Theoretical Issues in Deep Networks:

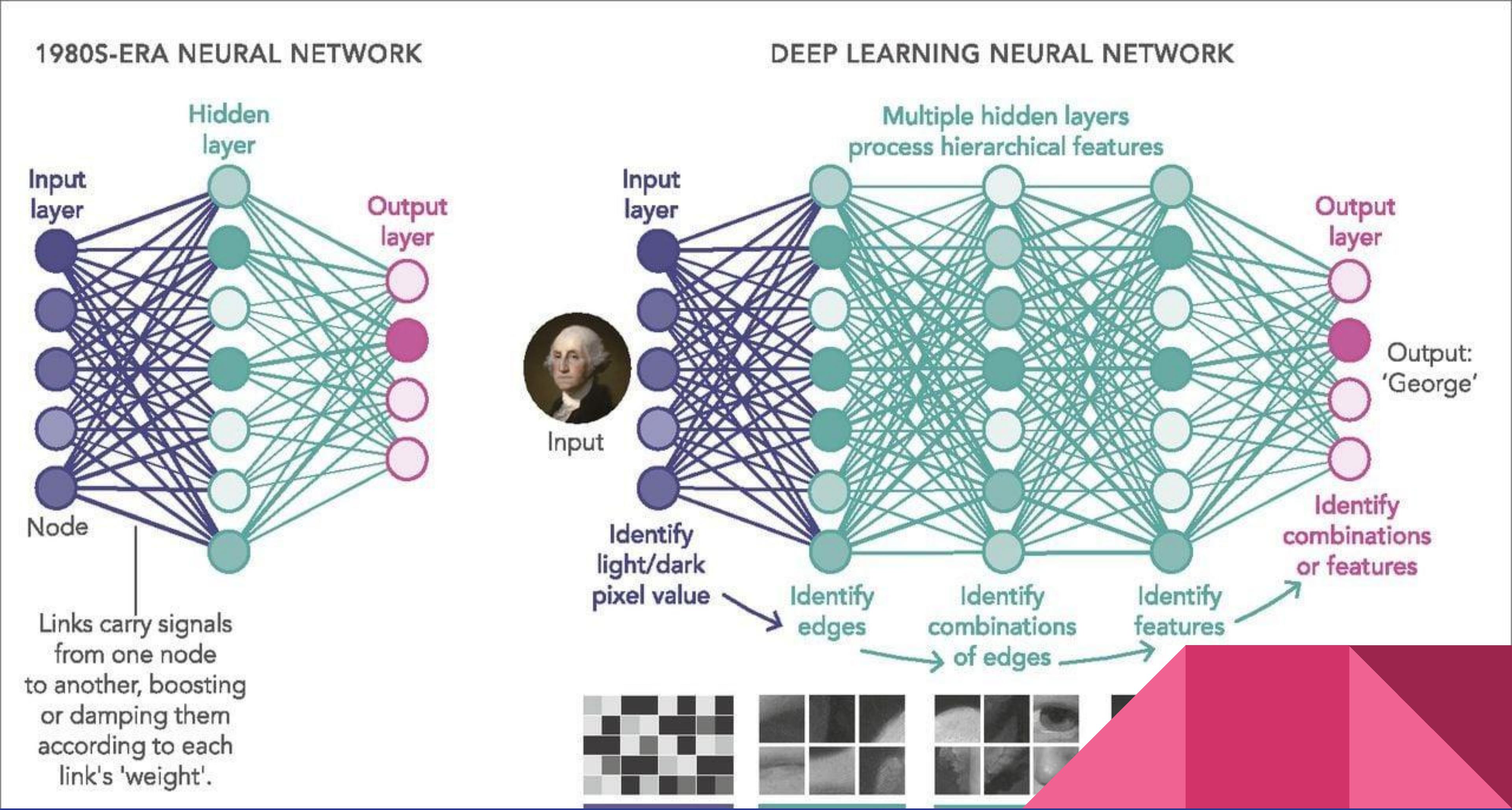
Approximation, Optimization and Generalization

Sergio Quiroga Sandoval
Luis Ángel Ballén Avellaneda
Manuel Fernando Valle Amortegui

UNIVERSIDAD
NACIONAL
DE COLOMBIA

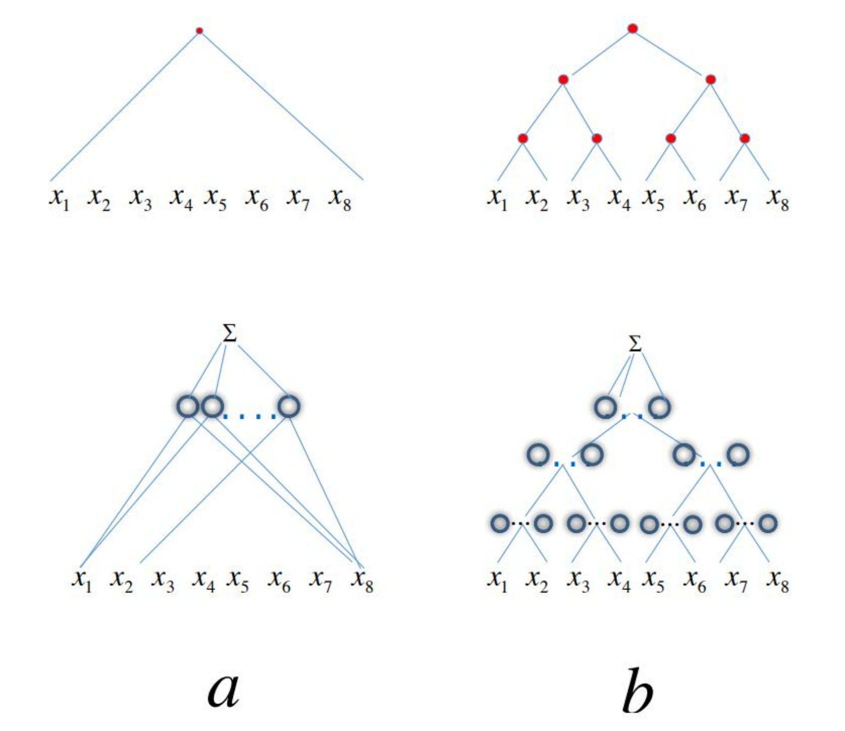
Contents

1. Representation power of deep networks.
2. Optimization of the empirical risk
3. Generalization properties of gradient descent techniques


1980S-ERA NEURAL NETWORK
Hidden layer
Input layer
Output layer
Node
Links carry signals from one node to another, boosting or damping them according to each link's 'weight'.
DEEP LEARNING NEURAL NETWORK
Multiple hidden layers process hierarchical features
Input layer
Input
Output layer
Output: 'George'
Identify light/dark pixel value
Identify edges
Identify combinations of edges
Identify features
Identify combinations or features

Retrieved from: https://doi.org/10.1073/pnas.1821594116

Shallow and deep networks



x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
Σ
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
Σ
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
a
b

The sequence of results is as follows:

- Both shallow and deep networks are universal.
- d is arbitrary but fixed and independent of t dimensionality n of the compositional function f.
- Same degree of accuracy but way fewer parameters for DN.

Degree of approximation

UNIVERSIDAD NACIONAL DE COLOMBIA

General paradigm: determining how complex a network ought to be to theoretically guarantee approximation of an unknown target function f up to a given accuracy $\epsilon > 0$.

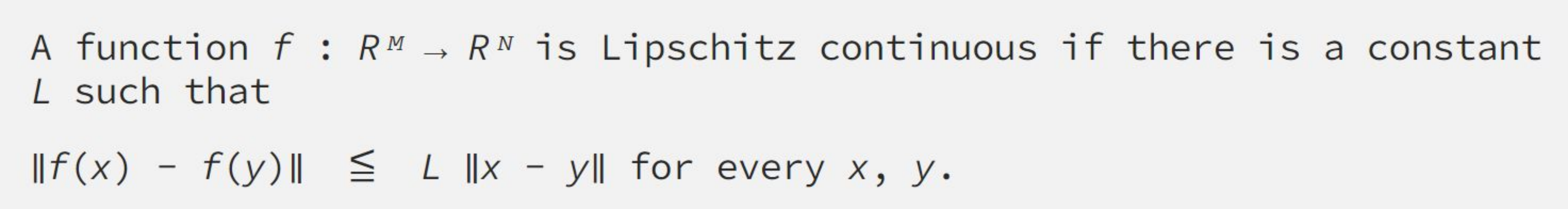
The degree of approximation is defined by

$$\text{dist}(f, V_N) = \inf_{P \in V_N} \|f - P\|$$

To measure we need a norm $| \cdot |$ on some normed linear space X.

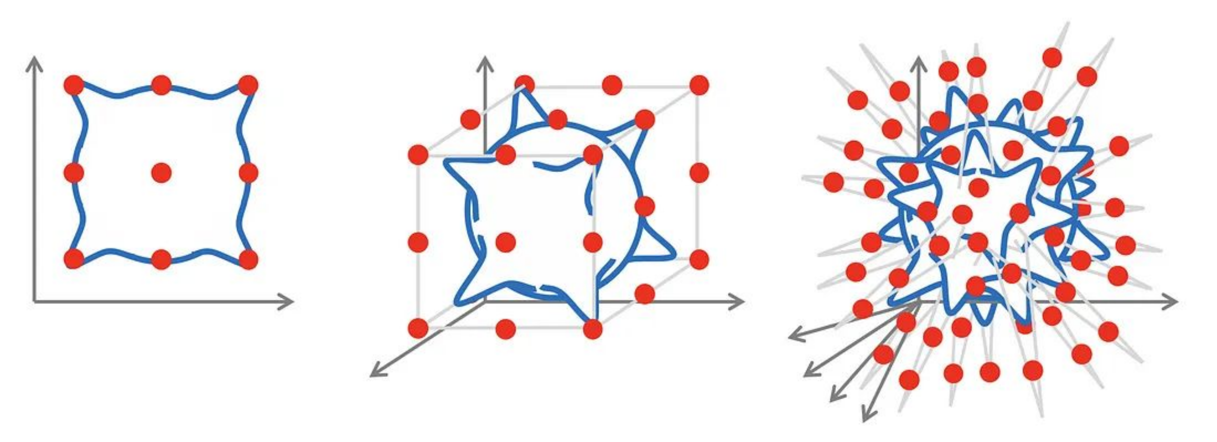
$$\text{dist}(f, V_N) = \mathcal{O}(N^{-\gamma}) \text{ for some } \gamma > 0$$

Curse of dimensionality

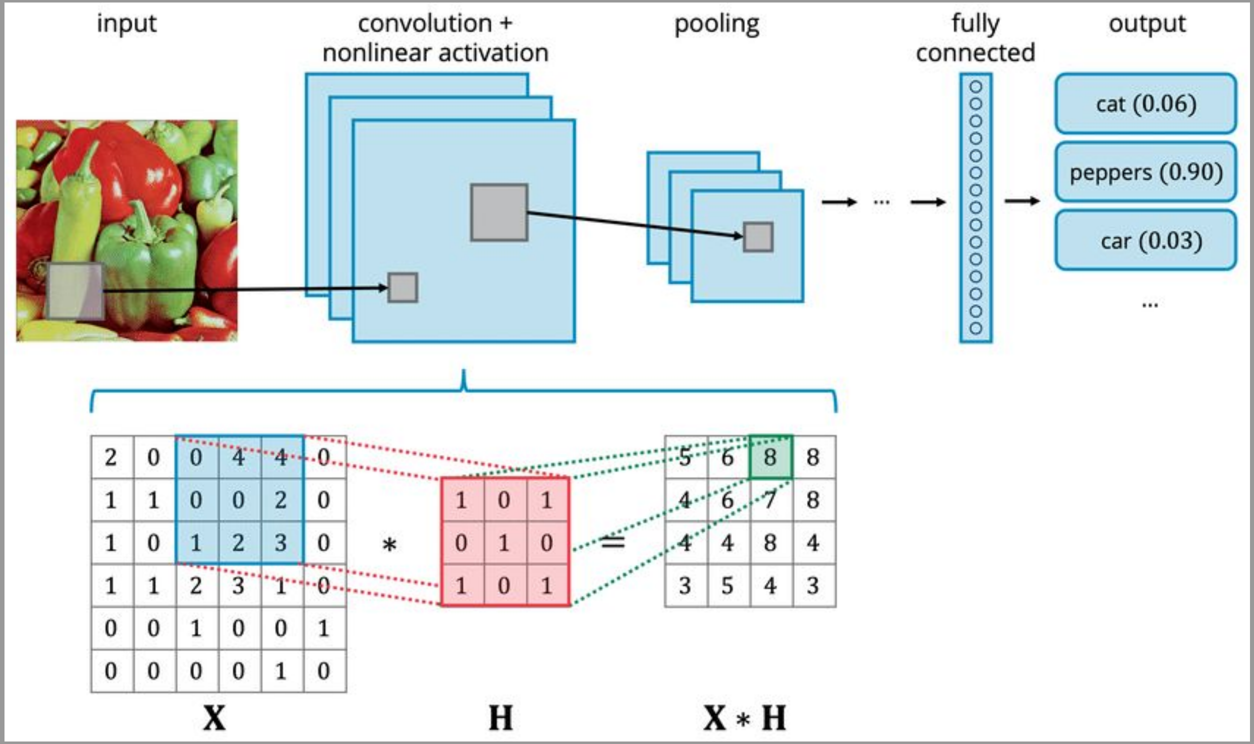


A function $f : R^M \rightarrow R^N$ is Lipschitz continuous if there is a constant L such that

$$\|f(x) - f(y)\| \leqq L \|x - y\| \text{ for every } x, y.$$



Glossary



$$\underbrace{\begin{array}{|c|c|c|c|c|c|}\hline 2 & 0 & 0 & 4 & 4 & 0 \\ \hline 1 & 1 & 0 & 0 & 2 & 0 \\ \hline 1 & 0 & 1 & 2 & 3 & 0 \\ \hline 1 & 1 & 2 & 3 & 1 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ \hline \end{array}}_{\mathbf{X}} * \underbrace{\begin{array}{|c|c|c|}\hline 1 & 0 & 1 \\ \hline 0 & 1 & 0 \\ \hline 1 & 0 & 1 \\ \hline \end{array}}_{\mathbf{H}} = \underbrace{\begin{array}{|c|c|c|c|}\hline 5 & 6 & 8 & 8 \\ \hline 4 & 6 & 7 & 8 \\ \hline 4 & 4 & 8 & 4 \\ \hline 3 & 5 & 4 & 3 \\ \hline \end{array}}_{\mathbf{X} * \mathbf{H}}$$

W_m^n $W_m^{n,\,2}$

$\mathcal{D}_{N,\,2}$ $\mathcal{S}_{N,\,n}$

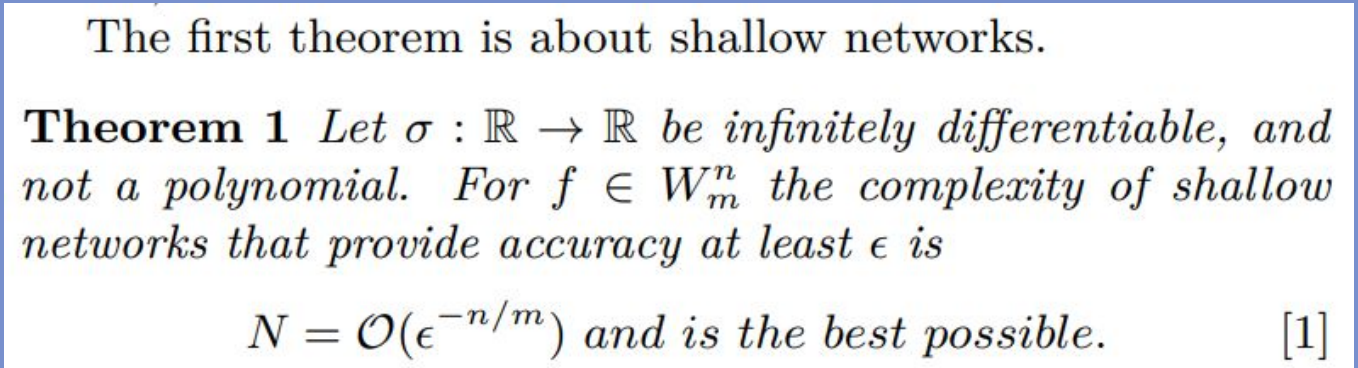
Compositional functions

Deep convolutional architectures have the theoretical guarantee that they can be much better than one layer architectures such as kernel machines for certain classes of problems

b) the problems for which certain deep networks are guaranteed to avoid the curse of dimensionality correspond to input-output mappings that are compositional with local constituent functions;

c) the key aspect of convolutional networks that can give them an exponential advantage is not weight sharing but locality at each level of the hierarchy.

Curse of dimensionality



The first theorem is about shallow networks.

Theorem 1 *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be infinitely differentiable, and not a polynomial. For $f \in W_m^n$ the complexity of shallow networks that provide accuracy at least ϵ is*

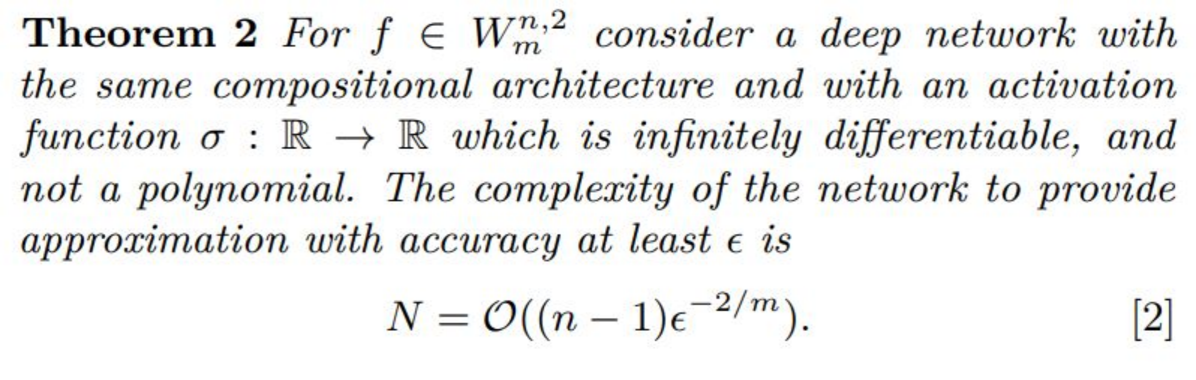
$$N = \mathcal{O}(\epsilon^{-n/m}) \text{ and is the best possible.}$$

[1]

What is it about?

The exponential dependence on the dimension **n** of the number of parameters needed to obtain an accuracy **ϵ** is known as the curse of dimensionality.

Curse of dimensionality



Theorem 2 *For $f \in W_m^{n,2}$ consider a deep network with the same compositional architecture and with an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ which is infinitely differentiable, and not a polynomial. The complexity of the network to provide approximation with accuracy at least ϵ is*

$$N = \mathcal{O}((n-1)\epsilon^{-2/m}).$$

[2]

For deep networks:

We formulate it in the binary tree case for simplicity but it extends immediately to functions that are compositions of constituent functions of a fixed number of variables d.

Curse of dimensionality

UNIVERSIDAD NACIONAL DE COLOMBIA

Trainable parameters needed in a shallow and a deep network respectively to guarantee an accuracy of ϵ:

$$\mathcal{O}(\epsilon^{-n/m}) \qquad \mathcal{O}(\epsilon^{-2/m})$$

The only a priori assumption on the target function is about the number of derivatives: $f \in W_m^n$

Optimization Landscape

UNIVERSIDAD
NACIONAL
DE COLOMBIA

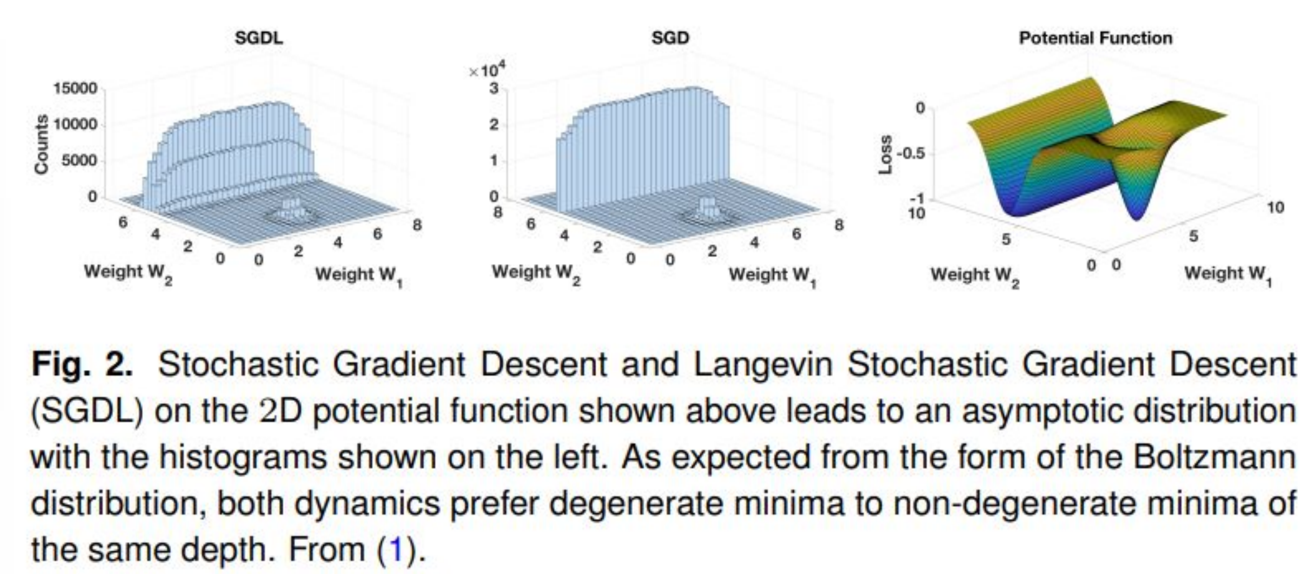
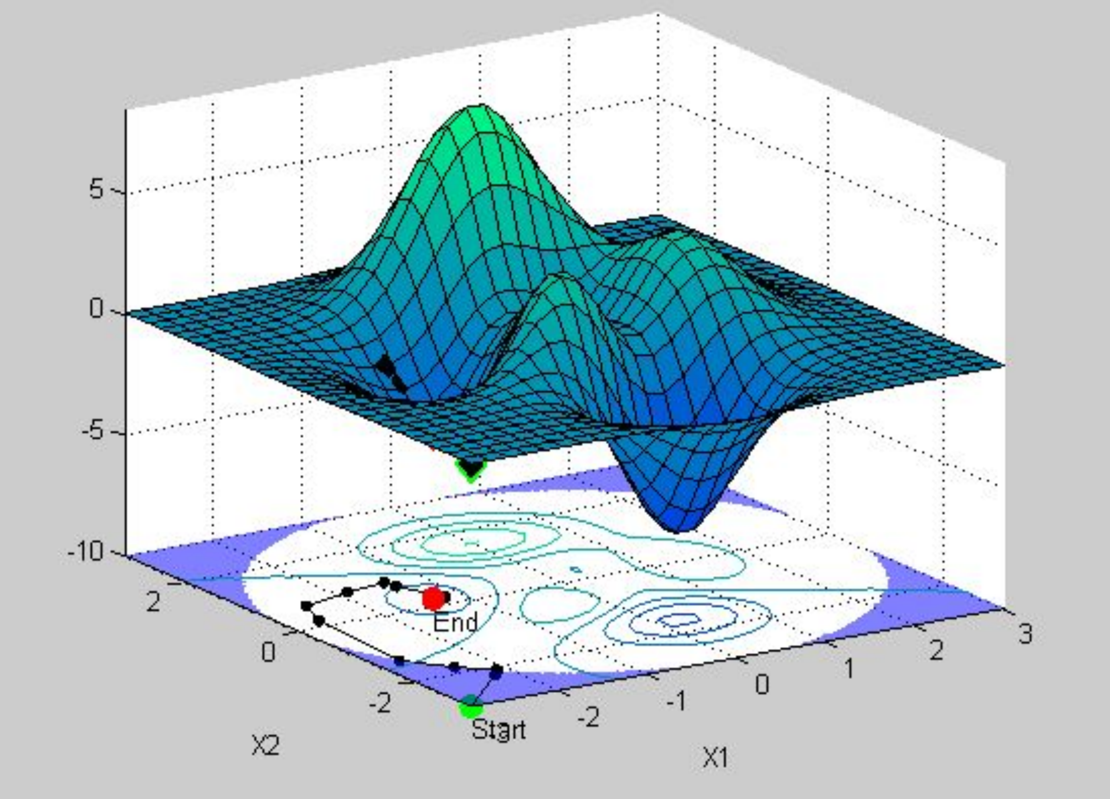


Fig. 2. Stochastic Gradient Descent and Langevin Stochastic Gradient Descent (SGDL) on the $2D$ potential function shown above leads to an asymptotic distribution with the histograms shown on the left. As expected from the form of the Boltzmann distribution, both dynamics prefer degenerate minima to non-degenerate minima of the same depth. From (1).

The other critical points of the gradient are less degenerate, with at least one – and typically N – nonzero eigenvalues

Optimization Landscape

UNIVERSIDAD NACIONAL DE COLOMBIA

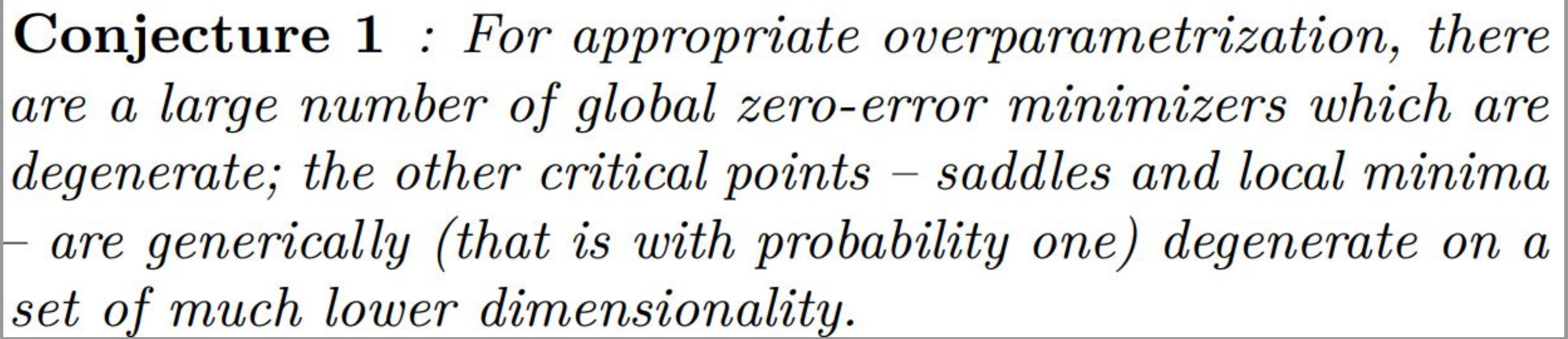


Under the exponential loss, global minima are completely degenerate with all eigenvalues of the Hessian (W of them with W being the number of parameters in the network) being zero..

Retrieved from: www.math.uwaterloo.ca

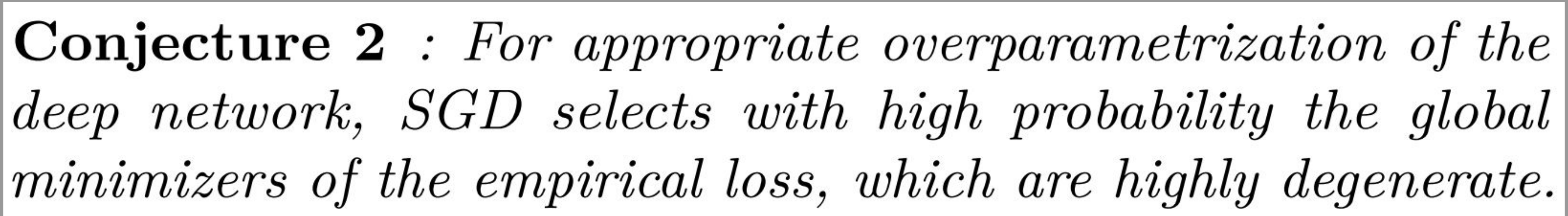
Optimization Landscape

UNIVERSIDAD
NACIONAL
DE COLOMBIA

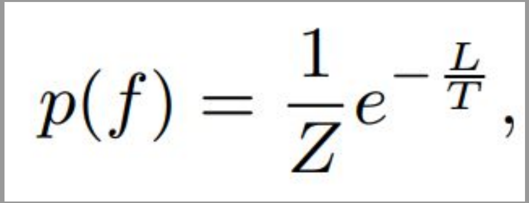


Conjecture 1 : *For appropriate overparametrization, there are a large number of global zero-error minimizers which are degenerate; the other critical points – saddles and local minima – are generically (that is with probability one) degenerate on a set of much lower dimensionality.*

Optimization Landscape



Conjecture 2 : *For appropriate overparametrization of the deep network, SGD selects with high probability the global minimizers of the empirical loss, which are highly degenerate.*



$$p(f) = \frac{1}{Z} e^{-\frac{L}{T}},$$

References

[1] Poggio, T., Banburski, A., & Liao, Q. (2019). Theoretical issues in deep networks: Approximation, optimization and generalization. arXiv preprint arXiv:1908.09375.

[2] Anselmi F, Rosasco L, Tan C, Poggio T (2015) Deep convolutional network are hierarchical kernel machines. Center for Brains, Minds and Machines (CBMM) Memo No. 35, also in arXiv.

[3] Waldrop, M. M. (2019). What are the limits of deep learning?. Proceedings of the National Academy of Sciences, 116(4), 1074-1077.

[4] Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478.