Karush-Kuhn-Tucker Conditions

Jianyong Sun

Xi'an Jiaotong University

jy.sun@xjtu.edu.cn

December 2, 2018

Duality gap

Given primal feasible x and dual feasible u, v, the quantity

$$f(x) - g(u, v)$$

is called the duality gap between x and u, v. Note that

$$f(x) - f^\star \leq f(x) - g(u, v)$$

so if the duality gap is zero, then x is primal optimal (and similarly u, v are dual optimal)

From an algorithmic viewpoint, provides a stopping criterion: if $f(x) - g(u, v) \leq \epsilon$, then we are guaranteed that $f(x) - f^\star \leq \epsilon$

Very useful, especially in conjunction with iterative methods.

Karush-Kuhn-Tucker conditions

Give general problem

$$\begin{aligned} \min \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, i = 1, 2, \cdots, m \\ & \ell_j(x) = 0, j = 1, 2, \cdots, r \end{aligned}$$

The Karush-Kuhn-Tucker conditions or KKT conditions are

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

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

In other words, all these inequalities are actually equalities.

Necessity

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 $\sum_i u_i^\star h_i(x^\star) = 0$, 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 $u^\star, v^\star$ be 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 prior about convexity of the 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

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

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 $x^\star, u^\star, v^\star$ are primal and dual optimal. Here we've shown

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

Putting it together

In summary KKT conditions are

- 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 constraints),

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

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

For unconstrained problem, the KKT conditions are nothing more than the subgradient optimality condition

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

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

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

Quadratic with equality constraints

Consider for $Q \succeq 0$,

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

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: support vector machine

Given $y \in \{-1, 1\}^n$, $X \in \mathbb{R}^{n \times p}$, rows $x_1, \cdots, x_n$, recall the support vector machine problem

$$\min_{\beta, \beta_0, \xi} \quad \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i$$

$$\text{subject to} \quad \xi_i \geq 0, i = 1, \cdots, n$$
$$y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, i = 1, \cdots, n$$

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

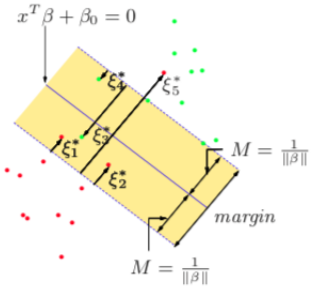
$$0 = \sum_{i=1}^n w_i y_i, \qquad \beta = \sum_{i=1}^n w_i y_i x_i, \qquad w = C\mathbf{1} - v$$

Complementary slackness

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

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 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

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

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

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

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

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

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

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

Constrained and Lagrange forms

(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:

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

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

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 $t = 0$, i.e.

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

(e.g. this is true if h is a norm), 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:

Theorem

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 $\hat{\beta} \in \mathbb{R}^p$ 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 $p \gg n$).

Solving the primal via the dual

One of the most important use 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 $u^\star, v^\star$, any primal solution $x^\star$ satisfies the stationarity condition

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

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

- 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

Example

Consider

$$\min_x \sum_{i=1}^{n} f(x) \text{ subject to } a^T x = b$$

where each $f_i : \mathbb{R} \to \mathbb{R}$ is smooth, strictly convex. Dual function

$$\begin{aligned} g(v) &= \min_x \sum_{i=1}^{n} f_i(x_i) + v(b - a^T x) \\ &= bv + \sum_{i=1}^{n} \min_{x_i \in \mathbb{R}} (f_i(x_i) - a_i v x_i) \\ &= bv - \sum_{i=1}^{n} f_i^*(a_i v) \end{aligned}$$

where f_i^* is the conjugate of f_i, to be defined shortly

Example

Therefore, the dual problem is

$$\max_{v \in \mathbb{R}} bv - \sum_{i=1}^{n} f_i^*(a_i v)$$

or equivalently

$$\min_{v \in \mathbb{R}} \sum_{i=1}^{n} f_i^*(a_i v) - bv$$

This is a convex minimization problem with scalar variable—much easier to solve than primal

Given $v^\star$, the primal solution $x^\star$ solves

$$\min_{x} \sum_{i=1}^{n} f_i(x_i) - a_i v^\star x_i$$

Strict convexity of each f_i implies that this has a unique solution, namely $x^\star$, which we compute by solving $\nabla f_i(x_i) = a_i v^\star$ for each i

Back to SVM

The SVM:

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

The Lagrangian

$$\mathcal{L}(\beta,\beta_0,\xi,w,v) = \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n w_i\left[y_i(x_i^T\beta+\beta_0)-1+\xi_i\right] - \sum_{i=1}^n v_i\xi$$

The dual problem

$$\max_w \quad \mathcal{D}(w) = \sum_{i=1}^n w_i - \frac{1}{2}\sum_{ij=1}^n y_i y_j w_i w_j \langle x_i, x_j\rangle$$
$$\text{s.t.} \quad 0 \leq w_i \leq C, i=1,\cdots,n \qquad \sum_{i=1}^n w_i y_i = 0$$

The primal solution:

$$\beta^\star = \sum_{i=1}^n w_i^\star y_i x_i, \beta_0^\star = \frac{\max_{i:y_i=-1}(w^\star)^T x_i + \min_{i:y_i=1}(w^\star)^T x_i}{2}$$

Summary

For the problem

$$\begin{aligned} \min \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, i = 1, 2, \cdots, m \\ & \ell_j(x) = 0, j = 1, 2, \cdots, r \end{aligned}$$

The KKT conditions are

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

These are necessary for optimality (of a primal-dual pair $x^\star$ and $u^\star, v^\star$ under strong duality, and always sufficient

Summary

Two key uses of duality

- For x primal feasible, and u, v dual feasible

$$f(x) - g(u, v)$$

  is called the duality gap between x and u, v, since

$$f(x) - f(x^\star) \leq f(x) - g(u, v)$$

  a zero duality gap implies optimality. Also, the duality gap can be used as a stopping criterion in algorithms
- Under strong duality, given dual optimal $u^\star, v^\star$, any primal solution minimizes $L(x, u^\star, v^\star)$ over all x (i.e. it satisfies stationarity condition). This can be used to characterize or compute primal solutions.

Summary

An important consequence of stationarity: under strong duality, given a dual solution $u^\star, v^\star$, any primal solution $x^\star$ solves

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

Often, solutions of this unconstrained problem can be expressed explicitly, giving an explicit characterization of primal solutions from dual solutions.

Furthermore, suppose the solution of this problem is unique; then it must be the primal solution $x^\star$

This can be very helpful when the dual is easier to solve than the primal.

Questions?