Mathematical Modelling for Computer Networks- Part I **Spring 2013 (Period 4)**

Math Appendix 3: Convex Optimization: From Convex Sets to KKT rules

Lecturers: Laila Daniel and Krishnan Narayanan *Date:15th March 2013*

If you optimize everything, you will always be unhappy.
- Donald Knuth

Abstract

This lesson provides a concise account of bare minimum needed to state the KKT rules in convex optimization problems. On the road to KKT rules, we describe several useful concepts such as convex sets, convex function, polyhedral convex set, hyperplane, halfspace, local vs global minima and convexity. Lagrangian function, Lagrangian multipliers, saddle point and optimality, KKT rules and KKT point.

The life-breath of modern optimization (also known as mathematical programming) is the notion of convexity as brilliantly portrayed in the quotation below by Rockafellar, one of the founding fathers of the theory of Convex Optimization.

In fact the great watershed in optimzation isn't between linearity and nonlinearity, but convexity and nonconvexity.
- R. Tyrrell Rockafellar

3.1 Convex set, Convex function

Definition 3.1 ***Convex set:***
A subset $\mathcal{X}$ of $\mathbb{R}^n$ is convex, if for any pair of points x_1 and $x_2 \in \mathcal{X}$, the line segment joining x_1 and x_2 lies entirely within the set $\mathcal{X}$.

Notation: A subset $\mathcal{X}$ of $\mathbb{R}^n$ is convex, if for any x_1 and $x_2 \in \mathcal{X} \implies \lambda x_1 + (1-\lambda)x_2 \in \mathcal{X}$ for $\forall \lambda \in [0,1]$. So a set of real vectors is convex if the entire line segment joining any pair of elements in the set lies entirely within the set.

Definition 3.2 ***Convex function:***
Informally a function is convex if the graph of the function resembles a cup, i.e., $\bigcup$. We can state this idea precisely as follows.

A function $f : \mathbb{R} \to \mathbb{R}$, is convex, if the graph of the function between any two points a and b, where $a < b$ lies entirely below the chord joining the points $(a, f(a)$ and $(b, f(b))$. This geometric idea is written formally as

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda)f(b) \quad \lambda \in [0,1]$$

We can extend this idea to define a convex function on an arbitrary convex set that lies in $\mathbb{R}^n$. Let $\mathcal{X}$ be a convex set in $\mathbb{R}^n$. A function $f : \mathcal{X} \to \mathbb{R}^n$ is convex if for any x_1 and $x_2 \in \mathcal{X}$, it holds that $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$, for every $\lambda \in (0,1)$. The function is strictly convex if the inequality is strict for any distinct x_1 and x_2 and any $\lambda \in (0,1)$.

A function f is concave if its negation $-f$ is convex. A function f is strictly concave if its negation $-f$ is strictly convex. So a concave function resembles a cap, i.e., $\bigcap$.

Note: The notion of a convex function and concave function are dual concepts; and we pass from one to other by taking the 'negative'; the key thing about the duality is that by taking the dual of the dual we end up with the original (function we stared with). Things which are dual to each other are roughly like 'opposites' of each other. A set and its complement are duals of each other. A nice consequence of duality is that for every result we prove, we get another for 'free' by invoking duality. For example, from $(A \cup B)^c = (A^c \cap B^c)$, by invoking duality we get $(A \cap B)^c = (A^c \cup B^c)$. So the well-known pair of DeMorgan laws in set theory can be regarded as duals of each other. In the linear algebra lesson, we saw that $ker(A) = ran(A^T)$ implies that $ker(A)$ and $ran(A^T)$ are duals of each other.

Duality is central to mathematical programming. When we study duality in linear programming and convex programming, we shall see that linear and convex programs come in pairs called *primal* and *dual*, which are duals of each other.

The only functions that are both convex and concave are affine functions. An affine function is a ' shifted' linear function that is a linear function with a fixed displacement. So in particular a linear function and a constant function are convex as well as concave.

Hyperplane: An equation of the form $ax = b$ where a is a nonzero vector in $\mathbb{R}^n$, b is an arbitrary scalar ($b \in \mathbb{R}$) and $x \in \mathbb{R}^n$ defines a hyperplane in $\mathbb{R}^n$. It is the set of all the points in $\mathbb{R}^n$ (also called vectors) which satisfy the above equation. Here the vector a appearing in the equation of the hyper plane is called the *normal* defining the hyperplane.

Let $H(a, b)$ denote the hyperplane defined by the non-zero vector $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$. The above hyperplane defines two closed halfspaces $H^+(a, k)$ and $H^-(a, k)$ which are defined as follows.

$$H^+(a, b) = \{x \in \mathbb{R}^n | ax \geq b\}$$

$$H^-(a, b) = \{x \in \mathbb{R}^n | ax \leq b\}$$

Geometrically $H^+(a, b)$ contains vectors which lie in the same side as pointed to by vector a and $H^-(a, b)$ contains vectors which lie in the opposite side to by vector a, i.e, the same side as pointed to vector $-a$ Note that the hyperplane $H(a, b)$ divides the space $\mathbb{R}^n$ into two half spaces, $H^+(a, b)$ and $H^-(a, b)$.

Note: $H^+(a, b)$ and $H^-(a, b)$ are convex sets.
The hyperplane $H(a, b)$ is a convex set. $\mathbb{R}^n = H^+(a, b) \cup H^-(a, b)$
$H(a, b) = H^+(a, b) \cap H^-(a, b)$

That $H(a, b)$ is a convex set can be seen in two ways. By directly applying the definition of convexity to the set and also by noting that the intersection of (arbitrary) convex sets is a convex set.

3.1.1 How to recognize Convex functions

1. The notion of a convex function is intimately related to the notion of a convex set. For a convex function f, the set that lies 'above' the graph of the function called the *epigraph* of the function, is a convex set. This epigraph

viewpoint characterizes a convex function. We state this formally as follows: The epigraph of the function f is defined by

$$epi(f) = \{(x, y) \mid x \in \text{ domain}(f) \text{ and } y \geq f(x)\}$$

. The function f is convex if and only if (iff) $epi(f)$ is a convex set.

2. The graph of a convex function $f : \mathbb{R} \to \mathbb{R}$ between any two points a and b in domain(f) such that the $a < b$ lies below the chord joining the points $(a, f(a))$ and $(b, f(b))$

3. From the graph of a convex function $f : \mathbb{R} \to \mathbb{R}$, for any $a < b < c$, the slope of the chord joining $(a, f(a))$ and $(b, f(b))$ is not greater than the slope of the chord joining $(b, f(b))$ and $(c, f(c))$. we can state this as follows:

$$\frac{(f(b) - f(a))}{b - a} \leq \frac{f(c) - f(a)}{c - a} \leq \frac{f(c) - f(b)}{c - b}$$

4. When a function is first or second order differentiable, convexity can be characterized entirely in terms of its differentiability.

5. Given a differentiable function f defined on an open convex set, then the function is convex (strictly convex) if and only if its derivative Df (also denoted by f') is a monotone (strictly monotone) increasing function on its domain.

6. Analogously, for a function which has a second derivative throughout its domain of definition given by an open convex set, f is convex iff its second derivative, $f'' \geq 0$ on it domain. If the second derivative is $f''(x) > 0$, then f is strictly convex.
   Characterizing a convex function by the behaviour of the first or second order derivatives can be easily generalised to differentiable convex functions defined in $\mathbb{R}^n$.

7. For a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, the above result has an analog which is expressed in terms of the Hessian matrix. f is convex iff the Hessian matrix $\nabla^2 f$ is positive-semidefinite denoted by $\nabla^2 f \succeq 0$, and if the Hessian matrix is strictly positive-semidefinite ($\nabla^2 f \succ 0$) , then f strictly convex function, .

   Note: Here the Hessian matrix $\nabla^2 f$ is an $n \times n$ matrix of second-order partial derivatives
   $\nabla^2 f = (f_{i,j})$ *where* $f_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$

8. The tangent drawn to a convex function at any point on its graph lies below the graph of the function. This elementary result leads to a very useful inequality that can be stated as follows. Assuming that a convex function defined on an open set $U \subset \mathbb{R}^n$ is differentiable at a point x_0 we obtain

$$f(x) \geq f(x_0) + \nabla f(x_0)(x - x_0)$$

   . Here $\nabla f(x_0)$ denotes the gradient of the function f at the point x_0.

   Note that this simple inequality leads to a dramatic conclusion that for a convex function, local minimum equals global minimum (by taking the point x_0 where $\nabla f(x_0) = 0$).

9. We can specialize the above result to a function defined on an open interval in $\mathbb{R}$

$$f(x) \geq f(x_0) + f'(x_0)(x - x_0)$$

10. And finally,the most famous inequality of convex analysis: *Jensen inequality*: For a convex function defined on an open (possibly infinite) interval (a, b), let $x_i \in (a, b)$.
    If $\alpha_i \leq 0$ and $\sum_{i=1}^{n} \alpha_i = 1$, $\forall i$, then

$$f(\sum_{i=1}^{n} \alpha_i x_i) \leq \sum_{i=1}^{n} \alpha_i f(x_i)$$

Note: Dualizing the above, we get corresponding results for a concave function. For example, for a differentiable concave function, the slope is decreases monotonically and so on.

The exponential function e^x is perhaps the most important convex function. The log function $log\, x$ (for $x > 0$) is an important example of concave function. Functions x^2, x^3, x^n are convex functions where as the function $-\frac{1}{x}$ is concave. Often the characterization of convex functions using their derivatives can be used to verify that the example functions are either convex or concave.

3.1.2 Some important examples of convex/concave functions on $\mathbb{R}$

Convex

- affine: $ax + b$ on $\mathbb{R}$, for any $a, b \in \mathbb{R}$
- exponential: e^{ax}, for any $a \in \mathbb{R}$
- powers: x^α on $\mathbb{R}_{++}$, for $\alpha \geq 1$ or $\alpha \leq 0$
- powers of absolute value: $|x|^p$ on $\mathbb{R}$, for $p \geq 1$
- negative entropy: $x\, log\, x$ on $\mathbb{R}_{++}$

Concave

- affine: $ax + b$ on $\mathbb{R}$, for any $a, b \in \mathbb{R}$
- powers: x^α on $\mathbb{R}_{++}$, for $0 \leq \alpha \geq 1$
- logarithm: $log\, x$ on $\mathbb{R}$
- entropy: $-x\, log\, x$ on $\mathbb{R}_{++}$

3.1.3 Examples of convex/concave functions on $\mathbb{R}^n$ and $\mathbb{R}^{m \times n}$

- affine functions are convex and concave
- all norms are convex
- Examples on $\mathbb{R}^n$
  - affine function $f(x) = a^T x + b$
  - norms: $\|x\|_p = (\sum_{i=1}^{n} |x|^p)$ for $p \geq 1$; $\|x\|_\infty = \max_k |x_k|$

  - norms: $\|x\|_p = (\sum_{i=1}^{n} |x|^p)^{1/p}$ for $p \geq 1$; $\|x\|_\infty = \max_k |x_k|$

- Examples on $\mathbb{R}^{m \times n}$ ($m \times n$ matrices)
  - affine function

$$f(X) = tr(A^T X) + b = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{i,j} X_{i,j} + b$$

  - spectral (maximum singular value)

$$f(X) = \|X\| = \sigma_{max}(X) = (\lambda_{max}(X^T X))^{1/2}$$

Constrained optimization problems

In constrained optimization problems, we optimize (maximize or minimize) an arbitrary real-valued function over an arbitrary set in $\mathbb{R}^n$which which may be implicitly defined using functional constraints. The function to be optimized may have multiple local optima and the algorithms designed to find the global optimum may get find it difficult to get past local optimum it may enter during the search. Solutions to nonlinear optimizations problems can be difficult to find in general. In constrained optimization problems, the set over which a real-valued function has to be maximized or minimized is of often specified by means of inequality constraints (such as $g_i(x) \leq 0$) based on functions (such as $g_i : \mathbb{R}^n \to \mathbb{R}$).

Convexity is a special structure that gives convex optimization problems well-developed theory, efficient algorithms and a variety of applications spanning numerous fields such as communication system design, optimal control, machine learning, statistics etc.

In a *convex optimization problem* we minimize a convex function (or equivalently maximize a concave function) defined over a convex set. This convex set is often implicitly defined by means of inequalities involving convex functions (such as the functions g_i described earlier being convex).

Example: Let $\mathcal{X}$ be a convex set in $\mathbb{R}^n$. Suppose the functions $f_i : \mathcal{X} \to \mathbb{R}$ are convex over $\mathcal{X}$ and b_i are real numbers for $i = 1, \ldots 1$. Then the set $\{f_i \leq b_i, i \in [n]\} \doteq \{x \in \mathbb{R}^n \mid f_i(x) \leq b_i, i \in [n]\}$ is a convex set. Here $[n]$ denotes the set $\{1, \ldots, n\}$.

Example: This example is a special case of the last example. If A is an $m \times n$ matrix with real entries, (we write this as $A \in M_{mn}(\mathbb{R})$) and $b \in \mathbb{R}^m$, then $\{x \in \mathbb{R}^n \mid Ax \leq b\}$ is a convex set. This set is called a *polyhedral convex set* and is denoted by $P(A, b)$.

Note: Geometrically, each inequality defined by a row of the matrix A yields a half space in $\mathbb{R}^n$ and the set given by $P(A, b)$ is the intersection of these half spaces. $P(A, b)$ is convex as it is a finite intersection of half spaces which are convex sets.

3.2 Local and Global Optima

Definition 3.3 *Feasible solution vs Optimal solution: In the problem of minimizing a function $f : \chi \to \mathcal{R}$ any x belonging to the domain χ is a feasible solution and an element x^* in χ is a global optimal solution (or just a solution) if $f(x^*) \leq f(x)$ $\forall x \in \chi$.*

Convexity implies local minimum = global minimum. If f is strictly convex over X, then the local minimum is a unique global minimum.

3.3 The Karush-Kuhn-Tucker (KKT) Conditions

In dealing with a convex optimization problem, we minimize a convex objective function (equivalently maximize a concave objective function) subject to constraints which are given by convex functions. So the constraints together define a convex set and so a convex optimization problem in one of minimizing a convex function over a convex set. Any point that lies within the convex set is called a *feasible point* or a *feasible solution* of a convex optimization problem. The *optimal solution* is a feasible solution which yields a bounded (finite) minimum over the set of all feasible solutions.

Primal Problem

$$\begin{array}{ll} & \min\ f(x) \\ \text{subject to} & \\ & g_i(\mathbf{x}) \leq 0, \quad 1 \leq i \leq m \\ & \mathbf{x} \in \mathbb{R}^n \end{array}$$

Assumption: $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$ are all convex and differentiable functions over $\mathbb{R}^n$.

The Lagrangian function we introduce helps to convert the constrained optimization problem in to an unconstrained optimization problem which can be simpler to solve at the expense of introducing additional variables called *Lagrangian multipliers*, denoted by $\lambda_i \in \mathbb{R}_{++}$, $i = 1, \ldots m$. Th Lagrangian multiplier λ_i corresponds to the ith constraint $g_i(x) \leq 0$. We form the Lagrangian function $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ given below and the basic philosophy of this approach is to obtain the optimal solution as the *saddle point* of the Lagrangian function. This requires the *stationarity condition* $\mathcal{S}$ and the *complementary slackness* (CS) condition to be satisfied at a feasible point x^*. The conditions $\mathcal{S}$ and CS together are called the *KKT conditions*.

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \doteq f + \boldsymbol{\lambda}\mathbf{g} = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

Given a feasible $x^* \in \mathbb{R}^n$, $\exists \boldsymbol{\lambda}^* \in \mathbb{R}^n_+$,

$$(\mathcal{S}) \quad \nabla f(x^*) + \sum \lambda_i^* \nabla g_i(x^*) = 0$$

$$(CS) \quad \sum \lambda_i g_i(x^*) = 0$$

Then x^* is a global optimal solution for the primal problem.

Here $\mathcal{S}$ is the Stationarity condition and CS is the Complementary Slackness. The conditions of stationarity and complementary slackness together is called the KKT conditions. If x^* satisfies the KKT conditions, then x^* is a *KKT point*. The vector $\boldsymbol{\lambda}^*$ which figures in the equations labelled $\mathcal{S}$ and CS is called a *KKT vector*. Therefore x^* in conjunction with $\boldsymbol{\lambda}^*$ yields the saddle point $(x^*, \boldsymbol{\lambda}^*)$ of the Lagrangian function $\mathcal{L}$

The complementary slackness condition implies that if at x^*, the primal constraint $g_i(x^*)$ is met with a strict inequality $(g_i(x^*) < 0)$, i.e., there is *slack* in the ith

constraint, then the corresponding $\lambda_i^* = 0$ as λ_i is non-negative. If at x^*, the primal constraint $g_i(x^*)$ is met with equality ($g_i(x^*) = 0$), the corresponding $\lambda_i^* > 0$. Hence the name complementary slackness. A constraint that is satisfied with equality such as ($g_i(x) = 0$ or $\lambda_i = 0$) is called a *binding* or *active*. Complementary slackness requires that corresponding to the saddle point (x^*, λ^*) of the Lagrangian, a component $\lambda_i^* = 0$ of the vector λ^* is positive precisely when constraint g_i is tight ($g_i(x^*) = 0$) and $\lambda_i = 0$ when the constraint g_i is inactive ($g_i(x^*) < 0$).

So at optimality, Slack in primal constraints $g_i \implies$ corresponding $\lambda_i = 0$
Constraint g_i met with equality $\implies$ corresponding $\lambda_i \geq 0$

Example
This example is taken from 'Appendix C: Convex Optimization' of the book [KMK04]

$$\min\ (x_1 - 5)^2 + (x_2 - 5)^2$$

subject to

$$x_1^2 + x_2^2 - 5 = 0$$

$$\frac{1}{2}x_1 + x_2^2 - 2 = 0$$

$$-x_1 \leq 0$$

$$-x_2 \leq 0$$

$$\mathbf{x} \in \mathbb{R}^2$$

Plug this to the standard primal form, we can see that

$$m = 4,\ n = 2$$

$$f(x_1, x_2) = (x_1 - 5)^2 + (x_2 - 5)^2$$

$$g_1(x1, x_2) = x_1^2 + x_2^2 - 5$$

$$g_2(x_1, x_2) = \frac{1}{2}x_1 + x_2^2 - 2$$

$$g_3(x_1, x_2) = -x_1, \quad g_4(x_1, x_2) = -x_2$$

Also we can find

$$\nabla f(x) = \begin{bmatrix} 2(x_1 - 5) \\ 2(x_2 - 5) \end{bmatrix}$$

$$\nabla g_1(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}, \quad \nabla g_2(x) = \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}$$

$$\nabla g_3(x) = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad \nabla g_4(x) = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

Consider the point

$$x^* = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

At this point the first and the second constraints are binding. We have

$$\nabla f(x^*) = \begin{bmatrix} -6 \\ -8 \end{bmatrix}, \quad \nabla g_1(x^*) = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \quad \nabla g_2(x^*) = \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}$$

$$\text{Take } \lambda = \begin{bmatrix} \frac{2}{3} \\ \frac{20}{3} \\ 0 \\ 0 \end{bmatrix},$$

We can see that at the point x^*

$\nabla f(x^*) + \lambda_1 \nabla g_1(x^*) + \lambda_2 \nabla g_2(x^*) = 0$ (Stationarity condition) and
(x^*, λ^*): $\sum_{i=1}^{4} \lambda_i g_i(x^*) = 0$, because $\lambda_1 = \lambda_2 = 0$ (Complementary slackness)

$$\text{The KKT conditions are satisfied at the point } x^* = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Since $f(x^*) \leq f(x)\ \forall x \in \chi$, x^* is a global optimal solution. The optimum value of $f(x)$ is 25. From Figure 3.1 we can see that $f(x) = 25$ passes through the x^*. As $f(x)$ is strictly convex, x^* is the unique global optimum.

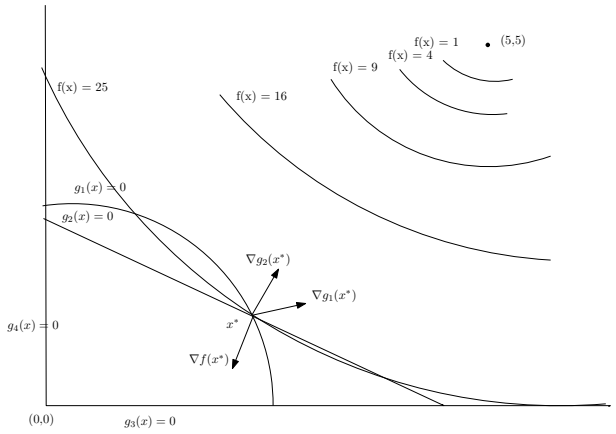

Figure 3.1: Geometry of the convex optimization example problem. The quarter area of circles centred at (5,5) are potions of contours of constant value of the objective function: that is, they have equations $(x_1 - 5)^2 + (x_2 - 5)^2 = h$, for various values of $h \geq 0$; shown are these curves for $h = 1, 4, 9, 25$.

References

[BV] Stephen Boyd and Lieven Vandenberghe, http://www.stanford.edu/ boyd/cvxbook/

[KMK04] Anurag Kumar, D. Manjunath, and Joy Kuri, "Communication Networking: An Analytical Approach", *Morgan Kaufmann* , 2004.