Approximation Error and Approximation Theory

Federico Girosi

Center for Basic Research in the Social Sciences
Harvard University
and
Center for Biological and Computational Learning
MIT

- Learning and generalization error
- Approximation problem and rates of convergence
- N-widths
- "Dimension independent" convergence rates

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

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 $\mathfrak{n}$-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.
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-Souliè and P. Gallinari, editors, Vol. 1, 295–302. Paris, France, October 1995.

- $I[f] = \int_{X \times Y} V(f(\mathbf{x}), y) p(\mathbf{x}, y) d\mathbf{x} dy$
- $I_{emp}[f] = \frac{1}{l} \sum_{i=1}^{l} V(f(\mathbf{x}_i), y_i)$
- $f_0 = \arg\min_f I[f] \quad , \qquad f_0 \in \mathcal{T}$
- $f_H = \arg\min_{f \in H} I[f]$
- $\hat{f}_{H,l} = \arg\min_{f \in H} I_{emp}[f]$

- $\mathrm{I}[\mathrm{f}_0]$ = how well we could possibly do
- $\mathrm{I}[\mathrm{f}_\mathrm{H}]$ = how well we can do in space H
- $\mathrm{I}[\hat{\mathrm{f}}_{\mathrm{H},\mathrm{l}}]$ = how well we do in space H with l data
- $|\mathrm{I}[\mathrm{f}] - \mathrm{I}_{\mathrm{emp}}[\mathrm{f}]| \leq \Omega(\mathrm{H}, \mathrm{l}, \delta) \quad \forall \mathrm{f} \in \mathrm{H}$ (from VC theory)
- $\mathrm{I}[\hat{\mathrm{f}}_{\mathrm{H},\mathrm{l}}]$ is called *generalization error*
- $\mathrm{I}[\hat{\mathrm{f}}_{\mathrm{H},\mathrm{l}}] - \mathrm{I}[\mathrm{f}_0]$ is also called *generalization error* ...

$$I[\hat{f}_{H,\iota}] - I[f_0] = I[\hat{f}_{H,\iota}] - I[f_H] + I[f_H] - I[f_0]$$

$$\text{generalization error} = \text{estimation error} + \text{approximation error}$$

When the cost function V is quadratic:

$$I[f] = \|f_0 - f\|^2 + I[f_0]$$

and therefore

$$\|f_0 - f_{H,\iota}\|^2 = I[\hat{f}_{H,\iota}] - I[f_H] + \|f_0 - f_H\|^2$$

$$\text{generalization error} = \text{estimation error} + \text{approximation error}$$

If, with probability $1 - \delta$

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

then

$$\left| I[\hat{f}_{H,l}] - I[f_H] \right| \leq 2\Omega(H, l, \delta)$$

You can prove it using the following observations:

- $I[f_H] \leq I[\hat{f}_{H,l}]$ (from the definition of f_H)
- $I_{emp}[\hat{f}_{H,l}] \leq I_{emp}[f_H]$ (from the definition of $\hat{f}_{H,l}$)

$$\|f_0 - f_{H,l}\|^2 \leq 2\Omega(H, l, \delta) + \|f_0 - f_H\|^2$$

Notice that:

- Ω has nothing to do with the target space $\mathcal{T}$, it is studied mostly in statistics;
- $\|f_0 - f_H\|$ has everything to do with the target space $\mathcal{T}$, it is studied mostly in approximation theory;

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

$$\mathrm{H}_0 \subset \mathrm{H}_1 \subset \ldots \mathrm{H}_n \subset \ldots$$

and define the approximation error as:

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

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

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 $\{H_n\}_{n=1}^{\infty}$

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

The interesting question is:

How fast does $\epsilon_{\mathcal{T}}(\mathrm{f}, \mathrm{H}_{\mathrm{n}})$ go to zero?

- The rate of convergence is a measure of the relative complexity of $\mathcal{T}$ with respect to the approximation scheme H.
- The rate of convergence determines how many samples we need in order to obtain a given generalization error.

- In the next slides we compute explicitly 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.

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:

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

The L_2 norm satisfies the equation:

$$\|f\|_{L_2}^2 = \sum_{k=0}^{\infty} {c_k}^2$$

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:

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

If f belongs to $W_{s,2}$ then Fourier coefficients c_k must go to zero at a rate which increases with s.

We choose as *hypothesis space* H_n the set of trigonometric polynomials of degree n:

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

Given a function of the form

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

the optimal hypothesis $f_n(x)$ 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:

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

- 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?**

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

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

$$\Downarrow$$

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

More smoothness $\Rightarrow$ faster rate of convergence

But what happens in more than one dimension?

It is enough to study $d = 2$. 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 $\|f\|^2_{W_{s,2}} < +\infty$

We choose as hypothesis space H_n the set of trigonometric polynomials of degree l:

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

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:

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

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

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

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

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

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

The previous calculations generalizes easily to the $\mathtt{d}$-dimensional case. Therefore we conclude that:

if we approximate functions of $\mathtt{d}$ *variables with* $\mathtt{s}$ *square integrable derivatives with a trigonometric polynomial with* $\mathtt{n}$ *coefficients, the approximation error satisfies:*

$$\epsilon_{\mathtt{n}} < \frac{\mathrm{C}}{\mathtt{n}^{\frac{2\mathtt{s}}{\mathtt{d}}}}$$

More smoothness $\mathtt{s}$ $\Rightarrow$ **faster rate of convergence**

Higher dimension $\mathtt{d}$ $\Rightarrow$ **slower rate of convergence**

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.

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:

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

Moreover, the coefficients a_k are linear functionals of f.

This rate is optimal

If the approximation error is

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

then 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;

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

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

Example: polynomial and spline approximation techniques, 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?

Let X be a normed space of functions, A a subset of X. We want to approximate elements of X with linear superposition of n basis functions $\{\phi_i\}_{i=1}^{n}$.

Some sets of basis functions are better than others: which are the best basis function? what error do they achieve?

To answer these questions we define the **Kolmogorov n-width** of A in X:

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

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

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

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

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):

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

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

$$X = L_2[I_d]$$

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

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

Theorem (from Pinkus, 1980)

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

Optimal basis functions are usually splines (or their relatives)

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

- 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 increase with the dimension d*.

Let f be a function such that its Fourier transform satisfies

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

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

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

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

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.

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

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

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 t_α and n variances σ_α such that

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

n

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

- There is a trade off between the size of the sample ($\mathtt{l}$) and the size of the hypothesis space $\mathtt{n}$;
- For a given pair of hypothesis and target space the approximation error depends on the trade off between dimensionality and smoothness;
- The trade off has a "generic" form and sets bounds on what can and cannot be done, both in linear and non-linear approximation;
- Suitable spaces, which trade dimensionality versus smoothness, can be defined in such a way that the rate of convergence of the approximation error is independent of the dimensionality.