10-725/36-725: Convex Optimization **Fall 2018**

Lecture 12: October 8

Lecturer: Lecturer: Ryan Tibshirani *Scribes: Cheng Cheng, Amanda Coston, Landon Settle*

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

This lecture's notes illustrate some uses of various LaTeX macros. Take a look at this and imitate.

Karush-Kuhn-Tucker (KKT) Conditions

For the optimization problem

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

the KKT conditions are

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

The KKT conditions are always sufficient for optimality. The KKT conditions are necessary for optimality if strong duality holds. We often use Slater's condition to prove that strong duality holds (and thus KKT conditions are necessary). Slater's condition implies that strong duality holds for a convex primal with all affine constraints . For a convex problem with nonaffine constraints, if there is a feasible x satisfying the strict inequalities, then strong duality holds.

We can prove that KKT conditions are necessary under strong duality denoting x^*, u^*, v^* as the primal and dual solutions. Since strong duality holds, $f(x^*) = g(u^*, v^*)$. By definition of Lagrangian dual $g(u^*, v^*) = \min_x f(x) + \sum_{i=1}^{m} u_i^* h_i(x) + \sum_{j=1}^{r} v_j^* l_j(x)$. By definition of min function this is
$\leq f(x^*) + \sum_{i=1}^{m} u_i^* h_i(x^*) + \sum_{j=1}^{r} v_j^* l_j(x^*) \leq f(x^*)$ where the last inequality holds by feasibility.

So all these inequalities must be strict equalities. Then we can see that $\sum_{i=1}^{m} u_i^* h_i(x) = 0$ and since each term of this sum is negative (by feasibility), then this requires complementary slackness. We can also see that x^*

minimizes the lagrangian so $0 \in \partial_x(L(x, u^*, v^*))$ at $x = x^*$, which is the stationarity condition. Primal and dual feasibility hold because they are required for optimality.

We can prove KKT conditions are sufficient for optimality. By stationarity, we know that x^* minimizes the Lagrangian so $g(u^*, v^*) = f(x^*) + \sum_{i=1}^{m} u_i^* h_i(x^*) + \sum_{j=1}^{r} v_j^* l_j(x^*) = f(x^*)$ where the last inequuality holds from complementary slackness.

Note that if the problem is unconstrained, then the KKT conditions are subgradient optimality condition $0 \in \partial(f(x))$. We can see KKT as a generalization of this for constrained problems.

Examples

Quadratic with equality constraints

Consider $Q \succeq 0$

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

By KKT conditions:

- stationarity: $\nabla L(x, u) = 0 => Qx + c + A^T u = 0$
- complementary slackness: $\emptyset$
- feasibility: $Ax = 0$

Combining them gives:

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

Water-filling

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

By KKT conditions:

- stationarity: $\nabla L(x, u, v) = 0 => -\frac{1}{\alpha_i + x_i - u_i + v}, i = 1, ..., n$
- complementary slackness: $u_i x_i = 0, i = 1, ..., n$
- feasibility: $x \geq 0, 1^T x = 1, u \geq 0$

These conditions lead to an algorithm for computing $x^\star$

Support vector machines

$$\begin{aligned} \min \quad & \frac{1}{2}||\beta||_2^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & \xi_i \geq 0 \\ & y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \end{aligned}$$

By KKT conditions:

- stationarity:
  $L(\beta, \beta_0, \xi, v, w) = \frac{1}{2}||\beta||_2^2 + C\sum \xi_i - \sum v_i\xi_i + \sum w_i(1 - \xi_i - y_i(x_i^T\beta + \beta_0))$
  $\nabla_\beta L = 0 => \beta - \sum w_i y_i x_i$
  $\beta = \sum w_i y_i x_i$
  $\nabla_{\beta_0} L = 0 => \sum w_i y_i = 0$
  $\nabla_\xi L = 0 => C1 - v - w = 0$

The results given by stationarity condition shows that at optimality we have $\beta = \sum w_i y_i x_i$. w_i is none zero only if $y_i(x_i^T\beta + \beta_0) = 1 - \xi_i$. Such points give the support vectors.

Constrained and Lagrange forms

$$\begin{aligned} &\text{constrained form:} && \min_x f(x) \text{ subject to } h(x) \leq t \\ &\text{Lagrange form:} && \min_x f(x) + \lambda h(x) \end{aligned}$$

We want to prove them to be equivalent.
C to L: $L(x, \lambda) = f(x) + \lambda(h(x) - t)$
By KKT condition (stationarity), $X^\star$ must solve

$$\min_x f(x) + \lambda(h(x) - t) <=> \min_x f(x) + \lambda(h(x)$$

L to C: By KKT condition (stationarity), $X^\star$ must solve

$$\min_x f(x) + \lambda(h(x) - t)$$

By complementary slackless: $\lambda(h(x^\star) - t) = -0$ which is satisfied by $t = h(x^\star)$
By feasibility: $h(x^\star) \leq t, \lambda \geq 0$
Therefore, KKT condition for C is satisfied by taking $t = h(x^\star)$, so $x^\star$ is a solution to C

Uniqueness in ℓ_1 penalized problems

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

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

  If the entries of X are drawn form 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

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

Notice that strict convexity of f implies the solution β^* is unique, and so the KKT conditions imply that $s \in \mathbb{R}^n$ is unique. Therefore the *equicorrelation set*

$$S = \{j \, : \, \left|X_j^T \nabla f(X\beta)\right| = \lambda\} = \{j \, : \, \beta_j \neq 0\}$$

is also unique, as any solution satisfies $\beta_i = 0$ for all $i \notin S$. Let $X_S \in \mathbb{R}^{n \times |S|}$ be the submatrix of X containing only those columns with indices in S. It can be shown that $\text{rank}(X_S) < |S|$ implies $s_i X_i$ is in the affine span of $s_j X_j$ for $j \in S \setminus \{i\}$, i.e.

$$s_i X_i = \sum_{j \in S \setminus \{i\}} a_i (s_j X_j) \text{ for some } a_j \in \mathbb{R}, j \in S \setminus \{i\}.$$

But this cannot happen almost surely (i.e. with probability one) if entries of X are drawn from a continuous probability distribution, and so any solution must satisfy $\text{rank}(X_S) = |S|$. Lastly, the KKT conditions and the equicorrelation set show that $|S|$, the number of nonzero components in a solution, is at most $\min\{n, p\}$.

Furthermore, we can now reduce the problem to

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

(where β_S is defined similarly to X_S), and strict convexity implies uniqueness of the solution in this problem and hence in the original problem as well.

Back to Duality

Under strong duality, we can use duality to characterize primal solutions from dual solutions. Given dual solutions u^* and v^*, since the KKT conditions are neccesary for optimality under strong duality, we know any primal solution x^* must minimize $L(x, u^*, v^*)$, i.e.

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

In particular, note that if there is only one such x^* which minimizes $L(x, u^*, v^*)$ then this must be the primal solution.