Lecture 21. Strong duality; Slater's condition.

A natural question may be. why don't we define $\phi(\lambda,\mu) = \inf_{x\in X} L(x,\lambda,\mu)$? Probably the most important reason is simplicity. If so. in our two examples. we need to solve.

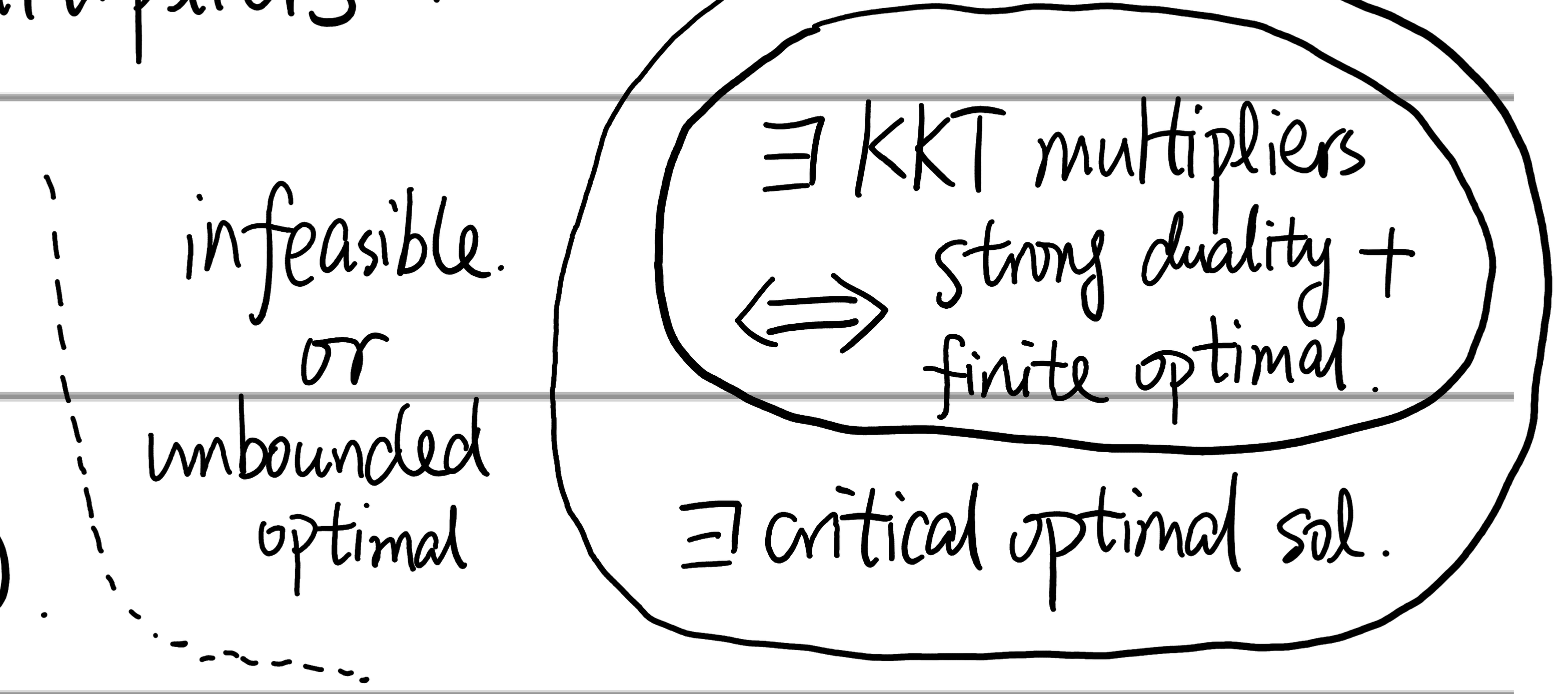
$\min\, (\lambda^T A_1 + \mu^T A_2 - c^T)\, x - \lambda^T b_1 - \mu^T b_2$ s.t. $\begin{matrix} A_1 x = b_1 \\ A_2 x \le b_1 \end{matrix}$. and

$\min\, x^T x + (\lambda^T A - \mu^T)\, x - \lambda^T b$. s.t. $Ax = b$. $x \ge 0$. (in particular, $\lambda=\mu=0$)

If we relax $\phi(\lambda,\mu) = \inf_{x\in D} L(x,\lambda,\mu)$. it becomes easier. Of course weak duality still holds. and under some conditions strong duality also holds. such as having KKT multipliers.

x^* has KKT multipliers λ^*, μ^* $\Rightarrow$

$\phi(\lambda^*,\mu^*) = \inf_x \tilde{L}(x) = L(x^*) = f(x^*)$.



Duality gap: $f(x) - \phi(\lambda,\mu) \ge 0$. $\forall x,\lambda,\mu$. note $\phi(\lambda,\mu) \le \phi^* \le f^* \le f(x)$.

Example. $\min\, x_1 + x_2$. s.t. $\begin{matrix} (x_1-1)^2 + x_2^2 \le 1 \\ (x_1+1)^2 + x_2^2 \le 1 \end{matrix}$ $\quad \phi(\mu_1,\mu_2) = \begin{cases} -\infty & \mu_1+\mu_2 \le 0. \\ \varphi(\mu_1,\mu_2) & \text{o.w.} \end{cases}$

$$\varphi(\mu_1,\mu_2) = \frac{-2(\mu_1-\mu_2)^2 + 2\mu_1 - 2\mu_2 - 1}{2(\mu_1+\mu_2)} \qquad \phi^* = \sup_{\mu\ge 0} \phi = \sup_{\mu_1+\mu_2>0} \varphi(\mu_1,\mu_2) = 0.$$

Example. $\min_{x,y>0}\, e^{-x}$ s.t. $x^2/y \le 0$. $\phi(\mu) = \inf_{x,y>0} (e^{-x} + \mu x^2/y) = \begin{cases} 0 & \mu \ge 0 \\ -\infty & \text{o.w.} \end{cases}$

Strong duality: $\phi^* = f^*$. Slater's condition: $\exists$ feasible x. $h(x) < 0$.

Why? A geometric interpretation is to consider the epigraph.

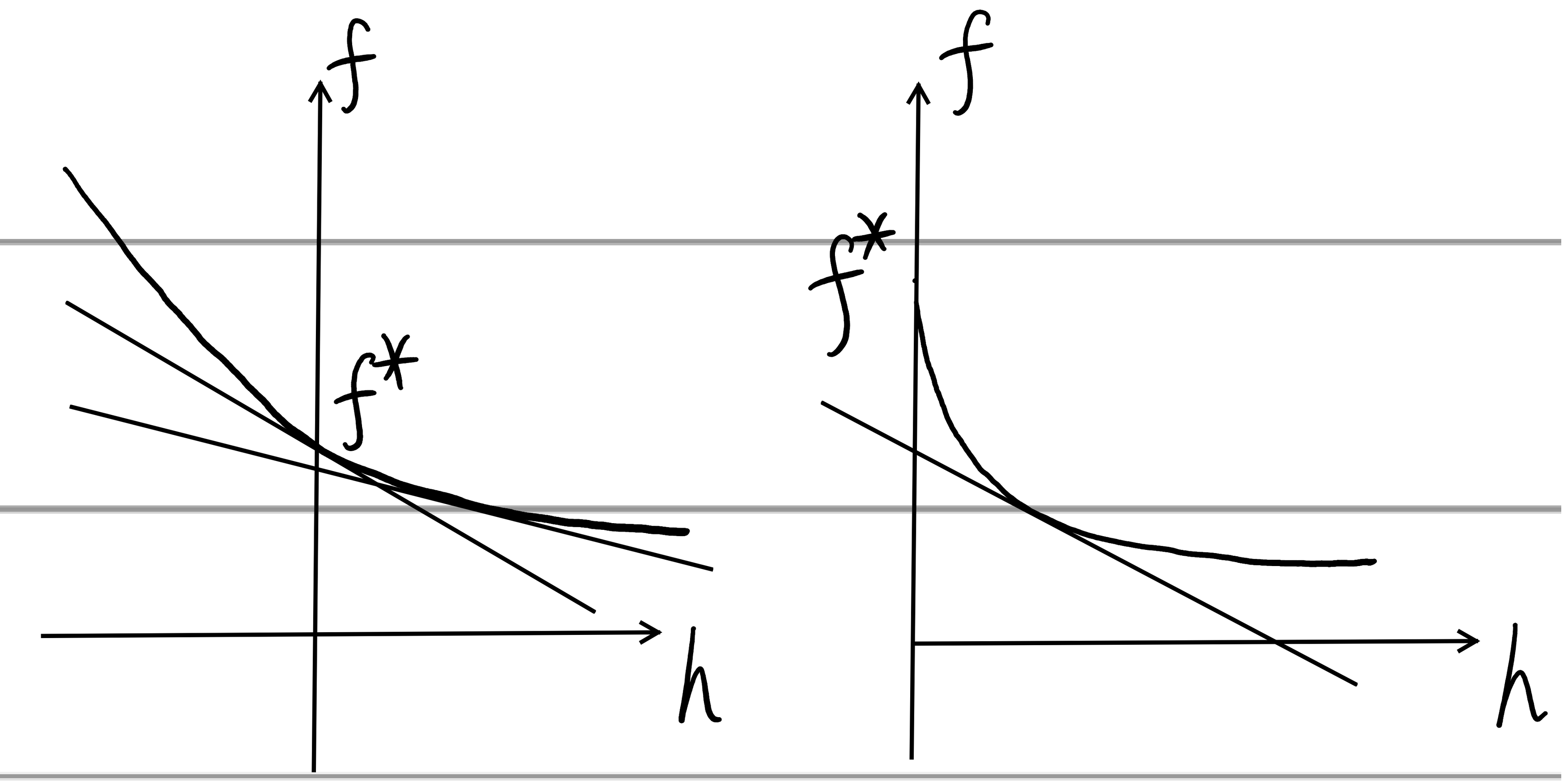
$C = \{(p, q, t) : \exists x,\ h_i(x) \le p_i,\ g_j(x) = q_j,\ f(x) \le t\}$. C is convex.

Note that f^* is the lowest intersection of C and t-axis.

$\phi(\lambda, \mu) = \inf\{(\mu, \lambda, 1)^T (p, q, t) : (p, q, t) \in C\}$ is the intersection of t-axis and a supporting hyperplane to C with normal vector $(\mu, \lambda, 1)^T$.

ϕ^* is the highest intersection.

Consider the supporting hyperplane passing through $(0, 0, f^*)$.



When does this argument fail? The unique supporting hyperplane is vertical.

Recall the example $\min_{x,y>0} e^{-x}$ s.t. $x^2/y \le 0$. $C = \left\{ \begin{matrix} (0, t) : t \ge 1 \\ (p, t) : p > 0,\ t > 0 \end{matrix} \right\}$.

Slater's condition: $\exists x \in \operatorname{int} D$, $\begin{matrix} g(x) = 0 \\ h(x) < 0 \end{matrix}$ $\Rightarrow \{(p, q, t) \in C : p < 0\} \ne \emptyset$.

$\Rightarrow$ the vertical hyperplane passing through $(0, 0, f^*)$ cannot be supporting.

Proof of convexity: Take two points (p_1, q_1, t_1), $(p_2, q_2, t_2) \in C$. So

$$\left.\begin{matrix} \exists x_1 \text{ s.t. } h(x_1) \le p_1,\ g(x_1) = q_1,\ f(x_1) \le t_1 \\ \exists x_2 \text{ s.t. } h(x_2) \le p_2,\ g(x_2) = q_2,\ f(x_2) \le t_2 \end{matrix}\right\} \Rightarrow \begin{matrix} \forall \theta \in [0,1].\ \text{let } y = \theta x_1 + \bar\theta x_2. \\ h(y) \le \theta h(x_1) + \bar\theta h(x_2) \le \theta p_1 + \bar\theta p_2 \\ g(y) = \theta g(x_1) + \bar\theta g(x_2) = \theta q_1 + \bar\theta q_2 \\ f(y) \le \theta f(x_1) + \bar\theta f(x_2) \le \theta t_1 + \bar\theta t_2 \end{matrix}$$

$\Rightarrow \theta(p_1, q_1, t_1) + \bar\theta(p_2, q_2, t_2) \in C$. $\square$

since f, h convex, g affine.

If $f^* = -\infty$. by weak duality. $\phi^* \le f^* \Rightarrow \phi^* = -\infty = f^*$.

Now we assume $f^* > -\infty$. under Slater's condition. feasible set $X \neq \phi$

so $f^* < \infty$. It suffices to show $\exists$ nonvertical. SH passing through $(0, 0, f^*)$

Proof of strong duality: We first show that $(0,0,f^*) \in \partial C$.

note that $(0,0,f^*)$ may not in C. but it is in $\mathrm{cl}\, C$.

$f^* = \inf_{x \in X} f \Rightarrow \forall \varepsilon > 0.\ \exists x.\ g(x) = 0.\ h(x) < 0.\ f(x) < f^* + \varepsilon$.

$\Leftrightarrow (0,0,f^* + \varepsilon) \in C \Rightarrow (0,0,f^*) \in \mathrm{cl}\, C$. In addition. $\forall \delta > 0$.

$\nexists$ feasible x. $f(x) \le f^* - \delta \Leftrightarrow (0,0,f^* - \delta) \notin C \Rightarrow (0,0,f^*) \notin \mathrm{int}\, C$.

$(0,0,f^*) \in \partial C \Rightarrow \exists$ supporting hyperplane passing through $(0,0,f^*)$.

$\Rightarrow \exists (\mu, \lambda, \xi) \neq 0.\ \forall (p, q, t) \in C.\ \mu^T p + \lambda^T q + \xi t \geq \xi f^*$.

$\forall t > f^*.\ (0,0,t) \in C \Rightarrow \xi \geq 0.\ \forall \vec{p} > \vec{0}.\ (p, q, f^*) \in C$

$\Rightarrow \vec{\mu} \geq 0$. we now show that $\xi \neq 0$. o.w. $\mu^T p + \lambda^T q \geq 0$.

By Slater's condition. $\exists \tilde{x}$. s.t. $g(\tilde{x}) = 0$. $h(\tilde{x}) < 0$. $\Rightarrow \exists \tilde{t}$. s.t.

$(h(\tilde{x}).\ g(\tilde{x}).\ \tilde{t}) \in C \Rightarrow \mu^T h(\tilde{x}) \geq 0$. since $h(\tilde{x}) < 0$. we have $\mu = 0$.

$(\mu, \lambda, \xi) \neq 0 \Rightarrow \lambda \neq 0$. and $\lambda^T g(x) \geq 0$. $\forall x \in D$. let $g(x) = Ax - b$.

A full rank $\Rightarrow \lambda^T A \neq 0$. $\tilde{x} \in \mathrm{int}\, D$. $g(\tilde{x}) = 0 \Rightarrow \exists \hat{x}.\ \lambda^T g(\hat{x}) < 0$.

Thus $\xi > 0$. Let $\tilde{\mu} = \mu/\xi$, $\tilde{\lambda} = \lambda/\xi$. Then SH: $\tilde{\mu}^T p + \tilde{\lambda}^T q + t \geq f^*$. $\forall (p, q, t) \in C$.

$\phi(\tilde{\lambda}, \tilde{\mu}) = \inf_{x \in D} L(x, \tilde{\lambda}, \tilde{\mu}) = \inf_{x \in D} f(x) + \tilde{\lambda}^T g(x) + \tilde{\mu}^T h(x)$.

$\forall x \in D$. $(h(x), g(x), f(x)) \in C \Rightarrow f(x) + \tilde{\lambda}^T g(x) + \tilde{\mu}^T h(x) \geq f^*$

$\Rightarrow \phi(\tilde{\lambda}, \tilde{\mu}) \geq f^* \Rightarrow \phi^* \geq f^*$ since $\tilde{\mu} \geq 0$. $\Rightarrow \phi^* = f^*$. $\square$.

Corollary. can relax $h_j(\tilde{x}) < 0$ to feasibility $h_j(\tilde{x}) \leq 0$. if h_j affine.

Recall KKT. $\min f(x)$ s.t. $\begin{matrix} g(x) = 0 \\ h(x) \leq 0 \end{matrix}$.

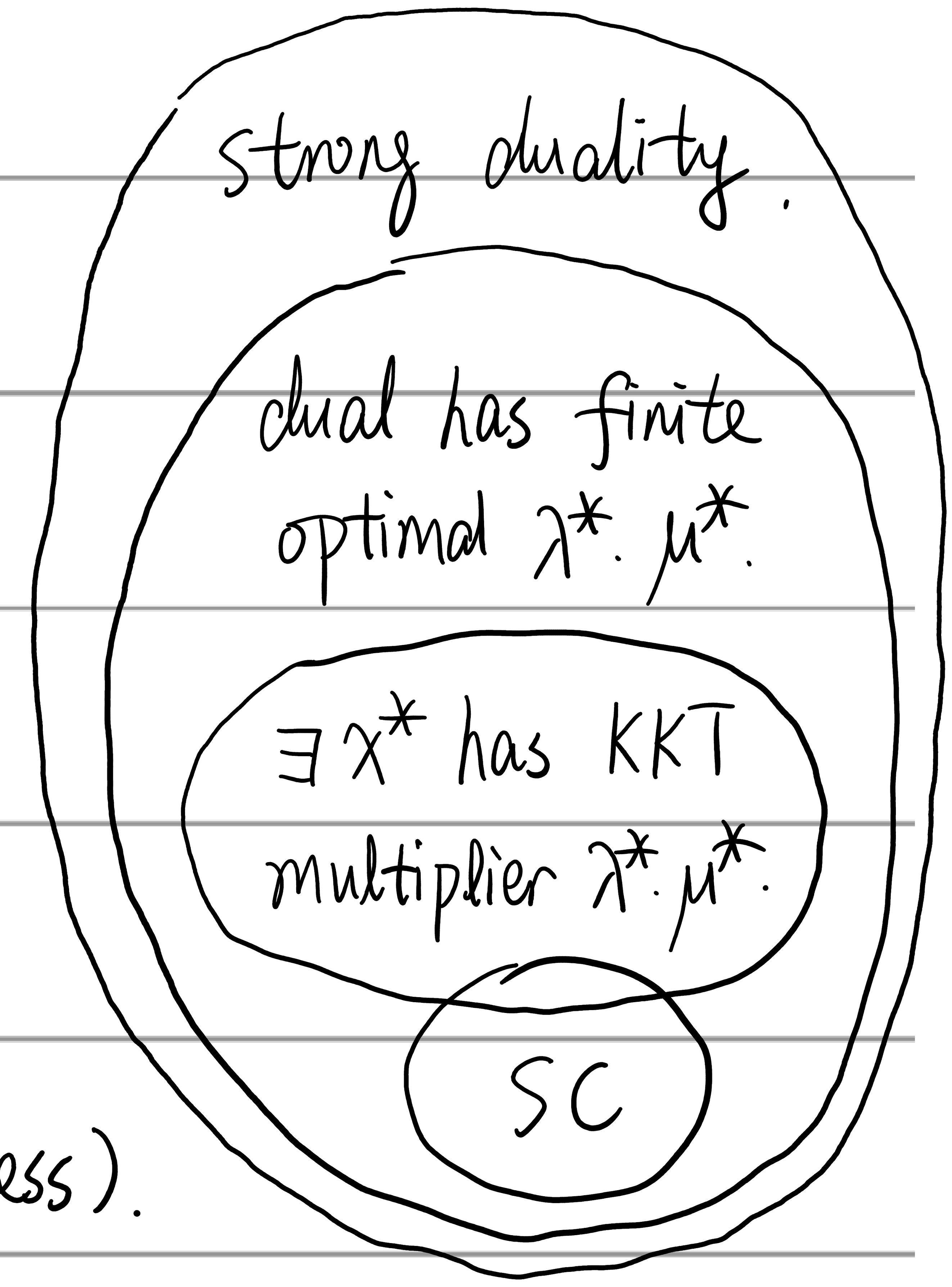
x^* has KKT multiplier λ^*, μ^*, if.

① $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$. (stationarity).

② $g(x^*) = 0$. $h(x^*) \leq 0$. (primal feasibility).

③ $\mu^* \geq 0$. (dual feasibility).

④ $\mu_j^* h_j(x^*) = 0$. $\forall j$. (complementary slackness).



We already proved: x^* optimal solution + regularity $\Rightarrow$ KKT.

for convex problem. x^* has KKT λ^*, μ^* $\Rightarrow \begin{cases} x^* \text{ optimal for primal} \\ \lambda^*, \mu^* \text{ optimal for dual} \\ \text{strong duality holds.} \end{cases}$

In fact, vice versa and we now prove it. ?

Assume x^* is optimal for primal. (λ^*, μ^*) is optimal for dual. then.

$f^* = \phi^* = \phi(\lambda^*, \mu^*) = \inf_x L(x, \lambda^*, \mu^*) \leq L(x^*, \lambda^*, \mu^*) \leq f(x^*)$.

$$\Rightarrow \inf_x L(x,\lambda^*,\mu^*) = L(x^*,\lambda^*,\mu^*) = f(x^*) \Rightarrow \begin{cases} \nabla_x L(x^*,\lambda^*,\mu^*) = 0 \\ \mu_j h_j(x^*) = 0,\ \forall j. \end{cases} \quad \square.$$

Example. Dual of the support vector machine.

SVM: separate data by a hyperplane. s.t. max minimum distance

$$\max_{w,b} \min_{1\le i\le m} \frac{|w^T x_i + b|}{\|w\|} \quad \text{s.t.} \quad y_i(w^T x_i + b) > 0. \quad y_i \in \{\pm 1\}.$$

equivalent to $\min \frac{1}{2}\|w\|^2$. s.t. $y_i(w^T x_i + b) \ge 1$.

if not linearly separable. $\min \frac{1}{2}\|w\|^2 + C \cdot \mathbf{1}^T \xi$. s.t. $\begin{array}{l} y_i(w^T x_i + b) \ge 1 - \xi_i \\ \xi \ge 0 \end{array}$.

Convex optimization with affine constraints. Slater's condition holds.

$$L(w,b,\xi,\mu,\nu) = \frac{1}{2}\|w\|^2 + C\mathbf{1}^T\xi + \sum \mu_i (1 - \xi_i - y_i(w^T x_i + b)) - \nu^T \xi.$$

$\tilde{x}_i = y_i x_i$ $$= \frac{1}{2}\|w\|^2 + (C\mathbf{1} - \mu - \nu)^T \xi - (\sum \mu_i y_i x_i)^T w - \mu^T y b + \mathbf{1}^T \mu.$$

$$\nabla_{w,b,\xi} L = (w - \sum \mu_i y_i x_i,\ \mu^T y,\ C\mathbf{1} - \mu - \nu) \Rightarrow w = \sum \mu_i y_i x_i. \text{ and}$$

$$\phi(\mu,\nu) = \inf L(w,b,\xi,\mu,\nu) = \begin{cases} \mathbf{1}^T\mu - \frac{1}{2}\mu^T \tilde{X}\tilde{X}^T \mu. & \text{if } \mu^T y = 0,\ \mu + \nu = C\mathbf{1} \\ -\infty & \text{o.w.} \end{cases}$$

Dual problem. $\max_{\mu\ge 0, \nu \ge 0} \phi(\mu,\nu) = \mathbf{1}^T\mu - \frac{1}{2}\mu^T \tilde{X}\tilde{X}^T\mu$. s.t. $\begin{array}{l} \mu^T y = 0 \\ \mu + \nu = C\mathbf{1}. \end{array}$

eliminating ν. $$\max_{\substack{\mu^T y = 0 \\ 0 \le \mu_i \le C.}} \mathbf{1}^T\mu - \frac{1}{2}\mu^T \tilde{X}\tilde{X}^T \mu = \mathbf{1}^T\mu - \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m \mu_i \mu_j y_i y_j x_i^T x_j$$