# MIT 9.520/6.7910
# Statistical Learning Theory and Applications

## Class 10: Learning bounds for ERM

Lorenzo Rosasco

# ML algorithms design so far

- ERM+bias/regularization+optimization
- Implicit regularization
- Linear models
- Nonlinear models: neural nets+kernels

# Theory so far

$$\min_{f \in \mathcal{T}} L(f) \quad \mapsto \quad \min_{f \in \mathcal{H}} \widehat{L}(f)$$

- Representation: $\mathcal{H}, \mathcal{B}$ etc.
- Optimization:

$$\min_{f \in \mathcal{H}} \widehat{L}(f).$$

- **Generalization**: beyond linear least squares.

# Outline

# The supervised learning problem

- $\mathcal{X} \times \mathbb{R}$ probability space, with distribution $P$.
- $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ *loss function*.

Problem: Solve

$$L(f_P) = \min_{f \in \mathcal{T}} L(f), \qquad L(f) = \mathbb{E}_{(x,y)\sim P}[\ell(y, f(x))],$$

given

$$S_n = (x_1, y_1), \ldots, (x_n, y_n) \sim P^n.$$

# Regularized ERM

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \widehat{L}(f) + \lambda R(f), \qquad \widehat{L}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)).$$

Note: an estimator is a random variable.

# How good is a solution?

$$S_n \to \widehat{f}$$

- Practice: test error[1]

$$\widehat{L}_{\text{Test}}(\widehat{f}) = \frac{2}{n} \sum_{i=n/2+1}^{n} \ell(y_i, \widehat{f}(x_i)).$$

- Theory: **excess risk**

$$L(\widehat{f}) - L(f_P).$$

Note: these are also random variables.

[1] Use $n/2$ points for $\widehat{f}$.

# Outline

# Loss functions

$$\ell(y, f(x))$$

- Lipschitz for all $y, a, a' \in \mathbb{R}$, $\exists C_\ell$ such that

$$|\ell(y, a) - \ell(y, a')| \leq C_\ell |a - a'|.$$

- Bounded in zero, $\exists C_0$ such that for all $y \in \mathbb{R}$

$$|\ell(y, 0)| \leq C_0.$$

Locally bounded: it follows that for all $y, a \in \mathbb{R}$

$$|\ell(y, a)| \leq C_\ell |a| + C_0.$$

# Regularization and representation

$$R(f)$$

s.t. $R(0) = 0$.

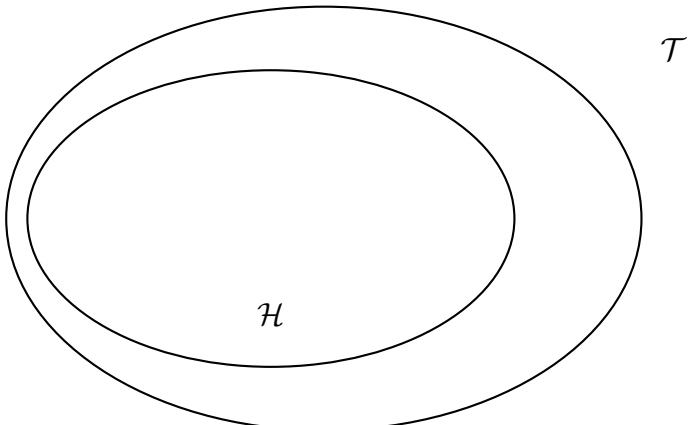
Later we consider

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

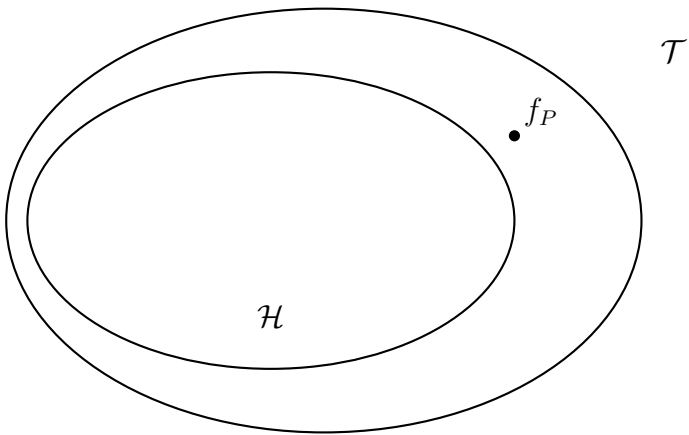
- Linear models $\Phi(x) = x, w \in \mathbb{R}^d$, and $R(f)$ e.g. $\|w\|, \|w\|_1$.
- Hilbert features $\Phi(x), w \in \mathcal{F}$, and $R(f)$ e.g. $\|f\|_{\mathcal{H}}$.
- Banach features $\Phi(x) \in \mathcal{F}^*, w \in \mathcal{F}$, and $R(f)$ e.g. $\|f\|_{\mathcal{B}}$.

# Outline

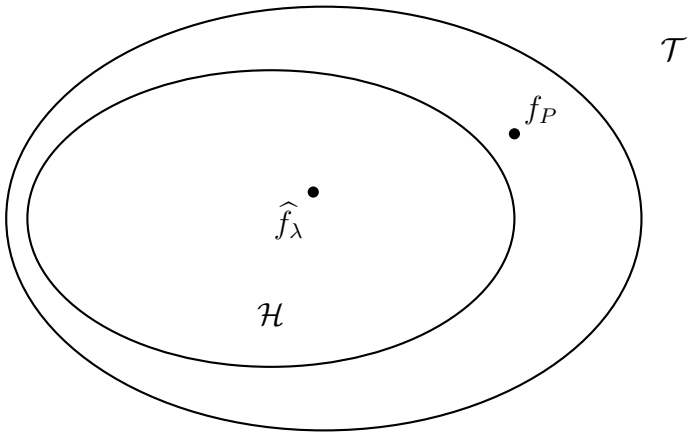
# Error decomposition



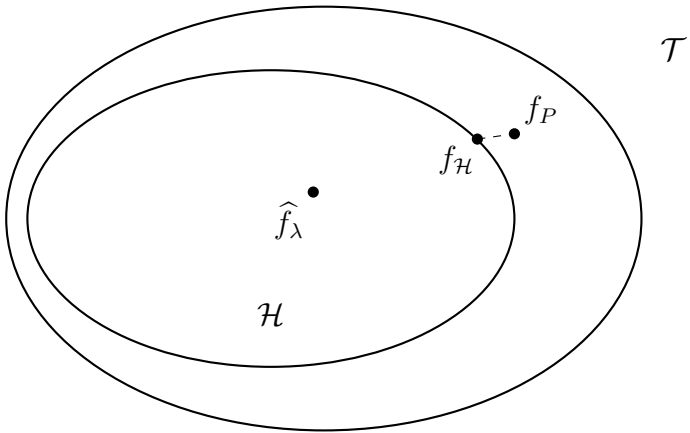
# Error decomposition



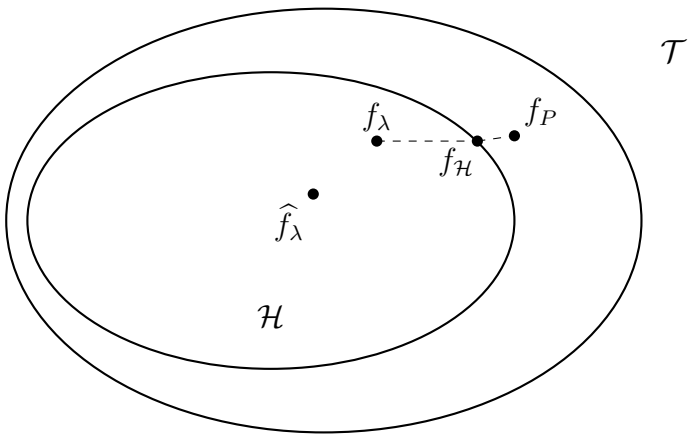
# Error decomposition



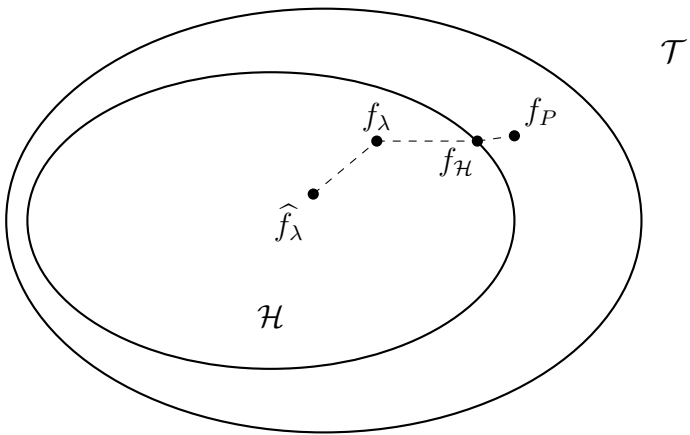
# Error decomposition



# Error decomposition



# Error decomposition



# Error decomposition

$$L(\widehat{f}_\lambda) - L(f_P) = L(\widehat{f}_\lambda) - L(f_\lambda) + L(f_\lambda) - \inf_{f \in \mathcal{H}} L(f) + \inf_{f \in \mathcal{H}} L(f) - L(f_P)$$

where

$$f_\lambda = \arg\min_{f \in \mathcal{H}} L(f) + \lambda R(f), \qquad \inf_{f \in \mathcal{H}} L(f),$$

and

$$\widehat{f}_\lambda = \arg\min_{f \in \mathcal{H}} \widehat{L}(f) + \lambda R(f), \qquad L(f_P) = \min_{f \in \mathcal{T}} L(f).$$

# Irreducible, approximation and estimation errors

$$L(\widehat{f}_\lambda) - L(f_P) = L(\widehat{f}_\lambda) - L(f_\lambda) + L(f_\lambda) - \inf_{f \in \mathcal{H}} L(f) + \inf_{f \in \mathcal{H}} L(f) - L(f_P)$$

- Estimation error (stochastic)

$$L(\widehat{f}_\lambda) - L(f_\lambda).$$

- Approximation error (deterministic)

$$L(f_\lambda) - \inf_{f \in \mathcal{H}} L(f).$$

- Irreducible error (deterministic)

$$\inf_{f \in \mathcal{H}} L(f) - L(f_P)$$

# Outline

# Irreducible error and universality

The error

$$\inf_{f \in \mathcal{H}} L(f) - L(f_P)$$

is irreducible once $\mathcal{H}$ is fixed and the problem is fixed.

For universal representations

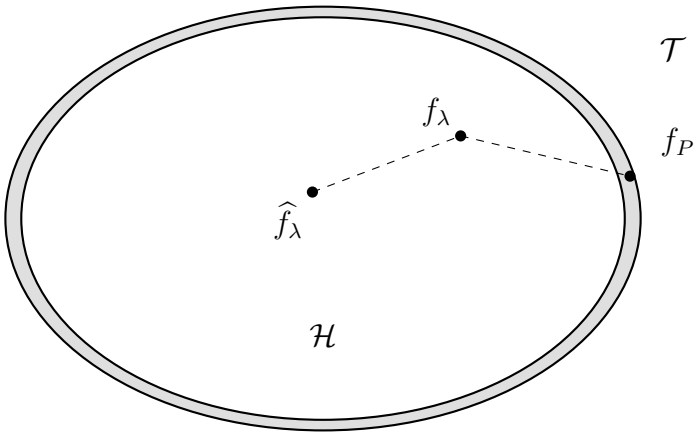
$$\inf_{f \in \mathcal{H}} L(f) = L(f_P)$$

for any $P$.

Note that, in general there is no $f_{\mathcal{H}}$ s.t.

$$\inf_{f \in \mathcal{H}} L(f) = L(f_{\mathcal{H}}).$$

# Universal representations



# Universality

$$\inf_{f \in \mathcal{H}} L(f) - L(f_P) = 0.$$

We can see this in two steps:

First (see next slide):

$$L(f) - L(f_P) \leq C_\ell \, \|f - f_P\|_{L^1(\mathcal{X})} .$$

Second:

$$\inf_{f \in \mathcal{H}} L(f) - L(f_P) \;\leq\; C_\ell \, \inf_{f \in \mathcal{H}} \|f - f_P\|_{L^1(\mathcal{X})} = 0.$$

When $\mathcal{X}$ is compact and $P_{\mathcal{X}}$ has full support, continuous functions are dense in $L^1(\mathcal{X})$, so density in $C(\mathcal{X})$ implies density in $L^1(\mathcal{X})$.

# Risks and norms

$$L(f) - L(f_P) \leq C_\ell \, \|f - f_P\|_{L^1(\mathcal{X})} .$$

Indeed[2]

$$\begin{aligned} L(f) - L(f_P) &= \left| \int \big(\ell(y, f(x)) - \ell(y, f_P(x))\big) \, dP(x,y) \right| \\ &\leq \int |\ell(y, f(x)) - \ell(y, f_P(x))| \, dP(x,y) \\ &\leq C_\ell \int |f(x) - f_P(x)| \, dP_{\mathcal{X}}(x) = C_\ell \, \|f - f_P\|_{L^1(\mathcal{X})} . \end{aligned}$$

[2]Here $L^1(\mathcal{X}) = L^1(\mathcal{X}, P_{\mathcal{X}})$ with norm $\|f\|_{L^1(\mathcal{X})} = \int |f(x)| \, dP_{\mathcal{X}}(x)$.

# Approximation and estimation errors

$$L(\widehat{f}_\lambda) - \inf_{f \in \mathcal{H}} L(f) = L(\widehat{f}_\lambda) - L(f_\lambda) + L(f_\lambda) - \inf_{f \in \mathcal{H}} L(f)$$

We ignore the irreducible error, which is either zero or fixed by $\mathcal{H}$ and the problem.

- Estimation error (stochastic)
$$L(\widehat{f}_\lambda) - L(f_\lambda).$$

- Approximation error (deterministic)
$$L(f_\lambda) - \inf_{f \in \mathcal{H}} L(f).$$

# Outline

# Preliminary estimation error bound

$$\mathbb{E}[L(\widehat{f}_\lambda) - L(f_\lambda)] \leq \mathbb{E}[L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)] + \lambda R(f_\lambda)$$

- Generalization gap (stochastic)
$$L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)$$

- We will add $\lambda R(f_\lambda)$ to the approximation error.

# Preliminary estimation error bound: derivation

$$\mathbb{E}[L(\widehat{f}_\lambda) - L(f_\lambda)] \leq \mathbb{E}[L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)] + \lambda R(f_\lambda)$$

$$L(\widehat{f}_\lambda) - L(f_\lambda) \leq L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda) + \widehat{L}(\widehat{f}_\lambda) + \lambda R(\widehat{f}_\lambda) - \widehat{L}(f_\lambda) - \lambda R(f_\lambda) + \widehat{L}(f_\lambda) + \lambda R(f_\lambda) - L(f_\lambda)$$

- By definition of ERM

$$\widehat{L}(\widehat{f}_\lambda) + \lambda R(\widehat{f}_\lambda) - \widehat{L}(f_\lambda) - \lambda R(f_\lambda) \leq 0.$$

  (This is the optimization error if $\widehat{f}_\lambda$ is not computed exactly.)
- $\mathbb{E}[\widehat{L}(f_\lambda) - L(f_\lambda)] = 0$, since

$$\mathbb{E}\big[\widehat{L}(f_\lambda)\big] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(f_\lambda(x_i), y_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{(x_i,y_i)\sim P}[\ell(f_\lambda(x_i), y_i)] = \underbrace{\mathbb{E}_{(x,y)\sim P}[\ell(f_\lambda(x), y)]}_{L(f_\lambda)}.$$

# Approximation error and generalization gap

$$\mathbb{E}[L(\widehat{f}_\lambda) - \inf_{f\in\mathcal{H}} L(f)] \leq \mathbb{E}[L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)] + L(f_\lambda) - \inf_{f\in\mathcal{H}} L(f) + \lambda R(f_\lambda)$$

- Generalization gap (stochastic)
$$L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)$$

- **Approximation error** (deterministic)
$$L(f_\lambda) - \inf_{f\in\mathcal{H}} L(f) + \lambda R(f_\lambda).$$

# Outline

# Approximation error: convergence

It can be shown that for any $P$

$$\lim_{\lambda \to 0} L(f_\lambda) - \inf_{f \in \mathcal{H}} L(f) + \lambda R(f_\lambda) = 0.$$

No rates with no further assumptions (either on $\mathcal{H}$ or the problem)!

# Approximation error with best in the model

Assume $\exists f_{\mathcal{H}} \in \mathcal{H}$ s.t.

$$L(f_{\mathcal{H}}) = \inf_{f \in \mathcal{H}} L(f).$$

Then

$$L(f_\lambda) - L(f_{\mathcal{H}}) + \lambda R(f_\lambda) \leq \lambda R(f_{\mathcal{H}}).$$

Easy to see since

$$L(f_\lambda) + \lambda R(f_\lambda) \leq L(f_{\mathcal{H}}) + \lambda R(f_{\mathcal{H}}).$$

# General approximation error

In general, we might assume that $\exists C_{P,\mathcal{H}} > 0, \alpha \in (0,1)$ s.t.

$$L(f_\lambda) - \inf_{f \in \mathcal{H}} L(f) + \lambda R(f_\lambda) \;\leq\; C_{P,\mathcal{H}}\; \lambda^\alpha.$$

Characterizations of the above condition depend on the considered representation.

# Approximation error and generalization gap

$$\mathbb{E}[L(\widehat{f}_\lambda) - \inf_{f\in\mathcal{H}} L(f)] \leq \mathbb{E}[L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)] + L(f_\lambda) - \inf_{f\in\mathcal{H}} L(f) + \lambda R(f_\lambda)$$

- **Generalization gap** (stochastic)
$$L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)$$

- Approximation error (deterministic) ✓
$$L(f_\lambda) - \inf_{f\in\mathcal{H}} L(f) + \lambda R(f_\lambda) \leq \lambda R(f_{\mathcal{H}})$$

# Outline

# Generalization gap bound

We will prove that

$$\mathbb{E}[L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)] \leq 2\frac{C_\ell C_0 C_x}{\lambda\sqrt{n}}$$

considering neural nets.

The argument is more general . . . but quite technical.

# Not so easy

$$\mathbb{E}[L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)]$$

Indeed, $\mathbb{E}[L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)] \neq 0$, since

$$\mathbb{E}\big[\widehat{L}(\widehat{f}_\lambda)\big] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(\widehat{f}_\lambda(x_i), y_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big[\ell(\widehat{f}_\lambda(x_i), y_i)\Big] \neq \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{(x_i,y_i)}\Big[\ell(\widehat{f}_\lambda(x_i), y_i)\Big]}_{L(\widehat{f}_\lambda)}.$$

# Bounding strategy

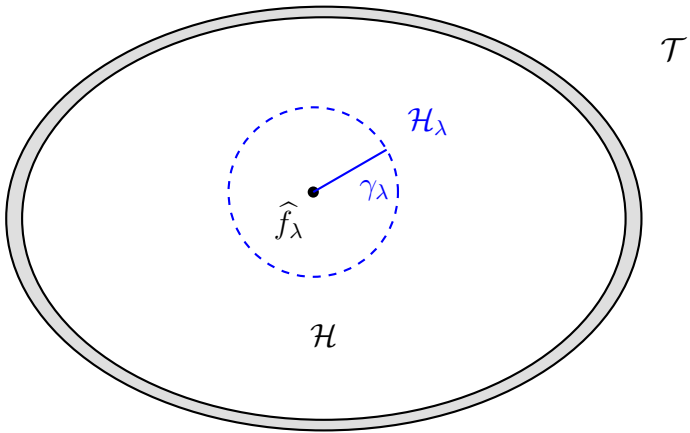
$$\mathbb{E}[L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)]$$

- Show that $R(\widehat{f}_\lambda) \leq \gamma_\lambda$ almost surely, for some $\gamma_\lambda$ depending on $\lambda$.

- Let $\mathcal{H}_\lambda = \{f \in \mathcal{H} \mid R(f) \leq \gamma_\lambda\}$, and consider

$$\mathbb{E}[L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)] \leq \mathbb{E}\left[\sup_{f \in \mathcal{H}_\lambda} \left(L(f) - \widehat{L}(f)\right)\right].$$

# Regularization ball around $\widehat{f}_\lambda$



# Regularization ball around $\widehat{f}_\lambda$

$$R(\widehat{f}_\lambda) \leq \frac{C_0}{\lambda}$$

Indeed,

$$\lambda R(\widehat{f}_\lambda) \leq \widehat{L}(\widehat{f}_\lambda) + \lambda R(\widehat{f}_\lambda) \leq \widehat{L}(0) + \lambda R(0) \leq \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, 0) \leq C_0.$$

# Uniform law of large numbers

$$\sup_{f \in \mathcal{H}_\lambda} \left( L(f) - \widehat{L}(f) \right)$$

The convergence to zero of this term is related to the *uniform law of large numbers*.

It is intimately connected to empirical process theory.

# Empirical processes

Let $Z, Z_1, \ldots, Z_n \sim P$ i.i.d. random variables in $\mathcal{Z}$, and let $\mathcal{F} \subset \{F \mid F : \mathcal{Z} \to \mathbb{R}\}$.

The *empirical process* indexed by $\mathcal{F}$ is

$$\left\{ \mathbb{E}_Z[F(Z)] - \frac{1}{n}\sum_{i=1}^{n} F(Z_i) \ : \ F \in \mathcal{F} \right\}.$$

Its *supremum* is

$$\sup_{F \in \mathcal{F}} \left( \mathbb{E}_Z[F(Z)] - \frac{1}{n}\sum_{i=1}^{n} F(Z_i) \right).$$

# Empirical processes: example

Let $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and define

$$\mathcal{F}_\lambda = \{\, F : F(z) = \ell(f(x), y),\ f \in \mathcal{H}_\lambda \,\}.$$

Then

$$\sup_{f \in \mathcal{H}_\lambda} \big(L(f) - \widehat{L}(f)\big) = \sup_{F \in \mathcal{F}_\lambda} \Big( \mathbb{E}[F(Z)] - \frac{1}{n} \sum_{i=1}^{n} F(Z_i) \Big).$$

# Empirical processes bound

$$\sup_{F \in \mathcal{F}} \left( \mathbb{E}_Z[F(Z)] - \frac{1}{n} \sum_{i=1}^{n} F(Z_i) \right)$$

is a classical quantity.

It can be bounded using:

- Symmetrization.
- Rademacher complexity.

# Symmetrization

Let $S = (Z_1, \dots, Z_n) \sim P^n$ and let $S' = (Z'_1, \dots, Z'_n) \sim P^n$ be an independent copy. Then

$$\mathbb{E}_S \left[ \sup_{F \in \mathcal{F}} \left( \mathbb{E}_Z[F(Z)] - \frac{1}{n} \sum_{i=1}^{n} F(Z_i) \right) \right] \leq \mathbb{E}_{S,S'} \left[ \sup_{F \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left( F(Z'_i) - F(Z_i) \right) \right].$$

This is the *symmetrization inequality*.
It can be derived noting that

$$\mathbb{E}_Z[F(Z)] = \mathbb{E}_{S'} \left[ \frac{1}{n} \sum_{i=1}^{n} F(Z'_i) \right],$$

so that

$$\mathbb{E}_S \left[ \sup_{F \in \mathcal{F}} \left( \mathbb{E}_{S'} \left[ \frac{1}{n} \sum_{i=1}^{n} F(Z'_i) - F(Z_i) \right] \right) \right] \leq \mathbb{E}_{S,S'} \left[ \sup_{F \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left( F(Z'_i) - F(Z_i) \right) \right].$$

(Jensen's inequality for the supremum).

# Symmetrization (cont.)

Introduce independent Rademacher variables $\{\sigma_i\}_{i=1}^n$,

$$\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}.$$

Then

$$\mathbb{E}_{S,S'}\left[\sup_{F\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (F(Z_i') - F(Z_i))\right] = 2\,\mathbb{E}_{S,\sigma}\left[\sup_{F\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i F(Z_i)\right].$$

Indeed,

$$\mathbb{E}_{S,S'} \sup_{F\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (F(Z_i') - F(Z_i)) = \mathbb{E}_{S,S'}\,\mathbb{E}_\sigma\left[\sup_{F\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i\big(F(Z_i') - F(Z_i)\big)\right],$$

and by symmetry of $\{\sigma_i\}$ and the i.i.d. nature of $(Z_i, Z_i')$,

$$\mathbb{E}_{S,S',\sigma}\left[\sup_{F\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i\big(F(Z_i') - F(Z_i)\big)\right] = 2\,\mathbb{E}_{S,\sigma}\left[\sup_{F\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i F(Z_i)\right].$$

# Rademacher complexity

The *empirical Rademacher complexity* of $\mathcal{F}$ is

$$\widehat{\mathcal{R}}(\mathcal{F}) = \mathbb{E}_\sigma \left[ \sup_{F \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i F(Z_i) \right].$$

The *expected Rademacher complexity* of $\mathcal{F}$ is

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_S \left[ \widehat{\mathcal{R}}_n(\mathcal{F}) \right] = \mathbb{E}_{S,\sigma} \left[ \sup_{F \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i F(Z_i) \right].$$

Finally, we obtain the bound

$$\mathbb{E}_S \left[ \sup_{F \in \mathcal{F}} \left( \mathbb{E}_Z[F(Z)] - \frac{1}{n} \sum_{i=1}^{n} F(Z_i) \right) \right] \leq 2\, \mathcal{R}_n(\mathcal{F}).$$

# Expected empirical processes bound

$$\mathbb{E}\left[\sup_{F\in\mathcal{F}}\left(\mathbb{E}_Z[F(Z)]-\frac{1}{n}\sum_{i=1}^{n}F(Z_i)\right)\right]$$

can be bounded using:

- **Symmetrization:**

$$\mathbb{E}_S\left[\sup_{F\in\mathcal{F}}\left(\mathbb{E}_Z[F(Z)]-\frac{1}{n}\sum_{i=1}^{n}F(Z_i)\right)\right]\leq\mathbb{E}_{S,S'}\left[\sup_{F\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\left(F(Z_i')-F(Z_i)\right)\right].$$

- **Rademacher complexity:**

$$\mathbb{E}_{S,S'}\left[\sup_{F\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\left(F(Z_i')-F(Z_i)\right)\right]\leq 2\,\mathcal{R}_n(\mathcal{F}),\quad \mathcal{R}_n(\mathcal{F})=\mathbb{E}_{S,\sigma}\left[\sup_{F\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_iF(Z_i)\right].$$

# Expected empirical processes bound

$$\mathbb{E}\left[\sup_{F\in\mathcal{F}}\left(\mathbb{E}_Z[F(Z)]-\frac{1}{n}\sum_{i=1}^{n}F(Z_i)\right)\right]\leq 2\,\mathcal{R}_n(\mathcal{F})$$

- **Symmetrization:**

$$\mathbb{E}_S\left[\sup_{F\in\mathcal{F}}\left(\mathbb{E}_Z[F(Z)]-\frac{1}{n}\sum_{i=1}^{n}F(Z_i)\right)\right]\leq \mathbb{E}_{S,S'}\left[\sup_{F\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\left(F(Z_i')-F(Z_i)\right)\right].$$

- **Rademacher complexity:**

$$\mathbb{E}_{S,S'}\left[\sup_{F\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\left(F(Z_i')-F(Z_i)\right)\right]\leq 2\,\mathcal{R}_n(\mathcal{F}),\quad \mathcal{R}_n(\mathcal{F})=\mathbb{E}_{S,\sigma}\left[\sup_{F\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i F(Z_i)\right].$$

# Generalization gap and Rademacher complexity

Back to learning:

$$\mathbb{E}[L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)] \leq \mathbb{E}_S \left[ \sup_{f \in \mathcal{H}_\lambda} \left( L(f) - \widehat{L}(f) \right) \right] \leq 2\, \mathcal{R}_n(\mathcal{F}_\lambda).$$

with

$$\mathcal{F}_\lambda = \{\, F : F(z) = \ell(f(x), y),\ f \in \mathcal{H}_\lambda \,\}.$$

# Contraction principle

Consider

$$\mathcal{R}_n(\mathcal{F}_\lambda), \quad \text{with} \quad \mathcal{F}_\lambda = \{\, F : F(z) = \ell(f(x), y),\ f \in \mathcal{H}_\lambda \,\}.$$

Then

$$\mathcal{R}_n(\mathcal{F}_\lambda) \leq C_\ell \mathcal{R}_n(\mathcal{H}_\lambda),$$

where $\mathcal{H}_\lambda = \{\, f \in \mathcal{H} : R(f) \leq \gamma_\lambda \,\}$.

# Rademacher complexity

$$\mathcal{R}_n(\mathcal{H}_\lambda) = \mathbb{E}_{S,\sigma}\left[\sup_{f\in\mathcal{H}_\lambda} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)\right].$$

with

$$\mathcal{H}_\lambda = \{\, f \in \mathcal{H} : R(f) \leq \gamma_\lambda \,\}.$$

Intuitively, it measures the richness of the class $\mathcal{H}_\lambda$ by how well it correlates with random noise.

# Generalization gap and Rademacher complexity (cont.)

Back to learning:

$$\mathbb{E}[L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)] \leq \mathbb{E}_S \left[ \sup_{f \in \mathcal{H}_\lambda} \left( L(f) - \widehat{L}(f) \right) \right] \leq 2\, C_\ell\, \mathcal{R}_n(\mathcal{H}_\lambda).$$

with

$$\mathcal{R}_n(\mathcal{H}_\lambda) = \mathbb{E}_{S,\sigma} \left[ \sup_{f \in \mathcal{H}_\lambda} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right].$$

and

$$\mathcal{H}_\lambda = \{\, f \in \mathcal{H} : R(f) \leq \gamma_\lambda \,\}.$$

# Rademacher complexities

$$\mathcal{R}_n(\mathcal{H}_\lambda) = \mathbb{E}_{S,\sigma}\left[\sup_{f\in\mathcal{H}_\lambda} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)\right], \qquad \mathcal{H}_\lambda = \{\, f \in \mathcal{H} : R(f) \leq \gamma_\lambda \,\}.$$

Examples:

$$f(x) = \langle v, \Phi(x)\rangle\,.$$

- Linear models: $\Phi(x) = x$, $v \in \mathbb{R}^d$, and $R(f)$ e.g. $\|v\|_2$, $\|v\|_1$.
- Hilbert features: $\Phi(x) \in \mathcal{F}$, $v \in \mathcal{F}$, and $R(f)$ e.g. $\|f\|_{\mathcal{H}}$.
- Banach features: $\Phi(x) \in \mathcal{F}^*$, $v \in \mathcal{F}$, and $R(f)$ e.g. $\|f\|_{\mathcal{B}}$.

# Rademacher complexities with Banach features

Consider

$$f(x) = \langle v, \Phi(x) \rangle, \quad v \in \mathcal{F},\ \Phi(x) \in \mathcal{F}^*, \quad \|f\|_{\mathcal{B}} = \inf\{\|v\|_{\mathcal{F}} : f(x) = \langle v, \Phi(x) \rangle\}.$$

Then

$$\mathcal{R}_n(\mathcal{H}_\lambda) \ \leq \ \frac{\gamma_\lambda}{n} \ \mathbb{E}_\sigma \left[ \left\| \sum_{i=1}^n \sigma_i \Phi(x_i) \right\|_{\mathcal{F}^*} \right].$$

Indeed, the following relationship holds between dual norms:

$$\sup_{\|v\|_{\mathcal{F}} \leq 1} |\, \langle v, u \rangle \,| = \|u\|_{\mathcal{F}^*}, \quad u \in \mathcal{F}^*.$$

So that

$$\mathcal{R}_n(\mathcal{H}_\lambda) \leq \mathbb{E}_{S,\sigma} \left[ \sup_{\|v\|_{\mathcal{F}} \leq \gamma_\lambda} \frac{1}{n} \sum_{i=1}^n \sigma_i \langle v, \Phi(x_i) \rangle \right] = \frac{\gamma_\lambda}{n} \ \mathbb{E}_{S,\sigma} \left[ \left\| \sum_{i=1}^n \sigma_i \Phi(x_i) \right\|_{\mathcal{F}^*} \right].$$

# Neural networks with $\ell^1$ weights

Assume $\|x_i\| \leq \kappa$, and consider

$$f(x) = \sum_{j=1}^{m} v_j \, |w_j^\top x + b_j|_+, \quad \|v\|_1 \leq \gamma_\lambda, \quad \|(w_j, b_j)\|_2 = 1.$$

Then,

$$\mathcal{R}_n(\mathcal{H}_\lambda) \leq \frac{\gamma_\lambda}{\sqrt{n}} \, (\kappa^2 + 1)^{1/2}.$$

Indeed, by definition and the previous result,

$$\mathcal{R}_n(\mathcal{H}_\lambda) \leq \frac{\gamma_\lambda}{n} \, \mathbb{E}_\sigma \left[ \sup_{\|(w,b)\|_2 \leq 1} \Big| \sum_{i=1}^{n} \sigma_i |w^\top x_i + b| \Big|_+ \right].$$

By the contraction principle and dual norm relation,

$$\sup_{\|(w,b)\|_2 \leq 1} \Big| \sum_i \sigma_i |w^\top x_i + b| \Big| \leq \Big\| \sum_i \sigma_i (x_i, 1) \Big\|_2.$$

# Neural networks with $\ell^1$ weights (cont.)

By Jensen's inequality,

$$\mathbb{E}_\sigma \Big[ \Big\| \sum_i \sigma_i (x_i, 1) \Big\|_2 \Big] \leq \Big( \mathbb{E}_\sigma \Big[ \Big\| \sum_i \sigma_i (x_i, 1) \Big\|_2^2 \Big] \Big)^{1/2}.$$

Expanding the square and using $\mathbb{E}[\sigma_i \sigma_j] = \delta_{ij}$,

$$\mathbb{E}_\sigma \Big[ \Big\| \sum_i \sigma_i (x_i, 1) \Big\|_2^2 \Big] = \sum_i \|(x_i, 1)\|_2^2 \leq n \, (\kappa^2 + 1).$$

Putting all together,

$$\mathcal{R}_n(\mathcal{H}_\lambda) \leq \frac{\gamma_\lambda}{n} \sqrt{n(\kappa^2 + 1)} = \frac{\gamma_\lambda}{\sqrt{n}} \, (\kappa^2 + 1)^{1/2}.$$

# Outline

# Approximation error and generalization gap

$$\mathbb{E}[L(\widehat{f}_\lambda) - \inf_{f\in\mathcal{H}} L(f)] \leq \mathbb{E}[L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda)] + L(f_\lambda) - \inf_{f\in\mathcal{H}} L(f) + \lambda R(f_\lambda)$$

- Generalization gap (stochastic) ✓

$$L(\widehat{f}_\lambda) - \widehat{L}(\widehat{f}_\lambda) \leq 2\frac{C_\ell C_0(\kappa^2+1)^{1/2}}{\lambda\sqrt{n}}.$$

- Approximation error (deterministic) ✓

$$L(f_\lambda) - \inf_{f\in\mathcal{H}} L(f) + \lambda R(f_\lambda) \leq \lambda R(f_{\mathcal{H}}).$$

# Final bound

$$\mathbb{E}[L(\widehat{f}_\lambda) - \inf_{f \in \mathcal{H}} L(f)] \leq 2\frac{C_\ell C_0(\kappa^2+1)^{1/2}}{\lambda\sqrt{n}} + \lambda R(f_{\mathcal{H}}).$$

- We can choose $\lambda$ balancing these terms.
- We can obtain a bound $O(\frac{1}{n^{1/4}})$.
- This choice of $\lambda$ depends on quantities not known in practice; hold-out can be used instead.

# Remarks

- The result is not optimal. More refined empirical process theory is needed.
- Better (fast) rates can be obtained with further assumptions (noise, smoothness etc.).
- In practice, the estimation error often does not blow up as $\lambda$ decreases!
  Our prediction followed from the strategy using a rough norm bound...

# Summary

- Generalization error for Lipschitz loss and general representations.
- Dimension-free analysis applying to infinite-dimensional models.
- Again only one of many possible regimes (faster/slower rates possible!).
- Many extensions (implicit regularization, deep networks, etc.).

# Neural networks in TV RKBS

Assume $\|x_i\| \leq \kappa$, and that $\mu$ is supported on $\{(w,b) : \|w\|^2 + b^2 = 1\}$. Consider

$$f(x) = \int |w^\top x + b| \, d\mu(w,b), \quad \mu \in \big(M(\mathbb{R}^d \times \mathbb{R}), \|\cdot\|_{\mathrm{TV}}\big), \quad \Phi(x) \in \big(\mathcal{C}_0(\mathbb{R}^d \times \mathbb{R}), \|\cdot\|_\infty\big).$$

Then

$$\mathcal{R}_n(\mathcal{B}_\lambda) \leq \frac{\gamma_\lambda}{\sqrt{n}} (\kappa^2 + 1)^{1/2}.$$

By definition and the previous result,

$$\mathcal{R}_n(\mathcal{B}_\lambda) \leq \frac{\gamma_\lambda}{n} \, \mathbb{E}_\sigma \left[ \sup_{\|(w,b)\|_2 \leq 1} \left| \sum_{i=1}^n \sigma_i |w^\top x_i + b| \right| \right].$$

By the contraction principle and dual norm relation,

$$\sup_{\|(w,b)\|_2 \leq 1} \left| \sum_i \sigma_i |w^\top x_i + b| \right|_+ \leq \sup_{\|(w,b)\|_2 \leq 1} \left| \sum_i \sigma_i (x_i, 1)^\top (w,b) \right| = \left\| \sum_i \sigma_i (x_i, 1) \right\|_2.$$

# Neural networks in TV RKBS (cont.)

Assume $\|x_i\| \leq \kappa$, and that $\mu$ is supported on $\{(w,b) : \|w\|^2 + b^2 = 1\}$. Consider

$$f(x) = \int |w^\top x + b| \, d\mu(w,b), \quad \mu \in \big(M(\mathbb{R}^d \times \mathbb{R}), \|\cdot\|_{\mathrm{TV}}\big), \quad \Phi(x) \in \big(\mathcal{C}_0(\mathbb{R}^d \times \mathbb{R}), \|\cdot\|_\infty\big).$$

Then

$$\mathcal{R}_n(\mathcal{B}_\lambda) \leq \frac{\gamma_\lambda}{\sqrt{n}} (\kappa^2 + 1)^{1/2}.$$

By Jensen's inequality,

$$\mathbb{E}_\sigma \Big[ \Big\| \sum_i \sigma_i (x_i, 1) \Big\|_2 \Big] \leq \Big( \mathbb{E}_\sigma \Big[ \Big\| \sum_i \sigma_i (x_i, 1) \Big\|_2^2 \Big] \Big)^{1/2}.$$

Expanding the square and using $\mathbb{E}[\sigma_i \sigma_j] = \delta_{ij}$,

$$\mathbb{E}_\sigma \Big[ \Big\| \sum_i \sigma_i (x_i, 1) \Big\|_2^2 \Big] = \underbrace{\mathbb{E}_\sigma \Big[ \sum_{i,j} \sigma_i \sigma_j (x_i, 1)^\top (x_j, 1) \Big]}_{\mathbb{E}[\sigma_i \sigma_j] = \delta_{ij}} = \sum_i \|(x_i, 1)\|_2^2 \leq n(\kappa^2 + 1).$$