

Lecture 7: Optimization

Yi, Yung (이웅)

Mathematics for Machine Learning

<https://yung-web.github.io/home/courses/mathml.html>

KAIST EE

April 7, 2021

- (1) Optimization Using Gradient Descent
- (2) Constrained Optimization and Lagrange Multipliers
- (3) Convex Sets and Functions
- (4) Convex Optimization
- (5) Convex Conjugate

- Training machine learning models = finding a good set of parameters
- A good set of parameters = Solution (or close to solution) to some optimization problem
- Directions: Unconstrained optimization, Constrained optimization, Convex optimization
- High-school math: A necessary condition for the optimal point: $f'(x) = 0$ (stationary point)
 - Gradient will play an important role

- (1) Optimization Using Gradient Descent
- (2) Constrained Optimization and Lagrange Multipliers
- (3) Convex Sets and Functions
- (4) Convex Optimization
- (5) Convex Conjugate

- Goal

$$\min f(\mathbf{x}), \quad f(\mathbf{x}) : \mathbb{R}^n \mapsto \mathbb{R}, \quad f \in C^1$$

- Gradient-type algorithms

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \gamma_k \mathbf{d}_k, \quad k = 0, 1, 2, \dots$$

- **Lemma.** Any direction $\mathbf{d} \in \mathbb{R}^{n \times 1}$ that satisfies $\nabla f(\mathbf{x}) \cdot \mathbf{d} < 0$ is a descent direction of f at \mathbf{x} . That is, if we let $\mathbf{x}_\alpha = \mathbf{x} + \alpha \mathbf{d}$, $\exists \bar{\alpha} > 0$, such that for all $\alpha \in (0, \bar{\alpha}]$, $f(\mathbf{x}_\alpha) < f(\mathbf{x})$.
- Steepest gradient descent¹. $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)^\top$.
- Finding a local optimum $f(\mathbf{x}_*)$, if the step-size γ_k is suitably chosen.
- **Question.** How do we choose \mathbf{d}_k for a constrained optimization?

¹In some cases, just gradient descent often means this steepest gradient descent.

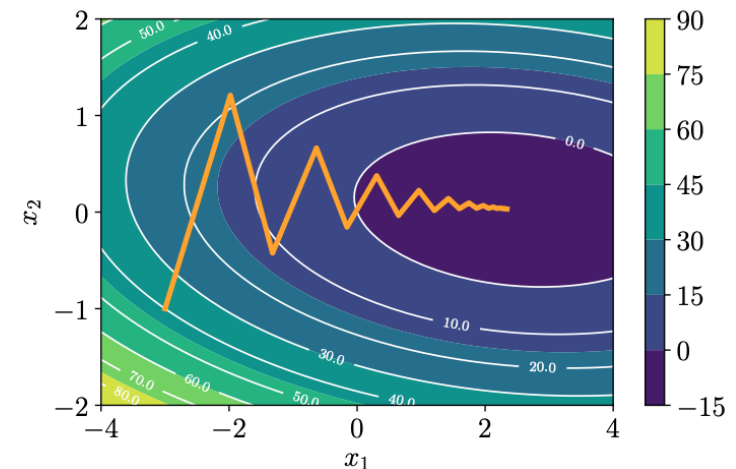
Example

- A quadratic function $f : \mathbb{R}^2 \mapsto \mathbb{R}$.

$$f \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right) = \frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} 2 & 1 \\ 1 & 20 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} 5 \\ 3 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

whose gradient is $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} 2 & 1 \\ 1 & 20 \end{pmatrix} - \begin{pmatrix} 5 \\ 3 \end{pmatrix}^T$

- $\mathbf{x}_0 = (-3 - 1)^T$
- constant step size $\alpha = 0.085$
- Zigzag pattern



- Goal: $\min L(\theta)$ for n training data
- Based on the **amount of training data** used for **each** iteration
 - Batch gradient descent (the entire n)
 - Mini-batch gradient descent ($k < n$ data)
 - Stochastic gradient descent (one sampled data)
- Based on the adaptive method of update
 - Momentum, NAG, Adagrad, RMSprop, Adam, etc
- <https://ruder.io/optimizing-gradient-descent/>

- Assume $L(\boldsymbol{\theta}) = \sum_{i=1}^n L_n(\boldsymbol{\theta})$ (which happens in many cases in machine learning, e.g., negative log-likelihood in regression)
- Gradient update

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \gamma_k \nabla L(\boldsymbol{\theta}_k)^\top = \boldsymbol{\theta}_k - \gamma_k \sum_{n=1}^N \nabla L_n(\boldsymbol{\theta}_k)^\top$$

- Batch gradient: $\sum_{n=1}^N \nabla L_n(\boldsymbol{\theta}_k)^\top$
 - Mini-batch gradient: $\sum_{n \in \mathcal{K}} \nabla L_n(\boldsymbol{\theta}_k)^\top$ for a suitable choice of \mathcal{K} , $|\mathcal{K}| < n$
 - Stochastic gradient: $\nabla L_n(\boldsymbol{\theta}_i)^\top$ for some (randomly chosen) i . Noisy approximation to the real gradient.
- Tradeoff: computation burden vs. exactness

- Step size.
 - Too small: slow update, Too big: overshoot, zig-zag, often fail to converge
- Adaptive update: smooth out the erratic behavior and dampens oscillations
- Gradient descent with **momentum**

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma_i \nabla f(\mathbf{x}_k)^\top + \alpha \Delta \mathbf{x}_k, \quad \alpha \in [0, 1]$$

$$\Delta \mathbf{x}_k = \mathbf{x}_k - \mathbf{x}_{k-1}$$

- Memory term: $\alpha \Delta \mathbf{x}_k$, where α is the degree of how much we remember the past
- Next update = a linear combination of current and previous updates

- (1) Optimization Using Gradient Descent
- (2) **Constrained Optimization and Lagrange Multipliers**
- (3) Convex Sets and Functions
- (4) Convex Optimization
- (5) Convex Conjugate

- An optimization problem in standard form:

minimize $f(\mathbf{x})$

subject to $g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \dots, m$ (*Inequality constraints*)

$h_j(\mathbf{x}) = 0, \quad j = 1, 2, \dots, p$ (*Equality constraints*)

- Variables: $\mathbf{x} \in \mathbb{R}^n$. Assume nonempty feasible set
- Optimal value: p^* . Optimizer: \mathbf{x}^*

- Duality Mentality
 - Bound or solve an optimization problem via a different optimization problem!
 - We'll develop the basic Lagrange duality theory for a general optimization problem, then specialize for convex optimization

- Idea: augment the objective with a weighted sum of constraints

- Lagrangian:

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

- Lagrange multipliers (dual variables): $\boldsymbol{\lambda} = (\lambda_i : i = 1, \dots, m) \succeq 0$, $\boldsymbol{\nu} = (\nu_1, \dots, \nu_p)$

- Lagrange dual function:

$$\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

- The dual function $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is the **lower bound** on the optimal value p^* .
- **Theorem.** $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*$, $\forall \boldsymbol{\lambda} \succeq 0, \boldsymbol{\nu}$
- **Proof.** Consider feasible $\tilde{\mathbf{x}}$. Then,

$$\mathcal{L}(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\tilde{\mathbf{x}}) + \sum_{i=1}^m \lambda_i g_i(\tilde{\mathbf{x}}) + \sum_{i=1}^p \nu_i h_i(\tilde{\mathbf{x}}) \leq f(\tilde{\mathbf{x}})$$

since $f_i(\tilde{\mathbf{x}}) \leq 0$ and $\lambda_i \geq 0$.

Hence, $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \mathcal{L}(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f(\tilde{\mathbf{x}})$ for all feasible $\tilde{\mathbf{x}}$. Therefore, $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*$.

- What's the relationship between d^* and p^* ?

Weak Duality

$$d^* \leq p^*$$

- Weak duality **always** hold (even if the primal problem is not convex):
- Optimal duality gap: $p^* - d^*$
- Efficient generation of the lower bounds through the dual problem

- (1) Optimization Using Gradient Descent
- (2) Constrained Optimization and Lagrange Multipliers
- (3) **Convex Sets and Functions**
- (4) Convex Optimization
- (5) Convex Conjugate

- Convex optimization problem

minimize $f(\mathbf{x})$

subject to $\mathbf{x} \in \mathcal{X}$,

where $f(\mathbf{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ is a convex function, and \mathcal{X} is a convex set.

- The watershed between easily solvable problem and intractable ones is not 'linearity', but '**convexity**'
- Let's overview the background of convex functions, convex sets, and their basic properties.

- Set \mathcal{C} is a **convex set** if the line segment between any two points in \mathcal{C} lies in \mathcal{C} , i.e., if for any $x_1, x_2 \in \mathcal{C}$ and any $\theta \in [0, 1]$, we have $\theta x_1 + (1 - \theta)x_2 \in \mathcal{C}$
- **Convex hull** of \mathcal{C} is the set of all **convex combinations** of points in \mathcal{C} :

$$\left\{ \sum_{i=1}^k \theta_i x_i \mid x_i \in \mathcal{C}, \theta_i \geq 0, i = 1, 2, \dots, k, \sum_{i=1}^k \theta_i = 1 \right\}$$

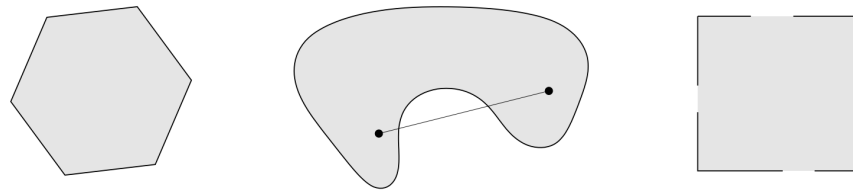
- What is k ? For all k ? For some k ?
- Generalize to infinite sums and integrals:

$$\sum_{i=1}^{\infty} \theta_i x_i \in \mathcal{C}, \quad \int_{\mathcal{C}} p(x) x dx \in \mathcal{C},$$

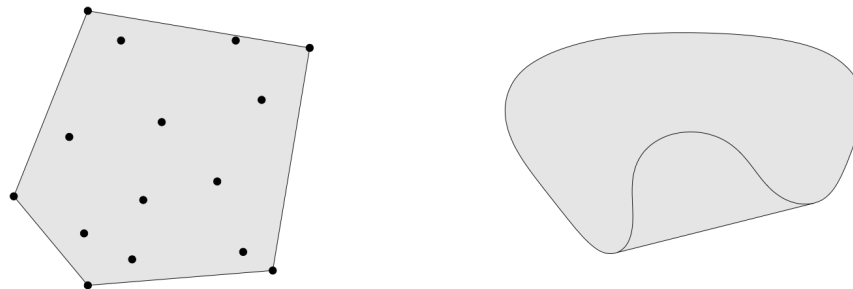
where $\sum_{i=1}^{\infty} \theta_i = 1$ and $p(x)$ is a pdf of some random variable.

Examples

- Convex and Non-convex sets

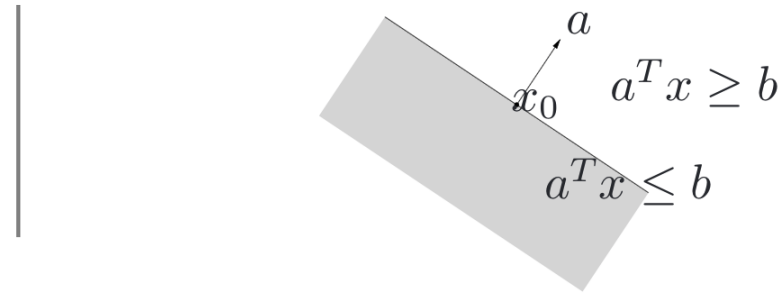


- Convex hulls

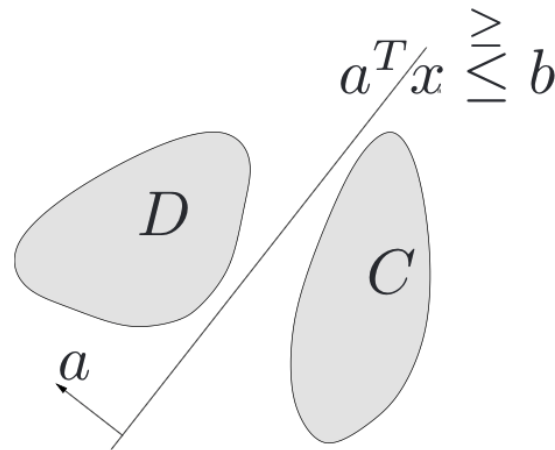


- **Hyperplane** in \mathbb{R}^n is a set: $\{x \mid a^T x = b\}$ where $a \in \mathbb{R}^n, a \neq 0, b \in \mathbb{R}$
In other words, $\{x \mid a^T(x - x_0) = 0\}$, where x_0 is any point in the hyperplane, i.e., $a^T x_0 = b$.

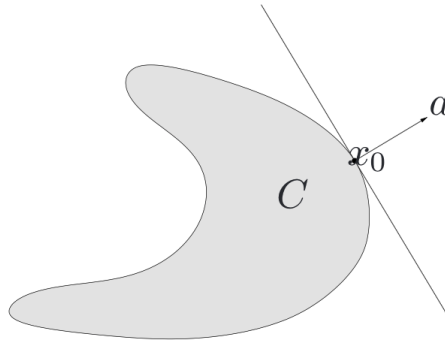
- Divides \mathbb{R}^n into two **halfspaces**:
 $\{x \mid a^T x \leq b\}$ and $\{x \mid a^T x > b\}$



- **Polyhedron** is the solution set of a finite number of linear equalities and inequalities (intersection of finite number of halfspaces and hyperplanes)
$$\mathcal{P} = \{x \mid a_j^T x \leq b_j, j = 1, \dots, m, c_j^T x = d_j, j = 1, \dots, p\} = \{x \mid Ax \leq b, Cx = d\}$$
- **Polytope**: a bounded polyhedron

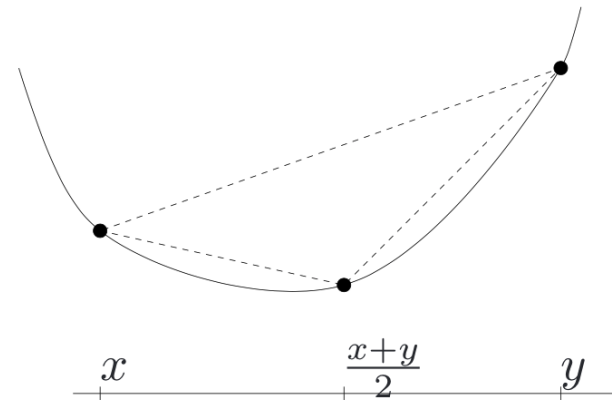


- \mathcal{C} and \mathcal{D} : non-intersecting convex sets, i.e., $\mathcal{C} \cap \mathcal{D} = \phi$.
- Then, there exist $a \neq 0$ and b such that $a^T x \leq b$ for all $x \in \mathcal{C}$ and $a^T x \geq b$ for all $x \in \mathcal{D}$.



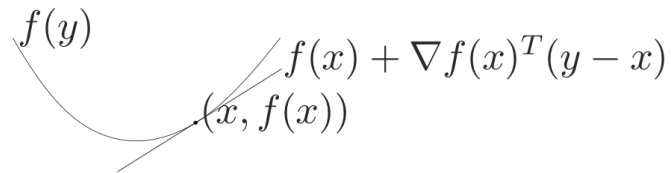
- Given a set $C \in \mathbb{R}^n$ and a point x_0 on its boundary, if $a \neq 0$ satisfies $a^T x \leq a^T x_0$ for all $x \in C$, then $\{x | a^T x = a^T x_0\}$ is called a **supporting hyperplane** to C at x_0
- For any nonempty convex set C and **any** x_0 on boundary of C , there exists a supporting hyperplane to C at x_0
- What happens if C is non-convex?

- $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a **convex function** if $\text{dom } f$ is a convex set and for all $x, y \in \text{dom } f$ and $\theta \in [0, 1]$, we have
$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$$
- f is **strictly convex** if the strict inequality in the above holds for all $x \neq y$ and $0 < \theta < 1$.
- f is **concave** if $-f$ is convex
- Affine functions are convex and concave
- **Jensen's inequality.** For a rv X ,
$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$



Conditions of Convex Functions (1)

- **First-order condition.** For differentiable functions, f is convex iff
$$f(y) - f(x) \geq \nabla f(x)^T (y - x), \quad \forall x, y \in \text{dom } f, \text{ and } \text{dom } f \text{ is convex}$$



- **Example.** $f(y) = y^2$.
- $f(y) \geq \tilde{f}_x(y)$ where $\tilde{f}_x(y)$ is the first order Taylor expansion of $f(y)$ at x .
- **Local** information (first order Taylor approximation) about a convex function provides **global** information (global underestimator).
- If $\nabla f(x) = 0$, then $f(y) \geq f(x)$, $\forall y$. Thus, x is a global minimizer of f

- **Second-order condition.** For twice differentiable functions, f is convex iff
 $\nabla^2 f(x) \succeq 0$
for all $x \in \text{dom } f$ (upward slope) and $\text{dom } f$ is convex
- Example: $f(x) = x^2$.
- Meaning: The graph of the function have positive (upward) curvature at x .

Examples of Convex or Concave Functions

- e^{ax} is convex on \mathbb{R} , for any $a \in \mathbb{R}$
- x^a is convex on \mathbb{R}_{++} when $a \geq 1$ or $a \leq 0$, and concave for $0 \leq a \leq 1$
- $|x|^p$ is convex on \mathbb{R} for $p \geq 1$
- $\log x$ is concave on \mathbb{R}_{++}
- $x \log x$ is strictly convex on \mathbb{R}_{++}
- Every norm on \mathbb{R}^n is convex
- $f(x) = \max\{x_1, \dots, x_n\}$ is convex on \mathbb{R}^n
- $f(x) = \log \sum_{i=1}^n e^{x_i}$ is convex on \mathbb{R}^n
- $f(x) = (\prod_{i=1}^n x_i)^{\frac{1}{n}}$ is concave on \mathbb{R}_{++}^n

- $f = \sum_{i=1}^n w_i f_i$ convex if f_i are all convex and $w_i \geq 0$
- $g(x) = f(Ax + b)$ is convex iff $f(x)$ is convex
- $f(x) = \max\{f_1(x), f_2(x)\}$ convex if f_i convex, e.g., sum of r largest components is convex
- $f(x) = h(g(x))$, where $h : \mathbb{R}^k \mapsto \mathbb{R}$ and $g : \mathbb{R}^n \mapsto \mathbb{R}^k$.

If $k = 1$: $f''(x) = h''(g(x))g'(x)^2 + h'(g(x))g''(x)$. So, f is convex if h is convex and nondecreasing and g is convex, or if h is convex and nonincreasing and g is concave ...

- $g(x) = \inf_{y \in \mathcal{C}} f(x, y)$ is convex if f is convex in (x, y) and \mathcal{C} is convex

- If $f(x, y)$ is convex in x for each $y \in \mathcal{A}$, then

$$g(x) = \sup_{y \in \mathcal{A}} f(x, y)$$

is **convex**. Similarly, if $f(x, y)$ is concave in x for each $y \in \mathcal{A}$, then

$$g(x) = \inf_{y \in \mathcal{A}} f(x, y)$$

is **concave**.

- **Example.** distance to farthest point in a set \mathcal{C} : $f(x) = \sup_{y \in \mathcal{C}} \|x - y\|$ is **convex**.
- **Example.** Lagrange dual function

$$\mathcal{D}(\lambda, \nu) = \inf_x \mathcal{L}(x, \lambda, \nu)$$

is **concave**.

- (1) Optimization Using Gradient Descent
- (2) Constrained Optimization and Lagrange Multipliers
- (3) Convex Sets and Functions
- (4) **Convex Optimization**
- (5) Convex Conjugate

- A **standard convex optimization** problem with variables \mathbf{x} :

$$\begin{aligned} & \text{minimize} && f(\mathbf{x}) \\ & \text{subject to} && g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \dots, m \\ & && a_i^\top \mathbf{x} = b_i, \quad i = 1, 2, \dots, p \end{aligned}$$

where f, f_1, \dots, f_m are convex functions.

- **Minimize convex** objective function (or maximize concave objective function)
- **Upper bound inequality** constraints on **convex** functions (\Rightarrow Constraint set is convex)
- **Equality** constraints must be **affine** (Only affine functions leads to a convex set for the equality constraints)

- Local optimality implies global optimality.
 - Given \mathbf{x} is **locally optimal** for a convex optimization problem, i.e., \mathbf{x} is feasible and for some $R > 0$,

$$f(\mathbf{x}) = \inf \{ f(\mathbf{z}) \mid \mathbf{z} \text{ is feasible}, \|\mathbf{z} - \mathbf{x}\|_2 \leq R \}$$

- **Theorem.** if \mathbf{x} is locally optimal in convex program, then globally optimal.
- Optimal condition for differentiable f
 - \mathbf{x} is optimal for a convex optimization problem iff \mathbf{x} is feasible and for all feasible \mathbf{y} :

$$\nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) \geq 0$$

- $-\nabla f(\mathbf{x})$ defines a supporting hyperplane to the feasible set $(\{\mathbf{y} \mid -\nabla f(\mathbf{x})^T \mathbf{y} \leq -\nabla f(\mathbf{x})^T \mathbf{x}\})$.
 - (Note) Unconstrained convex optimization: condition reduces to: $\nabla f(\mathbf{x}) = 0$

- Strong duality (zero optimal duality gap):
 $d^* = p^*$
- If strong duality holds, solving dual is ‘equivalent’ to solving primal. But strong duality does **not** always hold
- Convexity and **constraint qualifications** \implies Strong duality
- A simple constraint qualification: **Slater’s condition** (there exists strictly feasible primal variables $f_i(\mathbf{x}) < 0$ for non-affine f_i) (see Section 5.3.2 of Boyd’s book).
- Another reason why convex optimization is ‘easy’

- Assume **strong duality** holds. Then,

$$\begin{aligned} f(\mathbf{x}^*) = \mathcal{D}(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) &= \inf_{\mathbf{x}} \left(f(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}) \right) \\ &\leq f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*) \leq f(\mathbf{x}^*) \end{aligned}$$

- Thus, the two inequalities must hold with equality, implying: $\lambda_i^* f_i(\mathbf{x}^*) = 0, \forall i$
- Complementary slackness** condition:

$$\lambda_i^* > 0 \implies f_i(\mathbf{x}^*) = 0$$

$$f_i(\mathbf{x}^*) < 0 \implies \lambda_i^* = 0$$

- i -th optimal Lagrange multiplier is zero unless the i th constraint is active at the optimum.

- Since \mathbf{x}^* minimizes $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ over \mathbf{x} ,

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

Karush-Kuhn-Tucker optimality condition

$$f_i(\mathbf{x}^*) \leq 0, \quad h_i(\mathbf{x}^*) = 0, \quad \lambda_i^* \succeq 0$$

$$\lambda_i^* f_i(\mathbf{x}^*) = 0$$

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

- **Any** optimization with strong duality, KKT condition is necessary for primal-dual optimality
- **Convex** optimization with Slater's condition, KKT is also **sufficient** for primal-dual optimality.

- Minimization problem (**min-min-max** rule)
 - Problem: $\min f(x)$ s.t. $f_i(x) \leq 0, \quad g_i(x) = 0, \quad x$
 - $f(x)$: convex, $f_i(x)$: convex, $g_i(x)$: affine
 - $L(x, \lambda, \mu) = f(x) + \sum_i \lambda_i f_i(x) + \sum_i \mu_i g_i(x)$
 - $\inf_x L(x, \lambda, \mu) = \mathcal{D}(\lambda, \mu)$
 - $\max_{\lambda \geq 0} \mathcal{D}(\lambda, \mu)$
- Maximization problem (**max-max-min** rule)
 - Problem: $\max f(x)$ s.t. $f_i(x) \geq 0, \quad g_i(x) = 0, \quad x$
 - $f(x)$: concave, $f_i(x)$: concave, $g_i(x)$: affine
 - $L(x, \lambda, \mu) = f(x) + \sum_i \lambda_i f_i(x) + \sum_i \mu_i g_i(x)$
 - $\sup_x L(x, \lambda, \mu) = \mathcal{D}(\lambda, \mu)$
 - $\min_{\lambda \geq 0} \mathcal{D}(\lambda, \mu)$

- Primal problem

$$\begin{array}{ll}\min_{\mathbf{x} \in \mathbb{R}^d} & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} & \mathbf{A}\mathbf{x} \preceq \mathbf{b},\end{array}$$

where $\mathbf{A} \in \mathbb{R}^{m \times d}$ and $\mathbf{b} \in \mathbb{R}^m$.

- Dual problem

$$\begin{array}{ll}\max_{\boldsymbol{\lambda} \in \mathbb{R}^m} & -\mathbf{b}^\top \boldsymbol{\lambda} \\ \text{subject to} & \mathbf{c} + \mathbf{A}^\top \boldsymbol{\lambda} = 0, \boldsymbol{\lambda} \succeq 0,\end{array}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m$.

- The Lagrangian: $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = (\mathbf{c} + \mathbf{A}^\top \boldsymbol{\lambda})^\top \mathbf{x} - \boldsymbol{\lambda}^\top \mathbf{b}$, whose derivative w.r.t. \mathbf{x} becomes zero, when $\mathbf{c} + \mathbf{A}^\top \boldsymbol{\lambda} = 0$.
- The dual function: $\mathcal{D}(\boldsymbol{\lambda}) = -\boldsymbol{\lambda}^\top \mathbf{b}$

- Primal problem

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^d} \quad & \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{c}^T \mathbf{x} \\ \text{subject to} \quad & \mathbf{A} \mathbf{x} \preceq \mathbf{b}, \end{aligned}$$

where $\mathbf{A} \in \mathbb{R}^{m \times d}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^d$, the square matrix \mathbf{Q} is symmetric, positive definite.

- Dual problem

$$\begin{aligned} \max_{\boldsymbol{\lambda} \in \mathbb{R}^m} \quad & \left(-\frac{1}{2} (\mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda})^T \mathbf{A} \mathbf{Q}^{-1} (\mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda}) - \boldsymbol{\lambda}^T \mathbf{b} \right) \\ \text{subject to} \quad & \boldsymbol{\lambda} \succeq 0, \end{aligned}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m$.

- (1) Optimization Using Gradient Descent
- (2) Constrained Optimization and Lagrange Multipliers
- (3) Convex Sets and Functions
- (4) Convex Optimization
- (5) **Convex Conjugate**

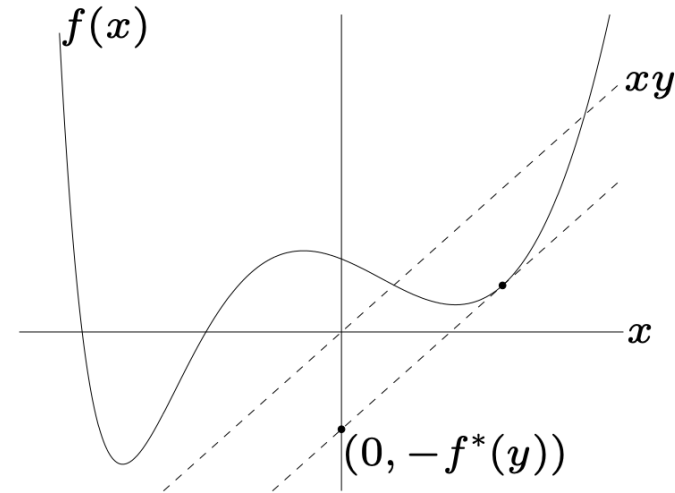
Conjugate Function: Definition and Meaning

- Given $f : \mathbb{R}^n \mapsto \mathbb{R}$, the **conjugate function** $f^* : \mathbb{R}^n \mapsto \mathbb{R}$ defined as:

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \text{dom } f} (\mathbf{y}^T \mathbf{x} - f(\mathbf{x}))$$

with domain consisting of $\mathbf{y} \in \mathbb{R}^n$ for which the supremum is finite

- Assume \mathbb{R}^1 .
- For a given slope of y , $yx - f(x)$ is the vertical distance between the line yx and $f(x)$.
- Thus, $f^*(y)$ is the maximum distance



- Given $f : \mathbb{R}^n \mapsto \mathbb{R}$, the conjugate function $f^* : \mathbb{R}^n \mapsto \mathbb{R}$ defined as:

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \text{dom } f} (\mathbf{y}^T \mathbf{x} - f(\mathbf{x}))$$

with domain consisting of $\mathbf{y} \in \mathbb{R}^n$ for which the supremum is finite

- $f^*(\mathbf{y})$: **always convex** (the pointwise supremum of a family of affine functions of \mathbf{y})
- $f^* = f$ if f is convex and closed
- Fenchel's inequality**: $f(\mathbf{x}) + f^*(\mathbf{y}) \geq \mathbf{x}^T \mathbf{y}$ for all \mathbf{x}, \mathbf{y} (by definition)
 - Example**. $f(x) = |x|^2/2$. Then, $f^*(y) = |y|^2/2$. Thus, F-inequality tells us:

$$\frac{1}{2}(|x|^2 + |y|^2) \geq xy$$

Examples of Conjugate Functions

- $f(x) = ax + b, f^*(a) = -b$
- $f(x) = -\log x, f^*(y) = -\log(-y) - 1$ for $y < 0$
- $f(x) = e^x, f^*(y) = y \log y - y$
- $f(x) = x \log x, f^*(y) = e^{y-1}$
- $f(x) = \frac{1}{2}x^T Qx, f^*(y) = \frac{1}{2}y^T Q^{-1}y$ (Q is positive definite)
- $f(x) = \log \sum_{i=1}^n e^{x_i},$
$$f^*(y) = \begin{cases} \sum_{i=1}^n y_i \log y_i & \text{if } y \succeq 0 \text{ and } \sum_{i=1}^n y_i = 1, \\ \infty & \text{otherwise} \end{cases}$$

- They are closely related. Consider the following problem:

$$\begin{array}{ll}\text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{Ax} \preceq \mathbf{b}, \\ & \mathbf{Cx} = \mathbf{d}\end{array}$$

- Using the conjugate of f , we can write the dual function as:

$$\begin{aligned}\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\nu}) &= \inf_{\mathbf{x}} \left(f(\mathbf{x}) + \boldsymbol{\lambda}^\top (\mathbf{Ax} - \mathbf{b}) + \boldsymbol{\nu}^\top (\mathbf{Cx} - \mathbf{d}) \right) \\ &= -\mathbf{b}^\top \boldsymbol{\lambda} - \mathbf{d}^\top \boldsymbol{\nu} + \inf_{\mathbf{x}} \left(f(\mathbf{x}) + (\mathbf{A}^\top \boldsymbol{\lambda} + \mathbf{C}^\top \boldsymbol{\nu})^\top \mathbf{x} \right) \\ &= -\mathbf{b}^\top \boldsymbol{\lambda} - \mathbf{d}^\top \boldsymbol{\nu} - f^* \left(-\mathbf{A}^\top \boldsymbol{\lambda} - \mathbf{C}^\top \boldsymbol{\nu} \right)\end{aligned}$$

Questions?

1)