Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 3: Karush-Kuhn-Tucker Conditions

Optimization

We consider the following form of constrained optimization problem

$$\min_x f(x) \tag{1}$$
$$\text{subject to } g_j(x) \leq 0 \quad (j = 1, \ldots, k)$$
$$\text{and } h_j(x) = 0 \quad (j = 1, \ldots, m),$$

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.

Lagrangian Functions

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 $g(x) = [g_1(x), \ldots, g_k(x)]$ and $h(x) = [h_1(x), \ldots, h_m(x)]$.

KKT Conditions

Theorem

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:*
$$g(x_*) \leq 0, \qquad h(x_*) = 0.$$
- *Dual Feasibility:*
$$\mu_j \geq 0 \qquad \forall j = 1, \ldots, k.$$
- *Complementary Slackness:*
$$\mu_j g_j(x_*) = 0 \qquad \forall j = 1, \ldots, k.$$

Counter Example of KKT

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

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

Proof of KKT with $k = 1$ and $m = 0$

To show that there exists μ_1 such that:

- Complementary Slackness:

$$\mu_1 g_1(x_*) = 0.$$

- Dual Feasibility:

$$\mu_1 \geq 0.$$

- Stationarity:

$$\nabla f(x_*) + \mu_1 \nabla g_1(x_*) = 0.$$

Proof of Complementary Slackness

Consider a local solution x_* of (1).
If $g_1(x_*) < 0$ then we can remove the constraint without affecting the local solution. This means we can set $\mu_1 = 0$, and the KKT conditions hold.

Therefore we only need to consider the case $g_1(x_*) = 0$.
This implies the complementary slackness condition $\mu_1 g_1(x_*) = 0$.

Dual Feasibility

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 $g(x') \leq 0$ when t is sufficiently small.

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

It follows that

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

for all Δx such that $\nabla g_1(x_*)^\top \Delta x < 0$.

Dual Feasibility

Since $\nabla g_1(x_*) \neq 0$ by the assumption of the theorem, we may define

$$\mu_1 = -\nabla f(x_*)^\top \nabla g_1(x_*) / \|\nabla g_1(x_*)\|_2^2.$$

We have $\mu_1 \geq 0$.

Proof of Stationarity

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

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

This implies the stationarity condition.

Example

Example

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

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

The solution is $x_1 = x_2 = x_3 = 1/3$.

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

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

KKT conditions for Convex Formulation

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

Theorem

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 of Sufficiency

Assume that the KKT conditions hold. Then there exists a subgradient $\nabla f(x_*)$ of $f(x)$ at x_* such that

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

and

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

and

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

Therefore

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

Example

Example

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

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

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

KKT for SVM

The Lagrangian function is

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

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

Simplified KKT for SVM

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

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

Taking derivative with respect to w and b, we obtain

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

Optimality condition for unconstrained SVM

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

We obtain

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

and

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

where ν_i is a subgradient of $(u_i)_+$ at $u_i = 1 - (w^\top x_i + b) y_i$.

Excercise: KKT conditions for Lasso

Example

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_*.