AIE6003: Optimization for Machine Learning

Xiaopeng Li
CUHK(SZ)
SAI

April 4, 2026

Outline

Duality Theory

Lagrangian Duality

From PGD to Duality

Recall: PGD solves $\min_{x\in\Omega} f(x)$ via projected gradient steps:

$$x_{k+1} = \mathrm{proj}_\Omega(x_k - \eta\nabla f(x_k)).$$

Limitation:

- projection onto Ω can be expensive;
- PGD gives no lower bound on f^*; no info on how close to optimal.

Goal of this lecture:

- develop KKT conditions;
- develop Lagrangian duality;
- apply duality to machine learning problems.

Equality-Constrained Optimization: Examples

We consider the general form

$$\min_x \; f(x) \quad \text{s.t.} \quad Ax = b.$$

Min-variance portfolio: invest in n assets with return covariance Σ,

$$\min_{x \in \mathbb{R}^n} \; \tfrac{1}{2} x^\top \Sigma \, x \quad \text{s.t.} \quad \mu^\top x = r_{\text{target}}, \;\; \mathbf{1}^\top x = 1.$$

Consensus in distributed optimization:

$$\min \; \sum_{i=1}^m f_i(x_i) \quad \text{s.t.} \quad x_1 = x_2 = \cdots = x_m.$$

ERM with linear model:

$$\min_{w,z} \; \sum_{i=1}^m \phi_i(z_i) + \tfrac{\lambda}{2} \|w\|^2 \quad \text{s.t.} \quad z_i = w^\top x_i.$$

Decoupling predictions z_i from model w via equality constraints enables Fenchel duality, which we will see later.

Optimality Conditions for Equality Constraints

Problem: let f be smooth and convex, $A \in \mathbb{R}^{m \times n}$ with $\operatorname{rank}(A) = m$,

$$\min_x \; f(x) \quad \text{s.t.} \quad Ax = b.$$

Recall: for a closed convex set Ω,

$$x^* = \arg\min_{x \in \Omega} f(x) \iff \nabla f(x^*)^\top (x - x^*) \geq 0 \quad \forall\, x \in \Omega.$$

Specialize to $\Omega = \{x : Ax = b\}$: any feasible direction $v = x - x^*$ satisfies $Av = 0$, so

$$\nabla f(x^*)^\top v \geq 0, \quad \forall\, v \in \ker(A).$$

Since $\ker(A)$ is a subspace, ≥ 0 for all v and $-v$ implies

$$\nabla f(x^*)^\top v = 0 \quad \forall\, v \in \ker(A) \implies \nabla f(x^*) \in \ker(A)^\perp = \operatorname{Im}(A^\top).$$

KKT Conditions

Since $\nabla f(x^*) \in \operatorname{Im}(A^\top)$, $\exists\, \lambda^* \in \mathbb{R}^m$ such that $\nabla f(x^*) = -A^\top \lambda^*$.

KKT Conditions for Equality Constraints

x^* solves $\min f(x)$ s.t. $Ax = b$ if and only if $\exists\, \lambda^* \in \mathbb{R}^m$:

$$\begin{cases} \nabla f(x^*) + A^\top \lambda^* = 0 & \text{(stationarity)} \\ Ax^* = b & \text{(feasibility)} \end{cases}$$

Example: quadratic program $\min\ \frac{1}{2} x^\top Q x - c^\top x$ s.t. $Ax = b$.

KKT becomes a linear system:

$$\begin{pmatrix} Q & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} c \\ b \end{pmatrix}.$$

Beyond Linear Equality Constraints

So far we have KKT for linear equality constraints. What about others?

$$\min \ f(x) \quad \text{s.t.} \quad g_j(x) \leq 0, \ \ j = 1, \ldots, p, \quad Ax = b.$$

Geometric viewpoint: pack all inequalities into a convex set

$$\Omega = \{x : g_j(x) \leq 0, \ \ j = 1, \ldots, p\},$$

which is convex when each g_j is convex. Then study

$$\min \ f(x) \quad \text{s.t.} \quad x \in \Omega, \ \ Ax = b.$$

Why no nonlinear equalities $h_i(x) = 0$?

- The set $\{x : h_i(x) = 0\}$ is generally not convex for nonlinear h_i.
- E.g., $\{x : x^2 = 1\} = \{-1, +1\}$, a discrete set!
- Only affine equalities preserve convexity, so they are kept separate.

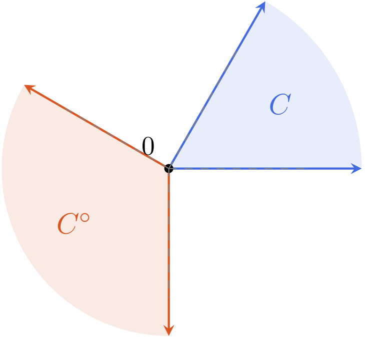
To study optimality over Ω, we need the notion of cones.

Cones and Polar Cones

Cone: a set C is a cone if $x \in C \implies \alpha x \in C$ for all $\alpha \geq 0$.

Conic hull: For a set S, $\mathrm{cone}(S) = \{\alpha x : x \in S,\ \alpha \geq 0\}$.

Polar cone: $C^\circ = \{u : \langle u, v\rangle \leq 0\ \ \forall\, v \in C\}$.



Bipolar Theorem: If C is a closed convex cone, then $C^{\circ\circ} = C$.

Dual cone: $C^* := \{u : \langle u, v\rangle \geq 0\ \ \forall\, v \in C\} = -C^\circ$.

Bipolar Theorem: $C^{\circ\circ} = C$

Bipolar Theorem

If C is a closed convex cone, then $C^{\circ\circ} = C$.

($\supseteq$) Take any $x \in C$. By definition of C°, every $u \in C^\circ$ satisfies $\langle u, x\rangle \le 0$, so $x \in C^{\circ\circ}$.

($\subseteq$) Take any $x \notin C$. Since C is closed and convex, by the separation theorem, $\exists\, u$ such that

$$\langle u, x\rangle > 0 \quad \text{and} \quad \langle u, v\rangle \le 0, \quad \forall\, v \in C.$$

- Second condition $\implies u \in C^\circ$.
- First condition $\implies \langle u, x\rangle > 0$ with $u \in C^\circ$, so $x \notin C^{\circ\circ}$.

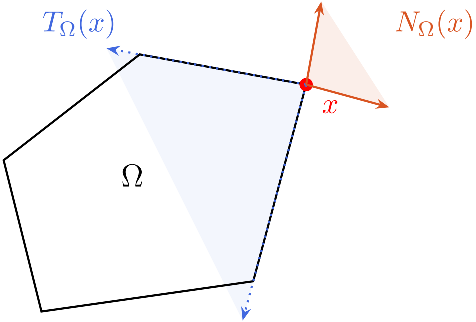
Thus, we have obtained $x \notin C \implies x \notin C^{\circ\circ}$, hence $C^{\circ\circ} \subseteq C$.

Remark: both closedness and convexity are essential.

Tangent and Normal Cones

Tangent cone of Ω at x: $T_\Omega(x) = \mathrm{cone}\{z - x : z \in \Omega\}$.

Normal cone of Ω at x: $N_\Omega(x) = \big(T_\Omega(x)\big)^\circ$.



At a corner: both $T_\Omega(x)$ and $N_\Omega(x)$ are proper cones (convex, closed, nonempty interior, contains no line).

Optimality: for f convex and differentiable,
$x^* = \arg\min_{x\in\Omega} f(x) \iff -\nabla f(x^*) \in N_\Omega(x^*)$.

Optimality via Normal Cone

Optimality Condition

Let f be convex and differentiable, Ω closed and convex. Then

$$x^* = \arg\min_{x \in \Omega} f(x) \quad \Longleftrightarrow \quad -\nabla f(x^*) \in N_\Omega(x^*).$$

($\Longrightarrow$): For any $x \in \Omega$, convexity of Ω gives $x^* + t(x - x^*) \in \Omega$ for $t \in [0, 1]$. Since x^* is a minimizer:

$$\frac{f(x^* + t(x - x^*)) - f(x^*)}{t} \geq 0 \quad \forall t \in (0, 1].$$

Let $t \to 0^+$: $\nabla f(x^*)^\top (x - x^*) \geq 0$ for all $x \in \Omega$, so $-\nabla f(x^*) \in N_\Omega(x^*)$.

($\Longleftarrow$): here we need convexity of f.
$-\nabla f(x^*) \in N_\Omega(x^*)$ gives $\nabla f(x^*)^\top (x - x^*) \geq 0$ for all $x \in \Omega$. Then:

$$f(x) \geq f(x^*) + \underbrace{\nabla f(x^*)^\top (x - x^*)}_{\geq 0} \geq f(x^*) \quad \forall x \in \Omega.$$

General & Equality Constraints

Problem: let Ω is closed convex, $A \in \mathbb{R}^{m\times n}$, $\mathrm{rank}(A) = m$.

$$\min \ f(x) \quad \text{s.t.} \quad x \in \Omega, \ \ Ax = b,$$

Optimality Condition

x^* is optimal if and only if $\exists\, \lambda^* \in \mathbb{R}^m$ such that

$$-\nabla f(x^*) - A^\top \lambda^* \in N_\Omega(x^*).$$

Normal Cone Decomposition

Let $\mathcal{A} = \{x : Ax = b\}$. For $x \in \Omega \cap \mathcal{A}$ with $\mathrm{relint}(\Omega) \cap \mathcal{A} \neq \emptyset$,

$$N_{\Omega\cap\mathcal{A}}(x) = N_\Omega(x) + \{A^\top \lambda : \lambda \in \mathbb{R}^m\}.$$

Consequence: $\nabla f(x^*) \in N_{\Omega\cap\mathcal{A}}(x^*)$ decomposes into a normal-cone part and a multiplier part.

Normal Cone Decomposition

($\supseteq$) Easy direction: take $u \in N_\Omega(x)$ and any $\lambda \in \mathbb{R}^m$.

- For any $z \in \Omega \cap \mathcal{A}$, we have $A(z-x) = 0$, so $(z-x)^\top A^\top \lambda = 0$.
- Also $(z-x)^\top u \leq 0$ since $u \in N_\Omega(x)$.
- Hence $(z-x)^\top (u + A^\top \lambda) \leq 0$, i.e., $u + A^\top \lambda \in N_{\Omega \cap \mathcal{A}}(x)$.

($\subseteq$) Harder direction: take $v \in N_{\Omega \cap \mathcal{A}}(x)$. Define in $\mathbb{R}^{d+1}$:

- Define two sets $C_1 = \{(y, \mu) : y = z - x,\ z \in \Omega,\ \mu \leq v^\top y\}$ and $C_2 = \{(y, \mu) : y \in \ker(A),\ \mu = 0\}$.
- Since $\mathrm{relint}(\Omega) \cap \mathcal{A} \neq \emptyset$, apply the separation theorem to C_1 and C_2.
- The separating hyperplane yields $w = A^\top \lambda$ for some λ.
- Conclude $v = (v + w) - w$ where $v + w \in N_\Omega(x)$ and $w \in \{A^\top \lambda\}$.

Key assumption: $\mathrm{relint}(\Omega) \cap \mathcal{A} \neq \emptyset$ is a constraint qualification ensuring the normal cones add properly.

Tangent Cone of a Polyhedron

Let $\Omega = \{x : Cx \leq d\}$. Fix $x \in \Omega$.

Active constraints: $I(x) = \{j : C_j x = d_j\}$.

Claim: $T_\Omega(x) = \{v : C_j v \leq 0 \ \ \forall\, j \in I(x)\}$.

($\subseteq$): take $v \in T_\Omega(x)$, so $v = \alpha(z - x)$ for some $z \in \Omega$, $\alpha \geq 0$. For $j \in I(x)$:

$$C_j v = \alpha(C_j z - C_j x) = \alpha(\underbrace{C_j z}_{\leq d_j} - \underbrace{d_j}_{= C_j x}) \leq 0.$$

($\supseteq$): take v with $C_j v \leq 0$ for all $j \in I(x)$. For small $t > 0$:

- Active ($j \in I(x)$): $C_j(x + tv) = \underbrace{C_j x}_{= d_j} + tC_j v \leq d_j$.
- Inactive ($j \notin I(x)$): $C_j x < d_j$, so $C_j(x + tv) < d_j$ for small t.

So $x + tv \in \Omega$, hence $v \in T_\Omega(x)$.

Normal Cone of a Polyhedron

Let $\Omega = \{x : Cx \le d\}$. Fix $x \in \Omega$.

Active constraints: $I(x) = \{j : C_j x = d_j\}$.

Tangent cone: $T_\Omega(x) = \{v : C_j v \le 0 \ \ \forall\, j \in I(x)\}$.

Define: $K = \mathrm{cone}\{C_j^\top : j \in I(x)\} = \left\{\sum_{j \in I(x)} \mu_j\, C_j^\top : \mu_j \ge 0\right\}$.

How to compute normal cone?

- Compute K°:
$$K^\circ = \left\{v : \left\langle C_j^\top, v \right\rangle \le 0 \ \ \forall\, j \in I(x)\right\} = T_\Omega(x).$$
- Identify $N_\Omega(x) = (T_\Omega(x))^\circ = K^{\circ\circ}$;
- K is closed and convex. By the bipolar theorem:
$$N_\Omega(x) = K^{\circ\circ} = K = \left\{\sum_{j \in I(x)} \mu_j\, C_j^\top : \mu_j \ge 0\right\}.$$
- Setting $\mu_j = 0$ for $j \notin I(x)$:
$N_\Omega(x) = \{C^\top \mu : \mu \ge 0,\ \mu_j (C_j x - d_j) = 0,\ \forall\, j\}$.

Recovering KKT from Normal Cones: Linear

Consider the problem with both equality and inequality constraints:

$$\min\ f(x) \quad \text{s.t.} \quad Ax = b, \quad Cx \le d.$$

Set $\Omega = \{x : Cx \le d\}$. The normal cone is

$$N_\Omega(x) = \{C^\top \mu : \mu \ge 0,\ \ \mu_j(C_j x - d_j) = 0\ \ \forall\, j\}.$$

Substituting into $-\nabla f(x^*) - A^\top \lambda^* \in N_\Omega(x^*)$ gives:

$$\begin{cases} \nabla f(x^*) + A^\top \lambda^* + C^\top \mu^* = 0 & \text{(stationarity)} \\ Ax^* = b & \text{(primal feasibility)} \\ Cx^* \le d & \text{(primal feasibility)} \\ \mu^* \ge 0 & \text{(dual feasibility)} \\ \mu_j^*(C_j x^* - d_j) = 0\ \ \forall\, j & \text{(complementary slackness)} \end{cases}$$

This is exactly the standard KKT conditions! The normal cone formulation is a compact geometric encoding of KKT.

Normal Cone: Convex Inequality Constraints

Now let $\Omega = \{x : g_j(x) \le 0,\ j = 1, \dots, p\}$ with each g_j convex and differentiable.

Active set: $I(x) = \{j : g_j(x) = 0\}$.

Define: $K = \operatorname{cone}\{\nabla g_j(x) : j \in I(x)\} = \left\{\sum_{j \in I(x)} \mu_j \nabla g_j(x) : \mu_j \ge 0\right\}$.

How to compute normal cone?

- $K^\circ = \{v : \nabla g_j(x)^\top v \le 0 \ \ \forall\, j \in I(x)\} \stackrel{?}{=} T_\Omega(x)$;
- if $K^\circ = T_\Omega(x)$, then $N_\Omega(x) = (T_\Omega(x))^\circ = K^{\circ\circ}$;
- K is finitely generated $\implies$ closed. By the bipolar theorem:

$$N_\Omega(x) = K^{\circ\circ} = K = \left\{ \sum_{j \in I(x)} \mu_j \nabla g_j(x) : \mu_j \ge 0 \right\}.$$

The question: the first step requires $K^\circ = T_\Omega(x)$. Is this always true?

Why Slater's Condition?

We always have $T_\Omega(x) \subseteq K^\circ$. The reverse needs care.

Issue: take v with $\nabla g_j(x)^\top v \le 0$ for $j \in I(x)$.

- If $\nabla g_j(x)^\top v < 0$: then $g_j(x + tv) < 0$ for small t.
- If $\nabla g_j(x)^\top v = 0$: first-order info says nothing.

Slater's condition: $\exists\, \bar{x}$ with $g_j(\bar{x}) < 0\ \forall\, j$. By convexity and $g_j(x) = 0$:

$$\nabla g_j(x)^\top (\bar{x} - x) \le g_j(\bar{x}) - g_j(x) = g_j(\bar{x}) < 0.$$

So $\bar{x} - x$ is a strictly feasible direction.

Perturbation: define $v_\varepsilon = (1-\varepsilon)v + \varepsilon(\bar{x} - x)$. For $j \in I(x)$:

$$\nabla g_j(x)^\top v_\varepsilon = (1-\varepsilon)\underbrace{\nabla g_j(x)^\top v}_{\le 0} + \varepsilon \underbrace{\nabla g_j(x)^\top (\bar{x} - x)}_{< 0} < 0.$$

Strict inequality $\implies$ $g_j(x + tv_\varepsilon) < 0$ for small t $\implies$ $v_\varepsilon \in T_\Omega(x)$.

Taking $\varepsilon \to 0$: $v \in T_\Omega(x)$. Hence $K^\circ \subseteq T_\Omega(x)$.

Constraint Qualifications

LICQ (Linear Independence CQ): $\{\nabla g_j(x^*) : j \in I(x^*)\}$ are linearly independent. Also guarantees unique multipliers μ^*.

MFCQ (Mangasarian–Fromovitz CQ): $\exists\, v$ such that
$$\nabla g_j(x^*)^\top v < 0 \quad \forall\, j \in I(x^*).$$
A strictly feasible direction exists at x^*. Same perturbation trick as Slater.

Slater's condition (convex only): $\exists\, \bar{x}$ with $g_j(\bar{x}) < 0 \;\; \forall\, j$. A global strictly feasible point. Implies MFCQ at every feasible x.

Relation:

$$\text{LICQ} \implies \text{MFCQ} \impliedby \text{Slater (convex)}$$
$$\text{All three} \implies K^\circ = T_\Omega(x) \implies \text{KKT holds.}$$

Recovering KKT: Nonlinear Constraints

Consider the general convex problem:

$$\min \; f(x) \quad \text{s.t.} \quad Ax = b, \quad g_j(x) \le 0, \;\; j = 1, \dots, p.$$

Set $\Omega = \{x : g_j(x) \le 0, \; j = 1, \dots, p\}$. Under a CQ, the normal cone is

$$N_\Omega(x) = \left\{ \sum_{j=1}^{p} \mu_j \, \nabla g_j(x) : \mu_j \ge 0, \;\; \mu_j \, g_j(x) = 0 \;\; \forall \, j \right\}.$$

Substituting into $-\nabla f(x^*) - A^\top \lambda^* \in N_\Omega(x^*)$ gives:

$$\begin{cases} \nabla f(x^*) + A^\top \lambda^* + \sum_{j=1}^{p} \mu_j^* \, \nabla g_j(x^*) = 0 & \text{(stationarity)} \\ Ax^* = b & \text{(equality feasibility)} \\ g_j(x^*) \le 0 \;\; \forall \, j & \text{(inequality feasibility)} \\ \mu_j^* \ge 0 \;\; \forall \, j & \text{(dual feasibility)} \\ \mu_j^* \, g_j(x^*) = 0 \;\; \forall \, j & \text{(complementary slackness)} \end{cases}$$

This is the general convex KKT conditions.

Linear vs Nonlinear Constraints

Recall: the bipolar argument requires $K^\circ = T_\Omega(x)$, i.e., the linearized tangent cone equals the true tangent cone.

Linear case: $g_j(x) = C_j x - d_j$, so linearization is exact:

$$C_j(x + tv) = \underbrace{C_j x}_{= d_j} + tC_j v \leq d_j \iff C_j v \leq 0.$$

No constraint qualification needed.

Nonlinear case: $g_j(x + tv) \approx g_j(x) + t\nabla g_j(x)^\top v$ is only first-order. The linearized cone K° may be strictly larger than $T_\Omega(x)$ without a CQ.

Example:

$$f(x) = x, \quad \Omega = \{x \in \mathbb{R} : \; x^2 \leq 0\} = \{0\}, \quad \bar{x} = 0.$$

Then

$$T_\Omega(\bar{x}) = \{0\}, \quad K^\circ = \{v : \nabla g(\bar{x})\, v \leq 0\} = \mathbb{R},$$

since $\nabla g(0) = 0$. Hence $K^\circ \supsetneq T_\Omega(\bar{x})$.

Outline

Duality Theory

Lagrangian Duality

Lagrangian and Dual Problem

Primal problem:

$$p^* = \inf_{x \in \mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad g(x) \le 0, \quad h(x) = 0.$$

Lagrangian:

$$\mathcal{L}(x, \lambda, \nu) = f(x) + \lambda^\top g(x) + \nu^\top h(x), \quad \lambda \in \mathbb{R}^m_+, \; \nu \in \mathbb{R}^p.$$

Observation: in the extended-value sense,

$$p^* = \inf_{x \in \mathbb{R}^n} \; \sup_{\lambda \ge 0, \; \nu \in \mathbb{R}^p} \mathcal{L}(x, \lambda, \nu).$$

Dual function:

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

Dual problem:

$$d^* = \sup_{\lambda \ge 0, \; \nu \in \mathbb{R}^p} \; q(\lambda, \nu).$$

Note: $q(\lambda, \nu)$ is always concave because it is the pointwise infimum of affine functions in (λ, ν).

Weak Duality

Weak Duality

For the primal-dual pair just defined,

$$d^* = \sup_{\lambda \geq 0,\ \nu \in \mathbb{R}^p} q(\lambda, \nu) \leq p^*.$$

Proof: for any primal feasible x and dual feasible (λ, ν),

$$\mathcal{L}(x, \lambda, \nu) = f(x) + \lambda^\top g(x) + \nu^\top h(x) \leq f(x),$$

because $g(x) \leq 0$, $h(x) = 0$, and $\lambda \geq 0$.

By definition of the dual function,

$$q(\lambda, \nu) = \inf_{z \in \mathbb{R}^n} \mathcal{L}(z, \lambda, \nu) \leq \mathcal{L}(x, \lambda, \nu) \leq f(x).$$

Taking $\sup_{\lambda \geq 0, \nu}$ on the left and $\inf$ over feasible x on the right,

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

Consequence: every dual feasible pair (λ, ν) gives a lower bound on the primal optimum.

Strong Duality

The duality gap is $p^* - d^* \geq 0$. When do we have equality?

Strong Duality

Under the convexity assumptions and Slater's condition,

$$d^* = p^*.$$

If the primal optimum is attained, then the dual optimum is also attained.

Meaning: solving the primal minimization problem is equivalent to solving the dual maximization problem:

$$\min_{\substack{x \in \mathbb{R}^n \\ g(x) \leq 0,\ h(x) = 0}} f(x) = \max_{\lambda \geq 0,\ \nu \in \mathbb{R}^p} q(\lambda, \nu).$$

Equivalently, (x^*, λ^*, ν^*) is a primal-dual optimal triple if and only if

$$\mathcal{L}(x^*, \lambda, \nu) \leq \mathcal{L}(x^*, \lambda^*, \nu^*) \leq \mathcal{L}(x, \lambda^*, \nu^*)$$

for all $x \in \mathbb{R}^n$, $\lambda \geq 0$, and $\nu \in \mathbb{R}^p$.

Proof Sketch of Strong Duality

Weak duality always gives $d^* \le p^*$.

Under convexity and Slater's condition, KKT holds at a primal optimum x^*: there exist $\lambda^* \ge 0$ and $\nu^* \in \mathbb{R}^p$ such that

$$\nabla_x \mathcal{L}(x^*, \lambda^*, \nu^*) = 0,$$

$g(x^*) \le 0$, $h(x^*) = 0$, and $(\lambda_i^*)\, g_i(x^*) = 0$.

Since $x \mapsto \mathcal{L}(x, \lambda^*, \nu^*)$ is convex, this means

$$x^* \in \arg\min_{x \in \mathbb{R}^n} \; \mathcal{L}(x, \lambda^*, \nu^*).$$

Evaluate the dual at (λ^*, ν^*):

$$\begin{aligned} q(\lambda^*, \nu^*) &= \inf_{x \in \mathbb{R}^n} \; \mathcal{L}(x, \lambda^*, \nu^*) = \mathcal{L}(x^*, \lambda^*, \nu^*) \\ &= f(x^*) + (\lambda^*)^\top g(x^*) + (\nu^*)^\top h(x^*) \\ &= f(x^*) = p^*, \end{aligned}$$

Conclusion: $d^* \ge q(\lambda^*, \nu^*) = p^*$, hence $d^* = p^*$.

Example: Soft-Margin SVM

Primal problem:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{s.t.} \quad 1-\xi_i - y_i(w^\top x_i + b) \le 0, \;\; -\xi_i \le 0.$$

Here $y_i \in \{\pm 1\}$ and ξ_i are slack variables.
Lagrangian: introduce multipliers $\alpha_i \ge 0$ and $\beta_i \ge 0$:

$$\mathcal{L} = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\big(1-\xi_i - y_i(w^\top x_i + b)\big) - \sum_{i=1}^{n}\beta_i\xi_i.$$

KKT stationarity:

$$\nabla_w\mathcal{L} = 0 \Rightarrow w = \sum_{i=1}^{n}\alpha_i y_i x_i, \qquad \partial_b\mathcal{L} = 0 \Rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0,$$

$$\partial_{\xi_i}\mathcal{L} = 0 \Rightarrow C - \alpha_i - \beta_i = 0 \quad \Longrightarrow \quad 0 \le \alpha_i \le C.$$

KKT complementarity:

$$\alpha_i\big(1-\xi_i - y_i(w^\top x_i + b)\big) = 0, \qquad \beta_i\xi_i = 0.$$

Hence $\alpha_i > 0$ means sample i lies on or inside the margin: it is a support vector.

SVM Duality

Strong duality: the primal is convex, the constraints are affine, and Slater holds (e.g. $w = 0,\ b = 0,\ \xi_i = 2$), so $d^* = p^*$.

Dual problem: minimizing $\mathcal{L}$ over w, b, ξ gives

$$\max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \, x_i^\top x_j$$

$$\text{s.t.} \quad 0 \le \alpha_i \le C, \qquad \sum_{i=1}^{n} \alpha_i y_i = 0.$$

Recover the classifier from the dual:

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i, \qquad 0 < \alpha_i^* < C \Rightarrow y_i(w^{*\top} x_i + b^*) = 1.$$

So only points with $\alpha_i^* > 0$ determine the classifier.

Sparsity and Kernel Trick

Sparsity from the dual: from stationarity,

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i.$$

So if $\alpha_i^* = 0$, sample i does not appear in w^*. Only points with $\alpha_i^* > 0$ matter; these are the support vectors.

Kernel trick: the dual objective depends on the data only through $x_i^\top x_j$. If in the primal, we use nonlinear classifier $w^\top \phi(x_i) + b$ instead of $w^\top x_i + b$, then the dual objective depends on the data only through $\phi(x_i)^\top \phi(x_j)$. We can replace them by a kernel

$$K(x_i, x_j) = \phi(x_i)^\top \phi(x_j),$$

without explicitly computing the feature map $\phi(x)$.

The classifier becomes

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i^* y_i \, K(x_i, x) + b^* \right).$$

Interpreting the Training Samples

KKT complementarity gives

$$\alpha_i^* \big(1 - \xi_i^* - y_i(w^{*\top} x_i + b^*)\big) = 0, \qquad (C - \alpha_i^*)\, \xi_i^* = 0,$$

together with

$$0 \le \alpha_i^* \le C.$$

These conditions tell us exactly what role each sample plays.

Case 1: $\alpha_i^* = 0 \implies y_i(w^{*\top} x_i + b^*) \ge 1.$ The point is correctly classified and outside the margin. It is inactive: removing it locally does not change w^*.

Case 2: $0 < \alpha_i^* < C \implies \xi_i^* = 0$ and $y_i(w^{*\top} x_i + b^*) = 1.$ The point lies exactly on the margin. These points define the decision boundary.

Case 3: $\alpha_i^* = C \implies y_i(w^{*\top} x_i + b^*) = 1 - \xi_i^*.$ The point is at or inside the margin. If $0 < \xi_i^* \le 1$, it violates the margin but is still correctly classified; if $\xi_i^* > 1$, it is misclassified.

Wrap-up and Next Week

Week 9: Key takeaways

- KKT conditions: first-order optimality for constrained problems.
- Normal cones generalize KKT to set constraints; decomposition under constraint qualification.
- Lagrangian duality: dual provides lower bound (weak); equals primal under Slater's condition (strong).
- ERM as a canonical application of duality in machine learning.

Thank you for your attendance!

Questions?

Next week (Week 10): Dual First-order Algorithm, ALM, ADMM