Karush-Kuhn-Tucker Conditions

1 Optimization Problem

We consider the following form of constrained optimization problem

$$\begin{aligned} &\min_x f(x) \\ &\text{subject to } g_j(x) \le 0 \quad (j = 1, \dots, k) \\ &\text{and } h_j(x) = 0 \quad (j = 1, \dots, m), \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^d$ is a parameter to be optimized. Each $g_j(x)$ and $h_j(x)$ is a real-valued function. Therefore $g_j(x) \le 0$ defines k inequality constraints, and $h_j(x) = 0$ defines m equality constraints. The condition of $h_j(x) = 0$ can also be written as $h_j(x) \le 0$ and $-h_j(x) \le 0$, which becomes $2m$ inequality constraints. However, it is convenient to explicitly include equality constraint.

We assume $g_j(x)$ and $h_j(x)$ are differentiable. We will first look at the general (not necessarily convex) situation, and in such case, we assume that $f(x)$ is differentiable.

Given a point $x \in \mathbb{R}^d$, we say x is feasible if it satisfies the constraints of (1). The active inequality constraints are the inequality constraints such that $g_j(x) = 0$.

2 General KKT conditions

In order to solve (1), we can form the Lagrangian function

$$L(x, \mu, \lambda) = f(x) + \mu^\top g(x) + \lambda^\top h(x),$$

where $\mu \in \mathbb{R}^k$ and $\lambda \in \mathbb{R}^m$. Here $g(x) = [g_1(x), \dots, g_k(x)]$ and $h(x) = [h_1(x), \dots, h_m(x)]$.

The variable x is called primal variable, and μ and λ are dual variables. The dual variables μ for the inequality constraints should satisfy the non-negativity constraints $\mu \ge 0$, which is required by the KKT conditions.

The KKT conditions are necessary conditions for a local optimal solution of (1). It can be stated as follows.

Theorem 1 *Assume that $f(x)$, $g(x)$, $h(x)$ are continuously differentiable. Assume x_* is a local optimal solution of* (1)*. If the gradients of the active inequality constraints and the gradients of the equality constraints are linearly independent at x_*, then the following KKT conditions hold.*

- *Stationarity*

$$\nabla_x L(x_*, \mu, \lambda) = 0.$$

- *Primal Feasibility:*

$$g(x_*) \le 0, \qquad h(x_*) = 0.$$

- *Dual Feasibility:*

$$\mu_j \ge 0 \qquad \forall j = 1, \ldots, k.$$

- *Complementary Slackness:*

$$\mu_j g_j(x_*) = 0 \qquad \forall j = 1, \ldots, k.$$

The assumption of the theorem, that is, "gradients of the active inequality constraints and the gradients of the equality constraints are linearly independent at x_*", is one of the possible regularity conditions for KKT conditions to hold. There are other regularity conditions. For example, if $g(x)$ and $h(x)$ are linear constraints, then KKT conditions hold [1].

As a counter example of KKT condition when the necessary regularity condition is violated, we consider the following problem:

$$\min_x [x_1 + x_2^2] \qquad \text{subject to} \quad x_1^2 \le 0.$$

The optimal solution is $[x_1, x_2] = [0, 0]$, and the gradient $\nabla_{x_1} f(x_*) = 1$ and $\nabla_{x_1} g_1(x_*) = 0$. Therefore for any μ_1, we have $\nabla_{x_1}[f(x_*) + \mu_1 g_1(x_*)] \neq 0$. This violation of KKT condition is due to the fact that at the optimal solution, $\nabla g_1(x_*) = 0$. This violates the regularity condition of the theorem.

The detailed proof of the theorem can be found in [1], and is beyond the scope of this class. However, to illustrate the key ideas, in the following, we present an informal proof of the KKT conditions by considering only a single inequality constraint $g_1(x) \le 0$, with $k = 1$ and $m = 0$.

Proof [of Theorem 1 with $k = 1$ and $m = 0$] Consider a local solution x_* of (1). If $g_1(x_*) < 0$ then we can remove this constraint without affecting the local solution. This means we can set $\mu_1 = 0$, and $\nabla f(x_*) = 0$. This implies that the KKT conditions hold in this case.

If $g_1(x_*) = 0$, then x_* is a local solution of

$$\min_x f(x) \quad \text{subject to } g_1(x) = 0.$$

In this case, we automatically have the complementary slackness condition $\mu_1 g_1(x_*) = 0$.

Now consider any direction Δx such that $\nabla g_1(x_*)^\top \Delta x < 0$. Consider the solution $x' = x_* + t\Delta x$ for $t \to 0_+$. We know that $g(x') \le 0$ when t is sufficiently small.

$$f(x') = f(x_*) + t\nabla f(x_*)^\top \Delta x + o(t) \ge f(x_*).$$

It follows that $\nabla f(x_*)^\top \Delta x \ge 0$ for all Δx such that $\nabla g_1(x_*)^\top \Delta x < 0$. Since $\nabla g_1(x_*) \neq 0$ by the assumption of the theorem, we may define $\mu_1 = -\nabla f(x_*)^\top \nabla g_1(x_*)/\|\nabla g_1(x_*)\|_2^2$.

If $\mu_1 \neq 0$, then

$$\nabla f(x_*)^\top [\mu_1 \nabla g_1(x_*)] = -\mu_1^2 \|\nabla g_1(x_*)\|_2^2 < 0.$$

Therefore we must have $\nabla g_1(x_*)^\top [\mu_1 \nabla g_1(x_*)] \ge 0$, which implies that $\mu_1 \ge 0$. This proves the dual feasibility.

Let $\Delta x = \nabla f(x_*) + (\mu_1 + t)\nabla g_1(x_*)$ for some $t \to 0_+$, we have

$$\nabla g_1(x_*)^\top [-\Delta x] = -t\|\nabla g_1(x_*)\|_2^2 < 0.$$

Therefore $\nabla f(x_*)^\top [-\Delta x] \geq 0$. Let $t \to 0$, we know that

$$\nabla f(x_*)^\top \Delta x \leq 0.$$

Since $\nabla g_1(x_*)^\top \Delta x = 0$, we have

$$\Delta x^\top \Delta x = \nabla f(x_*)^\top \Delta x \leq 0.$$

This implies that $\Delta x = 0$. That is, we obtain the stationarity condition $\nabla f(x_*) + \mu_1 \nabla g_1(x_*) = 0$. ■

A Geometric illustration of the two cases in the proof can be seen from Figure 1: the first case is when the constraint is inactive ($g_1(x_*) < 0$) and the second caseis when the constraint is active ($g_1(x_*) = 0$). When the constraint is active, the stationarity condition means that the constraint $g_1(x) = 0$ is tangent to the level-set of $f(x)$, which implies that there exists μ_1 such that $\nabla f(x_*) + \mu_1 \nabla g_1(x_*) = 0$. Moreover, along the direction $-\nabla g_1(x_*)$, which points into the feasibility region $g(x) \leq 0$, $f(x)$ should be non-decreasing due to the optimality of $f(x_*)$. This means that $\nabla f(x_*)^\top \nabla g_1(x_*) \leq 0$, and thus the gradients of $f(x_*)$ and $g_1(x_*)$ are in the opposite directions. This means that we have $\mu_1 \geq 0$.

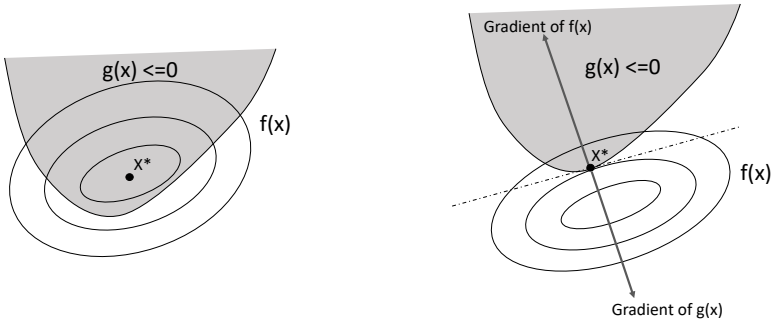

Figure 1: Illustration of KKT Conditions

Example 1 *Find the solution of the following optimization problem of* $x \in \mathbb{R}^2$*:*

$$\min_x [x_1^2 + x_2^2 + x_3^2] \qquad \textit{subject to } x_1 + x_2 + x_3 \geq 1.$$

Solution The constraint can be written as $1 - (x_1 + x_2 + x_3) \leq 0$. We define Lagrangian function as

$$L(x, \mu) = x_1^2 + x_2^2 + x_3^2 + \mu[1 - (x_1 + x_2 + x_3)].$$

The KKT conditions at the solution x are:

- $\nabla_x L(x, \mu) = 0$, which implies $2x_j - \mu = 0$ for $j = 1, 2, 3$.
- $x_1 + x_2 + x_3 \geq 1$
- $\mu(x_1 + x_2 + x_3 - 1) = 0$ and $\mu \geq 0$.

If $\mu = 0$, then we have $x = 0$, but this does not satisfy the inequality constraint. Therefore we must have $\mu > 0$ and $x_1 + x_2 + x_3 = 1$. Since $x_j = \mu/2$, we have $\mu = 2/3$. The solution is $x_1 = x_2 = x_3 = 1/3$. ■

3 Convex Formulation

In the convex formulation, we assume that $f(x)$ is a convex function but not necessarily differentiable. Each $g_j(x)$ is a continuously differentiable convex function. Each $h_j(x) = 0$ is a linear constraint, so that the set of constraints $h_j(x) = 0$ for $j = 1, \ldots, m$ can be reformulated as

$$Ax + b = 0. \tag{2}$$

For convex functions, KKT conditions are both necessary and sufficient, under mild regularity conditions.

Theorem 2 *Assume that* (1) *is convex with linear equality constraint as in* (2)*. Moreover, assume that there exists x satisfying* (2) *such that $g_j(x) < 0$ for all j. Then x_* is an optimal solution of* (1) *if and only if the KKT conditions of Theorem 1 are satisfied with a subgradient of $\nabla f(x_*)$.*

Proof The proof can be found in [1]. We will present a proof of sufficiency. Assume that the KKT conditions hold. Then there exists a subgradient $\nabla f(x_*)$ of $f(x)$ at x_* such that for some $\mu_j \geq 0$ and $\mu_j g_j(x_*) = 0$, we have

$$\nabla f(x_*) + \sum_{j=1}^{k} \mu_j \nabla g_j(x_*) + \sum_{j=1}^{m} \lambda_j \nabla h_j(x_*) = 0.$$

Given any $x \in C = \{x : g(x) \leq 0, h(x) = 0\}$, we have

$$\mu_j \nabla g_j(x_*)^\top (x - x_*) \leq \mu_j (g_j(x) - g_j(x_*)) = \mu_j g_j(x) \leq 0.$$

Moreover, $\lambda_j \nabla h_j(x_*)^\top (x - x_*) = 0$. Therefore

$$f(x) - f(x_*) \geq \nabla f(x_*)^\top (x - x_*) = -\sum_{j=1}^{k} \mu_j \nabla g_j(x_*)^\top (x - x_*) \geq 0.$$

■

In the following example, we can pose an unconstrained convex optimization problem with non-smooth objective function as a constrained but smooth problem. They lead to KKT conditions that are equivalent.

Example 2 *Consider the SVM method below with $C > 0$. Given $\{(x_i, y_i) : i = 1, \ldots, n\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$, we want to find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ to solve*

$$[w_*, b_*, \xi_*] = \arg\min_{w,b,\xi} \left[C \sum_{i=1}^{n} \xi_i + \frac{1}{2}\|w\|_2^2 \right], \tag{3}$$

$$\text{subject to } \xi_i \geq 0, \quad (w^\top x_i + b) y_i + \xi_i \geq 1 \qquad (i = 1, \ldots, n). \tag{4}$$

Solution The Lagrangian function is

$$L(w, b, \xi, \mu, \nu) = C \sum_{i=1}^{n} \xi_i + \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \mu_i \xi_i - \sum_{i=1}^{n} \nu_i [(w^\top x_i + b) y_I + \xi_i - 1].$$

The KKT conditions are

- $\mu_i \xi_i = 0$ and $\nu_i [(w^\top x_i + b) y_i + \xi_i - 1] = 0$ and $\mu_i \geq 0$ and $\nu_i \geq 0$.
- $\xi_i \geq 0$ and $(w^\top x_i + b) y_i + \xi_i \geq 1$
- $\nabla_{w,b,\xi} L(w, b, \xi, \mu, \nu) = 0$.

From $\nabla_\xi L(w, b, \xi, \mu, \nu) = 0$, we obtain $\mu_i = C - \nu_i$. Since $\mu_i \geq 0$, we have $\nu_i \in [0, C]$. We consider three cases:

- $(w^\top x_i + b) y_i < 1$. This implies that $\xi_i > 0$. We obtain from the complementary slackness condition $\mu_i = 0$ and thus $\nu_i = C$.
- $(w^\top x_i + b) y_i > 1$. We obtain from the complementary slackness condition $\nu_i = 0$.
- $(w^\top x_i + b) y_i = 1$. If $\xi_i > 0$, then the complementary slackness condition $\mu_i = 0$ and thus $\nu_i = C$. However, this violates the complementary slackness condition. Therefore $\xi_i = 0$, and we obtain $\nu_i \in [0, C]$.

In summary, at the optimal solution, we have $\xi_i = (1 - (w^\top x_i + b) y_i)_+$ and $L(\cdot)$ can be simplified as

$$L(w, b, \xi, \mu, \nu) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \nu_i [(w^\top x_i + b) y_I - 1].$$

It is easy to verify that ν_i given above is a subgradient of $C(u_i)_+$ at $u_i = 1 - (w^\top x_i + b) y_i$. Taking derivative with respect to w and b, we obtain

$$\nabla_{w,b} \left[\frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \nu_i [(w^\top x_i + b) y_I - 1] \right] = 0.$$

We note that by setting $\xi_i = (1 - (w^\top x_i + b) y_i)_+$, we obtain the unconstrained SVM formulation as

$$\min_{w,b} f(w, b) \qquad f(w, b) = \left[C \sum_{i=1}^{n} (1 - (w^\top x_i + b) y_i)_+ + \frac{1}{2}\|w\|_2^2 \right]. \tag{5}$$

The optimality condition for this unconstrained problem is that a subgradient is 0. Using subgradient algebra and let

$$\nu_i \in C \cdot \partial(u_i)_+|_{u_i=1-(w^\top x_i+b)y_i} = \begin{cases} C & u_i > 0 \\ 0 & u_i < 0 \\ \in [0, C] & u_i = 0 \end{cases},$$

we obtain

$$\nabla_w f(w, b) = -\sum_{i=1}^{n} \nu_i x_i y_i + w = 0,$$

and

$$\nabla_b f(w, b) = -\sum_{i=1}^{n} \nu_i y_i = 0.$$

These optimality conditions are equivalent to (5) derived from the KKT conditions for the constrained formulation. ■

Example 3 *Consider the Lasso method. Given $X \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$, we want to find $w \in \mathbb{R}^d$ to solve*

$$[w_*, \xi_*] = \arg\min_{w,b,\xi} \left[\|Xw - y\|_2^2 + \lambda \sum_{j=1}^{d} \xi_j \right], \tag{6}$$

$$\text{subject to } \xi_j \geq w_j, \quad \xi_j \geq -w_j \quad (j = 1, \ldots, d). \tag{7}$$

Lasso produces sparse solutions. Define the support of the solution as

$$S = \{j : w_{*,j} \neq 0\}.$$

Find and simplify the KKT conditions in terms of $S, X_S, X_{\bar{S}}, y, w_S$. Here X_S contains the columns of X in S, $X_{\bar{S}}$ contains the columns of X not in S, and w_S contains the nonzero components of w_*.

References

[1] Dimitri Bertsekas. *Nonlinear Programming (3rd ed.)*. Athena Scientific, 2016.