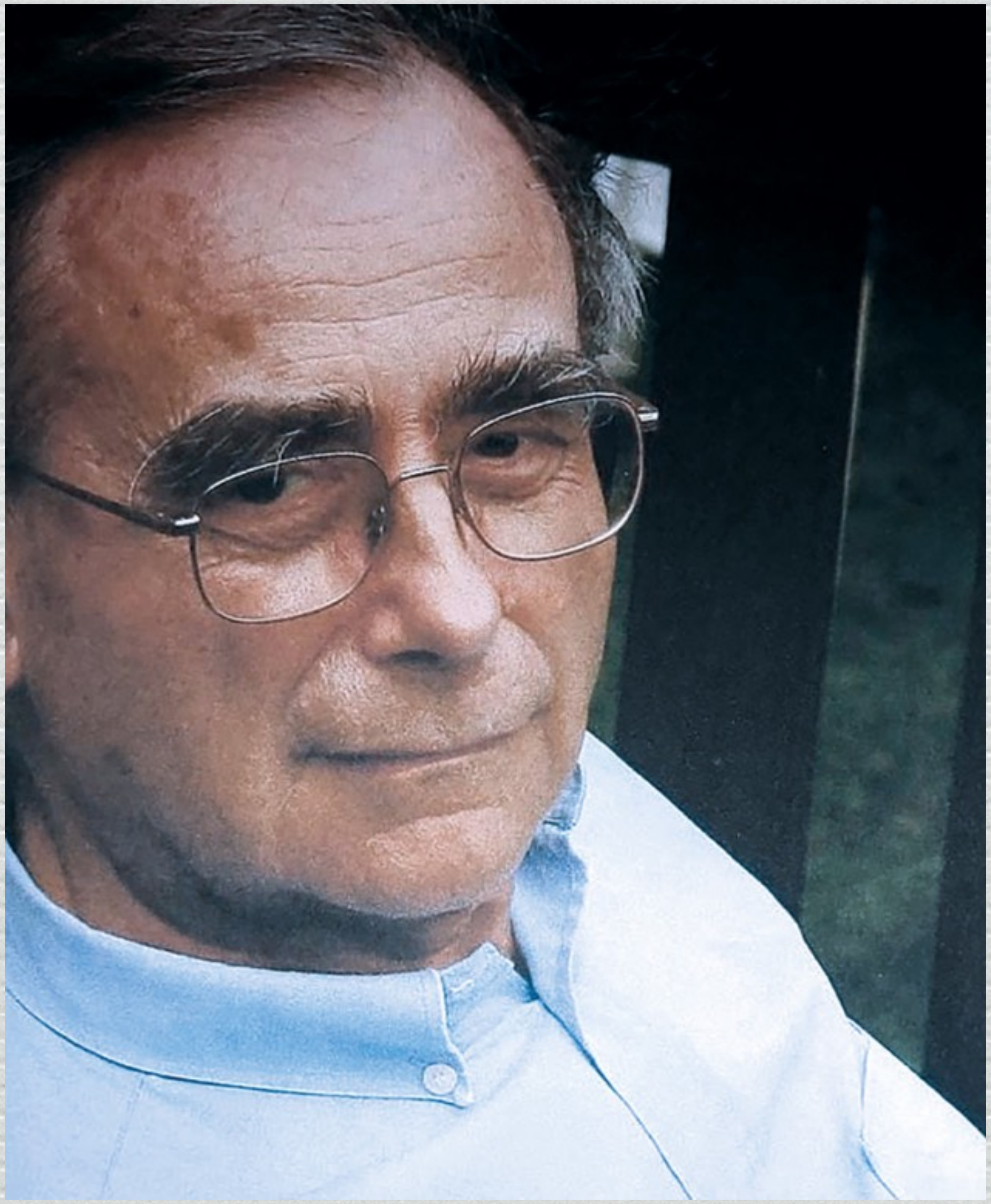
From Spin-Glass Theory to Machine Learning: An Odyssey

Marc Mézard
Bocconi University, Milan

Claude Bouchiat Memorial Conference
ENS, le 12 juillet 2023



PART I:

The spin glass revolution and its four challenges

Strongly interacting disordered systems with many components: a long-term perspective

- Maxwell, Boltzmann, etc., 150 years ago, create statistical physics

Give up deterministic description
Probabilistic approach

Strongly interacting disordered systems with many components: a long-term perspective

- Maxwell, Boltzmann, etc., 150 years ago, create statistical physics

Give up deterministic description
Probabilistic approach

- 50 years ago, creation of a new branch of statistical physics, strongly disordered systems, posing several formidable challenges

Spin glasses. Major developments in the last four decades, starting with Parisi's replica solution of the SK model in 1979.

Strongly interacting disordered systems with many components: a long-term perspective

- Maxwell, Boltzmann, etc., 150 years ago, create statistical physics

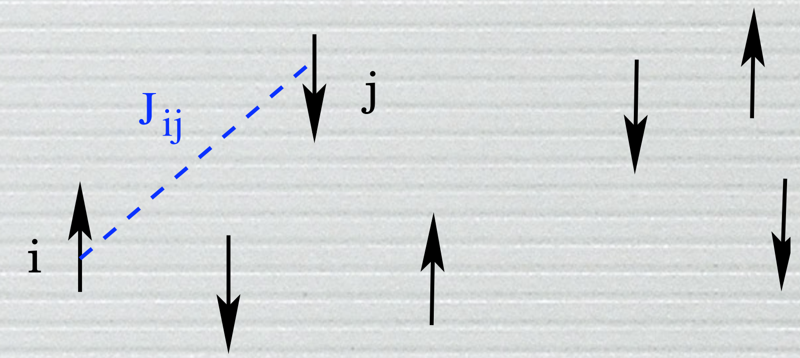
Give up deterministic description
Probabilistic approach

- 50 years ago, creation of a new branch of statistical physics, strongly disordered systems, posing several formidable challenges

Spin glasses. Major developments in the last four decades, starting with Parisi's replica solution of the SK model in 1979.

Four challenges -> new branch of statistical physics

Challenge 1: ensembles of samples



One sample of a spin glass= set of couplings J between $N \gg 1$ spins. Boltzmann probability measure on the spins $P_J(S)$

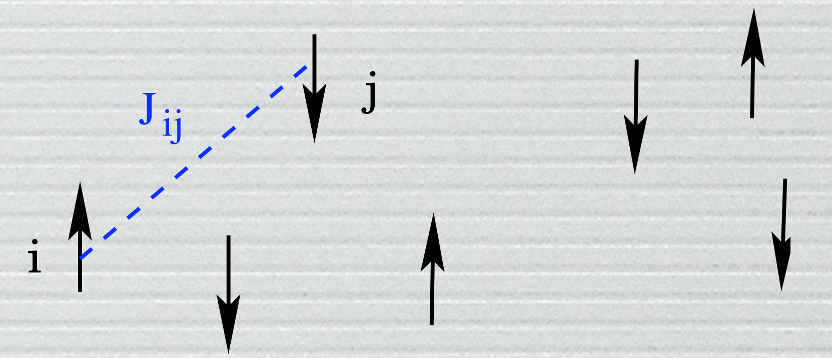
$$s_i = \pm 1$$

$$J = \{J_{ij}\}$$

$$E_J(S) = -\sum_{(i,j)} J_{ij} s_i s_j$$

$$P_J(S) = \frac{1}{Z_J} e^{-\beta E_J(S)}$$

Challenge 1: ensembles of samples



One sample of a spin glass= set of couplings J between $N \gg 1$ spins.
Boltzmann probability measure on the spins $P_J(S)$

$$s_i = \pm 1$$

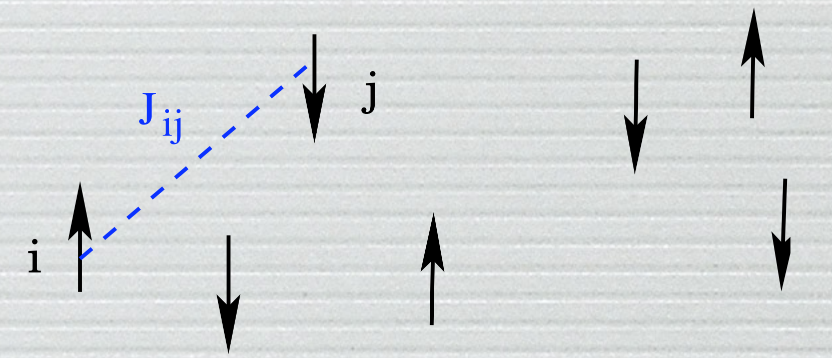
$$J = \{J_{ij}\}$$

$$E_J(S) = -\sum_{(i,j)} J_{ij} s_i s_j$$

$$P_J(S) = \frac{1}{Z_J} e^{-\beta E_J(S)}$$

Ensemble of samples = probability distribution on the set of couplings $\mathcal{P}(J)$

Challenge 1: ensembles of samples



One sample of a spin glass= set of couplings J between $N \gg 1$ spins. Boltzmann probability measure on the spins $P_J(S)$

$$s_i = \pm 1$$

$$J = \{J_{ij}\}$$

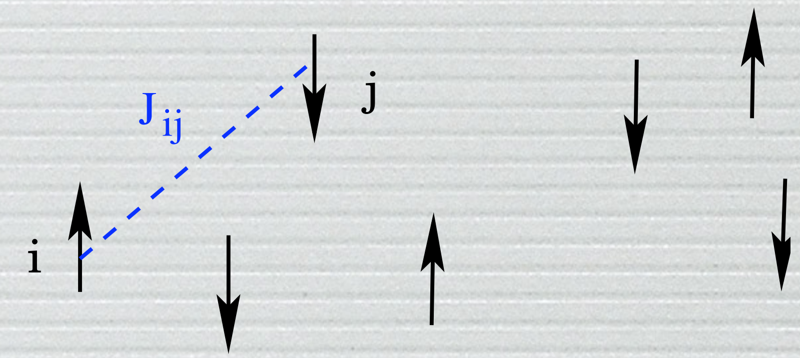
$$E_J(S) = -\sum_{(i,j)} J_{ij} s_i s_j$$

$$P_J(S) = \frac{1}{Z_J} e^{-\beta E_J(S)}$$

Ensemble of samples = probability distribution on the set of couplings $\mathcal{P}(J)$

Generate a sample with probability $\mathcal{P}(J)$
What are the properties of spin configurations sampled from $P_J(S)$?

Challenge 1: ensembles of samples



One sample of a spin glass= set of couplings J between $N \gg 1$ spins.
Boltzmann probability measure on the spins $P_J(S)$

$$s_i = \pm 1$$

$$J = \{J_{ij}\}$$

$$E_J(S) = -\sum_{(i,j)} J_{ij} s_i s_j$$

$$P_J(S) = \frac{1}{Z_J} e^{-\beta E_J(S)}$$

Ensemble of samples = probability distribution on the set of couplings $\mathcal{P}(J)$

Generate a sample with probability $\mathcal{P}(J)$
What are the properties of spin configurations sampled from $P_J(S)$?

Quenched disorder: each sample is different.
Thermal disorder: in a given sample, spins fluctuate.

Challenge 1: ensembles of samples

Disorder: each sample is different. Study sample **ensembles**. Find « self-averaging » quantities, which are identical in almost all samples. Understand differences (between samples)

eg Sherrington Kirkpatrick model

$$J_{ij} \sim \mathcal{N}(0, 1/N)$$

$$E_J(S) = -\sum_{(i,j)} J_{ij} s_i s_j$$

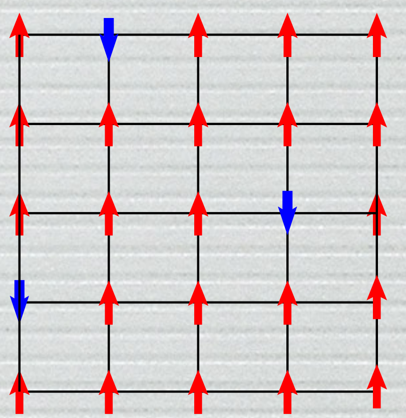
$$P_J(S) = \frac{1}{Z_J} e^{-\beta E_J(S)}$$

Self-averaging:

$$N \to \infty \qquad \frac{1}{N} \sum_i \langle s_i \rangle \qquad \frac{1}{N} \langle E_J(S) \rangle$$

Sample dependent: details of the landscape, ground state

Challenge 2: inhomogeneity



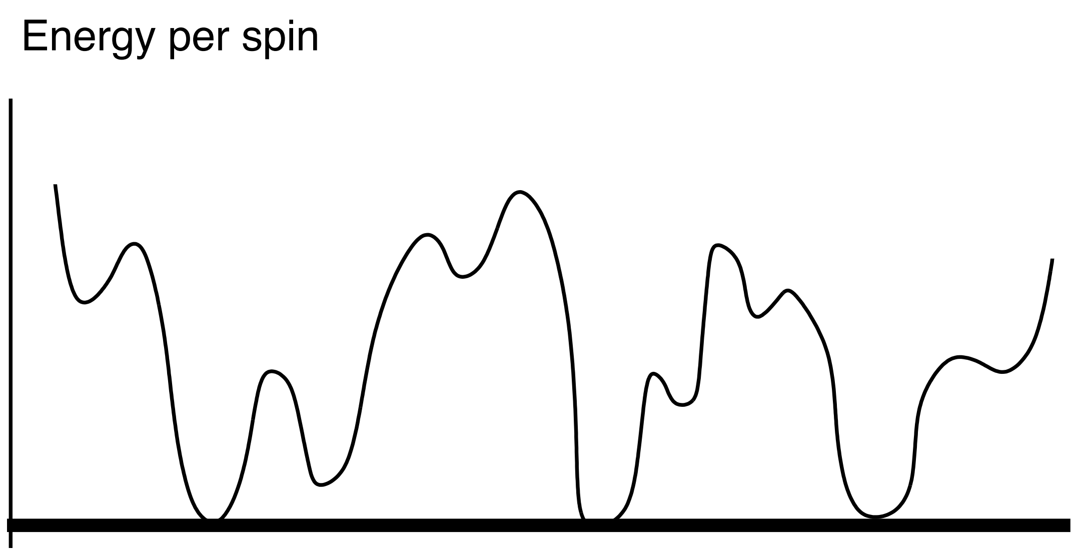
Every spin is in a different environment.
Different magnetizations.
No « representative agent ».

Mean-field equations= N coupled equations for the local magnetizations (Thouless, Anderson, Palmer 1976)
Major simplification from a probability over 2^N configurations

Statistical description of the magnetizations, the local fields: cavity method (M, Parisi, Virasoro 1986)

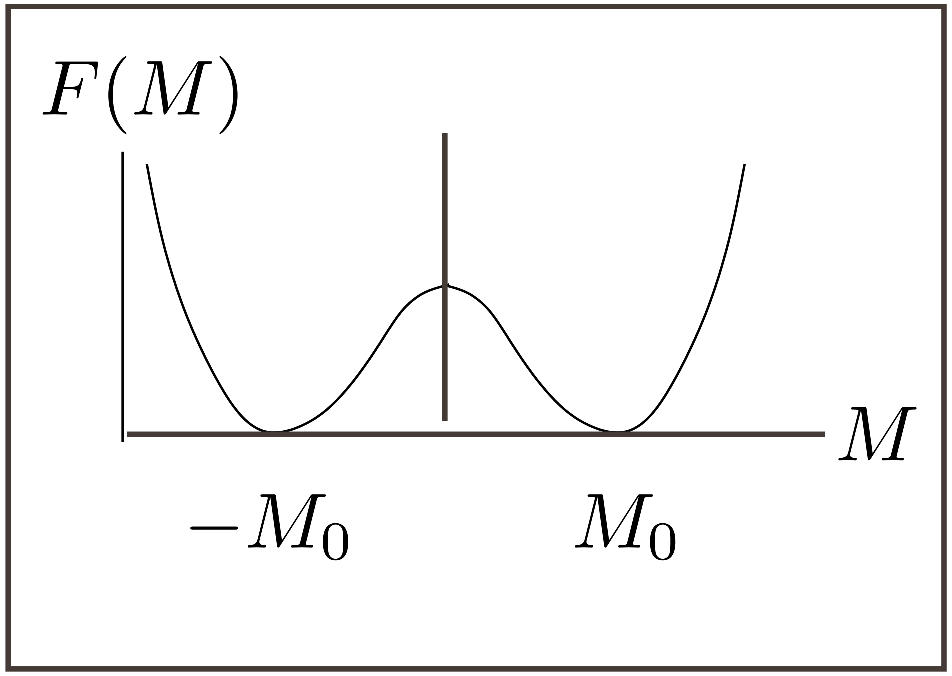
Challenge 3: rough landscape

Complicated landscape, many states in which the spin system can freeze. In SK: hierarchical (ultrametric) structure (MPSTV 85)



(sketch in a N-dimensional space)

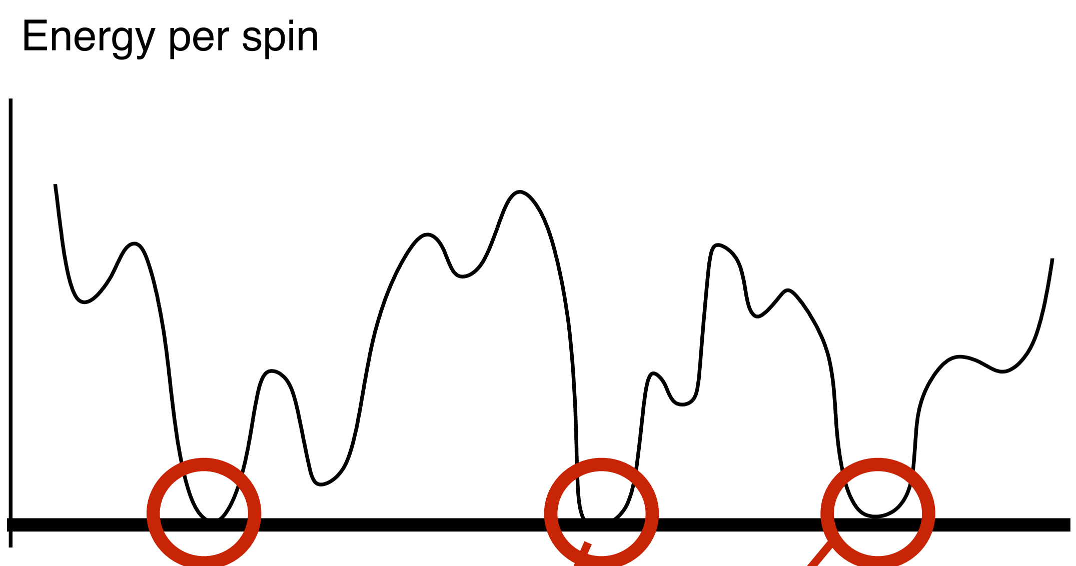
Spin glass



Ferromagnet

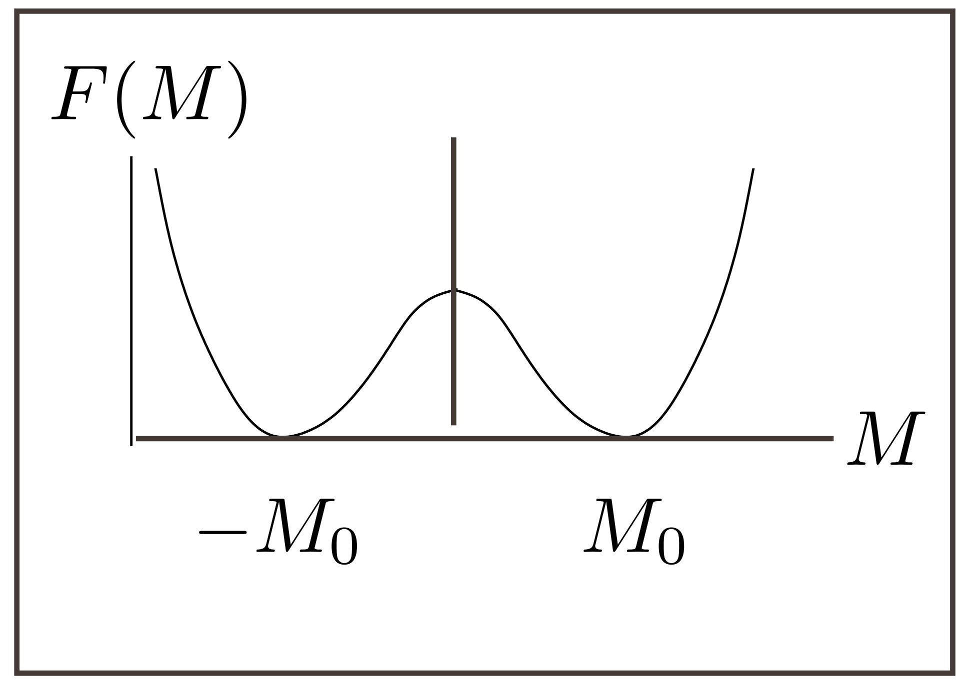
Challenge 3: rough landscape

Complicated landscape, many states in which the spin system can freeze. In SK: hierarchical (ultrametric) structure (MPSTV 85)



(sketch in a N-dimensional space)

Spin glass



Ferromagnet

Details of the landscape depend on the sample !

Landscape and **order parameters**

Ferromagnet: $$M^{\pm} = \lim_{B \to 0^{\pm}} \langle s_i \rangle_B$$

Spin glass: $$M_i^{\alpha} = \lim_{B_i \to 0^{\pm(\alpha)}} \langle s_i \rangle_B$$

Spontaneous symmetry breaking into an unknown, disordered state: unwieldy!

Landscape and **order parameters**

Ferromagnet: $M^{\pm} = \lim_{B \to 0^{\pm}} \langle s_i \rangle_B$

Spin glass: $M_i^{\alpha} = \lim_{B_i \to 0^{\pm(\alpha)}} \langle s_i \rangle_B$

Spontaneous symmetry breaking into an unknown, disordered state: unwieldy!

Use the system itself as a conjugate field: **replicas**

Landscape and **order parameters**

Ferromagnet: $M^{\pm} = \lim_{B \to 0^{\pm}} \langle s_i \rangle_B$

Spin glass: $M_i^{\alpha} = \lim_{B_i \to 0^{\pm(\alpha)}} \langle s_i \rangle_B$

Spontaneous symmetry breaking into an unknown, disordered state: unwieldy!

Use the system itself as a conjugate field: **replicas**

Overlap between two equilibrium configurations

$$q = \frac{1}{N} \sum_i s_i^1 s_i^2$$

Order parameter = Probability of overlap q :

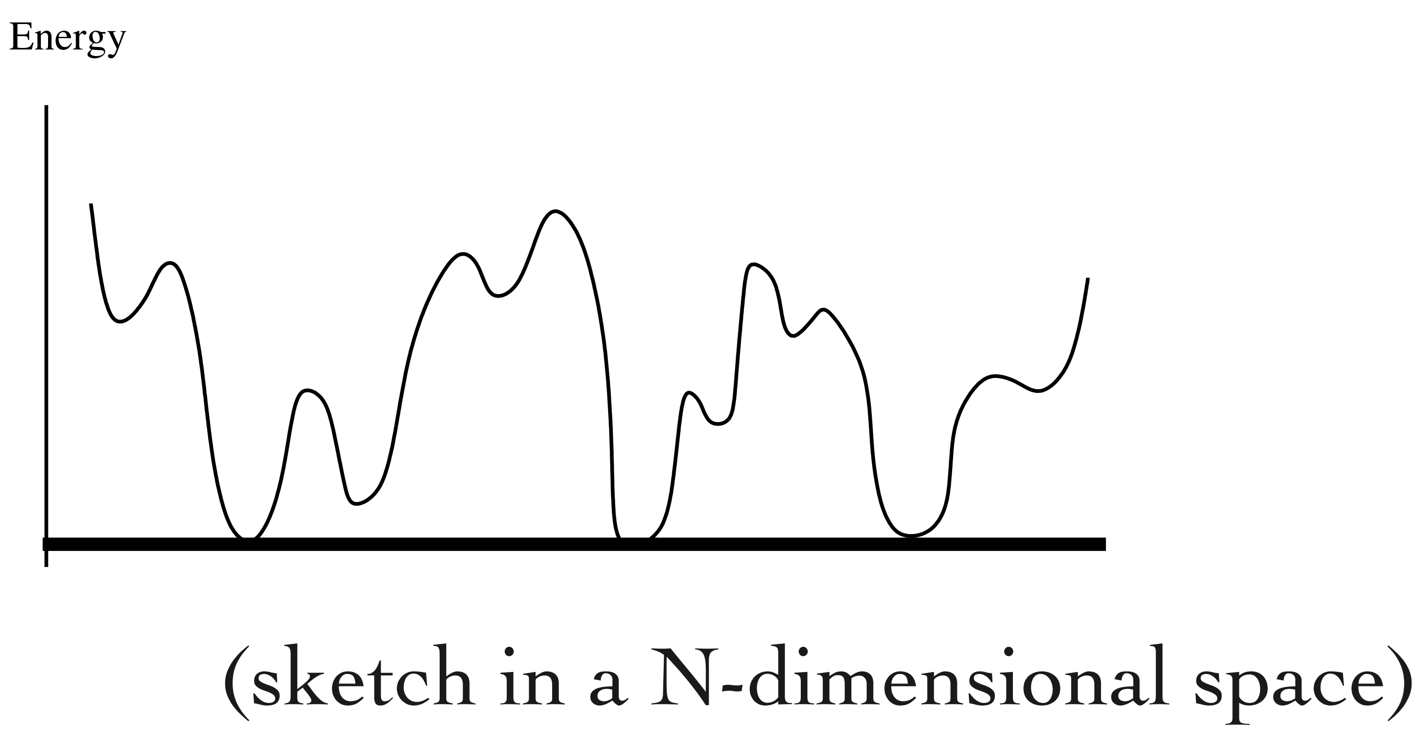
$$P_J(q)$$

Parisi 82

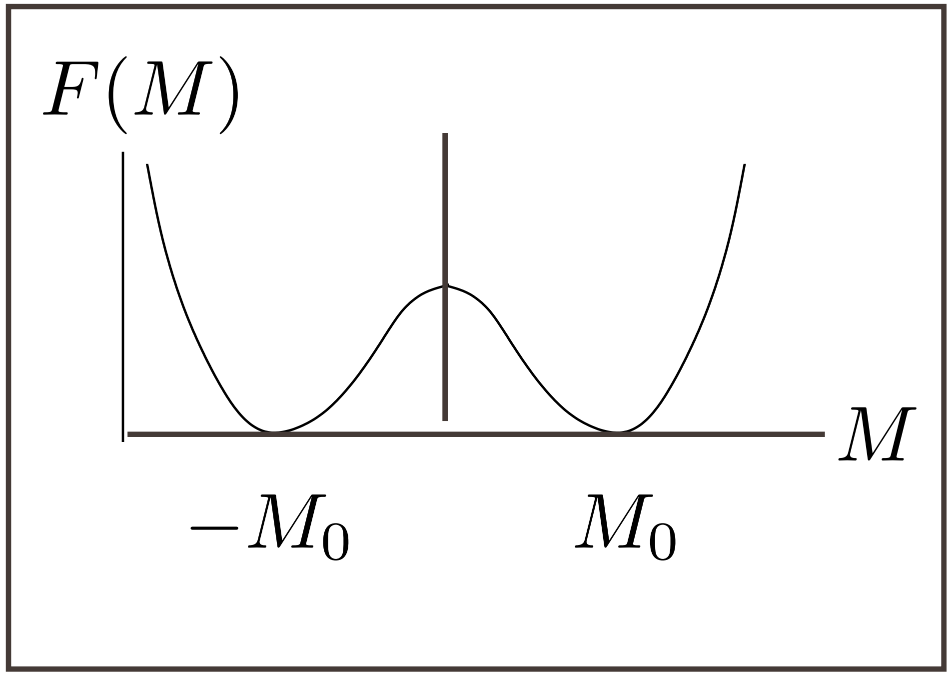
This order parameter depends on the sample: study its distribution over an ensemble of samples

Challenge 4: non equilibrium

Slow relaxation, aging. Non-equilibrium effects crucial



Spin glass



Ferromagnet

Relate equilibrium to non-equilibrium (landscape, fluctuation-dissipation relations)

A new branch of statistical physics

- Study ensembles of problems
- Each spin 'sees' a different local field

  Spins freeze in random directions
- Rough landscape: difficult to find min. of E
- Strong out of equilibrium dynamical effects

NB : beyond the simple mean field theory of the « representative agent »:
Statistics of agents. **Replicas, cavity**…

A new branch of statistical physics

- ➡ Study ensembles of problems
- ➡ Each spin 'sees' a different local field

Spins freeze in random directions

- ➡ Rough landscape: difficult to find min. of E
- ➡ Strong out of equilibrium dynamical effects

NB : beyond the simple mean field theory of the « representative agent »:
Statistics of agents. **Replicas, cavity**…

Useless, but « cornucopia »…

SK= Generic model of binary variables interacting by pairs

PART II:

Machine Learning and Large Dimensional Inference

Machine learning going deep: a decade of technological revolution

1- Image understanding.

In the last ten years, detection, segmentation and recognition of objects and regions in images. Image generation.

2- Language analysis: topic classification, question answering, language translation. Language generation

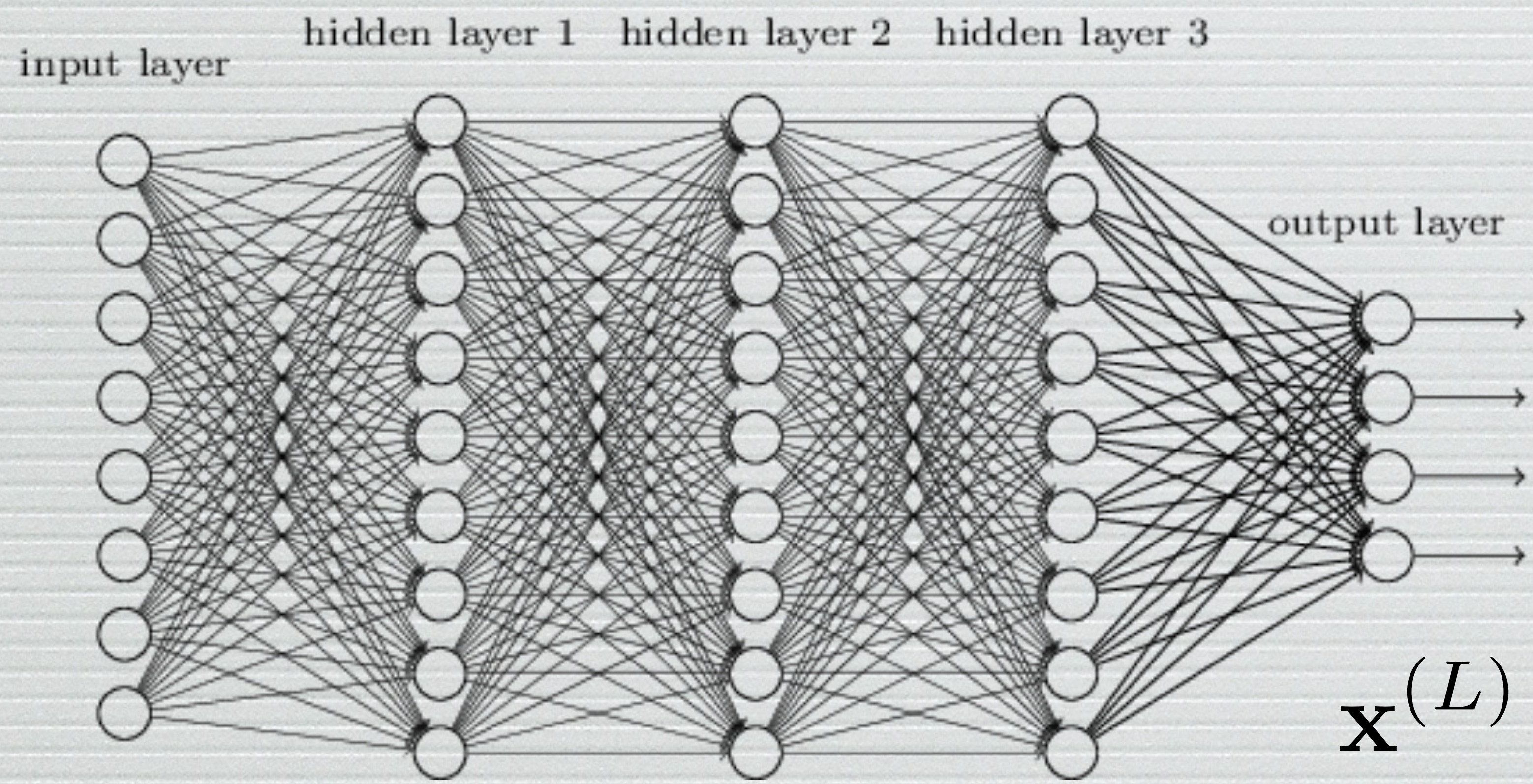
3- Science. Protein Folding. Predicting the activity of potential drug molecules. Algorithmic speedup, feature detection in data, quantum computing…

4- Playing games (chess, go, poker, video-games,…)

etc.

waiting for a general theoretical framework

The tool: Deep neural network



$\mathbf{x}^{(L)}$

$\mathbf{x}^{(1)}$ $\mathbf{x}^{(2)}$

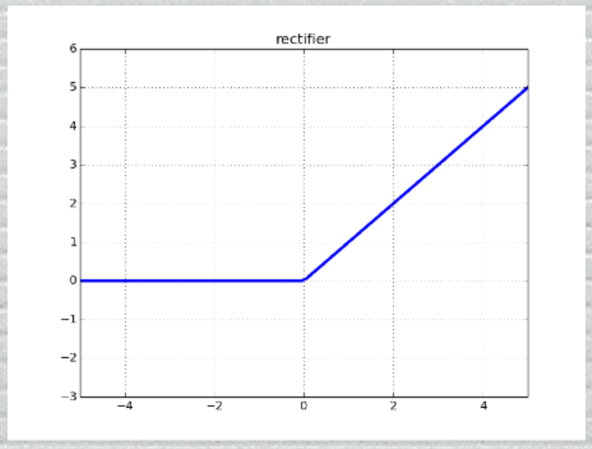
Parameters to be learnt: weights $\mathbb{W}$

$$\mathbf{x}^{(n+1)} = f\left(\mathbb{W}^{(n)}\mathbf{x}^{(n)}\right)$$

$$f = \mathrm{Sign},\ \mathrm{Relu}, \tanh\ \ldots$$

Artificial neuron

$$x_i^{(n+1)} = f\left(\sum_j \mathbb{W}_{ij}^{(n)}\ x_j^{(n)}\right)$$



NB : component-wise nonlinearity

Machine learning

$$\xi \in \mathbb{R}^N \rightarrow y = f(W, \xi) \begin{cases} \in \mathbb{R} \quad \text{or} \\ \in \{0, 1, ...q\} \end{cases}$$

Database = P examples of input-output $\quad (\xi_\mu, y_\mu) \quad \mu = 1, ..., P$

Machine learning

$$\xi \in \mathbb{R}^N \rightarrow y = f(W, \xi) \begin{cases} \in \mathbb{R} & \text{or} \\ \in \{0, 1, ...q\} \end{cases}$$

Database = P examples of input-output (ξ_μ, y_μ) $\mu = 1, ..., P$

Desired label (« supervised learning »)

Machine learning

$$\xi \in \mathbb{R}^N \rightarrow y = f(W, \xi) \begin{cases} \in \mathbb{R} & \text{or} \\ \in \{0, 1, ...q\} \end{cases}$$

Database = P examples of input-output (ξ_μ, y_μ) $\mu = 1, ..., P$

Desired label (« supervised learning »)

Learning = Optimization

Find W^* that minimizes the training error: $\sum_{\mu=1}^{P} [f(W, \xi_\mu) - y_\mu]^2$

(or other « loss function »)

Example stochastic gradient descent Very large dimensional landscape.

Machine learning

$$\xi \in \mathbb{R}^N \rightarrow y = f(W,\xi)\begin{cases}\in \mathbb{R} & \text{or} \\ \in \{0,1,...q\}\end{cases}$$

Database = P examples of input-output (ξ_μ, y_μ) $\mu = 1, ..., P$

Desired label (« supervised learning »)

Learning = Optimization

Find W^* that minimizes the training error: $$\sum_{\mu=1}^{P} [f(W,\xi_\mu) - y_\mu]^2$$

(or other « loss function »)

Example stochastic gradient descent Very large dimensional landscape.

The big Challenge: Generalization

Use the optimal W^*, test the machine on new data

Machine learning: learning phase

$$\xi \in \mathbb{R}^N \rightarrow y = f(W, \xi) \begin{cases} \in \mathbb{R} & \text{or} \\ \in \{0, 1, ...q\} \end{cases}$$

Database = P examples of input-output (ξ_μ, y_μ) $\mu = 1, ..., P$

Desired label (« supervised learning »)

Bayesian learning:

$$P(W|\{\xi_\mu, y_\mu\}) = \frac{1}{Z} P^0(W) \exp\left(-\beta \sum_\mu [f(W, \xi_\mu) - y_\mu)]^2\right)$$

Unknown Data Prior Loss

Effective inverse temperature allows to tune the importance of data wrt prior (annealing)

Machine learning: learning phase

Disordered system. Database = sample= disorder. For each database, study the properties of the probability measure on the weights

- Specific database, MNIST, CIFAR, etc
- Statistical ensemble of database. Generative models

Bayesian learning:

$$P(W|\{\xi_\mu, y_\mu\}) = \frac{1}{Z} P^0(W) \exp\left(-\beta \sum_\mu [f(W,\xi_\mu) - y_\mu)]^2\right)$$

Unknown — Data — Prior — Loss

Effective inverse temperature allows to tune the importance of data wrt prior (annealing)

The (old) ingredients

- ✴ Feedforward neural networks
- ✴ Trained with gradient descent learning, implemented with gradient back propagation

What is new in practice since the 80's ?

- ✴ Availability of very large data bases
- ✴ Much larger computing power
- ✴ Much deeper networks
- ✴ Numerous « tricks »:
  - Accumulated experience on structures (depth, width).
  - First layers = local convolutions
  - Activation functions (ReLu)
  - Stochastic gradient descent
  - Early stopping
  - Transfer learning
  - …

Surprises and questions

Surprises and questions

Training

Generalization

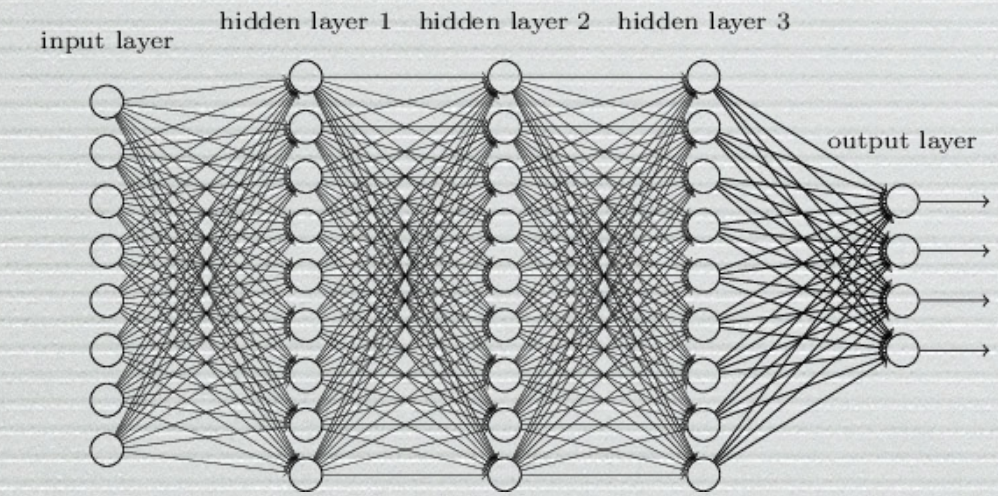
Mechanism

Surprises and questions

Training

Generalization

Mechanism



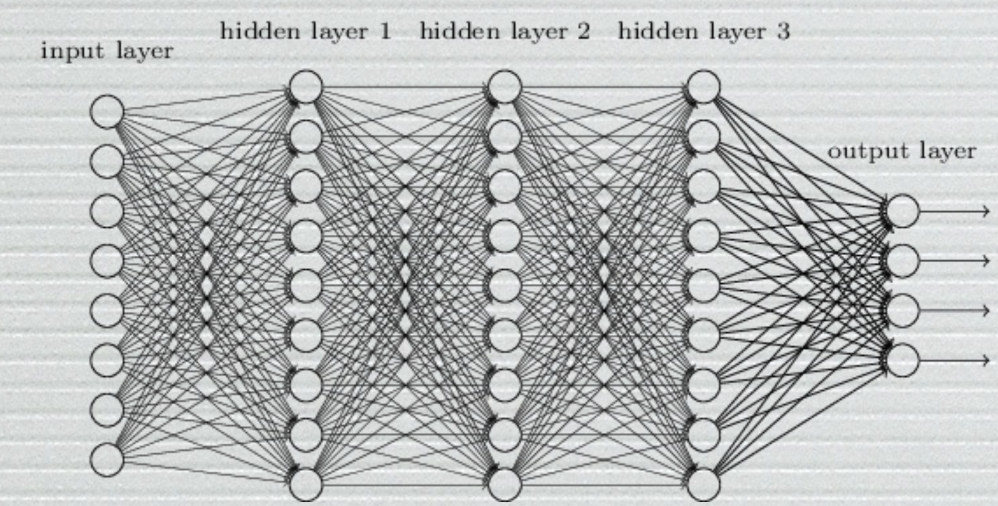
Training = optimization of a disordered system in a large dimensional space

Surprises and questions

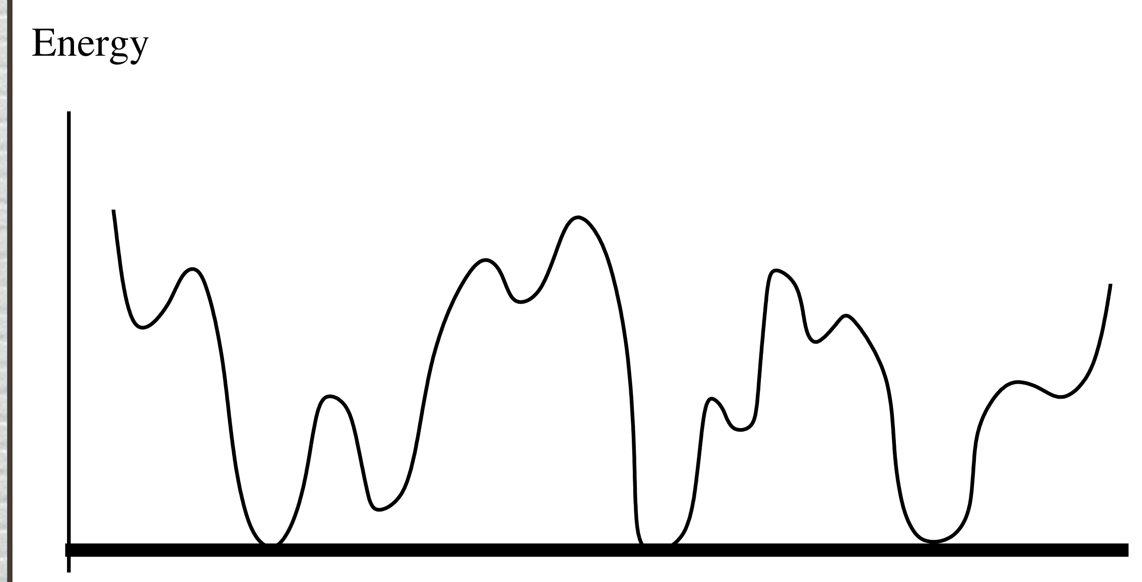
Training

Generalization

Mechanism



Training = optimization of a disordered system in a large dimensional space



(sketch in a N-dimensional spa

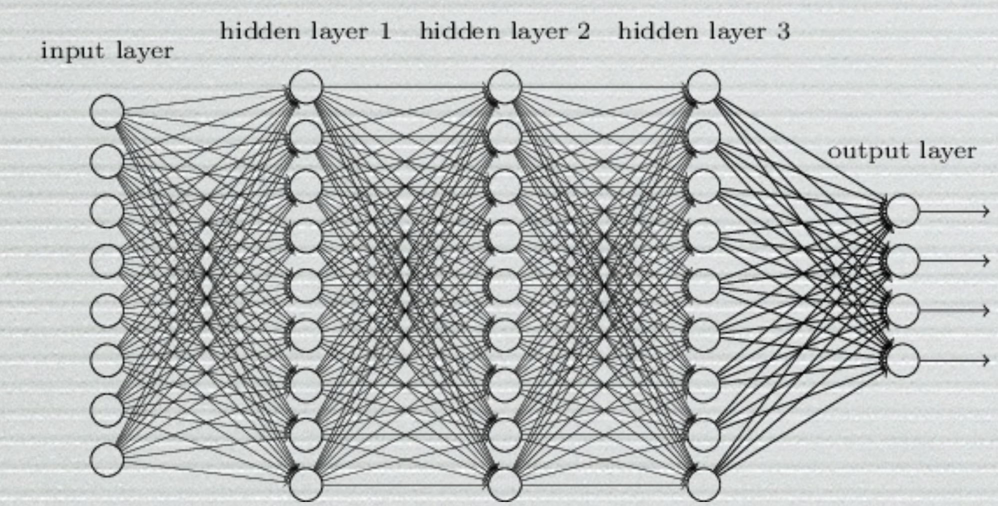
Spin glass

Surprises and questions

Training

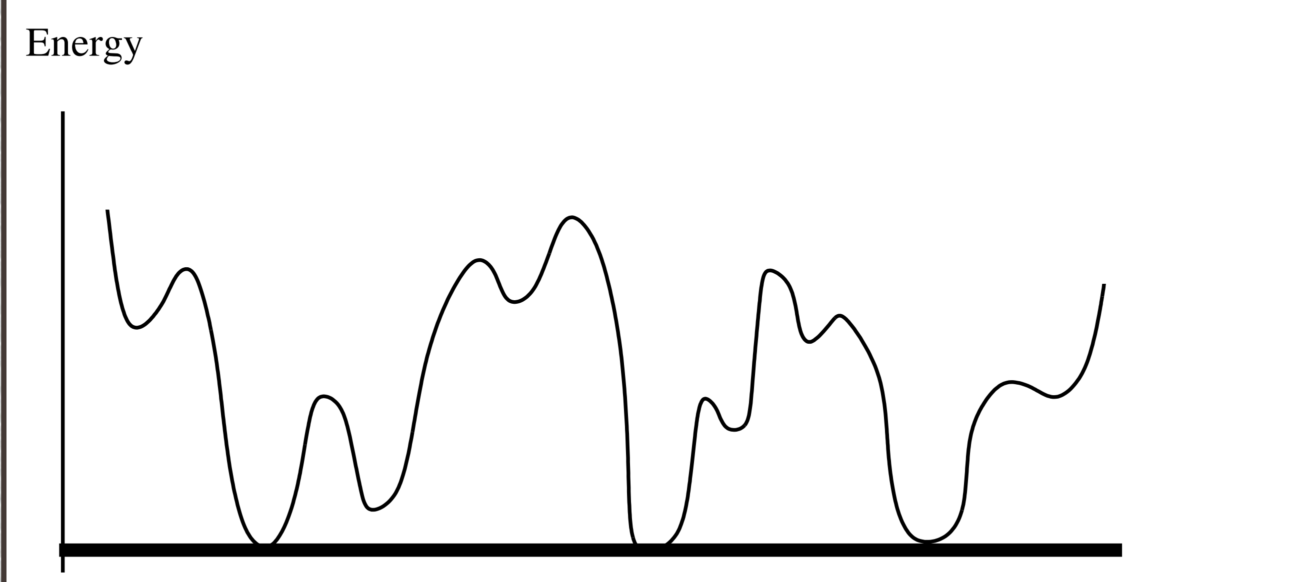
Generalization

Mechanism



Training = optimization of a disordered system in a large dimensional space

Experimentally: one can reach zero training error, using simple stochastic gradient descent, in the neighborhood of any random starting point provided the network is deep enough



(sketch in a N-dimensional spa

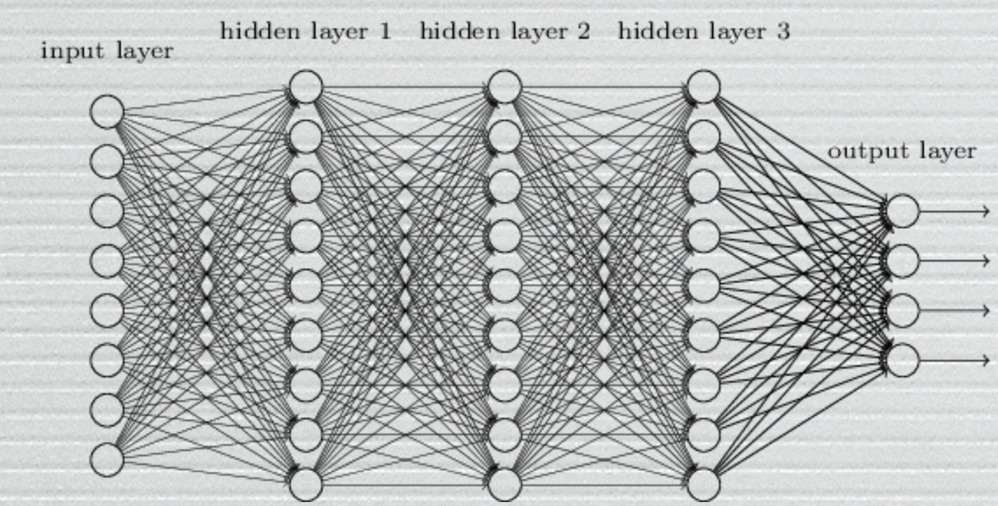
Spin glass

Surprises and questions

Training

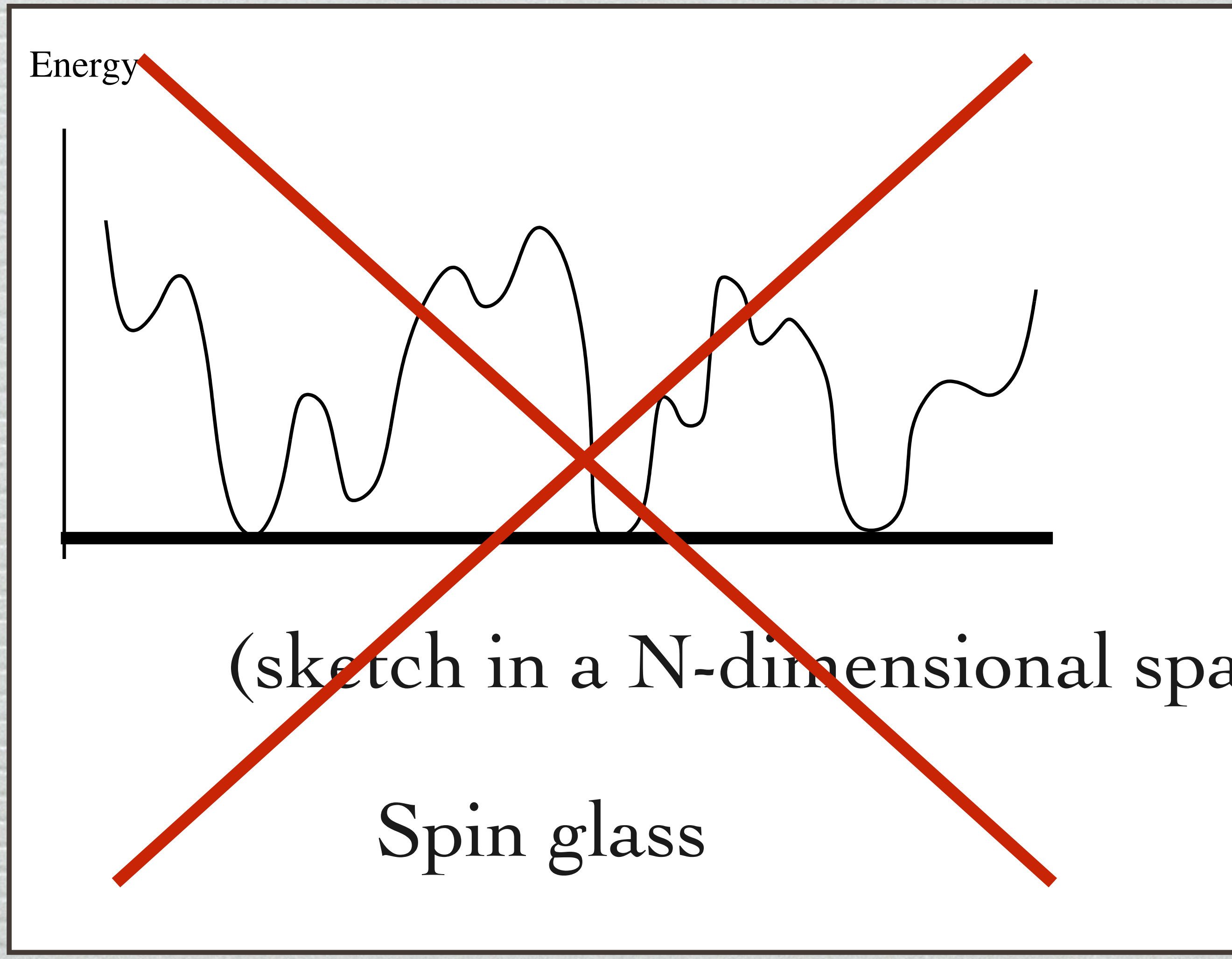
Generalization

Mechanism



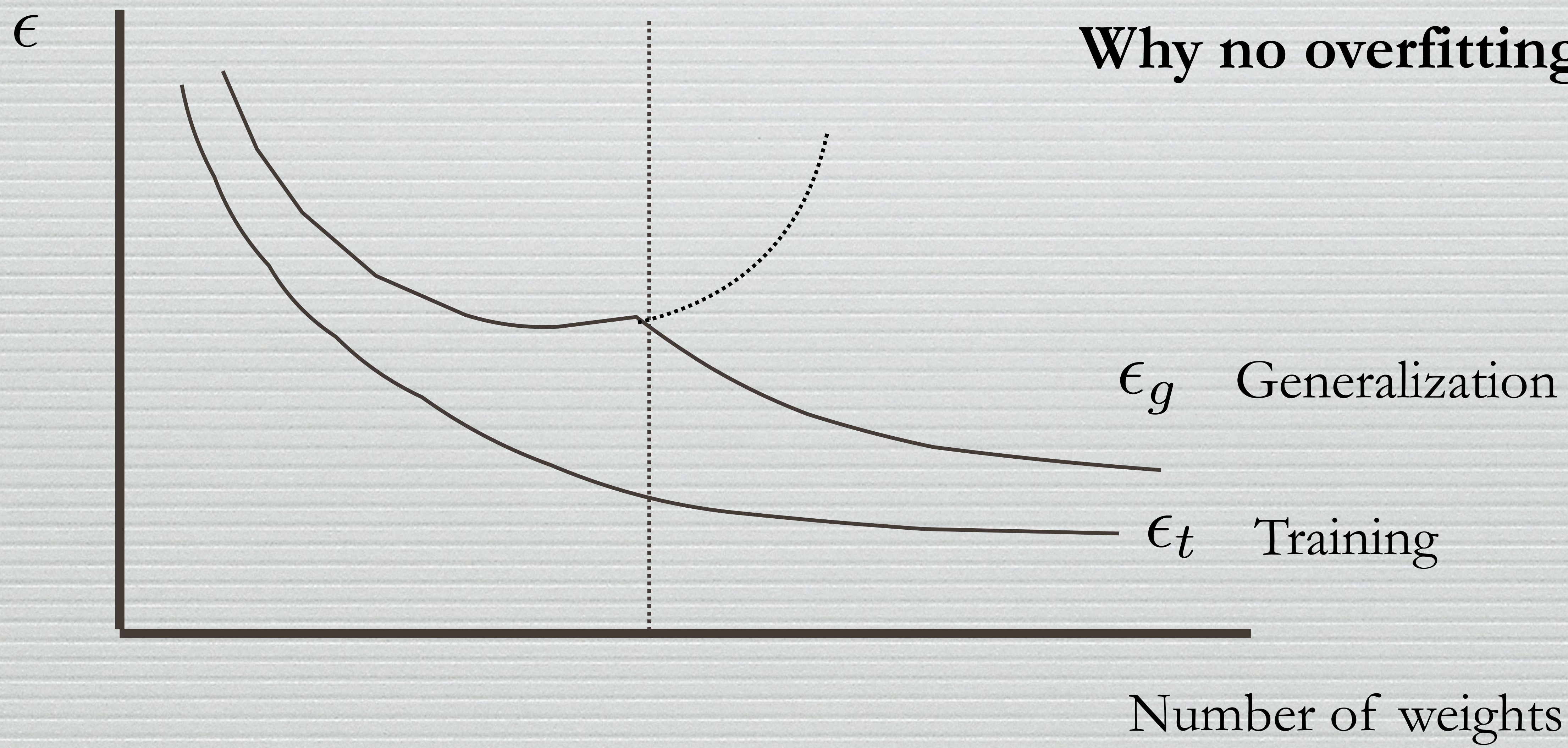
Training = optimization of a disordered system in a large dimensional space

Experimentally: one can reach zero training error, using simple stochastic gradient descent, in the neighborhood of any random starting point provided the network is deep enough

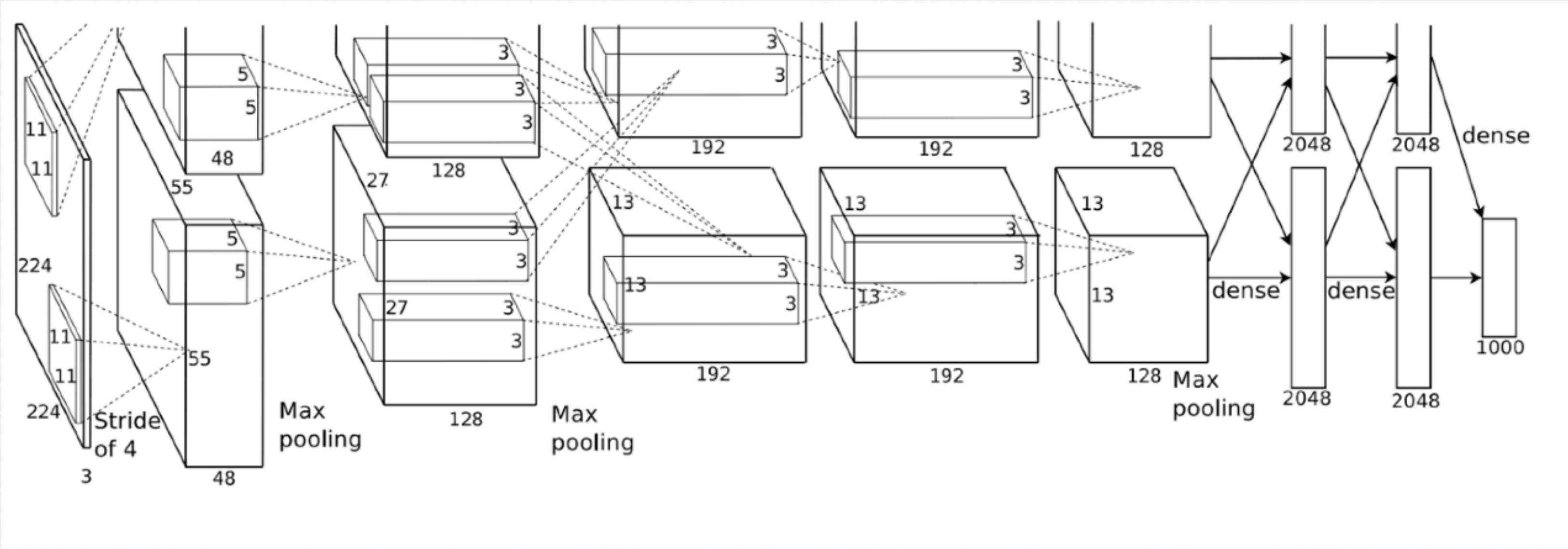


Surprises and questions

Generalization :
We train with billions of parameters.
Why no overfitting?



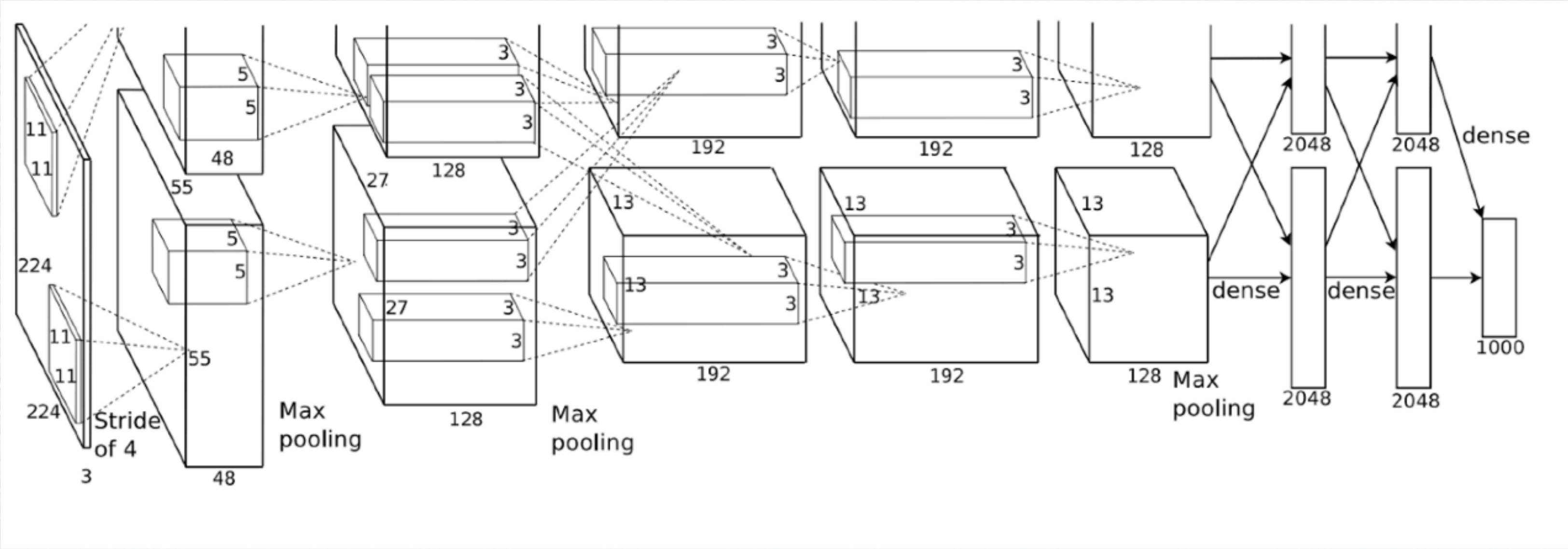
Surprises and questions



We know everything of the trained network (neuroscientist's dream)

We do not understand much. Emergent phenomenon

Surprises and questions



We know everything of the trained network (neuroscientist's dream)

We do not understand much. Emergent phenomenon

No guarantee
No explanation

Simple architecture, random iid data:
still far from deep networks

Ingredients of deep networks

Architecture

Art. Go deep, use convolutions in first layers, use pooling, etc…

Learning algorithms

Art. The (nearly) most naive algorithm, stochastic gradient descent initialized with small weights

« Simple » Data structure

Maybe the tasks that machine learning addresses are easier than expected because data has a lot more structure than our theories (worst case, or typical case with iid data) used so far

PART III:

The Challenge of Data Structure

Combinatorial
Hierarchical
Semantic
Low-dimensional Manifold

Combinatorial

Hierarchical

New York Times, July 10, 2023

President Vladimir Putin of Russia held a three-hour meeting with Yevgeny Prigozhin and his top Wagner commanders on June 29, according to the Kremlin.

Semantic: eg Large Language Models

President Vladimir Putin of Russia held a three-hour meeting with Yevgeny Prigozhin and his top Wagner commanders on June 29, according to the

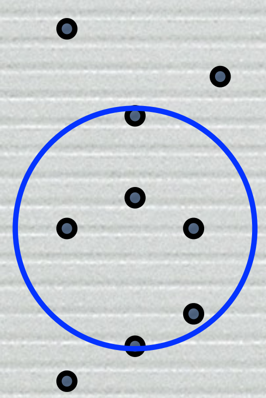
Attention Mechanism

Hidden manifold example: MNIST

Input space: dimension

$$28^2 = 784$$

Manifold of handwritten digits in MNIST:

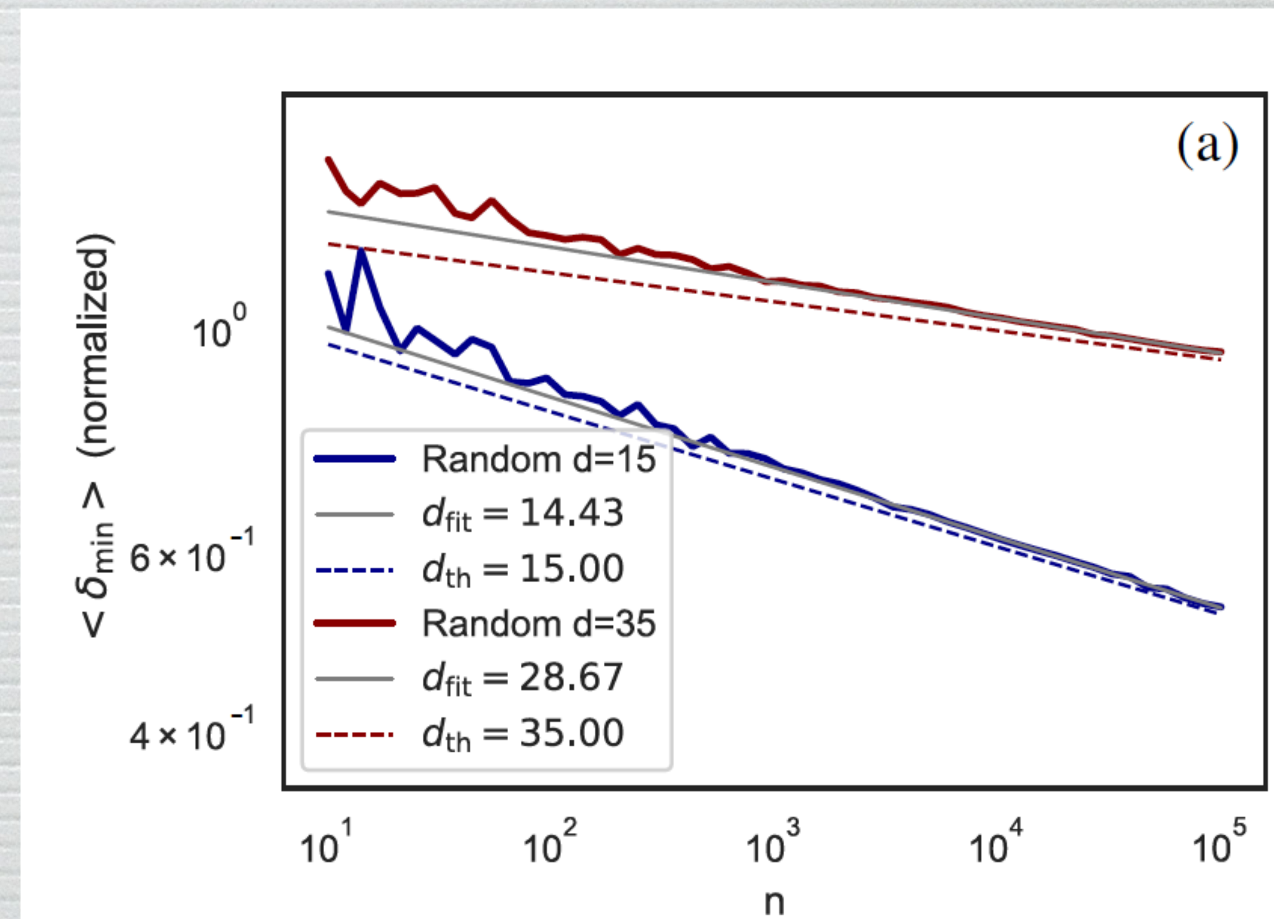


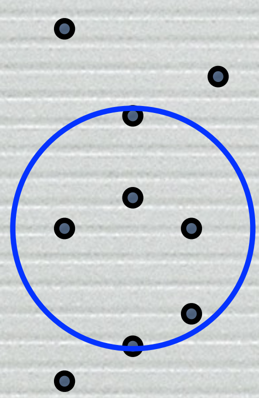
$$p \simeq cR^d$$

Nearest neighbors' distance : $R_{nn} \simeq p^{-1/d}$

Grassberger Procaccia 83, Costa Hero 05, Heinz Audibert 05, Facco et al. 17, Ansuini et al. 19, Spigler et al. 19…

Hidden manifold example: MNIST

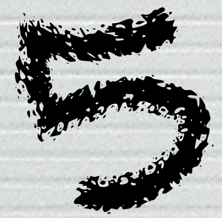




Nearest neighbors' distance : $R_{nn} \simeq p^{-1/d}$

$p \simeq cR^d$

$d_{\text{eff}} \simeq 15$



$d_{\text{eff}}(5) \simeq 12$

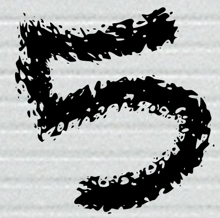
Table 7. Number of samples and estimated intrinsic dimensionality of the digits in MNIST.

1	2	3	4	5
7877	6990	7141	6824	6903
8/7/7	13/12/13	14/13/13	13/12/12	12/12/12

6	7	8	9	0
6876	7293	6825	6958	6903
11/11/11	10/10/10	14/13/13	12/11/11	12/11/11

Hein Audibert 05

MNIST problem: in the **15-dim manifold** of handwritten digits, identify the **10 perceptual sub manifolds** associated with each digit, of **dimensions between 7 and 13**…



$d_{\mathrm{eff}}(5) \simeq 12$

Table 7. Number of samples and estimated intrinsic dimensionality of the digits in MNIST.

1	2	3	4	5
7877	6990	7141	6824	6903
8/7/7	13/12/13	14/13/13	13/12/12	12/12/12
6	**7**	**8**	**9**	**0**
6876	7293	6825	6958	6903
11/11/11	10/10/10	14/13/13	12/11/11	12/11/11

Hein Audibert 05

MNIST problem: in the **15-dim manifold** of handwritten digits, identify the **10 perceptual sub manifolds** associated with each digit, of **dimensions between 7 and 13**…

… from an input in 784 dimensions!

Ensembles of structured data

- Combinatorial patterns in a Hidden Manifold

S. Goldt, F. Krzakala, MM, L. Zdeborova

S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, MM, L. Zdeborova

Latent space
iid variables

Data

$D, N \to \infty$

$C \in \mathbb{R}^D$ ⟶ $\xi \in \mathbb{R}^N$

$D/N < 1$

In a hidden D-dimensional manifold

$$\xi_i = f\left(\frac{1}{\sqrt{D}}\sum_{r=1}^{D} C_r F_{ir}\right)$$

Ensembles of structured data

- Combinatorial patterns in a Hidden Manifold

S. Goldt, F. Krzakala, MM, L. Zdeborova

S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, MM, L. Zdeborova

Latent space
iid variables

Data

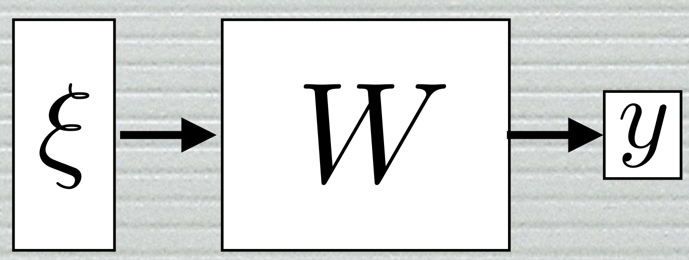
$D, N \to \infty$

$C \in \mathbb{R}^D \longrightarrow \xi \in \mathbb{R}^N$

$D/N < 1$

In a hidden D-dimensional manifold

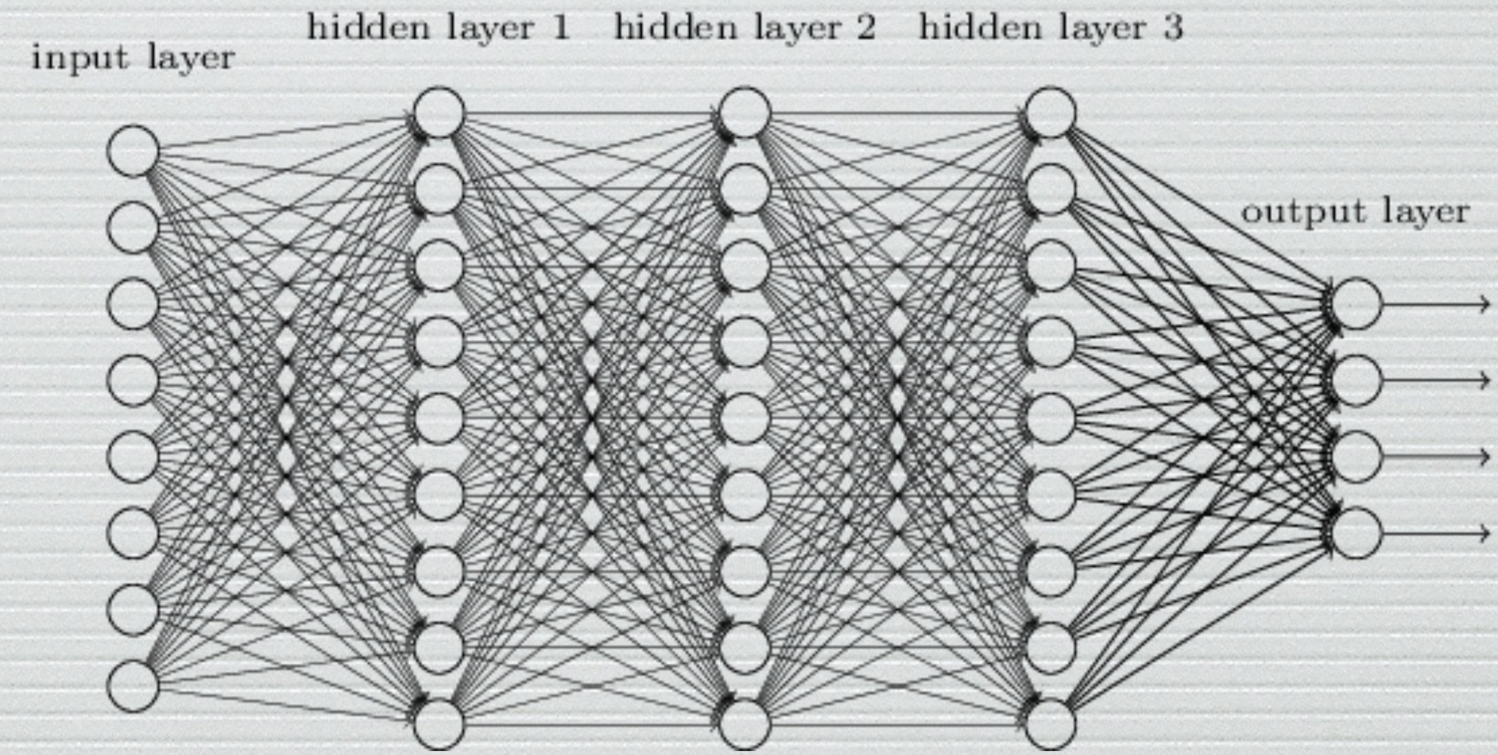
$$\xi_i = f\left(\frac{1}{\sqrt{D}} \sum_{r=1}^{D} C_r F_{ir}\right)$$



Deep generator: Generative Adversarial Network

Latent space
iid variables

$$C \in \mathbb{R}^D$$



$$\xi$$

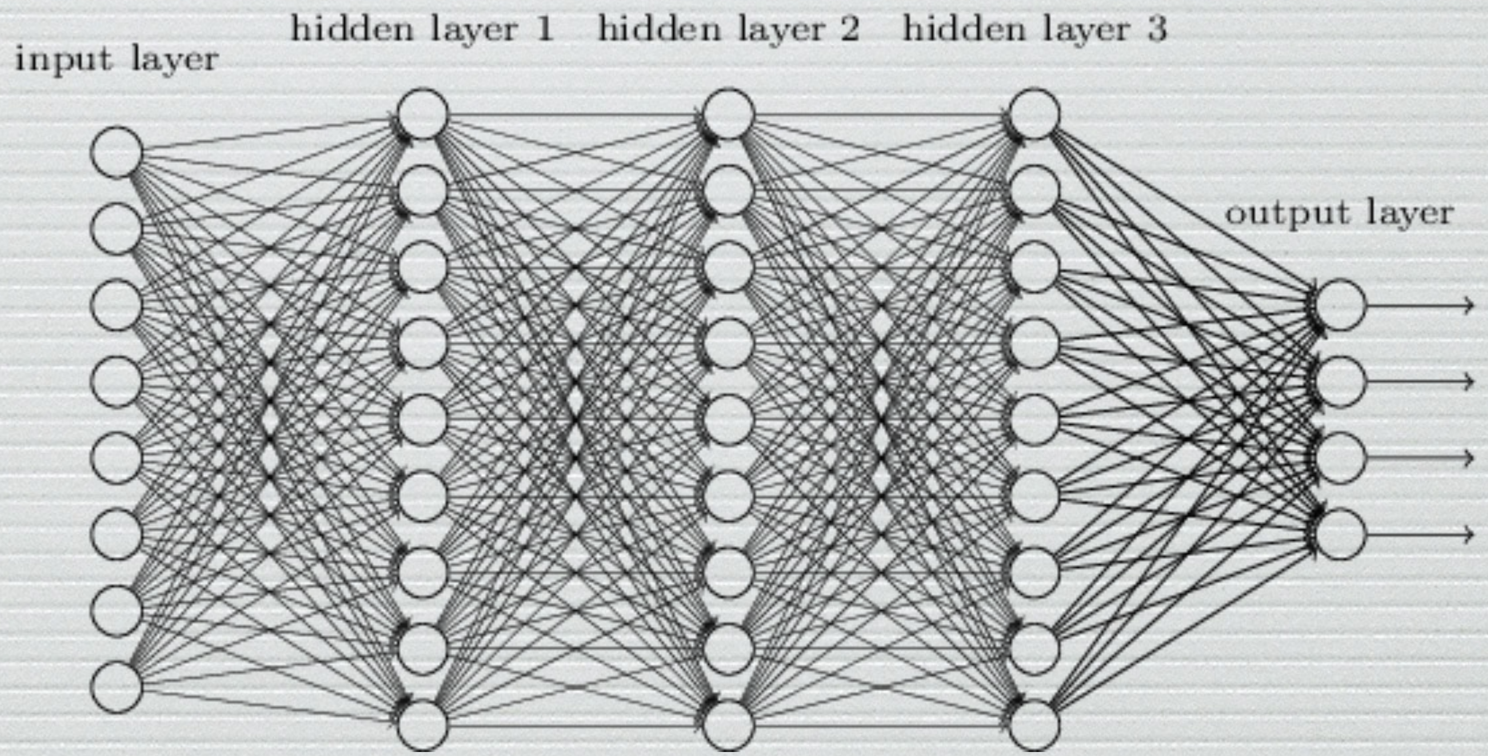
$$\xi = f\left[F^L f\left(F^{L-1}(....f(F^1 c))\right)\right]$$

Natural generalization: generate hidden manifold data through L layers of a generator

Deep generator: Generative Adversarial Network

Latent space
iid variables

$$C \in \mathbb{R}^D$$



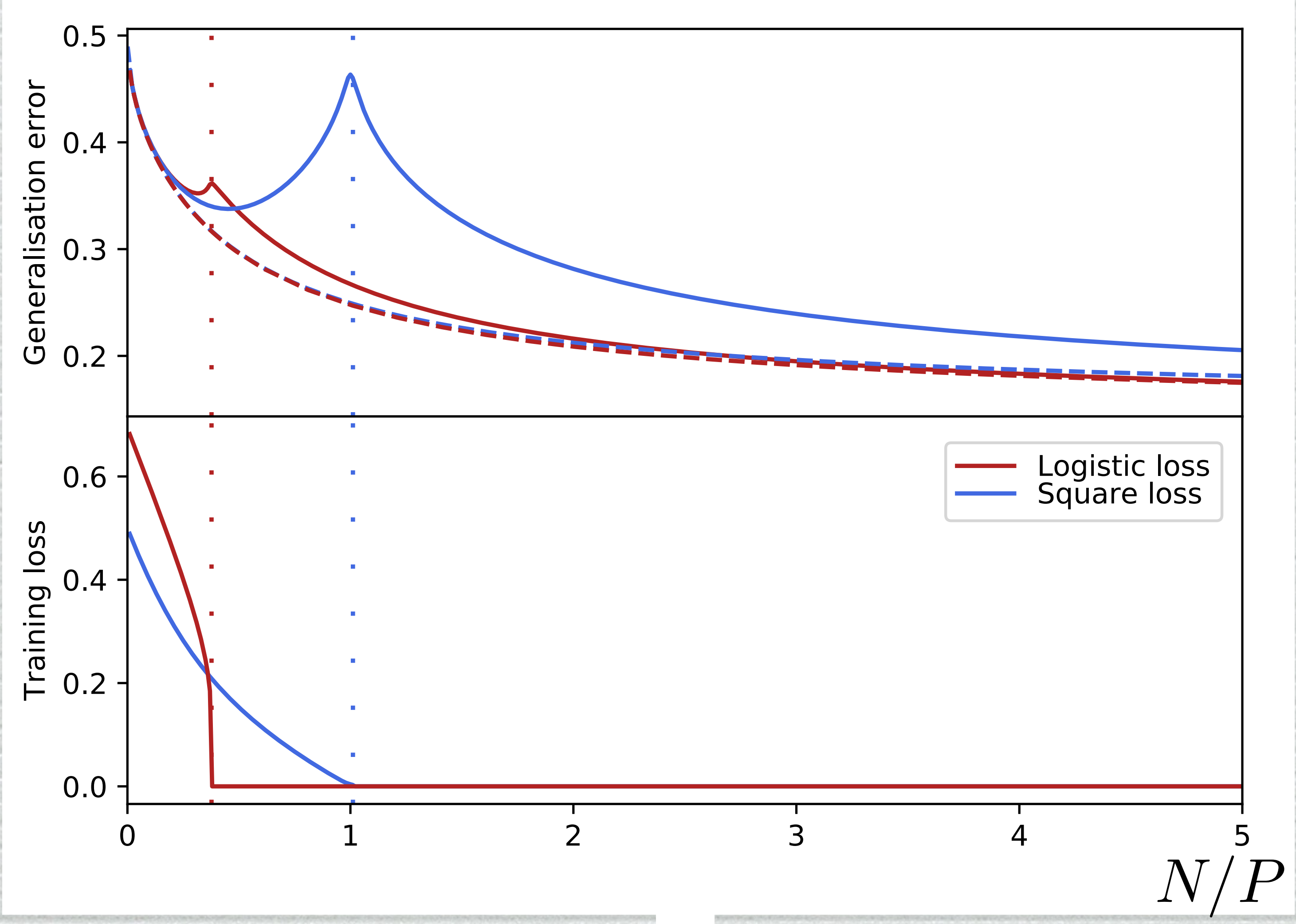
ξ

$$\xi = f\left[F^L f\left(F^{L-1}(\ldots.f(F^1 c))\right)\right]$$

Natural generalization: gene
manifold data through L lay

Example: Replica study of perceptron/regression with Hidden Manifold Model

F. Gerace, B. Loureiro, F. Krzakala, MM, L. Zdeborova



$$D/N = 1/3$$

$$\lambda = 10^{-4}$$

« Double descent »

Spigler et al. 2019
Belkin et al. 2019

Summary

Machine learning needs a general theory. One knows everything but understands little. Emergence.

Statistical physics can contribute to this theory, but it faces a new challenge : data structure

Idea: Use synthetic data ensembles with structure:

Input in low-dimensional Manifold
Combinatorial structure
Semantic/attention
Invariances of task with respect to some transformations (perceptual manifolds)

Summary

Machine learning needs a general theory. One knows everything but understands little. Emergence.

Statistical physics can contribute to this theory, but it faces a new challenge : data structure

Idea: Use synthetic data ensembles with structure:
Input in low-dimensional Manifold
Combinatorial structure
Semantic/attention
Invariances of task with respect to some transformations (perceptual manifolds)

Questions: Important properties of experimental data sets?
Generic? Relation to generative models.
New tools- or applicability of old tools