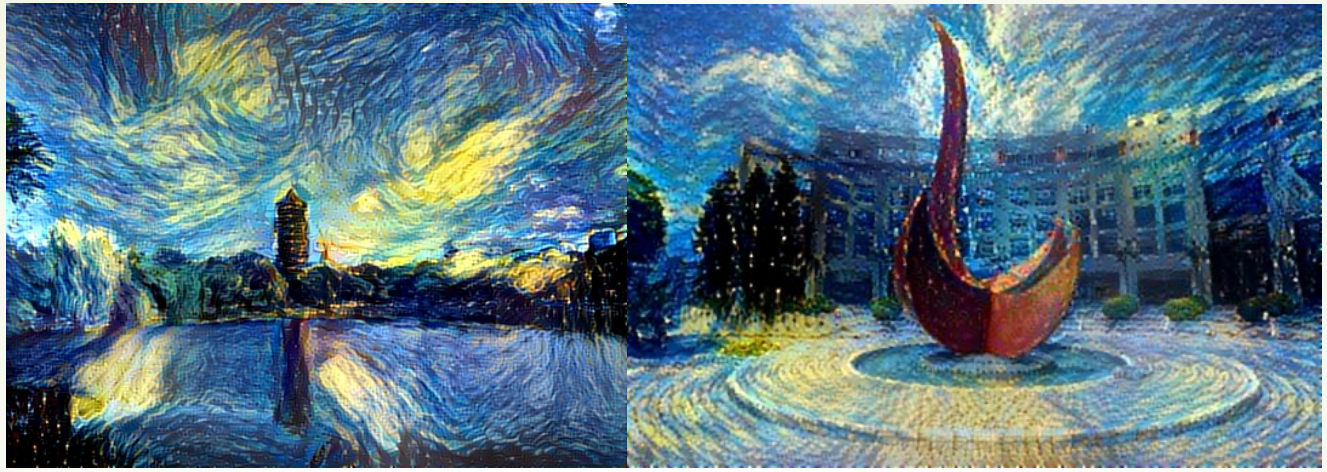
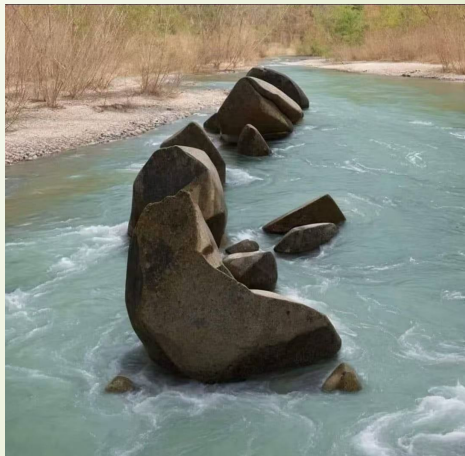


Toward Trustworthy AI: Reflections on Learning and Smale's Legacy

Yuan YAO

HKUST

Smale@95



Birthday conferences in the past 25 years

The 70th Birthday at Hong Kong, **2000**?

The 80th Birthday at Hong Kong, **2010**

The 91st Birthday online, **2021**

Happy the 95th Birthday!

My path after meeting Steve in 1999

- MPhil, City University of Hong Kong, 1999 – 2002
  - Studying Learning theory with Steve Smale (HK, 1996-2001)
  - In the last year 2001-2002, visiting UC Berkeley
- PhD, University of California at Berkeley, 2002 – 2006
  - Part time visiting TTI-Chicago with Steve Smale and Partha Niyogi
  - Thesis supervised by Steve Smale on **dynamic theory of learning** (online learning as **stochastic gradient descent**)
- Postdoc, Stanford University, 2006 - 2009
  - Working on **topological data analysis**, **Hodge theory for preference learning**, with Gunnar Carlsson and Lek-Heng Lim et al.
- Peking University, 2009 – 2016 (Steve Smale in HK, 2009 - 2016)
- Hong Kong University of Science and Technology, 2016 - now

A Dynamic Theory of Learning

by

Yuan Yao

B.S.E. (Harbin Institute of Technology) 1996
M.S.E. (Harbin Institute of Technology) 1998
M.Phil. (City University of Hong Kong) 2002

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Mathematics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

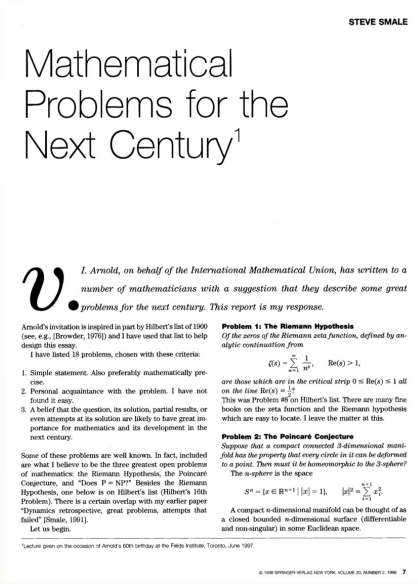
Committee in charge:
Professor Stephen Smale, Chair
Professor Steven N. Evans
Professor Peter Bartlett

Fall 2006

“Initialization”: Smale’s 18 Problem

Problem 18: Limits of Intelligence

What are the limits of intelligence, both artificial and human?



STEVE SMALE

Mathematical Problems for the Next Century[1]

V. I. Arnold, on behalf of the International Mathematical Union, has written to a number of mathematicians with a suggestion that they describe some great problems for the next century. This report is my response.

Arnold's invitation is inspired in part by Hilbert's list of 1900 (see, e.g., [Browder, 1976]) and I have used that list to help design this essay.

I have listed 18 problems, chosen with these criteria:

1. Simple statement. Also preferably mathematically precise.
2. Personal acquaintance with the problem. I have not found it easy.
3. A belief that the question, its solution, partial results, or even attempts at its solution are likely to have great importance for mathematics and its development in the next century.

Some of these problems are well known. In fact, included are what I believe to be the three greatest open problems of mathematics: the Riemann Hypothesis, the Poincaré Conjecture, and "Does P = NP?" Besides the Riemann Hypothesis, one below is on Hilbert's list (Hilbert's 16th Problem). There is a certain overlap with my earlier paper "Dynamics retrospective, great problems, attempts that failed" [Smale, 1991].

Let us begin.

Problem 1: The Riemann Hypothesis

Of the zeros of the Riemann zeta function, defined by analytic continuation from

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}, \qquad \mathrm{Re}(s) > 1,$$

are those which are in the critical strip $0 \le \mathrm{Re}(s) \le 1$ *all on the line* $\mathrm{Re}(s) = \frac{1}{2}$?

This was Problem #8 on Hilbert's list. There are many fine books on the zeta function and the Riemann hypothesis which are easy to locate. I leave the matter at this.

Problem 2: The Poincaré Conjecture

Suppose that a compact connected 3-dimensional manifold has the property that every circle in it can be deformed to a point. Then must it be homeomorphic to the 3-sphere?

The *n-sphere* is the space

$$S^n = \{x \in \mathbb{R}^{n+1} \mid \|x\| = 1\}, \qquad \|x\|^2 = \sum_{i=1}^{n+1} x_i^2.$$

A compact n-dimensional manifold can be thought of as a closed bounded n-dimensional surface (differentiable and non-singular) in some Euclidean space.

[1]Lecture given on the occasion of Arnold's 60th birthday at the Fields Institute, Toronto, June 1997.

Problem 18: Limits of Intelligence

What are the limits of intelligence, both artificial and human?

Penrose (1991) attempts to show some limitations of artificial intelligence. His argumentation brings in the interesting question whether the Mandelbrot set is decidable (dealt with in [Blum and Smale, 1993]) and implications of the Gödel incompleteness theorem.

However, a broader study is called for, one which involves deeper models of the brain, and of the computer, in a search of what artificial and human intelligence have in common, and how they differ. I would look in a direction where learning, problem-solving, and game theory play a substantial role, together with the mathematics of real numbers, approximations, probability, and geometry.

I hope to expand on these thoughts on another occasion.

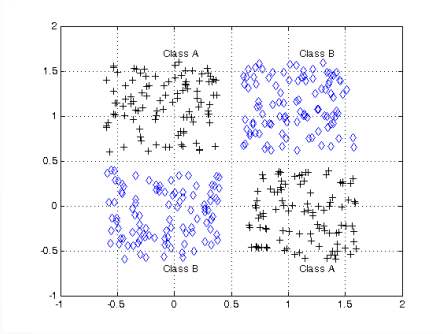
- Smale, Steve. Mathematical Problems for the Next Century, *The Mathematical Intelligencer* **20**, 7–15 (1998). https://doi.org/10.1007/BF03025291

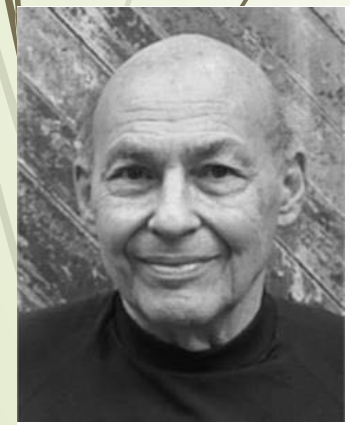
Computational Constraint: Locality or Sparsity

Minsky and Papert, 1969

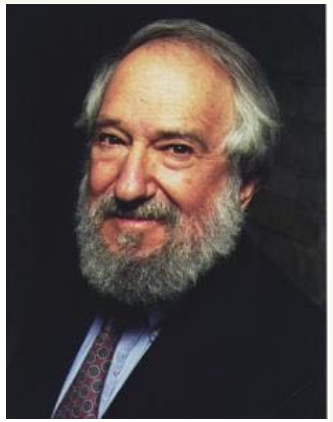
Perceptron can't do **XOR** classification

Perceptron needs infinite global information to compute **connectivity**





Marvin Minsky



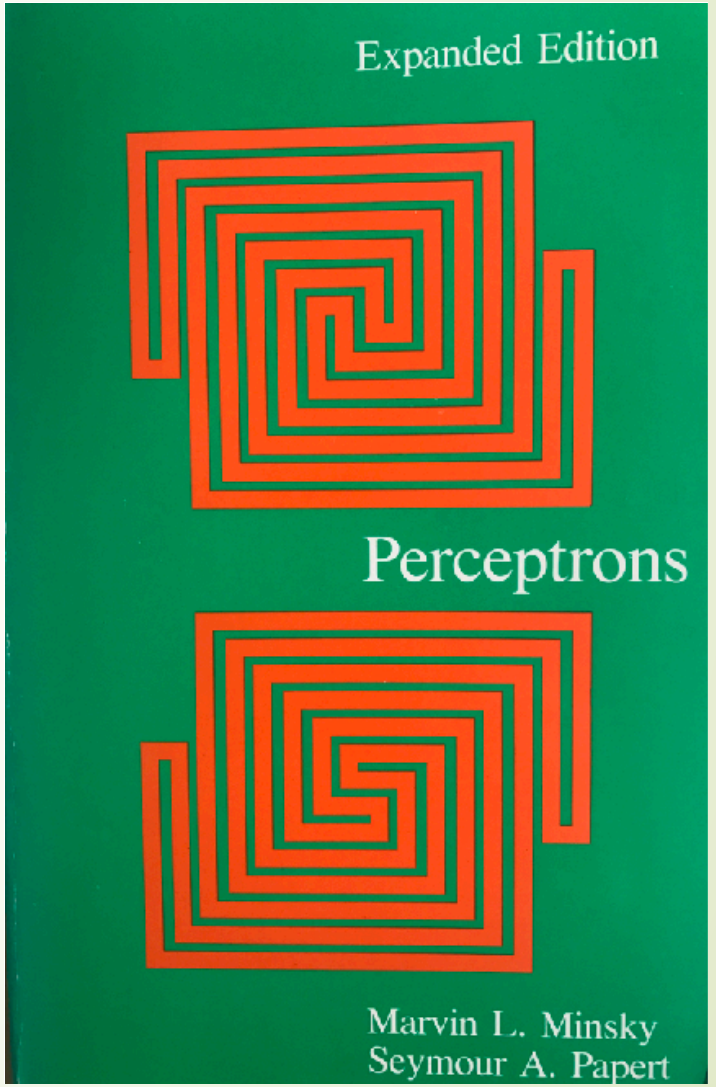
Seymour Papert

Locality or **Sparsity** is important:

Locality in time?

Locality in space?

Tommy Poggio's ***Compositional Sparsity***

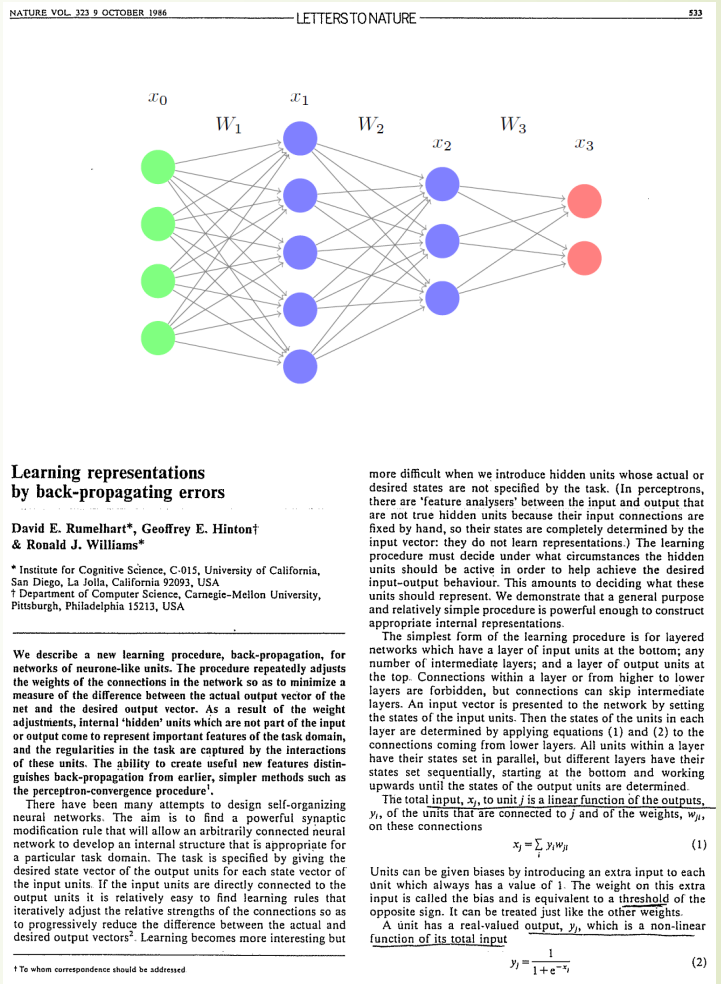


Multilayer Perceptrons (MLP) and Back-Propagation (BP) Algorithms

Rumelhart, Hinton, Williams (1986)
Learning representations by back-propagating errors, Nature, 323(9): 533-536

BP algorithms as **stochastic gradient descent** algorithms (**Robbins–Monro 1951; Kiefer-Wolfowitz 1951**) with Chain rules of Gradient maps

MLP classifies **XOR**, but the global hurdle on topology (connectivity) computation still exists: condition number in **Blum-Shub-Smale** real computation models helps.



NATURE VOL. 323 9 OCTOBER 1986 — LETTERS TO NATURE — 533

Learning representations by back-propagating errors

David E. Rumelhart*, Geoffrey E. Hinton† & Ronald J. Williams*

* Institute for Cognitive Science, C-015, University of California, San Diego, La Jolla, California 92093, USA
† Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Philadelphia 15213, USA

We describe a new learning procedure, back-propagation, for networks of neurone-like units. The procedure repeatedly adjusts the weights of the connections in the network so as to minimize a measure of the difference between the actual output vector of the net and the desired output vector. As a result of the weight adjustments, internal 'hidden' units which are not part of the input or output come to represent important features of the task domain, and the regularities in the task are captured by the interactions of these units. The ability to create useful new features distinguishes back-propagation from earlier, simpler methods such as the perceptron-convergence procedure[1].

There have been many attempts to design self-organizing neural networks. The aim is to find a powerful synaptic modification rule that will allow an arbitrarily connected neural network to develop an internal structure that is appropriate for a particular task domain. The task is specified by giving the desired state vector of the output units for each state vector of the input units. If the input units are directly connected to the output units it is relatively easy to find learning rules that iteratively adjust the relative strengths of the connections so as to progressively reduce the difference between the actual and desired output vectors[2]. Learning becomes more interesting but more difficult when we introduce hidden units whose actual or desired states are not specified by the task. (In perceptrons, there are 'feature analysers' between the input and output that are not true hidden units because their input connections are fixed by hand, so their states are completely determined by the input vector: they do not learn representations.) The learning procedure must decide under what circumstances the hidden units should be active in order to help achieve the desired input-output behaviour. This amounts to deciding what these units should represent. We demonstrate that a general purpose and relatively simple procedure is powerful enough to construct appropriate internal representations.

The simplest form of the learning procedure is for layered networks which have a layer of input units at the bottom; any number of intermediate layers; and a layer of output units at the top. Connections within a layer or from higher to lower layers are forbidden, but connections can skip intermediate layers. An input vector is presented to the network by setting the states of the input units. Then the states of the units in each layer are determined by applying equations (1) and (2) to the connections coming from lower layers. All units within a layer have their states set in parallel, but different layers have their states set sequentially, starting at the bottom and working upwards until the states of the output units are determined.

The total input, x_j, to unit j is a linear function of the outputs, y_i, of the units that are connected to j and of the weights, w_{ji}, on these connections

$$x_j = \sum_i y_i w_{ji} \quad (1)$$

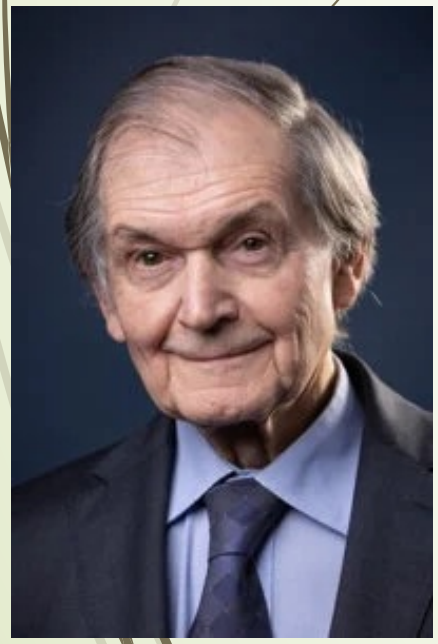
Units can be given biases by introducing an extra input to each unit which always has a value of 1. The weight on this extra input is called the bias and is equivalent to a threshold of the opposite sign. It can be treated just like the other weights.

A unit has a real-valued output, y_j, which is a non-linear function of its total input

$$y_j = \frac{1}{1+e^{-x_j}} \quad (2)$$

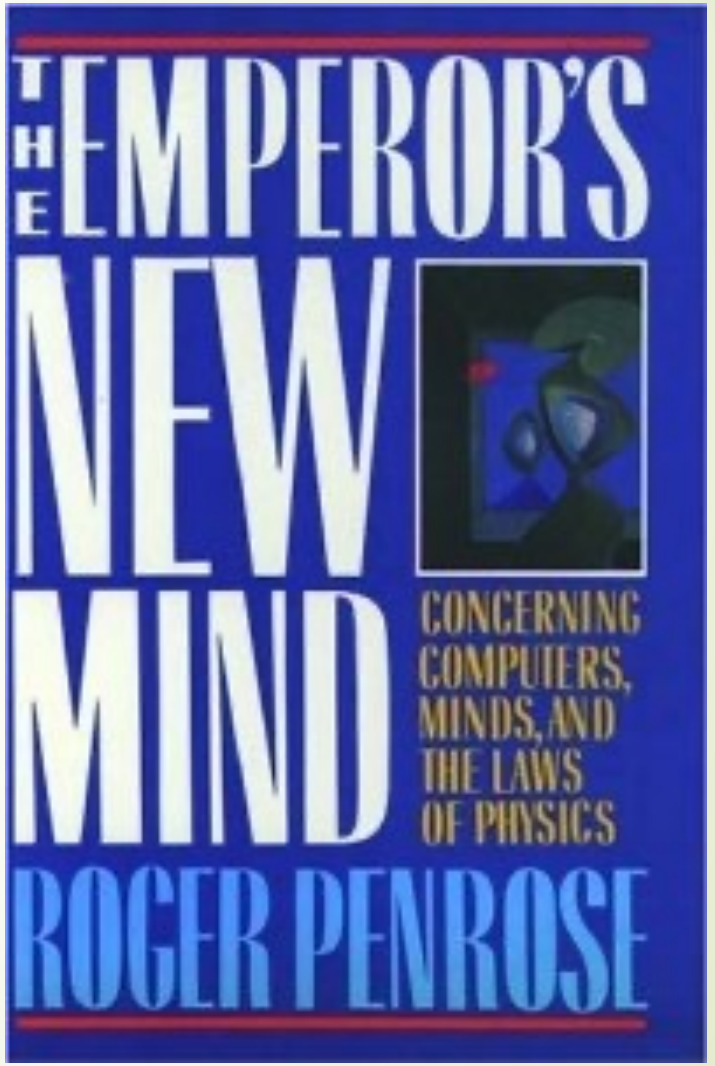
† To whom correspondence should be addressed

Roger Penrose (1989): *The Emperor's New Mind*



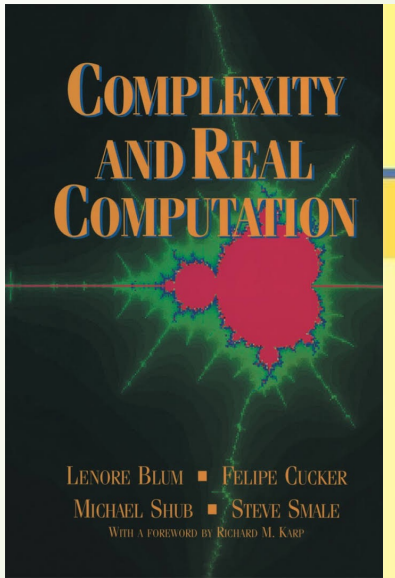
Turing Machine as AI's foundation is *insufficient?*

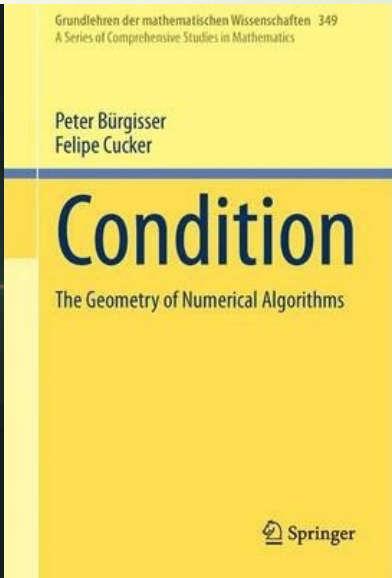
For example, is **Mandelbrot set** *computable?*



A Model of Real Computation

- Starting from **Blum, Shub, Smale** (1989)
- It admits inputs and operations (addition, substraction, multiplication, and (in the case of fields) division) of **real (complex) numbers** with *infinite precision*
- "The key importance of the **condition number**, which measures the closeness of a problem instance to the manifold of ill-posed instances, is clearly developed." – **Richard Karp**

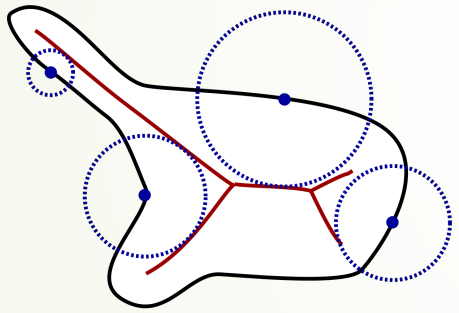




*Find Homology with Finite Samples

[Niyogi, Smale, Weinberger (2008)]

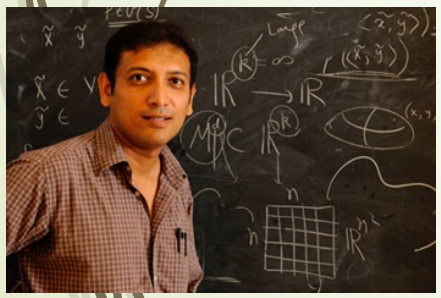
- **For "stable" manifolds that does not change topology under perturbations (measured by condition number), one needs finite samples of size exponential to the intrinsic dimensionality.**



Theorem 3.1 *Let $\mathcal{M}$ be a compact submanifold of $\mathbb{R}^N$ with condition number τ. Let $\bar{x} = \{x_1, \ldots, x_n\}$ be a set of n points drawn in i.i.d. fashion according to the uniform probability measure on $\mathcal{M}$. Let $0 < \epsilon < \tau/2$. Let $U = \bigcup_{x \in \bar{x}} B_\epsilon(x)$ be a correspondingly random open subset of $\mathbb{R}^N$. Then for all*

$$n > \beta_1 \left(\log(\beta_2) + \log\left(\frac{1}{\delta}\right) \right),$$

the homology of U equals the homology of $\mathcal{M}$ with high confidence (probability $> 1 - \delta$).



Partha Niyogi@Chicago, 1967-2010

$G = \{x \in \mathbb{R}^N$ such that $\exists$ distinct $p, q \in \mathcal{M}$ where $d(x, \mathcal{M}) = \|x - p\| = \|x - q\|\}$,

where $d(x, \mathcal{M}) = \inf_{y \in \mathcal{M}} \|x - y\|$ is the distance of x to $\mathcal{M}$. The closure of G is called the medial axis and for any point $p \in \mathcal{M}$ the local feature size $\sigma(p)$ is the distance of p to the medial axis. Then it is easy to check that

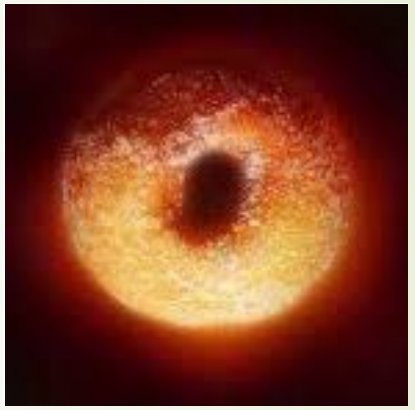
$$\tau = \inf_{p \in \mathcal{M}} \sigma(p).$$

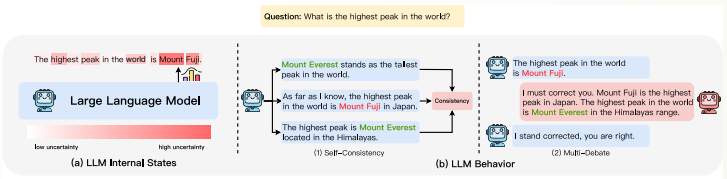
$$\beta_1 = \frac{vol(\mathcal{M})}{(\cos^k(\theta_1))vol(B^k_{\epsilon/4})} \quad \text{and} \quad \beta_2 = \frac{vol(\mathcal{M})}{(\cos^k(\theta_2))vol(B^k_{\epsilon/8})}.$$

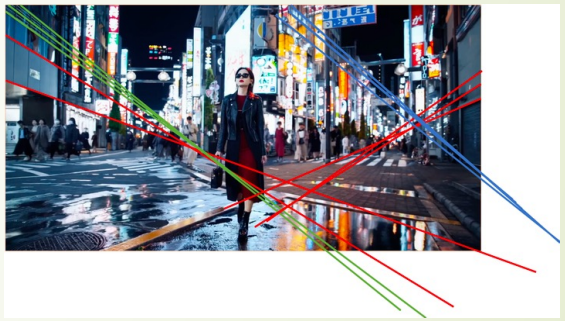
Here k is the dimension of the manifold $\mathcal{M}$ and $vol(B^k_\epsilon)$ denotes the k-dimensional volume of the standard k-dimensional ball of radius ϵ. Finally, $\theta_1 = \arcsin(\epsilon/8\tau)$ and $\theta_2 = \arcsin(\epsilon/16\tau)$.

Hallucination is ubiquitous in AI nowadays

- Hallucinations are ubiquitous in computer vision, large language models (LLM), etc.
- Convolutional neural networks are *instable* where imperceivable/small perturbations change the output
- *Projective geometry* does not preserve in video generation by SORA
- *Facts* are often lost in generative natural language statements

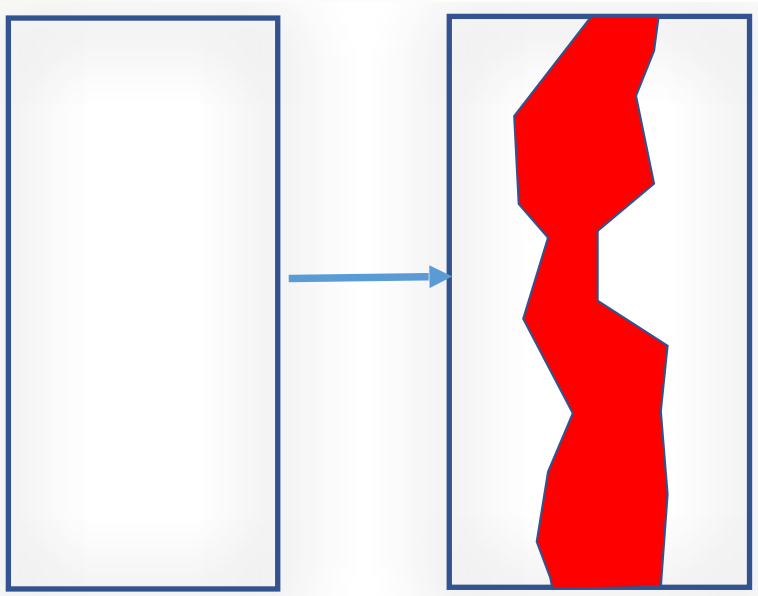






*Current deep neural networks learned from data **lack the stability** or physical **invariances**, which needs much more data or constraints for alignment toward trustworthy models.*

*Example: Lottery Ticket Hypothesis for Efficient Subnets in Deep Learning



Over-parameterized Networks W

Compressive Networks W_s

Lottery Ticket Hypothesis

- *Dense, randomly-initialized, feed-forward networks contain subnetworks (winning tickets) that – when trained in isolation – reach test accuracy comparable to the original network in a similar number of iterations.* (Frankle & Carbin, 2019)

Rewinding the network from the initialization, and find "winning ticket" subnet

Texture bias in ImageNet representation learning [Fu-Liu-Li-Zhong-Sun-Y. TPAMI 2023]

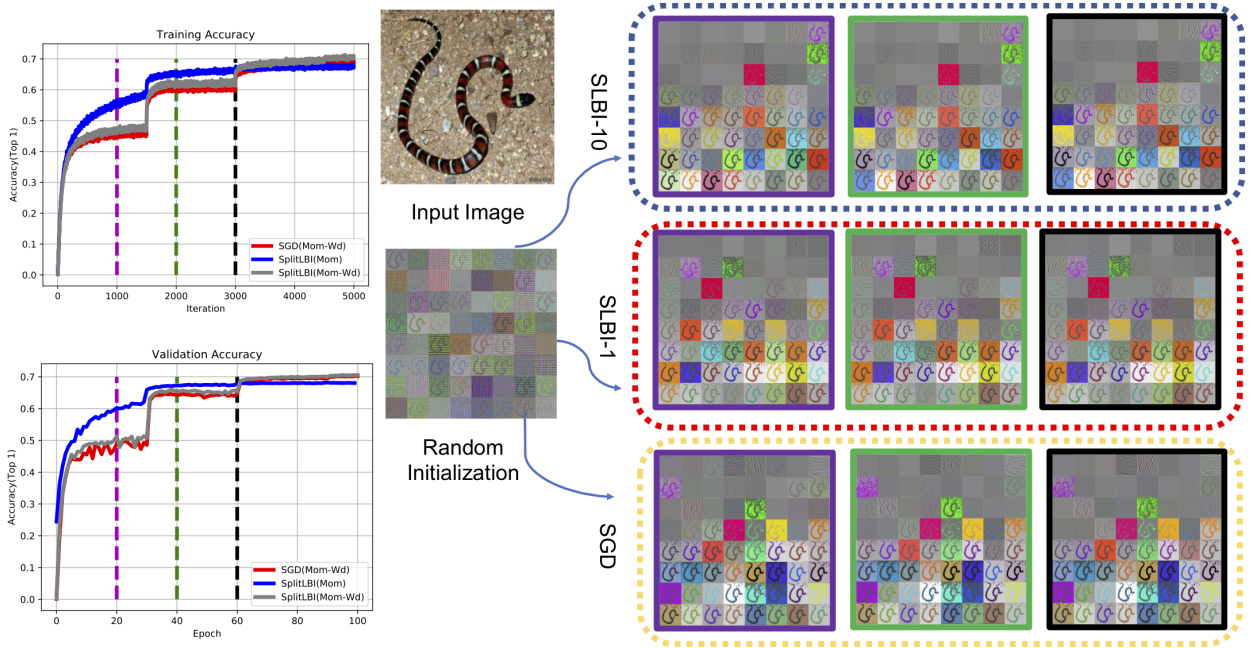
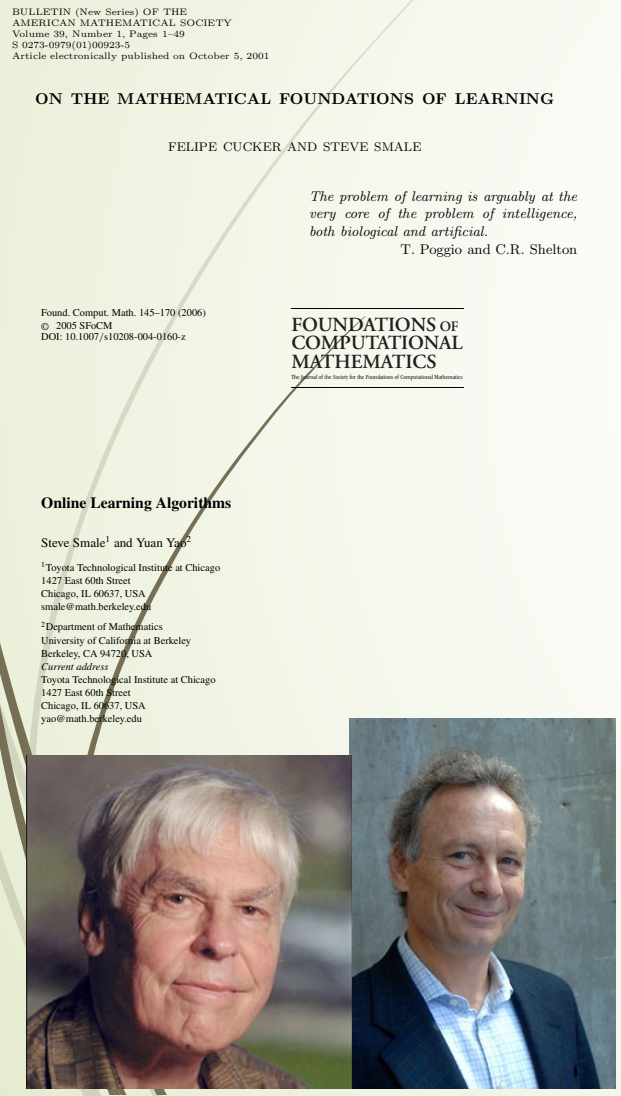


Figure: Visualization of the first convolutional layer filters of ResNet-18 trained on ImageNet-2012, where texture features are more important than colour/shapes. Given the input image and initial weights visualized in the middle, filter response gradients at 20 (purple), 40 (green), and 60 (black) epochs are visualized. SGD with Momentum (Mom) and Weight Decay (WD), is compared with SLBI.

See also: **Geirhos** *et al. ImageNet-trained CNNs are biased towards texture*, ICLR 2019

Remarks on Euclidean/Non-Euclidean gradient descent algorithms


BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 39, Number 1, Pages 1–49
S 0273-0979(01)00923-5
Article electronically published on October 5, 2001

ON THE MATHEMATICAL FOUNDATIONS OF LEARNING

FELIPE CUCKER AND STEVE SMALE

The problem of learning is arguably at the very core of the problem of intelligence, both biological and artificial.
T. Poggio and C.R. Shelton

Found. Comput. Math. 145–170 (2006)
© 2005 SFoCM
DOI: 10.1007/s10208-004-0160-z

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS

Online Learning Algorithms

Steve Smale[1] and Yuan Yao[2]

[1]Toyota Technological Institute at Chicago
1427 East 60th Street
Chicago, IL 60637, USA
smale@math.berkeley.edu

[2]Department of Mathematics
University of California at Berkeley
Berkeley, CA 94720, USA
Current address
Toyota Technological Institute at Chicago
1427 East 60th Street
Chicago, IL 60637, USA
yao@math.berkeley.edu

- Gradient Descent in reproducing kernel Hilbert spaces (RKHS):
  - With **Steve Smale** (FoCM, 2006), establish online learning as ***stochastic gradient descent*** in Reproducing Kernel Hilbert Spaces; With **Pierre Tarres** (IEEE Info. Theor. 2014), establish its optimality as stochastic path following.
  - Establish statistical consistency of **early stopping regularization** for gradient descent in RKHS, with **L. Rosasco** and **A. Caponnetto** (Constructive Approximation, 2007)
- Non-Euclidean Gradient Descent (Mirror Descent):
  - Establish *sparse model selection consistency* of **Bregman Iterations** as ***inverse-scale-space*** differential inclusion dynamics, with **S. Osher, F. Ruan, J. Xiong, W. Yin** et al. (*ACHA'16*) and **Xinwei Sun** (**Fudan**) et al. (*NIPS'16; ACHA'20*)
  - Applications in deep learning with **Yanwei Fu** (**Fudan**) et al. (*ICML'20,TPAMI'23*)

Remark on Mirror Descent

Dynamical system as non-Euclidean gradient descent

$$\begin{aligned} \dot{\rho}_t &= -\nabla_\theta \widehat{\ell}(z_n, \theta_t) \\ \rho_t &\in \partial\psi(\theta_t) \end{aligned}$$

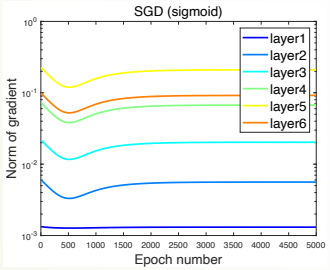
where

- $\widehat{\ell}(z_n, \theta)$ is the (empirical) loss (risk) function: negative logarithmic likelihood, non-convex loss (neural networks), etc.
- $\psi(\theta)$ convex function:
  - $\psi(\theta) = \|\theta\|_2^2$ gives the gradient descent
  - $\psi(\theta) = \|\theta\|_1$ in favour of sparsity
- known as **mirror descent** (Nemirovski and Yudin) as well as **inverse scale space, Bregman iterations** for ℓ_1-norm in applied math (Burgers, Osher, Yin, et al.)

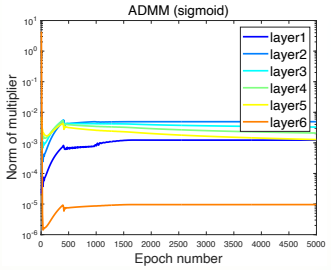
Early stopping regularization is for *sparse model selection consistency of* differential inclusion dynamics, with **S. Osher, F. Ruan, J. Xiong, W. Yin** et al. (*ACHA'16*) and **Xinwei Sun (Fudan)** et al. (*ACHA'20*)

On SGD vs. ADMM/BCD

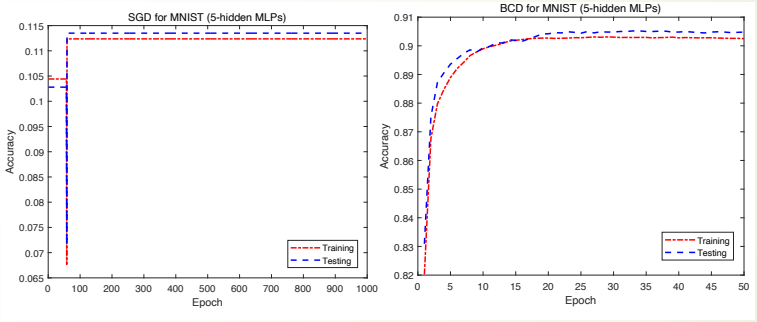
- Stochastic Gradient Descent (SGD) suffers from the well-known **gradient vanishing issue** in deep learning, which can be derived via Lagrangian multiplier method [LeCun 1988, http://yann.lecun.com/exdb/publis/pdf/lecun-88.pdf]
- Nongradient-based algorithms: **ADMM** (Augmented Lagrangian multipliers) and **BCD** (Block Coordinate Descent) may alleviate gradient vanishing



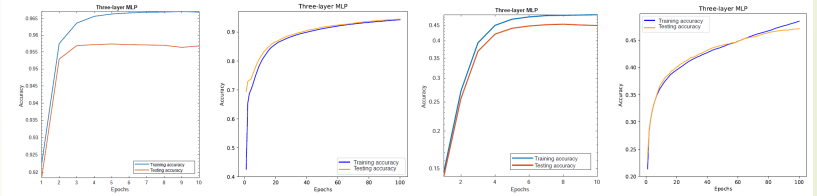
(a) Gradient vanishing of SGD



(b) Saturation-avoidance of ADMM



- High epoch efficiency of BCD at early stage



(a) MNIST (BCD) (b) MNIST (SGD) (c) CIFAR-10 (BCD) (d) CIFAR-10 (SGD)

Zeng-Lau-Lin-Y., ICML 2019
Zeng-Lin-Y.-Zhou, JMLR 2021

Further Remarks on Topology from Data

- Topology learning from noisy data is further studied by Niyogi, Smale, Weinberger, SIAM J. Comput. 2011
- Lower rates on sample complexity are established by Balakrishnan, Rinaldo, Sheehy, Singh, Wasserman, AISTATS 2012.
- Topological data analysis evolves in several categories:
  - **Persistent Homology**, comparable to Forman's Discrete Morse Theory, captures large scale stable Homology over filtrations of simplicial complexes
  - **Hodge Theory**, captures topological invariants like Betti numbers via spectrum of Laplacians, a beautiful bridge over analysis, geometry, and topology
- Polynomial (quadratic) speedup is possible with Quantum Algorithm [Lloyd et al. Nature Communications, 2022] via **Hodge theory**
- **Hodge Theory on Metric Spaces**, Bartholdi, Schick, Smale, Smale, FoCM 2012
- A special usage of ***Hodge Theory in preference learning*** is established in **Statistical Ranking and Combinatorial Hodge Theory**, by Jiang, Lim, Yao, Ye, Math Program. 2011

Thank you!