DS-GA 1003: Machine Learning

Lecture 4: Convex Optimization and SVMs

Slides adapted from material from David Rosenberg.

Logistics & Announcements

PS 1 grades/solutions. Grades will be released Wednesday, along with solutions.

PS 2 extension. Due in two weeks, Tuesday, Feb. 24 11:59 PM ET.

Lecture for Week 5 (02/17) is cancelled due to President's Day.

Lecture on Week 6 (02/24) will be remote and recorded. Sam out of town for conference :(

Projects. Group formation due Feb. 28th on Gradescope (full guidelines on website).

EdStem thread "Project group formation thread" for forming groups.

Midterm. March 10th during lecture. Details + practice problems coming this week.

Outline

Convexity Primer

Convex Optimization

Convex Optimization: Duality

Constraint Qualification & Complementary Slackness

SVM Optimization Problem

SVM Dual Optimization

Strong Duality applied to SVM

Why Convex Optimization?

Motivation

Historically

Linear programs (linear objectives & constraints) were the focus.

Nonlinear programs: some easy, some hard.

Early 2000s

Main distinction is between convex and non-convex problems.

Convex problems are the ones we know how to solve efficiently.

2010+

Many people begin to understand optimization / estimation / approximation error tradeoffs.

Accepted stochastic methods often faster to get good results (especially on "big data").

These days: nobody's scared of non-convex problems – SGD works well enough on problems of interest (i.e. neural networks).

Classification Losses

Convexity

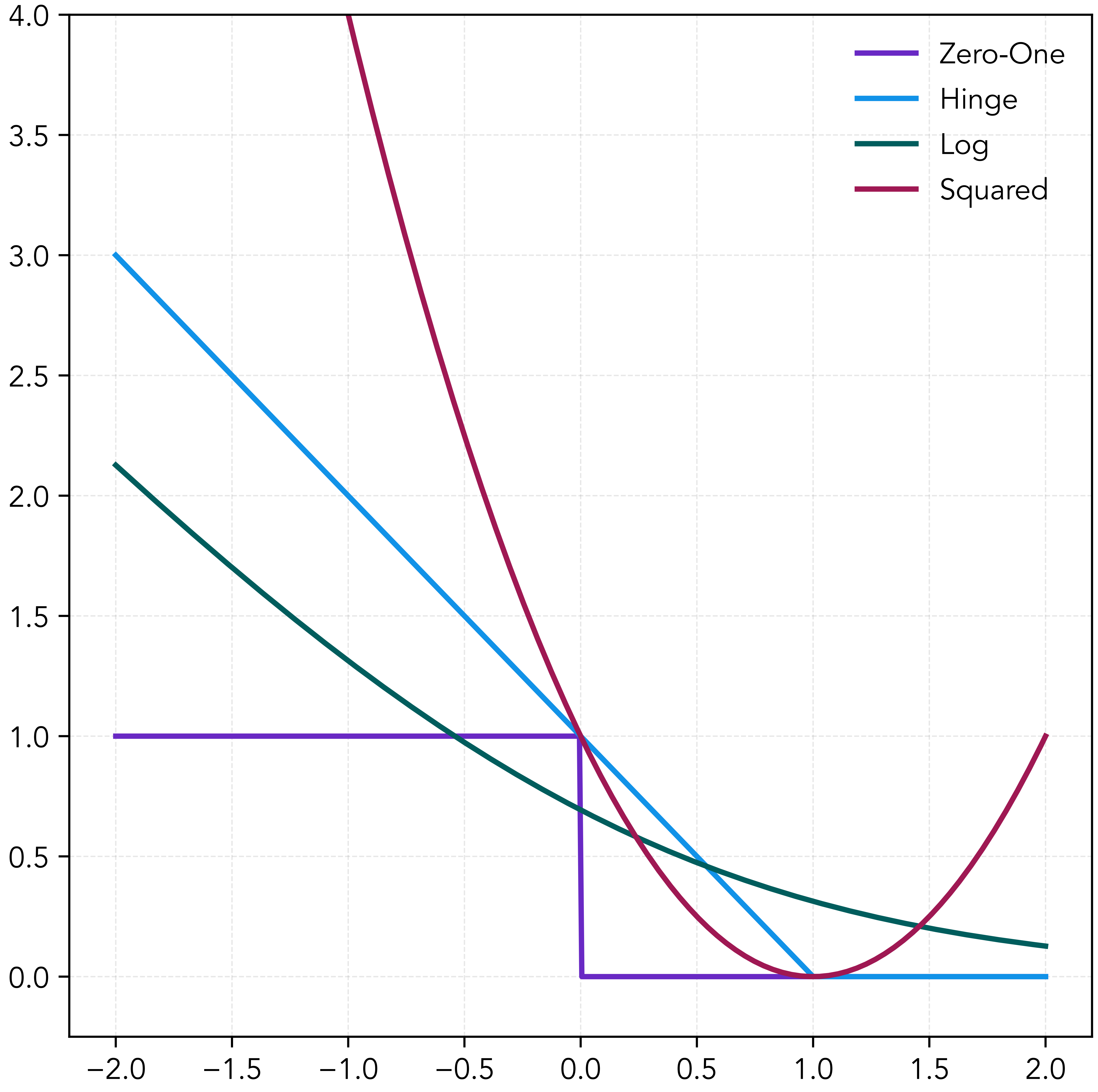
All of these losses have a property in common: **convexity**.

$$\ell_{\text{hinge}}(m) := \max(1 - m, 0)$$

$$\ell_{\text{perc}}(m) := \max(-m, 0)$$

$$\ell_{\text{log}}(m) := \log(1 + e^{-m})$$

$$\ell_{\text{square}}(m) := (1 - m)^2$$



Gradient Descent Guarantee

Convex, Smooth Functions

Recall: Convex functions are the functions where gradient descent is *guaranteed* to converge.

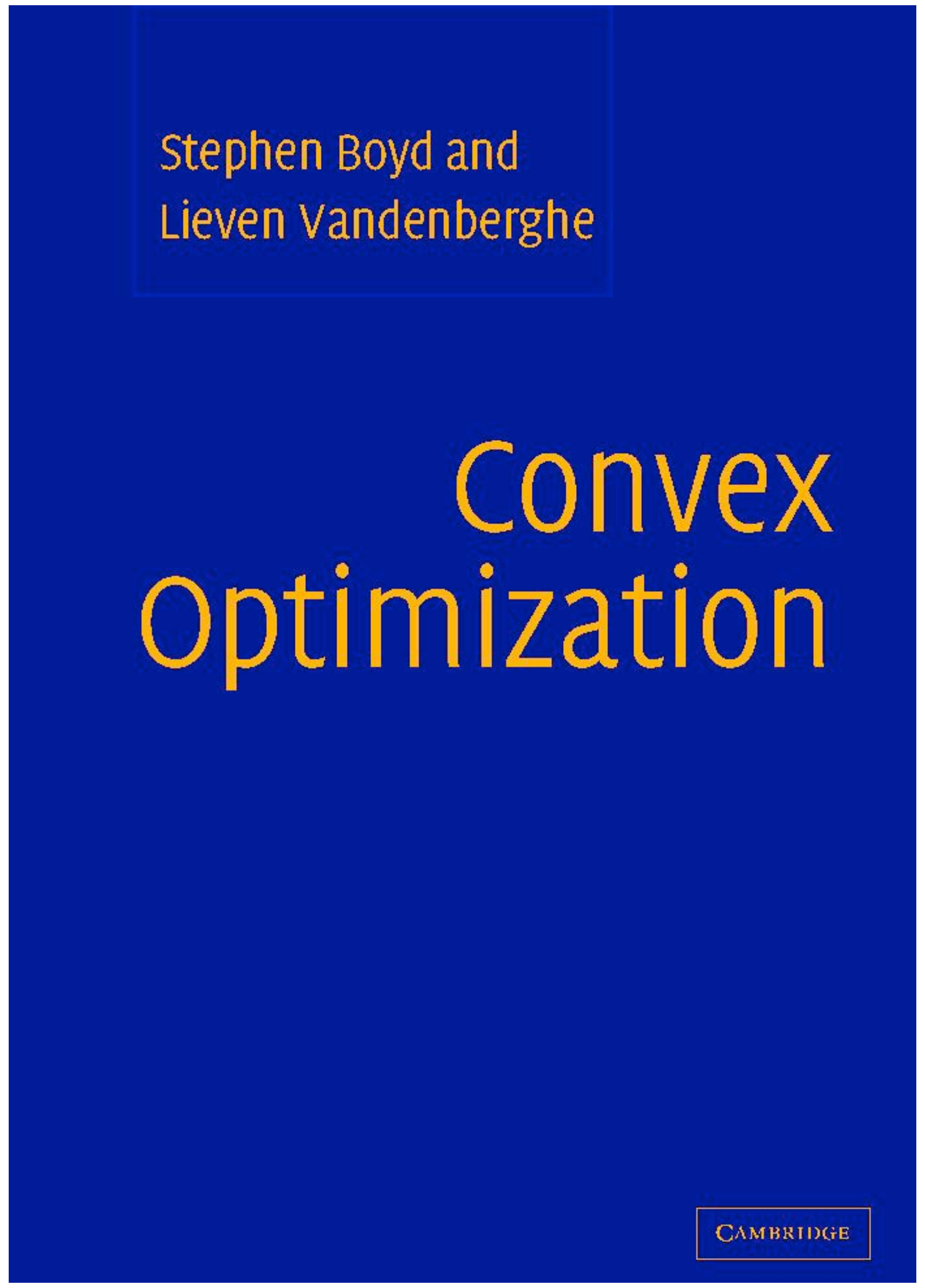
Theorem (GD on Convex, Smooth Functions). If $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex, differentiable, and L-smooth, then gradient descent with $\eta \leq 1/L$ converges:

$$F(w^{(T)}) - F(w*) \leq \frac{\|w^{(0)} - w*\|^2}{2\eta T} \text{ after } T \text{ steps.}$$

Convex Opt. Reference

Boyd & Vandenberghe (2004)

Standard, comprehensive reference for convex optimization is Boyd & Vandenberghe (2004).



Notation

From Boyd & Vandenberghe

$f : \mathbb{R}^d \to \mathbb{R}$ means that f maps from some *subset* of $\mathbb{R}^d$.

Write $\mathbf{dom}\, f \subset \mathbb{R}^d$, where $\mathbf{dom}\, f$ is the domain of f.

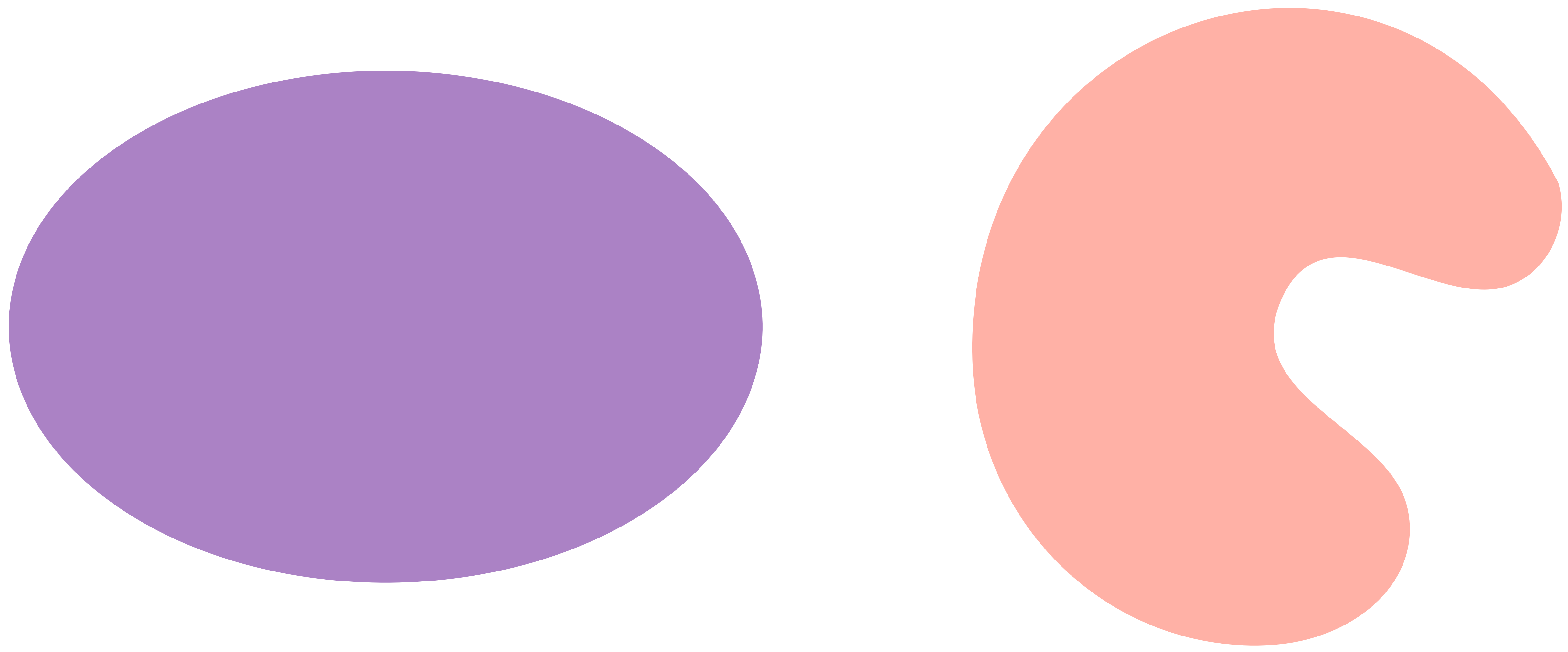
Convex Sets

Definition

A set C is convex if for any $x_1, x_2 \in C$ and any θ with $0 \leq \theta \leq 1$ we have

$$\theta x_1 + (1 - \theta)x_2 \in C.$$

"All line segments between points in the set are in the set."



Convex Functions

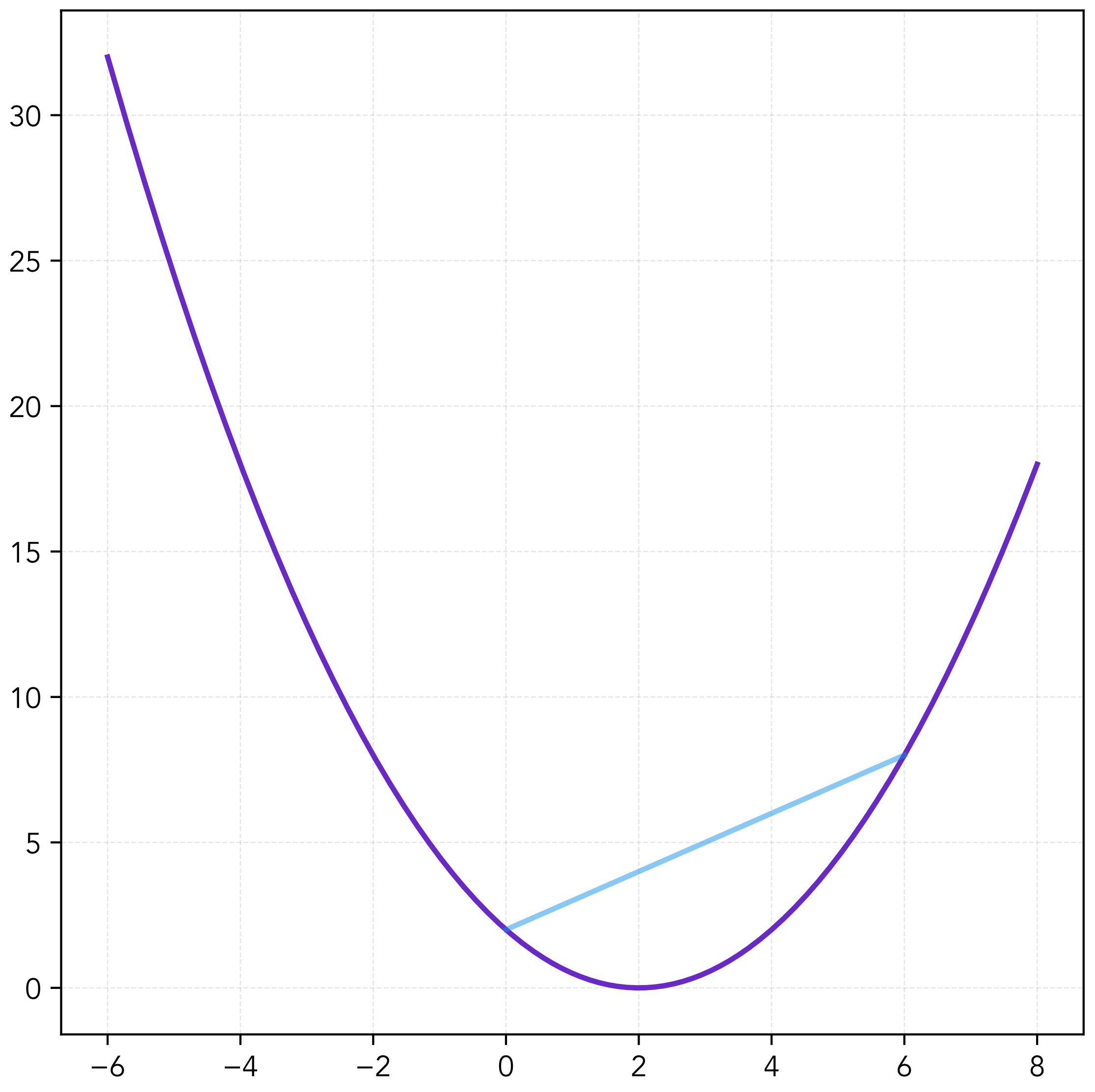
Definition

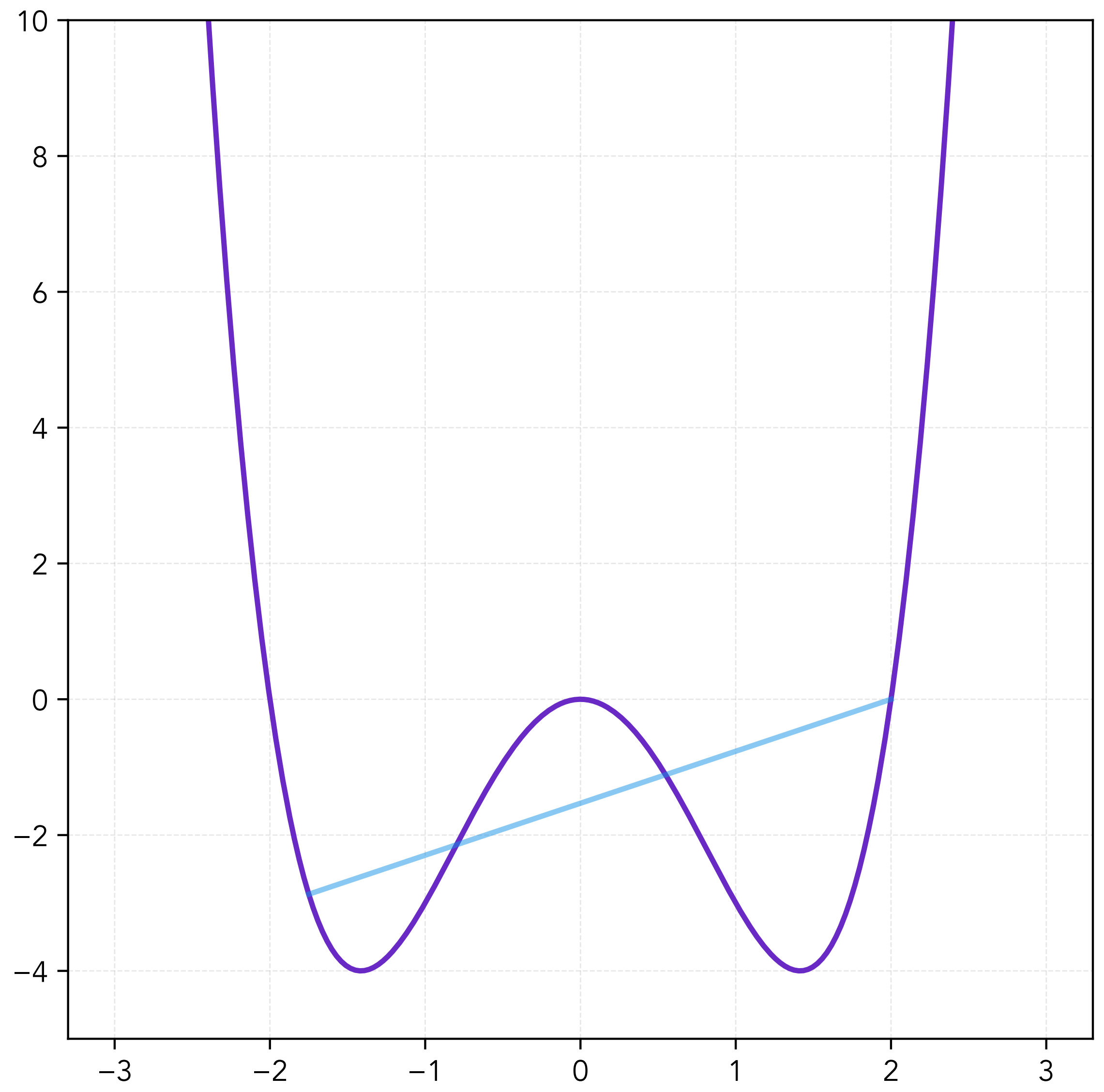
A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $\operatorname{dom} f$ is a convex set and if for all $x, y \in \operatorname{dom} f$ and $0 \leq \theta \leq 1$:

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y).$$

"All secant lines lie above the function."

Convex





Nonconvex

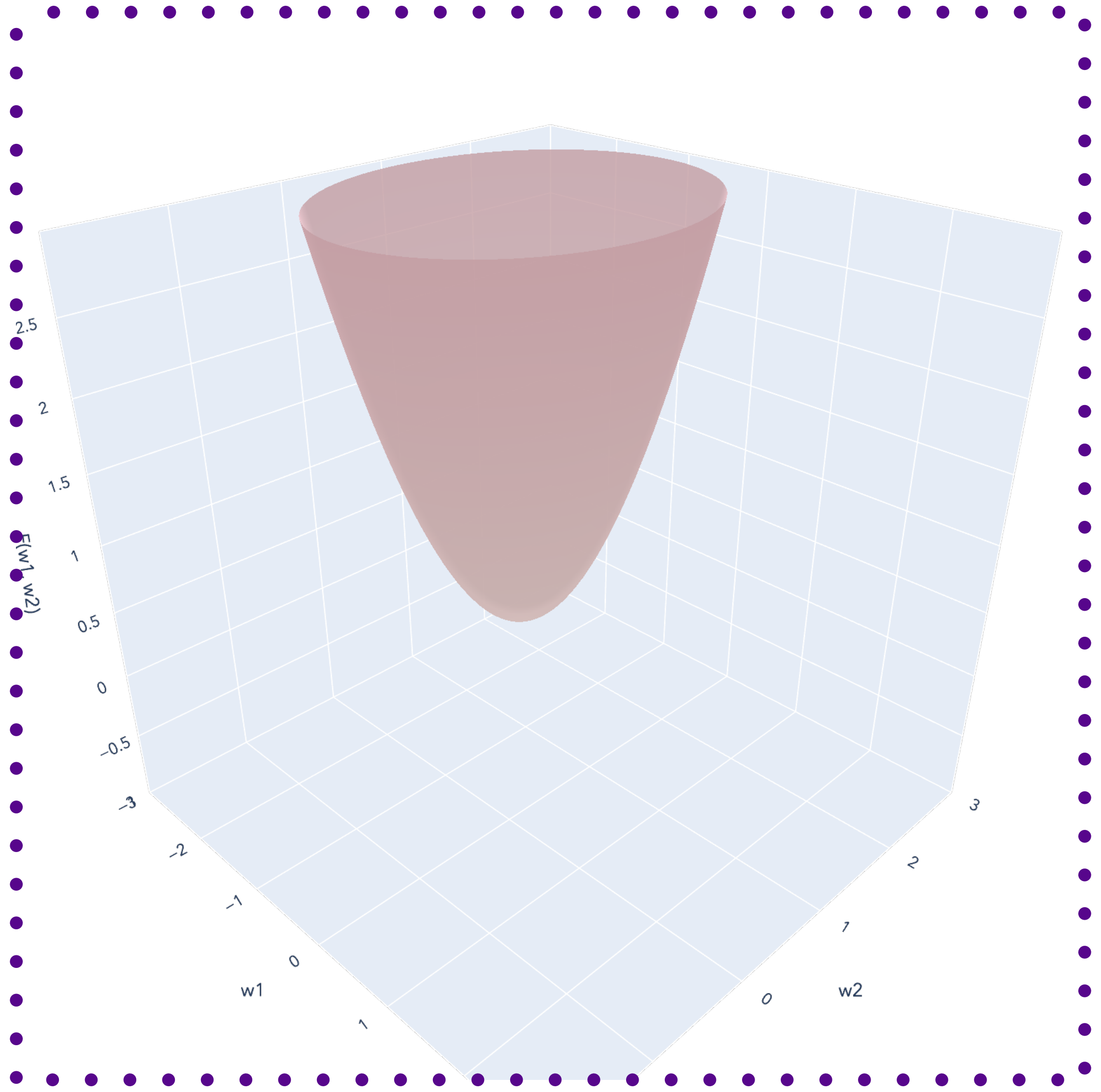
Convex Functions

Definition

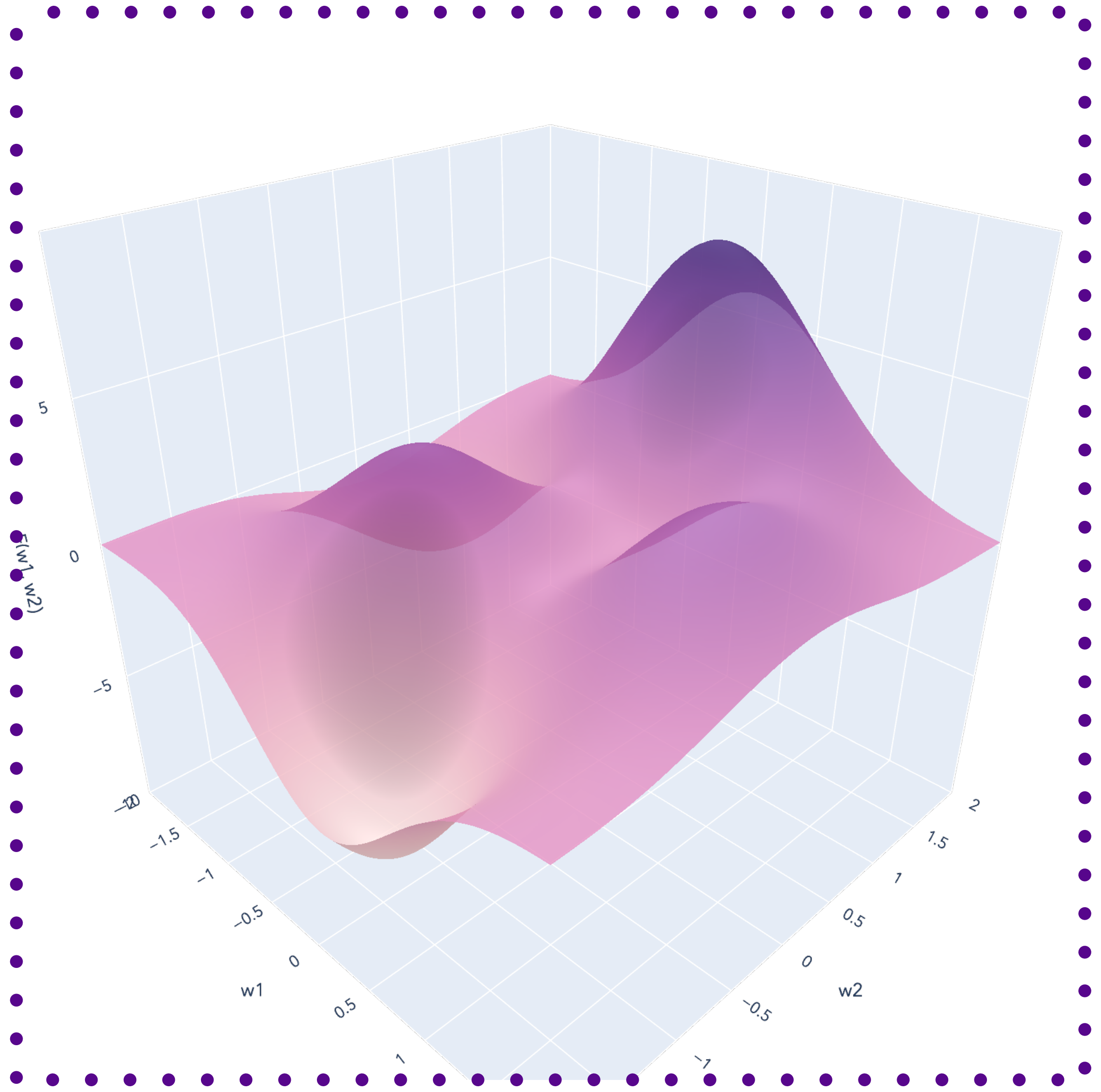
A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if $\mathrm{dom}\, f$ is a convex set and if for all $x, y \in \mathrm{dom}\, f$ and $0 \le \theta \le 1$:

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y).$$

"All secant lines lie above the function."



Convex



Nonconvex

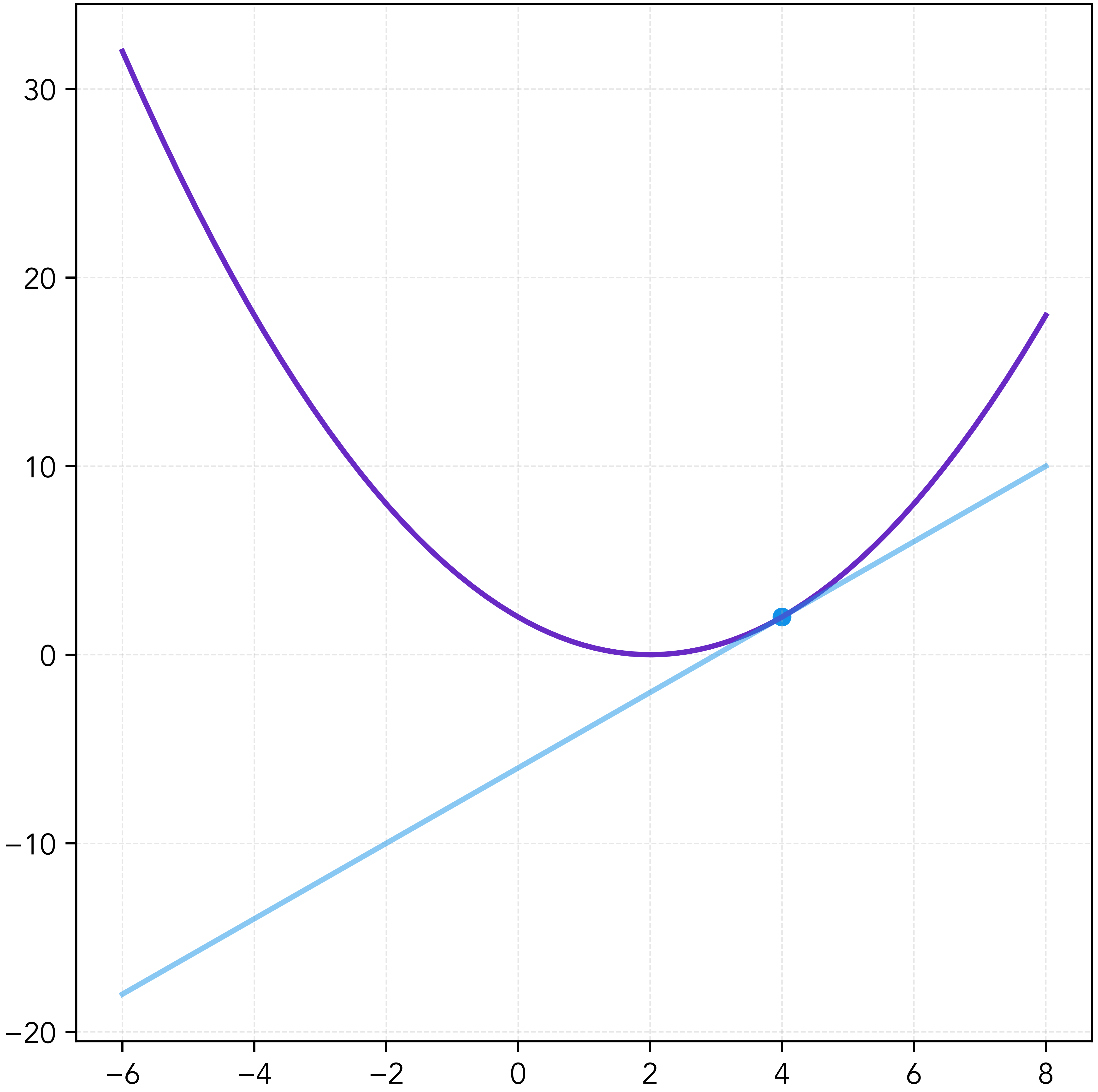
Convex Functions

First-order Condition

A *differentiable* function $f : \mathbb{R}^d \to \mathbb{R}$ is a convex if, for any $x, y \in \operatorname{dom} f$:

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)\,.$$

Tangent (*linear approximation*) at any x lies *below* the function.



Convex Functions

Second-order Condition

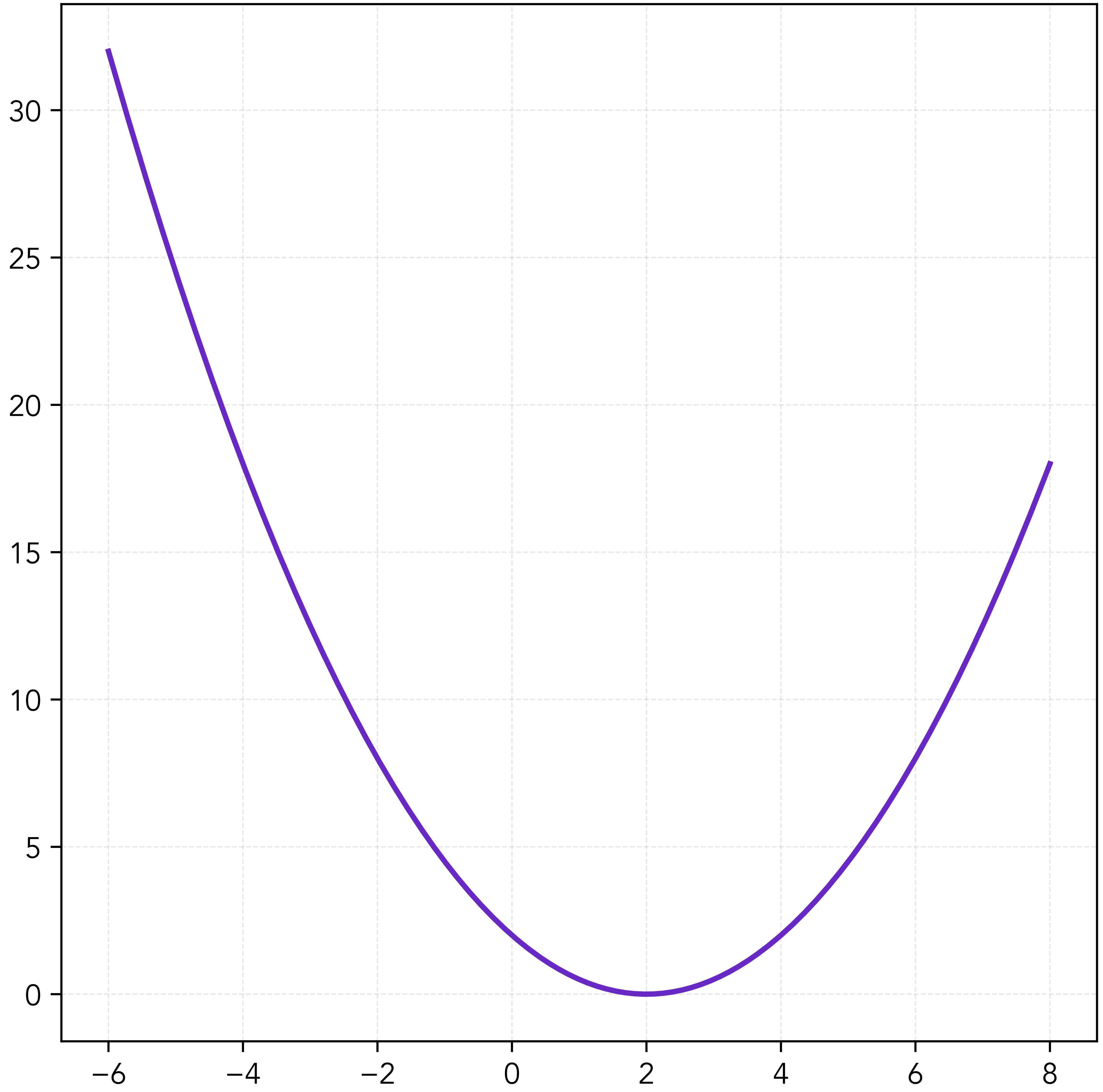
A *twice-differentiable* function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if, for any $x \in \operatorname{dom} f$, the Hessian $\nabla^2 f(x)$ is positive semidefinite:

$$d^\top \nabla^2 f(x) d \geq 0 \text{ for all } d \in \mathbb{R}^d.$$

$\iff$ Eigenvalues of $\nabla^2 f(x)$ are nonnegative.

$\iff$ There exists $A \in \mathbb{R}^{d \times r}$ s.t. $\nabla^2 f(x) = AA^\top$.

The function has a nonnegative "second derivative."



Common Convex Functions

Examples

Affine functions. $x \mapsto ax + b$ is both convex and concave on $\mathbb{R}$ for all $a, b \in \mathbb{R}$.

Powers. $x \mapsto \quad x \,^{p}$ for $p \geq 1$ is convex on $\mathbb{R}$.

Exponentials. $x \mapsto e^{ax}$ is convex on $\mathbb{R}$ for all $a \in \mathbb{R}$.

Logarithm. $x \mapsto \log x$ is concave for all $x \geq 0$.

Norms. All norms on $\mathbb{R}^d$ are convex (e.g. $\|x\|_1$ and $\|x\|_2$).

Maximum. $(x_1, \ldots, x_d) \mapsto \max\{x_1, \ldots, x_d\}$ is convex on $\mathbb{R}^d$.

Closure of Convex Functions

The “Algebra” of Convex Functions

We can also combine convex functions with operations that preserve convexity:

Nonnegative linear combination. If $f_1, \ldots, f_n$ convex, then $g(x) := \lambda_1 f_1(x) + \ldots + \lambda_n f_n(x)$ is convex.

Extends to infinite sums and integrals.

Pre-composition with affine function. If f is convex, so is $f(Ax + b)$.

Maximum. If $f_1, \ldots, f_n$ are convex, then $g(x) := \max\{f_1(x), \ldots, f_n(x)\}$ is convex.

Extends to pointwise supremum.

See *Boyd and Vandenberghe* Section 3.2 for comprehensive reference.

Outline

Convexity Primer

Convex Optimization

Convex Optimization: Duality

Constraint Qualification & Complementary Slackness

SVM Optimization Problem

SVM Dual Optimization

Strong Duality applied to SVM

Convex Optimization

Standard Form

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, \quad i = 1,\ldots,m, \\ & h_j(x) = 0, \quad j = 1,\ldots,k\,. \end{aligned}$$

where $x \in \mathbb{R}^d$ are the optimization/decision variables and f_0 is the objective function.

Convex Optimization

Terminology: Feasibility

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, \quad i = 1,\ldots,m, \\ & h_j(x) = 0, \quad j = 1,\ldots,k\,. \end{aligned}$$

The set of points satisfying the constraints is called the feasible set.

A point x in the feasible set is called a feasible point.

If x is feasible and $f_i(x) = 0$, then we say the equality constraint $f_i(x) \le 0$ is active at x.

Convex Optimization

Terminology: Optimality

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & h_j(x) = 0, \quad j = 1, \ldots, k\,. \end{aligned}$$

The optimal value $p*$ of the problem is defined as:

$$p* = \min\{f_0(x) : x \text{ satisfies all constraints}\}.$$

$x*$ is an optimal point (or a solution) if $x*$ is feasible and $f_0(x*) = p*$.

Convex Optimization

Equality Constraints

$$h(x) = 0 \iff h(x) \geq 0 \text{ AND } h(x) \leq 0.$$

Any equality-constrained problem

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & h(x) = 0 \end{aligned}$$

can be rewritten as:

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & h(x) \leq 0 \\ \text{s.t.} \quad & -h(x) \leq 0 \end{aligned}$$

So without loss off generality, we will only consider **inequality-constrained** optimization problems.

Outline

Convexity Primer

Convex Optimization

Convex Optimization: Duality

Constraint Qualification & Complementary Slackness

SVM Optimization Problem

SVM Dual Optimization

Strong Duality applied to SVM

Lagrangian

Definition

General (inequality-constrained) optimization problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^d} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \end{aligned}$$

The Lagrangian for this optimization problem is:

$$L(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

Each λ_i is the "price" we pay for violating constraint $f_i(x)$.

The λ_i are called the Lagrange multipliers (or dual variables).

Lagrangian

Encoding Constraints

Maximizing over the Lagrangian gives back encoding of objective and constraints:

$$\max_{\lambda \geq 0} L(x, \lambda) = \max_{\lambda \geq 0} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right)$$

$$= \begin{cases} f_0(x) & \text{when } f_i(x) \leq 0 \text{ for all } i \\ \infty & \text{otherwise} \end{cases}$$

Equivalent **primal form** of the optimization problem:

$$p* = \min_{x} \max_{\lambda \geq 0} L(x, \lambda).$$

Lagrangian

Primal and Dual

Original optimization problem in primal form:

$$p* = \min_{x} \max_{\lambda \geq 0} L(x, \lambda)$$

The Lagrangian dual problem comes from “swapping the $\min$ and the $\max$”:

$$d* = \max_{\lambda \geq 0} \min_{x} L(x, \lambda)$$

$p* \geq d*$ for *any* optimization problem (this is called weak duality).

Weak Max-Min Inequality

Theorem

Theorem (Weak Duality). For any $f : W \times Z \to \mathbb{R}$, we have:

$$\max_{z \in Z} \min_{w \in W} f(w, z) \leq \min_{w \in W} \max_{z \in Z} f(w, z).$$

Proof. For any $w_0 \in W$ and $z_0 \in Z$, by definition of $\min$ and $\max$:

$$\min_{w \in W} f(w, z_0) \leq f(w_0, z_0) \leq \max_{z \in Z} f(w_0, z).$$

Sine $\min_{w \in W} f(w, z_0) \leq \max_{z \in Z} f(w_0, z)$ for all w_0 and z_0, we must also have:

$$\max_{z_0 \in Z} \min_{w \in W} f(w, z_0) \leq \min_{w_0 \in W} \max_{z \in Z} f(w_0, z).$$

Weak Duality

Duality Gap

For any optimization problem, the weak max-min inequality implies weak duality:

$$p^* = \min_x \max_{\lambda \succeq 0} \left[f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right]$$

$$\geq \max_{\lambda \succeq 0} \min_x \left[f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right] = d^*$$

The difference $p^* - d^*$is called the duality gap.

For *convex problems*, we often have strong duality: $p^* = d^*$.

Dual Function

Definition

The Lagrangian dual problem:

$$d* = \max_{\lambda \succeq 0} \min_x L(x, \lambda).$$

The Lagrangian dual function (or just dual function) is:

$$g(\lambda) = \min_x L(x, \lambda) = \min_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right).$$

The dual function may take on the value $-\infty$ (one example: $f_0(x) = x$).

The dual function is always **concave** (it is pointwise minimum of affine functions).

Dual Function

Best Lower Bound

$$g(\lambda) = \min_x L(x, \lambda) = \min_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right)$$

In terms of the Lagrange dual function, we can write weak duality as:

$$p* \geq \max_{\lambda \geq 0} g(\lambda) = d*.$$

$p* \geq g(\lambda)$ for all $\lambda \geq 0.$

So any λ with $\lambda \geq 0$ in dual function gives a **lower bound** on the optimal solution.

Dual Function

Best Lower Bound

Weak duality: $p^* \geq \max_{\lambda \succeq 0} g(\lambda) = d^*$

The Lagrange dual problem is a search for the best lower bound on p^*:

$$\begin{aligned} \max \quad & g(\lambda) \\ \text{s.t.} \quad & \lambda \succeq 0 \end{aligned}$$

λ is dual feasible if $\lambda \succeq 0$ and $g(\lambda) > -\infty$ and dual optimal if, in addition, $g(\lambda) = d^*$.

Lagrange dual problem often easier to solve (simpler constraints) and can reveal structure.

d^* can be used as stopping criterion for primal optimization.

Outline

Convexity Primer

Convex Optimization

Convex Optimization: Duality

Constraint Qualification & Complementary Slackness

SVM Optimization Problem

SVM Dual Optimization

Strong Duality applied to SVM

Strong Duality

Convex Optimization

A convex optimization problem is a (possibly constrained) optimization problem

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \le 0, \quad i = 1,\ldots,m \end{aligned}$$

where $f_0, f_1, \ldots, f_m$ are all **convex functions.**

Strong Duality

Convex Optimization

For convex optimization problems, we *usually* have strong duality, but not always:

$$\begin{aligned} \min_{x,y} \quad & e^{-x} \\ \text{s.t.} \quad & x^2/y \leq 0 \\ & y > 0 \end{aligned}$$

The additional conditions needed for strong duality are called **constraint qualifications**.

Constraint Qualification

Slater's Conditions

When is $p^* = d^*$ (strong duality) for *convex optimization*?

Roughly: the problem must be **strictly** feasible (there is *some* solution).

Qualifications when problem domain $\mathscr{D} = \bigcap_{i=0}^{m} \mathrm{dom}\, f_i \subseteq \mathbb{R}^d$ is an open set:

Strict feasibility is sufficient (there exists x such that $f_i(x) < 0$ for all $i = 1,\ldots, m$).

For affine inequality constraints, finding x such that $f_i(x) \leq 0$ is sufficient.

If $\mathscr{D}$ is not open, see notes in B&V Section 5.2.3, pg. 226.

Complementary Slackness

Definition

If **strong duality** holds, we get an interesting relationship between:

Optimal Lagrange multiplier λ_i^* and

The ith constraint at the optimum: $f_i(x^*)$.

The relationship is called complementary slackness:

$$\lambda_i^* f_i(x^*) = 0$$

Always have Lagrange multiplier is zero **or** constraint is active at optimum **or** both.

Complementary Slackness

“Sandwich Proof”

Proof. Assume strong duality: $p^* = d^*$. Let x^* be primal optimal and let λ^* be dual optimal.

$$
\begin{aligned}
f_0(x^*) = g(\lambda^*) &= \min_x L(x, \lambda^*) \\
&\leq L(x^*, \lambda^*) \\
&= f_0(x^*) + \sum_{i=1}^{m} \underbrace{\lambda_i^* f_i(x^*)}_{\leq 0} \\
&\leq f_0(x^*)
\end{aligned}
$$

Each term in the sum $\sum_{i=1}^{m} \lambda_i^* f_i(x^*)$ must actually be 0. That is, $\lambda_i^* f_i(x_i^*) = 0$ for $i = 1,\ldots,m$.

Recipe for Using Dual

Summary

1. Unconstrain your constrained optimization problem by defining the Lagrangian.

2. Find the dual function $g(\lambda)$ by minimizing the Lagrangian over x.

3. Maximize the dual function over λ to get a **lower bound** on the primal (weak duality).

4. Check Slater's conditions to see if you have strong duality.

5. Strong duality $\Longrightarrow$ complementary slackness. Investigate complementary slackness for insights.

$$\begin{tikzcd}
L(x,\lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \arrow[d] \\
g(\lambda) = \min_x L(x,\lambda) \arrow[d, "\text{weak max-min duality}"] \\
p^* \geq \max_\lambda g(\lambda) = d^* \arrow[d, "\text{strong duality}"] \\
p^* = d^* \arrow[d] \\
\lambda_i^* f_i(x^*) = 0 \quad \forall i \in [m]
\end{tikzcd}$$

Outline

Convexity Primer

Convex Optimization

Convex Optimization: Duality

Constraint Qualification & Complementary Slackness

SVM Optimization Problem

SVM Dual Optimization

Strong Duality applied to SVM

Classification

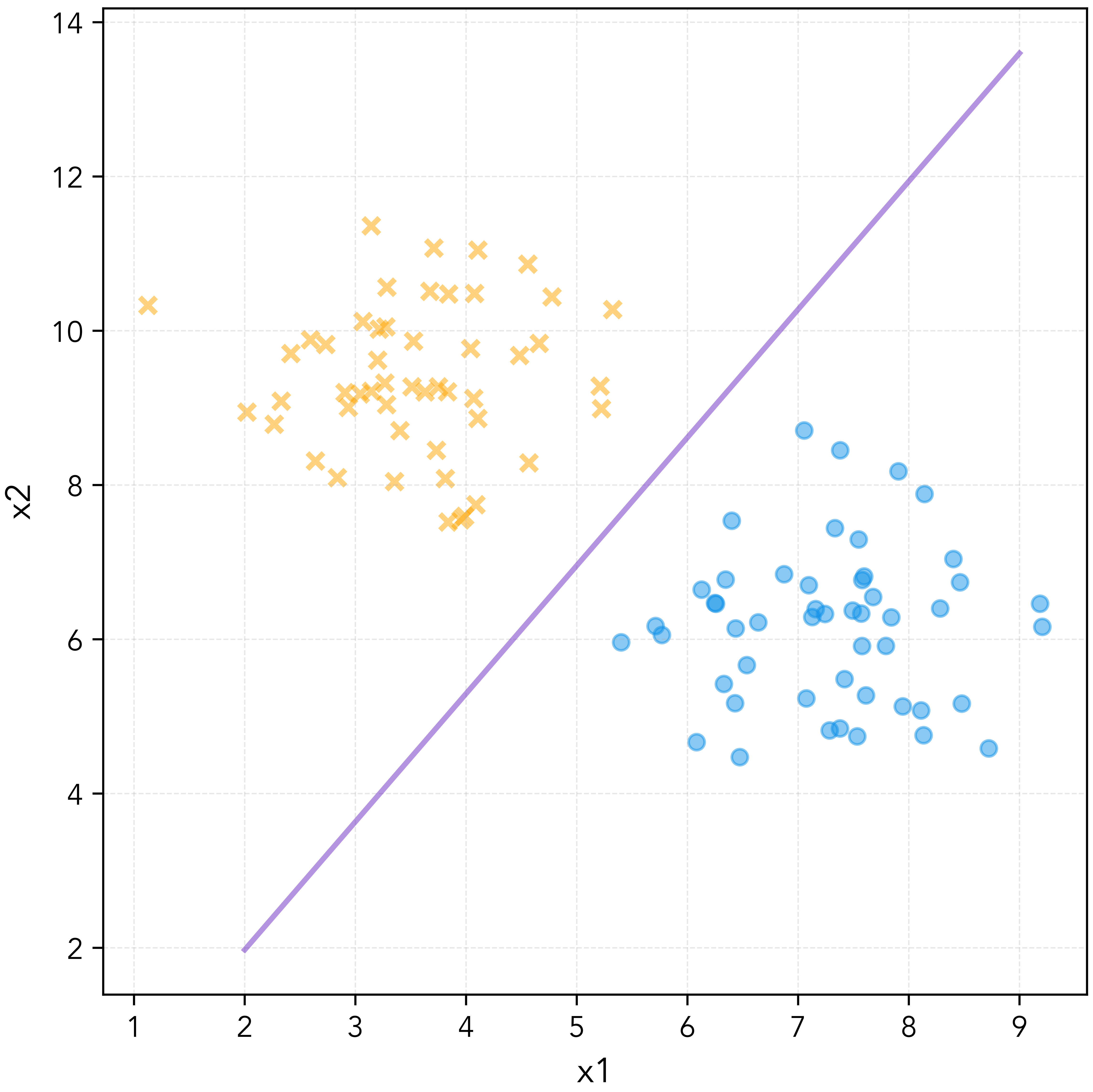
Geometric Picture

Input space: $\mathcal{X} = \mathbb{R}^d$

Action space: $\mathcal{A} = \{-1,1\}$

Outcome space: $\mathcal{Y} = \{-1,1\}$

We will focus on methods that induce **linear decision boundaries (hyperplanes).**



Classification

Problem Instance

Input space: $\mathcal{X} = \mathbb{R}^d$

Action space: $\mathcal{A} = \mathbb{R}$

Outcome space: $\mathcal{Y} = \{-1,1\}$

For a **linear** function $f(x) = w^\top x$, the semantics typically are:

$$w^\top x > 0 \implies \text{Predict } 1$$

$$w^\top x < 0 \implies \text{Predict } -1$$

Margin

Definition

The margin for a predicted score $\hat{y}$ and the true class $y \in \{-1,1\}$ is $y\hat{y}$.

With a score function $f: \mathcal{X} \to \mathbb{R}$, the margin is $yf(x)$.

If y and $\hat{y}$ are the same sign, prediction is **correct** and margin is **positive**.

If y and $\hat{y}$ have different sign, prediction is **incorrect** and margin is **negative**.

We want to find f that **maximizes** the margin.

Many classification losses only depend on the margin (margin-based losses).

Classification Losses

Convexity

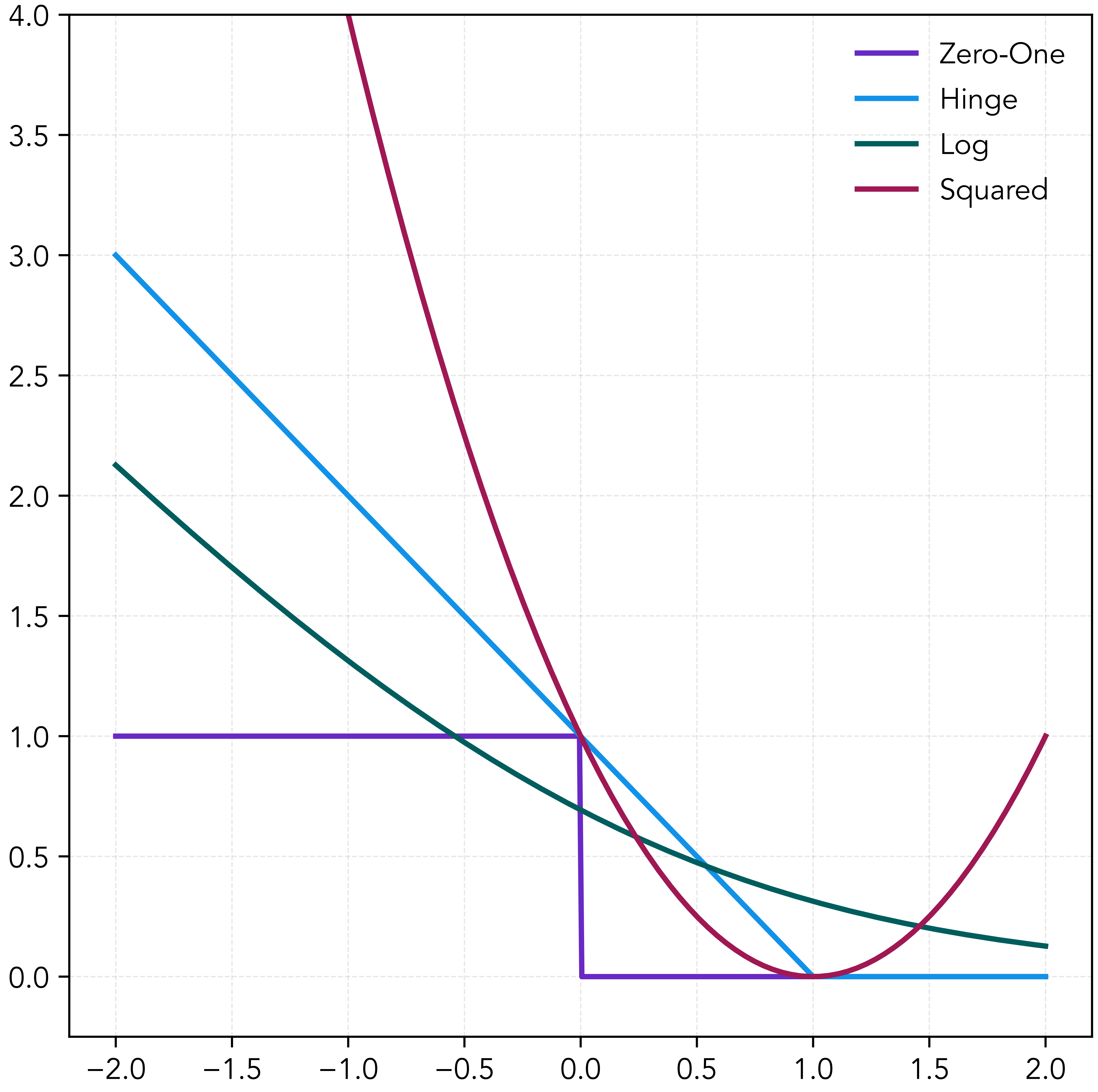
All of these losses have a property in common: **convexity**.

$$\ell_{\text{hinge}}(m) := \max(1 - m, 0)$$

$$\ell_{\text{perc}}(m) := \max(-m, 0)$$

$$\ell_{\log}(m) := \log(1 + e^{-m})$$

$$\ell_{\text{square}}(m) := (1 - m)^2$$



Classification Losses

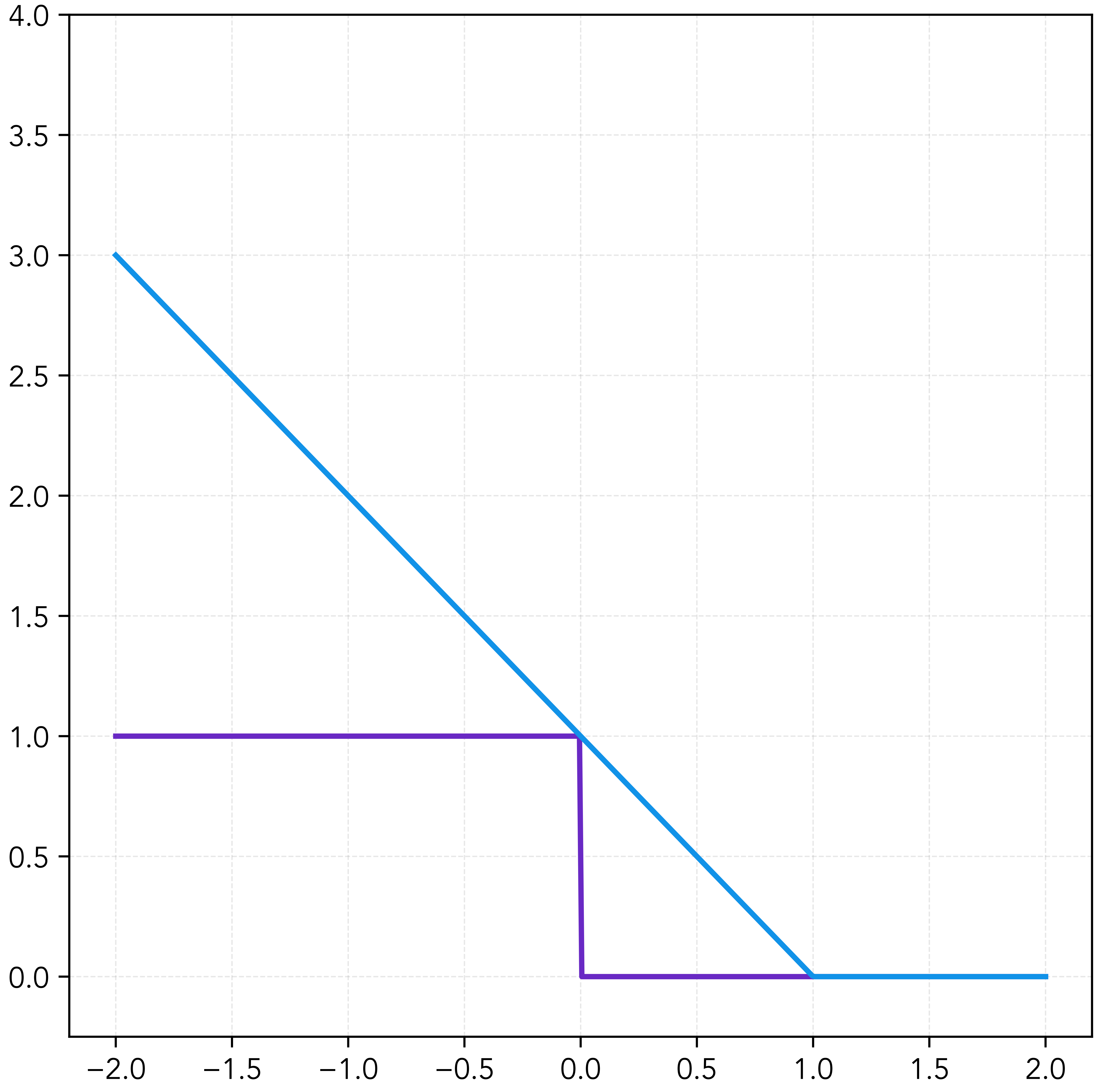
Hinge Loss

Margin: $m = \hat{y}y$

Hinge loss: $\ell_{\text{hinge}}(m) := \max(1 - m, 0)$

Hinge loss is **convex, upper bound** on zero-one loss.

Not differentiable at $m = 1$.



Hinge Loss

(Soft-Margin) Support Vector Machine

Hypothesis class: $\mathscr{H} = \{h_w(x) = w^\top x + b : w \in \mathbb{R}^d, b \in \mathbb{R}\}$

Loss: $\ell_{\text{hinge}}(m) = \max(1 - m, 0)$ (hinge loss)

Regularizer: ℓ_2

Empirical risk minimization:

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \max(1 - y^{(i)} h_w(x^{(i)}), 0) + \frac{C}{2} \|w\|_2^2$$

SVM Optimization Problem

Penalized ERM

Hypothesis class: $\mathscr{H} = \{h_w(x) = w^\top x + b : w \in \mathbb{R}^d, b \in \mathbb{R}\}$

Loss: $\ell_{\text{hinge}}(m) = \max(1 - m, 0)$ (hinge loss)

Regularizer: ℓ_2

Empirical risk minimization:

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{C}{n} \sum_{i=1}^{n} \max(1 - y^{(i)}(w^\top x^{(i)} + b), 0) + \frac{1}{2}\|w\|_2^2$$

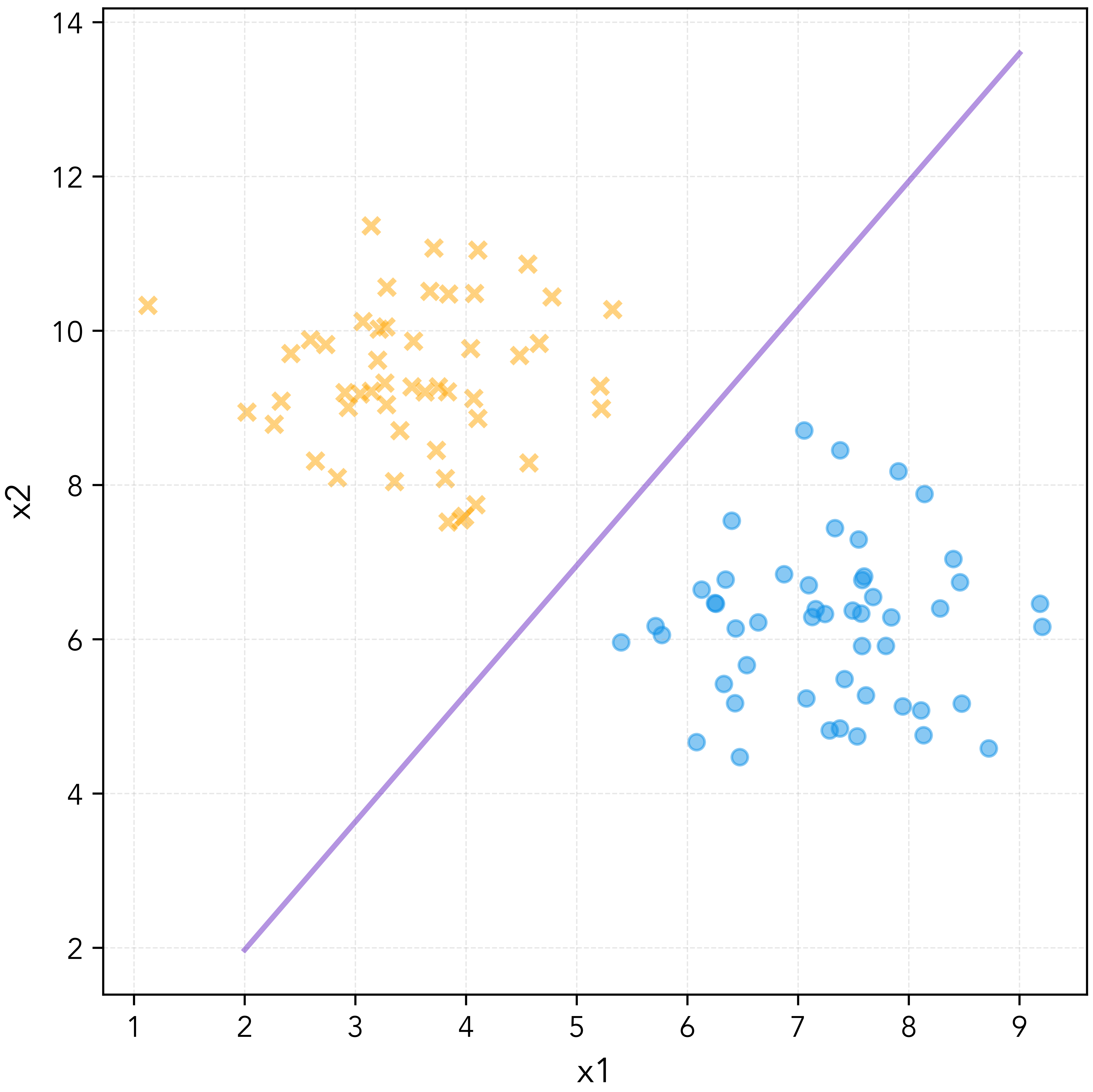
SVM Optimization

(Hyper)plane

The SVM hypothesis is the solution to:

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{C}{n}\sum_{i=1}^{n} \max(1 - y^{(i)}(w^\top x^{(i)} + b), 0) + \frac{1}{2}\|w\|_2^2$$

The w and b define an affine (hyper)plane in $\mathbb{R}^d$.



SVM Optimization

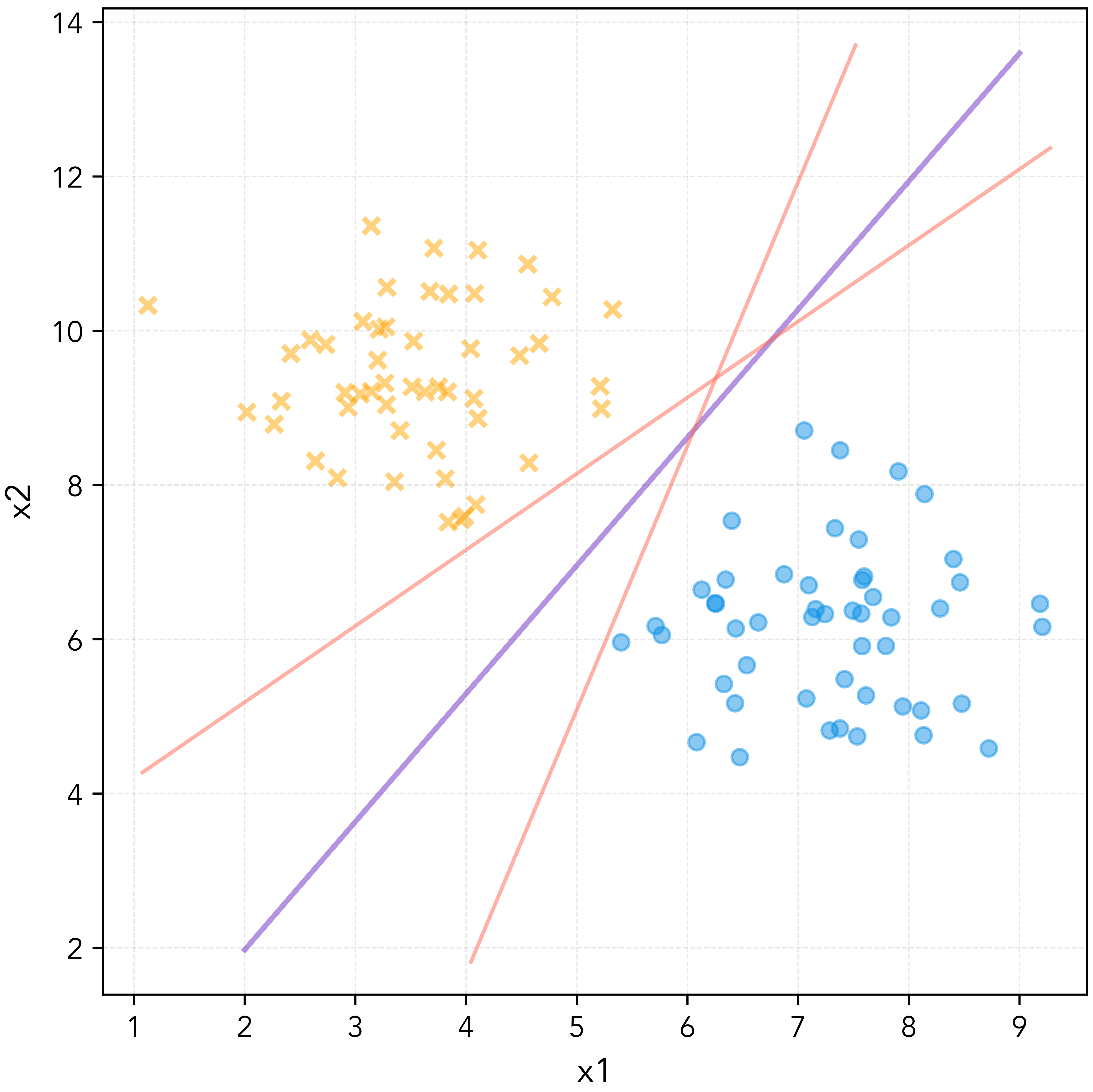
(Hyper)plane

The SVM hypothesis is the solution to:

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{C}{n} \sum_{i=1}^{n} \max(1 - y^{(i)}(w^\top x^{(i)} + b), 0) + \frac{1}{2}\|w\|_2^2$$

The w and b define an affine (hyper)plane in $\mathbb{R}^d$.

Turns out this has nice geometric properties (max geometric margin)!



SVM Optimization Problem

Penalized ERM

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{C}{n} \sum_{i=1}^{n} \max(1 - y^{(i)}(w^\top x^{(i)} + b), 0) + \frac{1}{2}\|w\|_2^2$$

Unconstrained optimization problem (penalized ERM).

Not differentiable because of the max (right at the “hinge” of the hinge loss).

Can we re-formulate into a differentiable problem?

SVM Optimization

Constrained ERM

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{C}{n} \sum_{i=1}^{n} \max(1 - y^{(i)}(w^\top x^{(i)} + b),\ 0) + \frac{1}{2} \|w\|_2^2$$

is equivalent to:

$$\begin{aligned} \min_{w \in \mathbb{R}^d, b \in \mathbb{R}, \xi \in \mathbb{R}^n} \quad & \frac{1}{2}\|w\|^2 + \frac{C}{n} \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & \xi_i \geq \max\left(1 - y^{(i)}(w^\top x^{(i)} + b),\ 0\right) \end{aligned}$$

SVM Optimization

Constrained ERM

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & \xi_i \geq \max\left(1 - y^{(i)}(w^\top x^{(i)} + b),\ 0\right) \end{aligned}$$

is equivalent to:

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & \xi_i \geq 1 - y^{(i)}(w^\top x^{(i)} + b) \quad \text{for } i = 1,\ldots,n \\ & \xi_i \geq 0 \quad \text{for } i = 1,\ldots,n \end{aligned}$$

SVM Optimization

…is just convex optimization

The SVM optimization problem is equivalent to the **convex optimization problem**:

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & \left(1 - y^{(i)}(w^\top x^{(i)} + b)\right) - \xi_i \le 0 \quad \text{for } i = 1,\ldots,n \\ & -\xi_i \le 0 \quad \text{for } i = 1,\ldots,n \end{aligned}$$

Objective function is differentiable and convex.

$n + d + 1$ unknowns and $2n$ affine constraints.

Now a quadratic program that can be solved using any off-the-shelf QP solver!

Outline

Convexity Primer

Convex Optimization

Convex Optimization: Duality

Constraint Qualification & Complementary Slackness

SVM Optimization Problem

SVM Dual Optimization

Strong Duality applied to SVM

Recipe for Using Dual

Summary

1. Unconstrain your constrained optimization problem by defining the Lagrangian.

2. Find the dual function $g(\lambda)$ by minimizing the Lagrangian over x.

3. Maximize the dual function over λ to get a **lower bound** on the primal (weak duality).

4. Check Slater's conditions to see if you have strong duality.

5. Strong duality $\implies$ complementary slackness. Investigate complementary slackness for insights.

$$\begin{tikzcd}
L(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \arrow[d] \\
g(\lambda) = \min_{x} L(x, \lambda) \arrow[d, "\text{weak max-min duality}"'] \\
p^* \geq \max_{\lambda} g(\lambda) = d^* \arrow[d, "\text{strong duality}"'] \\
p^* = d^* \arrow[d] \\
\lambda_i^* f_i(x^*) = 0 \quad \forall i \in [m]
\end{tikzcd}$$

Dual SVM Problem

Lagrange Multipliers

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & \left(1 - y^{(i)}(w^\top x^{(i)} + b)\right) - \xi_i \le 0 \quad \text{for } i = 1,\ldots,n \\ & -\xi_i \le 0 \quad \text{for } i = 1,\ldots,n \end{aligned}$$

Lagrange multiplier $\alpha_i \iff$ Constraint $\left(1 - y^{(i)}(w^\top x^{(i)} + b)\right) - \xi_i \le 0.$

Lagrange multiplier $\lambda_i \iff$ Constraint $-\xi_i \le 0.$

Lagrangian: $L(w, b, \xi, \alpha, \lambda) = \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - y^{(i)}(w^\top x^{(i)} + b) - \xi_i\right) + \sum_{i=1}^{n}\lambda_i(-\xi_i)$

Dual SVM Problem

Weak Duality

$$L(w, b, \xi, \alpha, \lambda) = \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - y^{(i)}(w^\top x^{(i)} + b) - \xi_i\right) + \sum_{i=1}^{n}\lambda_i(-\xi_i)$$

$$\iff L(w, b, \xi, \alpha, \lambda) = \frac{1}{2}w^\top w + \sum_{i=1}^{n}\xi_i\left(\frac{C}{n} - \alpha_i - \lambda_i\right) + \sum_{i=1}^{n}\alpha_i\left(1 - y^{(i)}\left(w^\top x^{(i)} + b\right)\right)$$

By **weak duality**: $p^* = \min_{w,\xi,b}\max_{\alpha,\lambda\geq 0} L(w, b, \xi, \alpha, \lambda) \geq \max_{\alpha,\lambda\geq 0}\min_{w,\alpha,b} L(w, b, \xi, \alpha, \lambda) = d^*$.

Do we have **strong duality**:

$$p^* = d^*?$$

Constraint Qualification

Recall: Slater's Conditions

When is $p^* = d^*$ (strong duality) for *convex optimization*?

Roughly: the problem must be **strictly** feasible (there is *some* solution).

Qualifications when problem domain $\mathscr{D} = \bigcap_{i=0}^{m} \mathrm{dom}\, f_i \subseteq \mathbb{R}^d$ is an open set:

Strict feasibility is sufficient (there exists x such that $f_i(x) < 0$ for all $i = 1, \ldots, m$).

For affine inequality constraints, finding x such that $f_i(x) \leq 0$ is sufficient.

If $\mathscr{D}$ is not open, see notes in B&V Section 5.2.3, pg. 226.

Checking Strong Duality

Slater's Condition

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & \left(1 - y^{(i)}(w^\top x^{(i)} + b)\right) - \xi_i \leq 0 \quad \text{for } i = 1,\ldots,n \\ & -\xi_i \leq 0 \quad \text{for } i = 1,\ldots,n \end{aligned}$$

Convex problem + affine constraints $\implies$ strong duality iff the problem is feasible.

Constraints are satisfied by $w = b = 0$ and $\xi_i = 1$ for $i = 1,\ldots,n$.

Therefore, we do have strong duality!

$$p* = \min_{w,\xi,b}\max_{\alpha,\lambda\geq 0} L(w,b,\xi,\alpha,\lambda) = \max_{\alpha,\lambda\geq 0}\min_{w,\alpha,b} L(w,b,\xi,\alpha,\lambda) = d*$$

Dual Function

Recall

$$g(\lambda) = \min_x L(x, \lambda) = \min_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right)$$

In terms of the Lagrange dual function, we can write weak duality as:

$$p^* \geq \max_{\lambda \geq 0} g(\lambda) = d^*.$$

$$p^* \geq g(\lambda) \text{ for all } \lambda \geq 0.$$

So any λ with $\lambda \geq 0$ in dual function gives a **lower bound** on the optimal solution.

If strong duality holds: $p^* = g(\lambda^*) = d^*$

Lagrangian Dual

How to find the Lagrangian dual?

Lagrangian dual is the $\min$ over primal variables of the Lagrangian:

$$
\begin{aligned}
g(\alpha, \lambda) &= \min_{w,b,\xi} L(w, b, \xi, \alpha, \lambda) \\
&= \min_{w,b,\xi} \left[\frac{1}{2} w^\top w + \sum_{i=1}^{n} \xi_i \left(\frac{C}{n} - \alpha_i - \lambda_i \right) + \sum_{i=1}^{n} \alpha_i \left(1 - y^{(i)} \left(w^\top x^{(i)} + b \right) \right) \right]
\end{aligned}
$$

Taking the $\min$ of convex and differentiable function of w, b, ξ.

Quadratic in w and linear in ξ and b.

Thus, optimal point iff $\partial_w L = 0$, $\partial_b L = 0$, and $\partial_\xi L = 0$.

Lagrangian Dual

Taking derivatives

$$
\begin{aligned}
g(\alpha, \lambda) &= \min_{w,b,\xi} L(w, b, \xi, \alpha, \lambda) \\
&= \min_{w,b,\xi} \left[\frac{1}{2} w^\top w + \sum_{i=1}^{n} \xi_i \left(\frac{C}{n} - \alpha_i - \lambda_i \right) + \sum_{i=1}^{n} \alpha_i \left(1 - y^{(i)} \left(w^\top x^{(i)} + b \right) \right) \right]
\end{aligned}
$$

$$
\partial_w L = 0 \iff w - \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} = 0 \iff w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}
$$

$$
\partial_b L = 0 \iff -\sum_{i=1}^{n} \alpha_i y^{(i)} = 0 \iff \sum_{i=1}^{n} \alpha_i y^{(i)} = 0
$$

$$
\partial_\xi L = 0 \iff \frac{C}{n} - \alpha_i - \lambda_i = 0 \iff \alpha_i + \lambda_i = \frac{C}{n}
$$

Lagrangian Dual

Plugging back in to the dual

$$w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}$$

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

$$\alpha_i + \lambda_i = \frac{C}{n}$$

$$\begin{aligned} g(\alpha, \lambda) &= \min_{w,b,\xi} L(w, b, \xi, \alpha, \lambda) \\ &= \min_{w,b,\xi} \left[\frac{1}{2} w^\top w + \sum_{i=1}^{n} \xi_i \left(\frac{C}{n} - \alpha_i - \lambda_i \right) + \sum_{i=1}^{n} \alpha_i \left(1 - y^{(i)} \left(w^\top x^{(i)} + b \right) \right) \right] \end{aligned}$$

Dual Optimization Problem

Maximum over the Lagrangian Dual

$$
\begin{aligned}
\max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(j)})^\top x^{(i)} \\
\text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y^{(i)} = 0 \\
& \alpha_i \in \left[0, \frac{C}{n}\right] \quad \text{for } i = 1, \ldots, n
\end{aligned}
$$

Given solution α^* to dual, the primal solution is $w^* = \sum_{i=1}^{n} \alpha_i^* y^{(i)} x^{(i)}$ (in the "span of the data")

Regularization parameter C controls the max weight put on each example: $\alpha_i^* \in \left[0, \frac{C}{n}\right]$.

SVM Optimization

Dual Optimization Problem

$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(j)})^\top x^{(i)}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

$$\alpha_i \in \left[0, \frac{C}{n}\right] \quad \text{for } i = 1, \ldots, n$$

Quadratic objective with n unknowns and $n + 1$ constraints.

What other insights can we get from the dual formulation?

SVM Optimization

Primal and Dual

$$
\begin{aligned}
\min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\
\text{s.t.} \quad & \left(1 - y^{(i)}(w^\top x^{(i)} + b)\right) - \xi_i \le 0 \quad \text{for } i = 1,\ldots,n \\
& -\xi_i \le 0 \quad \text{for } i = 1,\ldots,n
\end{aligned}
$$

$$
\begin{aligned}
\max_{\alpha} \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}(x^{(j)})^\top x^{(i)} \\
\text{s.t.} \quad & \sum_{i=1}^{n}\alpha_i y^{(i)} = 0 \\
& \alpha_i \in \left[0, \frac{C}{n}\right] \quad \text{for } i = 1,\ldots,n
\end{aligned}
$$

Outline

Convexity Primer

Convex Optimization

Convex Optimization: Duality

Constraint Qualification & Complementary Slackness

SVM Optimization Problem

SVM Dual Optimization

Strong Duality applied to SVM

Classification Losses

Hinge Loss

$$f^*(x) = x^\top w^* + b^*$$

Margin: $m = yf^*(x)$

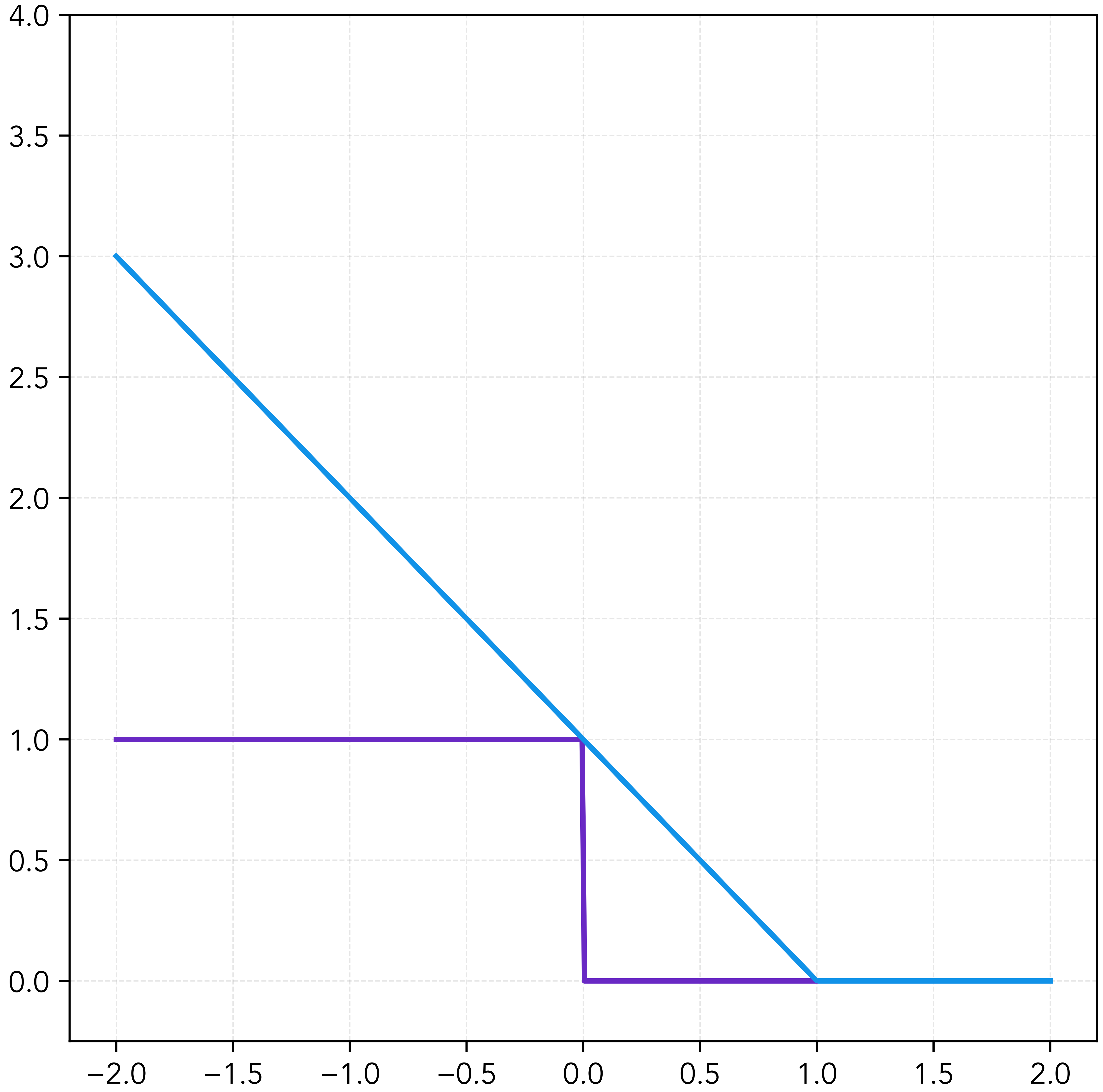
$$\ell_{\text{hinge}}(yf^*(x)) := \max(1 - yf^*(x), 0)$$

Incorrect: $yf^*(x) \leq 0.$

"Margin error": $yf^*(x) < 1.$

"On the margin": $yf^*(x) = 1$

"Good side of margin": $yf^*(x) > 1.$



Support Vectors

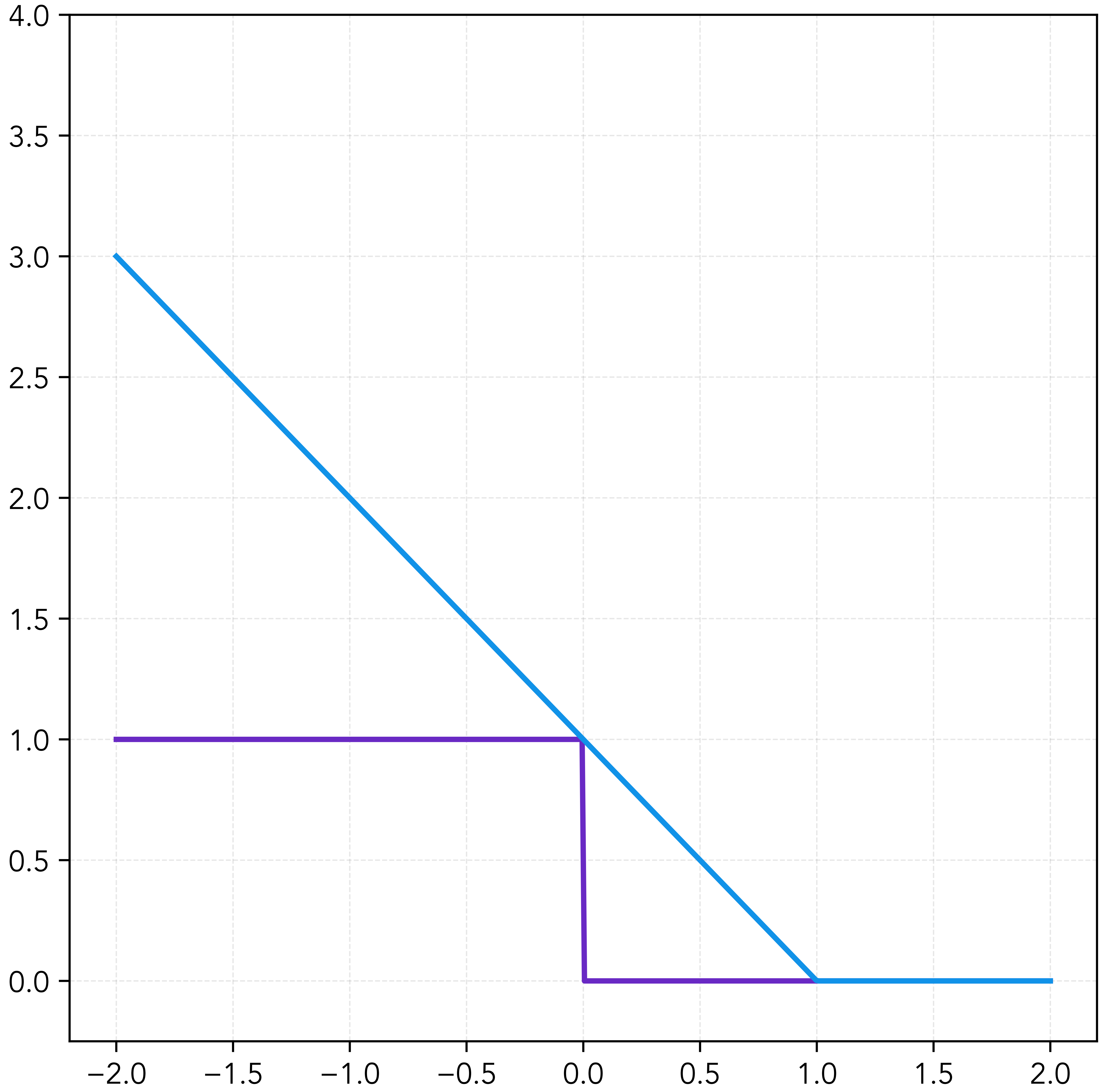
Relationship to margin

Slack variable $\xi_i^* = \max(1 - y^{(i)} f^*(x^{(i)}),\ 0)$ is the hinge loss on $(x^{(i)}, y^{(i)})$.

Suppose $\xi_i^* = 0$. Then, $y^{(i)} f^*(x^{(i)}) \geq 1$, i.e.

> "On the margin" $(= 1)$, or
>
> "On the good side" (> 1).

$$\xi_i^* = 0 \iff y^{(i)} f^*(x^{(i)}) \geq 1$$



Complementary Slackness

Recall

If **strong duality** holds, we get an interesting relationship between:

Optimal Lagrange multiplier λ_i^* and

The ith constraint at the optimum: $f_i(x^*)$.

The relationship is called complementary slackness:

$$\lambda_i^* f_i(x^*) = 0$$

Always have Lagrange multiplier is zero **or** constraint is active at optimum **or** both.

Strong Duality

Complementary Slackness

Lagrange multiplier $\lambda_i \iff$ Constraint $-\xi_i \leq 0.$

Lagrange multiplier $\alpha_i \iff$ Constraint $\left(1 - y^{(i)}(w^\top x^{(i)} + b)\right) - \xi_i \leq 0.$

Recall first-order condition $\partial_{\xi_i} L = 0$ gave us $\lambda_i^* = \dfrac{C}{n} - \alpha_i^*.$

By strong duality, **complementary slackness**:

$$\lambda_i^* \xi_i^* = \left(\frac{C}{n} - \alpha_i^*\right) \xi_i^* = 0$$

$$\alpha_i^*(1 - y^{(i)} f^*(x^{(i)}) - \xi_i^*) = 0$$

Strong Duality

Complementary Slackness

$$\lambda_i^* \xi_i^* = \left(\frac{C}{n} - \alpha_i^*\right) \xi_i^* = 0$$

$$\alpha_i^*(1 - y^{(i)} f^*(x^{(i)}) - \xi_i^*) = 0$$

If $y^{(i)} f^*(x^{(i)}) > 1 \implies$ margin loss $\xi_i^* = 0$ so we get $\alpha_i^* = 0$.

If $y^{(i)} f^*(x^{(i)}) < 1 \implies$ margin loss $\xi_i^* > 0$ so $\alpha_i^* = \frac{C}{n}$.

If $\alpha_i^* = 0 \implies \xi_i^* = 0$, which implies no loss, so $y^{(i)} f^*(x^{(i)}) \geq 1$.

If $\alpha_i^* \in \left(0, \frac{C}{n}\right) \implies \xi_i^* = 0$, which implies $1 - y^{(i)} f^*(x^{(i)}) = 0$.

Strong Duality

Summary of Complementary Slackness

$$\alpha_i^* = 0 \implies y^{(i)} f^*(x^{(i)}) \geq 1$$

$$\alpha_i^* \in \left(0, \frac{C}{n}\right) \implies y^{(i)} f^*(x^{(i)}) = 1$$

$$\alpha_i^* = \frac{C}{n} \implies y^{(i)} f^*(x^{(i)}) \leq 1$$

$$y^{(i)} f^*(x^{(i)}) < 1 \implies \alpha_i^* = \frac{C}{n}$$

$$y^{(i)} f^*(x^{(i)}) = 1 \implies \alpha_i^* \in \left[0, \frac{C}{n}\right]$$

$$y^{(i)} f^*(x^{(i)}) > 1 \implies \alpha_i^* = 0$$

When $y^{(i)} f^*(x^{(i)}) > 1$ (*good side of margin*), we are guaranteed $\alpha_i^* = 0$.

When $y^{(i)} f^*(x^{(i)}) = 1$ (*exactly on margin*), we *could* have $\alpha_i^* = 0$ or $\alpha_i^* > 0$.

When $y^{(i)} f^*(x^{(i)}) < 1$ (*bad side of margin*), we are guaranteed $\alpha_i^* > 0$.

Strong Duality

Support Vector Interpretation

If α^* is a solution to the dual problem, the primal solution is:

$$w^* = \sum_{i=1}^{n} \alpha_i^* y^{(i)} x^{(i)} \quad \text{with } \alpha_i^* \in \left[0, \frac{C}{n}\right]$$

The $x^{(i)}$'s corresponding to $\alpha_i^* > 0$ are called **support vectors**.

By comp. slackness, correspond to points *on the margin* or *on bad side of margin*.

Few margin errors or "on the margin" examples $\Longrightarrow$ **sparsity in input examples**.

Strong Duality

Getting $b*$

$$\lambda_i^* \xi_i^* = \left(\frac{C}{n} - \alpha_i^* \right) \xi_i^* = 0$$

$$\alpha_i^*(1 - y^{(i)} f^*(x^{(i)}) - \xi_i^*) = 0$$

Suppose there's an i such that $\alpha_i^* \in \left(0, \frac{C}{n}\right)$.

$$\lambda_i^* \xi_i^* = \left(\frac{C}{n} - \alpha_i^* \right) \xi_i^* = 0 \implies \xi_i^* = 0$$

$$\alpha_i^*(1 - y^{(i)} f^*(x^{(i)}) - \xi_i^*) = 0 \implies y^{(i)}((x^{(i)})^\top w* + b*) = 1 \iff (x^{(i)})^\top w* + b* = y^{(i)}$$

$$\iff b* = y^{(i)} - (x^{(i)})^\top w*$$

Strong Duality

Getting $b*$

Therefore, the optimal b is:

$$b* = y^{(i)} - (x^{(i)})^\top w*.$$

We get the same $b*$ for any choice of i with $\alpha_i^* \in \left(0, \frac{C}{n}\right)$.

If there are no $\alpha_i^* \in \left(0, \frac{C}{n}\right)$? Then we have a **degenerate SVM training problem** ($w* = 0$).

Dual Problem

Teaser for Kernelization

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(j)})^\top x^{(i)} \\ \text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y^{(i)} = 0 \\ & \alpha_i \in \left[0, \frac{C}{n}\right] \quad \text{for } i = 1, \ldots, n \end{aligned}$$

All dependence on inputs $x^{(i)}$ and $x^{(j)}$ is through the inner product $\langle x^{(j)}, x^{(i)} \rangle = (x^{(j)})^\top x^{(i)}$.

What if we replace $(x^{(j)})^\top x^{(i)}$ with some other inner product?

Outline

Convexity Primer

Convex Optimization

Convex Optimization: Duality

Constraint Qualification & Complementary Slackness

SVM Optimization Problem

SVM Dual Optimization

Strong Duality applied to SVM