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

Lecture 12: Oct 8

Lecturer: Lecturer: Ryan Tibshirani *Scribes: Scribes: Yuntian Deng, Zhiting Hu*

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.*

12.1 Recap on duality

Given a minimization problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \le 0, \;\; i = 1, \cdots, m \\ & l_j(x) = 0, \;\; j = 1, \cdots, r \end{aligned} \tag{12.1}$$

We define the Lagrangian:

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j l_j(x) \tag{12.2}$$

and Lagrange dual function:

$$g(u, v) = \min_x L(x, u, v) \tag{12.3}$$

The subsequent dual problem is

$$\begin{aligned} \max_{u,v} \quad & g(u, v) \\ \text{subject to} \quad & u \ge 0 \end{aligned} \tag{12.4}$$

Recall the important properties of dual problems from previous lectures:

- The dual problem is always convex no matter if the primal problem is convex, i.e., g is always concave.
- The primal and dual optimal values, f^* and g^*, always satisfy weak duality: $f^* \ge g^*$.
- Slater's condition: for convex primal, if there is an x such that

$$h_1(x) < 0, \cdots, h_m(x) < 0 \text{ and } l_1(x) = 0, \cdots, l_r(x) = 0 \tag{12.5}$$

  then strong duality holds: $f^* = g^*$. Note that the condition can be further relaxed to strict inequalities over the nonaffine h_i, $i = 1, \cdots, m$.

12.2 Karush-Kuhn-Tucker conditions

Given general problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \le 0, \;\; , i = 1, \cdots, m \\ & l_j(x) = 0 \;, j = 1, \cdots, r \end{aligned} \tag{12.6}$$

The Karush-Kuhn-Tucker conditions or KKT conditions are:

- $0 \in \partial f(x) + \sum_{i=1}^m u_i \partial h_i(x) + \sum_{j=1}^r v_j \partial l_j(x)$ (stationarity)
- $u_i \cdot h_i(x) = 0$ for all i (complementary slackness)
- $h_i(x) \leq 0$, $l_j(x) = 0$ for all i, j (primal feasibility)
- $u_i \geq 0$ for all i

An important conclusion of KKT conditions is: KKT conditions are

- always sufficient
- necessary under strong duality

Under strong duality assumption, KKT conditions are both sufficient and necessary, as stated in **Theorem 12.1**:

Theorem 12.1 *For a problem with strong duality (e.g., assume Slater's condition: convex problem and there exists x strictly satisfying non-affine 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: We first prove necessity:

Let x^* and u^*, v^* be primal and dual solutions with zero duality gap (strong duality holds, e.g., under Slater's condition). Then

$$\begin{aligned} f(x^*) \quad &= g(u^*, v^*) \\ &= \min_x f(x) + \textstyle\sum_{i=1}^m u_i^* h_i(x) + \sum_{j=1}^r v_j^* l_j(x) \\ &\leq f(x^*) + \textstyle\sum_{i=1}^m u_i^* h_i(x^*) + \sum_{j=1}^r v_j^* l_j(x^*) \\ &\leq f(x^*) \end{aligned} \tag{12.7}$$

Therefore, all these inequalities are actually equalities. Hence we have

- The point x^* minimizes $L(x, u^*, v^*)$ over $x \in \mathbb{R}^n$, i.e.,

  $$0 \in \partial_x L(x^*, u^*, v^*) \tag{12.8}$$
  $$0 \in \partial f(x^*) + \sum_{i=1}^m u_i^* \partial h_i(x^*) + \sum_{j=1}^r v_j^* \partial l_j(x^*) \tag{12.9}$$

  This is stationarity.
- $\sum_{i=1}^m u_i^* h_i(x^*) = 0$, since each term here is ≤ 0, this implies $u_i^* h_i(x^*) = 0$ for every i. This is exactly comlementary slackness.
- Primal and dual feasibility hold by virtue of optimality.

Now that we have proved necessity, we move on to prove sufficiency. If there exists x^*, u^*, v^* that satisfy the KKT conditions, then

$$\begin{aligned} g(u^*, v^*) \quad &= f(x^*) + \textstyle\sum_{i=1}^m u_i^* h_i(x^*) + \sum_{j=1}^r v_j^* l_j(x^*) \text{ (stationarity)} \\ &= f(x^*) \text{ (complementary slackness)} \end{aligned} \tag{12.10}$$

Therefore the duality gap is zero (and x^* and u^*, v^* are primal and dual feasible) so x^* and u^*, v^* are primal and dual optimal. Hence, we have shown the sufficiency. ■

Note that concerning the stationary condition, for a differentiable function f, we cannot use $\partial f(x) = \{\nabla f(x)\}$ unless f is convex.

For unconstrained problems, the KKT conditions are just the subgradient optimality condition.

For general problems, the KKT conditions could have been derived entirely from studying optimalit via subgradients

$$0 \in \partial f(x^*) + \sum_{i=1}^{m} \mathcal{N}_{\{h_i \le 0\}}(x^*) + \sum_{j=1}^{r} \mathcal{N}_{\{l_j = 0\}}(x^*) \tag{12.11}$$

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

12.3 Examples

12.3.1 Example: quadratic with equality constraints

Consider for $Q \succeq 0$,

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & \frac{1}{2} x^T Q x + c^T x \\ \text{subject to} & Ax = 0 \end{array} \tag{12.12}$$

E.g., as we will see, this corresponds to Newton step for equality-constrainted problem $\min_x f(x)$ subject to $Ax = b$.

This problem is a convex problem with no inequality constraints, so by KKT conditions, x is a solution if and only if

$$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} -c \\ 0 \end{bmatrix} \tag{12.13}$$

for some u. Eq. (12.13) is a linear system combining stationarity and primal feasibility. Complementary slackness and dual feasibility are vacuous in this case.

12.3.2 Example: water-filling

Consider the problem

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & -\sum_{i=1}^{n} \log(\alpha_i + x_i) \\ \text{subject to} & x \ge 0,\ 1^T x = 1 \end{array} \tag{12.14}$$

This problem arises from information theory, where for each channel i, x_i represents the allocated transmitter power and $\log(\alpha_i + x_i)$ is the communication rate. The problem is to maximize the total communication rate under a budget of total power one.

The KKT conditions are:

$$-1/(\alpha_i + x_i) - u_i + v = 0,\ i = 1, \cdots, n \tag{12.15}$$

$$u_i \cdot x_i = 0,\ i = 1, \cdots, n,\ x \ge 0,\ 1^T x = 1,\ u \ge 0 \tag{12.16}$$

After eliminating u, we have:

$$1/(\alpha_i + x_i) \le v,\ i = 1, \cdots, n \tag{12.17}$$

$$x_i(v - 1/(\alpha_i + x_i)) = 0,\ i = 1, \cdots, n,\ x \ge 0,\ 1^T x = 1 \tag{12.18}$$

We argue that if $v \geq 1/\alpha_i$, then x_i must be 0; if $v < 1/\alpha_i$, then $x_i = 1/v - \alpha_i$. Using the primal feasibility $1^T x = 1$ we need to solve the following problem to get v:

$$\sum_{i=1}^{n} \max\{0, 1/v - \alpha_i\} = 1 \tag{12.19}$$

This is a univariate equation, piecewise linear in $1/v$ and not hard to solve. The reduce problem is called water-filling (Figure 12.1).

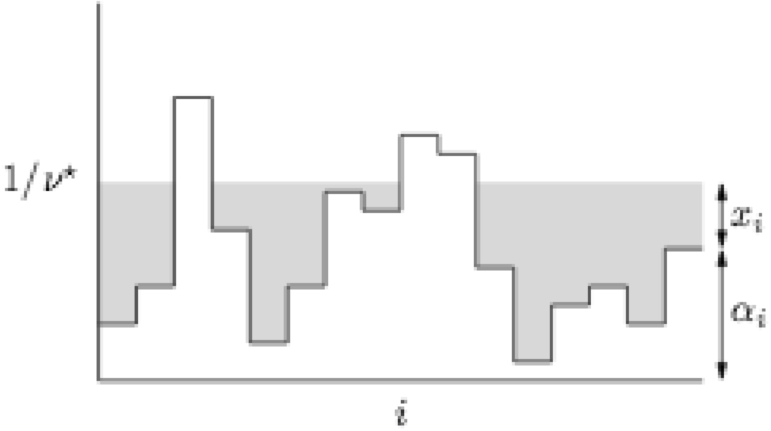

Figure 12.1: Water-filling illustration

The α_i can be thought as the ground level above patch i. If we flood the region with total amount of water 1, then the water depth $1/v$ satisfies $\sum_{i=1}^{n} \max\{0, 1/v - \alpha_i\} = 1$.

12.3.3 Example: support vector machines

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

$$\begin{array}{ll} \min\limits_{\beta, \beta_0, \xi} & \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} & \xi_i \geq 0,\ i = 1, \cdots, n \\ & y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i,\ i = 1, \cdots, n \end{array} \tag{12.20}$$

Introduce dual variables $v, w \geq 0$, from the KKT stationarity condition we have

$$0 = \sum_{i=1}^{n} w_i y_i,\ \beta = \sum_{i=1}^{n} w_i y_i x_i,\ w = C1 - v \tag{12.21}$$

From the complementary slackness we have

$$v_i \xi_i = 0,\ w_i(1 - \xi_i - y_i(x_i^T\beta + \beta_0)) = 0,\ i = 1, \cdots, n \tag{12.22}$$

Hence 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 i are calle the support points

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

Note that KKT conditions do not really give us a way to find solution, but gives a better understanding. In fact, we can use this to screen away non-support points before performing optimization.

12.4 Constrained and Lagrange forms

Often in statistics and machine learning we will switch back and forth between constrained form, where $t \in \mathbb{R}$ is a tuning parameter,

$$\min_x f(x) \text{ subject to } h(x) \leq t \tag{C}$$

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

$$\min_x f(x) + \lambda \cdot h(x) \tag{L}$$

and claim these are equivalent. We will show that this is almost always true assuming convexity of f and h.

(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

$$f(x) + \lambda \cdot (h(x) - t) \tag{12.23}$$

so x^* is also 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:

$$\begin{array}{ccc} \bigcup_{\lambda \geq 0} \{\text{solutions in (L)}\} & \subseteq & \bigcup_{t} \{\text{solutions in (C)}\} \\ \bigcup_{\lambda \geq 0} \{\text{solutions in (L)}\} & \supseteq & \bigcup_{\substack{t such that (C) \\ is strictly feasible}} \{\text{solutions in (C)}\} \end{array} \tag{12.24}$$

This is nearly a perfect equivalence. Note: when the only value of t that leads to a feasible but not strictly feasible constraint set is $t = 0$, i.e.,

$$\{x : h(x) \leq t\} \neq \phi, \ \{x : h(x) < t\} \neq \phi \ \Rightarrow t = 0 \tag{12.25}$$

(e.g., this is true if h is a norm) then we do get perfect equivalence.

12.5 Uniqueness in l_1 penalized problems

Using the KKT conditions and simple probability arguments, we have the following (perhaps surprising) result:

Theorem 12.2 *Let f be differentiable and strictly convex, let $X \in \mathbb{R}^{n \times p}$, $\lambda > 0$. Consider*

$$\min_{\beta \in \mathbb{R}^p} f(X\beta) + \lambda \|\beta\|_1 \tag{12.26}$$

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.

Proof: The KKT conditions are

$$s_i \in \begin{cases} \{\text{sign}(\beta_i)\} & \text{if} \beta_i \neq 0 \\ \text{[-1,1]} & \text{if} \beta_i = 0 \end{cases}, i = 1, \cdots, n \tag{12.27}$$

Note that $X\beta$, s are unique. Define $S = \{j : |X_j^T \nabla f(X\beta)| = \lambda\}$, then S is also unique as both X and $X\beta$ are unique. Note that any solution satisfies $\beta_i = 0$ for all $i \notin S$.

First assume that $\text{rank}(X_S) < |S|$ (here $X \in \mathbb{R}^{n \times |S|}$ is a submatrix of X corresponding to columns in S). Then for some $i \in S$,

$$X_i = \sum_{j \in S\{i\}} c_j X_j \tag{12.28}$$

for constants $c_j \in \mathbb{R}$, hence

$$s_i X_i = \sum_{j \in S\{i\}} (s_i s_j c_j) \cdot (s_j X_j) \tag{12.29}$$

Taking an inner product with $-\nabla f(X\beta)$,

$$\lambda = \sum_{j \in S\{i\}} (s_i s_j c_j) \lambda, \text{ i.e., } \sum_{j \in S\{i\}} s_i s_j c_j = 1 \tag{12.30}$$

In other words, we have proved that $\text{rank}(X_S) < |S|$ implies $s_i X_i$ is the affine span of $s_j X_j$, $j \in S$ $\{i\}$ (subspace of dimension ¡ n)

We say that the matrix X has columns in general position if any affine subspace L of dimension $k < n$ does not contain more than $k + 1$ elements of $\{\pm X_1, \cdots, \pm X_p\}$ (excluding antipodal pairs)

It is straightforward to show that, if the entries of X have a density over $\mathbb{R}^{np}$, then X is in general position with probability 1.

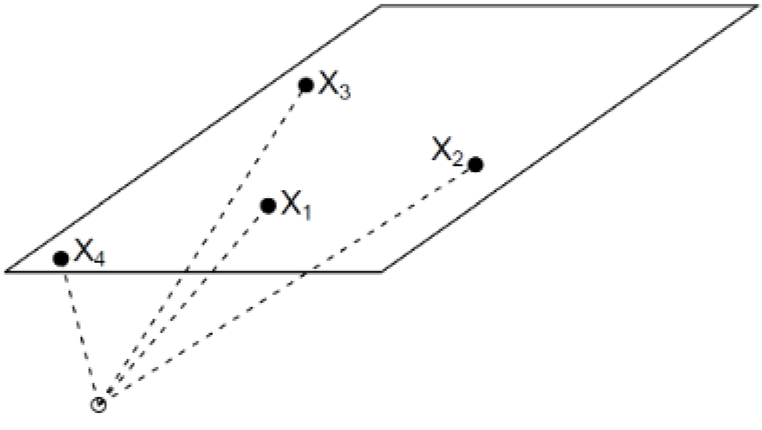

Figure 12.2: General position

Therefore, if entries of X are drawn from continuous probability distribution, any solution must satisfy $\text{rank}(X_S) = |S|$.

Recalling the KKT conditions, this means the number of nonzero components in any solution at most $\leq |S| \leq \min\{n, p\}$. Further, we can reduce our optimization problem (by partially solving) to

$$\min_{\beta_S \in \mathbb{R}^{|S|}} f(X_S \beta_S) + \lambda \|\beta_S\|_1 \tag{12.31}$$

Finally, strict convexity implies uniqueness of the solution in this problem, and hence in our original problem. ■

12.6 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

$$0 \in \partial f(x^*) + \sum_{i=1}^{m} u_i^* \partial h_i(x^*) + \sum_{j=1}^{r} v) j^* \partial l_j(x^*) \tag{12.32}$$

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