9.520: Class 21

Approximation Error and Approximation Theory

Federico Girosi

Plan of the class

1. Learning and generalization error;

2. Approximation problem and rates of convergence;

3. N-widths;

4. "Dimension independent" convergence rates.

Note

These slides cover more extensive material than what will be presented in class.

References

The background material on generalization error (first 8 slides) is explained at length in:

[1] P. Niyogi and F. Girosi. On the relationship between generalization error, hypothesis complexity, and sample complexity for Radial Basis Functions. *Neural Computation*, 8:819–842, 1996.

[2] P. Niyogi and F. Girosi. Generalization bounds for function approximation from scattered noisy data. *Advances in Computational Mathematics*, 10:51–80, 1999.

[1] has a longer explanation and introduction, while [2] is more mathematical and also contains a very simple probabilistic proof of a class of "dimension independent" bounds, like the ones discussed at the end of this class.

As far as I know it is A. Barron who first clearly spelled out the decomposition of the generalization error in two parts. Barron uses a different framework from what we use, and he summarizes it nicely in:

[3] A.R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14:115–133, 1994.

The paper is quite technical, and uses a framework which is different from what we use here, but it is important to read it if you plan to do research in this field.

The material on n-widths comes from:

[4] A. Pinkus. *N-widths in Approximation Theory*, Springer-Verlag, New York, 1980.

Although the book is very technical, the first 8 pages contain an excellent introduction to the subject. The other great thing about this book is that you do not need to understand every single proof to appreciate the beauty and significance of the results, and it is a mine of useful information.

References (continued)

[5] H.N. Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation*, 8:164–177, 1996.

[6] A.R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transaction on Information Theory*, 39:3, 930–945, 1993.

[7] F. Girosi and G. Anzellotti. Rates of convergence of approximation by translates A.I. Memo 1288, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1992.

For a curious way to prove dimension independent bounds *using VC theory* see:

[8] F. Girosi. Approximation error bounds that use VC-bounds. In *Proc. International Conference on Artificial Neural Networks*, F. Fogelman-Soulié and P. Gallinari, editors, Vol. 1, 295–302. Paris, France, October 1995.

Notations review

- $I[f] = \int_{X \times Y} V(f(\mathbf{x}), y) p(\mathbf{x}, y) d\mathbf{x} dy$

- $I_{\text{emp}}[f] = \frac{1}{\ell} \sum_{i=1}^{\ell} V(f(\mathbf{x}_i), y_i)$

- $f_0 = \arg\min_f I[f] \quad , \qquad f_0 \in \mathcal{T}$

- $f_H = \arg\min_{f \in H} I[f]$

- $\widehat{f}_{H,\ell} = \arg\min_{f \in H} I_{emp}[f]$

More notation review

- $I[f_0] =$ how well we could possibly do
- $I[\widehat{f}_{H,\ell}] =$ how well we do in space H and ℓ data
- $I[f_H] =$ how well we can do in space H
- $|I[f] - I_{emp}[f]| \leq \Omega(H, \ell, \delta) \quad \forall f \in H$ (from VC theory)
- $I[\widehat{f}_{H,\ell}]$ is called *generalization error*
- $I[\widehat{f}_{H,\ell}] - I[f_0]$ is also called *generalization error* ...

A General Decomposition

$$I[\widehat{f}_{H,\ell}] - I[f_0] = (I[\widehat{f}_{H,\ell}] - I[f_H]) + (I[f_H] - I[f_0])$$

generalization error = estimation error + approximation error

When the cost function V is quadratic:

$$I[f] = E[(f_0 - f)^2] + I[f_0]$$

and therefore

$$E[(f_0 - f_{H,\ell})^2] = (I[\widehat{f}_{H,\ell}] - I[f_H]) + E[(f_0 - f_H)^2]$$

generalization error = estimation error + approximation error

A useful inequality

If, with probability $1 - \delta$

$$|I[f] - I_{emp}[f]| \leq \Omega(H, \ell, \delta) \quad \forall f \in H$$

then

$$|I[\widehat{f}_{H,\ell}] - I[f_H]| \leq 2\Omega(H, \ell, \delta)$$

Bounding the Generalization Error

$$E[(f_0 - f_{H,\ell})^2] \leq 2\Omega(H,\ell,\delta) + E[(f_0 - f_H)^2]$$

Notice that:

- Ω has nothing to do with the target space $\mathcal{T}$, it is studied mostly in statistics;

- $E[(f_0 - f_H)^2]$ has everything to do with the target space $\mathcal{T}$, it is studied mostly in approximation theory;

Approximation Error

We consider a *nested family* of hypothesis spaces H_n:

$$H_0 \subset H_1 \subset \ldots H_n \subset \ldots$$

and define the approximation error as:

$$\epsilon_{\mathcal{T}}(f, H_n) \equiv \inf_{h \in H_n} \|f - h\|$$

$\epsilon_{\mathcal{T}}(f, H_n)$ is the smallest error that we can make if we approximate $f \in \mathcal{T}$ with an element of H_n (here $\| \cdot \|$ is the norm in $\mathcal{T}$).

Approximation error

For reasonable choices of hypothesis spaces H_n:

$$\lim_{n \to \infty} \epsilon_{\mathcal{T}}(f, H_n) = 0$$

This means that we can approximate functions of $\mathcal{T}$ arbitrarily well with elements of $\{H_n\}_{n=1}^{\infty}$

For example: $\mathcal{T} =$ continuous functions on compact sets, and $H_n =$ polynomials of degree at most n.

The rate of convergence

The interesting question is:

How fast does $\epsilon_{\mathcal{T}}(f, H_n)$ go to zero?

The rate of convergence to zero is a measure of the relative complexity of $\mathcal{T}$ with respect to the approximation scheme H.

In the next slides we compute explicitely the rate of convergence of approximation of a smooth function by trigonometric polynomials.

We are interested in studying how fast the approximation error goes to zero when the number of parameters of our approximation scheme goes to infinity.

The reason for this exercise is that the results are representative: more complex and interesting cases all share the basic features of this example.

Rates of convergence in 1 dimension

Consider the set of functions

$$C_2[-\pi,\pi] \equiv C[-\pi,\pi] \bigcap L_2[-\pi,\pi]$$

Functions in this set can be represented as a Fourier series:

$$f(x) = \sum_{k=0}^{\infty} c_k e^{ikx} \ , \quad c_k \propto \int_{-\pi}^{\pi} dx f(x) e^{-ikx}$$

The L_2 norm satisfies the equation:

$$\|f\|_{L_2}^2 = \sum_{k=1}^{\infty} c_k^2$$

We consider as target space the following Sobolev space of smooth functions:

$$W_{s,2} \equiv \left\{ f \in C_2[-\pi, \pi] \mid \left\| \frac{d^s f}{dx^s} \right\|_{L_2}^2 < +\infty \right\}$$

The (semi)-norm in this Sobolev space is defined as:

$$\|f\|_{W_{s,2}}^2 \equiv \left\| \frac{d^s f}{dx^s} \right\|_{L_2}^2 = \sum_{k=1}^{\infty} k^{2s} c_k^2$$

In order for a function to belong to the space $W_{s,2}$ its Fourier coefficients c_k must go to zero at a rate which depends on s.

The set $W_{s,2}$ will be our *target space*. The *hypothesis space* H_n is the set of trigonometric polynomials of degree n, that consists of functions of the form:

$$p(x) = \sum_{k=1}^{n} a_k e^{ikx}$$

Given a funtion of the form

$$f(x) = \sum_{k=0}^{\infty} c_k e^{ikx}$$

It is easy to see that the trigonometric polynomial of degree n which best approximates f in the L_2 norm is given by the first n terms of its Fourier series:

$$f_n(x) = \sum_{k=1}^{n} c_k e^{ikx}$$

For a given $f \in W_{s,2}$ we want to study the approximation error:

$$\epsilon_n \equiv \|f - f_n\|_{L_2}^2$$

Notice that n, the degree of the polynomial, is also the number of parameters that we use in the approximation.

Obviously ϵ_n goes to zero as $n \to +\infty$, but the key question is: **how fast?**

An easy estimate of ϵ_n

$$\epsilon_n \equiv \|f - f_n\|^2_{L_2} = \sum_{k=n+1}^{\infty} c_k^2 = \sum_{k=n+1}^{\infty} c_k^2 k^{2s} \frac{1}{k^{2s}} <$$

$$< \frac{1}{n^{2s}} \sum_{k=n+1}^{\infty} c_k^2 k^{2s} < \frac{1}{n^{2s}} \sum_{k=1}^{\infty} c_k^2 k^{2s} = \frac{\|f\|^2_{W_{s,2}}}{n^{2s}}$$

$$\Downarrow$$

$$\epsilon_n < \frac{\|f\|^2_{W_{s,2}}}{n^{2s}}$$

More smoothness $\Rightarrow$ faster rate of convergence

But what happens in more than one dimension?

Rates of convergence in d dimensions

It is enough to study $d = 2$ and a particular choice of Sobolev space to understand the problem. We proceed in full analogy with the 1-d case:

$$f(x, y) = \sum_{k,m=1}^{\infty} c_{km} e^{i(kx+my)}$$

$$\|f\|^2_{W_{s,2}} \equiv \left\| \frac{d^s f}{dx^s} \right\|^2_{L_2} + \left\| \frac{d^s f}{dy^s} \right\|^2_{L_2} = \sum_{k,m=1}^{\infty} (k^{2s} + m^{2s}) c^2_{km}$$

Here $W_{s,2}$ is defined as the set of functions such that $\|f\|^2_{W_{s,2}} < +\infty$

As before the hypothesis space H_n is the set of trigonometric polynomials of degree l, which consists of functions of the form:

$$p(x) = \sum_{k,m=1}^{l} a_{km} e^{(ikx+imy)}$$

A trigonometric polynomial of degree l in d variables has a number of coefficients $n = l^d$.

We are interested in the behavior of the approximation error as a function of n. The approximating function is:

$$f_n(x, y) = \sum_{k,m=1}^{l} c_{km} e^{(ikx+imy)}$$

An easy estimate of ϵ_n

$$\epsilon_n \equiv \|f - f_n\|^2_{L_2} = \sum_{k,m=l+1}^{\infty} c^2_{km} = \sum_{k,m=l+1}^{\infty} c^2_{km}(k^{2s} + m^{2s})\frac{1}{k^{2s} + m^{2s}} <$$

$$< \frac{1}{2l^{2s}} \sum_{k,m=l+1}^{\infty} c^2_{km}(k^{2s} + m^{2s}) < \frac{1}{2l^{2s}} \sum_{k,m=1}^{\infty} c^2_{km}(k^{2s} + m^{2s}) =$$

$$= \frac{\|f\|^2_{W_{s,2}}}{2l^{2s}}$$

Since $n = l^d$, then $l = n^{\frac{1}{d}}$ (with $d = 2$), and we obtain:

$$\epsilon_n < \frac{\|f\|^2_{W_{s,2}}}{2n^{\frac{2s}{d}}}$$

Now notice that the previous calculations carries over to the d-dimensional cases in straightforward manner, and so the following general result holds:

if we approximate functions of d variables with s square integrable derivatives with a trigonometric polynomial with n coefficients, tha approximation error satisfies:

$$\epsilon_n < \frac{C}{n^{\frac{2s}{d}}}$$

More smoothness $\Rightarrow$ faster rate of convergence

Higher dimension $\Rightarrow$ slower rate of convergence

Another Example: Generalized Translation Networks

Consider networks of the form:

$$f(\mathbf{x}) = \sum_{k=1}^{n} a_k \phi(A_k \mathbf{x} + \mathbf{b}_k)$$

where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{b}_k \in \mathbb{R}^m$, $1 \leq m \leq d$, A_k are $m \times d$ matrices, $a_k \in \mathbb{R}$ and ϕ is some given function.

For $m = 1$ this is a Multilayer Perceptron .

For $m = d$, A_k diagonal and ϕ radial this is a Radial Basis Functions network.

Theorem (Mhaskar, 1994)

Let $W_s^p(\mathbb{R}^d)$ be the space of functions whose derivatives up to order s are p-integrable in $\mathbb{R}^d$. Under very general assumptions on ϕ one can prove that there exists $d \times m$ matrices $\{A_k\}_{k=1}^n$ such that, for any $f \in W_s^p(\mathbb{R}^d)$, one can find $\mathbf{b}_k$ and a_k such that:

$$\|f - \sum_{k=1}^{n} a_k \phi(A_k \mathbf{x} + \mathbf{b}_k)\|_p \leq c n^{-\frac{s}{d}} \|f\|_{W_s^p}$$

Moreover, the coefficients a_k are linear functionals of f.

This rate is optimal

The curse of dimensionality (1)

If the approximation error is

$$\epsilon_n \propto \left(\frac{1}{n}\right)^{\frac{s}{d}}$$

that means that the number of parameters needed to achieve an error smaller than ϵ is:

$$n \propto \left(\frac{1}{\epsilon}\right)^{\frac{d}{s}}$$

the curse of dimensionality is the d factor;

the blessing of smoothness is the s factor;

Jackson type rates of convergence

It happens "very often" that rates of convergence for functions in d dimensions with "smoothness" of order s are of the **Jackson type**:

$$O\left(\left(\frac{1}{n}\right)^{\frac{s}{d}}\right)$$

This is true for example for polynomial and spline approximation techniques, and it is also true for many non-linear techniques.

Can we do better than this? Can we defeat the curse of dimensionality? Have we tried hard enough to find "good" approximation techniques?

N-widths: definition (from Pinkus, 1980)

Let X be a normed space of functions, A a subset of X, and X_n a linear subspace of X, so that

$$X_n = \{f \in X \mid f = \sum_{i=1}^{n} c_i \phi_i, \quad \phi_i \in X, c_i \in R, i = 1, \ldots, n\}$$

Define the **Kolmogorov n-width** of A in X as

$$d_n(A, X) = \inf_{X_n} \sup_{f \in A} \inf_{y \in X_n} \|f - y\|$$

This is equivalent to

$$d_n(A, X) = \inf_{\phi_1, \ldots \phi_n} \sup_{f \in A} \inf_{c_1, \ldots c_n} \|f - \sum_{i=1}^{n} c_i \phi_i\|$$

Example (Kolmogorov, 1936)

$X = L_2[0, 2\pi]$

$\tilde{W}_2^s \equiv \{f \mid f \in C^{s-1}[0, 2\pi], f^{(j)} \text{ periodic}, \quad j = 0, \ldots, s-1\}$

$A = \tilde{B}_2^s \equiv \{f \mid f \in \tilde{W}_2^s \ , \|f^{(s)}\|_2 \leq 1\} \subset X$

Then

$$d_{2n-1}(\tilde{B}_2^s, L_2) = d_{2n}(\tilde{B}_2^s, L_2) = \frac{1}{n^s}$$

and the following X_n is optimal (in the sense that it achieves the rate above):

$X_{2n-1} = \text{span}\{1, \sin(x), \cos(x), \ldots, \sin(n-1)x, \cos(n-1)x\}$

Example: multivariate case

$I_d \equiv [0,1]^d$

$X = L_2[I_d]$

$W_2^s[I_d] \equiv \{f \mid f \in C^{s-1}[I_d]\ ,\ f^{(s)} \in L_2[I_d]\}$

$B_2^s \equiv \{f \mid f \in W_2^s[I_d]\ ,\ \|f^{(s)}\|_2 \leq 1\}$

Theorem (from Pinkus, 1980)

$$d_n(B_2^s, L_2) \approx \left(\frac{1}{n}\right)^{\frac{s}{d}}$$

Dimensionality and smoothness

Classes of functions in d dimensions with smoothness of order s have an *intrinsic complexity* characterized by the ratio $\frac{s}{d}$:

- the curse of dimensionality is the d factor;

- the blessing of smoothness is the s factor;

We cannot expect to find an approximation technique that "beats the curse of dimensionality" , *unless we let the smoothness s change with the dimension d*.

Theorem (Barron, 1991)

Let f be a function such that its Fourier transform satisfies

$$\int_{\mathbb{R}^d} d\omega \ \|\omega\| |\tilde{f}(\omega)| < +\infty \ .$$

and let Ω be a bounded domain in $\mathbb{R}^d$. Then we can find a neural network with n coefficients c_i, n weights $\mathbf{w}_i$ and n biases θ_i such that

$$\|f - \sum_{i=1}^{n} c_i \sigma(\mathbf{x} \cdot \mathbf{w}_i + \theta_i)\|^2_{L_2(\Omega)} < \frac{c}{n}$$

The rate of convergence is **independent of the dimension** d.

Here is the trick...

The space of functions such that

$$\int_{\mathbb{R}^d} d\omega \ \|\omega\| |\tilde{f}(\omega)| < +\infty \ .$$

is the space of functions that can be written as

$$f = \frac{1}{\|\mathbf{x}\|^{d-1}} * \lambda$$

where λ is any function whose Fourier Transform is integrable.

Notice how the space becomes more constrained as the dimension increases.

Jones lemma (Maurey, 1981; Jones, 1990; Barron, 1991)

Let H be an Hilbert space and $G \subset H$ such that $\|g\| \leq b \;\; \forall g \in G$. Let $f \in \overline{\text{coG}}$.
Then $\forall c > b^2 - \|f\|^2$ we can find n numbers $\alpha_k > 0$, such that $\sum_{i=1}^{n} \alpha_k = 1$, and n elements $g_k \in G$ such that

$$\|f - \sum_{k=1}^{n} \alpha_k g_k\|^2 \leq \frac{c}{n}$$

Theorem (Girosi and Anzellotti, 1992)

Let $f \in H^{s,1}(\mathbb{R}^d)$, where $H^{s,1}(\mathbb{R}^d)$ is the space of functions whose partial derivatives up to order s are integrable, and let $K_s(\mathbf{x})$ be the Bessel-Macdonald kernel, that is the Fourier transform of

$$\tilde{K}_s(\omega) = \frac{1}{(1 + \|\omega\|^2)^{\frac{s}{2}}} \qquad s > 0 \ .$$

If $s > d$ and s is even, we can find a Radial Basis Functions network with n coefficients c_α and n centers $\mathbf{t}_\alpha$ such that

$$\|f - \sum_{\alpha=1}^{n} c_\alpha K_s(\mathbf{x} - \mathbf{t}_\alpha)\|^2_{L_\infty} < \frac{c}{n}$$

Theorem (Girosi, 1992)

Let $f \in H^{s,1}(\mathbb{R}^d)$, where $H^{s,1}(\mathbb{R}^d)$ is the space of functions whose partial derivatives up to order s are integrable. If $s > d$ and s is even, we can find a Gaussian basis function network with n coefficients c_α, n centers $\mathbf{t}_\alpha$ and n variances σ_α such that

$$\|f - \sum_{\alpha=1}^{n} c_\alpha e^{-\frac{(\mathbf{x}-\mathbf{t}_\alpha)^2}{2\sigma_\alpha^2}}\|_{L_\infty}^2 < \frac{c}{n}$$

Same rate of convergence: $O(\frac{1}{\sqrt{n}})$

Function space	Norm	Approximation scheme
$\int_{\mathbb{R}^d} d\omega \; \|\tilde{f}(\omega)\| < +\infty$ **(Jones)**	$L_2(\Omega)$	$f(\mathbf{x}) = \sum_{i=1}^{n} c_i \sin(\mathbf{x} \cdot \mathbf{w}_i + \theta_i)$
$\int_{\mathbb{R}^d} d\omega \; \|\omega\| \|\tilde{f}(\omega)\| < +\infty$ **(Barron)**	$L_2(\Omega)$	$f(\mathbf{x}) = \sum_{i=1}^{n} c_i \sigma(\mathbf{x} \cdot \mathbf{w}_i + \theta_i)$
$\int_{\mathbb{R}^d} d\omega \; \|\omega\|^2 \|\tilde{f}(\omega)\| < +\infty$ **(Breiman)**	$L_2(\Omega)$	$f(\mathbf{x}) = \sum_{i=1}^{n} c_i \|\mathbf{x} \cdot \mathbf{w}_i + \theta_i\|_+ + \; + \mathbf{a} \cdot \mathbf{x} + b$
$\tilde{f}(\omega) \in C_0^k, \; 2k > d$ **(Girosi and Anzellotti)**	$L_\infty(\mathbb{R}^d)$	$f(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha e^{-\|\mathbf{x} - \mathbf{t}_\alpha\|^2}$
$H^{2s,1}(\mathbb{R}^d), \; 2s > d$ **(Girosi)**	$L_\infty(\mathbb{R}^d)$	$f(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha e^{-\frac{\|\mathbf{x} - \mathbf{t}_\alpha\|^2}{\sigma_\alpha^2}}$