Karush-Kuhn-Tucker Conditions and Its Usages

Yu-Xiang Wang
CS292F

(Based on Ryan Tibshirani's 10-725)

Last time: duality

Given a minimization problem

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & h_i(x) \leq 0,\ i = 1, \ldots m \\ & \ell_j(x) = 0,\ j = 1, \ldots r \end{array}$$

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:

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

The subsequent dual problem is:

$$\begin{array}{ll} \max\limits_{u,v} & g(u,v) \\ \text{subject to} & u \geq 0 \end{array}$$

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, $i = 1, \ldots m$

Outline

Today:

- KKT conditions
- Examples
- Constrained and Lagrange forms
- Usages of Duality and KKT condition
- Dual norms, Conjugate functions, Dual cones
- Dual tricks and subtleties

Karush-Kuhn-Tucker conditions

Given general problem

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & h_i(x) \leq 0,\ i = 1, \ldots m \\ & \ell_j(x) = 0,\ j = 1, \ldots r \end{array}$$

The Karush-Kuhn-Tucker conditions or KKT conditions are:

- $0 \in \partial\left(f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x) \right)$ (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 $x^\star$ and $u^\star, v^\star$ be primal and dual solutions with zero duality gap (strong duality holds, e.g., under Slater's condition). Then

$$\begin{aligned} f(x^\star) &= g(u^\star, v^\star) \\ &= \min_x \; f(x) + \sum_{i=1}^m u_i^\star h_i(x) + \sum_{j=1}^r v_j^\star \ell_j(x) \\ &\leq f(x^\star) + \sum_{i=1}^m u_i^\star h_i(x^\star) + \sum_{j=1}^r v_j^\star \ell_j(x^\star) \\ &\leq f(x^\star) \end{aligned}$$

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 $\sum_{i=1}^m u_i^\star h_i(x^\star) = 0$, and since each term here is ≤ 0, this implies $u_i^\star h_i(x^\star) = 0$ for every i—this is exactly complementary slackness

Primal and dual feasibility hold by virtue of optimality. Therefore:

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

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

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

Therefore the duality gap is zero (and $x^\star$ and $u^\star, v^\star$ are primal and dual feasible) so $x^\star$ and $u^\star, v^\star$ are primal and dual optimal. Hence, 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

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 contraints),

$$x^\star \text{ and } u^\star, v^\star \text{ are primal and dual solutions}$$
$$\iff x^\star \text{ and } u^\star, v^\star \text{ satisfy the KKT conditions}$$

(Warning, concerning the stationarity condition: for a differentiable function f, we cannot use $\partial f(x) = \{\nabla f(x)\}$ unless f is convex! There are other versions of KKT conditions that deal with local optima.)

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

For unconstrained problems, the KKT conditions are nothing more than the subgradient optimality condition

For general convex problems, the KKT conditions could have been derived entirely from studying optimality via subgradients

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

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

Example: quadratic with equality constraints

Consider for $Q \succeq 0$,

$$\begin{array}{ll} \min\limits_{x} & \frac{1}{2} x^T Q x + c^T x \\ \text{subject to} & Ax = 0 \end{array}$$

E.g., as we will see, this corresponds to Newton step for equality-constrained problem $\min_x \ f(x) \ \text{subject to} \ Ax = b$

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)

Example: support vector machines

Given $y \in \{-1, 1\}^n$, and $X \in \mathbb{R}^{n \times p}$, the support vector machine problem is:

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} \quad & \xi_i \geq 0, \ i = 1, \ldots n \\ & y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \ i = 1, \ldots n \end{aligned}$$

Introduce dual variables $v, w \geq 0$. KKT stationarity condition:

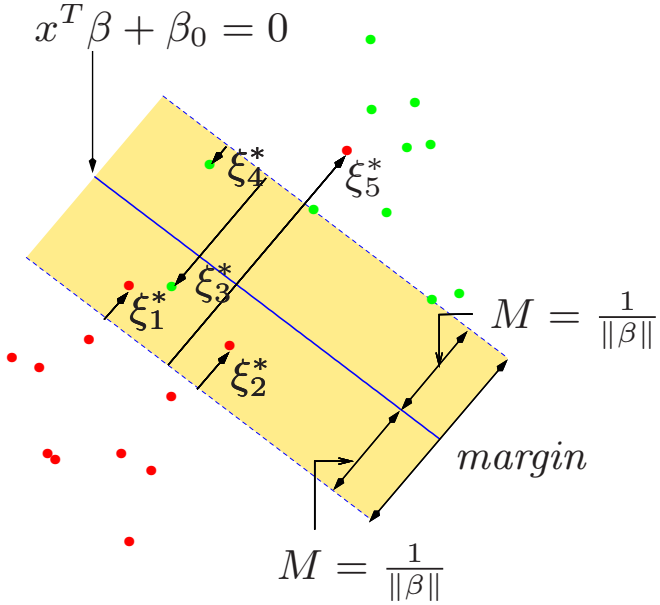
$$0 = \sum_{i=1}^n w_i y_i, \quad \beta = \sum_{i=1}^n w_i y_i x_i, \quad w = C1 - v$$

Complementary slackness:

$$v_i \xi_i = 0, \ w_i\big(1 - \xi_i - y_i(x_i^T\beta + \beta_0)\big) = 0, \quad i = 1, \ldots n$$

Hence at optimality we have $\beta = \sum_{i=1}^{n} w_i y_i x_i$, and w_i is nonzero only if $y_i(x_i^T \beta + \beta_0) = 1 - \xi_i$. Such points i are called the support points

- For support point i, if $\xi_i = 0$, then x_i lies on edge of margin, and $w_i \in (0, C]$;
- For support point i, if $\xi_i \neq 0$, then x_i lies on wrong side of margin, and $w_i = C$



KKT conditions do not really give us a way to find solution, but gives a better understanding

In fact, we can use this to screen away non-support points before performing optimization

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,

$$\min_x \ f(x) \ \text{ subject to } \ h(x) \leq t \tag{C}$$

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

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

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 $x^\star$ in (C) minimizes

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

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:

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

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

This is nearly a perfect equivalence. Note: when 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 $h \geq 0$, and (C), (L) are feasible for all $t, \lambda \geq 0$, then we do get perfect equivalence

Example: Lasso Support Recovery

Consider the standard Lasso:

$$\min_{\beta} \ \frac{1}{2}\|X\beta - y\|^2 + \lambda\|\beta\|_1$$

Suppose we assume that $y = X\beta_0 + \mathcal{N}(0, \sigma^2 I)$, and that β_0 is sparse. How can we prove properties of the optimal solutions β^*?

In particular, can we establish a condition under which there is no false discovery

$$\{i|\beta_i^* \neq 0\} \subset \{i|[\beta_0]_i \neq 0\}$$

or even that the discoveries are all correct (Sparsistency)

$$\{i|\beta_i^* \neq 0\} = \{i|[\beta_0]_i \neq 0\}.$$

The KKT conditions are

$$-X^T(X\beta - y) = \lambda v, \quad v_i \in \begin{cases} \{\text{sign}(\beta_i)\} & \text{if } \beta_i \neq 0 \\ [-1, 1] & \text{if } \beta_i = 0 \end{cases}, \ i = 1, \dots n$$

Prove (Sparsistency) using KKT condition!
Consider the fictitious optimization problem that knows the set of coordinates where β_0 is non-zero to begin with.

$$\min_{\beta_S} \ \frac{1}{2}\|X_S\beta_S - y\|^2 + \lambda\|\beta_S\|_1$$

The optimal solution to this fictitious problem provides us with a candidate primal solution to the original problem. It suffices to construct a dual solution v^* and check the KKT condition of the original problem. In fact, it suffices to show that the dual solution satisfy $v_i^* \in (-1, 1)$ for all $i \notin S$.

Uses of duality

Two key uses of duality:

- For x primal feasible and u, v dual feasible,

$$f(x) - g(u, v)$$

  is called the duality gap between x and u, v. Since

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

  a zero duality gap implies optimality. Also, the duality gap can be used as a stopping criterion in algorithms

- Under strong duality, given dual optimal $u^\star, v^\star$, any primal solution minimizes $L(x, u^\star, v^\star)$ over all x (i.e., it satisfies stationarity condition). This can be used to characterize or compute primal solutions

Solving the primal via the dual

An important consequence of stationarity: under strong duality, given a dual solution $u^\star, v^\star$, any primal solution $x^\star$ solves

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

or simply:

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

Often, solutions of this unconstrained problem can be expressed explicitly, giving an explicit characterization of primal solutions from dual solutions

Furthermore, suppose the solution of this problem is unique; then it must be the primal solution $x^\star$

This can be very helpful when the dual is easier to solve than the primal

Example from B & V page 249:

$$\min_x \sum_{i=1}^n f_i(x_i) \quad \text{subject to} \quad a^T x = b$$

where each $f_i : \mathbb{R} \to \mathbb{R}$ is smooth, strictly convex. Dual function:

$$\begin{aligned} g(v) &= \min_x \sum_{i=1}^n f_i(x_i) + v(b - a^T x) \\ &= bv + \sum_{i=1}^n \min_{x_i} \left\{ f_i(x_i) - a_i v x_i \right\} \\ &= bv - \sum_{i=1}^n f_i^*(a_i v) \end{aligned}$$

where f_i^* is the conjugate of f_i, to be defined shortly

Therefore the dual problem is

$$\max_v \; bv - \sum_{i=1}^n f_i^*(a_i v) \quad \Longleftrightarrow \quad \min_v \; \sum_{i=1}^n f_i^*(a_i v) - bv$$

This is a convex minimization problem with scalar variable—much easier to solve than primal

Given $v^\star$, the primal solution $x^\star$ solves

$$\min_x \; \sum_{i=1}^n \left(f_i(x_i) - a_i v^\star x_i \right)$$

Strict convexity of each f_i implies that this has a unique solution, namely $x^\star$, which we compute by solving $\nabla f_i(x_i) = a_i v^\star$ for each i

Dual norms

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

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

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 generalized Holder). Back to our examples,

- ℓ_p norm dual: $(\|x\|_p)_* = \|x\|_q$, where $1/p + 1/q = 1$
- Trace norm dual: $(\|X\|_{\mathsf{tr}})_* = \|X\|_{\mathsf{op}} = \sigma_1(X)$

Dual norm of dual norm: can show that $\|x\|_{**} = \|x\|$

Proof: consider the (trivial-looking) problem

$$\min_y \ \|y\| \ \ \text{subject to} \ \ y = x$$

whose optimal value is $\|x\|$. Lagrangian:

$$L(y, u) = \|y\| + u^T(x - y) = \|y\| - y^T u + x^T u$$

Using definition of $\|\cdot\|_*$,

- If $\|u\|_* > 1$, then $\min_y \{\|y\| - y^T u\} = -\infty$
- If $\|u\|_* \leq 1$, then $\min_y \{\|y\| - y^T u\} = 0$

Therefore Lagrange dual problem is

$$\max_u \ u^T x \ \ \text{subject to} \ \ \|u\|_* \leq 1$$

By strong duality $f^\star = g^\star$, i.e., $\|x\| = \|x\|_{**}$

Example: lasso dual

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, recall the lasso problem:

$$\min_{\beta} \ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

Its dual function is just a constant (equal to $f^\star$). Therefore we transform the primal to

$$\min_{\beta,z} \ \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 \ \text{ subject to } \ z = X\beta$$

so dual function is now

$$\begin{aligned} g(u) &= \min_{\beta,z} \ \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 + u^T(z - X\beta) \\ &= \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2 - I_{\{v \,:\, \|v\|_\infty \leq 1\}}(X^T u/\lambda) \end{aligned}$$

Therefore the lasso dual problem is

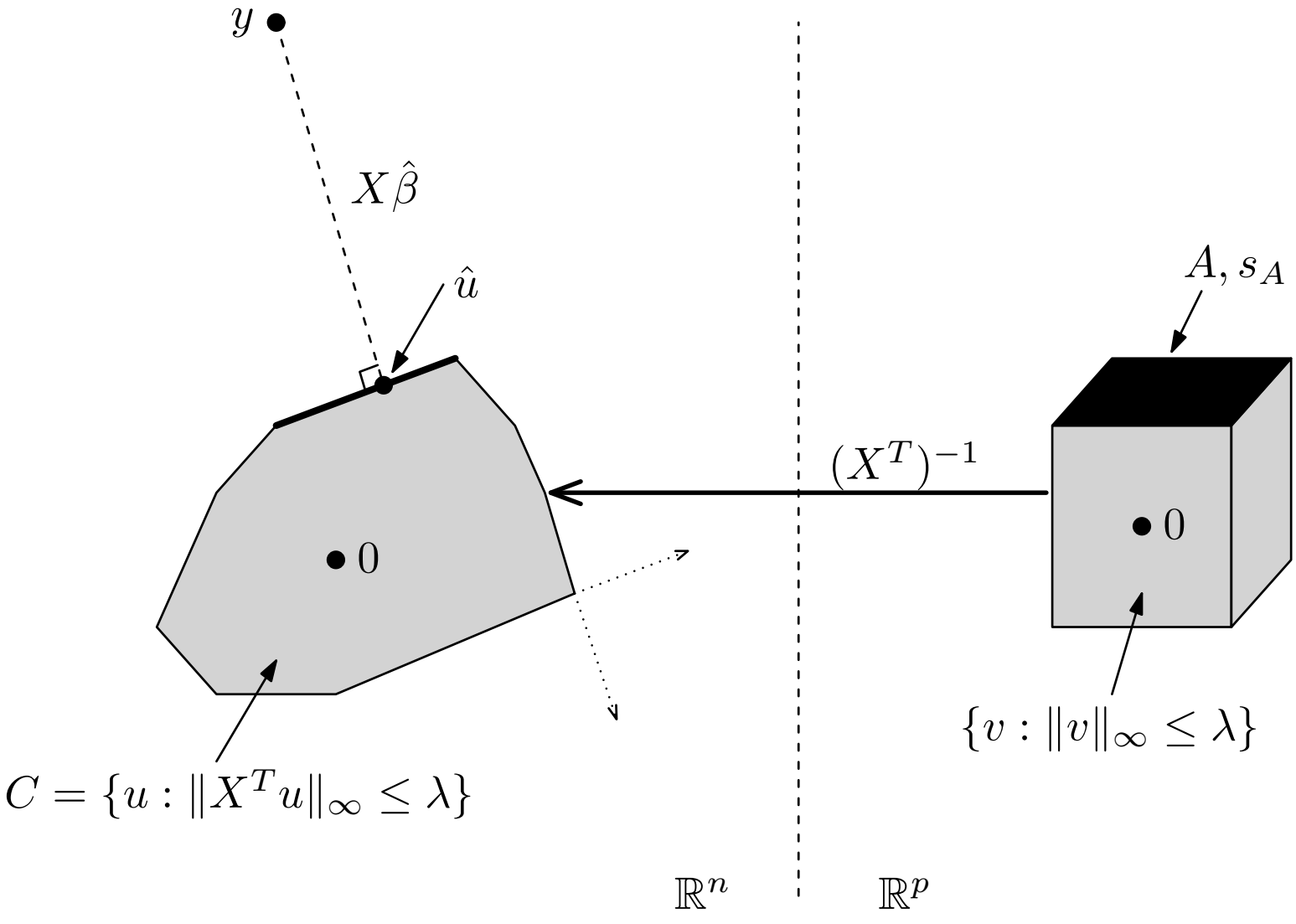
$$\max_u \ \frac{1}{2}\Big(\|y\|_2^2 - \|y-u\|_2^2\Big) \ \text{ subject to } \ \|X^T u\|_\infty \le \lambda$$
$$\iff \ \min_u \ \|y-u\|_2^2 \ \text{ subject to } \ \|X^T u\|_\infty \le \lambda$$

Check: Slater's condition holds, and hence so does strong duality. But note: the optimal value of the last problem is not the optimal lasso objective value

Further, note that given the dual solution u, any lasso solution β satisfies

$$X\beta = y - u$$

This is from KKT stationarity condition for z (i.e., $z - y + \beta = 0$). So the lasso fit is just the dual residual


y
$X\hat{\beta}$
$\hat{u}$
A, s_A
$(X^T)^{-1}$
0
0
$C = \{u : \|X^T u\|_\infty \leq \lambda\}$
$\{v : \|v\|_\infty \leq \lambda\}$
$\mathbb{R}^n$
$\mathbb{R}^p$

Conjugates and dual problems

Conjugates appear frequently in derivation of dual problems, via

$$-f^*(u) = \min_x \; f(x) - u^T x$$

in minimization of the Lagrangian. E.g., consider

$$\min_x \; f(x) + g(x)$$

Equivalently: $\min_{x,z} \; f(x) + g(z) \;\; \text{subject to} \;\; x = z$. Dual function:

$$g(u) = \min_x \; f(x) + g(z) + u^T(z - x) = -f^*(u) - g^*(-u)$$

Hence dual problem is

$$\max_u \; -f^*(u) - g^*(-u)$$

Examples of this last calculation:

- Indicator function:

$$\text{Primal}: \quad \min_x \; f(x) + I_C(x)$$
$$\text{Dual}: \quad \max_u \; -f^*(u) - I_C^*(-u)$$

  where I_C^* is the support function of C

- Norms: the dual of

$$\text{Primal}: \quad \min_x \; f(x) + \|x\|$$
$$\text{Dual}: \quad \max_u \; -f^*(u) \;\; \text{subject to} \;\; \|u\|_* \leq 1$$

  where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$

Shifting linear transformations: Fenchel Dual

Dual formulations can help us by "shifting" a linear transformation between one part of the objective and another. Consider

$$\min_x \; f(x) + g(Ax)$$

Equivalently: $\min_{x,z} \; f(x) + g(z) \;\; \text{subject to} \;\; Ax = z$. Like before, dual is:

$$\max_u \; -f^*(A^T u) - g^*(-u)$$

Example: for a norm and its dual norm, $\|\cdot\|$, $\|\cdot\|_*$:

$$\begin{aligned} \text{Primal}: \quad & \min_x \; f(x) + \|Ax\| \\ \text{Dual}: \quad & \max_u \; -f(A^T u) \;\; \text{subject to} \;\; \|u\|_* \leq 1 \end{aligned}$$

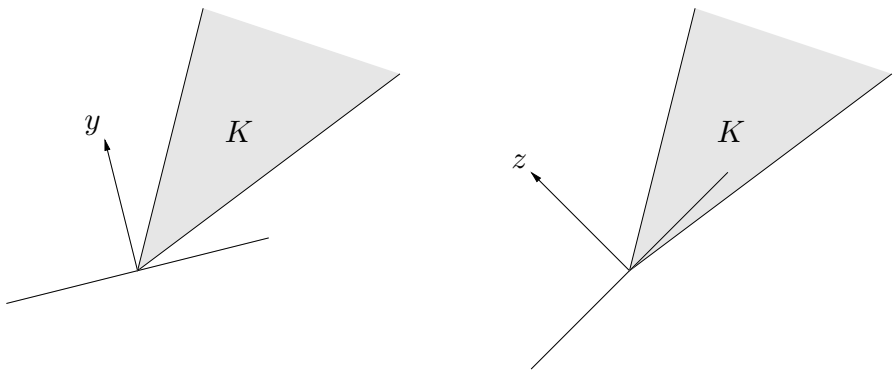
The dual can often be a helpful transformation here

Dual cones

For a cone $K \subseteq \mathbb{R}^n$ (recall this means $x \in K$, $t \geq 0 \Longrightarrow tx \in K$),

$$K^* = \{y : y^T x \geq 0 \text{ for all } x \in K\}$$

is called its dual cone. This is always a convex cone (even if K is not convex)



Notice that $y \in K^*$ $\Longleftrightarrow$ the halfspace $\{x : y^T x \geq 0\}$ contains K

(From B & V page 52)

Important property: if K is a closed convex cone, then $K^{**} = K$

Examples:

- Linear subspace: the dual cone of a linear subspace V is $V^\perp$, its orthogonal complement. E.g., $(\mathrm{row}(A))^* = \mathrm{null}(A)$

- Norm cone: the dual cone of the norm cone

$$K = \{(x,t) \in \mathbb{R}^{n+1} : \|x\| \le t\}$$

  is the norm cone of its dual norm

$$K^* = \{(y,s) \in \mathbb{R}^{n+1} : \|y\|_* \le s\}$$

- Positive semidefinite cone: the convex cone $\mathbb{S}^n_+$ is self-dual, meaning $(\mathbb{S}^n_+)^* = \mathbb{S}^n_+$. Why? Check that

$$Y \succeq 0 \iff \mathrm{tr}(YX) \ge 0 \text{ for all } X \succeq 0$$

  by looking at the eigendecomposition of X

Dual cones and dual problems

Consider the cone constrained problem

$$\min_x \ f(x) \ \text{ subject to } \ Ax \in K$$

Recall that its dual problem is

$$\max_u \ -f^*(A^T u) - I_K^*(-u)$$

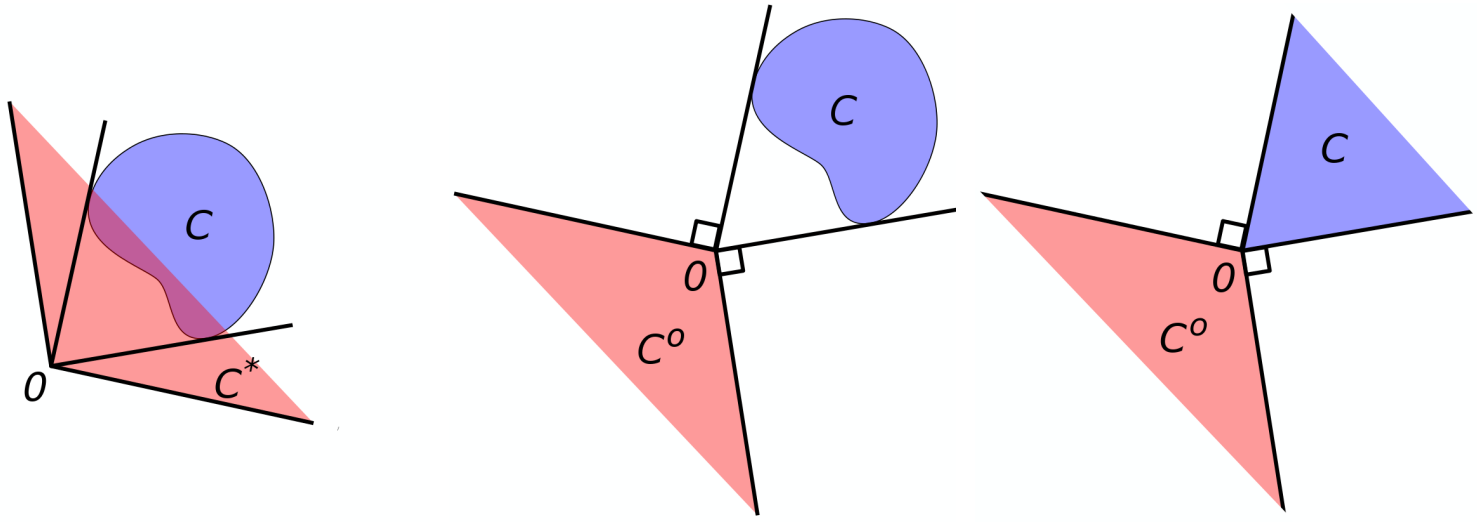
where recall $I_K^*(y) = \max_{z \in K} \ z^T y$, the support function of K. If K is a cone, then this is simply

$$\max_u \ -f^*(A^T u) \ \text{ subject to } \ u \in K^*$$

where K^* is the dual cone of K, because $I_K^*(-u) = I_{K^*}(u)$

This is quite a useful observation, because many different types of constraints can be posed as cone constraints

Dual Cone and Polar Cone



$$K^* = \{y : y^T x \geq 0 \text{ for all } x \in K\}$$

$$K^o = \{y : y^T x \leq 0 \text{ for all } x \in K\}$$

Any x satisfies that $x = \text{Proj}_K(x) + \text{Proj}_{K^o}(x)$. Recall the Moreau Decomposition:

$$x = \text{prox}_f(x) + \text{prox}_{f^*}(x)$$

Let $f = I_K$, its conjugate is $\text{supp}_K(y) = \max_{x \in K} \langle x, y \rangle$.

The prox of the support function of a cone is the projection to its polar cone!

Dual subtleties

- Often, we will transform the dual into an equivalent problem and still call this the dual. Under strong duality, we can use solutions of the (transformed) dual problem to characterize or compute primal solutions

  Warning: the optimal value of this transformed dual problem is not necessarily the optimal primal value

- A common trick in deriving duals for unconstrained problems is to first transform the primal by adding a dummy variable and an equality constraint

  Usually there is ambiguity in how to do this. Different choices can lead to different dual problems!

Double dual

Consider general minimization problem with linear constraints:

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & Ax \leq b,\ Cx = d \end{array}$$

The Lagrangian is

$$L(x, u, v) = f(x) + (A^T u + C^T v)^T x - b^T u - d^T v$$

and hence the dual problem is

$$\begin{array}{ll} \max\limits_{u,v} & -f^*(-A^T u - C^T v) - b^T u - d^T v \\ \text{subject to} & u \geq 0 \end{array}$$

Recall property: $f^{**} = f$ if f is closed and convex. Hence in this case, we can show that the dual of the dual is the primal (Exercise!)

Actually, the connection (between duals of duals and conjugates) runs much deeper than this, beyond linear constraints. Consider

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & h_i(x) \leq 0,\ i = 1, \ldots m \\ & \ell_j(x) = 0,\ j = 1, \ldots r \end{array}$$

If f and $h_1, \ldots h_m$ are closed and convex, and $\ell_1, \ldots \ell_r$ are affine, then the dual of the dual is the primal

This is proved by viewing the minimization problem in terms of a bifunction. In this framework, the dual function corresponds to the conjugate of this bifunction (for more, read Chapters 29 and 30 of Rockafellar)

References

- S. Boyd and L. Vandenberghe (2004), "Convex optimization", Chapters 2, 3, 5
- R. T. Rockafellar (1970), "Convex analysis", Chapters 12, 13, 14, 16, 28–30