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

Lecture 12: KKT Conditions

Lecturer: Ryan Tibshirani *Scribes: Fei Xia, Hao Zhang(haoz1), Jingwei Zhuo*

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

12.1 Recap on duality

For a minimization problem:

$$\begin{array}{ll} \min_x & f(x) \\ \text{subject to} & h_i(x) \leq 0, i = 1, ...m \\ & \ell_j(x) = 0, j = 1, ...r \end{array}$$

The Lagrangian is defined as:

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

The Lagrange dual function is:

$$g(u, v) = \min_x L(x, u, v)$$

The corresponding dual problem is:

$$\begin{array}{ll} \max_{u,v} & g(u, v) \\ \text{subject to} & u \geq 0 \end{array}$$

The Lagrange dual function can be viewd as a pointwise maximization of some affine functions so it is always concave. The dual problem is always convex even if the primal problem is not convex.

For any primal problem and dual problem, the weak duality always holds:

$$f^* \geq g^*$$

When the Slater's conditioin is satisfied, we have strong duality so $f^* = g^*$.

The dual problem sometime can be easier to solve compared with the primal problem and the primal solution can be constructed from the dual solution.

12.2 Karush-Kuhn-Tucker conditions

Given general problem

$$\begin{array}{ll} \min & f(x) \\ \text{subject to} & h_i(x) \leq 0, i = 1, ...m \\ & \ell_j(x) = 0, j = 1, ...r \end{array}$$

The Karush-Kuhn-Tucker conditions (KKT conditions) are:

- Stationarity: $0 \in \partial f(x) + \sum_{i=1}^m u_i \partial h_i(x) + \sum_{j=1}^r v_j \partial \ell_j(x)$
- Complementary: $u_i h_i(x) = 0$ for all i
- Primal feasibility: $h_i(x) \leq 0$, $\ell_j(x) = 0$ for all i, j
- Dual feasibility: $u_i \geq 0$ for all i

Warning: Concerning the stationarity condition: for a differentiable function f, we cannot use $\partial f(x) = \{\nabla f(x)\}$ unless f is convex.

Theorem 12.1 *For a problem with strong duality (e.g., assume Slaters condition: convex problem and there exists x strictly satisfying non-affine inequality contraints), x^* and u^*, v^* satisfy the KKT conditions if and only if x^* and u^*, v^* are primal and dual solutions.*

Proof: We first prove the necessity: Let x^* and u^*, v^* be primal and dual solutions with zero duality gap (i.e. strong duality holds), then

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

Thus, all inequalities above are actually equalities, which means:

- x^* minimizes $L(x, u^*, v^*)$ over x, i.e.,

  $$\begin{aligned} &0 \in \partial_x L(x^*, u^*, v^*) \\ &0 \in \partial f(x^*) + \sum u_i^* \partial h_i(x^*) + \sum v_j^* \partial \ell_j(x^*) \end{aligned}$$

  which is the stationary condition.
- $\sum u_i^* h_i(x^*) = 0$, i.e.,

  $$u_i^* h_i(x) = 0 \quad \text{for all } i$$

  which is the complementary slackness condition.
- The primal and dual feasibility of (x^*, u^*, v^*) hold.

Then we prove the sufficiency. If there exists x^*, u^*, v^* that satisfy the KKT conditions, then

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

where the first equality holds from stationarity, and the second holds from complementary slackness. Therefore, the duality gap is zero, so x^* and u^*, v^* are primal and dual optimal respectively. ■

It should be noticed that for unconstrained problems, KKT conditions are just the subgradient optimality condition.

For general problems, the KKT conditions can be derived entirely from studying optimality via subgradients:

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

12.3 Example

12.3.1 Quadratic with equality constraints

For any $Q \succeq 0$, the quadratic problem is defined as:

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

This is a convex problem only with equality constraints, so according to KKT conditions, x is a solution if and only if:

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

for some u. This linear system contains stationarity and primal feasibility. Because there is no inequality constrains the complementary slackness and dual feasibility are vacuous.

12.3.2 Water-filling

Consider the following optimization problem:

$$\begin{array}{ll} \min_{x \in R^n} & -\sum_{i=1}^{n} log(\alpha_i + x_i) \\ \text{subject to} & x \geq 0, 1^T x = 1 \end{array}$$

This problem arises from information theroy, where each variable x_i represents the transmitter power allocated to the i-th channel and $\log(\alpha_i + x_i)$ gives the apacity 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 Lagrangian is:

$$L(x, u, v) = -\sum_{i=1}^{n} log(\alpha_i + x_i) - \sum_{i=1}^{n} u_i x_i + v(\sum_{i=1}^{n} x_i - 1)$$

The stationarity is:

$$-\frac{1}{\alpha_i + x_i} - u_i + v = 0$$

The complementary slackness is:

$$u_i x_i = 0$$

The primal feasibility is:

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

The dual feasibility is:

$$u_i \geq 0$$

From the above result, we have:

$$v \geq \frac{1}{\alpha_i + x_i}$$

and

$$x_i(v - \frac{1}{\alpha_i + x_i}) = 0$$

We argue taht if $v \geq \frac{1}{\alpha_i}$, then x_i must be 0; if $v \leq \frac{1}{\alpha_i}$, then $v = \frac{1}{\alpha_i + x_i}$, we can solve x_i from above. Combine the primal feasibility $1^T x = 1$ we have the following problem:

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

This is a univariate equation and easy to solve. This reduced problem is called water-filling. Here the α_i can be thought as the ground level above patch i, and then we flood the region with water to a depth $\frac{1}{v}$. The total amount of water used is then $\sum_i^n max\{0, \frac{1}{v} - \alpha_i\}$. We can increase the flood level until we have used a total amount of water equal to one.

12.3.3 Support vector machines

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

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

There are no equality constraints, so we can introduce dual variables $v, w \geq 0$. From the KKT stationarity condition:

$$0 \in \partial \left\{ \frac{1}{2}||\beta||_2^2 + C\sum_{i=1}^{n} \xi_i \right\}$$

Note that the objective function is differentiable, so we have

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

The complementary slackness condition implies

$$v_i \xi_i = 0, \quad w_i \left(1 - \xi_i - y_i(x_i^{\mathrm{T}}\beta + \beta_0)\right) = 0, \quad i = 1, ..., n$$

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

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, as shown in figure.12.1, 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.

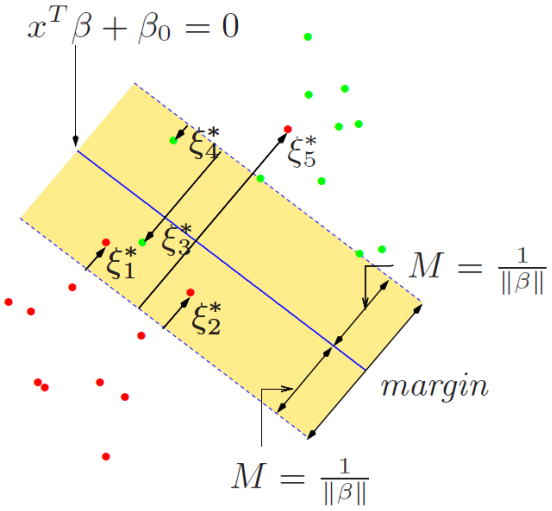

Figure 12.1: The decision boundary of SVM

12.3.4 Uniqueness in ℓ_1 penalized problems

Using the KKT coditions and simple probability arguments, we have the following result:

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

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

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

Proof: the KKT conditions are

$$-X^{\mathrm{T}}\nabla f(X\beta) = \lambda s,\ s_i \in \begin{cases} \{\mathrm{sign}(\beta_i)\}, & \beta_i \neq 0 \\ [-1, 1], & \beta_i = 0 \end{cases},\ i = 1, ..., n$$

Recall that f is strictly convex, $X\beta, s$ are unique. Define $S = \{j : |X_j^{\mathrm{T}}\nabla f(X\beta)| = \lambda\}$, which is also unique, and note that any solution satisfies $\beta_i = 0$ for all $i \notin S$.

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

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

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

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

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

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

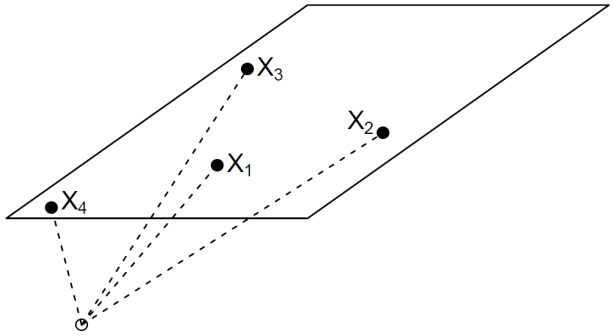

Figure 12.2: Uniqueness in ℓ_1 penalized problems

In other words, we have proved that $\text{rank}(X_s) < |S|$ implies $s_i X_i$ is in the affine span of $s_j X_j, j \in S \setminus \{i\}$ (subspace of dimension $< n$) as shown in figure.12.2.

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

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

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

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

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

12.4 Relation between constrained and Lagrange forms

Often in statistics and machine learning we'll switch back and forth between constrained form, where $t \in \mathbb{R}$ is a tuning parameter,

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

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

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

and claim these are quivalent. 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 Slater's condition implies strong duality holds, and there exists some $\lambda \geq 0$ (dual solution) such that any solution x^* in (C) minimizes

$$f(x) = \lambda(h(x) - t)$$

so x^* is also a solution in (L). So we can get the relationship:

$$\bigcup_{\lambda \geq 0} \{\text{solutions in (L)}\} \supseteq \bigcup_{\substack{t \text{ such that (C)} \\ \text{is strictly feasible}}} \{\text{solutions in (C)}\}$$

(L) to (C): if x^* is a solution in (L), then the KKT conditions for (C) are satisfied by taking $t = h(x^*)$, so x^* is a solution in (C). So we can get the relationship:

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

This is nearly a perfect equivalence-the only exception is the case (C) is not strictly feasible for some t. If we introduce an extra condition that the only value of t that leads to a feasible but not strictly feasible constraint set is $t = 0$, i.e.,

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

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

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

12.5 From dual solutions to primal solutions

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

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

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

- Generally, this reveals a characterization of primal solutions.
- In particular, if above problem has a unique minimizer, then the corresponding point must be the primal solution.

Reference

- S. Boyd and L. Vandenberghe (2004), "Convex Optimization", Chapter 5
- R. Tibshirani (2015), "Karush-Kuhn-Tucker Conditions"