when the $x^{(k)}$ are all mutually $\perp$ (so the matrix is diagonal), or furthermore, all these vectors have unit length and mutually orthogonal(so it is an identity matrix).

How do you orthogonalize and normalize a matrix?

Gram-Schmidt 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 $x^{(1)}$

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

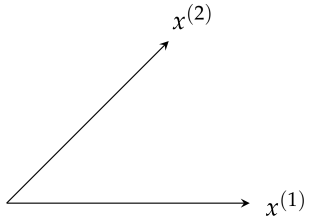


Figure 2.17: Problem setting

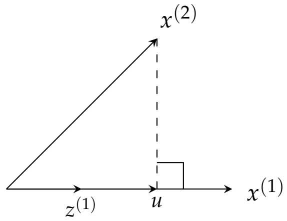


Figure 2.18: Step 1

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 $z^{(2)}$

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

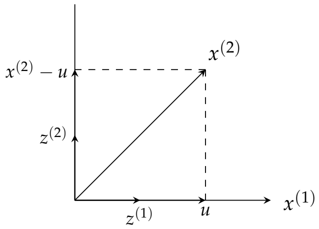


Figure 2.19: Step 2

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

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

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

$$= QR$$

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

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

Project onto affine set

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

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

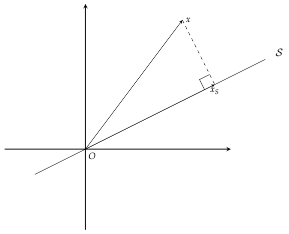


Figure 2.20: Subspace S

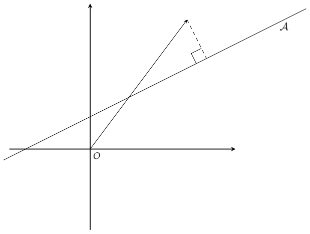


Figure 2.21: Affine set $\mathcal{A}$ as a shifted subspace

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

The idea of finding projection onto affine set:
Step (0) Goal: to project $x \in \Omega$ onto $\mathcal{A}$.
Step (1) Translate both x and $\mathcal{A}$ by $-x^{(0)}$, and note that the translation of $\mathcal{A}$ is S.
Step (2) Project(translate) $x - x^{(0)}$ onto S(as we did before), and shift result back by $+x^{(0)}$.
Step (3) Get the projection point in $\mathcal{A}$.

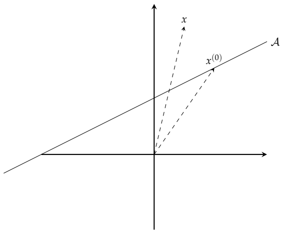


Figure 2.22: Step (0)

Theorem: Projection onto affine set

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

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

A necessary and sufficient(set of) conditions:
(1). $x^* \in \mathcal{A}$
(2). $(x - x^*) \perp S$

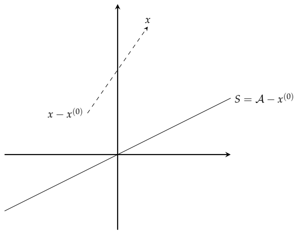


Figure 2.23: Step (1)

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

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

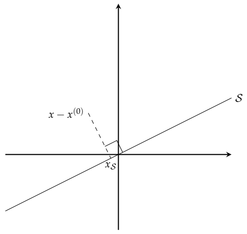


Figure 2.24: Step (2)

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

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

What are the conditions for optimality? That is,
$z^{(*)} - (x - x^{(0)}) \perp S$, where $z^* \in S$ is obtained by projection onto S.
Thus, in terms of optimal x^*,
(1) $x^{(*)} = z^{(*)} + x^{(0)} \in A$.
(2) $(z^{(*)} + x^{(0)} - x) \perp S \Leftrightarrow (x^{(*)} - x) \perp S$.

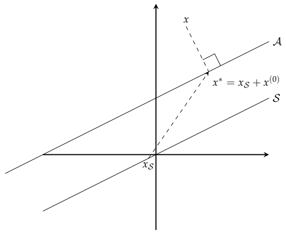


Figure 2.25: Step (3)

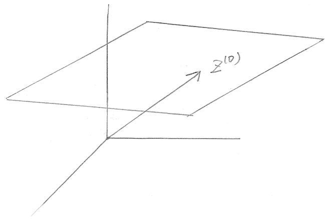
Example: Projection onto a hyperplane
A **hyperplane** is an affine set $\mathcal{H}$ specified by the pair $(a, b) \in \mathbb{R}^n \times \mathbb{R}$.

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

An equivalent definition is

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

where S is the subspace $S = \{z \in \mathbb{R}^n | a^T z = 0\}$.
Note:
(1) The "equivalent" definition is clearly that of an affine set.



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

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

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

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

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

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

Exercise: Prove equivalence of above 2 definition

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

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

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

Thus,

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

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

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

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

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

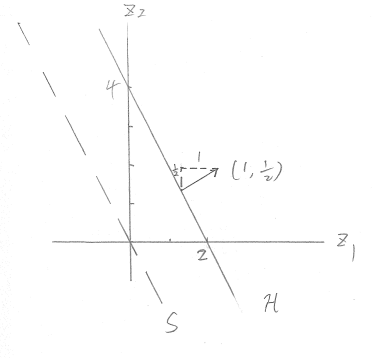
Now, let's start with $\mathcal{H} = \{z \in \mathbb{R}^n | z = u + z^{(0)}, u \in \mathcal{S}, z \in \mathcal{H}\}$.

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

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

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

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



when

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

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

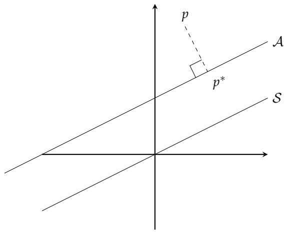
$$\mathcal{H} = \{z \in \mathbb{R}^n | z = u + z^{(0)}, u \in S, z^{(0)} = \begin{bmatrix} 0 \\ 4 \end{bmatrix}\}$$

or

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

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

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



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

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

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

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

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

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

.

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

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

Projection w.r.t other norms

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

In following we will

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

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

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

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

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

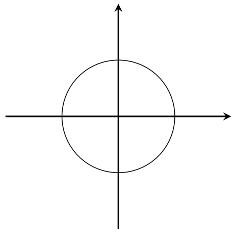


Figure 2.26:

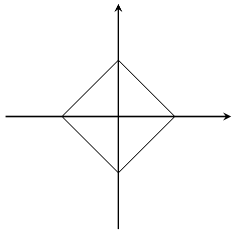


Figure 2.27:

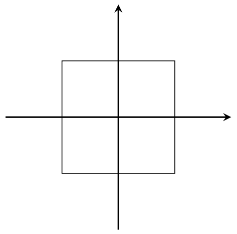
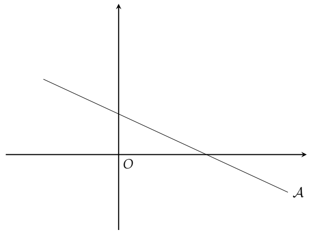
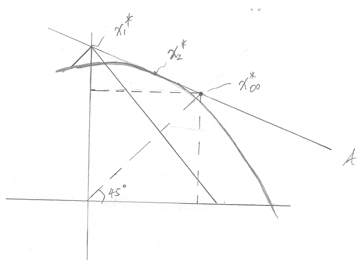


Figure 2.28:



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



Observe that:

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

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

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

Functions

Some terminologies will be used in this material:

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

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

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

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

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

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

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

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

Sets related to functions

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

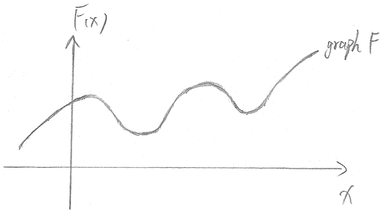
(1) The "graph" (a.k.a, "plot") of F is the set

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

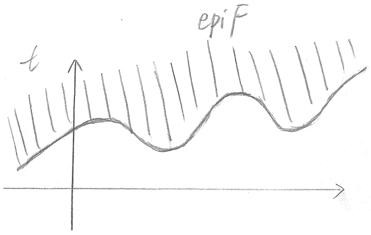
(2) The "epigraph" of F is the set

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

Graph



Epigraph



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

(3) The "level" set

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

(4) The "sub-level" set

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

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

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

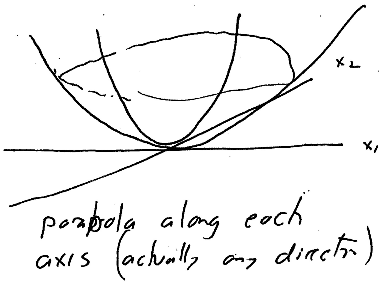


Figure 2.29: Graph 1

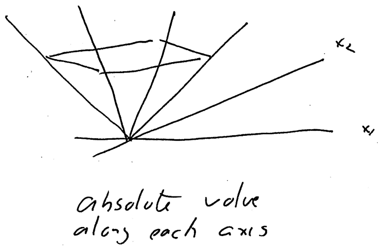


Figure 2.30: Graph 2

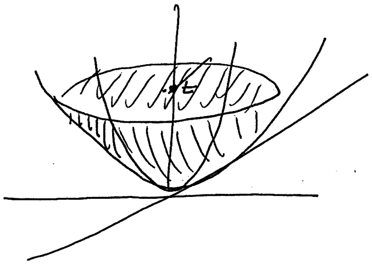


Figure 2.31: Epigraph 1

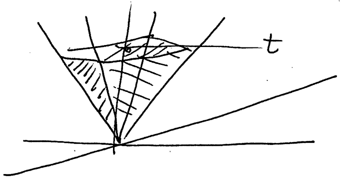


Figure 2.32: Epigraph 2

Linear and affine functions

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

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

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

Put together to get

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

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

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

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

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

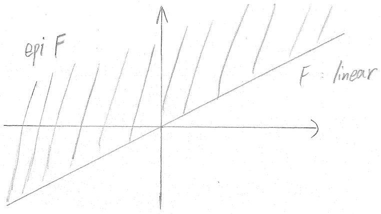
Sets and linear/affine functions

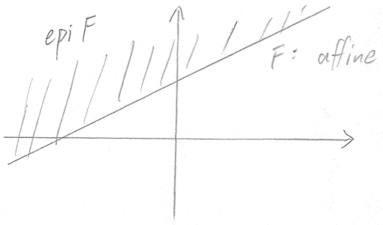
The graph of $F : \mathbb{R}^n \mapsto \mathbb{R}$ is a

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

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

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





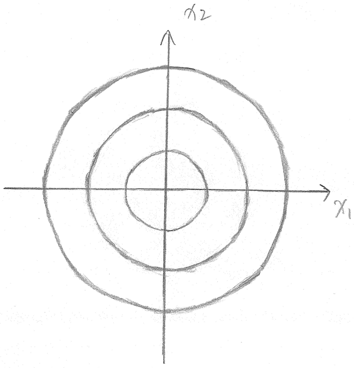


Figure 2.33: Level set 1

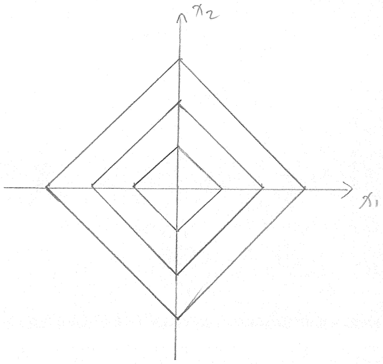


Figure 2.34: Level set 2

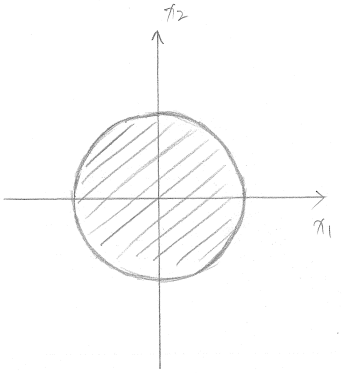


Figure 2.35: Sub-level set 1

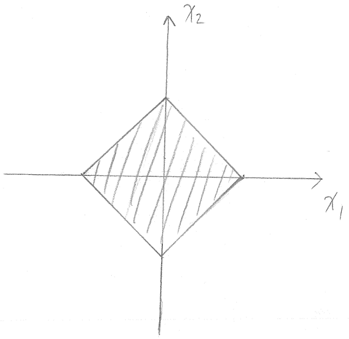
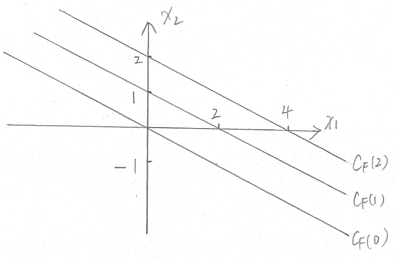


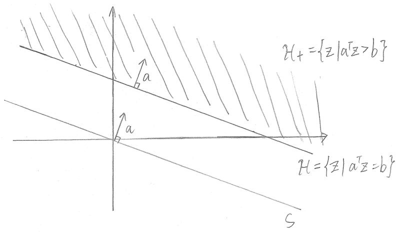
Figure 2.36: Sub-level set 2

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



Definition of a hyperplane:

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


$\mathcal{H}_+ = \{z | a^T z > b\}$
a
a
$\mathcal{H} = \{z | a^T z = b\}$
S

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

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

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

Gradient

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

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

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

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

Example

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

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

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

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

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

Affine approximations

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

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

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

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

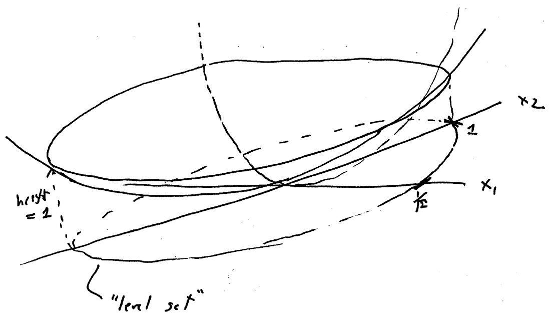
$$\nabla F(x) = \begin{bmatrix} 4x_1 \\ 2x_2 \end{bmatrix}$$

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

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

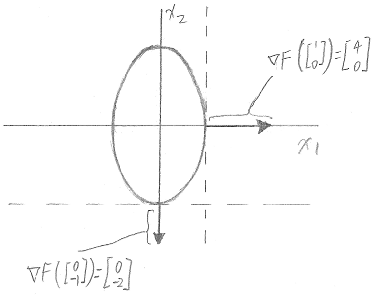
$$\nabla F\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

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



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

$\llcorner$ in above fig $t=1$


x_2
$\nabla F\left(\begin{bmatrix}1\\0\end{bmatrix}\right)=\begin{bmatrix}4\\0\end{bmatrix}$
x_1
$\nabla F\left(\begin{bmatrix}0\\-1\end{bmatrix}\right)=\begin{bmatrix}0\\-2\end{bmatrix}$

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

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

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

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

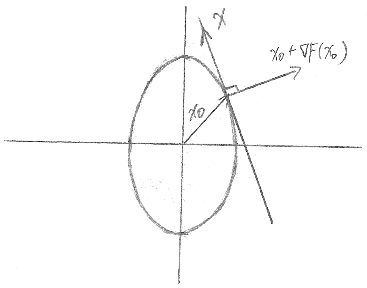
(1) Consider case $c = 0$

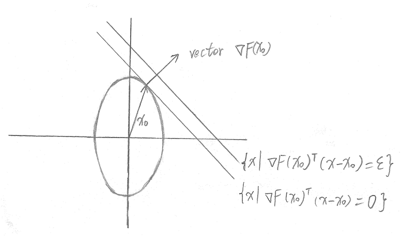
There are points s.t., to approximate $(*)$, has some level as $F(x_0)$ when $c = 0$, we have the set $\{x|\nabla F(x_0)^T(x - x_0) = 0\}$

(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 $x = x_0$.



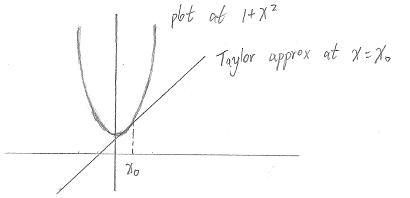

vector $\nabla F(x_0)$
x_0
$\{x \mid \nabla F(x_0)^T (x - x_0) = \varepsilon\}$
$\{x \mid \nabla F(x_0)^T (x - x_0) = 0\}$

In general the set

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

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

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



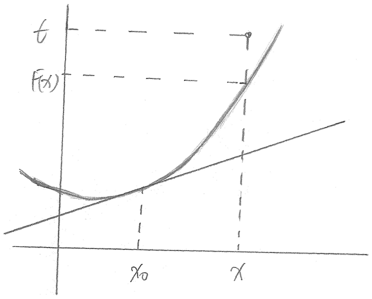
To develop the approximation we need to consider the plot or "graph" at the function F

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

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

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

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



Recap:

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

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

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

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

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

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

Next, we re-arrange

(6)

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

Observe that:

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

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

Now recall connection between angles and inner products.

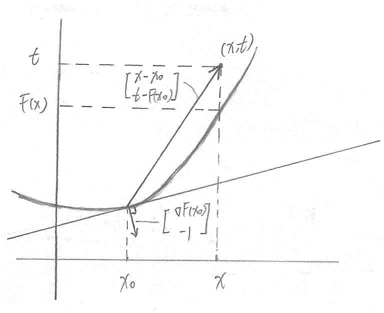
(1). If inner product of 2 vectors is negative, then the angles is obtuse.

(2). Matches picture.

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

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

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


t
F(x)
(x,t)
$\begin{bmatrix} x-x_0 \\ t-F(x_0) \end{bmatrix}$
$\begin{bmatrix} \nabla F(x_0) \\ -1 \end{bmatrix}$
x_0
x

The condition becomes

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

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

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

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

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

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

3 Matrices and eigen decomposition

3.1 Matrices: array of numbers

Matrices are rectangular arrays of numbers:

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

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

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

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

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

Operations of matrices:

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

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

Note: the origin is a all-zero matrix.

Definition 3.1. Inner product of matrices

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

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

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

Length of Matrix: Norm

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

Frobenius Norm:

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

Matrix inverse

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

Some properties for invertible matrices:

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

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

Notice that:

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

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

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

Matrices as linear & affine maps

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

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

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

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

Approximations

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

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

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

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

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

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

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

Orthogonal Matrices

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

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

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

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

(a) Length between vectors

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

(b) Angle between vectors

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

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

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

Range, rank and null space

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

Some important terminology regarding this map are illustrated as follows:

- Domain

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

- Range(or, Column space)

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

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

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

- Rank

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

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

- Nullspace (or, Kernel)

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

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

- Fundamental Theorem

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

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

Theorem 3.3. *Fundamental theorem of linear algebra*

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

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

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

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

PageRank

PageRank algorithm: More important web page should be ranked higher.

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

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

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

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

Let's consider following example

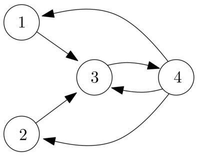


Figure 3.1 A simple example of importance score with 4 webpages and 6 hyperlinks. It is small with much symmetry, leading to a simple calculation of importance scores of the nodes.

Let the score of node 1 be x, and that of node 4 be y. Looking at node 1's incoming links, we see that there is only one such link, coming from node 4 that points to three nodes. So $x = \frac{y}{3}$ and $2x + 2y = 1$. The second equality comes from the normalization of importance scores, and it turns out node 1 and node 2 has the same importance score, and also node 3 and 4 have the same one as well. So the set of importance scores turns out to be [0.125, 0.125, 0.375, 0.375].

Let's define following terminology:

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

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

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

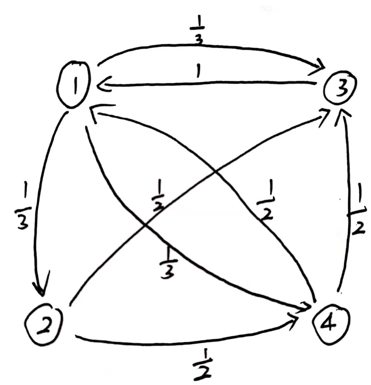
$$\pi^T[k] = \pi^T[k-1]H$$

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

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

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

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



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

The importance scores for each page can be computed as follows

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

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

Note:

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

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

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

Therefore, we have

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

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

[Further reading] In the terminology of stochastic process, the matrix H is called the column stochastic matrix (i.e., each column sum up to one and no entries is negative), and important score for each page we want to compute π^* is the stationary distribution (also equal to the limiting distribution when it is an irreducible Markov chain). The $\pi^*(i)$ is interpreted as, the proportion of time you stay in the state i in a long run.

If the matrix A have repeated eigenvalues,

$$\begin{aligned} Au^{(1)} &= \lambda_1 u^{(1)} \\ Au^{(2)} &= \lambda_2 u^{(2)} \end{aligned}$$

clearly:

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

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

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

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

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

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

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

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

Or equivalently,

$$\begin{aligned} AU &= U\Lambda \\ A &= U\Lambda U^{-1} \\ \Lambda &= U^{-1}AU \end{aligned}$$

Recall pagerank:

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

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

Determinant

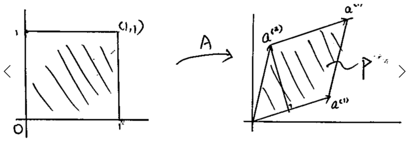
In previous eigenvalue decomposition, we need to solve for λ from $det(A - \lambda I) = 0$ (characteristic function).

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

Example:

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

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



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

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

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

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

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

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

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

4
Symmetric matrices and spectral decomposition

4.1 Symmetric Matrices

The set of n by n square matrix is defined as

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

Following are a few examples of symmetric matrix

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

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

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

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

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

$$\begin{aligned} \frac{d}{dx_k}F(x) &= \frac{d}{dx_k}(\sum_{i=1}^{n} c_i x_i) = c_k \\ \nabla F(x) &= \begin{bmatrix} \frac{\sigma F(x)}{\sigma x_1} \\ \vdots \\ \frac{\sigma F(x)}{\sigma x_n} \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = C \end{aligned}$$

2) Let

$$\begin{aligned}
F(x) &= x^{\mathrm{T}}Ax = \sum_{i=1}\sum_{j=1} a_{ij}x_ix_j \\
&= a_{11}x_1^2 + a_{12}x_1x_2 + \cdots + a_{21}x_2x_1 + \cdots \\
\frac{d}{dx_k}F(x) &= \frac{d}{dx_k}(a_{kk}x_k^2 + \sum_{l\neq k} x_lx_k(a_{lk}+a_{kl})) \\
&= (a_{kk}+a_{kk})x_k + \sum_{l\neq k} x_l(a_{lk}+a_{kl}) \\
&= \sum_{i=1}^{n}(a_{lk}+a_{kl})x_l \\
&= \sum_{i=1}^{n}([A]_{kl}+[A]_{lk})x_l
\end{aligned}$$

Hence,

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

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

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

Symmetric Matrices and Eigenvectors

Theorem 4.1. *4.18 & 4.2 in textbook*

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

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

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

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

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

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

Spectral Decomposition:

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

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

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

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

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

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

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

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

Proof.

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

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

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

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

□

Positive (Semi) Definite matrices (PD & PSD)

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

Alternatively, we denote the set of PSD matrix and the set of PD matrix as
PSD: $S^n_+ = \{A \in S^n | A \succeq 0\}$
PD: $S^n_{++} = \{A \in S^n | A \succ 0\}$
Note: The curled inequality symbol $\succeq$ (and its strict form $\succ$) is used to denote generalized inequality: between vectors, it represents component-wise inequality; between matrices, it represents matrix inequality.

Necessary and sufficient conditions:
(1) A symmetric matrix is PSD iff all its eigenvalues ≥ 0, or, all its **principal minors** are nonnegative.
(2) A symmetric matrix is PD iff all its eigenvalues > 0, or, all its **leading principal minors** are positive (Sylvester's criterion).

Now we prove the argument for PD:

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

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

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

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

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

where only the jth entry is 1.

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

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

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

□

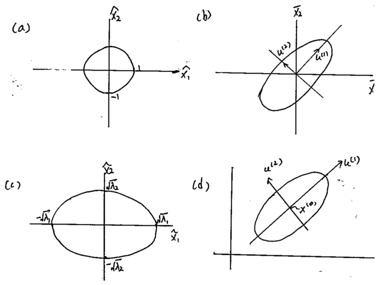
Recall some previous results and note that:
(1) $\det(A) = \prod_{i=1}^{n} \lambda_i$
(2) $\det(A) = 0$ iff there exist eigenvalue $\lambda_i = 0$:
(3) Combine (1) and (2) and our proof above, we have
→ PD matrices are invertible
→ PSD matrices are invertible only if PD

Ellipses

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

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

where matrix P is PD.



Note that the argument above is a quadratic function:

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

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

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

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

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

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

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

So we could compute the covariance matrix by

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

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

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

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

sample mean:

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

sample variance:

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

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

Square-root matrix and Cholesky decomposition

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

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

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

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

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

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

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

5
Singular value decomposition

5.1 The Singular Value Decomposition(SVD)

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

Eigen-decomposition

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

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

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

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

Spectral decomposition

For any n by n symmetric matrix A, we can express A as

$$A = U\Lambda U^{\mathrm{T}}$$

U: Orthogonal matrix ($\perp$ & normalized) $U^{(i)} \in \mathbb{R}^n$;
Λ: Diagonal matrix of $\lambda_i \in \mathbb{R}$.

Singular Value Decomposition(SVD)

For any matrix $A \in \mathbb{R}^{m\times n}$, we can be expressed A as

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

$U \in \mathbb{R}^{m\times m}$: An orthogonal matrix, so $UU^{\mathrm{T}} = U^{\mathrm{T}}U = I_m$
$V \in \mathbb{R}^{n\times n}$: An orthogonal matrix, so $VV^{\mathrm{T}} = V^{\mathrm{T}}V = I_n$

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

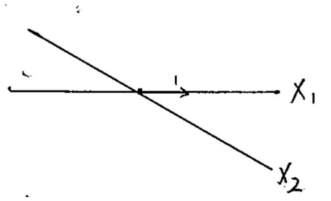
where $\Sigma = diag(\sigma_1, ..., \sigma_r) > 0$, **0** denotes part with all 0.

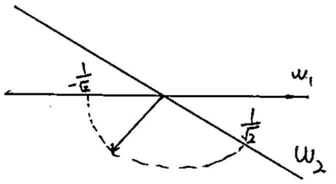
Comments on SVD:

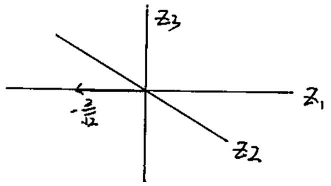
- Inherits $\perp$ matrices of spectral decomposition and all λ_i are real.
- Generalizes eigen decomposition and spectral decomposition to a non-square matrix.
- Lose the property of direction invariance of eigen-decomposition.

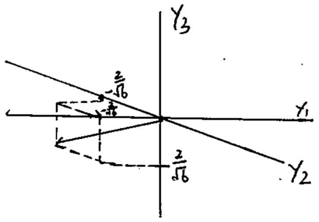
Example 5.1. Let's consider an example, $y = Ax = U\tilde{\Sigma}V^{\mathrm{T}}x$, and see how these matrices impose influences on a vector x(refer to the figures on the r.h.s)

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









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

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

$$3) z = \Sigma w = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ 0 \\ 0 \end{bmatrix}$$

$$4) y = Uz = \begin{bmatrix} -\frac{2}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \end{bmatrix}$$

This example and the figures show that the SVD of a matrix A has lost the property of direction invariance, compared to the eigen-decomposition.

Computing SVD

The idea of singular value decomposition(SVD) follows from eigen decomposition of $A^{\mathrm{T}}A$ and AA^{T}, since both of these two matrices are symmetric, so the spectral theorem is applicable(i.e., $A^{\mathrm{T}}A$ and AA^{T} can be orthogonally diagonalized).

Let A be an m by n matrix. First, we let λ_i denote eigenvalues of the symmetric matrix AA^{T} (or of matrix $A^{\mathrm{T}}A$), and arrange these eigenvalues in a decreasing order, that is, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$(there are n eigenvalues if we consider $A^{\mathrm{T}}A$). We define the singular value of a matrix A as the square root of the eigenvalues and arrange them in a decreasing order as well, i.e., $\sigma_i = \sqrt{\lambda_i}$ and suppose we have r singular values that are positive, so $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

By spectral theorem, we write out U and V as (note that rank$(A) = r$):

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

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

Thus,

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

where $\sigma_i = 0$ if $r + 1 \leq i \leq m$.

We may notice that the vector $u^{(k)}$ is k^{th} eigenvector of AA^{T} by showing that

$$\begin{aligned} (AA^{\mathrm{T}})u^{(k)} &= \sum_{i=1}^{m} \sigma_i^2 u^{(i)} (u^{(i)})^{\mathrm{T}} u^{(k)} \\ &= \sum_{i=1}^{m} \sigma_i^2 \mathbb{1}_{i=k} u^{(i)} \\ &= \sigma_k^2 u^{(k)} \\ &= \lambda_k u^{(k)} \end{aligned}$$

where λ_k is the k-th eigenvalue of AA^{T}, and the indicator function comes from the orthogonality of matrix $\mathcal{U}$.

The same logic can be applied to $A^{\mathrm{T}}A$ and shows that $v^{(k)}$ is k-th eigenvector of $A^{\mathrm{T}}A$.

Accordingly, we have already known how to obtain he SVD of a matrix.

We summarize the procedure that how we compute the SVD of a matrix m by n matrix A as follows:

1) Singular values: Compute eigenvalues of AA^{T} or $A^{\mathrm{T}}A$, and to find the r positive singular values $\sigma_i = \sqrt{\lambda_i(A^{\mathrm{T}}A)}$, so we have the matrix $\tilde{\Sigma}$ with diagonal entries are these positive singular values and other entries are zero.

2) Right-Singular vectors $v^{(i)}$: Find the eigenvectors of $A^{\mathrm{T}}A$, so we have an n by n orthogonal matrix V.

3) Left-Singular vectors $u^{(i)}$: Find the eigenvectors of AA^{T}, so we have an m by m orthogonal matrix $\mathcal{U}$.

4) Write down the expression $A = \mathcal{U}\tilde{\Sigma}V^{\mathrm{T}}$.

Bases for Fundamental Subspaces

Let's consider arbitrary $x \in \mathbb{R}^n$,

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

The above equalities show us that we have lost all components of x along $v^{(i)}$ directions when $r+1 \le i \le n$, i.e., columns of V_{nr}. We summarize the results as follows:

(1) All direction in output are in span $\{u^{(1)}, ..., u^{(n)}\}$
(2) Columns of V_r provide basis for $R(A^{\mathrm{T}})$
(3) Columns of V_{nr} provide basis for $N(A)$.
(4) Columns of U_r provide basis for $R(A)$.
(5) Columns of U_{mr} provide basis for $N(A^{\mathrm{T}})$

Condition number

Most numerical computations involving an equation $Ax = b$ are as reliable as possible when the SVD of A is used. The two orthogonal matrices U and V do not affect lengths of vectors or angles between vectors. Any possible instabilities in numerical calculations are identified in Σ. If the singular values of A are extremely large or small, roundoff errors are almost inevitable, but an error analysis is aided by knowing the entries in Σ and V.

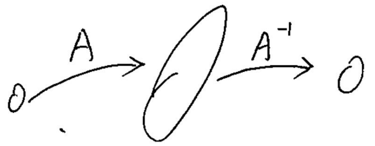
If A is an invertible n by n matrix, then the ratio $\frac{\sigma_1}{\sigma_n}$ of the largest and smallest singular values gives the **condition number** of A (Actually, a "condition number" of A can be computed in several ways).

Let's consider $Ax = b$ when A is invertible. We solve for x and obtain that $x = A^{-1}b$. What if $b = b_r + e$? how much does solution change (let's take b as the true value and e as the round-off error here) ?

Now, our solution is $\hat{x} = A^{-1}b_r + A^{-1}e$.

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

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



where $K(A)$ is defined as the condition number of matrix A.

Note that, by the SVD of A^{-1} we have

$$\begin{aligned}
A^{-1} &= (\mathcal{U}\Sigma V^{\mathrm{T}})^{-1} \\
&= V\Sigma^{-1}\mathcal{U}^{\mathrm{T}} \\
&= V\begin{bmatrix} \frac{1}{\sigma_1} & & \\ & \ddots & \\ & & \frac{1}{\sigma_n} \end{bmatrix}\mathcal{U}^{\mathrm{T}}
\end{aligned}$$

and by the definition of Rayleigh Quotients we have

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

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

Reduced SVD and Pseudo inverse

When matrix $\tilde{\Sigma}$ contains a zero row (or column) vector, we may have a simplified expression compared to the original SVD, namely, reduced SVD:

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

Noticed that the diagonal entries of matrix Σ are non-zero, so we may have the pseudo inverse(i.e., Moore–Penrose inverse) of matrix A, which is given by

$$A^+ = V_r \Sigma^{-1} \mathcal{U}_r^{\mathrm{T}}$$

SVD and matrix norms

Recall the definition of matrix norms (Frobenius norm), it turns out that there is connection between Frobenius norm and the singular values:

$$\begin{aligned} \|A\|_F^2 &= \sum_i \sum_j a_{ij}^2 \\ &= \text{trace}(A^{\mathrm{T}}A) \\ &= \text{trace}(V\tilde{\Sigma}^{\mathrm{T}}U^{\mathrm{T}}U\tilde{\Sigma}V^{\mathrm{T}}) \\ &= \text{trace}(V\tilde{\Sigma}^{\mathrm{T}}\tilde{\Sigma}V^{\mathrm{T}}) \\ &= \text{trace}(VV^{\mathrm{T}}\tilde{\Sigma}^{\mathrm{T}}\tilde{\Sigma}) \\ &= \text{trace}(\tilde{\Sigma}^{\mathrm{T}}\tilde{\Sigma}) \\ &= \text{trace}(\tilde{\Sigma}^2) \\ &= \sum_{i=1}^{r} \sigma_i^2 \end{aligned}$$

That is, the Frobenius norm of A equals to the sum of the square of positive singular values.

6
Linear equations and least squares

6.1 Least Squares

$$Ax = y \qquad (*)$$

where $A \in \mathbb{R}^{m \times n}$ is the coefficient data matrix (known), $y \in \mathbb{R}^m$ are constraints(known) and $x \in \mathbb{R}^n$ are parameters (need to choose).

There are three possibilities:

- a) No $x \in \mathbb{R}^n$ satisfies $(*)$
- b) An unique $x \in \mathbb{R}^n$ satisfies $(*)$
- c) Many $x \in \mathbb{R}^n$ satisfies $(*)$

(a) Existence:

Since $Ax \in R(A)$, a solution will exist if $y \in R(A)$

A simple test based on the ranks of augmented matrix and coefficient matrix:

1) $rank(\begin{bmatrix} A & y \end{bmatrix}) = rank(A)$: solution to $(*)$ exists

2) $rank(\begin{bmatrix} A & y \end{bmatrix}) > rank(A)$: no solution to $(*)$ exists

If a solution exists, is it unique?

Assume a solution $\bar{x}$ exists s.t. $A\bar{x} = y$, any there may have other solution x also satisfy $Ax = y$.

So

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

Any solution to $(*)$ can be expressed as:

$$x = \bar{x} + (x - \bar{x}) = \bar{x} + e$$

where $e \in N(A)$

So if there are many solutions to $(*)$, we will have a affine sets:

$$\mathcal{A} = \{x | x = \bar{x} + N(A), A\bar{x} = y\}$$

The solution is unique if the elements of $N(A)$ are all zero. $\bar{x}$ is any particular solution for $A\bar{x} = y$

The three cases typically bread down into a question about dimensions of $A \in \mathbb{R}^{m \times n}$

- 1) Overdetermined LS: more constraints than parameters $m > n$. Typically a solution does not exist.
- 2) Square: Equal # constraints & parameters, typically $\exists$ a unique solution.
- 3) Underdetermined LS: Fewer constraints than parameters, $m < n$, typically many solutions

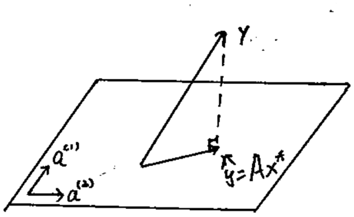
1) Overdetermined: $m > n$

Assume A is full (column) rank, $rank(A) = n$. A is a tall and thin matrix. $dim(R(A)) = n < m$. $y \in \mathbb{R}^m$.

We want to find x^* such that Ax^* is the 'closest' to y:

$$x^* = \arg\min_{x \in \mathbb{R}^n} \|Ax - y\|_2$$

$$\min_{x \in \mathbb{R}^n} ||Ax - y||_2 = \min_{\hat{y} \in R(A)} \|\hat{y} - y\|_2 = \prod_{R(A)} (y)$$



From chapter2,

$$y^* = \sum_{i=1}^{n} x_i^* a^{(i)}$$

where

$$A = \begin{bmatrix} a^{(1)} & \cdots & a^{(n)} \end{bmatrix}$$

Solve for x^* via

$$\sum_{i=1}^{n} x_i^* \langle a^{(k)}, a^{(i)} \rangle = \langle a^{(k)}, y \rangle, \quad \forall k \in \{1, 2, ..., m\}$$

Stack up to get

$$A^{\mathrm{T}}Ax^* = A^{\mathrm{T}}y$$

There are 2 possibilities: (1)$A^{\mathrm{T}}A$ is invertible; (2) $A^{\mathrm{T}}A$ is not invertible.

1) When $A^{\mathrm{T}}A$ is invertible

$$\begin{aligned} x^* &= (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}y \\ \hat{y}^* &= A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}y \end{aligned}$$

2) When $A^{\mathrm{T}}A$ is not invertible We apply SVD to $A^{\mathrm{T}}A$

$$\begin{aligned} \hat{y}^* &= A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}y \\ &= A(V\begin{bmatrix}\Sigma & \mathbf{0}\end{bmatrix}\mathcal{U}^{\mathrm{T}}\mathcal{U}\begin{bmatrix}\Sigma \\ \mathbf{0}\end{bmatrix}V^{\mathrm{T}})^{-1}A^{\mathrm{T}}y \\ &= A(V\Sigma^2V^{\mathrm{T}})^{-1}A^{\mathrm{T}}y \\ &= AV(\Sigma^{-1})^2V^{\mathrm{T}}A^{\mathrm{T}}y \\ &= \mathcal{U}\begin{bmatrix}\Sigma \\ \mathbf{0}\end{bmatrix}V^{\mathrm{T}}V\Sigma^{-2}V^{\mathrm{T}}V\begin{bmatrix}\Sigma & \mathbf{0}\end{bmatrix}\mathcal{U}^{\mathrm{T}}y \\ &= \mathcal{U}\begin{bmatrix}\Sigma \\ \mathbf{0}\end{bmatrix}\Sigma^{-1}\Sigma^{-1}\begin{bmatrix}\Sigma & \mathbf{0}\end{bmatrix}\mathcal{U}^{\mathrm{T}}y \\ &= \mathcal{U}\begin{bmatrix}I_r \\ \mathbf{0}\end{bmatrix}\begin{bmatrix}I_r & \mathbf{0}\end{bmatrix}\mathcal{U}^{\mathrm{T}}y \\ &= \mathcal{U}\begin{bmatrix}I_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathcal{U}^{\mathrm{T}}y \\ &= \mathcal{U}\begin{bmatrix}I_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\langle u^{(1)}, y\rangle \\ \langle u^{(2)}, y\rangle \\ \vdots \\ \langle u^{(m)}, y\rangle\end{bmatrix} \\ &= \sum_{i=1}^{r}\langle u^{(i)}, y\rangle u^{(i)} \end{aligned}$$

b) Uniquely determined: $(m = n)$

$$[A]x = y$$

So,

$$x^* = A^{-1}y$$

- A has full rank (columns & rows).
- Equal number of constraints and parameters.
- A is square matrix with full rank so A^{-1} exists.

c) Underdetermined($m < n$)

- A has full row-rank.

- rank$(A) = m$.

- A is a wide and short matrix.

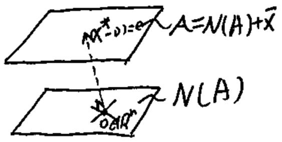
- More parameters than constraints.

- Hence there are many solutions.

Idea: pick solution x that satisfies($*$) with the shortest length in the sense of L_2 norm

$$x^* = \arg \min_{Ax=y, x \in \mathbb{R}^n} \|x\|_2$$

$$\min_{Ax=y, x \in \mathbb{R}^n} \|x\| = \min_{x \in \mathbb{R}^n, x \in \mathcal{A}} \|x - 0\| = \prod_{\mathcal{A}}(0)$$



To solve for x^*, note (from before) error vector

$$e = x^* - 0 = x^* \perp N(A)$$
$$x^* \in N(A)^{\perp} = R(A^{\mathrm{T}})$$

So we can write $x^* = A^{\mathrm{T}}\alpha$ for some $\alpha \in \mathbb{R}^m$. For x^* to be in $\mathcal{A}$ it must be satisfies that $Ax^* = y$

Substituting in we get

$$y = Ax^* = AA^{\mathrm{T}}\alpha$$

By assumption, A is full row rank so AA^{T} is invertible.

$$x^* = A^{\mathrm{T}}\alpha = A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}y$$

Furthermore, we apply SVD to AA^{T},

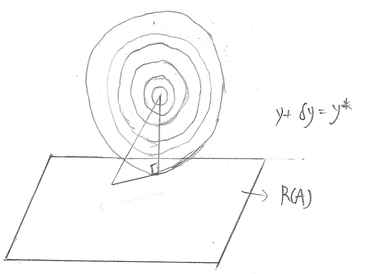
$$\begin{aligned}
x^* &= A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}y \\
&= A^{\mathrm{T}} \left[\mathcal{U} \begin{bmatrix} \Sigma & \mathbf{0} \end{bmatrix} V^{\mathrm{T}} V \begin{bmatrix} \Sigma \\ \mathbf{0} \end{bmatrix} \mathcal{U}^{\mathrm{T}} \right]^{-1} y \\
&= A^{\mathrm{T}}(\mathcal{U}\Sigma^2\mathcal{U}^{\mathrm{T}})^{-1}y \\
&= A^{\mathrm{T}}\mathcal{U}\Sigma^{-2}\mathcal{U}^{\mathrm{T}}y \\
&= V \begin{bmatrix} \Sigma \\ \mathbf{0} \end{bmatrix} \mathcal{U}^{\mathrm{T}}\mathcal{U}\Sigma^{-2}\mathcal{U}^{\mathrm{T}}y \\
&= V \begin{bmatrix} \Sigma^{-1} \\ \mathbf{0} \end{bmatrix} \mathcal{U}^{\mathrm{T}}y \\
&= V \begin{bmatrix} \Sigma^{-1} \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \langle u^{(1)}, y \rangle \\ \dots \\ \langle u^{(n)}, y \rangle \end{bmatrix} \\
&= \sum_{i=1}^{r} \frac{1}{\sigma_i} \langle u^{(i)}, y \rangle v^{(i)}
\end{aligned}$$

Interpretation of $x^* = \arg\min_x \|y - Ax\|_2$:

(1). Approximated solution to $y = Ax$

$y^* = Ax^*$ is the 'best' approximated solution in the sense of L_2 norm, which means, y^* is the closest point in $R(A)$ to y.

(2). Minimum perturbation of y to 'feasibility'



(3). Perturb both y and A to get 'feasibility'

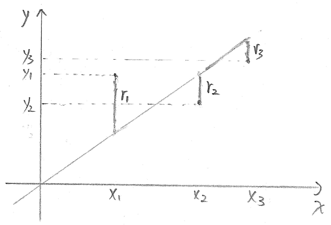
'Total least square'

$$\min_{\delta y, \delta A} \| \begin{bmatrix} \delta A & \delta y \end{bmatrix} \|_F$$

where δA is m by n matrix and thus $\begin{bmatrix} \delta A & \delta y \end{bmatrix}$ is m by $n+1$, $y + \delta y \in R(A + \delta A)$.

(4). Linear regression

$$\|y - Ax\|_2^2 = \sum_{i=1}^{m} (y_i - \langle a^{(i)}, x \rangle)^2 = \sum_{i=1}^{m} r_i^2$$



where r_i defined as the residual.

Example

Let's consider fitting a line to $\{(0,6),(1,0),(2,0)\} = (a_i, y_i)$.

The approximation takes the form of $y = x_1 + ax_2$, and we want to choose a vector x to minimize $\sum_{i=1}[y_i - (x_1 + a_i x_2)]^2 = \sum_{i=1} r_i^2$,

$$\begin{aligned}\|y - Ax\|_2^2 &= \left\| \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\|_2^2 \\ &= \|y - Ax^*\|_2^2 \\ &= 6\end{aligned}$$

Solve for x^*, we have $x^* = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}y = [5, -3]^{\mathrm{T}}$

Thus, the equation for this line is

$$\hat{y} = x_1^* + ax_2^* = 5 - 3a$$

Variants of least square

In the previous classical least square method, we do not consider the weights for each square of the residual(i.e., all of them are equally weighted), however, some residuals might be more important than the others. A very natural approach is to assign different weights to different r_i^2.

Weighted least square:

$$\begin{aligned}\min \sum_{i=1}^{n} w_i^2 r_i^2 &= \|W(y - Ax)\|_2^2 \\ &= \|Wy - WAx)\|_2^2 \\ &= \|\bar{y} - \bar{A}x\|_2^2\end{aligned}$$

where $W = \operatorname{diag}(w_1, w_2, \cdots, w_m)$ and each $w_i \geq 0$, $\bar{y} \triangleq Wy$ and $\bar{A} \triangleq WA$. From the last two equities, we may find that this is very similar with the classic one but now we need to solve for x in a transformed coordinate system.

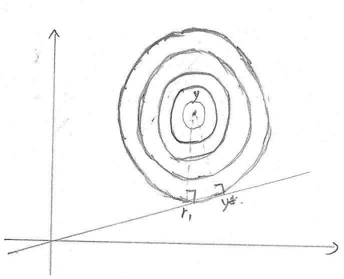
In fact, we can use a more general transform with PSD:

$$\|W(y - Ax)\|_2^2 = (y - Ax)^{\mathrm{T}} W^{\mathrm{T}} W(y - Ax) = r^{\mathrm{T}} W^{\mathrm{T}} W r$$

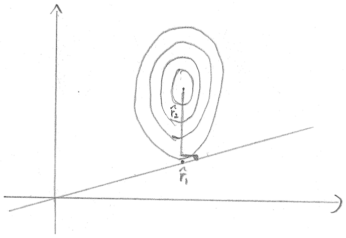
Note that r is the residual in the original coordinate system. Solve for x^*, we have

$$x^* = (A^{\mathrm{T}} W W^{\mathrm{T}} A)^{-1} A^{\mathrm{T}} W W^{\mathrm{T}} y$$

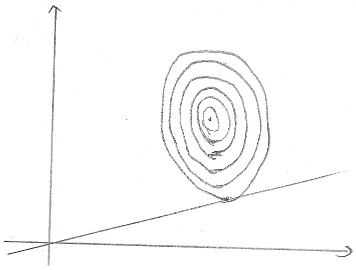
Standard LS



Weighted LS with W is diagonal



Weighted LS with W is PSD (rotation)



L_2- regularization least square

In the original least square,

$$x^* = \arg_{x \in \mathbb{R}_n} \min \|y - Ax\|_2^2$$

we do not have preference for any specific x over any other, and often x is a vector of resources consumed.

Regularized least square

$$x^* = \arg_{x \in \mathbb{R}_n} \min \|y - Ax\|_2^2 + \gamma \|x\|_2^2$$

where γ is a non negative scalar(so if $\gamma = 0$ we retrieve the original LS)

To solve regularized least square, first note that if we have

$$u \in \mathbb{R}^n, v \in \mathbb{R}^m$$

We can define

$$\bar{A} = \begin{bmatrix} A \\ \gamma I \end{bmatrix}, \bar{y} = \begin{bmatrix} y \\ 0_n \end{bmatrix}$$

$$\|Ax - y\|_2^2 + \gamma \|x\|_2^2 = \|\bar{A}x - \bar{y}\|_2^2$$

$$x^* = (\bar{A}^{\mathrm{T}} A)^{-1} \bar{A}^{\mathrm{T}} \bar{y} = (A^{\mathrm{T}} A + \gamma I)^{-1} A^{\mathrm{T}} y$$

'Tikhanov' regularization (also known as ridge regression)

$$\min_x \|W_1(Ax - y)\|_2^2 - \|W_2(x - x^{(0)})\|_2^2 = \min_x \|\bar{A}x - \bar{y}\|_2^2$$

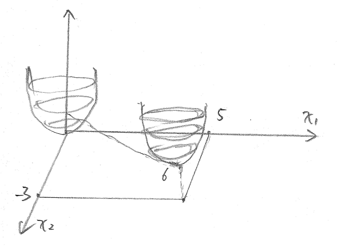
where $\bar{A} = \begin{bmatrix} W_1 A \\ W_2 \end{bmatrix}, \bar{y} = \begin{bmatrix} W_1 y \\ W_2 x^{(0)} \end{bmatrix}$, and W_1 and W_2 are PSD.

Visualize regularized LS: min $\|Ax - y\|_2^2 + \gamma \|x\|_2^2$

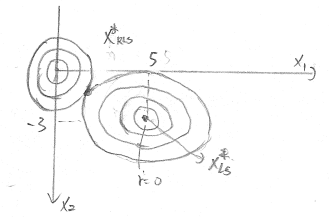
Recall our previous example,

$$x^* = \arg_{x \in \mathbb{R}^n} \min = \left\| \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\|_2^2 = \begin{bmatrix} 5 \\ -3 \end{bmatrix}$$

and $\|Ax^* - y\|_2^2 = 6$.



Draw level set for same picture

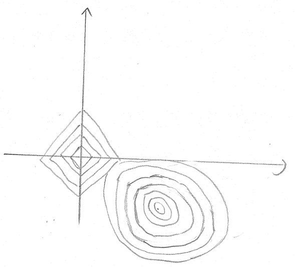


$$
\begin{aligned}
c_1 &= \|Ax - y\|_2^2 \\
&= \|A(x - x^*_{ls} + x^*_{ls}) - y\|_2^2 \\
&= \|(Ax^*_{ls} - y) - A(x - x^*_{ls})\|_2^2 \\
&= \|(Ax^*_{ls} - y)\|_2^2 - \|A(x - x^*_{ls})\|_2^2
\end{aligned}
$$

The first term on the last equality is a scalar 6(from previous example), and so we focus on the geometry of second term.

$$\|A(x - x^*_{ls})\|_2^2 = (x - x^*_{ls})^{\mathrm{T}} A^{\mathrm{T}} A (x - x^*_{ls})$$

. Note that $A^{\mathrm{T}}A$ is a PSD matrix. Understand geometry of level set of $\|A(x - x_{ls})\|_2^2$ via eigenvector of the PSD matrix $A^{\mathrm{T}}A$



$$A^{\mathrm{T}}A = \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix} = \begin{bmatrix} -0.81 & 0.58 \\ 058 & 0.81 \end{bmatrix} \begin{bmatrix} 0.84 & 0 \\ 0 & 7.14 \end{bmatrix} \begin{bmatrix} -0.81 & 0.58 \\ 058 & 0.81 \end{bmatrix}$$

Brief summary of Least Squares

$$x^* = \arg\min_{x \in \mathbb{R}^n} \|y - Ax\|_2^2 \qquad (*)$$

(1) Standard LS variant in (∗) weights all elements of error vector equally.(weighted LS)

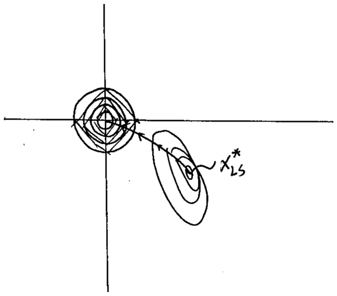
(2) Standard LS measures error along standard coordinate system.(change coordinate system)

(3) Standard LS ignores that certain elements of x may "cost" more than others.(regularization)

"Tikhanov regularization"

$$x^* = \arg\min_{x\in\mathbb{R}^n} \|w_1(y-Ax)\|_2^2 + \|w_2(x-x^{(0)})\|_2^2$$

We do a simple example:



$$x^* = \arg\min_{x\in\mathbb{R}^n} \|y-Ax\|_2^2 + \gamma\|x\|_2^2$$

Look at form of optional solution to

$$\begin{aligned} x^* &= \arg\min_{x\in\mathbb{R}^n} \|y-Ax\|_2^2 + \gamma\|x\|_2^2 \\ &= (A^{\mathrm{T}}A+\gamma I)^{-1}A^{\mathrm{T}}y \\ \hat{y} = Ax^* &= A(A^{\mathrm{T}}A+\gamma I)^{-1}A^{\mathrm{T}}y \end{aligned}$$

Apply SVD to A,

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

First thing is to analyze $(A^{\mathrm{T}}A+\gamma I)^{-1}$:

$$\begin{aligned} (A^{\mathrm{T}}A+\gamma I)^{-1} &= (V\tilde{\Sigma}^{\mathrm{T}}\mathcal{U}^{\mathrm{T}}\mathcal{U}\tilde{\Sigma}V^{\mathrm{T}}+\gamma I)^{-1} \\ &= (V\begin{bmatrix}\Sigma^{\mathrm{T}} & 0 \\ 0 & 0\end{bmatrix}\begin{bmatrix}\Sigma & 0 \\ 0 & 0\end{bmatrix}V^{\mathrm{T}}+\gamma I)^{-1} \\ &= (V\begin{bmatrix}\Sigma^2 & 0 \\ 0 & 0\end{bmatrix}V^{\mathrm{T}}+\gamma VV^{\mathrm{T}})^{-1} \\ &= (V(\begin{bmatrix}\Sigma^2 & 0 \\ 0 & 0\end{bmatrix}+\gamma I)V^{\mathrm{T}})^{-1} \\ &= V\left[\begin{array}{c|c}\Sigma^2+\gamma I_r & \\ \hline & I_{n-r}\end{array}\right]V^{\mathrm{T}} \\ &= V\left[\begin{array}{c|c}(\Sigma^2+\gamma I_r)^{-1} & \\ \hline & (I_{n-r})^{-1}\end{array}\right]V^{\mathrm{T}} \\ &= V\left[\begin{array}{ccc|ccc}\frac{1}{\sigma_1^2+\gamma} & & & & & \\ & \ddots & & & & \\ & & \frac{1}{\sigma_r^2+\gamma} & & & \\ \hline & & & \frac{1}{\gamma} & & \\ & & & & \ddots & \\ & & & & & \frac{1}{\gamma}\end{array}\right]V^{\mathrm{T}} \end{aligned}$$

$$
\begin{aligned}
y^* = Ax^* &= A(A^{\mathrm{T}}A+\gamma I)^{-1}A^{\mathrm{T}}y \\
&= \mathcal{U}\tilde{\Sigma}V^{\mathrm{T}} \left(V \left[\begin{array}{ccc|ccc} \frac{1}{\sigma_1^2+\gamma} & & & & & \\ & \ddots & & & & \\ & & \frac{1}{\sigma_r^2+\gamma} & & & \\ \hline & & & \frac{1}{\gamma} & & \\ & & & & \ddots & \\ & & & & & \frac{1}{\gamma} \end{array}\right] V^{\mathrm{T}} \right) V\tilde{\Sigma}\mathcal{U}^{\mathrm{T}}y \\
&= \mathcal{U} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \left[\begin{array}{ccc|ccc} \frac{1}{\sigma_1^2+\gamma} & & & & & \\ & \ddots & & & & \\ & & \frac{1}{\sigma_r^2+\gamma} & & & \\ \hline & & & \frac{1}{\gamma} & & \\ & & & & \ddots & \\ & & & & & \frac{1}{\gamma} \end{array}\right] \begin{bmatrix} \Sigma^{\mathrm{T}} & 0 \\ 0 & 0 \end{bmatrix} \mathcal{U}^{\mathrm{T}}y \\
&= \mathcal{U} \left[\begin{array}{ccc|c} \frac{\sigma_1^2}{\sigma_1^2+\gamma} & & & \\ & \ddots & & \\ & & \frac{\sigma_r^2}{\sigma_r^2+\gamma} & \\ \hline & & & 0 \end{array}\right] \mathcal{U}^{\mathrm{T}}y \\
&= \mathcal{U} \left[\begin{array}{ccc|c} \frac{\sigma_1^2}{\sigma_1^2+\gamma} & & & \\ & \ddots & & \\ & & \frac{\sigma_r^2}{\sigma_r^2+\gamma} & \\ \hline & & & 0 \end{array}\right] \begin{bmatrix} \langle u^{(1)}, y\rangle \\ \langle u^{(2)}, y\rangle \\ \vdots \\ \langle u^{(m)}, y\rangle \end{bmatrix} \\
&= \mathcal{U} \begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2+\gamma}\langle u^{(1)}, y\rangle \\ \vdots \\ \frac{\sigma_r^2}{\sigma_r^2+\gamma}\langle u^{(r)}, y\rangle \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\
&= \sum_{i=1}^{r} \frac{\sigma_i^2}{\sigma_i^2+\gamma}\langle u^{(i)}, y\rangle u^{(i)}
\end{aligned}
$$

From the last equality,

$$\sum_{i=1}^{r} \frac{\sigma_i^2}{\sigma_i^2+\gamma}\langle u^{(i)}, y\rangle u^{(i)}$$

We should note that:

- $\frac{\sigma_i^2}{\sigma_i^2+\gamma}$: scaling is changed by regularization. If $\gamma = 0$, then $\frac{\sigma_i^2}{\sigma_i^2+\gamma} = 1$ and get back standard LS. If $\gamma > 0$, it's shrinkage.
- $\langle u^{(i)}, y \rangle$: projection of data vector y along that i^{th} direction.
- $u^{(i)}$: component of approximation along i^{th} direction or i^{th} basis element.

7
Linear, quadratic, and quadratically-constrained quadratic programs

7.1 Linear Programs: An Optimization Problem

Terminology and concepts regarding LP problem

Consider following problem, i.e., Liner programming with equality and inequity constraints:

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} c^{\mathrm{T}}x + d \\ s.t. \quad &Ax = b \\ &Gx \leq h \end{aligned}$$

where $x \in \mathbb{R}^n, c \in \mathbb{R}^n, d \in \mathbb{R}, A \in \mathbb{R}^{q\times n}, b \in \mathbb{R}^q, G \in \mathbb{R}^{m\times n}, h \in \mathbb{R}^m$.

Note that, the function $p(x) = c^{\mathrm{T}}x + d$ is called the objective function and x is called the decision variable. The goal of this problem is to find x^* such that the optimal value p^* of the objective function is achieved.

This formulation is the general form, and let's write it in matrix form(list of vectors), that is,

$$A = \begin{bmatrix} \alpha^{(1)^{\mathrm{T}}} \\ ... \\ \alpha^{(q)^{\mathrm{T}}} \end{bmatrix} \qquad G = \begin{bmatrix} g^{(1)^{\mathrm{T}}} \\ ... \\ g^{(q)^{\mathrm{T}}} \end{bmatrix}$$

$$< \alpha^{(i)}, x >= b_i,\ i \in [q]$$

$$< G^{(i)}, x >\leq h_i,\ i \in [m]$$

Note that, there are also two other forms which are commonly used.

1. Inequality form (only contains inequity constraints)

$$\begin{aligned} min \quad & c^T x + d \\ s.t. \quad & Gx \leq h \end{aligned}$$

Given the general form, to get the inequality form, we simply break the equality

$$Ax = b \Leftrightarrow Ax \geq b, Ax \leq b$$

So we can get inequality form as follows:

$$\begin{aligned} \min c^T x + d \\ s.t. \quad Gx \leq h \\ Ax \leq b \\ -Ax \leq -b \end{aligned}$$

2. Standard form (only contains equality and all variables are non negative)

$$\begin{aligned} \min \quad & c^T x + d \\ s.t. \quad & Ax = b \\ & x \geq 0 \end{aligned}$$

Given the general form, we can also convert it into a standard form in 2 steps.

Step 1: Introducing the slack variables s

Given the general form,

$$\begin{aligned} \min \quad & c^T x + d \\ s.t. \quad & Gx \leq h \\ & Ax = b \end{aligned}$$

We add the slack variables s so that the formulation become:

$$\begin{aligned} \min \quad & c^T x + d \\ s.t. \quad & Gx + s = h \\ & Ax = b \\ & s \geq 0 \end{aligned}$$

Note that the slack variables s must be non negative here.

Step 2: We break the decision variable x by $x = x^+ - x^-$

$$\begin{aligned} \min \quad & c^T(x^+ - x^-) + d \\ s.t. \quad & G(x^+ - x^-) + s = h \\ & A(x^+ - x^-) = b \\ & s \geq 0 \\ & x^+ \geq 0 \\ & x^- \geq 0 \end{aligned}$$

Concepts that are frequently used in LP (and also optimization theory):

(1) Feasible set(or feasible region): The set of points S that are satisfying all the constraints, i.e.,

$$S = \{x \in \mathbb{R}^n | Ax = b, Gx \leq h\}$$

(2) Feasible solution: The points in the feasible set S.

(3) Polyhedron: intersection of finite number of half-spaces, i.e.,

$$\{x \in \mathbb{R}^n | Gx \leq h\}$$

(4) Polytope: bounded intersection of finitely many half-spaces.

Let p^* be the optimal value of the given objective function under the constraints, i.e.,

$$\begin{aligned} p^* = \min \quad & c^T x + d \\ s.t. \quad & Ax = b \\ & Gx \leq h \end{aligned}$$

Remarks on "optimal" value p^* of program:

- Lowest cost choice amongst all feasible x.
- Possible here is no minimal choice
- possible no feasible choice
- $p^* \in \mathbb{R}$

Let x^* be the optimal choice of the decision variable x, i.e.,

$$\begin{aligned} x^* = \arg\min \quad & c^T x + d \\ s.t. \quad & Ax = b \\ & Gx \leq h \end{aligned}$$

Remarks on "optimal" solution x^* of program:

- Sometimes x^* does not exist
- If exists, may not be unique
- $x^* \in \mathbb{R}^n$

Let's consider an example:

During the The Second World War, the US army is considering how to make their soldiers have enough nutrients...

Different nutrients in different foods and daily requirement:

Nutrients	Meat	Potatoes	Daily Requirement
Carbohydrates	40	200	400
Protein	100	20	200
Fiber	5	40	40

The price of meat and potatoes:

Resources	cost/kg
Meat	\$ 1
Potatoes	\$ 0.25

Let x_1 denotes meat(kg) and x_2 denotes potatoes(kg), and we formulate this LP as follows:

Objective function:

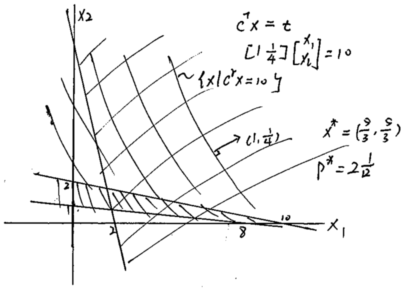
$$\min_{x_1,x_2} x_1 + \frac{1}{4}x_2 = \min_{x_1,x_2} \begin{bmatrix} 1 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Constrains:

$$\begin{aligned} 40x_1 + 200x_2 &\geq 400 \\ 100x_1 + 20x_2 &\geq 200 \\ 5x_1 + 40x_2 &\geq 40 \\ x_1 &\geq 0 \\ x_2 &\geq 0 \end{aligned}$$

Rewrite it as $Gx \leq h$, that is,

$$\begin{bmatrix} -\frac{1}{5} & -1 \\ -\frac{1}{8} & -1 \\ -5 & -1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} -2 \\ -1 \\ -10 \\ 0 \\ 0 \end{bmatrix}$$


x_2
$c^T x = t$
$[1\ \frac{1}{4}]\begin{bmatrix}x_1\\x_2\end{bmatrix} = 10$
$\sim \{x \mid c^T x = 10\}$
$(1, \frac{1}{4})$
$x^* = (\frac{9}{3}, \frac{5}{3})$
$p^* = 2\frac{1}{12}$
2
1
2
8
10
x_1

LP without constraints

Consider the LP does not have constraints, so we have

$$\begin{aligned} p^* &= \min c^{\mathrm{T}} x + d \\ x^* &= \arg \min_{x \in \mathbb{R}^n} \quad c^{\mathrm{T}} x + d \end{aligned}$$

Situation 1: $c = 0 \in \mathbb{R}^n$

$$\begin{aligned} p^* &= \min_{x \in \mathbb{R}^n} \quad d = d \\ x^* &= \arg \min_{x \in \mathbb{R}^n} \quad d = \mathbb{R}^n \end{aligned}$$

Situation 2: $c \neq 0 \in \mathbb{R}^n$

$$\begin{aligned} p^* &= -\infty \quad \text{by convention if no minimum} \\ x(\alpha) &= -\alpha c \quad \alpha \geq 0 \\ c^{\mathrm{T}} x + d = c^{\mathrm{T}}(-\alpha c) + d &= \alpha - \alpha c^{\mathrm{T}} c = \alpha - \alpha \|c\|_2^2 \\ x^* &\text{doesn't exist} \end{aligned}$$

Thus, for unconstrained LP we conclude that

$$p^* = \begin{cases} d & \text{if } c = 0 \\ -\infty & \text{otherwise} \end{cases} \qquad x^* = \begin{cases} \mathbb{R}^n & \text{if } c = 0 \\ \text{doesn't exist} & \text{otherwise} \end{cases}$$

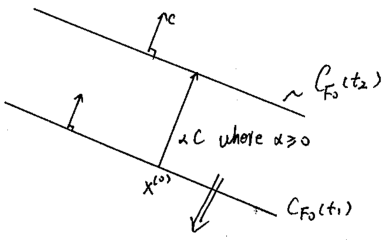
Let's think about the geometry of cost function:

$$F_0(x) = c^{\mathrm{T}} x + d$$

where $F_0(x)$ is the objective function and it turns out that it is also an affine function.

Recall that the level set for $F_0(x)$,

$$\begin{aligned} c_{F_0}(t) &= \{x \in \mathbb{R}^n | F_0(x) = c^{\mathrm{T}} x + d = t\} \\ &= \{x \in \mathbb{R}^n | C^{\mathrm{T}} x = (t - d)\} \end{aligned}$$



Obviously, the level set for $F_0(x)$ defines a hyperplane, and when $t = d$ it defines a subspace(go through the origin). Let's consider two level sets, for t_1 and t_2,

Note that c is the normal vector to x. Let's find the relationship between t_1 and t_2:

Approach (1)

$$\begin{aligned} t_2 &= c^{\mathrm{T}}(x^{(0)} + \alpha c) + d \\ t_1 &= c^{\mathrm{T}}x^{(0)} + d \\ t_2 - t_1 &= [c^{\mathrm{T}}x^{(0)} + d + \alpha\|c\|^2] - c^{\mathrm{T}}x^{(0)} = \alpha\|c\|^2 \end{aligned}$$

So apparently $t_2 > t_1$.

Approach (2)

$$\nabla F_0(x) = \begin{bmatrix} \frac{\sigma}{\sigma x_1}(c^{\mathrm{T}}x + d) \\ \vdots \\ \frac{\sigma}{\sigma x_n}(c^{\mathrm{T}}x + d) \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = c$$

The gradient points out that the direction of c is the direction of increase in $F_0(x)$. In fact, we could show that the direction of the gradient evaluated at a certain point, is the direction that the value of function increases more rapidly(remind yourself what you have learned in Calculus).

Hence, to minimize the objective function, we go in opposite direction of the gradient, that is, we should go as far as possible along the direction of $-c$ (unless $c = 0$).

Now, let's turn back to LP with following constraints as we specified before:

$$\begin{aligned} Ax &= b \\ Gx &\leq h \end{aligned}$$

Interesting results and interpretation for these two kind of constraints:

(1) $Ax = b$ (Equality constraints): force x^* into an affine set

$$\{x \in \mathbb{R}^n | Ax = b\} = \cap_{i=1}^{q} \{x \in \mathbb{R}^n | < \alpha^{(i)}, x >= b_i\}$$

(2) $Gx \leq h$ (Inequality constraints): force x^* to be in an intersection of half-spaces

$$\{x \in \mathbb{R}^n | Gx \leq h\} = \cap_{i=1}^{q} \{x \in \mathbb{R}^n | < g^{(i)}, x > \leq h_i\}$$

(3) The feasible set: intersection of half-spaces and hyperplanes

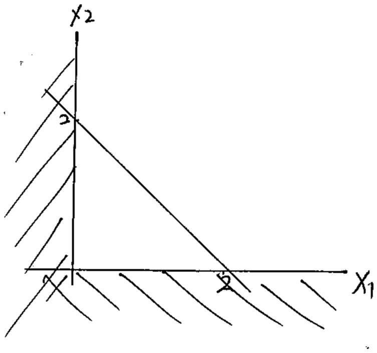
$$S = \left(\cap_{i=1}^{q} \{x \in \mathbb{R}^n | < \alpha^{(i)}, x >= b_i\}\right) \cap \left(\cap_{i=1}^{m} \{x \in \mathbb{R}^n | < g^{(i)}, x > \leq h_i\}\right)$$

A few remarks:

- Concepts of polyhedron and polytope.
- Ax = b → $Ax \leq b$, $Ax \geq b$

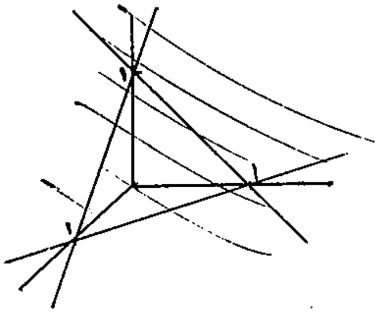
Example 7.1.

$$A = \begin{bmatrix} 1 & 1 \end{bmatrix} b = \begin{bmatrix} 2 \end{bmatrix}$$
$$G = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} h = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$
$$Ax = b \rightarrow x_1 + x_2 = 2$$
$$Gx \leq h \rightarrow x_1 \geq 0, x_2 \geq 0$$



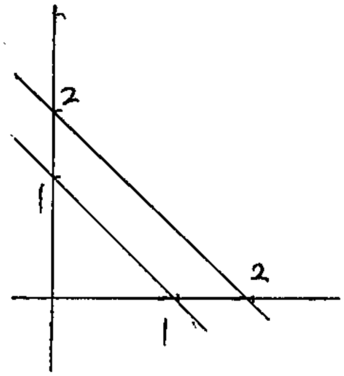
Example 7.2. Generally, equality constraints move you from a higher dimensional geometry in $\mathbb{R}^n$ to a slice, which is a lower dimensional geometry. See the following example of losing 1 dimension per linearly independent constraint:

$$A = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$$
$$B = \begin{bmatrix} 1 \end{bmatrix}$$



Example 7.3. In some cases, there may be no intersection between hyperplane:

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$



So the feasible set is empty, $S = \emptyset$.

This situation may also happen with inequalities constraints:

$$\begin{bmatrix} -1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$
$$S = \emptyset$$

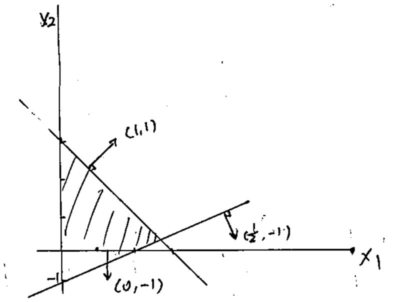
Example 7.4. Generally, to facilitate sketch will often just sketch inequity constraints:

$$\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ \frac{1}{2} & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} 0 \\ 0 \\ 1 \\ 3 \end{bmatrix}$$

So we have

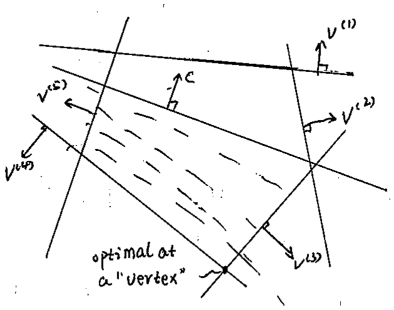
$$
\begin{aligned}
x_1 &\geq 0 \\
x_2 &\geq 0 \\
x_2 &\geq -1 + \frac{1}{2}x_1 \\
x_2 &\leq 3 - x_1
\end{aligned}
$$

Sketch the feasible set:



The rows of matrix G are the normal directions of the hyperplanes that define the half-spaces, and the normal directions point outward from the feasible set.

Example 7.5. Let's combine all these things together, liner objective function, linear equality constraints and linear inequity constraints.



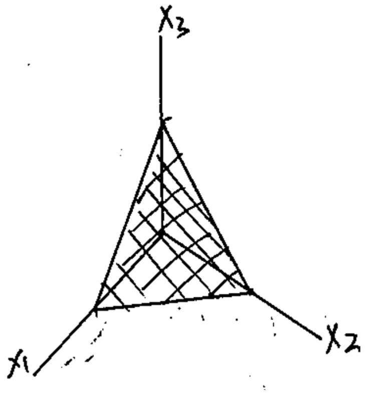
What's optimum? Looks like $x^* = v^{(3)}$.

We may also have a facet with the feasible set

$$S = \{x \in \mathbb{R}^3 | x_1 + x_2 + x_3 = 1, x_1 \geq 0, x_2 \geq 0, x_3 \geq 0\}$$

So there are various possibilities here for p^* and x^*:

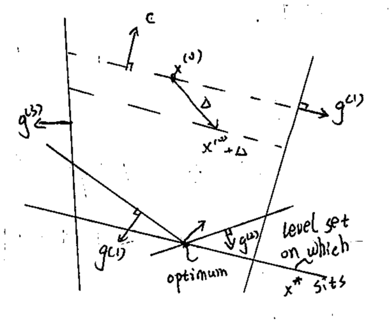
- 1) x^* is unique, p^* finite
- 2) x^* is not unique, p^* finite.



- 3) There is no x^*:

  a) $S = \emptyset$ (Feasible set is empty), constraint $p^* = \infty$. So we say that this LP problem is infeasible.

  b) S is unbounded & no minimum, constraint $p^* = -\infty$



Active constraints: An optimal solution that lies at the intersection point of two constraints causes both of those constraints to be considered active

Inactive constraints: If any of the constraint lines do not pass through the optimal point, those constraints are called inactive.

In this example(see picture above) constraints g(1) and g(2) are active at optimum.

Note that, we could improve(decrease) the cost if:

$$c^T(x^{(0)} + \triangle) + d < c^T x^{(0)} + d$$

That is, $\langle c, \triangle \rangle < 0$, the angle between displacement vector $\triangle$ and normal vector c is an obtuse angle(see the picture above).

Some observations:

- If you are at a vertex(doesn't have to be optimum).

- Any "move" that keeps you feasible must also let you move into the feasible set

  → opposite vector that define active constraints.

$$v - \alpha g^{(1)} - \beta g^{(2)}, \quad \alpha, \beta \geq 0$$

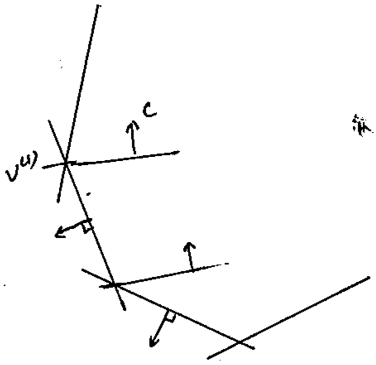
- Are these any choices of α, β that decrease the cost?

$$c^{\mathrm{T}}(v - \alpha g^{(1)} - \beta g^{(2)}) + d \leq c^{\mathrm{T}} x + d$$
$$-\alpha \langle c, g^{(1)} \rangle - \beta \langle c, g^{(2)} \rangle \leq 0$$

If:
1) $\langle c, g^{(1)} \rangle < 0$
2) $\langle c, g^{(2)} \rangle < 0$
no more into feasible set will decrease the cost.

Condition for optimality:

A feasible vertex v: $v \in \{x | Gx \leq h\}$ is an optimal solution to LP with cost $F_0(x) = c^{\mathrm{T}} x + d$ if $c^{\mathrm{T}} g^{(i)} < 0$, $\forall i \in$ active set.



Simplex Algorithm

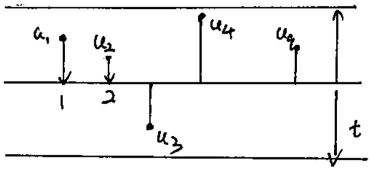
Simplex algorithm:

- 1) Start from a feasible vertex;
- 2) Identify direction of cost decrease along an edge;
- 3) Move on that direction until any further more would violate a previously inactive constraints.
- 4) Stop + add that new constraint(s) to active set.
- 5) Repeat

Example 7.6. Consider the problem:

$$\begin{aligned} &min\|Ax-b\|_\infty \\ &s.t.Gx \leq h \end{aligned}$$

Recall that $\|u\|_\infty = max_{i\in[n]}|u_i|$,

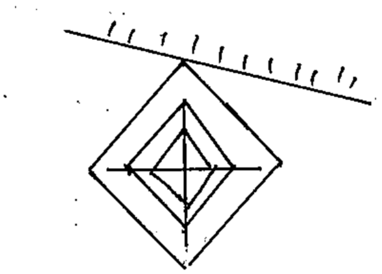


Introduce a helper(auxiliary) variable $t \in \mathbb{R}$ which corresponding to the value of the norm, so the we convert the original problem into following problem:

$$\begin{aligned} min \quad & t \\ s.t.Ax-b &\leq t\mathbf{1} \\ Ax-b &\geq (-t)\mathbf{1} \\ Gx &\leq h \end{aligned}$$

Example 7.7. Consider the problem:

$$\begin{aligned} &\min_x\|Ax-b\|_1, \qquad A \in \mathbb{R}^{q\times n} \\ &s.t.Gx \leq h \end{aligned}$$



Recall the definition of L_1 norm:

$$\|u\|_1 = \sum_{i=1}^{q}|u_i|$$

Let's introduce the helper vector $t \in \mathbb{R}^q$,

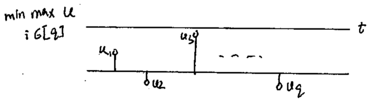
$$\begin{aligned} \min_{x,t} \sum_{i=1}^{q} t_i & \\ s.t. \quad Gx &\le h \\ Ax - b &\le t \\ Ax - b &\ge -t \end{aligned}$$

Example 7.8. Consider following problem:

$$\begin{aligned} min \quad max_{i \in [q]} (c^{(i)^{\mathrm{T}}} x + d_i) \\ s.t. Gx \le h \end{aligned}$$

This case is similar to the L_∞ norm case due to the inner max function, but it is one-sided(no lower bound). So we could convert the original one to the following:

$$\begin{aligned} min \quad & t \\ s.t. (c^{(i)^{\mathrm{T}}} x + d_i) \le t, \quad & \forall i \in [q] \\ Gx \le h & \end{aligned}$$



Remark:

In above three examples, the decision variables in initial formulation are $x \in \mathbb{R}^n$, but in reformulation they become:

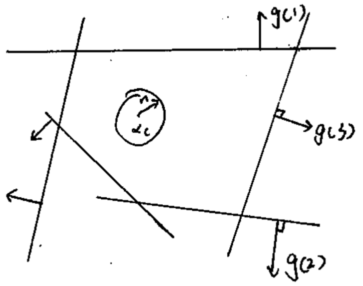
Example 7.6: $(x, t) \in \mathbb{R}^n \times \mathbb{R}$

Example 7.7: $(x, t) \in \mathbb{R}^n \times \mathbb{R}^n$

Example 7.8: $(x, t) \in \mathbb{R}^n \times \mathbb{R}$

Example 7.9. Finding the largest L_2 ball that fits in a polytope.

Let $p = \{x \in \mathbb{R}^n | Gx \le h\}$,



It turns out that, this problem can be formulated as an LP problem.

Note that:

- A sphere is fully parameterized by its center x_c and its radius r
- A sphere (x_c, r) fits in p if: $x_c + u \in p, \forall u \quad s.t. \|u\|_2 \leq r$

Now, let's formulate as an LP. To accomplish this, observe that $x_c + u \in p$ means that ${g^{(i)}}^{\mathrm{T}}(x_c + u) \leq h_i, \quad \forall i \in [q], \forall u$ s.t. $\|u\|_2 \leq r$.

Examine the constraint ${g^{(i)}}^{\mathrm{T}} x_c + {g^{(i)}}^{\mathrm{T}} u \leq h_i$ at a time:

As for ${g^{(i)}}^{\mathrm{T}} u$, what's the direction of u in order to make this term large as possible? It turns out, u must be aligned with $g^{(i)}$, the same direction with $g^{(i)}$.

Furthermore, what's the value of u that is aligned with $g^{(i)}$ and satisfies $|u|_2 = r$? It should be:

$$u^* = \frac{g^{(1)}}{\|g^{(1)}\|} r$$

So, if the following is satisfied:

$${g^{(i)}}^{\mathrm{T}} [x_c + u^*] \leq h_i$$

Then constraint i will be satisfied for all u such that $|u| \leq r$.

Substituting in u^*, the constraints become:

$${g^{(i)}}^{\mathrm{T}} x_c + \|g^{(i)}\|_2 r \leq h_i$$

Note that the constraint is linear in x_c and r.

The problem of finding the largest sphere becomes the following:

$$\begin{aligned} \max \quad & r \\ s.t. \quad & {g^{(i)}}^{\mathrm{T}} x_c + r \|g^{(i)}\|_2 \leq h_i \qquad \forall i \in [q] \end{aligned}$$

Remarks:

(1) Optimal value $(x_c, r) \in \mathbb{R}^{n+1}$.

(2) x_c is a variable that does not enter the objective.

(3) Possibly no solution if p is an empty set.

(4) Constraints and objective are linear in optimal variable so it is an LP problem.

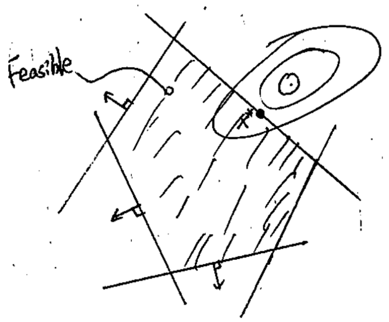
(5)Transformed some quadratic-like problem (quadratic as a sphere is involved) into an LP, because we were able to identify which direction of u the worst case for each constraint.

7.2 Quadratic program(QP)

A quadratic programming problem is formulated in a general way as follows

$$
\begin{aligned}
p^* = \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2}x^T Hx + c^T x + d \\
s.t. \quad & Ax = b \\
& Gx \leq h
\end{aligned}
$$

The feasible set(feasible region) is illustrated as follows



Connection between LS problem and QP problem

Recall the lest square problem we have studied, it turns out that we can convert the least square problem into a QP problem.

$$
\begin{aligned}
\|Ax - y\|_2^2 &= (Ax - y)^T (Ax - y) \\
&= x^T A^T Ax - 2y^T Ax + y^T y \\
&= \frac{1}{2}x^T (2A^T A)x - 2y^T Ax + \|y\|_2^2
\end{aligned}
$$

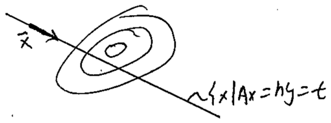
However, for the converse case, we can not always manipulate the objective of a QP into a LS problem.

Equality constrained QPs: Substitute back to the objective

A basic idea of solving such kind of problem is to substitute the equality constraints back to the objective function so that hopefully we can eliminate some variables(dimensions), and obtain an unconstrained QP problem.

Consider a formulation of such kind of problem,

$$
\begin{aligned}
p^* = min_{x \in \mathbb{R}^n} \quad & \frac{1}{2}x^T Hx + c^T x + d \\
s.t. \quad & Ax = b
\end{aligned}
$$



The feasible set could be expressed as follows

$$\mathcal{A} = \{x | x = \bar{x} + \xi, \quad \text{where } \xi \in N(A)\}$$

where $\bar{x}$ is a particular solution to the equation $Ax = b$.

Let N be a basis for $N(A)$, so that we can express $\xi \in N(A)$ as $\xi = Nz$, where $z \in \mathbb{R}^k$, $k = \dim(N(A)) \leq n$.

Substitute this expression for any feasible x into the objective $F_0(\cdot)$

$$\begin{aligned} F_0(x) &= F_0(\bar{x} + \xi) \\ &= F(\bar{x} + Nz) \\ &= \frac{1}{2}(\bar{x} + Nz)^{\mathrm{T}} H(\bar{x} + Nz) + c^{\mathrm{T}}(\bar{x} + Nz) + d \\ &= \frac{1}{2} z^{\mathrm{T}}[N^{\mathrm{T}} H N] z + [c^{\mathrm{T}} N + \bar{x}^{\mathrm{T}} H N] z + [\frac{1}{2} \bar{x}^{\mathrm{T}} H \bar{x} + c^{\mathrm{T}} x + d] \\ &= \frac{1}{2} z^{\mathrm{T}} \tilde{H} z + \tilde{c^{\mathrm{T}}} z + \tilde{d} \end{aligned}$$

where we newly define $\tilde{H}$, $\tilde{c^{\mathrm{T}}}$, and $\tilde{d}$ in the last equality. So now we have obtain an unconstrained QP problem which is formulated as

$$p^* = min_{z \in \mathbb{R}^k} \quad \frac{1}{2} z^{\mathrm{T}} \tilde{H} z + \tilde{c^{\mathrm{T}}} z + \tilde{d}$$

Hence, to summarize,

(1)We have got a newly lower-dimensional optimization problem since $k = \dim(N(A)) \leq n$.

(2)The new problem is still a QP, but is an unconstrained QP.

Example 7.10. Markowitz Portfolio optimization/Mean/variance" analysis Harry Markowitz(1990 Nobel Prize)

Problem formulation:

- Objective: For a fixed level of (expected) returns, we want to minimize the variance of returns.
- There are n stocks, and we only consider a single investment period
- Design an optimal investment strategy $p \in \mathbb{R}^n$, where the component $p_i =$ is the weight of your total wealth invested in the stock i.

- We require $\sum_{i=1}^{n} p_i = 1$, that is, you must invest all your money.
- We also require $p_i \geq 0$, that is, you are only allowed to take long positions, and short selling is not allowed.
- Your total wealth is normalized to 1. So p_i is not only weight but also the amount of money you invest on stock i now.

  Return and Variance:

- Let $x \in \mathbb{R}^n$ be a random vector denotes the return of n stocks, so component x_i is the return on the i-th stock in one period. That is, if we invest 1 RMB in a stock at the beginning, we will get x_i RMB back at the end of this period.
- Expected returns: $\bar{x}_i = \mathbb{E}[x_i]$. In general, it is a known estimator based on historical data.
- Your return (random) is $\sum_{i=1}^{n} p_i x_i = p^{\mathrm{T}} x$
- Your expected return $\mathbb{E}[\sum_{i=1}^{n} p_i x_i] = \sum_{i=1}^{n} p_i \mathbb{E}[x_i] = \sum_{i=1}^{n} p_i \bar{x}_i = p^{\mathrm{T}} \bar{x}$
- variance in your return:

$$
\begin{aligned}
var(p^{\mathrm{T}} x) &= \mathbb{E}[(p^{\mathrm{T}} x - p^{\mathrm{T}} \bar{x})^2] \\
&= \mathbb{E}[(p^{\mathrm{T}}(x - x^{\mathrm{T}}))^2] \\
&= \mathbb{E}[p^{\mathrm{T}}(x - \bar{x})(x - \bar{x})^{\mathrm{T}} p] \\
&= p^{\mathrm{T}} \mathbb{E}[(x - \bar{x})(x - \bar{x})^{\mathrm{T}}] p \\
&= p^{\mathrm{T}} \Sigma p
\end{aligned}
$$

With above terminology and model formulation, given a mean return $\bar{x}$ and the covariance matrix of returns $\Sigma \in \mathbb{R}^{n \times n}$, we want to find an optimal strategy/policy p to minimize the risk(represent by variance) subject to same minimal returns.:

The problem is formulated as:

$$
\begin{aligned}
min_{p \in \mathbb{R}^n} \quad & p^{\mathrm{T}} \Sigma p \\
s.t. \quad & p^{\mathrm{T}} \bar{x} \geq r_{min} \\
& \mathbf{1}^{\mathrm{T}} p = 1 \\
& p \geq 0
\end{aligned}
$$

Geometry of QP

Consider the general form of QP

$$
\begin{aligned}
p^* = min_{x \in \mathbb{R}^n} \quad & \frac{1}{2} x^{\mathrm{T}} H x + x^{\mathrm{T}} x + d \\
s.t. \quad & Ax = b \\
& Gx \leq h
\end{aligned}
$$

Follow the same as LP, we would like to discuss the geometry of QP.

(1) Geometry of feasible set: The same as LP.

(2) Geometry of objective: Without loss of generality(w.l.o.g), we can assume that $H \in S^n$, i.e., H is symmetric, since

$$\begin{aligned} x^{\mathrm{T}} H x &= \frac{1}{2}[x^{\mathrm{T}} H x + x^{\mathrm{T}} H^{\mathrm{T}} x] \\ &= \frac{1}{2} x^{\mathrm{T}} (H + H^{\mathrm{T}}) x \end{aligned}$$

clearly H is symmetric.

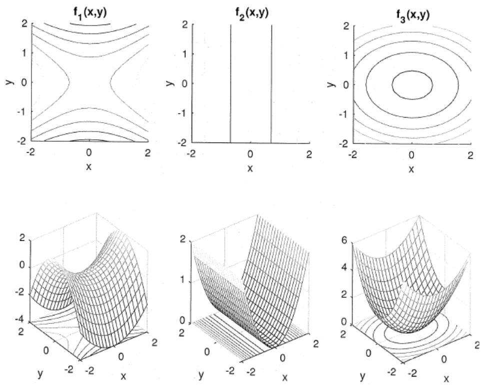


Figure 7.1: Plot 1

Recal that for symmetric matrices, we have

a) Eigenvalues: purely real eigenvalues (so we can arrange them in an order).

b) Eigenvectors: can be chosen to be $\perp$ and can always diagonalize H, i.e., we can write

$$H = \mathcal{U} \Lambda \mathbb{U}^{\mathrm{T}} = \sum_{i=1}^{n} \lambda_i u^{(i)} u^{(i)^{\mathrm{T}}}$$

Consider following 3 different cases for the matrix H:

- A) $H \in S^n$ but not PSD

$$H = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}$$

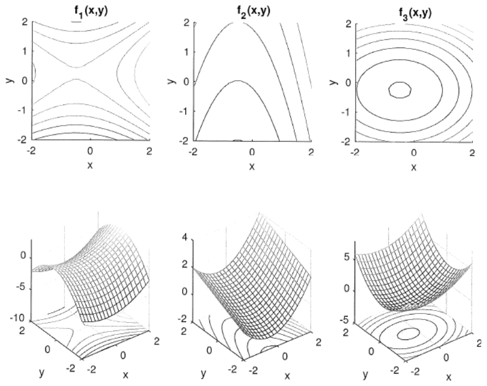


Figure 7.2: Plot 2

- B) $H \in S^n_+$ but not PD

$$H = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

- C) $H \in S^n_{++}$

$$H = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

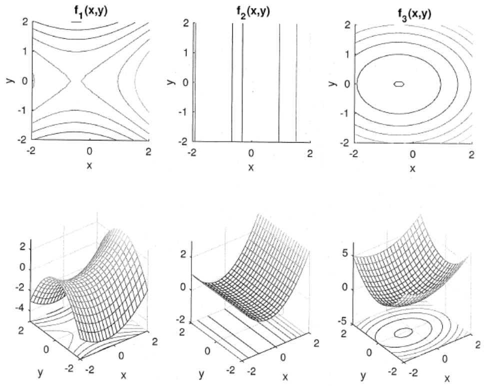


Figure 7.3: Plot 3

Consider the 3 cases for H given above and the following different objective functions, we plot these figures on r.h.s. for illustration.

Plot 1: $F_0(x) = \frac{1}{2} x^{\mathrm{T}} H x$

Plot 2: $F_0(x) = \frac{1}{2} x^{\mathrm{T}} H x + [0.5 \quad 0.5] x$

Plot 3: $F_0(x) = \frac{1}{2} x^{\mathrm{T}} H x + [0.5 \quad 0] x$

Let's consider 3 different cases for a general QP problem.

Case A: $H \in S^n$ but $H \notin S^n_+$ (Symmetric not PSD).

For such H, there must be an eigenvalue/vector pair (λ, u) s.t. $\lambda < 0$.

Set $x_\alpha = \alpha u$ for some $\alpha \in \mathbb{R}$, we have

$$\begin{aligned} F_0(\alpha u) &= \frac{1}{2}(\alpha u)^{\mathrm{T}} H(\alpha u) + c^{\mathrm{T}}(\alpha u) + d \\ &= \frac{\alpha^2}{2} u^{\mathrm{T}} [\sum_{i=1}^{n} \lambda_i u^{(i)} u^{(i)^{\mathrm{T}}}] u + \alpha c^{\mathrm{T}} u + d \\ &= \frac{\alpha^2}{2} \lambda + \alpha < c, u > + d \end{aligned}$$

Since $\lambda < 0$, let $\alpha \to \infty$ leads to an unbounded objective function, i.e., $p^* = -\infty$.

Case B: $H \in S_+^n$ but $H \notin S_{++}^n$ (H is PSD but not PD) $\to H$ has at least 1 zero eigenvalue.

Case B (i) $H \in S_+^n$ but $H \notin S_{++}^n$ and $c \notin R(H)$.

There is a complement of c in $R(H)^{\perp} = N(H^{\mathrm{T}}) = N(H)$, and thus we can move in that direction without affecting 2^{nd} order term while driving 1^{st} order term to $-\infty$.

Let $c_{\parallel} = \prod_{R(H)}(c)$, and $c_{\perp} = \prod_{N(H)}(c)$, where the first one is the component in $R(A)$ and second one is the component in $N(H)$.

Notice that $R(H)^{\perp} = N(H^{\mathrm{T}}) = N(H)$, since H is symmetric. By orthogonal decomposition lemma, there is an unique decomposition

$$c = c_{\parallel} + c_{\perp}$$

Now, let $x_\alpha = -\alpha c_{\perp}$, where $\alpha \in \mathbb{R}_+$.

$$\begin{aligned} F_0(x_\alpha) &= \frac{\alpha^2}{2} c_{\perp}^{\mathrm{T}} H c_{\perp} + c^{\mathrm{T}}(-\alpha c_{\perp}) + d \\ &= 0 - \alpha(c_{\parallel} + c_{\perp})^{\mathrm{T}} c_{\perp} + d \\ &= -\alpha(c_{\parallel}^{\mathrm{T}} c_{\perp} + c_{\perp}^{\mathrm{T}} c_{\perp}) + d \\ &= -\alpha \|c_{\perp}\|_2^2 + d \end{aligned}$$

Hence the function is unbounded below since we could take $\alpha \to \infty$. Therefore,

$$p^* = -\infty$$

Case B (ii) $H \in S_+^n$ but $H \notin S_{++}^n$ and $c \in R(H)$.

First, remind us of some results from previous chapter,

$$\begin{aligned} H &= \sum_{i=1}^{n} \lambda_i U^{(i)} U^{(i)^{\mathrm{T}}} \\ &= \sum_{i=1}^{r} \lambda_i U^{(i)} U^{(i)^{\mathrm{T}}} \\ &= U_r \Sigma Y_r^{\mathrm{T}} \end{aligned}$$

and,

$$\begin{aligned} H^{\frac{1}{2}} &= U_r \Sigma^{\frac{1}{2}} U_r^{\mathrm{T}} \\ H^{+} &= U_r \Sigma^{-1} U_r^{\mathrm{T}} \\ (H^{\frac{1}{2}})^{+} &= U_r \Sigma^{-\frac{1}{2}} U_r^{\mathrm{T}} \end{aligned}$$

where

$$\Sigma = \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_r} \end{bmatrix}, \; \Sigma^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{\lambda_r}} \end{bmatrix}$$

Observe that

$$C \in R(H) = R(H^{\frac{1}{2}}) = R(H^{+}) = R\left((H^{\frac{1}{2}})^{+}\right)$$

Since $C \in R(H)$ and $R(H) = R(H^{\frac{1}{2}})$, there exists $y \in \mathbb{R}^n$ such that

$$C = H^{\frac{1}{2}} y = H^{\frac{1}{2}}(y + \xi)$$

where $\xi \in N(H^{\frac{1}{2}})$. Note that choice of y is not unique and the second equality due to $\mathrm{rank}(H) = r < n$.

Explicitly, we have

$$C = (U_r \Sigma^{\frac{1}{2}} U_r^{\mathrm{T}}) y = \sum_{i=1}^{r} U^{(i)} \sqrt{\lambda_i} (U^{(i)^{\mathrm{T}}} y)$$

Now, the question is, which y we should pick? Let's pick the y with min-norm, that is,

$$\begin{aligned} y = \arg \min_{\bar{y} \in \mathbb{R}^n} \|\bar{y}\|_2 \\ s.t \; C = H^{\frac{1}{2}} \bar{y} \end{aligned}$$

It turns out that, this is an under determined and rank-deficient LS problem. So the solution to this question is given by

$$y = (H^{\frac{1}{2}})^{+} C = U_r \Sigma^{-\frac{1}{2}} U_r^{\mathrm{T}} C$$

So, when $C = H^{\frac{1}{2}} y$, we can pick $y = (H^{\frac{1}{2}})^{+} C$. Let's look back at the objective to understanding such choice.

$$\begin{aligned} \frac{1}{2} x^{\mathrm{T}} H x + c^{\mathrm{T}} x + d &= \frac{1}{2} x^{\mathrm{T}} H^{\frac{1}{2}} H^{\frac{1}{2}} x + y^{\mathrm{T}} H^{\frac{1}{2}} x + d \\ &= \frac{1}{2} \tilde{x}^{\mathrm{T}} \tilde{x} + y^{\mathrm{T}} \tilde{x} + d \\ &= \frac{1}{2} (\tilde{x}^{\mathrm{T}} x + 2 y^{\mathrm{T}} \tilde{x} + y^{\mathrm{T}} y) - \frac{1}{2} y^{\mathrm{T}} y + d \\ &= \frac{1}{2} \|\tilde{x} + y\|_2^2 - \frac{1}{2} \|y\|_2^2 + d \qquad (*) \end{aligned}$$

In the second equality we let $\tilde{x} = H^{\frac{1}{2}}x$.

Apparently, we would like to set $\tilde{x} = -y$ to minimize $(*)$, and the question is, is this always possible? The answer is yes.

Since $y = (H^{\frac{1}{2}})^{+}C \in R((H^{\frac{1}{2}})^{+}) = R(H) = R(H^{\frac{1}{2}})$, and $\tilde{x} = H^{\frac{1}{2}}x$ so $\tilde{x} \in R\left(H^{\frac{1}{2}}\right) = R(H)$. Since $x \in \mathbb{R}^n$ is unconstrained, so we can choose x to make $\tilde{x}$ equal to any element of $R(H)$, namely y.

Thus, to minimize $(*)$ we set $\tilde{x} = -y$ where $\tilde{x} = H^{\frac{1}{2}}x$, and yields

$$\begin{aligned}
F_0(x) \geq p^* &= d - \frac{1}{2}\|y\|_2^2 \\
&= d - \frac{1}{2}y^{\mathrm{T}}y \\
&= d - \frac{1}{2}C^{\mathrm{T}}(H^{\frac{1}{2}})^{+}(H^{\frac{1}{2}})^{+}C \\
&= d - \frac{1}{2}C^{\mathrm{T}}\left[U_r\right]\left[\Sigma^{-\frac{1}{2}}]\right]\left[U_r^{\mathrm{T}}\right]\left[U_r\right]\left[\Sigma^{-\frac{1}{2}}]\right]\left[U_r^{\mathrm{T}}\right]\left[C\right] \\
&= d - \frac{1}{2}C^{\mathrm{T}}(U_r\Sigma^{-1}U_r^{\mathrm{T}})C \\
&= d - \frac{1}{2}C^{\mathrm{T}}H^{+}C
\end{aligned}$$

What is the minimizing choice of $x \in \mathbb{R}^n$?

Since $y \in R(H) = R(H^{\frac{1}{2}})$ and $x \in \mathbb{R}^n$ is free, we can satisfy equality and thus there are many solutions.

We would like to pick min-norm solution for x, which is given by

$$\begin{aligned}
x^* &= (H^{\frac{1}{2}})^{+}\tilde{x} \\
&= -(H^{\frac{1}{2}})^{+}y \\
&= -(H^{\frac{1}{2}})^{+}(H^{\frac{1}{2}})^{+}C \\
&= -H^{+}C
\end{aligned}$$

Case C: $H \in S_{++}^n$ (H is PD)

This is indeed a special case of case B(ii), and that's because

(1)$H \in S_{++}^n$, then it is also in S_{+}^n, and so all eigenvalues are non negative.

(2)$H \in S_{++}^n$, then $\operatorname{rank}(H) = n$ and thus $R(H) = \mathbb{R}^n$, and certainly $C \in R(H)$.

From previous solution, we have

$$\begin{aligned}
x &= -H^{+}C \\
&= -U_r\Sigma^{-1}U_r^{\mathrm{T}}C \\
&= -U_r\Sigma^{-1}U_n^{\mathrm{T}}C \\
&= -H^{-1}C
\end{aligned}$$

So in this case $H^{+} = H^{-1}$. The optimal value is given by

$$p^* = d - \frac{1}{2}C^{\mathrm{T}}H^{-1}C$$

Summary

Consider the QP problem $p^* = min_{x \in \mathbb{R}^n} \frac{1}{2} x^T H x + c^T x + d$,

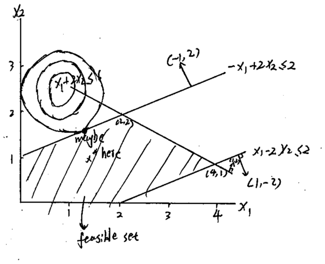
$$p^* = \begin{cases} d - \frac{1}{2} c^T H^+ c & H \text{ is PD, or, } H \text{ is PSD but not PD and } c \in R(H) \\ -\infty & H \text{ is not PSD, or, } H \text{ is PSD but not PD and } c \notin R(H) \end{cases}$$

$$x^* = \begin{cases} -H^+ c & H \text{ is PD, or, } H \text{ is PSD but not PD and } c \in R(H) \\ \text{not exist} & H \text{ is not PSD, or, } H \text{ is PSD but not PD and } c \notin R(H) \end{cases}$$

Solving QPs via "active set" methods

$$\min_{x \in \mathbb{R}^2} \quad (x_1 - 1)^2 + (x_2 - 2.5)^2$$

$$s.t. \quad \begin{bmatrix} -1 & 2 \\ 1 & 2 \\ 1 & -2 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} 2 \\ 6 \\ 2 \\ 0 \\ 0 \end{bmatrix}$$



- at optimum some subset of inequality constraints satisfied with equality "active" set
- Best "loose"
- If you know the active set or optimum, just solve an "equality"-constrained QP"

By illustrative example:

Step 1

Initialize at point $x^{(0)} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$, starting with initial "working set" of equality constraint(the fifth equality constraint):

$$w_0 = \{x | x_2 = 0\}$$

Step 2

$$\min_{x \in \mathbb{R}^2} \quad \frac{1}{2} x^T \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} x + \begin{bmatrix} -2 & -5 \end{bmatrix} + (1 + 2.5^2)$$

$$s.t. \quad \begin{bmatrix} 0 & 1 \end{bmatrix} x = 0$$

Recall the linearly constrained QP problem, we need a basis for $N(\begin{bmatrix} 0 \\ 1 \end{bmatrix}) = \{x | x = \alpha \begin{bmatrix} 1 \\ 0 \end{bmatrix}\}$ in this case. Let $N = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$,

$$\begin{aligned}\tilde{H} &= N^{\mathrm{T}}HN = 2\\ \tilde{C}^{\mathrm{T}} &= (c^{\mathrm{T}}N + \bar{x}^{\mathrm{T}}HN) = -2\\ \tilde{d} &= (d + c^{\mathrm{T}}x + \frac{1}{2}\bar{x}^{\mathrm{T}}H\bar{x}) = 0\end{aligned}$$

So we have converted the problem to :

$$\begin{aligned}z^* &= \arg\min_{z\in\mathbb{R}} \frac{1}{2}z^{\mathrm{T}}\tilde{H}z + \tilde{c}^{\mathrm{T}}z + \tilde{d}\\ &= \arg\min_{z\in\mathbb{R}} \frac{1}{2}2z^2 + (-2z) + 0\\ &= \arg\min_{z\in\mathbb{R}} z^2 - 2z\end{aligned}$$

Take the first derivative and set it to zero, we get $z^* = 1$.

So the optimum is:

$$x^{(1)} = \bar{x} + z^*N = \begin{bmatrix}0\\0\end{bmatrix} + 1\begin{bmatrix}1\\0\end{bmatrix} = \begin{bmatrix}1\\0\end{bmatrix}$$

Step 3

Note:

(1) At this point we recognize that the equality constraint $\{x|x_2 = 0\}$ is no longer "binding" because we just optimized along that set.

(2) We could drop the fifth constraint from "working set" left with $w_2 = \emptyset$

(3) We are facing with an unconstrained optimization problem. Clearly we would like to move to $x^{(1)} + \Delta = \begin{bmatrix}1\\0\end{bmatrix} + \begin{bmatrix}0\\2.5\end{bmatrix}$.

(4) But there may be a "blocking" constraint in way. In this example, this is the first constraint $\{x| - x_1 + 2x_2 \leq 2\}$

(5) Instead, we solve for the a step size γ to make the constraint light, i.e., $g^{(i)T}(x^{(1)} + \gamma\Delta) = h_1$.

So we have,

$$[-1\ 2]\begin{bmatrix}1\\0\end{bmatrix} + \gamma\begin{bmatrix}0\\2.5\end{bmatrix} = 2 \Leftrightarrow \gamma = \frac{3}{5}$$

$$x^{(2)} = x^{(1)} + \frac{3}{5}\Delta = \begin{bmatrix}1\\1.5\end{bmatrix}$$

Step 4

This time, $w_3 = \{x| \begin{bmatrix}-1 & 2\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} = 2\}$.

$$N(\begin{bmatrix}-1 & 2\end{bmatrix}) = \{x|x = \alpha\begin{bmatrix}2\\1\end{bmatrix}\},\quad \bar{x} \in \{x| \begin{bmatrix}-1 & 2\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} = 2\}$$

We pick $N = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \bar{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$

$$\begin{aligned} \tilde{H} &= N^{\mathrm{T}} H N = 10 \\ \tilde{C}^{\mathrm{T}} &= (c^{\mathrm{T}} N + \bar{x}^{\mathrm{T}} H N) = -7 \\ \tilde{d} &= (d + c^{\mathrm{T}} x + \frac{1}{2} \bar{x}^{\mathrm{T}} H \bar{x}) = -4 \end{aligned}$$

$$\begin{aligned} z^* &= \arg\min_{z \in \mathbb{R}} \frac{1}{2} z^{\mathrm{T}} \tilde{H} z + \tilde{c}^{\mathrm{T}} z + \tilde{d} \\ &= \arg\min_{z \in \mathbb{R}} \frac{1}{2} 10 z^2 + (-7z) - 4 \\ &= \arg\min_{z \in \mathbb{R}} 5z^2 - 7z - 4 \end{aligned}$$

Take the first derivative and set it to zero, we get $z^* = \frac{7}{10}$.

$$x^{(3)} = \bar{x} + z^* N = \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \frac{7}{10} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.4 \\ 1.7 \end{bmatrix}$$

Note $x^{(3)}$ is the global optimum so we end this algorithm here.

Quadratically Constrained Quadratic Program (QCQP)

Let's formulate such kind of problem first

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2} x^{\mathrm{T}} H_0 x + c_0^{\mathrm{T}} x + d_0 & \\ s.t. \quad & \frac{1}{2} x^{\mathrm{T}} H_i x + c_i^{\mathrm{T}} x + d_i \leq 0 & i \in [m] \\ & \frac{1}{2} x^{\mathrm{T}} \tilde{H}_i x + \tilde{c}_i^{\mathrm{T}} x + \tilde{d}_i = 0 & i \in [q] \end{aligned}$$

Note:

- If $H_i = 0, \forall i \in [n]$, $\tilde{H}_i = 0, \forall i \in [q]$, then we have a *QP*.
- Typically $H_0 \geq 0$, $H_i \geq 0$, $i \in [m]$, and $\tilde{H}_i = 0$, $\forall i \in [q]$, in which case the problem is a convex optimization problem, and thus it is easy to solve as we will discuss in next chapter.

To see that why $\tilde{H}_i \neq 0$ makes things difficult, let's consider a single equality constraint $q = 1$, and a scalar problem $x \in \mathbb{R}$:

$$\tilde{H}_1 = 1 \quad \tilde{c}_i = 0 \quad \tilde{d}_i = -\frac{1}{2}$$

So the equality constraint becomes:

$$\frac{1}{2} x_1^2 + 0 - \frac{1}{2} = 0 \Leftrightarrow x_1^2 = 1 \Leftrightarrow x_1 \in \{-1, 1\}$$

Notice that the feasible set is not a continuum of possibilities, i.e., it is a distinct set so the problem is "Combinatorial".

Example 7.11. Let's consider the QCQP with only three inequality constraints ($m = 3$).

Constraint 1:

$$H_1 = \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \qquad c_1^{\mathrm{T}} = \begin{bmatrix} 0 & 0 \end{bmatrix} \qquad d_1 = -1$$

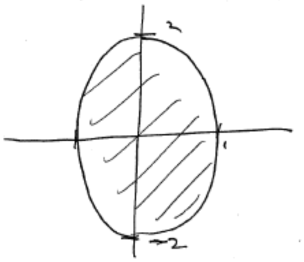


Figure 7.4: Feasible set of constraint 1

So this constraint can be written as

$$\begin{aligned} \frac{1}{2}x^{\mathrm{T}}H_1x + c_1^{\mathrm{T}}x + d_1 \leq 0 &\Leftrightarrow 2x_1^2 + \frac{1}{2}x_2^2 \leq 2 \\ &\Leftrightarrow \frac{x_1^2}{1} + \frac{x_2^2}{4} \leq 1 \end{aligned}$$

It is obviously the feasible set of this constraint corresponds to an ellipse, and by computing the eigenvalues(lengths of major and minor axis) and eigenvectors(directions of major and minor axis) of H_1, we are able to draw such feasible set as on the r.h.s.

Constraint 2:

$$H_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad c_2^{\mathrm{T}} = \begin{bmatrix} -1 & 1 \end{bmatrix} \qquad d_2 = -1$$

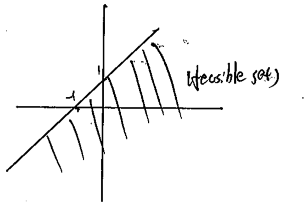


Figure 7.5: Feasible set of constraint 2

So this constraint can be written as

$$\begin{bmatrix} -1 & 1 \end{bmatrix} x \leq 1 \Leftrightarrow \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq 1 \Leftrightarrow x_2 \leq 1 + x_1$$

The corresponding feasible set of this constraint is illustrated on r.h.s.

Constraint 3:

$$H_3 = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \qquad c_3^{\mathrm{T}} = \begin{bmatrix} -1 & 0 \end{bmatrix} \qquad d_3 = -1$$

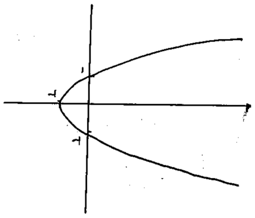


Figure 7.6: Feasible set of constraint 3

So this constraint can be written as

$$x_2^2 - x_1 - 1 \leq 0$$

The corresponding feasible set of this constraint is illustrated on r.h.s.

Put all these 3 constraints together(i.e., find the intersection of above 3 feasible sets), so we can obtain the desired feasible set for this QCQP problem

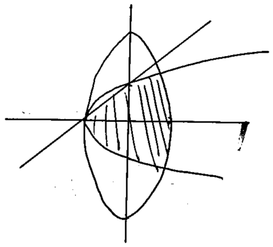


Figure 7.7: Feasible set for this QCQP

8
Convex sets and functions

8.1 Linear/affine/convex/conic hulls & sets

Given a set of points $x^{(i)} \in \mathbb{R}^n$, $i \in [m]$

$$P = \{x^{(1)}, x^{(2)}, ..., x^{(m)}\}$$

Consider combinations of form $\sum_{i=1}^{m} \lambda_i x^{(i)}$,
1) The "linear" hull:

$$\{x | x = \sum_{i=1}^{m} \lambda_i x^{(i)}, \lambda_i \in \mathbb{R}, \forall i \in [m]\}$$

2) The "affine" hull:

$$\{x | x = \sum_{i=1}^{m} \lambda_i x^{(i)}, \lambda_i \in \mathbb{R}, \sum_{i=1}^{n} \lambda_i = 1\}$$

3) The "convex" hull:

$$\{x | x = \sum_{i=1}^{m} \lambda_i x^{(i)}, \lambda_i \in \mathbb{R}, \lambda_i \geq 0, \sum_{i=1}^{m} \lambda_i = 1\}$$

4) The "conic" hull:

$$\{x | x = \sum_{i=1}^{m} \lambda_i x^{(i)}, \lambda_i \in \mathbb{R}, \lambda_i \geq 0\}$$

Summary

	$\lambda_i \geq 0$	$\sum_{i=1}^{m} \lambda_i = 1$
Linear	no	no
Affine	no	yes
Covex	yes	yes
Conic	yes	no

Example 8.1 (Linear Hull). Let $P = \{x^{(1)}, x^{(2)}\}$, linear hull of $P =$ span$\{x^{(1)}, \cdots, x^{(m)}\} =$ span(P).

Note that linear hull of P forms the smallest subspace that contains P.

Example 8.2 (Affine Hull)**.** Let $P = \{x^{(1)}, x^{(2)}\}$, the point of the affine hull is given by

$$\begin{aligned} x &= \lambda_1 x^{(1)} + \lambda_2 x^{(2)} \\ &= \lambda_1 x^{(1)} + (1-\lambda)_1 x^{(1)} \\ &= x^{(2)} + \lambda(x^{(1)} - x^{(2)}) \end{aligned}$$

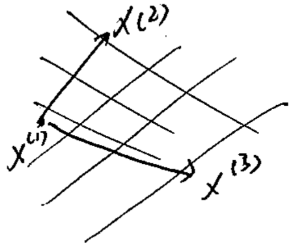
Hence, $\text{aff}(P) = x^{(2)} + \text{span}(x^{(1)} - x^{(2)})$.

Let $P = \{x^{(1)}, x^{(2)}, x^{(3)}\}$, the point of the affine hull is given by

$$\begin{aligned} x &= \lambda_1 x^{(1)} + \lambda_2 x^{(2)} + \lambda_3 x^{(3)} \\ &= (1 - \lambda_2 - \lambda_3) x^{(1)} + \lambda_2 x^{(2)} + \lambda_3 x^{(3)} \\ &= x^{(1)} + \lambda_2 (x^{(2)} - x^{(1)}) + \lambda_3 (x^{(3)} - x^{(1)}) \end{aligned}$$

Hence, $\text{aff}(P) = x^{(1)} + \text{span}(x^{(2)} - x^{(1)}) + \text{span}(x^{(3)} - x^{(1)})$.

Note that, the affine hull is the smallest affine set contains the set P.

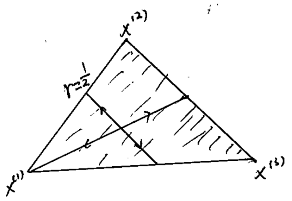


Example 8.3 (Convex hull)**.** Let $P = \{x^{(1)}, x^{(2)}\}$, the point of convex hull is given by

$$\begin{aligned} x &= \lambda_1 x^{(1)} + \lambda_2 x^{(2)} \\ &= (1-\lambda) x^{(1)} + \lambda x^{(2)} \\ &= x^{(1)} + \lambda(x^{(2)} - x^{(1)}) \end{aligned}$$

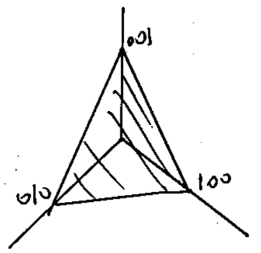
Let $P = \{x^{(1)}, x^{(2)}, x^{(3)}\}$, the point of convex hull is given by

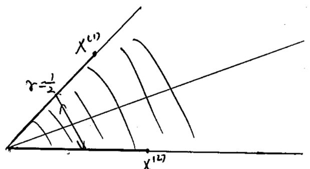
$$\begin{aligned} x &= \lambda_1 x^{(1)} + \lambda_2 x^{(2)} + \lambda_3 x^{(3)} \\ &= x^{(1)} + \lambda_2 (x^{(2)} - x^{(1)}) + \lambda_3 (x^{(3)} - x^{(1)}) \\ &= x^{(1)} + \lambda\gamma (x^{(2)} - x^{(1)}) + (1-\lambda)\gamma (x^{(3)} - x^{(1)}) \end{aligned}$$



Example 8.4 (Conic hull)**.** Let $P = \{x^{(1)}, x^{(2)}\}$, the point of conic hull is given by

$$\begin{aligned} x &= \lambda_1 x^{(1)} + \lambda_2 x^{(2)} \\ &= (\lambda_1 + \lambda_2)[\frac{\lambda_1}{\lambda_1 + \lambda_2} x^{(1)} + \frac{\lambda_2}{\lambda_1 + \lambda_2} x^{(2)}] \\ &= \gamma[\lambda x^{(1)} + (1-\lambda) x^{(2)}] \end{aligned}$$





Convex Sets

Definition 8.5 (Convex set)**.** A subset $C \subseteq \mathbb{R}^n$ is a convex set if $\forall x, y \in \mathcal{C}$, then $z \in \mathcal{C}$, $\forall z = \lambda x + (1-\lambda)y$, $\lambda \in [0,1]$.

Definition 8.6 (Strictly Convex)**.** A convex set is strictly convex if $\forall x, y \in \mathcal{C}$, $\forall \lambda \in (0,1)$, $z = \lambda x + (1-\lambda)y \in rel\,int(\mathcal{C})$(relative interior)

Objects with straight edges are not strictly convex sets.

Definition 8.7 (Cone)**.** A set $\mathcal{C} \subseteq \mathbb{R}^n$ is a cone if $\forall x \in \mathcal{C}$, then $\gamma x \in \mathcal{C}, \forall \gamma \geq 0$.

Typical convex sets

1) Hyper-planes are convex.

Proof. Consider the hyper-plane defined as $H = \{x | a^T x = b\}$, we pick arbitrary $x, y \in H$, and show that $z = \lambda x + (1-\lambda)y \in H$ $\forall \lambda \in [0,1]$.

$$\begin{aligned} a^T z &= a^T(\lambda x + (1-\lambda)y) \\ &= \lambda(a^T x) + (1-\lambda)y \\ &= \lambda b + (1-\lambda)b \\ &= b \end{aligned}$$

□

2) Half-spaces are convex.

Proof. Consider the half-space defined as $\{x | a^T x \leq b\}$, we use the similar proof of the hyper-planes case, except we replace the $'' = ''$ with $'' \leq ''$ as follows

$$\begin{aligned} a^T z &= a^T(\lambda x + (1-\lambda)y) \\ &= \lambda(a^T x) + (1-\lambda)y \\ &\leq \lambda b + (1-\lambda)b \\ &= b \end{aligned}$$

Thus, $a^T z \leq b$, the points $z = \lambda x + (1-\lambda)y$ form a convex set. So the half-space is convex. □

3) If $C_1, \cdots, C_n$ are convex sets, then the set $\mathcal{C} = \cap_{i=1}^m C_i$ is convex.

Proof. First we pick any $x, y \in \mathcal{C}$, therefore we have $x, y \in \mathcal{C}_i, \forall i \in [m]$, and we want to show that $z = \lambda x + (1-\lambda)y$ is in the set $\mathcal{C}$.

Note that $x, y \in \mathcal{C}_i \forall i \in [m]$ implies that $z \in \mathcal{C}_i\ \forall i \in [m]$, since $\mathcal{C}_i$ is a convex set $\forall i \in [m]$,

Therefore, $z \in \cap_{i=1}^m C_i = C$, the set $\mathcal{C}$ is a convex set. □

Example 8.8. Recall that in previous LP and QP problems, the feasible set

$$\{x|Ax = b\} \cap \{x|Gx \leq b\} = \{\cap_{i=1}^{q}\{x|a^{(i)^T}x = b_i\} \cap \{\cap_{i=1}^{m}\{x|g^{(i)^T}x \leq h_i\}$$

is the intersection of $m + q$ convex sets, so it is a convex set.

4) Affine transformations preserve the convexity of a set.

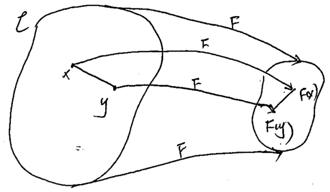
If a map $F : \mathbb{R}^n \to \mathbb{R}^m$ is affine (i.e., $F(x) = Ax + b$), and a set $\mathcal{C} \subseteq \mathbb{R}^n$ is convex, then the image of $\mathcal{C}$ under F is convex.

$$F(\mathcal{C}) = \{F(x)|x \in \mathcal{C}\} \subseteq \mathbb{R}^m$$

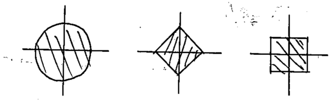
Conversely, the pre-image of a convex set $\tilde{e}$ in $\mathbb{R}^m$ is also convex

$$\{x|F(x) \in \mathcal{C}\} \subseteq \mathbb{R}^n$$

.



5) Norm balls are convex (recall that a norm is defined as $\|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$ for $p \geq 1$).



Proof. Take any two points u, v s.t. $\|u\| \leq 1$, $\|v\| \leq 1$, by utilizing the triangular inequality and scaling property of a norm, we show that

$$\begin{aligned}
\|\lambda u + (1-\lambda)v\| &\leq \|\lambda u\| + \|(1-\lambda)v\| \\
&= |\lambda|\|u\| + |(1-\lambda)|\|v\| \\
&= \lambda\|u\| + (1-\lambda)\|v\| \\
&\leq \lambda 1 + (1-\lambda)1 \\
&= 1
\end{aligned}$$

□

Example 8.9. The ellipsoids defined as

$$\xi(x_c, P) = \{x|(x - x_c)^T P^{-1}(x - x_c) \leq 1\}$$

is a convex set, where $P \in S_{++}^n$.

Proof. First recall that l_2 norm ball is convex, and consider following affine map

$$F(u) = P^{\frac{1}{2}}u + x_c$$

Therefore the set $\{F(u)|\|u\|_2 \leq 1\}$ is convex. We show that this set is equivalent to a ellipsoid,

$$\begin{aligned}
\{F(u)|\|u\|_2^2 \leq 1\} &= \{x|x = P^{\frac{1}{2}}u + x_c, \|u\|^2 \leq 1\} \\
&= \{x|x - x_c = P^{\frac{1}{2}}u, \|u\|^2 \leq 1\} \\
&= \{x|P^{-\frac{1}{2}}(x - x_c) = u, \|u\|^2 \leq 1\} \\
&= \{x|\|P^{-\frac{1}{2}}(x - x_c)\|_2^2 \leq 1\} \\
&= \{x|(x - x_c)^T P^{-1}(x - x_c) \leq 1\}
\end{aligned}$$

So the set $\xi(x_c, P) = \{x | (x - x_c)^T P^{-1} (x - x_c) \leq 1\}$ is a convex set.

Also, remind that in previous QP chapter, we intersect these shapes with polyhedron to get the feasible set of QCQP. □

Example 8.10. Consider the set

$$\{x | \|Ax - b\|_2^2 \leq 1\} = \{x | \|F(x)\|^2 \leq 1\}$$

where $F(x) = Ax - b$.

This set is the pre-image(inverse image) of a convex set(norm ball is a convex set) under an affine function, and so it's convex.

Cones and generalized inequalities

We introduce some important cones here and first recall the following definitions for set of symmetric matrices, set of PSD matrices and set of PD matrices,

$$\begin{aligned} S^n &= \{x \in \mathbb{R}^{n \times n} \ s.t. \ x = x^T\} \\ S^n_+ &= \{x \in S^n \ s.t. \ v^T X v \geq 0, \forall v \in \mathbb{R}^n\} \\ S^n_{++} &= \{x \in S^n \ s.t. \ v^T X v > 0, \forall v \in \mathbb{R}^n\} \end{aligned}$$

Sets of PSD and PD matrices are two class of important cones, and recall the definition of cone: Set $\mathcal{C}$ is a cone if $\forall x \in \mathcal{C}$ and $\theta \in \mathbb{R}_+$(i,e. $\theta \geq 0$), $\theta x \in \mathcal{C}$. In particular,

1) S^n_+ is a cone.

Since $\forall X \in S^n_+$ and $\forall \theta \geq 0$, we have

$$v^T(\theta X)v = \theta v^T X v > 0$$

We write:

$$\begin{aligned} X \in S^n_+ &\Leftrightarrow X \geq 0 \\ X \in S^n_{++} &\Leftrightarrow X > 0 \end{aligned}$$

2) S^n_+ is a convex cone.

Let $A \in S^n_+$, $B \in S^n_+$, and consider the combination $\lambda A + (1 - \lambda)B$ where $\lambda \in [0, 1]$.

First note that $(\lambda A + (1 - \lambda)B)^T = \lambda A^T + (1 - \lambda)B^T = \lambda A + (1 - \lambda)B \in S^n$, so this combination is still a symmetric matrix.

Secondly, we show that $v^T(\lambda A + (1 - \lambda B))v = \lambda(v^T A v) + (1 - \lambda)(v^T B v) \geq 0$, so this combination is still a PSD matrix.

Therefore, S^n_+ is convex and thus it is a convex cone.

3) S^n_{++} is also a convex cone.

The proof follows the same as previous case (2) except we replace $\geq$ with $>$.

Cones lead to "generalized" inequalities.

We want to extend idea of orderings to $\mathbb{R}^n$ (i.e., extend the comparison between two real numbers to two real vectors/matrices), and let's start with a "proper" cone $K \in \mathbb{R}^n$.

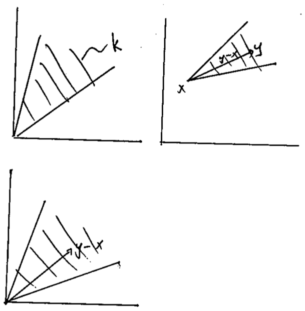
Definition 8.11 (Proper cone)**.** A cone $K \in \mathbb{R}^n$ is called a proper cone if it satisfies the following

- K is convex.
- K is closed.
- K is solid, which means it has nonempty interior.
- K is pointed, which means that it contains no line (or equivalently, $x \in K, -x \in K \Rightarrow x = 0$).

A proper cone K can be used to define a generalized inequality $\leq_K$, says "less-than-or-equal to with respect to cone K".

Interpretation of $\leq_K$ and $< K$:

$$
\begin{aligned}
x \leq_K y \Leftrightarrow 0 \leq_K (y - x) \Leftrightarrow y - x \in K \\
x <_K \Leftrightarrow y - x \in \text{int}(K)
\end{aligned}
$$



where the set $\text{int}(K)$ denotes the points in the interior of K.

Example 8.12. Let a proper cone $k = S^n_+$, which denote the ordering of matrices, whose elements of vector space are in S^n. So,

$$
X \leq_k Y \Leftrightarrow 0 \leq_k Y - X.
$$

Thus, $Y - X \in S^n_+$.

$\rightarrow$ It true since that $v^T(Y - X)v \geq 0, \forall v \in \mathbb{R}^n$

$\rightarrow$ All eigenvalues are non-negative.

Note: The interior of S^n_+ is S^n_{++}.

The following 2 generalized inequalities come up so often, so we assume that they are the default setting.

1. If we compare 2 vectors $x, y \in \mathbb{R}^n$, we write

$$
x \leq y \Leftrightarrow x \leq_{\mathbb{R}^n_+} y \Leftrightarrow y - x \in \mathbb{R}^n_+
$$

2. If we compare 2 symmetric matrices, we write:

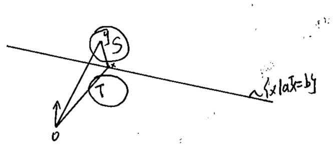
$$
\begin{aligned}
x \leq y \Leftrightarrow x \leq_{S^n_+} y \Leftrightarrow y - x \in S^n_+ \\
x < y \Leftrightarrow y - x \in S^n_{++}
\end{aligned}
$$

Example 8.13. Consider the set $\{x \in \mathbb{R}^n | x_1 A_1 + x_2 A_2 + ... + x_n A_n \leq B\}$, where $A_i \in S^m$ $B \in S^m$. The inequity here is called the "linear matrix inequality", and notice that $F(x) = B - \sum_{i=1}^{n} x_i A_i$ is an affine function of x.

Hence,

$$\begin{aligned} \{x \in \mathbb{R}^n | x_1 A_1 + x_2 A_2 + ... + x_n A_n \leq B\} &= \{x \in \mathbb{R}^n | F(x) \geq 0\} \\ &= \{x \in \mathbb{R}^n | F(x) \in S_+^n\} \\ &= F^{-1}(S_+^n) \end{aligned}$$

So, it is a convex set, since the set is the pre-images of a convex set S_+^n under an affine transform.



Properties of Convex Sets:

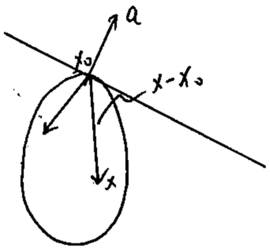
1. Separating hyperplane:

   If S, T are convex sets in $\mathbb{R}^n$ and disjoint, i.e, $S \cap T = \emptyset$, then there exists an $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ s.t.

$$\begin{aligned} a^T x \geq b, \forall x \in S \\ a^T x < b, \forall x \in T \end{aligned}$$

   Hence,

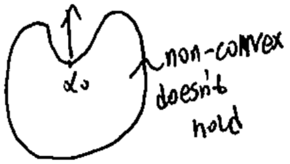
$$a^T y - a^T x = a^T (y - x) \geq 0$$



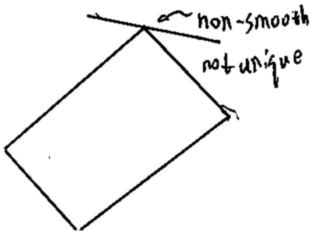
2. Supporting hyperplanes:

   If S is a convex set, then $\forall x_0 \in \delta\mathcal{S}$ (boundary of S) and $\forall x \in S$, $\exists a \in \mathbb{R}^n$ such that

$$a^T x \leq a^T x_0 \Leftrightarrow a^T (x - x_0) \leq 0$$



Convex Functions



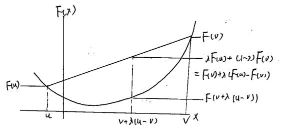
Let F have a convex domain, then $F : \mathbb{R}^n \to \mathbb{R}$ is a convex function if $\forall x, y \in \text{dom}(F)$:

$$F(\lambda x + (1 - \lambda) y) \leq \lambda F(x) + (1 - \lambda) F(y), \ \forall \lambda \in [0, 1]$$

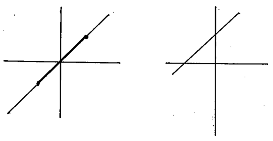
and F is strictly convex if

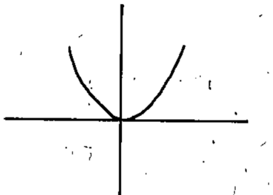
$$F(\lambda x + (1 - \lambda) y) < \lambda F(x) + (1 - \lambda) F(y), \forall \ \lambda \in (0, 1)$$

Note: F is a "concave" function if $-F$ is convex.
Definition of convex:



$$F(\lambda u + (1 - \lambda) v) \leq \lambda F(u) + (1 - \lambda) F(v)$$





And $F(\lambda u + (1-\lambda)v)$ can be written as $F(v + \lambda(u-v))$:

$$F(\lambda u + (1-\lambda)v) \leq \lambda F(v) + \lambda(F(u) - F(v))$$

line segment connecting $(u, F(u))$ to $(v, F(v))$ always above bottom of bowl.

Sometimes define an "extended value" function

$$\tilde{F}(x) = \begin{cases} F(x) & \text{if } x \in dom(F) \\ \infty & else \end{cases}$$

Example 8.14 (Examples of convex/concave functions)**.** Refer to the figures on r.h.s.

(1) Linear function and affine functions are both convex and concave.

(2) $F(x) = x^2$ is convex.

(3) $F(x) = \log x$ with dom $F = \mathbb{R}_{++}$ is a concave function.

(4) The norm function $\|x\|$ is convex.

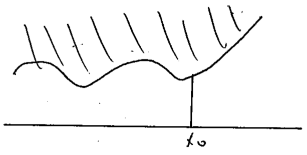
Since we have

$$\begin{aligned} \|\lambda x + (1-\lambda)y\| &\leq \|\lambda x\| + \|(1-\lambda)y\| \\ &= \lambda\|x\| + (1-\lambda)\|y\| \end{aligned}$$

(5) $F(x) = \frac{1}{x}$ is convex on $\mathbb{R}_{++}$, and is concave on $\mathbb{R}_{--}$.

Definition 8.15 (Epigraph)**.** Recall the epigraph of a function is a set of points lying on or above its graph:

$$\text{epi } F = \{(x,t) | t \geq F(x), x \in \text{dom } F, t \in \mathbb{R}\}$$

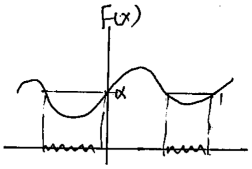


Definition 8.16. F is a convex function iff epi F is a convex set

Definition 8.17 (Sublevel sets)**.** Recall that, the α-sublevel set of a function F is defined as

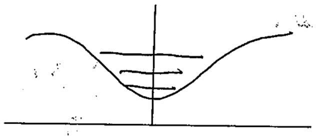
$$C(\alpha) = \{x | F(x) \leq \alpha, x \in domF\}$$

for $\alpha \in \mathbb{R}$.



Theorem 8.18. *If F is convex, then its sub-level sets are all convex sets.*

Note: The converse of this theorem is not true. If all sub-level sets of a function are convex sets, the function is "quasi-convex" but not necessarily convex.



Three kinds of convex functions:

1) Non-negative sums of convex functions are convex.

Let $F(x) = \sum_{i=1}^{m} a_i F_i(x)$, F_i are convex function and dom $F = \cap_{i=1}^{m}$dom F_i.

We show that such F is a convex function,

$$\begin{aligned}
F(\lambda x + (1-\lambda)y) &= \sum_{i=1}^{m} a_i F_i(\lambda x + (1-\lambda)y) \\
&\leq \sum_{i=1}^{m} a_i[\lambda F_i(x) + (1-\lambda)F_i(y)] \\
&= \lambda[\sum_{i=1}^{m} a_i F_i(\lambda)] + (1-\lambda)[\sum_{i=1}^{m} a_i F_i(y)] \\
&= \lambda[F(x)] + (1-\lambda)[F(y)]
\end{aligned}$$

2) Convex functions of affine transformations of variables is still a convex function.

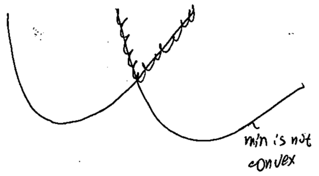
Let $g(x) = F(Ax+b)$, where $F(\cdot)$ is convex, and notice that dom $g = \{x | Ax + b \in \text{dom } F\}$.

We show that function g is convex in x,

$$\begin{aligned}
g(\lambda x + (1-\lambda)y) &= F(A(\lambda x + (1-\lambda)y) + b) \\
&= F(\lambda(Ax+b) + (1-\lambda)(Ay+b)) \\
&\leq \lambda F(Ax+b) + (1-\lambda)F(Ay+b) \\
&= \lambda g(x) + (1-\lambda)g(y)
\end{aligned}$$

3) The max of a pair of convex functions is a convex function:

$$g(x) = \max\{F_1(x), F_2(x)\}$$



More examples

Let's consider following three kinds of functions we have mentioned above,

$$\begin{aligned}
F(x) &= \sum \alpha_i F_i(x), \forall \alpha_i \geq 0 \\
g(x) &= F(Ax+b) \\
g(x) &= max\{F_1(x), F_2(x)\}
\end{aligned}$$

Example 8.19. Consider the function

$$\begin{aligned}
F(x) &= \sum_{i=1}^{n} log(b_i - a_i^T x)^{-1} \\
&= \sum_{i=1}^{n} -log(b_i - a_i^T x)
\end{aligned}$$

where $x \in \mathbb{R}^n$, $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$.

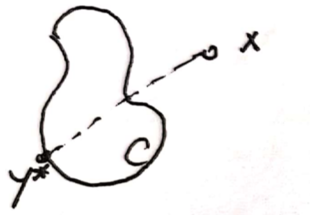
1) Note: $-log(\cdot)$ is a convex function, and

$$\text{dom} - log(\cdot) = R_{++}$$

$$\text{dom}F = \{x | b_i - a_i^T x > 0, \forall i \in [m]\} = \{x | b_i - a_i^T x \in \mathbb{R}_{++}, \forall i \in [m]\}$$

So, the domain of this function is the inverse image of $\mathbb{R}_{++}$(a convex set) under an affine transformation, and therefore it is a convex set.

2) Each function is a convex function of an affine transformation of x, therefore the sum of these functions is still a convex function.



Example 8.20.

$$F(x) = \sup_{y \in \mathcal{C}} \|y - x\|$$

Note that $\mathcal{C}$ is not necessarily to be a convex set.

1. Function $y - x$ is an affine function w.r.t x and therefore it is a convex function of x.

2. $\|\cdot\|$ is a norm function, so it is a convex function of its argument.

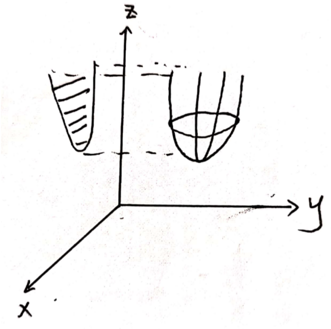
3. $F(x) = \sup_{y \in \mathcal{C}} \|y - x\|$, $y \in \mathcal{C}$ is the basic max of a bunch of convex functions, each indexed by a $y \in \mathcal{C}$

Example 8.21.

$$F(x) = \inf_{y \in \mathcal{C}} \|y - x\|$$

→ Generally this function is NOT convex in x.

→ If the set $\mathcal{C}$ is a convex set, then this function is convex in x.



Theorem 8.22 (Projection theorem). *If $h(x, y)$ is convex in (x, y), then $F(x) = \inf_y h(x, y)$ is convex in x.*

Idea: Shine light along y-axis, and get the shadow on $x - z$ plane, which is an epi F and it is a convex set.

Proof. Since $h(x, y)$ is convex in $\begin{bmatrix} x \\ y \end{bmatrix} \in \text{dom } h$, the epigraph of h is given by

$$\text{epi } h = \{(x, y, t) | t \geq h(x, y), \begin{bmatrix} x \\ y \end{bmatrix} \in \text{dom } h\}$$

That's the black bowl in the graph.

Now consider:

$$F(x) = \inf_{y : \begin{bmatrix} x & y \end{bmatrix}^T \in \text{dom } h} h(x, y)$$

So the domain is given by

$$\begin{aligned}\text{dom } F &= \{x | \exists\ y\ s.t. (x, y) \in \text{dom } h\} \\ &= \{ \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \mid \begin{bmatrix} x \\ y \end{bmatrix} \in \text{dom } h\}\end{aligned}$$

Note that this domain is an affine map of all points in a convex set, and therefore dom F is a convex set.

Consider the epigraph of F,

$$\begin{aligned}\text{epi } F &= \{(x, t) | t \geq \inf_{y: \begin{bmatrix} x & y \end{bmatrix}^T} \in \text{dom } h, x \in \text{dom } F\} \\ &= \{ \begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ t \end{bmatrix} | t \geq h(x, y), \begin{bmatrix} x \\ y \end{bmatrix} \in \text{dom } h\}\end{aligned}$$

So this set is a convex set, and since it is the epigraph of F, F must be a convex function.

□

Example 8.23. The function

$$F(x) = \inf_{y \in \mathcal{C}} \|x - y\|$$

is a convex function if $\mathcal{C}$ is a convex set.

1. $x - y$ is affine in x
2. $\| \cdot \|$ is a convex function for all norms.
3. Apply projection theorem where dom $h = \{ \begin{bmatrix} x \\ y \end{bmatrix} | x \in \mathbb{R}^n, y \in \mathcal{C}\}$ is a convex set and x is unconstrained.

Characterizing Convexity by Restricting to a Line

Theorem 8.24. *$F : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if $g : \mathbb{R} \to \mathbb{R}$,*

$$g(t) = F(x_0 + tv), dom\ g = \{t | x_0 + tv \in dom\ F\}$$

is convex for any $x_0 \in dom\ F$, $v \in \mathbb{R}^n$.

- $g(t)$ is function restricted to a line/slice
- If all possible slices convex then so is F.

Note: need $x_0 + tv \in$ dom F, also note dom F is a convex set.

$$g(t) = g_{x_0, v}(t) = F(x_0 + tv)$$

$$\mathrm{dom}(g_{x_0,v}) = \{t | x_0 + tv \in \mathrm{dom} F\}$$

Therefore $\mathrm{dom}(g_{x_0,v})$ is convex set for all x_0, v.

Proof. First, we show that, if F is convex then g is convex.

$$\begin{aligned}
&\forall t_1, t_2 \in \mathrm{dom}\, g, \lambda \in [0,1] \\
&\qquad g(\lambda t_1 + (1-\lambda)t_2) = F(x_0 + [\lambda t_1 + (1-\lambda)t_2]v) \\
&\qquad\qquad = F(\lambda[x_0 + t, v] + (1-\lambda)[x_0 + t_2 v]) \\
&\qquad\qquad \leq \lambda F(x_0 + t_1 v) + (1-\lambda)F(x_0 + t_2 v) \\
&\qquad\qquad = \lambda g(t_1) + (1-\lambda) g(t_2)
\end{aligned}$$

Second, we show that if g is convex in t, $\forall(x_0, v)$, then F is convex.

Pick arbitrary $x, y \in \mathrm{dom}\ F$, let $x_0 = x$, $v = (y - x)$, consider $g_{x_0,v}(t)$ for $t \in [0,1]$:

$$\begin{aligned}
g_{x_0,v}(t) &= F(x_0 + tv) \\
&= F(x + t(y - x)) \\
&= F((1-t)x + ty)
\end{aligned}$$

Since $g_{x_0,v}$ is convex in t, so F is convex in t. (t plays role of λ)

□

Example 8.25.

$$F(x) = \log\det(x^{-1})$$

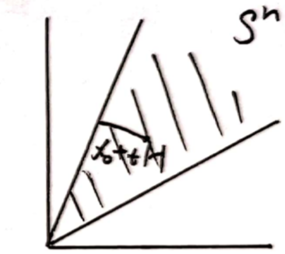
Note:

(1) dom $F = S^n_{++} \Leftrightarrow$ PD matrices.

(2) $S^n_{++} \subset S^n \leftarrow$ vector space of $n \times n$ symmetric matrices.

To show this function F is a convex function, we will show it is convex for all "lines".

The "line" in S^n is $x_0 + tH$, where x_0 is a symmetric PD matrix, $t \in \mathbb{R}$ and H is symmetric matrix in S^n. So $x_0 + tH$ is a family of symmetric matrices.



Notice that $x_0 \in S^n_{++}$, so x_0^{-1}exists and also $x_0^{\frac{1}{2}}$ exists.

$$\begin{aligned}
g(t) &= \log\det(x_0 + tH)^{-1} \\
&= \log\det[(x_0^{\frac{1}{2}} x_0^{\frac{1}{2}} + t x_0^{\frac{1}{2}} x_0^{-\frac{1}{2}} + H x_0^{-\frac{1}{2}} x_0^{\frac{1}{2}})^{-1}] \\
&= \log\det[x_0^{-\frac{1}{2}} (I + t x_0^{-\frac{1}{2}} H x_0^{-\frac{1}{2}})^{-1} x_0^{-\frac{1}{2}}] \\
&= \log\det x_0^{-1} + \log\det[(I + t x_0^{-\frac{1}{2}} H x_0^{-\frac{1}{2}})^{-1}] \\
&= \log\det x_0^{-1} + \log(\det(I + tM)^{-1}) \\
&= \log\det x_0^{-1} + \log[\prod_{i=1}^{n} (1 + t\lambda_i)^{-1}] \\
&= \log\det x_0^{-1} - \sum_{i=1}^{n} \log(1 + t\lambda_i)
\end{aligned}$$

In above inequalities we have utilized the property that $\det(AB) = \det(A) \cdot \det(B)$ and the determinant of a matrix equals to the product of its eigenvalues.

Also, notice that

$$(I + tM)v = v + t\lambda v = (1 + t\lambda)v$$

so the eigenvalues of $(I + tM)$ are $1 + t\lambda_i$.

Note:

1. $1 + t\lambda_i$ is an affine map in t.
2. Function $-log(\cdot)$ is convex.
3. Combine above results, since it is a sum of convex functions, $g(t)$ is convex in t (and thus $F(x)$ is convex in x).

Example 8.26. Consider the function of finding the max eigenvalue,

$$F(X) = \lambda_{\max}(X)$$

where dom $F = S^n$, and this function F is convex.

We illustrate this fact by two parts.

(a) First, we have

$$\lambda_{\max}(X) = \max_{v:\|v\|=1} v^T X v$$

By eigen-decomposition of symmetric matrices,

$$\begin{aligned}
v^T X v &= v^T Q \Lambda Q^T v \\
&= \tilde{v}^T \Lambda \tilde{v} \\
&= \sum_{i=1}^{n} (\tilde{v}_i)^2 \lambda_i \\
&\leq \lambda_{max}(X)
\end{aligned}$$

(b) Express x as following

$$v^T(\alpha X_1 + \beta X_2)v = \alpha v^T X_1 v + \beta v^T X_2 v$$

So $v^T X v$ is linear in X, and therefore $F(X)$ is the max of a bunch of functions that are linear in X, and thus it is convex.

Two more examples

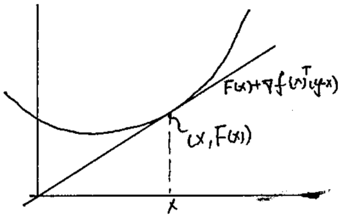
1. $F(x) = \sigma_{\max}(X)$ is convex on dom $F = \mathbb{R}^{n\times m}$.
2. $F(x) = (\det X)^{\frac{1}{n}}$, is concave on dom $F = S^n_{++}$.

"First-order" condition

Theorem 8.27. *A differentiable function F (i.e., dom F is open and gradients exist everywhere in domain F) is convex if and only if $\forall x, y \in$ dom F:*

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

and is strictly convex if $(*)$ is a strict inequality for all $x \neq y$.



Note that:

(1) The affine function of y given by $F(x) + \nabla F(x)^T(y - x)$ the First-order Taylor approximation, and the inequality above states that this approximation is an global underrestimator of F.

(2) There is a tangent plane that is a supporting hyperplane of epi F.

Proof. First, assume F is convex, and we show that $(*)$ holds.

Take any $(x, y) \in$ dom F, and by definition of convex function,

$$F((1-\lambda)x + \lambda y) \leq (1-\lambda)F(x) + \lambda F(y)$$

Rearrange yields

$$\frac{F(x + \lambda(y - x)) - F(x)}{\lambda} \leq F(y) - F(x)$$

Let $\lambda \to 0$ and observe that

$$lim_{\lambda\to 0}\frac{F(x + \lambda(y - x)) - F(x)}{\lambda} = \nabla F(x)^T(y - x)$$

Therefore,

$$\nabla F(x)^T(y - x) \leq F(y) - F(x)$$

Hence, the inequality $(*)$ holds (You can also do above procedure in 1-dimension, try it by yourself).

Secondly, we assume $(*)$ holds and show that F is convex.

Take any $(x, y) \in \text{dom } F$, then $\forall \lambda \in [0,1]$, $z = \lambda x + (1-\lambda)y \in \text{dom } F$ since dom F is convex.

Using $(*)$ for 2 times we get:

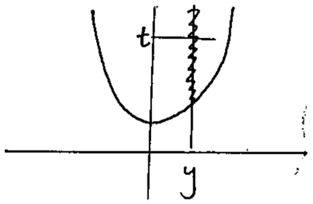
$$\begin{aligned}F(x) &\geq F(z) + \nabla F(x)^T(x - z)\\ F(y) &\geq F(z) + \nabla F(x)^T(y - z)\end{aligned}$$

Compute

$$\begin{aligned}\lambda F(x) + (1-\lambda)F(y) &\geq F(z) + \nabla F(z)^T[\lambda(x-z) + (1-\lambda)(y-z)]\\ &= F(z) + \nabla F(z)^T[\lambda x - \lambda z + y - z - \lambda y + \lambda z]\\ &= F(z) + \nabla F(z)^T[\lambda x + (1-\lambda)y - z]\\ &= F(z) + \nabla F(z)^T[z - z]\\ &= F(z)\\ &= F(\lambda x + (1-\lambda)y)\end{aligned}$$

Therefore, the function F is convex given that the inequality $(*)$ holds. □

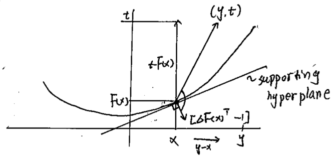
Connect 1-st order condition with epi F



Recall that $(x, t) \in \text{epi } F$ if $t \geq F(x)$, and the 1-st order condition: $\forall x, y \in \text{dom } F$, $F(y) \geq F(x) + \nabla F(x)^T(y - x)$.

Consider any $(y, t) \in \text{epi } F$:

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



"Second-order" condition

Theorem 8.28. *If F is everywhere twice differentiable, then F is convex if and only if its Hessian $\nabla^2 F(x) \geq 0$ (i.e., PSD) for all $x \in$ dom F.*

Proof. Similarly, we prove this theorem in two steps.

Firstly, assume F convex, and we show that $\nabla^2 F(x) \geq 0$.

Let $x_o \in \text{dom } F$(any point), $v \in \mathbb{R}^n$(a direction), then $z = x_0 + \lambda v$ is in dom F if $\lambda > 0$ sufficiently small.

By Taylor approximation,

$$F(z) = F(x_0) + \nabla F(x_0)^T(\lambda v) + \frac{1}{2}(\lambda_v)^T \nabla^2 F(x_0)\lambda v + O(\lambda^3)$$

Rearrange yields

$$\frac{1}{2}\lambda^2 v^T \nabla^2 F(x_0) v + O(\lambda^3) = F(z) - F(x_0) - \nabla F(x_0)^T (\lambda v)$$

The right-hand side is ≥ 0 by first-order-convexity result. Continuing, divide through by λ^2 to get:

$$\frac{1}{2} v^T \nabla^2 F(x_0) v + \frac{O(\lambda^3)}{\lambda^2} \geq 0$$

In above equation, $O(\lambda^3)$ means this is $\leq M\lambda^3$ (Big-O notation), so $\frac{O(\lambda^3)}{\lambda^2} \leq M\lambda$.

Letting $\lambda \to 0$, we get

$$\frac{1}{2} v^T \nabla^2 F(x_0) v \geq 0$$

So $\nabla^2 F(x_0)$ is PSD for all $x_0 \in \text{dom} f$, since $v \in \mathbb{R}^n$ can be chosen arbitrarily.

Secondly, assume $\nabla^2 F(x_0) \geq 0, \forall x_0 \in \text{dom } F$, and we show that F is convex.

Apply Taylor approximation with remainder $\forall x, y \in \text{dom } F$,

$$F(y) = F(x) + \nabla F(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 F(z)(y - x)$$

where Hessian is evaluated at some z between x and y(Mean value theorem).

Since $\nabla^2 F(z) \geq 0$, we have

$$F(y) \geq F(x) + \nabla F(x)^T (y - x)$$

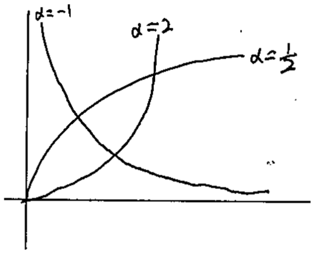
Then we could go back to 1^{st}−order-condition, which is true for $\forall x, y$. So F must be convex. □

Here we can give some examples:

1. $F(x) = x^2$, dom $F = \mathbb{R}$, $F''(x) = 2 > 0, \forall x \in \mathbb{R}$, so F is strictly convex.

2. $F(x) = x^3$, dom $F = \mathbb{R}$, $F'(x) = 3x^2$, $F''(x) = 6x$, so if we restrict the domain as dom $F = \mathbb{R}_+$ then F is convex.

3. $F(x) = x^\alpha$, dom $F = \mathbb{R}_+$, $F''(x) = \alpha(\alpha - 1)x^{\alpha-2}$, where $\alpha(\alpha - 1) > 0$ if $\alpha > 1$ or $\alpha < 0$ and $\alpha(\alpha - 1) < 0$ if $0 < \alpha < 1$. $x^{\alpha-2} \geq 0$ since dom $F = \mathbb{R}_+$.



4. $F(x) = \log x$, dom $F = \mathbb{R}_{++}$, $F'(x) = \frac{1}{x}$, $F''(x) = -\frac{1}{x^2} < 0$, so it is concave.

5. $F(x) = x \log x$, dom $F = \mathbb{R}_{++}$, $\frac{\alpha^2}{\alpha x^2} F(x) = \frac{\alpha}{\alpha x}[log_e(x) + \frac{x}{x}] = \frac{1}{x} > 0$, so it is convex.

6. $F(x) = e^{\alpha x}$, dom $F = \mathbb{R}$, $F''(x) = \alpha^2 e^{\alpha x}$, so it is convex for for $\alpha \neq 0$.

7. $F(x) = \frac{1}{2}x^T H x + c^T x + d$, the gradient and Hessian of F are given by

$$\nabla F(x) = \frac{1}{2}(H + H^T)x + c = \tilde{H}x + c$$
$$\nabla^2 F(x) = \tilde{H}$$

If $\tilde{H} \geq 0$, then it is convex; If $\tilde{H} \leq 0$, then it is concave.

If $\tilde{H}$ is neither PSD or negative semi-definite, then F is neither convex or concave.

Example 8.29. Consider following quadratic function,

$$\begin{aligned} F(x,y) &= x^2 + y^2 + 3xy \\ &= \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & \frac{3}{2} \\ \frac{3}{2} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \end{aligned}$$

Compute the eigenvalues of the matrix,

$$\begin{aligned} \det(\begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix} - \lambda I) &= \det \begin{bmatrix} 2-\lambda & 3 \\ 3 & 2-\lambda \end{bmatrix} \\ &= (2-\lambda)^2 - 9 \\ &= (\lambda - 5)(\lambda + 1) \end{aligned}$$

The matrix has one negative eigenvalue and one positive eigenvalue, so it is neither PSD or negative PSD.

Also, if we try $-45°$ line, we will see slice is not convex,

$$F(x, -x) = x^2 + (-x)^2 + 3x(-x) = 2x^2 - 3x^2 = -x^2.$$

Example 8.30. Geometric mean $\sqrt{x_1 x_2} = F(x_1 x_2)$, dom $F = \mathbb{R}_+ \times \mathbb{R}_+$, is concave.

The Hessian of F is given by

$$\nabla^2 F(x) = -\frac{\sqrt{x_1 x_2}}{y} \begin{bmatrix} \frac{1}{x_1^2} & -\frac{1}{x_1 x_2} \\ -\frac{1}{x_1 x_2} & \frac{1}{x_2^2} \end{bmatrix}$$

There are two ways to determine the type of the Hessian(e.g., whether it is PSD)

1) Calculate eigenvalues as we did in previous example.
2) Use definition of negative PSD (or PSD):

$$\begin{aligned} v^T\nabla^2 F(x)v &= -\frac{\sqrt{x_1x_2}}{2}[\frac{v_1^2}{x_1^2} - \frac{2v_1v_2}{x_1x_2} + \frac{v_2^2}{x_2^2}] \\ &= -\frac{\sqrt{x_1x_2}}{2}(\frac{v_1}{x_1} - \frac{v_2}{x_2}) \le 0 \end{aligned}$$

So it is concave in (x_1, x_2). More generally, the function $F(x) = (\prod_{i=1}^n x_i)^{\frac{1}{n}}$ is concave in $x \in \mathbb{R}^n$.

Consequence of convexity conditions for differentiable F

By 1-st order condition, if F is convex, then we have

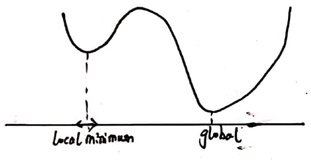
$$F(y) \ge F(x) + \nabla F(x)^T(y-x), \forall x, y \in \text{dom } F$$

What if $\nabla F(x^*) = 0 \in \mathbb{R}^n$ for some $x^* \in \text{dom } F$? In this case we may have

$$F(y) \ge F(x^*) + \nabla F(x^*)^T(y - x^*) = F(x^*),\ \forall y \in \text{dom } F$$

Therefore, if we can find an $x^* \in \text{dom } F$ such that $\nabla F(x^*) = 0$, then it is also a global minimum.

Local minima vs Global minima



Definition 8.31. x^* is a local minimum of F if $\exists \epsilon > 0$ such that for all x satisfying $\|x - x^*\| < \epsilon$ we will have $F(x) \ge F(x^*)$.

Theorem 8.32. *Suppose F is twice differentiable (not necessarily to be convex), then we have*
(1) If x^* is a local optimum, then $\nabla F(x^*) = 0$ and $\nabla^2 F(x^*) \ge 0$.
(2) If $\nabla F(x^*) = 0$ and $\nabla^2 F(x^*) > 0$, then x^* is a local minimum.

Proof of (1). Let x^* be a local optimum, consider any v:

$$lim_{t\to 0^+}\frac{F(x^*+tv) - F(x^*)}{t} = \nabla F(x^*)^T v \ge 0$$

It's non-negative since x^* is local minimum.

This implies that $\nabla F(x^*) = 0$ because v is arbitrary.

E.g. If $\nabla F(x^*) \neq 0$, then take $v = -\nabla F(x^*)$ and we would get a negative one unless $\nabla F(x^*) = 0$.

Consider the second derivative,

$$\begin{aligned} lim_{t\to 0^+}\frac{F(x^*+tv) - F(x^*)}{t^2} &= lim_{t\to 0^+}\frac{F(x^*) + \nabla F(x^*)^T(tv) + \frac{1}{2}(tv)^T\nabla F(x^*)(tv) + o(t^2) - F(x^*)}{t^2} \\ &= lim_{t\to 0^+}\frac{1}{2}v^T\nabla^2 F(x^*)v + \frac{\sigma(t^2)}{t^2} \\ &= \frac{1}{2}v^T\nabla^2 F(x^*)v \\ &\ge 0 \end{aligned}$$

Since v is arbitrary and by definition of PSD, the Hessian $\nabla^2 F(x^*) \geq 0$

□

For twice differentiable functions, what we are told here is:
x^* is a local optimum $\Rightarrow \nabla F(x^*) = 0$ and $\nabla^2 F(x^*) \geq 0$.
Furthermore, if the function F is convex, we may have
$\Rightarrow \nabla^2 F(x) \geq 0, \forall x \in \text{dom} F$
$\Rightarrow \nabla F(x^*) = 0 \Rightarrow x^*$ is global optimum.
Put together to say: If F is convex, then the local optimum is also the global optimum.

Proof of (2). If $\nabla F(x^*) = 0$ and $\nabla^2 F(x^*) > 0$, then x^* is a local optimum.

Again, use Taylor expansion:

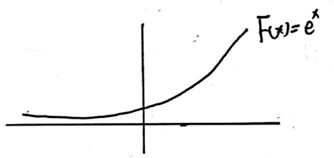
$$\begin{aligned} F(x) &= F(x^* + tv) \\ &= F(x^*) + \nabla F(x^*)^T (tv) + \frac{1}{2}(tv)^T \nabla^2 F(x^*)(tv) + o(\|v\|^2) \\ &= F(x^*) + \frac{1}{2} t^2 v^T \nabla^2 F(x^*) v + o(\|v\|^2) \end{aligned}$$

$\Rightarrow$ If t is sufficiently small, the quadratic term dominates the $o(\|v\|^2)$ term.
$\Rightarrow$ Any point in neighborhood sufficiently small (meaning t sufficiently small) has evaluation larger than $F(x^*)$.
$\Rightarrow x^*$ is local optimum. □

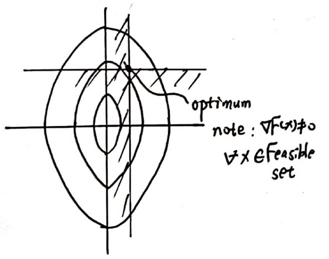
Remarks:

1. It is possible that there is no local or global optimum.

   This graph is on inf $F(x) = 0$ but there is no $x \in \text{dom} F = \mathbb{R}$ achieves $F(x) = 0$.



2. The above story is for unconstrained optimization problem, since we consider the entire domain above.

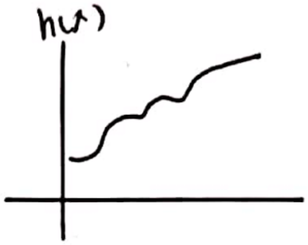


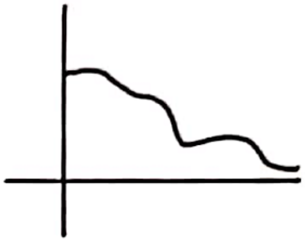
Composition of Functions

Let $F(x) = h(g(x))$ where $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$. Then $F : \mathbb{R}^n \to \mathbb{R}$ is convex if:

1. g is convex, h is convex and non-decreasing. Or,

2. g is concave, h is convex and non-increasing.





Prove: For differentiable functions, can also prove for non-differentiable.
(i)

- $F'(x) = h'(g(x))g'(x)$.
- $F''(x) = h''(g(x))g'(x)g'(x) + h'(g(x))g''(x) \geq 0$

Note: Doing derivative for $n = 1$ without loss of generality because we showed a convex function must be convex along all lines.
(ii)

$$F''(x) = h''(g(x))(g'(x))^2 + h'(g(x))g''(x) \geq 0$$

Can extend to multiple dimensions.

$$\begin{aligned} g_i : \mathbb{R}^n \to \mathbb{R}, h : \mathbb{R}^k \to \mathbb{R} \\ F(x) &= h(g(x)) \\ &= h(g_1(x)g_2(x)...g_k(x))\text{is convex} \end{aligned}$$

$\cdot$ g is convex and h is convex and non-decreasing in each of its arguments.

Example 8.33. The function $F(x) = \exp(g(x))$ is convex, where $g(x)$ is convex.

Apparently we could let $h(\cdot) = \exp(\cdot)$, and since it is convex and non-decreasing, $F(x)$ is convex.

Example 8.34. The function $F(x) = \frac{1}{g(x)}$ is convex if $g(x)$ is concave and positive, $\forall x$.

We let $F(x) = h(g(x))$, where $h(x) = \frac{1}{x}$. Since dom $h = \mathbb{R}_{++}$, h is convex.

Note that h is convex and $h(x) = \frac{1}{x}$ is non-increasing on $\mathbb{R}_{++}$, so if $g(\cdot)$ is concave, then F is a convex function.

Example 8.35. The function $F(x) = -\sum_{i=1}^{k} log(-F_i(x))$ is convex on $\{x|F_i(x) < 0\ \forall i \in \{1, \cdots, k\}\}$ if all F_i are convex.

Consider the domain of this F, note that for each $F_i(x) < 0 \Leftrightarrow -F_i(x) > 0$, so dom $F = \cap_{i=1}^{k}\{x|F_i(x) < 0\}$. It is the intersection of sublevel sets of convex functions, and therefore it is a convex set.

Since $F(x) = \sum_{i=1}^{k} -log(-F_i(x))$ is the positive sum of convex functions, it is convex eventually.

9
Convex optimization

Introduction to convex optimization problems

General form: Consider the functions $F_i(x), h_i(x) : \mathbb{R}^n \to \mathbb{R}$.

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & F_0(x) \quad \text{"objective function"} \\ s.t. \quad & F_i(x) \leq 0, i = 1...m \quad \text{inequality constraints} \\ & h_i(x) = 0, i = 1...p \quad \text{equality constraints} \end{aligned}$$

Feasible set for this question:

$$\mathcal{C} = \{x | F_i(x) \leq 0, i = 1...m, h_i(x) = 0, i = 1...p\}$$

Optimal value: $p^* = \inf_{x \in \mathcal{C}} F_0(x)$. Note that it could be ∞, and also could be empty.

Optimal points: $\{x \in \mathcal{C} | F_0(x) = p^*\}$. Note that it could be empty, and also could be not unique.

Example 9.1. Consider the optimization problem:

$$\begin{aligned} \min_x \quad & \min x_1 + x_2 \\ s.t. & - x_1 \leq 0 \\ & - x_2 \leq 0 \\ & 1 - x_1 x_2 \leq 0 \end{aligned}$$

So $x^* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $p^* = 2$, as illustrated in the figure.

Convex optimization problem:

$$\begin{aligned} \min_x \quad & F_0(x) \\ s.t. F_i(x) & \leq 0 \quad i = 1, \cdots, m \\ a_i^T x - b_i & = 0 \quad i = i, \cdots, p \end{aligned}$$

$a_i^T + b_i = 0$ is often written as:

$$\begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_p^T \end{bmatrix} x = \begin{bmatrix} b_1 \\ \vdots \\ b_P \end{bmatrix} \Leftrightarrow Ax = b$$

1. All F_i, $i \in \{0, 1, ...n\}$ are convex functions.

2. All equality constraints are affine.

Remarks:

1. Think about feasible set,

$$\mathcal{C} = (\cap_{i=1}^{m} \{x | F_i(x) \leq 0\}) \cap (\cap_{i=1}^{p} \{x | a_i^T x - b_i = 0\})$$

For the first part, each is a sublevel set of a convex function therefore convex.

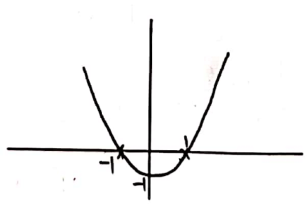
For the second part, each is an affine set and therefore convex.

So the feasible set $\mathcal{C}$ is an intersection of $p + m$ convex sets, and therefore it is a convex set.

2. Note: $h_i(x)$ are affine(and not more general convex) to keep the set $\{x | h_i(x) = 0\}$ a convex set.

Let $h_i(x) = x^2 - 1$:

$$\{x | x^2 - 1 = 0\} = \{x | x^2 = 1\} = \{\pm 1\}$$



$$\begin{aligned} \min_x \quad & F_0(x) \\ s.t. \quad & F_i(x) \leq 0 \quad i = 1, ..., m \\ & a_i^T x - b_i = 0 \quad i = 1, ..., p \end{aligned}$$

- $F_o, F_1, ..., F_m$ are convex
- $Ax - b = 0$

Definition 9.2. $x \in \mathcal{C}$ is local optimum for a constrained optimization if $\exists \epsilon > 0$, $s.t. \forall y \in \mathcal{C}$ and $\|x - y\| < \epsilon$, we have $F_0(y) \geq F_0(x)$

Theorem 9.3. *For a convex optimization problem a local minimum is also a global optimum.*

We prove this theorem for two particular instances:

1. Unconstrained optimization problem

2. Differentiable objective function F_0

Proof. For the first instance, suppose $x \in \mathcal{C}$ is not globally optimal but is locally minimal.

$\rightarrow$ Because not globally optimal, $\exists y \in \mathcal{C}$ s.t. $F_0(y) < F_0(x)$

$\rightarrow$ Consider $z = \lambda x + (1-\lambda)y \in \mathcal{C}$, because $\mathcal{C}$ is convex, so we have

$$\begin{aligned} F_0(z) &\leq \lambda F_0(x) + (1-\lambda)F_0(x) \\ &< \lambda F_0(x) + (1-\lambda)F_0(x) \\ &= F_0(x) \end{aligned}$$

$\rightarrow$ By picking λ sufficiently close to 1 (but < 1), $z \in \mathcal{C}$ is in neighborhood of x and has a lower cost, so x cannot be local minimum, and thus lead to a contradiction.

Hence, for a unconstrained convex optimization problem, a local minimum must also be global minimum.

As for the case F_0 is differentiable and convex

$$\begin{aligned} \min_x \quad & F_0(x) \\ s.t. \quad & F_i(x) \leq 0 \quad i = 1, ..., m \\ & a_i^T x - b_i = 0 \quad i = 1, ..., p \end{aligned}$$

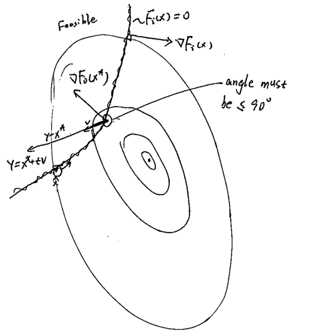
- For unconstrained case x^* is optimal iff $\nabla F(x^*) = 0$
- For constrained optimization it is very possible there is no $x \in \mathcal{C}$ satisfies $\nabla F(x) = 0$

Therefore, in this case a local minimum is also a global minimum(recall that first order condition is a necessary condition for a local optimum but not a sufficient condition). □

Theorem 9.4. *For a convex optimization problem with (convex) feasible set $\mathcal{C}$ and differentiable (convex) objective $F_0 : \mathbb{R}^n \rightarrow \mathbb{R}$, a point $x^* \in \mathcal{C}$ is optimal iff:*

$$\nabla F_0(x^*)^T(y - x^*) \geq 0 \quad \forall y \in \mathcal{C}$$

That is, start at the point $x^* \in \mathcal{C}$, move into feasible set in direction v and then evaluate $F_0(x^* + tv)$. $F_0(x^* + tv)$ must be non-decreasing for $t \geq 0$.



Proof. First assume the inequality holds, and we show that x^* is the global optimum. A

Apply 1-st order condition for optimality, i.e. $\forall y \in \mathcal{C}$:

$$\begin{aligned} F_0(y) &\geq F_0(x^*) + \nabla F_0(x^*)^T(y - x^*) \\ &\geq F_0(x^*) \end{aligned}$$

where the second term on the r.h.s. must be non-negative by our assumption.

So for $\forall y \in \mathcal{C}$, we have $F_0(y) \geq F_0(x^*)$. Hence x^* is a global optimum if the inequality holds.

Secondly, we show that, if x^* is the global optimum, then the inequality must holds. We prove this by contradiction.

Suppose that $\exists y \in \mathcal{C}$ such that $\nabla F_0(x^*)^T(y - x^*) < 0$.

Look at the point z defined as

$$\begin{aligned} z &= \lambda y + (1-\lambda)x^* \\ &= x^* + \lambda(y - x^*) \end{aligned}$$

All the points z defined above must be feasible $\forall \lambda \in [0,1]$, because the set $\mathcal{C}$ is a convex set.

We claim that, there exists some points z such that $F_0(z) < F_0(x^*)$, by showing that

$$\begin{aligned} \left.\frac{dF_0(z)}{d\lambda}\right|_{\lambda=0} &= \left.\frac{d}{d\lambda}F_0(x^* + \lambda(y - x^*))\right|_{\lambda=0} \\ &= \nabla F_0(x^*)^T(y - x^*) \\ &< 0 \end{aligned}$$

Since the slope(the gradient) is strictly negative(as we assume the inequality does not hold), the value of $F_0(z)$ will decreases as $\lambda > 0$ increase, so we will have a smaller value $F_0(z)$ compared to $F_0(x^*)$, which contradicts the global optimality of x^*.

Therefore, if the inequality does not hold, x^* will not be the global optimum. So x^* is the global optimum if and only if the inequality holds. □

Quasi-convex minimization

Definition 9.5. A function $F_0 : \mathbb{R}^n \to \mathbb{R}$ is called quasi-convex if its domain and all its sub-level sets are convex sets.

Definition 9.6. A function $F_0 : \mathbb{R}^n \to \mathbb{R}$ is called quasi-concave if $-F_0$ is quasi-convex, that is, all its super-level sets are convex sets.

Definition 9.7. If a function is both quasi-convex and quasi-concave, i.e., all its sub-level sets and super-levels are convex sets, then it is called quasi-linear.

Consider the Quasi-convex minimization problem as follows,

$$\begin{aligned} \min_x \quad & F_0(x) \quad \text{quasi-convex} \\ s.t. \quad & F_i(x) \leq 0 \quad i = 1, \cdots, m \quad \text{convex} \\ & a_i^T - b_i = 0 \quad i = 1, \cdots, m \quad \text{affine} \end{aligned}$$

We can rewrite this formulation into the form of feasibility problem,

$$\begin{aligned} &\min_x \\ s.t.\quad & F_0(x) \leq t \\ & F_i(x) \leq 0 \quad i = 1, \cdots, m \\ & a_i^T - b_i = 0 \quad i = 1, \cdots, m \end{aligned}$$

where the constraint $F_0(x) \leq t$ defines a convex set(sub-level set is a convex set).

Example 9.8. Consider the function $F_0(x) = \log x$ defined on R_{++}.

We can show that, the sub-level sets of $F_0(x)$ are convex sets, so it is quasi-convex; the super-level sets of $F_0(x)$ are convex sets, so it is also quasi-concave.

Therefore, we call the function $F_0(x) = \log x$ is quasi-linear.

Example 9.9. Consider the function $F_0(x) = \frac{P(x)}{Q(x)}$, where $P(x)$ is convex and non-negative(to make sure $t \geq 0$), $Q(x)$ is concave and dom $F_0 = \{x|Q(x) > 0\}$.

So the sub-level set of this function can be expressed as

$$\begin{aligned} \{x|F_0(x) \leq t\} &= \{x|\frac{P(x)}{Q(x)} \leq t\} \\ &= \{x|P(x) \leq tQ(x)\} \\ &= \{x|P(x) - tQ(x) \leq 0\} \end{aligned}$$

which is a convex set, since $P(x)$ and $-Q(x)$ are convex functions(so $P(x) - tQ(x)$ is convex as well), $t \geq 0$ and the set is the pre-image of a convex set under a convex function.

Hence, the function $F_0(x) = \frac{P(x)}{Q(x)}$ is quasi-convex since all its sub-level sets are convex sets.

Let's consider the special case for this kind of functions by setting $F_0(x) = \frac{a^T x + b}{c^T x + d}$ with dom $F_0 = \{x \mid c^T x + d > 0\}$.

We can show that(by similar approach above), the sub-level sets and super-level sets are convex sets, and thus $F_0(x) = \frac{a^T x + b}{c^T x + d}$ is quasi-linear(both quasi-convex and quasi-concave).

Important note:

(1) Quasi-convex optimization problems may have local optimum that are NOT global optimum (differ from the convex optimization problems).

Convex optimization problem with generalized inequality constraints

Convex optimization problem with generalized inequality constraints is an useful generalized version of convex optimization problem. By using the generalized inequalities in the constraints, the problem is formulated as

$$\begin{aligned} \min_x \quad & F_0(x) \\ s.t. \quad & F_i(x) \preceq_{k_i} 0 \quad i = 1, ..., m \\ & h_i(x) = 0 \quad i = 1, ..., m \end{aligned}$$

where $F_0 : \mathbb{R}^n \to \mathbb{R}$ is convex, $F_i : \mathbb{R}^n \to \mathbb{R}^l$ is "k_i-convex" and k_i are cones. Recall a bit about the k_i-convex, which means $\forall \lambda \in [0, 1]$ we have

$$F_i(\lambda x + (1 - \lambda) y) \preceq_{k_i} \lambda F_i(x) + (1 - \lambda) F_i(y)$$

Notice that, many of the results for ordinary convex optimization problems hold for problems with generalized inequalities. Some examples are:

(1) The feasible set, any sub-level set, and the optimal set are convex.

(2) Any point that is locally optimal for the problem is globally optimal.

(3) The optimality condition for differentiable F_0 in ordinary convex optimization problems still holds in this problem, without any change.

Semi-definite program(SDP)

When above cone k is S^n_+, the cone of PSD n by n matrices, a special case of convex optimization problem with generalized inequality constraints is, the Semi-definite program(SDP) problem,

$$\begin{aligned} \min_x \quad & c^T x \\ s.t. \quad & A_0 + A_1 x_1 + \cdots + A_n x_n \leq 0 \\ & Fx = g \end{aligned}$$

where the inequality constraint is defined by a linear matrix inequity (LMI), $A_i \in S^n$, and $-(A_0 + A_1 x_1 + \cdots + A_n x_n) \in S^n_+$. Note that here $\geq$ is understood as $\geq_k$, and thus for some symmetry matrices $Z \geq 0$ means that Z is PSD. We will return to SDP later in this chapter and have more discussion.

Second-Order Cone Program (SOCPs)

A SOCP problem is formulated as

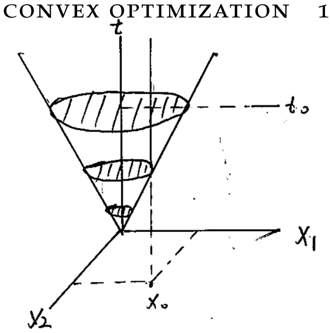
$$\begin{aligned} \min \quad & F^T x \\ s.t. \quad & \|A_i x + b_i\|_2 \leq c_i^T x + d_i \quad i = 1, ..., m \\ & Fx \leq g \end{aligned}$$

where $A_i \in \mathbb{R}^{n_i \times n}$, $x \in \mathbb{R}^n$, $b_i \in \mathbb{R}^{n_i}$, $F \in \mathbb{R}^{p \times n}$ and $g \in \mathbb{R}^p$.

Norm cone:

$$\mathcal{C} = \{(x,t) \in \mathbb{R}^{n+1} | \|x\| \leq t\} \subseteq \mathbb{R}^{n+1}$$

- First fix $t = t_0$, $\mathcal{C}_{t_0} = \{(x,t_0) | \|x\| \leq t_0\}$, "fill-in" slice
- Next fix $x = x_0$, $\mathcal{C}_{x_0} = \{(x_0,t) | \|x_0\| \leq t\}$
  → Fix point in $x \in \mathbb{R}^n$ to $x = x_0$ and "Fill up".



Proof. We first prove that the set $\mathcal{C}$ is a Cone, and then prove it is a convex cone on the second part.

1. Pick any $(x_0, t_0) \in \mathcal{C}$, we show that for $(\theta x_0, \theta t_0) \in \mathcal{C}$ and $\forall \theta \in \mathbb{R}_+$, we have
$$\|\theta x_0\| = |\theta| \|x_0\| = \theta \|x_0\| \leq \theta t_0$$
Therefore, $\mathcal{C}$ is a cone.

2. Pick $(x_0, t_0) \in \mathcal{C}$ and $(y_0, s_0) \in \mathcal{C}$, we show that the point
$$(\theta x_0 + (1-\theta) y_0, \theta t_0 + (1-\theta) s_0) \in \mathcal{C},\ \forall 0 \leq \theta \leq 1$$
That is,
$$\begin{aligned} \|\theta x_0 + (1-\theta) y_0\| &\leq \|\theta x_0\| + \|(1-\theta) y_0\| \\ &= |\theta| \|x_0\| + |(1-\theta)| \|y_0\| \\ &= \theta \|x_0\| + (1-\theta) \|y_0\| \\ &\leq \theta t_0 + (1-\theta) s_0 \end{aligned}$$
Therefore the set $\mathcal{C} = \{(x,t) | \|x\|_2 \leq t\}$ is convex, and thus it is a convex cone.

□

Example 9.10. Consider following affine map:

$$F_i(x) = \begin{bmatrix} A_i x + b_i \\ c_i^T x + d_i \end{bmatrix} \in \mathbb{R}^{n+1}$$

For i-th constraint, we want to have $\|A_i x + b_i\| \leq c_i^T x + d_i$, and this is equivalent to require

$$\{x | F_i(x) \in \mathcal{C}_{n+1}\} = F_i^{-1}(\mathcal{C}_{n+1})$$

Remarks:

1. If all $A_i = 0$, then we get an general LP;
2. If $c_i = 0$, then we get a QCQP (which is obtained by squaring each of the constraints);
3. The constraint is a second-order cone constraint, since we use l_2 norm for this cone.

Robust Linear Programs

We consider a linear program in inequality form,

$$\begin{aligned} \min \quad & c^T x \\ s.t. \quad & a_i^T x \le b_i \quad i = 1, \cdots, m \end{aligned}$$

in which there is some uncertainty or variation in the parameters c, a_i, b_i. To simplify the exposition we assume that c and b_i are fixed, and that a_i are the parameters with uncertainty.

There are two versions for the robustness of uncertainty in a_i, that is, (1) worst-case (2)statistical approach

(1) Worst Case: Assume a_i are known to lie in given ellipsoids.
In this case, we know that

$$a_i \in \xi_i = \{a | a = \bar{a}_i + P_i u, \|u\| \le 1\}$$

where $P_i \in \mathbb{R}^{n \times n}$ (If P_i is singular we obtain 'flat' ellipsoids, of dimension rank P_i; $P_i = 0$ means that a_i is known perfectly).

Hence, the robust linear program problem is formulated as

$$\begin{aligned} \min \quad & c^T x \\ s.t. \quad & a_i^T x \le b_i \quad \forall a_i \in \xi_i \quad i = 1, \cdots, m \end{aligned}$$

Notice that the constraint of above formulation can be expressed as

$$(\bar{a}_i + P_i u)^T x \le b_i \text{ and } \|u_i\| \le 1, \ \forall i = 1, \cdots, m$$

Or equivalently,

$$\sup_{\|u_i\| \le 1} (\bar{a}_i + P_i u)^T x \le b_i, \ \forall i = 1, \cdots, m$$

Furthermore, this constraint can be expressed as a second order cone constraint, since

$$\begin{aligned} (\bar{a}_i + P_i u)^T x &= \bar{a}_i{}^T x + u^T P_i^T x \\ &\le \bar{a}_i{}^T x + (\frac{P_i^T x}{\|P_i^T x\|})^T P_i^T x \\ &= \bar{a}_i{}^T x + \frac{x^T P_i P_i^T x}{\|P_i^T x\|_2} \\ &= \bar{a}_i{}^T x + \|P_i^T x\|_2 \end{aligned}$$

so this robust LP can be expressed as the SOCP formulated as

$$\begin{aligned} \min \quad & c^T x \\ s.t. \quad & \bar{a}_i{}^T + \|P_i^T x\|_2 \le b_i, \ \forall i = 1, \cdots, m \end{aligned}$$

Note that the additional norm terms act as regularization terms; they prevent x from being large in directions with considerable uncertainty in the parameters a_i.

(2) Statistical approach: Assume that a_i are independent Gaussian random vectors.

In this case, we assume that a_i are independent Gaussian random vectors such that $a_i \sim N(\bar{a}_i, \Sigma_i)$.

Recall a little bit regarding the statistics, since each component a_i are independent, we could consider the mean and variance for the random variables $a_i^T x - b_i$ respectively,

The mean is given by

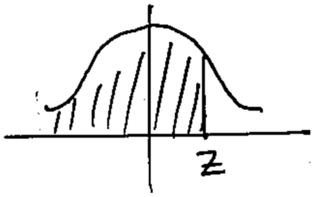
$$\mathbb{E}[a_i^T x - b_i] = \bar{a}_i{}^T x - b_i = \mu_i$$

The variance is given by

$$\begin{aligned}
\mathbb{E}[((a_i^T x - b_i) - (\bar{a}_i{}^T x - b_i))^2] &= \mathbb{E}[((a_i - \bar{a}_i)^T x)^2] \\
&= \mathbb{E}[x^T (a_i - \bar{a}_i)(a_i - \bar{a}_i)^T x] \\
&= x^T \mathbb{E}[(a_i - \bar{a}_i)(a_i - \bar{a}_i)^T] x \\
&= x^T \Sigma_i x \\
&= \sigma^2 \\
&= x^T \Sigma_i^{\frac{1}{2}} \Sigma_i^{\frac{1}{2}} x \\
&= \|\Sigma_i^{\frac{1}{2}} x\|_2^2
\end{aligned}$$

Since the constraints involve randomness, we would like to consider the probability such that the constraint holds, which is given by

$$\begin{aligned}
\mathbb{P}[a_i^T x \le b_i] &= \Phi\left(\frac{b_i - \bar{a}_i{}^T x}{\sigma_i}\right) \\
&= \Phi\left(\frac{b_i - \bar{a}_i{}^T x}{\|\Sigma_i^{\frac{1}{2}} x\|_2}\right)
\end{aligned}$$



Suppose now we require that each constraint should hold with a probability exceeding η, that is,

$$\begin{aligned}
&\mathbb{P}[a_i^T x \le b_i] \ge \eta \\
\Leftrightarrow &\Phi(\frac{b_i - \bar{a}_i{}^T x}{\sigma_i}) \ge \eta \\
\Leftrightarrow &\frac{b_i - \bar{a}_i{}^T x}{\sigma_i} \ge \Phi^{-1}(\eta)
\end{aligned}$$

By above argument, we have the problem formulated as

$$\begin{aligned}
\min \quad & c^T x \\
s.t. \quad & \mathbb{P}[a_i^T x \le b_i] \ge \eta, \ \forall i = 1, \cdots, m
\end{aligned}$$

and it turns out(we have showed this fact above) that, this is equivalent to a SOCP problem as the following

$$\begin{aligned} \min \quad & c^T x \\ s.t. \quad & b_i - \bar{a}_i{}^T x \geq \Phi^{-1}(\eta) \|\Sigma_i^{\frac{1}{2}} x\|_2, \ \forall i = 1, \cdots, m \end{aligned}$$

Geometric Program(GP)

Before introducing GP, we give some definitions first,

- "Monomial": $h(x) = cx_1^{\alpha_1} x_2^{\alpha_2} ... x_n^{\alpha_n}$, $c > 0, \alpha_i \in \mathbb{R}$, dom$h = (x | x_i > 0 \quad \forall i \in 1...n)$

- Posynomial: $F(x) = \sum_{k=1}^{k} c_k x_1^{\alpha_{1k}} x_2^{\alpha_{2k}} ... x_n^{\alpha_{nk}}$(sum of monomials)

  $\rightarrow$ note closed under addition, multiplication, non-negative scaling.

With these definitions in mind, a Geometric program problem is formulated as

$$\begin{aligned} \min \quad & F_0(x) \\ s.t. \quad & F_i(x) \leq 1, \quad i = 1, \cdots, m \\ & h_i(x) = 1, \quad i = 1, \cdots, p \end{aligned}$$

where $x \in \mathbb{R}^n$, $F_0, F_1, \cdots, F_m$ are posynomials, $h_0, h_1, \cdots, h_m$ are monomials.

Geometric programs are not (in general) convex optimization problems, but they can be transformed to convex problems by a change of variables and a transformation of the objective and constraint functions.

To get a GP into convex form, we set $y_i = log x_i$, so $x_i = e^{y_i}$ (recall $x_i > 0$). Hence, we have

Transformation of monomials:

$$\begin{aligned} & h(x_1, x_2, \cdots, x_m) = cx_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n} \\ \Leftrightarrow & \log h(x_1, x_2, \cdots, x_m) = \log c + \alpha_1 \log x_1 + \alpha_2 \log x_2 + \cdots + \alpha_n \log x_n \\ \Leftrightarrow & \log h(e^{y_1}, e^{y_2}, \cdots, e^{y_m}) = \log c + \alpha_1 y_1 + \alpha_2 y_2 + \cdots + \alpha_n y_n \end{aligned}$$

and this an affine function of y_i, and thus it is convex.

Transformation of posynomials:

$$
\begin{aligned}
&F(x_1, \cdots, x_n) = \sum_{i=1}^{k} c_k x_1^{\alpha_{1k}} x_2^{\alpha_{2k}} \cdots x_n^{\alpha_{nk}} \\
&\Leftrightarrow \log F(e^{y_1}, e^{y_2}, \cdots, e^{y_n}) = \log(\sum_{k=1}^{k} e^{\log c_k} e^{\alpha_{1k} y_1} \cdots e^{\alpha_{nk} y_n}) \\
&\Leftrightarrow \log F(e^{y_1}, e^{y_2}, \cdots, e^{y_n}) = \log(\sum_{k=1}^{k} e^{\alpha_{1k} y_1 + \cdots + \alpha_{nk} y_n + \log c_k})
\end{aligned}
$$

and we could show this is also a convex function of y_i.

With above transformations, we have the geometric program in convex form,

$$
\begin{aligned}
\min \quad & \log F_0(e^{y_1}, e^{y_2}, \cdots, e^{y_n}) \\
s.t. \quad & \log F_i(e^{y_1}, e^{y_2}, \cdots, e^{y_n}) \le \log(1) = 0, \ \forall i = 1, \cdots, m \\
& \log h_i(e^{y_1}, e^{y_2}, \cdots, e^{y_n}) = 0, \ \forall i = 1, \cdots, q
\end{aligned}
$$

Note that:

(1) To distinguish the GP from the original formulation and this convex one, we refer to the original as a geometric program in posynomial form, and the one after transformation as the geometric program in convex form.

(2) If the posynomial objective and constraint functions all have only one term, i.e., are monomials, then the convex form geometric program reduces to a (general) linear program.

Semi-definite Programs(SDPs)

Recall that, a SDP is formulated as

$$
\begin{aligned}
\min \quad & c^T x \\
s.t. \quad & F_0 + x_1 F_1 + x_2 F_2 + ... + x_n F_n \le 0 \\
& Gx = h
\end{aligned}
$$

where $F_0, F_1, \cdots, F_n \in S^m$, and inequality constraint here is a linear matrix inequality.

We show that, the inequality constraint above indeed defines a convex set, and more precisely, is a PSD cone. Since

$$F(x) := F_0 + x_1 F_1 + x_2 F_2 + \cdots + x_n F_n \le 0 \Leftrightarrow -F(x) \ge 0$$

so $-F(x) \in S^m_+$, and thus $\{x | -F(x) \in S^n_+\}$ is a convex set.

Alternatively, we could also consider the standard form of a SDP:

$$
\begin{aligned}
\min \quad & \text{trace}(CZ) \\
s.t. \quad & \text{trace}(A_i Z) = b_i, \quad i = 1, \cdots, m \\
& Z \ge 0
\end{aligned}
$$

To obtain the above standard form, similar with what we did in LP, we

(1) Introduce slack variables(variables that are non-negative) into inequality constraints and then obtain equality constraints;

(2) Decompose x_i by $x_i = x_i^+ - x_i^-$, $x_i^+ \geq 0$, $x_i^- \geq 0$.

Relaxation of homogeneous QCQP

First, recall that a QCQP problem is formulated as

$$\begin{aligned} \min \quad & \frac{1}{2}x^T P_0 x + q_0^T x + r_0 \\ s.t. \quad & \frac{1}{2}x^T P_i x + q_i^T x + r_i \leq 0, \quad i = 1, ..., m \\ & Fx = g \end{aligned}$$

Notice that,

- The QCQP is homogeneous if $q_i = 0, \forall i \in \{0, 1, ...m\}$;
- The QCQP is convex if all P_i are PSD;
- The QCQP is non-convex if some P_i are not PSD, or if we replace the inequality with a equality.

Now, lets consider the homogeneous QCQP with the following formulation

$$\begin{aligned} \min \quad & x^T C x \\ s.t. \quad & x^T F_i x \leq g_i \\ & x^T H_i x = l_i \end{aligned}$$

which is not necessary to be convex optimization problem.

Notice that

$$x^T C x = \text{trace}(x^T C x) = \text{trace}(C x x^T) = \text{trace}(CX)$$

where we let $X := xx^T$, rank$(X) = 1$, $X \geq 0$.

So, the above homogeneous QCQP can also be written as:

$$\begin{aligned} \min \quad & \text{trace}(CX) \\ s.t. \quad & \text{trace}(F_i X) \leq g_i, \quad i = 1, \cdots, m \\ & \text{trace}(H_i X) = l_i, \quad i = 1, \cdots, p \\ & \text{rank}(X) = 1 \\ & X \geq 0 \end{aligned}$$

Relaxation: Drop a constraint that is hard to deal with and thus we could solve an easier problem.

→ min is the lower bound to original problem's optimum value.

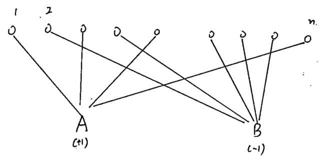
→ Maybe (if relaxation is good) you can figure out an X that is a good enough solution to original:

$$X^*_{relaxed} = \sum_{i=1}^{n} v_i v_i^T \lambda_i \Rightarrow \lambda_1 v_1 v_1^T$$

Example 9.11. Two-way partitioning problem

Suppose now we have n items, and we want to partition these n items into 2 different sets while minimizing the total cost of the whole partition procedure. We let $x \in \mathbb{R}^n$ be a partition for these n items, in particular, each $x_i = 1 \text{or} - 1$ for $i = 1, \cdots, n$, where 1 and -1 represent 2 different sets.

Let W_{ij} be the cost of placing the items i and j in the same set, and $-W_{ij}$ be the cost of having i and j in different sets. So we define a matrix W that contains these cost.



With this problem setting, this question can be formulated as a non convex optimization problem as follows

$$\begin{aligned} \min \quad & x^T W x \\ s.t. \quad & x_i \in \{1, -1\}, \ \forall i = 1, \cdots, n \end{aligned}$$

Equivalently, we have

$$\begin{aligned} \min \quad & x^T W x \\ s.t. \quad & x_i^2 = 1, \ \forall i = 1, \cdots, n \end{aligned}$$

Notice that

$$x^T W x = \sum_{i=1}^{n} W_{ii}(x_i)^2 + \sum_{i \neq j} (W_{ij} + W_{ji}) x_i x_j$$

and thus the original formulation can be expressed as

$$\begin{aligned} \min \quad & \text{trace}(WX) \\ s.t. \quad & X_{ii} = 1, \ i = 1, \cdots, n \\ & X \geq 0 \\ & \text{rank}(X) = 1 \end{aligned}$$

By relaxation, we drop the constraint $\text{rank}(X) = 1$ so that we have a SDP problem.

10
Duality

In optimization theory, duality or the duality principle is the principle that optimization problems may be viewed from either of two perspectives, that is, the primal problem or the dual problem.

Let's consider the primal problem formulated as follows,

$$\begin{aligned} \min \quad & F_0(x) \\ s.t. \quad & F_i(x) \leq 0, i = 1, ..., m \\ & h_i(x) = 0, i = 1, ..., p \end{aligned}$$

Note that we do not have any assumptions of convexity here.

So the feasible set for this problem is

$$D = (\cap_{i=1}^{m} \text{dom } F_i) \cap (\cap_{i=1}^{p} \text{dom } h_i)$$

and the optimal value is p^*, optimal variable is x^*.

Definition 10.1 (The Lagrangian Function)**.** We define the Lagrangian function as follows,

$$L(x, \lambda, \nu) := F_0(x) + \sum_{i=1}^{m} \lambda_i F_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

where

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{bmatrix}, \; \nu = \begin{bmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_p \end{bmatrix}$$

The pairs (λ, ν) are called the "Lagrange multipliers" or "dual variables", and the domain for the Lagrangian function is given by

$$\text{dom } L = D \times \mathbb{R}^m \times \mathbb{R}^p$$

Definition 10.2 (The "dual" function)**.** The dual function $g(\cdot, \cdot)$ is defined as

$$g(\lambda, \nu) = \min_{x \in D} \quad L(x, \lambda, \nu)$$

Note: removes dependence on x.

Definition 10.3. The dual optimization problem is formulated as

$$\begin{aligned} \max_{\lambda,\nu} \quad & g(\lambda, \nu) \\ s.t. \quad & \lambda \geq 0 \end{aligned}$$

Note: ν_i are unconstrained, and we denote the optimal value for dual problem as d^*, optimal dual variables as λ^* and ν^*.

Duality theory

The duality theory says that, **for most convex optimization problem**, we have $d^* = p^*$, that is, the primal optimum equals to the dual optimum.

Recall the problem formulations previously, the primal optimization problem is formulated as

$$\begin{aligned} \min \quad & F_0(x) \\ s.t. \quad & F_i(x) \leq 0, \quad i = 1, ..., m \\ & h_i(x) = 0, \quad i = 1, ..., p \end{aligned}$$

The Lagrange function is given by

$$L(x, \lambda, \nu) = F_0(x) + \sum_{i=1}^{m} \lambda_i F_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

The dual function is given by

$$g(\lambda, \nu) = \min_{x \in D} \quad L(x, \lambda, \nu)$$

So the dual optimization problem is formulated as

$$\begin{aligned} \max \quad & g(\lambda, \nu) \\ s.t. \quad & \lambda \geq 0 \end{aligned}$$

A few observations

1. $g(\lambda, \nu)$ is concave in (λ, ν) for all $F_0, ..., F_m$, $h_0, ..., h_p$.

   Proof. Recall that

   $$g(\lambda, \nu) = \min_{x \in D} [F_0(x) + \sum_{i=1}^{m} \lambda_i F_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)]$$

   First, notice that Lagrange function is an affine function in (λ, ν) so it is concave(of course it is convex at the same time). Secondly, note that the dual function is a pointwise infimum of a family of affine functions in (λ, ν), and thus $g(\lambda, \nu)$ is concave. □

2. For any primal feasible x (i.e., $F_i(x) \leq 0, \forall i = [m]$, $h_i(x) = 0, \forall i = [p]$ and dual feasible (λ, ν) (i.e., $\lambda \geq 0$), we have

$$g(\lambda, \nu) \leq F_0(x)$$

for any tuple $(x, \lambda, \nu) \in \mathcal{C} \times \mathbb{R}_+^m \times \mathbb{R}^p$, where $\mathcal{C}$ is the feasible set of the primal problem (contains all feasible x).

Proof. Notice that, we have

$$\begin{aligned} F_0(x) &\geq F_0(x) + \sum_{i=1}^{m} \lambda_i F_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \\ &\geq \min_{x \in D} [F_0(x) + \sum_{i=1}^{m} \lambda_i F_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)] \\ &= g(\lambda, \nu) \end{aligned}$$

where the first summation on r.h.s is negative due to $\lambda_i \geq 0$ and $F_i(x) \leq 0$, and the second summation equals to zero due to $h_i(x) = 0$.

Thus the desired result can be obtained by the definition of min function and dual function. □

The point of greatest interest is x^*, where $p^* = F_0(x^*)$.

Plug in to the above inequality, for all dual feasible (λ, ν) (i.e., for $\lambda \geq 0$), we have

$$p^* = F_0(x^*) \geq g(\lambda, \nu)$$

Optimize over (λ, ν) where $\lambda \geq 0$ in order to maintain dual feasibility, we can get the greatest lower bound,

$$p^* = F_0(x^*) \geq g(\lambda^*, \nu^*) = d^*$$

That is, we have the so called **weak duality**, $p^* \geq d^*$.

Furthermore, we refer to the difference $p^* - d^*$ as the **optimal duality gap**.

3. For convex primal optimization problems, (i.e., $F_i(x)$ are convex and $h_i(x)$ are affine) and under certain conditions called "constraint qualification" (i.e., not all constraint sets allowed), the **strong duality** holds, i.e.,

$$p^* = d^*$$

and thus the optimal duality gap is zero.

There are many types of constraint qualification, and we will introduce a simple one called Slater's condition in the next section.

Slater Conditions

Definition 10.4 (Slater conditions)**.** Consider a primal problem with a set of constraints $F_i(x) \leq 0, i = [m]$ and $Ax = b$, it is said to be satisfied the Slater's conditions if there exists an $x \in$ relint D such that

1. $F_i(x) < 0, \; \forall i = [m]$

2. $Ax = b$

Furthermore, if some of the inequality constraints are defined by affine functions, this conditions can be weaken a bit. Suppose F_i are affine for $i = 1, ..., k$, where $k < m$, then the Slater conditions requires that there exists an $x \in$ relint D such that

1. $F_i(x) \leq 0, \; \forall i = 1, ..., k$

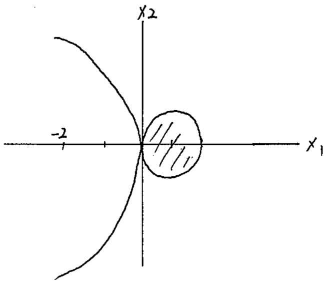
2. $F_i(x) < 0, \; \forall i = k + 1, ..., m.$

3. $Ax = b$

Example 10.5. Convex problem that doesn't satisfy Slater's:

$$\begin{bmatrix}(x_1 - 1) & x_2\end{bmatrix} \begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix} \begin{bmatrix}x_1 - 1 \\ x_2\end{bmatrix} \leq 1$$

$$\begin{bmatrix}(x_1 + 2) & x_2\end{bmatrix} \begin{bmatrix}1 & 0 \\ 0 & 0\end{bmatrix} \begin{bmatrix}x_1 + 2 \\ x_2\end{bmatrix} \leq y$$



Feasible set is $(x_1, x_2) = \{(0, 0)\}$

Theorem 10.6. *If the primal optimization problem is convex and satisfies Slater's conditions, then $p^* = d^*$, the strong duality holds.*

Proof. We propose a sketch proof for the case $m = 1$, i.e., 1 inequality constraint and there is no equality constraint so $p = 0$.

Given the basic setting for this case, the primal problem is given by

$$\begin{aligned} \min \quad & F_0(x) \\ s.t. \quad & F_1(x) \leq 0 \end{aligned}$$

and we let p^* be the optimal value of the primal problem.

The Lagrange function is:

$$L(x, \lambda) = F_0(x) + \lambda F_1(x)$$

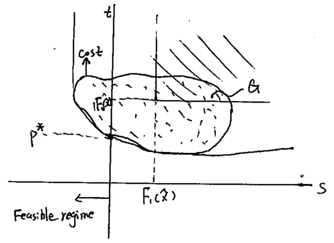
The dual function is $g(\lambda) = \min_{x \in D} L(x, \lambda) = \min_x F_0(x) + \lambda F_1(x)$

The dual optimal problem is formulated as

$$\begin{aligned} \max \quad & g(\lambda) \\ s.t. \quad & \lambda \geq 0 \end{aligned}$$

and we let d^* be the optimal value of the dual problem.

To start, we define a set



$$\begin{aligned} G &= \{(F_1(x), F_0(x)) | x \in D = \text{dom } F_1 \cap \text{dom } F_0\} \\ &= \cup_{x \in D} \{(F_1(x), F_0(x))\} \end{aligned}$$

and also define the set $\mathcal{A}$

$$\begin{aligned} \mathcal{A} &= G + \mathbb{R}_+ \times \mathbb{R}_+ \\ &= \{(s,t) | F_1(x) \leq s, F_0(x) \leq t, x \in D\} \\ &= \cup_{x \in D} \{(s,t) | F_1(x) \leq s, F_0(x) \leq t\} \end{aligned}$$

A few observations regarding these two sets:

- The set G contains all information about primal problem.
- The set $\mathcal{A}$ contains all points "above" and to "right" of each point in G.
- Each such point(above and to right) is less interesting than the point in G due to

  (1) Perhaps higher cost

  (2) Perhaps more resources

The boundary of $\mathcal{A}$ is specified by the function:

$$\begin{aligned} p(u) = \min \quad & F_0(x) \\ s.t. \quad & F_1(x) \leq u \end{aligned}$$

(p is the boundary, $\mathcal{A}$ lies above the boundary)

A few observations regarding $p(u)$:
(1) $p^* = p(0)$.
(2) p is non-increasing in u
(3) p is convex in u
(4) $\mathcal{A} = \text{epi } p$

proof of (1). When $u = 0$, we just get the original primal problem so certainly $p^* = p(0)$. □

proof of (2). As u gets larger, feasible set of the $p(u)$ optimization gets larger so objective cannot increase → therefore non-increasing. □

proof of (3). We want to prove the convexity of the problem

$$\begin{aligned} p(u) = \min \quad & F_0(x) \\ s.t. \quad & F_1(x) \leq u \end{aligned}$$

and this means that we need to show $\forall u_1, u_2 \in \text{dom } p$, $\forall \lambda \in [0,1]$, we have

$$p(\lambda u_1 + (1-\lambda)u_2) \leq \lambda p(u_1) + (1-\lambda)p(u_2).$$

Consider $i = 1, 2$, let

$$\begin{aligned} x_i = \arg\min \quad & F_0(x) \\ s.t. \quad & F_1(x) \leq u_i \end{aligned}$$

That is, $F_0(x_1) = p(u_1)$ and $F_0(x_2) = p(u_2)$.
Let $\tilde{x} = \lambda x_1 + (1-\lambda)x_2$, and note that

$$x_1 \in \text{dom } F_1 \cap \text{dom } F_0 \cap \{x | F_1(x) \leq u_1\}$$

$$x_2 \in \text{dom } F_1 \cap \text{dom } F_0 \cap \{x | F_1(x) \leq u_2\}$$

So we can write

$$\begin{aligned} F_1(\tilde{x}) &= F_1(\lambda x_1 + (1-\lambda)x_2) \\ &\leq \lambda F_1(x_1) + (1-\lambda)F_1(x_2) \\ &\leq \lambda u_1 + (1-\lambda)u_2 \end{aligned}$$

where the first equality is due to $\tilde{x} \in \text{dom } F_1$, the first inequality is due to convexity of F_1, and the second inequality is due to $F_1(x_i) \leq u_i, i = 1, 2$.

Hence,

$$\tilde{x} \in \text{dom } F_1 \cap \text{dom } F_0 \cap \{x | F_1(x) \leq \lambda u_1 + (1-\lambda)u_2\}$$

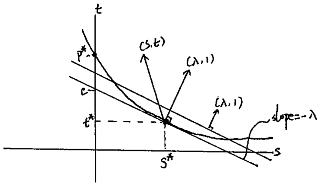
Therefore, $\tilde{x}$ is a feasible point for the optimization problem $p(\lambda u_1 + (1-\lambda)u_2)$ □

Think about the trade-off between $F_1(x)$ and $F_0(x)$ in a slightly different way:

$$\begin{aligned} \min_{(s,t)} \quad & \lambda s + t \\ \text{where} \quad & (s,t) \in \mathcal{A} \end{aligned}$$

which is equivalent to

$$\begin{aligned} \min_{(s,t)} \quad & \begin{bmatrix} \lambda & 1 \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} \\ \text{where} \quad & (s,t) \in \mathcal{A} \end{aligned}$$



For a given λ, the optimum is attained by some point (s^*, t^*) on boundary of $\mathcal{A}$. Point on boundary is a function of λ, so we can write

$$(s^*(\lambda), t^*(\lambda)) = (F_1(x^*(\lambda)), F_0(x^*(\lambda)))$$

Consider any point $(s,t) \in \mathcal{A}$:

$$\begin{aligned} &\begin{bmatrix} \lambda & 1 \end{bmatrix} \begin{bmatrix} F_1(x^*(\lambda)) \\ F_0(x^*(\lambda)) \end{bmatrix} \leq \begin{bmatrix} \lambda & 1 \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} \qquad (*) \\ \Leftrightarrow 0 \leq &\begin{bmatrix} \lambda & 1 \end{bmatrix} \left(\begin{bmatrix} s \\ t \end{bmatrix} - \begin{bmatrix} F_1(x^*(\lambda)) \\ F_0(x^*(\lambda)) \end{bmatrix} \right) \end{aligned}$$

1. This optimization yields a tangent plane, "supporting hyperplane"
2. In this $2-D$ picture, supporting hyper-plane is a line, change $'' \leq ''$ in $(*)$ to $'' = ''$ get a line

   $$c = \lambda s + t$$

   where c is the l.h.s of $(*)$.

   Rearrange yields $t = c - \lambda s$.

   This is the problem we just talked about:

$$\begin{aligned} \min_{(s,t)\in\mathcal{A}} \quad \lambda s + t &= \lambda s^* + t^* \\ &= \lambda F_1(x^*(\lambda)) + F_0(x^*(\lambda)) \\ &= \min_{x \in D} \quad [\lambda F_1(x) + F_0(x)] \\ &= g(\lambda) \end{aligned}$$

so it is the dual function.

Put these pieces all together:

(1) The dual function $g(\lambda)$ specifies the $y-$intercept of the tangent line of slope $-\lambda$.

(2) Last time proved $g(\lambda)$ is a lower bound on p^* as long as λ are dual-feasible (i.e., $\lambda \geq 0$).

(3) The $y-$intercept is a lower bound on p^*, i.e., $c \leq p^*$.

Get best lower bound by maximizing $g(\lambda)$ over $\lambda \geq 0$, that is, solve the following optimization problem

$$\begin{aligned} \max \quad & g(\lambda) \\ s.t. \quad & \lambda \geq 0 \end{aligned}$$

which exactly takes the form of dual problem.

□

"Pricing" Interpretation

There is an interesting and intuitive interpretation for the duality theory called the pricing interpretation. Suppose the variable x denotes how an company operates(i.e., "policy") and $F_0(x)$ denotes the cost of operating at policy x. Each constraint denotes representing some limit, such as a limit on resources, labor, etc.

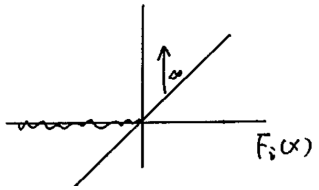
To optimize policy(i.e., x^*) with these constraints can be found by solving the problem(consider this is the primal problem)

$$\begin{aligned} \min \quad & F_0(x) \\ s.t. \quad & F_i(x) \leq 0, \quad i = 1, ..., m \\ & h_i(x) = 0, \quad i = 1, ..., p \end{aligned}$$

In the next step, we reformulate this problem as an unconstrained problem by introducing I and $\tilde{I}$, which are given by

$$I(x) = \begin{cases} 0, & x \leq 0 \\ \infty, & x > 0 \end{cases} \tag{10.1}$$

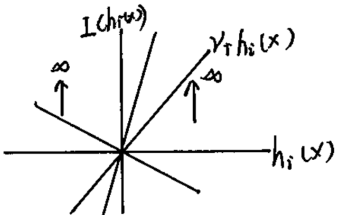
$$\tilde{I}(x) = \begin{cases} 0, & x = 0 \\ \infty, & \text{else} \end{cases} \tag{10.2}$$



So the primal problem can be reformulated as

$$\min_x \quad F_0(x) + \sum_{i=1}^{m} I(F_i(x)) + \sum_{i=1}^{p} \tilde{I}(h_i(x)) \qquad (*)$$

This unconstrained problem is the same as primal one, but with "hard" penalties introducing by I and $\tilde{I}$.



Now, let's consider the case that are more realistic. Suppose the company is allowed to break the limit on their resources by paying an additional cost which is linear in the amount of violation, measured by F_i and h_i. More precisely, the additional payment made by the company for the i-th constraint is $\lambda_i F_i(x)$, and payments are also made to the company for the constraints that are not right(i.e., $F_i(x) < 0$), the $\lambda_i F_i(x)$ represents a payment received by the company.

So λ_i is interpreted as the "price" for violating constraint F_i, and similarly we have μ_i as the "price" for violating constraint h_i.

Under this relaxed setting (i.e., not all constraints are satisfied), the problem can be formulated as

$$F_0(x) + \sum_{i=1}^{m} \lambda_i F_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

It's obviously that such formulation takes the from of Lagrange function. If we minimize this function to obtain the minimal total cost and then maximize over λ and ν, that is

$$\max_{\lambda,\nu} \min_{x} F_0(x) + \sum_{i=1}^{m} \lambda_i F_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

where we obtain the optimal cost to the company under the least favorable set of prices. We use d^* to denote this optimal value, and it readily follows that we have an interpretation for the weak duality, i.e.,

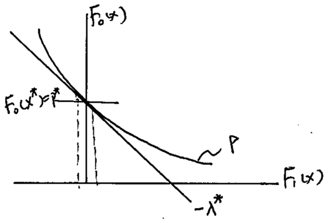
$$d^* \leq p^*$$

Furthermore, if strong duality holds, the problem $(*)$ (i.e., problem that is equivalent to the primal problem) becomes

$$\begin{aligned}(*) &= \max_{\lambda,\nu,\lambda \geq 0} \quad [\min_{x} \quad F_0(x) + \sum_{i=1}^{m} \lambda_i F_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)] \\ &= \max_{\lambda,\nu} g(\lambda, \nu) \quad \text{where } \lambda \geq 0\end{aligned}$$

Implication: Adjust the prices λ, ν so that the solution to relaxed problem matches the solution to the primal problem.

Sensitivity Analysis



→ At dual optimum slope of tangent is $-\lambda^*$

→ If change constraint by ϵ, optimum value will change by something like $(-\epsilon\lambda^*)$.

We consider the following perturbed version of the original optimization problem (unperturbed), says the perpetuated problem, as follows

$$\begin{aligned} p^*(u, v) = \min \quad & F_0(x) \\ s.t. \quad & F_i(x) \leq u_i, \quad i = 1, ..., m \\ & h_i(x) = v_i, \quad i = 1, ..., p \end{aligned}$$

where when $u = v = 0$ is exactly the same as the unperturbed one.

→ $u_i < 0$ "tighten" constraint, $u_i > 0$ loosen constraint, $v_i \neq 0$ change set-point.

→ Same as $p(u)$ in last lecture.

→ Note $p^*(0, 0) = p^*$ is the optimal value of unperturbed problem.

→ relate $p^*(u, v)$ to $p^*(0, 0)$

- Let (λ^*, ν^*) be optimal dual variables for unperturbed problem

- Consider a convex optimization problem satisfying Slater's conditions, i.e., strong duality holds,

$$\begin{aligned} p^*(0,0) &= g(\lambda^*,\nu^*) \\ &= \min_{x\in D} L(x,\lambda^*,\nu^*) \\ &\leq F_0(x) + \sum_{i=1}^{m} \lambda_i^* F_i(x) + \sum_{i=1}^{p} \nu_i^* h_i(x) \\ &\leq F_0(x) + \sum_{i=1}^{m} \lambda_i^* u_i + \sum_{i=1}^{p} \nu_i^* v_i \\ &= F_0(x) + (\lambda^*)^{\mathrm{T}} u + (\nu^*)^{\mathrm{T}} v \end{aligned}$$

Our focus in on $x \in D$ s.t. x is optimal for perturbed problem
i.e. $F_0(x) = p^*(u,v)$
$\Rightarrow p^*(u,v) \geq p^*(0,0) - (\lambda^*)^{\mathrm{T}} u - (\nu^*)^{\mathrm{T}} v$

1. E.g. If $\lambda_i >> 0$ and tighten constraint F_i slightly so that $F_i(x) \leq -\epsilon < 0$

$$p^*(u,v) \geq p^*(0,0) - (\lambda^*)^{\mathrm{T}} u - (\nu^*)^{\mathrm{T}} v$$

2. Note not symmetric in general, big relaxation in constraint doesn't necessary mean big drop in cost.

3. If $p(u,v)$ is differentiable, then have symmetry for small perturbations.

Lagrange Method

In order to solve the following primal problem,

$$\begin{aligned} \min_{x} \quad & F_0(x) \\ s.t. \quad & F_i(x) \leq 0, \quad i = 1, ..., m \end{aligned}$$

we propose the Lagrange method as follows:

Step 1. First write Lagrangian: $L(x,\lambda) = F_0(x) + \sum_{i=1}^{m} \lambda_i F_i(x)$.

Step 2. Solve for dual function $g(\lambda) = \min_{x\in D} L(x,\lambda)$, where D is the primal feasible set.

Step 3. Find $\lambda^* = \arg\max g(\lambda)$, s.t. $\lambda \geq 0$.

Step 4. Recover the primal optimal x^* by solving $\arg\min L(x,\lambda^*)$, that is, solve

$$\arg\min F_0(x) + \sum_{i=1}^{m} \lambda_i^* F_i(x).$$

Remarks:

- It is a nice approach if the problem has a nice structure, in particular if it is easy to solve for (λ^*,ν^*) analytically or numerically.

- Even if the dual optimum (λ^*, ν^*) is unique, the primal optimum x^* that minimized $L(x, \lambda^*, \nu^*)$ may not be unique.

Example 10.7 (Lagrange Duality for LS problems)**.** Consider the problem

$$\begin{aligned} \min_x \quad & \|x\|_2^2 \\ s.t. \quad & Ax = b \end{aligned}$$

where $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{p \times n}$, $\operatorname{rank}(A) = p < n$.

Recall the chapter least square, this an under determined LS problem, and the optimal solution is given by

$$x^* = A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}b$$

To verify that this solution is coincide with the one obtained by Lagrange's method, we proceed following procedure to solve this question.

(1) Form the Lagrange function:

$$L(x, \nu) = x^{\mathrm{T}}x + \nu^{\mathrm{T}}(Ax - b)$$

(2) Solve for the dual function $g(\lambda) = \min_x L(x, \nu)$:

Notice that Lagrange function here is a convex quadratic function of x(you may verify this), so simply by the first-order condition, we have

$$\frac{\partial}{\partial x}L(x, \nu) = 2x + A^{\mathrm{T}}\nu = 0 \Rightarrow x^*(\nu) = -\frac{1}{2}A^{\mathrm{T}}\nu$$

(3) Find the dual optimum ν^*:

Now, we have dual problem as

$$\begin{aligned} \max_\nu g(\nu) &= \max_\nu L(x^*(\nu), \nu) \\ &= \max_\nu [x^*(\nu)^{\mathrm{T}}x^*(\nu) + \nu^{\mathrm{T}}(Ax^*(\nu) - b)] \\ &= \max_\nu [\frac{1}{4}\nu^{\mathrm{T}}AA^{\mathrm{T}}\nu + \nu^{\mathrm{T}}(A(-\frac{1}{2}A^{\mathrm{T}}\nu) - b)] \\ &= \max_\nu [\frac{1}{4}\nu^{\mathrm{T}}AA^{\mathrm{T}}\nu - \frac{1}{2}\nu^{\mathrm{T}}AA^{\mathrm{T}}\nu - \nu^{\mathrm{T}}b] \\ &= \max_\nu [-\frac{1}{4}\nu^{\mathrm{T}}AA^{\mathrm{T}}\nu - \nu^{\mathrm{T}}b] \end{aligned}$$

Note that $g(\nu)$ is a concave quadratic function of ν, and therefore utilize the first-order condition we yields

$$\begin{aligned} & \frac{\partial}{\partial \nu}(-\frac{1}{4}\nu^{\mathrm{T}}AA^{\mathrm{T}}\nu - \nu^{\mathrm{T}}b) = 0 \\ \Leftrightarrow & -\frac{1}{4}2AA^{\mathrm{T}}\nu - b = 0 \\ \Leftrightarrow & (AA^{\mathrm{T}})\nu = -2b \\ \Leftrightarrow & \nu^* = -2(AA^{\mathrm{T}})^{-1}b \end{aligned}$$

(4) Substitute into $x^*(\nu)$ to get the primal optimum:

$$\begin{aligned} x^*(\nu^*) &= -\frac{1}{2}A^{\mathrm{T}}v^* \\ &= -\frac{1}{2}A^{\mathrm{T}}(-2(AA^{\mathrm{T}})^{-1}b) \\ &= A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}b \end{aligned}$$

Hence, the optimal solution is coincide with our previous result in LS chapter, that is, for under-determined LS problem we have $x^* = A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}b$.

Furthermore, at step (3), we have the problem

$$\max_{\nu}[-\frac{1}{4}\nu^{\mathrm{T}}AA^{\mathrm{T}}\nu - \nu^{\mathrm{T}}b]$$

which is equivalent to

$$\min_{\nu}[\frac{1}{4}\nu^{\mathrm{T}}AA^{\mathrm{T}}\nu + \nu^{\mathrm{T}}b]$$

and it turns out, this minimization problem is equivalent to the following norm minimization problem,

$$\begin{aligned} \min_{\nu} \quad & \|\frac{1}{2}A^{\mathrm{T}}\nu + x_0\|_2^2 \\ s.t. \quad & Ax_0 = b \end{aligned}$$

since they enjoy the same optimal solution x^*, and the difference of the optimal value is just a scalar.

More precisely, that's because

$$\begin{aligned} \|\frac{1}{2}A^{\mathrm{T}}\nu + x_0\|_2^2 &= (\frac{1}{2}A^{\mathrm{T}}\nu + x_0)^{\mathrm{T}}(\frac{1}{2}A^{\mathrm{T}}\nu + x_0) \\ &= \frac{1}{4}\nu^{\mathrm{T}}AA^{\mathrm{T}}\nu + 2\frac{1}{4}\nu^{\mathrm{T}}Ax_0 + x_0^{\mathrm{T}}x_0 \\ &= \frac{1}{4}\nu^{\mathrm{T}}AA^{\mathrm{T}}\nu + \nu^{\mathrm{T}}b + x_0^{\mathrm{T}}x_0 \end{aligned}$$

A final interpretation

Consider the problem

$$\begin{aligned} \min \quad & F_0(x) \\ s.t. \quad & F_i(x) \leq 0 \quad i = 1, ..., m \end{aligned}$$

We want connect to the problems with multiple (vector) objective $(F_0, F_1, \cdots, F_m)$, and one approach is to "scalarize" the objective as

$$F_0(x) + \lambda_1 F_1(x) + \cdots + \lambda_m F_m(x) = F_0(x) + \sum_{i=1}^{m} \lambda_i F_i(x)$$

Dual of LPs

Consider an LP problem as the primal problem, which is given by

$$\begin{aligned} \min \quad & c^{\mathrm{T}} x \\ s.t. \quad & Ax \le b \end{aligned}$$

where the constraint is the matrix form of $a_i^{\mathrm{T}} x \le b_i$, for $i = 1, ..., m$.

The Lagrange function is

$$\begin{aligned} L(x, \lambda) &= c^{\mathrm{T}} x + \sum_{i=1}^{m} \lambda_i (a_i^{\mathrm{T}} x - b_i) \\ &= c^{\mathrm{T}} x + \begin{bmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_m \end{bmatrix} \begin{bmatrix} a_i^{\mathrm{T}} x - b_1 \\ \vdots \\ a_m^{\mathrm{T}} x - b_m \end{bmatrix} \\ &= c^{\mathrm{T}} x + \lambda^{\mathrm{T}} (Ax - b) \\ &= -\lambda^{\mathrm{T}} b + (c^{\mathrm{T}} + \lambda^{\mathrm{T}} A) x \end{aligned}$$

The dual function is

$$\begin{aligned} g(\lambda) &= \min_x L(x, \lambda) \\ &= \min_x [-\lambda^{\mathrm{T}} b + (c^{\mathrm{T}} + \lambda^{\mathrm{T}} A) x] \\ &= \begin{cases} -\lambda^{\mathrm{T}} b & \text{if} \quad c^{\mathrm{T}} + \lambda^{\mathrm{T}} A = 0 \\ -\infty & \text{if} \quad c^{\mathrm{T}} + \lambda^{\mathrm{T}} A \neq 0 \end{cases} \end{aligned}$$

The dual optimization problem is formulated as

$$\begin{aligned} \max \quad & g(\lambda) \\ s.t. \quad & \lambda \ge 0 \end{aligned}$$

As we showed above, the dual problem is feasible iff $c^{\mathrm{T}} + \lambda^{\mathrm{T}} A = 0$, so in order to maintain the feasibility we may consider the problem with the form

$$\begin{aligned} \max \quad & g(\lambda) \\ s.t. \quad & \lambda \ge 0 \\ & c^{\mathrm{T}} + \lambda^{\mathrm{T}} A = 0 \end{aligned}$$

which is equivalent to

$$\begin{aligned} \max \quad & -\lambda^{\mathrm{T}} b \\ s.t. \quad & \lambda \ge 0 \\ & c^{\mathrm{T}} + \lambda^{\mathrm{T}} A = 0 \end{aligned}$$

Observations:

- The dual of an LP problem is still an LP problem.
- It maybe not clear from the form, but we have strong duality holds, so $p^* = d^*$.

	Primal	Dual
number of variables	n	m
number of constraints	m	$n+m$

To verify above results, we show that, for this LP case, the dual of the dual problem is just the given primal problem (i.e., we retrieve the original primal problem).

The dual optimization problem we obtain above can be expressed as follows

$$\begin{aligned} \min \quad & b^{\mathrm{T}}\lambda \\ s.t. \quad & -\lambda \leq 0 \\ & A^{\mathrm{T}}\lambda = -c \end{aligned}$$

which has the same optimal solution.

Let z_i be the inequalities multipliers, and y_i be the equality multipliers.

The Lagrange function is

$$L(\lambda, x, y) = b^{\mathrm{T}}\lambda + z^{\mathrm{T}}(-\lambda) + y^{\mathrm{T}}(A^{\mathrm{T}}\lambda + c)$$

The dual function is

$$\begin{aligned} g(z,y) &= \min_{\lambda} L(\lambda, z, y) \\ &= \min_{\lambda} y^{\mathrm{T}}c + (b^{\mathrm{T}} - z^{\mathrm{T}} + y^{\mathrm{T}}A^{\mathrm{T}})\lambda \\ &= \begin{cases} y^{\mathrm{T}}c & \text{if } b^{\mathrm{T}} - z^{\mathrm{T}} + y^{\mathrm{T}}A^{\mathrm{T}} = 0 \\ -\infty & \text{else} \end{cases} \end{aligned}$$

The dual problem is

$$\begin{aligned} \max \quad & g(z,y) \\ s.t. \quad & z \geq 0 \end{aligned}$$

To main the feasibility, we may have the form

$$\begin{aligned} \min \quad & y^{\mathrm{T}}c \\ s.t. \quad & b^{\mathrm{T}} - z^{\mathrm{T}} + y^{\mathrm{T}}A^{\mathrm{T}} = 0 \\ & z \geq 0 \end{aligned}$$

Notice that, the constraints

$$\begin{aligned} b^{\mathrm{T}} - z^{\mathrm{T}} + y^{\mathrm{T}}A^{\mathrm{T}} &= 0 \\ z &\geq 0 \end{aligned}$$

is equivalent to

$$b^{\mathrm{T}} + y^{\mathrm{T}} A^{\mathrm{T}} \geq 0$$

by a simple rearrangement and substitution.

Thus, this dual problem can also be written as:

$$\begin{aligned} \max \quad & y^{\mathrm{T}} c \\ s.t. \quad & b + Ay \geq 0 \end{aligned}$$

$\Leftrightarrow$

$$\begin{aligned} \max \quad & (-x)^{\mathrm{T}} c \\ s.t. \quad & b + A(-x) \geq 0 \end{aligned}$$

$\Rightarrow$

$$\begin{aligned} \min \quad & c^{\mathrm{T}} x \\ s.t. \quad & Ax \leq b \end{aligned}$$

Note that the second form is just utilizing the change of variable and let $-x = y$, and the third one has exactly the same optimal solution(but not optimal value), and the optimal value have a opposite sign.

By this argument, we show that, the dual of the dual problem is just the original LP problem, it implies the strong duality holds by the validation of weak duality(i.e., use weak duality for twice), since the two inequality holds only if the equality holds.

Karush-Kuhn-Tucker(KKT) conditions

Consider the optimization problem for which primal and dual optimal values are obtained and $p^* = d^*$ (i.e., strong duality holds).

Let x^* be the primal optimum, and (λ^*, ν^*) be the dual optimum.

Consider the primal problem:

$$\begin{aligned} \max \quad & F_0(x) \\ s.t. \quad & F_i(x) \leq 0 \quad i = 1, ..., m \\ & h_i(x) = 0 \quad i = 1, ..., p \end{aligned}$$

Note: we have no assumption of convexity for this problem.

The Lagrange function and the dual function are given by

$$\begin{aligned} L(x, \lambda, \nu) &= F_0(x) + \sum_{i=1}^{m} \lambda_i F_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \\ g(\lambda, \nu) &= \min_x L(x, \lambda, \nu) \end{aligned}$$

Since strong duality holds at (x^*, λ^*, ν^*),

$$\begin{aligned} F_0(x^*) &= g(\lambda^*, \nu^*) \\ &= \min_x [F_0(x) + \sum_{i=1}^{m} \lambda_i^* F_i(x) + \sum_{i=1}^{p} \nu_i^* h_i(x)] \\ &\leq F_0(x^*) + \sum_{i=1}^{m} \lambda_i^* F_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) \\ &\leq F_0(x^*) \end{aligned}$$

The first inequality is because the primal optimal x^* must minimize $L(x^*, \lambda^*, \nu^*)$, and the second inequality is due to the negativity of the second term on r.h.s.

We claim that the second inequity must hold at equality, since $F_0(x^*) = F_0(x^*)$, and an interesting result can be derived from this fact.

Complementary slackness:

The complementary slackness condition

$$\sum_{i=1}^{m} \lambda_i F_i(x^*) = 0, \ \forall i \in [m]$$

holds for any primal optimal x^* and any dual optimal λ^*, when the strong duality holds.

We can express the complementary slackness condition as

(1) If i-th constraint is "active" (i.e., $F_i(x) = 0$), then $\lambda_i^* = 0$ or $\lambda_i^* > 0$.

(2) If i-th constraint is "inactive" (i.e., $F_i(x) < 0$), then $\lambda_i^* = 0$.

Conversely,

(3) If $\lambda_i^* = 0$, then $F_i(x^*) = 0$ or $F_i(x^*) < 0$.

(4) If $\lambda_i^* > 0$, then $F_i(x^*) = 0$.

Recall the pricing interpretation,

→ perturb $F_i(x) \leq 0$ to $F_i(x) \leq \epsilon$

→ $\lambda_i^* = -\frac{d}{d\epsilon} p^*(0,0)$

If constraint is "inactive", then there is slack in resource i but since $\lambda_i^* = 0$, no gain by having more of that resource.

If constraint is "active", then resource is totally used, so cannot use more even if you want to.

We can say more if the problem is differentiable, assume that

- $F_i(x)$ and $h_i(x)$ are all differentiable.
- Strong convexity holds.

Observe that x^* minimizes $L(x, \lambda^*, \nu^*)$, and since Lagrangian function is differentiable, we have the first order condition:

$$\nabla_x L(x, \lambda^*, \nu^*)|_{x=x^*} = 0$$

Thus, we have the **KKT conditions**:

1. $\nabla_x L(x^*, \lambda^*, \nu^*) = \nabla F_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla F_i(x^*) + \sum_{i=1}^p \nu_i^* \nabla h_i(x^*) = 0$
2. $F_i(x^*) \le 0, \forall i = [m]$, $h_i(x^*) = 0, \forall i = [p]$.
3. $\lambda_i^* \ge 0, \forall i = [m]$.
4. $\lambda_i^* F_i(x_i^*) = 0, \forall i = [m]$.

Theorem 10.8. *If (x^*, λ^*, ν^*) are primal and dual optimal, for a differentiable problem for which strong duality holds, they must satisfy KKT conditions. Note that this is necessary but not sufficient.*

Theorem 10.9. *When the primal problem is convex, the KKT condtions are also sufficient for the points to be primal and dual optimal with zero duality gap, that is,*
If
(1) F_i and h_i are all differentiable, and
(2) F_i are convex functions and h_i are affine functions
Then for any points $(\tilde{x}, \tilde{\lambda}, \tilde{v})$ satisfy the KKT conditions, we may have
(1) $\tilde{x}$ is primal optimal, $(\tilde{\lambda}, \tilde{v})$ is dual optimal, and
(2) Duality gap is zero (i.e., Strong duality holds).

Proof. Let $(\tilde{x}, \tilde{\lambda}, \tilde{v})$ be a point that satisfy the KKT conditions.

Notice that, the Lagrange function

$$L(x, \tilde{\lambda}, \tilde{v}) = F_0(x) + \sum_{i=1}^m \tilde{\lambda}_i F_i(x) + \sum_{i=1}^p \tilde{v}_i h_i(x)$$

is a convex function in x.

Therefore, if we can find a point of zero-gradient, that is the global optimum.

The KKT-(1) tells us that , for the point $(\tilde{x}, \tilde{\lambda}, \tilde{v})$ we have

$$\nabla_x L(\tilde{x}, \tilde{\lambda}, \tilde{v}) = 0$$

Hence, $\tilde{x}$ minimizes $L(x, \tilde{\lambda}, \tilde{v})$.

So the dual function is given by

$$\begin{aligned} g(\tilde{\lambda}, \tilde{v}) &= \min_x L(x, \tilde{\lambda}, \tilde{v}) \\ &= L(\tilde{x}, \tilde{\lambda}, \tilde{v}) \\ &= F_0(\tilde{x}) + \sum_{i=1}^m \tilde{\lambda}_i F_i(\tilde{x}) + \sum_{i=1}^p \tilde{v}_i h_i(\tilde{x}) \\ &= F_0(\tilde{x}) \end{aligned}$$

where the third equality comes from the KKT-(2) and KKT-(4).

Since $g(\tilde{\lambda}, \tilde{v}) = F_0(\tilde{x})$, $\tilde{x}$ and $(\tilde{\lambda}, \tilde{v})$ have zero duality gap so the strong duality holds, and therefore $\tilde{x}$ is primal optimal and $(\tilde{\lambda}, \tilde{v})$ is dual optimal. □

So far, we have proposed two theorems regarding KKT conditions. One is necessary and the other one is sufficient. Certainly it is better if there is a theorem that is both necessary and sufficient.

To combine the above results such that KKT is necessary and sufficient, we need the following

(1) Optimization problem that is differentiable so that KKT conditions exist;

(2) Convex optimization problem so that we can apply Theorem 10.9 (sufficiently);

(3) Strong duality holds so that we can apply Theorem 10.8 (necessity).

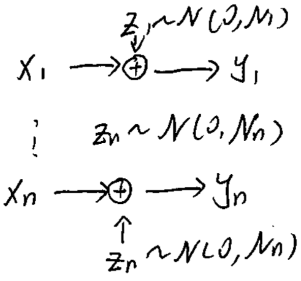
To summarize, we have the theorem,

Theorem 10.10. *If a convex optimization problem with differentiable objective and constraints functions satisfy Slater's condition(i.e., strong duality holds), then the KKT conditions provide necessary and sufficient conditions for optimality.*

Walter-pouring Problem

Given the maximization problem

$$\begin{aligned} \max_{P_1,\cdots,P_n} \quad & \sum_{i=1}^{n} log[1 + \frac{P_i}{N_i}] \\ s.t. \quad & 0 \leq P_i,\ i \in [m] \\ & \sum_{i=1}^{n} P_i \leq P_T \end{aligned}$$



We convert it to the general form (minimize the objective),

$$\begin{aligned} \min_{P_1,\cdots,P_n} \quad & -\sum_{i=1}^{n} log[1 + \frac{P_i}{N_i}] \\ s.t. \quad & 0 \leq P_i,\ i \in [m] \\ & \sum_{i=1}^{n} P_i \leq P_T \end{aligned}$$

The Lagrange function is formulated as

$$L(P, \lambda, \mu) = -\sum_{l=1}^{n} log_l[1 + \frac{P_l}{N_l}] + \lambda(\sum_{l=1}^{n} P_l - P_T) - \sum_{l=1}^{n} \mu_l P_l$$

By the KKT condtions, we have

$$(1)\quad \frac{\partial}{\partial P_i} L(P,\lambda,\mu) = -\frac{N_i}{N_i+P_i}\frac{1}{N_i} + \lambda - \mu_i = 0 \Leftrightarrow N_i + P_i = \frac{1}{\lambda-\mu_i}$$

$$(2)\quad P_i \geq 0,\ \forall i \in [m] \text{ and } \sum_{l=1}^{n} P_l \leq P_T$$

$$(3)\quad \lambda \geq 0,\ \mu \geq 0$$

$$(4)\quad \lambda(\sum_{l=1}^{n} P_l - P_T) = 0 \text{ and } \mu_i P_i = 0 \Leftrightarrow \text{if } P_i > 0, \text{then } \mu_i = 0$$

To solve this question,

1. First observe that you will use all power since F_0 is monotonically increasing in each P_i, so we have

$$\sum_{l=1}^{n} P_l = P_T$$

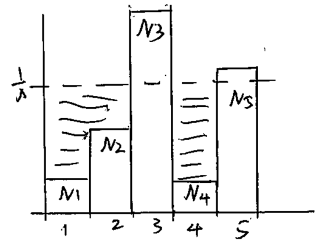
2. Look at KKT-(1):

(a) If $P_i > 0$ then $\mu_i = 0$ by KKT-(4), and therefore $N_i + P_i = \frac{1}{\lambda}$.

(b) If $P_i = 0$, by KKT-(3) $\mu_i \geq 0$, and therefore $N_i = \frac{1}{\lambda-\mu_i} \geq \frac{1}{\lambda}$.

(c) By KKT-(2), we require $P_i \geq 0$, so we conclude that

$$P_i = \max\{\frac{1}{\lambda} - N_i, 0\}$$



Let's give you $P_T + \epsilon$ power, says $|u^*| = n^*$, and we increase the power by $\frac{\epsilon}{n^*}$ to each channel (that is active),

So the change in rate for i-th channel can be computed as

$$\begin{aligned}\log[1+\frac{P_i+\frac{\epsilon}{n^*}}{N_i}] - \log[1+\frac{P_i}{N_i}] &= \log[\frac{N_i+P_i+\frac{\epsilon}{n^*}}{N_i+P_i}] \\ &= \log[1+\frac{\frac{\epsilon}{n^*}}{N_i+P_i}] \\ &= \log[1+\frac{\frac{\epsilon}{n^*}}{\frac{1}{\lambda^*}}] \\ &= \log[1+\frac{\epsilon}{n^*}\lambda^*] \\ &\approx \frac{\epsilon\lambda^*}{n^*}\end{aligned}$$

Hence, Total change in rate = $n^*\frac{\epsilon\lambda^*}{n^*} = \epsilon\lambda^*$, where λ^* can be computed by

$$\frac{\partial F_o(x^*)}{\partial P_T} = \lambda^*$$

Geometric interpretation of KKT

Consider the primal problem

$$\begin{aligned} \min \quad & F_0(x) \\ s.t. \quad & F_i(x) \leq 0 \quad i = 1, ..., m \\ & h_i(x) = 0 \quad i = 1, ..., p \end{aligned}$$

which is equivalent to

$$\begin{aligned} \min \quad & F_0(x) \\ s.t. \quad & F_i(x) = 0, \ \forall i \in \{i | F_i(x^*) = 0\} \\ & h_i(x) = 0 \quad i = 1, ..., p \end{aligned}$$

Stacking up the constraints, we may expressed this problem with a linear constraint (in matrix form),

$$\begin{aligned} \min \quad & F_0(x) \\ s.t. \quad & Ax = b \end{aligned}$$

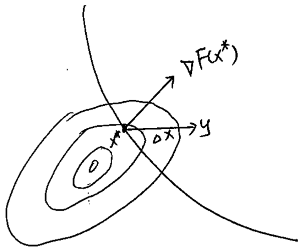
Consider the optimum point x^* and small perturbations about x^*, i.e., $x^* + \triangle x$ which perturbations stay feasible (so the point $x^* + \triangle x$ is still within feasible set).

By feasibility, we have $A(x^* + \triangle x) = b$, and

$$Ax^* + A\triangle x = b \Leftrightarrow A\triangle x = 0 \Leftrightarrow \triangle x \in N(A)$$

Apply first order condition,

$$\begin{aligned} & L(x, \nu) = F_0(x) + \nu^{\mathrm{T}}(Ax - b) \\ \Leftrightarrow & \nabla_x L(x, \nu) = \nabla F_0(x) + A^{\mathrm{T}}\nu = 0 \\ \Leftrightarrow & \nabla F_0(x) = -A^{\mathrm{T}}\nu \end{aligned}$$



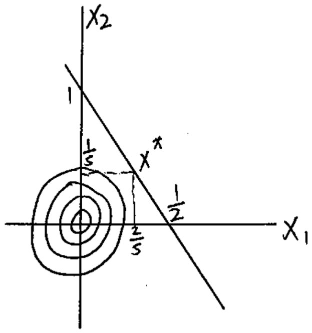
Because ν is unconstrained, $-A^T\nu$ can by any vector in $R(A^{\mathrm{T}})$.

Put together with the condition for optimality of constrained problem, a point $x^* \in \mathcal{C}$ is optimal iff

$$\langle \nabla F_0(x^*), (y - x^*) \rangle \geq 0, \quad \forall y \in \mathcal{C}$$

Since,

$$\begin{aligned} & \nabla F_0(x^*)^{\mathrm{T}} \triangle x \geq 0, \forall \triangle x \in N(A) \\ & \Leftrightarrow \nabla F_0(x^*) \perp N(A) \\ & \Leftrightarrow \nabla F_0(x^*) \in N(A)^{\perp} = R(A^{\mathrm{T}}) \end{aligned}$$



First-order for general problem

From the first-order condition,

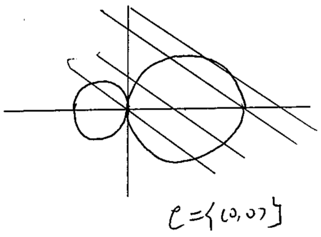
$$\nabla_x L(x,\lambda,\nu) = \nabla F_0(x) + \sum_{i=1}^{m} \lambda_i \nabla F_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0$$

We have

$$\nabla F_0(x) = \sum_{i=1}^{m} (-\lambda_i) \nabla F_i(x) - \sum_{i=1}^{p} \nu_i \nabla h_i(x)$$

Let's think about the case $m = 1$ and $p = 0$, so we have $\nabla F_0(x) = -\lambda_1 \nabla F_1(x)$.

However, this method not always work since we require the KKT condition holds, as we show in the following example.



Example 10.11. Consider the optimization problem,

$$\begin{aligned} \min \quad & x_1 + x_2 \\ s.t. \quad & (x_1+1)^2 + x_2^2 \le 1 \\ & (x_1-2)^2 + x_2^2 \le 4 \end{aligned}$$

Note that, this is a convex optimization problem with differentiable objective, and feasible set $\mathcal{C} = \{(0,0)\}$, a singleton.

If we compute the gradients for objective and constraints, and evaluate the gradient at the point $\{(0,0)\}$ (since x is only feasible at $\{(0,0)\}$), we may find that

$$\nabla F_0(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\nabla F_1(x) = \begin{bmatrix} 2(x_1+1) \\ 2x_2 \end{bmatrix} \Big|_{x=x^*=\begin{bmatrix} 0 \\ 0 \end{bmatrix}} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

$$\nabla F_2(x) = \begin{bmatrix} 2(x_1-2) \\ 2x_2 \end{bmatrix} \Big|_{x=x^*=\begin{bmatrix} 0 \\ 0 \end{bmatrix}} = \begin{bmatrix} -4 \\ 0 \end{bmatrix}$$

Apparently, we can not find λ_1 and λ_2 such that

$$\nabla F_0(x) = -\lambda_1 \nabla F_i(x) - \lambda_2 \nabla F_2(x)$$

In other words, the first order condition is not applicable here. The reason for this situation is, the Slater's condition does not hold for this problem.

The feasible set here is $\mathcal{C} = \{(0,0)\}$, it is a singleton. **For any finite set $\mathcal{C} \in \mathbb{R}^n$, there is no interior point for such set $\mathcal{C}$, that**

is, int$(\mathcal{C}) = \emptyset$**, and thus relint**$(\mathcal{C}) = \emptyset$ **as well.** Recall that, the Slater's condition requires there exist a relative interior point, so the condition fails here, and thus the strong duality and the KKT condition do not hold.

Extend duality theory to generalized inequality

Recall generalized inequalities are defined with respect to proper cones. A cone is proper if it is
(1) closed
(2) convex
(3) pointed (if $x \in k$ and $-x \in k$, then $x = 0$)
(4) Solid

As we want similar properties still hold when we extend to generalized inequality defined a proper cone, recall our previous derivation of weak duality:

$$\begin{aligned} g(\lambda) &= \min_x [F_0(x) + \sum_{i=1}^m \lambda_i F_i(x)] \\ &\leq F_0(x^*) + \sum_{i=1}^m \lambda_i F_i(x^*) \\ &\leq F_0(x^*) \end{aligned}$$

To accommodate generalized inequalities, we need to identify what set dual variables need to be restricted to keep the following holds (so that the weak duality holds),

$$\langle \lambda, F(x) \rangle \leq 0, \ \forall x \text{ feasible}$$

Now, we are considering the feasibility inequality constraints are defined by a cone k, and we say that the dual variable need to be restricted in the dual cone k^*.

A dual cone k^* is defined as

$$k^* = \{y | \langle x, y \rangle \geq 0, \ \forall x \in k\}$$

and the dual cone k^* is always closed and convex, whether the cone k is convex or not.

Also note that, a cone k is self dual if $k^* = k$. For instance, The non-negative orthant cone ($\mathbb{R}^m_+$), second order cone and PSD cone (S^m_+) are self dual.

With this definition for dual cone, apparently we could keep

$$\langle \lambda, F(x) \rangle \leq 0, \ \forall x \text{ feasible}$$

so that the weak duality holds.

Example 10.12. Recall the SDP problem,

$$\begin{aligned} \min \quad & c^{\mathrm{T}}x \\ s.t. \quad & x_1F_1 + x_2F_2 + \ldots + x_nF_n + G \leq_k 0 \end{aligned}$$

where $x \in \mathbb{R}^n$, $F_i \in S^m$, $G \in S^m$, the cone $k = S^m_+$.

Let $z \in S^m$ be the dual variable.

The Lagrangian function given by

$$\begin{aligned} L(x,z) &= c^{\mathrm{T}}x + \langle z, x_1F_1 + \ldots + x_nF_n + G \rangle \\ &= c^{\mathrm{T}}x + \sum_{i=1}^{n} \langle z, x_iF_i \rangle + \langle z, G \rangle \\ &= \sum_{i=1}^{n} x_i(c_i + \langle z, F_i \rangle) + \langle z, G \rangle \end{aligned}$$

is affine in x.

If you are confused where this inner product comes from, recall the definition of Frobenius inner product, for any $A, B \in \mathbb{R}^{m \times n}$, we have the inner product,

$$\langle A, B \rangle_F = \text{trace}(A^TB)$$

Furthermore, if $A, B \in S^m$, then

$$\langle A, B \rangle_F = \text{trace}(A^TB) = \text{trace}(AB)$$

and also remind yourself, the trace of a square matrix equals to sum of all its eigenvalues.

The dual function is give by

$$g(z) = \min_x L(x,z) = \begin{cases} -\infty & \text{if } \exists i \text{ s.t. } c_i + Tr(zF_i) \neq 0 \\ Tr(zG) & \text{else} \end{cases}$$

Hence, to maintain the feasibility, the dual optimization problem can be expressed as

$$\begin{aligned} \max_z \quad & tr(zG) \\ s.t. \quad & tr(zG) = -c_i \quad \forall i = 1, \cdots, n \\ & z \geq 0 \end{aligned}$$

where we utilize the fact that S^m_+ is self-dual, that is, $(S^m_+)^* = S^m_+$.

In summarize, to formulate the dual problem, what we need to do are

1. Given a cone k, find the dual cone k^*.
2. Restrict multipliers to be in k^*.

Bibliography