INSEAD PhD Course

Foundations of Machine Learning and AI

Theodoros Evgeniou - Nicolas Vayatis

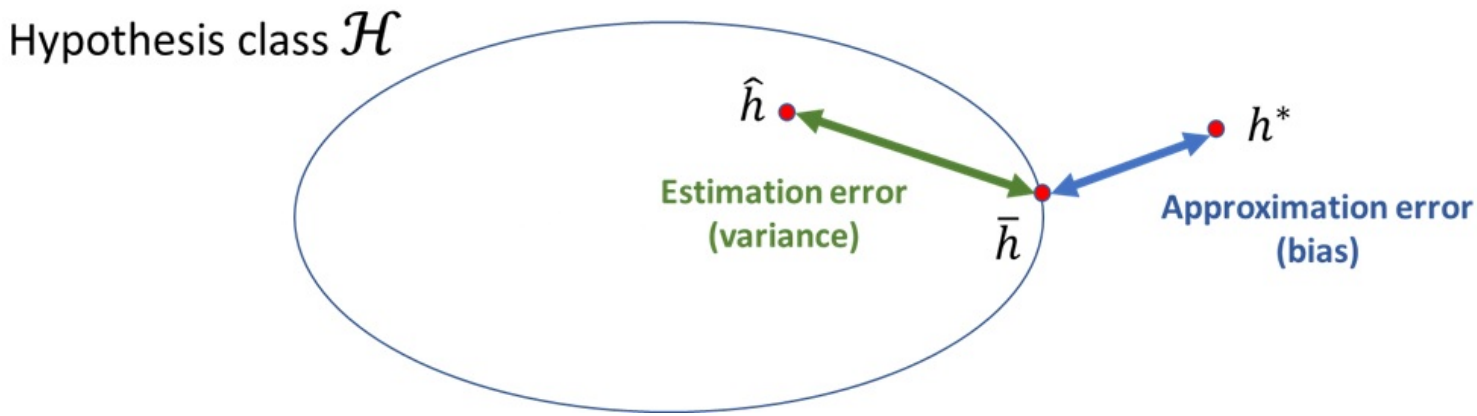
Sessions 5-6: Theory of Machine Learning

From previous session
Estimation vs. Approximation

- How to learn (select/estimate) a function from a set of functions: Approach is to use a penalized optimization with a chosen "regularization" (or "smoothing") parameter λ

$$\text{Criterion}(h) = \text{Training error}(h) + \lambda\, \text{Penalty}(h)$$

- **Key Question:** Can this formulation help us control the bias-variance trade off (through λ)?



"Shallow Learning"

- **Shallow Learning** are algorithms which will only depend on very few hyperparameters beyond the λ (such as the choice of a kernel). The methods seen so far fall in this category (and also those we will see in sessions 7-8).
- **Deep Learning** relies on many architectural hyperparameters (e.g., number of layers, nodes, etc - see Sessions 9-10) whose calibration is a very complex optimization problem. We will discuss this in sessions 9-10.
- The theory of Supervised Machine Learning applies to both shallow and deep learning.

Examples of Shallow Learning (1/2)

- Ridge regression

$$\hat{\beta}_\lambda \in \arg\min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_2^2 \right\}$$

- LASSO

$$\hat{\beta}_\lambda \in \arg\min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_1 \right\}$$

- Structured sparsity with $\|\beta\|_S$ being a sparsity inducing norm (group LASSO...)

$$\hat{\beta}_\lambda \in \arg\min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_S \right\}$$

Examples of Shallow Learning (2/2)

- SVM (hinge loss, L2 penalty)

$$\hat{\beta}_\lambda \in \arg\min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \cdot \beta^T X_i)_+ + \lambda \|\beta\|_2^2 \right\}$$

- Kernel ridge regression: K kernel and its parameter

$$\hat{\alpha}_\lambda = \min_{\alpha \in \mathbb{R}^n} \{\alpha^T \mathbf{K} \alpha - 2\alpha^T \mathbf{Y} + \lambda \alpha^T \alpha\}$$

where $\mathbf{K} = (K(X_i, X_j))_{i,j}$

Example of kernel K with one parameter μ:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{\mu} \right) \quad , \quad \mu > 0$$

Notations Reminder

- Hypothesis space $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathcal{Y}$ and loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty]$
- Empirical risk of a function h: this is a data-dependent quantity

$$\hat{L}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i)$$

- ERM = Empirical Risk Minimization

$$\hat{h}_n = \arg\min_{h \in \mathcal{H}} \hat{L}_n(h)$$

- Assuming (X, Y) has joint distribution P, the *prediction error* ("true loss"), assuming stationarity, for function h is given by: $L(h) = \mathbb{E}_P(\ell(h(X), Y))$

Estimation error for the ERM

- The estimation error of the ERM solution $\hat{h}_n$ on the hypothesis space $\mathcal{H}$ is:

$$L(\hat{h}_n) - \inf_{h \in \mathcal{H}} L(h)$$

[which is upper bounded by $2 \sup_{h \in \mathcal{H}} |L(h) - \hat{L}_n(h)|$ - see Sessions 1-2]

- **Main question for today**: understand the drivers of this quantity $\rightarrow$ dependence on: sample size n; "size" of $\mathcal{H}$; nature of stochastic fluctuations due to the training data (which is a random sample)

Upper bound on the estimation error bound for the ERM: Finite case

[See Sessions 1-2 and Exercise Set 1]

- Assume that the hypothesis space $\mathcal{H}$ of functions is finite and the loss function ℓ is bounded by 1 (any constant)
- Then, we have, for any $\delta > 0$, with probability at least $1 - \delta$:

$$L(\hat{h}_n) \leq \inf_{h \in \mathcal{H}} L(h) + \sqrt{\frac{\log |\mathcal{H}| + \log\left(\frac{1}{\delta}\right)}{2n}}$$

Theory of Machine Learning

A. From finite to infinite sets of functions

B1. VC Dimension ("Combinatorial Approach")

B2. Concentration inequalities

Complexity measures
Historical perspective

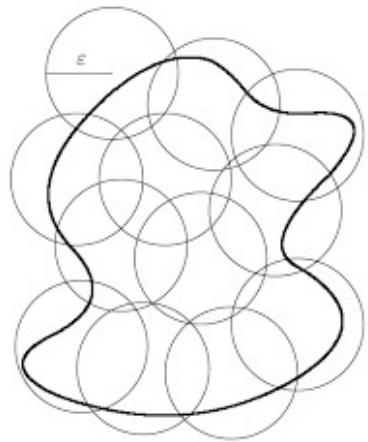
- Kolmogorov (1950's): developed metric concepts such as covering numbers, metric entropy... in mathematical analysis.
- Vapnik and Chervonenkis (1970's): discovered combinatorial concepts such as VC entropy, VC dimension and growth function in probability theory.
- Koltchinskii and Panchenko (2000) then Bartlett and Mendelson (2002): baptized a combinatorial quantity Rademacher complexity which was a variation of gaussian complexity in the continuous case to solve some technical issues in machine learning theory.

A. From finite to infinite sets of functions

Complexity measures based on metric concepts ("from dots to balls")

Covering numbers
Definition

- Consider a general space $\mathcal{H}$ (possibly space of functions) with a metric $\|\cdot\|$
- An ε-cover $\mathcal{T}$ is a set of elements of $\mathcal{H}$ such that for any $h \in \mathcal{H}$ there exists an element $t \in \mathcal{T}$ such that t is ε-close to h (i.e. $|h - t\| \leq \varepsilon$)



- The covering number $N(\varepsilon)$ is the cardinality of the smallest ε-cover of $\mathcal{H}$
- The metric entropy of $\mathcal{H}$ is the function $\varepsilon \mapsto \log N(\varepsilon)$

Covering numbers
Example

- Result: for the unit ball n $\mathbb{R}^d$, we have:

$$\left(\frac{1}{\varepsilon}\right)^d \leq N(\varepsilon) \leq \left(\frac{2}{\varepsilon} + 1\right)^d$$

Covering numbers
Upper bound on the error

Result by D. Pollard (1984)

- Notations: n sample size, ℓ loss function
- For bounded loss functions ($\|\ell(\cdot,\cdot)\| \leq M$), we have:

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} |L(h) - \hat{L}_n(h)| > \varepsilon\right) \leq N\left(\frac{\varepsilon}{8M}\right) \exp\left(-\frac{n\varepsilon^2}{2M^2}\right)$$

- Not easy to invert wrt to ε to obtain a clean error bound for $L(\hat{h}_n) - \inf_{h\in\mathcal{H}} L(h)$ (the variance part of the variance-bias for ERM)...

From finite to infinite sets of functions

Complexity concepts based on combinatorics (counting)

First: The classification problem
Notations

- Supervised data distributed as the pair (X, Y) over $\mathbb{R}^d \times \{-1, 1\}$
- Functions in $\mathcal{H}$ of the form: $h : \mathbb{R}^d \to \{-1, 1\}$ are equivalent to indicator functions of sets in $\mathbb{R}^d$
- Basic loss function: classification error $\ell(Y, h(X)) = \mathbb{I}\{Y \neq h(X)\}$
- Expected error and empirical error for function h:

$$L(h) = \mathbb{P}\{Y \neq h(X)\} \quad \text{and} \quad \hat{L}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Y_i \neq h(X_i)\}$$

- See regression later...

Vapnik-Chervonenkis (VC) entropy

- For a given sample $(X_1, \ldots, X_n)$ and for a given indicator function h, denote by $X_n(h)$ the $+1/$-1 (classification/"coloring") vector:

$$X_n(h) = (h(X_1), \ldots, h(X_n))^T \in \{-1, 1\}^n$$

- For this sample $(X_1, \ldots, X_n)$, denote by $\hat{N}(\mathcal{H})$ the cardinality of all possible vectors ("**colorings of the data**") achieved by the set of functions $h \in \mathcal{H}$. [Note that there are at most 2^n vectors but can be less than 2^n since some vectors ("colorings") may be un-achievable with functions in $\mathcal{H}$.]
- VC entropy (Expectation wrt Sample Data):

$$\mathcal{E}(\mathcal{H}) = \mathbb{E}(\log \hat{N}(\mathcal{H}))$$

- Relation with "over-fitting" and "Fooled by Randomness"?

Sufficient condition for estimation error to go to zero

- Finite case (reminder): convergence to zero of the estimation error ("Variance") if

$$\frac{\log |\mathcal{H}|}{n} \to 0 \ , \quad n \to \infty$$

- Similar role for the VC entropy: convergence to zero of the estimation error ("Variance") if

$$\frac{\mathcal{E}(\mathcal{H})}{n} = \frac{\mathbb{E}(\log \hat{N}(\mathcal{H}))}{n} \to 0 \ , \quad n \to \infty$$

- *Questions*: are there weaker conditions? Which sets of functions fulfill such a condition? What are the rates of convergence?

VC dimension
Definition

- The VC dimension of a function space $\mathcal{H}$ is the largest integer such that there exists a configuration of n points in $\mathbb{R}^d$ for which all its "colorings" (separations in $+1/$-1 classes) can be achieved by function in $\mathcal{H}$, *i.e.*

$$V(\mathcal{C}) = \max\{n \ \text{ integer } \ : \ \exists \ \text{Data s.t. } \ |\hat{N}(\mathcal{H})| = 2^n\}$$

- By comparison to the VC entropy, the VC dimension corresponds to a **worst case scenario** since the expectation is replaced by a maximum over all possible training samples.

VC dimension
Examples

- Hyperplanes in $\mathbb{R}^d$: $V = d + 1$
- Axis-aligned rectangles in $\mathbb{R}^2$: $V = 4$
- Just any rectangles in $\mathbb{R}^2$: $V = 7$
- Triangles in $\mathbb{R}^2$: $V = 7$
- Convex polygons in $\mathbb{R}^2$: $V = +\infty$

Observation: Number of parameters is irrelevant

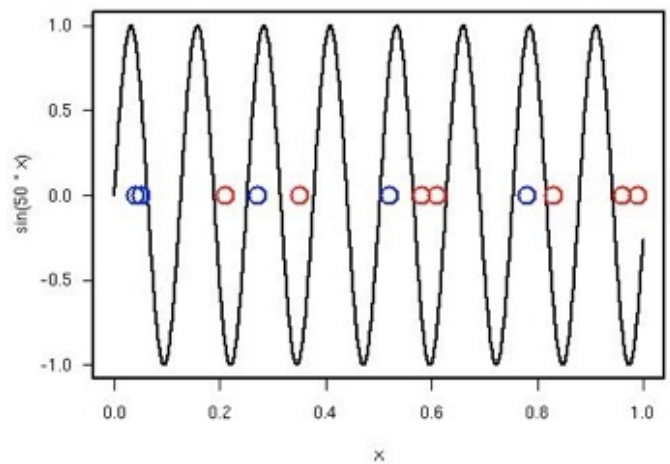
- Set of indicator functions parameterized by a single parameter ω:

$$h(x) = \mathbb{I}\{x : \sin(\omega x) > 0\}\ ,\ \text{where } \omega \in [0, 2\pi)$$

- VC dimension of this set is infinite, using:

$$\omega = \frac{1}{2}\left(1 + \sum_{i=1}^{n}\left(\frac{1 - y_i}{2}\right)10^i\right)$$

for a set of points $x_j = 2\pi 10^{-j}$



VC bound on classification error

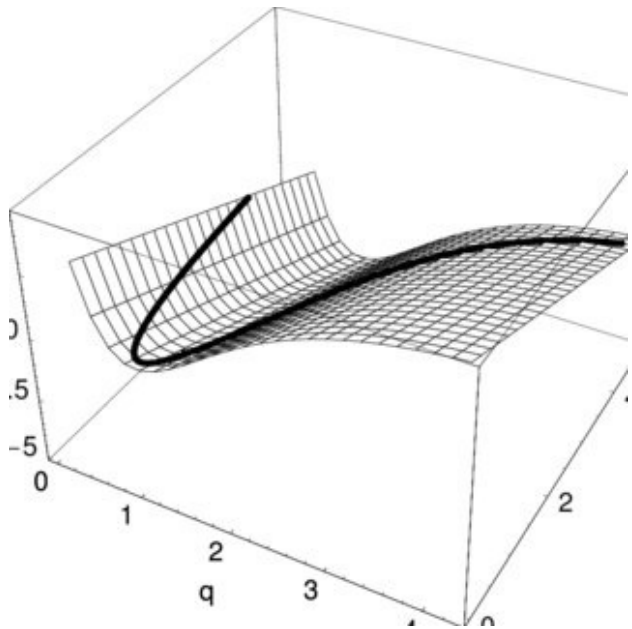
- Assume $\mathcal{H}$ has finite VC dimension V. Then, we have, for any δ, with probability at least $1 - \delta$:

$$L(\hat{h}_n) \leq \inf_{h \in \mathcal{H}} L(h) + \sqrt{\frac{2V \log\left(\frac{en}{V}\right)}{n}} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}}$$

- Behavior of the bound wrt V: as VC dimension V increases, the estimation error increases, but at the same time, it is expected that the approximation error goes down since the hypothesis space gets larger.

VC dimension for functions in the regression case

- Learning a function in the regression case boils down to learning a collection of indicator functions of the nested level sets of the function



What's next? Beyond VC Theory

1. Stability and Predictability
2. Rademacher Complexity

Background: Concentration inequalities

Talagrand (1996): "*A random variable that depends (in a "smooth way") on the influence of many independent variables (but not too much on any of them) is essentially constant.*"

Probability inequalities
Historical perspective

- Kolmogorov, Smirnov (1936): convergence of empirical cdf to its expectation
- Dvoretsky, Kiefer, Wolfowitz (DKW) (1956): nonasymptotic version of Kolmogorov-Smirnov
- Hoeffding (1963): deviation inequality (of average of IID from its expectation)
- Vapnik-Chervonenkis (1968): equivalent of DKW for general measures (not only 1D on half lines)
- Mc Diarmid (1981): first concentration inequality
- Massart (1990): exact constant in DKW
- Talagrand (1996): new concentration inequalities

Domains: uniform law of large numbers (and central limit theorem), empirical processes, large deviations, convex geometry, high dimensional probability

Reference: book by Boucheron-Lugosi-Massart (1996)

Reminder on Hoeffding's inequality

[see Exercise set 1]

- Assume $Z_1, \ldots, Z_n$ IID on $[0, 1]$ with common expectation $\mathbb{E}(Z_1)$
- Then we have, for any t:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}(Z_1) > t\right) \leq \exp(-2nt^2)$$

Basic concentration inequality
McDiarmid's inequality

- Here we replace the average of IID random variables by a general function of these IID variables.

- Consider $Z_1, \ldots, Z_n$ IID. Under a regularity assumption on the function f called the *bounded difference assumption*, we have, for any $t > 0$

$$\mathbb{P}\left(f(Z_1, \ldots, Z_n) - \mathbb{E}\big(f(Z_1, \ldots, Z_n)\big) > t\right) \leq \exp\left(-2nt^2\right)$$

- Take-home message: **Independence is more important/general than averaging**

Bounded difference assumption

- Consider a function f of n variables (the data in the case of learning). We say that f has bounded differences if the variations along each variables are uniformly bounded.
- Here we need to have:

$$\sup_{z_1,\ldots,z_n,z_i'} |f(z_1,\ldots,z_n) - f(z_1,\ldots,z_{i-1},z_i',z_{i+1},\ldots,z_n)| \leq \frac{1}{n}$$

[or $O(\frac{1}{n})$]

From finite to infinite sets of functions

Rademacher complexity: the modern approach to complexity

Learning Theory: Pre vs Post 2000

- Combinatorial complexity concepts (like VC-Dimension) were leading to loose bounds and raised technical difficulties. Cucker and Smale (2001); Evgeniou et al. (1999; 2000); Bartlett et al (1998) resolved various issues.
- Those complexity concepts also accounted for worst-case situations in terms of sample configuration. There was a challenge to develop data-dependent complexity measures (although it was possible).
- Two new approaches started in the late 1990s / early 2000s: **Stability** and **Rademacher complexity**.

Rademacher complexity
Why another concept?

- The concept was already there in 1968 (Vapnik-Chervonenkis paper) but was not identified as a key quantity except used in an intermediate step of a proof which had to be simplified in later stages of the proof.

- It was rediscovered in 2000 by Koltchinskii and Panchenko and led to neater bounds and theory to encompass all state-of-the-art methods such as SVM, boosting and bagging, as well as neural nets.

A data-dependent view on complexity

- VC entropy is about counting ("coloring") vectors *on average wrt the training data* in the hypercube of $\mathbb{R}^n$ defined by vectors of the form:

$$X_n(h) = (h(X_1), \ldots, h(X_n))^T \in \{-1, 1\}^n \ , \qquad \text{for all } h \in \mathcal{H}$$

- Rademacher complexity is about counting the possible label-prediction fitting outcomes *on average wrt the possible classification labels of a* **fixed** *training data set*

Rademacher complexity
Definition (1/2)

- Principle: Measures the expectation of the worst "misclassification" ("correlation" between randomly assigned labels and their predicted ones) with a random vector of Rademacher random variables:

$$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$$

where ε_i's are IID and independent of the training data such that $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$

Rademacher complexity
Definition (2/2)

- Consider a sample of $D_n = (X_1, \ldots, X_n)$ of IID random variables, then the Rademacher complexity of the set of functions $\mathcal{H}$ is the sample-dependent quantity:

$$\begin{aligned}\hat{R}_n(\mathcal{H}) &= \mathbb{E}\left(\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i h(X_i) \middle| D_n\right) \\ &= \frac{1}{n}\mathbb{E}\left(\sup_{h\in\mathcal{H}} (\varepsilon^T X_n(h)) \middle| D_n\right)\end{aligned}$$

- Exactly the same definition works for real-valued functions in the regression case.

Rademacher complexity for linear classes

- Consider a sample $x_1, \ldots, x_n$ which are all contained in a ball with radius R
- Denote by $\mathcal{H}$ the hypothesis space of linear functions such that $h(x) = \beta^T x$ where $\|\beta\|_2 \leq M$
- We then have:

$$\hat{R}_n(\mathcal{H}) \leq \frac{MR}{\sqrt{n}}$$

- Comment:
  1. similar bound in the case of kernel classes (with boundedness condition)
  2. Note relation with Ivanov Regularization (see Sessions 3-4)
  3. More examples in Sessions 7-8

Basic property
Rademacher complexity concentrates (Exercise)

- Set

$$f(X_1, \ldots, X_n) = \hat{R}_n(\mathcal{H}) = \mathbb{E}\left(\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(X_i) \middle| D_n \right)$$

- The function f satisfies the bounded differences assumption (why?)

- Therefore, we have, by McDiarmid's inequality, with probability at least $1 - \delta$:

$$\mathbb{E}(\hat{R}_n(\mathcal{H}) \leq \hat{R}_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Upper bound on the error of the ERM

- For the ERM solution $\hat{h}_n$, we have with probability at least $1 - \delta$:

$$L(\hat{h}_n) \leq \inf_{h \in \mathcal{H}} L(h) + \hat{R}_n(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

Link between Rademacher complexity and VC dimension

- Rademacher bounded by VC dimension

$$\hat{R}_n(\mathcal{H}) \leq \sqrt{\frac{2\big(1+\log(n/V)\big)}{(n/V)}}$$

- This means that finite VC dimension implies that Rademacher complexity is of the order of $\sqrt{(\log n)/n}$
- However, there are classes with infinite VC dimension which have Rademacher complexity of the same order of magnitude $\sqrt{(\log n)/n}$ (see sessions 7-8)

Theory of ML: From 2000 to Today

- Consistency/rates/fast rates of convergence of the estimation error for regularized learning methods: SVM (Steinwart, 2005), Boosting (Lugosi-Vayatis, 2004; Zhang, 2004), general surrogate losses (Bartlett, Jordan, McAuliffe, 2006)
- Theory of ranking and scoring algorithms with advanced concentration inequalities (Clémençon-Lugosi-Vayatis, 2008)
- Other tracks with theoretical advances: multiclass classification, ranking, multitask learning
- Other setups: online learning, learning view on game theory, transfer learning, active learning...

Stability

The principle of stability

- Stability is a property of the algorithm (e.g. ERM, KRR...) not of the hypothesis space
- Idea of stability of the function provided by an algorithm with respect to changes in the training set
- It builds upon the good old concept of robustness in statistics, revisited with modern tools from probability (concentration inequalities) and applied to the analysis of learning algorithms
- References: Bousquet and Elisseeff (2002) and Mukherjee, Niyogi, Poggio, and Rifkin (2006)

Definition of (uniform) stability

- Consider an algorithm which provides an estimator $\tilde{h}_n$ on a sample of size n and we denote $\tilde{h}'_n$ the estimator resulting from the same sample where one observation was changed.
- We say that the algorithm is *uniformly stable* if there exists a constant γ for which we have: for any pair (x, y),

$$|\ell(y, \tilde{h}_n(x)) - \ell(y, \tilde{h}'_n(x))| \leq \gamma$$

- Note: There are also other definitions of stability (e.g., *hypothesis stability*, *pointwise hypothesis stability*, *error stability*)

Error bound based on stability

- We have, with probability at least $1 - \delta$:

$$L(\tilde{h}_n) \leq \hat{L}_n(\tilde{h}_n) + (2n\gamma + M)\sqrt{\frac{\log(1/\delta)}{2n}}$$

- Proof based on concentration inequality
- Reminder: M is a bound on the values of the loss

Stability of Ridge Regression

- Assume the training data are all in a sphere of radius R and M is a bound on the loss function
- Ridge regression is stable with parameter γ such that:

$$\gamma \leq \frac{4MR^2}{n\lambda}$$

- *The smaller the regularization parameter λ, the less stable the algorithm*

Next session

- Other methods (e.g., random forests)
- Algorithms based on resampling and combinations (e.g. bagging, boosting)