Convex Optimization Full Structure Summary

0. Global Structure

$$\text{Convex set} \to \text{Convex function} \to \text{Line above graph} \to \text{Tangent below graph}$$
$$\to \text{Local} = \text{Global minimum} \to \text{Optimization becomes clean}$$

With smoothness:

$$\text{Smooth} \to \text{(12.5) Quadratic upper bound} \to \text{GD one-step decrease}$$
$$\to \text{Self-bounded} \to \text{ML convergence analysis}$$

1. Convex Set

Definition.

A set C is convex iff

$$\alpha u + (1-\alpha)v \in C, \quad \forall u, v \in C,\ \alpha \in [0,1].$$

This represents the entire line segment between u and v.

2. Convex Function

Definition.

$$f(\alpha x_1 + (1-\alpha)x_2) \le \alpha f(x_1) + (1-\alpha) f(x_2).$$

Meaning: the graph lies below the line connecting two points.
Example:

$$f(x) = x^2 \quad \text{convex}$$

$$f(x) = -x^2 \quad \text{concave}$$

3. First-Order Characterization

Convex $\iff$

$$f(v) \geq f(w) + \langle \nabla f(w), v - w \rangle.$$

The graph always lies above its tangent plane.

4. 1D Characterization

$$f \text{ convex} \iff f' \text{ increasing} \iff f'' \geq 0.$$

5. Epigraph

$$\mathrm{epi}(f) = \{(x, \beta) : \beta \geq f(x)\}.$$

$$f \text{ convex} \iff \mathrm{epi}(f) \text{ convex}.$$

6. Local = Global

For convex f:

$$\nabla f(w^*) = 0 \quad \Rightarrow \quad w^* \text{ is global minimum}.$$

7. Closure Properties

(1) Maximum

If each f_i is convex, then

$$\max_i f_i \text{ is convex}.$$

Key inequality:

$$\max_i(\alpha a_i + (1-\alpha) b_i) \leq \alpha \max_i a_i + (1-\alpha) \max_i b_i.$$

(2) Nonnegative Weighted Sum

$$g = \sum_i w_i f_i, \quad w_i \geq 0 \Rightarrow g \text{ convex}.$$

8. Lipschitz

$$|f(w_1) - f(w_2)| \le \rho \|w_1 - w_2\|.$$

1D case:

$$|f'| \le \rho.$$

9. Smoothness

$$\|\nabla f(v) - \nabla f(w)\| \le \beta \|v - w\|.$$

1D:

$$|f''| \le \beta.$$

10. Smooth $\Rightarrow$ Quadratic Upper Bound (12.5)

$$f(v) \le f(w) + \langle \nabla f(w), v - w \rangle + \frac{\beta}{2}\|v - w\|^2.$$

11. Gradient Descent Decrease

GD update:

$$v = w - \frac{1}{\beta}\nabla f(w).$$

Plugging into (12.5):

$$f(v) \le f(w) - \frac{1}{2\beta}\|\nabla f(w)\|^2.$$

So

$$f(w) - f(v) \ge \frac{1}{2\beta}\|\nabla f(w)\|^2.$$

12. Self-Bounded Property

If $f \ge 0$, then

$$\|\nabla f(w)\|^2 \le 2\beta f(w).$$

13. Composition and $\beta\|x\|^2$

Let

$$f(w) = g(\langle w, x \rangle).$$

Chain rule:

$$\nabla f = g'(\cdot)x.$$

Smoothness gives:

$$\beta\|x\|^2.$$

14. ML Loss Examples

Squared loss:

$$(\langle w, x \rangle - y)^2 \quad \text{is } 2\|x\|^2\text{-smooth.}$$

Logistic loss:

$$\log(1 + e^{-y\langle w, x \rangle}) \quad \text{is } \frac{\|x\|^2}{4}\text{-smooth.}$$

15. Why Scaling Matters

$$\beta \propto \|x\|^2.$$

GD condition:

$$\eta \leq \frac{1}{\beta}.$$

Large data scale $\Rightarrow$ large β $\Rightarrow$ small step size $\Rightarrow$ slower convergence.

16. Convex Learning Problem

$$L_S(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(w, z_i).$$

If each ℓ convex $\Rightarrow$ L_S convex.
If each ℓ smooth $\Rightarrow$ L_S smooth.

Final Core Insight

Convex:

Mixing never decreases below structure.

Smooth:

Curvature is controlled.

Together:

Optimization is fully analyzable.

PAC Learnability: Full Structure

1. True Meaning of PAC Learnability

A hypothesis class $\mathcal{H}$ is PAC learnable if:

$$\forall D, \quad \text{for sufficiently large } m, \quad L_D(A(S)) \leq \min_{w \in \mathcal{H}} L_D(w) + \varepsilon$$

with probability at least $1 - \delta$.
Core requirement:

Capacity must be controlled.

Typical sufficient conditions:

- Finite VC dimension
- Finite Rademacher complexity
- Uniform convergence holds

These are different formalizations of the same idea.

2. Finite Hypothesis Class

If $|\mathcal{H}| < \infty$, then

$$m = O\left(\frac{\log |\mathcal{H}| + \log(1/\delta)}{\varepsilon}\right)$$

is sufficient.
Hence:

$$\text{Finite } \mathcal{H} \Rightarrow \text{Always PAC learnable.}$$

3. Infinite but Learnable

Example: Halfspaces in $\mathbb{R}^d$.

$$\mathcal{H} = \{x \mapsto \text{sign}(\langle w, x \rangle)\}$$

Even though w is infinite-dimensional:

$$\text{VCdim} = d + 1$$

Finite $\Rightarrow$ PAC learnable.

4. Why the Counterexample Fails

Consider:

$$\mathcal{H} = \mathbb{R}$$

Squared loss:

$$\ell(w, (x, y)) = (wx - y)^2$$

Key issue:

$$\sup_w |L_S(w) - L_D(w)| \not\to 0$$

Why?

- Hypothesis space unbounded
- Loss unbounded
- Rare points can explode loss

Uniform convergence fails.
Therefore:

$$\forall A, \exists D \text{ such that } A \text{ fails.}$$

Not PAC learnable.

Convex Alone Is Not Enough

Convexity ensures optimization is easy.
But it does NOT ensure generalization.
What is needed:

- Bounded hypothesis space
- Lipschitz or smooth loss

These prevent loss explosion.

Convex–Lipschitz–Bounded Setting

Assume:

- $\mathcal{H}$ convex and $\|w\| \leq B$
- $\ell(w, z)$ convex
- ℓ is ρ-Lipschitz:
$$|\ell(w_1, z) - \ell(w_2, z)| \leq \rho\|w_1 - w_2\|$$

This ensures stability.

Convex–Smooth–Bounded Setting

Assume:

- $\mathcal{H}$ convex and bounded
- ℓ convex, nonnegative
- ℓ is β-smooth:
$$\|\nabla\ell(w_1, z) - \nabla\ell(w_2, z)\| \leq \beta\|w_1 - w_2\|$$

Smoothness controls curvature.

Why This Guarantees Learnability

Key step: Uniform convergence.

$$\sup_{w \in \mathcal{H}} |L_D(w) - L_S(w)| \leq \alpha$$

If this holds, then ERM generalizes.
Let

$$\hat{w} = \arg\min_w L_S(w), \quad w^* = \arg\min_w L_D(w).$$

Then:

$$L_D(\hat{w}) \leq L_S(\hat{w}) + \alpha \leq L_S(w^*) \leq L_D(w^*) + \alpha$$

Thus:

$$L_D(\hat{w}) \leq L_D(w^*) + 2\alpha.$$

Choosing $\alpha = \varepsilon/2$ gives:

$$L_D(\hat{w}) \leq \min_w L_D(w) + \varepsilon.$$

Surrogate Loss

Why Surrogate Is Needed

0–1 loss:

$$\ell^{0-1}(w,(x,y)) = \mathbf{1}[y\langle w,x\rangle \le 0]$$

Problems:

- Non-convex
- Non-differentiable
- ERM is NP-hard

Hinge Loss

$$\ell^{hinge}(w,(x,y)) = \max\{0, 1 - y\langle w,x\rangle\}$$

Properties:

- Convex
- Upper bound on 0–1 loss

$$\ell^{0-1} \le \ell^{hinge}.$$

Generalization with Surrogate

We obtain:

$$L_D^{hinge}(A(S)) \le \min_w L_D^{hinge}(w) + \varepsilon.$$

Using upper bound:

$$L_D^{0-1}(A(S)) \le L_D^{hinge}(A(S)).$$

Therefore:

$$L_D^{0-1}(A(S)) \le \min_w L_D^{hinge}(w) + \varepsilon.$$

Error Decomposition

$$L_D^{0-1}(A(S)) \le \min_w L_D^{0-1}(w) + \underbrace{\left(\min_w L_D^{hinge}(w) - \min_w L_D^{0-1}(w)\right)}_{\text{surrogate gap}} + \varepsilon.$$

Three sources of error:

- Approximation error
- Estimation error
- Surrogate gap

Final Core Insight

- Finite $\mathcal{H} \Rightarrow$ PAC
- Infinite $\mathcal{H}$ is fine if capacity controlled
- Convex alone does NOT imply learnability
- Convex + bounded + Lipschitz/smooth

$$\Rightarrow \text{Uniform convergence} \Rightarrow \text{ERM generalizes}$$