10-725/36-725: Convex Optimization **Fall 2016**

Lecture 12: KKT conditions

Lecturer: Ryan Tibshirani *Scribes: Jayanth Krishna Mogali, Hsu-Chieh Hu*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

12.1 KKT Conditions

Consider the general optimization problem (**P**) shown below, where we have not assumed anything regarding the functions f, g, h (like convexity). We define **G** as the dual of **P**.

P

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0,\ i = 1, \ldots, m \\ & l_j(x) = 0,\ j = 1, \ldots, r \end{aligned}$$

G

$$\begin{aligned} \max_{u,v} \ \min_x \quad & f(x) + \textstyle\sum_{i=1}^m u_i\, h_i(x) + \sum_{j=1}^r v_j l_j(x) \\ \text{subject to} \quad & u \geq 0 \end{aligned}$$

We define the function $g(u, v) = \min_x L(x, u, v)$ where $L(x, u, v) \triangleq f(x) + \sum_{i=1}^m u_i\, h_i(x) + \sum_{j=1}^r v_j l_j(x)$. We first state the KKT conditions associated with problem **P**, they are:

1. Stationarity Condition
$$0 \in \partial \left(f(x) + \sum_{i=1}^m u_i\, h_i(x) + \sum_{j=1}^r v_j l_j(x) \right)$$
2. Complementary Slackness
$$u_i \cdot h_i(x) = 0,\ i = 1, \ldots, m$$
3. Primal feasibility
$$\begin{aligned} h_i(x) &\leq 0\,, i = 1, \ldots, m \\ l_j(x) &= 0,\ j = 1, \ldots, r \end{aligned}$$
4. Dual Feasibility
$$u_i \geq 0\,, i = 1, \ldots, m$$

We begin by first explaining what each of the KKT conditions state, and later discuss their implications. We begin by noting that the KKT conditions apply to a triplet x, u, v. The stationarity condition tells us that for the given dual variable pair u, v, the point x minimizes the lagrangian $L(x, u, v)$. For convex f, g, h the stationarity condition can be alternatively written as

$$0 \in \partial f(x) + \sum_{i=1}^{m} u_i \, \partial h_i(x) + \sum_{j=1}^{r} v_j \partial l_j(x)$$

The complementary slackness condition applies only to inequality constraints. For the i^{th} inequality constraint, complementary slackness tells us that at x, either $h_i(x) = 0$ **or** the corresponding dual variable $u_i = 0$. If $h_i(x) = 0$, we say that the inequality constraint is tight at x.

Primal feasibility basically tells us that x must satisfy all the constraints specified in problem **P**.

Dual feasibility tells us that the dual variables associated with the inequality constraints must be non-negative.

We will next see that the KKT conditions on x, u, v are in a very broad sense equivalent to having an optimal primal solution x and optimal dual solution u, v at the same time. In other words, x is a solution to the primal problem **P** and u, v is a solution to the dual problem **G** at the same time.

12.1.1 Necessity

If x^* and u^*, v^* are the primal and dual solutions respectively with zero duality gap, we will show that x^*, u^*, v^* satisfy the KKT conditions. It is important to note that we are assuming zero duality gap, for example if **P** was a convex problem, then, strong duality is implied if Slater's condition holds for **P**.

$$\begin{aligned}
f(x^*) &= g(u^*, v^*) \text{ by zero duality gap assumption} \\
&= \min_x f(x) + \sum_{i=1}^{m} u_i^* h_i(x) + \sum_{j=1}^{r} v_j^* l_j(x), \text{ by definition} && (12.1) \\
&\leq f(x^*) + \sum_{i=1}^{m} u_i^* h_i(x^*) + \sum_{j=1}^{r} v_j^* l_j(x^*) && (12.2) \\
&\leq f(x^*) && (12.3)
\end{aligned}$$

Eqn (12.2) follows from eqn (12.1) since minimization of $L(x, u^*, v^*)$ in eqn (12.1) is carried with respect to all $x \in \mathbb{R}^n$, $L(x^*, u^*, v^*)$ must be atleast greater than $g(u^*, v^*)$. Eqn (12.3) follows from that the fact that since x^* is optimal to our primal problem **P**, it must satisfy $h_i(x^*) \leq 0, \forall\, i$ and $l_j(x^*) = 0, \forall\, j$. Similarly since u^*, v^* are solutions to dual problem **G**, we should have $u_i^* \geq 0, \forall\, i$. Combining these two facts yeilds the relation $u_i^* h_i(x^*) \leq 0, \forall\, i$ and $v_j^* l_j(x^*) = 0, \forall\, j$. Looking at equations eqn (12.1) and (12.3), we can conclude that the inequalities appearing in eqns (12.2) and (12.3) can actually be replaced by equalities. Two things that we have learnt from the above set of equations:

1. By the equality of eqns (12.1) and (12.2), the point x^* minimizes $L(x, u^*, v^*)$ over $x \in \mathbb{R}^n$. Hence the subdifferential of $L(x, u^*, v^*)$ must contain 0 at $x = x^*$ —is exactly the stationarity condition.

2. From equality of eqns (12.2) and (12.3), we must have $\sum_{i=1}^{m} u_i^* h_i(x^*) = 0$, and since each term here is ≤ 0, this implies $u_i^* h_i(x^*) = 0, \ \forall\, i$ —this is exactly complementary slackness.

12.1.2 Sufficiency

We will show that if there exists x^*, u^*, v^* that satisfy KKT conditions, then x^* and u^*, v^* are primal and dual optimal. Assume there exists x^*, u^*, v^* that satisfies KKT conditions for problem **P**, then,

$$g(u^*, v^*) = f(x^*) + \sum_{i=1}^{m} u_i^* h_i(x^*) + \sum_{j=1}^{r} v_j^* l_j(x^*), \text{ from stationarity condition} \quad (12.4)$$

$$= f(x^*) \quad (12.5)$$

Eqn (12.5) follows from the fact that $u_i^* h_i(x^*) = 0, \forall\, i$ due to complementary slackness and $l_j(x^*) = 0$ due to primal feasibility. Therefore we have shown that the duality gap is 0 at x^* and u^*, v^* (and x^* and u^*, v^* are primal and dual feasible from KKT conditions). Recall from last lecture, if we ever have a zero duality gap then we necessarily have the solutions, hence x^* and u^*, v^* are primal and dual optimal.

12.1.3 Putting it together

In summary,

1. For any optimization problem, if x^* and u^*, v^* satisfy KKT conditions for the problem, then satisfying those KKT conditions is sufficient to imply that x^* and u^*, v^* are the optimal solutions for the primal and it's dual. This statement is equivalent to saying satisfying KKT conditions is always sufficient for optimality.
2. If strong duality holds and we have solutions for the problem, then those solutions must necessarily satisfy KKT conditions.

An easy way to remember the above equivalence is suppose we know that strong duality holds (for example a convex problem satisfying Slater's conditions) then:
x^* and u^*, v^* are primal and dual solutions $\iff$ x^* and u^*, v^* satisfy the KKT conditions.
An important warning concerning the stationarity condition: for a differentiable function f, we cannot use $\partial f(x) = \{\nabla f(x)\}$ unless f is convex. The motivation for this warning is from the fact that for non-convex problems, the gradient of the function cannot be substituted in place of the sub-differential as a general rule to satisfy the stationarity condition. In fact, a sub-differential may not even exist for a differentiable non-convex function.
Another warning concerning stationarity condition is when atleast one of f, g, h is non-convex, we cannot assume $\partial \left(f(x) + \sum_{i=1}^{m} u_i\, h_i(x) + \sum_{j=1}^{r} v_j l_j(x) \right) = \partial f(x) + \sum_{i=1}^{m} u_i\, \partial h_i(x) + \sum_{j=1}^{r} v_j \partial l_j(x)$.

12.1.4 Origins Of KKT Conditions

1. KKT conditions first appeared in a publication by Kuhn and Tucker in 1951. KKT conditions were originally called KT conditions until recently.
2. Later people found out that Karush had the conditions in his unpublished master's thesis of 1939, so KT conditions have since been referred to as KKT conditions to acknowledge the contribution by Karush.

A side point, for unconstrained problems, the KKT conditions are nothing more than the subgradient optimality condition. Another side-point, for general constrained convex optimization problems, recall we

could have pushed the constraints into the objective through their indicator functions and obtained an equivalent convex problem. The KKT conditions for the constrained problem could have been derived from studying optimality via subgradients of the equivalent problem, i.e.

$$0 \in \partial f(x^*) + \sum_{i=1}^{m} \mathcal{N}_{h_i \leq 0}(x^*) + \sum_{j=1}^{r} \mathcal{N}_{l_j = 0}(x^*)$$

where $\mathcal{N}_C(x)$ is the normal cone of C at x.

12.2 Examples

12.2.1 Example: Quadratic with equality constraints

Consider the problem below for $Q \succeq 0$,

$$\begin{aligned} \min_x \quad & \frac{1}{2} x^T Q x + c^T x \\ \text{subject to} \quad & Ax = 0 \end{aligned}$$

We will derive the KKT conditions for the above quadratic problem. Lets start by noting that the problem is convex and Slater's condition definitely holds (check at x = 0, Ax = 0 so feasible), hence strong duality holds. We start with the Lagrangian,

$$L(x, u) = \frac{1}{2} x^T Q x + c^T x + u^T (Ax)$$

Since the Lagragian is differentiable we can arrive at the stationarity condition by setting $\nabla_x L(x, u) = 0$. Hence, the stationarity condition is equivalent to

$$Qx + c + A^T u = 0 \tag{12.6}$$

Since there are no inequality constraints in the problem, we do not have any equations for complementary slackness. For primal feasibility, we need to satisfy

$$Ax = 0 \tag{12.7}$$

Note, there are no constraints for dual feasibility since u is allowed to be unconstrained. Eqns (12.6) and (12.7) can be succintly written as

$$\underbrace{\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix}}_{\text{KKT matrix}} \begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} -c \\ 0 \end{bmatrix} \tag{12.8}$$

Eqn (12.8) can be solved in closed form. The KKT matrix will reappear when we discuss Newton's method.

12.2.2 Example: water-filling

Consider the following power allocation optimization problem:

$$\begin{aligned} \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad & -\textstyle\sum_{i=1}^{n} \log(\alpha_i + x_i) \\ \text{subject to} \quad & x \geq 0, \quad 1^T x = 1 \end{aligned} \tag{12.9}$$

This problem arises from information theory, where each variable xi represents the transmitter power allocated to the i-th channel and $\log(\alpha_i + x_i)$ gives the capacity or communication rate of the channel. The problem can be regarded as allocating a total power of one to the channels in order to maximize the total communication rate.

The KKT conditions are:

- Stationarity:
$$\frac{-1}{\alpha_i + x_i} - u_i + v = 0, \quad i = 1, \ldots, n \tag{12.10}$$
- Complementary slackness:
$$u_i \cdot x_i = 0, i = 1, \ldots, n \tag{12.11}$$
- Primal feasibility:
$$x \geq 0, \quad 1^T x = 1 \tag{12.12}$$
- Dual feasibility:
$$u_i \geq 0 \tag{12.13}$$

From above results, we know that

$$v - \frac{1}{\alpha_i + x_i} v = u_i \geq 0, \quad i = 1, \ldots, n \tag{12.14}$$

and

$$(v - \frac{1}{\alpha_i + x_i}) \cdot x_i = 0, i = 1, \ldots, n \tag{12.15}$$

We argue that if $v \geq 1/\alpha_i$, then x_i must be 0; otherwise $v = 1/(\alpha_i + x_i)$. Using the primal feasibility we need to solve the following problem to get v:

$$\sum_{i=1}^{n} \max\{0, 1/v - \alpha_i\} = 1 \tag{12.16}$$

This is a univariate equation and easy to solve. This reduced problem is called water-filling. Here the i can be thought as the ground level above patch i, and then we flood the region with water to a depth $1/v$. We can increase the flood level until we have used a total amount of water equal to one.

12.2.3 Example: support vector machines

$$\begin{array}{ll} \underset{\beta \in \mathbb{R}^p,\ \beta_0 \in \mathbb{R},\ \xi \in \mathbb{R}^n}{\text{minimize}} & (1/2)\|\beta\|_2^2 + C \sum_{i \in \mathcal{S}} \xi_i \\ \text{subject to} & \xi_i \geq 0, \quad y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \quad i = 1, \ldots, n, \end{array} \tag{12.17}$$

The KKT stationarity conditions we have

$$0 = \sum_{i=1}^{n} w_i y_i, \quad \beta = \sum_{i=1}^{n} w_i y_i x_i, \quad w = C1 - v \tag{12.18}$$

The complementary slackness is

$$v_i \xi_i = 0, \quad w_i(1 - \xi_i - y_i(x_i^T \beta + \beta_0)) = 0, \quad i = 1, \ldots, n, \tag{12.19}$$

Hence at optimality we have $\beta = \sum_{i=1}^{n} w_i y_i x_i$ and w_i is nonzero only if $1 - \xi_i - y_i(x_i^T \beta + \beta_0)$ Such points i are called the support points.

We note that KKT conditions does not give a way to find solution of primal or dual problem-the discussion above is based on the assumption that the dual optimal solution is known. However, it gives a better understanding of SVM: the dual variable w_i acts as an indicator of whether the corresponding point contributes to the decision boundary. This fact can give us more insight when dealing with large-scale data: we can screen away some non-support points before performing optimization.

12.3 Constrained and Lagrange forms

Often in statistics and machine learning well switch back and forth between constrained form, where $t \in R$ is a tuning parameter,

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & h(x) \leq t, \end{aligned} \tag{12.20}$$

and Lagrange form, where $\lambda \geq 0$ is a tuning parameter

$$\min f(x) + \lambda \cdot h(x) \tag{12.21}$$

and claim these are equivalent. We will show this claim is almost always true given the condition that f and h are both convex.

(C) to (L): if problem (C) is strictly feasible, then strong duality holds, and there exists some $\lambda > 0$ (dual solution) such that any solution x in (C) minimizes

$$\min f(x) + \lambda \cdot (h(x) - t) \tag{12.22}$$

so x^* is a solution in (L).

(L) to (C): if x is a solution in (L), then the KKT conditions for (C) are satisfied by taking $t = h(x)$, so x is a solution in (C).
conclusion:

$$\cup_{\lambda \geq 0}\{\text{solutions in (L)}\} \subseteq \cup_t\{\text{solutions in (C)}\} \tag{12.23}$$

$$\cup_{\lambda \geq 0}\{\text{solutions in (L)}\} \supseteq \cup_{\text{(C) is strict feasible}}\{\text{solutions in (C)}\} \tag{12.24}$$

12.4 Back to duality

One of the most important uses of duality is that, under strong duality, we can characterize primal solutions from dual solutions. Recall that under strong duality, the KKT conditions are necessary for optimality. Given dual solutions u^*, v, any primal solution x^* satisfies the stationarity condition

$$0 \in \partial f(x^*) + \sum_{i=1}^{m} u_i^* \, \partial h_i(x^*) + \sum_{j=1}^{r} v_j^* \partial l_j(x^*)$$

In other words, x solves $\min_x L(x, u, v)$

- Generally, this reveals a characterization of primal solutions

- In particular, if this is satisfied uniquely (i.e., above problem has a unique minimizer), then the corresponding point must be the primal solution

References

- S. Boyd and L. Vandenberghe (2004), "Convex optimization", Chapter 5.
- R. T. Rockafellar (1970), "Convex analysis", Chapters 28-30.