Introduction

- An artificial 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 artificial neural networks and projection pursuit algorithms.
- This poster considers fundamental complexity-theoretical limitations for learning one artificial neuron in the uniform norm \( L_\infty \). We ask what a-priori knowledge is necessary to guarantee a certain degree of tractability 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 artificial 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
  \[ \text{error}(n, F_d) = \inf_{A \text{ adaptive algorithm}} \sup_{f \in F_d} \| f - A(f) \|_\infty : \text{cost}(A) \leq n. \]
  Assume \( \text{cost}(A) = \| A \| \) is used for function samples (cost to obtain labels dominates computational cost).
- The worst-case information complexity is given by
  \[ n(e, F_d) = \min \{ n \in \mathbb{N} : \text{error}(n, F_d) \leq e \}. \]

Classes of artificial neurons (a-priori knowledge)

For \( d \in \mathbb{N} \), let \( D \) be the \( d \)-dimensional Euclidean ball, \( B^d \), or the \( d \)-dimensional cube, \( D = [-1, 1]^d \). Let \( p_{B^d} = 2 \), \( p_{[1,1]^d} = 1 \). We consider classes \( F_d \) of artificial neurons
  \[ f(x) = g(a^T x), \quad x \in D, \]
with
- activation function \( g : [-1, 1] \rightarrow \mathbb{R} \)
- weight vector \( a \in B^d \)

G1 The activation \( g \) has Lipschitz regularity \( r > 0 \) and
  \[ \| g \|_{Lip(r)} = \max \{ ||g||_{\infty}, \|g^{(1)}\|_{\infty}, \ldots, \|g^{(r)}\|_{\infty} \} \leq 1. \]
G2 The activation \( g \) is nondegenerate: \( |g(0)| \geq c \) for absolute \( c > 0 \).
A1 (relative) compressibility: \( \|a\|_2 \leq 1 \) for all \( 0 < p \leq p_0 \).
A2 (approximate) sparsity: for \( 0 < p \leq p_0 \) and \( S \in [d] \) fixed
  \[ \|a|_p \leq 1 \quad \text{and} \quad \|a|_\infty \leq \frac{1}{\sqrt{n/d-p}}. \]

The easy case: nondegenerate activations [3]

- Simple three-step procedure:
  1. Search for \( x_0 \in D \) such that \( g(x_0) \) is large.
  2. Approximate gradient \( \hat{g} = \nabla g(x_0) \) to obtain estimate \( \hat{a} = \hat{g} / \|\hat{g}\|_\infty \) 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 \( \text{sign}(a_1), \ldots, \text{sign}(a_d) \) are known [1].
- If we assume G3, then the simple three-step procedure is optimal for \( n \geq d + 1 \).
We have
  \[ \text{error}(n, F_d) = O(1) \quad \text{for} \ 1 \leq n \leq d + 1 \]
and
  \[ c_n^{-2} \leq \text{error}(n, F_d) \leq C_n(d) n^{-2} \quad \text{for} \ n \geq d + 1. \]

Acknowledgments

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

Metric entropy

Let \( 0 < p \leq q \leq \infty \). The entropy number \( \varepsilon_n(B^d, \ell^p) \) is the minimal \( \varepsilon > 0 \) such that the unit ball \( B^d \) can be covered by \( n \varepsilon^p \) balls of radius \( \varepsilon \).

The activation \( g \) is used in the construction of a ball of radius \( \varepsilon \) in \( \mathbb{R}^d \). The log of the number of balls needed to cover the ball is equal to \( \varepsilon_n(B^d, \ell^p) \).

Two-sided error bounds [2, 3]

- Consider \( D = B^d \) and assume G1 and A1. Then, for all \( r > 0 \) and \( 0 < p \leq 2 \),
  \[ c_n \varepsilon^p \varepsilon_n(B^d, \ell^p) \leq \text{error}(n, F_d) \leq c_n \varepsilon^p \varepsilon_n(B^d, \ell^p)^2 \]
  \( (\varepsilon^p): \) dual index of \( p \).
- Consider \( D = [-1, 1]^d \). Assume G1 and A1. Let \( r > 0 \) and \( 0 < p \leq 1 \). Then, for \( 1 \leq n \leq 2^d \),
  \[ \text{error}(n, F_d) \leq c_n \varepsilon^p \varepsilon_n(B^d, \ell^p)^2 \]
  \( (\varepsilon^p): \) dual index of \( p \).
If we allow randomized algorithms, the log(\text{error}/\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 \( 1 \leq n \leq 2^d \),
  \[ \text{error}(n, F_d) \leq c_n \varepsilon^p \varepsilon_n(B^d, \ell^p)^2 \]
  \( (\varepsilon^p): \) dual index of \( p \).

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
  \[ n(e, F_d) \leq C (1/e)^d \]
  for all \( 0 < e < 1 \) and all \( d \in \mathbb{N} \).
- Quasi-polynomial tractability (QPT) if there exist constants \( C, p, q > 0 \) such that
  \[ n(e, F_d) \leq C (1/e)^{\log_2(1/e)} \]
  for all \( 0 < e < 1 \) and all \( d \in \mathbb{N} \).
- Uniform weak tractability (UWT) if for all \( \alpha, \beta > 0 \)
  \[ \lim_{d \rightarrow \infty} \log_2 C_n(d) \]
  \( (\alpha) \)
- Weak tractability (WT) if
  \[ \lim_{d \rightarrow \infty} \log_2 C_n(d) \]
  \( (\alpha) \)
- Intractability is not WT.
- The curse of dimensionality (CURSE): if there are \( c_0, c, \gamma > 0 \) such that
  \[ n(e, F_d) \geq c_0 (1+\gamma)^d \]
  for all \( 0 < e \leq c_0 \) and infinitely many \( d \in \mathbb{N} \).

References

- Benjamin Doerr, Sebastian Mayer, and Daniel Rudolf, Tractability of recovering ridge functions on the cube, work in progress.

http://ullrich.ins.uni-bonn.de/people/mayer.html
mayer@ins.uni-bonn.de, sebastian.mayer@scai.fraunhofer.de