COM S 578X: Optimization for Machine Learning

Lecture Note 5: Optimality Conditions

Jia (Kevin) Liu

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

Fall 2019

Recap Last Lecture

Given a minimization problem

$$\begin{array}{lllll} \text{Minimize} & f(\mathbf{x}) & & & \\ \text{subject to} & g_i(\mathbf{x}) \le 0, & i = 1, \ldots, m & & \leftarrow u_i \ge 0 \\ & h_j(\mathbf{x}) = 0, & j = 1, \ldots, p & & \leftarrow v_j \text{ unconstrained} \end{array}$$

We define the Lagrangian:

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

and the Lagrangian dual function:

$$\Theta(\mathbf{u}, \mathbf{v}) = \min_{\mathbf{x}} L(\mathbf{x}, \mathbf{u}, \mathbf{v})$$

Recap Last Lecture

The subsequent Lagrangian dual problem is:

$$\begin{aligned} \text{Maximize} \quad & \Theta(\mathbf{u}, \mathbf{v}) \\ \text{subject to} \quad & \mathbf{u} \geq \mathbf{0} \end{aligned}$$

Important properties:

- Dual problem is always convex (or Θ 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 $\exists\ \mathbf{x}$ such that

$$f_1(\mathbf{x}) < 0, \ldots, f_m(\mathbf{x}) < 0 \text{ and } h_1(\mathbf{x}) = 0, \ldots, h_p(\mathbf{x}) = 0.$$

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

Outline

Today:

- KKT conditions
- Geometric interpretation
- Relevant examples in machine learning and other areas

Karush-Kuhn-Tucker Conditions

Given general problem

$$\begin{array}{lll} \text{Minimize} & f(\mathbf{x}) & \\ \text{subject to} & g_i(\mathbf{x}) \le 0, \quad i = 1, \ldots, m & \leftarrow u_i \ge 0 \\ & h_j(\mathbf{x}) = 0, \quad j = 1, \ldots, p & \leftarrow v_j \text{ unconstrained} \end{array}$$

$L(\underline{x}, \underline{u}, \underline{v}) = f(\underline{x}) + \underline{u}^T g(\underline{x}) + \underline{v}^T h(\underline{x})$ the grad of $L(\underline{x}, \underline{u}, \underline{v})$ w.r.t. $\underline{x}$ is $\underline{0}$

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

- Stationarity (ST): $\nabla_{\mathbf{x}} f(\mathbf{x}) + \sum_{i=1}^{m} u_i \nabla_{\mathbf{x}} g_i(\mathbf{x}) + \sum_{j=1}^{p} v_j \nabla_{\mathbf{x}} h_j(\mathbf{x}) = 0$
- Complementary slackness (CS): $u_i g_i(\mathbf{x}) = 0$, $\forall i$ either $u_i = 0$ or $g_i(\underline{x}) = 0$
- Primal feasibility (PF): $g_i(\mathbf{x}) \le 0$, $h_j(\mathbf{x}) = 0$, $\forall i, j$
- Dual feasibility (DF): $u_i \ge 0$, $\forall i$

KKT Necessity

If $\underline{x}^*$ primal opt., $(\underline{u}^*, \underline{v}^*)$ dual opt. $\Big\}\xrightarrow{\text{strong duality}}$ $(\underline{x}^*, \underline{u}^*, \underline{v}^*)$ are KKT.

Theorem 1

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

WTS: (ST), (CS), (PF), (DF)

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

$$\underbrace{f(\mathbf{x}^*) \overset{\text{strong duality}}{=} \Theta(\mathbf{u}^*, \mathbf{v}^*)} \overset{\text{def of dual fn.}}{=} \min_{\mathbf{x}} \left\{ f(\mathbf{x}) + \sum_{i=1}^{m} u_i^* g_i(\mathbf{x}) + \sum_{j=1}^{p} v_j^* h_j(\mathbf{x}) \right\}$$

def of "min" $\leq L$ evaluated at any $\underline{x}$ (including $\underline{x}^*$) (=).

$$\leq f(\mathbf{x}^*) + \underbrace{\sum_{i=1}^{m} \underbrace{\overset{\text{PF}\ \geq 0}{u_i^*}\, \overset{\text{PF}\ \leq 0}{g_i(\mathbf{x}^*)}}_{\leq 0} + \sum_{j=1}^{p} \underbrace{v_j^*\, \overset{\text{PF}\ =0}{h_j(\mathbf{x}^*)}}_{=0}}_{\leq 0\ (=0)} \leq f(\mathbf{x}^*)$$

That is, all these inequalities are equalities. Then:

- $\mathbf{x}^*$ minimizes $L(\mathbf{x}, \mathbf{u}^*, \mathbf{v}^*)$ over $\mathbf{x} \in \mathbb{R}^n$ (unconstrained) $\Rightarrow$ Gradient of $L(\mathbf{x}, \mathbf{u}^*, \mathbf{v}^*)$ must be 0 at x^*, i.e., the stationarity condition. (ST)
- Since $u_i^* g_i(\mathbf{x}^*) \leq 0$ (PF & DF), we must have each $u_i^* g_i(\mathbf{x}^*) = 0$, i.e., complementary slackness condition. (CS)

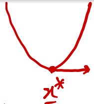
□

KKT Sufficiency

If $\left.\begin{array}{l}(\underline{x}^*, \underline{u}^*, \underline{v}^*) \text{ is KKT} \\ \text{primal is convex}\end{array}\right\} \Rightarrow \left\{\begin{array}{l}\underline{x}^* \text{ is primal opt.} \\ (\underline{u}^*, \underline{v}^*) \text{ is dual opt.}\end{array}\right.$

Theorem 2

If the primal problem is convex and $\mathbf{x}^*$ *and* $(\mathbf{u}^*, \mathbf{v}^*)$ *satisfy KKT conditions, then* $\mathbf{x}^*$ *and* $(\mathbf{u}^*, \mathbf{v}^*)$ *are primal and dual optimal solutions, respectively.*



Proof. If $\mathbf{x}^*$ and $(\mathbf{u}^*, \mathbf{v}^*)$ satisfy KKT conditions, then

Lagrangian: $L = f(\underline{x}) + \underline{u}^T g(\underline{x}) + \underline{v}^T h(\underline{x})$. From (ST): $\nabla_{\underline{x}} L(\underline{x}^*, \underline{u}^*, \underline{v}^*) = \underline{0}$ $\Downarrow$ $\underline{x}^*$ is a minimizer of $L(\underline{x}, \underline{u}^*, \underline{v}^*)$.

$$\Theta(\mathbf{u}^*, \mathbf{v}^*) \overset{(a)}{=} f(\mathbf{x}^*) + \underbrace{\sum_{i=1}^{m} \underbrace{u_i^* g_i(\mathbf{x}^*)}_{=0\ (CS)} + \sum_{j=1}^{p} \underbrace{v_j^* h_j(\mathbf{x}^*)}_{=0\ (PF)}}_{=0}$$

(ST)

$$\overset{(b)}{=} f(\mathbf{x}^*),$$

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

Therefore, the duality gap is zero. Note that $\mathbf{x}^*$ and $(\mathbf{u}^*, \mathbf{v}^*)$ are PF and DF. Hence, they are primal and dual optimal, respectively. □

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):

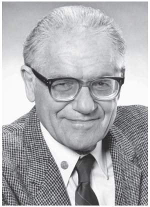
$$\mathbf{x}^* \textit{ and } (\mathbf{u}^*, \mathbf{v}^*) \textit{ are primal and dual solutions}$$
$$\iff \mathbf{x}^* \textit{ and } (\mathbf{u}^*, \mathbf{v}^*) \textit{ satisfy KKT conditions}$$

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

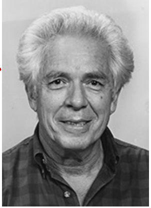
- First appeared in a publication by Kuhn and Tucker in 1951
- Kuhn & Tucker shared the John von Neumann Theory Prize in 1980
- Later people realized that Karush had the same conditions in his unpublished master's thesis in 1939.



Univ. of Chicago
Ph.D. advisor:
Magnus Hestenes.
(Conj. GD).
↓
Lipschitz
↓
Gauss.

William Karush

1975



Princeton.
Friend with
John Nash
Hungarian Alg.
$O(n^4)$ ($O(n^3)$).

Harold W. Kuhn

Princeton:
John Nash
Lloyd Shapley
(shapley value...
stochastic games).
Marvin Minsky

Albert W. Tucker

- A Fun Read: R. W. Cottle, "William Karush and the KKT Theorem," Documenta Mathematica, 2012, pp. 255-269.

RAND. Friend with Richard Bellman. → Assoc. Prof at Chicago. → Cal State University → Died 1997.

Other Optimality Conditions

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

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

  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 conditiosn (SONC,SOSC) – also in "KKT style"

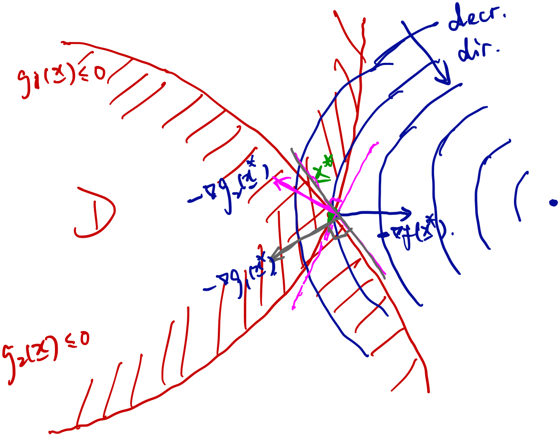
- For an excellent treatment on optimality conditions, see [BSS, Ch.4–Ch.6]

Geometric Interpretation of KKT

Set of binding (active/tight) constraints: $\mathcal{I}(\underline{x}^*) \triangleq \{ i : g_i(\underline{x}^*) = 0 \}$.

(CS): $u_i^* g_i(\underline{x}^*) = 0 \Rightarrow \underline{u}_i^* \geq 0$

(ST): $\underline{\nabla f(\underline{x}^*)} + \sum_i \underline{u}_i^{*T} \nabla g_i(\underline{x}^*) + \cancel{\sum_j \underline{v}_j^{*T} \nabla h_j(\underline{x}^*)} = 0$

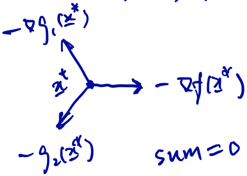


physics interpretation

$-\nabla f(\underline{x}^*)$: "pulling force".

$-\nabla g_i(\underline{x}^*)$, $i \in \mathcal{I}(\underline{x}^*)$.

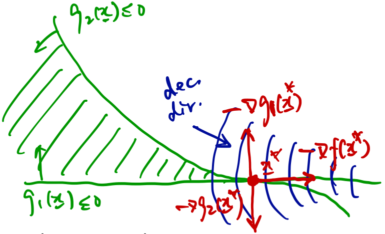
"repelling force".



sum = 0

When is KKT neither sufficient nor necessary?

- (Not necc.): $\mathbf{x}^*$ is a (local) minimum $\not\Rightarrow$ $\mathbf{x}^*$ is a KKT point



$\underline{x}^*$ opt. but not KKT.

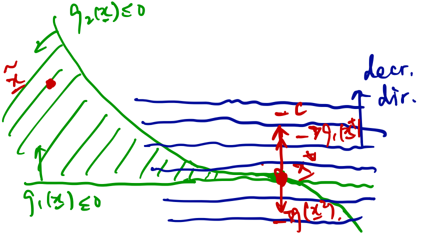
b/c: $\nabla f(\underline{x}^*) \neq u_1 \nabla g_1(\underline{x}^*) + u_2 \nabla g_2(\underline{x}^*)$.

u_0

$\forall u_1, u_2 \geq 0$.

($\underline{x}^*$ is Fritz John pt. for $u_0 = 0$).

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



obj: min $\underline{c}^T \underline{x}$

$\underline{x}^*$ is KKT: $\exists\, u_1, u_2 \geq 0$

s.t. $-\underline{c} = u_1 \nabla g_1(\underline{x}^*) + u_2 \nabla g_2(\underline{x}^*)$

But $\underline{x}^*$ NOT opt.

Example 1: Quadratic Problems with Equality Constraints

- Consider for $\mathbf{Q} \succeq 0$, the following quadratic programming problem is:

Lagrangian: $\frac{1}{2}x^T Q x + c^T x + u^T(Ax)$.

$$\begin{array}{lll} \underset{\mathbf{x}}{\text{Minimize}} & \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{c}^\top \mathbf{x} & \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{0} & \leftarrow \mathbf{u} \end{array}$$

- A convex problem w/o inequality constraints. By KKT, $\mathbf{x}$ is primal optimal iff

(ST): $Qx + c + A^T u = 0$

(PF): $Ax = 0$

(DF) & (CS): Implied by (PF)

$$\Rightarrow \begin{bmatrix} \mathbf{Q} & \mathbf{A}^\top \\ \mathbf{A} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} -\mathbf{c} \\ \mathbf{0} \end{bmatrix}$$

for some dual variable $\mathbf{u}$. A linear equation system combines ST & PF (CS and DF vacuous)

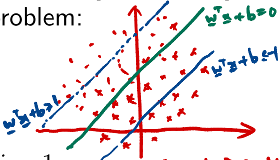
- Often arises from using Newton's method to solved equality-constrained problems $\{\min_{\mathbf{x}} f(\mathbf{x}) | \mathbf{A}\mathbf{x} = \mathbf{b}\}$

By Taylor's SO expansion: $f(x) \approx \underbrace{f(\bar{x})}_{\text{const.}} + \nabla f(\bar{x})^T \underbrace{(x-\bar{x})}_{\tilde{x}} + \frac{1}{2}\underbrace{(x-\bar{x})^T}_{\tilde{x}} \underbrace{H(\bar{x})}_{Q} \underbrace{(x-\bar{x})}_{\tilde{x}}$ $\frac{1}{2}\tilde{x}^T Q \tilde{x}$ $+ o(\|x-\bar{x}\|^2)$

Note: $A\bar{x} = b \quad Ax = b \Rightarrow A\underbrace{(x-\bar{x})}_{\tilde{x}} = 0 \Rightarrow A\tilde{x} = 0$

Example 2: Support Vector Machine

Given labels $y \in \{-1, 1\}^n$, feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$. Let $\mathbf{X} \triangleq [\mathbf{x}_1, \ldots, \mathbf{x}_m]^\top$
Recall from Lecture 1 that the support vector machine problem:



$$\underset{\mathbf{w}, b, \boldsymbol{\epsilon}}{\text{Minimize}} \qquad \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \epsilon_i$$

$$\text{subject to} \quad (\text{PF}) \begin{cases} y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \epsilon_i, & i = 1, \ldots, m \quad \leftarrow u_i \geq 0, \forall i \\ \epsilon_i \geq 0, & i = 1, \ldots, m \quad \leftarrow \underbrace{v_i \geq 0, \forall i}_{(\text{DF})} \end{cases}$$

Slater's cond. hold.

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

Lagrangian: $\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m} \epsilon_i + \sum_{i=1}^{m} u_i[1 - \epsilon_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)] - \sum_{i=1}^{m} v_i \epsilon_i$ Qudratic in $\mathbf{w}$

Affine in ϵ, b

$$(\text{ST}): 0 = \sum_{i=1}^{m} u_i y_i, \quad \mathbf{w} = \sum_{i=1}^{m} u_i y_i \mathbf{x}_i, \quad \mathbf{u} = C\mathbf{1} - \mathbf{v}$$

$$(\text{CS}): v_i \epsilon_i = 0, \quad u_i\big(1 - \epsilon_i - y_i(\mathbf{x}_i^\top \mathbf{w} + b)\big) = 0, \quad i = 1, \ldots, m$$

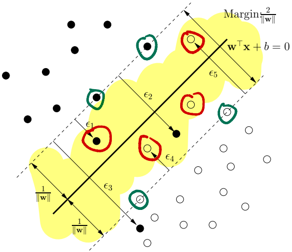
Take der. w.r.t.: $\mathbf{w}$: $\mathbf{w} - \sum_{i=1}^{m} u_i y_i \mathbf{x}_i = 0$, b: $-\sum_{i=1}^{m} u_i y_i = 0$, ϵ_i: $C - u_i - v_i = 0, \forall i$

Example 2: Support Vector Machine

(ST) $\underline{w} = \boxed{\underline{X}\,\mathrm{Diag}\{y_1 \cdots y_m\}}\,\underline{u} = \underline{\tilde{X}}\,\underline{u}$

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

- For support point i, if $\epsilon_i = 0$, then $\mathbf{x}_i$ lies on the edge of margin and $u_i \in (0, C]$ $\varepsilon_i = 0 \overset{(CS)}{\Longrightarrow} v_i \geq 0 \overset{(ST)}{\Longrightarrow} \underline{u} \leq C\underline{1}$
- For support point i, if $\epsilon_i \neq 0$, then $\mathbf{x}_i$ lies on wrong side of margin, and $u_i = C$ $\varepsilon_i \neq 0 \overset{(CS)}{\Longrightarrow} v_i = 0 \overset{(ST)}{\Longrightarrow} \underline{u} = C\underline{1}$



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)

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

Special case: $(t=0)$
cannot find $\underline{x}$
s.t. $g(\underline{x})<0$.
$g(\underline{x})=0, \forall \underline{x}$
set $u=\infty$

$$\text{(C): } \min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad g(\mathbf{x}) \le t$$

and Lagrange form, where $u \ge 0$ is a tuning parameter

$$\text{(L): } \min_{\mathbf{x}} f(\mathbf{x}) + u \cdot g(\mathbf{x})$$

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

WTS: $\underline{x}^* \in (C)$ w/ $t \Rightarrow \underline{x}^* \in (L)$ w/ u.

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

$\exists$ Dual var. u s.t. $\underline{x}^*$ solves $f(\underline{x}) + u(g(\underline{x}) - t)$

$$f(\mathbf{x}) + u \cdot \big(g(\mathbf{x}) - t\big). \;= \underbrace{f(\underline{x}) + u \cdot g(\underline{x})} - \underbrace{ut}_{\text{const}}$$

Clearly, $\mathbf{x}^*$ is also a solution in (L).

Constrained and Lagrange Forms

(L) to (C): If x^* is a solution in (L), then the KKT conditions for (C) are satisfied by taking $t = g(\mathbf{x}^*)$, so $\mathbf{x}^*$ is a solution in (C).

Putting things together: (CS): $u(g(\underline{x}^*) - t) = u(g(\underline{x}^*) - g(\underline{x}^*)) = 0.$

$$\bigcup_{u \geq 0} \Big\{ \text{solutions in (L)} \Big\} \quad \subseteq \quad \bigcup_{t} \Big\{ \text{solutions in (C)} \Big\}$$

$$\bigcup_{u \geq 0} \Big\{ \text{solutions in (L)} \Big\} \quad \supseteq \quad \bigcup_{t\text{: (C) is strictly feasible}} \Big\{ \text{solutions in (C)} \Big\}$$

WTS: $\underline{x}^* \in (L)$ w/ $u \Rightarrow \underline{x}^* \in (C)$ w/ t, i.e., Given $u \geq 0$, WTF $\exists t$, s.t. KKT is satisfied

Try $t = g(\underline{x}^*)$. check: (ST) $\underline{x}^*$ is soln to (L) $\Rightarrow \nabla_{\underline{x}} f(\underline{x}^*) + u \nabla g(\underline{x}^*) = 0$. (PF) & (DF)

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

$\left.\begin{array}{l} g(\underline{x}) \geq 0 \\ g(\underline{x}) \leq t \end{array}\right\} \Rightarrow$ $\left.\begin{array}{l} 1^\circ \text{ when } t \neq 0, \text{ (C) is strictly feas.} \\ 2^\circ \text{ when } t = 0, \ g(\underline{x}) = 0. \end{array}\right\}$ perfect equivalence.

If $g(\underline{x})$ is some norm.

Next Class

Gradient Descent