

Lecture notes on Optimization for MDI210

Robert M. Gower

October 12, 2020

Abstract

These are my notes for my lectures for the MDI210 Optimization and Numerical Analysis course. These notes are a work in progress, and will probably contain several mistakes (let me know?). If you are following my lectures you may find them useful to recall what we covered in class. Otherwise, I recommend you read these lectures notes [1] for the part on linear programming the excellent book [3] for the nonlinear optimization part. In particular, this book [3] contains all the subjects covered in these notes, and is a much better reference than these notes.

Contents

1	Linear Programming	2
1.1	A first example	2
1.2	Fundamental Theorem of linear programming	3
1.3	Notation and definitions	6
1.4	The simplex algorithm	8
1.5	Degeneracy and Cycling	9
1.6	Initialization using a first phase problem	9
1.7	Duality	11
1.8	How to compute dual solution: Complementary slackness	13
2	Nonlinear programming without constraints	14
2.1	The gradient, Hessian and the Taylor expansion	15
2.2	Level sets and geometry	17
2.3	Optimality conditions	19
2.4	Convex functions	20
2.5	Gradient method	22
2.6	Newton's method	24

3	Nonlinear programming with constraints	26
3.1	Admissable and Feasible directions	28
3.2	Lagrange's condition	31
3.3	Karush, Kuhn and Tuckers condition	31
3.4	Deducing Duality using KKT	35
3.5	Project Gradient Descent	36
3.6	The constrained descent method	39
3.7	Introduction to Constrained Optimization TP	39
3.8	Characterizing the constraint qualification	40
3.9	Farkas Lemma and the Geometry of Polyhedra	43

1 Linear Programming

Consider the problem

$$\begin{aligned}
 \max_x z &\stackrel{\text{def}}{=} c^\top x \\
 \text{subject to } Ax &\leq b, \\
 x &\geq 0,
 \end{aligned}
 \tag{LP}$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$. We assume that $m \leq n$, otherwise the feasible set might be trivial. For instance if $n \leq m$ and full rank, then there exists only one feasible point or no feasible points.

1.1 A first example

First we start with a simple 2D graphic example. We want to maximize the production of products x and y . Each unit of product x and y gives us 4 and 2 profit, respectively. To produce one unit of product x we need 30 minutes on machine 1 and 40 minutes on machine 2. To produce one unit of product y we need 20 minutes on machine 1 and 10 minutes on machine 2. Thus our problem is to maximize $4x + 2y$ subject to.

$$\begin{aligned}
 3x + 2y &\leq 6 \\
 4x + 1y &\leq 4
 \end{aligned}$$

See Figures 1 for a graphical illustration. This graphical example leads us to believe that is the solution is attainable, then it is always at a vertex. Indeed, this is the case as we show in Theorem 1.1.

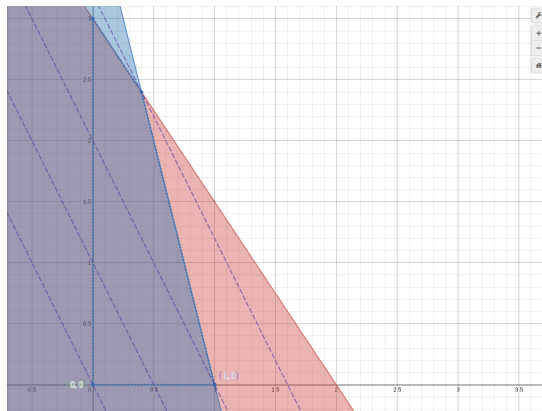


Figure 1: An graphical example of the constraint set together with the level sets of $4x + 2y = z$ with $z \in \{0, 2, 4, 6.4\}$ as dashed lines.

1.2 Fundamental Theorem of linear programming

Theorem 1.1 (Fundamental Theorem of Linear Programming). Let $P = \{x \mid Ax = b, x \geq 0\}$. One of the three circumstances must hold

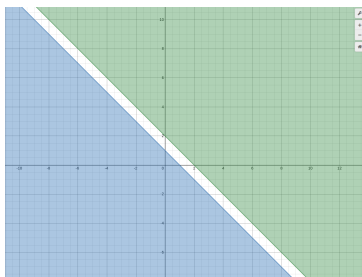
1. $P = \{\emptyset\}$
2. $P \neq \{\emptyset\}$ and there exists a vertex v of P such that $v \in \arg \min_{x \in P} c^\top x$
3. There exists $x, d \in \mathbb{R}^n$ such that $x + td \in P$ for all $t \geq 0$ and $\lim_{t \rightarrow \infty} c^\top (x + td) = \infty$.

Proof: A formal proof can be found in Section 2.1.2 [1]. Here we will illustrate with examples.

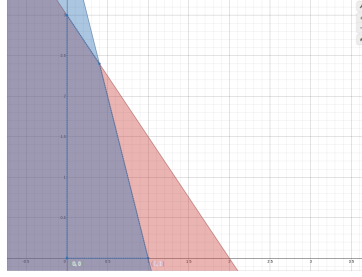
1. The case $P = \{\emptyset\}$ is evidenced with the simple example

$$x + y \leq 1$$

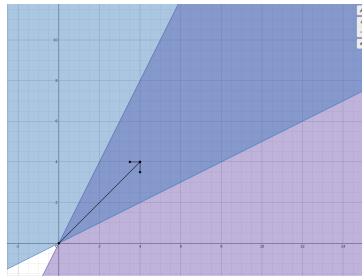
$$x + y \geq 2$$



2. $P \neq \{\emptyset\}$ and there exists a vertex v of P such that $v \in \arg \min_{x \in P} c^\top x$



3. There exists $x, d \in \mathbb{R}^n$ such that $x + td \in P$ for all $t \geq 0$ and $\lim_{t \rightarrow \infty} c^\top(x + td) = \infty$.



$$\begin{aligned} \max \quad & x + y \\ & 2x - y \geq 0 \\ & x - 2y \leq 2 \end{aligned}$$

This theorem suggests how we should develop an algorithm to solve the LP. First we should take care of finding a feasible point and assuring that the domain is non-empty. This is called **initialization phase** or **phase 1**. After determining a feasible point, we should traverse the vertices of P searching for the optimal point. If we find one, we say that the solution is **attainable**. If we find a direction in which $c^\top x \rightarrow \infty$, we say that there is an **unbounded** solution.

Let us transform this insight into a method for solving. First we transform a problem given in the standard form

$$\begin{aligned} \max \quad & 4x_1 + 2x_2 \\ & 3x_1 + 2x_2 \leq 600 \\ & 4x_1 + 1x_2 \leq 400 \\ & x_1 \geq 0, x_2 \geq 0. \end{aligned} \tag{1}$$

into a maximization over linear constraints with non-negativity constraints

$$\begin{aligned} \max \quad & 4x_1 + 2x_2 \\ x_3 = 600 \quad & - 3x_1 - 2x_2 \\ x_4 = 400 \quad & - 4x_1 - x_2. \end{aligned} \tag{2}$$

We arrived at the above (1) by simply adding x_3 and x_4 on the 1st and 2nd rows to fill in the *slack* of the inequality. Accordingly, the variables x_3 and x_4 are referred to as the slack variables.) It is not hard to show that (1) and (2) are equivalent.

Exercise 1.2. Show that if (x_1^*, x_2^*) is a solution to (1) then there exists x_3^* and x_4^* such that $(x_1^*, x_2^*, x_3^*, x_4^*)$ is a solution to (2). Furthermore, if $(x_1^*, x_2^*, x_3^*, x_4^*)$ is a solution to (2), show that (x_1^*, x_2^*) is a solution to (1).

Proof: Straightforward.

The transformation (1) into (2) has effectively turned the constraint $Ax \leq b$ into a linear system $Ax - b = I\hat{x}$ or equivalently $Ax - I\hat{x} = b$. So now we can apply row transformations (left multiplying by an invertible matrix). Next we re-write (2) in the *Dictionary* format

$$\begin{array}{rclcl} x_3 & = & 600 & - & 3x_1 & - & 2x_2 \\ x_4 & = & 400 & - & 4x_1 & - & x_2 \\ \hline z & = & & & 4x_1 & + & 2x_2 \end{array}$$

The non-negativity constraints are non longer explicitly included, so we must take care so as to ensure they hold. Note that the above has a convenient *feasible point*, that is, the point $(x_1^*, x_2^*, x_3^*, x_4^*) = (0, 0, 600, 400)$ satisfies the above equality constraints. The feasible point corresponds to the $(0, 0)$ point in our original 2D constraint space. We will now try to move from this vertex to a neighbouring vertex that has a larger objective value.

Note that the *cost coefficient* of x_1 in the objective function is positive, which shows that increasing x_1 will increase the objective function. But increasing x_1 decreases x_3 and x_4 . Indeed we have

$$\begin{aligned} x_3 \geq 0 &\Rightarrow 600 - 3x_1 \geq 0 \Rightarrow x_1 \leq 200, \\ x_4 \geq 0 &\Rightarrow 400 - 4x_1 \geq 0 \Rightarrow x_1 \leq 100. \end{aligned}$$

Thus $x_1 \leq 100$ otherwise x_4 will become negative. We will now perform row operations on (3) so that x_1 appears isolated on the left-hand side taking x_4 's position. This is called *pivoting* on the element $(4, 1)$. This gives

$$\begin{array}{rclcl} x_3 & = & 300 & & 0 & - & \frac{5}{4}x_2 \\ x_1 & = & 100 & - & \frac{x_4}{4} & - & \frac{x_2}{4} \\ \hline z & = & 400 & - & x_4 & + & x_2 \end{array}$$

Now we are at the vertex $(x_1^*, x_2^*) = (100, 0)$. We refer to this last operation as x_4 *leaving the basis* and x_1 *entering the basis*. Next we see that increasing x_2 increases the objective value but

$$\begin{aligned} x_3 \geq 0 &\Rightarrow 240 \geq x_2, \\ x_1 \geq 0 &\Rightarrow 400 \geq x_4. \end{aligned}$$

Consequently we can increase x_2 upto 240 while respecting the positivity constraints of x_3 . Thus x_3 will leave the basis and x_2 will enter the basis. Performing a row elimination again, we have

that

$$\begin{array}{rcccc} x_2 & = & 240 & & 0 & - & \frac{4}{5}x_3 \\ x_1 & = & 40 & - & \frac{x_4}{4} & - & \frac{1}{5}x_3 \\ \hline z & = & 640 & - & x_4 & - & \frac{4}{5}x_3 \end{array}$$

Now increasing x_4 or x_3 will decrease the objective value, thus we can make no further improvement. The final optimal vertex is given by $(x_1^*, x_2^*) = (40, 240)$ and the optimal objective value is $z^* = 640$.

1.3 Notation and definitions

Before formalizing a method for solving the above, we will establish a standard representation of linear programs. The standard form is given in (LP). Not all LPs fit the format (LP), but all LPs can be re-written in the form (LP). A few of the standard tricks we use to re-write an LP in the standard form are listed as here

1. (**Equality**) Replace $\sum_{j=1}^n a_{ij}x_j = b_i$ by

$$\sum_{j=1}^n a_{ij}x_j \leq b_i,$$

and

$$-\sum_{j=1}^n a_{ij}x_j \leq -b_i.$$

2. (**Box constraints**) Replace

$$\alpha \leq x_i \leq \beta,$$

by

$$y = x - \alpha,$$

$$y \leq \beta - \alpha$$

and

$$y \geq 0.$$

3. (**Unconstrained**) If x has no positivity constraint, then replace $x = x^+ - x^-$ and add the constraints $x^+ \geq 0$ and $x^- \geq 0$.

We will now formalize the definitions we introduced in the examples.

- The objective is to maximize the linear objective function $z = \sum_{j=1}^n c_j x_j$
- There are m inequality constraints in the standard form given by

$$\sum_{j=1}^n a_{ij}x_j \leq b_i, \text{ for } i \in \{1, \dots, m\}.$$

- There are n positivity constraints given by $x_j \geq 0$, for $j \in \{1, \dots, n\}$.
- We call $(x_1^*, \dots, x_n^*) \in \mathbb{R}^n$ a feasible solution if it satisfies the inequality and positivity constraints.

When passing from the standard form to the dictionary form, we had to introduce some additional notation

- We call the additionally introduced variable $(x_{n+1}, \dots, x_{n+m}) \in \mathbb{R}^m$ the slack variables (“variables d’écart”)
- We refer to

$$\begin{aligned} x_{n+1} &= b_1 - \sum_{j=1}^n a_{1j}x_j \\ &\vdots \\ x_{n+i} &= b_i - \sum_{j=1}^n a_{ij}x_j \\ &\vdots \\ x_{n+m} &= b_m - \sum_{j=1}^n a_{mj}x_j \\ z &= \sum_{j=1}^n c_jx_j, \end{aligned}$$

as the initial dictionary.

- We say that the system of equations with non-negativity constraints forms a valid dictionary if m of the variables (x_1, \dots, x_{n+m}) can be expressed as an explicit function of the remaining n variables.
- We refer to the m variables isolated on the left-hand side as the *basic variable* (*variable de base*) and the remaining n variables as *non-basic* (*variable hors-base*) or *outside* the basis.

After writing the problem in the dictionary format, we then performed several row elimination operations to change the basic variables. These row operations altered the coefficients of the constraints a_{ij} and the cost vector c_j . We introduce notation to accommodate for these changing coefficients.

- Let $I \subset \{1, \dots, n+m\}$ be the set of indices of the basic variables and let $J = \{1, \dots, n+m\} \setminus I$ be the non-basic variables.
- For a given basis determined by I there is a corresponding dictionary

$$\begin{aligned} x_i &= b'_i + \sum_{j \in J} a'_{ij}x_j, \text{ for } i \in I \\ z &= z^* + \sum_{j \in J} c'_jx_j, \end{aligned}$$

where $a'_{ij}, b'_i, z^* \in \mathbb{R}$ are coefficients resulting from the row operations. For this to be a feasible dictionary we require that $b'_i \geq 0$.

1.4 The simplex algorithm

Let us now formalize the operations of the simplex algorithm. First, we choose a variable x_{j_0} that has a positive cost c'_{j_0} to enter the basis, see Algorithm 1. Otherwise if all costs are negative, we have found the optimal. Next we must choose a variable x_{i_0} that will leave the basis. We determine i_0 as the first variable whose value equals zero as we increase $x_{j_0}^*$. That is, since $x_{j_0}^*$ will be the only non-zero non-basic variable, we have

$$x_i^* = b'_i + \sum_{j \in J} a'_{ij} x_j^* = b'_i + a'_{ij_0} x_{j_0}^*.$$

Since we require $x_i^* \geq 0$ this imposes that $b'_i + a'_{ij_0} x_{j_0}^* \geq 0$. In other words

$$x_{j_0}^* \leq -\frac{b'_i}{a'_{ij_0}} \quad \text{if } a'_{ij_0} < 0, \quad \text{for } i \in I, \quad (3)$$

$$x_{j_0}^* \geq -\frac{b'_i}{a'_{ij_0}} \quad \text{if } a'_{ij_0} \geq 0, \quad \text{for } i \in I. \quad (4)$$

Since we are only interested in increasing the value of $x_{j_0}^*$, the case (4) where $a'_{ij_0} \geq 0$ does not restrict $x_{j_0}^*$ from increasing (since $b'_i \geq 0$). Thus only (3) constrains the value $x_{j_0}^*$. With the preceding definitions, we can now state the Simplex method in Algorithm 1. In particular, the pivoting step in Algorithm 1 can be stated using elementwise operations as

for $j \in J$ **do**

for $i \in I \setminus \{i_0\}$ **do**

$$a'_{ij} \leftarrow a'_{ij} - \frac{a'_{ij_0}}{a'_{i_0j_0}} a'_{i_0j} \quad \# \text{ Row elimination on pivot } (i_0, j_0).$$

$$c_j \leftarrow c_j - \frac{c'_{j_0}}{a'_{i_0j_0}} a'_{i_0j}$$

for $j \in J$ **do**

$$a_{i_0j} \leftarrow -\frac{a_{i_0j}}{a'_{i_0j_0}}$$

$$a'_{i_0j_0} = 1/a'_{i_0j_0}$$

What was left unclear is how do we choose j_0 to enter the basis. There are four common choices.

1. The mad hatter rule: Choose the first one you see.
2. Dantzig's 1st rule: $j_0 = \arg \max_{j \in J} c_j$.
3. Dantzig's 2nd rule: Choose $j_0 \in \{j \in J : c_j > 0\}$ that results in the largest increase in the objective value. Let $t = \min_{i \in I, a_{ij_0} < 0} \left\{ -\frac{b_i}{a_{ij_0}} \right\}$. The variable entering the basis will have value $x_{j_0}^* = t$. Consequently the objective value increases by tc_{j_0} . Choosing j_0 that maximizes this increase is equivalent to choosing via

$$j_0 = \arg \max_{j \in J} \left\{ c_j \min_{i \in I, a_{ij} < 0} \left\{ -\frac{b_i}{a_{ij}} \right\} \right\}.$$

This effective but computationally expensive.

Algorithm 1 One Simplex iteration

Input: A basic index set $I \subset \{1, \dots, n+m\}$, $J = I \setminus \{1, \dots, n+m\}$, constraint coefficients $a'_{ij} \in \mathbb{R}$, $b'_i \geq 0$ and $c'_i \in \mathbb{R}$.

if $c_i \leq 0$ for all $i \in J$ **then**

STOP; # Optimal point found.

Choose a variable j_0 to **enter the basis** from the set $j_0 \in \{j \in J : c'_j > 0\}$.

if $a'_{ij_0} \geq 0$ for all $i \in I$ **then**

STOP; # The problem is unbounded.

Choose a variable i_0 to **leave the basis** from the set $i_0 \in \arg \min_{i \in I, a'_{ij_0} < 0} \left\{ -\frac{b'_i}{a'_{ij_0}} \right\}$.

$aux \leftarrow a'_{i_0 j_0}$

for $i \in I \setminus \{i_0\}$ **do**

$a'_i \leftarrow a'_i - a'_{ij_0} a'_{i_0} / a_{i_0 j_0}$ # Row elimination on pivot (i_0, j_0) .

$c' \leftarrow c' - c'_{j_0} a'_{i_0} / a_{i_0 j_0}$ # Update the cost coefficients.

$a'_{i_0} \leftarrow -a'_{i_0} / a'_{i_0 j_0}$ # Row normalization .

$a_{i_0 j_0} \leftarrow 1 / aux$

$I \leftarrow (I \setminus \{i_0\}) \cup \{j_0\}$. # Update basic variables set

$J \leftarrow (J \setminus \{j_0\}) \cup \{i_0\}$. # Update non-basic variable set

Output: I, a'_{ij}, b'_j, c'_j .

4. Bland's rule: Choose the smallest indices j_0 and i_0 . That is, choose

$$j_0 = \arg \min \{j \in J : c_j > 0\}.$$

If the set $\arg \min_{i \in I, a_{ij_0} < 0} \left\{ \frac{b_i}{a_{ij_0}} \right\}$ has more than one element, choose the smallest

$$i_0 = \min \left\{ \arg \min_{i \in I, a_{ij_0} < 0} \left\{ -\frac{b_i}{a_{ij_0}} \right\} \right\}.$$

Dantzig's rules were designed to maximize the objective function in a greedy manner. While Bland's rule, though apparently mundane, was designed to avoid *cycling*.

1.5 Degeneracy and Cycling

If any of the basic variables have zero value, we say that it is a *degenerate* basis. Degenerate basis require extra care because they may lead to the simplex algorithm cycling. See example on board and your alternative french notes.

1.6 Initialization using a first phase problem

Not always will we have that $b_i \geq 0$ for $i = 1, \dots, m$. Simply including slack variables will not lead to a feasible basic solution. To find a feasible solution we will use the simplex method on an

auxiliary *first phase* problem.

First assume we are given a problem

$$\begin{aligned} \max_x z &\stackrel{\text{def}}{=} c^\top x \\ \text{subject to } Ax &\leq b, \\ x &\geq 0, \end{aligned} \tag{5}$$

where at least one $b_i < 0$ where $i \in \{1, \dots, m\}$. Consequently $x^* = 0$ is not a feasible solution. To remedy this we add an additional variable x_0 and **change the objective**

$$\begin{aligned} \max_x -x_0 \\ \text{subject to } Ax &\leq b + Ix_0, \\ x \geq 0, x_0 &\geq 0. \end{aligned} \tag{LP-1st}$$

There now exists x_0 for which there is a feasible solution, for instance choosing $x_0^* = \max_{i=1, \dots, m} |b_i|$ and $x_i^* = 0$ for $i = 1, \dots, n$. The problem (LP-1st) is known as the *1st phase* simplex method. We can use the simplex method to solve (LP-1st). If the solution to (LP-1st) is such that $x_0^* \neq 0$, then we know that the original problem (5) is infeasible. If the solution to (LP-1st) is such that $x_0^* = 0$, then we can use the remaining variables $x_i^* \neq 0$ as a starting basis.

To do this, first we setup a dictionary with slack variables as the basis, even though they do not form a feasible basis.

$$\begin{aligned} x_{n+1} &= b_1 - \sum_{j=1}^n a_{1j}x_j + x_0 \\ &\vdots \\ x_{n+m} &= b_m - \sum_{j=1}^n a_{mj}x_j + x_0 \\ z &= \phantom{b_m - \sum_{j=1}^n a_{mj}x_j} - x_0 \end{aligned} \tag{6}$$

Next, different from our previous examples, we will make x_0 enter the basis even though it has a negative cost. Suppose w.l.o.g that x_{n+1} leaves the basis as x_0 enters. Thus after pivoting on row 1, column $n + 1$ and performing the row operations

$$r_i \rightarrow r_i - r_1 \text{ for the } i = 2, \dots, m,$$

the next dictionary would be

$$\begin{aligned} x_0 &= -b_1 + \sum_{j=1}^n a_{1j}x_j + x_{n+1} \\ &\vdots \\ x_{n+m} &= b_m - b_1 - \sum_{j=1}^n (a_{mj} - a_{1j})x_j + x_{n+1} \\ z &= b_1 - \sum_{j=1}^n a_{1j}x_j - x_{n+1} \end{aligned} \tag{7}$$

Now, so long as there are positive elements in $(a_{1j})_j$, we can proceed with the simplex method as usual. Note that this indicates we should have chosen an element i_0 to leave basis such that $(a_{i_0j})_j$ has many positive coefficients. If we continue to iterate and x_0 leaves the basis, then we have found a feasible basis point and we can drop the x_0 variable.

1.7 Duality

Consider again the LP in standard form

$$\begin{aligned} \max_x z &\stackrel{\text{def}}{=} c^\top x \\ \text{subject to } Ax &\leq b, \\ x &\geq 0, \end{aligned} \tag{P}$$

Though we now have a technique for solving (P), at any given moment we do not know how far we are from the solution. Will we need another few minutes of computing resources or days of computing resources? This is a troublesome question. For this, and other reasons, we will now develop an alternative and equivalent formulation of (P) that will help determine if we are near the solution, among other insights. This equivalent formulation is known as the *dual* formulation. We can first derive the dual problem as a means of finding an upper bound to the solution of (P).

Say we wish to upper bound our objective $z = c^\top x$ and we want to do this by combining rows of the constraints $Ax \leq b$. That is, let $y \geq 0 \in \mathbb{R}^m$. If we could determine such a y so that $y^\top A \approx c^\top$ then we would have that

$$c^\top x \approx (y^\top A)x \leq y^\top b,$$

thus $y^\top b$ is an approximate upper bound of $c^\top x$. What is more, $y^\top b$ does not depend on x , ergo this bound holds for all x , including the optimal solution. But just being approximate upper bound is no good. Instead, assume we have $0 \leq y \in \mathbb{R}^m$ such that $y^\top A \geq c^\top$ or equivalently $A^\top y \geq c$. Then indeed the upper bound holds since

$$c^\top x \leq (y^\top A)x \leq y^\top b.$$

Now say we want the upper bound to be as tight as possible. We can do this by choosing $y \geq 0$ so that $y^\top b$ is small as possible. That is, we need to the following *dual* linear program.

$$\begin{aligned} \min_y w &\stackrel{\text{def}}{=} y^\top b \\ \text{subject to } A^\top y &\geq c, \\ y &\geq 0. \end{aligned} \tag{D}$$

Exercise 1.3. Show that the dual of the dual is the primal program. In other words, this is a *reflexive* transformation.

By construction of the dual program, the following lemma holds.

Lemma 1.4 (Weak Duality). If $x \in \mathbb{R}^n$ is a feasible point for (P) and $y \in \mathbb{R}^m$ is a feasible point for (D) then

$$c^\top x \leq y^\top Ax \leq y^\top b. \quad (8)$$

Consequently

- If (P) has an unbounded solution, that is $c^\top x \rightarrow \infty$, then there exists no feasible point y for (D)
- If (D) has an unbounded solution, that is $y^\top b \rightarrow -\infty$, then there exists no feasible point x for (P)
- If x and y are primal and dual feasible, respectively, and $c^\top x = y^\top b$, then x and y are the primal and dual optimal points, respectively.

What is even more remarkable is that, not only does (D) provide an upper bound for (P), but they are equivalent problems, in the following sense

Theorem 1.5 (Strong Duality). If (P) or (D) is feasible, then $z^* = w^*$. Moreover, if c' is the cost vector of the optimal dictionary of the primal problem, that is, if

$$z = z^* + \sum_{i=1}^{n+m} c'_i x_i, \quad (9)$$

then $y_i^* = -c'_{n+i}$ for $i = 1, \dots, m$.

Proof: First note that $c'_i \leq 0$ for $i = 1, \dots, m+n$ otherwise the dictionary would not be optimal. Consequently $y_i^* = -c'_{n+i} \geq 0$ for $i = 1, \dots, m$. Furthermore, by the definition of the slack variables we have that

$$x_{n+i} = b_i - \sum_{j=1}^n a_{ij} x_j, \quad \text{for } i = 1, \dots, m. \quad (10)$$

Consequently, setting $y_i^* = -c'_{n+i}$, we have that

$$\begin{aligned} z &\stackrel{(9)}{=} z^* + \sum_{j=1}^n c'_j x_j + \sum_{i=n+1}^{n+m} c'_i x_i \\ &\stackrel{(10)}{=} z^* + \sum_{j=1}^n c'_j x_j - \sum_{i=1}^m y_i^* (b_i - \sum_{j=1}^n a_{ij} x_j) \\ &= z^* - \sum_{i=1}^m y_i^* b_i + \sum_{j=1}^n \left(c'_j + \sum_{i=1}^m y_i^* a_{ij} \right) x_j \\ &= \sum_{j=1}^n c_j x_j, \quad \forall x_1, \dots, x_n. \end{aligned} \quad (11)$$

where the last line followed by definition of the objective function $z = \sum_{j=1}^n c_j x_j$. Since the above holds for all $x \in \mathbb{R}^n$, we can match the coefficients to obtain

$$z^* = \sum_{i=1}^m y_i^* b_i \quad (12)$$

$$c_j = c'_j + \sum_{i=1}^m y_i^* a_{ij}, \quad \text{for } j = 1, \dots, n. \quad (13)$$

Since $c'_j \leq 0$ for $j = 1, \dots, n$, the above is equivalent to

$$z^* = \sum_{i=1}^m y_i^* b_i \quad (14)$$

$$\sum_{i=1}^m y_i^* a_{ij} \leq c_j, \quad \text{for } j = 1, \dots, n. \quad (15)$$

The inequalities (15) prove that y_i^* 's satisfies the constraints in (D), and thus is feasible. The equality (14) shows that $z^* = \sum_{i=1}^m y_i^* b_i = w$, and consequently by weak duality the y_i^* 's are dual optimal. \square

Calculating the dual optimal variables y^* using Theorem 1.5 requires knowing the cost vector c' of the optimal tableau. But it turns out that we do not need c' to calculate y^* . We can instead recover the y^* by only knowing x^* , and the complementary slackness theorem shows (see notes).

1.8 How to compute dual solution: Complementary slackness

Let $x^* \in \mathbb{R}^n$ be an optimal solution of (P). Then $y^* \in \mathbb{R}^m$ is an optimal dual solution if $c^\top x^* = (y^*)^\top b$. Thus by the weak duality theorem we have that

$$c^\top x^* = (y^*)^\top A x^* = (y^*)^\top b.$$

Subtracting $(y^*)^\top A x^*$ from all sides of the above gives

$$\underbrace{(c - A^\top y^*)^\top}_{\geq 0} x^* = 0 = (y^*)^\top \underbrace{(b - A x^*)}_{\geq 0}.$$

Re-writing the above in element form we have that

$$\sum_{j=1}^n (c_j - \sum_{i=1}^m a_{ij} y_i^*) x_j^* = 0 = \sum_{i=1}^m y_i^* (b_i - \sum_{j=1}^n a_{ij} x_j^*).$$

On both sides we have a sum over the product of positive numbers. Thus the total sum is only zero if the individual products are zero, that is

$$y_i^* (b_i - \sum_{j=1}^n a_{ij} x_j^*) = 0, \quad \forall i = 1, \dots, m.$$

$$x_j^* (c_j - \sum_{i=1}^m a_{ij} y_i^*) = 0, \quad \forall j = 1, \dots, n.$$

This gives the following rule for computing y^* .

$$\begin{aligned} \sum_{i=1}^n a_{ij}y_i^* &= c_j, \quad \forall j \in \{1, \dots, n\}, \quad x_j^* > 0. \\ y_i &= 0, \quad \forall i \in \{1, \dots, m\}, \quad b_i > \sum_{j=1}^n a_{ij}x_j^*. \end{aligned}$$

Thus we need a single system solve to compute y^* . Finally since $b_i > \sum_{j=1}^n a_{ij}x_j^*$ implies that the slack variable $x_{n+i}^* > 0$ we have the more succinct rule

$$\begin{aligned} \sum_{i=1}^n a_{ij}y_i^* &= c_j, \quad \forall j \in \{1, \dots, n\}, \quad x_j^* > 0. \\ y_i &= 0, \quad \forall i \in \{1, \dots, m\}, \quad x_{n+i}^* > 0. \end{aligned} \tag{16}$$

Definition 1.6. We refer to the method for computing the dual variables using (16) as the primal dual map. Indeed for any primal feasible values x^* we can compute a dual feasible y^* using (16).

2 Nonlinear programming without constraints

Robert: I recommend reading Chapter 1 in [2] as an introduction into nonlinear optimization.

We now move onto nonlinear optimization, that is, we wish to minimize a possibly nonlinear differentiable function $f : x \in \mathbb{R}^n \mapsto f(x) \in \mathbb{R}$. First we will consider the unconstrained optimization problem

$$x^* \in \arg \min_{x \in \mathbb{R}^n} f(x). \tag{17}$$

All the methods we develop are iterative, in that they produce a sequence of iterates x^1, \dots, x^k, \dots , in the hope that they converge to the solution with

$$\lim_{k \rightarrow \infty} x^k = x^*.$$

Furthermore, we will focus on *descent methods* where the iterates are calculated via

$$x^{k+1} = x^k + s_k d^k, \tag{18}$$

where $s_k > 0$ is a *step size* and $d^k \in \mathbb{R}^n$ is *search direction*. The search direction and step size will satisfy the *descent condition* given by

$$f(x^k) < f(x^{k+1}). \tag{19}$$

This in turn guarantees that, little by little, we get closer to the solution of (17).

The key tool in ensuring that the descent condition holds is the gradient.

2.1 The gradient, Hessian and the Taylor expansion

For a continuously differentiable function $f : x \in \mathbb{R}^n \mapsto f(x) \in \mathbb{R}$, we refer to $\nabla f(x)$ as the gradient evaluated at x defined by

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_n} \right]^\top.$$

Note that $\nabla f(x)$ is a column-vector.

For any vector valued function $F : x \in \mathbb{R}^d \rightarrow F(x) = [f_1(x), \dots, f_n(x)]^\top \in \mathbb{R}^n$ we define the *Jacobian matrix* by

$$\nabla F(x) \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial f_1(z)}{\partial x_1} & \frac{\partial f_1(z)}{\partial x_2} & \frac{\partial f_1(z)}{\partial x_3} & \cdots & \frac{\partial f_1(z)}{\partial x_d} \\ \frac{\partial f_2(z)}{\partial x_1} & \frac{\partial f_2(z)}{\partial x_2} & \frac{\partial f_2(z)}{\partial x_3} & \cdots & \frac{\partial f_2(z)}{\partial x_d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n(z)}{\partial x_1} & \frac{\partial f_n(z)}{\partial x_2} & \frac{\partial f_n(z)}{\partial x_3} & \cdots & \frac{\partial f_n(z)}{\partial x_d} \end{bmatrix} = \begin{bmatrix} \nabla f_1(x)^\top \\ \nabla f_2(x)^\top \\ \nabla f_3(x)^\top \\ \vdots \\ \nabla f_n(x)^\top \end{bmatrix}.$$

The gradient is useful because we can use it to build a linear approximation of $f(x)$ using the 1st order Taylor expansion

$$f(x^0 + d) = f(x^0) + \nabla f(x^0)^\top d + \epsilon(d) \|d\|_2, \quad (20)$$

where $\epsilon(d)$ is a continuous real valued such that

$$\lim_{d \rightarrow 0} \epsilon(d) = 0. \quad (21)$$

Furthermore, from definition of limit, given any constant $c > 0$ there exists $\delta > 0$ such that

$$\|d\| < \delta \Rightarrow |\epsilon(d)| < c. \quad (22)$$

We will use the expansion (20) to motivate and understand the steepest descent method later.

If $f(x)$ is twice differentiable, we refer to $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$ as the Hessian matrix:

$$\nabla^2 f(x) \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial^2 f(z)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(z)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(z)}{\partial x_1 \partial x_3} & \cdots & \frac{\partial^2 f(z)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(z)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(z)}{\partial x_2 \partial x_2} & \frac{\partial^2 f(z)}{\partial x_2 \partial x_3} & \cdots & \frac{\partial^2 f(z)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(z)}{\partial x_n \partial x_1} & \frac{\partial^2 f(z)}{\partial x_n \partial x_2} & \frac{\partial^2 f(z)}{\partial x_n \partial x_3} & \cdots & \frac{\partial^2 f(z)}{\partial x_n \partial x_n} \end{bmatrix}.$$

If $f \in C^2$ (twice continuously differentiable) then

$$\frac{\partial^2 f(z)}{\partial x_i \partial x_j} = \frac{\partial^2 f(z)}{\partial x_j \partial x_i}, \quad \forall i, j \in \{1, \dots, n\},$$

consequently the Hessian matrix is symmetric, that is, $\nabla^2 f(x) = \nabla^2 f(x)^\top$. Furthermore, for $f \in C^2$, with the Hessian matrix we can build a quadratic approximation to our function using the 2nd order Taylor expansion.

$$f(x^0 + d) = f(x^0) + \nabla f(x^0)^\top d + \frac{1}{2} d^\top \nabla^2 f(x^0) d + \epsilon(d) \|d\|_2^2. \quad (23)$$

We will use this expansion to motivate the Newton method later on.

To calculate the gradient and the Hessian, we need to use the chain-rule and product-rule.

Chain-rule. For every differentiable map $F(y) : \mathbb{R}^d \rightarrow \mathbb{R}^n$ and function $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ we have that the gradient of the *composition*

$$f(F(x)) = f(F_1(x), \dots, F_n(x)), \quad (24)$$

is given by

$$\nabla(f(F(x))) = \nabla F(x)^\top \nabla f(F(x)). \quad (25)$$

Example 2.1. Given a vector $a \in \mathbb{R}^n$ let $f(y) = y^\top a$. Show that

$$\nabla(f(F(x))) = \nabla F(x)^\top \frac{dy^\top a}{dy} = \nabla F(x)^\top a. \quad (26)$$

Product-rule. For any two vector valued functions $F^1 : \mathbb{R}^d \mapsto \mathbb{R}^n$ and $F^2 : \mathbb{R}^d \mapsto \mathbb{R}^n$ we have that

$$\nabla(F^1(x)^\top F^2(x)) = \nabla F^1(x)^\top F^2(x) + \nabla F^2(x)^\top F^1(x). \quad (27)$$

Proof: The product-rule can be deduced from the chain-rule. For this, let $f(z, y) = z^\top y$, for $z, y \in \mathbb{R}^n$. Thus $F^1(x)^\top F^2(x) = f(F^1(x), F^2(x))$. Now note that

$$\nabla f(z, y) = \begin{pmatrix} y \\ z \end{pmatrix}.$$

Now let $F(x) = (F^1(x), F^2(x))$. Thus by the definition of Jacobian we have that

$$DF(x) = \begin{pmatrix} DF^1(x) \\ DF^2(x) \end{pmatrix}.$$

The above combined with the chain rule gives that

$$\begin{aligned} \nabla f(F^1(x), F^2(x)) &= DF(x)^\top \nabla f(F^1(x), F^2(x)) \\ &= \begin{pmatrix} DF^1(x)^\top & DF^2(x)^\top \end{pmatrix} \begin{pmatrix} F^2(x) \\ F^1(x) \end{pmatrix} \\ &= \nabla F^1(x)^\top F^2(x) + \nabla F^2(x)^\top F^1(x). \end{aligned} \quad (28)$$

Exercise 2.2. Let $f(x) = \frac{1}{2}x^\top Ax + x^\top b + c$, where $A \in \mathbb{R}^{n \times n}$ is a symmetric positive definite matrix, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Calculate the gradient and Hessian of $f(x)$.

Proof: Taking the gradient we have

$$\begin{aligned}
\nabla f(x) &= \frac{1}{2}\nabla(x^\top Ax) + \nabla(x^\top b) + \nabla(c) \\
&\stackrel{(26)}{=} \frac{1}{2}\nabla(x^\top Ax) + b \\
&\stackrel{(27)}{=} \frac{1}{2}\nabla(x)^\top Ax + \frac{1}{2}\nabla(Ax)^\top x + b \\
&= \frac{1}{2}Ax + \frac{1}{2}A^\top x + b = Ax + b,
\end{aligned}$$

where we used in the last two sets that $\nabla(Ax) = A$ by the definition of the Jacobian and that $A = A^\top$.

2.2 Level sets and geometry

For a fixed constant $C \in \mathbb{R}$ we refer to the following surface

$$\{x \in \mathbb{R}^n \mid f(x) = C\}, \quad (29)$$

as the C level set of $f(x)$. Each level sets implicitly define a surface where $f(x)$ has the same value.

Exercise 2.3. Draw the level sets

$$4 = f(x_1, x_2),$$

$$9 = f(x_1, x_2),$$

where $f(x_1, x_2) = x_1^2 + x_2^2$.

The gradient direction is intimately linked to the geometry of the levels sets. Indeed, the gradient points orthogonal to the level sets. Thus we can also use the gradient to define the the tangent space, as follows.

Lemma 2.4. The tangent directions to a point on the level set of $f(x)$ are orthogonal to the gradient at that point. That is, for every

$$v \in T_f(x^0) = \left\{ v \in \mathbb{R}^n : v = \lim_{\substack{y \rightarrow x^0, \\ f(y) = f(x^0)}} \frac{y - x^0}{\|y - x^0\|_2} \right\}, \quad (30)$$

we have that $v^\top \nabla f(x^0) = 0$.

Proof: Let $y \in \mathbb{R}^n$ be such that $f(x^0) = f(y)$, consequently using (20) we have that

$$f(y) = f(x^0) + \nabla f(x^0)^\top (y - x^0) + \epsilon(y - x^0)\|y - x^0\|_2 = f(x^0).$$

Thus $\nabla f(x^0)^\top (y - x^0) + \epsilon(y - x^0)\|y - x^0\|_2 = 0$. Dividing by $\|y - x^0\|_2$ and taking the limit $y \rightarrow x^0$, gives that

$$0 = \lim_{y \rightarrow x^0} \left(\nabla f(x^0)^\top \frac{y - x^0}{\|y - x^0\|_2} + \epsilon(y - x^0) \right) \stackrel{(30)+(21)}{=} \nabla f(x^0)^\top v. \quad \square$$

Another way of phrasing the above result is that $y = f(x^0) + \nabla f(x^0)^\top d$ is the tangent line of $f(x)$ at $(x^0, f(x^0))$.

The gradient is also the direction that maximizes the local change in the objective function, while the negative gradient is the direction that minimizes the local change. This is the reason why gradient methods are often referred to as steepest descent methods.

Lemma 2.5 (Steepest Descent). For a give $d \in \mathbb{R}^n$ we refer to

$$\Delta(d) \stackrel{\text{def}}{=} \lim_{s \rightarrow 0^+} \frac{f(x^0 + sd) - f(x^0)}{s}, \quad (31)$$

as the *local change* of $f(x)$ around x^0 . Let $v = \nabla f(x^0) / \|\nabla f(x^0)\|_2$ be the normalized gradient. It follows that

$$\begin{aligned} v &= \arg \max_{d \in \mathbb{R}^n} \Delta(d) \\ &\text{subject to } \|d\|_2 = 1. \end{aligned} \quad (32)$$

That is, the normalized gradient is the direction maximizes the *local change* of $f(x)$ around x^0 . Furthermore the negative normalized gradient minimizes the local change.

Proof: Using (20) we have that

$$f(x^0 + sd) - f(x^0) = s \nabla f(x^0)^\top d + \epsilon(sd)s.$$

Dividing by s and taking the limit $s \rightarrow 0$ we have

$$\Delta(d) = \lim_{s \rightarrow 0^+} \frac{f(x^0 + sd) - f(x^0)}{s} = \nabla f(x^0)^\top d + \lim_{s \rightarrow 0^+} \epsilon(sd) = \nabla f(x^0)^\top d.$$

Now using that $\|d\|_2 = 1$ together with the Cauchy inequality

$$-\|\nabla f(x^0)\|_2 \leq \Delta(d) = \nabla f(x^0)^\top d \leq \|\nabla f(x^0)\|_2. \quad (33)$$

The upper and lower bound is achieved when $d = \nabla f(x^0) / \|\nabla f(x^0)\|_2$ and $d = -\nabla f(x^0) / \|\nabla f(x^0)\|_2$, respectively. \square

As a corollary, we have that the search direction is a *descent direction* if it has an obtuse angle with the gradient

Corollary 2.6 (Descent Condition). If $d^\top \nabla f(x_0) < 0$ then there exists $s > 0$ such that

$$f(x_0 + sd) < f(x_0).$$

Proof: From (33) we have that $\Delta(d) = \nabla f(x_0)^\top d < 0$. Let $c = -\nabla f(x_0)^\top d > 0$. Let $s > 0$ be such that $\epsilon(sd) < \frac{c}{2}$. Consequently from (20) we have that

$$\frac{f(x_0 + sd) - f(x_0)}{s} = \nabla f(x_0)^\top d + \epsilon(sd) \leq -\frac{c}{2} < 0. \quad \square$$

2.3 Optimality conditions

Let $f \in C^2$ be twice continuously differentiable throughout this section.

In most cases it is impossible to solve a general optimization problem such as (17). A much more modest, and often attainable, objective is to find a local minima instead.

Definition 2.7. The point $x^* \in \mathbb{R}^n$ is a *local minima* of $f(x)$ if there exists $r > 0$ such that

$$f(x^*) \leq f(x), \quad \forall \|x - x^*\|_2 < r. \quad (34)$$

Thus a local minima point x^* is one where all neighbouring points are worse. This is often good enough for most applications (though not *all* applications). Finding local minima is made possible through the following necessary and sufficient conditions.

Theorem 2.8 (Necessary optimality conditions). If x^* is a local minima of $f(x)$ then

1. $\nabla f(x^*) = 0$
2. $d^\top \nabla^2 f(x^*) d \geq 0, \quad \forall d \in \mathbb{R}^n.$

Proof: The fact that $\nabla f(x^*) = 0$ follows from Corollary 2.6.

Now suppose there exists $d \in \mathbb{R}^n$ such that $d^\top \nabla^2 f(x^*) d < 0$. Suppose w.l.o.g that $\|d\|_2 = 1$. Using the second order Taylor expansion we have that

$$f(x^* + sd) = f(x^*) + \frac{s^2}{2} d^\top \nabla^2 f(x^*) d + \epsilon(sd)s^2.$$

Let $\delta > 0$ be such that for $s \leq \delta$ we have that $\epsilon(sd) < |d^\top \nabla^2 f(x^*) d|/4$. Dividing the above by s^2 , for $s \leq \delta$ we have that

$$\begin{aligned} \frac{f(x^* + sd)}{s^2} &= \frac{f(x^*)}{s^2} + \frac{1}{2} d^\top \nabla^2 f(x^*) d + \epsilon(sd) \\ &< \frac{f(x^*)}{s^2} + \frac{1}{4} d^\top \nabla^2 f(x^*) d, \end{aligned}$$

thus $f(x^* + sd) < f(x^*)$ for all $s \leq \delta$ which contradicts the definition of local minima. \square

Not only are the conditions in Theorem 2.8 necessary, with a slight modification, they are also sufficient.

Theorem 2.9 (Sufficient Local Optimality conditions). If $x^* \in \mathbb{R}^n$ is such that

1. $\nabla f(x^*) = 0$
2. $d^\top \nabla^2 f(x^*) d > 0, \quad \forall d \in \mathbb{R}^n$ with $d \neq 0$,

then x^* is a local optima.

Proof: Let $d \in \mathbb{R}^n$ such that $\|d\|_2 = 1$. Because $\nabla^2 f(x^*)$ is positive definite, the smallest non-zero eigenvalue must also be strictly positive. Consequently

$$\lambda_{\min}(\nabla^2 f(x^*)) \leq d^\top \nabla^2 f(x^*) d, \quad \forall \|d\|_2 = 1.$$

Using the second-order Taylor expansion, we have that

$$\begin{aligned} f(x^* + d) &= f(x^*) + \frac{1}{2} d^\top \nabla^2 f(x^*) d + \epsilon(d) \|d\|_2^2 \\ &\geq f(x^*) + \frac{\|d\|_2^2}{2} \lambda_{\min}(\nabla^2 f(x^*)) + \epsilon(d) \|d\|_2^2. \end{aligned}$$

Let $r > 0$ be such that every d with $\|d\| \leq r$ we have that

$$|\epsilon(d)| < \lambda_{\min}(\nabla^2 f(x^*)) / 4 \quad \Rightarrow \quad \epsilon(d) > -\lambda_{\min}(\nabla^2 f(x^*)) / 4.$$

Thus for $\|d\| \leq r$ we have

$$\begin{aligned} f(x^* + d) &\geq f(x^*) + \frac{\|d\|_2^2}{2} \lambda_{\min}(\nabla^2 f(x^*)) + \epsilon(d) \|d\|_2^2 \\ &\geq f(x^*) + \frac{\|d\|_2^2}{4} \lambda_{\min}(\nabla^2 f(x^*)) > f(x^*). \quad \square \end{aligned}$$

Exercise 2.10. Let $f(x) = \frac{1}{2} x^\top A x - x^\top b + c$, with A symmetric positive definite. How many local/global minimas can $f(x)$ have? Find a formula for the minima using only the *data* A and b .

Proof: By the sufficient conditions x^* is a local minima if $Ax = b$ and $A \succ 0$. Since $Ax = b$, there exists only one local minima.

2.4 Convex functions

In some cases, we can even find a global minima of (17). Indeed, for the *convex functions*, all local minima are global minima. We say a function is convex if

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y), \quad \forall x, y \in \mathbb{R}^d, t \in [0, 1]. \quad (35)$$

Theorem 2.11. If f is a convex function, then every local minima of f is also a global minima.

Proof: Let x^* be a local minima and suppose there exists $\bar{x} \in \mathbb{R}^n$ such that $f(\bar{x}) < f(x^*)$. Now consider the line segment $z_t = t\bar{x} + (1-t)x^*$ for $t \in [0, 1]$. By the definition of convexity we have that

$$f(z_t) = f((1-t)\bar{x} + tx^*) \leq (1-t)f(\bar{x}) + tf(x^*) < (1-t)f(x^*) + tf(x^*) = f(x^*). \quad (36)$$

This shows that x^* cannot be a local minima. Indeed, for any $r > 0$ with $r \leq \|\bar{x} - x^*\|_2$, we have that by choosing $t = 1 - r/\|\bar{x} - x^*\|_2$ we have that

$$\|z_t - x^*\|_2 = (1-t)\|\bar{x} - x^*\|_2 \leq r.$$

Yet from (36) we have that $f(z_t) < f(x^*)$. A contradiction. Thus there exists no \bar{x} with $f(\bar{x}) < f(x^*)$. \square

When f is twice differentiable, there are three equivalent ways of defining convexity.

Theorem 2.12. If f is twice continuously differentiable, then the following three statements are equivalent

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y), \quad \forall x, y \in \mathbb{R}^d, t \in [0, 1]. \quad (37)$$

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x), \quad \forall x, y \in \mathbb{R}^d. \quad (38)$$

$$0 \leq d^\top \nabla^2 f(x) d, \quad \forall x, d \in \mathbb{R}^d. \quad (39)$$

Proof: We only prove that (37) \Rightarrow (38) \Rightarrow (39). The remaining proof of (39) \Rightarrow (37) is left as an exercise.

((37) \Rightarrow (38)) We can deduce (38) from (37) by dividing by t and re-arranging

$$\frac{f(y + t(x - y)) - f(y)}{t} \leq f(x) - f(y).$$

Now taking the limit $t \rightarrow 0$ gives

$$\langle \nabla f(y), x - y \rangle \leq f(x) - f(y).$$

((38) \Rightarrow (39)) First we prove this holds for 1-dimensional functions $\phi : \mathbb{R} \rightarrow \mathbb{R}$. For $x, y \in \mathbb{R}$ and from (38) we have that

$$\phi(y) \geq \phi(x) + \phi'(x)(y - x),$$

$$\phi(x) \geq \phi(y) + \phi'(y)(x - y).$$

Combining the above two we have that

$$\phi'(x)(y - x) \leq \phi(y) - \phi(x) \leq \phi'(y)(y - x).$$

Dividing by $(y - x)^2$ we have

$$\frac{\phi'(y) - \phi'(x)}{y - x} \geq 0, \quad \forall x, y, x \neq y.$$

Taking the limit of $y \rightarrow x$ we have that

$$\phi''(x) \geq 0, \quad \forall x.$$

To then extend to any n -dimensional function, consider the 1-dimensional function $\phi(t) = f(x + tv)$.

Now note that

$$\phi''(t)|_{t=0} = \left. \frac{d^2 f(x + tv)}{dv^2} \right|_{t=0} = v^\top \nabla^2 f(x) v \geq 0, \quad \forall v \neq 0.$$

□

We also say that $f(x)$ is strictly convex when $d^\top \nabla^2 f(x) d > 0$, $\forall d \in \mathbb{R}^d$ with $d \neq 0$.

Exercise 2.13 (Homework). Let $A \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix. Prove Theorem (2.12) for quadratic functions $f(x) = \frac{1}{2}x^\top Ax + x^\top b + c$. What can we say about A when f is convex or strictly convex?

2.5 Gradient method

Since the negative gradient is the direction that minimizes the local change, it is only natural to use the negative gradient as a descent direction. This suggests the following method

$$x^{k+1} = x^k - s^k \nabla f(x^k).$$

But the question still remains how to choose s^k the stepsize? One natural strategy is to choose s^k that results in the maximum descent. That is,

$$s^k = \arg \min_{s \geq 0} f(x^k + sd^k). \quad (40)$$

Finally, we also need a criteria to determine when we should stop iterating. Typically the user will choose a *tolerance to error* $\epsilon > 0$, and then stop the algorithm either when we are at an approximate stationary point

$$\|\nabla f(x^k)\|_2 \leq \epsilon \quad (41)$$

or when the rate in decrease is small

$$f(x^{k+1}) - f(x^k) \leq \epsilon. \quad (42)$$

This gives the steepest descent method in Algorithm (2).

To give some insight into these questions, we now prove that Algorithm 2 converges for a constant stepsize and quadratic functions

Algorithm 2 Steepest Descent

- 1: Choose $x^0 \in \mathbb{R}^n$.
 - 2: **while** $\|\nabla f(x^k)\|_2 > \epsilon$ or $f(x^{k+1}) - f(x^k) \leq \epsilon$ **do**
 - 3: Calculate $d^k = -\nabla f(x^k)$
 - 4: Calculate $s^k = \arg \min_{s \geq 0} f(x^k + sd^k)$
 - 5: Update $x^{k+1} = x^k + s^k d^k$.
-

Theorem 2.14. Let $A \in \mathbb{R}^{n \times n}$ is symmetric positive definite. Let $f(x) = \frac{1}{2}x^\top Ax - x^\top b + c$. If we choose a fixed stepsize of $s^k = 1/\sigma_{\max}(A)$ then the iterates of Algorithm 2 converge according to

$$\|\nabla f(x^{k+1})\|_2 \leq \left(1 - \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)}\right)^k \|\nabla f(x^0)\|. \quad (43)$$

Proof:

$$\begin{aligned} \nabla f(x^{k+1}) &= Ax^{k+1} - b \\ &= A(x^k - s\nabla f(x^k)) - b \\ &= A(x^k - s(Ax^k - b)) - b \\ &= Ax^k - b - sA(Ax^k - b) \\ &= (I - sA)\nabla f(x^k). \quad \square \end{aligned}$$

Taking norms

$$\|\nabla f(x^{k+1})\|_2 \leq \|I - sA\|_2 \|\nabla f(x^k)\|.$$

By choosing $s = 1/\sigma_{\max}(A)$ we have that $I - sA$ is symmetric positive definite and

$$\|I - sA\|_2 = 1 - s\sigma_{\min}(A) = 1 - \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)} < 1.$$

Robert: Homework: Prove this last step!

Thus finally

$$\|\nabla f(x^{k+1})\|_2 \leq \left(1 - \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)}\right) \|\nabla f(x^k)\| \leq \left(1 - \frac{\sigma_{\min}(A)}{\sigma_{\max}(A)}\right)^k \|\nabla f(x^0)\|.$$

The rate of convergence in (43) may be slow. Indeed, it depends on how far the smallest and largest singular are from each other. Perhaps the issue is that we used a fixed stepsize instead of the “best” stepsize (40)? It turns out that also this greedy strategy of choosing the best step size in (40) can also lead slow convergence. This is partly due to a phenomenon called *zig-zagging*. When using the optimal stepsize (40), sequential gradients are orthogonal to one another, that is

$$\nabla f(x^{k+1})^\top \nabla f(x^k) = 0. \quad (44)$$

To prove this, note that by the definition of s^k in (40) we have that

$$\frac{d}{ds} f(x^k - s\nabla f(x^k)) \Big|_{s=s^k} = 0.$$

Using the chain-rule we have that

$$\frac{d}{ds} f(x^k - s\nabla f(x^k)) \Big|_{s=s^k} = -s^k \nabla f(x^k - s\nabla f(x^k))^\top \nabla f(x^k) = 0.$$

Thus so long as $s^k \neq 0$, then (44) holds. This zigzagging can make the global convergence of steepest descent painfully slow. One way to fix this zigzagging is to use a constant stepsize. Yet another that works well in practice is to substitute the exact line search with an approximate line search such as the *Backtracking line search*, see Algorithm 3. This line search method tries increasingly smaller stepsizes until a sufficient decrease in the function is obtained.

Algorithm 3 Backtracking Line Search

- 1: Choose $\bar{\alpha} > 0, \rho, c \in (0, 1)$. Set $\alpha = \bar{\alpha}$
 - 2: **while** $f(x^k + \alpha d^k) \leq f(x^k) + c\alpha \nabla f(x^k)^\top d^k$ **do**
 - 3: Update $\alpha = \rho\alpha$.
-

Algorithm 2 also converges for all *smooth* functions, that is when there exists $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in \mathbb{R}^n. \quad (45)$$

See the notes here http://users.ece.utexas.edu/~cmcaram/EE381V_2012F/Lecture_4_Scribe_Notes.final.pdf for a similar proof for all smooth functions.

2.6 Newton's method

Using the second order Taylor expansion we can motivate the Newton method. At each iteration, the *Newton Method* minimizes the local quadratic approximation

$$q_k(x) = f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{2}(x - x^k)^\top \nabla^2 f(x^k)(x - x^k), \quad (46)$$

as a proxy for minimizing $f(x)$. Assuming $f(x)$ is strictly convex, the stationary point x^{k+1} characterized by $\nabla_x q_k(x) = 0$ of (46) is given by

$$\nabla_x q_k(x) = \nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k) = 0.$$

Isolating x^{k+1} we have

$$x^{k+1} = x^k - \nabla^2 f(x^k)^{-1} \nabla f(x^k).$$

The advantage that Newton's method has over the steepest descent is that it converges quadratically when x is close to the minimum.

Theorem 2.15. Let $f(x)$ be a μ -strongly convex function, that is, we have a global lower bound on the Hessian given by

$$\nabla^2 f(x) \succeq \mu I, \quad \forall x \in \mathbb{R}^n. \quad (47)$$

Furthermore, if the Hessian is also Lipschitz

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L\|x - y\|_2 \quad (48)$$

then Newton's method converges locally and quadratically fast according to

$$\|x^{k+1} - x^*\|_2 \leq \frac{L}{2\mu} \|x^k - x^*\|_2^2. \quad (49)$$

In particular if $\|x^0 - x^*\|_2 \leq \frac{\mu}{L}$, then for $k \geq 1$ we have that

$$\|x^k - x^*\|_2 \leq \frac{1}{2^{2^k}} \frac{\mu}{L}. \quad (50)$$

Proof:

$$\begin{aligned} x^{k+1} - x^* &= x^k - x^* - \nabla^2 f(x^k)^{-1} \left(\nabla f(x^k) - \nabla f(x^*) \right) \\ &= x^k - x^* - \nabla^2 f(x^k)^{-1} \int_{s=0}^1 \nabla^2 f(x^k + s(x^* - x^k))(x^k - x^*) ds \quad (\text{Mean value theorem}) \\ &= \nabla^2 f(x^k)^{-1} \int_{s=0}^1 \left(\nabla^2 f(x^k) - \nabla^2 f(x^k + s(x^* - x^k)) \right) (x^k - x^*) ds \end{aligned}$$

Taking norms we have that

$$\begin{aligned} \|x^{k+1} - x^*\|_2 &\leq \|\nabla^2 f(x^k)^{-1}\|_2 \int_{s=0}^1 \|\nabla^2 f(x^k) - \nabla^2 f(x^k + s(x^* - x^k))\|_2 \|x^k - x^*\|_2 ds \\ &\stackrel{(48)+(47)}{\leq} \frac{L}{\mu} \int_{s=0}^1 s \|x^k - x^*\|_2^2 ds \\ &= \frac{L}{2\mu} \|x^k - x^*\|_2^2. \end{aligned}$$

Finally if $\|x^0 - x^*\|_2 \leq \frac{\mu}{L}$, then by induction and assuming that (50) holds we have that

$$\|x^{k+1} - x^*\| \leq \frac{L}{2\mu} \|x^k - x^*\|_2^2 \leq \frac{L}{2\mu} \frac{1}{2^{2^k}} \frac{1}{2^{2^k}} \left(\frac{\mu}{L} \right)^2 < \frac{1}{2^{2^{k+1}}} \frac{\mu}{L},$$

which concludes the induction proof. \square

The assumptions on for this proof can be relaxed, since we only require the Hessian is Lipschitz and lower bounded in a $\frac{\mu}{2L}$ -ball around x^* .

3 Nonlinear programming with constraints

For this section, I highly recommend reading Chapter 12 in [3].

The objectives of this section are

1. To introduce nonlinear optimization with constraints and provide a geometric intuition.
2. Establish necessary first order optimality conditions which can be easily verified with linear algebra.

Some prerequisites for understanding the section:

- Closed, open and compact sets and their interaction with continuous functions.
- The extreme value theorem.

Let f, g_i and h_j be C^1 continuous functions, for $i = 1, \dots, m$ and $j = 1, \dots, p$. Consider the *constrained* optimization problem

$$\begin{aligned} & \min_{x \in \mathbb{R}^n} && f(x) \\ \text{subject to} && g_i(x) \leq 0, & \text{ for } i \in I. \\ && h_j(x) = 0, & \text{ for } j \in J, \end{aligned} \tag{51}$$

where $I = \{1, \dots, m\}$, $J = \{1, \dots, p\}$, $g_i(x) \leq 0$ are the *inequality constraints* and $h_j(x) = 0$ are the *equality constraint*. If a point $x \in \mathbb{R}^n$ satisfies all the inequality and equality constraints we say that x is a *feasible* point. We refer to the set of all feasible points as the *feasible set* which we denote by

$$X \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : g_i(x) \leq 0, h_j(x) = 0, \text{ for } i = 1, \dots, m, \text{ and } j = 1, \dots, p\}.$$

If x is such that $g_i(x) = 0$, we say that the i th inequality constraint is *saturated* or *active* at x .

Exercise 3.1. Solve the following problem graphically

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & (x_1 - 3)^2 + (x_2 - 2)^2 \\ \text{subject to} \quad & x_1^2 - x_2 - 3 \leq 0, \quad (52) \\ & x_2 - 1 \leq 0, \quad (53) \\ & -x_1 \leq 0. \quad (54) \end{aligned}$$

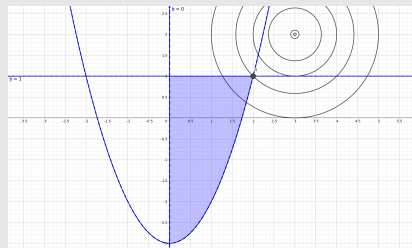


Figure 2: Plot of feasible set

Adding constraints can make the problem easier to solve as compared to the unconstrained problem (17), since now we do not need to search the whole of \mathbb{R}^n for a solution, but rather only a subset X . Indeed, if $X = \{x_0\}$ then clearly the solution is x_0 . But constraints can add many difficulties. Indeed, even with smooth differentiable functions g_i and h_j , the frontier of X may be non-smooth (linear constraints define a polyhedron!). Thus it can be hard to guarantee feasibility and find descent directions that are feasible. That is why first we focus on characterizing feasible descent directions. See the examples in the beginning of Chapter 12 in [3] for examples on the difficulties that constraints introduce.

Theorem 3.2 (Existence). If the feasible set X is bounded and non-empty, then there exists a solution to (51).

Proof: Given that the sets $\mathbb{R}_- = [-\infty, 0]$ and $\{0\}$ are closed, by the continuity of g_i and h_j we have that X is closed. Indeed,

$$X = \left(\bigcap_{i=1}^m g_i^{-1}([-\infty, 0]) \right) \cap \left(\bigcap_{j=1}^p h_j^{-1}(\{0\}) \right),$$

and thus is a finite intersection of closed sets. By assumption X is bounded, thus it is compact. By the continuity of f we have that $f(X)$ is also compact (The Extreme value theorem). Consequently there exists a minimum in $f(X)$. \square

Definition 3.3. We say that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is coercive if $\lim_{\|x\| \rightarrow \infty} f(x) = \infty$.

Theorem 3.4. If X is non-empty and f is coercive, then there exists a solution to (51)

Proof: Let $x_0 \in X$. Define $B_r := \{x : \|x\| \leq r\}$. Since f is coercive, there exists r such that for each x with $\|x\| \geq r$ we have that $f(x) \geq f(x_0)$. Otherwise we would be able to construct a sequence x_k with $\|x_k\| \rightarrow \infty$ such that $f(x_k) \leq f(x_0)$, which contracts the coercivity of f .

Thus clearly the minimum of f is in B_r . Since B_r is bounded and closed, we have that $x_0 \in B_r \cap X$ thus it is bounded, closed and nonempty. Again by the extreme value theorem, $f(x)$ attains its minimum in $B_r \cap X$, which is also the minimum in X . \square

3.1 Admissible and Feasible directions

To design iterative methods for solving constrained optimization problems, we need to know how to move from one given point and still remain within the feasible set X . For instance if X was a polyhedra, then we would say that d is a *feasible* or an *admissible* direction at $x_0 \in X$ if there exists $\epsilon > 0$ such that $x_0 + td \in X$ for all $0 \leq t \leq \epsilon$. To account for the case that the frontier of the feasible set is nonlinear (not a straight line), we need to consider a more general notion of feasible directions.

Definition 3.5. We say that d is an *admissible* direction at $x_0 \in X$ if there exists a C^1 differentiable curve $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}^n$ such that

1. $\phi(0) = x_0$
2. $\phi'(0) = d$
3. There exists $\epsilon > 0$ such that $t \leq \epsilon$ we have $\phi(t) \in X$

We denote by $A(x_0)$ the set of admissible directions at x_0 .

Some examples of admissible sets

- As a straight forward example, given $d \in \mathbb{R}^n$ let $X = \{x \mid \exists \alpha \in \mathbb{R}, x = \alpha d\}$. For any $x_0 \in X$ we have that $A(x_0) = X$.
- Consider the circle $X = \{(\cos(\theta), \sin(\theta)) \mid 0 \leq \theta \leq 2\pi\} \subset \mathbb{R}^2$. Then for every $x_0 = ((\cos(\theta_0), \sin(\theta_0)))$ we have that

$$A(x_0) = \{(-\alpha \sin(\theta), \alpha \cos(\theta)), \forall \alpha \in \mathbb{R}\}.$$

We will often compose differentiable functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with this curve and make use of the following representation of the first order Taylor expansion

Lemma 3.6. Let $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}^n$ be a C^1 curve as defined in Definition 3.5. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable. Then the first order Taylor expansion of the composition $f(\phi(t))$ around x_0 can be written as

$$f(\phi(t)) = f(x_0) + td^\top \nabla f(x_0) + t\hat{\epsilon}(t), \tag{55}$$

where $\lim_{t \rightarrow 0} \hat{\epsilon}(t) = 0$.

Proof:

Since both f and ϕ are C^1 , their composition is also C^1 . Thus $f(\phi(t))$ has a first order Taylor expansion around $t = 0$ which is

$$f(\phi(t)) = f(\phi(0)) + t \frac{df(\phi(t))}{dt} \Big|_{t=0} + t\epsilon(t).$$

Now it is just a matter of plugging in $\phi(0) = x_0$ and using the chain-rule (25):

$$\frac{df(\phi(t))}{dt} \Big|_{t=0} = (\phi'(t))^\top \nabla f(\phi(t)) \Big|_{t=0} = (d^\top \nabla f(x_0)). \quad \square$$

Thus Lemma 3.6 shows that if we are interested in the local behaviour $\phi(t)$ can be replaced by a line segment up to terms that go to zero faster than t .

A necessary condition on a direction being admissible is given in the following.

Proposition 3.7. Let $I_0(x_0) = \{i : g_i(x_0) = 0, i \in I\}$ be the indexes of saturated inequalities. If $d \in A(x_0)$ is an admissible direction then

1. For every $i \in I(x_0)$ we have that $d^\top \nabla g_i(x_0) \leq 0$.
2. For every $j \in J$ we have that $d^\top \nabla h_j(x_0) = 0$.

Let $B(x_0)$ be the set of directions that satisfy the above two conditions. Thus $A(x_0) \subset B(x_0)$.

Proof: 1. Let $i \in I(x_0)$. Let $\phi(t)$ be the curve associated to d . Consider the 1st order Taylor expansion of g_i around x_0 in the d direction which is

$$\begin{aligned} g_i(\phi(t)) &\stackrel{(55)}{=} g_i(x_0) + td^\top \nabla g_i(x_0) + t\epsilon(t) \\ &= td^\top \nabla g_i(x_0) + t\epsilon(t) \\ &\leq 0, \end{aligned}$$

where we used $g_i(\phi(t)) \leq 0$ for t sufficiently small. Dividing by t we have that

$$d^\top \nabla g_i(x_0) + \epsilon(t) \leq 0.$$

Letting $t \rightarrow 0$ we have that $d^\top \nabla g_i(x_0) \leq 0$.

2. Using the first order Taylor expansion of h_j around x_0 we have that

$$h_j(\phi(t)) \stackrel{(55)}{=} h_j(x_0) + td^\top \nabla h_j(x_0) + t\epsilon(t) = td^\top \nabla h_j(x_0) + t\epsilon(t) = 0. \quad (56)$$

Dividing by t and then taking the limit as $t \rightarrow 0$ gives $d^\top \nabla h_j(x_0) = 0$. \square

Note that $B(x_0)$ is a cone, and we will refer to $B(x_0)$ as the cone of feasible directions. Because the cone $B(x_0)$ is easy to work with, we would like to use $B(x_0)$ instead $A(x_0)$. Indeed, it only depends on linear constraints involving gradients. But sometimes $B(x_0)$ is not equivalent to $A(x_0)$ when there exists certain degeneracies.

Example 3.8. Consider the constraint given by

$$h_1(x) = (x_1^2 + x_2^2 - 2)^2 = 0.$$

Thus

$$\nabla h_1(x) = 2(x_1^2 + x_2^2 - 2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Consequently for every feasible point we have that $\nabla h_1(x) = 0$. As such, we have that $B(x) = \mathbb{R}^2$ for every feasible point. Yet $h_1(x) = 0$ describes a circle, and clearly $A(x)$ is the tangent line at x . Thus we cannot use $\nabla h_1(x)$ to describe feasible directions.

Note that we would not have this problem if instead we used the equivalent constraint

$$h_1(x) = (x_1^2 + x_2^2 - 2) = 0.$$

To exclude the possibility of such degeneracies, we impose the *Constraint qualifications*, which are the conditions that ensure we can use $B(x_0)$ instead of $A(x_0)$.

Definition 3.9. We say that the constraint qualifications hold at x_0 if for every $d \in B(x_0)$ there exists a sequence $(d_t)_{t=1}^{\infty} \in A(x_0)$ such that $d_t \rightarrow d$.

An example where the constraint qualifications do not hold, and consequently $B(x_0) \neq A(x_0)$ is given in the example above.

The following theorem justifies why constraint qualifications are important.

Theorem 3.10 (Necessary conditions). Let x^* be a local minimum of (51) and suppose the constraint qualification holds at x^* . It follows that for every $d \in B(x^*)$ we have that $\nabla f(x^*)^\top d \geq 0$. In other words, all direction in the feasible cone are not descent directions.

Proof: Let $d_k \in A(x_*)$ be a sequence such that $d_k \rightarrow d$. Let ϕ_k be the curve associated to d_k . According to Lemma 3.6 we have that the first order Taylor expansion of $f(\phi_k(t))$ can be written as

$$f(\phi_k(t)) = f(x_*) + t\nabla f(x_*)^\top d_k + t\epsilon_k(t). \quad (57)$$

Since x_* is a local minima, there exists T for which $t \leq T$ we have that $f(x_*) \leq f(\phi_k(t))$. Consequently

$$t\nabla f(x_*)^\top d_k + t\epsilon_k(t) = f(\phi_k(t)) - f(x_*) \geq 0, \quad \text{for } t \leq T.$$

Dividing by t and taking the limit we have

$$\lim_{t \rightarrow 0} \nabla f(x_*)^\top d_k + \epsilon_k(t) = \nabla f(x_*)^\top d_k \geq 0.$$

Taking the limit in k concludes the proof. □

3.2 Lagrange's condition

Consider the problem (51) but without inequality constraints, that is,

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{subject to} \quad & h_j(x) = 0, \quad \text{for } j \in J, \end{aligned} \tag{58}$$

The following theorem gives us necessary conditions for a point to be a local minima that requires only using linear algebra.

Theorem 3.11. Let $x^* \in X$ be a local minima and suppose that the constraint qualifications hold at x^* for (51). It follows that the gradient of the objective is a linear combination of the gradients of constraints at x^* , that is, there exists $\mu_j \in \mathbb{R}$ for $j \in J$ such that

$$\nabla f(x^*) = \sum_{j \in J} \mu_j \nabla h_j(x^*). \tag{59}$$

Proof: Let $E = \text{span}(\{\nabla h_1(x^*), \dots, \nabla h_p(x^*)\})$. Let us re-write $\nabla f(x^*) = y + z$ where $y \in E$ and $w \in E^\perp$, thus

$$-z^\top \nabla h_j(x^*) = 0, \quad \forall j \in J.$$

Thus by definition $-z \in B(x^*)$. Consequently by Theorem 3.10 we have that

$$-z^\top \nabla f(x^*) \geq 0.$$

It follows that

$$-z^\top \nabla f(x^*) = -z^\top y - \|z\|_2^2 = -\|z\|_2^2 \geq 0.$$

Consequently $z = 0$ and $\nabla f(x^*) = y \in E$. □

3.3 Karush, Kuhn and Tuckers condition

Consider again the general problem with inequality constraints

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \quad \text{for } i \in I. \\ & h_j(x) = 0, \quad \text{for } j \in J, \end{aligned} \tag{60}$$

The extension of Lagrange's condition to allow to inequality constraints is known at the Karush, Kuhn Tuckers condition.

Theorem 3.12 (Necessary conditions). Let $x^* \in X$ be a local minima and suppose that the constraint qualifications hold at x^* for (58). It follows that there exists $\mu_j \in \mathbb{R}$ and $\lambda_i \in \mathbb{R}_+$ for

$j \in J$ and $i \in I(x^*)$ such that

$$\nabla f(x^*) = \sum_{j \in J} \mu_j \nabla h_j(x^*) - \sum_{i \in I(x^*)} \lambda_i \nabla g_i(x^*). \quad (61)$$

Proof: From Theorem 3.10 we know that for every $d \in B(x^*)$ we have that $\nabla f(x^*)^\top d \geq 0$. That is, said more explicitly, we have that for every $d \in \mathbb{R}^n$ that satisfies

$$\begin{aligned} -d^\top \nabla g_i(x^*) &\geq 0, & \text{for } i \in I(x^*) \\ d^\top \nabla h_j(x^*) &= 0, & \text{for } j \in J, \end{aligned}$$

we have that $d^\top \nabla f(x^*) \geq 0$. This statement fits perfectly the 2nd version of Farkas Theorem 3.28. That is, let

$$A = [-\nabla g_1(x^*), \dots, -\nabla g_m(x^*)] \quad \text{and} \quad B = [\nabla h_1(x^*), \dots, \nabla h_p(x^*)],$$

and $b = \nabla f(x^*)$. Thus for every

$$d \in \{d : A^\top d \geq 0, \quad B^\top d = 0\}$$

we have that $d^\top b \geq 0$. By Farkas Theorem 3.28 we have that this statement can be equivalently re-written as there exists $(\lambda, \mu) \in P$ where

$$P = \{(\lambda, \mu) : A\lambda + B\mu = b, \quad \lambda \geq 0\}.$$

Through which we conclude that the set

$$P = \{\mu \in \mathbb{R}^p, \lambda \in \mathbb{R}_+^m : \sum_{j \in J} \mu_j \nabla h_j(x^*) - \sum_{i \in I(x^*)} \lambda_i \nabla g_i(x^*) = \nabla f(x^*)\},$$

is non-empty, which concludes the proof. \square

Theorem 3.13 (Sufficient conditions). Let f and g_i for $i \in I$ be convex functions. Let h_j be linear for $j \in J$. Suppose the constraint qualifications hold at $x^* \in X$ and that (61) holds. Then x^* is a local minima.

Proof: Since the KKT conditions hold, there exist $\mu_j \in \mathbb{R}$ and $\lambda_i \in \mathbb{R}_+$ for $j \in J$ and $i \in I(x^*)$ such that (61) holds. Let $x \in X$. Since $f(x)$ is convex, we have that

$$\begin{aligned} f(x) &\geq f(x^*) + \nabla f(x^*)^\top (x - x^*) \\ &\stackrel{(61)}{=} f(x^*) + \sum_{j \in J} \mu_j \nabla h_j(x^*)^\top (x - x^*) - \sum_{i \in I(x^*)} \lambda_i \nabla g_i(x^*)^\top (x - x^*). \end{aligned} \quad (62)$$

Using the linearity of each h_j we have that

$$\nabla h_j(x^*)^\top (x - x^*) = h_j(x) - h_j(x^*) = 0. \quad (63)$$

Since each g_i is convex, we have that

$$\nabla g_i(x^*)^\top (x - x^*) \leq g_i(x) - g_i(x^*) \stackrel{i \in I(x^*)}{=} g_i(x) \leq 0. \quad (64)$$

Plugging the above into (62) gives

$$\begin{aligned} f(x) &\stackrel{(63)}{\geq} f(x^*) - \sum_{i \in I(x^*)} \lambda_i \nabla g_i(x^*)^\top (x - x^*). \\ &\stackrel{(64)}{\geq} f(x^*) - \sum_{i \in I(x^*)} \lambda_i g_i(x) \geq f(x^*). \quad \square \end{aligned} \quad (65)$$

From now on we will refer to conditions given in the preceding Theorem 3.13 as the KKT conditions. That is,

KKT conditions. There exists x that is feasible $x \in X$ and $\mu \in \mathbb{R}^{|J|}$ and $\lambda \in \mathbb{R}^{|I|}$ such that (61) holds.

The KKT conditions are often described with the help of an auxiliary function called the Lagrangian function

$$L(x, \mu, \lambda) = f(x) - \langle \mu, h(x) \rangle + \langle \lambda, g(x) \rangle, \quad (66)$$

where we are using $h(x) \stackrel{\text{def}}{=} (h_j(x))_{j \in J}$ and $g(x) \stackrel{\text{def}}{=} (g_i(x))_{i \in I}$ for shorthand. Using the Lagrangian we can describe the sufficient conditions for local optimality more succinctly using the following theorem.

Theorem 3.14. Let $x \in \mathbb{R}^n$, $\mu \in \mathbb{R}^{|J|}$ and $\lambda \in \mathbb{R}^{|I|}$. If

$$\nabla_x L(x, \mu, \lambda) = 0 \quad (67)$$

$$\nabla_\mu L(x, \mu, \lambda) = 0 \quad (68)$$

$$\nabla_\lambda L(x, \mu, \lambda) \leq 0 \quad (69)$$

then the KKT conditions holds.

Proof: Differentiating we have that

$$\nabla_x L(x, \mu, \lambda) = \nabla f(x^*) - \sum_{j \in J} \mu_j \nabla h_j(x^*) + \sum_{i \in I(x^*)} \lambda_i \nabla g_i(x^*) \quad (70)$$

$$\nabla_\mu L(x, \mu, \lambda) = h(x) \quad (71)$$

$$\nabla_\lambda L(x, \mu, \lambda) = g(x) \quad (72)$$

Setting the first equation to zero is equivalent to (61). Setting the second equation to zero and restricting the third to be less than zero gives $h(x) = 0$ and $g(x) \leq 0$ and thus x is feasible, and the KKT conditions hold.

Using the KKT conditions, what is the largest sphere that fits in an ellipsoid?

Exercise 3.15.

$$\begin{aligned} \min & -x^2 - y^2 \\ \text{subject to} & ax^2 + by^2 \leq 1, \end{aligned}$$

where $a > b > 0$. Assume that constraint qualifications hold.

Using the KKT conditions and assuming that the constraint is not active, that is $ax^2 + by^2 < 1$, we quickly arrive at the solution $(x, y) = (0, 0)$.

Now assuming the constraint is active, that is $ax^2 + by^2 = 1$, we have the KKT conditions

$$\begin{aligned} 2x &= 2a\lambda x \\ 2y &= 2b\lambda y. \\ ax^2 + by^2 &= 1. \end{aligned}$$

We can enumerate all the solutions as follows. Consider the case that 1) $x \neq 0$. Thus from the KKT we have that $1 = a\lambda$ and consequently $\lambda = a^{-1}$. From $2y = 2b\lambda y$, since $b\lambda \neq 1$ we have that $y = 0$. The feasibility constraint now gives us that $x = \pm a^{-1/2}$. Now consider the case that 2) $x = 0$. If $y \neq 0$, then necessarily $\lambda = b^{-1}$, and feasibility gives us that $y = \pm b^{-1/2}$. In the first case, we have that $f(x, y) = -x^2 - y^2 = -a^{-1}$. Alternatively in the second case we have $f(x, y) = -b^{-1}$. Since $-b^{-1} < -a^{-1} \leq 0$, we have that $(x, y) = (0, \pm b^{-1/2})$ are the two minimum. What is the maximum?

Example 3.16. Let $A \in \mathbb{R}^{n \times n}$ be symmetric positive definite, $B \in \mathbb{R}^{n \times n}$ be invertible and $b, y \in \mathbb{R}^n$. Consider the problem

$$\begin{aligned} \min & \frac{1}{2}x^\top Ax - b^\top x \\ \text{subject to} & Bx = y. \end{aligned}$$

Write the solution x^* to the above as a function of A, B, b and y .

By Theorem 3.11 we have that there exists $\mu \in \mathbb{R}^n$ such that

$$\begin{aligned} Ax^* - b &= B^\top \mu \\ Bx^* &= y \end{aligned}$$

Rearranging gives

$$\begin{pmatrix} A & -B^\top \\ B & 0 \end{pmatrix} \begin{pmatrix} x^* \\ \mu \end{pmatrix} = \begin{pmatrix} b \\ y \end{pmatrix} \tag{73}$$

Thus

$$\begin{pmatrix} x^* \\ \mu \end{pmatrix} = \begin{pmatrix} A & -B^\top \\ B & 0 \end{pmatrix}^{-1} \begin{pmatrix} b \\ y \end{pmatrix}.$$

3.4 Deducing Duality using KKT

We can also use the KKT condition to deduce the strong duality theorem for linear programming. For this we will use the following primal form

$$\begin{aligned} \max_x \quad & c^\top x \\ \text{subject to} \quad & Ax = b, \\ & x \geq 0, \end{aligned} \tag{P}$$

which is the standard form after slack variables are introduced.

Exercise 3.17. Show that

$$\begin{aligned} \min_y \quad & b^\top y \\ \text{subject to} \quad & A^\top y \leq c \end{aligned} \tag{D}$$

is the dual of (P).

Proof: If we change the min for a max in (P) and write down the KKT equations with $\lambda \geq 0$ for the inequalities and y variables for the equalities we get

$$\begin{array}{ll} A^\top y + \lambda = c & \text{Colinear gradients} \\ Ax = b & \text{Enforcing equality constraints} \\ x \geq 0 & \text{Enforcing inequality constraints} \\ \lambda \geq 0 & \text{Positive Lagrange multipliers} \\ x_i \lambda_i = 0, \quad i = 1, \dots, n. & \text{Testing if } x_i \text{ is active} \end{array} \tag{74}$$

We have added one special constraint $x_i \lambda_i = 0$ that checks if the $x_i \geq 0$ constraint is active or not. In particular, if $x_i > 0$ then $\lambda_i = 0$ and thus λ has no effects on the remaining equations. On the hand, when $x_i = 0$ then λ_i can be any positive number. Furthermore, since both x and λ are positive we can rewrite (74) as $x^\top \lambda \geq 0$.

We can also write down the KKT equations of (D). For this, let $x \geq 0$ be the Lagrange parameters associated to the constraint so that

$Ax = b$	Colinear gradients	
$A^\top y \leq c$	Enforcing inequality constraints	
$x \geq 0$	Positive Lagrange multipliers	
$x^\top (A^\top y - c) = 0$	Testing if constraints are active	(75)

Now by renaming $\lambda = c - A^\top y$ we have that the above equations becomes

$A^\top y + \lambda = c$		(76)
$Ax = b$	Colinear gradients	
$\lambda \geq 0$	Enforcing inequality constraints	
$x \geq 0$	Positive Lagrange multipliers	
$x^\top \lambda = 0$	Testing if constraints are active	(77)

are identical to (74). □

The only thing that has changed are the roles of the variables. For the primal problem we have that x are the optimal variables and λ and y are the Lagrangian multipliers. For the dual problem, we have that y is the optimal variable and x and λ are Lagrangian multipliers.

3.5 Project Gradient Descent

Now we come back to designing algorithms that fit the format

$$x^{k+1} = x^k + s_k d^k, \tag{78}$$

such that $f(x_{k+1}) < f(x_k)$ and $x^{k+1} \in X$. In the constrained setting we have the additional problem of enforcing $x^{k+1} \in X$.

Divide tasks: Take one step to decrease f and another to become feasible. For this we need the *Projection Operator*.

$$P_X(z) \stackrel{\text{def}}{=} \arg \min \frac{1}{2} \|x - z\|^2$$

subject to $x \in X$.

With the projection operator we can now define the *projected gradient descent* method

$$x^{k+1} = P_X(x^k - s_k \nabla f(x^k)). \tag{79}$$

First, let us study some examples of projections.

Exercise 3.18. If $X = \{x : \|x\| \leq r\}$ where $r > 0$ show that

$$P_X(z) = r \frac{z}{\|z\|}.$$

Proof: We can solve this project problem

$$\begin{aligned} \min \quad & \frac{1}{2} \|x - z\|^2 \\ \text{subject to} \quad & \|x\|^2 \leq r^2. \end{aligned}$$

Suppose that $\|z\| \leq r$. Clearly $x = z$ is the solution.

Suppose instead $\|z\| > r$. Since $\{x : \|x\|^2 \leq r^2\}$ is a closed set, we know the projection will be on the boundary given by $\|x\| = r$. Let $h(x) = \|x\|^2 - r^2$. Using the KKT conditions we have that

$$\nabla f(x) = -\mu \nabla h(x) \quad \implies \quad (x - z) = -2\mu x.$$

Isolating x gives

$$x = \frac{z}{1 + 2\lambda}. \tag{80}$$

Since $\|x\| = r$ we have that

$$\frac{\|z\|}{1 + 2\mu} = r \quad \implies \quad \frac{1}{1 + 2\mu} = \frac{r}{\|z\|}.$$

Plugging the above into (80) gives

$$x = r \frac{z}{\|z\|}.$$

Exercise 3.19. Let $A \in \mathbb{R}^{n \times n}$ be an invertible matrix and let $b \in \mathbb{R}^n$. If $X = \{x : Ax = b\}$. Show that

$$P_X(z) = z - A^\top (AA^\top)^{-1} (Az - b).$$

Proof: The Lagrangian function associated to the projection is given by

$$L(x, \mu) = \frac{1}{2} \|x - z\|^2 + \mu^\top (Ax - b). \tag{81}$$

Taking the derivative in x and setting to zero gives

$$\nabla_x L(x, \mu) = x - z + A^\top \mu = 0 \quad \Leftrightarrow \quad x = z - A^\top \mu \tag{82}$$

Now using that $Ax = b$ and left multiplying the above by A gives

$$b = Ax = Az - AA^\top \mu = 0.$$

Since A is invertible, we have that AA^\top is invertible. Thus isolating μ in the above gives

$$\mu = (AA^\top)^{-1} (Az - b).$$

Inserting this value for μ into (82) gives

$$x = z - A^\top (AA^\top)^{-1} (Az - b). \quad \square$$

Remark 3.20. We did not need A to be square or invertible to define the projection onto $Ax = b$. Indeed, no matter what A is the set $\{x : Ax = b\}$ is a closed set, and thus there must exist a solution to the projection optimization problem. In general, the projection of z onto $Ax = b$ is given by

$$P_X(z) = z - A^\dagger(Az - b),$$

where A^\dagger is known as the Moore-Penrose Pseudoinverse. Infact, the pseudoinverse of a matrix can be defined as the operator that gives this solution!

The Project Gradient Methods (PGD) has some advantages and disadvantages. On the *good side*, it is general and can be applied to any closed convex set X . When the projection onto X is known in closed form, it can be very efficient to apply. On the *down side* depending on X the projection is not known and can be costly. Furthermore, PGD can converge slowly and suffer from excessive zig-zagging, see Figure 3.

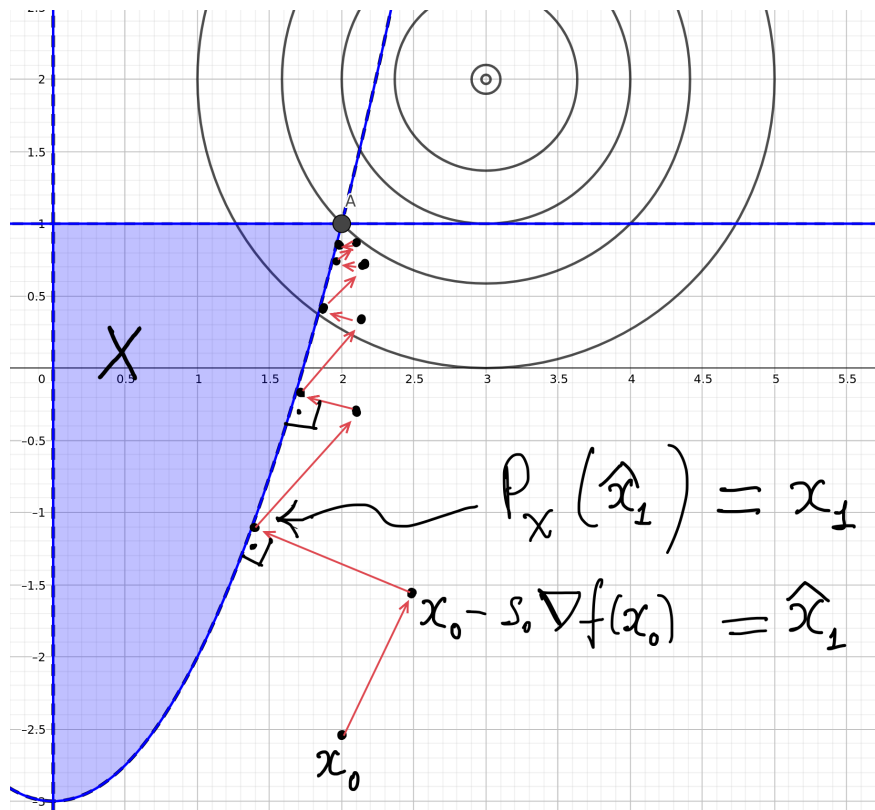


Figure 3: The constraint set and level sets of the problem (54). The PGD can zig-zag and thus converge slowly.

3.6 The constrained descent method

Let us turn the intuition we developed in the previous sections into an algorithm for solving

$$\begin{aligned} & \min_{x \in \mathbb{R}^n} && f(x) \\ & \text{subject to} && g_i(x) \leq 0, \quad \text{for } i \in I. \end{aligned} \tag{83}$$

For this we can use the necessary conditions from Theorem 3.10. That is, say we are given a current feasible point $x^k \in \mathbb{R}^n$ and we wish to find x^{k+1} such that

$$f(x^k) \leq f(x^{k+1})$$

and for which $g_i(x^{k+1}) \leq 0$ for all $i \in I$. To do this, we will look for an admissible direction $d \in \mathbb{R}^n$ that is also a descent direction, that is $\nabla f(x^k)^\top d \leq 0$. From Proposition 3.7 we have that all admissible satisfy $d^\top \nabla g_i(x^k) \leq 0$ for all the active constraints $i \in I(x^k)$.

This suggests the following iterative method in Algorithm 4.

Algorithm 4 Descent Algorithm

- 1: Choose $x^0 \in X$ and $\epsilon > 0$. Set $k = 0$.
 - 2: **while** KKT(x^k) conditions not verified or $\|\nabla f(x^k)\| > \epsilon$ **do**
 - 3: Find d such that $d^\top \nabla f(x^k) \leq 0$ and $d^\top \nabla g_i(x^k) \leq 0, \forall i \in I$ ▷ Find feasible direction
 - 4: Find $s \in \mathbb{R}_+$ such that $f(x^k + sd) < f(x^k)$ and $x^k + sd \in X$ ▷ Descent in set
 - 5: $x^{k+1} = x^k + sd$ ▷ Take a step
 - 6: $k = k + 1$
-

In finding a direction that satisfies line 3, we can solve a minimization problem such as

$$\begin{aligned} & \min_{x \in \mathbb{R}^n} && d^\top \nabla f(x^k) \\ & \text{subject to} && d^\top \nabla g_i(x^k) \leq 0, \quad \forall i \in I(x^k) \\ & && -1 \leq d \leq 1 \end{aligned} \tag{84}$$

3.7 Introduction to Constrained Optimization TP

In the TP2 you will implement a method based on Algorithm 4 but only for problems where g_j are affine functions and in 2D. That is,

$$\begin{aligned} & \min_{x \in \mathbb{R}^n} && f(x) \\ & \text{subject to} && a_i x + b_i y + c_i \leq 0, \quad \text{for } i \in I, \end{aligned} \tag{85}$$

where $g_i(x, y) = a_i x + b_i y + c_i \leq 0$, for $a_i, b_i, c_i \in \mathbb{R}$ for all $i \in I$.

Exercise 3.21. What is $\nabla g_j(x, y) = ?$ What is the solution to (84) if $I(x^k) = \emptyset$? What is the solution if $I(x^k) = \{i\}$?

The situation in which solving (84) can be tricky is when x^k is on a corner. In 2D, this means that there are two active constraints. Let us call them g_1 and g_2 thus $I(x^k) = \{1, 2\}$. In the TP you will be given access to the gradients of the constraints. Let $n_1 := -\nabla g_1(x^k) = -(a_1, b_1)^\top$ and $n_2 := -\nabla g_2(x^k) = -(a_2, b_2)^\top$ so that they point to the interior of their respective constraint. Let $u_1 \in \mathbb{R}^2$ and $u_2 \in \mathbb{R}^2$ be the border vectors of g_1 and g_2 . In other words $u_1^\top(x, y) = 0$ is equivalent to $g_1(x, y) = 0$. We choose u_1 and u_2 such that $u_1^\top n_2 \geq 0$ and $u_2^\top n_1 \geq 0$, respectively.

Based on the inner products $p_1 := \langle u_1, \nabla f(x^k) \rangle$ and $p_2 := \langle u_2, \nabla f(x^k) \rangle$, we now distinguish between three different cases.

1. If $\langle n_1, \nabla f(x^k) \rangle \leq 0$ and $\langle n_2, \nabla f(x^k) \rangle \leq 0$ then $d = -\frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}$ is feasible
2. Else if $p_1 \leq 0$ and $p_2 \leq 0$ then both border vectors are feasible and we choose $i^* = \arg \min_i \{\langle u_i, \nabla f(x^k) \rangle\}$ and $d = u_{i^*}$.
3. Else if $p_1 \geq 0$ and $p_2 \geq 0$ then all descent directions point outside the feasible region (indeed the KKT conditions are verified) and there is no descent direction.
4. Else if $p_1 \leq 0$ and $p_2 \geq 0$ then $d = u_1$.
5. Else $p_2 \leq 0$ and $p_1 \geq 0$ then $d = u_2$.

References

- [1] Marco Chiarandini. “Linear and Integer Programming Lecture Notes”. In: (2015).
- [2] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. 1st ed. Springer Publishing Company, Incorporated, 2014.
- [3] J Nocedal and S J Wright. *Numerical Optimization*. Ed. by Peter Glynn and Stephen M Robinson. Vol. 43. Springer Series in Operations Research 2. Springer, 1999. Chap. 5, pp. 164–75.

Appendix

3.8 Characterizing the constraint qualification

Robert: I didn't teach it last year

First we show that sufficient conditions to characterize an admissible direction. We then move on to use this to establish sufficient and easily verifiable conditions which guarantee that the constraint qualification holds.

Lemma 3.22. Suppose that h_j is an affine function for $j \in J$. Let $x^0 \in X$ and $d \in \mathbb{R}^n$ be such that

$$\begin{aligned} \tilde{d}^\top \nabla g_i(x^0) &< 0, & \text{for } i \in I(x^0) \\ \tilde{d}^\top \nabla h_j(x^0) &= 0, & \text{for } j \in J. \end{aligned}$$

Then d is an admissible direction, that is $d \in A(x_0)$.

Proof: Let $\phi(t) = x^0 + td$. Since h_j is affine we have that

$$h_j(\phi(t)) = h_j(x^0) + td^\top \nabla h_j(x^0).$$

Furthermore, given that x^0 is feasible we have that $h_j(x^0) = 0$ and by assumption $\tilde{d}^\top \nabla h_j(x^0) = 0$, consequently $h_j(\phi(t)) = 0$. Furthermore, for $i \in I(x_0)$ we have that

$$g_i(\phi(t)) = g_i(x_0) + t\tilde{d}^\top \nabla g_i(x^0) + t\epsilon(t).$$

Since $g_i(x^0) = 0$ and $\tilde{d}^\top \nabla g_i(x^0) < 0$, there exists $t > 0$ sufficiently small so that $g_i(\phi(t)) \leq 0$. Thus $x^0 + td \in X$ for $t > 0$ sufficiently small, which proves that d is a feasible direction. \square

We now give sufficient condition for the constraint qualifications to hold.

Lemma 3.23. Suppose h_j is affine for all $j \in J$. Let $x_0 \in X$. If there exists \tilde{d} such that

$$\begin{aligned} \tilde{d}^\top \nabla g_i(x^0) &< 0, & \text{for } i \in I(x^0) \\ \tilde{d}^\top \nabla h_j(x^0) &= 0, & \text{for } j \in J, \end{aligned}$$

then the constraint qualification holds at x^0 .

Proof: Let $d \in B(x^0)$. We will now show that

$$d_t = td + (1-t)\tilde{d},$$

for $t \in [0, 1[$ we have that $d_t \in A(x_0)$ and for $t \rightarrow 0$ we have that d_t converges to d , which in turn shows that the constraint qualifications hold. Indeed, for $i \in I(x_0)$ we have that

$$d_t^\top \nabla g_i(x^0) = t \underbrace{\tilde{d}^\top \nabla g_i(x^0)}_{\leq 0, d \in B(x^0)} + (1-t)\tilde{d}^\top \nabla g_i(x^0) < 0.$$

Furthermore

$$d_t^\top \nabla h_j(x^0) = t\tilde{d}^\top \nabla h_j(x^0) + (1-t)\tilde{d}^\top \nabla h_j(x^0) = 0.$$

Consequently, by Lemma 3.22, we have that for all $t \in [0, 1[$ we have shown that $d_t \in A(x_0)$. Now taking any sequence $t_n \rightarrow 1$ we have that $d_{t_n} \rightarrow d$, thus by definition the constraint qualifications hold at x_0 . \square

We will now use the previously Lemma to develop sufficient condition for the constraint qualifications to hold in the entire feasible set X .

Proposition 3.24. If g_i for $i \in I$ are convex and h_j for $j \in J$ are affine and there exists a single point $\tilde{x} \in X$ such that

$$g_i(\tilde{x}) < 0, \quad h_j(\tilde{x}) = 0 \quad \forall i \in I, j \in J, \quad (86)$$

then the constraint qualification holds for all feasible points in X .

Proof: Let $x^0 \in X$ and let \tilde{x} satisfy (86). Using the 1st order Taylor expansion around \tilde{x} we have that

$$0 > g_i(\tilde{x}) \stackrel{(86)}{\underset{\text{convexity}}{\geq}} g_i(x^0) + \nabla g_i(x^0)^\top (\tilde{x} - x^0).$$

If $i \in I(x^0)$ then $g_i(x^0) = 0$ and from the above we have that $\nabla g_i(x^0)^\top (\tilde{x} - x^0) < 0$. Let $\tilde{d} \stackrel{\text{def}}{=} (\tilde{x} - x^0)$. Furthermore, for the equality constraints we have that

$$h_j(\tilde{x}) \stackrel{\text{affine}}{=} h_j(x^0) + \nabla h_j(x^0)^\top \tilde{d}.$$

Since $h_j(x^0) = 0$ and thus $\nabla h_j(x^0)^\top \tilde{d} = 0$. This shows that \tilde{d} satisfies the conditions of Lemma 3.23, and thus the constraint qualifications hold at x^0 . \square

Yet another alternative characterization of constraint qualifications for a given x_0 , which is easier to verify, is given in the following proposition.

Proposition 3.25. Suppose that h_j are affine functions for $j \in J$. For a given feasible point $x_0 \in X$ we have that the set

$$\{\nabla g_i(x_0) : i \in I_0(x_0)\} \cup \{\nabla h_j(x_0) : j \in J\},$$

is a linearly independent set, then the constraint qualifications hold at x_0 .

Proof: Let $x^0 \in X$ and consider the following two Linear Programs

$$\begin{aligned} \max_{\lambda, \mu} z &= \sum_{i \in I(x^0)} \lambda_i + \sum_{j \in J} 0 \times \mu_j. \\ \text{subject to} \quad &\sum_{j \in J} \mu_j \nabla h_j(x^0) - \sum_{i \in I(x^0)} \lambda_i \nabla g_i(x^0) = 0, \\ &\lambda_i \geq 0, \text{ for } i \in I(x^0). \end{aligned} \quad (\text{P})$$

and

$$\begin{aligned} \min_d w &\equiv 0 \\ \text{subject to} \quad &d^\top \nabla g_i(x^0) \leq -1, \text{ for } i \in I(x^0), \\ &d^\top \nabla h_j(x^0) = 0, \text{ for } j \in J. \end{aligned} \quad (\text{D})$$

It is not hard to show that (P) is the primal and (D) is its associated dual problem. Indeed, to see this, first we right (P) in the standard form by introducing variables $\mu^+, \mu^- \in \mathbb{R}^p$ with $\mu = \mu^+ - \mu^-$

and including the constraints $\mu^+, \mu^- \geq 0$. Then write $c = (\underbrace{1, \dots, 1}_{|I(x^0)|}, \underbrace{0, \dots, 0}_{2p})$, $b = 0 \in \mathbb{R}^n$, and

$$A = [-\nabla g_1(x^0), \dots, -\nabla g_m(x^0), \nabla h_1(x^0), \dots, \nabla h_p(x^0), -\nabla h_1(x^0), \dots, -\nabla h_p(x^0).]$$

From (D) the dual constraints are given by

$$A^\top d \geq c \Rightarrow \begin{cases} d^\top \nabla h_j(x^0) \geq 0, \\ -d^\top \nabla h_j(x^0) \geq 0, \\ -d^\top \nabla g_i(x^0) \geq 1, \quad \text{for } i \in I(x^0), j \in J. \end{cases}$$

Which exactly the constraint of (D). The rest follows by examining (D).

Since (P) is non-empty, since it admits the zero solution. Now suppose that (P) admits a solution λ_i^*, μ^* that is not zero. In which case, we have that the gradients of the constraints must be linearly independent. This contradicts our assumption, thus $(\lambda_i^*, \mu^*) = (0, 0)$ is the optimal solution. Thus (P) is bounded and by strong duality Theorem 1.5 we have that (D) is feasible. Let \tilde{d} be a feasible solution to (D). Note that \tilde{d} satisfies the assumptions of Lemma 3.23, consequently the constraint qualification holds at x_0 . \square

3.9 Farkas Lemma and the Geometry of Polyhedra

Theorem 3.26 (Separating Hyperplane theorem). Let $X, Y \subset \mathbb{R}^n$ be two disjoint convex sets. Then there exists a hyperplane defined by $v \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$ such that

$$\langle v, x \rangle \leq \beta \quad \text{and} \quad \langle v, y \rangle \geq \beta, \quad \forall x \in X, \forall y \in Y.$$

Consider the cone given by

$$K \stackrel{\text{def}}{=} \{A\lambda + B\mu \mid \forall \lambda \geq 0, \forall \mu\}. \quad (87)$$

Now consider the problem of determining is a given vector b is in this cone K or not. Since $\{b\}$ is a convex set and so if the cone K we have by the Separating Hyperplane theorem that if b is not in K , then there exists a hyperplane separating K and $\{b\}$. Furthermore, this hyperplane happens to pass through the origin. We formalize this in the following theorem.

Theorem 3.27 (Geometric Farkas). Consider a given vector b and the cone

$$K \stackrel{\text{def}}{=} \{A\lambda + B\mu \mid \forall \lambda \geq 0, \forall \mu\}. \quad (88)$$

Then either $b \in K$ or there exists a vector y such that

$$\langle y, b \rangle \leq 0 \quad \text{and} \quad \langle y, k \rangle \geq 0, \quad \forall k \in K. \quad (89)$$

Proof: Since K is a cone, $b \in K$ if and only if $\alpha b \in K$ for every $\alpha \geq 0$. Consequently the problem of determining if b is in K is equivalent to determining if the hyperplane $\ell_b = \{\alpha b \mid \forall \alpha \geq 0\}$ intersects with K at a point other than the origin. Fortunately, excluding the origin from K results in a convex set $K \setminus \{0\}$. Thus by the Separating Hyperplane theorem we have that if b is not in K , then there exists a hyperplane separating $K \setminus \{0\}$ and ℓ_b . Clearly this hyperplane must pass through the origin.

Theorem 3.28 (2nd Version of Farkas). Consider the set

$$P = \{(\lambda, \mu) : A\lambda + B\mu = b, \quad \lambda \geq 0\}$$

and

$$Q = \{y : A^\top y \geq 0, \quad B^\top y = 0\}.$$

The set P is non-empty if and only if every $y \in Q$ is such that $b^\top y \geq 0$.

Proof: [Based on Separating Hyperplane] The proof follows by simply reinterpreting the Geometric Farkas Theorem 3.27. Indeed, if $b \in P$ is equivalent to saying that b and the cone (88) intersect. On the other hand, if b is not in K then there exists a separating hyperplane that passes through the origin parametrized by a vector y (89). Consequently

$$\langle y, A\lambda + B\mu \rangle = \langle A^\top y, \lambda \rangle + \langle B^\top y, \mu \rangle \geq 0, \quad \forall \lambda \geq 0, \forall \mu. \quad (90)$$

Since this has to hold for every vector μ it is easy to see that $B^\top y = 0$. Otherwise, fix $\lambda = 0$. If the i th row of $B^\top y$ is non-zero we can choose $\mu = e_i$ and then $\mu = -e_i$ which when inserted into (90) gives

$$\langle B^\top y, e_i \rangle \geq 0 \quad \text{and} \quad \langle B^\top y, e_i \rangle \leq 0,$$

which gives a contradiction and shows that $B^\top y = 0$.

Furthermore $A^\top y \geq 0$. This follows by simply choosing λ as the i th coordinate vector. The converse is also true, since if $A^\top y \geq 0$ and $\lambda \geq 0$ then clearly their inner product is positive. Finally, from (89) we also have that $b^\top y \geq 0$.

Proof: [Old version based on duality theorem] **Part (1) If P non-empty then $y \in Q \Rightarrow b^\top y \geq 0$:** Let $\mu = z^+ - z^-$ where $z^+, z^- \geq 0$. Then clearly P is non-empty if and only if

$$\hat{P} = \{(\lambda, z^+, z^-) : A\lambda + Bz^+ - Bz^- = b, \quad \lambda, z^+, z^- \geq 0\},$$

is non-empty. The dual program of \hat{P} if we consider a zero objective function (with $c = 0$) is given by

$$\begin{aligned} & \min_y y^\top b \\ & \text{subject to } A^\top y \geq 0, \quad B^\top y \geq 0, \quad -B^\top y \geq 0. \end{aligned} \quad (91)$$

See Lemma 3.30 for how to deduce this version of duality. The feasible region of (91) is equivalent to Q . By the weak duality Lemma 1.4 we have that if there exists a finite feasible primal point $y \in Q$ then we have that

$$b^\top y \geq c^\top(\lambda, z^+, z^-) = 0.$$

Part (2) If $y \in Q \Rightarrow b^\top y \geq 0$ then P non-empty:

Clearly $0 \in Q$ and thus the dual (91) is feasible. Since $y \in Q$ implies that $b^\top y \geq 0$, we have that (91) is feasible and bounded. This proves by duality that P is feasible. \square

Proof: [Using 1st Farkas Theorem] Let $\mu = z^+ - z^-$ where $z^+, z^- \geq 0$. Then clearly P is non-empty if and only if

$$\hat{P} = \{(\lambda, z^+, z^-) : A\lambda + Bz^+ - Bz^- = b, \quad \lambda, z^+, z^- \geq 0\},$$

is non-empty. Applying the 1st Farkas Theorem 3.29, we have that \hat{P} is non-empty if every $y \in \hat{Q}$ implies that $y^\top b \geq 0$ where

$$\hat{Q} = \{y : A^\top y \geq 0, \quad B^\top y \geq 0, \quad -B^\top y \geq 0\}.$$

Clearly $Q = \hat{Q}$ which concludes the proof. \square

Theorem 3.29 (Farkas Version 1). Consider the set

$$P = \{x : Ax = b, \quad x \geq 0\}$$

and

$$Q = \{y : A^\top y \geq 0\}.$$

The set P is non-empty if and only if every $y \in Q$ is such that $b^\top y \geq 0$.

Proof: Any text on optimization.

Lemma 3.30 (Duality 2nd version). Consider the following LP with equality constraints

$$\begin{aligned} & \max_x c^\top x \\ & \text{subject to } Ax = b, \\ & \quad x \geq 0, \end{aligned} \tag{LP}$$

Then the dual is given by

$$\begin{aligned} & \min_y y^\top b \\ & \text{subject to } A^\top y \geq c. \end{aligned} \tag{DP}$$

Proof: First we re-write (LP) as

$$\begin{aligned} & \max_x c^\top x \\ & \text{subject to } Ax \leq b, \\ & \quad -Ax \leq -b, \\ & \quad x \geq 0, \end{aligned} \tag{LP}$$

The dual program (D) of the above is given by

$$\begin{aligned} & \min w^\top b - z^\top b \\ & \text{subject to } A^\top w - A^\top z \geq c, \\ & \quad w, z \geq 0. \end{aligned}$$

Substituting $y = w - z$ gives the result.