Geometric Rates of Approximation by Neural Networks

Věra Kůrková
Institute of Computer Science
Academy of Sciences of the Czech Republic
Prague

Marcello Sanguineti
Universita di Genova, Genova

estimates of **model complexity of neural networks**

derived using **tools from approximation theory**

Learning = optimization problem

$$(M, \Phi)$$

hypothesis set
input/output functions

functional
defined by data

minimize Φ over M

$\mathrm{span}_n G$ = linear combinations of n functions corresponding to the type of computational units

expected error functional $\mathcal{E}_\rho$
empirical error functional $\mathcal{E}_z$
data: sample z or measure ρ

n = number of network units = measure of network complexity

Problem

for given data (defined either by a probability measure or by a sample) find a suitable type of computational units (defined by a parameterized set of functions G)

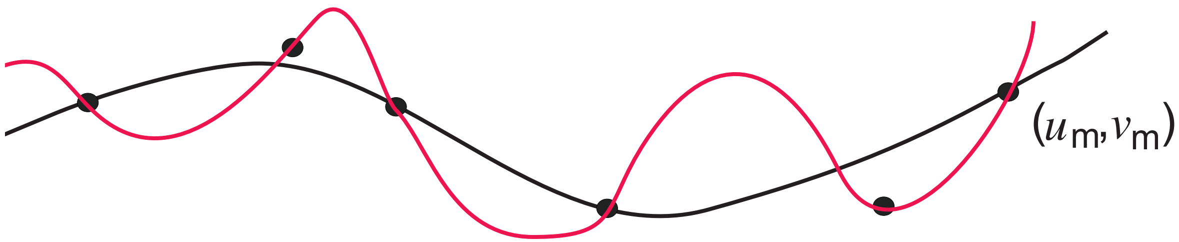
the better choice of units, the smaller number n of computational units

Functional defined by a sample of data

$z = \{(u_i, v_i) \mid i = 1, \ldots, m\} \subseteq \mathbb{R}^d \times \mathbb{R}$ sample of data

Empirical error functional

$$\mathcal{E}_z(f) = \frac{1}{m} \sum_{i=1}^{m} (f(u_i) - v_i)^2$$



Minimization of empirical error functional =

the least square method Gauss 1809, Legendre 1806

Functional defined by a probability measure

ρ = non degenerate (no nonempty open set has measure zero)
probability measure on $Z = X \times Y$ $\quad \rho(Z) = 1$
$X \subset \mathbb{R}^d$ compact $\qquad Y \subset \mathbb{R}$ bounded

Expected error functional

$$\mathcal{E}_\rho(f) = \int_{X \times Y} (f(u) - v)^2 d\rho$$

Traditional applications of the least square method

best fitting functions were searched for in
LINEAR hypothesis spaces

$\Rightarrow$ limitations on applications to high-dimensional data!

CURSE OF DIMENSIONALITY

the dimension n of a linear space needed for approximation of smooth functions of d variables within accuracy ε is

$$\mathcal{O}\left(\left(\frac{1}{\varepsilon}\right)^d\right)$$

$\Rightarrow$ model complexity n of LINEAR models grows EXPONENTIALLY with the data dimension d

Hypothesis sets in neurocomputing

the best fitting functions are searched for in
NONLINEAR and NONCONVEX hypothesis spaces

$$\mathrm{span}_n\, G = \{\sum_{i=1}^{n} \omega_i g_i \,|\, \omega_i \in \mathbb{R}, g_i \in G\}$$

= set of functions computable by a network with one linear output and n hidden units computing functions from G

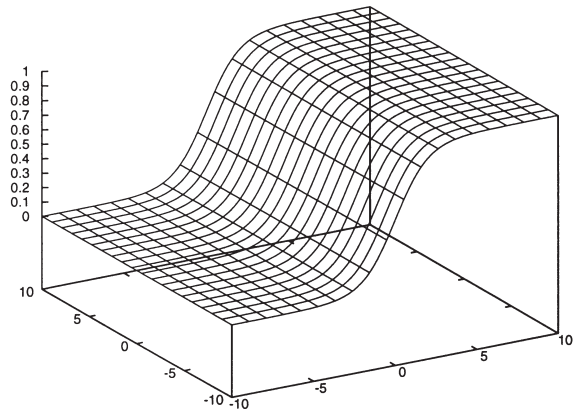
a nested family $\quad .... \subseteq \mathrm{span}_n\, G \subseteq \mathrm{span}_{n+1}\, G \subseteq ....$

variable-basis approximation scheme
approximation from a dictionary

Computational units

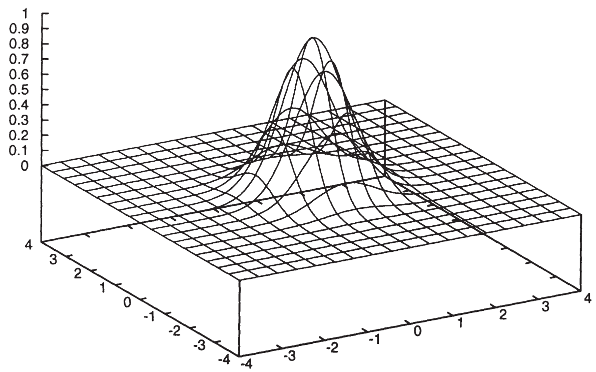
perceptrons:

$G = \mathcal{P}_d(\sigma) = \{\sigma(v.x + b) \mid v \in \mathbb{R}^d, b \in \mathrm{R}\}$ PLANE WAVES



radial-basis function (RBF) units:

$G = \mathcal{B}_d(\psi) = \{\psi(b\|x - v\|) \mid v \in \mathbb{R}^d, b \in \mathbb{R}\}$ SPHERE WAVES



Optimal solution

Global minimum of expected error $\mathcal{E}_\rho$

Regression function

$$f_\rho(x) = \int_Y y \, d\rho(y|x)$$

$\rho(y|x)$ =conditional (w.r.t. x) probability measure on Y
ρ_X = marginal probability measure on X ($\forall S \subseteq X \quad \rho_X(S) = \rho(\pi_X^{-1}(S))$, $\quad \pi_X : X \times Y \to X$ projection)

$$\min_{f \in \mathcal{L}^2_{\rho_X}} \mathcal{E}_\rho(f) = \mathcal{E}_\rho(f_\rho)$$

the regression function f_ρ is global minimizer of $\mathcal{E}_\rho$ over $\mathcal{L}^2_{\rho_X}$

Optimal solution

Global minimum of empirical error $\mathcal{E}_z$

$\forall$ sample of data z of size m
$\exists$ interpolating function f^o computable by a network with m units $\quad f^o \in \mathrm{span}_m G$

$$\min_{f \in \mathrm{span}_m G} \mathcal{E}_z(f) = \mathcal{E}_z(f^o) = 0$$

holds for sigmoidal perceptrons and RBF and kernel units with suitable kernels

Approximate minimization

optimal solutions f^o and the regression function f_ρ
may not be computable by networks with a reasonably small number of hidden units

BUT they can be approximated by suboptimal solutions
= minima over $\mathrm{span}_n G$ with $n << m$ number of units

approximation of the problems $(\mathrm{span}_m G, \mathcal{E}_z)$ and $(\mathrm{span}_m G, \mathcal{E}_\rho)$

by a sequence of problems

$\{(\mathrm{span}_n G, \mathcal{E}_z)|\, n = 1, \ldots, m\}$ and $\{(\mathrm{span}_n G, \mathcal{E}_\rho)|\, n = 1, \ldots, m\}$

speed of convergence as a measure of complexity

$\inf_{f \in \mathrm{span}_n G} \mathcal{E}_z(f) \longrightarrow 0$ and $\inf_{f \in \mathrm{span}_n G} \mathcal{E}_\rho(f) \longrightarrow \mathcal{E}_\rho(f_\rho)$

Tools from approximation theory

minimization of expected error $\mathcal{E}_\rho$ is equivalent to
minimization of the $\mathcal{L}^2_{\rho_X}$-distance from the regression function f_ρ

minimization of empirical error $\mathcal{E}_z$ is equivalent to
minimization of the l^2-distance from f_z
$f_z(u_i) = v_i$

$\Rightarrow$ we can use tools from approximation theory to
estimate speed of convergence of infima (minima) of
error functionals over $\mathrm{span}_n G$ with n increasing

Upper bound on rates of variable-basis approximation

Maurey (1981), Jones (1992), Barron (1993)

G a bounded subset of a Hilbert space $(X, \|.\|)$, $s_G = \sup_{g \in G} \|g\|$

$\forall f \in \mathrm{conv}\, G \qquad \forall\ n$

$$\|f - \mathrm{conv}_n G\| \leq \sqrt{\frac{{s_G}^2 - \|f\|^2}{n}}$$

$\mathrm{conv}_n\, G = \{\sum_{i=1}^n \omega_i g_i \,|\, \omega_i \in [0,1], \sum_{i=1}^n = 1, g_i \in G\}$

Corollary: $\forall f \in X \qquad \forall\ n$

$$\|f - \mathrm{span}_n G\| \leq \frac{s_G \|f\|_G}{\sqrt{n}}$$

$\|.\|_G \quad = \quad$ norm tailored to G

$\|f\|_G = \inf\{b > 0 \,|\, \frac{f}{b} \in \mathrm{cl\,conv}(G \cup -G)\}$

Comparison with linear approximation

number of hidden units = model complexity of the network
needed for approximation within ε grows as

$$\mathcal{O}\left(\frac{1}{\varepsilon}\right)^2$$

in contrast to $\mathcal{O}\left(\left(\frac{1}{\varepsilon}\right)^d\right)$ in linear approximation

d = number of variables of functions in G
= number of network inputs

Norm tailored to a set of functions G

$(X, \|.\|)$ normed linear space, G bounded subset of X

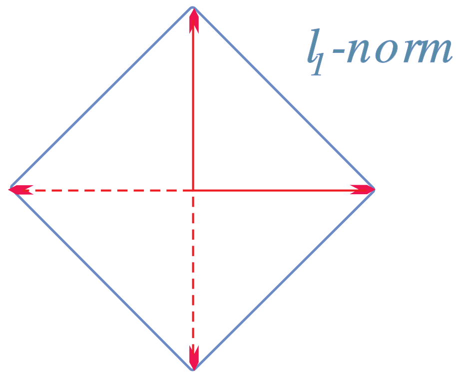
G-variation = Minkowski functional of the closed convex symmetric hull of G

$$\|f\|_G = \inf\{b > 0 \,|\, \tfrac{f}{b} \in \mathrm{cl\,conv}(G \cup -G)\}$$

(1) G orthogonal

$\|f\|_G = \|f\|_{1,G} = \sum_{g \in G} |f.g|$

l_1-norm wrt G



(2) G characteristic functions of half-spaces (perceptrons)

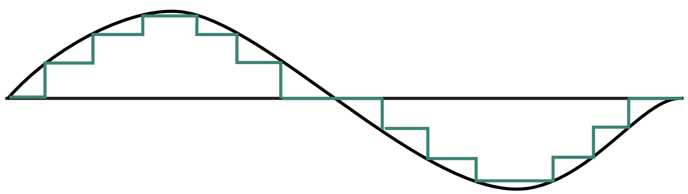
variation wrt half-spaces

(generalization of total variation)

$T(f) = \int |f'| \qquad d = 1$

$\approx$ sum of "heights of steps"

θ Heaviside activation function



Tightness of Maurey-Jones-Barron's theorem

Maurey-Jones-Barron's theorem is a worst-case result
holds for all functions in a ball in variational norm

Tightness results:
G orthonormal (constructive proof)
G sigmoidal perceptrons (proof by contradiction based on comparison of covering numbers)

Improvements of Maurey-Jones-Barron's theorem

better rates of approximation for
suitable subsets of balls in variational norms

Lavretsky, 2002
defined a subset $F_\delta(G)$ of $\mathrm{conv}G$ (for $\delta \in (0,1]$)

$\forall f \in F_\delta(G)$
$\|f - \mathrm{conv}_n G\| \le (1-\delta)^{n-1}({s_G}^2 - \|f\|^2)$

missing characterization of $F_\delta(G)$, no examples
? is $F_\delta(G)$ non-empty ?

non transparent definition
$F_\delta(G) = \Big\{ f \in \mathrm{cl}\,\mathrm{conv}\,G \,|\, (\forall h \in \mathrm{conv}\,G,\ f \ne h)(\exists g \in G)$
$\Big((f-g)\cdot(f-h) \le -\delta\, \|f-g\|\, \|f-h\| \Big) \Big\}$

Geometric rate for all functions in $\operatorname{conv} G$

Kůrková, Sanguineti

$(X, \|.\|)$ a Hilbert space, G its bounded subset

$\forall f \in X \ \exists \delta_f \in (0,1]$

$$\|f - \operatorname{conv}_n G\| \leq (1-\delta_f)^{n-1}({s_G}^2 - \|f\|^2)$$

constructive proof, δ_f and incremental approximants are not defined uniquely

Sets of functions with the same geometric rate

we can define unique $\delta(f)$

$$\delta(f) = \max\left\{\delta > 0 \,|\, (\forall n) \|f - \mathrm{conv}_n G\| \leq (1-\delta^2)^{n-1}({s_G}^2 - \|f\|^2)\right\}$$

$$A_\delta(G) = \{f \in \mathrm{conv}\, G \,|\, \delta(f) = \delta\}$$

$$\mathrm{conv} G = \cup_{\delta \in (0,1]} A_\delta(G)$$

? geometry of sets $A_\delta(G)$?

Geometry of sets $A_\delta(G)$

$(X, \|.\|)$ infinite-dimensional separable Hilbert space
G orthornormal basis

$\forall k \geq 3 \ \exists h_k \in \operatorname{conv} G$ with $\|h_k\| = \frac{1}{\sqrt{2k}}$

$\delta(h_k)^2 \leq 1 - 5^{-\frac{1}{k-1}} \, e^{-\frac{\ln(k-1)}{k-1}}$

$A_\delta(G)$ are not convex and do not contain any ball (even any sphere) in $\|.\|$

in finite dimensional spaces sets $A_\delta(G)$ contain balls in $\|.\|$

Conclusion

every function f in a Hilbert space can be approximated by $\mathrm{span}_n G$ with a rate bounded from above by $\frac{\|f\|_G}{\sqrt{n}}$

G-variation can be estimated using various methods (integral representations, smoothing operators, maxima of partial derivatives...)

MOREOVER

every function f in $\mathrm{conv}\, G$ can be approximated by $\mathrm{span}_n G$ with a rate bounded from above by $(1 - \delta(f))^{n-1}({s_G}^2 - \|f\|^2)$ where $\delta(f) \in (0, 1]$ is specific for each f

BUT geometry of sets with the same $\delta(f)$ is complicated, characterization is difficult