Week 7: Real-Valued Losses and Scale-Sensitive Complexity

Rademacher complexity, norm bounds, margins, and kernels

Tianhao Wang
tianhaowang@ucsd.edu

UCSD · Spring 2026
DSC 190/291 Topics: Learning Theory

Contents

Bridge from Week 6

The course story so far

Weeks 1-2

Complexity = count of labelings. VC dimension, growth function, shattering, Halving. No free lunch, inductive bias, finite classes.

Weeks 3-4

Non-uniform complexity (MDL / SRM / PAC-Bayes). PAC learning, uniform convergence, fundamental theorem. Computational hardness.

Weeks 5-6

Real-valued scores begin to appear. Agnostic PAC, surrogate losses, neural networks. Weak learning, AdaBoost, margin.

Generalization is controlled by the complexity of how the class behaves on the sample.

- Binary classification: behaviors are ± 1 labelings $\rightarrow$ VC / growth function (counting).

Why binary VC theory is not enough

Most modern learning is real-valued:

- regression: predict real targets (prices, ratings, scores);
- scale-sensitive: bound generalization by $\|w\|$ and margin γ, not just $\mathsf{sign}(f(x))$;
- convex surrogate losses: hinge, logistic, squared (week 5).

Where behaviors live on the sample:

- Binary: $\{-1, +1\}^n$. The growth function counts them (at most 2^n).
- Real-valued: $\mathbb{R}^n$. Uncountably many. Counting fails.

Plan for today

Each part: what aspect of the behaviors do we measure?

Part	**What we measure**	**Tool**
Generalized learning	the loss class $\mathcal{F}$ on S	empirical risk L_S
Counting route	behaviors up to resolution α	covering numbers, pseudo-dimension
Random-sign route	correlation with random signs	Rademacher complexity
Norm and margin	constrained by $\|w\|$ and margin γ	fat-shattering, ℓ_2, ℓ_1, kernels

In real-valued learning, generalization is controlled by scale-sensitive complexity.

Generalized Learning

General learning problem

A learning problem is specified by:

- a space of examples $\mathcal{Z}$;
- a hypothesis space $\mathcal{H}$;
- a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$;
- a distribution $\mathcal{D}$ over $\mathcal{Z}$, accessed only through an i.i.d. sample $S = (z_1, ..., z_n) \sim \mathcal{D}^n$.

Population and empirical objectives:

$$L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)], \quad L_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, z_i).$$

The agnostic learning goal is still:

$$L_{\mathcal{D}}(A(S)) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

> Same PAC philosophy, but ℓ may be **real-valued** and h may **not be a classifier**.

Examples of general learning

Binary classification

labeled example: $z = (x, y) \in \mathcal{X} \times \{-1, +1\}$
classifier: $h : \mathcal{X} \to \{-1, +1\}$
$0/1$ loss: $\ell(h, z) = \mathbf{1}[h(x) \neq y]$

Regression

labeled example: $z = (x, y) \in \mathcal{X} \times \mathbb{R}$
predictor: $h : \mathcal{X} \to \mathbb{R}$
squared loss: $\ell(h, z) = (h(x) - y)^2$

Clustering (k-means)

point: $z \in \mathbb{R}^d$
k centers: $h = (\mu_1, ..., \mu_k)$
k-means loss: $\ell(h, z) = \min_i \|\mu_i - z\|^2$

Density estimation

observation: $z \in \mathcal{Z}$
density model: h specifies a density p_h
negative log-likelihood: $\ell(h, z) = -\log p_h(z)$

Binary surrogates (with $h : \mathcal{X} \to \mathbb{R}$): hinge $(1 - yh(x))_+$, logistic $\log\left(1 + e^{-yh(x)}\right)$, exp $e^{-yh(x)}$.

Real-valued ℓ is the rule binary $0/1$ is the special case.

ERM: the natural learning rule

We cannot see the population risk $L_{\mathcal{D}}(h)$, but we can compute the empirical risk:

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

Empirical Risk Minimization (ERM) returns the h with smallest empirical risk:

$$\hat{h} = \mathsf{ERM}_{\mathcal{H}}(S) \in \arg\min_{h \in \mathcal{H}} L_S(h).$$

If L_S is close to $L_{\mathcal{D}}$ throughout $\mathcal{H}$, then $L_{\mathcal{D}}\left(\hat{h}\right)$ should be close to $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$.

ERM is our algorithm. When does the closeness hold?

- Previously: uniform convergence of 0/1 loss on binary labels

The loss class

To analyze the closeness, look at L_S:

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

Each h enters only through the function $z \to \ell(h, z)$.

Collect these per-hypothesis functions: the **loss class** is

$$\mathcal{F} = \{z \to \ell(h, z) : h \in \mathcal{H}\}.$$

So $L_S(h)$ is the sample average of the corresponding $f \in \mathcal{F}$.

Empirical risk is controlled by the behavior of $\mathcal{F}$ on S.

Two standard ingredients

Concentration for fixed h. If $0 \le \ell(h, z) \le a$, Hoeffding gives

$$|L_{\mathcal{D}}(h) - L_S(h)| \le a\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}} \quad \text{w.p.} \ge 1 - \delta.$$

But $\hat{h} = \mathsf{ERM}_{\mathcal{H}}(S)$ depends on S, so we need a **uniform** bound.

Uniform concentration implies ERM works. If $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \le \frac{\epsilon}{2}$, then for any $h^* \in \mathcal{H}$,

$$\begin{aligned} L_{\mathcal{D}}\left(\hat{h}\right) &\le L_S\left(\hat{h}\right) + \frac{\epsilon}{2} && \text{(concentration at } \hat{h}\text{)} \\ &\le L_S(h^*) + \frac{\epsilon}{2} && \text{(ERM: } \hat{h} \text{ minimizes } L_S\text{)} \\ &\le L_{\mathcal{D}}(h^*) + \epsilon && \text{(concentration at } h^*\text{)}. \end{aligned}$$

Taking inf over h^*: $L_{\mathcal{D}}\left(\hat{h}\right) \le \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$.

The proof pipeline

Three stages to control $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$:

1. Fixed h

Hoeffding for one h.

$\rightarrow$

2. Sup over $\mathcal{H}$

Bound the sup via behavior of $\mathcal{F}$ on S.

$\rightarrow$

3. ERM

Sup bound makes ERM ϵ-optimal.

Stages 1 and 3 are familiar. Stage 2 is the **new work this week**, with two routes:

- **covering numbers and pseudo-dimension** (Part 2);
- **symmetrization and Rademacher complexity** (Part 3).

Now we dive into Stage 2.

Counting Real-Valued Behaviors

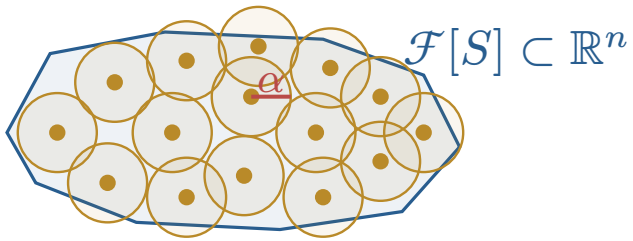
Why the growth function is no longer enough

For binary classes, a sample $S = (z_1, \ldots, z_n)$ induces finitely many behaviors:

$$\mathcal{F}[S] = \{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\} \subset \{-1, +1\}^n.$$

For real-valued losses, $\mathcal{F}[S] \subset \mathbb{R}^n$, and there may be **infinitely many** such vectors.

Covering as a remedy. Pick a finite set $V \subset \mathbb{R}^n$ so that every vector lies within α of some $v \in V$:



Replace exact counting by counting at resolution α.

Empirical covering numbers

An **α-cover** of $\mathcal{F}$ on sample $S = (z_1, \ldots, z_n)$ in ℓ_p is a finite subset $\mathcal{F}_\alpha \subset \mathcal{F}$ such that every $f \in \mathcal{F}$ has some anchor $g \in \mathcal{F}_\alpha$ with

$$\left(\frac{1}{n} \sum_{i=1}^{n} |f(z_i) - g(z_i)|^p \right)^{\frac{1}{p}} \leq \alpha.$$

The **empirical covering number** is the size of the smallest such cover:

$$\mathcal{N}_p(\mathcal{F}, \alpha, S) = \min\{|\mathcal{F}_\alpha| : \mathcal{F}_\alpha \text{ is an } \alpha\text{-cover of } \mathcal{F} \text{ on } S\}, \quad \mathcal{N}_p(\mathcal{F}, \alpha, n) = \sup_{|S|=n} \mathcal{N}_p(\mathcal{F}, \alpha, S).$$

Two common choices:

- $p = \infty$: every coordinate within α;
- $p = 2$: average squared coordinate error $\leq \alpha^2$.

Counts the α-resolutions of behavior that $\mathcal{F}$ admits on S.

From covering numbers to generalization

Sketch. Pick an α-cover $\mathcal{F}_\alpha = \{g_1, ..., g_N\} \subset \mathcal{F}$ of size $N = \mathcal{N}_\infty$, with $g_k = \ell(h_k, \cdot)$. For any h, let h_k be its anchor. Triangle inequality:

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \underbrace{|L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h_k)|}_{\text{pop. approx.}} + \underbrace{|L_{\mathcal{D}}(h_k) - L_S(h_k)|}_{\text{concentration}} + \underbrace{|L_S(h_k) - L_S(h)|}_{\text{sample approx.}}.$$

- **Sample-side approximation** (from the cover): $\leq \alpha$.
- **Anchor concentration** (Hoeffding + union over N anchors): $\leq O\left(a\sqrt{\frac{\log N + \log\left(\frac{1}{\delta}\right)}{n}}\right)$.
- **Population-side approximation needs symmetrization** (omitted)

Combining: with probability $\geq 1 - \delta$,

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq 2\alpha + O\left(a\sqrt{\frac{\log \mathcal{N}_\infty + \log\left(\frac{1}{\delta}\right)}{n}}\right).$$

Optimizing α trades the two terms: **how does $\mathcal{N}_\infty$ grow in $\frac{1}{\alpha}$ and n?**

Pseudo-dimension: VC for real-valued classes

How to extend VC dimension? For binary classes, shatter = realize every ± 1 pattern. For real-valued $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$, the values are continuous, so what does "shatter" mean?

Idea. Place a threshold θ_i at each z_i and ask: is $f(z_i)$ above or below?

Shattering. $\mathcal{F}$ shatters $z_1, \ldots, z_k$ if there exist thresholds $\theta_1, \ldots, \theta_k \in \mathbb{R}$ such that **every** binary pattern $s \in \{-1, +1\}^k$ is realized:

$$\exists f \in \mathcal{F} \quad \text{with} \quad \operatorname{sign}(f(z_i) - \theta_i) = s_i \quad \forall i.$$

The **pseudo-dimension** $\operatorname{Pdim}(\mathcal{F})$ is the largest such k. Equivalently, it is the VC dimension of the binary **subgraph class**:

$$\operatorname{Pdim}(\mathcal{F}) = \operatorname{VCdim}(\{(z, \theta) \to \mathbf{1}[f(z) \le \theta] : f \in \mathcal{F}\}).$$

Real-valued shattering = above/below patterns at chosen levels.

Pseudo-dimension controls covering

Covering via pseudo-dimension, Pollard

If $\mathcal{F} \subset [-a, a]^{\mathcal{Z}}$ and $\mathrm{Pdim}(\mathcal{F}) \le D$, then for $0 < \alpha \le a$,

$$\mathcal{N}_\infty(\mathcal{F}, \alpha, n) \le \left(\frac{ena}{D\alpha}\right)^D.$$

How to read it. A class with pseudo-dim D behaves like a **D-parameter family**:

- taking log, $\log \mathcal{N}_\infty(\mathcal{F}, \alpha, n) \le D \log(ena/(D\alpha))$;
- linear in D, logarithmic in $\frac{1}{\alpha}$ and n.

Compare to the binary Sauer–Shelah bound $\tau_{\mathcal{H}}(n) \le \left(e\frac{n}{d}\right)^d$: same shape, with α-resolution replacing scale 0.

Pseudo-dim D controls $\log \mathcal{N}_\infty$ at rate $D \log\left(n\frac{a}{\alpha}\right)$.

Example: linear real-valued predictors

Take $\mathcal{H} = \{x \to \langle w, x\rangle : w \in \mathbb{R}^d\}$. The subgraph $\mathbf{1}[f(x) \le \theta]$ becomes

$$\langle w, x\rangle - \theta \le 0 \quad \equiv \quad \langle (w, -1), (x, \theta)\rangle \le 0,$$

a halfspace through the origin in $\mathbb{R}^{d+1}$.

Halfspaces in $\mathbb{R}^{d+1}$ have VC dimension at most $d+1$, so $\mathrm{Pdim}(\mathcal{H}) \le d+1$.

Plugging into the covering theorem:

$$\mathcal{N}_\infty(\mathcal{H}, \alpha, n) \le \left(e\frac{n}{\alpha}\right)^{d+1}.$$

Pseudo-dim scales with the number of parameters, just like VC.

Which class gets pseudo-dimension?

ERM averages the **loss class**

$$\mathcal{F} = \{(x, y) \to \ell(h(x), y) : h \in \mathcal{H}\},$$

so we need $\mathrm{Pdim}(\mathcal{F})$, not $\mathrm{Pdim}(\mathcal{H})$.

Composition rules. If $\ell(\hat{y}, y)$ is well-behaved in the prediction $\hat{y} = h(x)$:

- ℓ **monotone** in $\hat{y}$ for each y (e.g., hinge $\max(0, 1 - y\hat{y})$): $\mathrm{Pdim}(\mathcal{F}) \leq \mathrm{Pdim}(\mathcal{H})$.
- ℓ **unimodal** in $\hat{y}$ for each y (e.g., squared $(\hat{y} - y)^2$, absolute $|\hat{y} - y|$): $\mathrm{Pdim}(\mathcal{F}) \leq 2\,\mathrm{Pdim}(\mathcal{H})$.

Pseudo-dim of $\mathcal{H}$ propagates to $\mathcal{F}$ up to a factor of 1 or 2.

Real-valued generalization bound

Plug Pollard's covering bound $\log \mathcal{N}_\infty(\mathcal{F}, \alpha, n) \le D \log\left(en\frac{a}{D\alpha}\right)$ into the covering $\to$ generalization bound and optimize α:

ERM bound via pseudo-dimension

If $0 \le \ell \le a$ and $\mathrm{Pdim}(\mathcal{F}) \le D$, then with probability $\ge 1-\delta$,

$$L_{\mathcal{D}}\left(\hat{h}\right) \le \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + ca\sqrt{\frac{D\log\left(\frac{n}{D}\right) + \log\left(\frac{1}{\delta}\right)}{n}}.$$

Real-valued analog of the VC bound, with $D = \mathrm{Pdim}(\mathcal{F})$ in place of $\mathrm{VCdim}(\mathcal{H})$.

Covering route recap

The pipeline:

Bounded loss

$0 \le \ell(h, z) \le a$

$\rightarrow$

Pseudo-dim

$\mathrm{Pdim}(\mathcal{F}) \le D$ bounds $\mathcal{N}_\infty$.

$\rightarrow$

Uniform conv.

Covering + concentration + **symmetrization** $\rightarrow$ ERM works.

- Pseudo-dim applies to the loss class $\mathcal{F} = \{(x, y) \to \ell(h(x), y) : h \in \mathcal{H}\}$, not directly to $\mathcal{H}$.

Counting gives a parametric bound. Parts 3–4 give a scale-sensitive alternative.

Rademacher Complexity

Why a second route?

Counting via Pdim gives a parametric bound: gen. error $\leq O\left(a\sqrt{\frac{D}{n}}\right)$, $D = \text{Pdim}(\mathcal{F})$.

Where it fails. Consider linear predictors in $\mathbb{R}^d$ with bounded norm:

$$\mathcal{H}_B = \{x \to \langle w, x\rangle : \|w\|_2 \leq B\}.$$

- $\text{Pdim}(\mathcal{H}_B) = \Theta(d)$ regardless of B — counting bound scales with the ambient dimension d.
- When $d \gg n$ (e.g., overparameterized models, kernels), $\sqrt{d/n}$ is vacuous.
- But intuitively, the norm constraint $\|w\|_2 \leq B$ should make the class simpler when B is small — predictions live in $[-BR, BR]$ if $\|x\|_2 \leq R$.

Counting fails when **parameter count $\gg$ true complexity**.

We need a complexity measure that depends on the actual constraint.

Symmetrization: deriving Rademacher complexity

Plan. Run a symmetrization proof on $\mathbb{E}_S \sup_h (L_{\mathcal{D}}(h) - L_S(h))$ and let the new complexity measure **emerge as the right-hand side**.

Step 1 (ghost sample). Let $S' \sim \mathcal{D}^n$ be independent of S. Since $L_{\mathcal{D}}(h) = \mathbb{E}_{S'} L_{S'}(h)$, Jensen gives:

$$\mathbb{E}_S \sup_h (L_{\mathcal{D}}(h) - L_S(h)) \le \mathbb{E}_{S,S'} \sup_h \frac{1}{n} \sum_i (\ell(h, z_i') - \ell(h, z_i)).$$

Step 2 (symmetrize). z_i, z_i' exchangeable, so multiplying by $\xi_i \in \{-1, +1\}$ leaves the expectation:

$$= \mathbb{E}_{S,S',\xi} \sup_h \frac{1}{n} \sum_i \xi_i (\ell(h, z_i') - \ell(h, z_i)).$$

Step 3 (split). $\sup(a - b) \le \sup a + \sup(-b)$, both halves equal by symmetry:

$$\le 2\mathbb{E}_{S,\xi} \sup_h \frac{1}{n} \sum_i \xi_i \ell(h, z_i).$$

The RHS is a **correlation of $\ell(h, \cdot)$ with random signs**: this is **Rademacher complexity**.

Empirical Rademacher complexity

Definition. Fix $S = (z_1, \ldots, z_n)$; draw $\xi_1, \ldots, \xi_n$ i.i.d. uniform on $\{-1, +1\}$. Assume $|f(z)| \le a$ for all $f \in \mathcal{F}, z \in \mathcal{Z}$ (e.g., a is the loss range).

$$\mathcal{R}_S(\mathcal{F}) := \mathbb{E}_\xi \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(z_i).$$

Game intuition. The signs ξ propose random labels; the player picks $f \in \mathcal{F}$ to maximize correlation:

- rich $\mathcal{F}$ (fits any labeling): $\mathcal{R}_S(\mathcal{F}) \approx a$ (the maximum possible);
- constrained $\mathcal{F}$: $\mathcal{R}_S(\mathcal{F})$ is small.

Immediate consequence of the proof. With $f = \ell(h, \cdot)$, applying the chain to both $\mathcal{F}$ and $-\mathcal{F}$:

$$\mathbb{E}_S \sup_h |L_{\mathcal{D}}(h) - L_S(h)| \le 2\mathbb{E}_S[\mathcal{R}_S(\mathcal{F})] =: 2\mathcal{R}_{\mathcal{D}^n}(\mathcal{F}).$$

$\mathcal{R}_{\mathcal{D}^n}(\mathcal{F})$ **bounds expected uniform convergence.**

From expectation to high probability

So far we have an **expected** bound. Turn it into a high-probability statement by concentration

- Apply McDiarmid's bounded-differences inequality to $g(S) = \sup_h (L_{\mathcal{D}}(h) - L_S(h))$

For $0 \le \ell \le a$, with probability $\ge 1 - \delta$,

$$\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_S(h)) \le 2\mathcal{R}_{\mathcal{D}^n}(\mathcal{F}) + a\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

Applying to $-\mathcal{F}$ for the reverse direction and plugging into the ERM chain:

$$L_{\mathcal{D}}\left(\hat{h}\right) \le \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 4\mathcal{R}_{\mathcal{D}^n}(\mathcal{F}) + O\left(a\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}\right).$$

Symmetrization (expected) + McDiarmid (concentration) → high-prob ERM bound.

Next: how to bound $\mathcal{R}_{\mathcal{D}^n}(\mathcal{F})$?

Bounding $\mathcal{R}$, case 1: finite classes (Massart)

For a finite class $\mathcal{F}$ with $-a \le f(z) \le a$ for all f, z, **Massart's finite-class lemma** gives

$$\forall S, \mathcal{R}_S(\mathcal{F}) \le a\sqrt{\frac{2\log|\mathcal{F}|}{n}} \quad \Rightarrow \quad \mathcal{R}_{\mathcal{D}^n}(\mathcal{F}) \le a\sqrt{\frac{2\log|\mathcal{F}|}{n}}.$$

Plugging into the ERM bound: with probability $\ge 1-\delta$,

$$L_{\mathcal{D}}\left(\hat{h}\right) \le \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + O\left(a\sqrt{\frac{\log|\mathcal{F}|}{n}} + a\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}\right).$$

Note $\log|\mathcal{F}| \le \log|\mathcal{H}|$: recovers the **finite-class uniform convergence** from Week 3 as a special case.

Bounding $\mathcal{R}$, case 2: via covering numbers

Second way to upper-bound $\mathcal{R}$, useful for **infinite** classes whose behavior on S has a finite cover.

One scale (Pollard)

$$\mathcal{R}_S(\mathcal{F}) \le \alpha + a\sqrt{\frac{\log \mathcal{N}_1(\mathcal{F}, \alpha, S)}{n}}.$$

Pick one resolution α; pay α for approximation and $\sqrt{\log \frac{\mathcal{N}_1}{n}}$ for the cover size.

All scales (Dudley)

$$\mathcal{R}_S(\mathcal{F}) \le \frac{C}{\sqrt{n}} \int_0^a \sqrt{\log \mathcal{N}_2(\mathcal{F}, \alpha, S)}\, \mathrm{d}\alpha.$$

Integrate over scales; useful when $\log \mathcal{N}_2$ grows polynomially in $\frac{1}{\alpha}$.

Take $\sup_S$ on the covering number to bound $\mathcal{R}_{\mathcal{D}^n}(\mathcal{F})$.

Covering numbers (Part 2's tool) feed directly into Rademacher bounds.

Comparing the two routes

We now have two complexity measures bounding generalization. How do they compare?

Counting route

Pdim $\to$ covering $\mathcal{N} \to$ uniform conv.

Random-sign route

$\mathcal{R} \to$ uniform conv.

Covers $\to$ Rademacher (previous slide). So any bound from counting yields a Rademacher bound.

Rademacher is strictly more refined. For ℓ_2-norm-constrained linear predictors in $\mathbb{R}^d$:

- $\mathrm{Pdim}(\mathcal{H}_B) = \Theta(d)$ — useless when d is large;
- $\mathcal{R}_{\mathcal{D}^n}(\mathcal{H}_B) \le BR/\sqrt{n}$ — dimension free (Part 4).

Counting needs small Pdim; Rademacher only needs a **structural constraint** (e.g., norm bound).

Next: develop Rademacher bounds for norm-constrained classes.

Scale-Sensitive Complexity

Why parameter count misses something

Motivating example: norm-constrained linear predictors $\mathcal{H}_B = \{x \to \langle w, x\rangle : w \in \mathbb{R}^d\}$.

The problem.

- $\mathrm{Pdim}(\mathcal{H}_B) \geq d$ **regardless of** B (next slide).
- Part 2′s bound is at least $\sqrt{\frac{d}{n}}$, which is loose when $d \gg n$.

Goal of this part. Build complexity measures that depend on:

- the norm bound B: $\|w\|_2 \leq B$,
- the data scale R: $\|x\|_2 \leq R$,
- the loss's Lipschitz constant L: $|\ell(a) - \ell(b)| \leq L|a - b|$.

Together: a d-**free** bound.

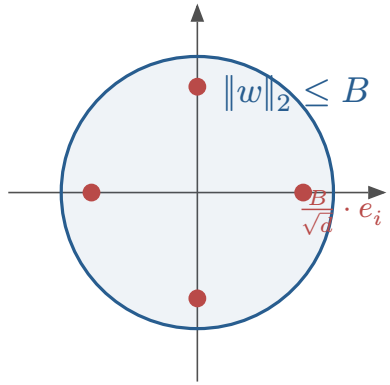
Norm constraints still have large pseudo-dimension

Claim. For any $B > 0$ and any d, $\mathcal{H}_B$ shatters the standard basis $e_1, \ldots, e_d$ at thresholds $\theta_i = 0$.

Recall: shattering at θ_i means every sign pattern of $\operatorname{sign}(f(z_i) - \theta_i)$ is realizable as f varies in $\mathcal{F}$.

Proof. Given $y_i \in \{-1, +1\}$, set $w = \left(\frac{B}{\sqrt{d}}\right)(y_1, \ldots, y_d)$. Then $\|w\|_2 = B$ and

$$\langle w, e_i \rangle = B\frac{y_i}{\sqrt{d}}, \quad \operatorname{sign}(\langle w, e_i \rangle) = y_i.$$



- $\operatorname{Pdim}(\mathcal{H}_B) \geq d$, regardless of B.
- Norm constraint changes the **magnitudes** $\frac{B}{\sqrt{d}}$, not the sign patterns.

- $\operatorname{Pdim}(\mathcal{H}_B)$: independent of B.
- **Margin** $|f(e_i) - \theta_i| = \frac{B}{\sqrt{d}}$: depends on B.
- Need a complexity measure that depends on this **margin**.

Fat-shattering dimension

Idea. For linear predictors $f(z) = \langle w, z \rangle$ with $\|w\|_2 \leq B$ and $\|z\|_2 \leq R$, Cauchy-Schwarz gives $|f(z)| \leq BR$, so predictions live in $[-BR, BR]$.

- At resolution α, this interval has only $O(BR/\alpha)$ distinguishable values per point.
- Pdim counts at resolution 0: every sign pattern, even with vanishing margin.
- Counting at resolution α instead gives a finite count, depending on B, R, α, not d.

Definition. $\mathcal{F}$ **α-shatters** $z_1, \ldots, z_k$ if there exist thresholds $\theta_1, \ldots, \theta_k$ such that every binary pattern $y \in \{-1, +1\}^k$ is realized with margin α:

$$y_i = +1 \quad \Rightarrow \quad f(z_i) \geq \theta_i + \alpha,$$

$$y_i = -1 \quad \Rightarrow \quad f(z_i) \leq \theta_i - \alpha.$$

The **fat-shattering dimension** $\mathrm{fat}_\alpha(\mathcal{F})$ is the largest such k.

- $\alpha = 0$: recovers pseudo-dimension.
- $\alpha > 0$: only patterns realizable with **margin α** count; larger $\alpha \Rightarrow$ smaller $\mathrm{fat}_\alpha(\mathcal{F})$.

Fat-shattering = pseudo-dimension at positive resolution α.

Fat-shattering for ℓ_2 linear predictors

Let $\mathcal{X}_R = \{x \in \mathbb{R}^d : \|x\|_2 \le R\}$ and $\mathcal{H}_B = \{x \to \langle w, x\rangle : \|w\|_2 \le B\}$.

Fat-shattering bound

For every $\alpha > 0$, if the input domain is $\mathcal{X}_R$, then $\mathrm{fat}_\alpha(\mathcal{H}_B) \le \left(B\frac{R}{\alpha}\right)^2$.

Two readings.

- **Dimension-free.** No d appears — bound is independent of the ambient dimension.
- **Scale-sensitive.** Halving α quadruples the bound; smaller gaps allow more patterns.

Contrast with pseudo-dim, which gave d: shrinking B helps only at **positive** resolution α.

Norm + margin replaces d in the complexity bound: $\left(B\frac{R}{\alpha}\right)^2$.

Direct Rademacher bound for ℓ_2 linear predictors

Fat-shattering gave a scale-sensitive complexity. To plug into Part 3's generalization bound, we need $\mathcal{R}_{\mathcal{D}^n}(\mathcal{H}_B)$. Compute the empirical version directly:

$$\begin{aligned}
\mathcal{R}_S(\mathcal{H}_B) &= \mathbb{E}_\xi \sup_{\|w\|_2 \le B} \frac{1}{n} \langle w, \sum_i \xi_i x_i \rangle && \text{(definition)} \\
&= \frac{B}{n} \mathbb{E}_\xi \left\| \sum_i \xi_i x_i \right\|_2 && \text{(Cauchy–Schwarz: } \sup_{\|w\|_2 \le B} \langle w, v \rangle = B\|v\|_2 \text{)} \\
&\le \frac{B}{n} \sqrt{\mathbb{E}_\xi \left\| \sum_i \xi_i x_i \right\|_2^2} && \text{(Jensen, } \sqrt{\cdot} \text{ concave)} \\
&= \frac{B}{n} \sqrt{\sum_i \|x_i\|_2^2} && \text{(cross terms vanish)} \\
&\le \frac{B}{n} \sqrt{nR^2} = \frac{BR}{\sqrt{n}} && (\ \|x_i\|_2 \le R\ ).
\end{aligned}$$

$\mathcal{R}_S(\mathcal{H}_B) \le \boldsymbol{BR/\sqrt{n}}$**, dimension-free.**

From hypotheses to losses: contraction

$\mathcal{R}_S(\mathcal{H}_B)$ controls the **prediction class**, but ERM operates on the **loss class**. Bridge via contraction.

Recall: $\varphi : \mathbb{R} \to \mathbb{R}$ is **G-Lipschitz** if $|\varphi(a) - \varphi(b)| \leq G|a - b|$ for all a, b.

Contraction (Talagrand)

If $\varphi_i : \mathbb{R} \to \mathbb{R}$ is G-Lipschitz for each i, then for any class $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ and sample $S = (x_1, ..., x_n)$,

$$\mathbb{E}_\xi \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_i \xi_i \varphi_i(h(x_i)) \leq \boldsymbol{G} \cdot \mathbb{E}_\xi \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_i \xi_i h(x_i).$$

Applied with $\varphi_i(\hat{y}) = \ell(\hat{y}, y_i)$ (Lipschitz in the prediction):

$$\mathcal{R}_S(\{(x, y) \to \ell(h(x), y) : h \in \mathcal{H}\}) \leq \boldsymbol{G} \cdot \mathcal{R}_S(\mathcal{H}).$$

Examples with $G = 1$: hinge, logistic, absolute. Squared loss is **$2a$-Lipschitz** on $[-a, a]$.

Lipschitz losses preserve Rademacher control, up to a factor of the Lipschitz constant G.

Norm-constrained generalization bound

For $\mathcal{H}_B = \{x \to \langle w, x\rangle : \|w\|_2 \le B\}$, data with $\|x\|_2 \le R$, and a G-Lipschitz loss bounded in $[0, a]$, with probability $\ge 1-\delta$, $\hat{w} = \mathsf{ERM}_{\mathcal{H}_B}(S)$ satisfies

$$L_{\mathcal{D}}(\hat{w}) \le \inf_{\|w\|_2 \le B} L_{\mathcal{D}}(w) + O\left(\frac{GBR}{\sqrt{n}} + a\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}\right).$$

Two terms: **Rademacher** $\frac{GBR}{\sqrt{n}}$ (Talagrand + previous slide), and **confidence** $a\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}$ (concentration).

Sample complexity: $n = O\left(\frac{G^2B^2R^2 + a^2\log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)$, **no** d.

Guarantee depends on G, B and R, not on the ambient dimension.

Margins, Surrogates, and Kernels

Margin normalization

- Part 4 bound: for **norm-constrained** linear classes.
- Binary classification provides a natural norm via the **margin**.

Margin assumption. Data is separable with margin γ by a unit direction:

$$\exists w_* : y\langle w_*, x\rangle \geq \gamma \quad \text{with} \quad \|w_*\|_2 = 1, \quad \forall (x, y) \sim \mathcal{D}.$$

Rescaling. Set $\tilde{w} = \frac{w_*}{\gamma}$. Then $y\langle \tilde{w}, x\rangle \geq 1$ with $\|\tilde{w}\|_2 = \frac{1}{\gamma}$.

- Margin assumption $\Leftrightarrow$ norm bound $B = \frac{1}{\gamma}$ at unit-margin threshold.
- Part 4 gives a **scale-sensitive** complexity $BR/\sqrt{n}$.
- But 0/1 loss depends only on $\text{sign}(\langle w, x\rangle)$: it is **scale-insensitive**. Mismatch.
- Next: a scale-sensitive surrogate for 0/1 loss.

> Margin $\gamma \Leftrightarrow$ norm bound $B = \frac{1}{\gamma}$ with unit-margin threshold.

Scale-sensitive surrogates

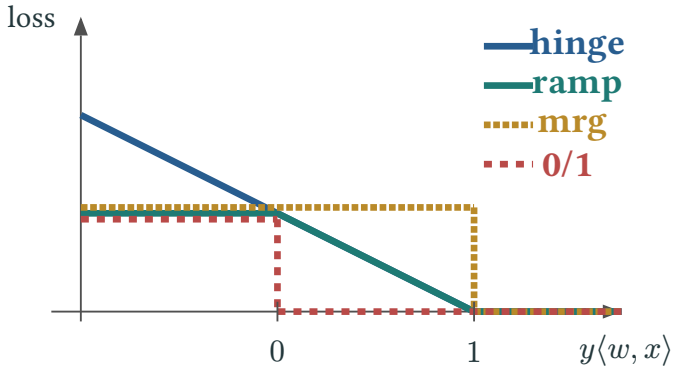
In the signed score $z = y\hat{y}$:

$$\text{loss}_{0/1}(z) = \mathbf{1}[z \le 0]$$

$$\text{loss}_{\text{mrg}}(z) = \mathbf{1}[z < 1]$$

$$\text{loss}_{\text{ramp}}(z) = \min(1, (1 - z)_+)$$

$$\text{loss}_{\text{hinge}}(z) = (1 - z)_+$$



- $\text{loss}_{0/1}$: target; **scale-insensitive**.
- loss_{mrg}: **scale-sensitive analog** of $\text{loss}_{0/1}$.
- $\text{loss}_{\text{ramp}}$: **Lipschitz smoothing** of loss_{mrg}.
- $\text{loss}_{\text{hinge}}$: convex relaxation (Part 6).

Ordering. $\text{loss}_{0/1} \le \text{loss}_{\text{ramp}} \le \text{loss}_{\text{mrg}}$; also $\text{loss}_{\text{ramp}} \le \text{loss}_{\text{hinge}}$.

mrg: scale-sensitive proxy for $\text{loss}_{0/1}$. ramp: Lipschitz smoothing for Talagrand.

Margin-based generalization

Assume $\|x\|_2 \le R$ a.s., and let $\hat{w} = \mathsf{ERM}_B^{\mathrm{mrg}}(S)$ minimize empirical margin loss over $\|w\|_2 \le B$.

Margin generalization bound

With probability $\ge 1 - \delta$,

$$L_{\mathcal{D}}^{0/1}(\hat{w}) \le \inf_{\|w\|_2 \le B} L_{\mathcal{D}}^{\mathrm{mrg}}(w) + O\left(\sqrt{\frac{B^2R^2 + \log\left(\frac{1}{\delta}\right)}{n}}\right).$$

Separable case. If $\mathcal{D}$ is separable with margin γ and $\|x\|_2 \le R$, take $B = \frac{1}{\gamma}$; then $\inf L_{\mathcal{D}}^{\mathrm{mrg}} = 0$ and

$$n = O\left(\frac{R^2}{\gamma^2\epsilon^2}\right) \quad \text{samples suffice for excess risk } \epsilon.$$

After normalization, margin bounds = norm bounds with $B = \frac{1}{\gamma}$.

Why the margin bound follows

Let $\hat{w} = \mathsf{ERM}_B^{\mathrm{mrg}}(S)$ and fix a comparator w^* with $\|w^*\|_2 \le B$. Let $\Delta_B = O\left(B\frac{R}{\sqrt{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{n}}\right)$ be the uniform-convergence deviation.

$$
\begin{aligned}
L_{\mathcal{D}}^{0/1}(\hat{w}) &\le L_{\mathcal{D}}^{\mathrm{ramp}}(\hat{w}) && (\mathrm{ramp} \ge 0/1 \ \ \text{pointwise}) \\
&\le L_S^{\mathrm{ramp}}(\hat{w}) + \Delta_B && (\text{uniform conv. on Lipschitz ramp}) \\
&\le L_S^{\mathrm{mrg}}(\hat{w}) + \Delta_B && (\mathrm{mrg} \ge \mathrm{ramp} \ \ \text{pointwise}) \\
&\le L_S^{\mathrm{mrg}}(w^*) + \Delta_B && (\hat{w} \ \ \text{is ERM on} \ \ \mathrm{mrg}) \\
&\le L_{\mathcal{D}}^{\mathrm{mrg}}(w^*) + 2\Delta_B. && (\text{Hoeffding for fixed} \ \ w^*)
\end{aligned}
$$

Why the ramp detour. ERM acts on $\mathrm{loss}_{\mathrm{mrg}}$ (not Lipschitz), but uniform convergence via Talagrand needs a Lipschitz function. Ramp sits between: $\mathrm{loss}_{0/1} \le \mathrm{loss}_{\mathrm{ramp}} \le \mathrm{loss}_{\mathrm{mrg}}$, with ramp 1-Lipschitz.

Ramp is the Lipschitz stepping stone between $\mathrm{loss}_{0/1}$ and $\mathrm{loss}_{\mathrm{mrg}}$ in the chain.

Why hinge loss matters

Ramp loss is Lipschitz and upper-bounds $0/1$, but it is **nonconvex**. Hinge loss fixes this:

- **upper bound** on $0/1$: $\ell_{\text{hinge}}(z) \geq \mathbf{1}[z \leq 0]$;
- **1-Lipschitz** in z; **convex** in $\hat{y}$; compatible with norm regularization.

Convexity makes hinge ERM **tractable**.

Hinge ERM bound

If $\|x\|_2 \leq R$ a.s. and $\hat{w} = \mathsf{ERM}_B^{\text{hinge}}(S)$, then with probability $\geq 1 - \delta$,

$$L_{\mathcal{D}}^{0/1}(\hat{w}) \leq \inf_{\|w\|_2 \leq B} L_{\mathcal{D}}^{\text{hinge}}(w) + O\left(BR\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}} \right).$$

Hinge ERM competes with the best **hinge-risk** comparator, not the best $0/1$ or margin-risk one.

Hard margin and soft margin

Hinge ERM with norm bound $\|w\|_2 \leq B$ is the **Support Vector Machine (SVM)**, the classical max-margin linear classifier. Two equivalent forms:

Hard margin (separable case):

$$\min_w \|w\|_2 \quad \text{subject to} \quad y_i \langle w, x_i \rangle \geq 1 \quad \forall i.$$

Points with $y_i \langle w^*, x_i \rangle = 1$ pin down w^*: the **support vectors** (origin of "SVM").

Soft margin (general):

$$\min_w L_S^{\text{hinge}}(w) + \lambda \|w\|_2^2.$$

Trades off **fit** (low empirical hinge loss) vs **complexity** (small $\|w\|_2$).

Hard margin is the $\lambda \to 0^+$ limit (under feasibility).

Restrict norm via constraint (hard) or regularizer (soft).

SVM as ℓ_2-controlled linear prediction

- SVM uses only **inner products $\langle w, x\rangle$ and norm $\|w\|_2$**.
- Replace x with $\varphi(x)$ in a Hilbert space $\mathbb{H}_K$ (possibly infinite-dim).
- SVM machinery transfers wholesale.

Two issues when φ is high-dim:

Statistical.

- Infinite-dim $\to$ infinite VCdim.
- ℓ_2 + Lipschitz $\to$ **dim-free** Rademacher rate $B\sqrt{\mathbb{E}\|\varphi(X)\|_2^2/n}$.

Computational/memory.

- Cannot store/optimize an infinite-dim w.
- **Representer**: $w = \sum_i \alpha_i \varphi(x_i)$, finite $\alpha \in \mathbb{R}^n$ suffices.
- **Kernel trick**: only $K(x, x') = \langle \varphi(x), \varphi(x')\rangle$ appears; φ never instantiated.

Two solutions: dim-free bound (statistics) + representer/kernel (computation).

Kernels: definitions and examples

Notation. $\mathbb{H}_K$: Hilbert space **generated by bumps** $K(\cdot, x)$, with inner product set by $\langle K(\cdot, x), K(\cdot, x')\rangle_{\mathbb{H}_K} := K(x, x')$ and **RKHS norm** $\|\cdot\|_K$. (SVM "weight" $w = h$ as vector.)

- **Feature**: $\varphi(x) := K(\cdot, x) \in \mathbb{H}_K$ (the function $y \to K(y, x)$).
- **Kernel**: $K(x, x') = \langle \varphi(x), \varphi(x')\rangle_{\mathbb{H}_K}$.
- **Reproducing property**: $h(x) = \langle h, \varphi(x)\rangle_{\mathbb{H}_K}$ for any $h \in \mathbb{H}_K$ (the kernel **reproduces** function evaluations; the R in RKHS).

Rademacher bound: $\mathcal{R}_{\mathcal{D}^n}(\{h : \|h\|_K \le B\}) \le B\sqrt{\frac{\mathbb{E}K(X,X)}{n}}$

- Linear: $K(x, x') = \langle x, x'\rangle$.
- Polynomial: $K(x, x') = (1 + \langle x, x'\rangle)^d$.
- Gaussian (RBF): $K(x, x') = \exp\left(-\|x - x'\|_2^2/(2\sigma^2)\right)$. $\varphi(x) = K(\cdot, x)$ is a **Gaussian bump** at x. Bumps at distinct centers are **linearly independent** $\to \mathbb{H}_K$ infinite-dim.

Complexity = RKHS norm $\|h\|_K$, not ambient feature dimension.

Representer theorem: why finite memory suffices

Computational obstacle. $w \in \mathbb{H}_K$ may have infinitely many coordinates: how to store or optimize?

Setup. SVM in $\mathbb{H}_K$:

$$\min_{w \in \mathbb{H}_K} L_S^{\text{hinge}}(w) + \lambda \|w\|_K^2.$$

Objective depends on w only through $\langle w, \varphi(x_i) \rangle_{\mathbb{H}_K}$ and $\|w\|_K^2$.

Representer

Optimum has the form $w^* = \sum_{i=1}^n \alpha_i \varphi(x_i)$, with $\alpha \in \mathbb{R}^n$.

Consequence. $h_{w^*}(x) = \sum_i \alpha_i K(x_i, x)$, $\|w^*\|_K^2 = \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j)$.

Support vectors. For hinge loss, $\alpha_i \neq 0$ only for training points on/inside the margin: these are the **support vectors**. w^* is supported on (i.e., a linear combination of) those points — hence "SVM".

Dim-free bound + finite $\alpha \to$ kernel SVM feasible.

ℓ_1 norm: weak-rule ensembles

ℓ_2 story closed: dim-free bound + representer/kernel. Consider a different ℓ_1 geometry:

$$\mathcal{H}_B^1 = \{x \to \langle w, \varphi(x)\rangle : \|w\|_1 \le B\}, \quad \varphi(x) \in \mathbb{R}^d.$$

If $\sup_x \|\varphi(x)\|_\infty \le R_\infty$, Rademacher complexity: $\mathcal{R}_S(\mathcal{H}_B^1) \le BR_\infty\sqrt{\frac{2\log(2d)}{n}}$.

- ℓ_2: $B\frac{R_2}{\sqrt{n}}$ with $\|x\|_2 \le R_2$. **No d.**
- ℓ_1: $BR_\infty\sqrt{\frac{\log(2d)}{n}}$ with $\|x\|_\infty \le R_\infty$. **Logarithmic in d.**

AdaBoost connection. Let $h_1, ..., h_d : \mathcal{X} \to \{-1, +1\}$ be the d weak rules (week 6; d can be huge).

- Feature map: $\varphi(x) = (h_1(x), ..., h_d(x))$. Since each $h_t \in \{-1, +1\}$, $\|\varphi(x)\|_\infty = 1$.
- AdaBoost ensemble = linear in this feature space:

$$f_T(x) = \sum_t \alpha_t h_t(x) = \langle w, \varphi(x)\rangle, \quad w_t = \alpha_t, \quad \|w\|_1 = \sum_t |\alpha_t|.$$

- Bound becomes $\sim \left(\sum_t |\alpha_t|\right)\sqrt{\log d/n}$: **logarithmic** in the number of weak rules.

ℓ_1 restricts to sparse or convex-combination hypotheses; pays $\log d$ instead of d.

Week 7 in one picture

Tool	What it restricts	Where it appears
Loss class $\mathcal{F}$	functions ERM averages	general learning
Pseudo-dimension	thresholded behavior at scale 0	bounded parametric classes
Rademacher	correlation with random labels	uniform convergence
Fat-shattering	thresholded behavior at margin α	scale-sensitive classes
ℓ_2 / RKHS norm	B, R, Lipschitz G	linear / kernel prediction
ℓ_1 norm	$\|w\|_1$, $\|\varphi(x)\|_\infty$	weak-rule ensembles

Every row is a different complexity measure driving the same uniform-convergence story.

Summary

Week 7 summary

Setting	Restricted complexity	Bound shape
General ERM	loss class $\mathcal{F}$	uniform convergence
Counting route	$\mathrm{Pdim}(\mathcal{F}) = D$	$\sqrt{D \log n / n}$
Random-sign route	$\mathcal{R}_S(\mathcal{F})$	sample-adaptive
ℓ_2 predictors	B, R, Lipschitz G	$GBR/\sqrt{n}$
Margins / kernels	$B = 1/\gamma$ and $\|h\|_K$	dimension-free
ℓ_1 predictors	$\|w\|_1$, $\|\varphi(x)\|_\infty$	$\sqrt{\log d / n}$

Generalization comes from the complexity the learning rule restricts — different settings, different names.

Looking ahead

Next steps in the course.

- **Stability.** Generalization through properties of a specific learning rule (rather than uniform convergence).
- **Regularized ERM.** Why adding $\lambda\Psi(w)$ stabilizes learning.
- **Online learning.** Regret bounds: FTL, FTRL, online gradient descent.
- **Implicit bias.** Different optimizers prefer different solutions among empirical minimizers.

> Week 7 finishes the **statistical-complexity** story; the next lectures connect it to **algorithms**.