Wonders of high-dimensions:
the maths and physics of ML

Bruno Loureiro
Département d'Informatique
École Normale Supérieure & CNRS

brloureiro@gmail.com

Archimedes Workshop on the Foundations of Modern AI
03-04.07.2024

Wonders of high-dimensions:
the maths and physics of ML

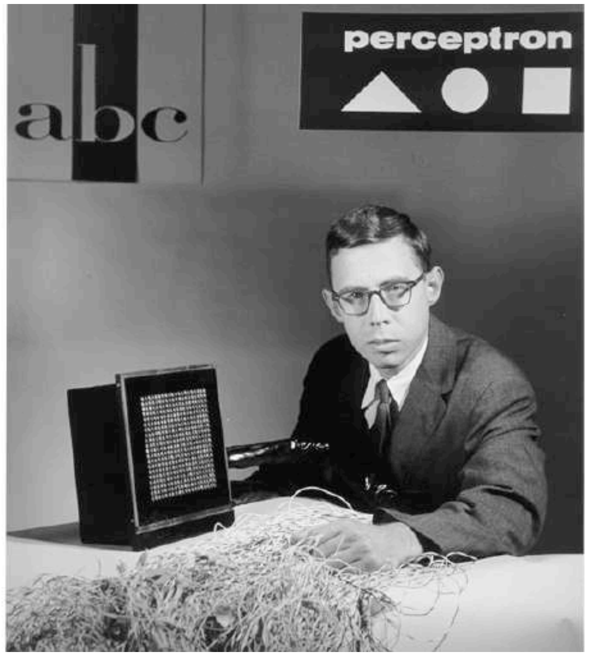
Bruno Loureiro
Département d'Informatique
École Normale Supérieure & CNRS

brloureiro@gmail.com

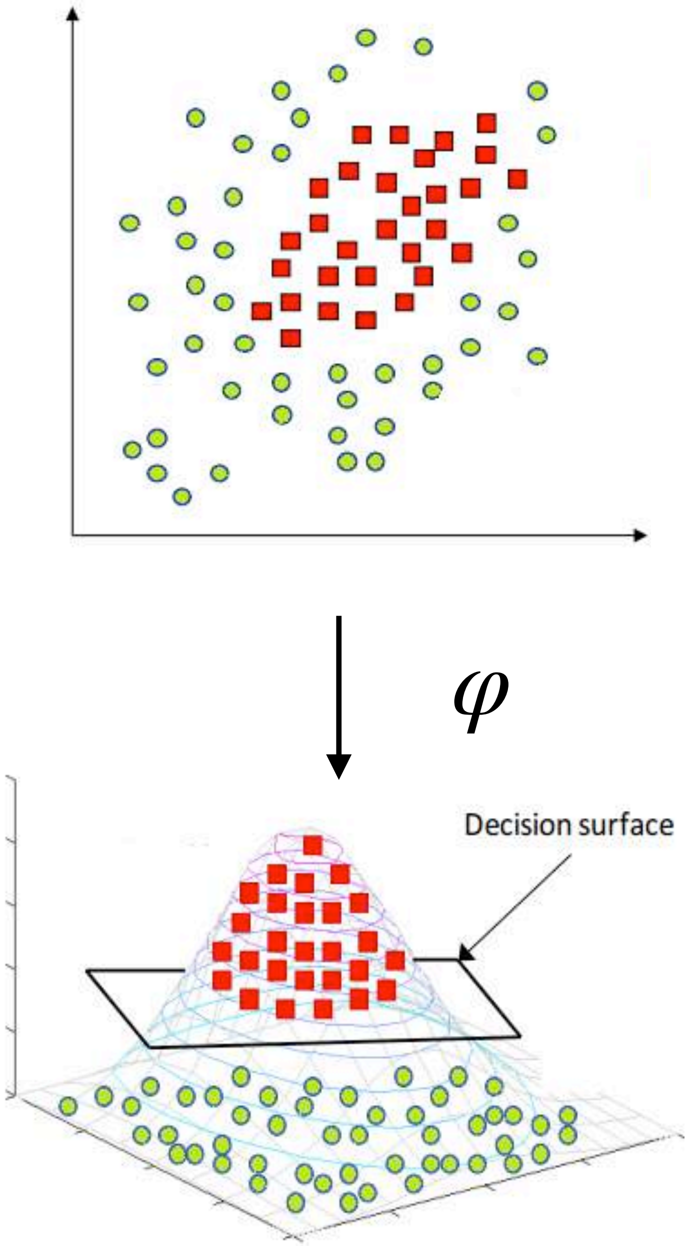
Archimedes Workshop on the Foundations of Modern AI
03-04.07.2024

Menu for this tutorial

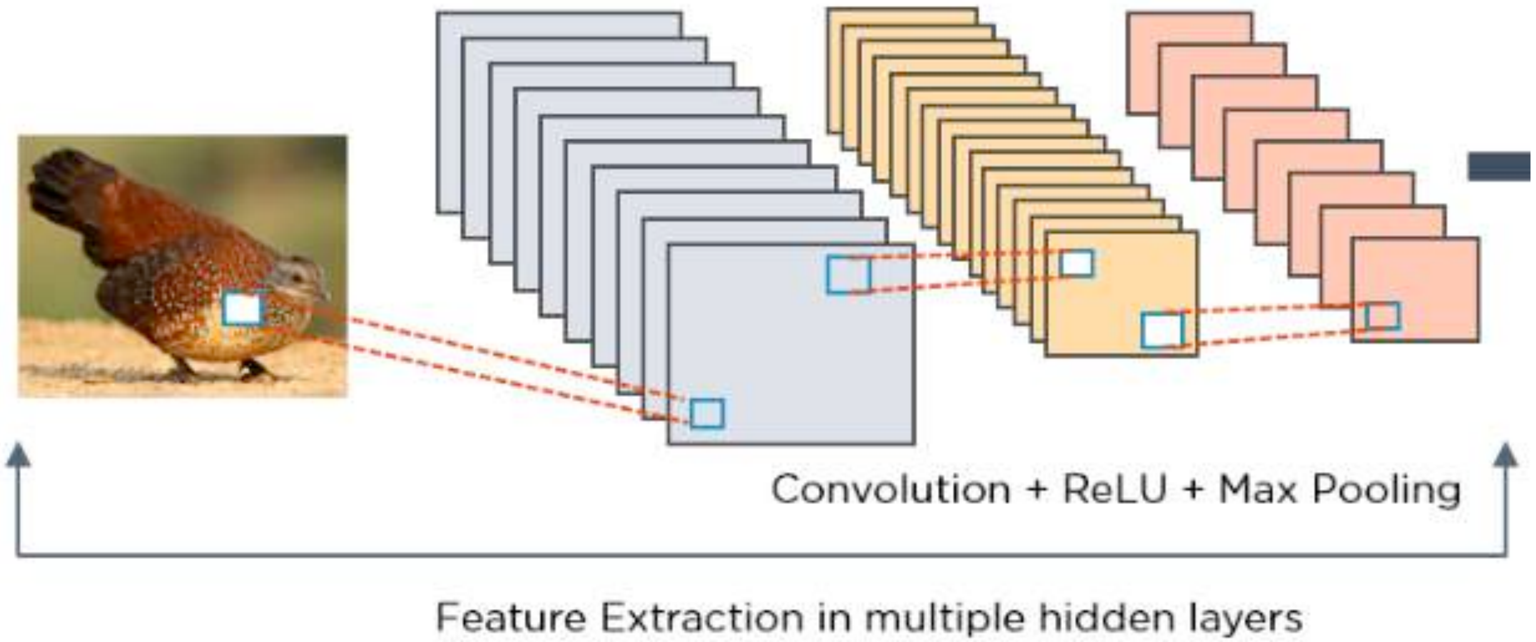
Part I: Statistical Physics of Computation



Part II: Neural Networks at initialisation (a.k.a. kernel methods)

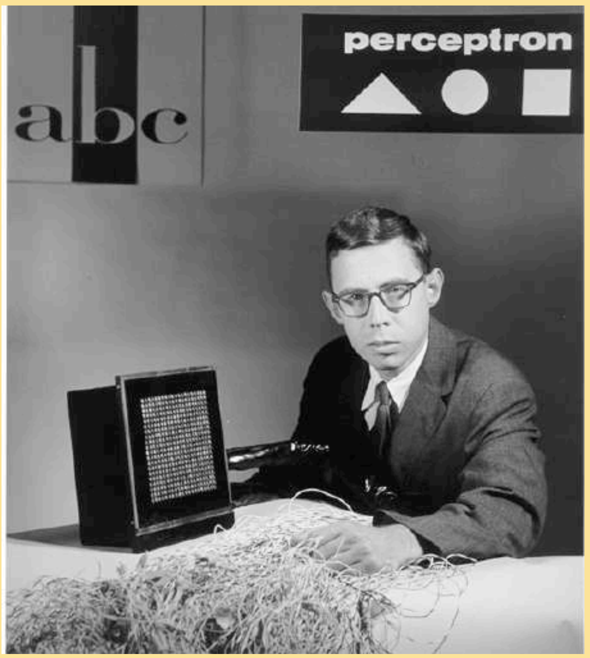


Part III: Feature learning

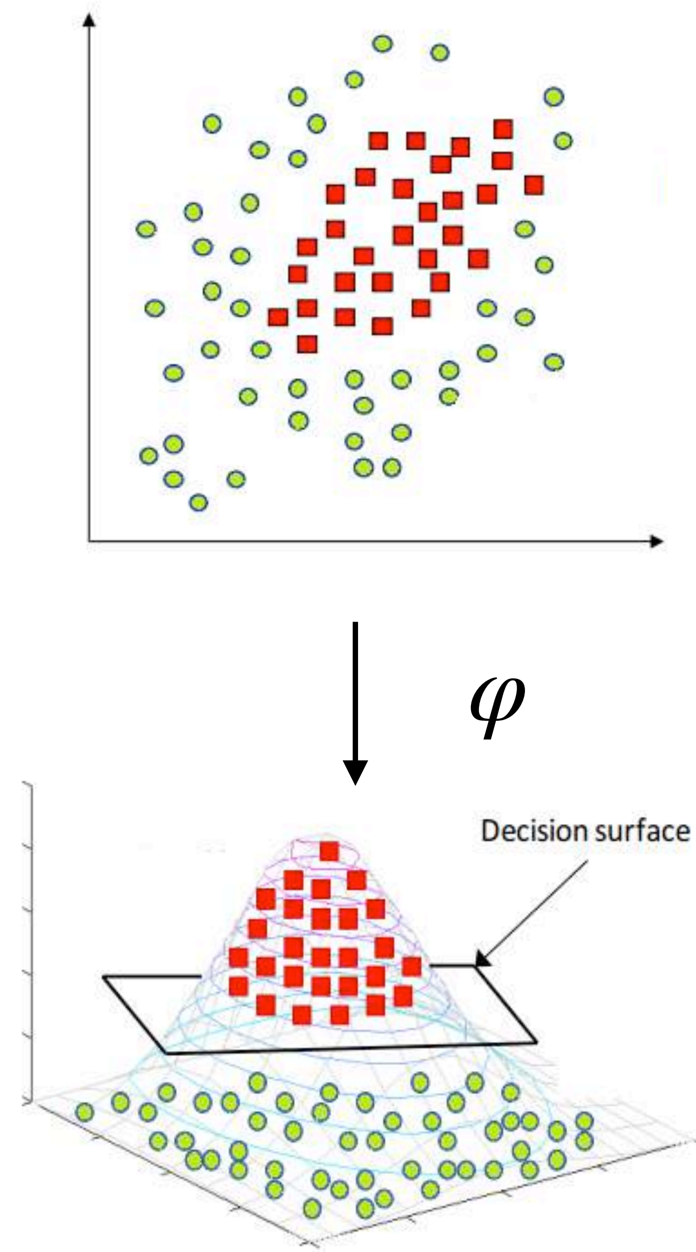


Menu for this tutorial

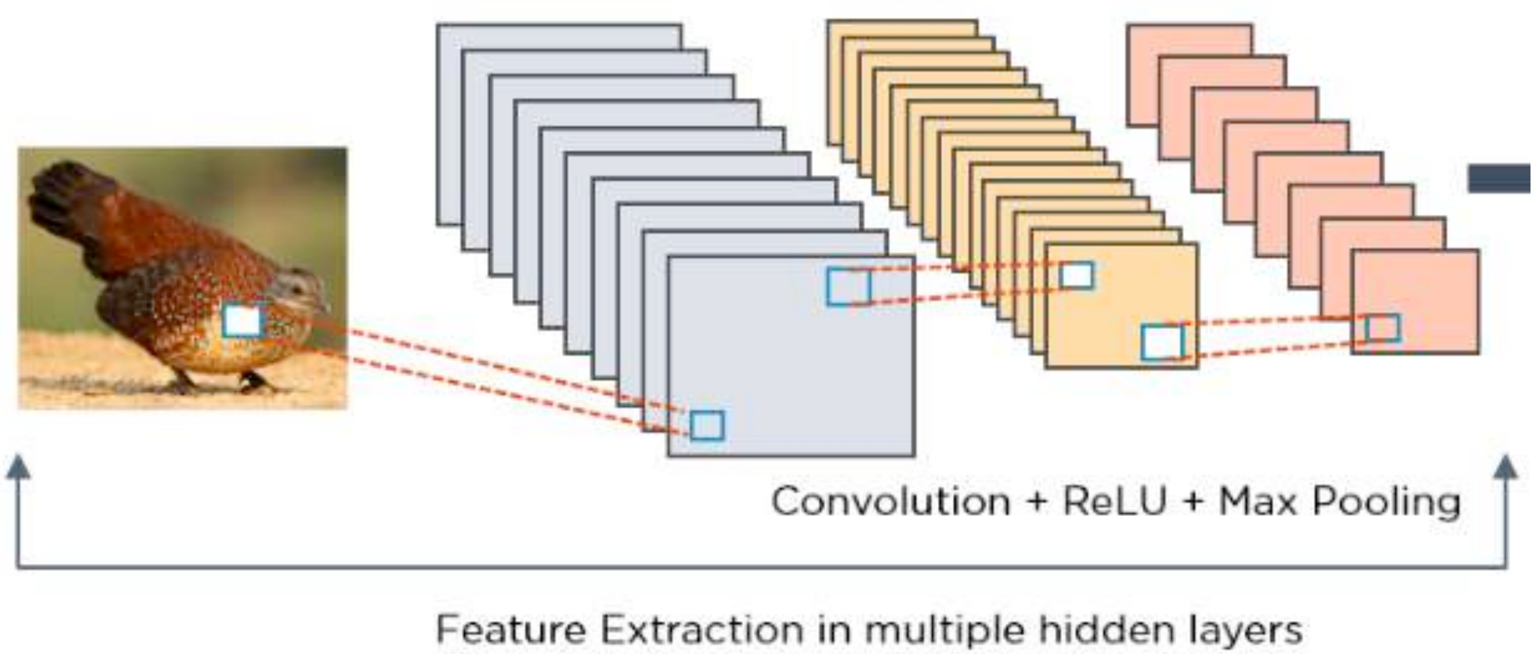
Part I: Statistical Physics of Computation



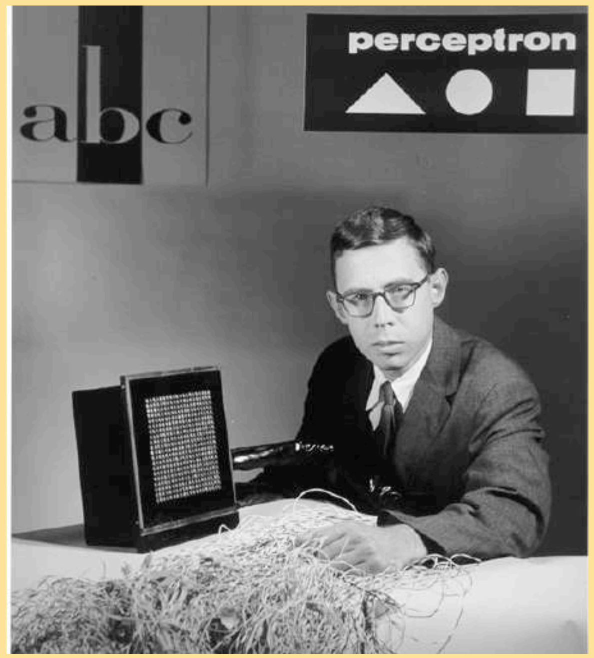
Part II: Neural Networks at initialisation (a.k.a. kernel methods)



Part III: Feature learning



<u>Part I:</u> Statistical Physics of Computation



1. Why Stat. Phys. and ML were made for each other

2. A brief history of the physics & computer science marriage

3. The relationship between phase transitions and computational hardness

Stat. Phys. 101

Classical Mechanics:

$$\dot{x} = \nabla_p H(x, p)$$

$$\dot{p} = -\nabla_x H(x, p)$$

"Hamiltonian"

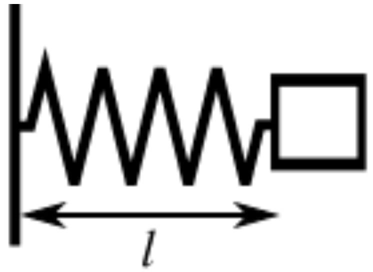
Stat. Phys. 101

Classical Mechanics: $\dot{x} = \nabla_p H(x,p)$

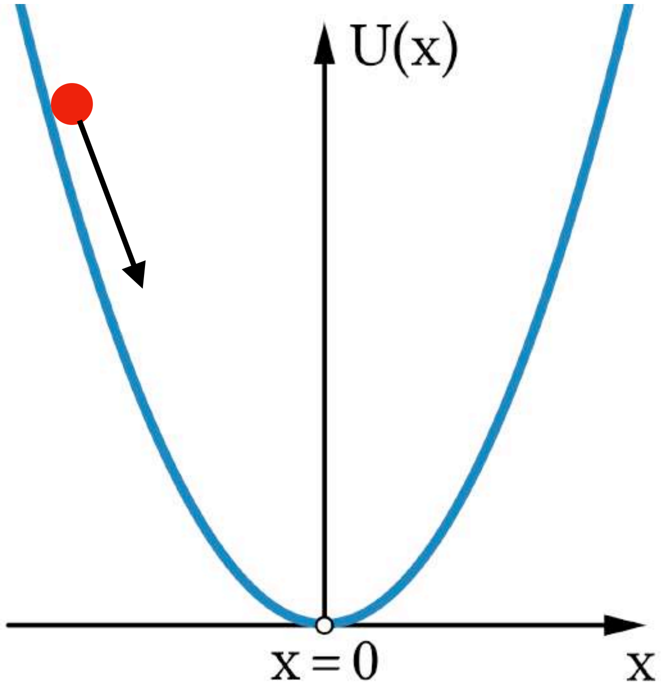
$$\dot{p} = -\nabla_x H(x,p)$$

"Hamiltonian"

Example: $H(x,p) = \dfrac{p^2}{2m} + \dfrac{kx^2}{2}$



$$\dot{x} = \frac{p}{m} \qquad \dot{p} = -kx$$



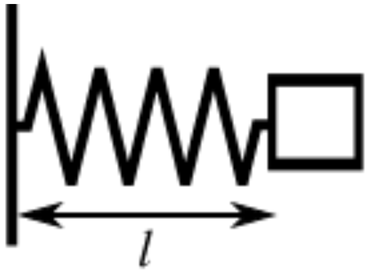
Stat. Phys. 101

Classical Mechanics: $\dot{x} = \nabla_p H(x,p)$

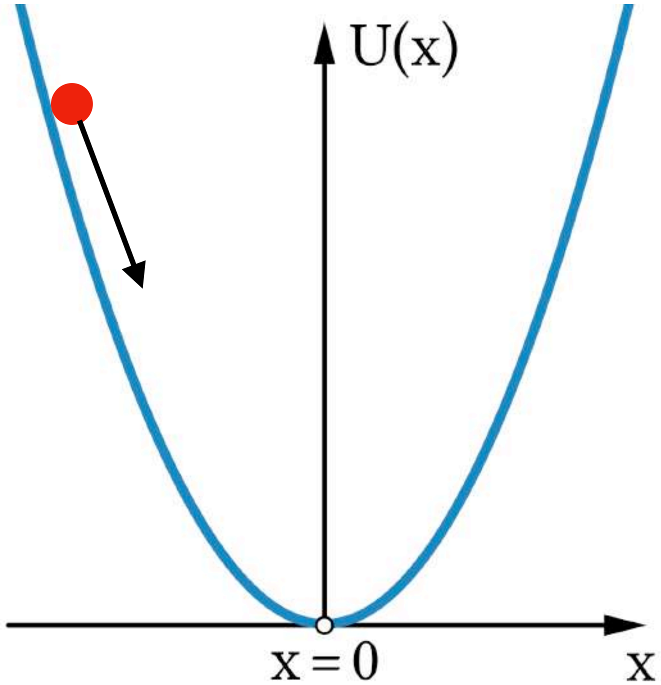
$$\dot{p} = -\nabla_x H(x,p)$$

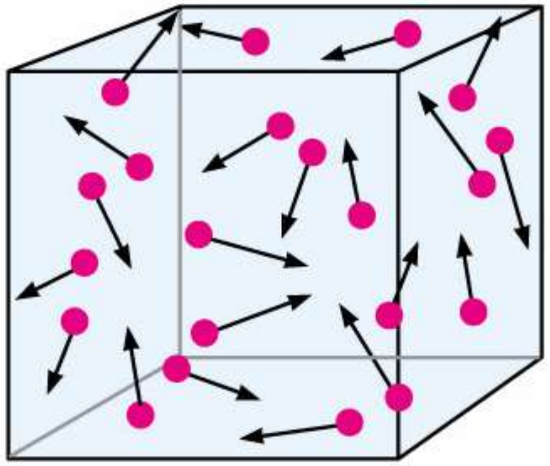
"Hamiltonian"

Example: $$H(x,p) = \frac{p^2}{2m} + \frac{kx^2}{2}$$



$$\dot{x} = \frac{p}{m} \qquad \dot{p} = -kx$$





What about n particles in dimension d ?

Requires solving $2dn$ coupled ODEs!

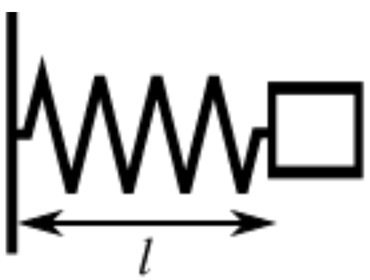
Stat. Phys. 101

Classical Mechanics: $\dot{x} = \nabla_p H(x,p)$

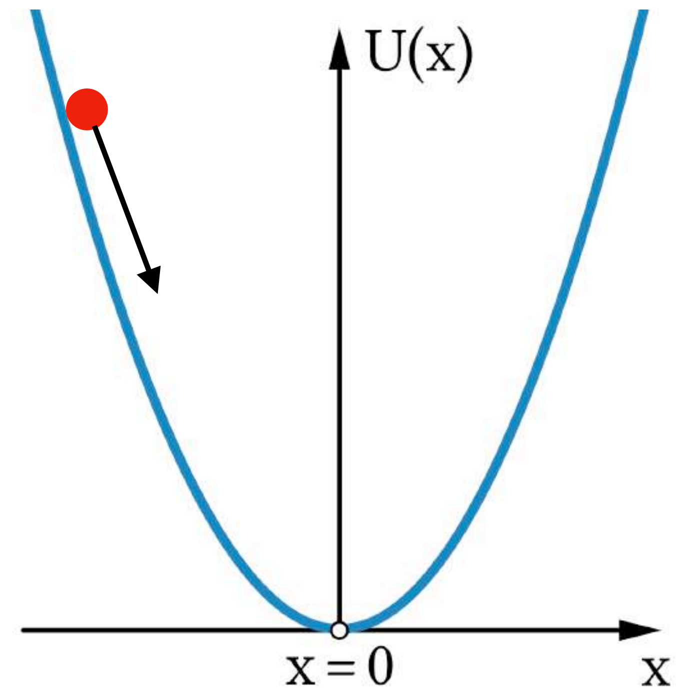
$\dot{p} = -\nabla_x H(x,p)$

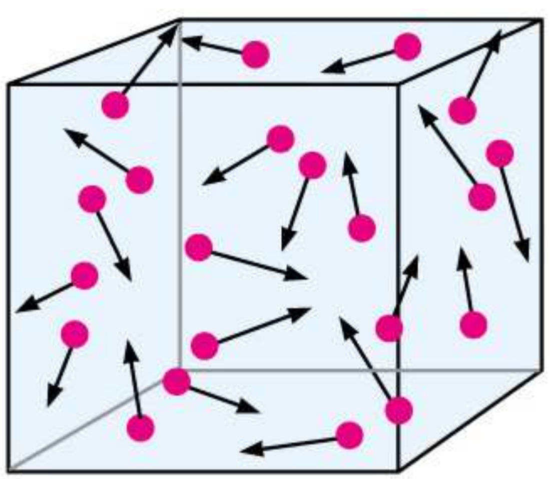
"Hamiltonian"

Example: $H(x,p) = \frac{p^2}{2m} + \frac{kx^2}{2}$



$\dot{x} = \frac{p}{m}$ $\dot{p} = -kx$





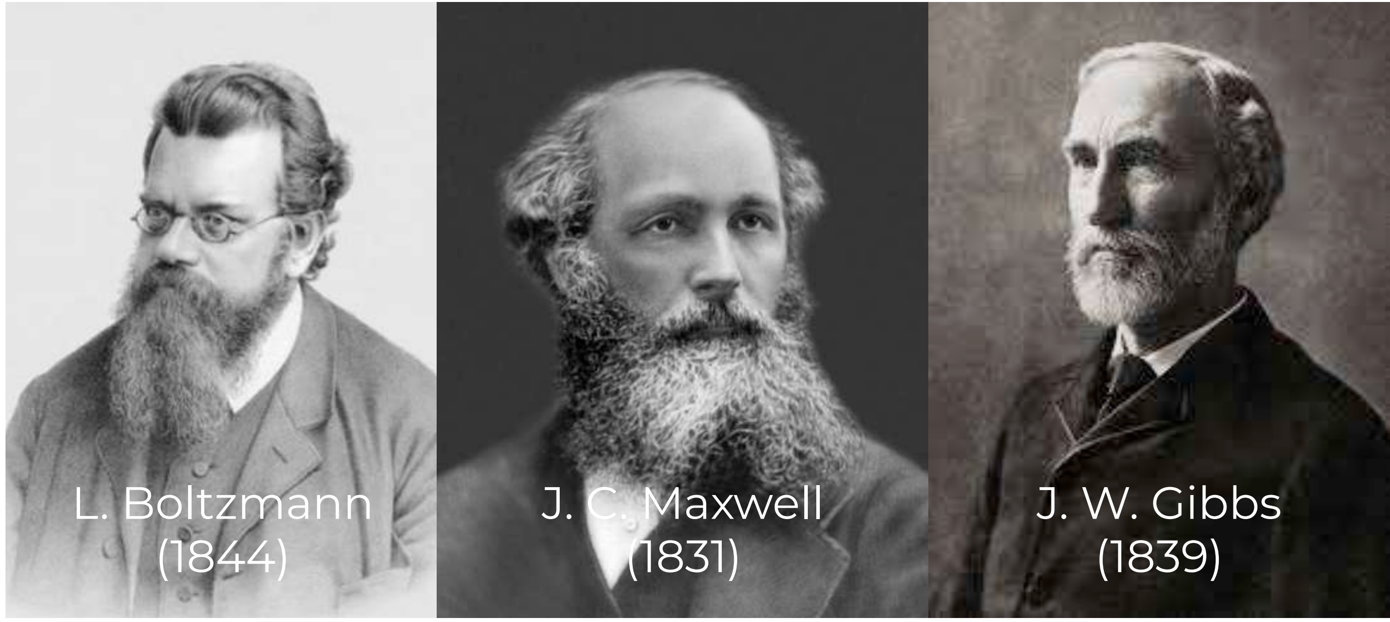
What about n particles in dimension d ?

Requires solving $2dn$ coupled ODEs!

analytically & computationally intractable!!!

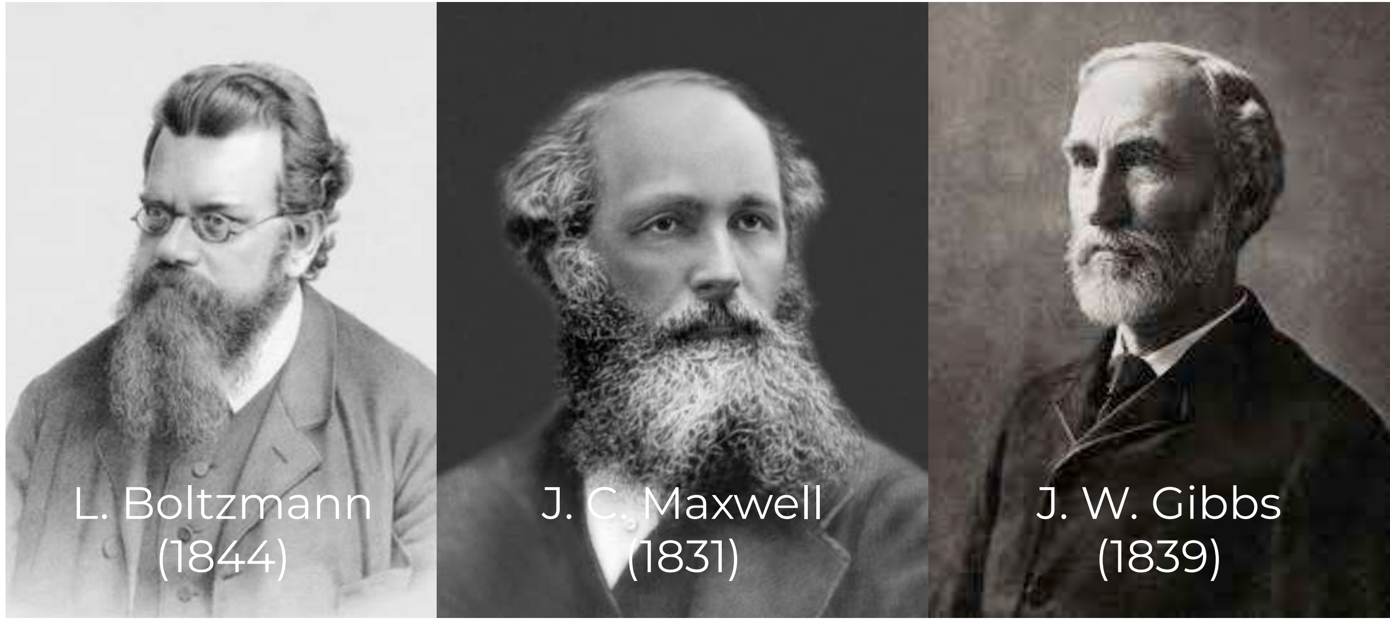
Stat. Phys. 101

Key idea of Statistical Physics:
Take a probabilistic approach.



Stat. Phys. 101

Key idea of Statistical Physics:
Take a probabilistic approach.



Define a probability measure over $\{(x_i, p_i) \in \mathbb{R}^{2d} : i \in [n]\}$ "Configuration space"

Boltzmann-Gibbs distribution

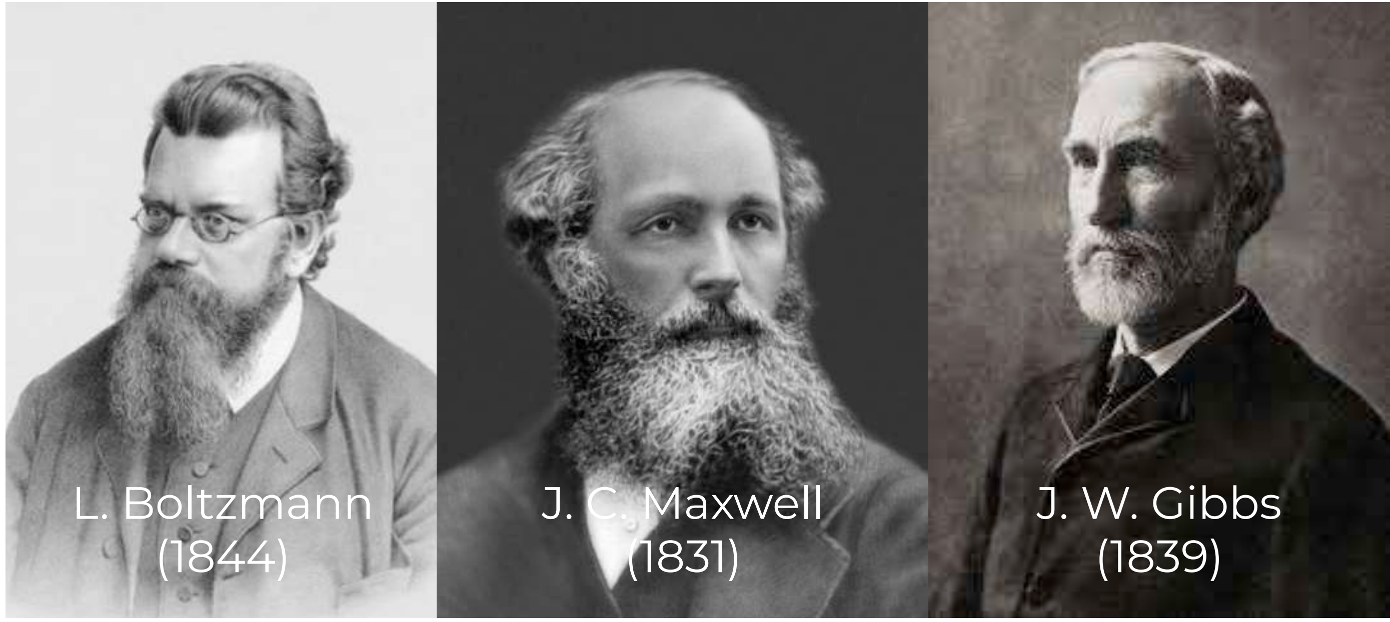
$$\mu_\beta(\{(x_i, p_i)\}) = \frac{e^{-\beta H\left(\{(x_i, p_i)\}\right)}}{\int dp \int dx \; e^{-\beta H\left(\{(x_i, p_i)\}\right)}}$$

Free energy (density)

$$-\beta f_\beta = \frac{1}{dn} \log \int dp \int dx \; e^{-\beta H\left(\{(x_i, p_i)\}\right)}$$

Stat. Phys. 101

Key idea of Statistical Physics:
Take a probabilistic approach.



Define a probability measure over $\{(x_i, p_i) \in \mathbb{R}^{2d} : i \in [n]\}$ "Configuration space"

Boltzmann-Gibbs distribution

$$\mu_\beta(\{(x_i, p_i)\}) = \frac{e^{-\beta H(\{(x_i, p_i)\})}}{\int dp \int dx \; e^{-\beta H(\{(x_i, p_i)\})}}$$

Free energy (density)

$$-\beta f_\beta = \frac{1}{dn} \log \int dp \int dx \; e^{-\beta H(\{(x_i, p_i)\})}$$

Remarks

- At $\beta \to \infty$, μ_β peaks at

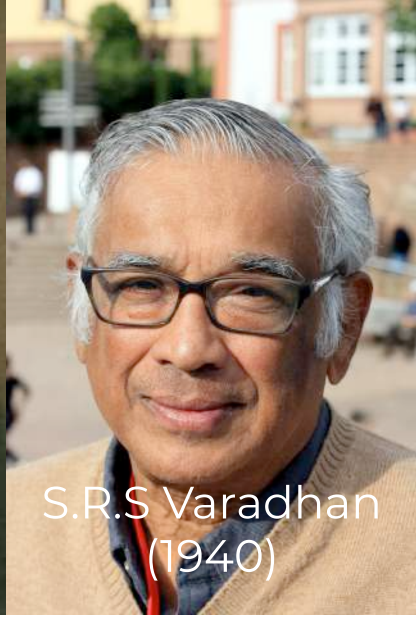
$$\operatorname{argmin} \; H(\{(p_i, q_i)\})$$

"Ground state"

- f_β is the moment generating function (MdF) of μ_β

Stat. Phys. 101


P-S. Laplace (1749)


S.R.S Varadhan (1940)

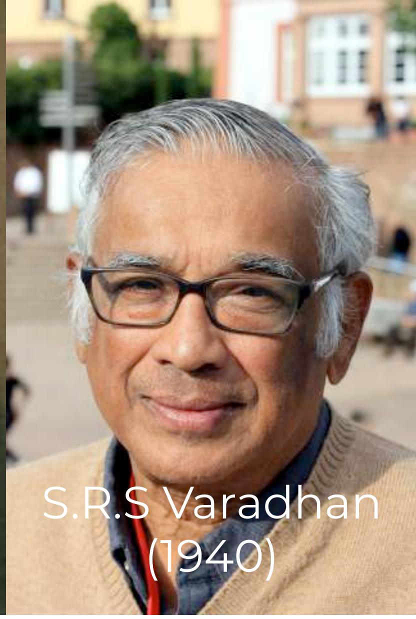
The central idea is to identify key "macroscopic" quantities

Order parameters / summary statistics

$$m : \{(x_i, p_i)\} \in \mathbb{R}^{2dn} \mapsto m(\{(x_i, p_i)\}) \in \mathbb{R}^k$$

Stat. Phys. 101

The central idea is to identify key "macroscopic" quantities



Order parameters / summary statistics

$$m : \{(x_i, p_i)\} \in \mathbb{R}^{2dn} \mapsto m(\{(x_i, p_i)\}) \in \mathbb{R}^k$$

Such that the free energy satisfy a large deviation principle

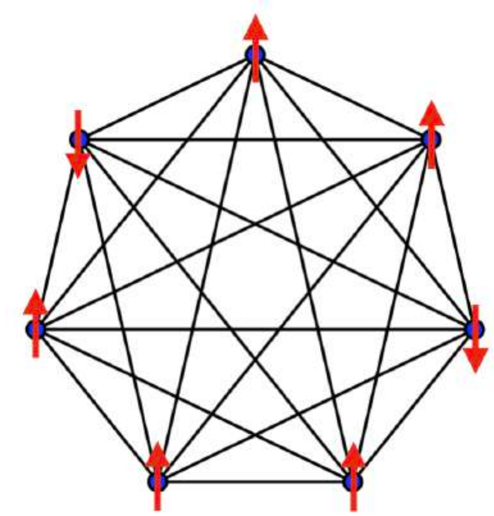
$$-\beta f_\beta = \frac{1}{dn} \log \int dp \int dx \; e^{-\beta H(\{x_i, p_i\})} \underset{n\to\infty}{\asymp} \underset{m \in \mathbb{R}^k}{\text{extr}} \, \Phi(m)$$

$$k = \Theta_n(1)$$

Curie-Weiss model

Curie-Weiss Model

$$H(s) = -\frac{J}{2n}\sum_{i,j=1}^{n} s_i s_j \qquad s \in \{-1,+1\}^n$$



"fully connected" or "complete"

Curie-Weiss model

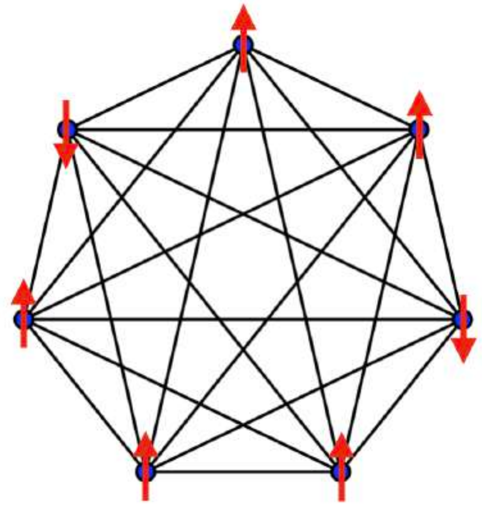
Curie-Weiss Model

$$H(s) = -\frac{J}{2n}\sum_{i,j=1}^{n} s_i s_j \qquad s \in \{-1,+1\}^n$$

Note that we can rewrite:

$$H(s) = -\frac{nJ}{2}\left(\frac{1}{n}\sum_{i=1}^{n} s_i\right)^2 = -nJm^2$$

"magnetisation"



"fully connected" or "complete"

Curie-Weiss model

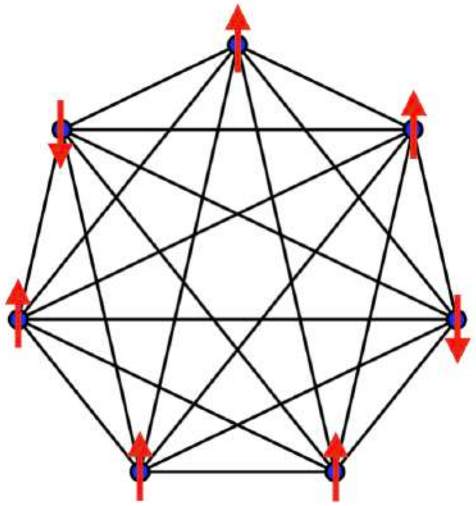
Curie-Weiss Model

$$H(s) = -\frac{J}{2n}\sum_{i,j=1}^{n} s_i s_j \qquad s \in \{-1, +1\}^n$$

Note that we can rewrite:

$$H(s) = -\frac{nJ}{2}\left(\frac{1}{n}\sum_{i=1}^{n} s_i\right)^2 = -nJm^2$$

"magnetisation"



"fully connected" or "complete"

$$\mathbb{P}[\bar{s}_n = m] = \frac{\Omega(m,n)}{Z_n(\beta)} e^{\frac{\beta n}{2}m^2}$$

$$\Omega(m,n) = \frac{n!}{\left(\frac{n(1-m)}{2}\right)!\left(\frac{n(1+m)}{2}\right)!}$$

# of configurations with $\bar{s}_n = \sum_{i=1}^{n} \frac{s_i}{n} = m$

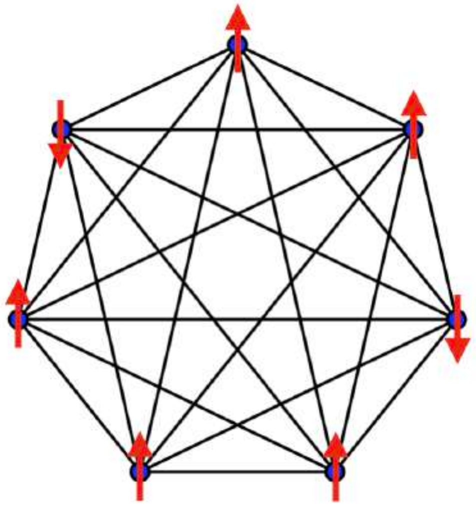
Curie-Weiss model

Curie-Weiss Model

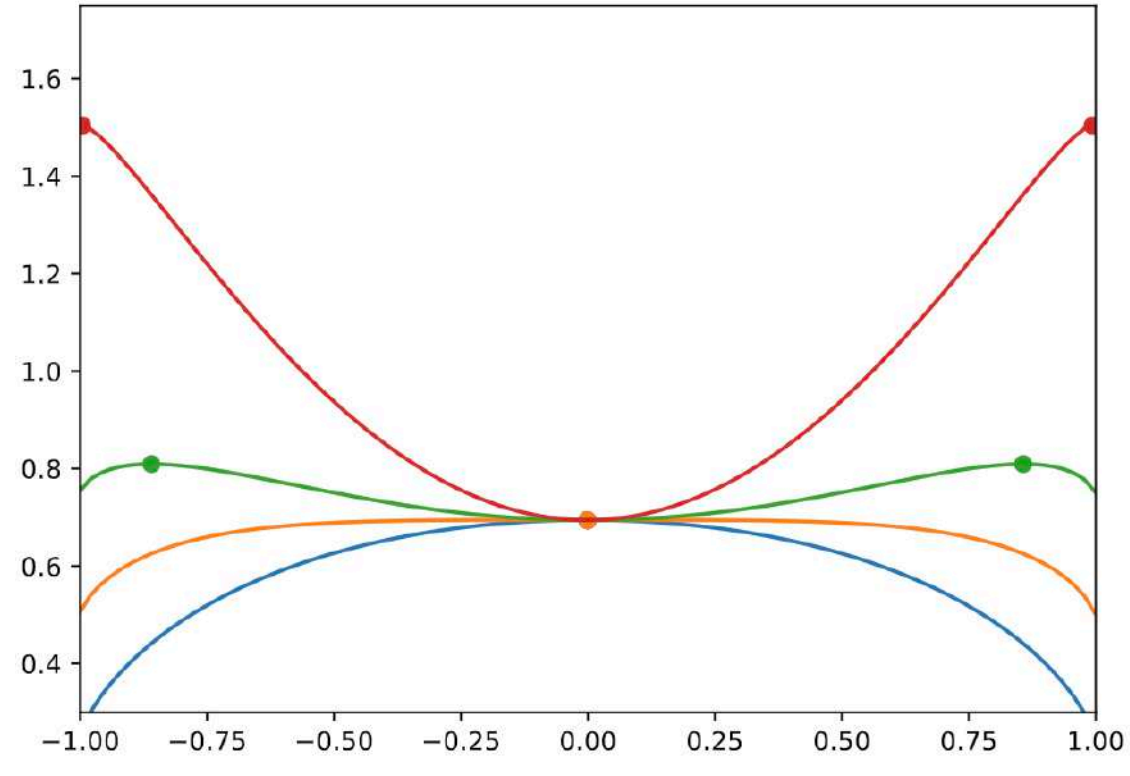
$$H(s) = -\frac{J}{2n}\sum_{i,j=1}^{n} s_i s_j \qquad s \in \{-1,+1\}^n$$

Theorem

$$\log \frac{1}{n} \lim_{n\to\infty} \mathbb{P}[\bar{s}_n = m] = \phi_\beta(m) - \phi_\beta(m^\star)$$



$\phi_\beta(m)$



m

$\beta = 3$

$\beta = 1.5$

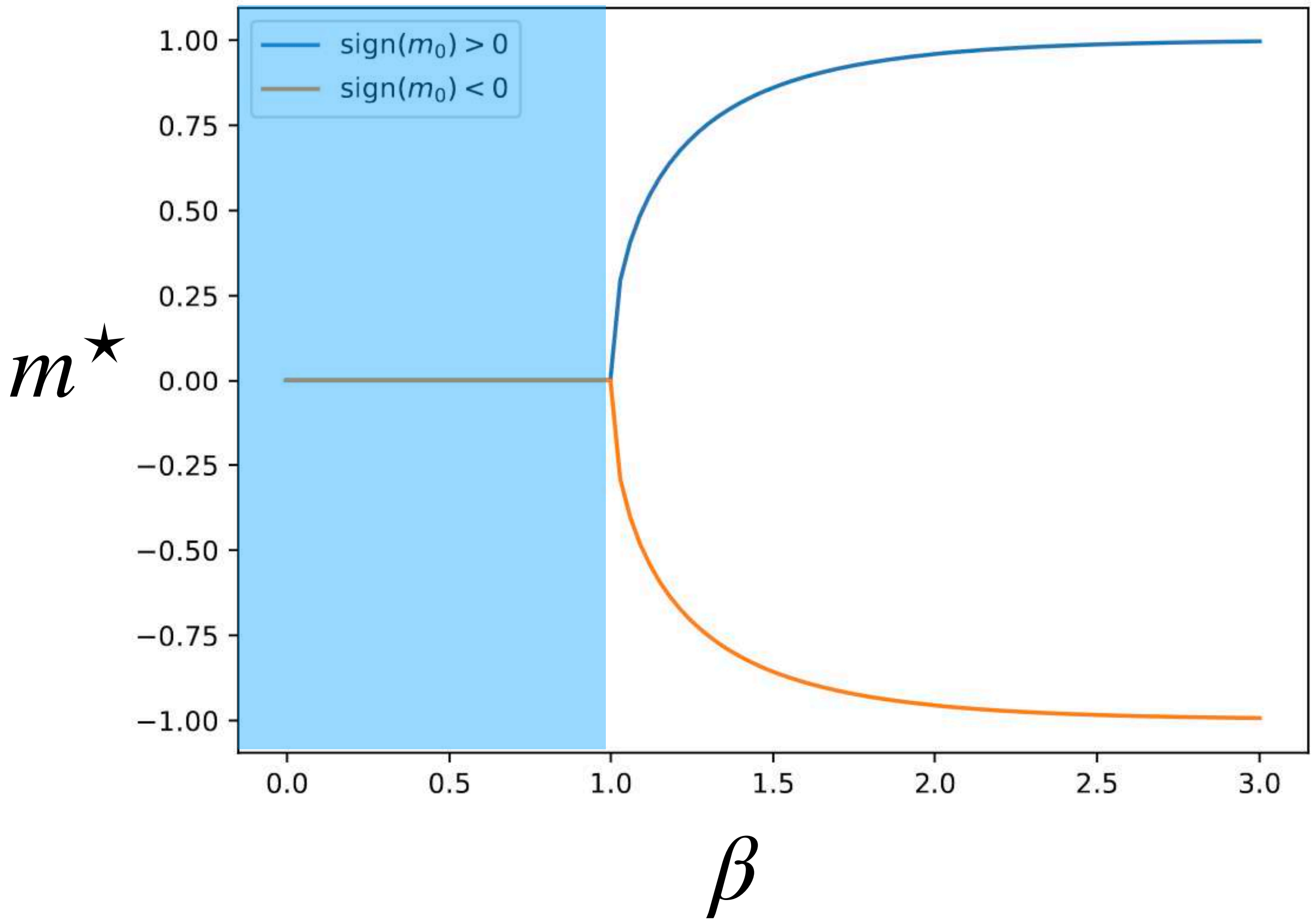
$\beta = 1$

$\beta = 0.5$

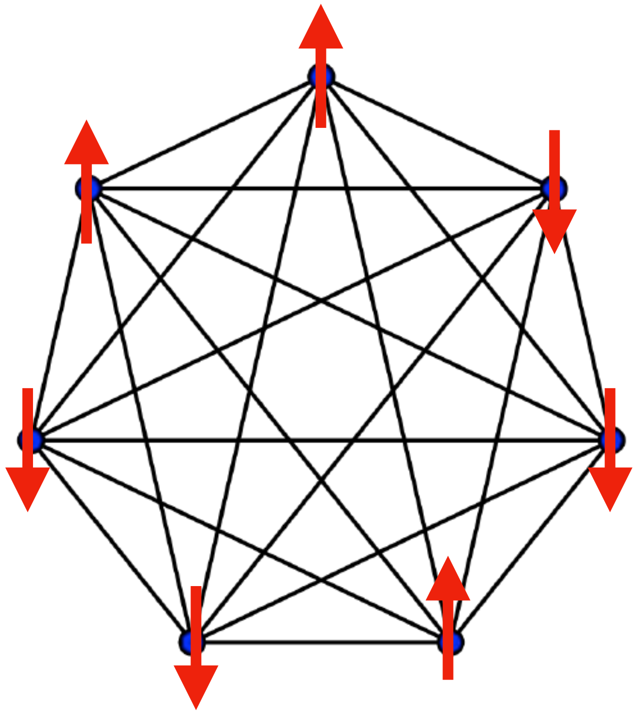
Curie-Weiss model

Curie-Weiss Model

$$H(s) = -\frac{J}{2n}\sum_{i,j=1}^{n} s_i s_j \qquad s \in \{-1, +1\}^n$$



$\beta < \beta_c$

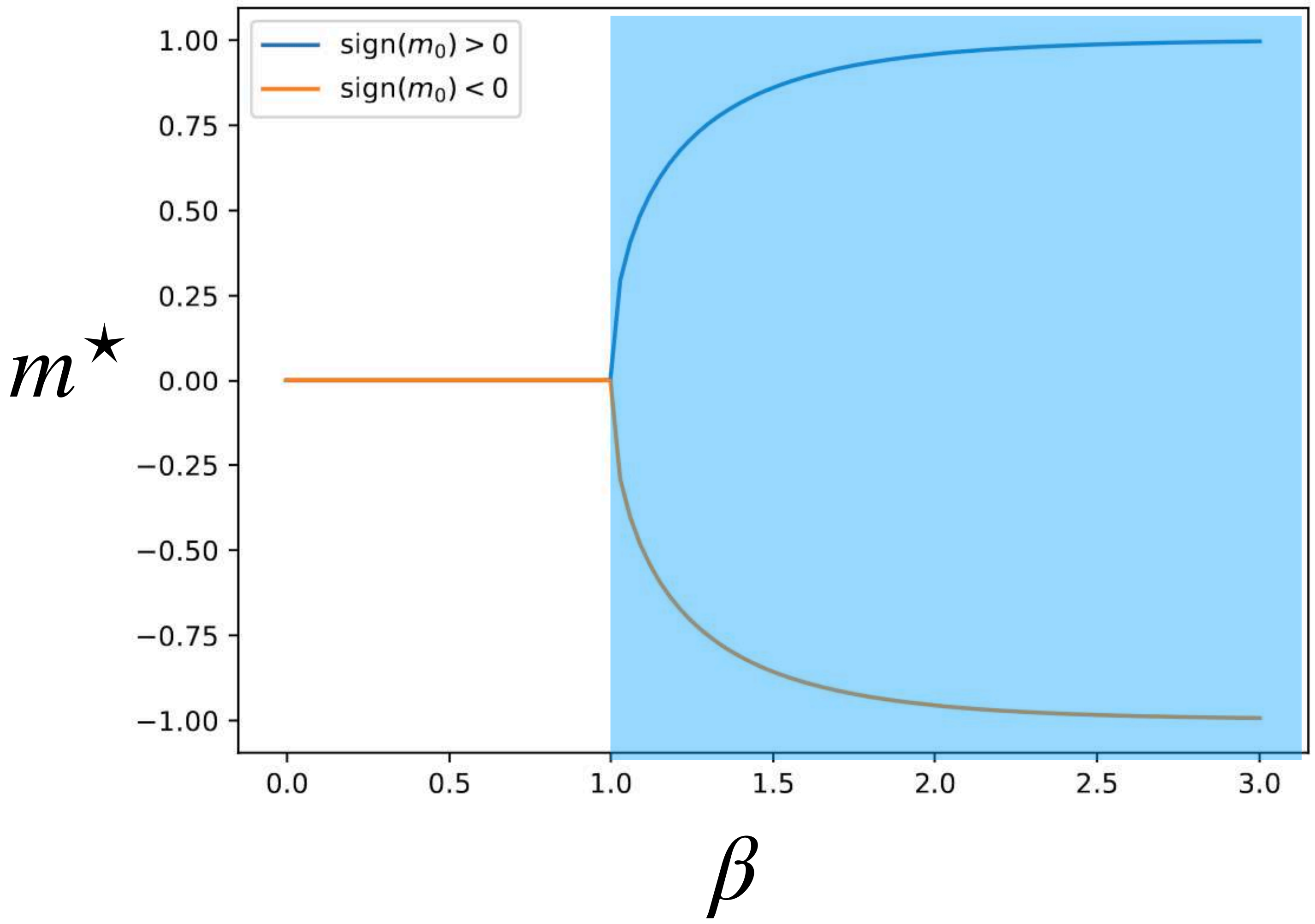


Paramagnetic
"disordered"
phase $m^\star = 0$

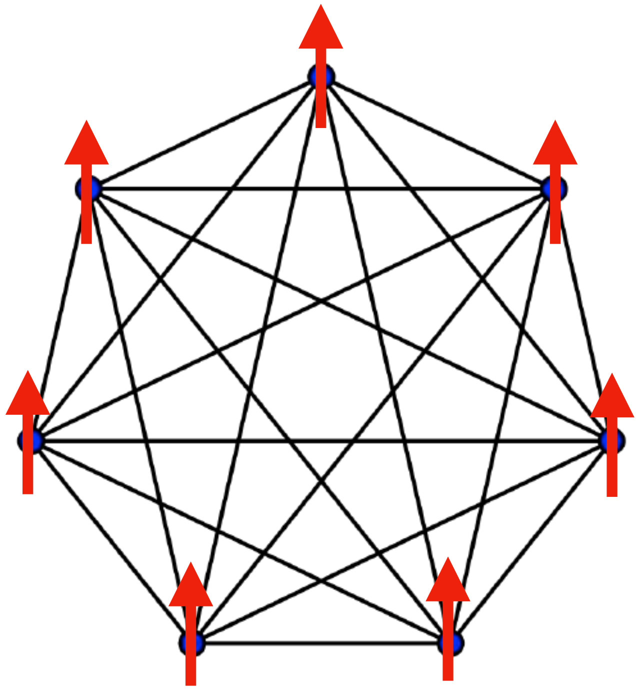
Curie-Weiss model

Curie-Weiss Model

$$H(s) = -\frac{J}{2n}\sum_{i,j=1}^{n} s_i s_j \qquad s \in \{-1,+1\}^n$$



$\beta > \beta_c$



Ferromagnetic
"ordered" phase
$|m^\star| > 0$

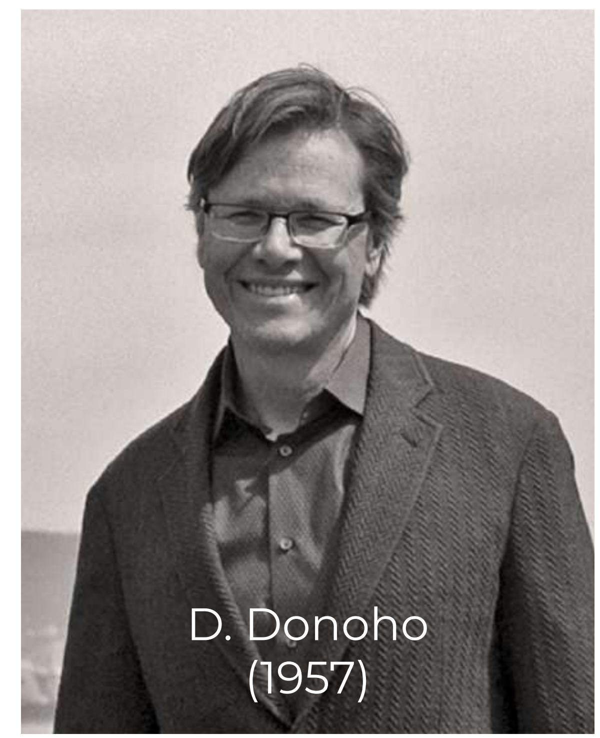
Blessing of dimensionality

High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality

Mathematicians are ideally prepared for appreciating the abstract issues involved in finding patterns in such high-dimensional data. Two of the most influential principles in the coming century will be principles originally discovered and cultivated by mathematicians: the blessings of dimensionality and the curse of dimensionality.

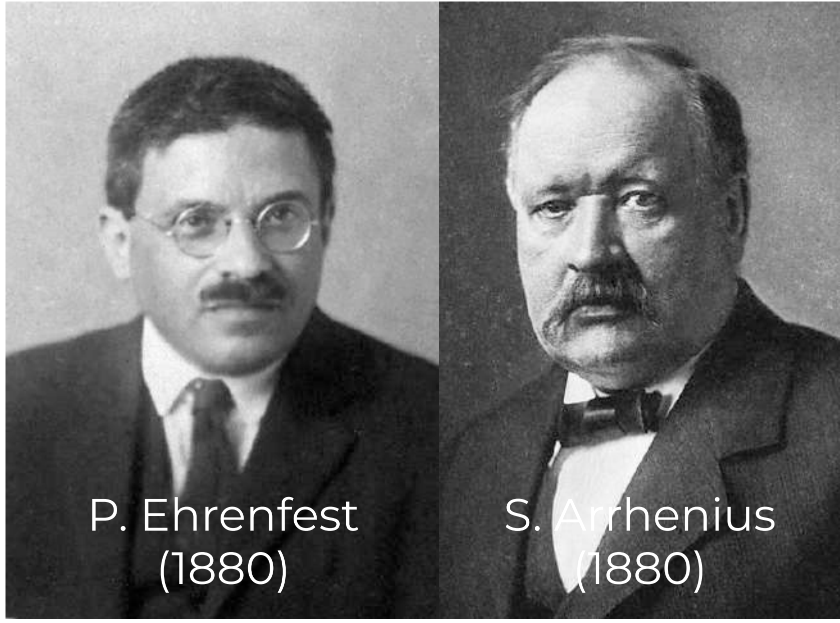
The curse of dimensionality is a phrase used by several subfields in the mathematical sciences; I use it here to refer to the apparent intractability of systematically searching through a high-dimensional space, the apparent intractability of accurately approximating a general high-dimensional function, the apparent intractability of integrating a high-dimensional function.

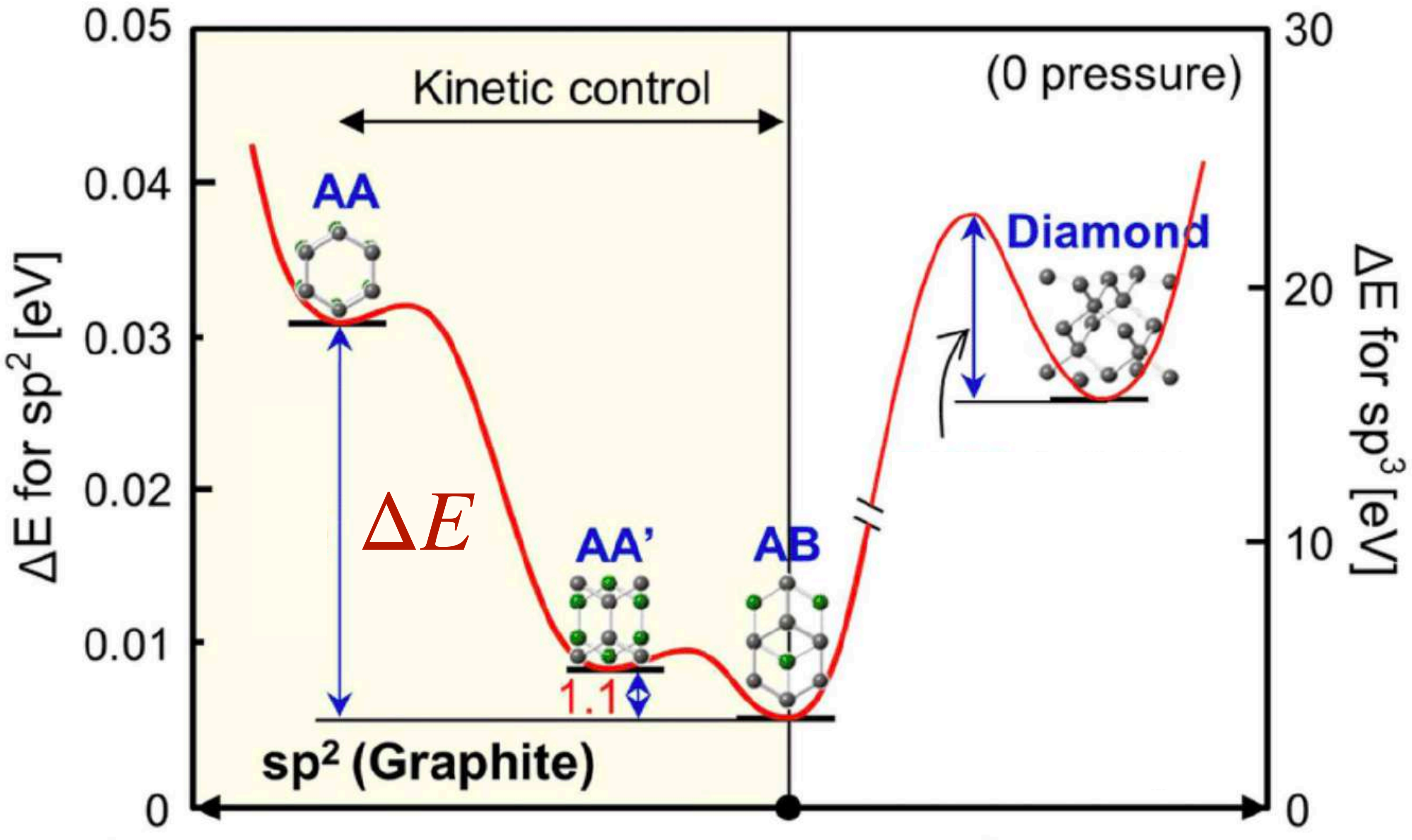
The blessings of dimensionality are less widely noted, but they include the concentration of measure phenomenon (so-called in the geometry of Banach spaces), which means that certain random fluctuations are very well controlled in high dimensions and the success of asymptotic methods, used widely in mathematical statistics and statistical physics, which suggest that statements about very high-dimensional settings may be made where moderate dimensions would be too complicated.



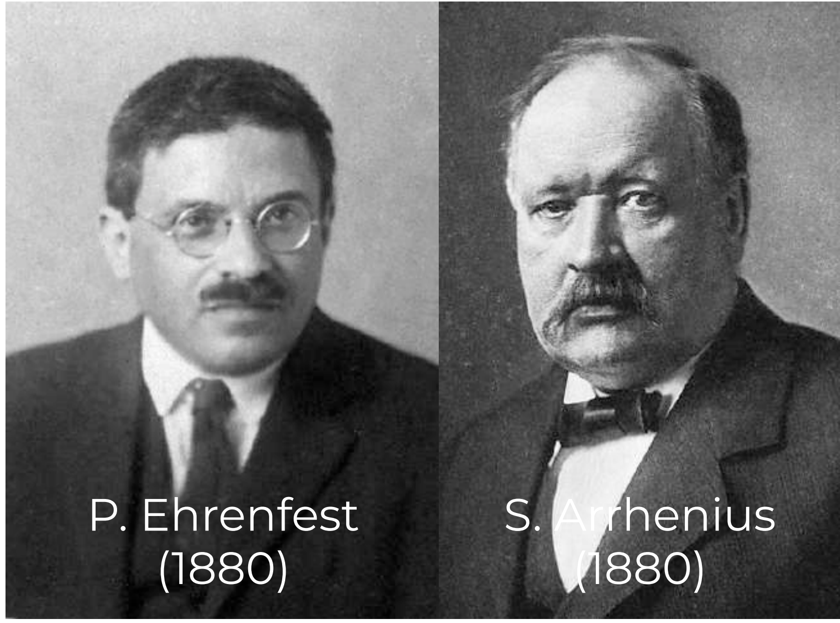
David Donoho, *AMS CONFERENCE ON MATH CHALLENGES OF THE 21ST CENTURY, 2000*

The energy landscape

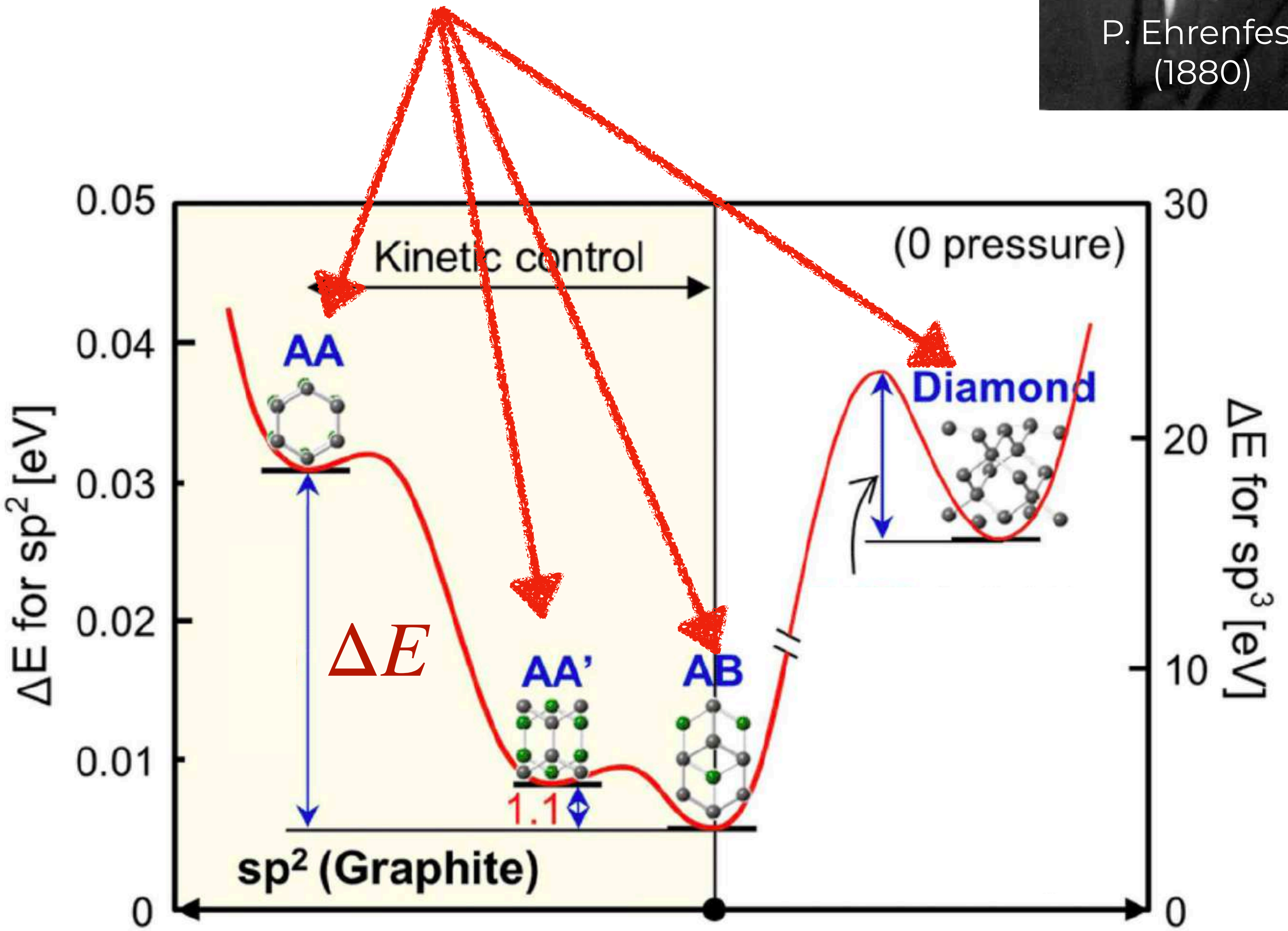




The energy landscape

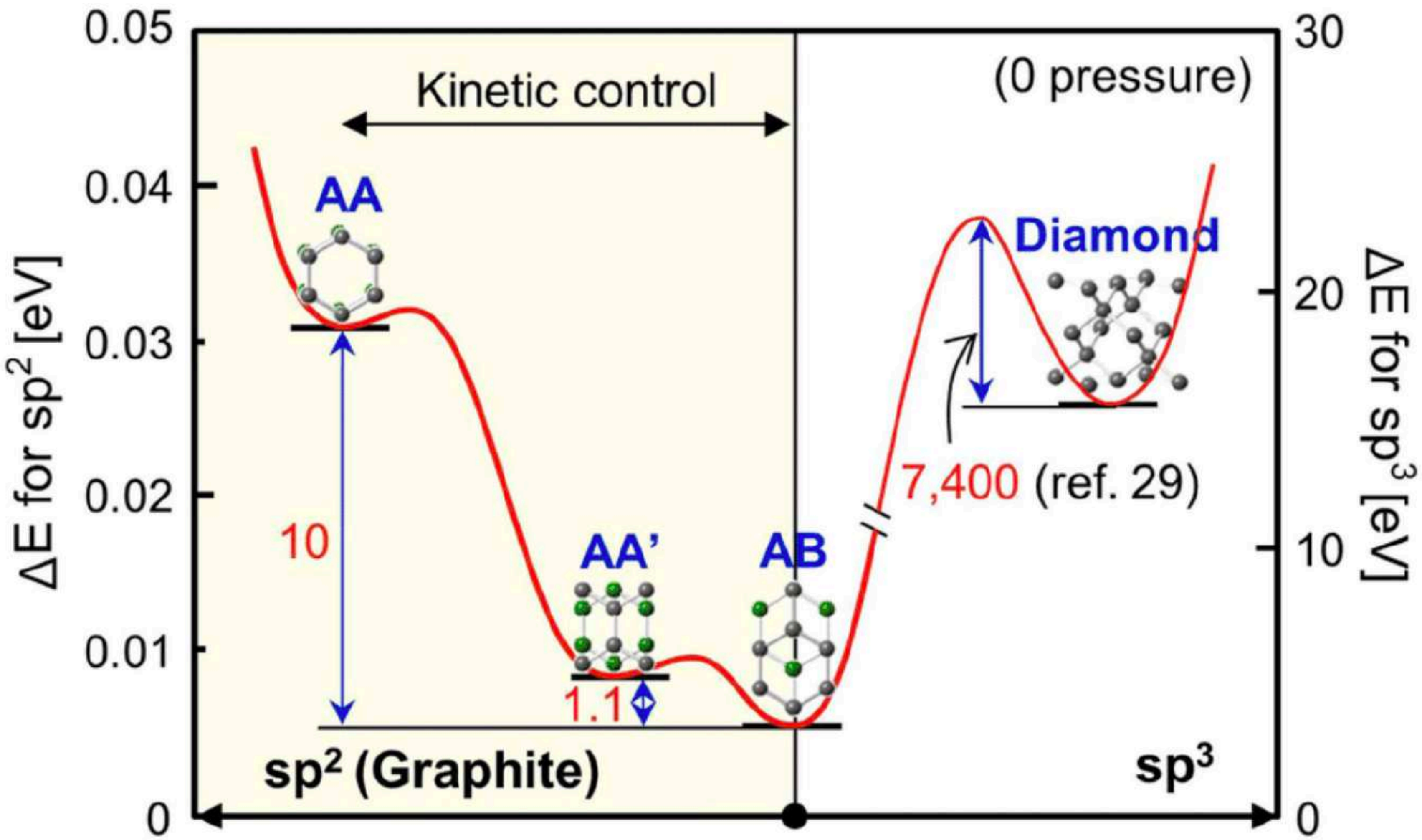


Phases
(minima of energy)



$$\tau \propto e^{\Delta E}$$ "Arrhenius law"

The energy landscape



Particles	⟷	Parameters
Hamiltonian	⟷	Loss function
Interactions	⟷	Data
Dynamics	⟷	Algorithms

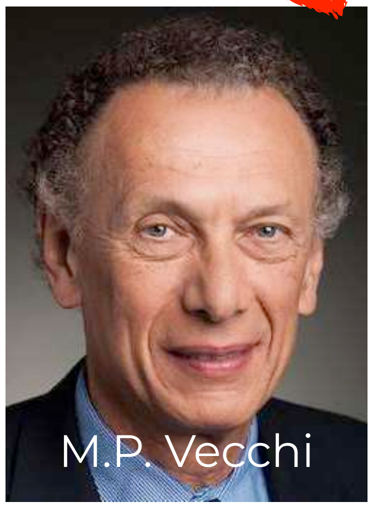
Simulated annealing

1983

13 May 1983, Volume 220, Number 4598 **SCIENCE**

Optimization by Simulated Annealing

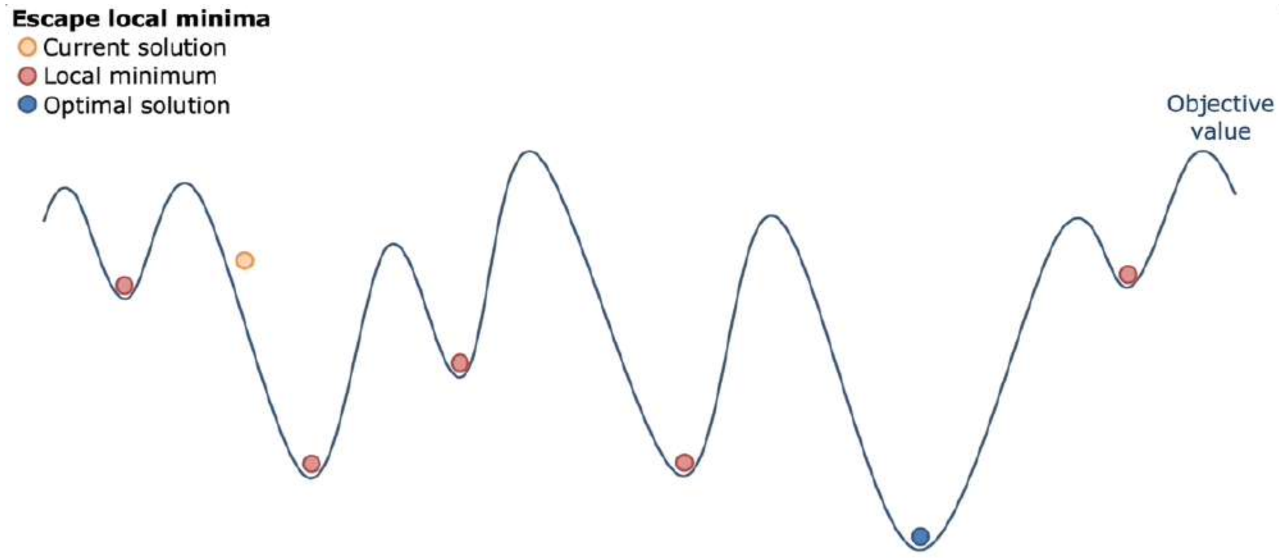
S. Kirkpatrick, C. D. Gelatt, Jr., M. P. Vecchi



Summary. There is a deep and useful connection between statistical mechanics (the behavior of systems with many degrees of freedom in thermal equilibrium at a finite temperature) and multivariate or combinatorial optimization (finding the minimum of a given function depending on many parameters). A detailed analogy with annealing in solids provides a framework for optimization of the properties of very large and complex systems. This connection to statistical mechanics exposes new information and provides an unfamiliar perspective on traditional optimization problems and methods.

The analogy between cooling a fluid and optimization may fail in one important respect. In ideal fluids all the atoms are alike and the ground state is a regular crystal. A typical optimization problem will contain many distinct, noninterchangeable elements, so a regular solution is unlikely.

The physical properties of spin glasses at low temperatures provide a possible guide for understanding the possibilities of optimizing complex systems subject to conflicting (frustrating) constraints.



The Hopfield Model

1982

Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 2554–2558, April 1982
Biophysics

Neural networks and physical systems with emergent collective computational abilities

(associative memory/parallel processing/categorization/content-addressable memory/fail-soft devices)

J. J. HOPFIELD

Division of Chemistry and Biology, California Institute of Technology, Pasadena, California 91125; and Bell Laboratories, Murray Hill, New Jersey 07974

Contributed by John J. Hopfield, January 15, 1982

ABSTRACT **Computational properties of use to biological organisms or to the construction of computers can emerge as collective properties of systems having a large number of simple equivalent components (or neurons). The physical meaning of content-addressable memory is described by an appropriate phase space flow of the state of a system. A model of such a system is given, based on aspects of neurobiology but readily adapted to integrated circuits. The collective properties of this model produce a content-addressable memory which correctly yields an entire memory from any subpart of sufficient size. The algorithm for the time evolution of the state of the system is based on asynchronous parallel processing. Additional emergent collective properties include some capacity for generalization, familiarity recognition, categorization, error correction, and time sequence retention. The collective properties are only weakly sensitive to details of the modeling or the failure of individual devices.**

calized content-addressable memory or categorizer using extensive asynchronous parallel processing.

The general content-addressable memory of a physical system

Suppose that an item stored in memory is "H. A. Kramers & G. H. Wannier *Phys. Rev.* **60**, 252 (1941)." A general content-addressable memory would be capable of retrieving this entire memory item on the basis of sufficient partial information. The input "& Wannier, (1941)" might suffice. An ideal memory could deal with errors and retrieve this reference even from the input "Vannier, (1941)". In computers, only relatively simple forms of content-addressable memory have been made in hardware (10, 11). Sophisticated ideas like error correction in accessing information are usually introduced as software (10).

There are classes of physical systems whose spontaneous behavior can be used as a form of general (and error-correcting)

The Hopfield Model

1982

$$H(s) = -\frac{1}{2}\sum_{i,j=1}^{d} J_{ij}s_is_j \ \left(= \langle s, Js\rangle\right)$$

$$s \in \{-1, +1\}^d$$

"configurations"

The Hopfield Model

1982

$$H(s) = -\frac{1}{2}\sum_{i,j=1}^{d} J_{ij}s_is_j \ \left(= \langle s, Js \rangle \right)$$

$$s \in \{-1, +1\}^d$$

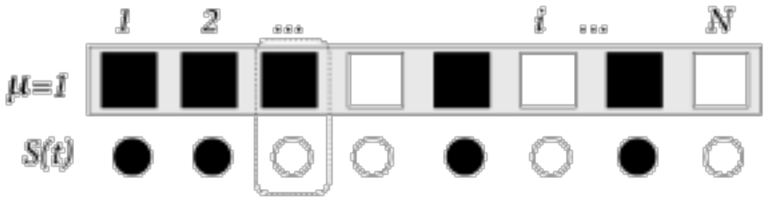
"configurations"

$$J_{ij} = \frac{1}{n}\sum_{\mu=1}^{n} x_i^\mu x_j^\mu \ \left(= \frac{1}{n}X^\top X \right)$$

"Hebbian rule"

$$x^\mu \sim \text{Unif}(\{-1, +1\}^d)$$

"patterns"

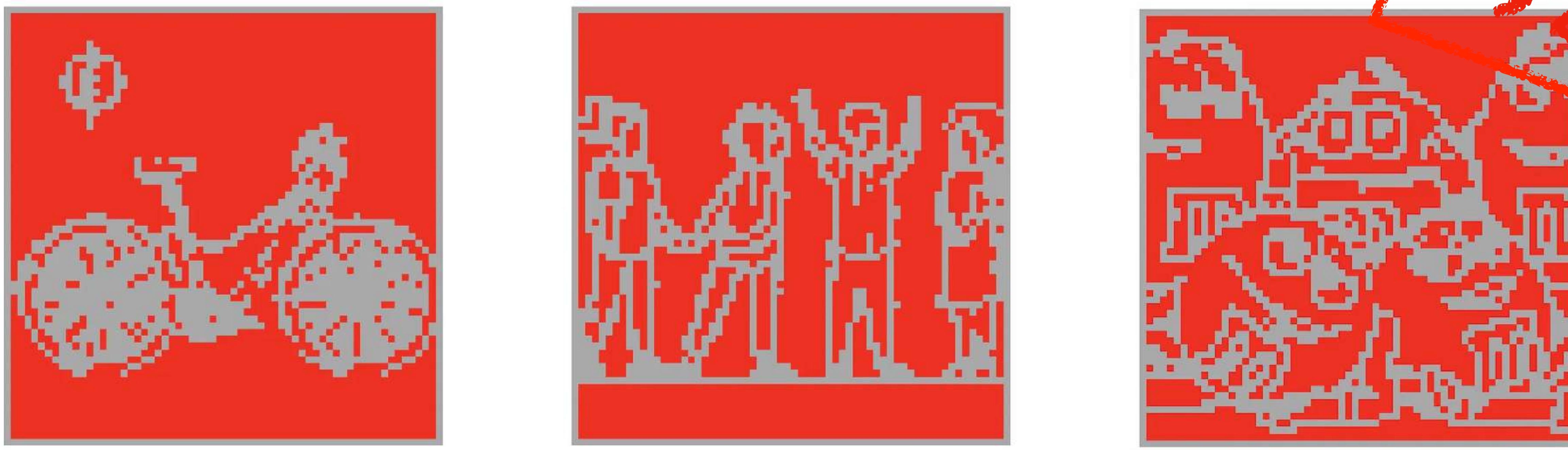


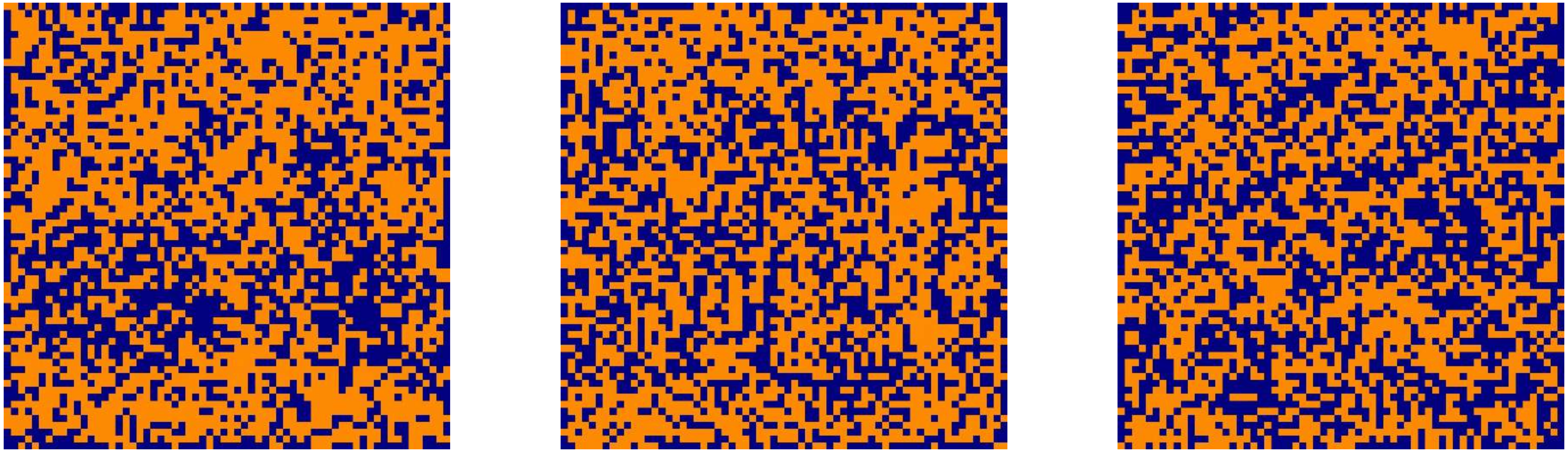
$$s(t+1) = \tanh(\beta J s(t))$$

GD-like algorithm
(Goes down in energy)

The Hopfield Model

1982





The Hopfield Model

1985

PHYSICAL REVIEW A VOLUME 32, NUMBER 2 AUGUST 1985

Spin-glass models of neural networks

Daniel J. Amit and Hanoch Gutfreund
Racah Institute of Physics, Hebrew University, 91904 Jerusalem, Israel

H. Sompolinsky
Department of Physics, Bar-Ilan University, 52100 Ramat-Gan, Israel
(Received 22 March 1985)

Two dynamical models, proposed by Hopfield and Little to account for the collective behavior of neural networks, are analyzed. The long-time behavior of these models is governed by the statistical mechanics of infinite-range Ising spin-glass Hamiltonians. Certain configurations of the spin system, chosen at random, which serve as memories, are stored in the quenched random couplings. The present analysis is restricted to the case of a finite number p of memorized spin configurations, in the thermodynamic limit. We show that the long-time behavior of the two models is identical, for all temperatures below a transition temperature T_c. The structure of the stable and metastable states is displayed. Below T_c, these systems have $2p$ ground states of the Mattis type: Each one of them is fully correlated with one of the stored patterns. Below $T \sim 0.46T_c$, additional dynamically stable states appear. These metastable states correspond to specific mixings of the embedded patterns. The thermodynamic and dynamic properties of the system in the cases of more general distributions of random memories are discussed.

The Hopfield Model

1985

PHYSICAL REVIEW A VOLUME 32, NUMBER 2 AUGUST 1985

Spin-glass models of neural networks

Daniel J. Amit and Hanoch Gutfreund
Racah Institute of Physics, Hebrew University, 91904 Jerusalem, Israel

H. Sompolinsky
Department of Physics, Bar-Ilan University, 52100 Ramat-Gan, Israel
(Received 22 March 1985)

Two dynamical models, proposed by Hopfield and Little to account for the collective behavior of neural networks, are analyzed. The long-time behavior of these models is governed by the statistical mechanics of infinite-range Ising spin-glass Hamiltonians. Certain configurations of the spin system, chosen at random, which serve as memories, are stored in the quenched random couplings. The present analysis is restricted to the case of a finite number p of memorized spin configurations, in the thermodynamic limit. We show that the long-time behavior of the two models is identical, for all temperatures below a transition temperature T_c. The structure of the stable and metastable states is displayed. Below T_c, these systems have $2p$ ground states of the Mattis type: Each one of them is fully correlated with one of the stored patterns. Below $T \sim 0.46T_c$, additional dynamically stable states appear. These metastable states correspond to specific mixings of the embedded pat-

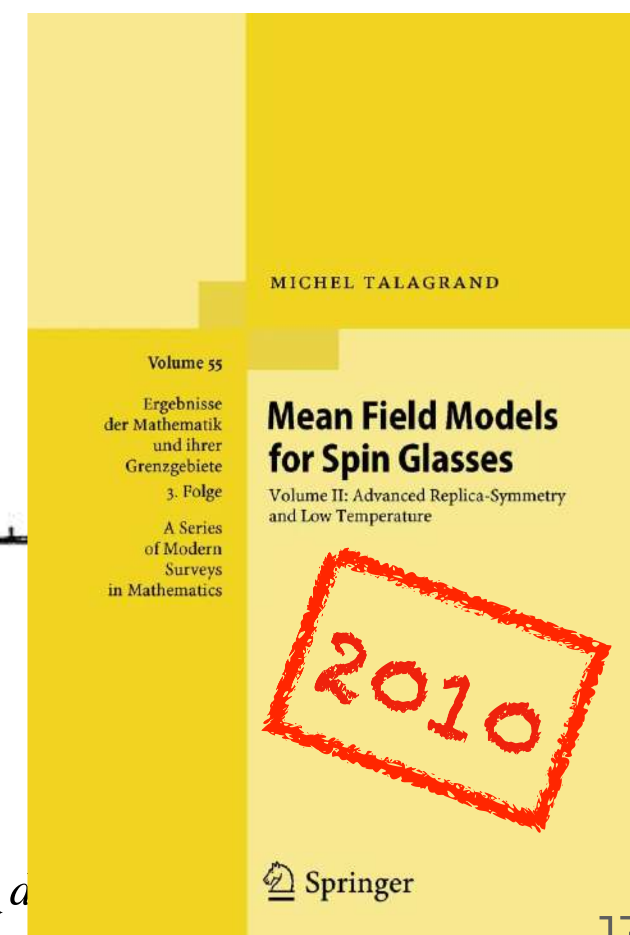
$k \leq M$. This however is not really interesting. The fascinating fact is that when N is large and $M/N \simeq \alpha$, if $\alpha > 2$ the set $\mathbb{S}_N \cap_{k \leq M} U_k$ is typically empty (a classical result), while if $\alpha < 2$, with probability very close to 1, we have

$$\frac{1}{N} \log \mu_N \Big(\mathbb{S}_N \bigcap_{k \leq M} U_k \Big) \simeq \mathrm{RS}(\alpha) \,. \qquad (0.2)$$

Here,

$$\mathrm{RS}(\alpha) = \min_{0<q<1} \left(\alpha \, \mathsf{E} \log \mathcal{N}\left(\frac{z\sqrt{q}}{\sqrt{1-q}} \right) + \frac{1}{2} \frac{q}{1-q} + \frac{1}{2} \log(1-q) \right) ,$$

where $\mathcal{N}(x)$ denotes the probability that a standard Gaussian r.v. g is $\geq x$, and where $\log x$ denotes (as everywhere through the book) the natural logarithm of x. Of course you should rush to require medical attention if this formula seems transparent to you. We simply give it now to demonstrate



2010

And they were not alone...

1985

ÉCOLE DE PHYSIQUE DES HOUCHES

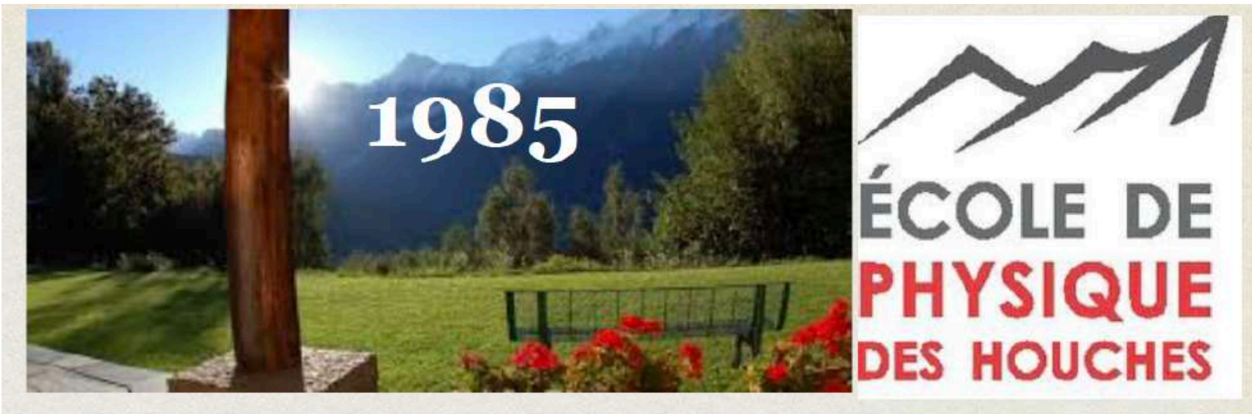
Disordered Systems and Biological Organization

"Only physicists were interested in neural networks at the time [...] My professional life truly shifted in February 1985 during a physics symposium in Les Houches, in the French Alps. There, I met the crème de la crème of international research interested in neural networks and gave my very first talk (in English!)."

From *"Quand la Machine Apprend"*

And they were not alone...

1985



Disordered Systems and Biological Organization

I benchmarked neural networks against kernel methods with my Ph.D advisors Gerard Dreyfus and Leon Personnaz. The same year, two physicists working close-by (Marc Mezard & Werner Krauth) published a paper on an optimal margin algorithm called 'minover,' which attracted my attention.... but it was not until I joined Bell Labs that I put things together and we created support vector machines.

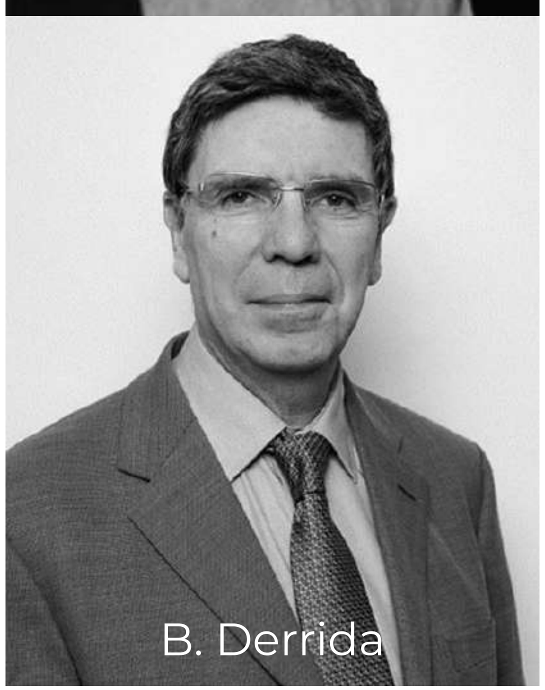
From *"Data Mining History: The Invention of Support Vector Machines"*

Towards a theory for typical-case algorithmic hardness

The Perceptron

1987


E. Gardner


B. Derrida

Optimal storage properties of neural network models

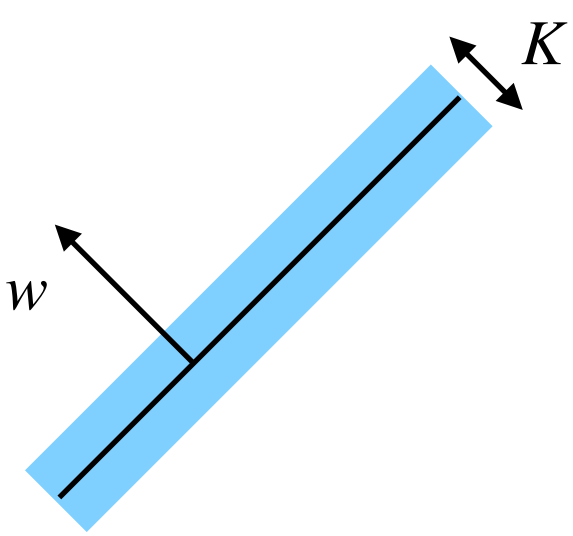
E Gardner† and B Derrida‡

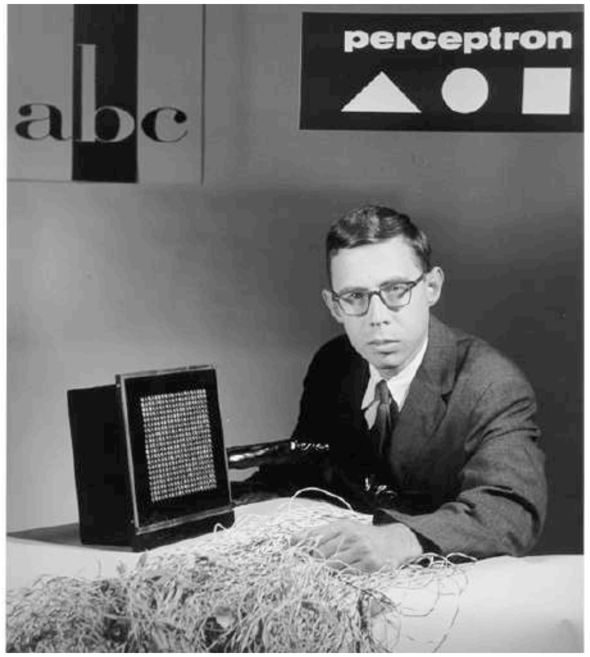
† Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK
‡ Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

Received 29 May 1987

Abstract. We calculate the number, $p=\alpha N$ of random N-bit patterns that an optimal neural network can store allowing a given fraction f of bit errors and with the condition that each right bit is stabilised by a local field at least equal to a parameter K. For each value of α and K, there is a minimum fraction $f_{\min}$ of wrong bits. We find a critical line, $\alpha_c(K)$ with $\alpha_c(0)=2$. The minimum fraction of wrong bits vanishes for $\alpha<\alpha_c(K)$ and increases from zero for $\alpha>\alpha_c(K)$. The calculations are done using a saddle-point method and the order parameters at the saddle point are assumed to be replica symmetric. This solution is locally stable in a finite region of the K,α plane including the line, $\alpha_c(K)$ but there is a line above which the solution becomes unstable and replica symmetry must be broken.

Given $(x_i, y_i)_{i\in[n]}$, wants: $y_i(w^\top x_i) \geq K$





[Rosenblatt 1958]

The Perceptron

1987

Optimal storage properties of neural network models

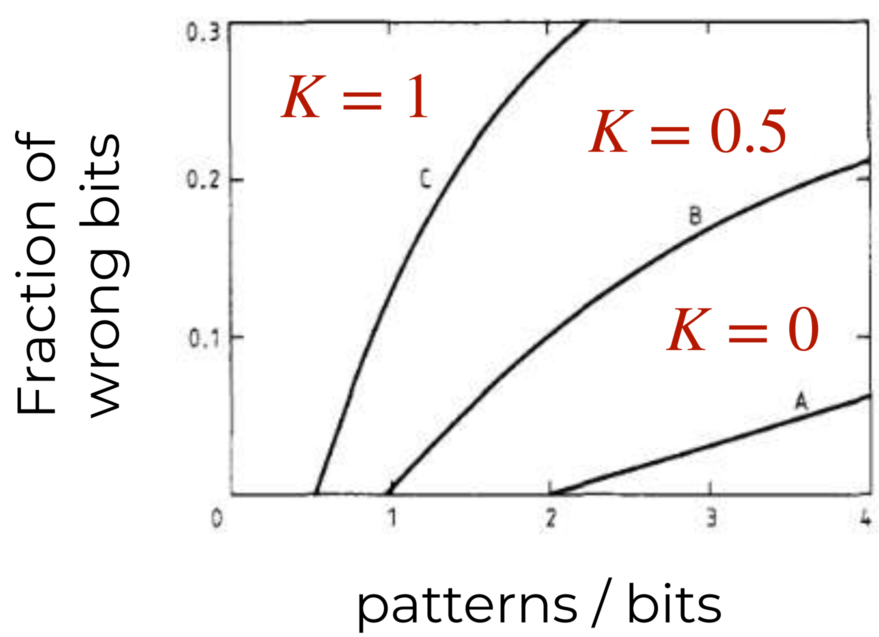
E Gardner† and B Derrida‡

† Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK
‡ Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

Received 29 May 1987

Abstract. We calculate the number, $p = \alpha N$ of random N-bit patterns that an optimal neural network can store allowing a given fraction f of bit errors and with the condition that each right bit is stabilised by a local field at least equal to a parameter K. For each value of α and K, there is a minimum fraction $f_{\min}$ of wrong bits. We find a critical line, $\alpha_c(K)$ with $\alpha_c(0) = 2$. The minimum fraction of wrong bits vanishes for $\alpha < \alpha_c(K)$ and increases from zero for $\alpha > \alpha_c(K)$. The calculations are done using a saddle-point method and the order parameters at the saddle point are assumed to be replica symmetric. This solution is locally stable in a finite region of the K, α plane including the line, $\alpha_c(K)$ but there is a line above which the solution becomes unstable and replica symmetry must be broken.

$$H(w) = \frac{1}{2}\sum_{\mu=1}^{n} \mathbb{I}\left[y^\mu \neq \mathrm{sign}(w^\top x^\mu - \kappa)\right]$$

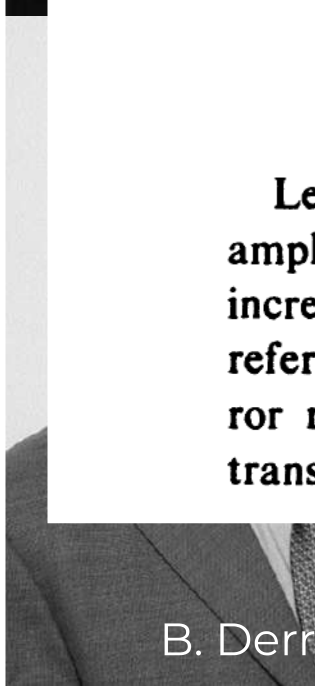


- Prefigures most of research that followed in “Statistical Physics of Learning”
- Precursor to “High-d” statistics (Donoho, Candès, Montanari, El Karoui)

c.f. [Cover 1967]

The Perceptron

1987



Optimal storage properties of neural network models

E Gardner† and B Derrida‡

† Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK
‡ Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

Received 29 May 1987

Abstract. We calculate the number, $p = \alpha N$ of random N-bit patterns that an optimal neural network can store allowing a given fraction f of bit errors and with the condition that each right bit is stabilised by a local field at least equal to a parameter K. For each value of α and K, there is a minimum fraction $f_{\min}$ of wrong bits. We find a critical line,

First-order transition to perfect generalization in a neural network with binary synapses

Géza Györgyi*
School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332-0430
(Received 9 February 1990)

Learning from examples by a perceptron with binary synaptic parameters is studied. The examples are given by a reference (teacher) perceptron. It is shown that as the number of examples increases, the network undergoes a first-order transition, where it freezes into the state of the reference perceptron. When the transition point is approached from below, the generalization error reaches a minimal positive value, while above that point the error is constantly zero. The transition is found to occur at $\alpha_{GD} = 1.245$ examples per coupling.

ratio, $\alpha = p/N$, of the value $\kappa(>0)$ of the product of the spin and the magnetic field at each site and of the magnetisation, m. Here m may vary between 0 (no correlation) and 1 (completely correlated). The capacity increases with the correlation between patterns from $\alpha = 2$ for correlated patterns with $\kappa = 0$ and tends to infinity as m tends to 1. The calculations use a saddle-point method and the order parameters at the saddle point are assumed to be replica symmetric. This solution is shown to be locally stable. A local iterative learning algorithm for updating the interactions is given which will converge to a solution of given κ provided such solutions exist.

c.f. [Cover 1967]

The Perceptron

1987

Optimal storage properties of neural network models

E Gardner† and B Derrida‡

† Department of Physics, Edinburgh University, Mayfield Road, Edinburgh, EH9 3JZ, UK
‡ Service de Physique Theorique, CEN Saclay, F 91191 Gif sur Yvette, France

Received 29 May 1987

Abstract. We calculate the number, $p = \alpha N$ of random N-bit patterns that an optimal neural network can store allowing a given fraction f of bit errors and with the condition that each right bit is stabilised by a local field at least equal to a parameter K. For each value of α and K, there is a minimum fraction $f_{\min}$ of wrong bits. We find a critical line,

First-order transition to perfect generalization in a neural network with binary synapses

Géza Györgyi*
School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332-0430
(Received 9 February 1990)

Learning fr
amples are gi
increases, the
reference perc
ror reaches a
transition is fc

Learning from Examples in Large Neural Networks

H. Sompolinsky[a] and N. Tishby
AT&T Bell Laboratories, Murray Hill, New Jersey 07974

H. S. Seung
Department of Physics, Harvard University, Cambridge, Massachusetts 02138
(Received 29 May 1990)

A statistical mechanical theory of learning from examples in layered networks at finite temperature is studied. When the training error is a smooth function of continuously varying weights the generalization error falls off asymptotically as the inverse number of examples. By analytical and numerical studies of single-layer perceptrons we show that when the weights are discrete the generalization error can exhibit a discontinuous transition to perfect generalization. For intermediate sizes of the example set, the state of perfect generalization coexists with a metastable spin-glass state.

The Perceptron

1987

Optimal storage properties of neural network models

E Gardner† and B Derrida‡

UK

nal
ion
ach
ne,

The statistical mechanics of learning a rule

Timothy L. H. Watkin* and Albrecht Rau†

Department of Physics, University of Oxford, Oxford OX1 3NP, United Kingdom

Michael Biehl

Physikalisches Institut, Julius-Maximilians-Univer

A summary is presented of the statistical mechanical the
rapidly advancing area which is closely related to other in
cists. By emphasizing the relationship between neural net
such as spin glasses, the authors show how learning theo
new, exact analytical techniques.

Learning fr
amples are gi
increases, the
reference perc
ror reaches a
transition is fo

Learn

A1

a discontinuous transition
of perfect generalization c

Basins of Attraction in a Perceptron-like Neural Network

Werner Krauth
Marc Mézard
Jean-Pierre Nadal
Laboratoire de Physique Statistique,
*Laboratoire de Physique Théorique de l'E.N.S.,**
24 rue Lhomond, 75231 Paris Cedex 05, France

ork of the per-
ters which ren-
s of attraction)
s and study the
size of the basins of attraction (the maximal allowable noise level still ensuring recognition) for sets of random patterns. The relevance of our results to the perceptron's ability to generalize are pointed out, as is the role of diagonal couplings in the fully connected Hopfield model.

Information storage and retrieval in synchronous neural networks

José F. Fontanari and R. Köberle

Phys. Rev. A **36**, 2475 – Published 1 September 1987

The Perceptron

1987

Optimal storage properties of neural network models

E Gardner† and B Derrida‡

UK

UK

The statistical mechanics of learning a rule

Timothy L. H. Watkin* and Albrecht Rau†

Department of Physics, University of Oxford, Oxford OX1 3NP, United Kingdom

Michael Biehl

Physikalisches Institut, Julius-Maximilians-Univer...

A summary is presented of the statistical mechanical the... rapidly advancing area which is closely related to other in... cists. By emphasizing the relationship between neural net... such as spin glasses, the authors show how learning theo... new, exact analytical techniques.

Learning from examples by...

Learning from Examples in Large ...

H. Sompolinsky[a] and N. T...

AT&T Bell Laboratories, Murray Hill, N...

Basins of Attraction in a Perceptron-like Neural Network

Werner Krauth
Marc Mézard
Jean-Pierre Nadal
Laboratoire de Physique Statistique,
*Laboratoire de Physique Théorique de l'E.N.S.,**
24 rue Lhomond, 75231 Paris Cedex 05, France

Information storage and retrieval in synchronous neural networks

José F. Fontanari and R. Köberle

Phys. Rev. A **36**, 2475 – Published 1 September 1987

A stat... studied. ... error falls off asymptotically as the inverse number of examples. ... single-layer perceptrons we show that when the weights are disc... a discontinuous transition to perfect generalization. For interme... of perfect generalization coexists with a metastable spin-glass sta...

...vork of the per- ...ters which ren- ...s of attraction) ...s and study the size of the basins of attraction (the maximal allowable noise level still ensuring recognition) for sets of random patterns. The relevance of our results to the perceptron's ability to generalize are pointed out, as is the role of diagonal couplings in the fully connected Hopfield model.

The CSP years

90's–10's

These works have triggered a wave of interest of Physicists for TCS, in particular random constraint satisfaction problems (CSP)

- Travelling Salesman Problem: Kirkpatrick **1981**, Mézard, Parisi 1985.
- Graph Colouring: Wu **1982**; Biroli, Monasson, Weigt 1999; Mulet, Pagnani, Weigt, Zecchina 2003
- Graph Matching Problem: Parisi, Mézard **1987**
- Error correcting codes: Sourlas **1989**
- K-SAT: Monasson, Zecchina **1997**; Mézard, Zecchina, Parisi 2002
- Compressive sensing: Donoho, Maleki, Montanari **2009**
- Stochastic Block Model: Decelle, Krzakala, Moore, Zdeborová **2011**

1995

Leo Breiman
Statistics Department, University of California, Berkeley, CA 94305;
e-mail: leo@stat.berkeley.edu

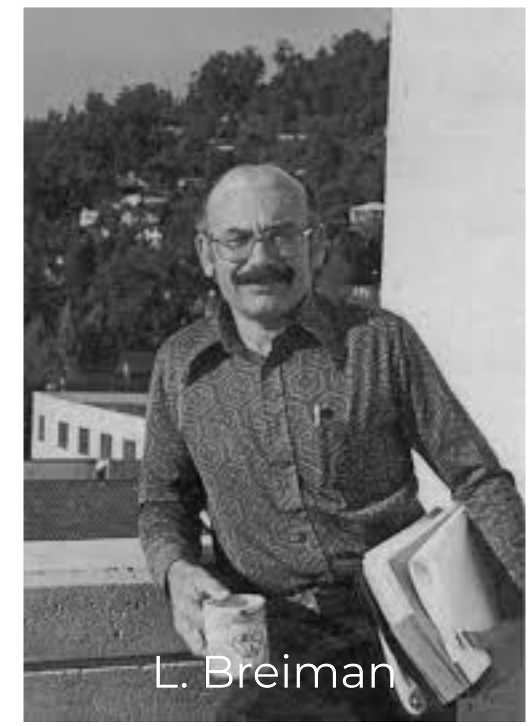
Reflections After Refereeing Papers for NIPS

Our fields would be better off with far fewer theorems, less emphasis on faddish stuff, and much more scientific inquiry and engineering. But the latter requires real thinking.

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?
- When should one stop the backpropagation and use the current parameters?

Mathematical theory is not critical to the development of machine learning. *But scientific inquiry is.*



Leo Breiman
Statistics Department, University of California, Berkeley, CA 94305;
e-mail: leo@stat.berkeley.edu

1995

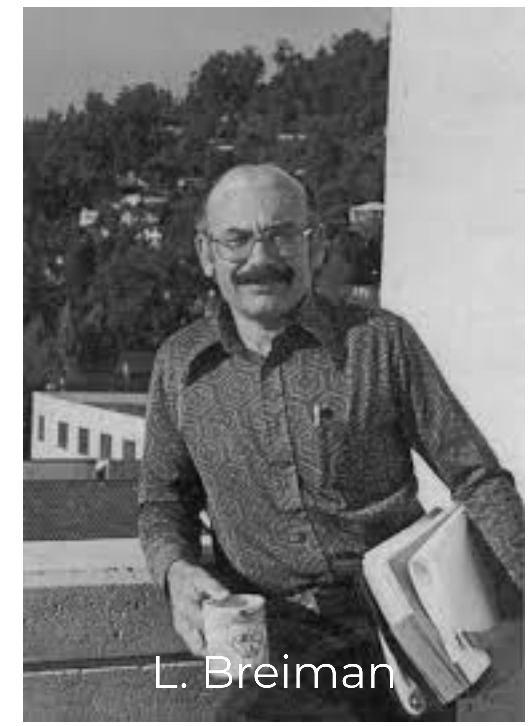
Reflections After Refereeing Papers for NIPS

Our fields would be better off with far fewer theorems, less emphasis on faddish stuff, and much more scientific inquiry and engineering. But the latter requires real thinking.

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?
- When should one stop the backpropagation and use the current parameters?

Mathematical theory is not critical to the development of machine learning. *But scientific inquiry is.*



3.5 INQUIRY

INQUIRY = sensible and intelligent efforts to understand what is going on. For example:

- mathematical heuristics
- simplified analogies (like the Ising Model)
- simulations
- comparisons of methodologies
- devising new tools
- theorems where useful (rare!)
- shunning panaceas

Case study: the GLM

Signal	Likelihood	Observation	
$w_\star \sim P_\star$ $\longrightarrow$	$P(y \mid Xw_\star)$ $\longrightarrow$	$y \in \mathbb{R}^n$	👀
	X Gaussian.		

Case study: the GLM

Signal	Likelihood	Observation
$w_\star \sim P_\star$ $\longrightarrow$	$P(y \mid X w_\star)$ $\longrightarrow$	$y \in \mathbb{R}^n$ 👀
	X Gaussian.	

$$\text{mmse} = \text{argmin}\ \mathbb{E}[\,||w - w_\star||_2^2] = \mathbb{E}[w \mid X, y]$$

$$p(w \mid X, y) \propto P_\star(w) \prod_{i=1}^{n} P(y_i \mid \langle w, x_i \rangle)$$

Posterior distribution

Case study: the GLM

[Barbier, Krzakala, Macris, Miolane, Zdeborová '17]

Signal | Likelihood | Observation

$$w_\star \sim P_\star \longrightarrow P(y \mid X w_\star) \longrightarrow y \in \mathbb{R}^n$$

X Gaussian.

$$\text{mmse} = \text{argmin}\ \mathbb{E}[\,||w - w_\star||_2^2] = \mathbb{E}[w \mid X, y]$$

$$p(w \mid X, y) \propto P_\star(w) \prod_{i=1}^{n} P(y_i \mid \langle w, x_i \rangle)$$

Posterior distribution

Theorem $\text{mmse} = \rho - m^\star$, where $\rho = \text{Var}\ P_\star$ $m^\star$ minimiser of

$$\Phi(m^*, \hat{m}^*) = \sup_m \inf_{\hat{m}} \Phi_{\text{RS}}(m, \hat{m})$$

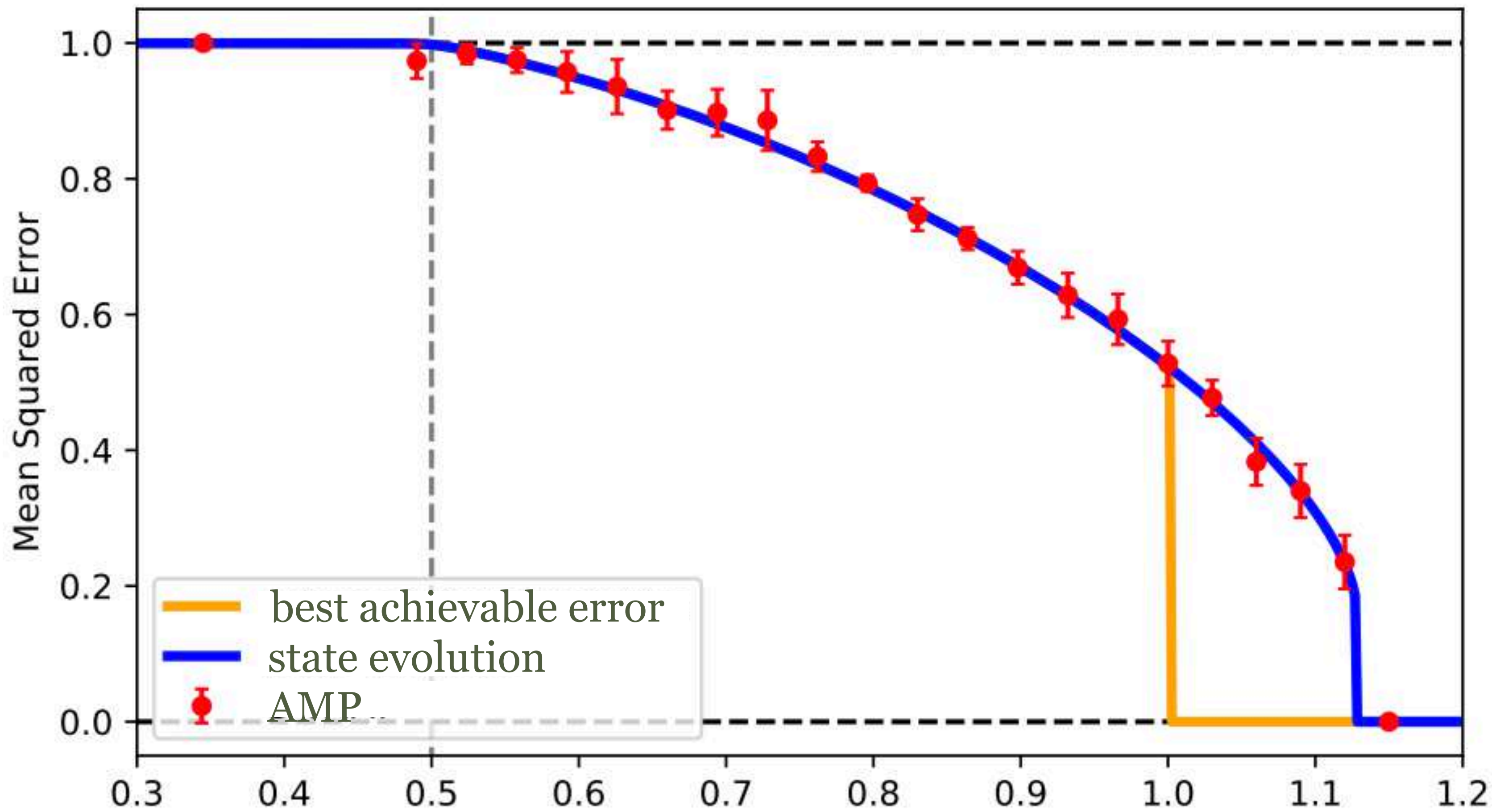
$$\Phi_{P_{\text{out}}}(m; \rho) \equiv \mathbb{E}_{v,z}\Big[\int \mathrm{d}y P_{\text{out}}(y \mid \sqrt{m}v + \sqrt{\rho - m}z) \ln \mathbb{E}_\xi[P_{\text{out}}(y \mid \sqrt{m}v + \sqrt{\rho - m}\xi)]\Big]$$

$$\Phi_{P_w}(\hat{m}) \equiv \mathbb{E}_{z,w_0}\Big[\ln \mathbb{E}_w\Big(e^{\hat{m}ww_0 + \sqrt{\hat{m}}wz - \hat{m}w^2/2}\Big)\Big]$$

Case study: the GLM

[Barbier et al. '17; Mondelli, Montanari '17; Maillard, **BL**, Krzakala, Zdeborová '20;]

$$y_i = |\langle w_\star, x_i \rangle|^2 \qquad P_\star = \mathcal{N}(0, I_d)$$

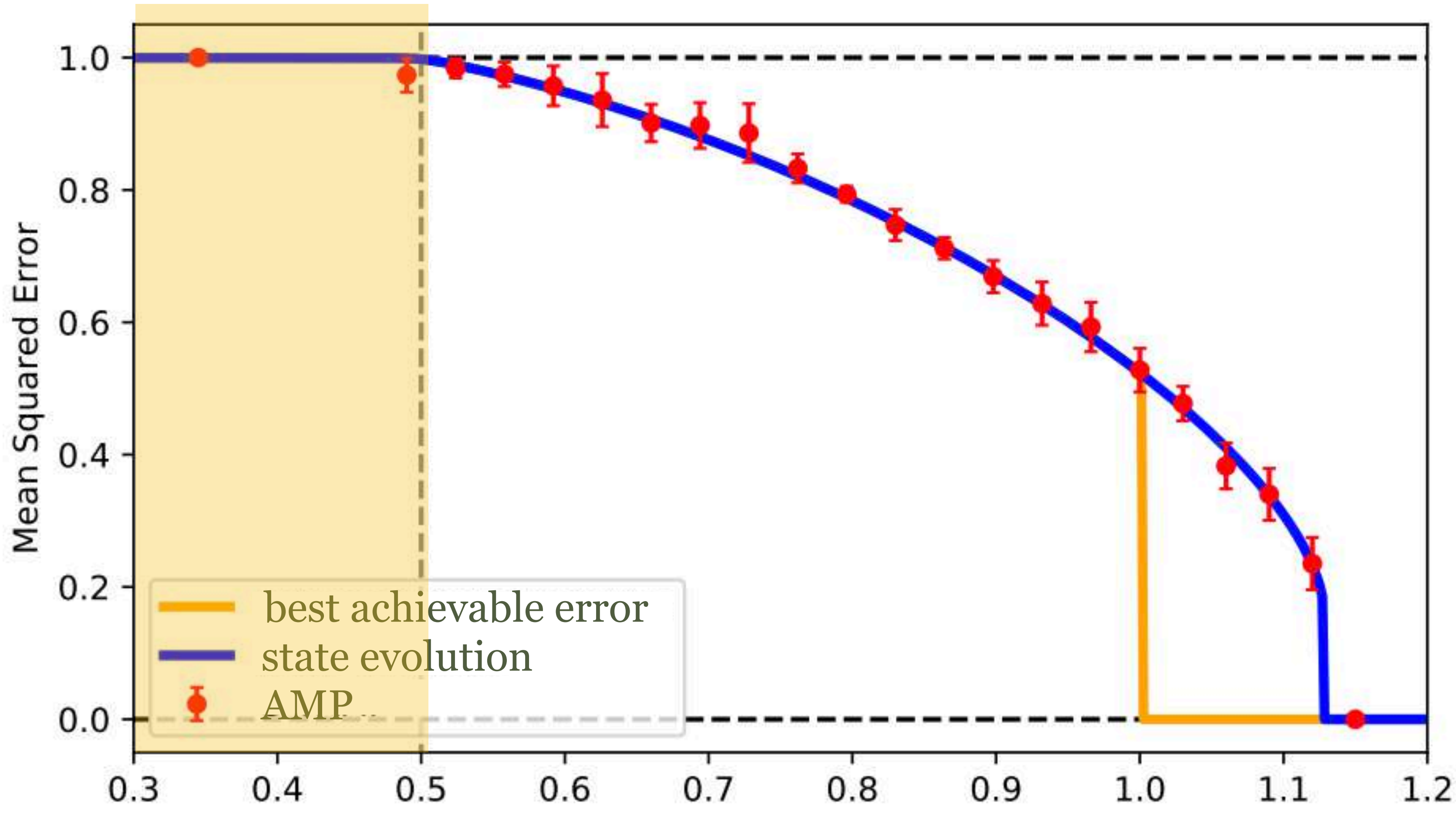


$\alpha = n/d$

Case study: the GLM

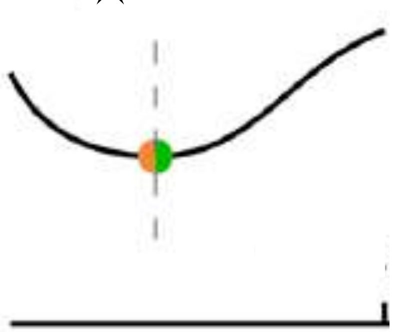
[Barbier et al. ’17; Mondelli, Montanari ’17; Maillard, **BL**, Krzakala, Zdeborová ’20;]

$$y_i = |\langle w_\star, x_i \rangle|^2 \qquad P_\star = \mathcal{N}(0, I_d)$$



$\alpha = n/d$

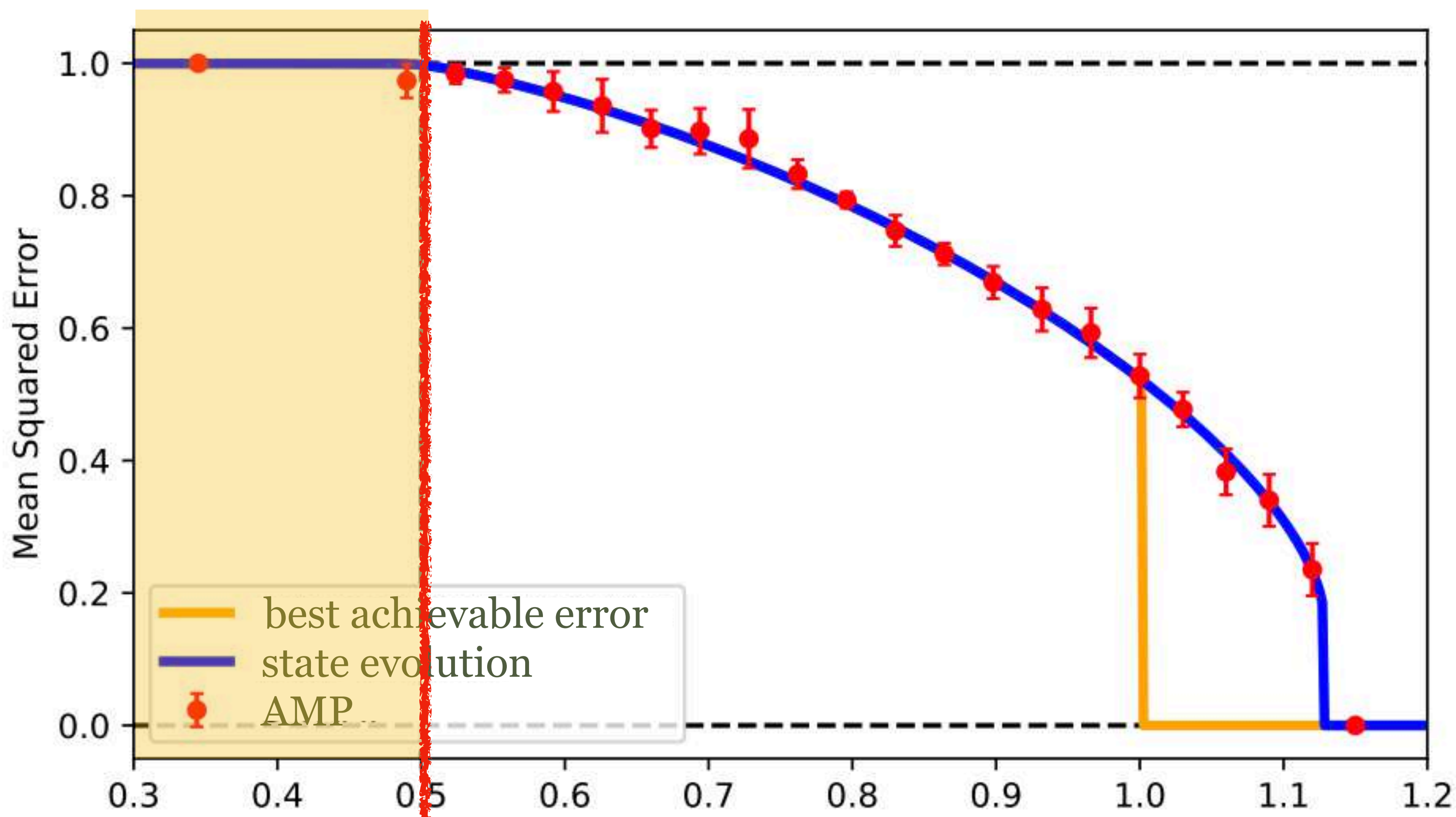
$m_\star = 0$



Case study: the GLM

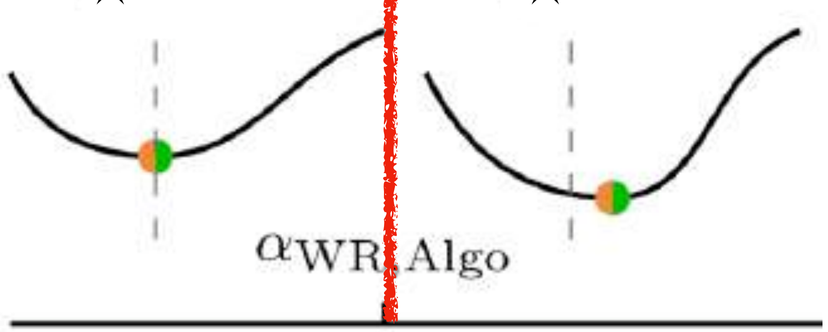
[Barbier et al. '17; Mondelli, Montanari '17; Maillard, **BL**, Krzakala, Zdeborová '20;]

$$y_i = |\langle w_\star, x_i \rangle|^2 \qquad P_\star = \mathcal{N}(0, I_d)$$



$\alpha = n/d$

$m_\star = 0$ $\quad m_\star > 0$

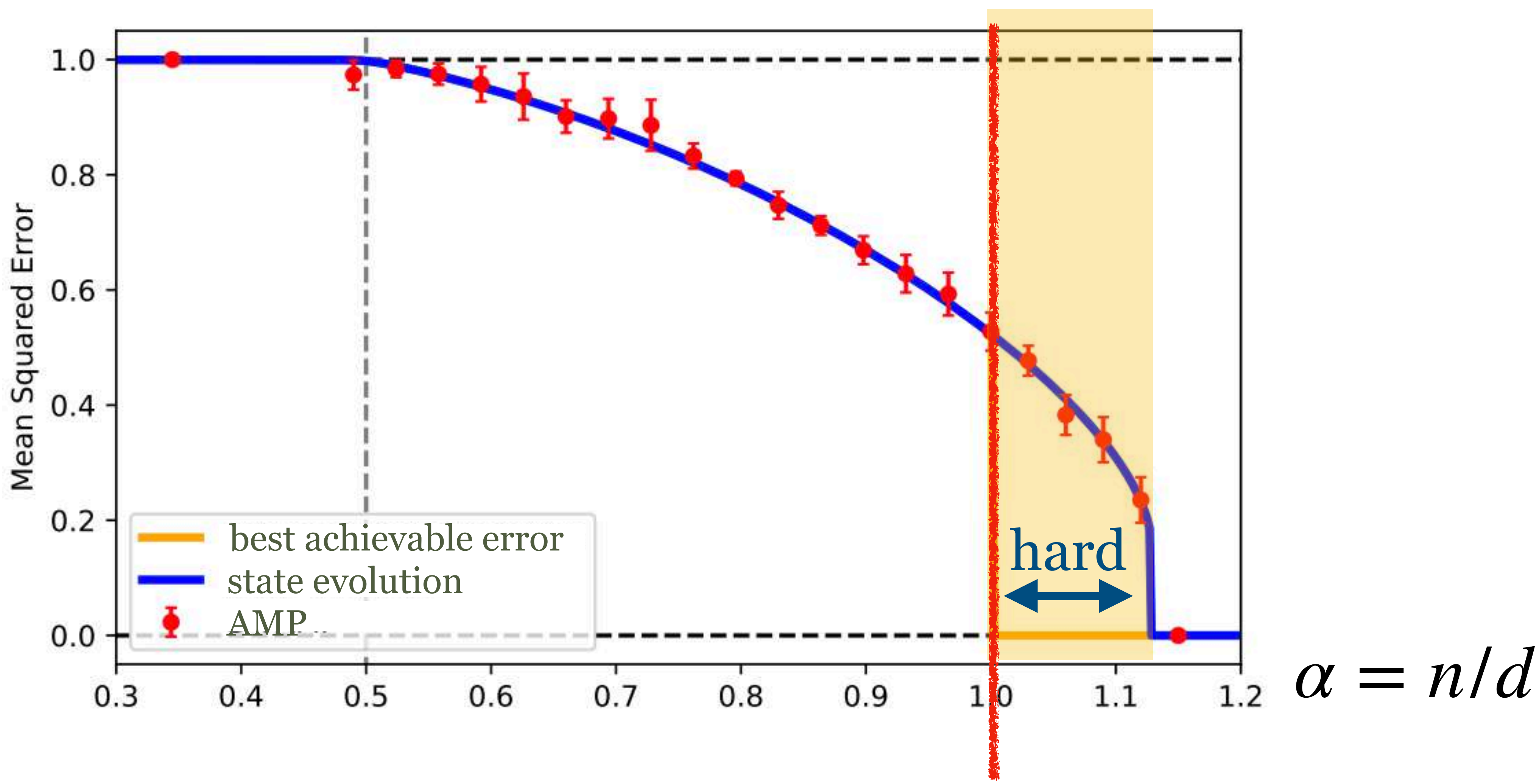


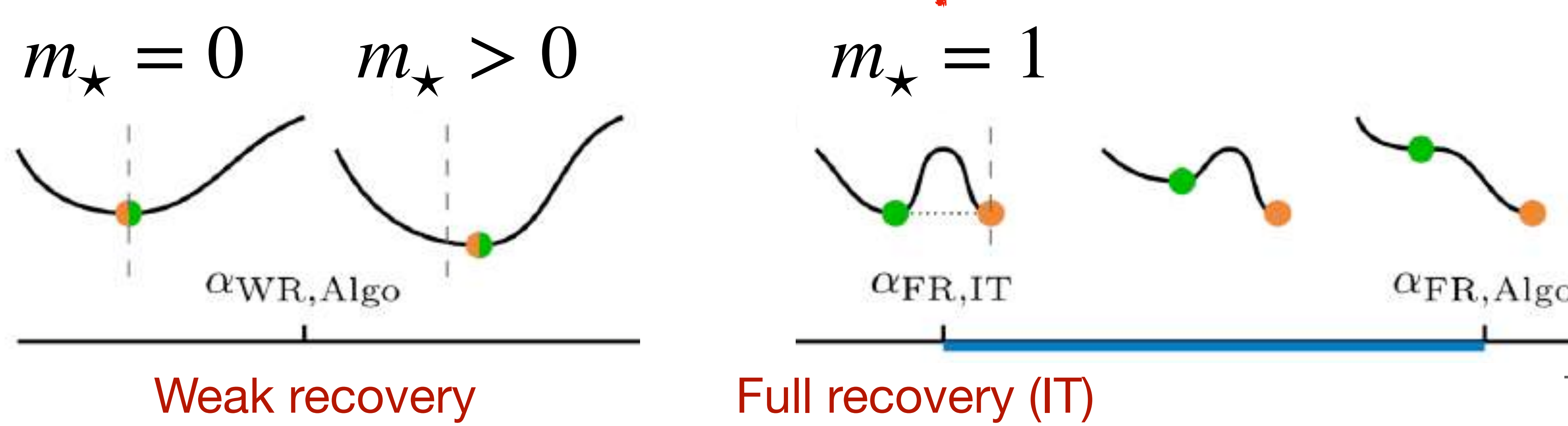
Weak recovery

Case study: the GLM

[Barbier et al. '17; Mondelli, Montanari '17; Maillard, **BL**, Krzakala, Zdeborová '20;]

$$y_i = |\langle w_\star, x_i \rangle|^2 \qquad P_\star = \mathcal{N}(0, I_d)$$

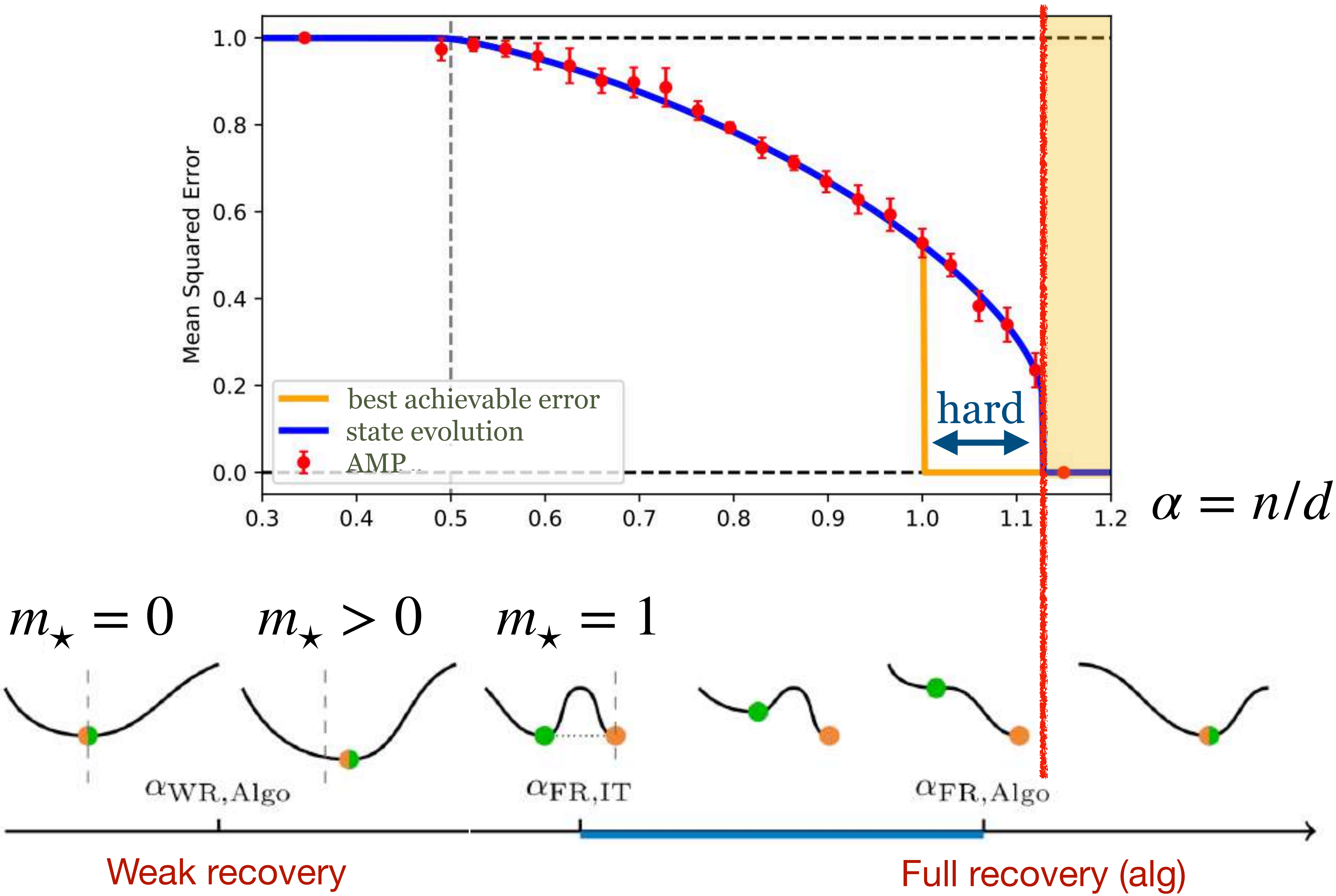




Case study: the GLM

[Barbier et al. ’17; Mondelli, Montanari ’17; Maillard, **BL**, Krzakala, Zdeborová ’20;]

$$y_i = |\langle w_\star, x_i \rangle|^2 \qquad P_\star = \mathcal{N}(0, I_d)$$



G-AMP algorithm

[Mézard 1989; Kabashima 2008; Donoho, Montanari 2009; Rangan 2011; Krzakala, Mézard, Sausset, Sun, Zdeborová 2011]

Key idea: split in two estimation problems

$$y = P(y \mid z)$$

1-d denonising problem

$$z = Xw_\star$$

Linear estimation

G-AMP algorithm

[Mézard 1989; Kabashima 2008; Donoho, Montanari 2009; Rangan 2011; Krzakala, Mézard, Sausset, Sun, Zdeborová 2011]

Key idea: split in two estimation problems

$$y = P(y \,|\, z)$$

1-d denonising problem

$$z = Xw_\star$$

Linear estimation

First estimate:

$$\hat{z} \,|\, y$$

Then estimate:

$$\hat{w} \,|\, \hat{z}$$

G-AMP algorithm

[Mézard 1989; Kabashima 2008; Donoho, Montanari 2009; Rangan 2011; Krzakala, Mézard, Sausset, Sun, Zdeborová 2011]

Key idea: split in two estimation problems

$$y = P(y\,|\,z) \qquad\qquad z = Xw_\star$$

1-d denonising problem — Linear estimation

First estimate: $\hat{z}\,|\,y$

Then estimate: $\hat{w}\,|\,\hat{z}$

$$\left\{\begin{array}{lcl}
V^t & = & \overline{\mathbf{v}^{t-1}} \\
\boldsymbol{\omega}^t & = & \mathbf{\Phi}\widehat{\mathbf{x}}^{t-1}/\sqrt{n} - V^t\mathbf{g}^{t-1} \\
g^t_\mu & = & g_{P_{\rm out}}(Y_\mu, \omega^t_\mu, V^t) \\
\lambda^t & = & \alpha\, \overline{g^2_{P_{\rm out}}(\mathbf{Y}, \boldsymbol{\omega}^t, V^t)} \\
\mathbf{R}^t & = & \widehat{\mathbf{x}}^{t-1} + (\lambda^t)^{-1}\mathbf{\Phi}^\intercal\mathbf{g}^t/\sqrt{n} \\
\widehat{x}^t_i & = & g_{P_0}(R^t_i, \lambda^t) \\
\mathrm{v}^t_i & = & (\lambda^t)^{-1}\,\partial_R g_{P_0}(R, \lambda^t)|_{R=R^t_i}
\end{array}\right.$$

G-AMP algorithm

[Mézard 1989; Kabashima 2008; Donoho, Montanari 2009; Rangan 2011; Krzakala, Mézard, Sausset, Sun, Zdeborová 2011]

Key idea: split in two estimation problems

$$y = P(y \,|\, z)$$

1-d denonising problem

$$z = Xw_\star$$

Linear estimation

First estimate: $\hat{z} \,|\, y$

Then estimate: $\hat{w} \,|\, \hat{z}$

$$\begin{cases} V^t = \overline{\mathbf{v}^{t-1}} \\ \boldsymbol{\omega}^t = \boldsymbol{\Phi}\widehat{\mathbf{x}}^{t-1}/\sqrt{n} - V^t \mathbf{g}^{t-1} \\ g^t_\mu = g_{P_{\text{out}}}(Y_\mu, \omega^t_\mu, V^t) \\ \lambda^t = \alpha\, \overline{g^2_{P_{\text{out}}}(\mathbf{Y}, \boldsymbol{\omega}^t, V^t)} \\ \mathbf{R}^t = \widehat{\mathbf{x}}^{t-1} + (\lambda^t)^{-1}\boldsymbol{\Phi}^\intercal \mathbf{g}^t/\sqrt{n} \\ \widehat{x}^t_i = g_{P_0}(R^t_i, \lambda^t) \\ \mathrm{v}^t_i = (\lambda^t)^{-1}\, \partial_R g_{P_0}(R, \lambda^t)|_{R=R^t_i} \end{cases}$$

Remarks

- In Bayes-optimal setting, use optimal denoiser
- Runs in linear time in nd
- Proven to be optimal over class of first-order methods

[Celentano, Montanari, Wu 2020]

Take away I:

Statistical Physics
=
study of high-d probability

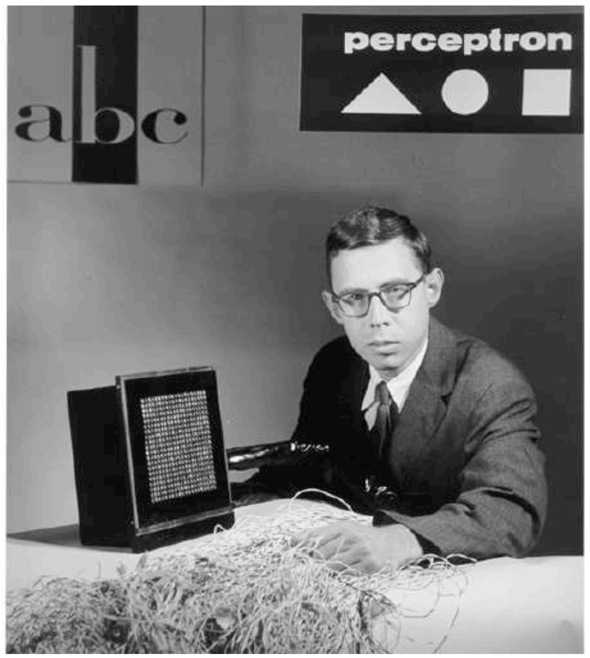
Statistical Physics provides both a
conceptual framework and a toolbox
to approach
high-dimensional optimisation problems

Close relationship between typical-case
computational hardness and landscape

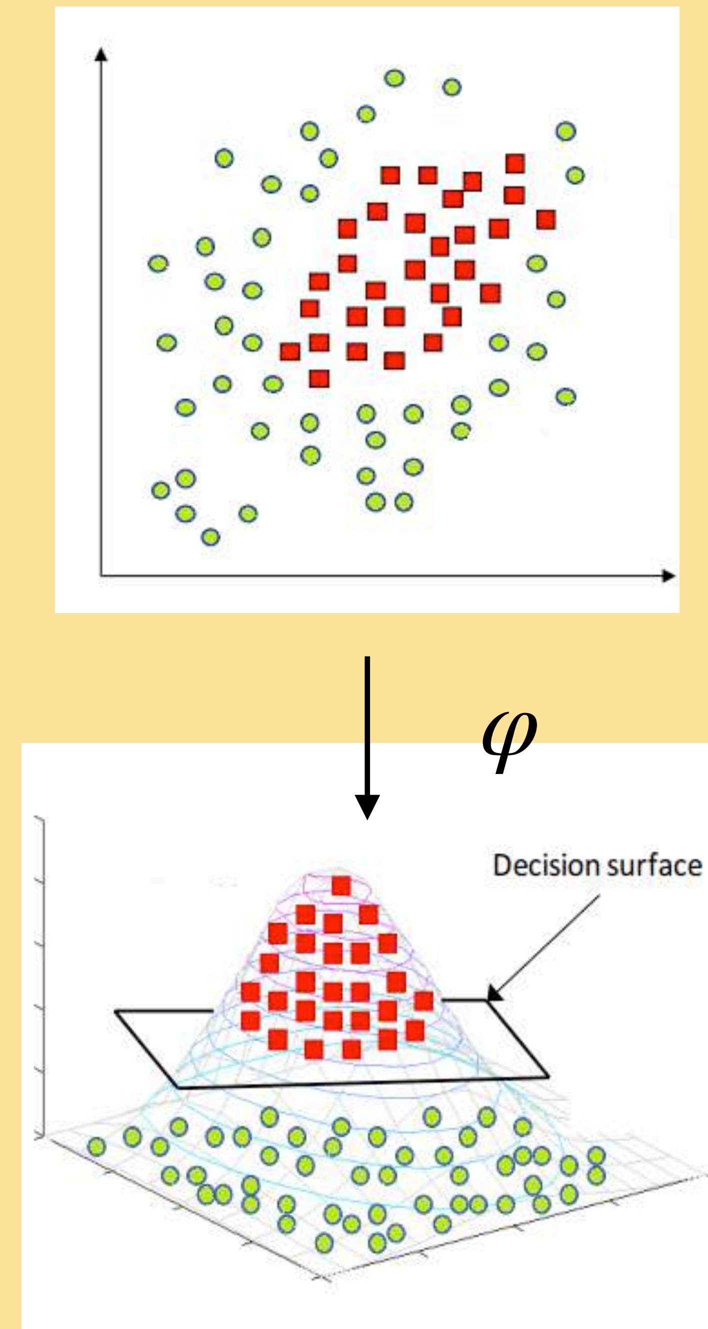
Fruitful history dating back from
(at least) the 80's

Menu for this tutorial

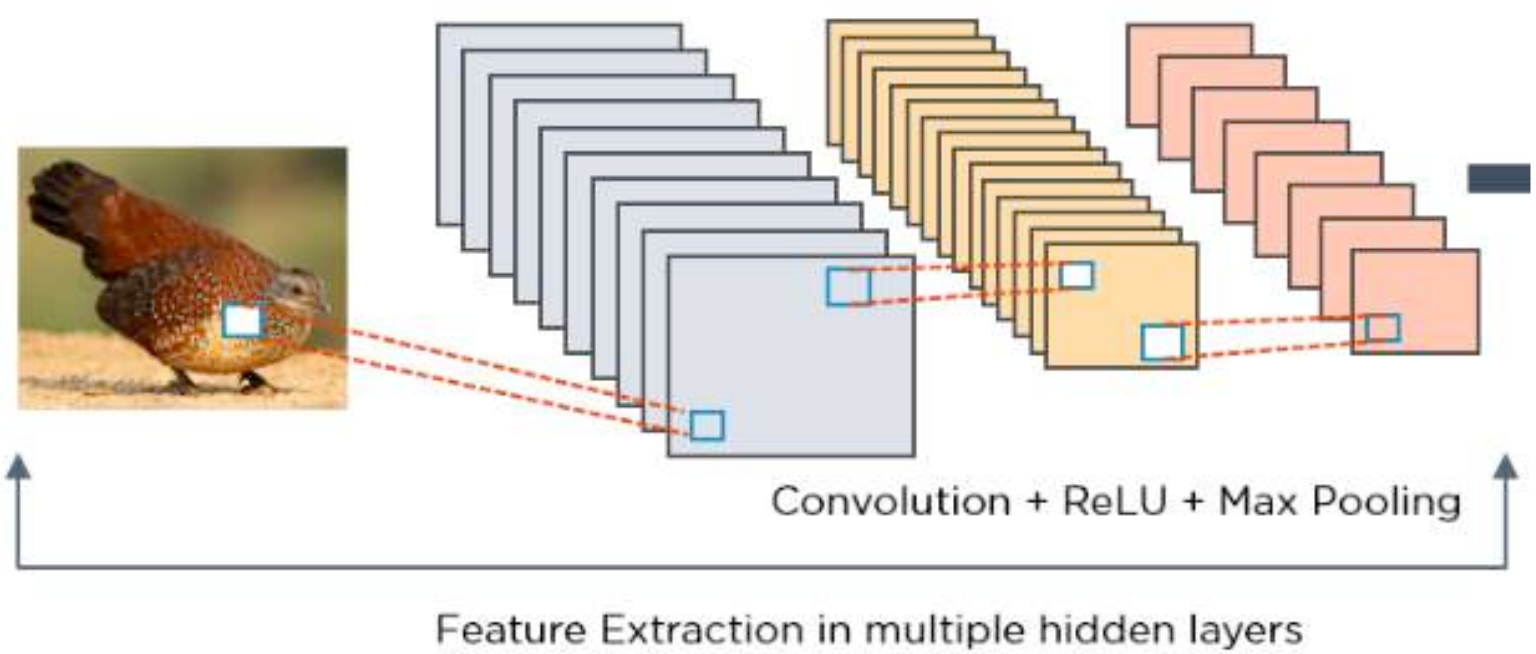
Part I: Statistical Physics of Computation



Part II: Neural Networks at initialisation (a.k.a. kernel methods)



Part III: Feature learning



Leo Breiman
Statistics Department, University of California, Berkeley, CA 94305;
e-mail: leo@stat.berkeley.edu

Reflections After Refereeing Papers for NIPS

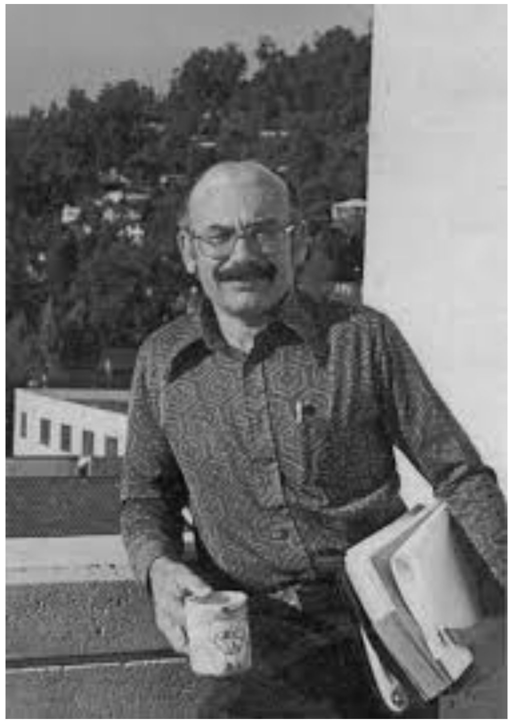
Our fields would be better off with far fewer theorems, less emphasis on faddish stuff, and much more scientific inquiry and engineering. But the latter requires real thinking.

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?
- When should one stop the backpropagation and use the current parameters?

Mathematical theory is not critical to the development of machine learning.

But scientific inquiry is.



3.5 INQUIRY

INQUIRY = sensible and intelligent efforts to understand what is going on. For example:

- mathematical heuristics
- simplified analogies (like the Ising Model)
- simulations
- comparisons of methodologies
- devising new tools
- theorems where useful (rare!)
- shunning panaceas

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$

🤔 $$f(x) = \begin{cases} y_i & \text{if} \quad x \in \mathscr{D} \\ 0 & \text{otherwise} \end{cases}$$

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$

$$f(x) = \begin{cases} y_i & \text{if} \quad x \in \mathscr{D} \\ 0 & \text{otherwise} \end{cases}$$

Memorisation,
not learning!

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$

$$f(x) = \begin{cases} y_i & \text{if} \quad x \in \mathscr{D} \\ 0 & \text{otherwise} \end{cases}$$

Memorisation, not learning!

Introduce a "cost function" $\ell(y, f(x)) \geq 0$

minimise $R(f) = \mathbb{E}_{(x,y)\sim\rho}[\ell(y, f(x))]$

Population Risk

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$

$$f(x) = \begin{cases} y_i & \text{if} \quad x \in \mathscr{D} \\ 0 & \text{otherwise} \end{cases}$$

Memorisation, not learning!

Introduce a "cost function" $\ell(y, f(x)) \geq 0$

minimise $R(f) = \mathbb{E}_{(x,y)\sim\rho}[\ell(y, f(x))]$

Population Risk

Challenges:

- In practice, does't know ρ, only $\mathscr{D}$
- How to minimise over $\{f : \mathbb{R}^d \to \mathbb{R}\}$?

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$

minimise $R(f) = \mathbb{E}_{(x,y)\sim\rho}[\ell(y, f(x))]$ — Population Risk

minimise $\hat{R}_n(f) = \frac{1}{n} \sum_{i \in [n]} [\ell(y_i, f(x_i))]$ — Empirical Risk

Challenges:

- In practice, does't know ρ, only $\mathscr{D}$ ✓
- How to minimise over $\{f : \mathbb{R}^d \to \mathbb{R}\}$?

Supervised Learning

Let $\mathscr{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : i \in [n]\}$ ind. sampled from ρ .

Want: Learn $f_\Theta : \mathbb{R}^d \to \mathbb{R}$ from data $\mathscr{D}$

minimise $R(\Theta) = \mathbb{E}_{(x,y)\sim\rho}[\ell(y, f(x))]$

Population Risk

minimise $\hat{R}_n(\Theta) = \frac{1}{n} \sum_{i\in[n]} [\ell(y_i, f(x_i))]$

Empirical Risk

Challenges:

- In practice, does't know ρ, only $\mathscr{D}$
- How to minimise over $\{f : \mathbb{R}^d \to \mathbb{R}\}$?

Bias-Variance decomposition

For $\ell(y, f_\Theta(x)) = (y - f_\Theta(x))^2$:

$$f_\star(x) = \underset{f}{\operatorname{argmin}} \; R(f) = \mathbb{E}[y \mid x]$$

"Bayes risk"

Bias-Variance decomposition

For $\ell(y, f_\Theta(x)) = (y - f_\Theta(x))^2$:

$$f_\star(x) = \operatorname*{argmin}_f \ R(f) = \mathbb{E}[y \mid x]$$

"Bayes risk"

Hence, for $\hat{\Theta} = \hat{\Theta}(X, y)$ the excess risk is given by:

$$R(\hat{\Theta}) - R(f_\star) = \mathbb{E}[(f_\star(x) - f(x; \Theta))^2]$$

Bias-Variance decomposition

For $\ell(y, f_\Theta(x)) = (y - f_\Theta(x))^2$:

$$f_\star(x) = \underset{f}{\operatorname{argmin}}\ R(f) = \mathbb{E}[y \,|\, x]$$

"Bayes risk"

Hence, for $\hat{\Theta} = \hat{\Theta}(X, y)$ the excess risk is given by:

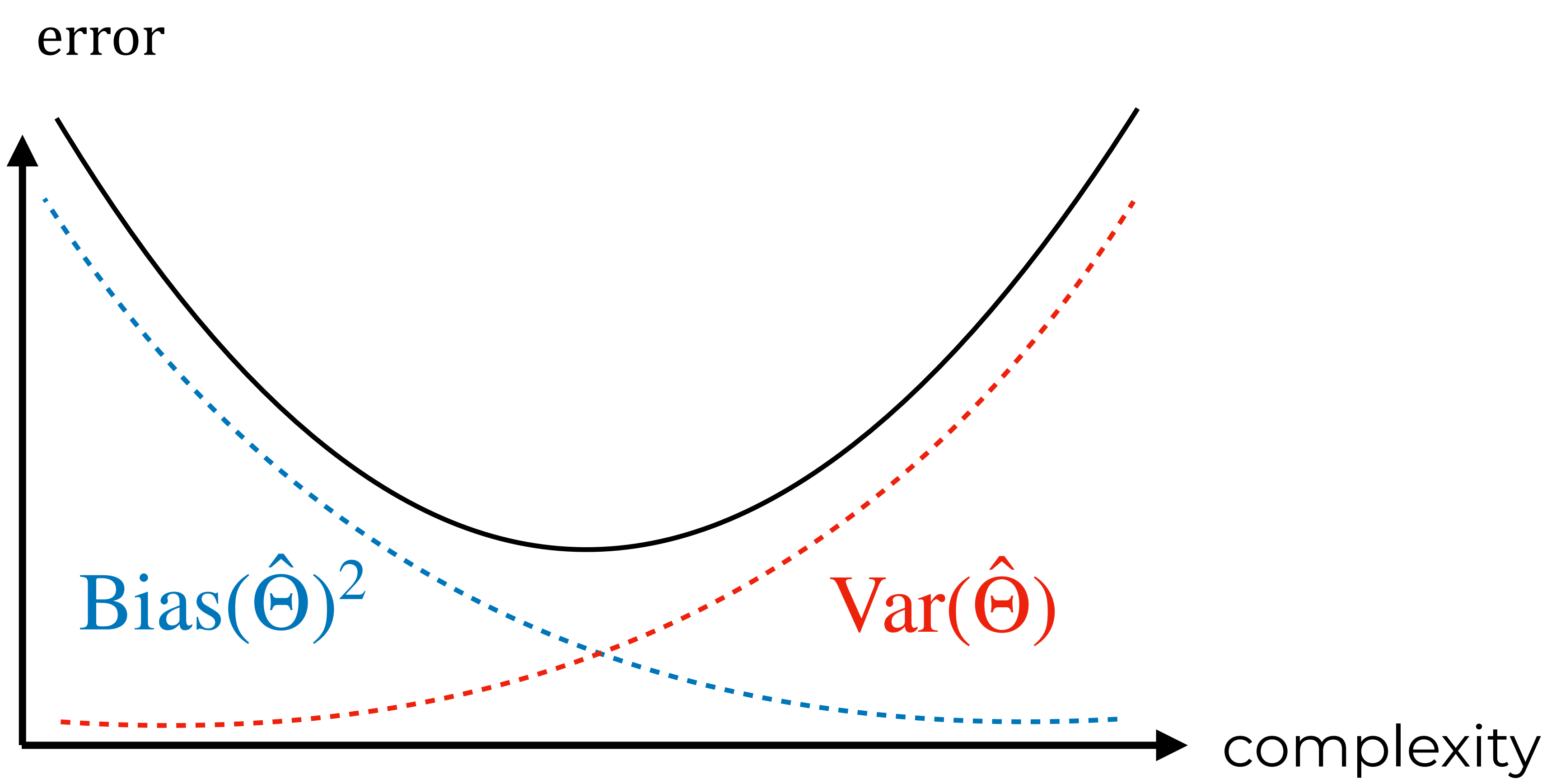
$$R(\hat{\Theta}) - R(f_\star) = \mathbb{E}[(f_\star(x) - f(x; \Theta))^2]$$
$$= \mathbb{E}_X[\text{Bias}(\hat{\Theta})^2] + \mathbb{E}_X[\text{Var}(\hat{\Theta})]$$

Where:

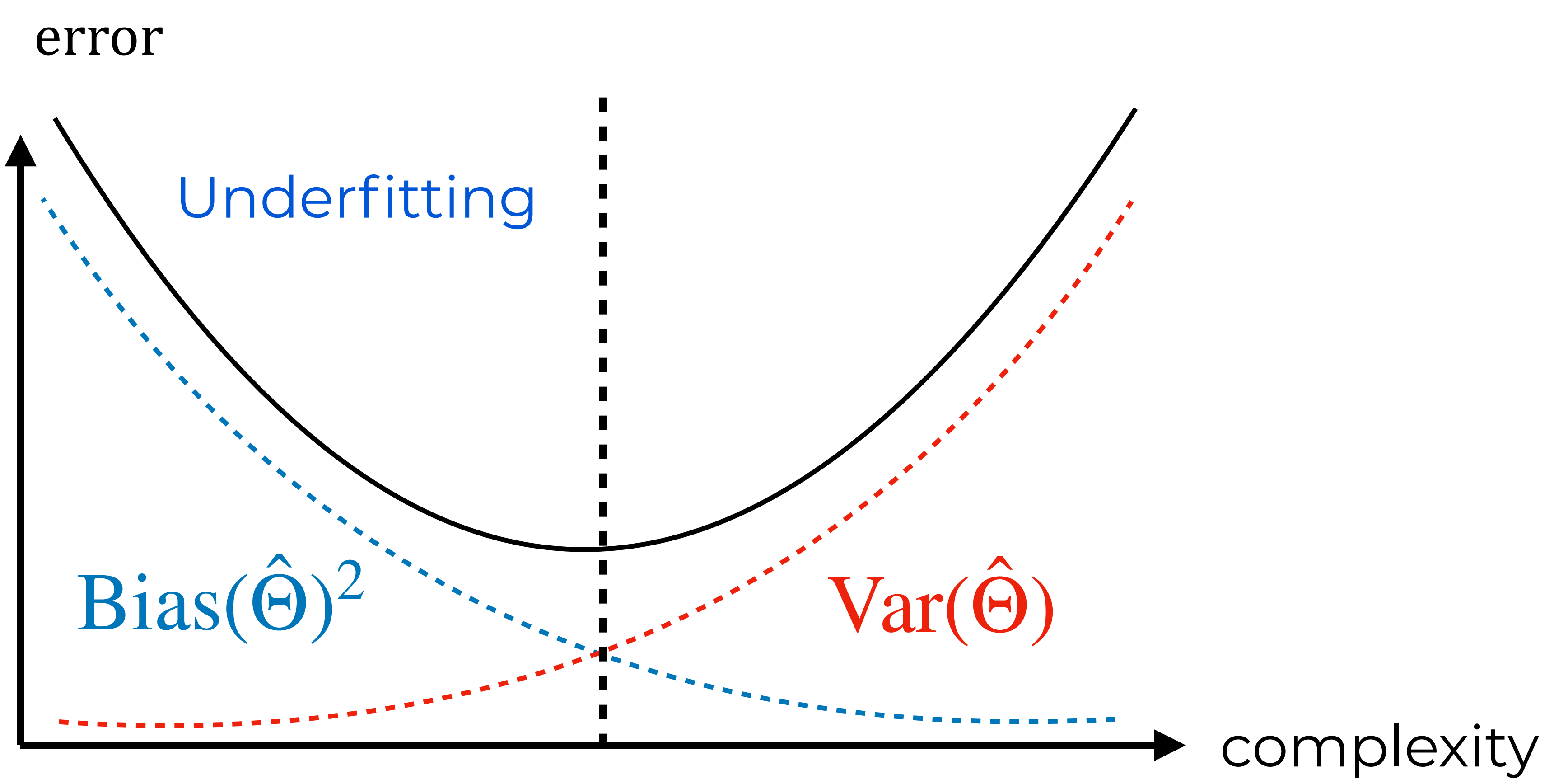
$$\text{Bias}(\hat{\Theta})^2 = \mathbb{E}_x\left[\left(f_\star(x) - \mathbb{E}_y\left[f(x; \hat{\Theta})\right]\right)^2\right]$$

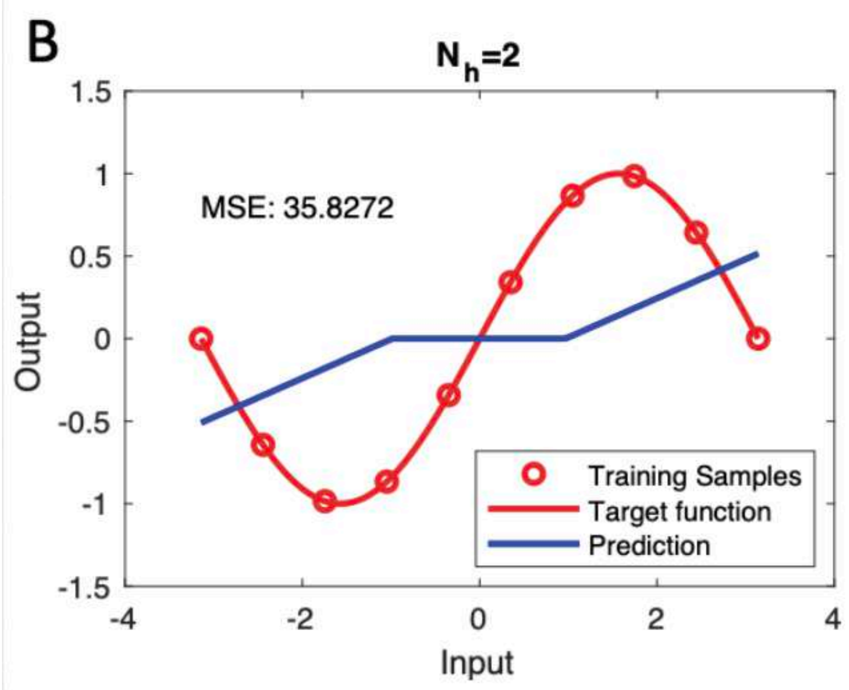
$$\text{Var}(\hat{\Theta}) = \mathbb{E}_{x,y}\left[\left(f(x; \hat{\Theta}) - \mathbb{E}_y\left[f(x; \hat{\Theta})\right]\right)^2\right]$$

Bias-variance trade-off



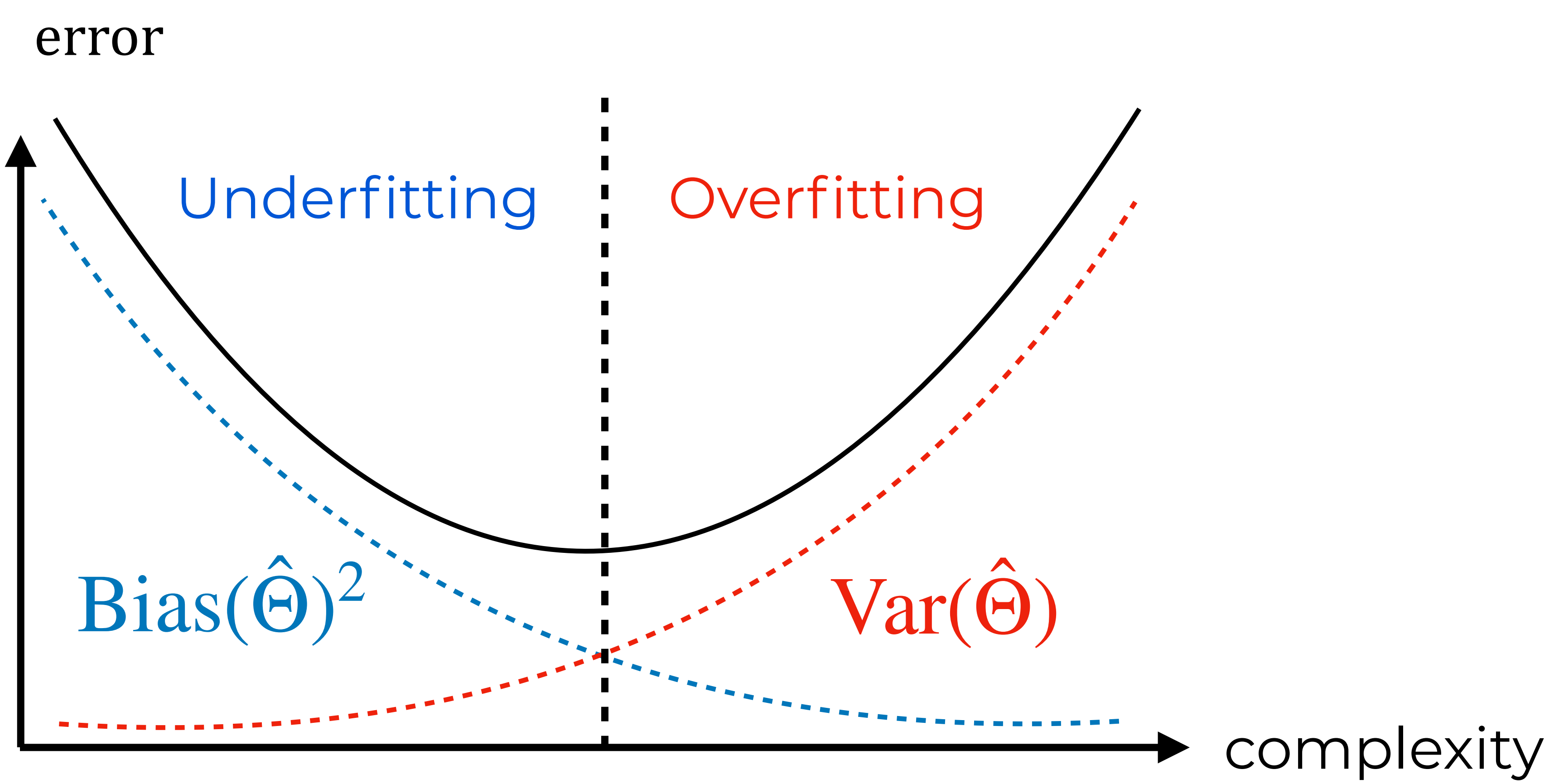
Bias-variance trade-off

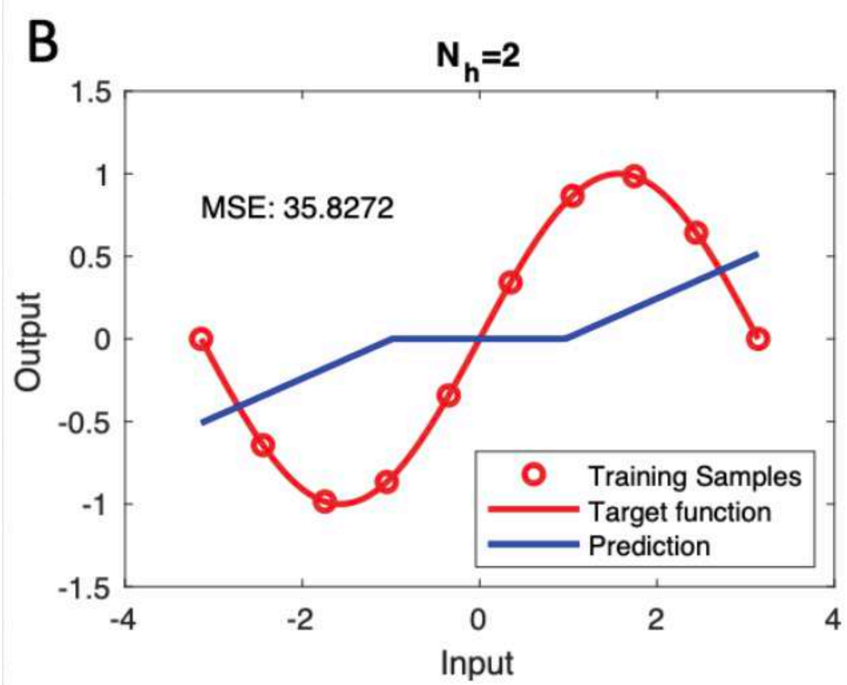


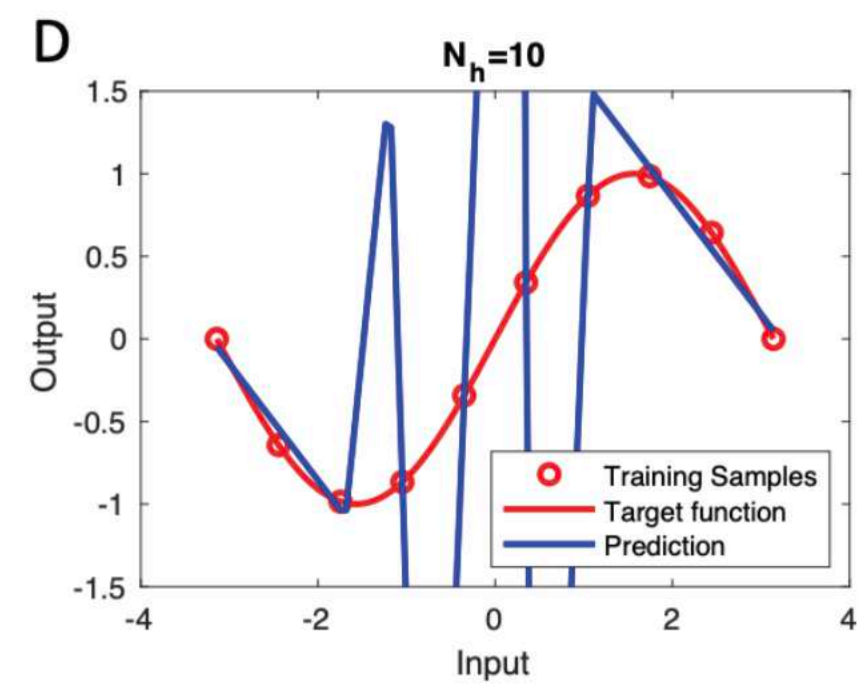


From [Advani, Saxe 17']

Bias-variance trade-off

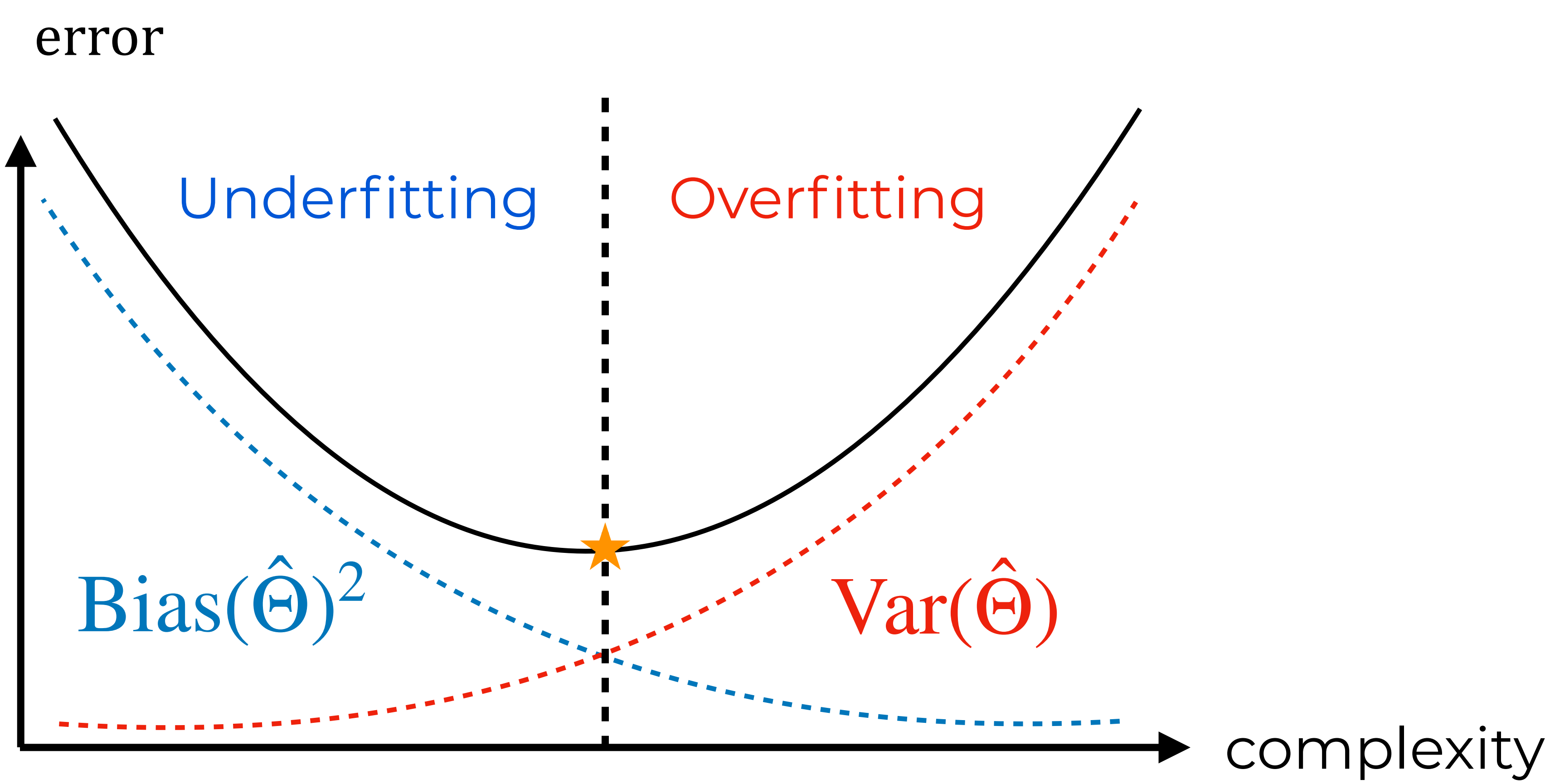


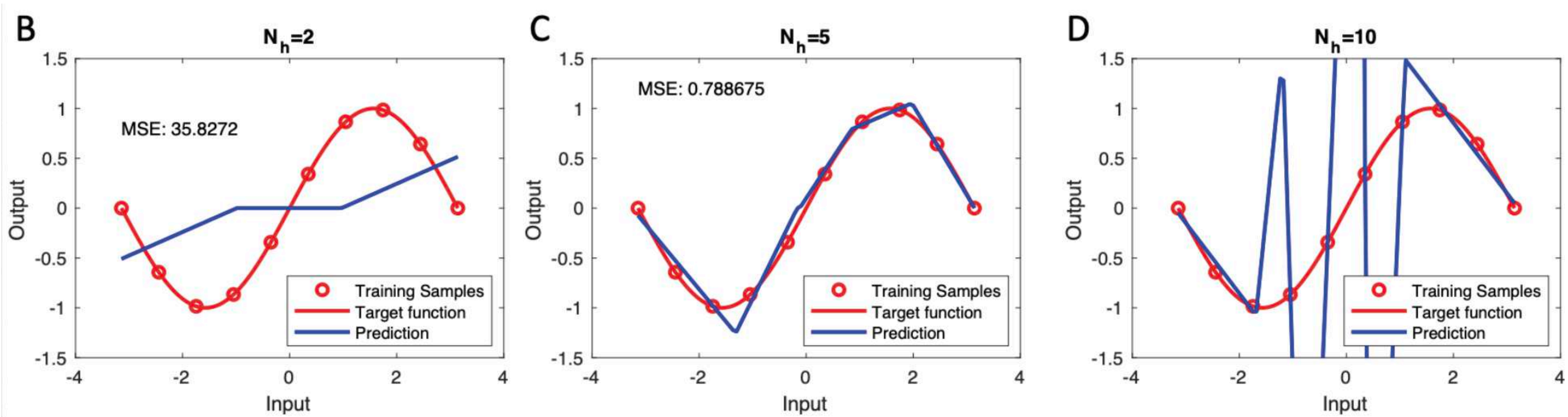




From [Advani, Saxe 17']

Bias-variance trade-off

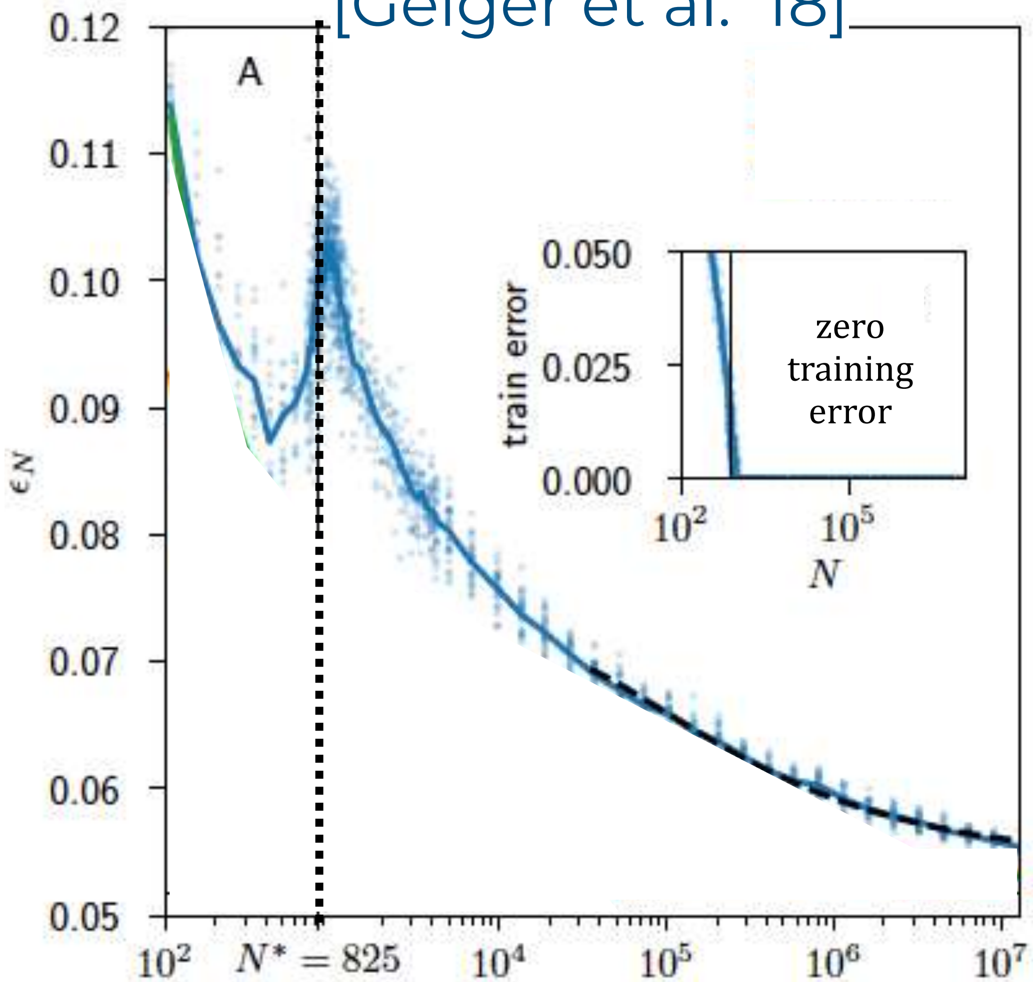




From [Advani, Saxe 17']

“Double descent” [Belkin ’18]

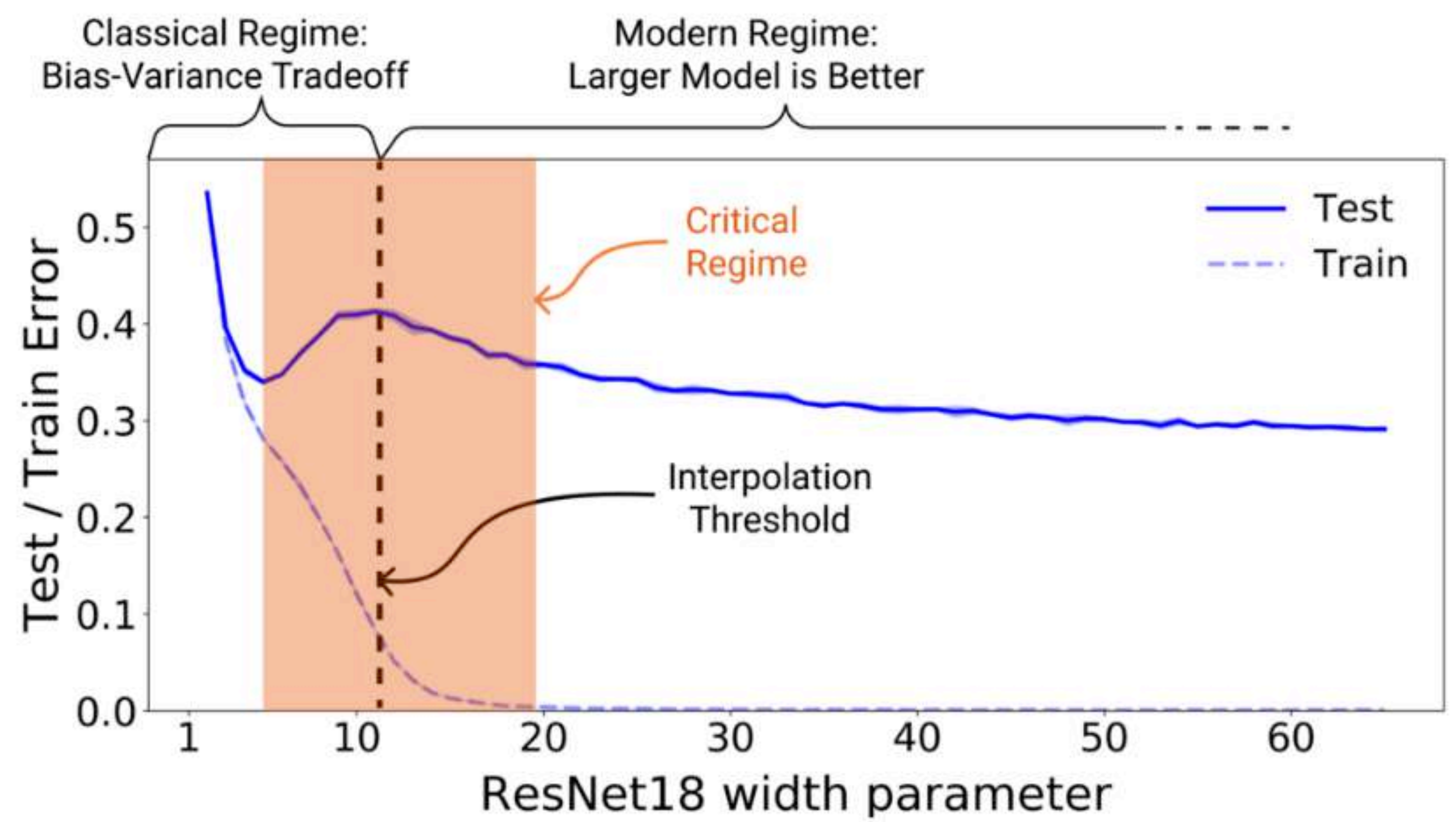
[Geiger et al. ’18]



Number of parameters

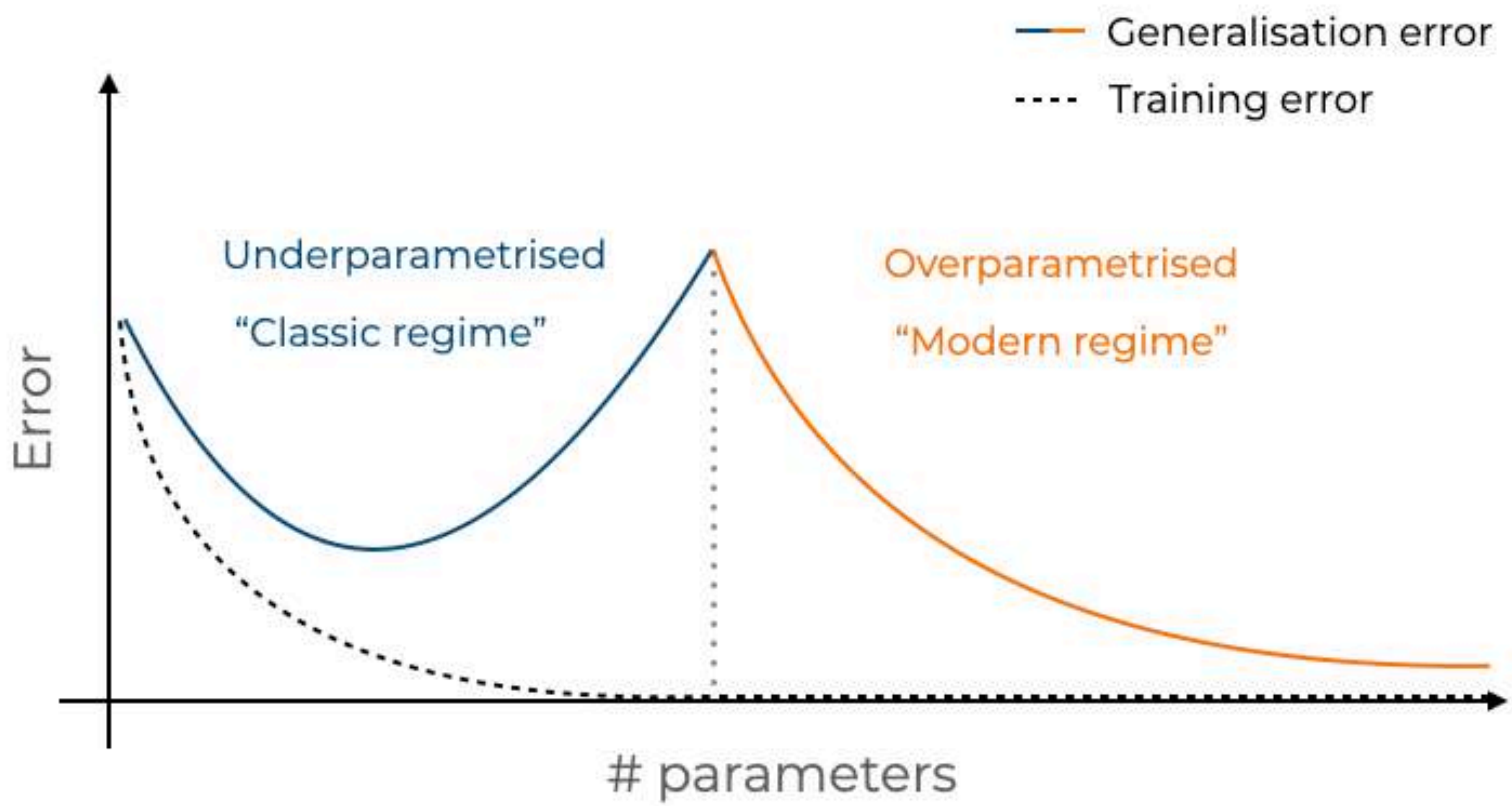
Parity-MNIST, 5 layers, fully-connected, no regularisation

[Nakkiran et al. ’19]



CIFAR10, no regularisation

“Double descent” [Belkin et al. ’18]



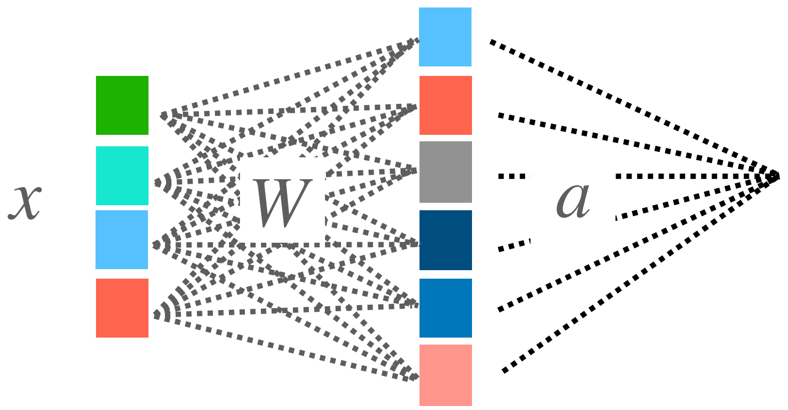
How to make sense of that?

See also [Geman et al. ’92; Opper ’95; Neyshabur, Tomyoka, Srebro, 2015; Advani-Saxe 2017; Belkin, Hsu, Ma, Soumik, Mandal 2019;]

Setting

Consider the hypothesis class of fully-connected two-layer neural networks

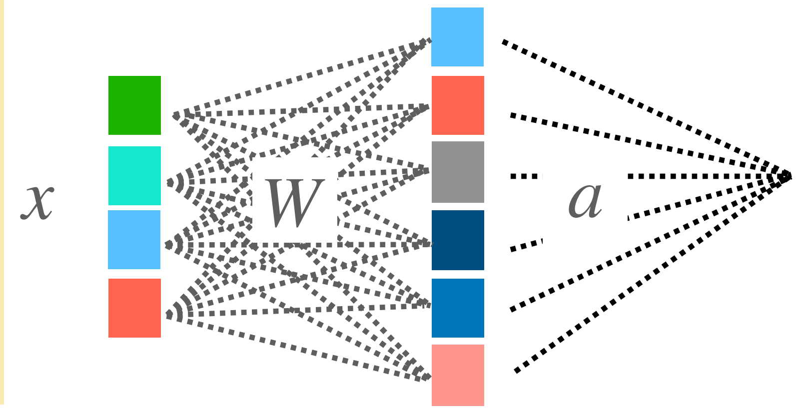
$$f(x; a, W) = \frac{1}{\sqrt{p}} \sum_{k=1}^{p} a_k \sigma\left(\langle w_k, x\rangle\right)$$



Setting

Consider the hypothesis class of fully-connected two-layer neural networks

$$f(x; a, W) = \frac{1}{\sqrt{p}} \sum_{k=1}^{p} a_k \sigma\left(\langle w_k, x \rangle\right)$$



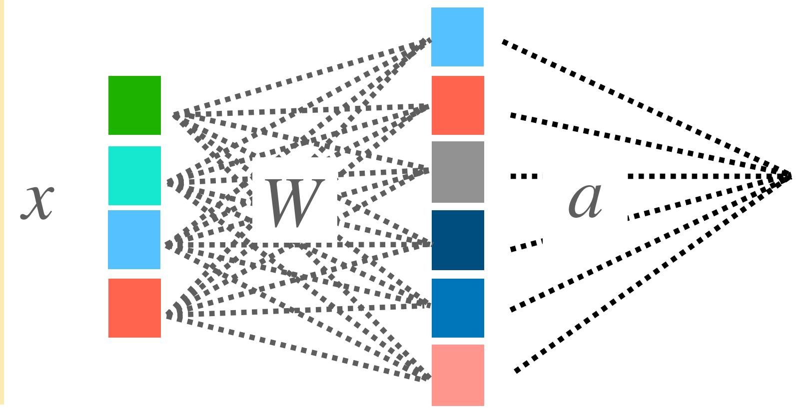
Given a training set $(x_i, y_i)_{i \in [n]} \in \mathbb{R}^{d+1}$ we are interested in the ERM problem :

$$\min_{a,W} \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i; a, W))^2 + \lambda r(a, W)$$

Setting

Consider the hypothesis class of fully-connected two-layer neural networks

$$f(x; a, W) = \frac{1}{\sqrt{p}} \sum_{k=1}^{p} a_k \sigma\left(\langle w_k, x \rangle\right)$$



Given a training set $(x_i, y_i)_{i \in [n]} \in \mathbb{R}^{d+1}$ we are interested in the ERM problem :

$$\min_{a,W} \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i; a, W))^2 + \lambda r(a, W)$$

And in particular, in characterising the risk:

$$R(a, W) = \mathbb{E}[(y - f(x; a, W))^2]$$

$$\hat{R}_n(a, W) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i; a, W))^2]$$

Uniform bounds

Supervised binary classification $(x_i, y_i) \in \mathbb{R}^d \times \{-1,1\}, \quad i \in [n]$

Uniform bounds

Supervised binary classification $(x_i, y_i) \in \mathbb{R}^d \times \{-1,1\}, \quad i \in [n]$

Theorem (Uniform convergence) with probability at least $1-\delta$

$$\forall f_\Theta \in \mathscr{H} \qquad R(\Theta) - \hat{R}_n(\Theta) \leq \mathrm{Rad}(\mathscr{H}) + \sqrt{\frac{\log(1/\delta)}{n}}$$

Where $$\mathrm{Rad}(\mathscr{H}) = \frac{1}{n}\mathbb{E}\left[\sup_{f_\Theta \in \mathscr{H}} \sum_{i\in[n]} y_i f_\Theta(x_i)\right]$$

[Bartlett, Mendelson '03]

Uniform bounds

Supervised binary classification $(x_i, y_i) \in \mathbb{R}^d \times \{-1,1\}, \quad i \in [n]$

Theorem (Uniform convergence) with probability at least $1-\delta$

$$\forall f_\Theta \in \mathcal{H} \qquad R(\Theta) - \hat{R}_n(\Theta) \leq \mathrm{Rad}(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{n}}$$

More generally, $\mathrm{Rad}(\mathcal{H}) \propto \#\mathrm{parameters}$

Model Name	n_{params}
GPT-3 Small	125M
GPT-3 Medium	350M
GPT-3 Large	760M
GPT-3 XL	1.3B
GPT-3 2.7B	2.7B
GPT-3 6.7B	6.7B
GPT-3 13B	13.0B
GPT-3 175B or "GPT-3"	175.0B

[Brown et al 2020]

All models were trained for a total of 300 billion tokens.

Uniform bounds

Supervised binary classification $(x_i, y_i) \in \mathbb{R}^d \times \{-1,1\}, \quad i \in [n]$

Theorem (Uniform convergence) with probability at least $1 - \delta$

$$\forall f_\Theta \in \mathscr{H} \qquad R(\Theta) - \hat{R}_n(\Theta) \leq \mathrm{Rad}(\mathscr{H}) + \sqrt{\frac{\log(1/\delta)}{n}}$$

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

assignments. While we consider multiclass problems, it is straightforward to consider related binary classification problems for which the same experimental observations hold. Since our randomization tests suggest that many neural networks fit the training set with random labels perfectly, we expect that $\hat{\mathfrak{R}}_n(\mathcal{H}) \approx 1$ for the corresponding model class $\mathcal{H}$. This is, of course, a trivial upper bound on the Rademacher complexity that does not lead to useful generalization bounds in realistic settings.

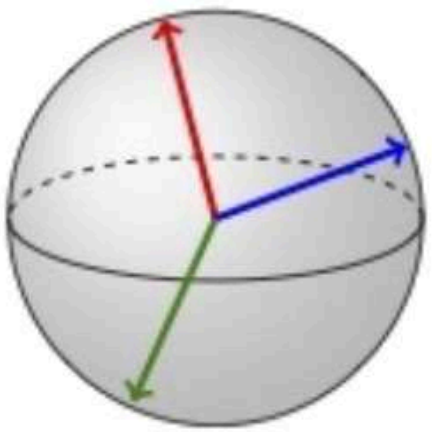
[Zhang, Bengio, Hardt, Recht, Vinyals 17']

Data model

We assume data $(x_i, y_i) \in \mathbb{R}^{d+1}$ is drawn i.i.d. from a multi-index model

$$y_i = g(w_1^\star x_i, \cdots, w_r^\star x_i)$$

$$x_i \sim \mathcal{N}(0, I_d/d) \qquad w_k \in \mathbb{S}^{d-1}(\sqrt{d})$$

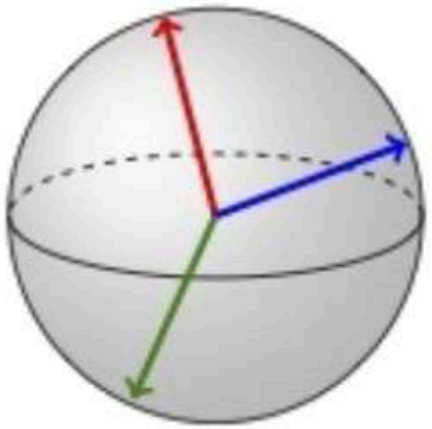


Data model

We assume data $(x_i, y_i) \in \mathbb{R}^{d+1}$ is drawn i.i.d. from a multi-index model

$$y_i = g(w_1^\star x_i, \cdots, w_r^\star x_i)$$

$$x_i \sim \mathcal{N}(0, I_d/d) \qquad w_k \in \mathbb{S}^{d-1}(\sqrt{d})$$



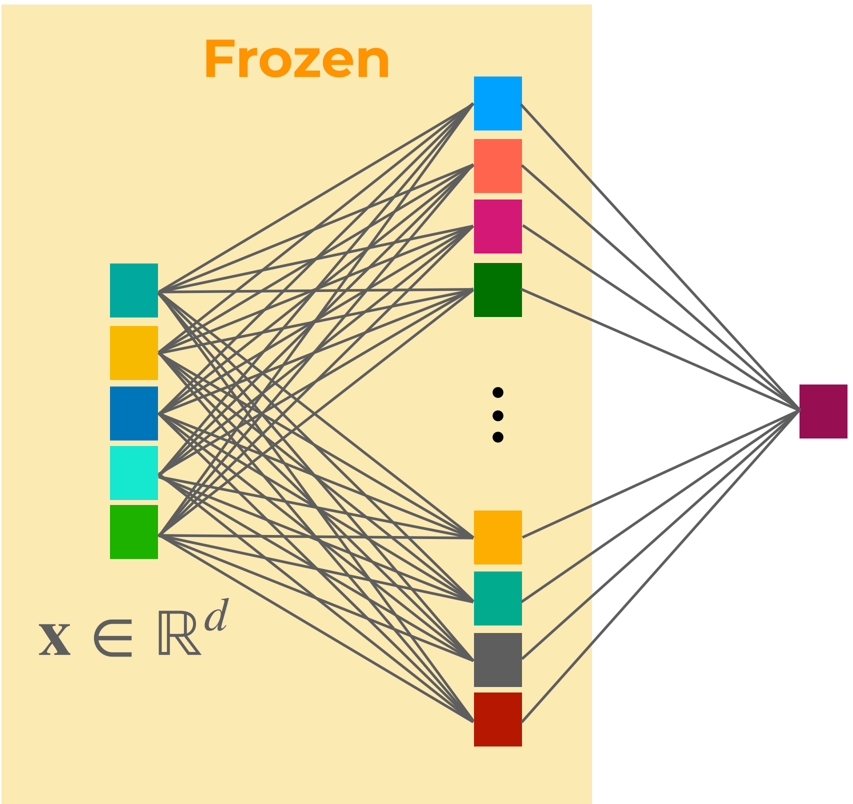
Examples:

$r = 1$	$r > 1$
$g(z) = z$	$g(z) = z_1 z_2 z_3 z_4$
$g(z) = z^2$	$g(z) = \mathrm{sign}(z_1 z_2 z_3)$
$g(z) = \mathrm{sign}(z)$	$g(z) = \sum_{k=1}^{r} a_k \sigma(z_k)$

Initialisation

"Random features model"

[Rahimi & Recht '07]

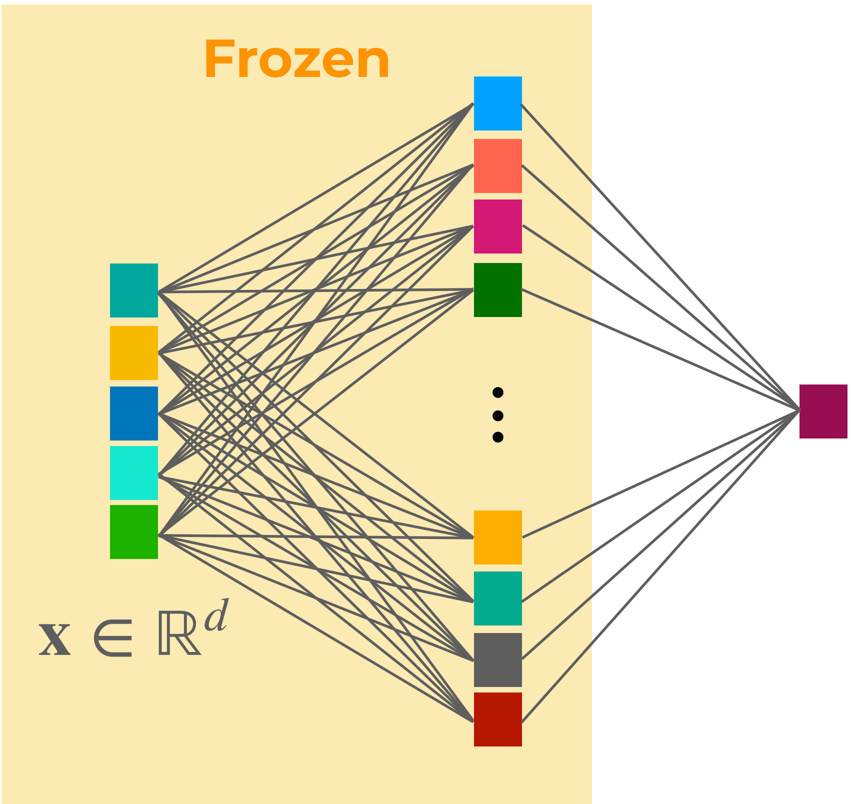


$$\hat{a}_\lambda = \underset{a \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \sum_{i \in [n]} \left(y_i - \frac{1}{\sqrt{p}} \langle a, \sigma(W_0 x_i) \rangle \right)^2 + \frac{\lambda}{2} ||a||_2^2$$

Initialisation

"Random features model"

[Rahimi & Recht '07]



$$\hat{a}_\lambda = \underset{a \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \sum_{i \in [n]} \left(y_i - \frac{1}{\sqrt{p}} \langle a, \sigma(W_0 x_i) \rangle \right)^2 + \frac{\lambda}{2} ||a||_2^2$$

$$= \left(\frac{\alpha}{n} \Phi^\top \Phi + \lambda I_p \right)^{-1} \Phi^\top y$$

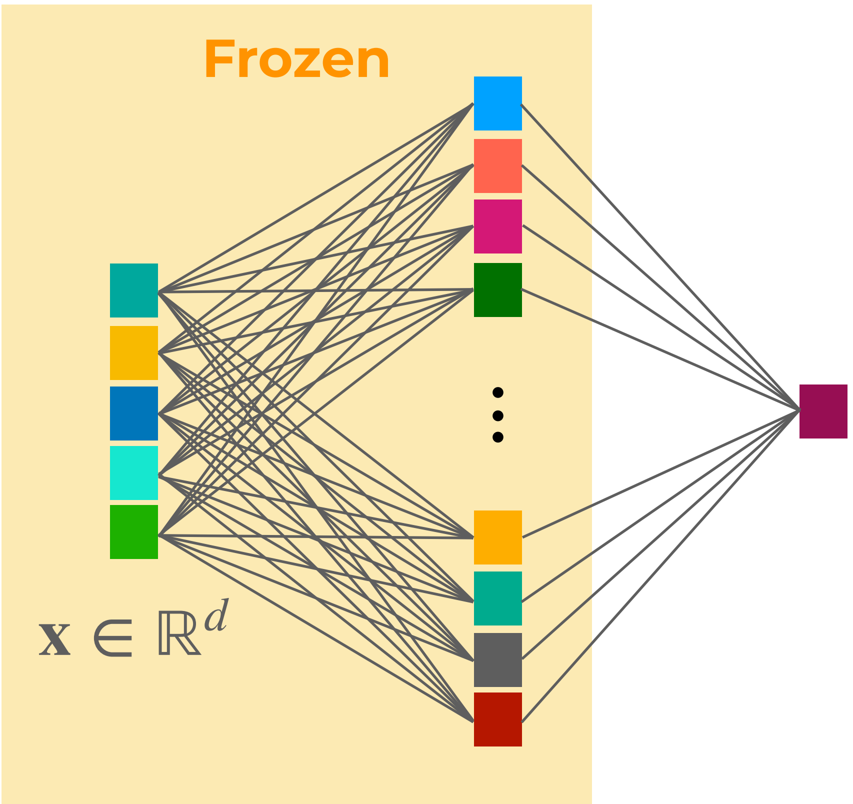
$$\Phi = \sigma(X W_0^\top) \in \mathbb{R}^{n \times p}$$

"Feature matrix"

Initialisation

"Random features model"

[Rahimi & Recht '07]



$$\hat{a}_\lambda = \underset{a \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \sum_{i \in [n]} \left(y_i - \frac{1}{\sqrt{p}} \langle a, \sigma(W_0 x_i) \rangle \right)^2 + \frac{\lambda}{2} ||a||_2^2$$

$$= \left(\frac{\alpha}{n} \Phi^\top \Phi + \lambda I_p \right)^{-1} \Phi^\top y$$

$$\Phi = \sigma(X W_0^\top) \in \mathbb{R}^{n \times p}$$

"Feature matrix"

[Ledoit, Péché 11'; Dobriban, Wager '15]

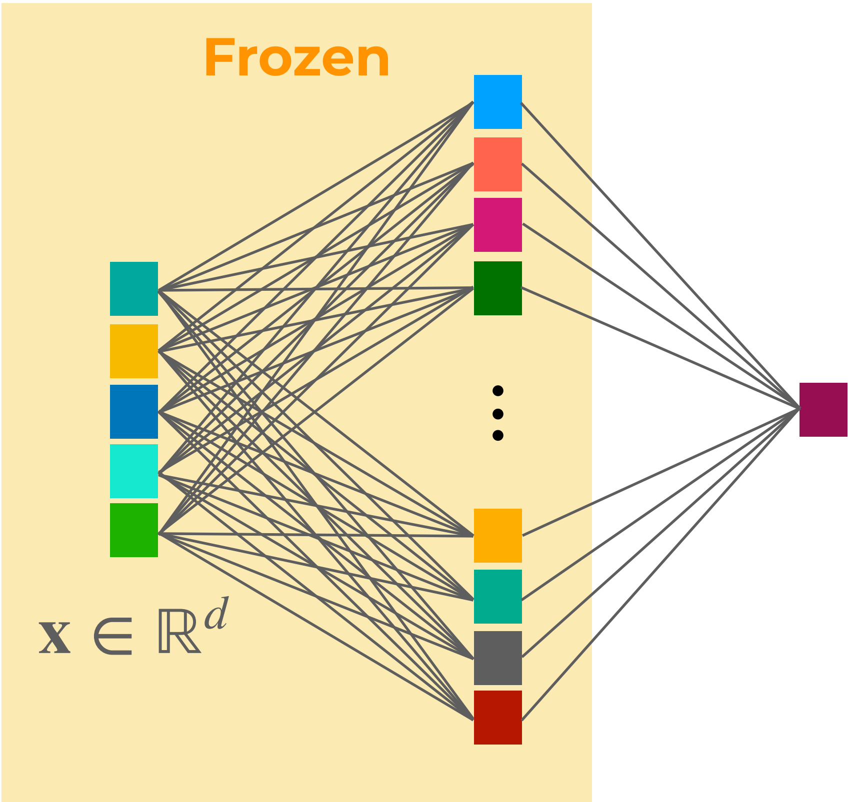
Several known results for the case Φ is a Gaussian matrix.

Challenge: Φ is not a Gaussian matrix!

Initialisation

"Random features model"

[Rahimi & Recht '07]



$$\hat{a}_\lambda = \underset{a \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \sum_{i \in [n]} \left(y_i - \frac{1}{\sqrt{p}} \langle a, \sigma(W_0 x_i) \rangle \right)^2 + \frac{\lambda}{2} ||a||_2^2$$

$$= \left(\frac{\alpha}{n} \Phi^\top \Phi + \lambda I_p \right)^{-1} \Phi^\top y \qquad \Phi = \sigma(X W_0^\top) \in \mathbb{R}^{n \times p}$$

"Feature matrix"

Several known results for the case Φ is a Gaussian matrix.

[Ledoit, Péché 11'; Dobriban, Wager '15]

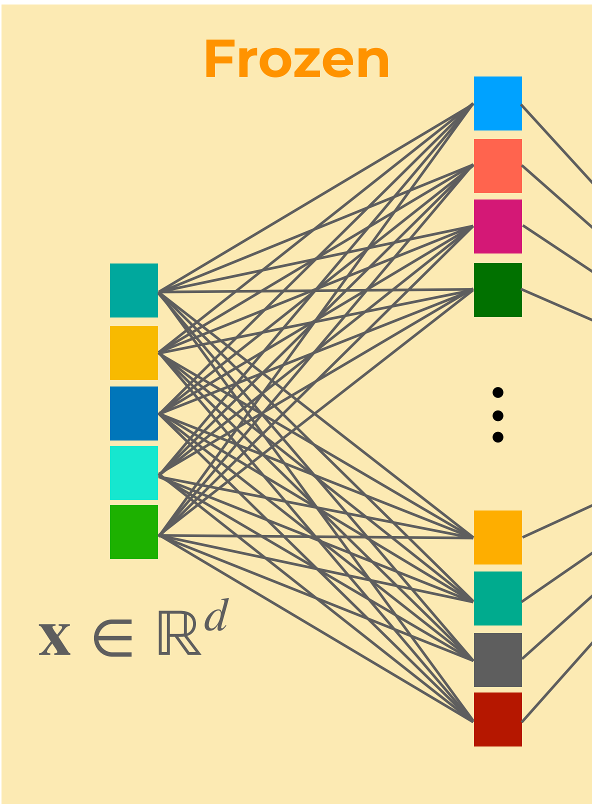
<u>Challenge:</u> Φ is not a Gaussian matrix!

$$\Phi_{ik} = \sigma(\langle w_{0,k}, x_i \rangle) = \sum_{\alpha \geq 0} \mu_\alpha \mathrm{He}_\alpha(\langle w_{0,k}, x_i \rangle)$$

Initialisation

"Random features model"

[Rahimi & Recht '07]



$$\hat{a}_\lambda = \underset{a \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \sum_{i \in [n]} \left(y_i - \frac{1}{\sqrt{p}} \langle a, \sigma(W_0 x_i) \rangle \right)^2 + \frac{\lambda}{2} ||a||_2^2$$

$$= \left(\frac{\alpha}{n} \Phi^\top \Phi + \lambda I_p \right)^{-1} \Phi^\top y \qquad \Phi = \sigma(X W_0^\top) \in \mathbb{R}^{n \times p}$$

"Feature matrix"

Several known results for the case Φ is a Gaussian matrix.

[Ledoit, Péché 11'; Dobriban, Wager '15]

Challenge: Φ is not a Gaussian matrix!

$$\Phi_{ik} = \sigma(\langle w_{0,k}, x_i \rangle) = \sum_{\alpha \geq 0} \mu_\alpha \mathrm{He}_\alpha(\langle w_{0,k}, x_i \rangle)$$

$= \Theta(d^{-1/2})$

Feature moments

Look at the moments of Φ w.r.t. $x \sim \mathcal{N}(0, I_d/d)$

$$\mathbb{E}[\Phi_{ik}] = \mu_0$$

Feature moments

Look at the moments of Φ w.r.t. $x \sim \mathcal{N}(0, I_d/d)$

$$\mathbb{E}[\Phi_{ik}] = \mu_0$$

$$\mathbb{E}[\Phi_{ik}\Phi_{jl}] = \mathbb{E}\left[\sum_{\alpha \geq 0} \mu_\alpha \mathrm{He}_\alpha(\langle w_{0,k}, x_i \rangle) \sum_{\beta \geq 0} \mu_\beta \mathrm{He}_\beta(\langle w_{0,k}, x_i \rangle)\right]$$

Feature moments

Look at the moments of Φ w.r.t. $x \sim \mathcal{N}(0, I_d/d)$

$$\mathbb{E}[\Phi_{ik}] = \mu_0$$

$$\mathbb{E}[\Phi_{ik}\Phi_{jl}] = \mathbb{E}\left[\sum_{\alpha \geq 0} \mu_\alpha \mathrm{He}_\alpha(\langle w_{0,k}, x_i \rangle) \sum_{\beta \geq 0} \mu_\beta \mathrm{He}_\beta(\langle w_{0,k}, x_i \rangle)\right]$$

$$= \sum_{\alpha, \beta \geq 0} \mu_\alpha \mu_\beta \mathbb{E}\left[\mathrm{He}_\alpha(\langle w_{0,k}, x_i \rangle)\mathrm{He}_\beta(\langle w_{0,k}, x_i \rangle)\right]$$

Feature moments

Look at the moments of Φ w.r.t. $x \sim \mathcal{N}(0, I_d/d)$

$$\mathbb{E}[\Phi_{ik}] = \mu_0$$

$$\mathbb{E}[\Phi_{ik}\Phi_{jl}] = \mathbb{E}\left[\sum_{\alpha\geq 0}\mu_\alpha \mathrm{He}_\alpha(\langle w_{0,k}, x_i\rangle)\sum_{\beta\geq 0}\mu_\beta \mathrm{He}_\beta(\langle w_{0,k}, x_i\rangle)\right]$$

$$= \sum_{\alpha,\beta\geq 0}\mu_\alpha\mu_\beta \mathbb{E}\left[\mathrm{He}_\alpha(\langle w_{0,k}, x_i\rangle)\mathrm{He}_\beta(\langle w_{0,k}, x_i\rangle)\right]$$

$$= \sum_{\alpha,\beta\geq 0}\mu_\alpha\mu_\beta\left(\frac{\langle w_{0,k}, w_{0,l}\rangle}{d}\right)^\alpha \delta_{\alpha\beta}$$

Feature moments

Look at the moments of Φ w.r.t. $x \sim \mathcal{N}(0, I_d/d)$

$$\mathbb{E}[\Phi_{ik}] = \mu_0$$

$$\mathbb{E}[\Phi_{ik}\Phi_{jl}] = \mu_0^2 + \mu_1^2 \frac{\langle w_{0,k}, w_{0,l} \rangle}{d} + \sum_{\alpha \geq 2} \mu_\alpha^2 \left(\frac{\langle w_{0,k}, w_{0,l} \rangle}{d} \right)^\alpha$$

$$= \begin{cases} \Theta(1) & k = l \\ \Theta(d^{-\alpha/2}) & d \neq l \end{cases}$$

Feature moments

Look at the moments of Φ w.r.t. $x \sim \mathcal{N}(0, I_d/d)$

$$\mathbb{E}[\Phi_{ik}] = \mu_0$$

$$\mathbb{E}[\Phi_{ik}\Phi_{jl}] \approx \mu_0^2 + \mu_1^2 \frac{\langle w_{0,k}, w_{0,l} \rangle}{d} + \delta_{kl} \sum_{\alpha \geq 2} \mu_\alpha^2$$

Exercise: check q-moment are $\Theta(d^{-q/2})$, hence negligible to order $\Theta(d^{-1})$

Gaussian Universality

Look at the moments of Φ w.r.t. $x \sim \mathcal{N}(0, I_d/d)$

$$\mathbb{E}[\Phi_{ik}] = \mu_0$$

$$\mathbb{E}[\Phi_{ik}\Phi_{jl}] \approx \mu_0^2 + \mu_1^2 \frac{\langle w_{0,k}, w_{0,l} \rangle}{d} + \delta_{kl} \sum_{\alpha \geq 2} \mu_\alpha^2$$

<u>Exercise</u>: check q-moment are $\Theta(d^{-q/2})$, hence negligible to order $\Theta(d^{-1})$

Gaussian equivalence theorem

Consider two models: (a) $\hat{a}_\lambda(\Phi, y)$ $\quad \Phi = \sigma(XW_0^\top)$

(b) $\hat{a}_\lambda(G, y)$ $\quad G = \mu_0 1_n 1_p^\top + \mu_1 W_0 X^\top + \mu_\star Z$

Then: $|R(\hat{a}_\lambda(\Phi, y)) - R(\hat{a}_\lambda(G, y))| \to 0 \quad d \to \infty \quad n, p = \Theta(d)$

Proofs in [Mei & Montanari '19; Goldt, **BL** et al. '20; Hu, Lu '20].

Gaussian Universality

Gaussian equivalence theorem

Consider two models: (a) $\hat{a}_\lambda(\Phi, y)$ $\Phi = \sigma(XW_0^\top)$

(b) $\hat{a}_\lambda(G, y)$ $G = \mu_0 1_n 1_p^\top + \mu_1 W_0 X^\top + \mu_\star Z$

Then: $|R(\hat{a}_\lambda(\Phi, y)) - R(\hat{a}_\lambda(G, y))| \to 0$ $d \to \infty$ $n, p = \Theta(d)$

Proofs in [Mei & Montanari ’19; Goldt, **BL** et al. ’20; Hu, Lu ’20].

Several extensions:

- Deep random features [Schroder, Cui, Dmitriev, **BL** ’23,’24; Bosch, Panahi, Hassibi ’23].
- Polynomial scaling [Lu, Yau ’23; Hu, Lu, Misiakiewicz ’24; Defilippis, **BL**, Misiakiewicz ’24].
- Multi-modal features [Refinetti, Goldt, Krzakala, Zdeborová ’21; Dandi, Stephan, Krzakala, **BL**, Zdeborová ’23].
- Application to real data [**BL,** Gerbelot, Cui, Goldt, Krzakala, Mezard, Zdeborová ’21; Wei, Hu, Steinhardt ‘22].
- Beyond Gaussian [El Karoui ’18; Adomaityte, Defilippis, **BL**, Sicuro ’23; Pesce, Krzakala, **BL**, Stephan ’22; Tsironis, Moustakas ‘24].

Definitions:

[Gerace, **BL**, Krzakala, Zdeborová '20] ($r = 1$)

Consider the unique fixed point of the following system of equations

$$\begin{cases} \hat{V}_s = \frac{\alpha}{\gamma}\kappa_1^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\partial_\omega \eta(y,\omega_1)}{V}\right], \\ \hat{q}_s = \frac{\alpha}{\gamma}\kappa_1^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\left(\eta(y,\omega_1)-\omega_1\right)^2}{V^2}\right], \\ \hat{m}_s = \frac{\alpha}{\gamma}\kappa_1 \mathbb{E}_{\xi,y}\left[\partial_\omega\mathcal{Z}\left(y,\omega_0\right)\frac{\left(\eta(y,\omega_1)-\omega_1\right)}{V}\right], \\ \hat{V}_w = \alpha\kappa_\star^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\partial_\omega \eta(y,\omega_1)}{V}\right], \\ \hat{q}_w = \alpha\kappa_\star^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\left(\eta(y,\omega_1)-\omega_1\right)^2}{V^2}\right], \end{cases}
\quad
\begin{cases} V_s = \frac{1}{\hat{V}_s}\left(1 - z\, g_\mu(-z)\right), \\ q_s = \frac{\hat{m}_s^2+\hat{q}_s}{\hat{V}_s}\left[1-2zg_\mu(-z)+z^2g'_\mu(-z)\right] \\ \quad -\frac{\hat{q}_w}{(\lambda+\hat{V}_w)\hat{V}_s}\left[-zg_\mu(-z)+z^2g'_\mu(-z)\right], \\ m_s = \frac{\hat{m}_s}{\hat{V}_s}\left(1-z\,g_\mu(-z)\right), \\ V_w = \frac{\gamma}{\lambda+\hat{V}_w}\left[\frac{1}{\gamma}-1+zg_\mu(-z)\right], \\ q_w = \gamma\frac{\hat{q}_w}{(\lambda+\hat{V}_w)^2}\left[\frac{1}{\gamma}-1+z^2g'_\mu(-z)\right], \\ \quad +\frac{\hat{m}_s^2+\hat{q}_s}{(\lambda+\hat{V}_w)\hat{V}_s}\left[-zg_\mu(-z)+z^2g'_\mu(-z)\right], \end{cases}
\quad
\begin{cases} \eta(y,\omega) = \underset{x\in\mathbb{R}}{\text{argmin}}\left[\frac{(x-\omega)^2}{2V}+\ell(y,x)\right] \\ \mathcal{Z}(y,\omega) = \int\frac{dx}{\sqrt{2\pi V^0}}e^{-\frac{1}{2V^0}(x-\omega)^2}\delta\left(y-g(x)\right) \end{cases}$$

where $V = \kappa_1^2 V_s + \kappa_\star^2 V_w$, $V^0 = \rho - \dfrac{M^2}{Q}$, $Q = \kappa_1^2 q_s + \kappa_\star^2 q_w$, $M = \kappa_1 m_s$, $\omega_0 = M/\sqrt{Q}\xi$, $\omega_1 = \sqrt{Q}\xi$ and $g_\mu$is the Stieltjes transform of FF^T

$$\kappa_0 = \mathbb{E}\left[\sigma(z)\right], \kappa_1 \equiv \mathbb{E}\left[z\sigma(z)\right], \kappa_\star \equiv \mathbb{E}\left[\sigma(z)^2\right] - \kappa_0^2 - \kappa_1^2, \text{ and } \mathbf{z}^\mu \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_p)$$

In the high-dimensional limit:

$$\epsilon_{gen} = \mathbb{E}_{\lambda,\nu}\left[(f^0(\nu)-\hat{f}(\lambda))^2\right]$$

$$\text{with} \quad (\nu,\lambda) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\rho & M^\star \\ M^\star & Q^\star\end{pmatrix}\right)$$

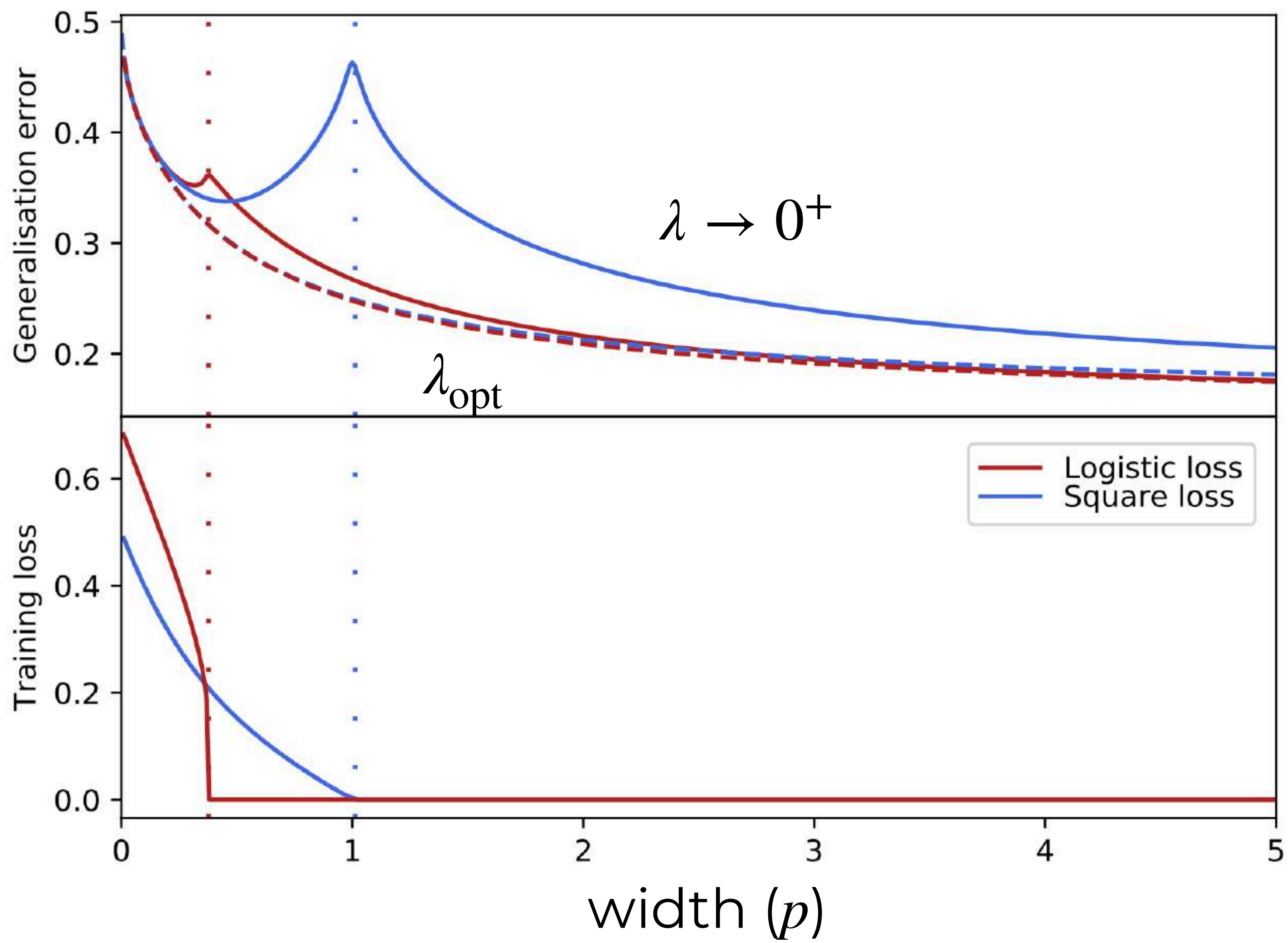
$$\mathcal{L}_{\text{training}} = \frac{\lambda}{2\alpha}q_w^\star + \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0^\star\right)\ell\left(y,\eta(y,\omega_1^\star)\right)\right]$$

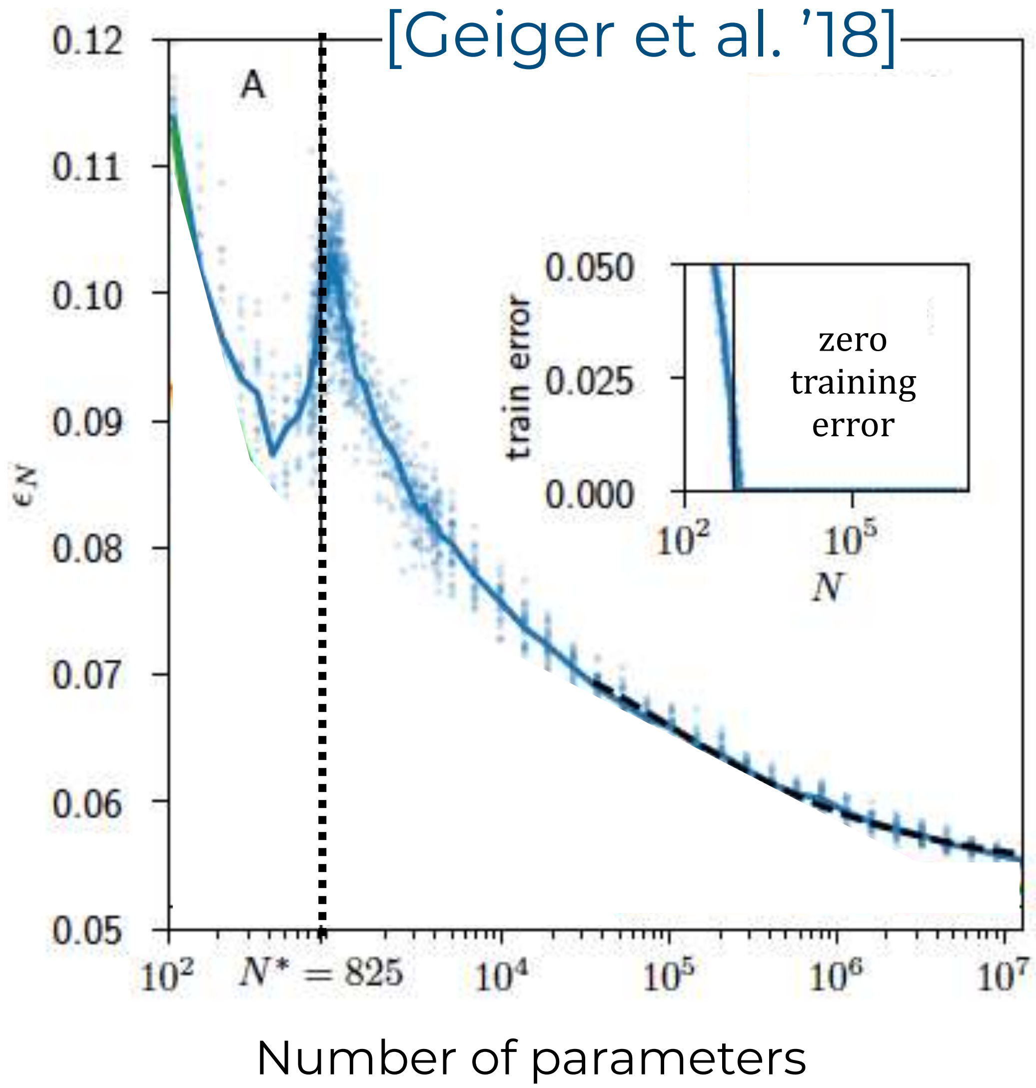
$$\text{with} \quad \omega_0^\star = M^\star/\sqrt{Q^\star}\xi, \omega_1^\star = \sqrt{Q^\star}\xi$$

See also [Mei & Montanari '19; Gerace, **BL** et al. '20; Hu, Lu, '20; Dhifallah, Lu '20; Hu, Lu, Misiakiewicz '24]

Double descent

$$g(z) = \text{sign}(z) \qquad \sigma(t) = \text{erf}(t)$$





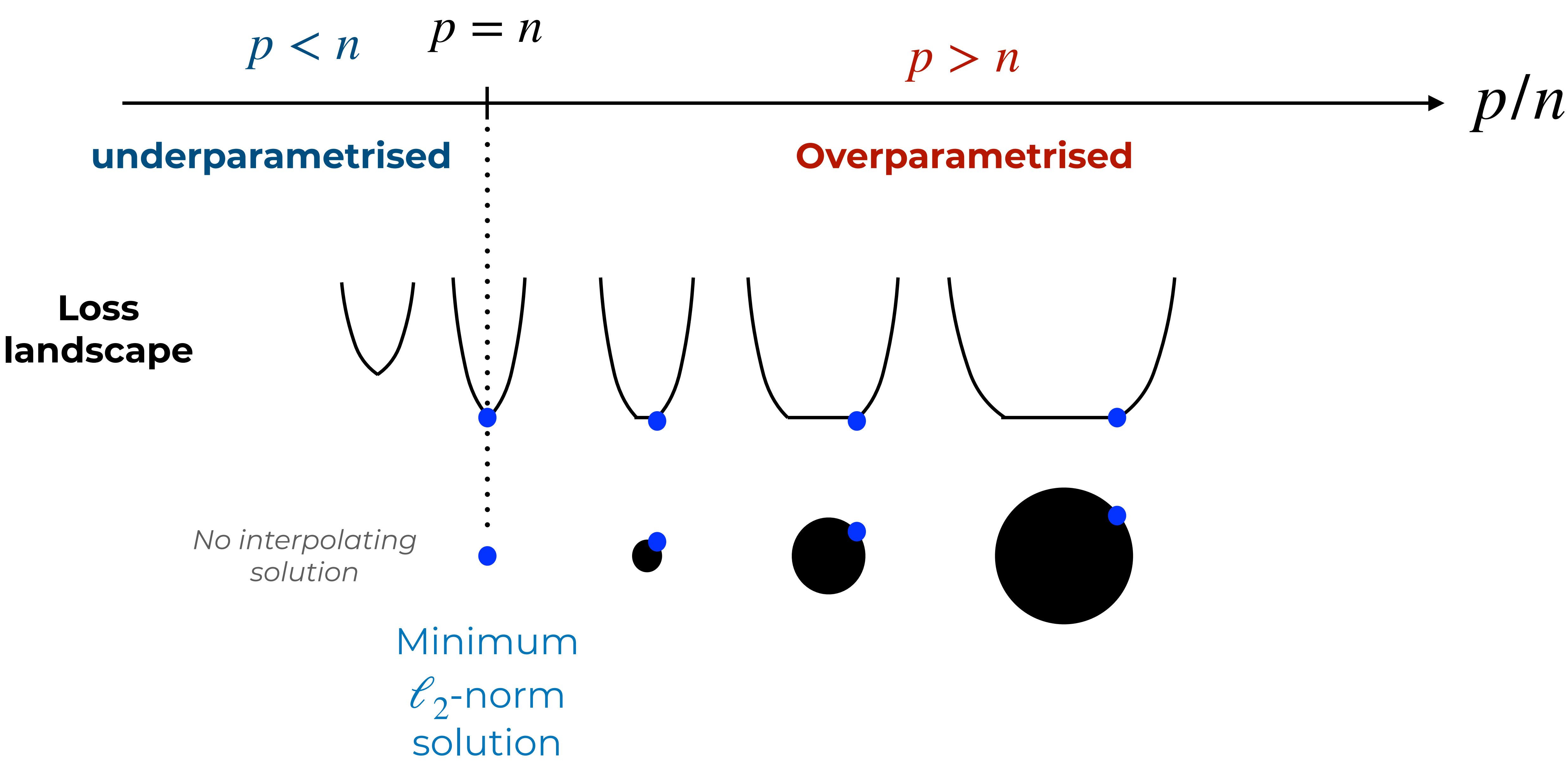
[Mei, Montanari '19; Gerace, **BL**, Krzakala, Mézard, Zdeborová '20]

What's going on?

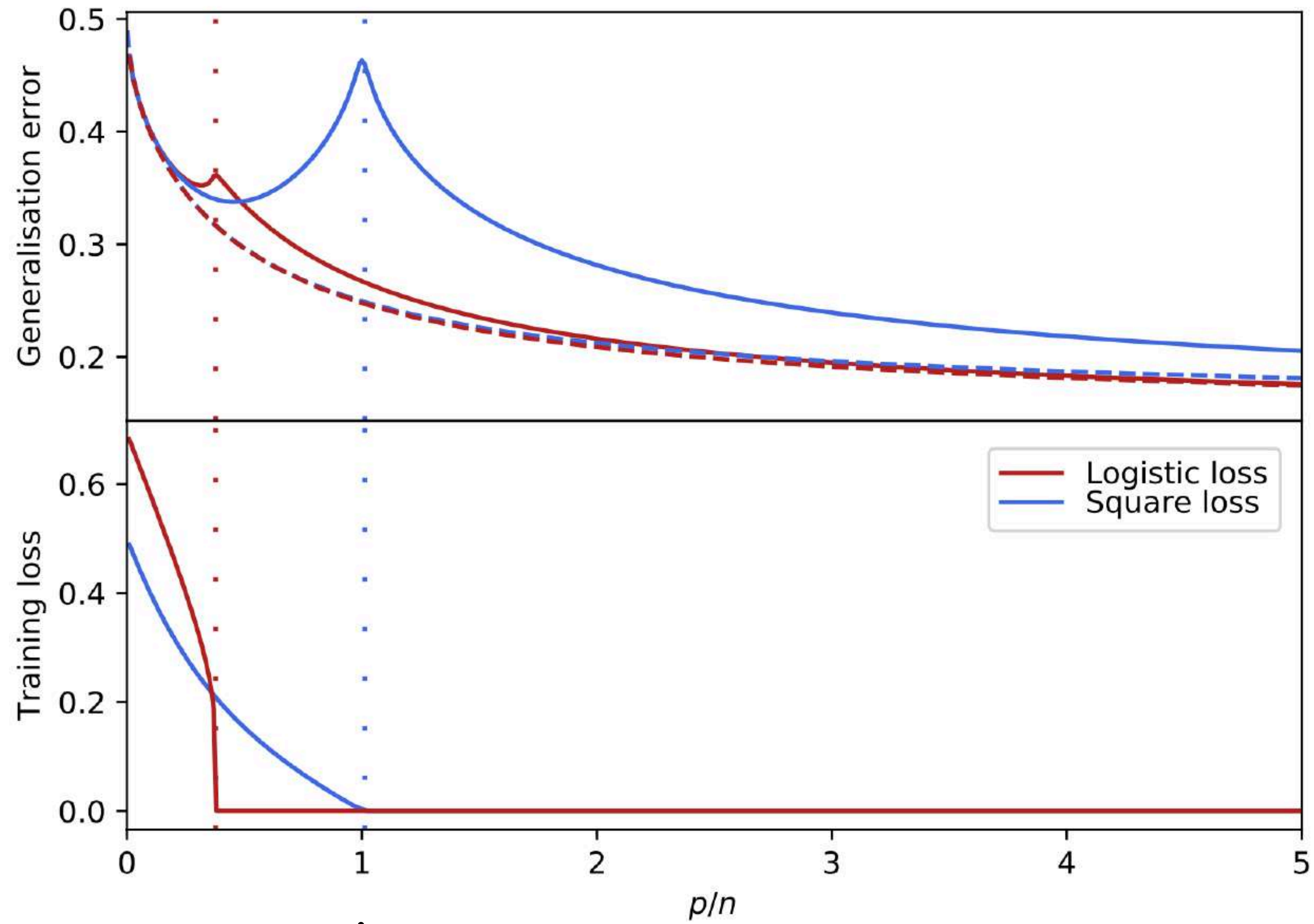
Focus on ℓ_2 loss $\lambda \to 0^+$.

What's going on?

Focus on ℓ_2 loss $\lambda \to 0^+$.

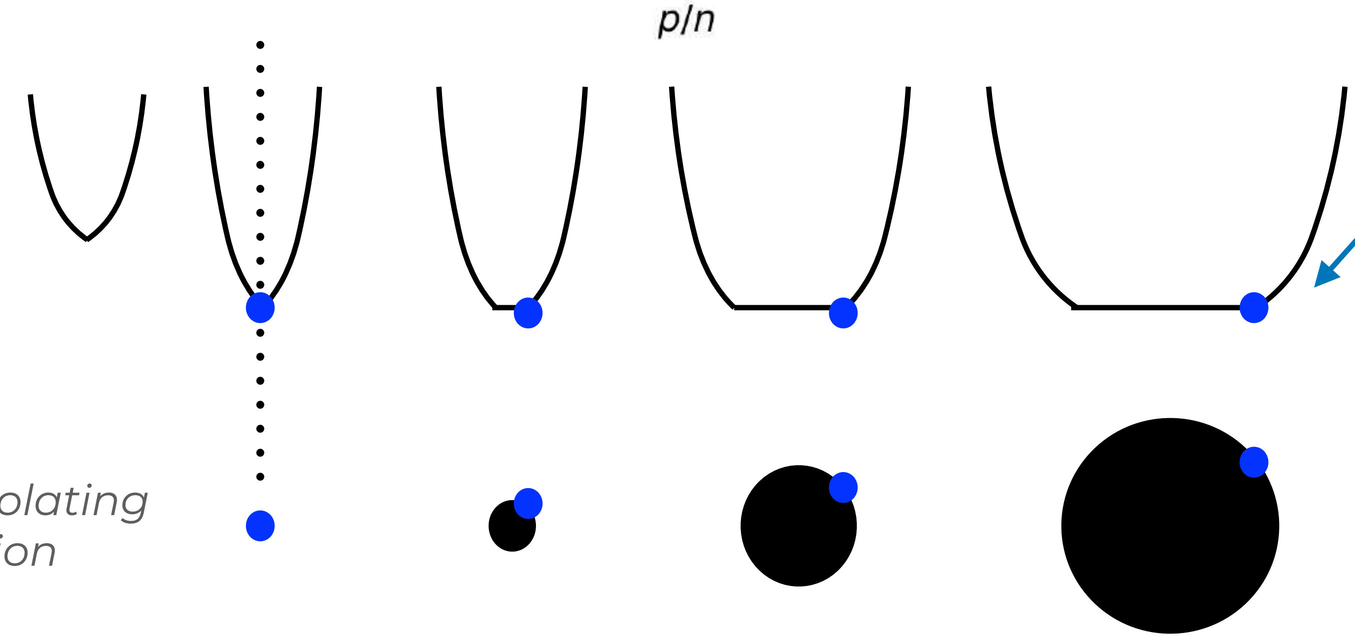


What’s going on?



ℓ_2 loss
$\lambda \to 0^+$.

Minimum ℓ_2-norm solution

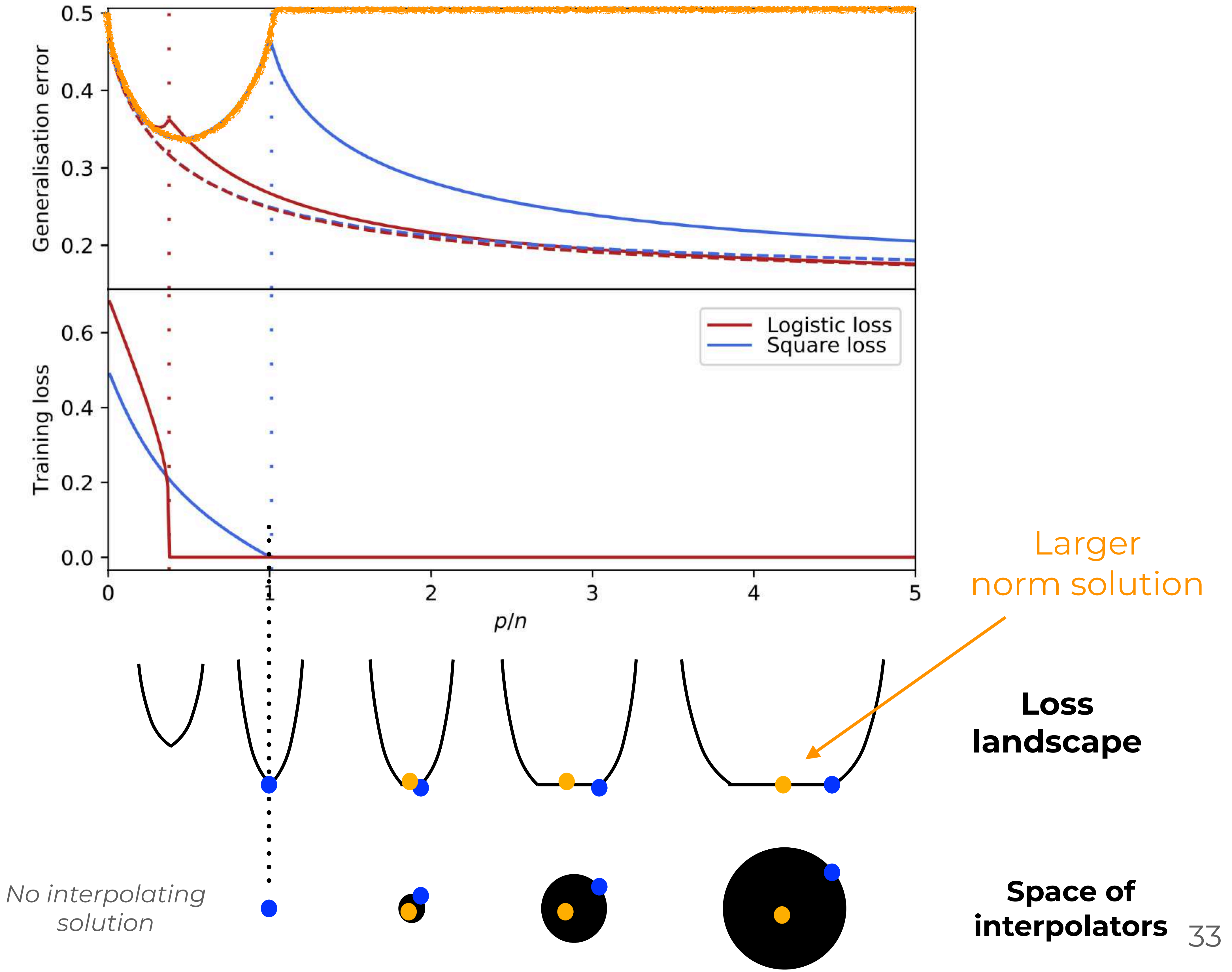


Loss landscape

No interpolating solution

Space of interpolators

What's going on?



ℓ_2 loss
$\lambda \to 0^+$.

Take away II:

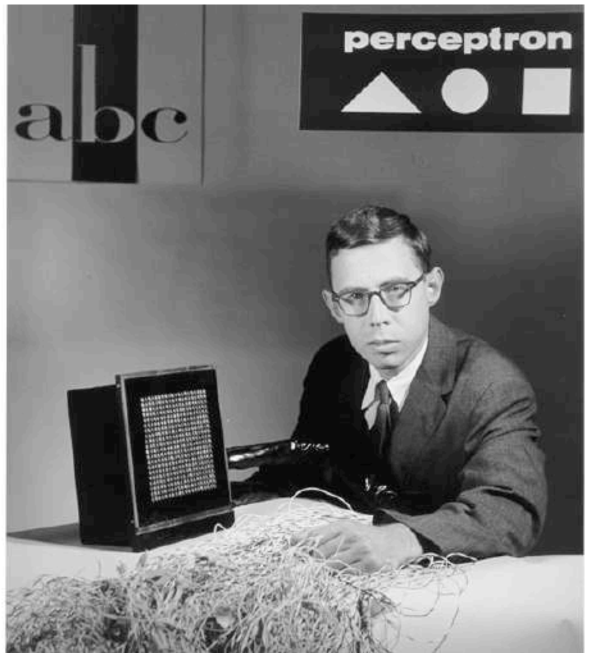
Overparametrisation is not at odds with generalisation

Benign overfitting can be understood from simple linear model

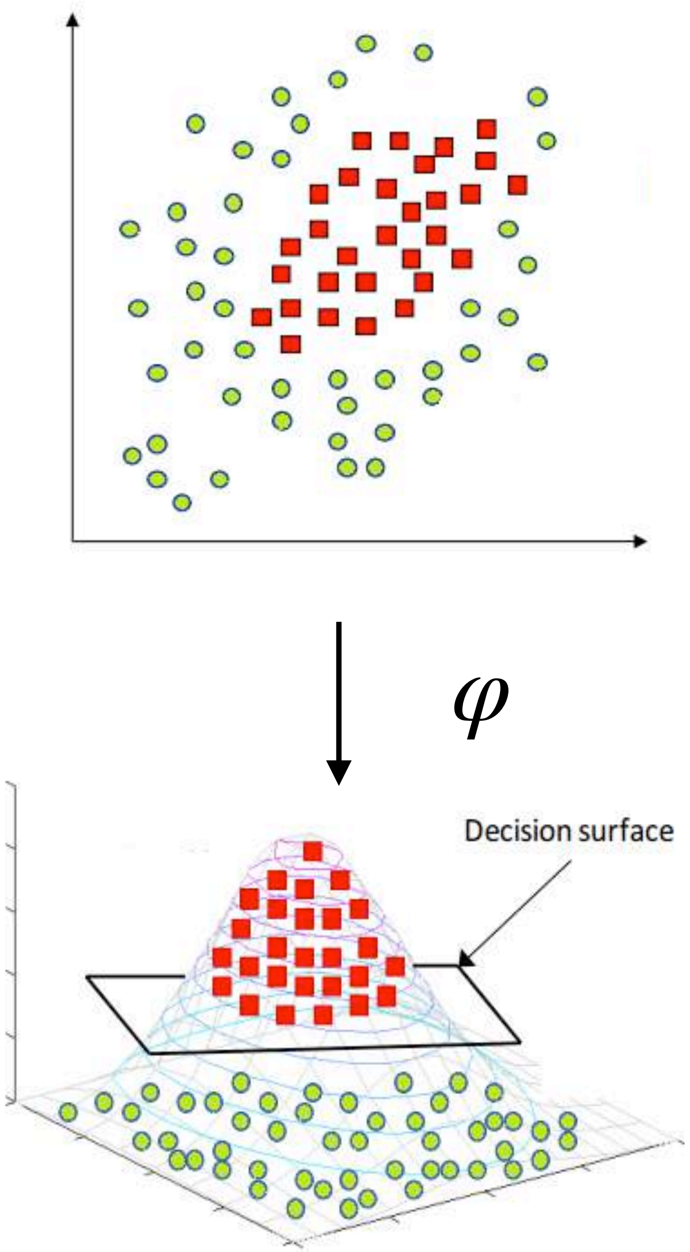
Implicit bias of algorithms

Menu for this tutorial

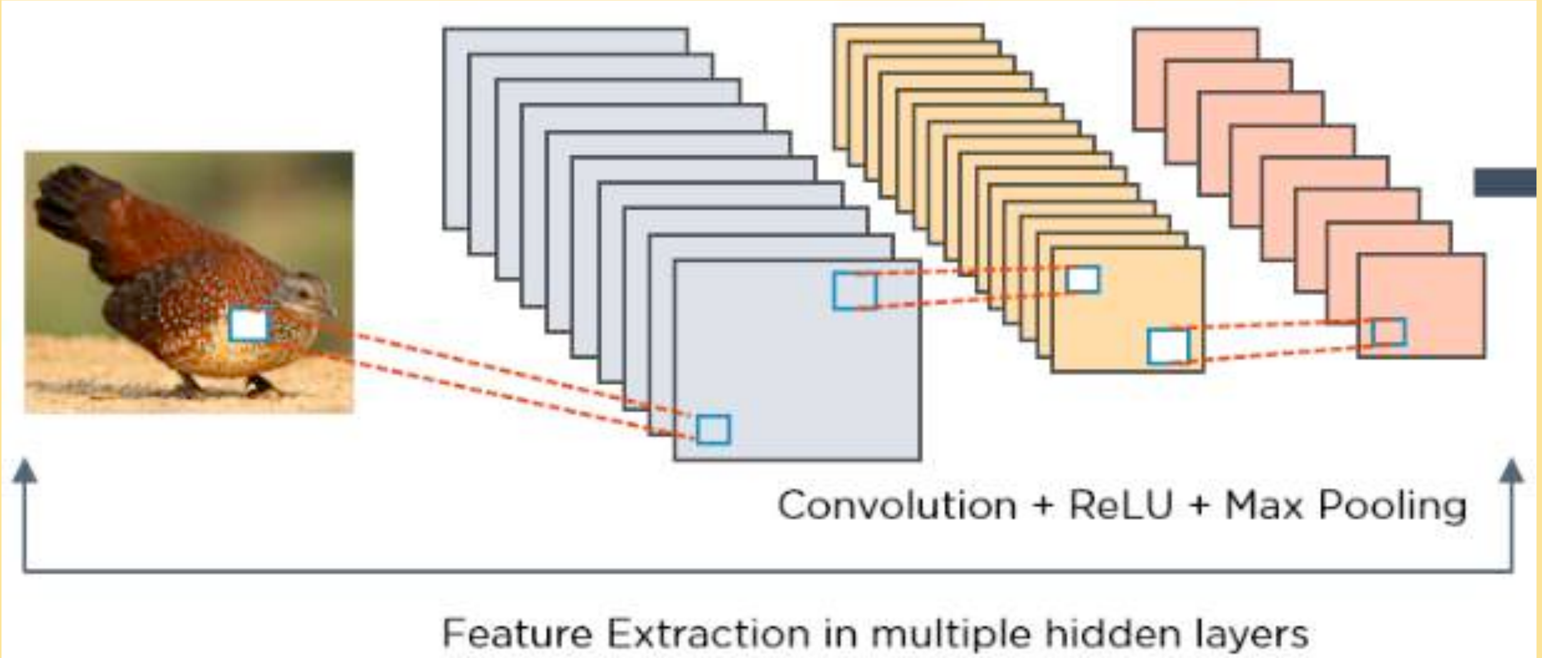
Part I: Statistical Physics of Computation



Part II: Neural Networks at initialisation (a.k.a. kernel methods)



Part III: Feature learning



Limitations of RF

Close connection between Gaussian universality and expressivity

Linear function of x

$$R(\hat{a}_\lambda) = \mathbb{E}[(y - \langle \hat{a}_\lambda, \sigma(W_0 x)\rangle)^2] \approx \mathbb{E}[(y - \langle \hat{a}_\lambda, \mu_0 1_p + W_0 x + \mu_\star z\rangle)^2]$$

Limitations of RF

Close connection between Gaussian universality and expressivity

Theorem [Mei, Misiakiewicz, Montanari '22, informal]

For isotropic data (e.g. $x \sim \mathrm{Unif}(\mathbb{S}^{d-1})$), with $n, p = \Theta(d^{\kappa})$ one can learn at best a polynomial approximation of degree κ of the target $f_\star(x)$

$$\mathbb{E}\,||f_\star(x) - f(x; \hat{a}_\lambda, W^0)||_2^2 = ||P_{>\kappa} f_\star||_{L_2}^2 + o_d(1)$$

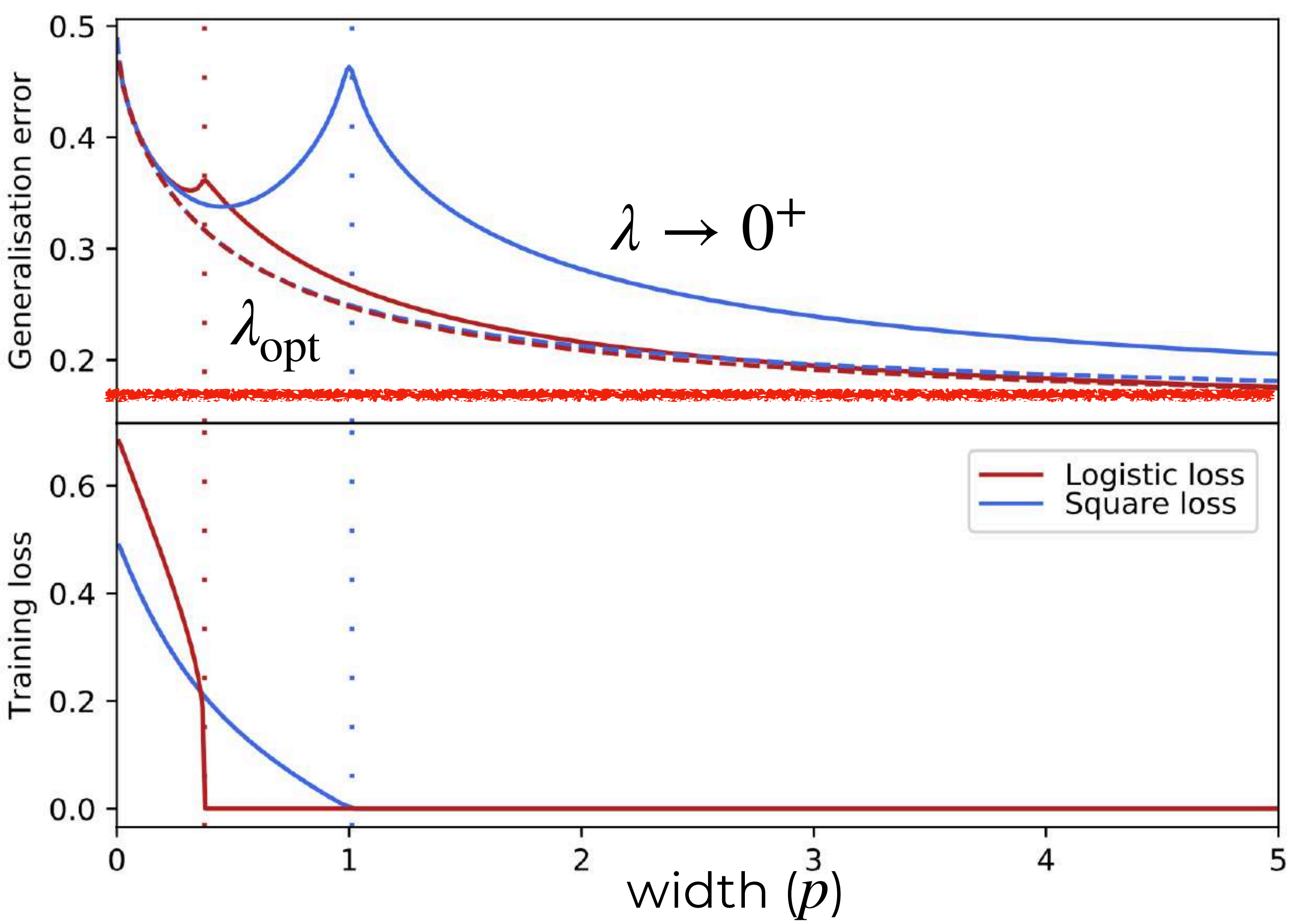
Limitations of RF

Close connection between Gaussian universality and expressivity

Theorem [Mei, Misiakiewicz, Montanari '22, informal]

For isotropic data (e.g. $x \sim \mathrm{Unif}(\mathbb{S}^{d-1})$), with $n, p = \Theta(d^\kappa)$ one can learn at best a polynomial approximation of degree κ of the target $f_\star(x)$

$$\mathbb{E}\,||f_\star(x) - f(x; \hat{a}_\lambda, W^0)||_2^2 = ||P_{>\kappa} f_\star||_{L_2}^2 + o_d(1)$$



In particular, for $n, p = \Theta(d)$, can learn at best a linear approximation of $f_\star$

$$f_\star(x) = \langle \theta_\star, x \rangle + f_{NL}(x)$$

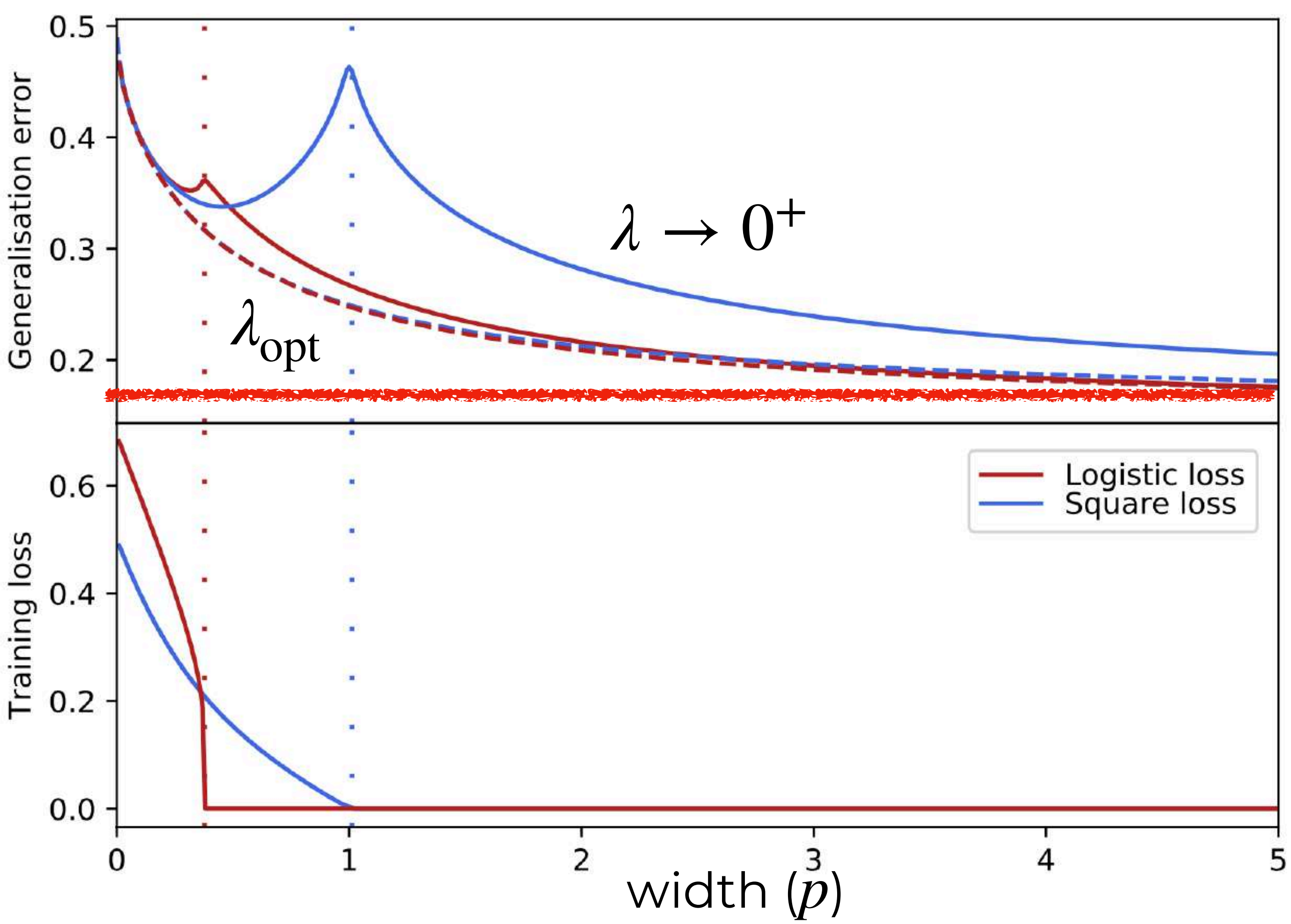
Limitations of RF

Close connection between Gaussian universality and expressivity

Theorem [Mei, Misiakiewicz, Montanari '22, informal]

For isotropic data (e.g. $x \sim \mathrm{Unif}(\mathbb{S}^{d-1})$), with $n, p = \Theta(d^\kappa)$ one can learn at best a polynomial approximation of degree κ of the target $f_\star(x)$

$$\mathbb{E}\,||f_\star(x) - f(x; \hat{a}_\lambda, W^0)||_2^2 = ||P_{>\kappa} f_\star||_{L_2}^2 + o_d(1)$$



In particular, for $n, p = \Theta(d)$, can learn at best a linear approximation of $f_\star$

$$f_\star(x) = \langle \theta_\star, x \rangle + f_{NL}(x)$$

To do better, need to learn features

One step of GD

Consider one step of GD from initialisation a^0, W^0 with fresh batch $(x_i, y_i)_{i \in [n_0]}$

$$W^{t+1} = W^t - \frac{\eta}{2|b_t|} \sum_{i \in b_t} \nabla_w (y_i - f(x_i; a_0, W^t))^2$$

One step of GD

Consider one step of GD from initialisation a^0, W^0 with fresh batch $(x_i, y_i)_{i\in[n_0]}$

$$W^{t+1} = W^t - \frac{\eta}{2|b_t|}\sum_{i\in b_t}\nabla_w(y_i - f(x_i; a_0, W^t))^2$$

Weak learnability: obtain non-trivial correlation with features:

$$\frac{\langle w_i^t, w_k^\star\rangle}{||w_i^1|| \cdot ||w_k^\star||} \xrightarrow{d\to\infty} M_{ik}^t > 0$$

One step of GD

Consider one step of GD from initialisation a^0, W^0 with fresh batch $(x_i, y_i)_{i\in[n_0]}$

$$W^{t+1} = W^t - \frac{\eta}{2|b_t|}\sum_{i\in b_t}\nabla_w (y_i - f(x_i; a_0, W^t))^2$$

Weak learnability: obtain non-trivial correlation with features:

$$\frac{\langle w_i^t, w_k^\star\rangle}{||w_i^1||\cdot||w_k^\star||} \xrightarrow{d\to\infty} M_{ik}^t > 0$$

Sample complexity depends on leap index ℓ of g:

$$g(z_1, \ldots, z_r) = \mu_0 + \sum_i \mu_i^{(1)} z_i + \sum_{ij} \mu_{ij}^{(2)} h_2(z_1, z_2) + \ldots$$

Morally: Smallest non-zero coefficient in Hermite expansion

What you learn in one-step of SGD?

Sample complexity determined by the leap index ℓ of g:

$$g(z_1, \dots, z_r) = \mu_0 + \sum_i \mu_i^{(1)} z_i + \sum_{ij} \mu_{ij}^{(2)} h_2(z_1, z_2) + \dots$$

Morally: Smallest non-zero coefficient in Hermite expansion

[Dandi, Krzakala, **BL**, Pesce, Stephan ’23]

What you learn in one-step of SGD?

Sample complexity determined by the leap index ℓ of g:

$$g(z_1, \dots, z_r) = \mu_0 + \sum_i \mu_i^{(1)} z_i + \sum_{ij} \mu_{ij}^{(2)} h_2(z_1, z_2) + \dots$$

Morally: Smallest non-zero coefficient in Hermite expansion

Examples:

$g(z) = z \qquad \ell = 1$

$g(z) = z^2 \qquad \ell = 2$

$g(z) = \tanh(z) \qquad \ell = 1$

$g(z) = z_1 z_2 \qquad \ell = 2$

$g(z) = z_1 \cdots z_r \qquad \ell = r$

[Dandi, Krzakala, **BL**, Pesce, Stephan ’23]

What you learn in one-step of SGD?

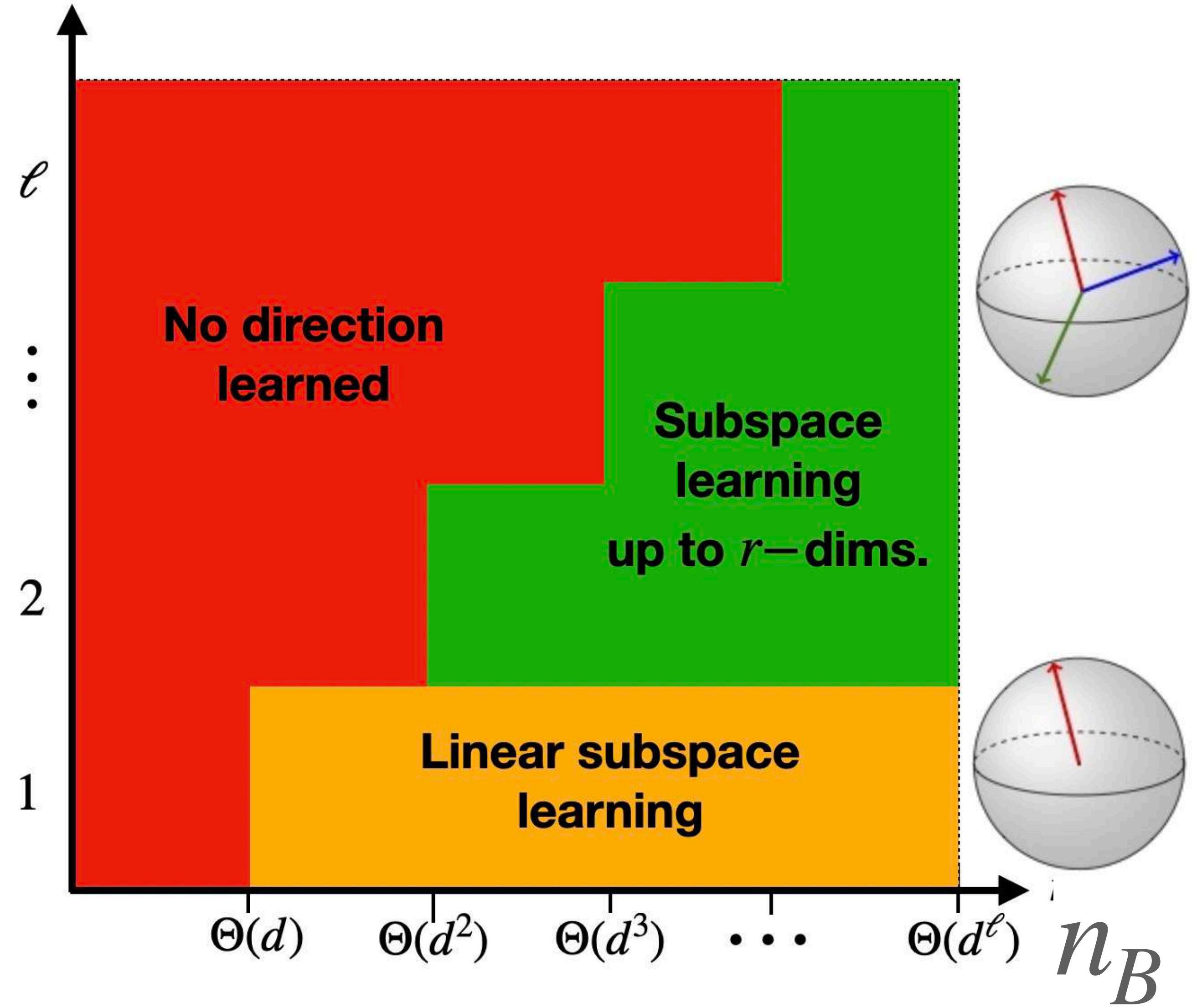
Sample complexity determined by the leap index ℓ of g:

$$g(z_1, \ldots, z_r) = \mu_0 + \sum_i \mu_i^{(1)} z_i + \sum_{ij} \mu_{ij}^{(2)} h_2(z_1, z_2) + \ldots$$

Morally: Smallest non-zero coefficient in Hermite expansion

Examples:

$g(z) = z$	$\ell = 1$
$g(z) = z^2$	$\ell = 2$
$g(z) = \tanh(z)$	$\ell = 1$
$g(z) = z_1 z_2$	$\ell = 2$
$g(z) = z_1 \cdots z_r$	$\ell = r$



[Dandi, Krzakala, **BL**, Pesce, Stephan '23]

Generalisation

We can show that at best learn non-linear functions along learned subspace:

Theorem [Dandi, Krzakala, BL, Pesce, Stephan '23, informal]

Let $U \subset \mathrm{span}(w_1^\star, \ldots, w_r^\star)$ be the space learned after a single SGD step. Then, for any a such that $||a||_\infty = \Theta_d(1)$:

$$\mathbb{E}||f_\star(x) - f(x; a, W^1)||_2^2 \geq \mathrm{Var}(f_\star(z)|P_U z) - o_d(1)$$

Generalisation

We can show that at best learn non-linear functions along learned subspace:

Theorem [Dandi, Krzakala, BL, Pesce, Stephan '23, informal]

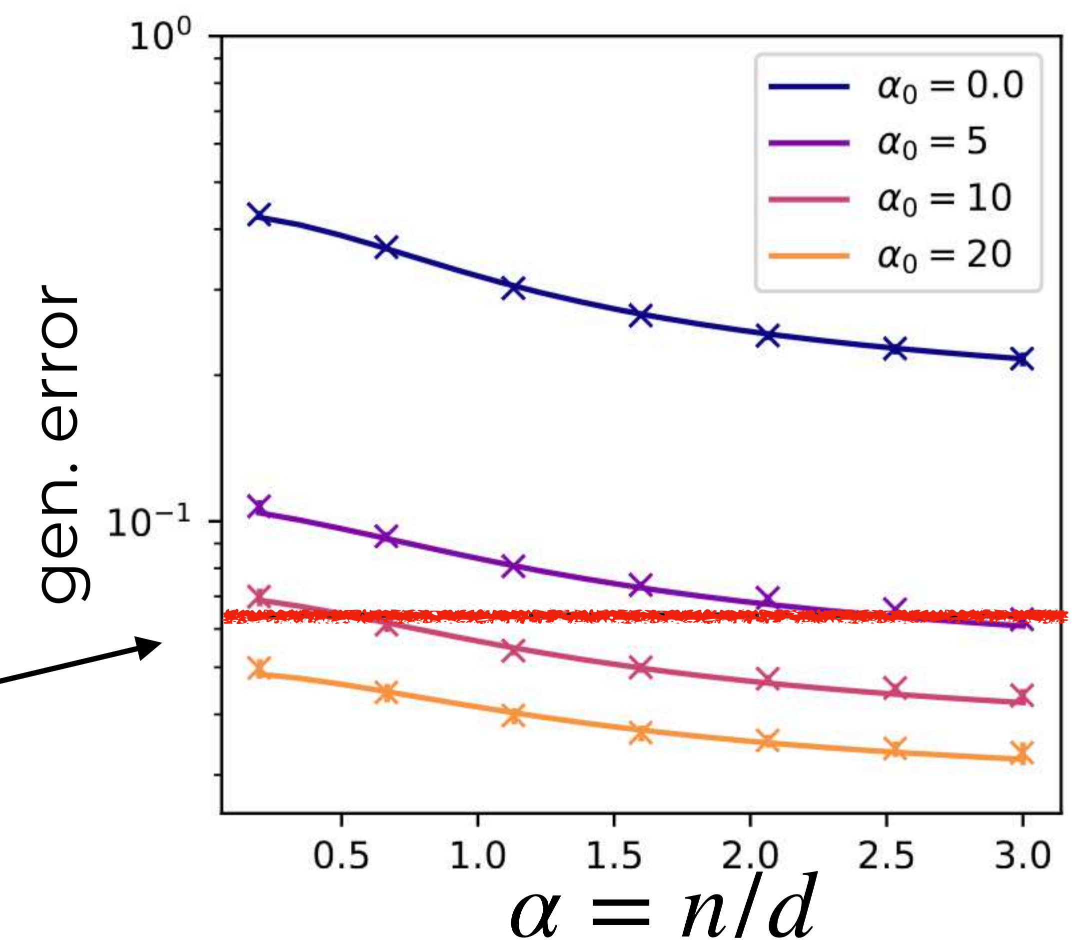
Let $U \subset \mathrm{span}(w_1^\star, \ldots, w_r^\star)$ be the space learned after a single SGD step. Then, for any a such that $||a||_\infty = \Theta_d(1)$:

$$\mathbb{E}||f_\star(x) - f(x; a, W^1)||_2^2 \geq \mathrm{Var}(f_\star(z)|P_U z) - o_d(1)$$

Can get exact results with $n, p = \Theta_d(d)$.

[Cui, Dandi, Pesce, Zdeborová, Lu, Krzakala, **BL** '24]

Best linear predictor $||P_{\kappa \leq 1} f_\star||^2$



Generalisation

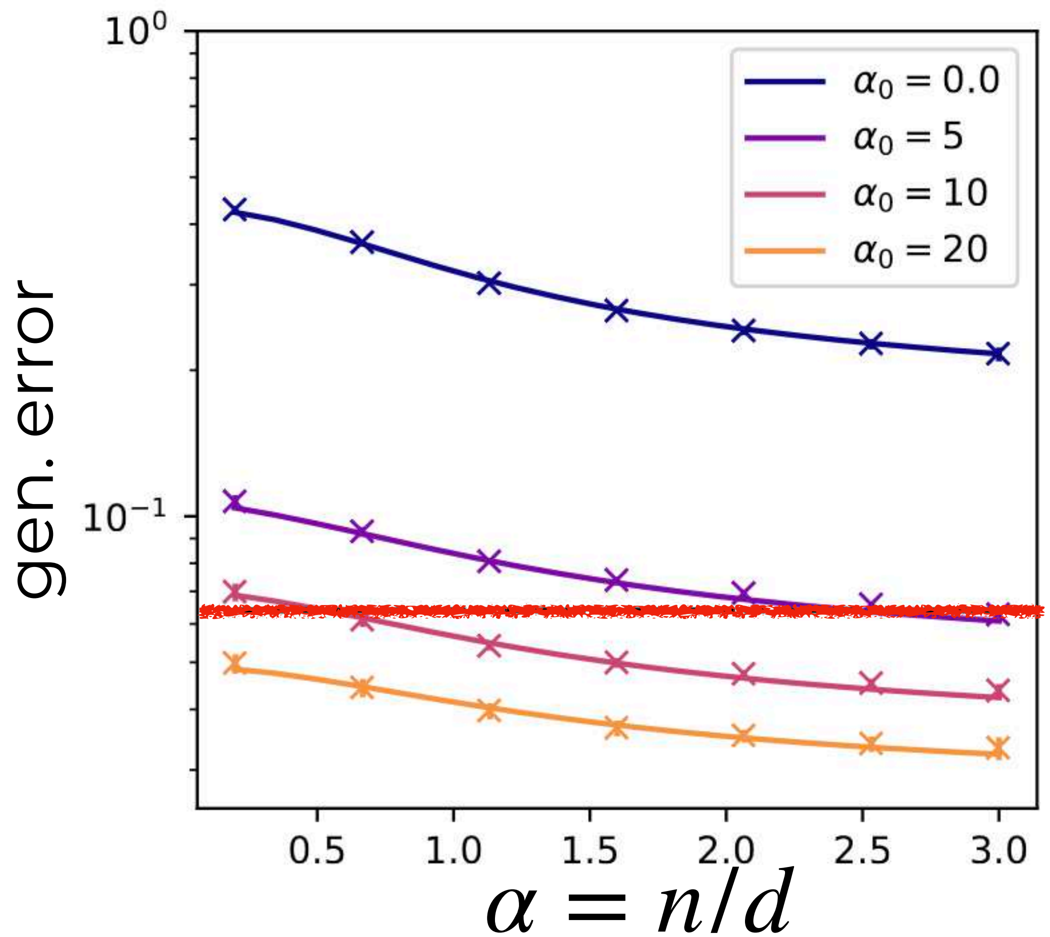
[Cui, Dandi, Pesce, Zdeborová, Lu, Krzakala, **BL** '24]

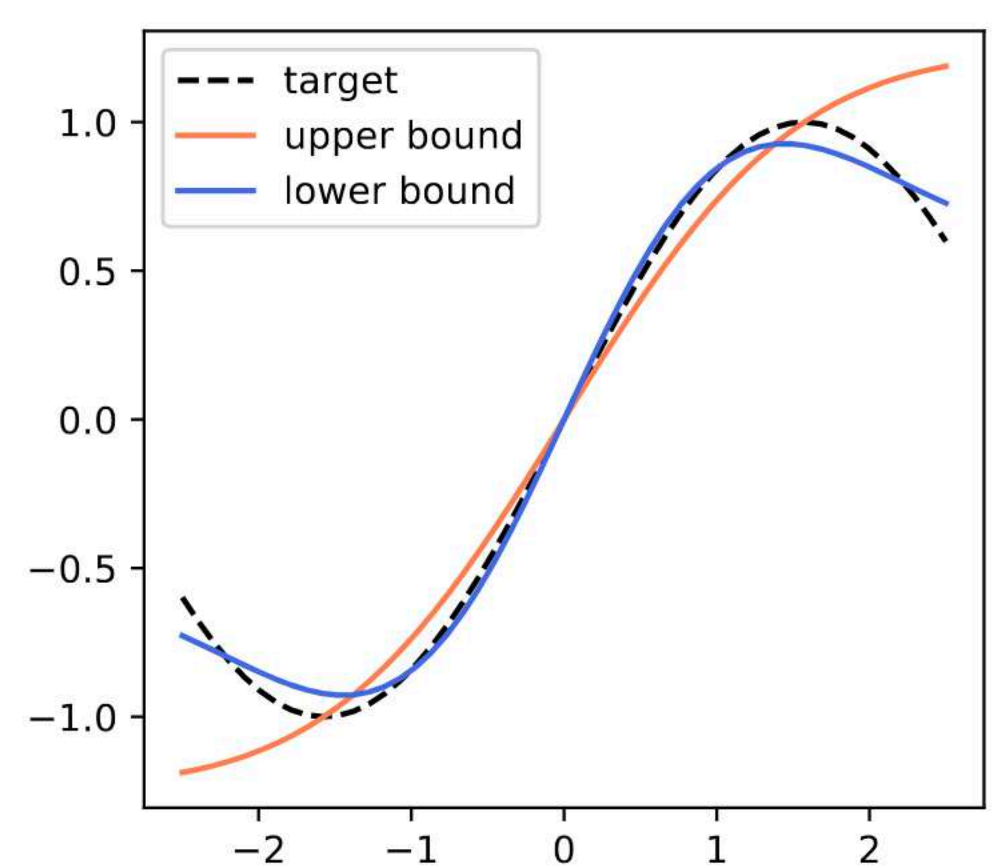
We can show that at best learn non-linear functions along learned subspace:

Theorem [Dandi, Krzakala, BL, Pesce, Stephan '23, informal]

Let $U \subset \mathrm{span}(w_1^\star, \dots, w_r^\star)$ be the space learned after a single SGD step. Then, for any a such that $||a||_\infty = \Theta_d(1)$:

$$\mathbb{E}||f_\star(x) - f(x; a, W^1)||_2^2 \geq \mathrm{Var}(f_\star(z)|P_U z) - o_d(1)$$





Take away III:

After one SGD step, first layer weights correlate with the relevant target directions

- For $n = \Theta(d)$, learn only averaged direction.
- At least $n = \Theta(d^2)$ required to learn more directions
- Exact $n = \Theta(d^\ell)$ depends on leap exponent of target.

🤔 What about multiple steps?

What you learn in multiple steps of GD?

Morally, it depends on how directions "interact".

In particular, there is a class of "easy" functions that can be sequentially learned with $n = \Theta(d)$

[Abbe et al. '22]

What you learn in multiple steps of GD?

Informally :

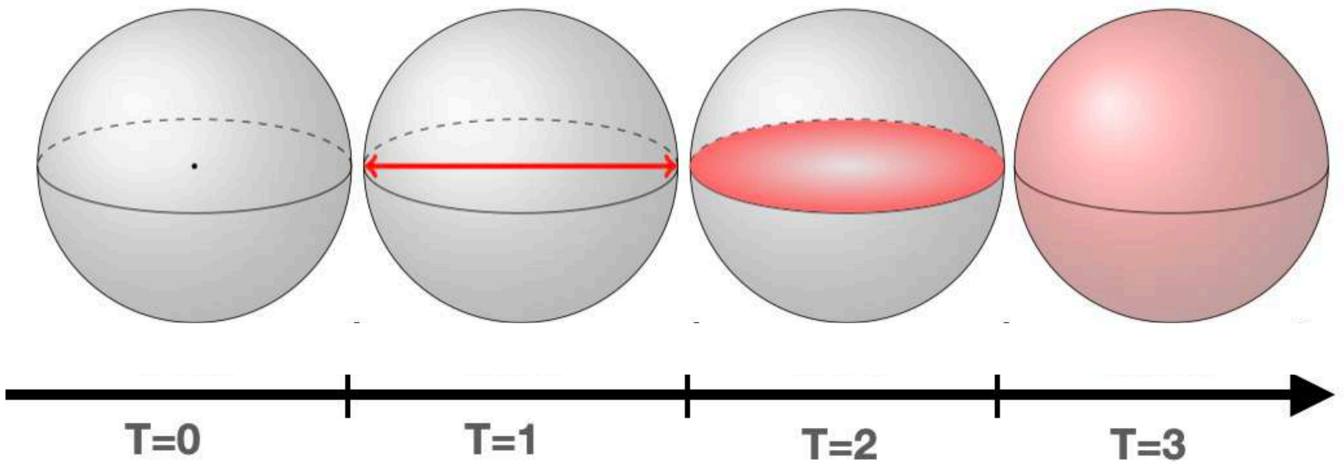
At each additional step, can learn a ***new directions*** each time, iff they are ***linear conditioned*** on the previously learned ones.

What you learn in multiple steps of GD?

Informally :

At each additional step, can learn a ***new directions*** each time, iff they are ***linear conditioned*** on the previously learned ones.

$$f^* = z_1 + g(z_1)z_2 + g(z_2)z_3$$

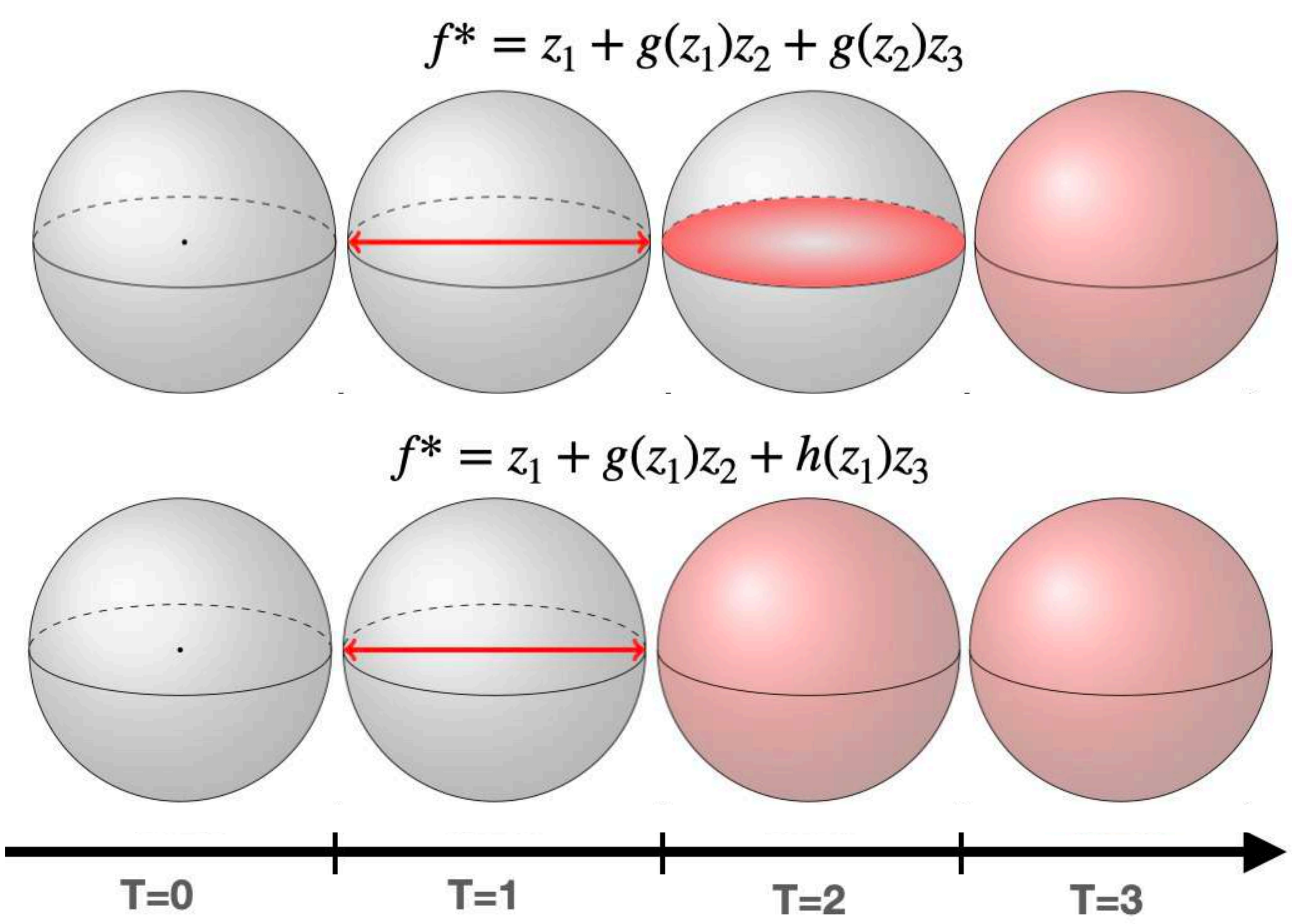


[Dandi, Krzakala, **BL**, Pesce, Stephan '23]

What you learn in multiple steps of GD?

Informally :

At each additional step, can learn a ***new directions*** each time, iff they are ***linear conditioned*** on the previously learned ones.

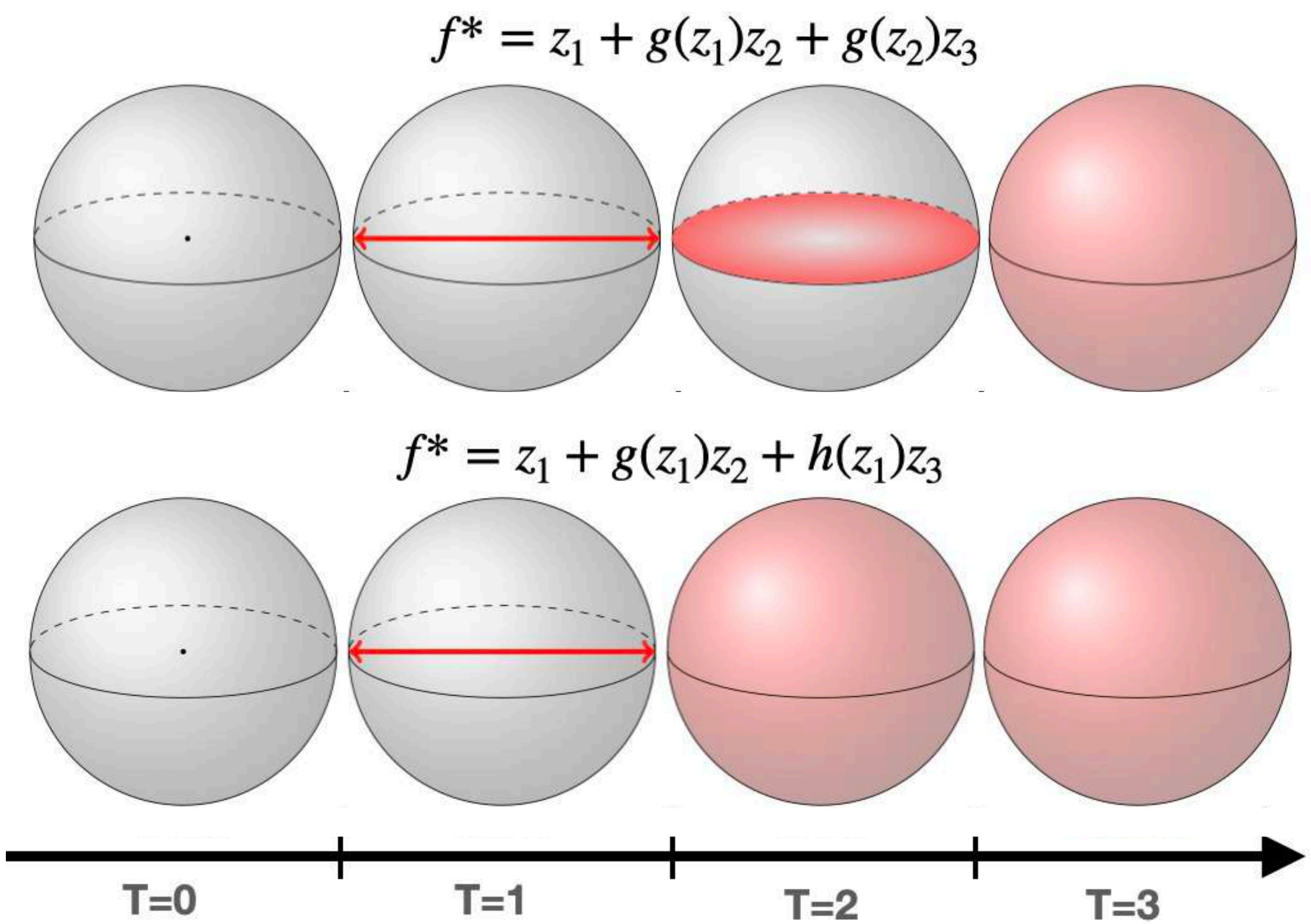


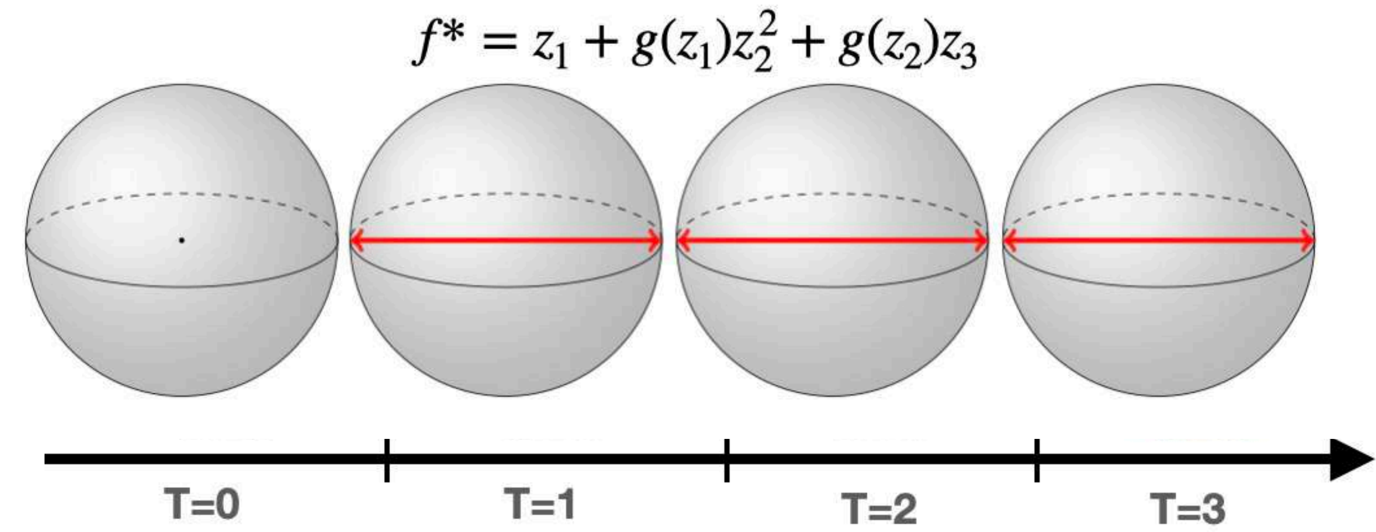
[Dandi, Krzakala, **BL**, Pesce, Stephan ’23]

What you learn in multiple steps of GD?

Informally :

At each additional step, can learn a ***new directions*** each time, iff they are ***linear conditioned*** on the previously learned ones.





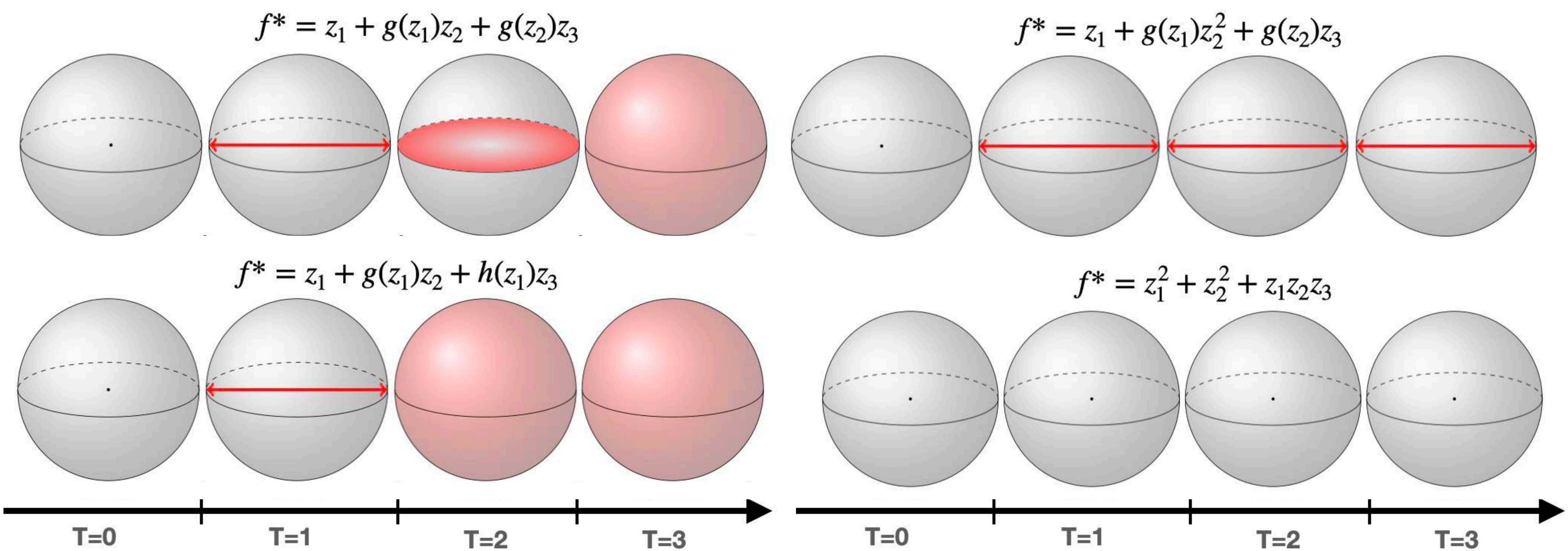
"SGD easy"

[Dandi, Krzakala, **BL**, Pesce, Stephan '23]

What you learn in multiple steps of GD?

Informally :

At each additional step, can learn a ***new directions*** each time, iff they are ***linear conditioned*** on the previously learned ones.



"SGD easy"   "SGD hard"

[Dandi, Krzakala, **BL**, Pesce, Stephan '23]

Take away IV:

With more than one step, might learn linearly correlated subspaces.

In particular, there are classes of multi-index models that can be learned in with $n = \Theta_d(d)$

🤔 Better than kernels, but fundamental computational barrier?

Fundamental limitation?

SGD learning on neural networks:
leap complexity and saddle-to-saddle dynamics

Emmanuel Abbe[*], Enric Boix-Adserà[†], Theodor Misiakiewicz[‡]

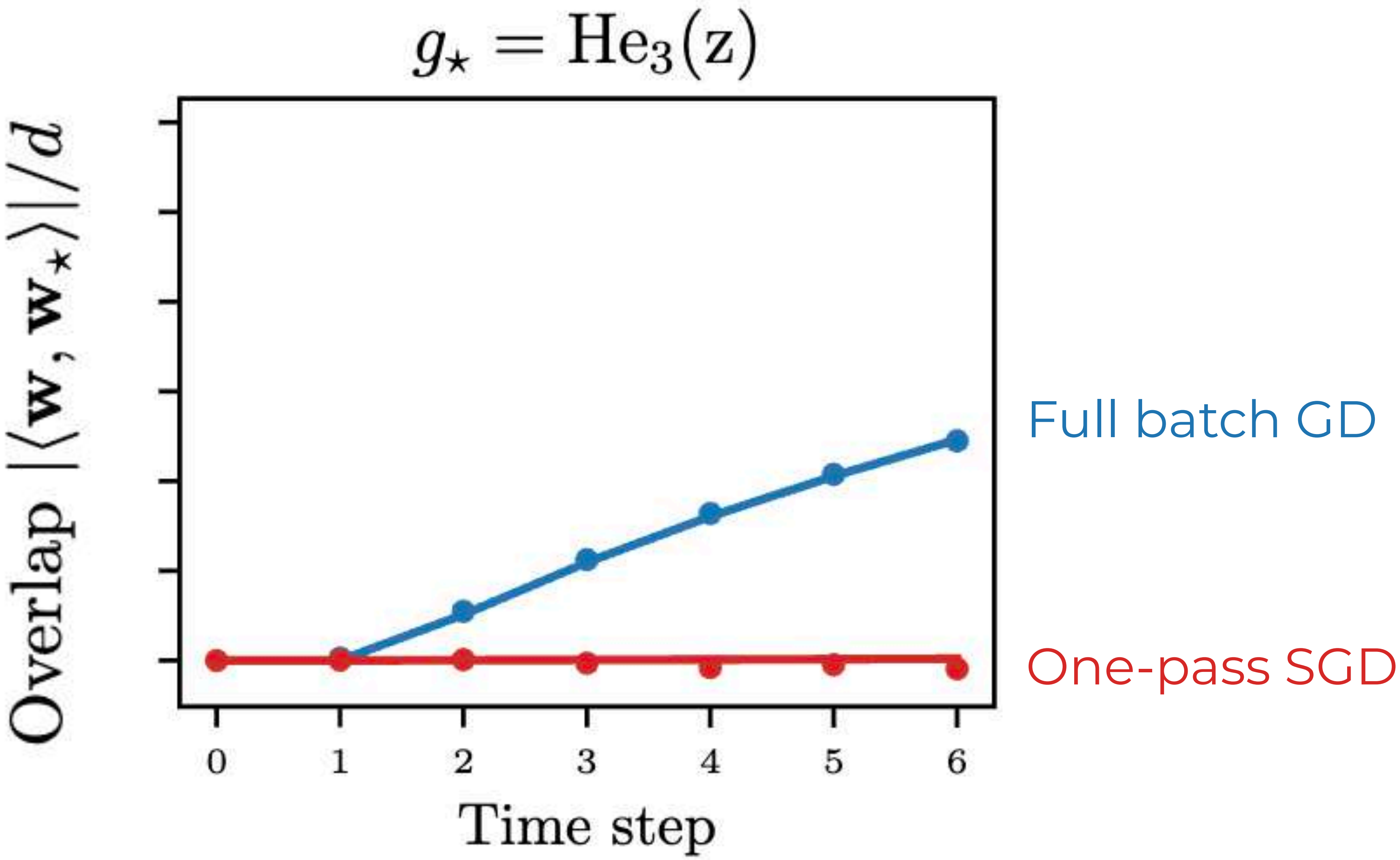
September 4, 2023

Finally we note that we considered here the setting of online-SGD, and a natural question is to consider how the picture may change under ERM (several passes with the same batch of samples). The ERM setting is however harder to analyze. We consider this to be an important direction for future works. Note that our results imply a sample complexity equal to the number of SGD steps $n = t = \widetilde{\Theta}(d^{\max(\mathrm{Leap}-1,1)})$. In ERM, we reuse samples and consequently reduce the sample complexity. We conjecture in fact that $n = \widetilde{\Theta}(d^{\max(\mathrm{Leap}/2,1)})$ is optimal for ERM. Furthermore,

See also [Biroli, Cammarota, Ricci-Tersenghi '19; Damian, Niching, Ge, Lee '23]

Fundamental limitation?

Single index $r = 1$



Recall: kernel requires $n = \Theta(d^3)$

[Dandi, Troiani, Arnaboldi, Pesce, Zdeborová, Krzakala '24;
Arnaboldi, Dandi, Krzakala, Pesce, Stephan 24']

Closer look at 1st step

Consider one step of SGD $w_k^1 = w_k^0 - \eta g_k^0$ $y = \varphi(\langle w_\star, x\rangle)$

$$g_k^0 = -\frac{1}{|b_0|p}\sum_{i\in b_0}\left(y_i - \frac{1}{p}\sum_{l=1}^{p} a_l^0 \sigma(\langle w_l^0, x_i\rangle)\right) a_k^0 \sigma'(\langle w_k^0, x_i\rangle)x_i$$

Closer look at 1st step

Consider one step of SGD $\quad w_k^1 = w_k^0 - \eta g_k^0 \qquad y = \varphi(\langle w_\star, x \rangle)$

$$g_k^0 = -\frac{1}{|b_0|p} \sum_{i \in b_0} \left(y_i - \frac{1}{p} \sum_{l=1}^{p} a_l^0 \sigma(\langle w_l^0, x_i \rangle) \right) a_k^0 \sigma'(\langle w_k^0, x_i \rangle) x_i$$

Therefore, on expectation:

$$\mathbb{E}[g_k^0] = -\frac{a_k^0}{p} \mathbb{E}[y_i \sigma'(\langle w_k^0, x_i \rangle) x_i] + \text{ Other (important) stuff}$$

Ind. from (x_i, y_i)

Closer look at 1st step

Consider one step of SGD $\quad w_k^1 = w_k^0 - \eta g_k^0 \qquad y = \varphi(\langle w_\star, x \rangle)$

$$g_k^0 = -\frac{1}{|b_0|p} \sum_{i \in b_0} \left(y_i - \frac{1}{p} \sum_{l=1}^{p} a_l^0 \sigma(\langle w_l^0, x_i \rangle) \right) a_k^0 \sigma'(\langle w_k^0, x_i \rangle) x_i$$

Therefore, on expectation:

$$\mathbb{E}[g_k^0] = -\frac{a_k^0}{p} \mathbb{E}[y_i \sigma'(\langle w_k^0, x_i \rangle) x_i] + \text{ Other (important) stuff}$$

Ind. from (x_i, y_i)

$$= -\frac{a_k^0}{p} \mu_\ell^\star \mu_{\ell+1} \frac{\langle w_k^0, w_\star \rangle^\ell}{d} + \cdots$$

$= \Theta(d^{-\ell})$

The first only access data through a CSQ query $\mathbb{E}[y\phi(x)]$

Closer look at 2nd step

Consider one step of SGD $\quad w_k^2 = w_k^1 - \eta g_k^1 \qquad y = \varphi(\langle w_\star, x \rangle)$

$$g_k^1 = -\frac{1}{|b_1|p} \sum_{i \in b_1} \left(y_i - \frac{1}{p} \sum_{l=1}^{p} a_l^0 \sigma(\langle w_l^1, x_i \rangle) \right) a_k^0 \sigma'(\langle w_k^1, x_i \rangle) x_i$$

Therefore, on expectation:

$$\mathbb{E}[g_k^1] = -\frac{a_k^0}{p} \mathbb{E}[y_i \sigma'(\langle w_k^1, x_i \rangle) x_i] + \text{ Other (important) stuff}$$

Closer look at 2nd step

Consider one step of SGD $w_k^2 = w_k^1 - \eta g_k^1$ $y = \varphi(\langle w_\star, x \rangle)$

$$g_k^1 = -\frac{1}{|b_1|p} \sum_{i \in b_1} \left(y_i - \frac{1}{p} \sum_{l=1}^{p} a_l^0 \sigma(\langle w_l^1, x_i \rangle) \right) a_k^0 \sigma'(\langle w_k^1, x_i \rangle) x_i$$

Therefore, on expectation:

$$\mathbb{E}[g_k^1] = -\frac{a_k^0}{p} \mathbb{E}[y_i \sigma'(\langle w_k^1, x_i \rangle) x_i] + \text{ Other (important) stuff}$$

Distinguish <u>2 cases:</u> 1. Fresh batch: $b_1 \perp b_0 \Rightarrow \mathbb{E}[\langle w^1, x_i \rangle] = 0 \quad i \in b_1$

Same as before!

Closer look at 2nd step

Consider one step of SGD $\quad w_k^2 = w_k^1 - \eta g_k^1 \qquad y = \varphi(\langle w_\star, x \rangle)$

$$g_k^1 = -\frac{1}{|b_1|p}\sum_{i\in b_1}\left(y_i - \frac{1}{p}\sum_{l=1}^{p} a_l^0 \sigma(\langle w_l^1, x_i\rangle)\right) a_k^0 \sigma'(\langle w_k^1, x_i\rangle) x_i$$

Therefore, on expectation:

$$\mathbb{E}[g_k^1] = -\frac{a_k^0}{p}\mathbb{E}[y_i \sigma'(\langle w_k^1, x_i\rangle) x_i] + \text{Other (important) stuff}$$

Distinguish 2 cases: 1. Fresh batch: $b_1 \perp b_0 \;\Rightarrow\; \mathbb{E}[\langle w^1, x_i\rangle] = 0 \quad i \in b_1$

Same as before!

2. Correlated batch: e.g. $b_1 = b_0 \;\Rightarrow\; \mathbb{E}[\langle w_k^1, x_i\rangle] \neq 0$

g_1 access data through a more general SQ query $\mathbb{E}[\phi(x, y)]$.

Leap exponent not invariant under $y \mapsto \phi(y)$!

What are the fundamental barriers for weak learnability in these models?

Remember?

[Barbier et al. '17; Mondelli, Montanari '17; Maillard, **BL**, Krzakala, Zdeborová '20;]

Signal		Likelihood		Observation
$w_\star \sim P_\star$	$\longrightarrow$	$P(y \mid X w_\star)$	$\longrightarrow$	$y \in \mathbb{R}^n$

X Gaussian.

Remember?

[Barbier et al. '17; Mondelli, Montanari '17; Maillard, **BL**, Krzakala, Zdeborová '20;]

Estimate $w_\star \in \mathbb{R}^d$ from n observations:

$$y_i = g(w^\star x_i)$$

$$x_i \sim \mathcal{N}(0, I_d/d)$$

+

Knowledge of
$w^\star \in \mathbb{S}^{d-1}(\sqrt{d})$ and $P(y \,|\, W_\star x)$

Remember?

[Barbier et al. '17; Mondelli, Montanari '17; Maillard, **BL**, Krzakala, Zdeborová '20;]

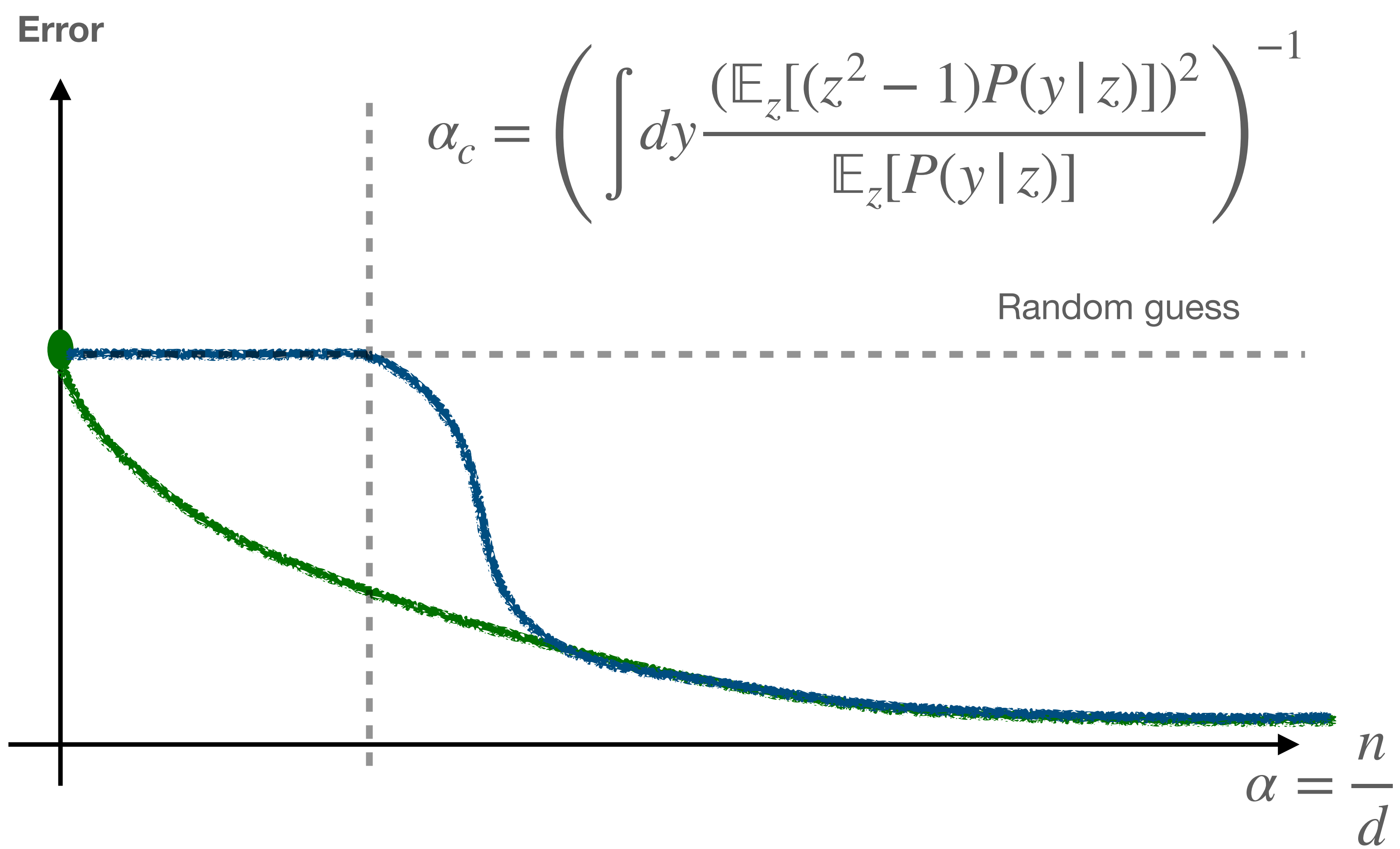
Estimate $w_\star \in \mathbb{R}^d$ from n observations:

$$y_i = g(w^\star x_i)$$

$$x_i \sim \mathcal{N}(0, I_d/d)$$

+ Knowledge of $w^\star \in \mathbb{S}^{d-1}(\sqrt{d})$ and $P(y \mid W_\star x)$

Recall: G-AMP achieves optimal weak recovery threshold.



"Generic" Likelihoods:

For any $n > 0$, beat random guess

e.g. $g(z) = z^3 - 3z$

Symmetric Likelihood:

Need $n = \Theta(d)$, large enough

e.g. $g(z) = z^2 - 1$

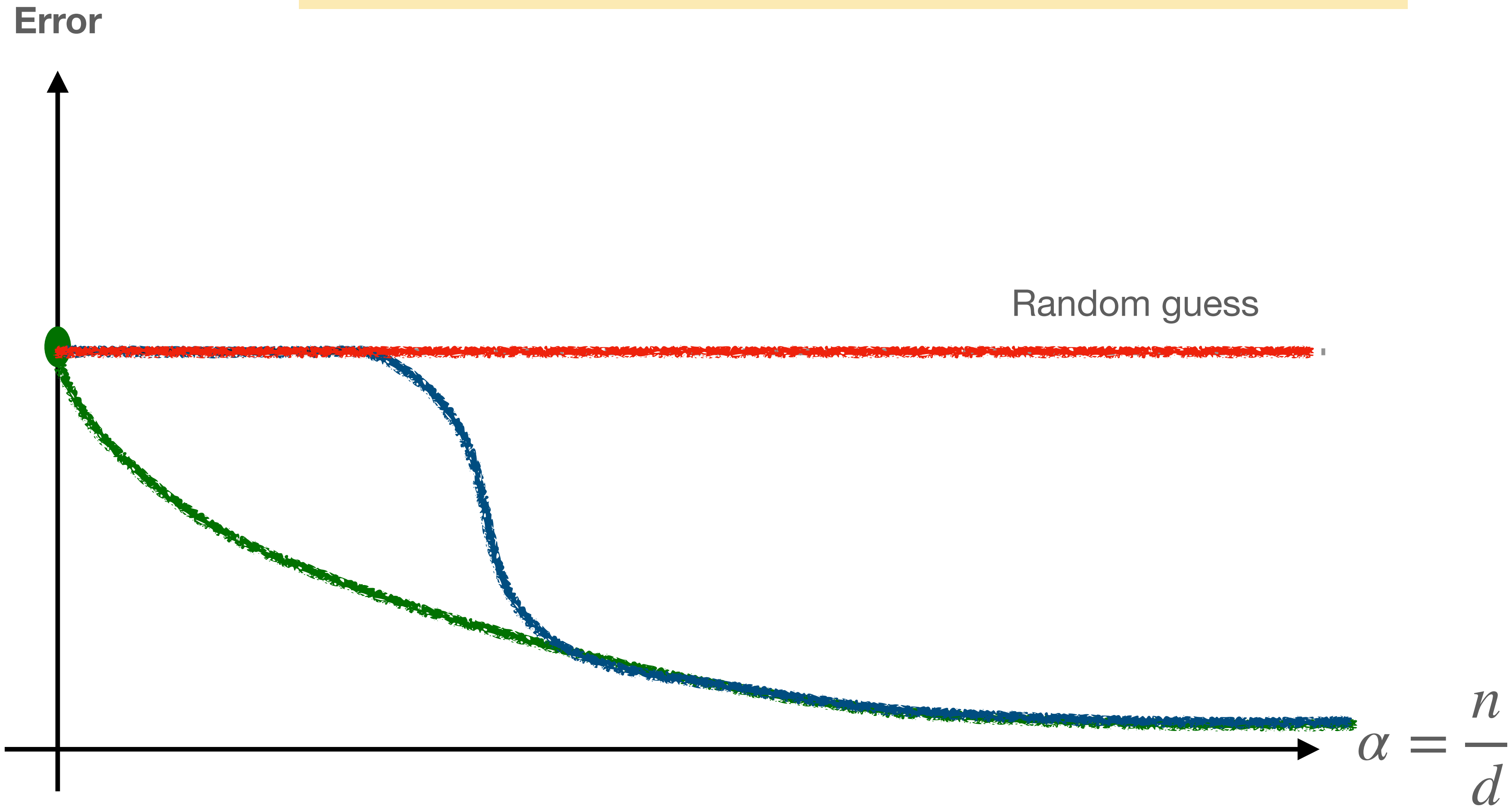
$(\alpha_c = 1/2)$

Multi-index

[Aubin et al. '18;
Troiani, Dandi, Defilippis, Zdeborová, **BL**, Krzakala '24;]

Similar story for $r > 1$.

$$y_i = g(z_1, \ldots, z_r) \qquad z_k = \langle w_k^\star, x \rangle$$



<u>Trivial subspaces:</u>
For any $n > 0$,
beat random guess
e.g. $g(z) = \tanh(z_1 z_2 z_3)$

<u>Easy subspaces:</u>
Need $n = \Theta(d)$,
large enough
e.g. $g(z) = z_1 z_2 z_3$
$\alpha_c \approx 3.725$

<u>Hard subspaces:</u>
Need $n > \Theta(d)$,
e.g. $g(z) = \text{sign}(z_1 z_2 z_3)$
"Parity-like"
$\alpha_c \to \infty$

A TCS point of view

One-pass SGD

Access data through $\mathbb{E}[y\phi(x)]$

Full-batch GD

Access data through $\mathbb{E}[\mathcal{T}_{GD}(y)\phi(x)]$

G-AMP

Access data through $\mathbb{E}[\mathcal{T}_{AMP}(y)\phi(x)]$

A TCS point of view

One-pass SGD	Full-batch GD	G-AMP
Access data through $\mathbb{E}[y\phi(x)]$	Access data through $\mathbb{E}[\mathcal{T}_{GD}(y)\phi(x)]$	Access data through $\mathbb{E}[\mathcal{T}_{AMP}(y)\phi(x)]$

Theorem [Troiani, Dandi, Defilippis, Zdeborová, BL, Krzakala '24, informal]

If $\mathbb{E}[\mathcal{T}_{AMP}(y)\phi(x)] = 0$ then $\mathbb{E}[\mathcal{T}(y)\phi(x)] = 0$

For any measurable $\mathcal{T} : \mathbb{R} \to \mathbb{R}$

Moreover, for single index models ($r = 1$) $\alpha_{c,AMP}$ matches SQ sample complexity.

Computational-Statistical Gaps in Gaussian Single-Index Models

Alex Damian[1], Loucas Pillaud-Vivien[2], Jason D. Lee[3], and Joan Bruna[4,5]

March 14, 2024

Hierarchical learning

$$y_i = z_1^2 + \text{sign}(z_1 z_2 z_3) \qquad z_k = \langle w_k^\star, x \rangle$$

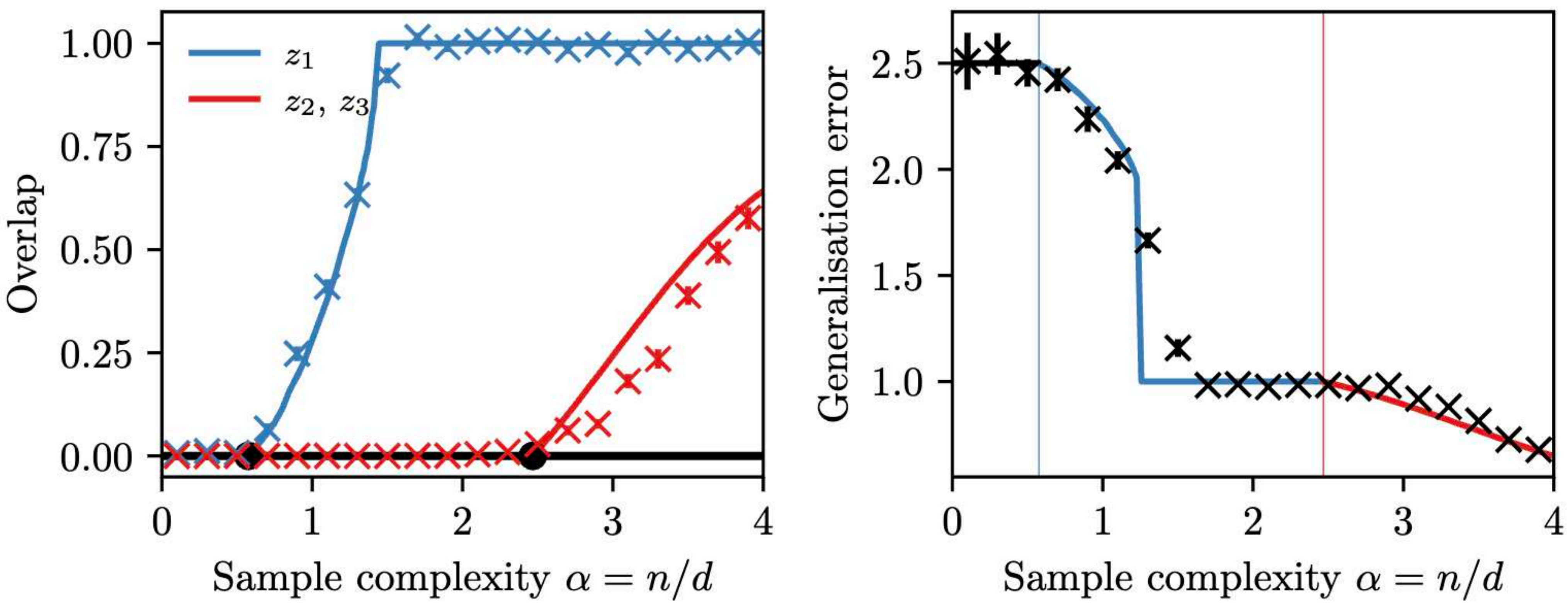


Figure 2: Hierarchical weak learnability for the staircase function $g(z_1, z_2, z_3) = z_1^2 + \text{sign}(z_1 z_2 z_3)$. (**Left**): Overlaps with the first direction $|M_{11}|$ (blue), and with the second and third one $1/2(M_{22} + M_{33})$ (red) as a function of the sample complexity $\alpha = n/d$, with solid lines denoting state evolution curves Equation (8), and crosses/dots finite-size runs of AMP Algorithm 1 with $d = 500$ and averaged over 72 seeds. All other overlaps are zero (black). The two black dots indicate the critical thresholds at $\alpha_1 \approx 0.575$ and $\alpha_2 = \pi^2/4$. (**Right**) Corresponding generalization error as a function of the sample complexity. Details on the numerical implementation are discussed in Appendix D.

Take away V:

Learning features improves allow shallow networks to learn more efficiently

Benefit of multi-pass over single-pass SGD for weak learnability

G-AMP classification of trivial, easy and hard subspaces

Overview

Part I: Statistical physics point of view on computational complexity: a landscape point of view

Part II: Shallow networks at initialisation: Double descent and benign overfitting in a convex, linear model.

Part III: Benefits of feature learning in shallow networks: Sample complexity and hierarchical learning phenomena

But this is only the tip of an iceberg…

brloureiro@gmail.com

Thank you

L. Zdeborová
(EPFL)

F. Krzakala
(EPFL)

L. Stephane
(EPFL -> ENSAI)

L. Defilippis
(DI-ENS)

G. Reeves
(Duke)

H. Cui
(Harvard)

E. Troiani
(EPFL)

L Pesce
(EPFL)

Y. Dandi
(EPFL)

C. Gerbelot
(Courant)

Y.M. Lu
(Harvard)

T. Misiakiewicz
(Yale)

S. Goldt
(SISSA)

F. Gerace
(Bologna)

M. Mézard
(Bocconi)