Estimates of Data Complexity in Neural Network Learning

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

vera@cs.cas.cz

Learning = optimization problem

(M, Φ)

hypothesis set
of network I/O
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

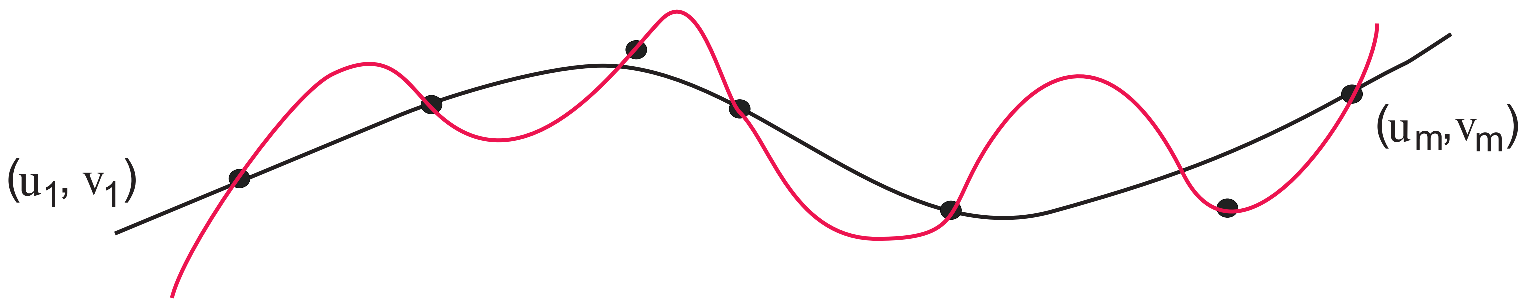
expected error functional $\mathcal{E}_\rho$
empirical error functional $\mathcal{E}_z$

Functional defined by a sample of data

$z = \{(u_i, v_i) : i = 1, \ldots, m\} \subseteq \mathbb{R}^d \times \mathbb{R}$ 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 = nondegenerate (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

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

The least square method: statistical inference, pattern recognition, function approximation, curve or surface fitting, etc.

the best fitting function was searched for in
LINEAR hypothesis spaces

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

CURSE OF DIMENSIONALITY

the dimension of a linear space needed for approximation of a function of d variables within accuracy ε is

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

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

Hypothesis sets in neurocomputing

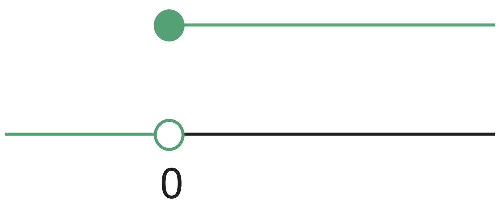
$$\mathrm{span}_n\, G = \{\sum_{i=1}^{n} \omega_i g_i \,|\, \omega_i \in \mathbb{R}, g_i \in G\}$$

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

NONLINEAR and NONCONVEX

Computational units: Heaviside perceptrons

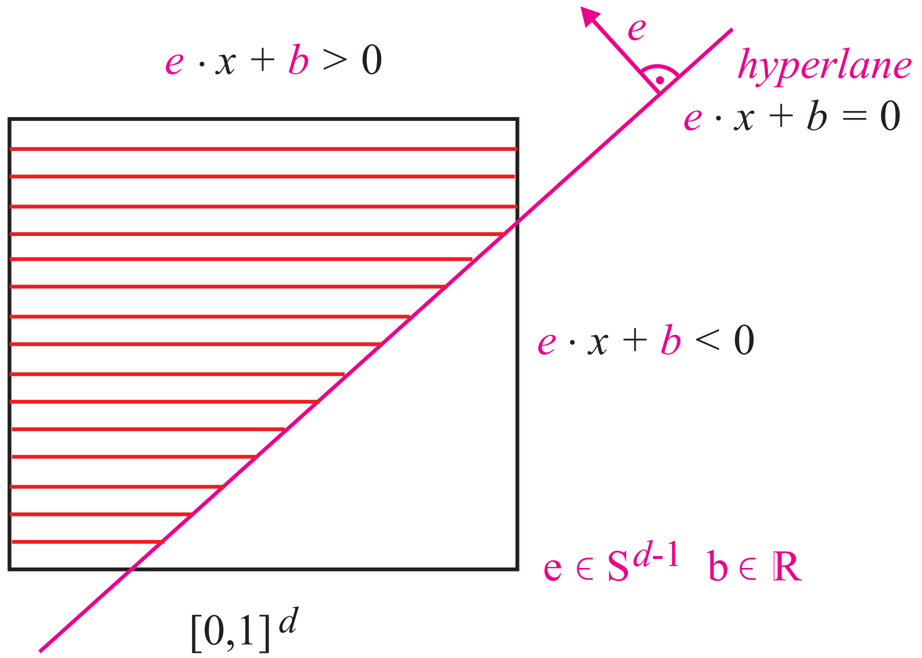
ϑ Heaviside function



$H_d(X) = \{\vartheta(e \cdot x + b) : X \to \mathbb{R} \mid e \in S^{d-1}, b \in \mathbb{R}\}$
set of characteristic functions of half-spaces of $X \subseteq \mathbb{R}^d$

$\mathrm{span}_n H_d(X)$ = set of functions on $X \subset \mathbb{R}^d$ computable by neural networks with n Heaviside perceptrons and one linear output

Heaviside perceptrons



compute functions of the form $\vartheta(e \cdot x + b)$

= characteristic functions of half-spaces

Optimal solution

Global minimum of expected error

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

Existence of the global minimum of empirical error over a set of functions computable by perceptron networks

Ito (92) $\forall$ sample of data z of size m

$\exists$ interpolating function f^o computable by a network with m perceptrons $f^o \in \mathrm{span}_m H_d$

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

similar results for RBF and kernel units

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 ?

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

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

$\Rightarrow$ 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

Rates of convergence of infima of expected error functional over networks with n units

$$\inf_{f \in \mathrm{span}_n G} \mathcal{E}_\rho(f) - \mathcal{E}_\rho(f_\rho) \leq \frac{\|f_\rho\|_G^2}{n}.$$

$\|f_\rho\|_G$ =norm tailored to G
variation with respect to G

value of the variation norm at f_ρ = measure of complexity ("smoothness") of data wrt the class of networks with units computing functions from G

Rates of convergence of minima of empirical error functional over networks with n units

for every h interpolating the sample z

$$\inf_{f \in \mathrm{span}_n G} \mathcal{E}_z(f) \leq \frac{\|h\|_G^2}{n}.$$

the smallest value of the variational norm of
a function interpolating z
$=$ measure of complexity ("smoothness") of data wrt the class
of networks with units computing functions from G

Comparison with linear approximation

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

$$\left(\frac{\|f_\rho\|_G}{\varepsilon}\right)^2 \quad \text{or} \quad \left(\frac{\|h\|_G}{\varepsilon}\right)^2$$

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

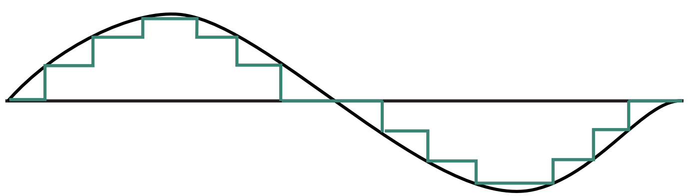
? dependence of variational norm on dimensionality d
= number of variables of functions in G = number of network inputs

Variation with respect to half-spaces

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

$$\|f\|_{H_d} = \inf\{b > 0 : \tfrac{f}{b} \in cl\ conv(H_d \cup -H_d)\}$$

ϑ Heaviside activation function



$d = 1$

generalization of total variation

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

$\approx$ sum of "heights of steps"

Smooth functions have small variations wrt half-paces

Kainen, Kůrková, Vogt (2005)

$$\|f\|_{H_d(\mathbb{R}^d)} \leq k_d \ \|f\|_{d,1,\infty}$$

$k_d < \frac{1}{\sqrt{d}} 2^{-\frac{d}{2}}$

k_d decreases with the number of variables d exponentially fast

Sobolev-type seminorm

$\|f\|_{d,1,\infty} = \max_{|\alpha|=d} \|D^\alpha f\|_{\mathcal{L}_1(\mathbb{R}^d)}$

$\|f\|_{d,1,\infty}$ is much smaller than $\|f\|_{d,1} = \sum_{|\alpha|\leq d} \|D^\alpha f\|_{\mathcal{L}_1(\mathrm{R}^d)}$

$\|f\|_{d,1,\infty}$ is maximum of partial derivatives,
while $\|f\|_{d,1}$ is sum of 2^d partial derivatives

$D^\alpha f = \frac{\partial^{\alpha_1}}{\partial x_1} \dots \frac{\partial^{\alpha_d}}{\partial x_d} f \qquad |\alpha| = \sum_{i=1}^d \alpha_i$

Integral representation as a perceptron network with a continuum of hidden units

Kůrková Kainen, Kreinovich (1997)
Kainen, Kůrková, Vogt (2005)
$\forall d$ odd $\forall f : \mathbb{R}^d \to \mathbb{R}$ sufficiently quickly vanishing at infinity

$$f(x) = \int_{S^{d-1}} \int_{\mathbb{R}} w_f(e,b)\vartheta(e \cdot x + b)dedb$$

$$\omega_f(e,b) = a_d \int_{H_{e,b}} (D_e^d f)(y)dy \qquad a_d = \frac{(-1)^{\frac{d-1}{2}}}{2}(2\pi)^{1-d}$$

$\omega_f(e,b)$ is orthogonal flow of order d through hyperplane $H_{e,b} = \{x \in \mathbb{R}^d, e \cdot x + b = 0\}$
a_d is exponentially decreasing

$$\|f\|_{H_d,\sup} \leq \int_{S^{d-1}} \int_{\mathbb{R}} |\omega_f(e,b)|dedb = \|\omega_f\|_{\mathcal{L}_1(S^{d-1}\times\mathbb{R})}$$

Fast rates of approximate minimization of empirical error over perceptron networks

If a sample of data z determining the empirical error $\mathcal{E}_z$ can be interpolated by a function h with the Sobolev seminorm $\|h\|_{d,1,\infty} \leq \sqrt{d}\, 2^{d/2}$, then

$$\min_{f \in \operatorname{span}_n H_d} \mathcal{E}_z(f) \leq \frac{1}{n}$$

Fast rates $\frac{1}{n\,m}$ of approximate minimization of empirical error $\mathcal{E}_z$ over Heaviside perceptron networks are guaranteed for samples of data z that can be interpolated by functions with quite large Sobolev seminorms (bounded from above by $\sqrt{d}\, 2^{d/2}$)

Example: Samples chosen from the Gaussian function

z sample chosen from the Gaussian function

$$\gamma(x) = e^{-\|x\|^2} : \mathbb{R}^d \to \mathbb{R}$$

$$\|\gamma\|_{H_d} \leq 2d \quad \Rightarrow \quad \min_{f \in span_n H_d} \mathcal{E}_z(f) \leq \frac{4d^2}{n}$$

relationship between two types of geometrically opposite units: perceptrons and radial-basis functions

Samples of data that cannot be interpolated by sufficiently smooth functions

every Boolean function $f : \{0,1\}^d \to \mathbb{R}$ determines a sample $z = \{(u_i, v_i) | i = 1, \dots, 2^d\}$ defined as $\{u_1, \dots, u_{2^d}\} = \{0,1\}^d$ and $v_i = f(u_i)$

for every function $h : X \to \mathbb{R}$ interpolating data z

$$\|f\|_{H_d(\{0,1\}^d)} \leq \|h\|_{H_d(X)}$$

$\Rightarrow$ a lower bound on variation wrt half-spaces of the Boolean function f is also a lower bound on variation of every function h interpolating the data z defined by f

$\Rightarrow$ we can use lower bounds on variations of Boolean functions

Functions with variations wrt half-spaces depending on the number of variables d exponentially

$\operatorname{card} H_d(\{0,1\}^d) < 2^{d^2}$

BUT

$\dim_\varepsilon 2^d$ is large $\quad \dim_\varepsilon 2^d = e^{\frac{2^d \varepsilon^2}{2}}$

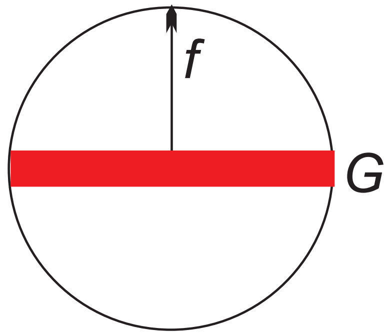
$\Rightarrow$ there exist functions with variations wrt half-spaces depending on the number of variables d exponentially

Example:
inner product modulo 2 has $H_d(\{0,1\}^d)$-variation at least $\mathcal{O}(2^{d/6})$

Geometric characterization of G-variation

Kůrková, Savický, Hlaváčková 98

$$\|f\|_G \geq \frac{\|f\|^2}{\sup_{g \in G} |f \cdot g|}$$



functions that are "almost orthogonal" to G have large G-variation

Hahn-Banach Theorem

Functions with large variation and covering numbers

$S_1 = S_1(\|\cdot\|)$ unit sphere in a Hilbert space $(X, \|\cdot\|)$
μ_X pseudometrics on S_1

$$\mu_X\ (f,g) = \arccos |f \cdot g|$$

minimum of two angles: between f and g and between f and $-g$

$\alpha > 0$ $\quad \mathcal{N}_\alpha(S_1)$ $\quad \alpha$-covering number of S_1 with respect to μ_X
(smallest number of balls of radius α covering S_1)

---

if $\operatorname{card} \mathrm{G} < \mathcal{N}_\alpha(\mathrm{S}_1) \Rightarrow$
S_1 contains a function with G-variation greater than $\frac{1}{\cos\alpha}$

Set of characteristic functions of Boolean half-spaces is small wrt covering numbers of S^{2^d-1}

samples of data z represented by Boolean functions

$\{f : \{0,1\}^d \to \mathbb{R}\} = \mathbb{R}^{2^d}$

hypothesis set $=$ set of characteristic functions of half-spaces of the Boolean cube $H_d(\{0,1\}^d)$

$\mathrm{card}\, H_d(\{0,1\}^d)$ is small

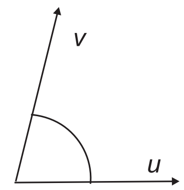
Shläfli $\quad \mathrm{card}\, H_d(\{0,1\}^d) = 2^{d^2 - d\log_2 d + \mathcal{O}(d)} < 2^{d^2}$ as $d \to \infty$

Quasiorthogonal dimension of Euclidean spaces

$\varepsilon > 0 \qquad u, v \in \mathbb{R}^m$

(u, v) are ε-quasiorthogonal if

$|u \cdot v| \leq \varepsilon \|u\| \, \|v\|$



$\alpha = \arccos \varepsilon$

$$\dim_\varepsilon m$$

= maximal number of pairwise ε-quasiorthogonal vectors

$\dim_\varepsilon m$ is large $\Rightarrow$ $\mathcal{N}_{\arccos \varepsilon}(S^{m-1})$ is large

$G \subseteq S^{m-1} \subseteq \mathbb{R}^m \qquad \mathrm{card\,G} \leq \mathrm{dim_\varepsilon\,m} \Rightarrow \exists \mathrm{f} \in \mathrm{S^{m-1}} \qquad \|f\|_G \geq \frac{1}{\varepsilon}$

Kainen, Kůrková 93 $\qquad \dim_\varepsilon m \geq e^{\frac{m\varepsilon^2}{2}}$ as $\varepsilon \to 0$ and $m \to \infty$