# Week 8: Stability and Regularized ERM

## Regularization as algorithmic stability

Tianhao Wang
tianhaowang@ucsd.edu

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

# Contents

# Bridge from Week 7

# What Week 7 gave us, and did not

Week 7 measured the **hypothesis class**: if the class is not too rich, then

- $L_S(h)$ is close to $L_{\mathcal{D}}(h)$ for **all** $h$ at once,
- so **any** ERM over that class is statistically reliable.

**Uniform convergence**

Control the gap $\sup_{h \in \mathcal{H}}(L_{\mathcal{D}}(h) - L_S(h))$ by measuring the loss class $\mathcal{F}$.

Tools: pseudo-dimension, covering numbers, Rademacher complexity, fat-shattering.

**Scale-sensitive control**

For real-valued predictors the bound depends on scale, e.g. $\|w\|$ or margin, not on ambient dimension.

The bound is uniform over the class:

- the **same** gap bound holds for every $h \in \mathcal{H}$,
- so it depends on the **class**, not on the **rule** that selects $h$.

**Week 7 analyzed the class; Week 8 analyzes the learning rule.**

# Two routes to generalization

Can a property of the **rule itself** certify generalization?

**Class-level route**

Show $L_{\mathcal{D}}(h) \approx L_S(h)$ for **many hypotheses at once**.

Then any ERM over the class learns.

Tools: VC, Rademacher, fat-shattering.

**Rule-level route**

Show the **returned predictor** barely changes when one training point changes.

Then $L_S(A(S))$ estimates $L_{\mathcal{D}}(A(S))$.

This property of the rule is called **stability**.

**A stable learning rule generalizes, and regularization is what makes a rule stable.**

# Generalization from Stability

# Learning rule and notation

Recall the loss and its population and empirical risk:

$$\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}, \quad L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}} \ell(h, z), \quad L_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, z_i).$$

Here $h \in \mathcal{H}$ is a hypothesis and $z \in \mathcal{Z}$ an example.

This week's object of study is the **learning rule** itself:

$$A : \mathcal{Z}^n \to \mathcal{H}, \quad h_S = A(S).$$

$A$ takes a sample of size $n$ to its returned hypothesis.

**Stability**, defined next, will be a property of this map $A$.

# Replacement stability

A stable rule's output is barely affected when **one training point** is replaced.

Write $S$ and its **one-point replacement** $S'$:

$$S = (z_1, ..., z_i, ..., z_n), \quad S' = (z_1, ..., z_i', ..., z_n),$$

so $S'$ is $S$ with $z_i$ replaced by another point $z_i'$.

**Definition: uniform replacement stability**

$A$ is $\beta(n)$-stable if, for every $S$, every $i \in [n]$, and replacement $z_i'$,

$$|\ell(A(S), z_i) - \ell(A(S'), z_i)| \leq \beta(n).$$

**Small $\beta(n)$: replacing one training point barely moves the loss at that point.**

# Stability implies generalization

**Theorem**

If $A$ is $\beta(n)$-stable, then for every distribution $\mathcal{D}$,

$$\mathbb{E}_{S\sim\mathcal{D}^n} L_{\mathcal{D}}(A(S)) \leq \mathbb{E}_{S\sim\mathcal{D}^n} L_S(A(S)) + \beta(n).$$

Uniform convergence holds with probability $1-\delta$, for every $h \in \mathcal{H}$:

$$\forall h \in \mathcal{H} : \quad L_{\mathcal{D}}(h) \leq L_S(h) + \varepsilon.$$

The stability bound differs on two counts:

- it is **in expectation** over $S$, a weaker form than high-probability;
- it needs **no assumption on the class**, and controls the **returned** $A(S)$.

**Stability gives generalization for the returned predictor, not the whole class.**

# Proof setup: exchangeability

Let

$$S = (z_1, ..., z_n) \sim \mathcal{D}^n$$

and draw an independent copy $z_i' \sim \mathcal{D}$.

Replace coordinate $i$:

$$S' = (z_1, ..., z_i', ..., z_n).$$

Since $z_i$ and $z_i'$ have the same distribution and are independent of the other coordinates,

$$(S', z_i) \text{ has the same distribution as } (S, z)$$

for a fresh test point $z \sim \mathcal{D}$.

**Probability is used only here; the rest of the proof is algebra.**

# Proof: applying stability

For each coordinate $i$:

$$\begin{aligned}
\mathbb{E}_S L_{\mathcal{D}}(A(S)) &= \mathbb{E}_{S,z}\ell(A(S), z) && \text{(definition of } L_{\mathcal{D}}\text{)} \\
&= \mathbb{E}_{S,z_i'}\ell(A(S'), z_i) && \text{(exchangeability)} \\
&\leq \mathbb{E}_S\ell(A(S), z_i) + \beta(n) && \text{(stability)}
\end{aligned}$$

The left side is constant in $i$:

$$\begin{aligned}
\mathbb{E}_S L_{\mathcal{D}}(A(S)) &\leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_S\ell(A(S), z_i) + \beta(n) && \text{(average over } i\text{)} \\
&= \mathbb{E}_S L_S(A(S)) + \beta(n) && \text{(definition of } L_S\text{)}
\end{aligned}$$

**Generalization follows when replacing one point barely changes the returned loss.**

# Stability is not enough by itself

Stability controls the generalization gap; learning needs more.

**Stability gives**

$$\mathbb{E}L_{\mathcal{D}}(A(S)) \leq \mathbb{E}L_S(A(S)) + \beta(n)$$

Empirical loss tracks population loss.

**Still needed**

- the rule must make $L_S(A(S))$ **small**
- a rule ignoring the sample is stable but useless

Plain ERM minimizes the empirical loss by construction, **but is it stable?**

**Learning needs a rule that is stable and drives the empirical loss down.**

# ERM Can Be Unstable

# ERM may jump when one point changes

Consider linear prediction in $\mathbb{R}^2$ with absolute loss:

$$\ell(w, (x, y)) = |\langle w, x \rangle - y|, \quad \|w\|_2 \leq 2.$$

Define $A(S)$ to be an empirical risk minimizer for the sample.

The first $n - 1$ points are

$$z_1, ..., z_{n-1} : x = (1, 0), y = 1,$$

and the final point is either

$$z_n : x = (0, 1), y = 1, \quad z'_n : x = (0, 1), y = -1.$$

Replacing just one point changes the empirical minimizer from $(1, 1)$ to $(1, -1)$.

# Checking the ERM calculation

Recall the two samples, sharing $n-1$ points at $x=(1,0), y=1$:

- $S$: final point $z_n$ at $x=(0,1), y=1$
- $S'$: final point $z'_n$ at $x=(0,1), y=-1$

For the original sample, empirical absolute loss is

$$L_S(w) = \frac{n-1}{n}|w_1 - 1| + \frac{1}{n}|w_2 - 1|.$$

The minimizer in the ball $\|w\|_2 \le 2$ is $A(S) = (1,1)$.

For the replaced sample,

$$L_{S'}(w) = \frac{n-1}{n}|w_1 - 1| + \frac{1}{n}|w_2 + 1|,$$

and the minimizer is $A(S') = (1,-1)$.

Both minimizers fit their own sample, but their losses differ on the same test point.

# Consequence for stability

Recall the minimizers: $A(S) = (1, 1)$ and $A(S') = (1, -1)$.

On the original last point $z_n$ (with $x = (0, 1)$, $y = 1$),

$$\ell(A(S), z_n) = 0, \quad \ell(A(S'), z_n) = 2.$$

Thus the loss can change by a constant even as $n$ grows.

> Plain ERM can be unstable even where **uniform convergence holds**.

**We need a mechanism that makes the returned predictor insensitive to one training example.**

# Regularized ERM

# Add a stabilizing term

In the example, one replacement moved the empirical minimizer from $(1, 1)$ to $(1, -1)$.

Regularized empirical risk minimization adds a **penalty** to resist such jumps:

$$A_\lambda(S) = \mathsf{RERM}_\lambda(S) = \arg\min_{w \in \mathcal{W}} L_S(w) + \lambda \Psi(w).$$

Here $\mathcal{W}$ is the parameter domain, $\lambda > 0$, and $\Psi : \mathcal{W} \to \mathbb{R}$ is the regularizer.

**Regularization changes the learning rule, not just the set of predictors.**

# Two views of regularization

The same low-norm bias can support two different generalization arguments.

| View | What is controlled | Generalization certificate |
|---|---|---|
| Uniform convergence | A whole class, e.g. $\|w\| \leq B$ | $L_S(w) \approx L_{\mathcal{D}}(w)$ for all $w$ in the class |
| Stability | The rule $A_\lambda(S)$ | Replacing one point barely changes the returned loss |

**Week 7 controlled a class; Week 8 controls the learning rule.**

# Warm-up: regularized absolute loss

Return to linear prediction with absolute loss and $\|x\|_2 \leq R$:

$$\ell(w, (x, y)) = |\langle w, x \rangle - y|.$$

Use the regularized rule

$$A_\lambda(S) = \arg\min_{w \in \mathbb{R}^d} L_S(w) + \frac{\lambda}{2}\|w\|_2^2.$$

**Claim for now**

The regularized rule is replacement stable with $\beta(n) \leq \frac{2R^2}{\lambda n}$.

Recall: this means that for every one-point replacement $S'$ of $S$,

$$|\ell(A_\lambda(S), z_i) - \ell(A_\lambda(S'), z_i)| \leq \frac{2R^2}{\lambda n}.$$

**Regularization gives a quantitative stability bound for the returned rule.**

# Warm-up: learning a norm ball

Suppose we want to compete with all $w$ satisfying $\|w\|_2 \le B$. For any such $w$:

$$\begin{aligned}
\mathbb{E}L_{\mathcal{D}}(A_\lambda(S)) &\le \mathbb{E}L_S(A_\lambda(S)) + \frac{2R^2}{\lambda n} && \text{(stability)} \\
&\le \mathbb{E}\left(L_S(A_\lambda(S)) + \frac{\lambda}{2}\|A_\lambda(S)\|_2^2\right) + \frac{2R^2}{\lambda n} && \text{(nonnegative penalty)} \\
&\le \mathbb{E}\left(L_S(w) + \frac{\lambda}{2}\|w\|_2^2\right) + \frac{2R^2}{\lambda n} && \text{(optimality)} \\
&\le L_{\mathcal{D}}(w) + \frac{\lambda}{2}B^2 + \frac{2R^2}{\lambda n} && \text{(fixed comparator)}
\end{aligned}$$

Choose $\lambda = \Theta\left(\frac{R}{B\sqrt{n}}\right)$: $\mathbb{E}L_{\mathcal{D}}(A_\lambda(S)) \le \inf_{\|w\|_2 \le B} L_{\mathcal{D}}(w) + O\left(\frac{BR}{\sqrt{n}}\right)$.

RERM **learns the norm ball through stability, not uniform convergence.**

# Convex Lipschitz bounded problems

The warm-up used three facts about absolute-loss linear prediction:

- $L_S(w)$ is convex in $w$;
- small movement in $w$ gives small change in each loss;
- we compare against predictors of bounded size.

We isolate these facts so the same stability argument applies beyond Euclidean absolute loss.

A convex Lipschitz bounded problem has:

- **Convexity**: $\ell(w, z)$ convex in $w$;
- **Lipschitzness**: $|\ell(w, z) - \ell(w', z)| \leq G\|w - w'\|$;
- **Bounded comparison**: comparator set such as $\|w\| \leq B$ or $\Psi(w) \leq B^2$.

# Linear prediction as a convex Lipschitz problem

For linear prediction, let

$$\ell(w, (x, y)) = \mathrm{loss}(\langle w, \varphi(x)\rangle, y).$$

Assume:

- **Scalar convexity**: $\mathrm{loss}(\hat{y}, y)$ is convex in $\hat{y}$;
- **Scalar Lipschitzness**: $|\mathrm{loss}(a, y) - \mathrm{loss}(b, y)| \le g|a - b|$;
- **Feature scale**: $\|\varphi(x)\|_2 \le R$.

Then

$$\begin{aligned} |\ell(w, z) - \ell(w', z)| &\le g|\langle w - w', \varphi(x)\rangle| \\ &\le gR\|w - w'\|_2. \end{aligned}$$

So the problem is convex and $G = gR$-Lipschitz in $\|\cdot\|_2$.

**Scalar Lipschitzness plus feature scale gives the parameter Lipschitz bound $G = gR$.**

# Euclidean convex Lipschitz theorem

The warm-up extends from absolute loss to any convex $G$-Lipschitz objective.

**Euclidean RERM bound**

Suppose $\ell(w, z)$ is convex and $G$-Lipschitz with respect to $\|\cdot\|_2$.

Consider the regularized empirical risk minimization $A_\lambda(S) = \arg\min_w L_S(w) + \frac{\lambda}{2}\|w\|_2^2$.

Then $A_\lambda$ is replacement stable with $\beta(n) \le \frac{2G^2}{\lambda n}$.

Competing with $\|w\|_2 \le B$ and choosing $\lambda = \Theta\left(\frac{G}{B\sqrt{n}}\right)$ gives

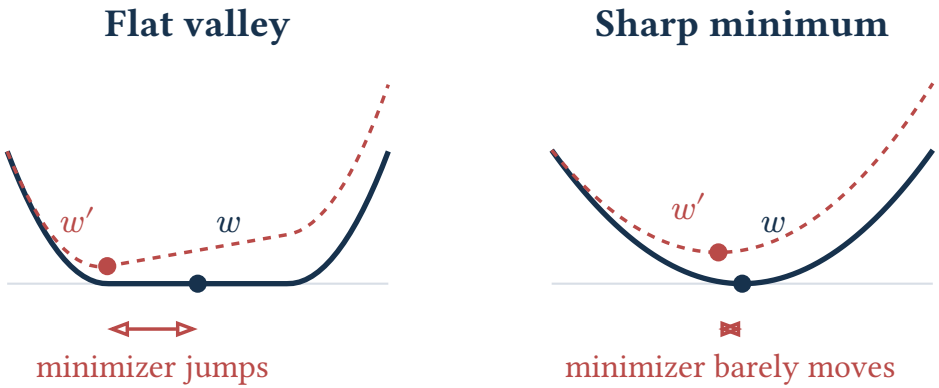
$$\mathbb{E}L_{\mathcal{D}}(A_\lambda(S)) \le \inf_{\|w\|_2 \le B} L_{\mathcal{D}}(w) + O\left(\frac{GB}{\sqrt{n}}\right).$$

**The Euclidean theorem replaces the absolute-loss scale $R$ by the general Lipschitz scale $G$.**

# Strong Convexity and Stability

# Stability needs a sharp minimum

- A rule is **stable** when swapping one example barely moves its minimizer.
- Whether the minimizer moves depends on the **shape of the objective near its minimum**:




Solid: the objective. Dashed: after one example is swapped. $w, w'$: the two minimizers.

**A flat valley lets the minimizer jump; a sharp minimum pins it down.**

# Strong convexity

We name the property that forces a **sharp minimum**:

**Definition: strong convexity**

$\Psi$ is $\alpha$-strongly convex with respect to $\|\cdot\|$ if for every $w, w'$,

$$\Psi(w') \geq \Psi(w) + \langle \nabla\Psi(w), w' - w \rangle + \frac{\alpha}{2}\|w' - w\|^2.$$

- **Convexity** is the same inequality without the last term.
- **Strong convexity** adds the gap $\frac{\alpha}{2}\|w' - w\|^2$.
- $\alpha > 0$ measures that gap; a larger $\alpha$ means a sharper minimum.

# Strong convexity around a minimizer

What does the definition say at the minimum? Let $w_0$ minimize $\Psi$, so $\nabla\Psi(w_0) = 0$.

Setting the base point to $w_0$, the linear term drops out, and for every $w$,

$$\Psi(w) \geq \Psi(w_0) + \frac{\alpha}{2}\|w - w_0\|^2.$$

This is the **sharp minimum** made precise: leaving $w_0$ raises $\Psi$ at a guaranteed rate.

Read in reverse, a small regularizer value forces a small distance from $w_0$:

$$\Psi(w) - \Psi(w_0) \leq B^2 \quad \rightarrow \quad \|w - w_0\| \leq \sqrt{\frac{2}{\alpha}}B.$$

**Strong convexity turns a small regularizer value into a bound on distance from the minimizer.**

# Example: squared Euclidean norm

Does a natural regularizer have this property? Take the one from the Euclidean theorem,

$$\Psi(w) = \frac{1}{2}\|w\|_2^2, \quad \nabla\Psi(w) = w.$$

Write $w' = w + (w' - w)$ and expand the square:

$$\begin{aligned}
\Psi(w') &= \frac{1}{2}\|w + (w' - w)\|_2^2 \\
&= \frac{1}{2}\|w\|_2^2 + \langle w, w' - w\rangle + \frac{1}{2}\|w' - w\|_2^2 \\
&= \Psi(w) + \langle \nabla\Psi(w), w' - w\rangle + \frac{1}{2}\|w' - w\|_2^2.
\end{aligned}$$

This matches the definition exactly, with **equality** and $\alpha = 1$.

So $\Psi$ is 1-strongly convex with respect to $\|\cdot\|_2$.

# Strong convexity: regularizer and objective

If $\Psi$ is $\alpha$-strongly convex:

- $\lambda\Psi$ is $\lambda\alpha$-strongly convex.
- If $f$ is convex, then $f + \lambda\Psi$ is $\lambda\alpha$-strongly convex.

So for any convex loss, $L_S(w) + \lambda\Psi(w)$ is **$\lambda\alpha$-strongly convex**.

The regularizer $\Psi$ is a free choice; strongly convex options for different norms:

| Regularizer | Strongly convex in |
|---|---|
| $\Psi(w) = \frac{1}{2}\|w\|_2^2$ | $\|\cdot\|_2$ |
| $\Psi(w) = \frac{1}{2}\|w\|_p^2 \quad (1 < p \leq 2)$ | $\|\cdot\|_p$ |
| negative entropy: $\Psi(w) = \sum_j w_j \log w_j$ | $\|\cdot\|_1$ |

# RERM stability theorem

The Euclidean theorem extends to any $\alpha$-strongly convex $\Psi$:

**Theorem: stability of RERM**

Assume:

- $\ell(w, z)$ is convex and $G$-Lipschitz with respect to $\|\cdot\|$.
- $\Psi$ is $\alpha$-strongly convex with respect to $\|\cdot\|$.

Let

$$A_\lambda(S) = \arg\min_{w \in \mathcal{W}} L_S(w) + \lambda \Psi(w).$$

Then $A_\lambda$ is replacement stable with

$$\beta(n) \leq \frac{2G^2}{\lambda \alpha n}.$$

At $\Psi = \frac{1}{2}\|w\|_2^2$ and $\alpha = 1$, this recovers the Euclidean bound $\beta(n) \leq 2\frac{G^2}{\lambda n}$.

# Proof setup: compare two minimizers

Let $S$ and $S'$ differ only in coordinate $i$. Set

$$F_S(u) = L_S(u) + \lambda\Psi(u), \quad w = A_\lambda(S), \quad w' = A_\lambda(S').$$

Each $F_S, F_{S'}$ is $\lambda\alpha$-**strongly convex** with $w, w'$ as its minimizer:

$$F_S(w') \geq F_S(w) + \frac{\lambda\alpha}{2}\|w - w'\|^2,$$

$$F_{S'}(w) \geq F_{S'}(w') + \frac{\lambda\alpha}{2}\|w - w'\|^2.$$

They share $\lambda\Psi$, so $F_S - F_{S'}$ depends only on coordinate $i$:

$$F_S(u) - F_{S'}(u) = \frac{1}{n}(\ell(u, z_i) - \ell(u, z_i')).$$

**Both objectives are strongly convex, and their shared $\lambda\Psi$ cancels under subtraction.**

# Proof: from minimizer shift to loss

**Step 1.** Add the two inequalities. The shared $\lambda\Psi$ **cancels**:

$$\lambda\alpha\|w - w'\|^2 \leq \frac{1}{n}(\ell(w', z_i) - \ell(w, z_i) + \ell(w, z_i') - \ell(w', z_i')).$$

**Step 2.** By **Lipschitzness**, each of the two loss differences is at most $G\|w - w'\|$:

$$\lambda\alpha\|w - w'\|^2 \leq \frac{2G}{n}\|w - w'\| \quad \Rightarrow \quad \|w - w'\| \leq \frac{2G}{\lambda\alpha n}.$$

**Step 3.** The loss at any test point shifts by at most $G\|w - w'\|$:

$$|\ell(A(S), z) - \ell(A(S'), z)| \leq \frac{2G^2}{\lambda\alpha n} = \beta(n).$$

**Strong convexity bounds $\|w - w'\|$; Lipschitzness converts this to a loss bound.**

# Stability plus empirical optimality

We now use $\Psi \geq 0$.

**Optimality.** For any comparator $w$, $A_\lambda(S)$ minimizes $L_S + \lambda\Psi$:

$$L_S(A_\lambda(S)) + \lambda\Psi(A_\lambda(S)) \leq L_S(w) + \lambda\Psi(w).$$

Drop $\lambda\Psi(A_\lambda(S)) \geq 0$, then take $\mathbb{E}_S$ (using $\mathbb{E}L_S(w) = L_{\mathcal{D}}(w)$):

$$\mathbb{E}L_S(A_\lambda(S)) \leq L_{\mathcal{D}}(w) + \lambda\Psi(w).$$

**Stability** gives $\mathbb{E}L_{\mathcal{D}}(A_\lambda(S)) \leq \mathbb{E}L_S(A_\lambda(S)) + \beta(n)$, so

$$\mathbb{E}L_{\mathcal{D}}(A_\lambda(S)) \leq L_{\mathcal{D}}(w) + \lambda\Psi(w) + \beta(n).$$

**Optimality + $\Psi \geq 0$ + stability $\rightarrow$ comparator bound for any $w$.**

# RERM learning theorem

Substituting $\beta(n) \le \frac{2G^2}{\lambda\alpha n}$ into the previous bound:

**Theorem: learning from stable RERM**

Assume $\ell$ convex and $G$-Lipschitz, $\Psi$ **$\alpha$-strongly convex and nonnegative**. Then for every comparator $w \in \mathcal{W}$,

$$\mathbb{E}_{S\sim\mathcal{D}^n} L_{\mathcal{D}}(A_\lambda(S)) \le L_{\mathcal{D}}(w) + \lambda\Psi(w) + \frac{2G^2}{\lambda\alpha n}.$$

Optimizing over a comparator set $\mathcal{C} \subseteq \mathcal{W}$ with $\sup_{w\in\mathcal{C}} \Psi(w)$ finite:

$$\mathbb{E} L_{\mathcal{D}}(A_\lambda(S)) \le \inf_{w\in\mathcal{C}} L_{\mathcal{D}}(w) + \lambda \sup_{w\in\mathcal{C}} \Psi(w) + \frac{2G^2}{\lambda\alpha n}.$$

**Two pieces: the comparator's value $L_{\mathcal{D}}(w) + \lambda\Psi(w)$, plus the stability cost.**

# Learning Convex Lipschitz Bounded Problems

# Main learning guarantee

Recall: for a comparator set $\mathcal{C} \subseteq \mathcal{W}$,

$$\mathbb{E} L_{\mathcal{D}}(A_\lambda(S)) \leq \inf_{w \in \mathcal{C}} L_{\mathcal{D}}(w) + \lambda \sup_{w \in \mathcal{C}} \Psi(w) + \frac{2G^2}{\lambda \alpha n}.$$

Let $B^2 := \sup_{w \in \mathcal{C}} \Psi(w)$, the regularizer's range on $\mathcal{C}$.

Balance the two $\lambda$-dependent terms by setting $\lambda B^2 = 2G^2/(\lambda \alpha n)$:

$$\lambda = \sqrt{\frac{2G^2}{\alpha B^2 n}}.$$

Plugging back:

$$\mathbb{E} L_{\mathcal{D}}(A_\lambda(S)) \leq \inf_{w \in \mathcal{C}} L_{\mathcal{D}}(w) + O\left(\frac{GB}{\sqrt{\alpha n}}\right).$$

**Optimal $\lambda$ balances bias and stability, giving the $1/\sqrt{n}$ rate.**

# Choosing the geometry

To learn with stability, choose $\Psi$ so that:

- $\Psi$ is $\alpha$-strongly convex with respect to the same norm used for Lipschitzness.
- $0 \le \Psi(w) \le B^2$ on the comparator set.

Then the stability-based excess term scales as

$$O\left(\frac{GB}{\sqrt{\alpha n}}\right).$$

| Comparator geometry | Regularizer | Effect |
|---|---|---|
| $\|w\|_2 \le B$ | $\frac{1}{2}\|w\|_2^2$ | dimension-free Euclidean scale |
| $\|w\|_p \le B$ | $\frac{1}{2}\|w\|_p^2$ | $\alpha = p-1$, cost $\sim 1/\sqrt{p-1}$ |
| simplex / $\ell_1$ | negative entropy: $\sum_j w_j \log w_j$ | logarithmic dependence on dimension |

**Each choice of $\Psi$ stabilizes the learning rule in a chosen geometry.**

# Data scale: the dual norm

Changing geometry changes $G$, the loss's Lipschitz constant in the chosen norm.

Scalar Lipschitzness, then **Cauchy-Schwarz** in $\ell_2$:

$$\begin{aligned}|\ell(w,z) - \ell(w',z)| &\le g|\langle w - w', \varphi(x)\rangle| \\ &\le g\|w - w'\|_2 \|\varphi(x)\|_2.\end{aligned}$$

For a general norm $\|\cdot\|$, the **second step** becomes Hölder's inequality:

$$\langle u, v\rangle \le \|u\| \cdot \|v\|_*, \quad \|v\|_* := \sup_{\|u\|\le 1} \langle u, v\rangle.$$

A $g$-Lipschitz scalar loss is then **$G$-Lipschitz** in $\|\cdot\|$ with $G = g\|\varphi(x)\|_*$.

Dual pairings: $\ell_2 \leftrightarrow \ell_2$, $\ell_p \leftrightarrow \ell_q$ (conjugate $\frac{1}{p} + \frac{1}{q} = 1$), $\ell_1 \leftrightarrow \ell_\infty$.

**The feature scale entering $G$ is measured in the dual norm.**

# Geometry example: quadratic norm

**Generalize the Euclidean row**: replace $\frac{1}{2}\|w\|_2^2$ by $\frac{1}{2}w^\top Q w$ for positive definite $Q$.

Define the $Q$-norm and regularizer

$$\|w\|_Q = \sqrt{w^\top Q w}, \quad \Psi(w) = \frac{1}{2} w^\top Q w.$$

Then $\Psi$ is **1-strongly convex** with respect to $\|\cdot\|_Q$.

The data scale is measured in the **dual norm**: $\|\varphi(x)\|_{Q,*} = \sqrt{\varphi(x)^\top Q^{-1} \varphi(x)}$.

If $\|w\|_Q \le B$ and $\|\varphi(x)\|_{Q,*} \le R_Q$, then the excess scale is

$$O\left(\frac{B R_Q}{\sqrt{n}}\right).$$

**$Q$ scales $w$ and $\varphi(x)$ anisotropically, and $Q = I$ recovers $\ell_2$.**

# Geometry example: $\ell_p$ norm

**Interpolate between $\ell_2$ and $\ell_1$**: take the comparator in $\ell_p$ for $1 < p \le 2$.

The regularizer $\Psi(w) = \frac{1}{2}\|w\|_p^2$ is **$(p-1)$-strongly convex** in $\|\cdot\|_p$.

**Why $p-1$?** Hölder bounds the diagonal Hessian term (rank-one part $\ge 0$):

$$\langle u, \nabla^2 \Psi(w) u \rangle \ge (p-1)\|u\|_p^2.$$

If $\|w\|_p \le B$ and $\|\varphi(x)\|_q \le R_q$ ($q$ conjugate to $p$), the excess scale is

$$O\left(\frac{BR_q}{\sqrt{(p-1)n}}\right).$$

**Smaller $p$ moves the geometry toward $\ell_1$, at a stability cost of $1/(p-1)$.**

# Geometry example: entropy on the simplex

Comparator: a **distribution over $d$ items** on $\Delta_d$; want $d$ only via $\log d$.

Entropy regularizer $\Psi(w) = \sum_{j=1}^{d} w_j \log(dw_j)$: the factor $d$ gives $0 \le \Psi \le \log d$.

**1-strongly convex** in $\|\cdot\|_1$ (so $\alpha = 1$): the strong-convexity gap is $\mathrm{KL}(w' \,\|\, w)$, and Pinsker gives

$$\mathrm{KL}(w' \,\|\, w) \ge \frac{1}{2}\|w' - w\|_1^2.$$

Dual is $\ell_\infty$: with $\|\varphi(x)\|_\infty \le R$, the excess scale is $O\left(R\sqrt{\frac{\log d}{n}}\right)$.

Over the **same simplex**, $\ell_p$ at $p = 1 + 1/\log d$ has $B = O(1)$ but $\alpha = 1/\log d$.

**Same simplex: the $\sqrt{\log d}$ sits in entropy's range $B^2$, in $\ell_p$'s small $\alpha$.**

# What changed from Week 7?

| Question | Week 7 answer | Week 8 answer |
|---|---|---|
| Why generalization? | Small class complexity | **Stable** learning rule |
| Main object | Loss class $\mathcal{F}$ | Algorithm $A(S)$ |
| Tool | Rademacher / fat-shattering | **Replacement stability** |
| Regularization role | Bounds scale of competitors | Creates **strong convexity** and stability |

**Week 8's new mechanism is strong convexity making the learning rule stable.**

# Examples and Takeaways

# Example: hinge loss classification

Let

$$\ell(w,(x,y)) = \max(0, 1 - y\langle w, x\rangle), \quad y \in \{\pm 1\}, \quad \|x\|_2 \le R.$$

Hinge loss is convex and 1-Lipschitz in the score $\langle w, x\rangle$, hence **$R$-Lipschitz** in $w$ ($G = R$).

The **soft-margin SVM** is $\ell_2$-regularized hinge minimization:

$$A_\lambda(S) = \arg\min_{w\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i\langle w, x_i\rangle) + \frac{\lambda}{2}\|w\|_2^2.$$

By the Euclidean theorem, it satisfies the stability-based guarantee against all $\|w\|_2 \le B$.

**The soft-margin SVM generalizes because its regularizer makes the rule stable.**

# Missing-data example: setup

**Can stability learn a problem where uniform convergence fails?** This example shows it can.

Predictors $w$ live in the unit ball $\|w\|_2 \le 1$ of an **infinite-dimensional** Hilbert space.

Each data point reveals a random subset $I$ of coordinates, all observed values 0:

$$z = (I, z_I), \quad z_j = 0, \quad \ell(w, z) = \sum_{j \in I} \big(w_j - z_j\big)^2.$$

Each coordinate lies in $I$ independently with probability $\frac{1}{2}$, so $L_{\mathcal{D}}(w) = \frac{1}{2}\|w\|_2^2$.

**The population problem is trivial: the minimizer is $w = 0$, with $L_{\mathcal{D}}(0) = 0$.**

# Missing-data example: uniform convergence fails

For a sample of size $n$, each coordinate is unseen with probability $2^{-n} > 0$.

In **infinite dimension**, some coordinate $j$ is then **never observed**.

The spike $e_j$ has zero empirical loss but population loss $\frac{1}{2}$:

$$L_S(e_j) = 0, \quad L_{\mathcal{D}}(e_j) = \frac{1}{2}.$$

So at **every** sample size, uniform convergence cannot hold:

$$\sup_{\|w\|_2 \le 1} (L_{\mathcal{D}}(w) - L_S(w)) \ge \frac{1}{2}.$$

**The class-level route fails, and an arbitrary ERM can pick $e_j$ and never generalize.**

# Missing-data example: RERM still learns

The loss is **convex** (a sum of squares) and **2-Lipschitz** on the unit ball:

$$|\ell(w,z) - \ell(u,z)| = \left|\sum_{j\in I}(w_j - u_j)(w_j + u_j)\right|$$
$$\leq \|w-u\|_2\|w+u\|_2 \leq 2\|w-u\|_2.$$

A convex Lipschitz bounded problem with $G \leq 2$, $B = 1$, so RERM's bound is **dimension-free**:

$$\mathbb{E}L_{\mathcal{D}}(A_\lambda(S)) \leq L_{\mathcal{D}}(0) + O\left(\frac{GB}{\sqrt{n}}\right) = O(1/\sqrt{n}).$$

It holds even in infinite dimension, where RERM picks the **min-norm** zero-loss solution $w = 0$.

**Stability learns where uniform convergence fails: the gain of the Week 8 route.**

# Summary

# Week 8 summary

| Concept | Takeaway |
|---|---|
| Stability | A stable rule is insensitive to replacing one sample |
| Generalization | $\mathbb{E}L_{\mathcal{D}}(A(S)) \leq \mathbb{E}L_S(A(S)) + \beta(n)$ |
| ERM | May be unstable even for convex Lipschitz problems |
| RERM | Adds **strongly convex regularization** |
| Strong convexity | Converts one replacement into small solution movement |
| Convex Lipschitz bounded problems | RERM learns with excess $O\left(G\frac{B}{\sqrt{\alpha n}}\right)$ |

# One-slide mental model

strongly convex regularization

↓

one replacement changes the minimizer little

↓

returned predictor is stable

↓

empirical loss estimates population loss

**Strongly convex regularization makes the returned predictor insensitive to one training example. This stability gives generalization.**