Assignment Project Exam Help

Karush-Kuhn-Tucker conditions

https://eduassistpro.github.i

Add WeChat edu_assist_pr

Remember duality

Given a minimization problem

$$\min_{x \in \mathbb{R}^n} \; f(x)$$

$$\text{subject to} \;\; h_i(x) \le 0, \;\; i = 1, \ldots m$$

we de

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h$$

and **Lagrange dual function:**

$$g(u, v) = \min_{x \in \mathbb{R}^n} L(x, u, v)$$

The subsequent **dual problem** is:

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

Important properties:

- ven

- atisfy weak duality: $f^\star \geq g^\star$
- Slater's condition: for convex primal, if the at

$$h_1(x) < 0, \ldots h_m(x) < 0 \quad \text{and} \quad {}_1 \quad {}_r$$

then **strong duality** holds: $f^\star = g^\star$. (Can be further refined to strict inequalities over 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

so if the … rly,
u, v are dual optimal)

From an algorithmic viewpoint, provides a stop…
$f(x) - g(u, v) \leq \epsilon$, then we are guaranteed th…

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

Dual norms

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

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

We define its **dual norm** x as

Gives us the inequality $|z^T x| \leq \|z\| \|x$

Back to our examples

- ℓ_p norm dual: $(\|x\|_p)_* = \|x\|_q$, wh
- Nuclear norm dual: $(\|X\|_{\text{nuc}})_* = \|X\|_{\text{spec}} = \sigma_{\max}(X)$

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

Today:

-
-
-
- Uniqueness with 1-norm penalties

Karush-Kuhn-Tucker conditions

Given general problem

$$\min_{x \in \mathbb{R}^n} \; f(x)$$
$$\text{subject to } \; h_i(x) \le 0, \;\; i = 1, \ldots m$$

The … are:

- $0 \in \partial f(x) + \sum_{i=1}^{m} u_i \partial h_i(x) + \sum_{j=1}^{r} v$ … rity)
- $u_i \cdot h_i(x) = 0$ for all i … ss)
- $h_i(x) \le 0, \; \ell_j(x) = 0$ for all i, j (primal feasibility)
- $u_i \ge 0$ for all i (dual feasibility)

Necessity

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

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

$$\leq f(x^\star) + \sum_{i=1}^{m} u_i^\star h_i(x^\star \quad \sum^{r}$$

$$\leq f(x^\star)$$

In other words, all these inequalities are actually equalities

Two things to learn from this:

- The point $x^\star$ minimizes $L(x, u^\star, v^\star)$ over $x \in \mathbb{R}^n$. Hence the subdifferential of $L(x, u^\star, v^\star)$ must contain 0 at $x = x^\star$—this is exactly the **stationarity** condition
- We must have $\,^{m}_{i=1} u_i^\star h_i(x^\star) = 0$, and since each term here is $\;0$, this implies $u^\star h_i(x^\star) = 0$ for every i—this is exactly

Prim

If $x^\star$ and $u^\star, v^\star$ are primal and dual solutio
gap, then $x^\star, u^\star, v^\star$ satisfy the KKT condi

(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 $x^\star, u^\star, v^\star$ that satisfy the KKT conditions, then

$$g(u^\star, v^\star) = f(x^\star) + \sum_{i=1}^{m} u_i^\star h_i(x^\star) + \sum_{j=1}^{r} v_j^\star \ell_j(x^\star)$$

$\star$

wher
hold

Therefore duality gap is zero (and $x^\star$
dual feasible) so $x^\star$ and $u^\star, v^\star$ are prim
we've shown:

> If $x^\star$ and $u^\star, v^\star$ satisfy the KKT conditions, then $x^\star$ and $u^\star, v^\star$ are primal and dual solutions

Putting it together

In summary, KKT conditions:

- always sufficient
- necessary under strong duality

Putt

For a p
tion c n-
affine inequality contraints),

$x^\star$ and $u^\star, v^\star$ are primal and du

$\Leftrightarrow \quad x^\star$ and $u^\star, v^\star$ satisfy the

(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

Man

for un

Note that we could have alternatively derived the

from studying optimality entirely via subgradie

$$0 \in \partial f(x^\star) + \sum_{i=1}^{m} \mathcal{N}_{\{h_i \le 0\}}(x^\star) + \sum_{j=1} \mathcal{N}_{\{\ell_j = 0\}}(x^\star)$$

where recall $\mathcal{N}_C(x)$ is the normal cone of C at x

Quadratic with equality constraints

Consider for $Q \succeq 0$,

$$\min_{x \in \mathbb{R}^n} \ \frac{1}{2} x^T Q x + c^T x$$
$$\text{subject to} \ \ Ax = 0$$

E.g., a b

Conv

x is a solution if and only if

$$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix}$$

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

$$\min_{x \in \mathbb{R}^n} \; -\sum_{i=1}^{n} \log(\alpha_i + x_i)$$
$$\text{subject to } \; x \quad 0, \; 1^T x = 1$$

Infor ... ate of
ith ch

$$-1/(\alpha_i + x_i) - u_i + v = 0$$
$$u_i \cdot x_i = 0, \quad i = 1, \ldots n, \quad x \geq$$

Eliminate u:

$$1/(\alpha_i + x_i) \leq v, \quad i = 1, \ldots n$$
$$x_i(v - 1/(\alpha_i + x_i)) = 0, \quad i = 1, \ldots n, \quad x \geq 0, \quad 1^T x = 1$$

Can argue directly stationarity and complementary slackness imply

$$x_i = \begin{cases} 1/v - \alpha & \text{if } v \le 1/\alpha \\ 0 & \text{if } v > 1/\alpha \end{cases} = \max\{0, 1/v - \alpha\}, \quad i = 1, \ldots n$$

Still need x to be feasible, i.e., $1^T x = 1$, and this gives

Univariate equation, piecewise linear in

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 $A \in \mathbb{R}^{n \times p}$ (columns $A_1, \ldots A_p$), solve

$$\min_{x \in \mathbb{R}^p} \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1$$

KKT

wher $_1$

$$s_i \in \begin{cases} \{1\} & \text{if} \\ \{-1\} & \text{if} \\ [-1, 1] & \text{if } x_i = 0 \end{cases}$$

Now we read off important fact: if $|A_i^T(y - Ax)| < \lambda$, then $x_i = 0$

... we'll return to this problem shortly

Group lasso

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

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

(From Yuan and Lin (2006), "Model selection and estimation in regression with grouped variables")

KKT conditions:

$$A_{(i)}^T(y - Ax) = \lambda\sqrt{p_{(i)}}\, s_{(i)}, \quad i = 1, \ldots G$$

where each $s_{(i)} \in \partial\|x_{(i)}\|_2$, i.e.,

$$x_{(i)}/\|x_{(i)}\|_2 \qquad \text{if } x_{(i)} \neq 0$$

Henc

r

hand, if $x_{(i)} \neq 0$, then

$$x_{(i)} = \left(A_{(i)}^T A_{(i)} + \frac{\lambda\sqrt{p_{(i)}}}{\|x_{(i)}\|_2} I\right)^{-1} A_{(i)}^T$$

$$\text{where} \quad r_{-(i)} = y - \sum_{j \neq i} A_{(j)} x_{(j)}$$

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,

$$\min_{x \in \mathbb{R}^n} f(x) \text{ subject to } h(x) \le t \tag{C}$$

and

(L)

and claim these are equivalent. Is this true (assumi)?

(C) to (L): if problem (C) is strictly feasible, then stro holds, and there exists some $\lambda \ge 0$ (d solution $x^\star$ in (C) minimizes

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

so $x^\star$ is also a solution in (L)

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

Conclusion:

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

$$\}$$

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

$$\{x : g(x) \leq t\} \neq \emptyset,\ \{x : g(x) = t\} = \emptyset \quad \Rightarrow \quad t = 0$$

(e.g., this is true if g is a norm) then we do get perfect equivalence

Uniqueness in 1-norm 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, $A \in \mathbb{R}^{n\times p}$, $\lambda > 0$. Consider

$$\min \ f(Ax) + \lambda \|x\|_1$$

If the e

trib $\in \mathbb{R}^p$

is unique and has at most $\min\{n, p\}$ nonzero components

Remark: here f must be strictly convex, but n

dimensions of A (we could have $p \gg$

Proof: the KKT conditions are

$$-A^T \nabla f(Ax) = \lambda s, \quad s_i \in \begin{cases} \{\mathrm{sign}(x_i)\} & \text{if } x_i \neq 0 \\ [-1, 1] & \text{if } x_i = 0 \end{cases}, \ \ i = 1, \ldots n$$

Note that Ax, s are unique. Define $S = \{j : |A_j^T \nabla f(Ax)| = \lambda\}$, also unique, and note that any solution satisfies $x_i = 0$ for all $i \notin S$

First assume that $\mathrm{rank}(A_S) < |S|$ (here $A \in \mathbb{R}^{n \times |S|}$, submatrix of A corresponding to columns in S). Then for some $i \in S$,

for co

$$s_i A_i = \sum_{j \in S \setminus \{i\}} (s_i s_j c$$

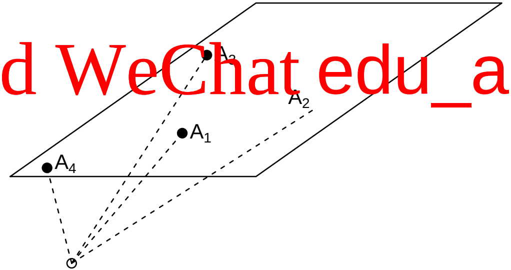
Taking an inner product with $-\nabla f(Ax)$,

$$\lambda = \sum_{j \in S \setminus \{i\}} (s_i s_j c_j)\lambda, \quad \text{i.e.,} \quad \sum_{j \in S \setminus \{i\}} s_i s_j c_j = 1$$

In other words, we've proved that $\mathrm{rank}(A_S) < |S|$ implies $s_i A_i$ is in the affine span of $s_j A_j$, $j \in S \setminus \{i\}$ (subspace of dimension $< n$)

We say that the matrix A has columns in general position if any affine subspace L of dimension $k < n$ does not contain more than $k+1$ elements; of $\{\pm A_1, \dots \pm A_p\}$ (excluding antipodal pairs)

It is str ensity over



Therefore, if entries of A are drawn from continuous probability distribution, any solution must satisfy $\mathrm{rank}(A_S) = |S|$

Recalling the KKT conditions, this means the number of nonzero com

Furt

solvi

$$\min_{x_S \in \mathbb{R}^{|S|}} f(A_S x_S) +$$

Finally, strict convexity implies uniqueness of th problem, and hence in our original problem $\square$

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 $u^\star, v^\star$, any primal solution $x^\star$ satis

$$\sum_{i=1} \quad \sum_{j=1}$$

In other words, $x^\star$ achieves the minimum i $\star)$

- Generally, this reveals a characterizatio
- 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

- *tion*,

- University Press, Chapters 28–30