This week: We prove Theorem 16.1 in the notes.

$$\text{OP}: \quad \min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to:} \quad \begin{cases} c_i(x) = 0 & i \in \mathcal{E} \\ c_i(x) \geq 0 & i \in \mathcal{I} \end{cases} \qquad (1)$$

Lagrangian:

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(x)$$

Suppose x_* is a minimizer of the OP (1). Furthermore suppose that f and the c_i's are continuously differentiable and that the LICQ holds at x_*. Then there exists a vector λ_* with components λ_i^* $(i \in \mathcal{E} \cup \mathcal{I})$ such that the following conditions hold:

KKT conditions

$$\begin{aligned} \nabla_x \mathcal{L}(x_*, \lambda_*) &= 0 \\ c_i(x_*) &= 0 && i \in \mathcal{E} \\ c_i(x_*) &\geq 0 && i \in \mathcal{I} \\ \lambda_i^* &\geq 0 && i \in \mathcal{I} \\ \lambda_i^* c_i(x_*) &= 0 && i \in \mathcal{E} \cup \mathcal{I} \end{aligned}$$

Remark: Since $c_i(x_*) > 0$ for $i \in \mathcal{I} \setminus \mathcal{A}(x_*)$, by KKT 5 we have $\lambda_i^* = 0$ for $i \in \mathcal{I} \setminus \mathcal{A}(x_*)$.

Hence, KKT 1 can be re-written as:

$$\nabla f(x_*) - \sum_{i \in \mathcal{A}(x_*)} \lambda_i^* \nabla c_i(x_*)$$

$$\nabla f(x_*) = \sum_{i \in \mathcal{A}(x_*)} \lambda_i^* \nabla c_i(x_*)$$

Some final intermediate results.

Lemma 16.1 (Fundamental Necessary Condition)

If x_* is a minimizer of the OP (1), then

$$d \cdot \nabla f(x_*) \geq 0 \quad \forall \; d \in T_2(x_*)$$

Proof by contradiction: Suppose that there exists a $d \in T_2(x_*)$ such that $d \cdot \nabla f(x_*) < 0$. (*) Let z_k be a feasible sequence tending to x_* such that:

$$d = \lim_{k \to \infty} \frac{z_k - x_*}{t_k}.$$

$$\lim_{t_k \to 0} \frac{\|\epsilon_k\|_2}{t_k} = 0$$

$$\Rightarrow \|\epsilon_k\|_2 \sim t_k^{1+\delta}$$

Re-arrange:

$$z_k = x_* + t_k d + \epsilon_k, \qquad \epsilon_k = o(t_k)$$

$$f(z_k) = f(x_* + t_k d + \epsilon_k)$$

$$= f(x_*) + t_k \underbrace{d \cdot \nabla f(x_*)}_{\text{negative}} + \underbrace{o(t_k)}_{\text{negligible}}$$

$\Rightarrow f(z_k) < f(x_*)$ for t_k sufficiently small.

Contradiction, since x_* is a local minimizer.

Hence, statement (*) is false, so

$$\nabla f(x_*) \cdot d \geq 0 \quad \forall \; d \in T_2(x_*)$$

Hence, statement (*) is [illegible], so

$$\underline{d} \cdot \nabla f(\underline{x}_*) \geq 0 \quad \forall\, \underline{d} \in T_\Omega(\underline{x}_*)$$

Another key result: Hyperplane Separation Theorem (Theorem 16.2)

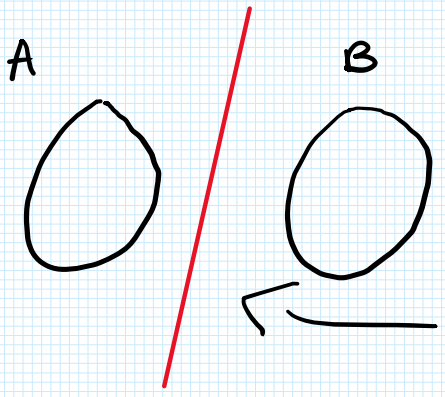
Let A and B be two disjoint non-empty convex subsets of $\mathbb{R}^n$. Then there exists a plane which separates A and B. Mathematically, there exists a non-zero vector $\underline{n}$ and a constant c such that:

$$\underline{x} \cdot \underline{n} \geq c \qquad \underline{y} \cdot \underline{n} \leq c$$

for all $\underline{x} \in A$ and all $\underline{y}$ in B.

No proof in this module, but see the pictures below for the idea.

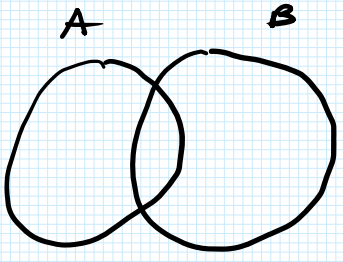
i)



$A \cap B = \emptyset$

Separating hyperplane

ii)



No separating hyperplane

Extension: Let A and B be disjoint non-empty convex subsets of $\mathbb{R}^n$ such that:

i) A and B are closed → contains its own boundary

ii) At least one of A and B is bounded.

Then there exists a non-zero vector $\underline{n}$ and a constant c such that:

$$\underline{x} \cdot \underline{n} > c \qquad \underline{y} \cdot \underline{n} < c$$

for all $\underline{x} \in A$, $\underline{y} \in B$.

The H.S.T. enables us to prove the final intermediate result called Farkas's Lemma.

Lemma 16.2 (Farkas's Lemma)

Let $K = \{ B\underline{y} \mid y_i \geq 0, \; i = 1, \dots, m \}$

be a cone, where $B \in \mathbb{R}^{n \times m}$. Given any vector $\underline{g} \in \mathbb{R}^n$, either

- $\underline{g} \in K$
- Or there exists a vector $\underline{d} \in \mathbb{R}^n$ such that $\underline{g} \cdot \underline{d} < 0$ and $(B^T \underline{d})_i \geq 0, \; i = 1, \dots, m$

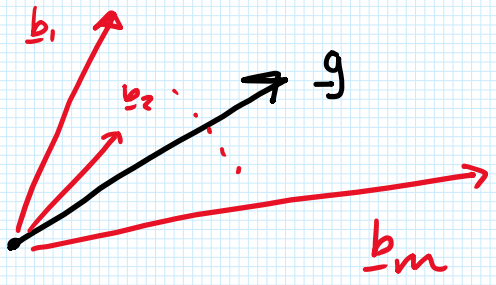
Idea:

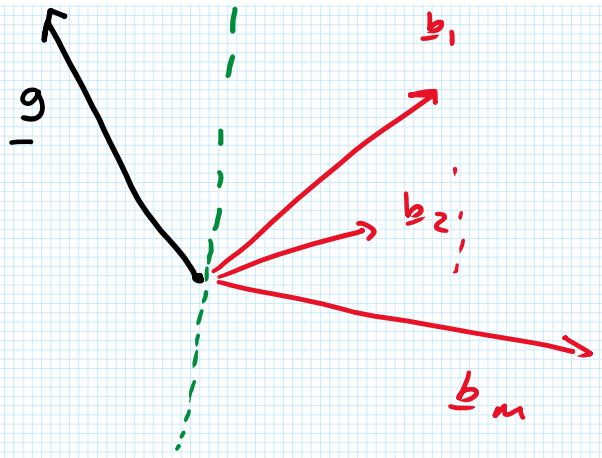
Case 2

Separating hyperplane

$\underline{b}_1$

Case 1





Proof: We show first that the alternatives are mutually exclusive, i.e. that either

1. $g \in K$
2. $\exists\, d \in \mathbb{R}^n$ s.t. $g \cdot d < 0$ AND $(B^T d)_i \geq 0$, $i = 1, \ldots, m$

but not both at once.

Therefore, take $g \in K$:

$$g = By, \quad y_i \geq 0. \Big\} \text{ \#1 is true.}$$

Assume for contradiction that #2 is also true.

Hence, $\exists\, d \in \mathbb{R}^n$ s.t.

$$g \cdot d < 0 .$$

$$\Rightarrow \langle g, d \rangle < 0$$

$$\overset{\#1}{\Rightarrow} \langle By, d \rangle < 0$$

$$\Rightarrow \langle y, B^T d \rangle < 0$$

$$\text{Conv} \overset{m}{\ldots} (B^T d)$$

$\Rightarrow \langle y, \ldots \rangle$

Components $\Rightarrow$ $$\sum_{i=1}^{m} y_i (B^T \underline{d})_i < 0 \quad ☆$$

But the y_i's are non-neg., and the $(B^T\underline{d})_i$'s are non-negative (#2 is true), hence:

$$\sum_{i=1}^{m} y_i (B^T \underline{d})_i \geq 0 \quad . \quad ☆☆$$

Contradiction between ☆ and ☆☆. Reasoning back up, if #1 is true, then #2 can't be true at the same time, so we have that #1 and #2 are mutually exclusive.

We now show that one option or the other must necessarily hold. If $g \in \mathcal{K}$ then we are done. ☑

So we look at $g \notin \mathcal{K}$. We have:

$$A = \{ g \}$$
$$B = \mathcal{K} = \{ By \mid y_i \geq 0, i = 1, \ldots, m \}$$

A is bounded, B is closed and convex (cone is closed and convex). $A \cap B = \emptyset$ ($g \notin \mathcal{K}$).

Hence, the H.S.T. applies (strict form). There exists a constant c and a vector $\underline{d} \in \mathbb{R}^n$ such that:

A: $\underline{d} \cdot \underline{g} < c$

B: $\underline{d} \cdot \underline{s} > c \quad \forall \ \underline{s} \in \mathcal{R}$

Notice: $\underline{0} \in \mathcal{R}$, $0 > c$ (c is negative).

Hence:

$$\underline{d} \cdot \underline{s} > c \qquad (c \text{ negative})$$

$\underline{s} \in \mathcal{R}$
$\Rightarrow$
$$\underline{d} \cdot (B\underline{y}) > c$$

$$\Rightarrow \quad \langle \underline{d}, B\underline{y} \rangle > c$$

$$\Rightarrow \quad \langle B^T \underline{d}, \underline{y} \rangle > c$$

$$\Rightarrow \quad \langle \underline{y}, B^T \underline{d} \rangle > c$$

Components
$\Rightarrow$
$$\sum_{i=1}^{m} y_i (B^T \underline{d})_i > c$$

True for all $\underline{s} \in \mathcal{R}$, therefore true for all $y_i \geq 0$.

So we can rescale the y_i's, taking $y_i \to \lambda y_i$,

where λ is positive:

$$\sum_{i=1}^{m} \lambda\, y_i (B^T \underline{d})_i > c$$

$$\Rightarrow \quad \sum_{i=1}^{m} y_i (B^T \underline{d})_i > \frac{c}{\lambda}$$

~~For~~ True for all $\lambda > 0$, so take $\lambda \to \infty$:

$$\sum_{i=1}^{m} y_i (B^T \underline{d})_i \geq 0$$

This is exactly Option # 2. H.S.T. gives us $\underline{d}$ $(g\underline{d} < 0)$

and we have shown :

$$\sum_{i=1}^{m} y_i (B^T \underline{d})_i \geq 0 \quad \forall \quad y_i \geq 0$$

$$\therefore \quad (B^T \underline{d})_i \geq 0 \qquad i = 1, \cdots, m .$$

Hence, Option # 2 holds. ∎

Start with Farkas's Lemma (from Tuesday)

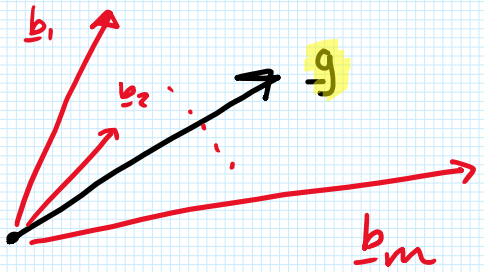
Let

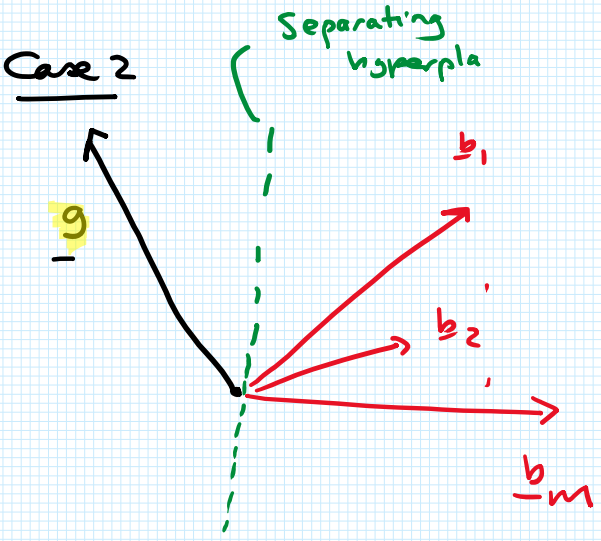
$$K = \{By \mid y_i \geq 0,\ i = 1, \dots, m\}$$

be a cone in $\mathbb{R}^n$, where $B \in \mathbb{R}^{n \times m}$. Let g be any vector in $\mathbb{R}^n$. Then, either:

- $g \in K$
- There exists a vector d in $\mathbb{R}^n$ such that:
  - $g \cdot d < 0$
  - $(B^T d)_i \geq 0, \quad i = 1, \dots, m$

Idea:





Extension of Farkas's Lemma:

Let

$$K = \left\{ By + Cw \;\middle|\; \begin{array}{l} y_i \geq 0,\ i = 1, \dots, m \\ w_i \in \mathbb{R},\ i = 1, \dots, p \end{array} \right\} \quad (1)$$

where B and C are matrices of the appropriate dimension.

For any vector $g \in \mathbb{R}^n$, either:

For any vector $g \in \mathbb{R}^n$, either:

- $g \in R$ or
- There exists a vector d in $\mathbb{R}^n$ such that
  - $g \cdot d < 0$
  - $(A^T d)_i \geq 0, \quad i = 1, \ldots, m$
  - $(C^T d)_i = 0, \quad i = 1, \ldots, p$

We take this extension of Farkas's Lemma as given (Nocedal & Wright).

We apply the extended Farkas's Lemma.

Let

$$N = \left\{ \sum_{i \in A(x_*) \cap I} \lambda_i \nabla c_i(x_*) + \sum_{i \in A(x_*) \cap E} \lambda_i \nabla c_i(x_*) \;\middle|\; \lambda_i \geq 0, \; i \in A(x_*) \cap I; \; \lambda_i \in \mathbb{R}, \; i \in A(x_*) \cap E \right\}$$

where x_* is a generic feasible point.

Re-write:

$$N = \left\{ A^T \lambda \;\middle|\; \lambda_i \geq 0, \; i \in A(x_*) \cap I \right\} \quad (2)$$

where

$$A = \begin{pmatrix} \frac{\partial c_1}{\partial x_1}(x_*) & \cdots & \frac{\partial c_1}{\partial x_n}(x_*) \\ \vdots & & \\ \frac{\partial c_m}{\partial x_1}(x_*) & \cdots & \frac{\partial c_m}{\partial x_n}(x_*) \end{pmatrix} \in \mathbb{R}^{m \times n}$$

where we have relabelled constraints such that

$$\mathcal{A}(\underline{x}_*) = \{1, \dots, m\}.$$

The set (2) is of the same form as (1). So extended Farkas's Lemma applies.

Take $-\underline{g} = \nabla f(\underline{x}_*)$. Then, either:

- $\nabla f(\underline{x}_*) \in N$ or
- there exists a $\underline{d}$ in $\mathbb{R}^n$ such that:
  - $\underline{d} \cdot \nabla f(\underline{x}_*) < 0$
  - $(A\underline{d})_i \geq 0, \quad i \in \mathcal{A}(\underline{x}_*) \cap I$
  - $(A\underline{d})_i = 0, \quad i \in \mathcal{A}(\underline{x}_*) \cap E.$

Hence, $\underline{d} \in \mathcal{F}_\Omega(\underline{x}_*)$.

Proof of KKT conditions

Let $\underline{x}_*$ be a local minimizer of the OP.

Then, by the fundamental necessary condition (Lemma 16.1)

$$\underline{d} \cdot \nabla f(\underline{x}_*) \geq 0, \quad \forall \ \underline{d} \in T_\Omega(\underline{x}_*).$$

By the hypothesis of the KKT theorem, the LICQ holds at $\underline{x}_*$, hence $T_\Omega(\underline{x}_*) = \mathcal{F}_\Omega(\underline{x}_*)$, hence

holds at x_*, hence $T_\Omega(x_*) = \mathcal{F}_\Omega(x_*)$, hence

$$\underline{d} \cdot \nabla f(x_*) \geq 0 \qquad \forall\, \underline{d} \in \mathcal{F}_\Omega(x_*).$$

So $\underline{d} \cdot \nabla f(x_*)$ can't be negative. This rules out option #2 in Farkas's Lemma (for $\underline{g} = \nabla f(x_*)$).

So option #1 in (extended) Farkas's Lemma applies:

$$\underline{g} = \nabla f(x_*) \in N.$$

Hence, there exist scalars λ_i such that:

$$\begin{aligned} &- \lambda_i \geq 0 \qquad i \in \mathcal{A}(x_*) \cap \mathcal{I} \\ &- \lambda_i \in \mathbb{R} \qquad i \in \mathcal{A}(x_*) \cap \mathcal{E} \end{aligned} \tag{3}$$

such that

$$\underline{g} = \nabla f(x_*) = \sum_{i \in \mathcal{A}(x_*)} \lambda_i \nabla c_i(x_*)$$

We may take:

$$\lambda_i^* = \begin{cases} \lambda_i, & i \in \mathcal{A}(x_*) \\ 0, & i \in \mathcal{I} \cap \mathcal{A}(x_*) \end{cases} \tag{4}$$

Hence:

$$\nabla f(x_*) = \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i^* \nabla c_i(x_*)$$

KKT 1

Since x_* is feasible by assumption:

$$c_i(x_*) = 0, \quad i \in \mathcal{E} \qquad \text{KKT 2}$$

$$c_i(x_*) \geq 0, \quad i \in \mathcal{I} \qquad \text{KKT 3}$$

From (3) and (4), it follows that

$$\lambda_i^* \geq 0, \quad i \in \mathcal{I} \qquad \text{KKT 4}$$

Finally:

$i \in \mathcal{A}(x_*) \cap \mathcal{E}$ $\quad c_i(x_*) = 0 \Rightarrow \lambda_i^* c_i(x_*) = 0$

$i \in \mathcal{A}(x_*) \cap \mathcal{I}$ $\quad c_i(x_*) = 0 \Rightarrow \lambda_i^* c_i(x_*) = 0$

$i \in \mathcal{I} \setminus \mathcal{A}(x_*) \overset{\text{Eq (3)}}{\Rightarrow} \lambda_i^* = 0 \Rightarrow \lambda_i^* c_i(x_*) = 0$

This covers all $i \in \mathcal{I} \cup \mathcal{E}$, therefore:

$$\lambda_i^* c_i(x_*) = 0, \quad i \in \mathcal{I} \cup \mathcal{E} \qquad \text{KKT 5}$$

This establishes the KKT conditions. ∎

Plan:

- Today
  - Second-derivative test for constrained optimization
  - Dual formulation
- Next Tuesday
  - Recorded lecture (website) Exercises #6

- Next Thursday
  - Exercises #6 wrap-up (duality)
  - Structure of final exam

Second-derivative test (Ch. 17)

Critical cone:

$$\mathcal{C}(x_*, \lambda_*) = \left\{ w \in \mathcal{F}_\Omega(x_*) \;\middle|\; \begin{array}{ll} w \cdot \nabla c_i(x_*) = 0, & i \in \mathcal{E} \\ w \cdot \nabla c_i(x_*) = 0, & i \in \mathcal{A}(x_*) \cap \mathcal{I} \text{ and } \lambda_i^* > 0 \\ w \cdot \nabla c_i(x_*) \geq 0 & i \in \mathcal{A}(x_*) \cap \mathcal{I} \text{ and } \lambda_i^* = 0 \end{array} \right\}$$

Since $\lambda_i^* = 0$ for $i \in \mathcal{I} \setminus \mathcal{A}(x_*)$

$$\lambda_i^* \, w \cdot \nabla c_i(x_*) = 0$$

for all $i \in \mathcal{I} \cup \mathcal{E}$ and all $w \in \mathcal{C}(x_*, \lambda_*)$.

Suppose that the KKT conditions are satisfied at x_*:

$$\nabla f(x_*) = \sum_{i \in \mathcal{A}(x_*)} \lambda_i^* \nabla c_i(x_*)$$

Dot both sides with $w \in \mathcal{C}(x_*, \lambda_*)$:

$$w \cdot \nabla f(x_*) = \sum_{i \in \mathcal{A}(x_*)} \lambda_i^* \, w \cdot \nabla c_i(x_*)$$

KKT conditions can be re-written as a first-derivative test:

$$w \cdot \nabla f(x_*) = 0, \quad \forall \; w \in \mathcal{C}(x_*, \lambda_*)$$

$$\underline{w} \cdot \nabla f(\underline{x}_*) = 0, \quad \forall \ \underline{w} \in \mathcal{C}(\underline{x}_*, \underline{\lambda}_*)$$

Theorem 17.1 Suppose that $\underline{x}_*$ is a local s/n of the OP and that the LICQ holds at $\underline{x}_*$. Let $\underline{\lambda}^*$ be the Lagrange multiplier for which the KKT conditions hold. Then:

$$\langle \underline{w}, \nabla_{xx} \mathcal{L}(\underline{x}_*, \underline{\lambda}_*) \underline{w} \rangle \geq 0 \quad \forall \ \underline{w} \in \mathcal{C}(\underline{x}_*, \underline{\lambda}_*)$$

where $\nabla_{xx}\mathcal{L}$ is the Hessian matrix.

Theorem 17.2: Suppose that for some feasible point $\underline{x}_*$ there exists a vector of Lagrange multipliers satisfying the KKT conditions. Suppose also that

$$\langle \underline{w}, \nabla_{xx} \mathcal{L}(\underline{x}_*, \underline{\lambda}_*) \underline{w} \rangle > 0 \quad \forall \underline{w} \in \mathcal{C}(\underline{x}_*, \underline{\lambda}_*).$$

Then $\underline{x}_*$ is a strict local minimizer of the OP.

Duality (§ 17.2)

To fix ideas, assume:

- Inequality constraints only:
  $$c_i(x) \geq 0, \quad i \in \{1, \ldots, m\}$$
- f and $-c_i$ are convex functions (essential).

- f and $-c_i$ are convex functions (essential).

OP:

$$\min_{\underline{x} \in \mathbb{R}^n} f(\underline{x}) \quad \text{subject to} \quad c_i(\underline{x}) \geq 0, \; i \in \{1, \ldots, m\} \qquad (1)$$

Introduce a general (no constraints) Lagrangian:

$$\mathcal{L}(\underline{x}, \underline{\lambda}) = f(\underline{x}) - \sum_{i=1}^{m} \lambda_i c_i(\underline{x})$$

$$\underline{\lambda} = (\lambda_1, \ldots, \lambda_m)^T .$$

Dual objective function:

$$q : \mathbb{R}^m \longrightarrow \mathbb{R}$$
$$\underline{\lambda} \longmapsto q(\underline{\lambda})$$

where

$$q(\underline{\lambda}) = \inf_{\underline{x} \in \mathbb{R}^n} \mathcal{L}(\underline{x}, \underline{\lambda})$$

"Legendre transformation"

We require q to be bounded below, so we restrict to $\underline{\lambda} \in \mathcal{D}$ where

$$\mathcal{D} = \{ \underline{\lambda} \in \mathbb{R}^m \mid q(\underline{\lambda}) > -\infty \}$$

Under certain conditions, solving the OP (1) is equivalent to solving the dual problem:

$$\max \; q(\underline{\lambda}) \quad \text{subject to} \quad \lambda_i \geq 0 \quad i \in \{1, \ldots, m\}$$

$$\max_{\underline{\lambda} \in \mathbb{R}^m} q(\underline{\lambda}) \quad \text{subject to} \quad \lambda_i \geq 0 \quad i \in \{1, \ldots, m\}$$

R — Eq. (2)

Terminology:

- OP (1) is called the primal problem.
- OP (2) is called the dual problem.

Application: Support-Vector Machines, in Machine Learning.

Theorem 17.4: Suppose that $\underline{x}^*$ solves the OP (1). Then any $\underline{\lambda}$ for which $(\underline{x}^*, \underline{\lambda})$ solve the KKT conditions solves the dual problem (2).

The other way around is true as well: A solution $\underline{\lambda}^*$ to the dual problem gives rise to a sln of the primal problem (Theorem 17.5).

Why? High-dimensional parameter space ($\mathbb{R}^n$). Solution of KKT conditions requires solving for $n+m$ unknowns (O.P. in n dimensions). Solution of dual problem is an O.P. in m dimensions. So it makes sense to solve the dual problem when $m < n$.

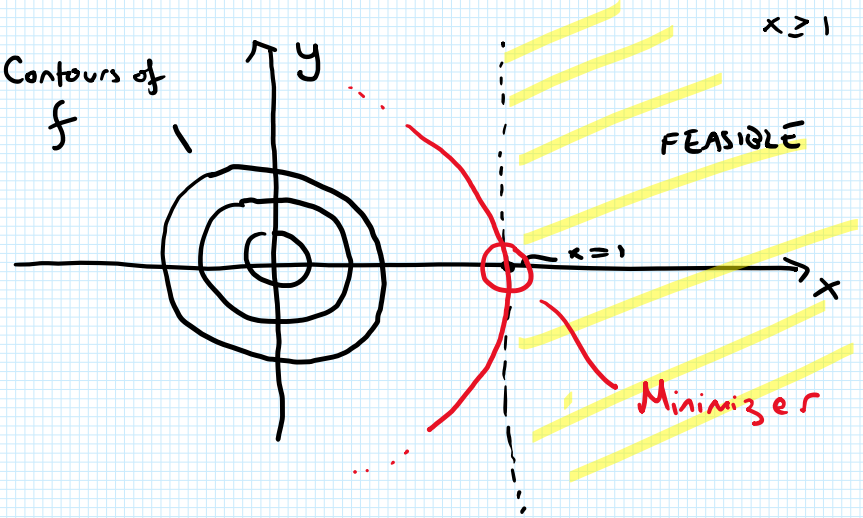
Example (§ 17.2.1)

$$\min_{\underline{x} \in \mathbb{R}^2} \ \tfrac{1}{2}(x^2+y^2) \quad \text{subject to} \quad x \geq 1 \ \ (x-1 \geq 0).$$

Primal problem : 2D optimization

Dual problem (λ) : 1D optimization

Graphical solution :



$$\underline{x}_* = (1,0)^T .$$

$$\mathcal{L}(\underline{x}, \lambda) = \tfrac{1}{2}(x^2+y^2) - \lambda(x-1)$$

$$q(\lambda) = \inf_{\underline{x} \in \mathbb{R}^2} \mathcal{L}(\underline{x}, \lambda)$$

Compute:

$$\frac{\partial \mathcal{L}}{\partial x} = 0, \qquad \frac{\partial \mathcal{L}}{\partial y} = 0 .$$

$$\downarrow$$

$$x - \lambda = 0 \qquad y = 0 . \implies \begin{cases} x = \lambda \\ y = 0 \end{cases}$$

$$\therefore \ q(\lambda) = \mathcal{L}(x=\lambda, y=0, \lambda)$$

$$= \left[\tfrac{1}{2}(x^2+y^2) - \lambda(x-1)\right]_{(x=\lambda, y=0)}$$

$= \left[\frac{1}{2}(x^2+y^2) - \lambda(x-1)\right]_{(x=\lambda, y=0)}$

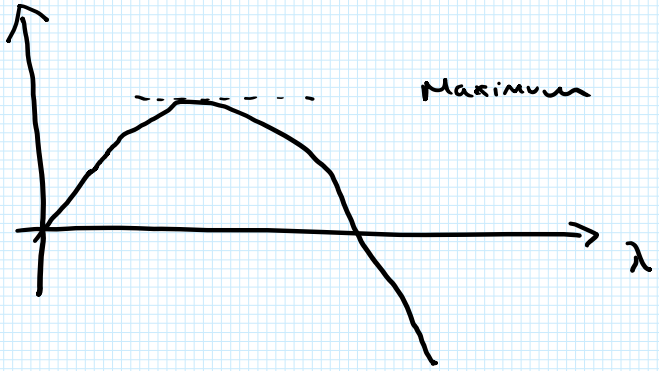
$= \frac{1}{2}\lambda^2 - \lambda(\lambda-1)$

$= \frac{1}{2}\lambda^2 - \lambda^2 + \lambda$

$\Rightarrow \boxed{q(\lambda) = -\frac{1}{2}\lambda^2 + \lambda}$

Dual problem:

$\max_{\lambda \in \mathbb{R}} q(\lambda)$, $\lambda \geq 0$.



$\frac{dq}{d\lambda} = -\lambda + 1$, $\frac{dq}{d\lambda} = 0 \Rightarrow \lambda = 1$

Solution: $\lambda_* = 1$.

Back to: $x = \lambda$, $y = 0$.

Therefore: $x_* = 1$, $y_* = 0$. ∎