KKT Conditions

October 8, 2024

1 KKT conditions

Definition: given general problem

 $\min_{x} f(x)$ subject to $h_i(x) \le 0, i = 1, \dots, m$ $\ell_j(x) = 0, j = 1, \dots, r,$

and corresponding dual problem

 $\max_{u,v} g(u,v)$
subject to $u \ge 0$.

The KKT conditions for x, u, v are:

1. Stationarity:

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

2. Complementary slackness:

$$u_i \cdot h_i(x) = 0 \; \forall i$$

3. Primal feasibility:

$$h_i(x) \le 0, \ \ell_j(x) = 0 \ \forall i, j$$

4. Dual feasibility:

 $u_i \ge 0 \; \forall i$

Necessity of KKT conditions: If x^* and u^* , v^* are primal and dual solutions, with strong duality holds (with zero duality gap), then x^* and u^* , v^* satisfy the KKT conditions.

Proof. We have

$$f(x^{\star}) = g(u^{\star}, v^{\star})$$

= $\min_{x} f(x) + \sum_{i=1}^{m} u_{i}^{\star} h_{i}(x) + \sum_{j=1}^{r} v_{j}^{\star} \ell_{j}(x)$ (*)
 $\leq f(x^{\star}) + \sum_{i=1}^{m} u_{i}^{\star} h_{i}(x^{\star}) + \sum_{j=1}^{r} v_{j}^{\star} \ell_{j}(x^{\star})$ (**)
 $\leq f(x^{\star}).$ (* * *)

Thus all these inequalities are actually equalities.

• from (*) to (**), we see that $L(x^*, u^*, v^*) = \min_x L(x, u^*, v^*)$. The point x^* minimizes $L(x, u^*, v^*)$ over $x \in \mathbb{R}^n$. Hence we have the stationarity

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

Sometimes this can be used to characterize or compute primal solutions

from (★★) to (★ ★ ★), we must have ∑_{i=1}^m u_i^{*}h_i(x^{*}) = 0, and since each term here is ≤ 0, this implies u_i^{*}h_i(x^{*}) = 0 for every *i*. This is complementary slackness.

Sufficiency of KKT conditions: When the primal problem is convex, and if x^* and u^* , v^* satisfy the KKT conditions, then x^* and u^* , v^* are primal and dual solutions.

Proof. We have

(by definition)
$$g(u^*, v^*) = \min_x f(x) + \sum_i u_i^* h_i(x) + \sum_j v_j^* \ell_j(x)$$

(by convexity+stationarity) $= f(x^*) + \sum_i u_i^* h_i(x^*) + \sum_j v_j^* \ell_j(x^*)$

(by comp. slackness and prime/dual feasibility) = $f(x^*)$.

The second line follows since $L(x, u^*, v^*)$ is convex in x and therefore the stationarity condition $0 \in \partial f(x^*) + \sum_{i=1}^m u_i^* \partial h_i(x^*) + \sum_{j=1}^r v_j^* \partial \ell_j(x^*)$ implies that x^* is the minimizer of $L(x, u^*, v^*)$. Therefore the duality gap is zero (and x^* and u^* , v^* are primal and dual feasible), thus by Proposition 3 of Duality notes, x^* and u^* , v^* are primal and dual optimal.

2 Constrained and Lagrange forms

Often in statistics and machine learning we will switch back and forth between constrained form, where $t \in \mathbb{R}$ is a tuning parameter,

$$\min_{x} f(x) \qquad \text{subject to} \qquad h(x) \le t \qquad (C)$$

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

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

and claim these are equivalent. Now we prove the equivalence solution \hat{x} of (C) and \tilde{x} of (L).

Proof. Denote the solution of (C) as \hat{x} and the solution of (L) as \tilde{x} .

<u>To show (C) to (L)</u>. If problem (C) is strictly feasible (satisfies Slate's condition – there exists at least one strictly feasible x such that h(x) < t), then strong duality holds, then \hat{x} should satisfy the KKT conditions of (C)

$$0 \in \partial f(\hat{x}) + \lambda \partial (h(\hat{x}) - t) \qquad \text{(stationarity of C)}$$
$$\iff 0 \in \partial f(\hat{x}) + \lambda \partial h(\hat{x}) \qquad \text{(stationarity of L)}$$

which implies \hat{x} satisfies the KKT condition of (L), so \hat{x} is also a solution in (L).

<u>To show (L) to (C)</u>. if \tilde{x} is a solution of (L), then we can show that \tilde{x} satisfies the KKT conditions for (C), since

$$0 \in \partial f(\tilde{x}) + \lambda \partial h(\tilde{x}) \qquad \text{(stationarity of L)}$$
$$\iff 0 \in \partial f(\tilde{x}) + \lambda \partial (h(\tilde{x}) - t) \qquad \text{(stationarity of C)}$$

which is the stationarity condition for (C). The complementary slackness condition

 $\lambda(h(\tilde{x}) - t) = 0$ (complementary slackness of C)

is also satisfied if we set $t = h(\tilde{x})$. Or if t is large enough so that $h(\tilde{x}) < t$, in order for both to be equivalent one must set $\lambda = 0$. Taking $t = h(x^*)$ in (C), we see that (i.e. " = " strictly holds in $h(x^*) \le t$). Therefore all the KKT conditions for (C) are satisfied and x^* is a solution in (C).

3 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

$$\min_{\beta} f(X\beta) + \lambda \|\beta\|_1$$

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 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 $p \gg n$)

Proof. the KKT conditions are

$$-X^T \nabla f(X \boldsymbol{\beta}) = \lambda s,$$

where

$$s_j \in \begin{cases} \{\operatorname{sign}(\beta_j)\} & \beta_j \neq 0 \\ [-1,1] & \beta_j = 0 \end{cases} \qquad j = 1, \dots, p$$

Define the equicorrelation set

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

The equicorrelation set S is named as such because when y, X have been standardized, S contains the variables that have equal (and maximal) absolute correlation with the residual $-\nabla f(X\beta) = r$. Note that for any solution $j \notin S$, $\beta_j = 0$.

First assume that X_S is not full column rank, i.e. $\operatorname{rank}(X_S) \leq |S|$. (here $X_S \in \mathbb{R}^{n \times |S|}$,

submatrix of X corresponding to columns in S). Then for some $k \in S$, X_j can be expressed as the linear combination of other X_k 's where $k \in S \setminus \{j\}$:

$$X_j = \sum_{k \in S \setminus \{j\}} c_k X_k$$

for constants $c_j \in \mathbb{R}$, so that

$$s_j X_j = s_j \sum_{k \in S \setminus \{j\}} c_k X_k = \sum_{k \in S \setminus \{j\}} s_j c_k X_k = \sum_{k \in S \setminus \{j\}} (s_j s_k c_k) \cdot (s_k X_k) \quad \text{since } s_k s_k = 1, s_k = \text{sign}(\beta_k)$$

By definition of the equicorrelation set, $X_k^{\top} r = s_j \lambda$ for any $j \in S$. $-X_j^T \nabla f(X\beta) = \lambda s_j$. Hence taking an inner product with $-\nabla f(X\beta)$,

$$-s_j X_j^T \nabla f(X\boldsymbol{\beta}) = \sum_{k \in S \setminus \{j\}} s_j s_k c_k (-s_k X_k^T \nabla f(X\boldsymbol{\beta}))$$
$$s_j s_j \lambda = \sum_{k \in S \setminus \{j\}} s_j s_k c_k (s_k s_k \lambda)$$
$$\lambda = \sum_{k \in S \setminus \{j\}} s_j s_k c_k \lambda$$
$$1 = \sum_{k \in S \setminus \{j\}} (s_j s_k c_k),$$

assuming that $\lambda > 0$. Therefore, we have shown that if $\operatorname{null}(X_S) \neq \{0\}$, then for some $j \in S$

$$s_j X_j = \sum_{k \in S \setminus \{j\}} a_k(s_k X_k)$$

with $\sum_{k \in S \setminus \{j\}} a_k = 1$, which means that $s_j X_j$ lies in the affine span of $s_k X_k$, $k \in S \setminus \{j\}$.