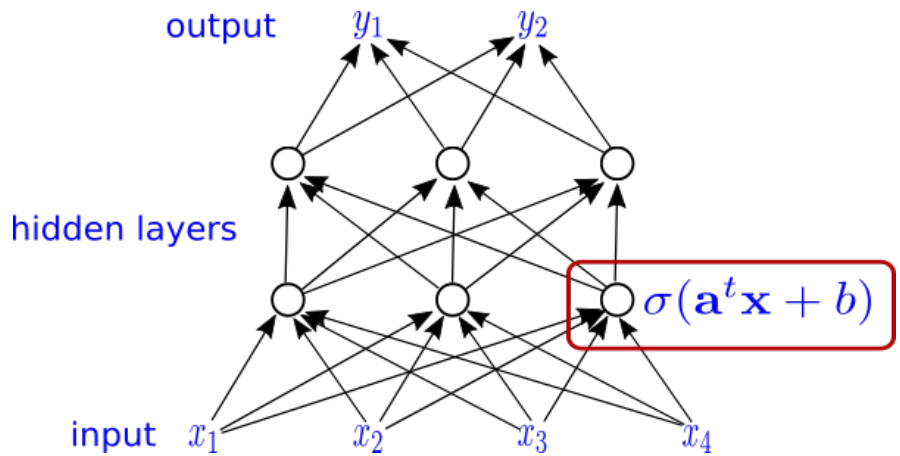
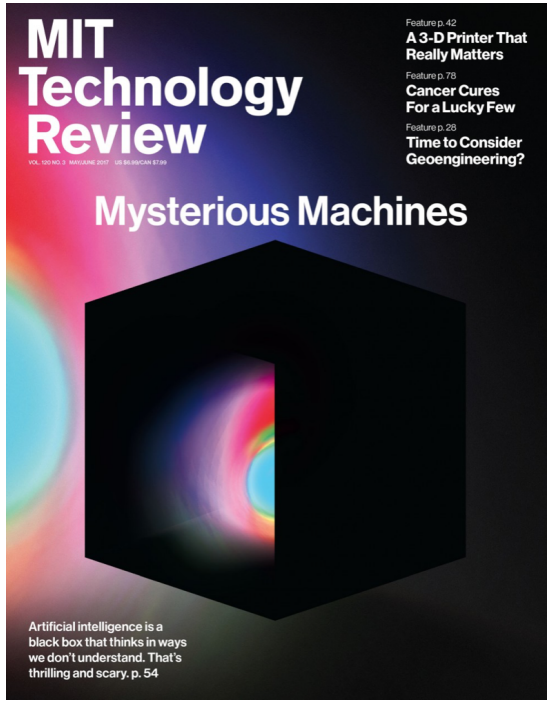
Towards a statistical foundation of deep learning



Johannes Schmidt-Hieber

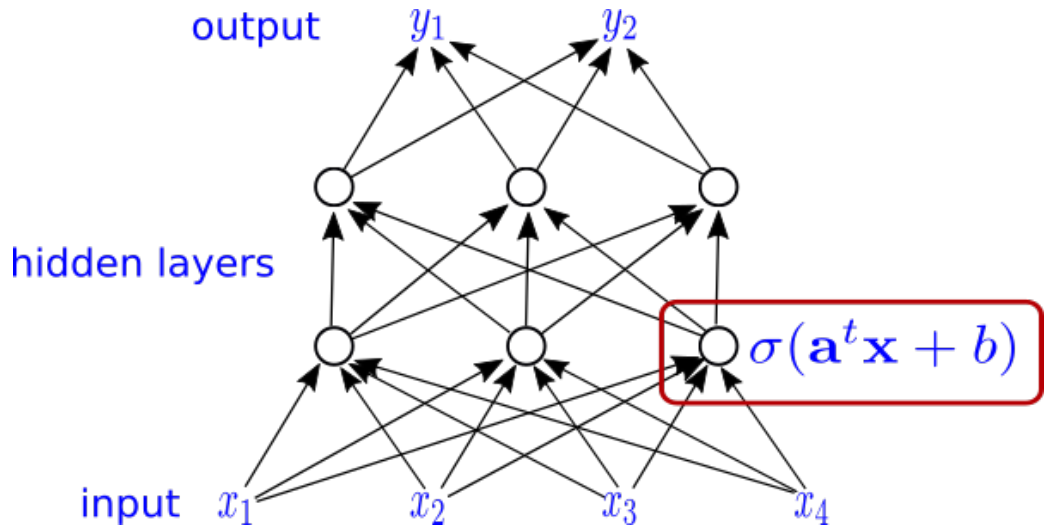
Challenges for a mathematical foundation of deep learning

- parameter selection
- theoretical guarantees
- understanding how signal is processed



Answers to these questions relate to statistical problems.

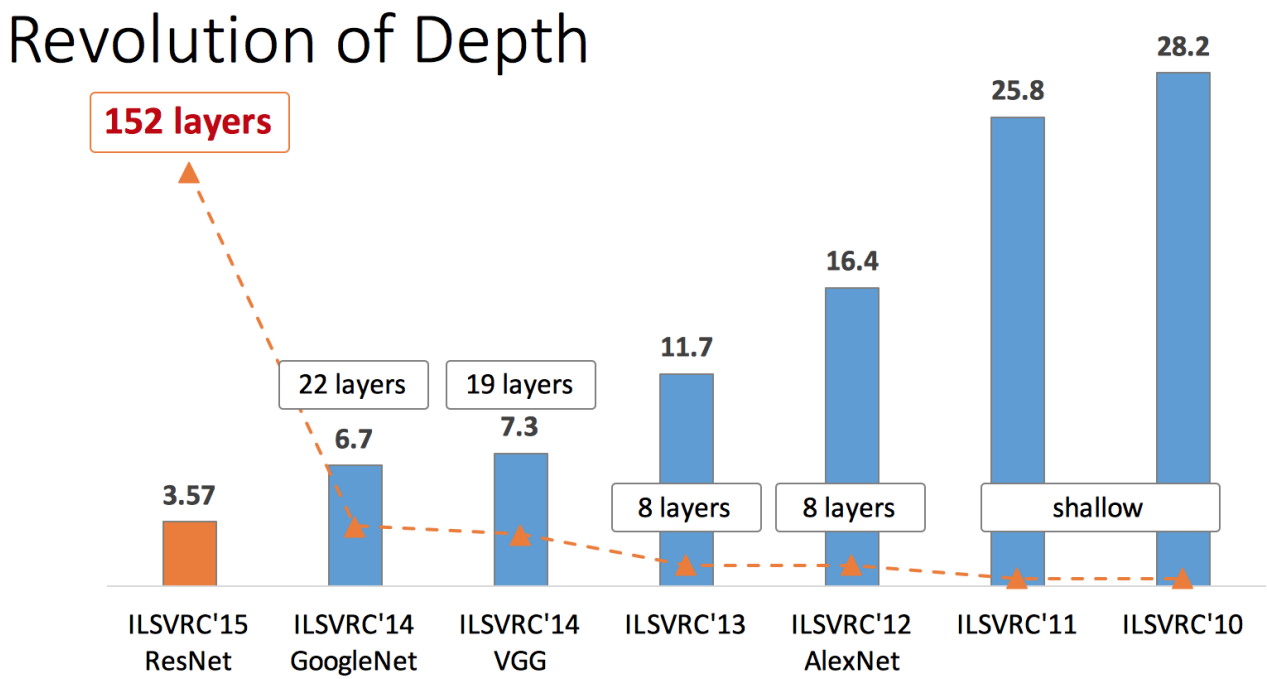
neural networks



$$f(\mathbf{x}) = W_L \sigma W_{L-1} \ldots \sigma W_1 \sigma W_0 \mathbf{x}$$

- $\sigma(x) = \max(x, 0)$ is the ReLU activation function
- L is the network depth or number of hidden layers
- $L = 1$ shallow, $L > 1$ deep
- matrices W_i are the free parameters

depth



Source: Kaiming He, Deep Residual Networks

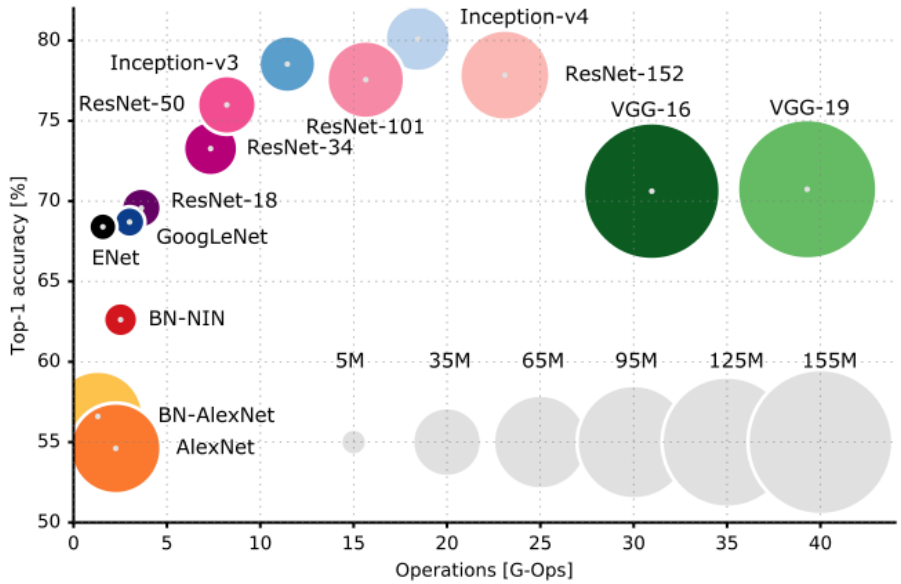
- Networks are deep
  - version of ResNet with 152 hidden layers
  - networks become deeper

why deep?

- for $L > 1$, we can localize
- Kolmogorov-Arnold representation theorem (Braun '09): Fix $d \geq 2$. There are real numbers a, b_p, c_q and a continuous and monotone function $\psi : \mathbb{R} \to \mathbb{R}$, such that for any continuous function $f : [0,1]^d \to \mathbb{R}$, there exists a continuous function $g : \mathbb{R} \to \mathbb{R}$ with

$$f(x_1, \ldots, x_d) = \sum_{q=0}^{2d} g\Big(\sum_{p=1}^{d} b_p \psi(x_p + qa) + c_q \Big).$$

high-dimensionality



Source: arxiv.org/pdf/1605.07678.pdf

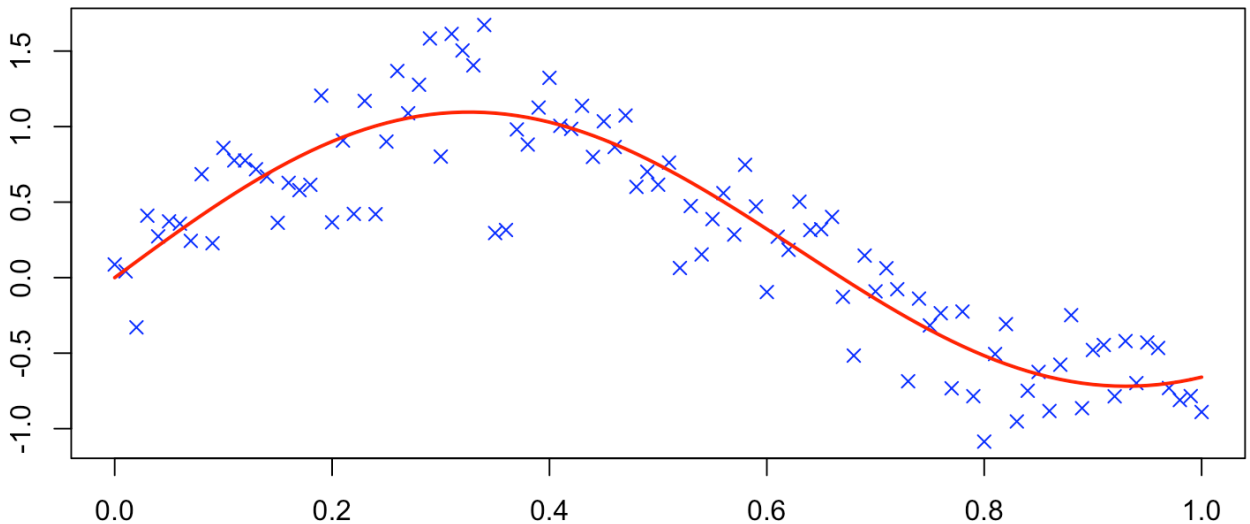
- Number of network parameters is larger than sample size
- AlexNet uses 60 million parameters for 1.2 million training samples

classification and regression

X = {images}

f : X —> Y

Y = {“chihuahua”,“muffin”}



Mathematical problem: Given n data points, how well can a given machine learning method approximate the unknown function f?

Our setup:

- study regression only
- data are n independent copies $(\mathbf{X}_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$

$$Y_i = f(\mathbf{X}_i) + \text{noise}$$

- assume that networks are sparsely connected
- sparsity s is chosen

loss and learning

- ideally we want to find a network function f minimizing

$$\sum_{i=1}^{n} \big(Y_i - f(\mathbf{X}_i)\big)^2$$

- this is the cross-entropy loss in regression
- finding a global minimum is computationally infeasible
- therefore (stochastic) gradient descent is employed
- for a given method $\widehat{f}_n$ returning a s-sparse network, we define the quantity

$$\Delta_n := E\Big[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{f}_n(\mathbf{X}_i))^2 - \inf_{f\in\mathcal{F}(s)} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i))^2\Big]$$

with $\mathcal{F}(s)$ the class of all s-sparse network functions

hierarchical structure

- curse of dimensionality: no method exists that can reconstruct regression function in high dimensions well
- deep learning outperforms other methods only for complex problems
- want to identify setups, where deep learning is better

- only few objects are combined on deeper abstraction level
  - few letters in one word
  - few words in one sentence
- generalizes many structural constraints

design supported on unknown manifold

- as before,

$$Y_i = f(\mathbf{X}_i) + \text{noise}$$

  but now we also suppose that $\mathbf{X}_i \in \mathbb{R}^d$ lies on unknown d^*-dimensional manifold

- if ψ_j denote the local coordinate maps, for any $\mathbf{x}$ on the manifold,

$$\sum_{j=1}^{K} \underbrace{\tau_j(\mathbf{x})}_{\text{partition of unity}} \cdot \underbrace{\left(f \circ \psi_j^{-1}\right)}_{\text{function on } \mathbb{R}^{d^*}} \circ \psi_j(\mathbf{x}) = f(\mathbf{x})$$

rate of convergence

Let $(\phi_n)_n$ be a sequence determined by properties of the function class.

Theorem: If

(i) depth $\asymp \log n$

(ii) width $\geq$ network sparsity $\asymp n\phi_n \log n$

Then, for any network reconstruction method $\widehat{f}_n$,

$$\textit{squared prediction error of } \widehat{f}_n \asymp \phi_n + \Delta_n$$

(up to $\log n$-factors).

- potential number of parameters is large
- right level of sparsity is crucial
- reproduces recent approximation theoretic result as a corollary
- it can be proved that that wavelet methods perform much worse for the same data

on the proof

- neural networks are often referred to as black boxes
- no explicit formulae are available on how parameters depend on data
- for that reason theory seems to be hopeless

building on existing tools in mathematical statistics, we can determine the quality of the output without really understanding what happens in the interior

on the depth

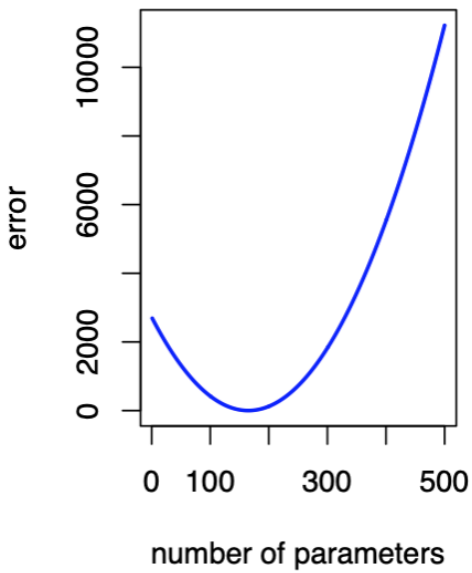
- $\log(n)$ scaling of the depth occurs naturally
- possibility to restrict to small parameters
- doubling property by adding layers
- learning a function representation
- learning composition structure in function

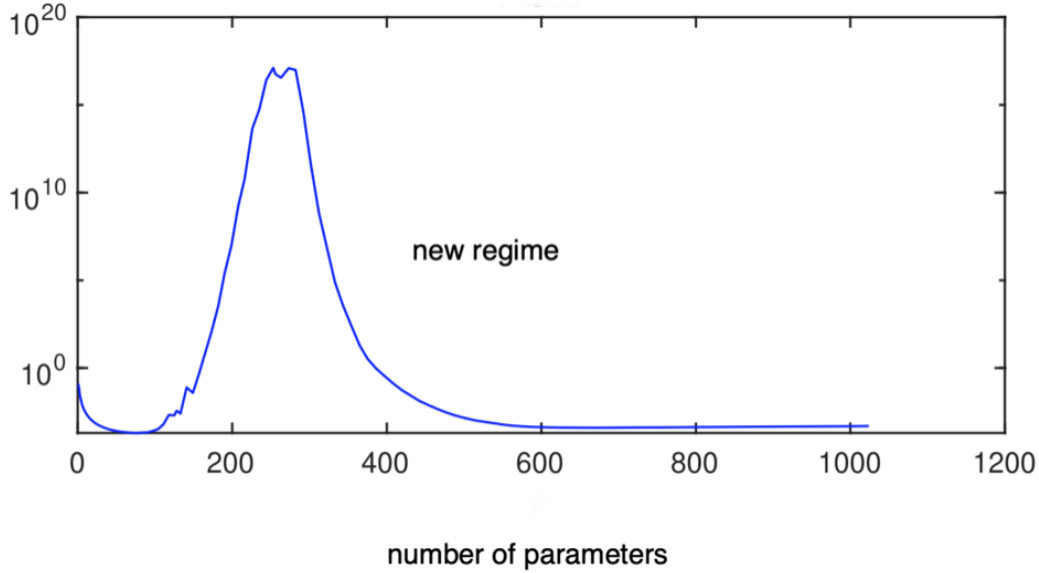
sparsely connected networks

Network sparsity is crucial in the proof but classical deep learning produces dense networks. Recently many new methods have been proposed generating sparsely connected networks.

- sparsifying as post-processing step $\rightsquigarrow$ compression
- starting with sparse network topology
- evolutionary methods inspired by human brain

double descent and implicit regularization





overparametrization generalizes well $\rightsquigarrow$ implicit regularization

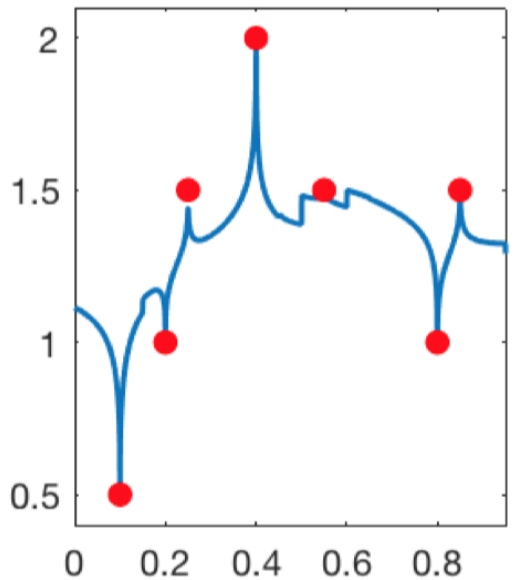
overfitting

- training error $= 0$ implies that $\Delta_n = 0$
- Δ_n does not fully characterize the statistical properties anymore
- because of implicit regularization, SGD will pick interpolant with good statistical properties

can implicit regularization avoid sparsity?

we conjecture the answer is **no** for the regression problem!

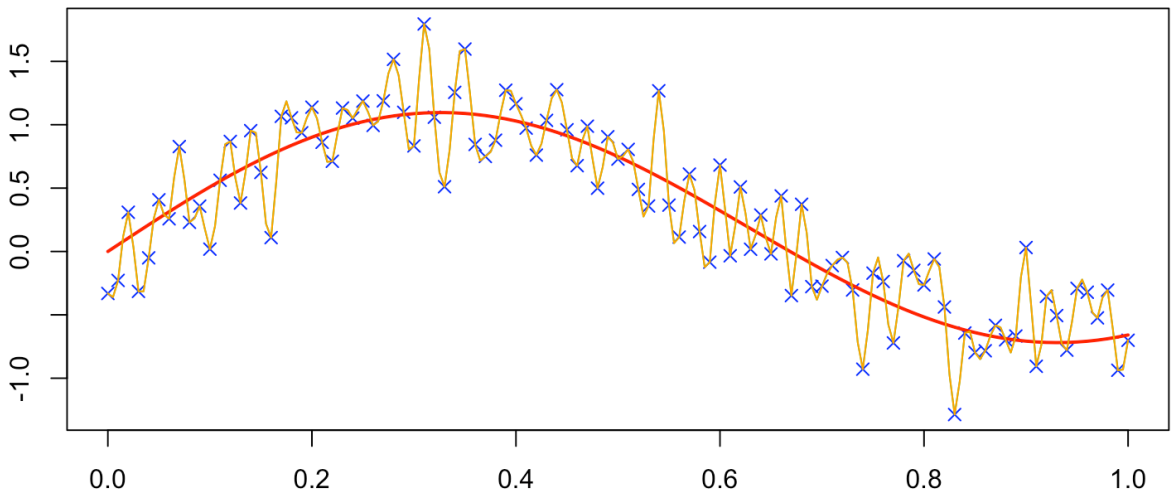
does data interpolation contradict statistical optimality?



Source: Belkin, Rakhlin, Tsybakov, 2018

in principle it is possible to interpolate and to denoise simultaneously

more details



we can show that for a simplified model and properly chosen learning rate, SGD converges to natural cubic spline interpolant

$\rightsquigarrow$ inconsistent estimator

main idea

A shallow network ($L = 1$) can be written as

$$x \mapsto \sum_{j=1}^{m} a_j (b_j x - c_j)_+, \quad a_j, b_j, c_j \in \mathbb{R}.$$

Taylor expansion in one dimension

$$g(x) = g(0) + x g'(0) + \int g''(u)(x-u)_+ du$$

If, say $g(0) = g'(0) = 0$, we have that approximately

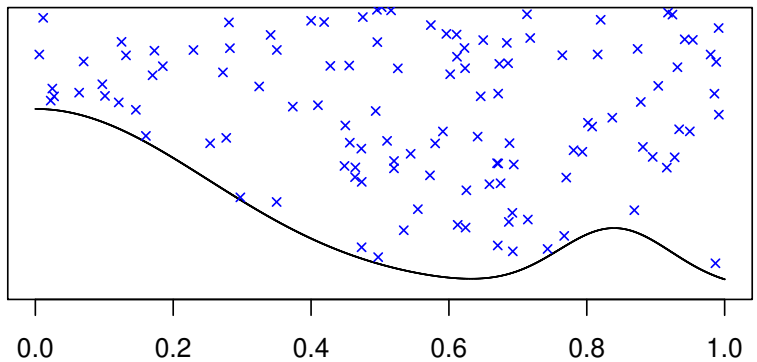
$$g(x) \approx \frac{1}{m} \sum_{j=1}^{m} g''\Big(\frac{j}{m}\Big)\Big(x - \frac{j}{m}\Big)_+.$$

denoising vs. interpolation

- implicit regularization is not sufficient to do denoising
- it still works in practice because standard datasets have a lot of structure in common (classification with few misclassified data points)

All statements that start with "In deep learning . . ." are wrong, what matters is the structure of the data. To describe for which data structures such claims are true is a major challenge for research in statistics.

outlook: binary classification with one class



- in practice we often face unbalanced datasets
- most extreme case is if we only sample from one class
- data are supposed to be correctly labeled
- by making assumptions on the design, it is still possible to come up with consistent classifiers
- problem is completely different than denoising, does deep learning work?

further challenges

- explainability
- energy landscape
- convolutional neural networks, autoencoders, . . .