Karush-Kuhn-Tucker conditions

Barnabas Poczos & Ryan Tibshirani
Convex Optimization 10-725/36-725
Recall duality

Given a minimization problem

\[
\min_{x \in \mathbb{R}^n} f(x)
\]

subject to \( h_i(x) \leq 0, \ i = 1, \ldots m \)
\( \ell_j(x) = 0, \ j = 1, \ldots r \)

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:

\[
g(u, v) = \min_{x \in \mathbb{R}^n} L(x, u, v)
\]
The subsequent dual problem is:

$$\max_{u \in \mathbb{R}^m, v \in \mathbb{R}^r} g(u, v)$$
subject to \( u \geq 0 \)

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 \( h_i, i = 1, \ldots, m \))
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

\[
f(x) - f^* \leq f(x) - g(u, v)
\]

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 \( f(x) - f^* \leq \epsilon \)

Very useful, especially in conjunction with iterative methods ... more dual uses in coming lectures
Dual norms

Let $\|x\|$ be a norm, e.g.,

- $\ell_p$ norm: $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$, for $p \geq 1$
- Nuclear norm: $\|X\|_{\text{nuc}} = \sum_{i=1}^{r} \sigma_i(X)$

We define its dual norm $\|x\|_*$ as

$$\|x\|_* = \max_{\|z\|\leq 1} z^T x$$

Gives us the inequality $|z^T x| \leq \|z\| \|x\|_*$, like Cauchy-Schwartz.

Back to our examples,

- $\ell_p$ norm dual: $(\|x\|_p)_* = \|x\|_q$, where $1/p + 1/q = 1$
- Nuclear norm dual: $(\|X\|_{\text{nuc}})_* = \|X\|_{\text{op}} = \sigma_{\text{max}}(X)$

Dual norm of dual norm: it turns out that $\|x\|_{**} = \|x\|$ ... connections to duality (including this one) in coming lectures
Outline

Today:

- KKT conditions
- Examples
- Constrained and Lagrange forms
- Uniqueness with $\ell_1$ penalties
Karush-Kuhn-Tucker conditions

Given general problem

\[
\min_{x \in \mathbb{R}^n} \ f(x)
\]

subject to \( h_i(x) \leq 0, \ i = 1, \ldots, m \)
\( \ell_j(x) = 0, \ j = 1, \ldots, r \)

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

\[
f(x^*) = g(u^*, v^*)
\]

\[
= \min_{x \in \mathbb{R}^n} f(x) + \sum_{i=1}^{m} u^*_i h_i(x) + \sum_{j=1}^{r} v^*_j \ell_j(x)
\]

\[
\leq f(x^*) + \sum_{i=1}^{m} u^*_i h_i(x^*) + \sum_{j=1}^{r} v^*_j \ell_j(x^*)
\]

\[
\leq f(x^*)
\]

In other words, all these inequalities are actually equalities.
Two things to learn from this:

- The point $x^*$ minimizes $L(x, u^*, v^*)$ over $x \in \mathbb{R}^n$. Hence the subdifferential of $L(x, u^*, v^*)$ must contain 0 at $x = x^*$—this is exactly the stationarity condition.

- We must have $\sum_{i=1}^{m} u^*_i h_i(x^*) = 0$, and since each term here is $\leq 0$, this implies $u^*_i h_i(x^*) = 0$ for every $i$—this is exactly complementary slackness.

Primal and dual feasibility obviously hold. Hence, we’ve shown:

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

$$g(u^*, v^*) = f(x^*) + \sum_{i=1}^{m} u^*_i h_i(x^*) + \sum_{j=1}^{r} v^*_j \ell_j(x^*)$$

$$= f(x^*)$$

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

Therefore duality gap is zero (and $x^*$ and $u^*, v^*$ are primal and dual feasible) so $x^*$ and $u^*, v^*$ are primal and dual optimal. I.e., we’ve shown:

If $x^*$ and $u^*, v^*$ satisfy the KKT conditions, then $x^*$ and $u^*, v^*$ are primal and dual solutions
Putting it together

In summary, KKT conditions:

- 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),

$$ x^* \text{ and } u^*, v^* \text{ are primal and dual solutions} $$

$$ \iff \quad x^* \text{ and } u^*, v^* \text{ satisfy the KKT conditions} $$

(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

Many people use the term the KKT conditions when dealing with unconstrained problems, i.e., to refer to stationarity condition

Note that we could have alternatively derived the KKT conditions from studying optimality entirely via subgradients

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

where recall \( \mathcal{N}_C(x) \) is the normal cone of \( C \) at \( x \)
Quadratic with equality constraints

Consider for $Q \succeq 0$,

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T Q x + c^T x$$

subject to $Ax = 0$

E.g., as in Newton step for $\min_{x \in \mathbb{R}^n} f(x)$ 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)
Water-filling

Example from B & V page 245: consider problem

\[ \min_{x \in \mathbb{R}^n} - \sum_{i=1}^{n} \log(\alpha_i + x_i) \]

subject to \( x \geq 0, \ 1^T x = 1 \)

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

\[-1/(\alpha_i + x_i) - u_i + v = 0, \ i = 1, \ldots n\]

\[ u_i \cdot x_i = 0, \ i = 1, \ldots n, \ x \geq 0, \ 1^T x = 1, \ u \geq 0 \]

Eliminate \( u \):

\[ 1/(\alpha_i + x_i) \leq v, \ i = 1, \ldots n \]

\[ x_i(v - 1/(\alpha_i + x_i)) = 0, \ i = 1, \ldots n, \ x \geq 0, \ 1^T x = 1 \]
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)
Lasso

Let's return the lasso problem: given response \( y \in \mathbb{R}^n \), predictors \( X \in \mathbb{R}^{n \times p} \) (columns \( X_1, \ldots, X_p \)), solve

\[
\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \| y - X \beta \|_2^2 + \lambda \| \beta \|_1
\]

KKT conditions:

\[
X^T (y - X \beta) = \lambda s
\]

where \( s \in \partial \| \beta \|_1 \), i.e.,

\[
s_i \in \begin{cases} 
\{1\} & \text{if } \beta_i > 0 \\
\{-1\} & \text{if } \beta_i < 0 \\
[-1, 1] & \text{if } \beta_i = 0
\end{cases}
\]

Now we read off important fact: if \( |X_i^T (y - X \beta)| < \lambda \), then \( \beta_i = 0 \) ... we’ll return to this problem shortly
Group lasso

Suppose predictors $X = [X_{(1)} \ X_{(2)} \ \ldots \ X_{(G)}]$, split up into groups, with each $X_{(i)} \in \mathbb{R}^{n \times p(i)}$. If we want to select entire groups rather than individual predictors, then we solve the group lasso problem:

$$
\min_{\beta=(\beta_{(1)},\ldots,\beta_{(G)}) \in \mathbb{R}^p} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \sum_{i=1}^{G} \sqrt{p(i)} \|\beta_{(i)}\|_2
$$

(From Yuan and Lin (2006), “Model selection and estimation in regression with grouped variables”)
KKT conditions:

\[ X_{(i)}^T(y - X\beta) = \lambda \sqrt{p(i)} s(i), \quad i = 1, \ldots G \]

where each \( s(i) \in \partial \| \beta(i) \|_2 \), i.e.,

\[ s(i) \in \begin{cases} \{ \beta(i)/\| \beta(i) \|_2 \} & \text{if } \beta(i) \neq 0 \\ \{ z \in \mathbb{R}^{p(i)} : \| z \|_2 \leq 1 \} & \text{if } \beta(i) = 0 \end{cases} \quad i = 1, \ldots G \]

Hence if \( \| X_{(i)}^T(y - X\beta) \|_2 < \lambda \sqrt{p(i)} \), then \( \beta(i) = 0 \). On the other hand, if \( \beta(i) \neq 0 \), then

\[ \beta(i) = \left( X_{(i)}^T X(i) + \frac{\lambda \sqrt{p(i)}}{\| \beta(i) \|_2} I \right)^{-1} X_{(i)}^T r_{-}(i), \]

where \( r_{-}(i) = y - \sum_{j \neq i} X(j) \beta(j) \)
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,

$$\min_{x \in \mathbb{R}^n} f(x) \text{ subject to } h(x) \leq t$$  \hspace{1cm} (C)

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

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda \cdot h(x)$$  \hspace{1cm} (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 solution $x^*$ in (C) minimizes

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

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

**Conclusion:**

\[
\bigcup_{\lambda \geq 0} \{ \text{solutions in (L)} \} \subseteq \bigcup_t \{ \text{solutions in (C)} \}
\]

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

Strictly speaking this is not a perfect equivalence (albeit minor nonequivalence). Note: when the only value of \( t \) that leads to a feasible but not strictly feasible constraint set is \( t = 0 \), i.e.,

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

(e.g., this is true if \( g \) is a norm) then we do get perfect equivalence
Uniqueness in $\ell_1$ penalized problems

Using the KKT conditions and simple probability arguments, we can produce 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 $\mathbb{R}^{np}$), 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 $p \gg n$)

Proof: the KKT conditions are

$$-X^T \nabla f(X\beta) = \lambda s, \quad s_i \in \begin{cases} \{\text{sign}(\beta_i)\} & \text{if } \beta_i \neq 0 \\ [-1, 1] & \text{if } \beta_i = 0 \end{cases}, \quad i = 1, \ldots n$$
Note that $X\beta, s$ are unique. Define $S = \{ j : |X_j^T \nabla f(X\beta)| = \lambda \}$, also unique, and note that any solution satisfies $\beta_i = 0$ for all $i \notin S$.

First assume that $\text{rank}(X_S) < |S|$ (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 $c_j \in \mathbb{R}$, hence

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

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 \( \text{rank}(X_S) < |S| \) implies \( s_i X_i \) is in the affine span of \( s_j X_j, j \in S \setminus \{i\} \) (subspace of dimension \(< n\))

We say that the matrix \( X \) has columns in general position if any affine subspace \( L \) of dimension \( k < n \) does not contain more than \( k + 1 \) elements; of \( \{\pm X_1, \ldots \pm X_p\} \) (excluding antipodal pairs)

It is straightforward to show that, if the entries of \( X \) have a density over \( \mathbb{R}^{np} \), then \( X \) is in general position with probability 1

![Diagram](image-url)
Therefore, if entries of $X$ are drawn from continuous probability distribution, any solution must satisfy $\text{rank}(X_S) = |S|$

Recalling the KKT conditions, this means the number of nonzero components in any solution is $\leq |S| \leq \min\{n, p\}$

Furthermore, 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

One of the most important uses 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 $u^*, v^*$, any primal solution $x^*$ satisfies the stationarity condition

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

In other words, $x^*$ achieves the minimum in $\min_{x \in \mathbb{R}^n} L(x, u^*, v^*)$

- 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.
References