6.Convex Problems

Duality, Slater's Conditions, KKT Conditions

Sun Woo Lim

Jan 29, 2021

Convex / Concave Functions defined on the domain of a convex set

For $f : R^n \to R$ defined for $x \in dom(f)$ and assume f takes ∞ outside the domain. For f defined on convex domain, f is convex if $\lambda f(x_1) + (1-\lambda)f(x_2) \geq f(\lambda x_1 + (1-\lambda)x_2)$ and concave if $\lambda f(x_1) + (1-\lambda)f(x_2) \leq f(\lambda x_1 + (1-\lambda)x_2)$

✓ **Properties of Convex Functions**

1. **Pointwise Supremum of convex functions is a convex function**
2. **Nonnegative linear combination of Convex Functions is a convex function**
3. $f : \mathbb{R}^n \to \mathbb{R}$ **is a convex function** $\leftrightarrow epi(f)$ **is a convex set : "Epigraph Characterization of a Convex Function"**
   $epi(f) := \{(x,t) \in \mathbb{R}^n \times \mathbb{R} : x \in dom(f), t \in \mathbb{R}, f(x) \leq t\}$
4. Jensen's Inequality : For a convex function f and a random variable X, $E(f(X)) \geq f(E(X))$
5. Local Minimum of a convex function is a global minimum and strictly convex function has at most 1 global minimum

✓ **Iff conditions for differentiable convex functions**

For f which has an open domain and differentiable on $dom(f)$,

1. **First order (gradient) condition for convexity**
   f convex $\leftrightarrow f(y) \geq f(x) + \nabla f(x)^T(y-x), \forall x, y \in dom(f)$.
2. **Second order (Hessian) condition for convexity**
   f convex $\leftrightarrow \nabla^2 f(x) \succeq 0, \forall x \in dom(f)$. "Hessian is PSD".

General Convex Optimization Problem

$min_x f_0(x)$: Objective Function
subject to (s.t.) $f_i(x) \leq 0, i = 1, 2, ...m,$: m inequality constraints
$h_i(x) = 0, i = 1, 2, ..., p$: p equality constraints

$\rightarrow$ This is a **Convex Optimization Problem** if $f_0, f_1, ..., f_m$ are convex functions and $h_1, ..., h_p$ are affine functions.

Terminologies

$\{x \in \{\cap_{i=0}^{m} dom(f_i) \cap \cap_{j=1}^{p} dom(h_i)\} | f_i(x) \leq 0, \forall i = 1, 2, ..., m, h_j(x) = 0, \forall j = 1, 2, ..., p\}$ is called a **Feasible Set**

An optimization problem having **an empty feasible set** is **infeasible**.

The **infimum** of the objective function over the **feasible set** is called the **primal optimal value**, denoted as $p*$.

If $\exists$ feasible x satisfying $f_0(x) = p*$, say x **attains** the optimum and x^* is called **primal optimal point**.

The set of feasible points at which the optimum is attained is called an **Optimal Set**

Constraints f_i or h_j is(are) **active** at feasible point x if $f_i(x) = 0$ or $h_j(x) = 0$ respectively. Else, they are **inactive** at x.

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex function with $dom(f)$. For $\alpha \in \mathbb{R}$, $S_\alpha := \{x \in \mathbb{R}^n | f(x) \leq \alpha\}$ is called a **Sublevel set** of f.

Dual Norm

Dual Norm of an arbitrary Norm $||x||* := sup_{\{z \in \mathbb{R}^n : ||z|| \leq 1\}} z^T x$

✓ Note that every norm on $\mathbb{R}^n$ is a **convex function**.
pf) For all norm on $\mathbb{R}^n$, $\forall x_1, x_2 \in \mathbb{R}^n, \forall \lambda \in [0,1], ||\lambda x_1 + (1-\lambda)x_2|| \leq ||\lambda x_1|| + ||(1-\lambda)x_2|| = \lambda ||x_1|| + (1-\lambda)||x_2||$

✓ Note that a **dual norm** is a **norm** so, is a **convex function**.
pf)

1. If $||x||^* = 0$, it means that $z^T x = 0$, $\forall z : ||z|| \leq 1$. $\therefore x = 0$. If $x \neq 0 \leftrightarrow ||x|| \neq 0$, since $z = \frac{x}{||x||}$, $||x||^* \geq \frac{||x||_2^2}{||x||} > 0$.
2. $||\alpha x||^* = max_{||z|| \leq 1} |\alpha| z^T x = |\alpha| ||x||^*$
3. Similarly check $||x_1 + x_2||^* \leq ||x_1||^* + ||x_2||^*$

✓ For $p = 1, 2, 3, ..., \infty$, dual of l_p norm is l_q norm for p, q satisfying $\frac{1}{p} + \frac{1}{q} = 1 \because$ using **Holder's Inequality.**

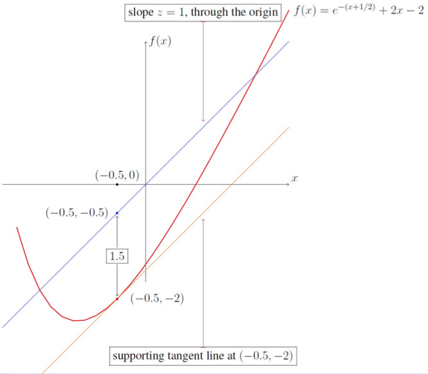
- dual of l_1 norm is l_∞ norm and dual of l_∞ norm is l_1 norm $||z||_1 = max_{u:||u||_\infty \leq 1} u^T z = ||z||_\infty^*$
- dual of l_2 norm is l_2 norm : can prove this by **Cauchy-Schwarz Ineqaulity.** $||z||_2 = max_{u:||u||_2 \leq 1} u^T z = ||z||_2^*$: self dual

Convex Conjugate of a Function

Let $f : \mathbb{R}^n \to \mathbb{R}$ **(need not be cvx ftn)** having a nonempty domain **(need not be cvx set)**.

$f^*(z) := sup_{x \in \mathbb{R}^n}(z^T x - f(x))$: **Convex Conjugate (Fenchel Conjugate)**

✓ Property 1) f^* is always **convex ftn** and **lower semicontinuous : epigraph is a closed set**.
✓ Property 2) If f is convex and **lower semicontinuous** them $f^{**} = f$



Convex Conjugate of a Function

Examples

1. $f(x) = e^x \rightarrow f^*(z) = zlnz - z$
2. $f(x) = a^T x + b \rightarrow$
$$f^*(z) = \begin{cases} -b & z = a \\ \infty & z \neq a \end{cases}$$
3. $f(x) = ||x|| \rightarrow$ conjugate of a norm is "the indicator function of unit dual norm ball"
$$f^*(z) := sup_{x \in \mathbb{R}^m}(z^T x - ||x||) = sup(0, sup_{x \in \mathbb{R}^n - \{0\}}(z^T x - ||x||)$$
$$= \text{sup}(0,\ \text{sup}_{L>0} L(sup_{||x||=1}(z^T x - 1)) = sup(0, sup_{L>0} L(||z||^* - 1)) = \begin{cases} 0 & ||z||_* \leq 1 \\ \infty & ||z||_* > 1 \end{cases}$$
4. $I^*_{\mathbb{R}_-}(z) = I_{\mathbb{R}_+}(z)$. Also, $I_{\mathbb{R}_-}(x) = sup_{z \geq 0} zx$: either by direct calculation or applying dual of dual
5. $I^*_{\{0\}}(z) = 0$. Also, $I_{\{0\}}(x) = sup_{z \in \mathbb{R}} zx$: either by direct calculation or applying dual of dual
6. $I^*_B(z) = sup_{x \in \mathbb{R}^n}(z^T x - I_B(x)) = sup_{x \in B} z^T x = I_{B^\perp}$. Why? consider 1) z is orthogonal to x 2) otherwise.

✓ Used for transforming **primal** problem to its **dual**. Very useful in deriving the **Lagrange Dual Function**.

Example 1 : Slack Variables

$min_x(max_{i=1,2,...,n}x_i - min_{j=1,2,...,n}x_j)$ s.t. $Ax = b$ for $A \in \mathbb{R}^{m,n}$ and $b \in \mathbb{R}^m$.

Introduce slack variables t and u which represent the maximum and minmum respectively.

$min_{x,t,u}(t - u)$ s.t. $Ax = b$, $t \geq x_i, i = 1, 2, ..., n$, $u \leq x_i, i = 1, 2, ..., n$.

Example 2 : Slack Variables, Relaxation and Dual Norms

$p^* = min_{x \in \mathbb{R}^n} ||Ax - y||_1 + \mu||x||_2$, $A \in \mathbb{R}^{m,n}$ and $y \in \mathbb{R}^m, \mu > 0$.

Use a slack variable $z \in \mathbb{R}^m$ that is elementwise bigger than or equal to elements of the absolute value of $Ax - y$.

Use a slack variable $t \in \mathbb{R}$ that is bigger than or equal to $||x||_2$.

Relaxation in the **Feasible Region**. Why??

$min_{x,z,t} z^T 1 + \mu t$ s.t. $|(Ax)_i - y_i| \leq z_i$, $i = 1, 2, ..., m$ and $||x||_2 \leq t$.

Hint : use the dual norm : $||z||_2 = max_{u:||u||_2 \leq 1} u^T z$ and $||z||_1 = max_{u:||u||_\infty \leq 1} u^T z$

$||Ax - y||_1 + \mu||x||_2 = max_{u:||u||_\infty \leq 1}\{u^T(Ax - y) + \mu \cdot max_{v:||v||_2 \leq 1} v^T x\}$.

$p^* = min_x max_{u,v:||u||_\infty \leq 1, ||v||_2 \leq 1}\{u^T(Ax - y) + \mu v^T x\}$.

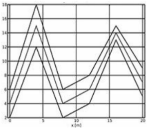
$d^* = max_{u,v:||u||_\infty \leq 1, ||v||_2 \leq 1} min_x\{u^T(Ax - y) + \mu v^T x\}$

Optimization Problem Solving Skills : Slack Variables and Relaxation

Example 3 : Slack Variables and Relaxation

Find the path that minimizes the total length of the path (Form this problem as an SOCP).

Exercise 9.2 (A slalom problem) A two-dimensional skier must slalom down a slope, by going through n parallel gates of known position (x_i, y_i), and of width c_i, $i = 1, \dots, n$. The initial position (x_0, y_0) is given, as well as the final one, (x_{n+1}, y_{n+1}). Here, the x-axis represents the direction down the slope, from left to right, see Figure 9.24.



Problem from Optimization Models (Calafiore El Ghaoui).

Example 4 : Slack Variables and Relaxation

Warehouse Location Problem : minimize the maximum distance (Form this problem as an SOCP).

$min_x max||x - y_i||_2, i = 1, 2, ..., m.$

Using slack variable relaxation, $min_{x,t} t$ s.t. $||x - y_i||_2 \leq t, i = 1, 2, ..., m.$

The Lagrangian

✓ If the original problem is easy to solve, find solution directly. Unless,

✓ Find a **Convex optimization** primal problem to another **dual** problem, wanting that the **Dual** is easier to solve! Actually, can dualize non-convex problems to make a convex dual, but not dealt here.

✓ Formulate a function called **The Lagrangian** that integrates all the **constraints** into an **unconstrained problem**.

$L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ having $dom(L) = D \times \mathbb{R}^m \times \mathbb{R}^p$, $L(x, \lambda, \nu) := f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_j h_j(x)$
1) **L is convex in x.**
2) **L is affine in λ, ν.**

Key Idea behind the Lagrangian : Pay infinite price for disobeying the constrints.

$p* = inf_x[f_0(x) + \sum_{i=1}^m I_{\mathbb{R}_-}(f_i(x)) + \sum_{j=1}^p I_{\{0\}}(h_j(x))]$:**Pay infinite price for disobeying the constrints**

Then, use indicator functions : $I_{\mathbb{R}_-}(f_i(x)) = sup_{\lambda_i \geq 0}\{\lambda_i f_i(x)\}$ and $I_{\{0\}}(h_j(x)) = sup_{\nu_j \in \mathbb{R}}\{\nu_j h_j(x)\}$

$\rightarrow p* = inf_x[f_0(x) + \sum_{i=1}^m sup_{\lambda_i \geq 0}\{\lambda_i f_i(x)\} + \sum_{j=1}^p sup_{\nu_j \in \mathbb{R}}\{\nu_j h_j(x)\}$

$\rightarrow p* = inf_x sup_{\lambda \geq 0, \nu}[f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{j=1}^p \nu_j h_j(x)]$

$\rightarrow p* = inf_x sup_{\lambda \geq 0, \nu} L(x, \lambda, \nu)$

Lagrangian Duality

$p* = min_x max_{\lambda \geq 0, \nu} [f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x)]$

The Lagrange Dual Function

$g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}, g(\lambda, \nu) := inf_{x \in D} L(x, \lambda, \nu)$

g is a **Concave Extended Real valued Function** possibly taking $-\infty$ as a function value or a function that is ∞ everywhere.

Why? Note that L is **affine : concave and convex** in λ, ν and **infimum over concave functions is concave**.

So what? In most cases, g has a global max! (examples we deal with has).

Dual Optimization Problem

If g is not an everywhere ∞ function, the **Dual Problem**

$d^* := sup_{\lambda \succeq 0, \nu} g(\lambda, \nu) = max_{\lambda \succeq 0, \nu} min_x L(x, \lambda, \nu)$
$\leftrightarrow$ a convex problem : $-d^* := inf_{\lambda \succeq 0, \nu} \{-g(\lambda, \nu)\}$ a.k.a $inf_{\lambda, \nu} \{-g(\lambda, \nu)\}$ s.t. $\lambda \succeq 0$

✓ Always, $d^* \leq p^*$: **weak duality**

✓ Under "good" conditions, $d^* = p^*$: **strong duality**

If $d^* = p^*$: **strong duality** and let x^* be a **primal optimal point** and let (λ^*, ν^*) be a **dual optimal point**.

Then, $f_0(x^*) = g(\lambda^*, \nu^*) = inf_{x \in D} L(x, \lambda^*, \nu^*)$

Slater's Condition for Strong Duality

1. Basic Version of the Slater's Condition

✓ A **Convex Problem** is given.
✓ If $\exists x \in relint(D)$ s.t. $f_i(x) < 0$, for $i = 1, 2, ..., m$ and $h_j(x) = 0$, for $j = 1, 2, ..., p$,

then $d^* = p^*$.

Moreover, if $p^* > -\infty$, $\exists$ dual optimal point (λ^*, ν^*).

2. Stronger Version of the Slater's Condition

✓ A **Convex Problem** is given.
✓ If $\exists x \in relint(D)$ s.t. $f_i(x) \leq 0$, for f_i : affine, $f_i(x) < 0$, for f_i not affine, and $h_j(x) = 0$, for $j = 1, 2, ..., p$,

then $d^* = p^*$.

Moreover, if $p^* > -\infty$, $\exists$ dual optimal point (λ^*, ν^*).

Relative Interior

✓$relint(S) := \{x \in S | \exists \epsilon > 0 : B_\epsilon(x) \cap aff(S)\}$: Interior of S as a subset of its affine hull.
OK to just regard this just as the interior of the domain at this level.
✓ What is the $relint(\{(1.5, 2), (3, 1)\})$?

Slater's Condition : Proof

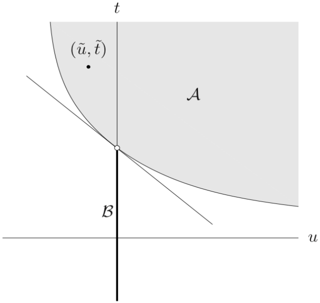


Figure 5.6 Illustration of strong duality proof, for a convex problem that satisfies Slater's constraint qualification. The set $\mathcal{A}$ is shown shaded, and the set $\mathcal{B}$ is the thick vertical line segment, not including the point $(0, p^\star)$, shown as a small open circle. The two sets are convex and do not intersect, so they can be separated by a hyperplane. Slater's constraint qualification guarantees that any separating hyperplane must be nonvertical, since it must pass to the left of the point $(\tilde{u}, \tilde{t}) = (f_1(\tilde{x}), f_0(\tilde{x}))$, where $\tilde{x}$ is strictly feasible.

Figure: from Boyd and Vandenburghe

Slater's Condition : Proof

KKT Conditions, Necessity

✓ Assume strong duality holds : $p^* = d^*$.

✓ $\exists$ Primal Optimal Point x^*

✓ $\exists$ Dual Optimal Point (λ^*, ν^*).

✓ $f_0, f_1, ..., f_i, h_1, ..., h_p$ are all differentiable.

Then, 4 conditions are satisfied.

1. Primal Feasibility of x^* : $f_i(x^*) \leq 0, h_j(x^*) = 0, i = 1, ..., m, j = 1, ..., p.$
2. Dual Feasibility of (λ^*, ν^*) : $\lambda_i^* \geq 0$
3. Complementary Slackness : $\lambda_i^*(f_i(x^*) = 0$
4. Lagrangian Stationarity : $\nabla_x L(x, \lambda^*, \nu^*)|_{x=x^*} = 0$

KKT Conditions : Sufficiency

✓ Assume the problem is convex.

✓ $f_0, f_1, ..., f_m, h_1, ..., h_p$ are all differentiable.

✓ x^*, λ^*, ν^* satisfy KKT conditions (primal feasible, dual feasible, complementary slackness, lagrangian stationarity)

Then, 3 conditions are satisfied.

1. x^* is primal optimal
2. (λ^*, ν^*) : dual optimal
3. $p^* = d^*$

Understanding of Complementary Slackness and Lagrangian Stationarity

Steps to arrive at p^* using the dual

Duality, KKT Conditions Examples

Example 1) Least Squares with Equality Constraints

✓ For the model matrix $X \in \mathbb{R}^{m,n}$, response vector $y \in \mathbb{R}^m, G \in \mathbb{R}^{p,n}, h \in \mathbb{R}^p$, assume X is full column rank.
Think of this optimization problem $min_{\beta \in \mathbb{R}^n} ||X\beta - y||_2^2$ s.t. $G\beta = h$

$L(x, \nu) = (X\beta - y)^T(X\beta - y) + \nu^T(G\beta - h) = \beta^T X^T X\beta + (G^T\nu - 2X^Ty)^T\beta - \nu^T h + y^T y$
Finding $\beta^* \in \mathbb{R}^n$, $\nu^* \in \mathbb{R}^p$ satisfying KKT conditions,since this problem : convex, β^* : primal optimal and ν^* : dual optimal.

Find such β^*, ν^* by KKT conditions.

✓ 1) Primal Feasibility : $G\beta^* = h$
✓ 2) Dual Feasibility : Just $\exists \nu^*$, no condition on ν^*
✓ 3) Complementary Slackness : No, because there is no inequality constraint
✓ 4) Lagrangian Stationarity : $2X^TX\beta^* + G^T\nu^* - 2X^Ty = 0$

✓ Use Lagrangian Stationarity along with $rank(X) = n \to \beta^* = (X^TX)^{-1}(X^Ty - 0.5G^T\nu^*)$

✓ Put this β^* into primal feasibility $\to \nu^* = -2(G(X^TX)^{-1}G^T)^{-1}(h - G(X^TX)^{-1}X^Ty)$.

✓ Put ν^* again into Lagrangian Stationarity $\to \beta^* = (X^TX)^{-1}(X^Ty + G^T(G(X^TX)^{-1}G^T)^{-1}(h - G(X^TX)^{-1}X^Ty))$

Duality, KKT Conditions Examples

Example 2) Entropy in Information Theory
Concepts of Code, Codeword, Prefix Free
✓ Want to find the shortest expected message length among **prefix free codes**.
$min_{l \in \mathbb{N}^n} p^T l$ s.t. $\sum_{i=1}^n 2^{-l_i} \leq 1$. Approximate this problem into :
$min_{l \in \mathbb{R}_+^n} p^T l$ s.t. $\sum_{i=1}^n 2^{-l_i} \leq 1 \leftrightarrow min_{l \in \mathbb{R}^n p^T l}$ s.t. $\sum_{i=1}^n 2^{-l_i} \leq 1$ using $2^{-negative} > 1$

Prove that $p^* = \sum_{i=1}^n p_i log_2 \frac{1}{p_i}$, which is **the entropy of the probability distribution p**.
✓ If I find l^*, a primal feasible point, λ^*, a dual feasible point, using the convexity of this problem, l^* : primal optimal, λ^* : dual optimal.

Find these by KKT conditions :
✓ 1) Primal Feasibility : $\sum_{i=1} n 2^{-l_i^*} \leq 1$
✓ 2) Dual Feasibility : $\lambda^* \geq 0$. Note that λ^* is a scalar.
✓ 3) Complementary Slackness : $\lambda^* \sum_{i=1}^n (2^{-l_i^*} - 1) = 0$
✓ 4) Lagrangian Stationarity : $p_i - ln2 \cdot \lambda^* \cdot 2^{-l_i^*} = 0, i \in \{1, 2, ..., n\}$

Sum over indices i in Lagrangian Stationarity $\rightarrow ln2 \cdot \lambda^* \cdot \sum_{i=1}^n 2^{-l_i^*} = 1$
✓ Use the Complementary Slackness : $\lambda^* = \lambda^* \sum_{i=1}^n 2^{-l_i^*} \rightarrow \sum_{i=1}^n 2^{-l_i^*} = 1$
✓ Come back to the Lagrangian Stationarity $\rightarrow l_i^* = log_2 \frac{1}{p_i}$

✓ Soon I get $p^* = p^T l^*$ (vector notation) $= \sum_{i=1}^n p_i log_2 \frac{1}{p_i}$

Duality, KKT Conditions Examples

Example 3) Support Vector Machines

✓ 1st) A hyperplane in $\mathbb{R}^n$ $H = \{x | w^T x + b = 0\}$. What is the distance from $x_0 \in \mathbb{R}^n$ to H?

✓ 2nd) Let $\{X_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ be train data points. $X_i \in \mathbb{R}^n$ and $y_i \in \{1, -1\}$. Then, **hard margin SVM** : Find H, or w, α where H satisfies 1. perfectly separating two classes, 2. all points are at least m distance away from H.

✓ 3rd) But, the optimization problem you've set in 2nd) has infinite number of solutions!

✓ 4th) Prove that the problem in 2nd) can be expressed as $min_{w,\alpha} \frac{1}{2}||w||_2^2$ s.t. $y_i(w^T x_i + b) \geq 1$, $i = 1, 2, ..., n$. Identify the relationship between m and w.

✓ 5th) Show that the problem in 4th) is convex. So what?

✓ 6th) Show that w^*, α^*, λ^* satisfy $\lambda_i^*(y_i(w^{*T} x_i + b^*) - 1) = 0$, $\forall i = 1, 2, ..., n$.

(Idea from Akash Velu)