KKT condition and optimality

Jong-June Jeon

October 26, 2023

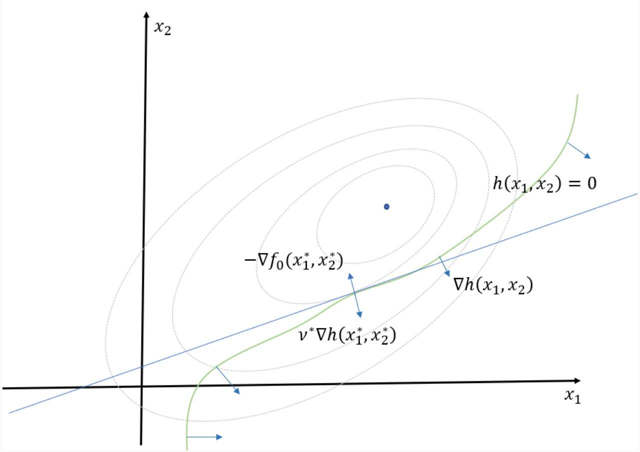
Department of Statistics, University of Seoul

Lagrangian

Consider a convex optimization problem with an equality constraint:

$$\begin{aligned} \min \quad & f_0(x_1, x_2) \\ \text{subject to} \quad & h(x_1, x_2) = 0 \end{aligned}$$

Recall that $-\nabla f_0(x_1, x_2)$ is the directional derivative with which $f_0(x_1, x_2)$ decreases most rapidly. Also $\nabla f_0(x_1, x_2)$ is orthogonal to the tangent line at (x_1, x_2).



If (x_1^*, x_2^*) is the optimal point, there exists a scalar $\nu^* \in \mathbb{R}$ such that

$$\nabla f_0(x_1^*, x_2^*) + \nu^* \nabla h(x_1^*, x_2^*) = 0 \tag{1}$$

If ν^* is known,

$$\min_{x_1, x_2} \; f_0(x_1, x_2) + \nu^* h(x_1, x_2),$$

gives the solution of the constrained problem. Note that $\nabla h(x_1^*, x_2^*)$ is proportional to $\nabla f_0(x_1^*, x_2^*)$.

Thus, we can apply an unconstrained optimization algorithm to

$$f_0(x_1, x_2) + \nu^* h(x_1, x_2).$$

Optimization problem

- Consider the following optimization problem in the standard form.

$$\begin{aligned} \text{min} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \dots, m \\ & h_j(x) = 0, \quad j = 1, \dots, p \end{aligned} \tag{2}$$

- Denote the domain of optimization problem by

$$\mathcal{D} = \left(\cap_{i=0}^{m} \mathsf{dom}(f_i) \right) \bigcap \left(\cap_{j=1}^{p} \mathsf{dom}(h_j) \right)$$

Lagrangian

- Define the Largrangian $L : \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associated with the problem (5):

$$L(x, \lambda, \nu) = \begin{cases} f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{j=1}^p \nu_j h_j(x), & \text{for } \lambda \succeq 0 \\ -\infty, & \text{otherwise,} \end{cases}$$

where $\lambda = (\lambda_1, \cdots, \lambda_m)^\top$ and $\nu = (\nu_1, \cdots, \nu_p)^\top$.

Example 1 (Linear programming)

$$
\begin{aligned}
\min \quad & 12x_1 + 16x_2 \\
\text{subject to} \quad & -x_1 - 2x_2 \leq -40 \\
& -x_1 - x_2 \leq 30 \\
& -x_1 \leq 0, \ -x_2 \leq 0
\end{aligned}
$$

More generally, we consider

$$
\begin{aligned}
\min \quad & c^\top x \\
\text{subject to} \quad & Ax - b \leq 0
\end{aligned}
$$

In our example, $c = (12, 16)^\top$, $b = (-40, 30, 0, 0)^\top$ and

$$A = \begin{pmatrix} -1 & -2 \\ -1 & -2 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}$$

$L(x, \lambda) = c^\top x + \lambda^\top (Ax - b)$.

Example 2 (Quadratic function)

$$\begin{aligned} \min \qquad & f_0(x_1, x_2) = \frac{1}{2}(x_1 - 2)^2 + 2(x_2 - 1)^2 + (x_1 - 2)(x_2 - 1) \\ \text{subject to} \qquad & f_1(x_1, x_2) = x_1^2 + x_2^2 - 0.5 \leq 0 \end{aligned}$$

The Lagrangian function is given by

$$L(x_1, x_2, \lambda) = \frac{1}{2}x_1^2 + 2x_2^2 + x_1 x_2 - 3x_1 - 6x_2 + 6 + \lambda(x_1^2 + x_2^2 - 0.5)$$

if $\lambda \geq 0$.

Example 3 (Ridge regression)

Let $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^p$ for $i = 1, \cdots, n$ be response-predict pairs and $Y = (y_1, \cdots, y_n)^\top$ and $X = (x_1, \cdots, x_n)^\top \in \mathbb{R}^{n \times p}$.

$$
\begin{aligned}
\min \quad & \frac{1}{2n}\|Y - X\beta\|^2 \\
\text{subject to} \quad & \|\beta\|^2 - C \leq 0 \\
& \mathbf{1}^\top \beta = 0
\end{aligned}
$$

The Lagrangian function is

$$
L(\beta, \lambda, \nu) = \frac{1}{2n}\|Y - X\beta\|^2 + \lambda\left(\|\beta\|^2 - C\right) + \nu(\mathbf{1}^\top \beta).
$$

if $\lambda \geq 0$

Lagrange dual function

- Define the Lagrange dual function $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$

$$\begin{aligned} g(\lambda, \nu) &= \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \\ &= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x) \right). \qquad (3) \end{aligned}$$

- (λ, ν) is called dual variable or Lagrange multiplier vector and $\{(\lambda, \nu) : \lambda \succeq 0,\ \nu \in \mathbb{R}^p\}$ is called dual feasible set.

Example 4 (Dual function in Example 1)

Let $L(x, \lambda) = c^\top x + \lambda^\top (Ax - b)$.

$$g(\lambda) = \inf_x (c + A^\top \lambda)^\top x - \lambda^\top b = \begin{cases} -\lambda^\top b & \text{if } A^\top \lambda + c = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

The Lagrangian dual is finite only when $A^\top \lambda + c = 0$.

Example 5 (Dual function in Example 2)

The Lagrangian function is given by

$$L(x_1, x_2, \lambda) = \frac{1}{2}x_1^2 + 2x_2^2 + x_1x_2 - 3x_1 - 6x_2 + \lambda(x_1^2 + x_2^2 - 0.5).$$

For a fixed λ $L(x_1, x_2, \lambda)$ is a quadratic function such that $\inf_{x_1,x_2} L(x_1, x_2, \lambda)$ is easily computed.

$$\begin{aligned}
\frac{\partial L(x_1, x_2, \lambda)}{\partial x_1} &= (1 + 2\lambda)\, x_1 + x_2 - 3 = 0 \\
\frac{\partial L(x_1, x_2, \lambda)}{\partial x_2} &= (4 + 2\lambda)\, x_2 + x_1 - 6 = 0
\end{aligned}$$

$$\begin{pmatrix} 1+2\lambda & 2 \\ 1 & 4+2\lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \end{pmatrix}$$

Thus,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1+2\lambda & 2 \\ 1 & 4+2\lambda \end{pmatrix}^{-1} \begin{pmatrix} 3 \\ 6 \end{pmatrix} = \begin{pmatrix} 6\lambda/(4\lambda^2+10\lambda+2) \\ (12\lambda-3)/(4\lambda^2+10\lambda+2) \end{pmatrix}$$

Since $L(x_1, x_2, \lambda)$ with $\lambda \geq 0$ is strictly convex, the solution is the minimizer of $L(x_1, x_2, \lambda)$ for the fixed λ.

By plugging $x_1 = 6\lambda/(4\lambda^2 + 10\lambda + 2)$ and $x_2 = (12\lambda - 3)/(4\lambda^2 + 10\lambda + 2)$ into

$$L(x_1, x_2, \lambda) = \frac{1}{2}x_1^2 + 2x_2^2 + x_1x_2 - 3x_1 - 6x_2 + \lambda(x_1^2 + x_2^2 - 0.5),$$

its dual function g is obtained. (note that g is a function of λ.)

Example 6 (Ridge regression)

The Lagrangian function is

$$\begin{aligned} L(\beta,\lambda,\nu) &= \frac{1}{2n}\|Y - X\beta\|^2 + \lambda\left(\|\beta\|^2 - C\right) + \nu(\mathbf{1}^\top\beta) \\ &= \frac{1}{2}\beta^\top\left(\frac{X^\top X}{n} + 2\lambda I\right)\beta - \left(\frac{X^\top Y}{n} - \nu\mathbf{1}\right)^\top\beta + \frac{1}{2n}Y^\top Y - \lambda C \end{aligned}$$

For a fixed β and ν, $L(\beta,\lambda,\nu)$ is minimized if

$$\beta = \left(\frac{X^\top X}{n} + 2\lambda I\right)^{-1}\left(\frac{X^\top Y}{n} - \nu\mathbf{1}\right).$$

If $\lambda > 0$ then $\left(\frac{X^\top X}{n} + 2\lambda I\right)^{-1}$ exists.

(Continue with the example)

Thus, the dual function is given by

$$
\begin{aligned}
g(\lambda, \nu) \quad = \quad & -\frac{1}{2}\left(\frac{X^\top Y}{n} - \nu \mathbf{1}\right)^\top \left(\frac{X^\top X}{n} + 2\lambda I\right)^{-1} \left(\frac{X^\top Y}{n} - \nu \mathbf{1}\right) \\
& -C\lambda + \frac{1}{2n} Y^\top Y.
\end{aligned}
$$

When $X^\top X/n = I$, the dual function is simply written. Denote $X^\top Y/n$ by r. Then,

$$
g(\lambda, \nu) = -\frac{1}{(1+2\lambda)}\left(n\nu^2 - 2(r^\top \mathbf{1})\nu + r^\top r\right) - C\lambda + \frac{1}{2n} Y^\top Y
$$

Proposition 1 (convexity of dual function)

The dual function is always concave even when all functions in (5) are nonconvex.

(proof) For convenience of notations, let $m = p = 1$.

$$
\begin{aligned}
& g(t\lambda + (1-t)\tilde{\lambda}, t\nu + (1-t)\tilde{\nu}) \\
= \;& \inf_x \left(\underbrace{f_0(x)}_{=tf_0(x)+(1-t)f_0(x)} + \left(t\lambda + (1-t)\tilde{\lambda}\right) f_1(x) + (t\nu + (1-t)\tilde{\nu})h_1(x) \right) \\
\geq \;& t \inf_x \left(f_0(x) + \lambda f_1(x) + \nu h_1(x)\right) + (1-t) \inf_x \left(f_0(x) + \tilde{\lambda} f_1(x) + \tilde{\nu} h_1(x) \right) \\
= \;& tg(\lambda, \nu) + (1-t)g(\tilde{\lambda}, \tilde{\nu}).
\end{aligned}
$$

Proposition 2 (feasibility)

If x is feasible then $f_0(x) \geq L(x, \lambda, \nu)$.

(proof) If x is feasible then $f_i(x) \leq 0$ for $i = 1, \cdots, m$ and $h_j(x) = 0$ for $j = 1, \cdots, p$. Since $\lambda_i \geq 0$ and $\nu_j \in \mathbb{R}$, $\lambda_i f_i(x) \leq 0$ and $\nu_j h_j(x) = 0$. Thus,

$$f_0(x) \geq f_0(x) + \sum_{i=1}^{n} \lambda_i f_i(x) + \sum_{j=1}^{m} \nu_j h_j(x).$$

Proposition 3 (Weak Duality)

The lower bounds on the optimal value p^* of the problem (5) is attained by the Lagrange dual function as

$$g(\lambda, \nu) \leq p^* \tag{4}$$

for any $\lambda \geq 0$ and any ν.

(proof)

Let $\tilde{x}$ be an arbitrary feasible point for the (5). Then, $f_i(\tilde{x}) \leq 0$ for $\forall i$ and $h_j(\tilde{x}) = 0$ for $\forall j$ such that

$$\sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{j=1}^{p} \nu_j h_j(\tilde{x}) \leq 0$$

for $\lambda \geq 0$. $L(\tilde{x}, \lambda, \nu) \leq f_0(\tilde{x})$ implies that

$$f_0(\tilde{x}) \geq L(\tilde{x}, \lambda, \nu) \geq \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = g(\lambda, \nu).$$

Since $g(\lambda, \nu)$ does not depend on $\tilde{x}$, we can conclude that

$$\inf_{\tilde{x} \in \mathcal{D}} f_0(\tilde{x}) \geq g(\lambda, \nu),$$

Interpretation

- Consider the following optimization problem in the standard form.

$$\begin{aligned} \text{min} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_j(x) = 0, \quad j = 1, \ldots, p \end{aligned} \tag{5}$$

- Unconstrained form:

$$\min \qquad f_0(x) + \sum_{i=1}^{m} I_-(f_i(x)) + \sum_{j=1}^{p} I_0(h_j(x))$$

where $I_-(u) = 0$ if $u \leq 0$ and ∞ otherwise and $I_0(u) = 0$ if $u = 0$ and ∞ otherwise. Thus, $\lambda_i f_i(x)$ and $\nu_j h_j(x)$ are linear approximation of $I_-(f_i(x))$ and $I_0(h_j(x))$.

Conjugate function

Recall that the conjugate f^* of a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is given by

$$f^*(y) = \sup_{x \in \mathsf{dom}(f)} y^\top x - f(x)$$

(See slide 2 of the convex function)

Proposition 4 (Dual function)

Consider an optimization problem,

$$\begin{aligned} \min \quad & f_0(x) \\ \textit{subject to} \quad & Ax \preceq b \\ & Cx = d \end{aligned}$$

Then, the dual function is

$$g(\lambda, \nu) = -b^\top \lambda - d^\top \nu - f_0^*(-A^\top \lambda - C^\top \nu)$$

and $\textit{dom}(g) = \{(\lambda, \nu) : -A^\top \lambda - C^\top \nu \in \textit{dom}(f_0^*)\}$

proof) See p221.

Example 7 (Equality constrained norm minimization)

$$\begin{aligned} \min \quad & \|x\| \\ \text{subject to} \quad & Ax = b, \end{aligned}$$

where $\|\cdot\|$ is any norm. Then,

$$g(\nu) = -b^\top \nu - f_0(-A^\top \nu) = \begin{cases} -b^\top \nu & \text{if } \|A^\top \nu\|_* \le 1 \\ -\infty & \text{otherwise.} \end{cases}$$

(Continue with the example)

Definition 8 (Dual norm)

Let $x \in X = \mathbb{R}^n$ and $\|\cdot\|$ be a norm of $\mathbb{R}^n$. Let $L : X \mapsto \mathbb{R}$ be a linear functional. Denote a collection of all Ls by X^*. Dual norm is defined by

$$\|L\|_* = \sum\{|Lx| : \|x\| \leq 1, x \in X\}$$

Consider the Euclidean norm on X. Let a linear functional indexed by $y \in X$ be

$$L_y : x \in X \mapsto y^\top x.$$

Then the dual norm of y (actually a linear function L_y) is define by $\|y\|_* = \sup\{y^\top x : \|x\| \leq 1\}$.

Proposition 5 (Conjugate function of norm)

The conjugate function of $f(x) = \|x\|$ is

$$f^*(y) = \begin{cases} 0 & \|y\|_* \le 1 \\ \infty & \textit{otherwise} \end{cases}$$

Suppose that $\|y\|_* > 1$. Then there exists z such that $y^\top z > 1$ and $\|z\| \le 1$. Let $x = tz$ for $t > 0$ then

$$f^*(y) \ge y^\top x - \|x\| = t(y^\top z - \|z\|) \to \infty \text{ as } t \to \infty.$$

If $\|y\|_* \le 1$ then $y^\top x \le \|x\|\|y\|_* \le \|x\|$. That is, $y^\top x - \|x\| \le 0$. Then $f^*(y) = 0$.

(proof of example) By Proposition 4,

$$g(\nu) = -b^\top \nu - f_0^*(-A^\top \nu) = \begin{cases} -b^\top \nu & \|A^\top \nu\| \leq 1 \\ -\infty & \text{otherwise} \end{cases}$$

Lagrangian dual problem

- Define the Lagrange dual problem associated with the problem (5)

$$\begin{aligned} \text{max} \quad & g(\lambda, \nu) \\ \text{subject to} \quad & \lambda \geq 0 \end{aligned} \tag{6}$$

- (6) is convex optimization problem although the primal problem (5) is not convex.
- Denote dual optimal or optimal Lagrange multipliers by (λ^*, ν^*).

Definition 9 (Duality Gap)

- Let the optimal value of the Lagrange dual problem (6) be d^*
- By (4), it is known that

$$d^* \leq p^*. \tag{7}$$

- The optimal duality gap $(p^* - d^*)$ is always nonnegative.

Strong duality

- Suppose that there exist x^* and (λ^*, ν^*) such that

$$f_0(x^*) = g(\lambda^*, \nu^*). \tag{8}$$

  - For any (λ, ν), $g(\lambda, \nu)$ is lower bound of the primal optimal value. That is, the equality means that $f(x^*)$ achieves the primal optimal value and x^* is primal solution.
  - In the view of the dual function $g(\lambda, \nu)$ the $f(x^*)$ is the upper bound of the function. Hence, (λ^*, ν^*) is a dual solution.
- In conclusion, the solution satisfying (8) implies that they are primal and dual optimal solution.

Definition 10 (Strong duality)

If duality gap is zero, the we call this condition the strong duality.

- The strong duality is crucial condition under which the primal problem could be solved by the dual problem. (Note that the dual problem sometimes is easier to solve that the primal problem.)
- Hence, the sufficient conditions for the strong duality is very useful.

Theorem 11 (Strong duality under Slater's condition)

Assume that $f_i(x)$ for $i = 1, \cdots, m$ are convex and $h_j(x)$ for $j = 1, \cdots, p$ are linear. Let A be $p \times n$ matrix corresponding to coefficients of $h_j(x)$'s. If the rank of A is p and there exists an x such that

$$f_i(x) < 0,\ i = 1, \ldots, m, \quad h_j(x) = 0,\ j = 1, \cdots, p,$$

then strong duality holds.

proof) See the appendix.

Example 12 (Linear programming: standard form)

$$
\begin{aligned}
\min \quad & c^\top x \\
\text{subject to} \quad & Ax = b \\
& x \preceq 0
\end{aligned}
$$

The Lagrangian is

$$L(x, \lambda, \nu) = c^\top x - \lambda^\top x + \nu^\top (Ax - b).$$

$$L(x, \lambda, \nu) = (c - \lambda + A^\top \nu)^\top x - b^\top \nu$$

If $c - \lambda + A^\top \nu = 0$ then $\inf_x L = -b^\top \nu$. Thus, $g(\lambda, \nu) = -\nu^\top b$. If the problem has a feasible solution, the strong duality holds by Slater's condition. The dual problem is written by

$$\begin{aligned} \max \quad & -b^\top \nu \\ \text{subject to} \quad & A^\top \nu + c = \lambda \\ & \lambda \succeq 0. \end{aligned}$$

Reparametrization $y = -\nu$ leads to

$$
\begin{aligned}
\max \quad & b^\top y \\
\text{subject to} \quad & A^\top y + \lambda = c \\
& \lambda \succeq 0.
\end{aligned}
$$

Example 13 (Linear programming: profit maximization)

$$
\begin{array}{rcl}
\max & c^\top x & \text{(profit)} \\
\text{subject to} & Ax - b \leq 0 & \text{(resource constraints)} \\
& x \preceq 0 & \text{(nonnegative production)}
\end{array}
$$

An equivalent problem is $\min -c^\top x$ with the same constraints. The Lagrangian is given by

$$L(x, \lambda_1, \lambda_2) = -c^\top x + \lambda_1^\top (Ax - b) - \lambda_2^\top x$$

for $\lambda_1 \succeq 0$ and $\lambda_2 \succeq 0$.

Its dual function is given by

$$g(\lambda_1, \lambda_2) = \begin{cases} -b^\top \lambda_1 & \text{if } A^\top \lambda_1 - \lambda_2 - c = 0 \\ -\infty & \text{otherwise} \end{cases}$$

Thus, the dual problem is written by

$$\begin{aligned} \max \quad & -b^\top \lambda_1 \\ \text{subject to} \quad & A^\top \lambda_1 - \lambda_2 = c \\ & \lambda_1 \succeq 0 \\ & \lambda_2 \succeq 0 \end{aligned}$$

Equivalently, the dual problem is written by

$$\begin{aligned} \min \quad & b^\top \lambda_1 \\ \text{subject to} \quad & A^\top \lambda_1 \succeq c \\ & \lambda_1 \succeq 0. \end{aligned}$$

(See the example of a dual problem in the textbook of the Management of science.)

Theorem 14 (Nonconvex quadratic problem with strong duality)

$$\begin{aligned} \min \quad & x^\top A x + 2b^\top x \\ \textit{subject to} \quad & x^\top x \le 1 \end{aligned}$$

where $A \in \mathcal{S}^n$ and $A \not\succeq 0$, and $b \in \mathbb{R}^n$. This problem has a dual problem with no-gap dual optimality:

$$\begin{aligned} \max \quad & -\sum_{i=1}^{n} (q_i^\top b)^2/(\lambda_i + \lambda) - \lambda \\ \textit{subject to} \quad & \lambda \ge -\lambda_{\min}(A), \end{aligned}$$

where λ_i and q_i are eigenvalues and corresponding eigenvectors of A.

Geometric interpretation

Consider a standard form of convex optimization.

$$\mathcal{G} = \{(f_1(x), \cdots, f_m(x), h_1(x), \cdots, h_p(x), f_0(x)) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}\}$$

The optimal value is given by

$$p^* = \inf\{t : (u, v, t) \in \mathcal{G}, u \preceq 0, h = 0\}.$$

Lagrangian function is written by the terms of dual variables and $(u, v, t) \in \mathcal{G}$

$$(\lambda, \nu, 1)^\top (u, v, t) = \sum_{i=1}^{m} \lambda_i u_i + \sum_{j=1}^{p} \nu_j v_j + t,$$

where $(u, v, t) \in \mathcal{G}$

Dual function is

$$g(\lambda, \nu) = \inf\{(\lambda, \nu, 1)^\top (u, v, t) : (u, v, t) \in \mathcal{G}\}$$

Thus, for $(u, v, t) \in \mathcal{G}$

$$(\lambda, \nu, 1)^\top (u, v, t) \geq g(\lambda, \nu).$$

Here, the inequality can be viewed as the supporting hyperplane for $\mathcal{G}$ ($a^\top x \geq b$ for $\forall x \in C$).

Suppose that $\lambda \succeq 0$. If $u \preceq 0$ and $\nu = 0$ then $t \geq (\lambda, \nu, 1)^\top (u, v, 1)$. Therefore,

$$\begin{aligned}
p^* &= \inf\{t : (u, v, t) \in \mathcal{G}, u \preceq 0, v = 0\} \\
&\geq \inf\{(\lambda, \nu, 1)^\top (u, v, t) : (u, v, t) \in \mathcal{G}, u \preceq 0, v = 0\} \\
&\geq \inf\{(\lambda, \nu, 1)^\top (u, v, t) : (u, v, t) \in \mathcal{G}\} \\
&= g(\lambda, \nu)
\end{aligned}$$

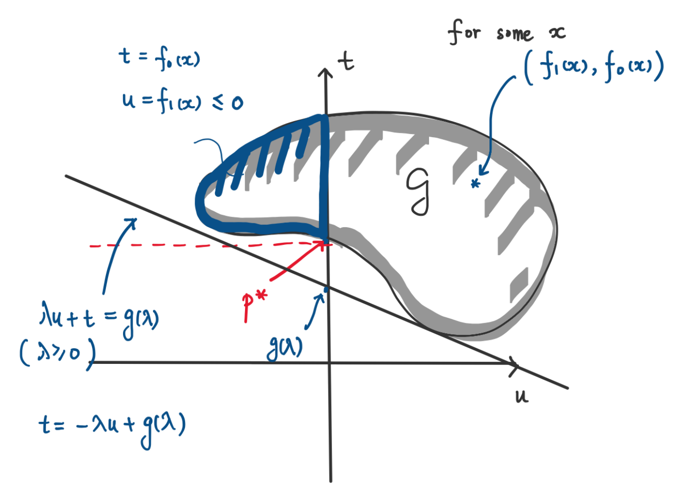


Figure 1: Geometric interpretation of dual function

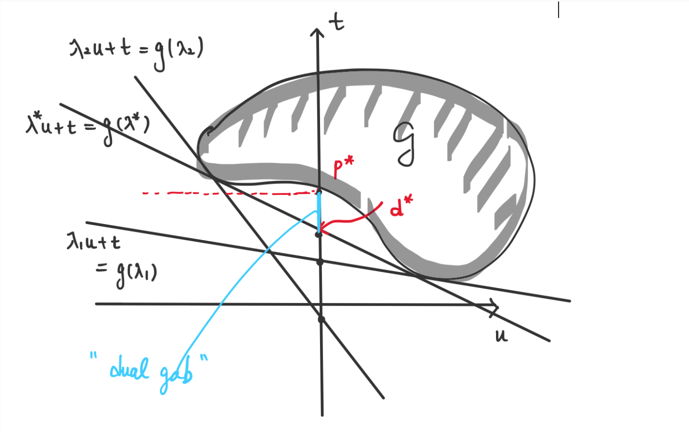


Figure 2: Geometric interpretation of dual gab

Let $\mathcal{A} = \{(u, v, t) : \exists x \in \mathcal{D}, f_i(x) \le u_i, i = 1, \cdots, m, h_i(x) = v_i, i = 1, \cdots, p, f_0(x) \le t\}$
Then, $p^* = \inf\{t : (0, 0, t) \in \mathcal{A}\}$

For $\lambda \succeq 0$,

$$g(\lambda, \nu) = \inf\{(\lambda, \nu, 1)^\top (u, v, t) : (u, v, t) \in \mathcal{A}\},$$

because $(\lambda, \nu, 1)^\top (u, v, t)$ is affine function.

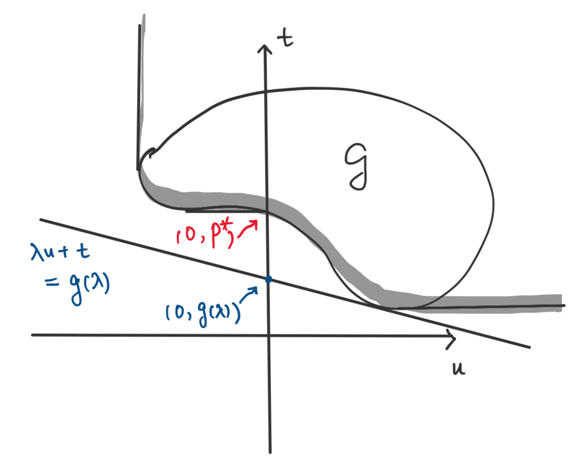


Figure 3: Shaded region denotes $\mathcal{A}$

For $(u, v, t) \in \mathcal{A}$,

$$(\lambda, \nu, 1)^\top (u, v, t) \geq g(\lambda, \nu)$$

Since $(0, 0, p^*) \in bd(\mathcal{A})$, we have $p^* = (\lambda, \nu, 1)^\top (0, 0, p^*) \geq g(\lambda, \nu)$. (weak duality) If there exists $(\lambda^*, \nu^*, 1)$ such that $p^* = (\lambda^*, \nu^*, 1)^\top (0, 0, p^*) = g(\lambda^*, \nu^*)$ then the strong duality holds. (the existence of a nonvertical support hyperplane to $\mathcal{A}$ and its boundary point $(0, 0, p^*)$

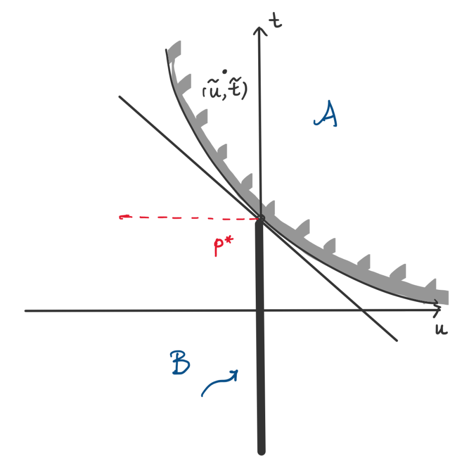


Figure 4: Strong duality

KKT optimality conditions

Sufficiency

- Assume that the functions $f_0, \dots, f_m, h_1, \dots, h_p$ are differentiable. f_is are convex and h_js are affine.
- KKT conditions

$$
\begin{aligned}
f_i(x^*) &\leq 0, \qquad & i = 1, \dots, m \\
h_j(x^*) &= 0, \qquad & j = 1, \dots, p \\
\lambda_i^* &\geq 0, \qquad & i = 1, \dots, m \\
\lambda_i^* f_i(x^*) &= 0, \qquad & i = 1, \dots, m
\end{aligned}
$$

$$
\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{j=1}^{p} \nu_j^* \nabla h_j(x^*) = 0,
$$

- For convex optimization problem, if there exists x^* and (λ^*, ν^*) satisfying the KKT condition, then x^* and (λ^*, ν^*) are primal and dual optimal solution, respectively.

Sufficiency

proof)

$$f_0(x^*) = f_0(x^*) + \sum_{i=1}^{n} \lambda_i^* f_i(x^*) + \sum_{j=1}^{p} \nu_j^* h_j(x^*) \geq g(\lambda^*, \nu^*)$$

Since $L(x, \lambda^*, \nu^*) = f_0(x) + \sum_{i=1}^{n} \lambda_i^* f_i(x) + \sum_{j=1}^{p} \nu_j^* h_j(x)$ is convex function of x, x^* is the minimizer of $L(x, \lambda^*, \nu^*)$ and

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

That is, the proof is completed by (8).

Necessity

Assume that the strong duality holds. If x^* and (λ^*, ν^*) are primal and dual solution of (5). Then these solutions satisfy the KKT conditions.

proof)

$$
\begin{aligned}
f_0(x^*) &= g(\lambda^*, \nu^*) \\
&= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{j=1}^{p} \nu_j^* h_j(x) \right) \\
&\leq f_0(x^*) + \sum_{i=1}^{m} \underbrace{\lambda_i^* f_i(x^*)}_{\leq 0} + \sum_{j=1}^{p} \underbrace{\nu_j^* h_j(x^*)}_{=0} \\
&\leq f_0(x^*)
\end{aligned}
$$

Necessity

From the above inequality we know that

$$f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{j=1}^{p} \nu_j^* h_j(x^*) = \inf_{x \in \mathcal{D}} L(x, \lambda^*, \nu^*).$$

That is, x^* is a minimizer of $L(x, \lambda^*, \nu^*)$ such that

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{j=1}^{p} \nu_j^* \nabla h_j(x^*) = 0.$$

KKT conditions

- Primal feasibility: $f_i(x^*) \leq 0,\ i = 1, \ldots, m$, $h_j(x^*) = 0, j = 1, \ldots, p$.
- Dual feasibility: $\lambda_i^* \geq 0, i = 1, \ldots, m$.
- Complementary slackness : $\lambda_i^* f_i(x^*) = 0, i = 1, \ldots, m$.
- Stationarity: $\nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{j=1}^p \nu_j^* \nabla h_j(x^*) = 0,$

KKT conditions

- If we find the (x^*, λ^*, ν^*) satisfying the KKT conditions, then they are the primal and dual optimal solutions by sufficiency.
- If the Slater's condition holds, the optimal solutions necessarily satisfy the KKT conditions. However, without the strong duality, the optimal solutions may not satisfies the KKT conditions.

Example 15 (quadratic optimization)

- Consider the convex problem

$$\begin{aligned} \text{min} \quad & f(x) = \frac{1}{2}x^\top P x + q^\top x + r \\ \text{subject to} \quad & Ax = b \end{aligned}$$

where $P \in S_+^n$, $A \in R^{p \times n}$ (S_+^n : positive semidefinite $n \times n$ matrices)

- The Lagrangian is

$$L(x, \nu) = \frac{1}{2}x^\top P x + q^\top x + r + \nu^\top (Ax - b)$$

- The KKT conditions:

$$Ax^* = b, \qquad Px^* + q + A^\top \nu^* = 0,$$

where x^* is the primal optimal and ν is the dual optimal solution.

- The equation

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \nu^* \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}$$

is called the KKT system. By solving the equation, we can obtain the primal and dual solutions.

Appendix

(proof of strong duality under Slater's condition)

- Let $b \in \mathbb{R}^m$ be constant vector in $h_j(x)$s.
- Let p^* be primal optimal value, $\mathbf{u} = (u_1, \cdots, u_m)$ and $\mathbf{v} = (v_1, \cdots, v_p)$.
- Define two subgraphs (epigraphs) as

$$
\begin{aligned}
\mathcal{A} &= \bigcup_{x \in \mathcal{D}} \{(\mathbf{u}, \mathbf{v}, t) : u_i \geq f_i(x), i = 1, \ldots, m, \\
&\quad v_j = h_j(x), j = 1, \ldots, p, t \geq f_0(x)\} \\
\mathcal{B} &= \{(\mathbf{0}, \mathbf{0}, t) \in R^m \times R^p \times R \mid t < p^*\}
\end{aligned}
$$

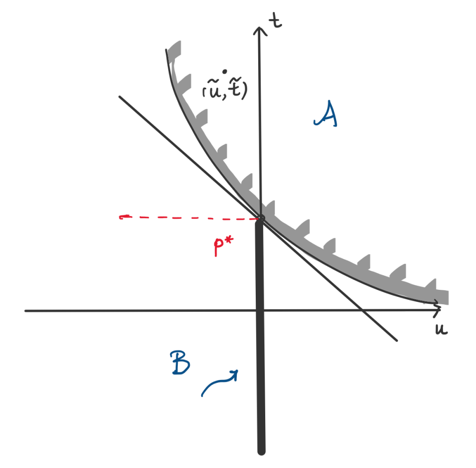


Figure 5: Regions of $\mathcal{A}$ and $\mathcal{B}$

- It can be shown that both $\mathcal{A}$ and $\mathcal{B}$ are convex and $\mathcal{A} \cap \mathcal{B}$ is empty (see p 235.)

- Because $\mathcal{A} \cap \mathcal{B}$ is empty, there is a nonzero vector $(\lambda, \nu, \mu) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}$ and $\alpha \in \mathbb{R}$ such that

$$\lambda^\top \mathbf{u} + \nu^\top \mathbf{v} + \mu t \geq \alpha \text{ for } (\mathbf{u}, \mathbf{v}, t) \in \mathcal{A} \tag{9}$$

$$\lambda^\top \mathbf{u} + \nu^\top \mathbf{v} + \mu t \leq \alpha \text{ for } (\mathbf{u}, \mathbf{v}, t) \in \mathcal{B} \tag{10}$$

by the separating hyperplane theorem. (see p 45)

- Since $\mathbf{u}$ and t is unbounded above, $\lambda \geq \mathbf{0}$ and $\mu \geq 0$ by (9). In addition, because $\mu t \leq \alpha$ for all $t < p^*$ in (10), $\mu p^* \leq \alpha$. Choose an arbitrary $x \in \mathcal{D}$ and let $\mathbf{u} = (f_1(x), \cdots, f_m(x))^\top$, $\mathbf{v} = (h_1(x), \cdots, h_p(x))^\top$ and $t = f_0(x)$ then, $(\mathbf{u}, \mathbf{v}, t) \in \mathcal{A}$ which implies

$$\sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x) + \mu f_0(x) \geq \alpha \geq \mu p^*.$$

- That is, the (λ, ν, μ) obtained by separating hyperplane theorem satisfies

$$\sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x) + \mu f_0(x) \geq \mu p^*.$$

for all $x \in \mathcal{D}$

- First assume that $\mu > 0$ then $x \in \mathcal{D}$

$$f_0(x) + \sum_{i=1}^{m} \frac{\lambda_i}{\mu} f_i(x) + \sum_{j=1}^{p} \frac{\nu_j}{\mu} h_j(x) \geq p^*.$$

  Note that the left hand side is Lagrangian $L(x, \lambda/\mu, \nu/\mu)$

- Minimizing over $x \in \mathcal{D}$ leads to $g(\lambda/\mu, \nu/\mu) \geq p^\star$ such that $g(\lambda/\mu, \nu/\mu) = p^*$ by weak duality.

- Assume that $\mu = 0$, we can obtain for all $x \in \mathcal{D}$

$$\sum_{i=1}^{m} \lambda_i f_i(x) + \nu^\top (Ax - b) \geq 0. \tag{11}$$

Let $\tilde{x}$ be a feasible point satisfying the Slater's condition, then we have $\sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) \geq 0$. Since $f_i(\tilde{x}) < 0$ and $\lambda \geq \mathbf{0}$, we know that $\lambda = \mathbf{0}$.

- From $(\lambda, \nu, \mu) \neq \mathbf{0}$, $\lambda = 0$ and $\mu = 0$, it is known that $\nu \neq 0$ and $\nu^\top(Ax - b) \geq 0$ for all $x \in \mathcal{D}$ from (11). However, the $\tilde{x}$ also satisfies $\nu^\top(A\tilde{x} - b) = 0$, since $\tilde{x} \in$ relint $\mathcal{D}$. However, there exists a point $x \in \mathcal{D}$ such that $\nu^\top(Ax - b) < 0$ unless $A^\top \nu = 0$. This is contradiction of our assumption that rank $A = p$. That is, $\mu = 0$ is impossible.