ECE 271B – Winter 2022

The Karush–Kuhn–Tucker Conditions and Duality

Disclaimer:
This class will be recorded
and made available to students asynchronously.

Manuela Vasconcelos

ECE Department, UCSD

Optimization

- **goal**: find **maximum** or **minimum** of a function
- **Definition**: given functions f, $g_i, i = 1, \dots, r$ and $h_i, i = 1, \dots, m$ defined on some domain $\Omega \in \mathbb{R}^n$

$$\begin{aligned} \min_{\mathbf{w}} \quad & f(\mathbf{w}), \mathbf{w} \in \Omega \\ \text{subject to} \quad & g_i(\mathbf{w}) \le 0, \forall i \\ & h_i(\mathbf{w}) = 0, \forall i \end{aligned}$$

- for compactness, we write $g(\mathbf{w}) \le 0$ instead of $g_i(\mathbf{w}) \le 0, \forall i$ and similarly $h(\mathbf{w}) = 0$
- we derived necessary and sufficient conditions for (local) **optimality**
  - in the **absence of constraints** (**unconstrained**)
  - with **equality constraints only**

Minima Conditions (Unconstrained)

- **Theorem:** Let $f(\mathbf{w})$ be continuously differentiable. $\mathbf{w}^*$ is a local minimum of $f(\mathbf{w})$ if and only if
  - f has zero gradient at $\mathbf{w}^*$

$$\boxed{\nabla f(\mathbf{w}^*) = 0}$$

  - and the Hessian of f at $\mathbf{w}^*$ is positive−semidefinite

$$\boxed{\mathbf{d}^T \nabla^2 f(\mathbf{w}^*)\mathbf{d} \geq 0, \forall \mathbf{d} \in \mathbb{R}^n}$$

where

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_0^2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_0 \partial x_{n-1}}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_{n-1} \partial x_0}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_{n-1}^2}(\mathbf{x}) \end{bmatrix}$$

Maxima Conditions (Unconstrained)

- **Theorem:** Let $f(\mathbf{w})$ be continuously differentiable. $\mathbf{w}^*$ is a local maximum of $f(\mathbf{w})$ if and only if
  - f has zero gradient at $\mathbf{w}^*$

$$\boxed{\nabla f(\mathbf{w}^*) = 0}$$

  - and the Hessian of f at $\mathbf{w}^*$ is negative−semidefinite

$$\boxed{\mathbf{d}^T \nabla^2 f(\mathbf{w}^*)\mathbf{d} \leq 0, \forall \mathbf{d} \in \mathbb{R}^n}$$

where

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_0^2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_0 \partial x_{n-1}}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_{n-1} \partial x_0}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_{n-1}^2}(\mathbf{x}) \end{bmatrix}$$

Constrained Optimization

- with equality constraints only

- **Theorem:** Consider the problem

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad h(\mathbf{x}) = 0$$

where the constraint gradients $\nabla h_i(\mathbf{x}^*)$ are linearly independent. Then, $\mathbf{x}^*$ is a solution if and only if there exits a unique vector $\boldsymbol{\lambda}$ such that

i) $\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}^*) = 0$

ii) $\mathbf{y}^T[\nabla^2 f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(\mathbf{x}^*)]\mathbf{y} \geq 0, \forall \mathbf{y} \text{ s.t. } \nabla h(\mathbf{x}^*)^T \mathbf{y} = 0$

i) $\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}^*) = 0$

ii) $\mathbf{y}^T[\nabla^2 f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(\mathbf{x}^*)]\mathbf{y} \geq 0, \forall y$ s.t. $\nabla h(\mathbf{x}^*)^T \mathbf{y} = 0$

Alternative Formulation

- stating the conditions through the Lagrangian

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_i(\mathbf{x})$$

- the theorem can be compactly written as

i) $\nabla L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \begin{bmatrix} \nabla_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) \\ \nabla_{\boldsymbol{\lambda}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) \end{bmatrix} = 0$

ii) $\mathbf{y}^T \nabla^2_{\mathbf{xx}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*)\mathbf{y} \geq 0, \forall \mathbf{y}$ s.t. $\nabla h(\mathbf{x}^*)^T \mathbf{y} = 0$

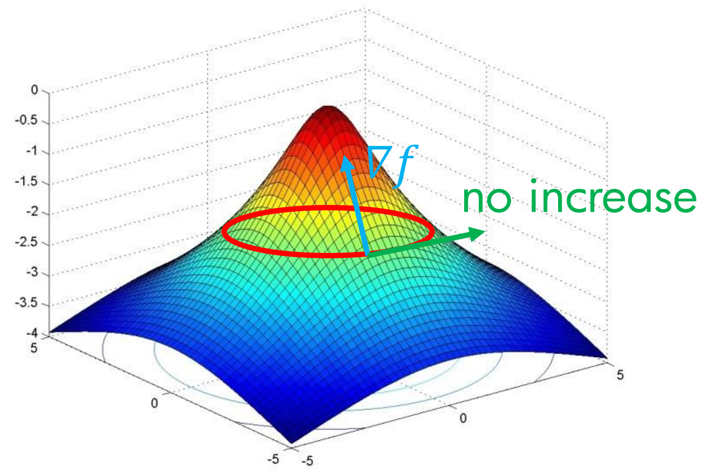
- the entries of λ are referred to as Lagrange multipliers

Geometric Interpretation

- derivative of f along $\mathbf{d}$ is

$$\boxed{\lim_{\alpha \to 0} \frac{f(\mathbf{w} + \alpha \mathbf{d}) - f(\mathbf{w})}{\alpha} = \mathbf{d}^T \nabla f(\mathbf{w}) = \|\mathbf{d}\| \|\nabla f(\mathbf{w})\| \cos\big(\mathbf{d}, \nabla f(\mathbf{w})\big)}$$

- this means that
  - greatest increase when $\mathbf{d} \parallel \nabla f$
  - no increase when $\mathbf{d} \perp \nabla f$ since there is no increase when $\mathbf{d}$ is tangent to iso–contour $f(\mathbf{x}) = k$
  - the gradient is perpendicular to the tangent of the iso–contour



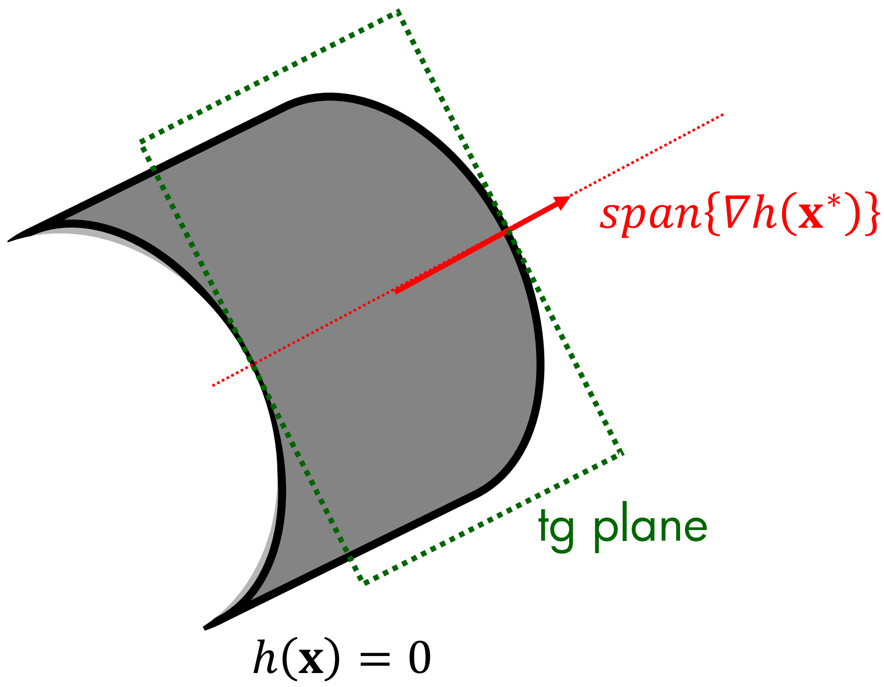
- allows <u>geometric interpretation</u> of the Lagrangian conditions

Lagrangian Optimization

i) $\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}^*) = 0$

ii) $\mathbf{y}^T[\nabla^2 f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(\mathbf{x}^*)]\mathbf{y} \geq 0, \forall \mathbf{y}$ s.t. $\nabla h(\mathbf{x}^*)^T \mathbf{y} = 0$

- geometric interpretation:
  - since $h(\mathbf{x}) = 0$ is an iso–contour of $h(\mathbf{x})$, $\nabla h(\mathbf{x}^*)$ is perpendicular to the iso–contour
  - $\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}^*) = 0$ says that $\nabla f(\mathbf{x}^*) \in span\{\nabla h_i(\mathbf{x}^*)\}$
  - i.e., $\nabla f \perp$ to tangent space of the constraint surface $h(\mathbf{x}) = 0$
  - intuitively
    - direction of largest increase of f is $\perp$ to constraint surface
    - the gradient is zero along the constraint
    - no way to give an infinitesimal gradient step, without violating the constraint
    - it is impossible to increase f and still satisfy the constraint

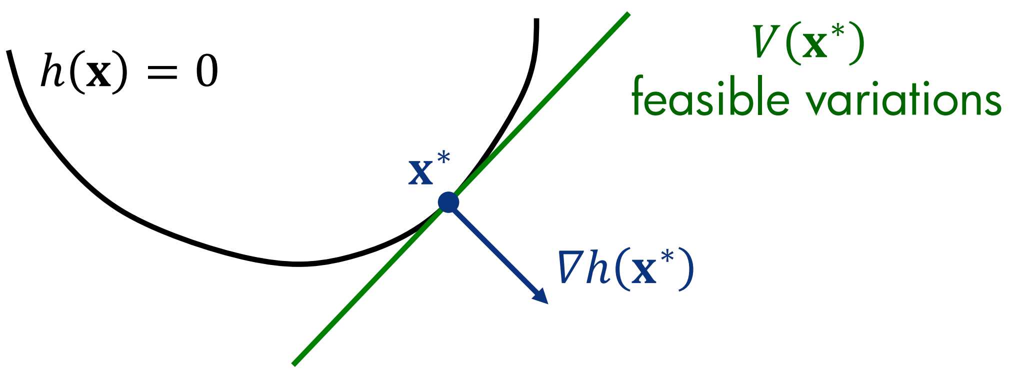


Alternative View

- consider the tangent space to the iso−contour $h(\mathbf{x}) = 0$
- this is the **subspace** of **first−order feasible variations**

$$\boxed{V(\mathbf{x}^*) = \left\{\Delta\mathbf{x} \mid \nabla h_i^T(\mathbf{x}^*)\,\Delta\mathbf{x} = 0, \forall i\right\}}$$

i.e., space of $\Delta\mathbf{x}$ for which a **step** $\mathbf{x} + \Delta\mathbf{x}$ satisfies the constraints $h_i(\mathbf{x})$ up to first−order approximation



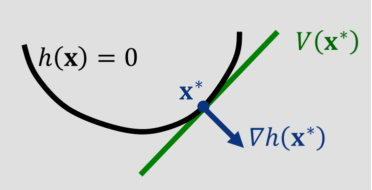
$$V(\mathbf{x}^*) = \{\Delta\mathbf{x} \mid \nabla h_i^T(\mathbf{x}^*)\,\Delta\mathbf{x} = 0, \forall i\}$$

Feasible Variations

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(\mathbf{x}^*) = 0$$

- multiplying our first Lagrangian condition by $\Delta\mathbf{x}$

$$\nabla f^T(\mathbf{x}^*)\,\Delta\mathbf{x} + \sum_{i=1}^{m} \lambda_i \nabla h_i^T(\mathbf{x}^*)\,\Delta\mathbf{x} = 0$$



- it follows that

$$\boxed{\nabla f^T(\mathbf{x}^*)\,\Delta\mathbf{x} = 0, \forall \Delta\mathbf{x} \in V(\mathbf{x}^*)}$$

- this is a **generalization** of $\boxed{\nabla f(\mathbf{x}^*) = 0}$ in the **unconstrained case**
  - here, all that matters is that $\nabla f(\mathbf{x}^*)$ has **no** projection in $V(\mathbf{x}^*)$
  - implies that $\nabla f(\mathbf{x}^*) \perp V(\mathbf{x}^*)$ and, therefore, $\nabla f(\mathbf{x}^*) \parallel \nabla h(\mathbf{x}^*)$
  - **note**:

$$\mathbf{y}^T\,\nabla^2_{\mathbf{xx}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*)\,\mathbf{y} \geq 0, \forall \mathbf{y} \text{ s.t. } \nabla h(\mathbf{x}^*)^T\mathbf{y} = 0$$

    - Hessian constraint **only** defined for $\mathbf{y}$ in $V(\mathbf{x}^*)$
    - explains the "extra stuff" in the Hessian condition (compared to **unconstrained**)
    - **makes sense**: we cannot move anywhere else – does not really matter what Hessian is outside $V(\mathbf{x}^*)$

Inequality Constraints

- what happens when we introduce inequalities?

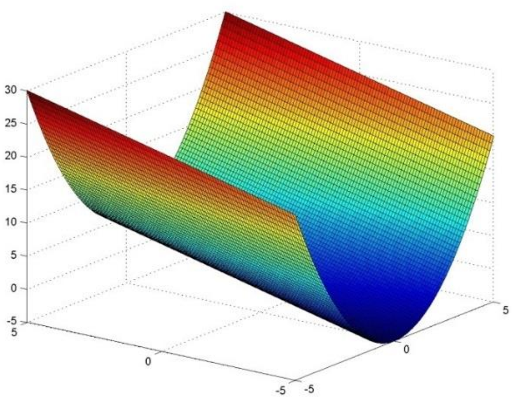
$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad h(\mathbf{x}) = 0, g(\mathbf{x}) \leq 0$$

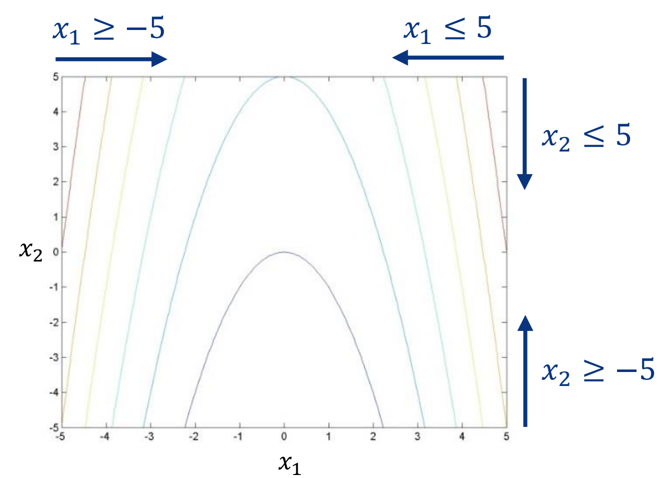
- we start by defining the set $A(\mathbf{x})$ of **active** inequality constraints

$$A(\mathbf{x}) = \{\, j \mid g_j(\mathbf{x}) = 0 \}$$

- example:

$$f(x_1, x_2) = {x_1}^2 + x_2, -5 \leq x_1 \leq 5, -5 \leq x_2 \leq 5$$





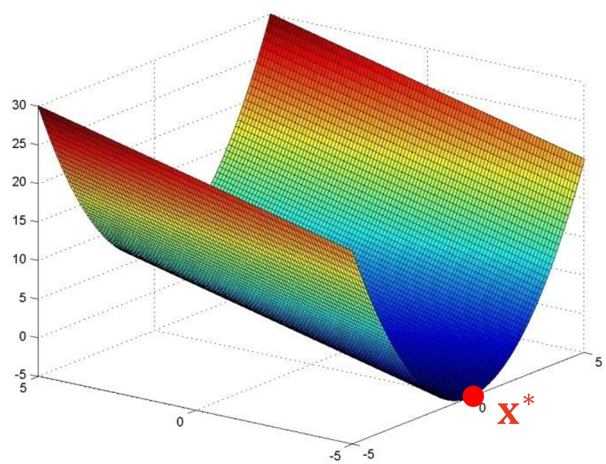
$f(x_1, x_2) = {x_1}^2 + x_2, -5 \le x_1 \le 5, -5 \le x_2 \le 5$

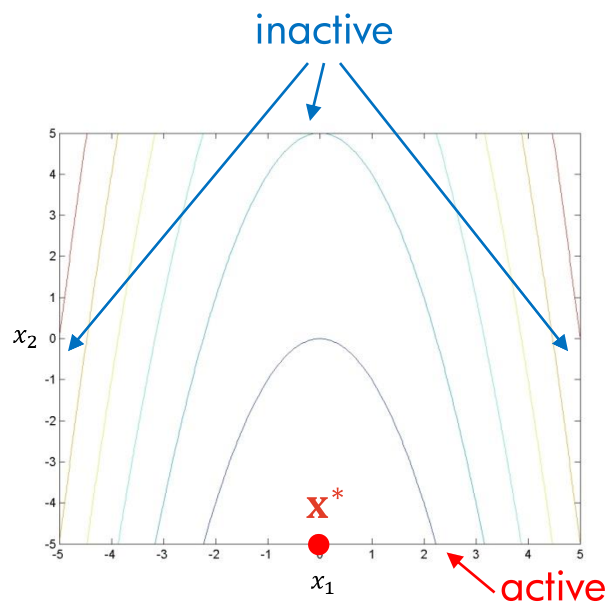
$A(\mathbf{x}) = \{ j \mid g_j(\mathbf{x}) = 0 \}$

Active Inequality Constraints

- we have a minimum at $\mathbf{x}^* = (0, -5)$
  - $x_1^* - 5 < 0, -x_1^* - 5 < 0$, and $x_2^* - 5 < 0$ are inactive
  - $-x_2^* - 5 = 0$ is active ($x_2^* = -5$)

- note that a local minimum for this problem would **still** be a local minimum if we **removed** the inactive constraints
  - inactive constraints do **not** do **anything**
  - active constraints are **equalities**





Constrained Optimization

- hence, the problem

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad h(\mathbf{x}) = 0, g(\mathbf{x}) \leq 0$$

- is <u>**equivalent**</u> to

$A(\mathbf{x}) = \{\, j \mid g_j(\mathbf{x}) = 0\}$

active constraints

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad h(\mathbf{x}) = 0, g_i(\mathbf{x}) = 0, \forall i \in A(\mathbf{x}^*)$$

- this is a problem with <u>equality constraints</u>: there must be a $\boldsymbol{\lambda}^*$ and μ_j^*, $j \in A(\mathbf{x}^*)$, such that

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(\mathbf{x}^*) + \sum_{j \in A(\mathbf{x}^*)} \mu_j^* \nabla g_j(\mathbf{x}^*) = 0$$

which does <u>**not**</u> **change** if we assign a <u>**zero**</u> Lagrange multiplier to the inactive constraints

$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$ subject to $h(\mathbf{x}) = 0, g_i(\mathbf{x}) \leq 0$

$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$ subject to $h(\mathbf{x}) = 0, g_i(\mathbf{x}) = 0, \forall i \in A(\mathbf{x}^*)$

Constrained Optimization

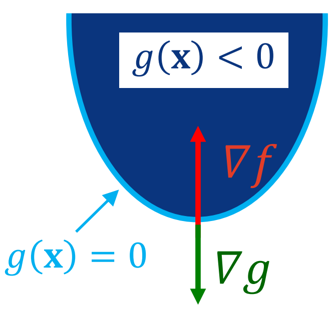
$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(\mathbf{x}^*) + \sum_{j \in A(\mathbf{x}^*)} \mu_j^* \nabla g_j(\mathbf{x}^*) = 0$$

- letting $\boxed{\mu_j^* = 0,\ j \notin A(x^*)}$ zero Lagrange multiplier for inactive constraints

$\mu_j^* = 0, j \notin A(x^*)$

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(\mathbf{x}^*) + \sum_{j=1}^{r} \mu_j^* \nabla g_j(\mathbf{x}^*) = 0$$

- there is **one** final **constraint**, which is $\boxed{\mu_j^* \geq 0,\ \forall j}$ due to the following picture



- ∇f has to point inward (otherwise, we would have a maximum of f)
- ∇g has to point outward (otherwise, g would increase inward, i.e. g would be non−negative inside)

- when we put all these together, we obtain the famed **Karush−Kuhn−Tucker (KKT) conditions**

W. Karush; Minima of Functions of Several Variables with Inequalities as Side Constraints. MS Dissertation. Dept. of Mathematics, Univ. of Chicago, 1939.

Kuhn, H.W.; Tucker, A.W.; Nonlinear programming. *Proceedings of 2nd Berkeley Symposium,* 1951.

The Karush–Kuhn–Tucker (KKT) Conditions

- **Theorem:** for the problem

$$\boxed{\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad h(\mathbf{x}) = 0, g(\mathbf{x}) \leq 0}$$

$\mathbf{x}^*$ is a local minimum if and only if there exist $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^*$ such that

i) $$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(\mathbf{x}^*) + \sum_{j=1}^{r} \mu_j^* \nabla g_j(\mathbf{x}^*) = 0$$

these conditions would be the <u>same</u> if all constraints were equalities

ii) $\mu_j^* \geq 0, \ \forall j$ condition on <u>all</u> inequality constraints

iii) $\mu_j^* = 0, \ \forall j \notin A(\mathbf{x}^*)$ this condition eliminates <u>inactive</u> constraints

iv) $h(\mathbf{x}^*) = 0$

v) $$\mathbf{y}^T \nabla \left[\nabla f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(\mathbf{x}) + \sum_{j=1}^{r} \mu_j^* \nabla g_j(\mathbf{x})\right]_{\mathbf{x}=\mathbf{x}^*} \mathbf{y} \geq 0, \forall \mathbf{y} \in V(\mathbf{x}^*)$$

where $V(\mathbf{x}^*) = \{\mathbf{y} \mid \nabla h_i^T(\mathbf{x}^*)\mathbf{y} = 0, \forall i \text{ and } \nabla g_j^T(\mathbf{x}^*)\mathbf{y} = 0, \forall j \in A(\mathbf{x}^*)\}$

Geometric Interpretation

- let's <u>forget</u> the **equality constraints for now**
- later, we will see that they do <u>not</u> change much

- consider the problem

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to } g(\mathbf{x}) \le 0$$

- from the **KKT conditions**, the **solution** satisfies

i) $\nabla L(\mathbf{x}^*, \boldsymbol{\mu}^*) = 0$

ii) $\mu_j^* \ge 0, \ \forall j$

iii) $\mu_j^* = 0, \forall j \notin A(\mathbf{x}^*)$

$\rightarrow$

this implies that

$\mu_j^* g_j(\mathbf{x}^*) = 0, \forall j$

active: $g_j(\mathbf{x}^*) = 0$

inactive: $\mu_j^* = 0$

and

$$L(\mathbf{x}^*, \boldsymbol{\mu}^*) = f(\mathbf{x}^*) + \sum_{j=1}^{r} \mu_j^* g_j(\mathbf{x}^*) = f(\mathbf{x}^*)$$

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to } g(\mathbf{x}) \leq 0$$

$$L(\mathbf{x}^*, \boldsymbol{\mu}^*) = f(\mathbf{x}^*) + \sum_{j=1}^{r} \mu_j^* g_j(\mathbf{x}^*)$$

i) $\nabla[L(\mathbf{x}^*, \boldsymbol{\mu}^*)] = 0$

ii) $\mu_j^* \geq 0, \ \forall j$

iii) $\mu_j^* = 0, \forall j \notin A(\mathbf{x}^*)$

Geometric Interpretation

- which is equivalent to

$$\boxed{\begin{array}{c} L^* = \min_{\mathbf{x}}[L(\mathbf{x}, \boldsymbol{\mu}^*)] = \min_{\mathbf{x}}[f(\mathbf{x}) + (\boldsymbol{\mu}^*)^T g(\mathbf{x})] \\ \text{with } \mu_j^* \geq 0, \forall j \text{ and } \mu_j^* = 0, \forall j \notin A(\mathbf{x}^*) \end{array}}$$

- we thus have
  - $\mathbf{x} = \mathbf{x}^* \Rightarrow f(\mathbf{x}) + (\boldsymbol{\mu}^*)^T g(\mathbf{x}) - L^* = 0$
  - $\mathbf{x} \neq \mathbf{x}^* \Rightarrow f(\mathbf{x}) + (\boldsymbol{\mu}^*)^T g(\mathbf{x}) - L^* \geq 0$
- or
  - $\mathbf{x} = \mathbf{x}^* \Rightarrow (\mathbf{w}^*)^T \mathbf{z} - b = 0$
  - $\mathbf{x} \neq \mathbf{x}^* \Rightarrow (\mathbf{w}^*)^T \mathbf{z} - b \geq 0$

plane in z – space
normal $\mathbf{w}^*$, bias b

$\mathbf{z}$ is in half–space
pointed to by $\mathbf{w}^*$

where

$$b = L^* \qquad \mathbf{w}^* = \begin{bmatrix} 1 \\ \boldsymbol{\mu}^* \end{bmatrix} \qquad \mathbf{z} = \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}$$

$\mu_j^* g_j(\mathbf{x}^*) = 0, \forall j$

active: $g_j(\mathbf{x}^*) = 0$

inactive: $\mu_j^* = 0$

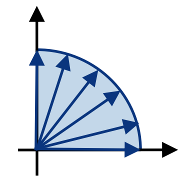
$\mathbf{x} = \mathbf{x}^* \Rightarrow (\mathbf{w}^*)^T \mathbf{z} - b = 0$

$\mathbf{x} \neq \mathbf{x}^* \Rightarrow (\mathbf{w}^*)^T \mathbf{z} - b \geq 0$

Geometric Interpretation

- from

$$b = L^* \qquad \mathbf{w}^* = \begin{bmatrix} 1 \\ \boldsymbol{\mu}^* \end{bmatrix} \qquad \mathbf{z} = \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}$$



- we have
  - since $\mu_j^* \geq 0, \forall j$, $\mathbf{w}^*$ is always in the **first quadrant** of $\mathbf{z}$ – space
  - since first coordinate is 1, $\mathbf{w}^*$ is **never parallel** to $g(\mathbf{x})$ –"axis"
- can be visualized in "z – space" as
- also, two cases:

active constraints

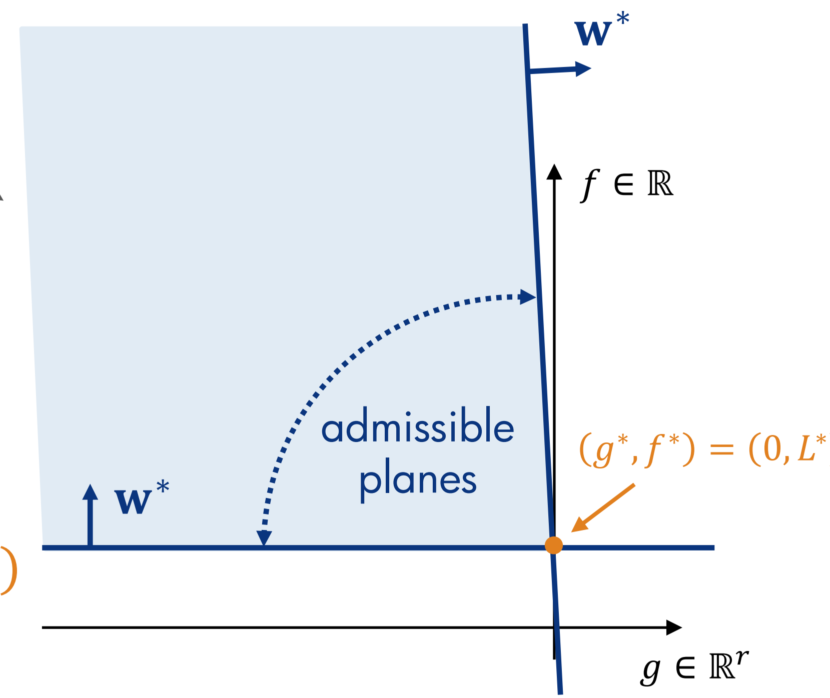
case 1) $g(\mathbf{x}^*) = 0$

- $\mathbf{x} = \mathbf{x}^* \Rightarrow (\mathbf{w}^*)^T \mathbf{z} - b = 0$

$\mathbf{z} = \begin{bmatrix} f(\mathbf{x}) \\ 0 \end{bmatrix}$

$$\Rightarrow f(\mathbf{x}^*) = b = L^*$$

- the f – intercept is $(0, L^*) = (0, f^*)$ and is the **minimum** of $L(\mathbf{x}, \boldsymbol{\mu}^*)$



$\mu_j^* g_j(\mathbf{x}^*) = 0, \forall j$

active: $g_j(\mathbf{x}^*) = 0$
inactive: $\mu_j^* = 0$

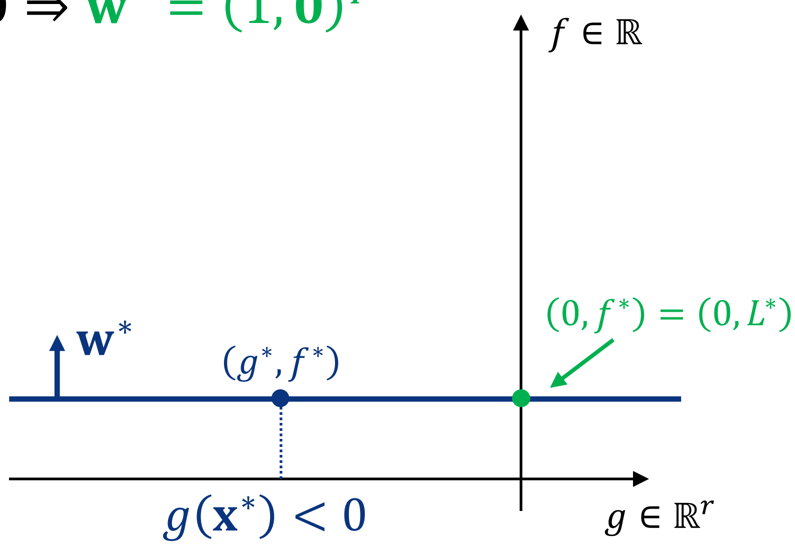
$\mathbf{x} = \mathbf{x}^* \Rightarrow (\mathbf{w}^*)^T \mathbf{z} - b = 0$
$\mathbf{x} \neq \mathbf{x}^* \Rightarrow (\mathbf{w}^*)^T \mathbf{z} - b \geq 0$

Geometric Interpretation

$$b = L^* \quad \mathbf{w}^* = \begin{bmatrix} 1 \\ \boldsymbol{\mu}^* \end{bmatrix} \quad \mathbf{z} = \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}$$

inactive constraints

- case 2) $g(\mathbf{x}^*) < 0$
  - the constraints are inactive $\Rightarrow \boldsymbol{\mu}^* = \mathbf{0} \Rightarrow \mathbf{w}^* = (1, \mathbf{0})^T$
  - plane is "horizontal"
  - $\mathbf{x} = \mathbf{x}^* \Rightarrow (\mathbf{w}^*)^T \mathbf{z} - b = 0$
    $\Rightarrow f(\mathbf{x}^*) = b = L^*$
    $\mathbf{w}^* = \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix}$
  - the $f-$intercept is $(0, L^*) = (0, f^*)$ and is the **minimum** of $L(\mathbf{x}, \boldsymbol{\mu}^*)$



- in **both cases**, the $f-$intersect is $(0, L^*)$
- in general, <u>mix</u> of active and inactive but behavior is <u>one of these two</u>

In Summary

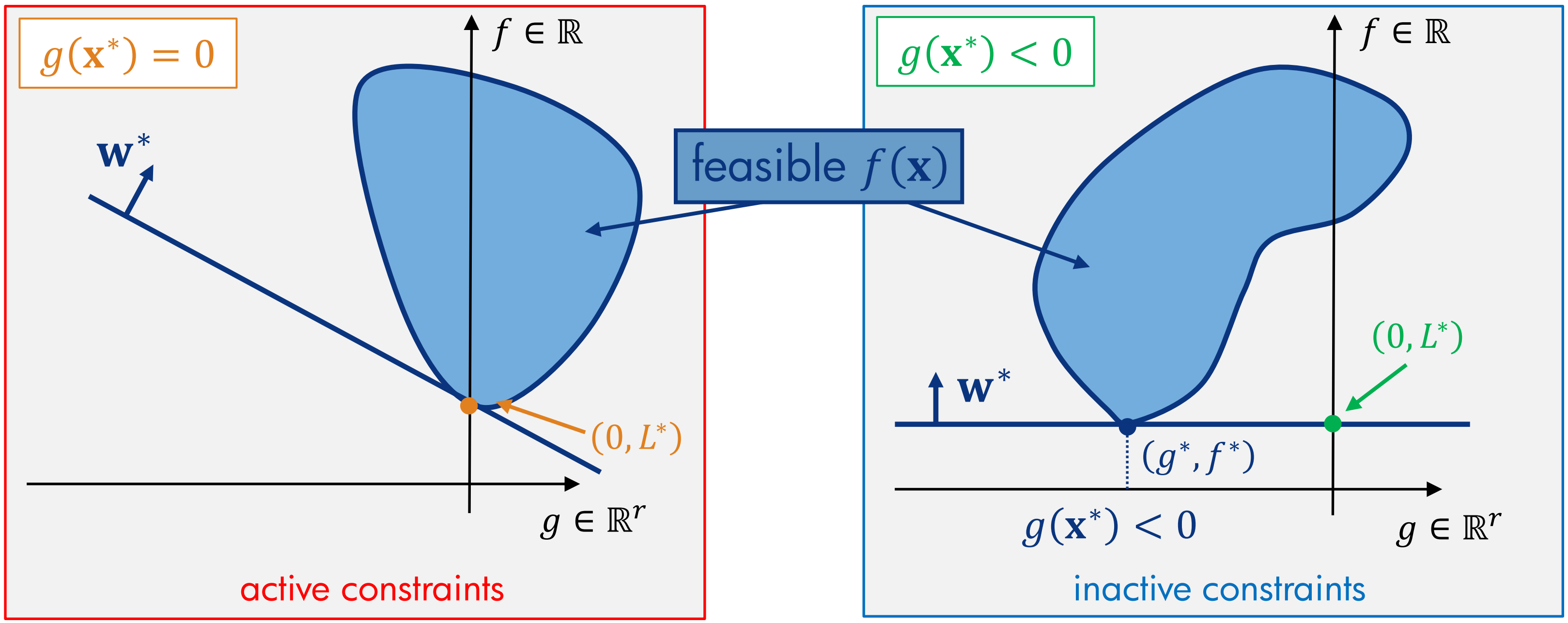
$$L^* = \min_{\mathbf{x}}[L(\mathbf{x}, \boldsymbol{\mu}^*)] = \min_{\mathbf{x}}[f(\mathbf{x}) + (\boldsymbol{\mu}^*)^T g(\mathbf{x})]$$

with $\mu_j^* \geq 0,\ \forall j$ and $\mu_j^* = 0, \forall j \notin A(\mathbf{x}^*)$

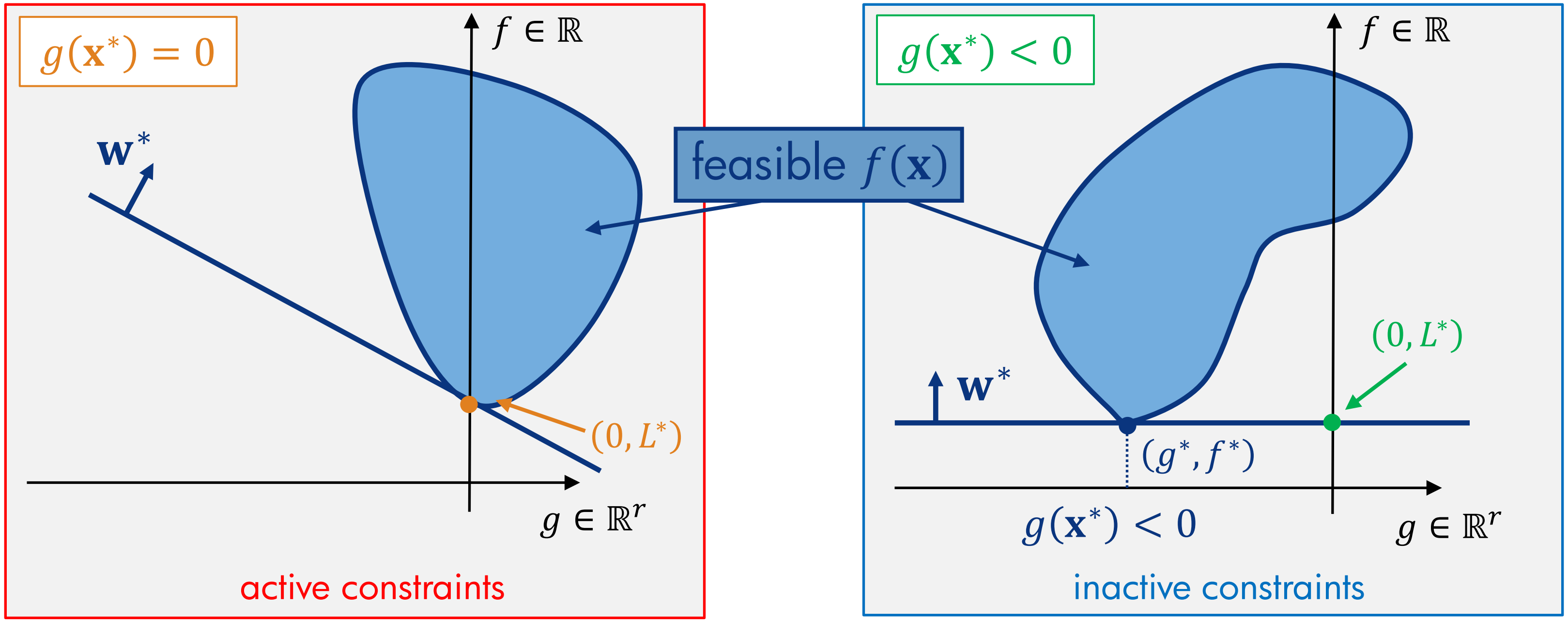
- is equivalent to
  - $\mathbf{x} = \mathbf{x}^* \Rightarrow (\mathbf{w}^*)^T \mathbf{z} - b = 0$
  - $\mathbf{x} \neq \mathbf{x}^* \Rightarrow (\mathbf{w}^*)^T \mathbf{z} - b \geq 0$

$$b = L^* \quad \mathbf{w}^* = \begin{bmatrix} 1 \\ \boldsymbol{\mu}^* \end{bmatrix} \quad \mathbf{z} = \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}$$

- can be visualized as



In Summary



- in <u>both cases</u>, the **plane with normal $\mathbf{w}^*$**
  - goes through $(0, L^*)$
  - supports the feasible set of $f(\mathbf{x})$
- the <u>difference</u> is the **direction of $\mathbf{w}^*$** and **what the feasible set needs to look like**
  - in one case (active), the point of support is in the f – axis
  - in the other (inactive), it is not

Duality

$$L^* = \min_{\mathbf{x}}[L(\mathbf{x}, \boldsymbol{\mu}^*)] = \min_{\mathbf{x}}[f(\mathbf{x}) + (\boldsymbol{\mu}^*)^T g(\mathbf{x})]$$

with $\mu_j^* \geq 0, \forall j$ and $\mu_j^* = 0, \forall j \notin A(\mathbf{x}^*)$

- does not appear terribly difficult <u>**once we know $\boldsymbol{\mu}^*$**</u>
- but **how do I find the value of $\boldsymbol{\mu}^*$?** Consider the function $q(\boldsymbol{\mu}), \forall \boldsymbol{\mu} \geq \mathbf{0}$

$$q(\boldsymbol{\mu}) = \min_{\mathbf{x}}[L(\mathbf{x}, \boldsymbol{\mu})] = \min_{\mathbf{x}}[f(\mathbf{x}) + \boldsymbol{\mu}^T g(\mathbf{x})]$$

with $\boldsymbol{\mu} \geq \mathbf{0}$

- this is equivalent to
  - $\mathbf{x} = \mathbf{x}^* \Rightarrow \mathbf{w}^T\mathbf{z} - b = 0$
  - $\mathbf{x} \neq \mathbf{x}^* \Rightarrow \mathbf{w}^T\mathbf{z} - b \geq 0$

$$b = q(\boldsymbol{\mu}) \quad \mathbf{w} = \begin{bmatrix} 1 \\ \boldsymbol{\mu} \end{bmatrix} \quad \mathbf{z} = \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}$$

- the picture is the same as before with

$\boldsymbol{\mu}^*$ replaced by $\boldsymbol{\mu}$ and L^* replaced by $q(\boldsymbol{\mu})$

Duality

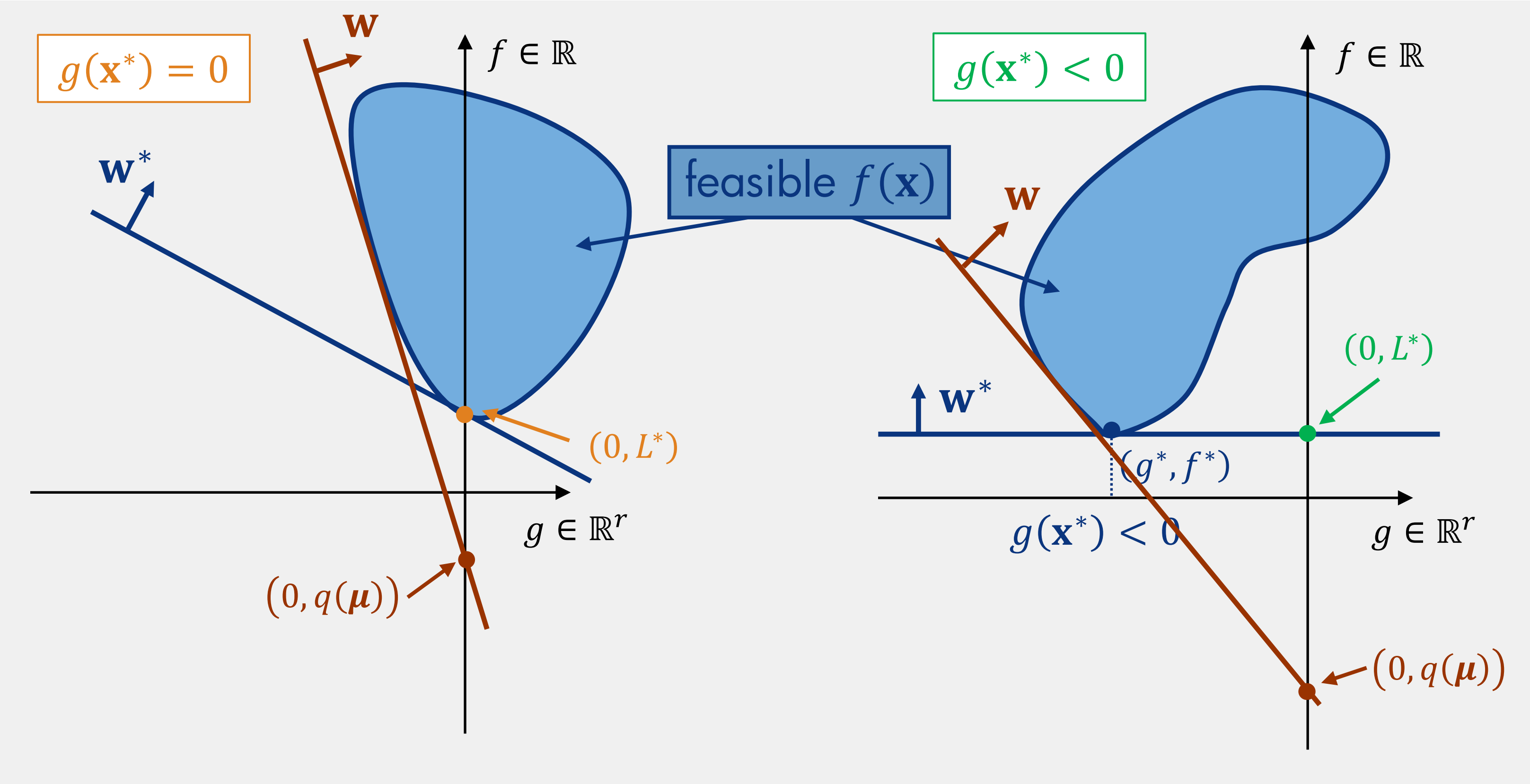
$$q(\boldsymbol{\mu}) = \min_{\mathbf{x}}[L(\mathbf{x},\boldsymbol{\mu})] = \min_{\mathbf{x}}[f(\mathbf{x}) + \boldsymbol{\mu}^T g(\mathbf{x})]$$
with $\boldsymbol{\mu} \geq \mathbf{0}$

$\mathbf{x} = \mathbf{x}^* \Rightarrow \mathbf{w}^T\mathbf{z} - b = 0$
$\mathbf{x} \neq \mathbf{x}^* \Rightarrow \mathbf{w}^T\mathbf{z} - b \geq 0$

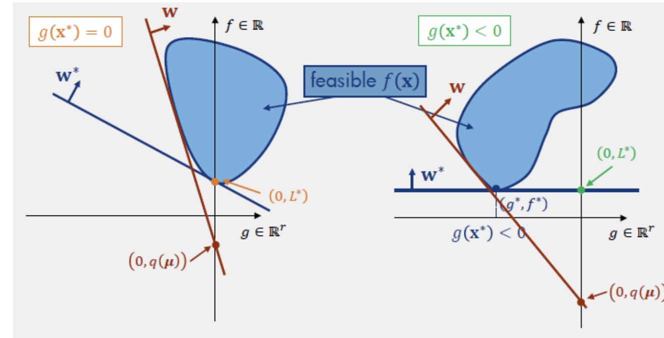
$b = q(\boldsymbol{\mu}) \quad \mathbf{w} = \begin{bmatrix} 1 \\ \boldsymbol{\mu} \end{bmatrix} \quad \mathbf{z} = \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}$

- noting that
  - hyperplane $(\mathbf{w}, b)$ still has to support the set of feasible $f(\mathbf{x})$
  - we still have $\boldsymbol{\mu} \geq \mathbf{0}$

this leads to



Duality



- note that
  - $q(\boldsymbol{\mu}) \leq L^* = f^*$
  - if we keep increasing $q(\boldsymbol{\mu})$, we will get $q(\boldsymbol{\mu}) = L^*$
  - we cannot go beyond L^*

- this is exactly the definition of the **dual problem**

$$\max_{\boldsymbol{\mu} \geq \mathbf{0}} q(\boldsymbol{\mu}) \qquad q(\boldsymbol{\mu}) = \min_{\mathbf{x}}[L(\mathbf{x}, \boldsymbol{\mu})] = \min_{\mathbf{x}}[f(\mathbf{x}) + \boldsymbol{\mu}^T g(\mathbf{x})]$$

$q(\boldsymbol{\mu})$ – Lagrangian dual function

- **note**:
  - $q(\boldsymbol{\mu})$ may go to $-\infty$ for some $\boldsymbol{\mu}$, which means that there is **<u>no</u> Lagrange multiplier (plane would be vertical)**
  - this is avoided by introducing the **constraint**

$$\boldsymbol{\mu} \in D_q = \{\boldsymbol{\mu} \mid q(\boldsymbol{\mu}) > -\infty\}$$

Duality

- Therefore, we have a **two–step** recipe to find the optimal solution

1. for any $\boldsymbol{\mu}$, solve

$$\boxed{q(\boldsymbol{\mu}) = \min_{\mathbf{x}}[L(\mathbf{x},\boldsymbol{\mu})] = \min_{\mathbf{x}}[f(\mathbf{x}) + \boldsymbol{\mu}^T g(\mathbf{x})]}$$

2. then solve

$$\boxed{\max_{\boldsymbol{\mu}\geq \mathbf{0},\, \boldsymbol{\mu}\in D_q} q(\boldsymbol{\mu})} \qquad D_q = \{\boldsymbol{\mu} \mid q(\boldsymbol{\mu}) > -\infty\}$$

- one of the reasons why this is interesting is that 2. turns out to be quite manageable (we will see why)
- 2. is called the dual problem
- 1. is similar to the Lagrangian of an equality constraint problem, but **easier** because we do not need to solve for $\boldsymbol{\mu}$

Duality

- **Theorem:** D_q is a convex set and $q(\boldsymbol{\mu})$ is concave on D_q.

- Proof:
  - for any $\mathbf{x}$, $\boldsymbol{\mu}$, $\overline{\boldsymbol{\mu}}$, and $\alpha \in [0,1]$

$L(\mathbf{x}, \boldsymbol{\mu}) = f(\mathbf{x}) + \boldsymbol{\mu}^T g(\mathbf{x})$

$$\begin{aligned}
L(\mathbf{x}, \alpha\boldsymbol{\mu} + (1-\alpha)\overline{\boldsymbol{\mu}}) &= f(\mathbf{x}) + (\alpha\boldsymbol{\mu} + (1-\alpha)\overline{\boldsymbol{\mu}})^T g(\mathbf{x}) \\
&= f(\mathbf{x}) + \alpha\boldsymbol{\mu}^T g(\mathbf{x}) + (1-\alpha)\overline{\boldsymbol{\mu}}^T g(\mathbf{x}) \\
&= \alpha[f(\mathbf{x}) + \boldsymbol{\mu}^T g(\mathbf{x})] + (1-\alpha)[f(\mathbf{x}) + \overline{\boldsymbol{\mu}}^T g(\mathbf{x})] \\
&= \alpha\, L(\mathbf{x}, \boldsymbol{\mu}) + (1-\alpha) L(\mathbf{x}, \overline{\boldsymbol{\mu}})
\end{aligned}$$

and taking the minimum on both sides

$$\begin{aligned}
\min_{\mathbf{x}} L(\mathbf{x}, \alpha\boldsymbol{\mu} + (1-\alpha)\overline{\boldsymbol{\mu}}) &= \min_{\mathbf{x}}[\alpha\, L(\mathbf{x}, \boldsymbol{\mu}) + (1-\alpha) L(\mathbf{x}, \overline{\boldsymbol{\mu}})] \\
&\geq \alpha \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\mu}) + (1-\alpha) \min_{\mathbf{x}} L(\mathbf{x}, \overline{\boldsymbol{\mu}})
\end{aligned}$$

we have

$$q(\alpha\boldsymbol{\mu} + (1-\alpha)\overline{\boldsymbol{\mu}}) \geq \alpha q(\boldsymbol{\mu}) + (1-\alpha) q(\overline{\boldsymbol{\mu}})$$

$q(\boldsymbol{\mu}) = \min_{\mathbf{x}}[L(\mathbf{x}, \boldsymbol{\mu})]$

Duality

Recall:

Definition: A set Ω is convex if

$\forall\, \mathbf{w}, \mathbf{u} \in \Omega$ and $\lambda \in [0,1]$ then $\lambda \mathbf{w} + (1-\lambda)\mathbf{u} \in \Omega$

Definition: $f(\mathbf{w})$ is concave if $\forall\, \mathbf{w}, \mathbf{u} \in \Omega$ and $\lambda \in [0,1]$

$f(\lambda \mathbf{w} + (1-\lambda)\mathbf{u}) \geq \lambda f(\mathbf{w}) + (1-\lambda) f(\mathbf{u})$

$D_q = \{\boldsymbol{\mu} \mid q(\boldsymbol{\mu}) > -\infty\}$

- we have

$$q(\alpha\boldsymbol{\mu} + (1-\alpha)\bar{\boldsymbol{\mu}}) \geq \alpha q(\boldsymbol{\mu}) + (1-\alpha) q(\bar{\boldsymbol{\mu}}) \quad (*)$$

- from which two conclusions follow
  - if $\boldsymbol{\mu} \in D_q$ and $\bar{\boldsymbol{\mu}} \in D_q \Rightarrow q(\boldsymbol{\mu}) > -\infty, q(\bar{\boldsymbol{\mu}}) > -\infty \Rightarrow$ $\alpha\boldsymbol{\mu} + (1-\alpha)\bar{\boldsymbol{\mu}} \in D_q \Rightarrow D_q$ is convex
  - by definition of concavity, $(*)$ implies that q is concave over D_q ■

- note that the dual is <u>always</u> concave, <u>irrespective</u> of the primal optimization problem

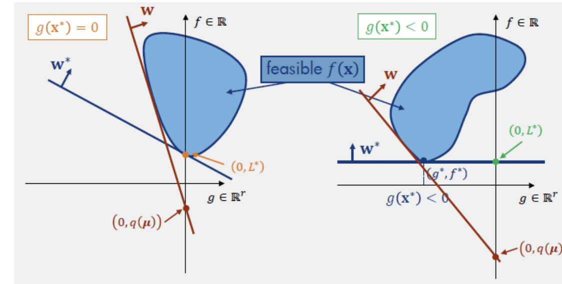
- $\max_{\boldsymbol{\mu} \geq \mathbf{0},\, \boldsymbol{\mu} \in D_q} q(\boldsymbol{\mu})$

  <u>dual problem is always concave</u> (even if **primal problem is not**) → very appealing result since convex optimization problems are among the <u>**easiest**</u> to solve

- the <u>next result</u> only proves what we already have inferred from the geometric interpretation

Duality

weak duality:
maximum of the dual is <u>**never**</u> larger than the minimum of the primal



- **Theorem:** (weak duality) it is always true that

$$\boxed{q^* \leq f^*}$$

- **Proof:**
  - for <u>any</u> $\boldsymbol{\mu} \geq \mathbf{0}$ and $\mathbf{x}$ with $g(\mathbf{x}) \leq 0$, since $\mu_j g_j(\mathbf{x}) \leq 0$, $\forall j$,

$$q(\boldsymbol{\mu}) = \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\mu}) \leq f(\mathbf{x}) + \sum_j \mu_j g_j(\mathbf{x}) \leq f(\mathbf{x})$$

$$\begin{aligned} q(\boldsymbol{\mu}) &= \min_{\mathbf{x}}[L(\mathbf{x}, \boldsymbol{\mu})] \\ &= \min_{\mathbf{x}}[f(\mathbf{x}) + \boldsymbol{\mu}^T g(\mathbf{x})] \end{aligned}$$

  - Hence,

$$q^* = \max_{\boldsymbol{\mu} \geq \mathbf{0}} q(\boldsymbol{\mu}) \leq f(\mathbf{x})$$

  - and, since this holds for any $\mathbf{x}$,

$$q^* \leq \min_{\mathbf{x},\, g(\mathbf{x}) \leq 0} f(\mathbf{x}) \quad \blacksquare$$

Duality Gap

- we say that
  - if $q^* = f^*$, there is <u>no</u> **duality gap**
  - otherwise, there is a **duality gap**
- the duality gap <u>constrains</u> the existence of Lagrange multipliers
- **Theorem:**
  - if there is **<u>no</u> duality gap**, the set of Lagrange multipliers is the set of optimal dual solutions;
  - if there is a **duality gap**, there are **<u>no</u>** Lagrange multipliers.
- **Proof:**
  - by definition, $\boldsymbol{\mu}^* \geq \mathbf{0}$ is a Lagrange multiplier if and only if

$$f^* = q(\boldsymbol{\mu}^*) \leq q^*$$

which, from the previous theorem, holds if and only if $q^* = f^*$, i.e. if there is no duality gap ■

Duality Gap

primal:			dual:		
$b = L^*$	$\mathbf{w}^* = \begin{bmatrix} 1 \\ \boldsymbol{\mu}^* \end{bmatrix}$	$\mathbf{z} = \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}$	$b = q(\boldsymbol{\mu})$	$\mathbf{w} = \begin{bmatrix} 1 \\ \boldsymbol{\mu} \end{bmatrix}$	$\mathbf{z} = \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}$

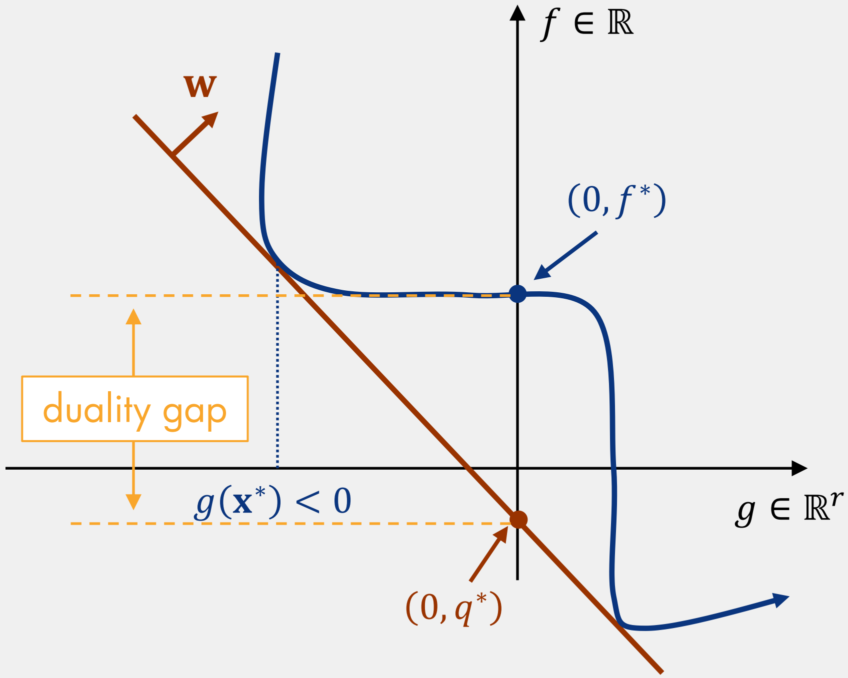
- note that there are situations in which the **dual problem** has a **solution**, but for which there is <u>no</u> Lagrange multiplier

- example:
  - this is a <u>valid</u> dual problem
  - however, the constraint

$$\mu_i^* g(\mathbf{x}_i^*) = 0$$

is not satisfied† and

$$\boxed{q^* \neq f^*}$$



† $g(\mathbf{x}^*) < 0$ but $\boldsymbol{\mu} \neq \mathbf{0}$ because $\mathbf{w} \neq (1, \mathbf{0})^T$

- in summary, duality is interesting <u>only</u> when there is <u>no</u> duality gap