10-725/36-725: Convex Optimization **Fall 2015**

Lecture 11: October 7th

Lecturer: Lecturer: Ryan Tibshirani *Scribes: Scribes: Cinnie Hsiung, Yuyan Wang, Zicheng Cai*

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

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

11.1 Recap from Last Time

Given the following convex minimization problem:

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \le 0, \quad i = 1, \ldots, m \\ & \ell_j(x) = 0, \quad j = 1, \ldots, r \end{aligned} \tag{11.1}$$

The Lagrangian is defined as $L(x, u, v) = f(x) + \sum_{u=1}^m u_i h_i(x) + \sum_{j=1}^r v_j \ell_j(x)$. The Lagrange dual function is defined as $g(u, v) = \min_x L(x, u, v)$. The dual problem is

$$\begin{aligned} \max_{u,v} \quad & g(u, v) \\ \text{subject to} \quad & u \ge 0 \end{aligned}$$

There are a few important things to note.

(1) The Lagrange dual function $g(u, v)$ is always concave regardless of whether the primal problem is convex or not.

(2) *Weak duality*: $f^* \ge g^*$ holds for all problems, where f^* and g^* are primal and dual optimal values, respectively.

(3) Slaters's condition, which says the primal has at least one strictly feasible point, is a sufficient condition for *strong duality* to hold. If $\exists x$ such that $h_i(x) < 0, i = 1, \ldots, m$ and $\ell_j(x) = 0, j = 1, \ldots, r$, $f^* = g^*$. This condition can be further refined to $h_i(x) < 0$ for all i such that h_i is nonaffine. As a result, Slater's condition is reduced to feasibility for LP's.

11.2 Karush-Kuhn-Tucker (KKT) Conditions

For the given problem (11.1), the KKT conditions are:

(1) $0 \in \partial_x \Big(f(x) + \sum_{i=1}^m u_i h_i(x) + \sum_{j=1}^r v_j \ell_j(x) \Big)$ (stationary)

(2) $u_i \cdot h_i(x) = 0$ for all i (complementary slackness)

(3) $h_i(x) \leq 0, \ell_j(x) = 0$ for all i, j (primal feasibility)

(4) $u_i \geq 0$ for all i (dual feasibility)

Theorem 11.1 *For x^* and u^*, v^* to be primal and dual solutions, KKT conditions are sufficient.*

Proof: Sufficiency: if $\exists x^*$ and u^*, v^* that satisfy the KKT conditions, $g(u^*, v^*) = f(x^*) + \sum_{i=1}^m u_i^* h_i(x^*) + \sum_{j=1}^r v_j^* \ell_j(x^*) = f(x^*)$ The first equality holds from stationarity, since $f(x) + \sum_{i=1}^m u_i h_i(x) + \sum_{j=1}^r v_j \ell_j(x)$ is convex, so any stationary point is a minimizer. and the second holds by complementary slackness. By weak duality, x^* and u^*, v^* are optimal. It always implies that the duality gap is 0. ■

Theorem 11.2 *For a problem with strong duality (e.g. assume Slater's condition: convex problem and there exists x strictly satisfying nonaffine inequality constraints),*

$$x^* \text{ and } u^*, v^* \text{ are primal and dual solutions} \iff x^* \text{ and } u^*, v^* \text{ satisfy the KKT conditions}$$

Proof: Sufficiency: Follows from Theorem 11.1.

Necessity: Let x^* and u^*, v^* be primal and dual solutions, and suppose we know strong duality holds.Then

$$\begin{aligned} f(x^*) &= g(u^*, v^*) \\ &= \min_x \Big(f(x) + \sum_{i=1}^m u_i^* h_i(x) + \sum_{j=1}^r v_j^* \ell_j(x) \Big) \\ &\leq f(x^*) + \sum_{i=1}^m u_i^* h_i(x^*) + \sum_{j=1}^r v_j^* \ell_j(x^*) \\ &\leq f(x^*) \end{aligned}$$

The LSH equals RHS, so all inequalities in the equation must be equalities. Looking at the KKT conditions one by one, primal and dual feasibility holds, by virtue of optimality. Stationarity comes from the fact that x^* minimizes $f(x) + \sum_{i=1}^m u_i^* h_i(x) + \sum_{j=1}^r v_j^* \ell_j(x)$. Since x^* is the minimizer, it must be a stationary point for this function. Complementary slackness comes from $f(x^*) + \sum_{i=1}^m u_i^* h_i(x^*) + \sum_{j=1}^r v_j^* \ell_j(x^*) = f(x^*)$, since we must have $\sum_{i=1}^m u_i^* h_i(x^*) = 0$ and they are each non-negative.

■

One thing to note: for a differentiable function f, we cannot use $\partial f(x) = \{\nabla f(x)\}$ unless f is convex when applying the stationarity conditions.

11.3 History of KKT

Previously known as the KT (Kuhn-Tucker) conditons:

- Appeared in publication by Kuhn and Tucker in 1951
- Later people discovered that Karush derived the conditions in his unpublished master's thesis of 1939

For **unconstrained problems**, the KKT conditions are nothing more than the subgradient optimality condition. For **general convex problems**, the KKT conditions could have been derived entirely from studying optimality via subgradients

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

where $N_C(x)$ is the normal cone of C at x

11.4 KKT Examples

This section steps through some examples in applying the KKT conditions.

11.4.1 Quadratic with Equality Constraints

Consider for $Q \geq 0$,

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

(For example, this corresponds to Newton step for the constrained problem $\min_x f(x)$ subject to $Ax = b$
Problem is convex, with no inequality constraints.
So by KKT conditions: x is a solution if and only if

$$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} -c \\ 0 \end{bmatrix}$$

for some u.

11.4.2 Water-Filling

Consider the following example from B&V p245:

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

The problem arises from information theory:
This problem arises from information theory, where each variable x_i 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.

KKT conditions give:

$$\begin{gathered} -\frac{1}{\alpha_i + x_i} - u_i + v = 0, \quad i = 1, \cdots, n \\ u_i \cdot x_i = 0, \quad i, \cdots, n \\ x \geq 0, \quad 1^T x = 1, \quad u \geq 0 \end{gathered}$$

Eliminate u:

$$-\frac{1}{\alpha_i + x_i} \leq v, \quad i = 1, \cdots, n$$

$$x_i \left(v - \frac{1}{\alpha_i + x_i} \right) = 0, \quad i, \cdots, n$$

$$x \geq 0, \quad 1^T x = 1$$

Then we can argue that stationarity and complementary slackness imply

$$x_i = \begin{cases} \frac{1}{v} - \alpha_i & \text{if } v < \frac{1}{-\alpha_i} \\ 0 & \text{if } v < \frac{1}{-\alpha_i} \end{cases} = \max\left\{0, \frac{1}{v} - \alpha_i\right\}, i = 1, \cdots, n$$

To guarantee feasibility of x so that $1^T x = 1$ holds, need

$$\sum_{i=1}^{n} \max\left\{0, \frac{1}{v} - \alpha_i\right\} = 1$$

Results in a univariate equation, which is piece-wise linear in $\frac{1}{v}$ and not hard to solve.
The problem is referred to as **water-filling**

11.4.3 Support Vector Machines

Given $y \in \{-1, 1\}^n$, and $X \in \mathbb{R}^{n \times p}$, the support vector machine problem is:

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C \sum_{i=1}^{N} \xi_i \\ \text{s.t.} \quad & \xi \geq 0, \quad i = 1, \cdots, n \\ & y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \quad i = 1, \cdots, n \end{aligned}$$

Introduce dual variables $v, w \geq 0$. KKT stationarity condition:

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

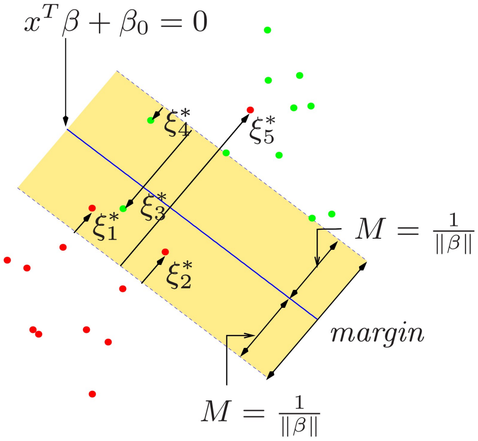
Complementary slackness:

$$v_i \xi_i = 0, \quad w_i(1 - \xi_i - y_i(x_i^T \beta + \beta_0)) = 0, \quad i = 1, \cdots, n$$

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

- if $\xi_i = 0$, then x_i lies on the edge of margin, and $w_i \in (0, C]$
- if $\xi_i \neq 0$, then x_i lies on the wrong side of the margin, and $w_i = C$.

KKT conditions is not a solution method, but gives a better understanding of the solution. In fact, we can apply KKT conditions to screen away non-support points before performing optimization.



11.5 Constrained and Lagrange Forms

Lemma 11.3 *For $t \in \mathbb{R}$ and $\lambda \geq 0$, the following forms are equivalent assuming convex f, h, and that the constrained form is strictly feasible:*

Constrained Form (C)	$\min_x f(x)$ *subject to* $h(x) \leq t$
Lagrange Form (L)	$\min_x f(x) + \lambda \cdot h(x)$

Proof: We will show that the constrained form and the lagrange form are equivalent given convex f, h, where (C) is strictly feasible.

(C) to (L) If (C) is strictly feasible ($\forall x$ s.t. $h(x) < t$), then strong duality holds. By the stationary condition, we have that

$$f(x^*) + \lambda \cdot (h(x^*) - t),$$

for x^* that minimizes the Lagrangian of (C) for some $\lambda \geq 0$. We can see that x^* also minimizes (L). Hence, x^* is also a solution in (L).

We have shown that

$$\bigcup_{\lambda \geq 0} \{\text{ solutions in } (L)\} \quad \supseteq \quad \bigcup_{t \in \{t \,:\, h(x) < t,\, \forall x\}} \{\text{solutions in } (C)\}$$

(L) to (C) If x^* is a solution in (L), then the KKT conditions for (C) are satisfied by taking $t = h(x^*)$.

$$\bigcup_{t \geq 0} \{\text{ solutions in } (C)\} \quad \supseteq \quad \bigcup_{\lambda \geq 0} \{\text{solutions in } (L)\}$$

Hence, we have shown a nearly perfect equivalence. ■

Note that when the only value of t that leads to a feasible but not strictly feasible constraint set is $t = 0$, then this is a perfect equivalence.

11.6 Uniqueness in ℓ_1 Penalized Problems

Theorem 11.4 *Let f be differentiable and* ***strictly convex****, let $X \in \mathbb{R}^{n \times p}$, $\lambda > 0$. Consider*

$$\min_{\beta} f(X\beta) + \lambda \|\beta\|_1.$$

If the entries of X are drawn from a continuous probability distribution (on $\mathbb{R}^{np}$), then with probability 1, there is a unique solution and it has at most $\min\{n, p\}$ *nonzero components.*

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[2] R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1970.

[3] Ryan J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Statist.*, 7:1456–1490, 2013.