COM S 672: Advanced Topics in Computational Models of Learning – Optimization for Learning

Lecture Note 4: Optimality Conditions

Jia (Kevin) Liu

Assistant Professor
Department of Computer Science
Iowa State University, Ames, Iowa, USA

Fall 2017
Recap Last Lecture

Given a minimization problem

Minimize \( f(x) \)
subject to \( g_i(x) \leq 0, \quad i = 1, \ldots, m \) \( \leftarrow u_i \geq 0 \)
\( h_j(x) = 0, \quad j = 1, \ldots, p \) \( \leftarrow v_j \) unconstrained

We define the Lagrangian:

\[
L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i g_i(x) + \sum_{j=1}^{p} v_j h_j(x)
\]

and the Lagrangian dual function:

\[
\Theta(u, v) = \min_{x} L(x, u, v)
\]
Recap Last Lecture

The subsequent **Lagrangian dual problem** is:

Maximize $\Theta(u, v)$

subject to $u \geq 0$

Important properties:

- Dual problem is always convex (or $\Theta$ is always concave), even if the primal problem is non-convex
- The weak duality property always holds, i.e., the primal and dual optimal values $p^*$ and $d^*$ satisfy $p^* \geq d^*$
- Slater’s condition: for convex primal, if $\exists$ $x$ such that $f_1(x) < 0, \ldots, f_m(x) < 0$ and $h_1(x) = 0, \ldots, h_p(x) = 0$.

then **strong duality** holds: $p^* = d^*$. 
Outline

Today:

- KKT conditions
- Geometric interpretation
- Relevant examples in machine learning and other areas
Karash-Kuhn-Tucker Conditions

Given general problem

\[ \text{Minimize} \quad f(x) \]
\[ \text{subject to} \quad g_i(x) \leq 0, \quad i = 1, \ldots, m \quad \leftarrow u_i \geq 0 \]
\[ h_j(x) = 0, \quad j = 1, \ldots, p \quad \leftarrow v_j \text{ unconstrained} \]

The Karash-Kuhn-Tucker (KKT) conditions are:

- **Stationarity (ST):** \( \nabla_x f(x) + \sum_{i=1}^{m} u_i \nabla_x g_i(x) + \sum_{j=1}^{p} v_j \nabla_x h_j(x) = 0 \)

- **Complementary slackness (CS):** \( u_i g_i(x) = 0, \forall i \) either \( u_i = 0 \) or \( g_i(x) = 0 \).

- **Primal feasibility (PF):** \( f_i \leq 0, h_j = 0, \forall i, j \)

- **Dual feasibility (DF):** \( u_i \geq 0, \forall i \)

\[ JKL \]
KKT Necessity

Theorem 1

If $x^*$ and $u^*, v^*$ be primal and dual solutions w/ zero duality gap (e.g., implied by Slater’s condition), then $(x^*, u^*, v^*)$ satisfy KKT conditions.

Proof. We have PF and DF for free from the assumption. Also, $x^*$ and $(u^*, v^*)$ are primal & dual solutions with strong duality $\Rightarrow$

$$f(x^*) = \Theta(u^*, v^*) = \min_x \left\{ f(x) + \sum_{i=1}^{m} u_i^* g_i(x) + \sum_{j=1}^{p} v_j^* h_j(x) \right\}$$

$$\leq f(x^*) + \sum_{i=1}^{m} u_i^* g_i(x^*) + \sum_{j=1}^{p} v_j^* h_j(x^*) \leq f(x^*)$$

That is, all these inequalities are equalities. Then:

- $x^*$ minimizes $L(x, u^*, v^*)$ over $x \in \mathbb{R}^n$ (unconstrained) $\Rightarrow$ Gradient of $L(x, u^*, v^*)$ must be 0 at $x^*$, i.e., the stationarity condition.
- Since $u_i^* h_i(x^*) \leq 0$ (PF & DF), we must $\sum_{i=1}^{m} u_i^* h_i(x^*) = 0$, i.e., complementary slackness condition.
KKT Sufficiency

\((x^*, u^*, v^*)\) is KKT pt. \(\Rightarrow\) \(\begin{cases} x^* \text{ is primal opt} \\ u^*, v^* \text{ is dual opt} \end{cases}\)

Theorem 2

*If \(x^*\) and \((u^*, v^*)\) satisfy KKT conditions, then \(x^*\) and \((u^*, v^*)\) are primal and dual optimal solutions, respectively.*

**Proof.** If \(x^*\) and \((u^*, v^*)\) satisfy KKT conditions, then from \((ST)\), \(\forall L(x^*, u^*, v^*) = 0 \Rightarrow x^* \text{ is a minimizer of } L(x, u^*, v^*)\)

\[
\Theta(u^*, v^*) \overset{(a)}{=} f(x^*) + \sum_{i=1}^{m} u^*_i g_i(x^*) + \sum_{j=1}^{p} v^*_j h_j(x^*)
\]

\[
\overset{(b)}{=} f(x^*),
\]

where \((a)\) follows from \(ST\) and \((b)\) follows from \(CS\).

Therefore, the duality gap is zero. Note that \(x^*\) and \((u^*, v^*)\) are PF and DF. Hence, they are primal and dual optimal, respectively. \(\square\)
In Summary

So putting things together...

**Theorem 3**

*For a convex optimization problem with strong duality (e.g., implied by Slater’s conditions or other constraints qualifications):*

\[ x^* \text{ and } (u^*, v^*) \text{ are primal and dual solutions} \]

\[ \iff x^* \text{ and } (u^*, v^*) \text{ satisfy KKT conditions} \]

**Warning:** This statement is only true for convex optimization problems. For non-convex optimization problems, KKT conditions are neither necessary nor sufficient! (more on this shortly)
Where Does this Name Come From?

Older books/papers referred to this as the KT (Kuhn-Tucker) conditions

- First appeared in a publication by Kuhn and Tucker in 1951
- Kuhn & Tucker shared the John von Neumann Theory Prize in 1980
- Later people realized that Karush had the same conditions in his unpublished master's thesis in 1939

William Karush
Harold W. Kuhn
Albert W. Tucker

Other Optimality Conditions

- KKT conditions are a special case of the more general Fritz John Conditions:

\[ u_0 \nabla f(x^*) + \sum_{i=1}^{m} u_i \nabla g_i(x^*) + \sum_{j=1}^{p} v_j \nabla h_j(x^*) = 0 \]

where \( u_0 \) could be 0

- In turn, Fritz John conditions (hence KKT) belong to a wider class of the first-order necessary conditions (FONC), which allow for non-smooth functions using subderivatives

- Further, there are a whole class second-order necessary & sufficient conditions (SONC, SOSC) – also in “KKT style”

- For an excellent treatment on optimality conditions, see [BSS, Ch.4–Ch.6]
Geometric Interpretation of KKT

\[ (ST): \sum_{i \in I^*} \left[ \nabla g_i (x^*) \right] u_i = -\nabla f (x^*) \]

\[ g_i (x) \leq 0 \]

\[ g_i (x) = 0 \]

\[ \mathcal{I} (x^*) = \{ i : g_i (x^*) = 0 \} \]

Physics interpretation:
- \(-\nabla f (x^*)\): "pulling force"
- \(-\nabla g_i (x^*)\): "repelling force"
- \(-\nabla g_i (x^*)\) is contained in the cone defined by the normal vectors of the binding constr.
When is KKT neither sufficient nor necessary?

- (Not necc.): $x^*$ is a (local) minimum $\not\Rightarrow$ $x^*$ is a KKT point
  
  $g_2(x)$ is non-convex.

  $g_1(x) \leq 0$

- (Not suff.): $x^*$ is a KKT point $\not\Rightarrow$ $x^*$ is a (local) minimum

  $z^*$ is opt. but NOT KKT.
  
  $b/c \nabla f(x^*) \neq u_1 \nabla g_1(x^*) + u_2 \nabla g_2(x^*)$
  
  $\forall u_1, u_2 \geq 0$.

  (Note: $z^*$ is called Fritz John point, must take $u_0 = 0$)

  $g_2(x)$ is non-convex.

  $g_1(x) \leq 0$

  $z^*$ is KKT pt: $\exists u_1, u_2 \geq 0$

  such that
  
  $-z = u_1 \nabla g_1(x^*) + u_2 \nabla g_2(x^*)$.

  But, $z^*$ is not opt.

  e.g., $z \neq z^* < z^*$.
Example 1: Quadratic Problems with Equality Constraints

Consider for $Q \succeq 0$, the following quadratic programming problem is:

Lagrangian: $\frac{1}{2} x^T Q x + c^T x + \frac{1}{2} \eta^T (A x - \xi) + \eta^T (A x - \xi)$

Minimize $\frac{1}{2} x^T Q x + c^T x$

subject to $A x = 0$ \hspace{1cm} \leftarrow \eta$

A convex problem w/o inequality constraints. By KKT, $x$ is primal optimal iff

\[
\begin{cases}
Q x + c + A^T \eta = 0 \\
A x = 0 \\
A \eta = 0
\end{cases}
\]

KKT system.

Often arises from using Newton’s method to solved equality-constrained problems $\{\min_x f(x) | A x = b\}$

Let $x^* = x - x^*$, $Q = H(x^*)$

By Taylor's 2nd expansion: $f(x) \approx f(x^*) + \nabla f(x^*) (x - x^*) + \frac{1}{2} (x - x^*)^T H(x^*) (x - x^*)$

Further, we want $A x = b$. Note $A \Delta x = b \Rightarrow A \Delta x = 0$.
Example 2: Support Vector Machine

Given labels $y \in \{-1, 1\}^n$, feature vectors $x_1, \ldots, x_m$. Let $X \triangleq [x_1, \ldots, x_m]^\top$

Recall from Lecture 1 that the support vector machine problem:

Minimize $w, b, \epsilon$

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \epsilon_i$$

subject to $y_i(w^\top x_i + b) \geq 1 - \epsilon_i, \quad i = 1, \ldots, m \leftarrow u_i \geq 0, \forall i$

$$\epsilon_i \geq 0, \quad i = 1, \ldots, m \leftarrow v_i \geq 0, \forall i$$

Introducing dual variables $u, v \geq 0$ to obtain the KKT system:

(ST): $0 = \sum_{i=1}^{m} u_i y_i, \quad w = \sum_{i=1}^{m} u_i y_i x_i, \quad u = C1 - v$

(CS): $v_i \epsilon_i = 0, \quad u_i (1 - \epsilon_i - y_i(x_i^\top w + b)) = 0, \quad i = 1, \ldots, m$

Lagrangian: $\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \epsilon_i + \sum_{i=1}^{m} u_i [1-\epsilon_i - y_i(x_i^\top w + b)] - \sum_{i=1}^{m} v_i \epsilon_i$: Quadratic in $w$, Affine in $x, b$.

Take der. w.r.t. $w$: $w - \sum_{i=1}^{m} u_i y_i x_i = 0$, w.r.t. $b$: $-\sum_{i=1}^{m} u_i y_i = 0$, w.r.t. $\epsilon_i$: $C - u_i - v_i = 0, \forall i$. 


Example 2: Support Vector Machine

\[ w = \sum_{i=1}^{m} y_i x_i \mathbf{u} = \mathbf{X} \mathbf{u} \]

Hence, at optimality, we have \( w = \sum_{i=1}^{m} u_i y_i x_i \), and \( u_i \) is nonzero only if \( y_i(x_i^T w + b) = 1 - \epsilon_i \). Such points are called the support points.

- For support point \( i \), if \( \epsilon_i = 0 \), then \( x_i \) lies on the edge of margin and \( u_i \in (0, C] \).
- For support point \( i \), if \( \epsilon_i \neq 0 \), then \( x_i \) lies on wrong side of margin, and \( u_i = C \).

KKT conditions do not really give us a way to find solution here, but gives better understanding & useful in proofs.

In fact, we can use this to screen away non-support points before performing optimization (lower-complexity).
Example 3: Water-filling

Example from [BV]: Consider the problem

\[
\begin{align*}
\text{Minimize} & \quad - \sum_{i=1}^{n} \log(\alpha_i + x_i) \\
\text{subject to} & \quad x \geq 0, \quad 1^T x = 1, \quad i = 1, \ldots, m
\end{align*}
\]

In Information Theory: \(\log(\alpha_i + x_i)\) is the communication rate of \(i\)th channel.

Introducing dual variables \(u, v \geq 0\) to obtain the KKT system:

\[
\begin{align*}
\text{(ST)}: & \quad -1/(\alpha_i + x_i) - u_i + v = 0, \quad i = 1, \ldots, n \\
\text{(CS)}: & \quad u_i x_i = 0, \quad i = 1, \ldots, n \\
\text{(PF)}: & \quad x \geq 0, \quad 1^T x = 1 \\
\text{(DF)}: & \quad u \geq 0, \quad v \text{ unconstrained}
\end{align*}
\]

Eliminating \(u\) yields:

\[
\begin{align*}
1/(\alpha_i + x_i) & \leq v, \quad i = 1, \ldots, n \\
x_i(v - 1/(\alpha_i + x_i)) & = 0, \quad i = 1, \ldots, n, \quad x \geq 0, \quad 1^T x = 1 \\
\end{align*}
\]
Example 3: Water-filling

ST and CS implies that:

\[
x_i = \begin{cases} 
1/v - \alpha_i & \text{if } v < 1/\alpha_i \\
0 & \text{if } v \geq 1/\alpha_i
\end{cases}
\]

Also, from PF, i.e., \( \mathbf{1}^\top \mathbf{x} = 1 \), we have:

\[
\sum_{i=1}^{n} \max\{0, 1/v - \alpha_i\} = 1
\]

Univariate equation, piecewise linear in \( 1/v \) and not hard to solve

This reduced problem is referred to as the water-filling solution

(From [BV], pp. 246)
Constrained and Lagrange Forms

Often in ML and Stats, we’ll switch back and forth between constrained form, where \( t \in \mathbb{R} \) is a tuning parameter

\[
(C): \min_{x} f(x) \quad \text{subject to} \quad g(x) \leq t
\]

and Lagrange form, where \( u \geq 0 \) is a tuning parameter

\[
(L): \min_{x} f(x) + u \cdot g(x)
\]

and claim these are equivalent. Is this true (assuming \( f \) and \( g \) convex)?

Proof. \((C)\) to \((L)\): If Problem \((C)\) is strictly feasible, then strong duality holds (why?), and there exists some \( u \geq 0 \) (dual solution) such that any solution \( x^* \) in \((C)\) minimizes

\[
\frac{\partial \Psi(x)}{\partial x} + u g(x^*) + ut
\]

\[
f(x) + u \cdot (g(x) - t).
\]

Clearly, \( x^* \) 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 \( t = g(x^*) \), so \( x^* \) is a solution in (C).

Putting things together:

\[
\bigcup_{u \geq 0} \begin{cases} \text{solutions in (L)} \end{cases} \subseteq \bigcup_{t} \begin{cases} \text{solutions in (C)} \end{cases}
\]

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

I.e., nearly perfect equivalence. Note: If the only value of \( t \) that leads to a feasible but not strictly feasible constraint set is \( t = 0 \), then we do get perfect equivalence.

So, e.g., if \( g \geq 0 \) and (C) and (L) are feasible for all \( t, u \geq 0 \), then we do get perfect equivalence.
Next Class

Gradient Descent