Karush-Kuhn-Tucker Conditions

Ryan Tibshirani
Convex Optimization 10-725

Last time: duality

Given a minimization problem

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

we defined the Lagrangian:

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

and Lagrange dual function:

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

The subsequent dual problem is:

$$\max_{u,v} \quad g(u,v)$$
$$\text{subject to} \quad u \geq 0$$

Important properties:

- Dual problem is always convex, i.e., g is always concave (even if primal problem is not convex)
- The primal and dual optimal values, $f^\star$ and $g^\star$, always satisfy weak duality: $f^\star \geq g^\star$
- Slater's condition: for convex primal, if there is an x such that

  $$h_1(x) < 0, \ldots, h_m(x) < 0 \quad \text{and} \quad \ell_1(x) = 0, \ldots, \ell_r(x) = 0$$

  then strong duality holds: $f^\star = g^\star$. Can be further refined to strict inequalities over the nonaffine h_i, $i = 1, \ldots m$

Outline

Today:

- KKT conditions
- Examples
- Constrained and Lagrange forms
- Uniqueness with ℓ_1 penalties

Karush-Kuhn-Tucker conditions

Given general problem

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

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

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

Necessity

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

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

In other words, all these inequalities are actually equalities

Two things to learn from this:

- The point $x^\star$ minimizes $L(x, u^\star, v^\star)$ over $x \in \mathbb{R}^n$. Hence the subdifferential of $L(x, u^\star, v^\star)$ must contain 0 at $x = x^\star$—this is exactly the stationarity condition
- We must have $\sum_{i=1}^m u_i^\star h_i(x^\star) = 0$, and since each term here is ≤ 0, this implies $u_i^\star h_i(x^\star) = 0$ for every i—this is exactly complementary slackness

Primal and dual feasibility hold by virtue of optimality. Therefore:

> If $x^\star$ and $u^\star, v^\star$ are primal and dual solutions, with zero duality gap, then $x^\star, u^\star, v^\star$ satisfy the KKT conditions

(Note that this statement assumes nothing a priori about convexity of our problem, i.e., of f, h_i, ℓ_j)

Sufficiency

If there exists $x^\star, u^\star, v^\star$ that satisfy the KKT conditions, then

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

where the first equality holds from stationarity, and the second holds from complementary slackness

Therefore the duality gap is zero (and $x^\star$ and $u^\star, v^\star$ are primal and dual feasible) so $x^\star$ and $u^\star, v^\star$ are primal and dual optimal. Hence, we've shown:

> If $x^\star$ and $u^\star, v^\star$ satisfy the KKT conditions, then $x^\star$ and $u^\star, v^\star$ are primal and dual solutions

Putting it together

In summary, KKT conditions are equivalent to zero duality gap:

- always sufficient
- necessary under strong duality

Putting it together:

For a problem with strong duality (e.g., assume Slater's condition: convex problem and there exists x strictly satisfying non-affine inequality contraints),

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

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

What's in a name?

Older folks will know these as the KT (Kuhn-Tucker) conditions:

- First appeared in publication by Kuhn and Tucker in 1951
- Later people found out that Karush had 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

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

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

Example: quadratic with equality constraints

Consider for $Q \succeq 0$,

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

(For example, this corresponds to Newton step for the constrained problem $\min_x \; f(x) \;\text{ subject to }\; Ax = b$)

Convex problem, 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. Linear system combines stationarity, primal feasibility (complementary slackness and dual feasibility are vacuous)

Example: water-filling

Example from B & V page 245: consider problem

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

Information theory: think of $\log(\alpha_i + x_i)$ as communication rate of ith channel. KKT conditions:

$$-1/(\alpha_i + x_i) - u_i + v = 0, \quad i = 1, \ldots, n$$
$$u_i \cdot x_i = 0, \quad i = 1, \ldots, n, \quad x \geq 0, \quad 1^T x = 1, \quad u \geq 0$$

Eliminate u:

$$1/(\alpha_i + x_i) \leq v, \quad i = 1, \ldots, n$$
$$x_i(v - 1/(\alpha_i + x_i)) = 0, \quad i = 1, \ldots, n, \quad x \geq 0, \quad 1^T x = 1$$

Can argue directly stationarity and complementary slackness imply

$$x_i = \begin{cases} 1/v - \alpha_i & \text{if } v < 1/\alpha_i \\ 0 & \text{if } v \geq 1/\alpha_i \end{cases} = \max\{0, 1/v - \alpha_i\}, \quad i = 1, \ldots, n$$

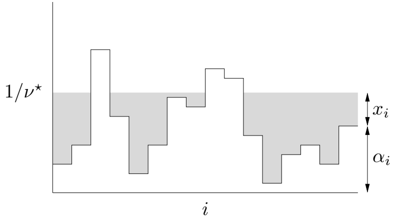
Still need x to be feasible, i.e., $1^T x = 1$, and this gives

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

Univariate equation, piecewise linear in $1/v$ and not hard to solve

This reduced problem is called water-filling

(From B & V page 246)



Example: support vector machines

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

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

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

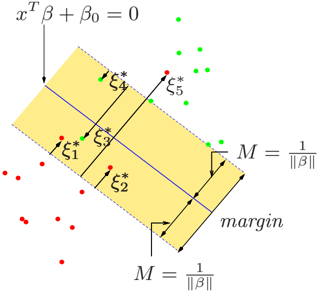
$$0 = \sum_{i=1}^n w_i y_i, \quad \beta = \sum_{i=1}^n w_i y_i x_i, \quad w = C1 - v$$

Complementary slackness:

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

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



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

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,

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

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

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

and claim these are equivalent. Is this true (assuming convex f, h)?

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

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

so $x^\star$ is also a solution in (L)

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

Conclusion:

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

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

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 $t = 0$, then we do get perfect equivalence

For example, if $h \geq 0$, and problems (C), (L) are feasible for $t \geq 0$, $\lambda \geq 0$, respectively, then we do get perfect equivalence

Uniqueness in ℓ_1 penalized problems

Using the KKT conditions and simple probability arguments, we have the following (perhaps surprising) result: [1]

Theorem: 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 w.p. 1 there is a unique solution and it has at most $\min\{n, p\}$ nonzero components

Remark: here f must be strictly convex, but no restrictions on the dimensions of X (we could have $p \gg n$)

[1] For example, Tibshirani (2013), "The lasso problem and uniqueness"

Proof: the KKT conditions are

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

Basic but important observations:

- $X\beta$ is unique by strict convexity of f
- The KKT conditions hence imply s is unique

Thus we can define equicorrelation set

$$S = \{j : |X_j^T \nabla f(X\beta)| = \lambda\}$$

This is also unique, 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 $c_j \in \mathbb{R}$, so that

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

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

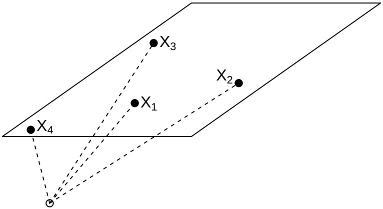
$$\lambda = \sum_{j \in S \setminus \{i\}} (s_i s_j c_j) \lambda, \quad \text{i.e.,} \quad \sum_{j \in S \setminus \{i\}} s_i s_j c_j = 1$$

In other words, we've proved that $\mathrm{rank}(X_S) < |S|$ implies

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

i.e., $s_i X_i$ is in the affine span of $s_j X_j$, $j \in S \setminus \{i\}$ (subspace of dimension $< n$)

It is easy to show that, if the entries of X have a density over $\mathbb{R}^{np}$, then almost surely, this cannot happen



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

Conclusions:

- Recalling the KKT conditions, we see the number of nonzero components in any solution at most $|S| \le \min\{n, p\}$
- 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

□

Back to duality

A key use of duality: under strong duality, can characterize primal solutions from dual solutions

Recall that under strong duality, the KKT conditions are necessary for optimality. Given dual solutions $u^\star, v^\star$, any primal solution $x^\star$ satisfies the stationarity condition

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

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

In particular, if this is satisfied uniquely (above problem has unique minimizer), then corresponding point must be the primal solution ... very useful when dual is easier to solve than primal

References

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