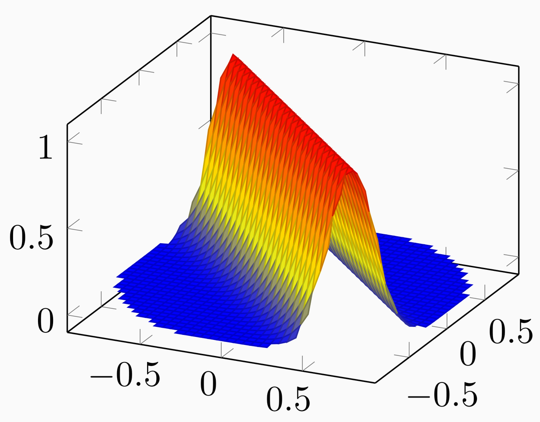
Pathological artifical neurons

Sebastian Mayer - mayer@ins.uni-bonn.de

Institute for Numerical Simulation and Fraunhofer SCAI

Introduction

- An artifical neuron (ridge function) is a multivariate function
$$f(x) = g(a^T x), \quad x \in \mathbb{R}^d$$
that varies only along one direction in space (given by the weight vector a).
- Ridge functions are the building blocks of artifical neural networks and projection pursuit algorithms.



Example ridge function in $d = 2$

- This poster considers fundamental complexity-theoretical limitations for learning *one* artifical neuron in the uniform norm L_∞. We ask what a-priori knowledge is necessary to guarantee a certain degree of tractablility for the learning problem when the number of weights d becomes large.
- Though it is nowadays possible to learn huge networks in many practical situations, we find circumstances where learning one simple artifical neuron suffers from the curse of dimensionality.

Worst-case information complexity

- For given function class F_d, the worst-case learning error is defined by
$$\mathrm{error}(n, F_d) = \inf_{A \text{ adaptive algorithm}} \sup\{\|f - A(f)\|_\infty : \mathrm{cost}(A) \leq n\}.$$
Assume $\mathrm{cost}(A) = \sharp$ used function samples (cost to obtain labels dominate computational cost).
- The worst-case information complexity is given by
$$n(\varepsilon, F_d) = \min\{n \in \mathbb{N} : \mathrm{error}(n, F_d) \leq \varepsilon\}.$$

Classes of artificial neurons (a-priori knowledge)

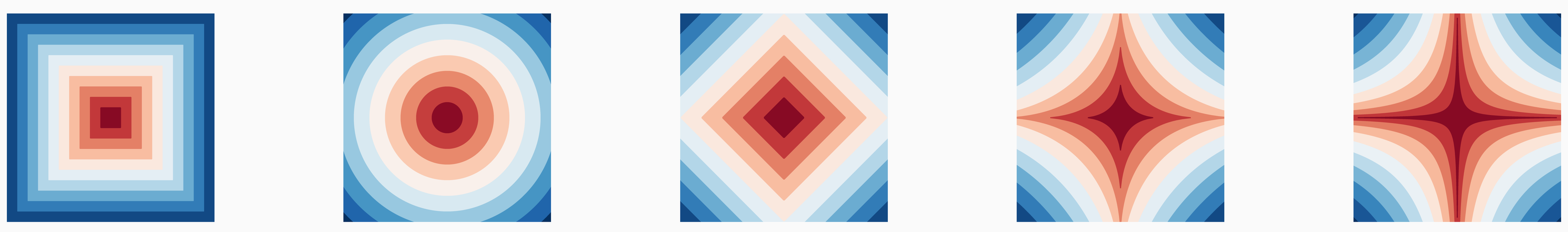
For $d \in \mathbb{N}$, let D be the d-dimensional Euclidean ball, $D = B_2^d$, or the d-dimensional cube, $D = [-1,1]^d$. Let $p_{B_2^d} = 2$, $p_{[-1,1]^d} = 1$. We consider classes F_d of artifical neurons
$$f(x) = g(a^T x), \quad x \in D,$$
with
$$\text{activation function } g : [-1,1] \to \mathbb{R}, \quad \text{weight vector } a \in B_{p_D}^d,$$
fulfilling some of the following conditions.

G1 The activation g has Lipschitz regularity $r > 0$ and
$$\|g\|_{\mathrm{Lip}(r)} = \max\{\|g\|_\infty, \|g^{(1)}\|_\infty, \dots, \|g^{(k)}\|_\infty, |g^{(k)}|_{r-k}\} \leq 1.$$

G2 The activation g is nondegenerate: $|g'(0)| \geq c$ for absolute $c > 0$.

A1 (relative) compressibility: $\|a\|_p \leq 1$ for $0 < p \leq p_D$.

A2 approximate sparsity: for $0 < p \leq p_D$ and $S \in [d]$ fixed
$$\|a\|_p \leq 1 \text{ and } \|a\|_{p_D} \geq S^{1/p_D - 1/p}.$$



$p = \infty$ $p = 2$ $p = 1$ $p = 0.5$ $p = 0.1$

The easy case: nondegenerate activations [3]

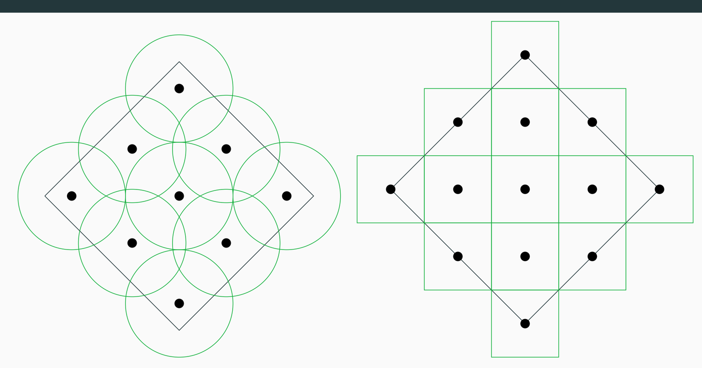
- Simple three-step procedure:
  1. Search for $x_0 \in D$ such that $|g'(a^T x_0)|$ large.
  2. Approximate gradient $\tilde{a} \approx \nabla f(x_0)$ to obtain estimate $\hat{a} = \tilde{a}/\|\tilde{a}\|_{p_D}$ of the ridge direction.
  3. Sample along $\hat{a}$ and approximate g with a spline.
- only works if g is nondegenerate (Assumption G3) or, in case $D = [-1,1]^d$, if $\mathrm{sign}(a_1), \dots, \mathrm{sign}(a_d)$ are known [1].
- If we assume G3, then the simple three-step procedure is optimal for $n \geq d + 1$. We have
$$\mathrm{error}(n, F_d) = O(1) \quad \text{for } 1 \leq n \leq d+1$$
and
$$c_r n^{-r} \leq \mathrm{error}(n, F_d) \leq C_r (n-d)^{-r} \quad \text{for } n \geq d+1.$$

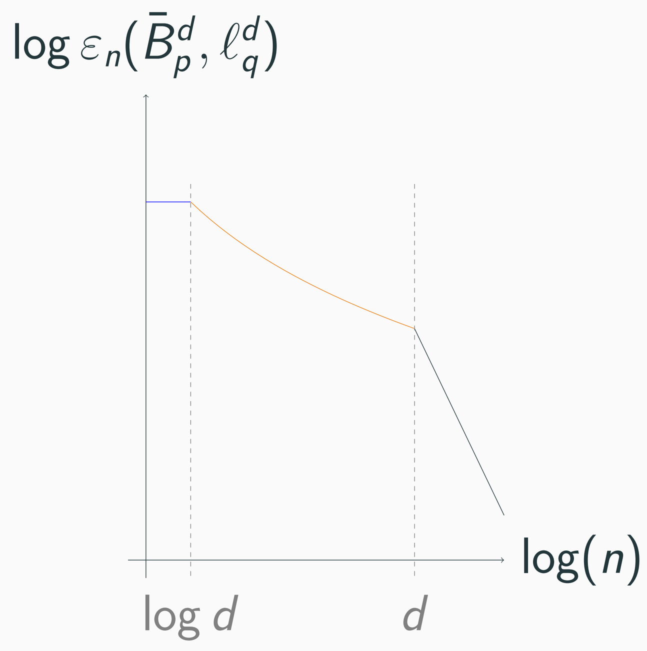
Acknowledgments

The results presented here are based on joint work with Benjamin Doerr (École Polytechnique), Danie Rudolf (Jena), Tino Ullrich (Bonn), and Jan Vybíral (Prag).

Metric entropy

Let $0 < p \leq q \leq \infty$. The entropy number $\varepsilon_n(B_p^d, \ell_q^d)$ is the minimial radius $\varepsilon > 0$ such that the unit ball B_p^d can be covered by n ℓ_q^d-balls of radius ε.





A by now classical result in approximation theory:
$$\varepsilon_n(B_p^d, \ell_q^d) \asymp \begin{cases} 1 & : 1 \leq n \leq d, \\ \left(\frac{\log(ed/\log(n))}{\log(n)}\right)^{1/p - 1/q} & : d \leq n \leq 2^d, \\ d^{1/q - 1/p} n^{-1/d} & : n \geq 2^d. \end{cases}$$
Schütt (1984), Kühn (2001), Edmunds and Triebel (1996).

Two-sided worst-case error bounds [2, 3]

- Consider $D = B_2^d$ and assume G1 and A1. Then, for all $r > 0$ and all $0 < p \leq 2$,
$$c_{r,p} \varepsilon_n(B_p^d, \ell_2^d)^{2r} \leq \mathrm{error}(n, F_d) \leq C_{r,p} \varepsilon_{n/\binom{d+s}{s}}(B_2^d, \ell_{p'}^d)^r$$
(p': dual index of p).
- Consider $D = [-1,1]^d$. Assume G1 and A1. Let $r > 0$ and $0 < p \leq 1$. Then, for $1 \leq n \leq 2^d$,
$$\mathrm{error}(n, F_d) \geq c_{r,p} \begin{cases} 1 & : 1 \leq n \leq d \\ \left(\frac{\log(ed/\log(n))}{\log(n)}\right)^{r(1/p - 1)} & : d \leq n \leq 2^d \end{cases}$$
If we allow ranodmized algorithms, the $\log(ed/\log(n))$-term disappears in the bound.
- Consider $D = [-1,1]^d$. Assume G1 and A2. Let $r > 0$ and $0 < p \leq 1$. Then, for $1 \leq n \leq 2^d$,
$$\mathrm{error}(n, F_d) \leq c_{r,p} \begin{cases} 1 & : 1 \leq n \leq d \\ \left(\frac{1}{\log(n)}\right)^{r(1/p - 1)} & : d \leq n \leq 2^d d \\ 2^{rd} n{-}r & : n \geq 2^d d. \end{cases}$$
The result holds both for deterministic and randomized algorithms.

Results on tractability

Tractability measures to what degree exponential dependencies are absent.

- Polynomial tractability (PT) if there exist constants $C, p, q > 0$ such that $n(\varepsilon, F_d) \leq C\,(1/\varepsilon)^p\, d^q$ for all $0 < \varepsilon < 1$ and all $d \in \mathbb{N}$;
- Quasi-polynomial tractability (QPT) if there exist constants $C, p, q > 0$ such that $n(\varepsilon, F_d) \leq C\,(1/\varepsilon)^{p(1+\log d)}\, d^q$ for all $0 < \varepsilon < 1$ and all $d \in \mathbb{N}$;
- Uniform weak tractability (UWT) if for all $\alpha, \beta > 0$ $\lim_{1/\varepsilon + d \to \infty} \frac{\log n(\varepsilon, F_d)}{1/\varepsilon^\alpha + d^\beta} = 0$;
- Weak tractability (WT) if $\lim_{1/\varepsilon + d \to \infty} \frac{\log n(\varepsilon, F_d)}{1/\varepsilon + d} = 0$;
- Intractability = not WT;
- The curse of dimensionality (CURSE): if there are $\varepsilon_0, c, \gamma > 0$ such that $n(\varepsilon, F_d) \geq c(1+\gamma)^d$ for all $0 < \varepsilon \leq \varepsilon_0$ and infinitely many $d \in \mathbb{N}$.

	ball	cube
CURSE	$p = 2$	$p = 1$
WT	$r > \left(\frac{1}{\max\{1,p\}} - \frac{1}{2}\right)^{-1}$	$r > \left(\frac{1}{p} - 1\right)^{-1}$
(almost) UWT	strong compressibility ($p \ll 1$)	
QPT	$r = \infty$	?
PT	$\lvert g'(0)\rvert > c$	sparsity, $\lvert g'(0)\rvert > c$, $a \geq 0$ [1]

References

- Albert Cohen, Ingrid Daubechies, Ronald DeVore, Gerard Kerkyacharian, and Dominique Picard, *Capturing ridge functions in high dimensions from point queries*, Constructive Approximation **35** (2012), no. 2, 225–243.
- Benjamin Doerr, Sebastian Mayer, and Daniel Rudolf, *Tractability of recovering ridge functions on the cube*, work in progress.
- Sebastian Mayer, Tino Ullrich, and Jan Vybiral, *Entropy and sampling numbers of classes of ridge functions*, Constructive Approximation **42** (2015), no. 2, 231–264.

mayer@ins.uni-bonn.de, sebastian.mayer@scai.fraunhofer.de