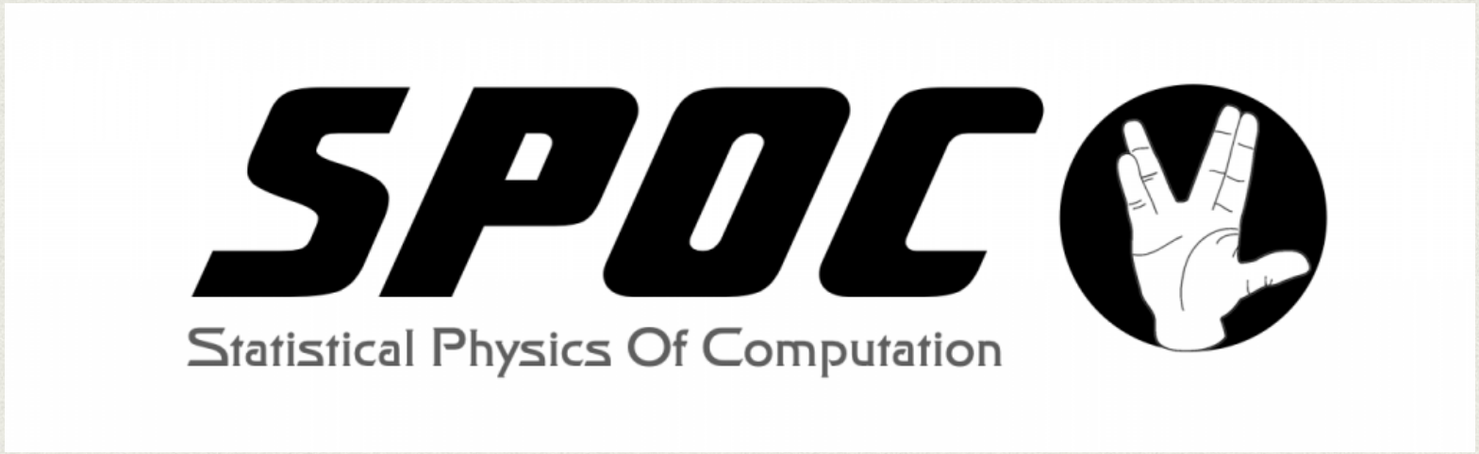
EPFL

STATISTICAL PHYSICS OF COMPUTATION



Lenka Zdeborová

Lecture Series of Scientific Machine Learning, Sep 19th, 2024

PHYSICS & ML

- ML is largely engineering success. But it remains a black box - not clear when one can make it work, what is the reliability/ uncertainty, how much data is required, etc.

- To fully develop scientific machine learning we may need better understanding and control of ML methods.

- Number of early concepts in machine learning come from physics (Boltzmann machine, Gibbs sampling, etc.)

- Physics is the main tool we have so far to study such complex systems.

comment

Check for updates

nature physics

Understanding deep learning is also a job for physicists

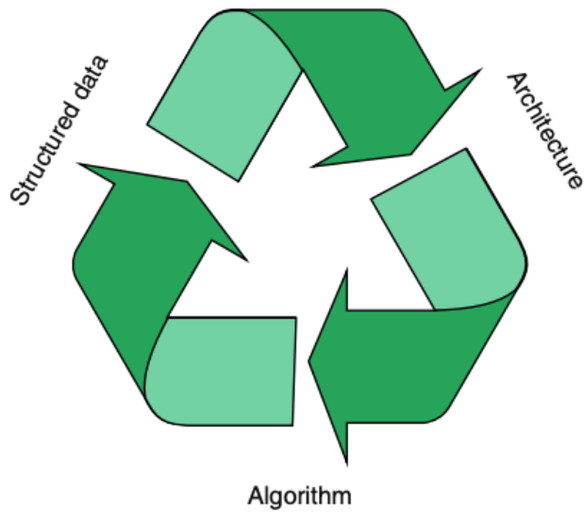
Automated learning from data by means of deep neural networks is finding use in an ever-increasing number of applications, yet key theoretical questions about how it works remain unanswered. A physics-based approach may help to bridge this gap.

Lenka Zdeborová

Imagine an event for which thousands of tickets get sold out in under 12 minutes. We are not speaking of a leading show on Broadway or a concert of a rockstar, but about the Conference on Neural Information Processing Systems (NeurIPS) — the principal gathering for research in machine learning and artificial intelligence. The fields related to automated learning from data are experiencing a surge in research activity, as well as in investment. This is largely thanks to developments in a subfield called deep learning, which has led to a physicists it is a matter of sitting tight waiting for tools and answers that we can subsequently put to use. In this Comment, I argue that, instead, we need to join the race of searching for these answers, because it is precisely the physicists' perspective and approach that is needed to enable progress in this endeavour.

Three ingredients to decipher deep learning

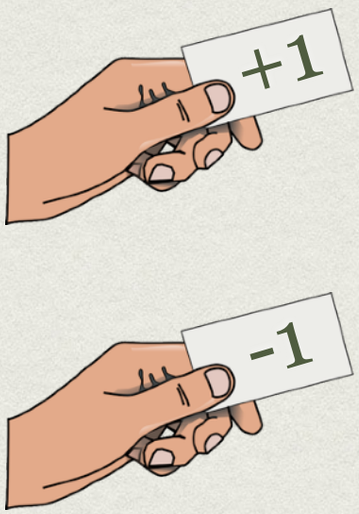
The engineering details of current deep-learning systems, such as the ones



Physics Nobel Prize 2021 to Giorgio Parisi

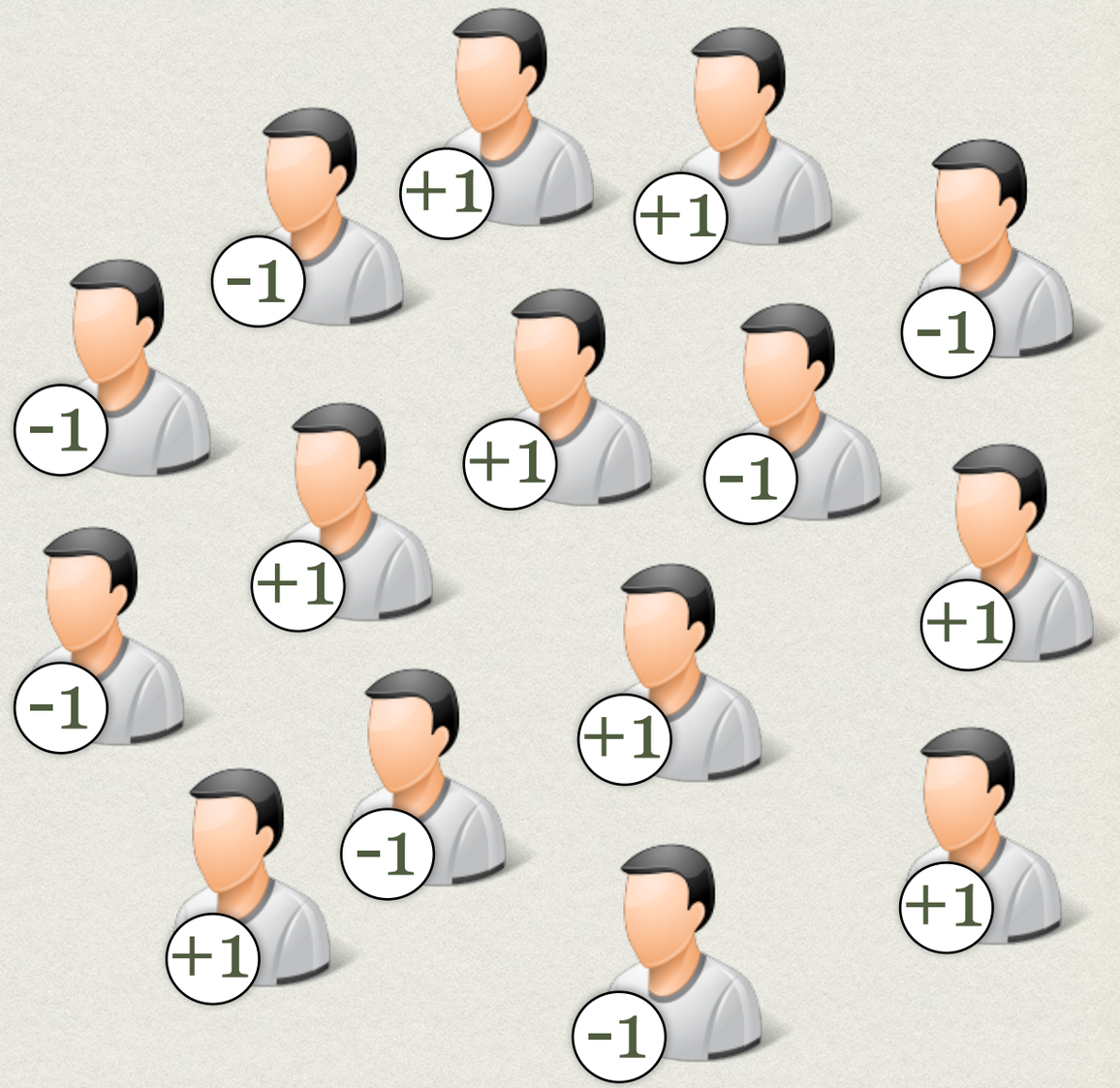
“Parisi's discoveries make it possible to understand and describe many different and apparently entirely **random complex materials and phenomena**, not only in physics but also in other, very different areas, such as mathematics, biology, neuroscience and **machine learning**.”

LET'S PLAY A GAME

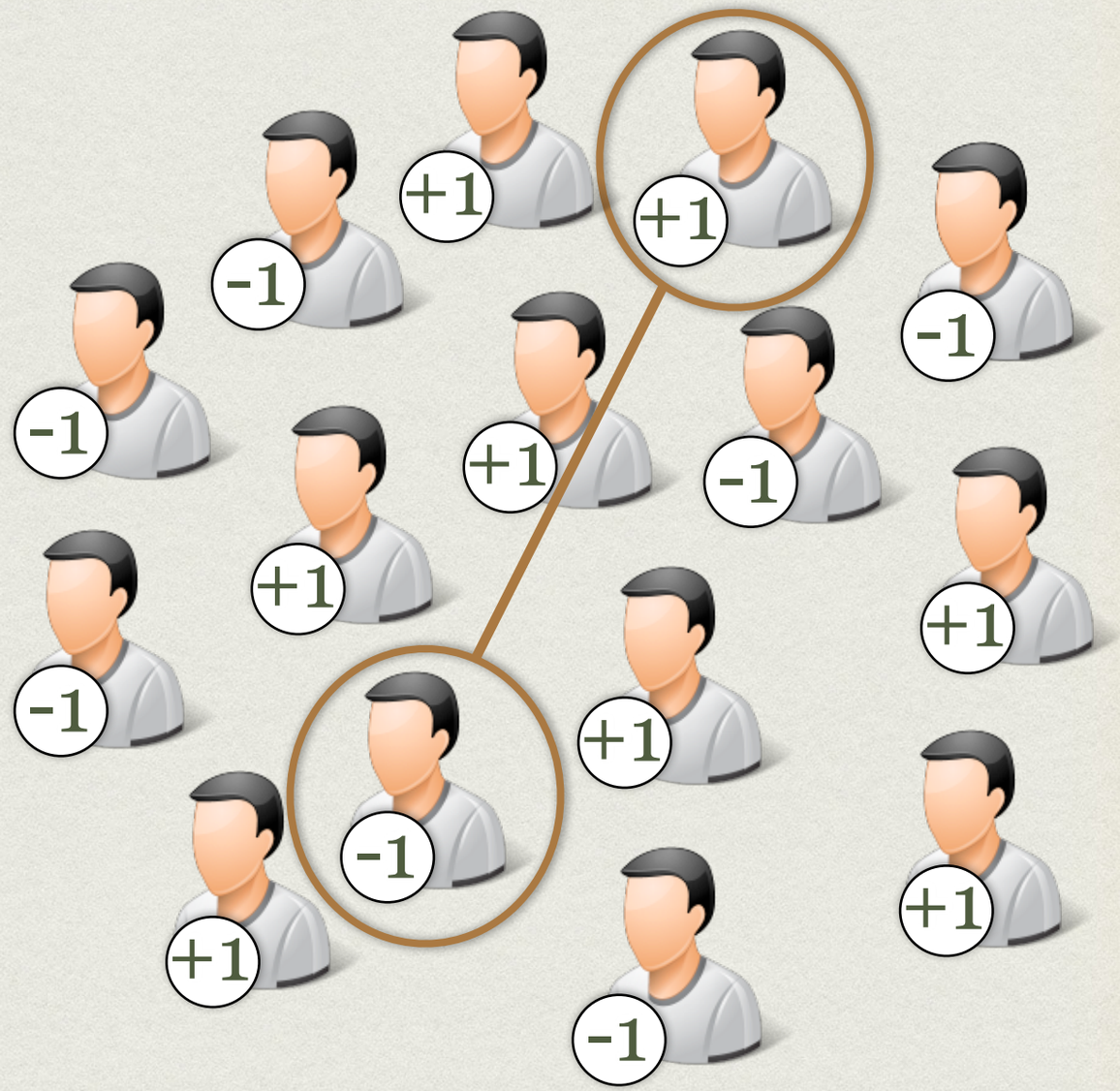


N=15 people

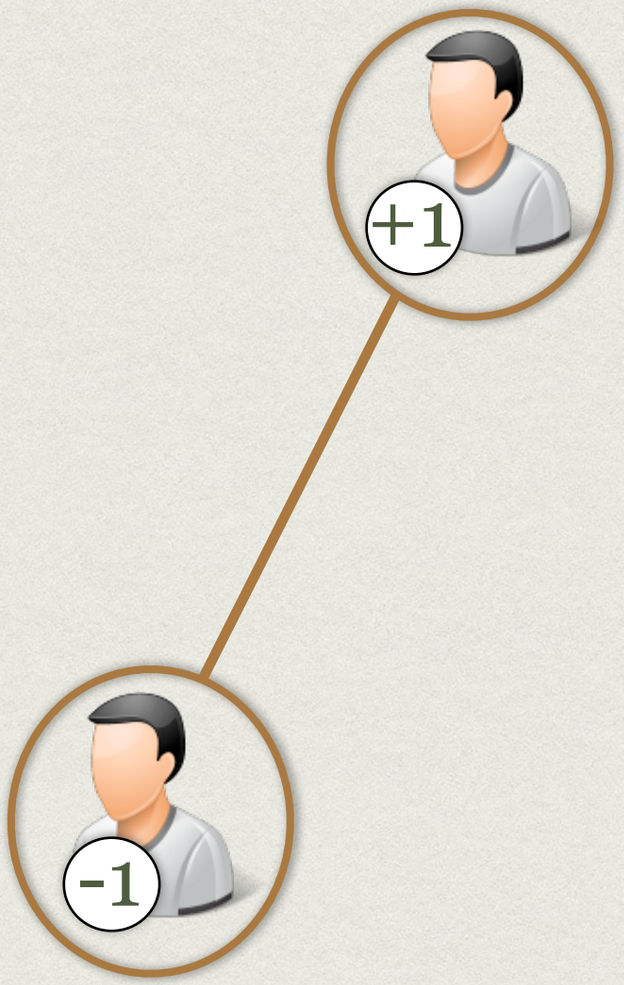
LET'S PLAY A GAME



LET'S PLAY A GAME



LET'S PLAY A GAME



- Generate a random Gaussian variable W (zero mean and variance Δ^*)
- Report:
  - $Y=W+1/\sqrt{N}$ if the cards were the same.
  - $Y=W-1/\sqrt{N}$ if the cards were different.

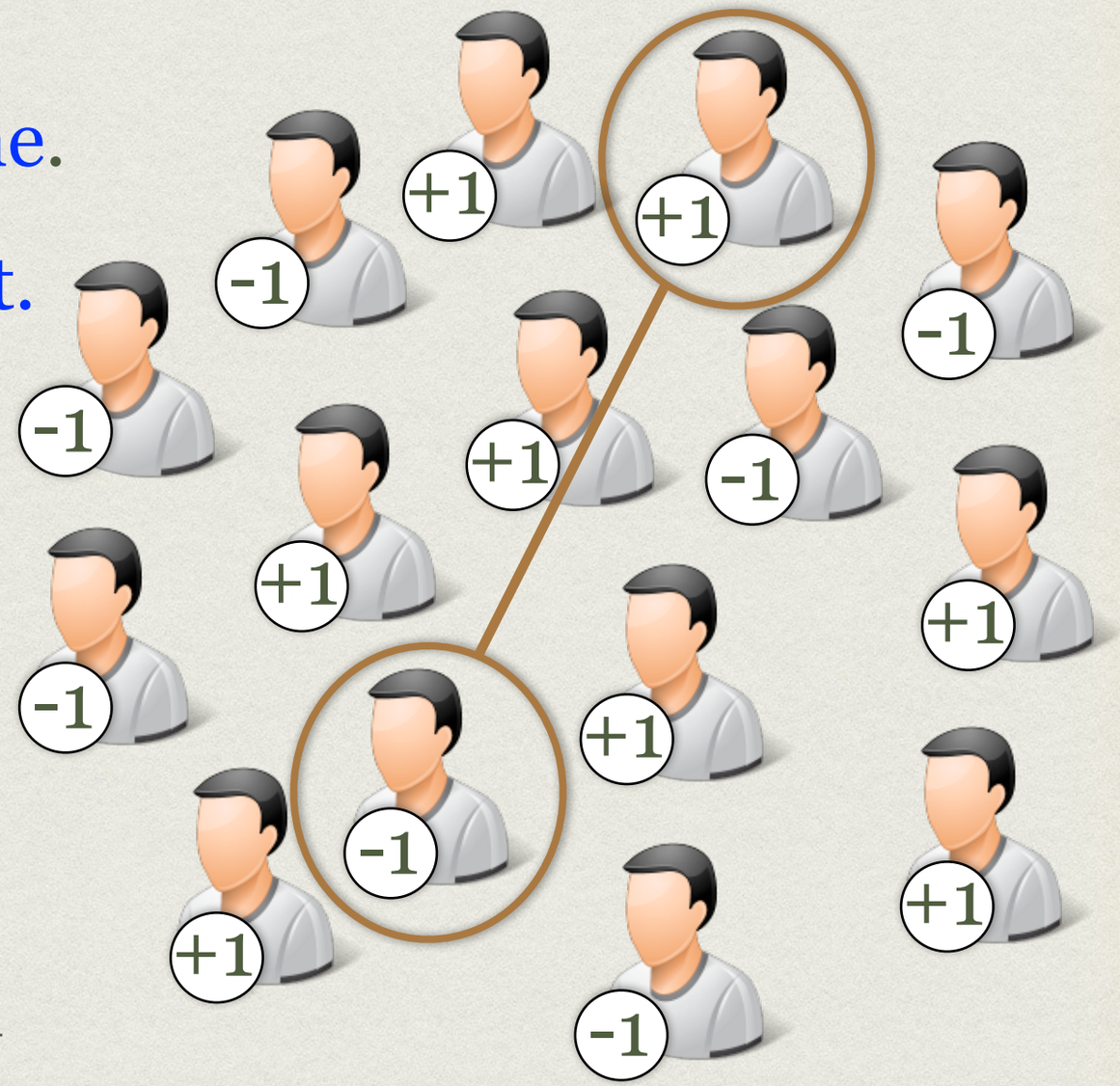
LET'S PLAY A GAME

- Each pair reports:
  - $Y_{ij}=W_{ij}+1/\sqrt{N}$ if cards the same.
  - $Y_{ij}=W_{ij}-1/\sqrt{N}$ if cards different.

$$W_{ij} \sim \mathcal{N}(0, \Delta^*)$$

Collect Y_{ij} for every pair (ij).

Goal: Recover cards purely from the knowledge of $Y = \{Y_{ij}\}_{i<j}$



HOW TO SOLVE THIS?

$$Y_{ij} = \frac{1}{\sqrt{N}} x_i^* x_j^* + W_{ij}$$

true values of cards:

$$W_{ij} \sim \mathcal{N}(0, \Delta^*)$$

$$x^* \in \{-1, +1\}^N$$

$$x_i^* \in \{-1, +1\}$$

Principal component analysis

x_{PCA} = leading eigenvector of Y estimates x* (up to a sign).

BBP phase transition: $\Delta > 1$ $\quad x_{\mathrm{PCA}} \cdot x^* \approx 0$

Watkin, Nadal'94
Baik, BenArous, Pechet'04

$\Delta < 1$ $\quad |x_{\mathrm{PCA}} \cdot x^*| > 0$

What is the minimal achievable estimation error on x*?

(Is it possible to do better than PCA?)

What is the minimal efficiently achievable estimation error on x*?

BAYESIAN INFERENCE

$$P(x|Y) = \frac{P(x)P(Y|x)}{P(Y)}$$

Values of cards: $x \in \{-1, +1\}^N$

$x_i \in \{-1, +1\}$

Posterior distribution:

$$P(x|Y) = \frac{1}{Z(Y,\Delta)} \prod_{i=1}^{N} [\delta(x_i + 1) + \delta(x_i - 1)] \prod_{i<j} e^{-\frac{(Y_{ij} - x_i x_j/\sqrt{N})^2}{2\Delta}}$$

Bayes-optimal inference = computation of marginals:

$$\mu(x_i) = \sum_{\{x_j\}_{j \neq i}} P(x \,|\, Y)$$

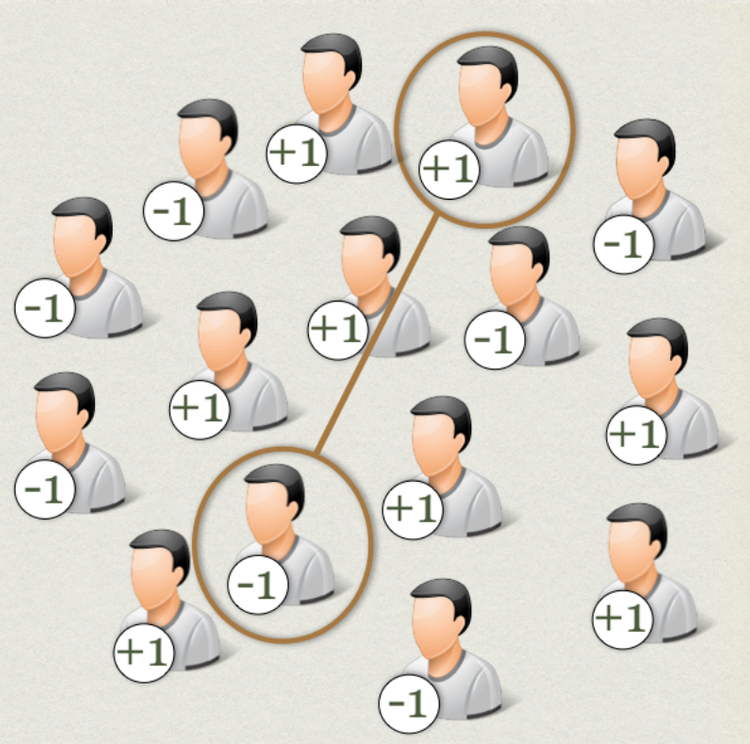
Computationally expensive in general (#P-hard)

How do we compute the Bayes-optimal performance?

Map to a spin glass?

$$Y \to J$$

$$x_i \to S_i$$



BACK TO THE CARD GAME

$$P(S|J) = \frac{1}{Z(J,\Delta^*)} \prod_{i<j} e^{-\frac{1}{2\Delta^*}(J_{ij} - S_i S_j/\sqrt{N})^2} \qquad S_i \in \{-1, +1\}$$

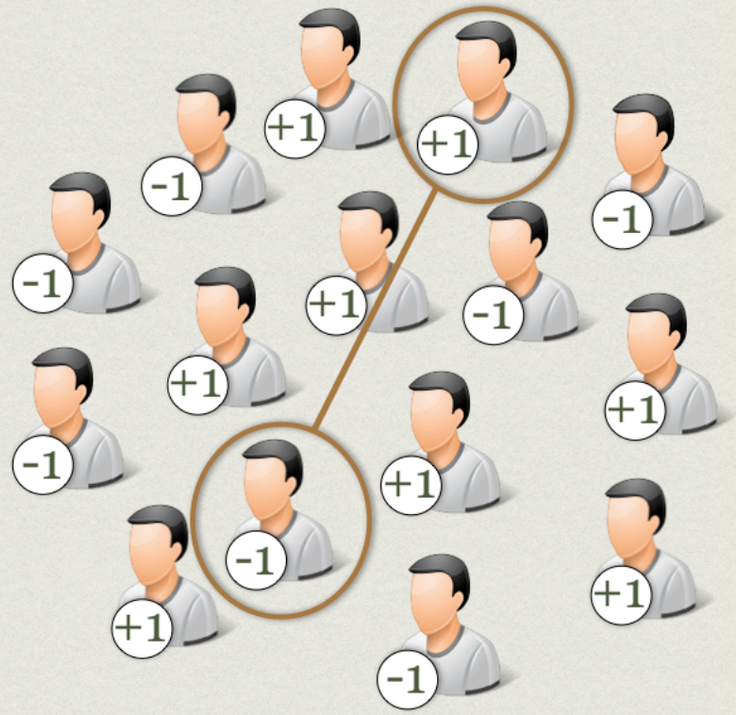
Hamiltonian

Boltzmann distribution → $P(S|J) = \frac{1}{\tilde{Z}(J,\Delta^*)} e^{-\frac{1}{\Delta^*\sqrt{N}} \sum_{i<j} J_{ij} S_i S_j}$

Partition function

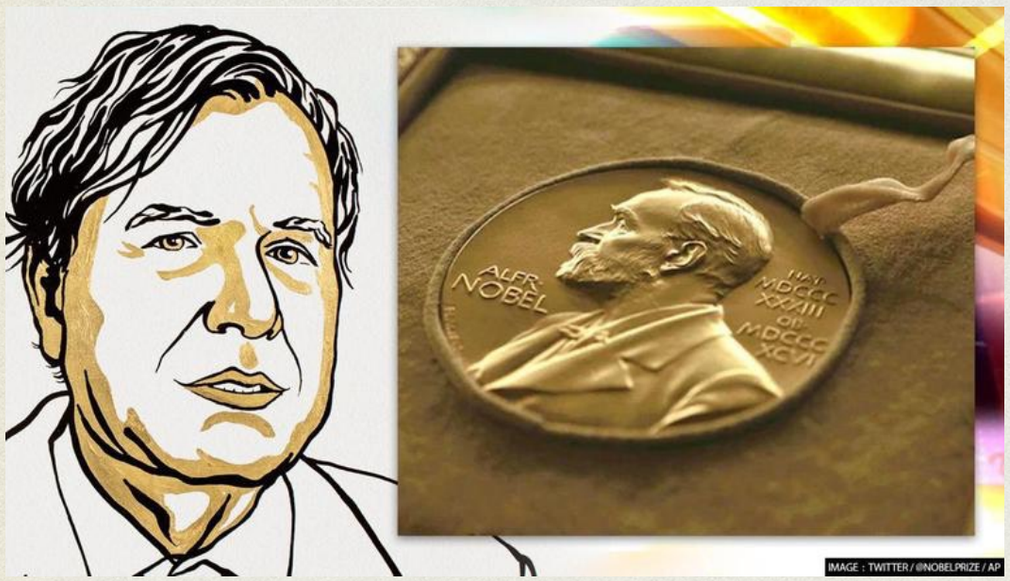
Mean-field Ising spin glass
(Sherrington-Kirkpatrick'75 model)

J_{ij} conditioned on S_i^*: *planted* disorder



HOW TO SOLVE THIS?

- Mean-field spin glass models are exactly solvable using replica method / cavity method. (Mezard, Parisi, Nishimori, Watkin, Nadal, Sompolinsky, many many others 70s-80s.)



“Parisi's discoveries make it possible to understand and describe many different and apparently entirely **random complex materials and phenomena**, not only in physics but also in other, very different areas, such as mathematics, biology, neuroscience and **machine learning**.”

MEAN-FIELD SPIN GLASS

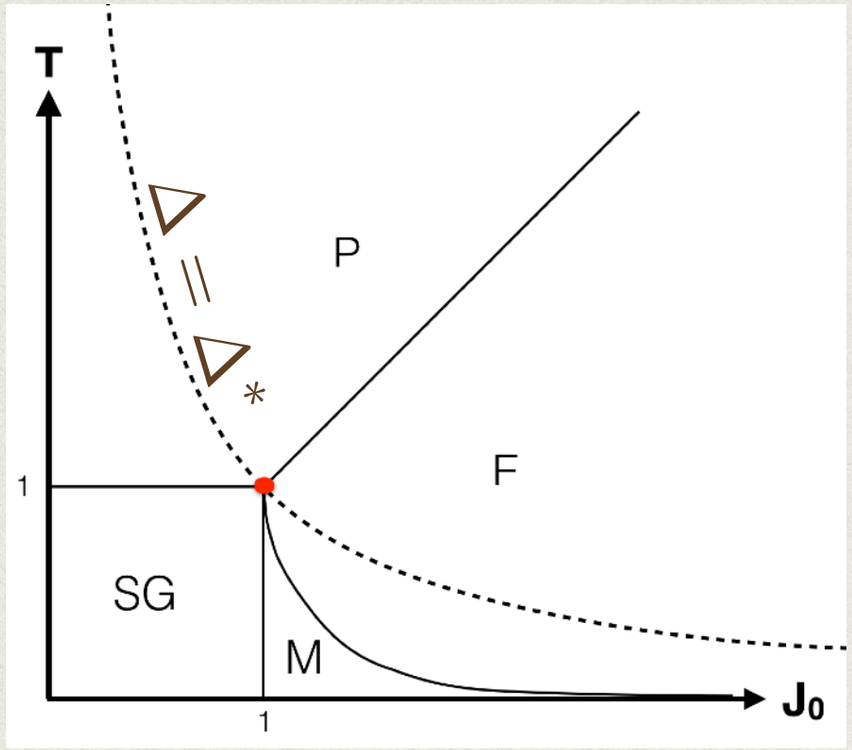
- For Ising spins, planting is equivalent to ferromagnetic bias

$$J_0 = 1/\sqrt{\Delta^*} \qquad T = \sqrt{\Delta}$$

Ill: N. Elmehed. © Nobel Media 2016
David J. Thouless

De Almeida; Thouless'78:

$\sqrt{\Delta}$

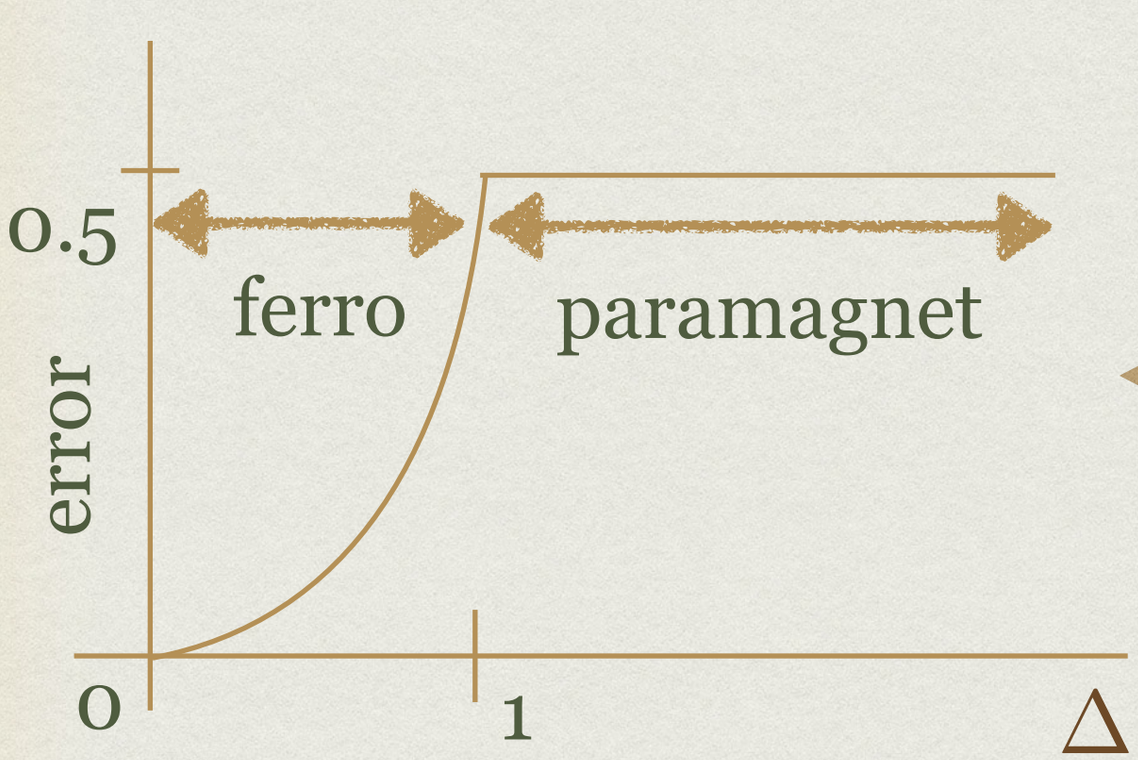


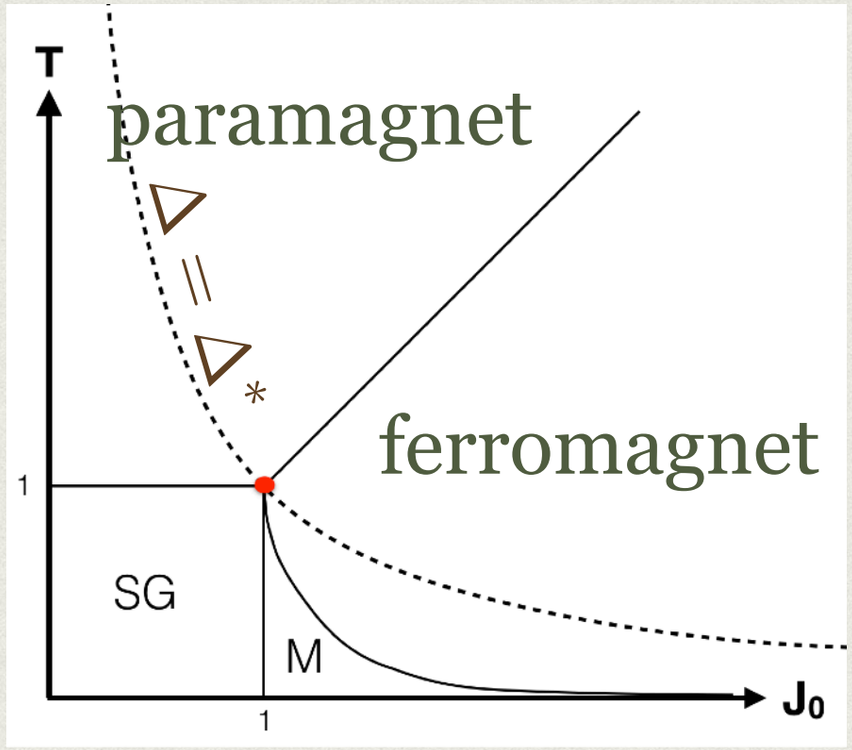
$\dfrac{1}{\sqrt{\Delta^*}}$

MEAN-FIELD SPIN GLASS

- For Ising spins, planting is equivalent to ferromagnetic bias

$$J_0 = 1/\sqrt{\Delta^*} \qquad T = \sqrt{\Delta}$$





What can such an analogy be good for?

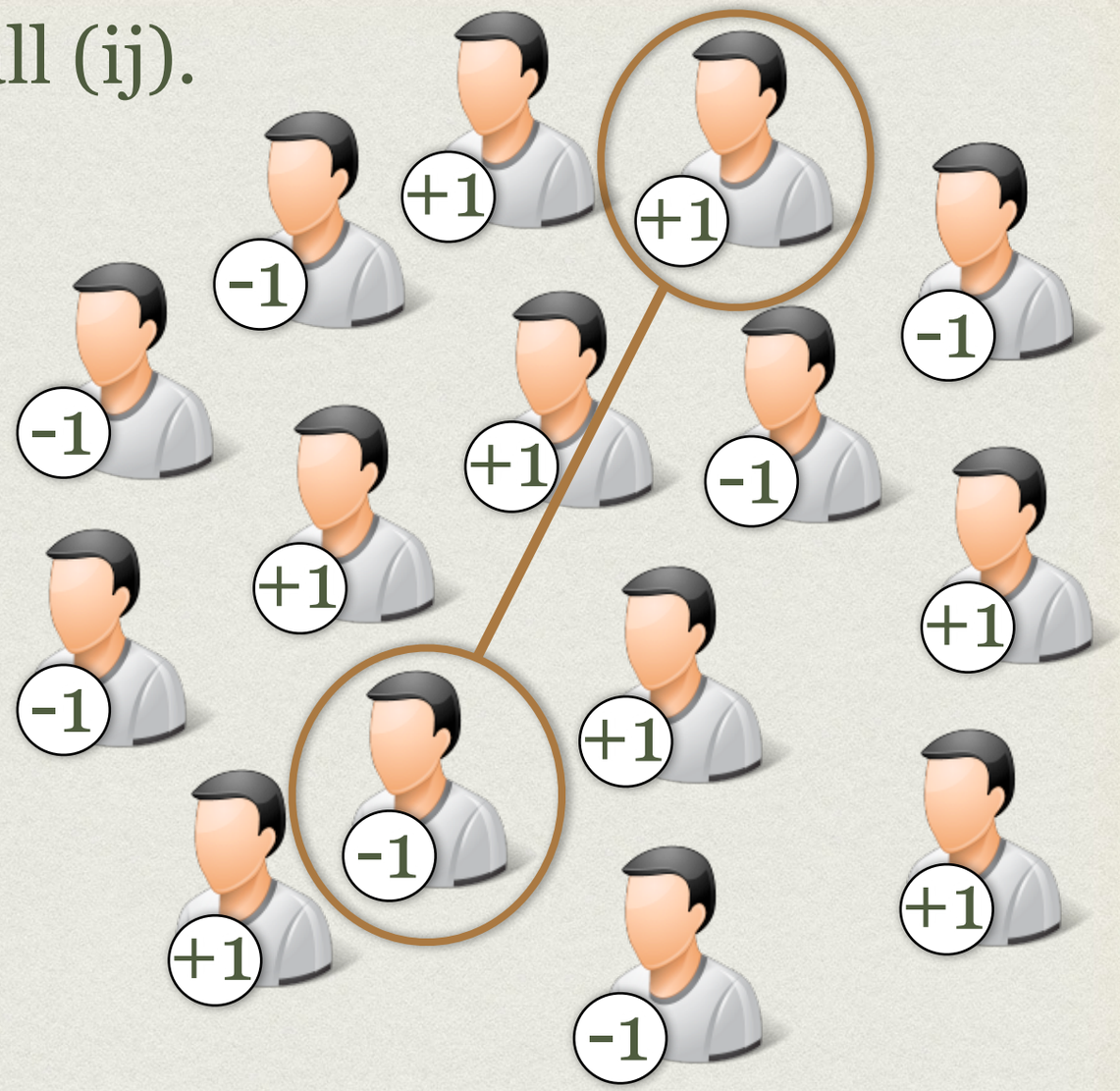
Designing new algorithms.

(Krzakala, Moore, Mossel, Neeman, Sly, LZ, Zhang, PNAS‘2013)

BACK TO THE GAME

Collect $J_{ij} = W_{ij} + S_i^* S_j^* / \sqrt{N}$, for all (ij).

Goal: Recover cards purely from the knowledge of $\mathbf{J} = \{J_{ij}\}_{i<j}$



Simple spectral algorithm:

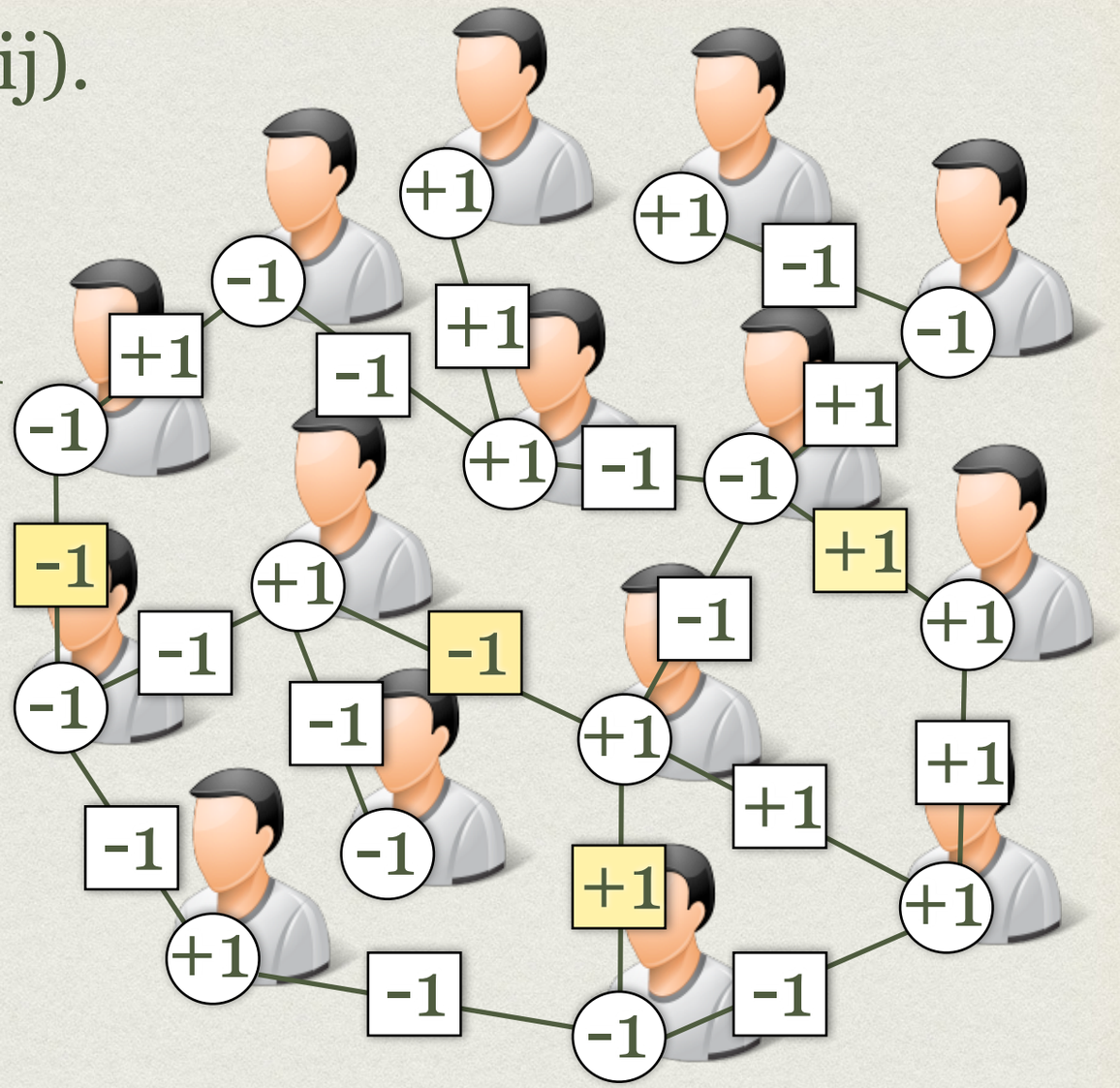
Leading eigenvector of J correlated to S*. Phase transition matches (overlap smaller).

BACK TO THE GAME

Collect $J_{ij} = S_i^* S_j^*$, for cN/2 of (ij).

Flip with probability ρ

Goal: Recover cards purely from the knowledge of $\mathbf{J} = \{J_{ij}\}_{i<j}$



Simple spectral algorithm:

Leading eigenvector of J correlated to S*. Phase transition does not match.

Caveat: Spectral algorithms for sparse data fail

(do not work down to the easy/hard phase transition).

SPARSE DATA: EXAMPLES

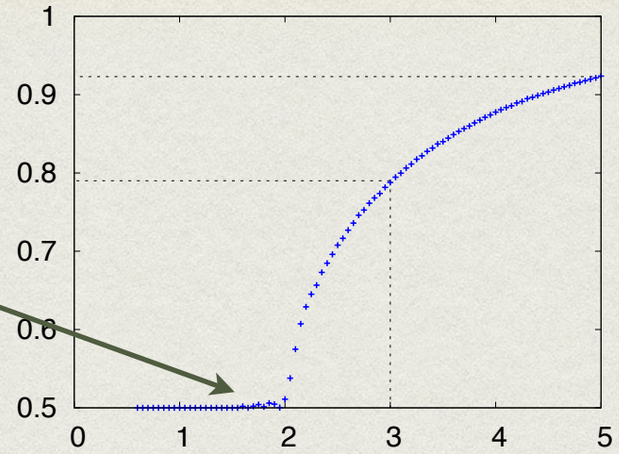
- Clustering sparse networks: Number of friends does not grow with the size of the world.
- Similarity based clustering: Obtaining similarities is costly. (lengthy experiment, or cost of information).
- Recommendation systems: Only some users ranked some movies. Goal: reconstruct the rest of ratings.
- Big data: the full matrix can not be stored nor analyzed.

Sparse matrices:

Leading eigenvalues arise from local heterogeneity/impurity not from global structure (Lisfhitz tails).

HOW CAN PHYSICS HELP?

- Analyze the phase transition



Ill: N. Elmehed. © Nobel Media 2016
David J. Thouless

VOLUME 56, NUMBER 10 PHYSICAL REVIEW LETTERS 10 MARCH 1986

Spin-Glass on a Bethe Lattice

D. J. Thouless
Department of Physics, University of Washington, Seattle, Washington 98195
(Received 27 November 1985)

The Ising spin-glass in a magnetic field is studied for the Bethe lattice. There is an instability that agrees with the replica-symmetry–breaking transition found for the infinite-range model. Correlation lengths are finite on both sides of the transition, but there is a correlation length that diverges at the transition. Some features are different from those of the infinite-range model, and in particular the magnetic susceptibility and internal energy vary smoothly through the transition. An analogy with the localization transition on the Bethe lattice is pointed out.

THOULESS' CALCULATION

Bethe recursion:

$$u_t^{i\to j} = \frac{1}{\beta}\operatorname{atanh}\left[\tanh(\beta J_{ij})\tanh\left(\beta\sum_{k\in\partial i\setminus j} u_{t-1}^{k\to i}\right)\right]$$

paramagnetic fixed point $\quad u^{i\to j} = 0 \quad \forall (ij)\in G$

small-u expansion $\quad u_t^{i\to j} = \tanh(\beta J_{ij})\sum_{k\in\partial i\setminus j} u_{t-1}^{k\to i}$

square and average over disorder

$$\langle u_t^2\rangle = c\langle\tanh^2(\beta J_{ij})\rangle_J \langle u_{t-1}^2\rangle$$

critical value: $\quad 1 = c\tanh^2(\beta_c) \quad$ for $\quad J_{ij}=\pm 1$

(c is average degree, c=2M/N)

So where is the spectral method?

ANOTHER VIEW

(Krzakala, Moore, Mossel, Neeman, Sly, LZ, Zhang, PNAS'2013)

$$u_t^{i\to j} = \frac{1}{\beta}\operatorname{atanh}\left[\tanh(\beta J_{ij})\tanh\left(\beta\sum_{k\in\partial i\setminus j} u_{t-1}^{k\to i}\right)\right]$$

paramagnetic fixed point $\quad u^{i\to j}=0 \quad \forall (ij)\in G$

small-u expansion $\quad u_t^{i\to j} = \tanh(\beta J_{ij})\sum_{k\in\partial i\setminus j} u_{t-1}^{k\to i}$

~~square and average over disorder~~

~~$\langle u_t^2\rangle = c\langle\tanh^2(\beta J_{ij})\rangle_J \langle u_{t-1}^2\rangle$~~

~~critical value: $1 = c\tanh^2(\beta_c)$ for $J_{ij}=\pm 1$~~

ANOTHER VIEW

(Krzakala, Moore, Mossel, Neeman, Sly, LZ, Zhang, PNAS'2013)

$$u_t^{i\to j} = \frac{1}{\beta}\mathrm{atanh}\left[\tanh(\beta J_{ij})\tanh\left(\beta\sum_{k\in\partial i\setminus j} u_{t-1}^{k\to i}\right)\right]$$

paramagnetic fixed point $\quad u^{i\to j} = 0 \quad \forall (ij)\in G$

small-u expansion

$$u_t^{i\to j} = \tanh(\beta J_{ij})\sum_{k\in\partial i\setminus j} u_{t-1}^{k\to i}$$

$$\vec{u}_t = B\vec{u}_{t-1}$$

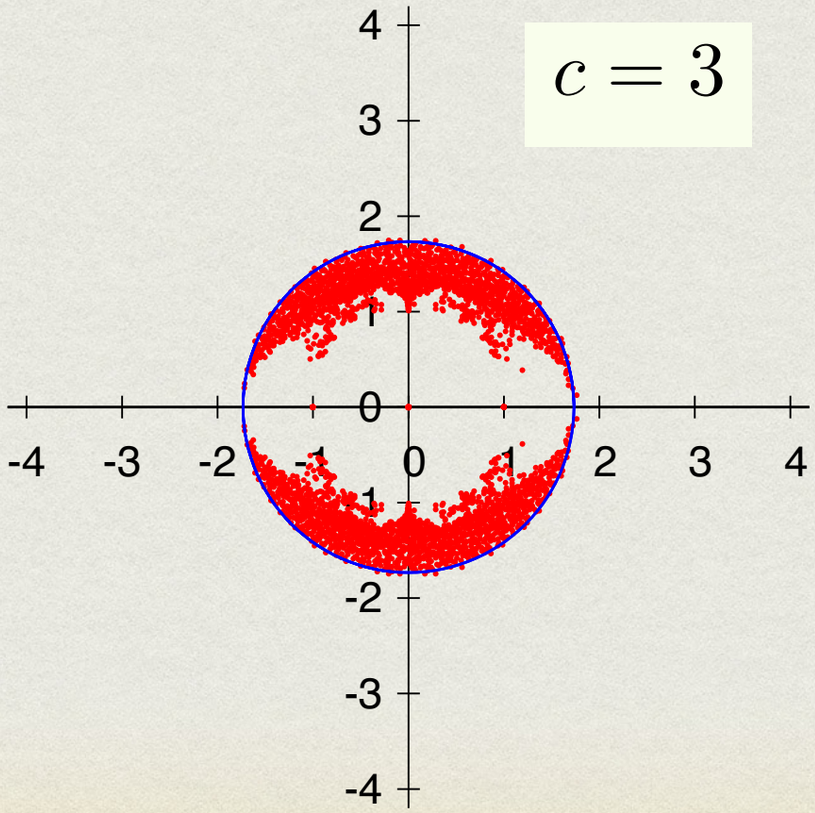
non-backtracking operator

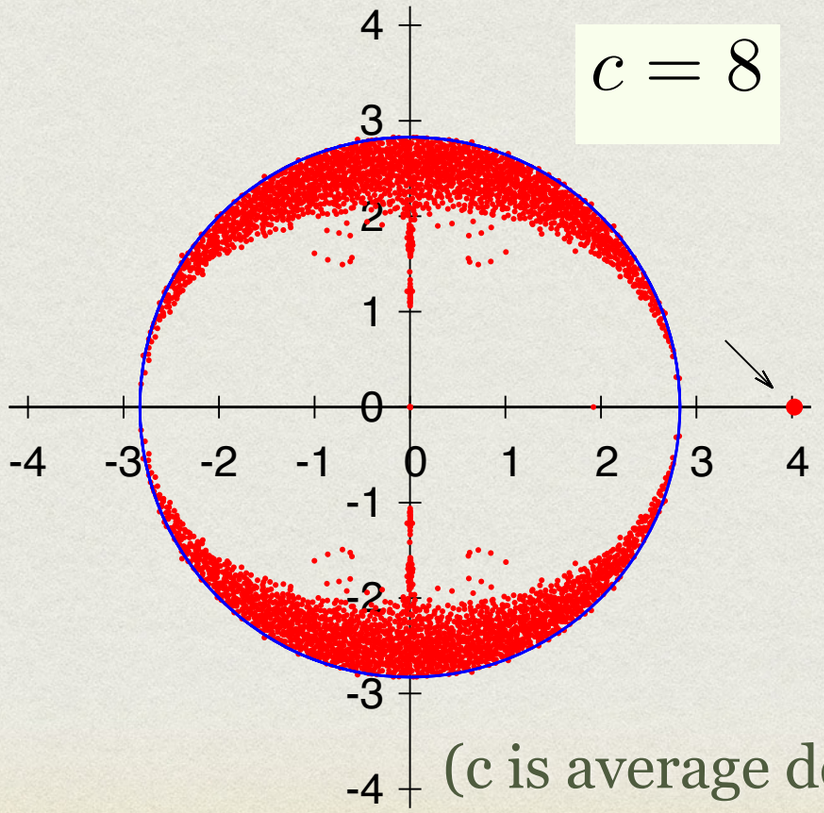
$$B_{i\to j,k\to l} = \delta_{j,k}(1-\delta_{i,l})a_{ij} \quad \text{if} \quad (ij),(kl)\in E$$
$$= 0 \quad \text{otherwise}$$

NON-BACKTRACKING SPECTRAL ALGORITHM

$c < \dfrac{1}{(2\rho-1)^2}$ all eigenvalues inside a circle of radius $\sqrt{c}$

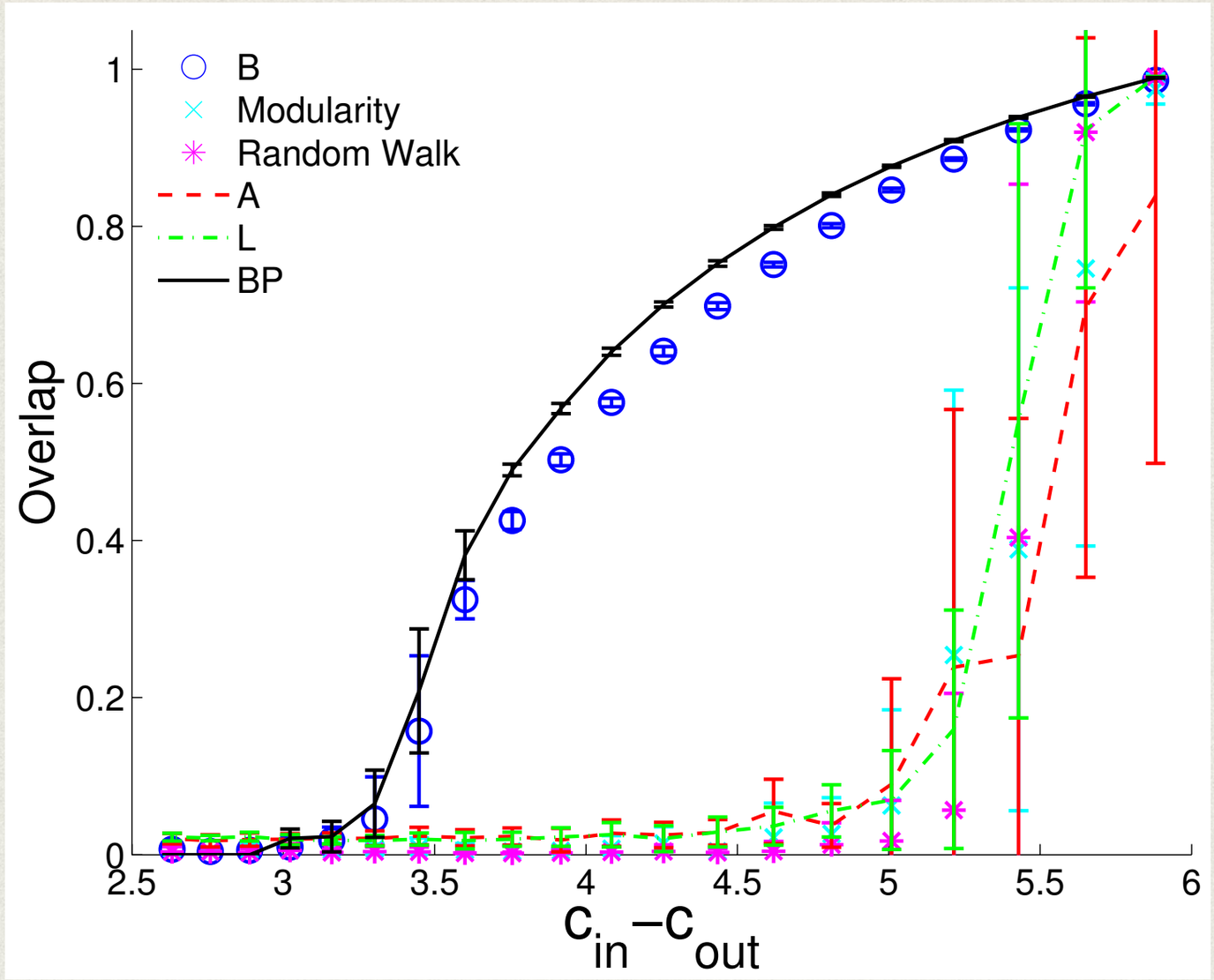
$c > \dfrac{1}{(2\rho-1)^2}$ additional real eigenvalue at $\lambda = c(2\rho-1) > \sqrt{c}$ eigenvector correlated to the ground truth.





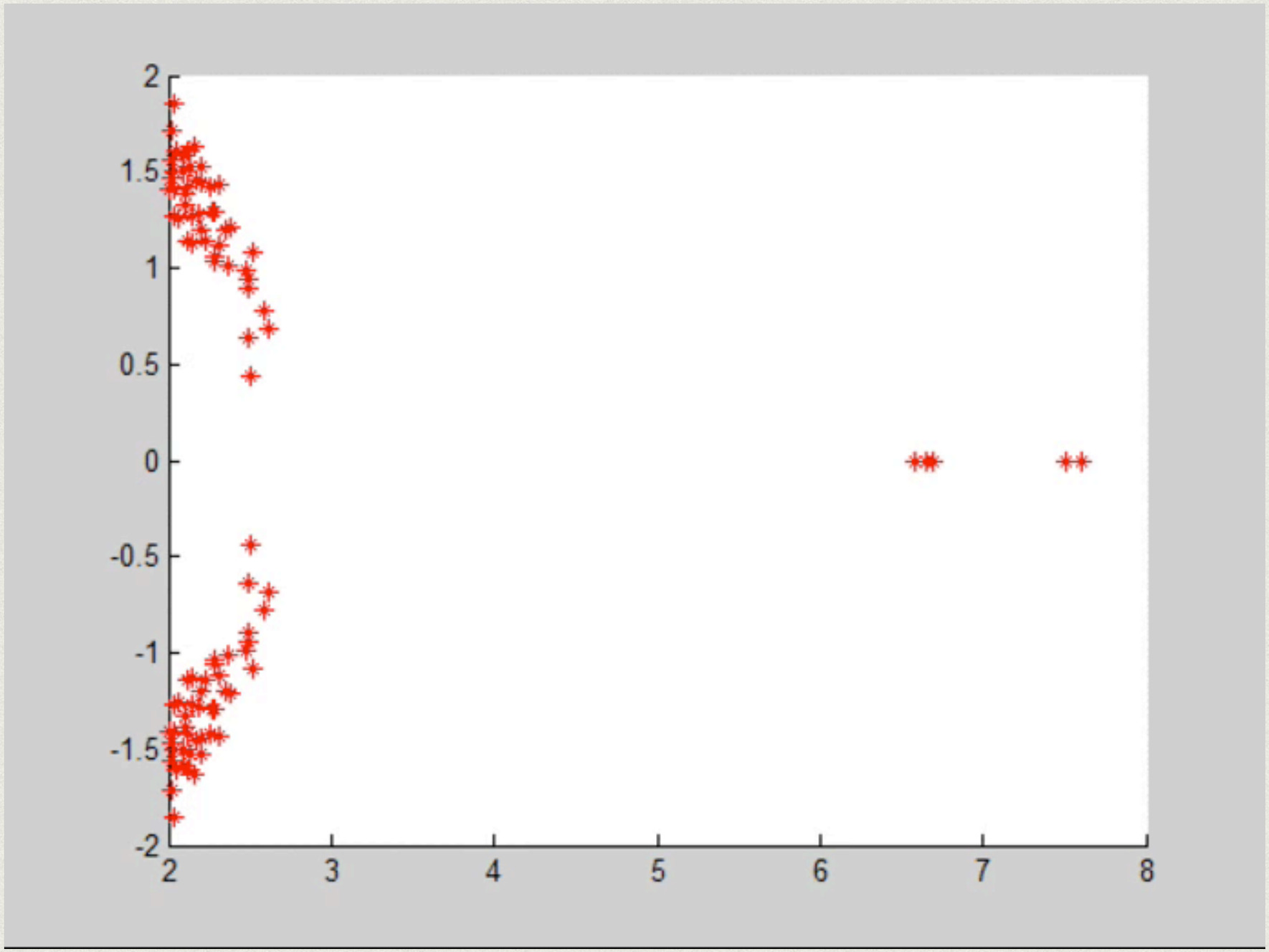
(c is average degree, c=2M/N)

NON-BACKTRACKING SPECTRAL ALGORITHM FOR CLUSTERING OF NETWORKS



Non-backtracking spectral method improves over traditional ones.

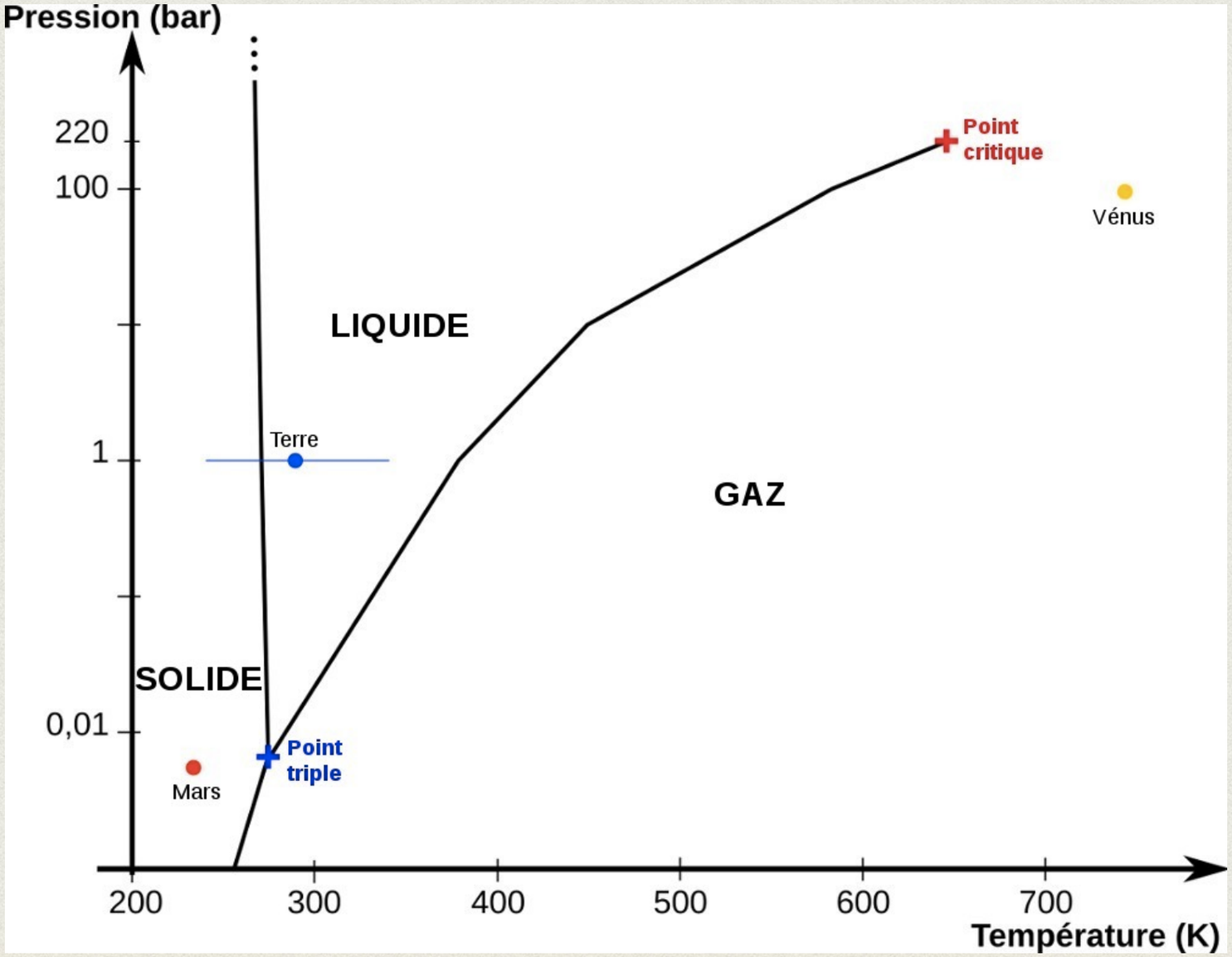
Spectrum of a graph with 5 communities as edges are added between groups.



1ST ORDER PHASE TRANSITIONS

PHASE TRANSITIONS

Phase diagram of water

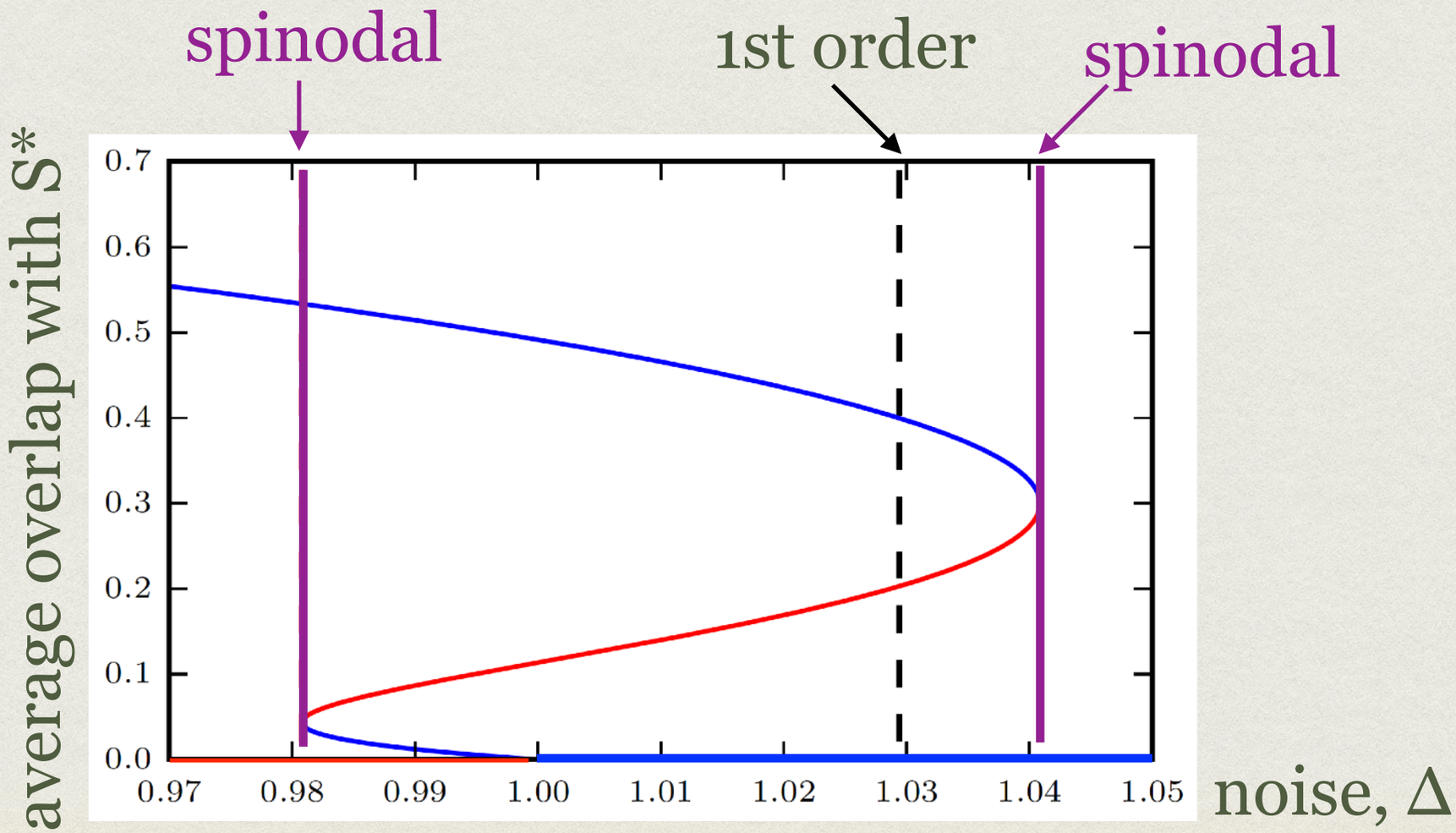


1ST ORDER PHASE TRANSITIONS

Slight change of the rules of the game:

$$P(S^*) = \rho[\delta(S^* - 1) + \delta(S^* + 1)]/2 + (1 - \rho)\delta(S^*)$$

(sparse PCA - relevant in data-science applications to learn relevant dimensions)

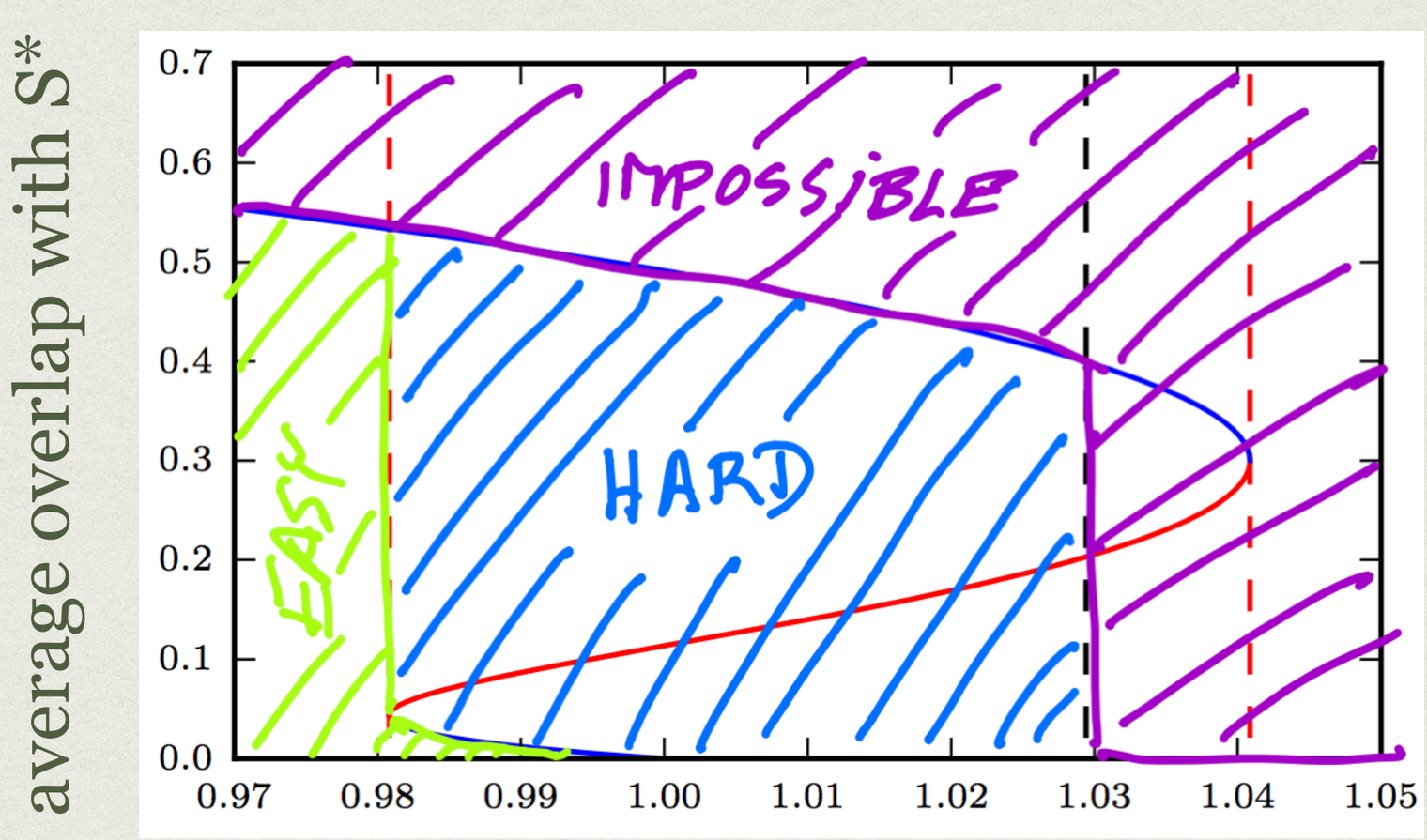


1ST ORDER PHASE TRANSITION

Slight change of the rules of the game:

$$P(S^*) = \rho[\delta(S^* - 1) + \delta(S^* + 1)]/2 + (1 - \rho)\delta(S^*)$$

(sparse PCA - relevant in data-science applications to learn relevant dimensions)

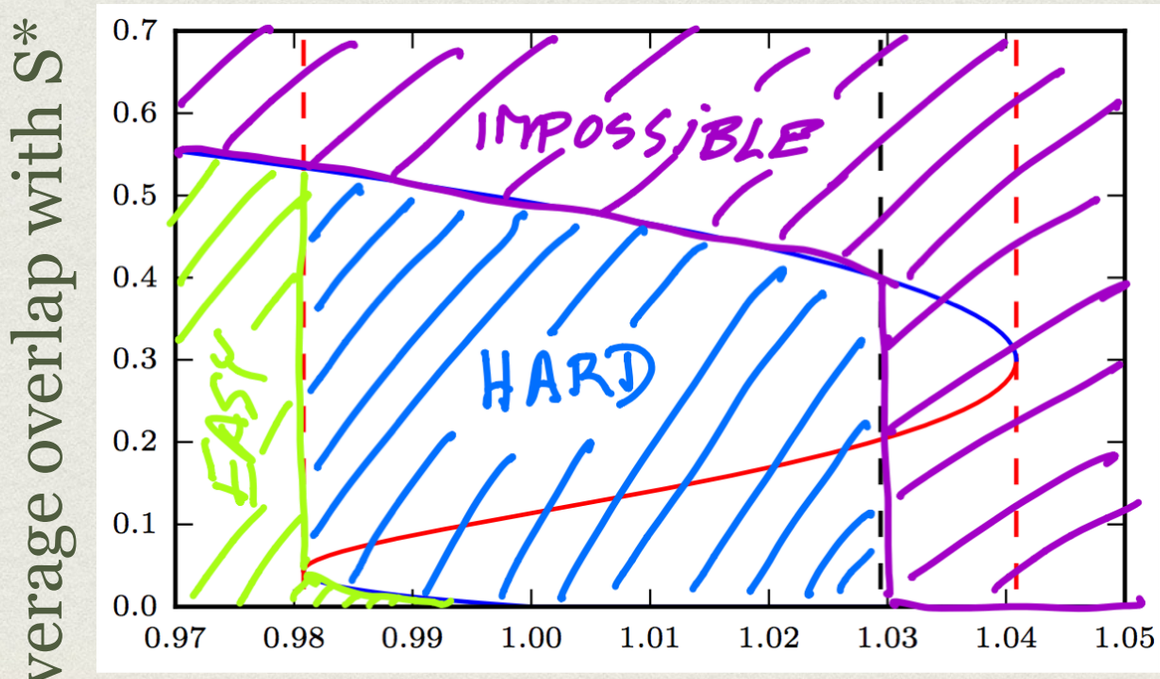


ALGORITHMIC INTERPRETATION

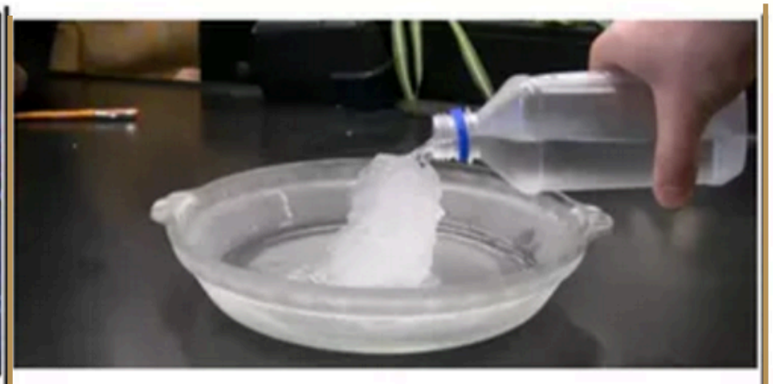
- Easy by approximate message passing algorithms.
- Impossible information theoretically.
- Hard phase conjecture: No polynomial algorithm works.

Mathematically wide open.

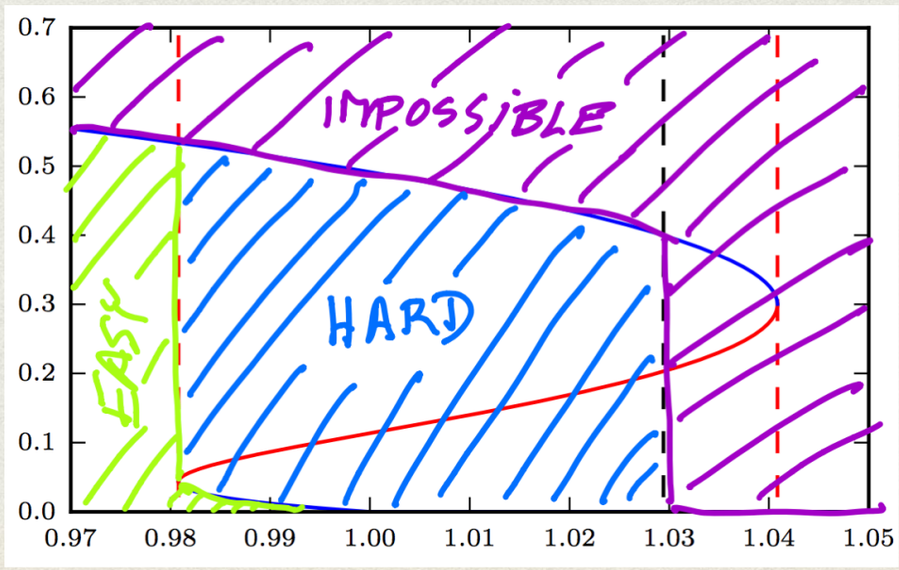
Physically sensible.



PHYSICS VS LEARNING

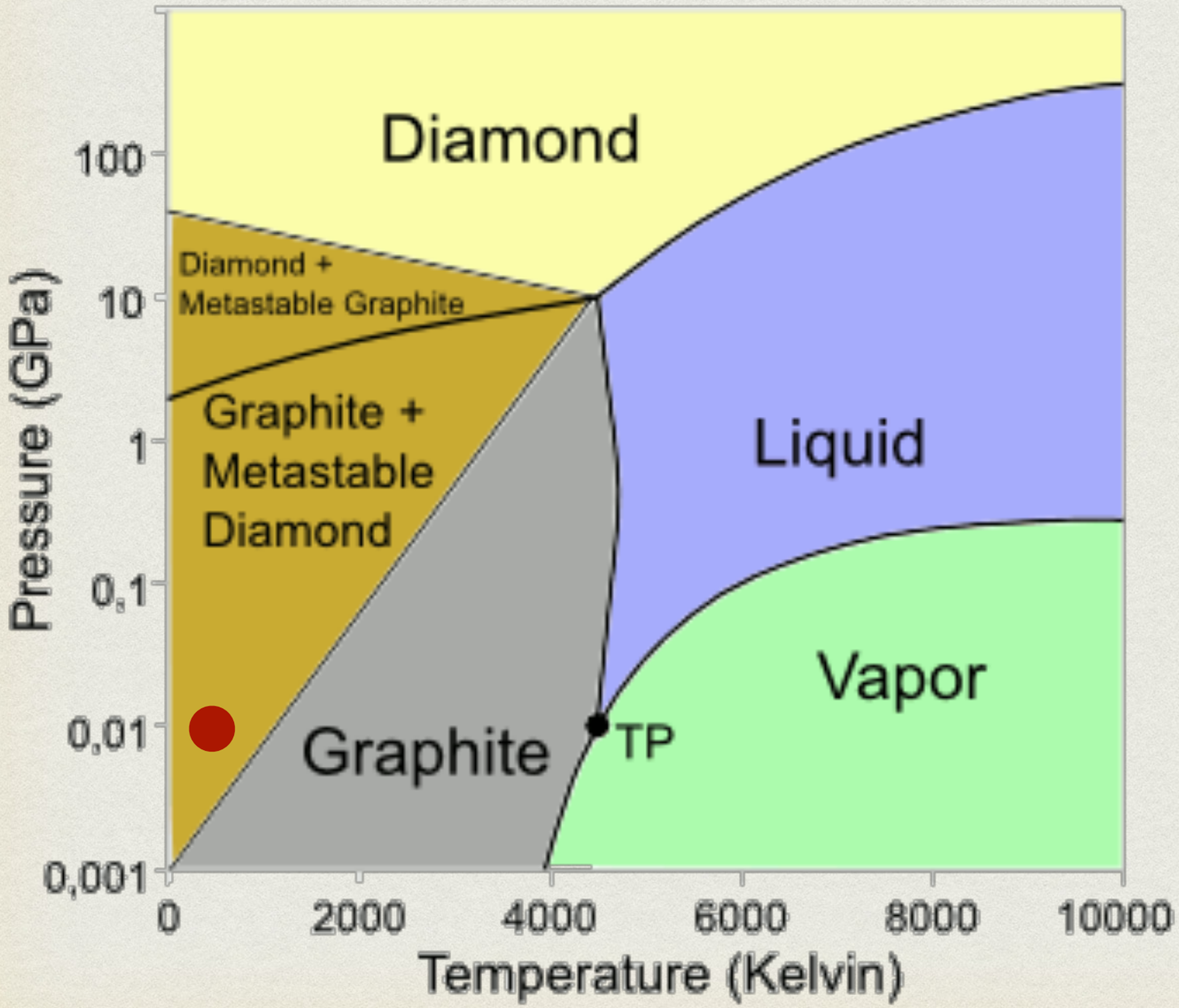


liquid	supercooled liquid	ice
impossible	computationally hard	possible



HARD REGIME

Hard phase: Algorithms “stuck” at low accuracy for exponential time.



Metastable diamond
= low accuracy.

Equilibrium graphite
= high accuracy.


AURELE

comment

Check for updates

nature physics

Understanding deep learning is also a job for physicists

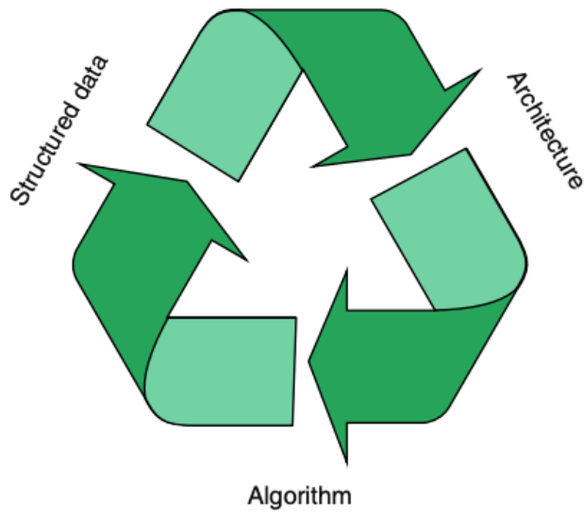
Automated learning from data by means of deep neural networks is finding use in an ever-increasing number of applications, yet key theoretical questions about how it works remain unanswered. A physics-based approach may help to bridge this gap.

Lenka Zdeborová

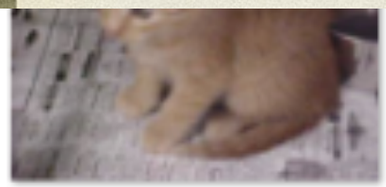
Imagine an event for which thousands of tickets get sold out in under 12 minutes. We are not speaking of a leading show on Broadway or a concert of a rockstar, but about the Conference on Neural Information Processing Systems (NeurIPS) — the principal gathering for research in machine learning and artificial intelligence. The fields related to automated learning from data are experiencing a surge in research activity, as well as in investment. This is largely thanks to developments in a subfield called deep learning, which has led to a physicists it is a matter of sitting tight waiting for tools and answers that we can subsequently put to use. In this Comment, I argue that, instead, we need to join the race of searching for these answers, because it is precisely the physicists' perspective and approach that is needed to enable progress in this endeavour.

Three ingredients to decipher deep learning

The engineering details of current deep-learning systems, such as the ones



SUPERVISED LEARNING

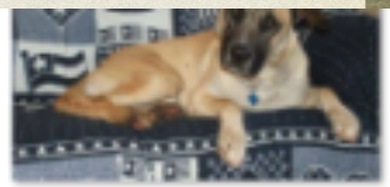

701.jpg

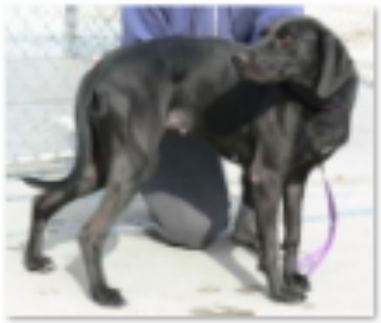

702.jpg


703.jpg


704.jpg


705.jpg


706.jpg


707.jpg


708.jpg


709.jpg


710.jpg


711.jpg


712.jpg


713.jpg

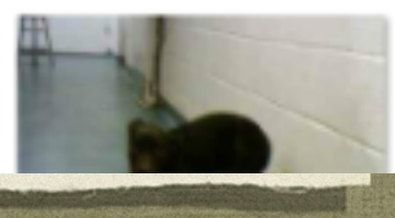

714.jpg


715.jpg


716.jpg


717.jpg


718.jpg



LEARNING A RULE

$= X_\mu =$ (01001010 01110011 10001100 01001011
01110000 10001100 ..... all the pixels ....)

Goal: Find a function f so that

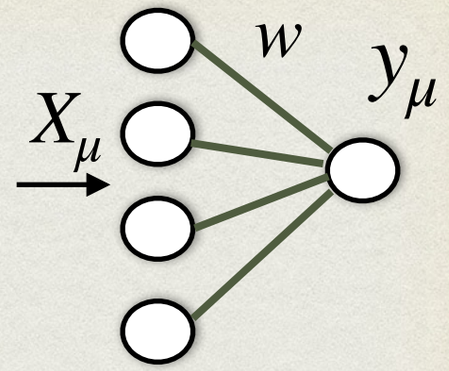
$f(X_\mu) = +1$ for a new picture of a cat.

$f(X_\mu) = -1$ for a new picture of a dog.

Today this is routinely done with deep neural networks.

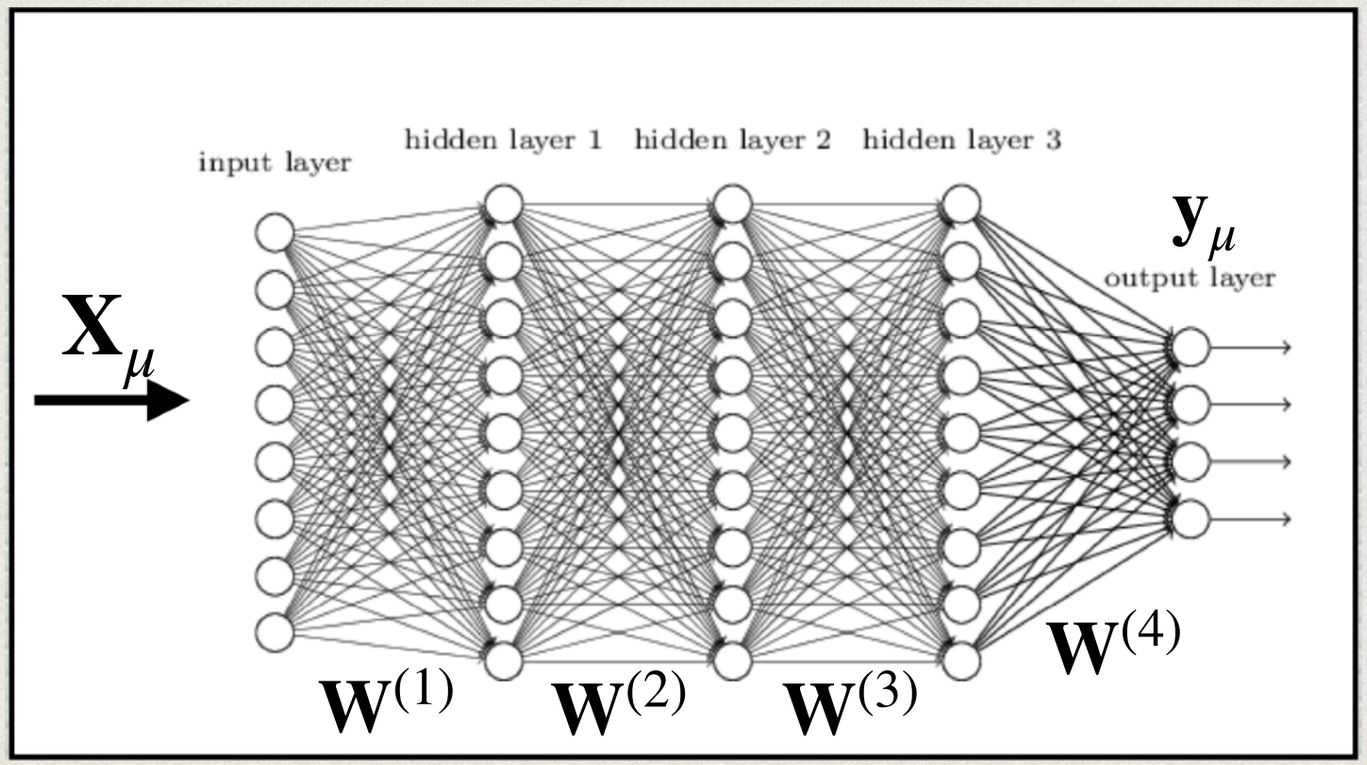
HOW DOES IT WORK?

- Linear regression: $f_w(\mathbf{X}_\mu) = \mathbf{w} \cdot \mathbf{X}_\mu$



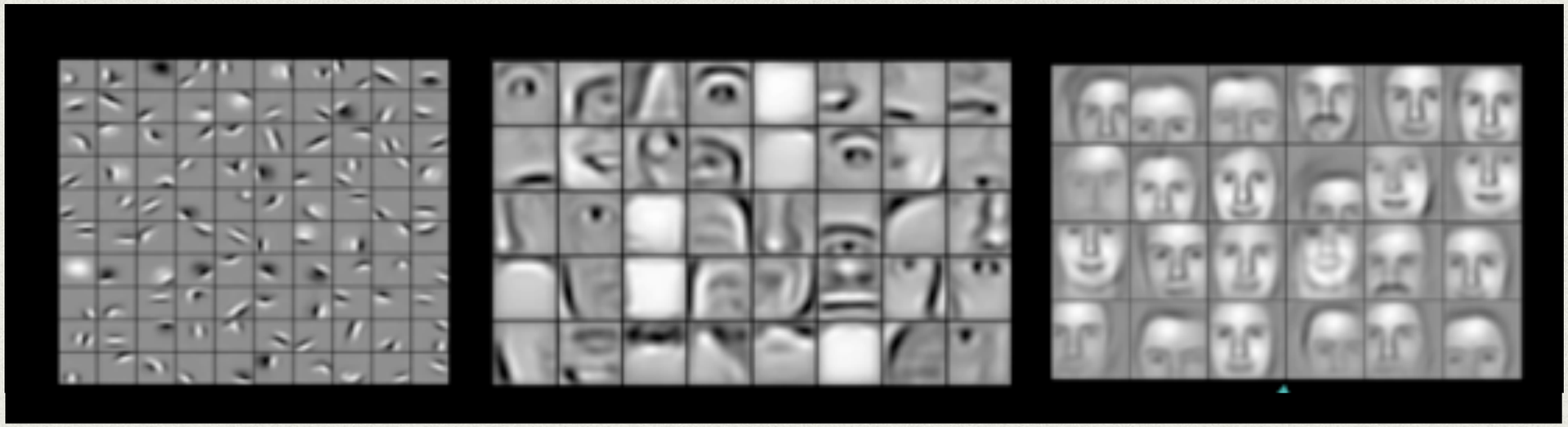
- Generalized linear regression: $f_w(\mathbf{X}_\mu) = \varphi(\mathbf{w} \cdot \mathbf{X}_\mu)$

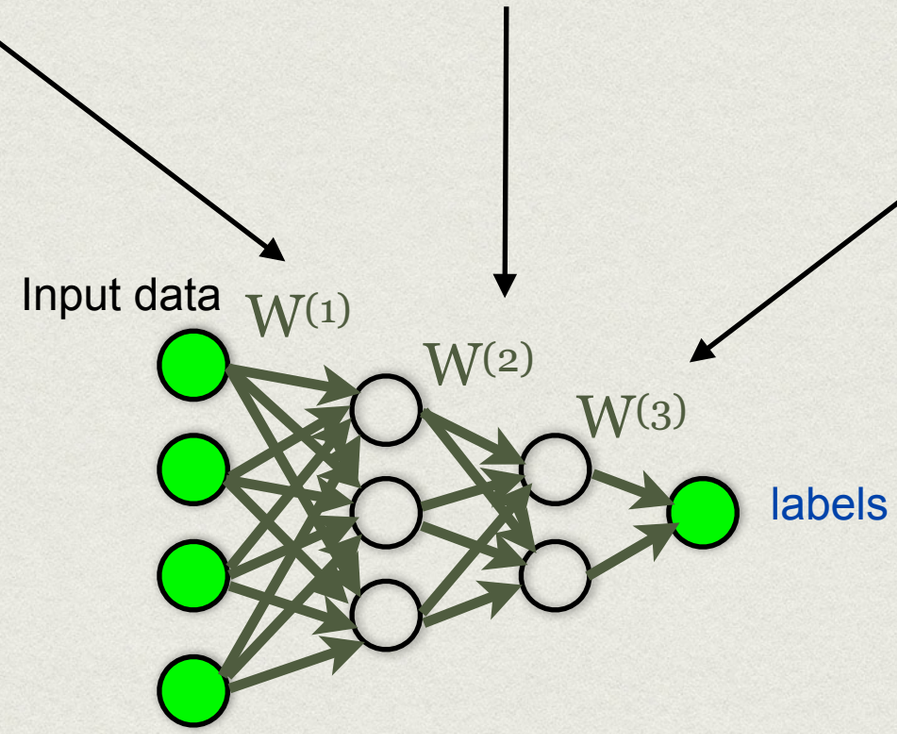
- Neural networks: $f_w(\mathbf{X}_\mu) = \varphi^{(4)}(\mathbf{W}^{(4)}\varphi^{(3)}(\mathbf{W}^{(3)}\varphi^{(2)}(\mathbf{W}^{(2)}\varphi^{(1)}(\mathbf{W}^{(1)}\mathbf{X}_\mu))))$



activation function,
e.g. sign, or relu.

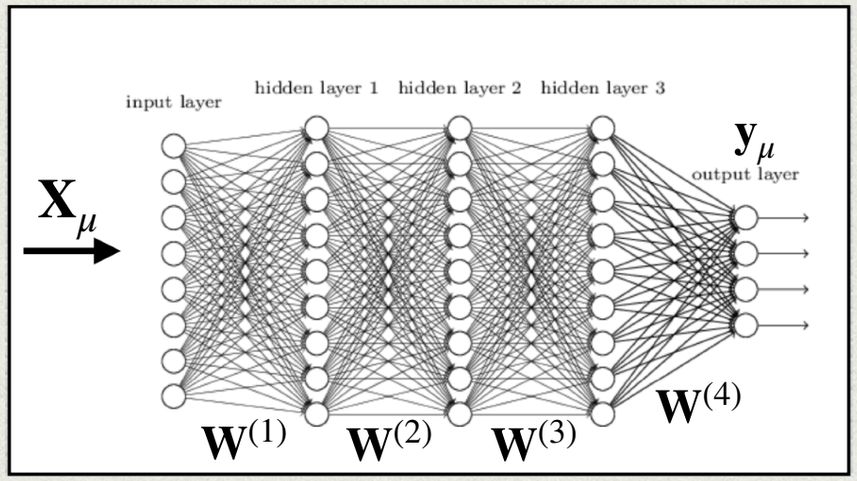
Hierarchy of features





HOW DOES IT WORK?

- Neural networks: $f_w(\mathbf{X}_\mu) = \varphi^{(4)}(\mathbf{W}^{(4)}\varphi^{(3)}(\mathbf{W}^{(3)}\varphi^{(2)}(\mathbf{W}^{(2)}\varphi^{(1)}(\mathbf{W}^{(1)}\mathbf{X}_\mu))))$



- Core of ML today: many labeled examples + GPUs + stochastic gradient descent minimisation of

$$\min_w \sum_{\mu=1}^{n} \mathscr{L}(y_\mu, f_w(X_\mu))$$

Loss function, e.g. least squares, or cross-entropy

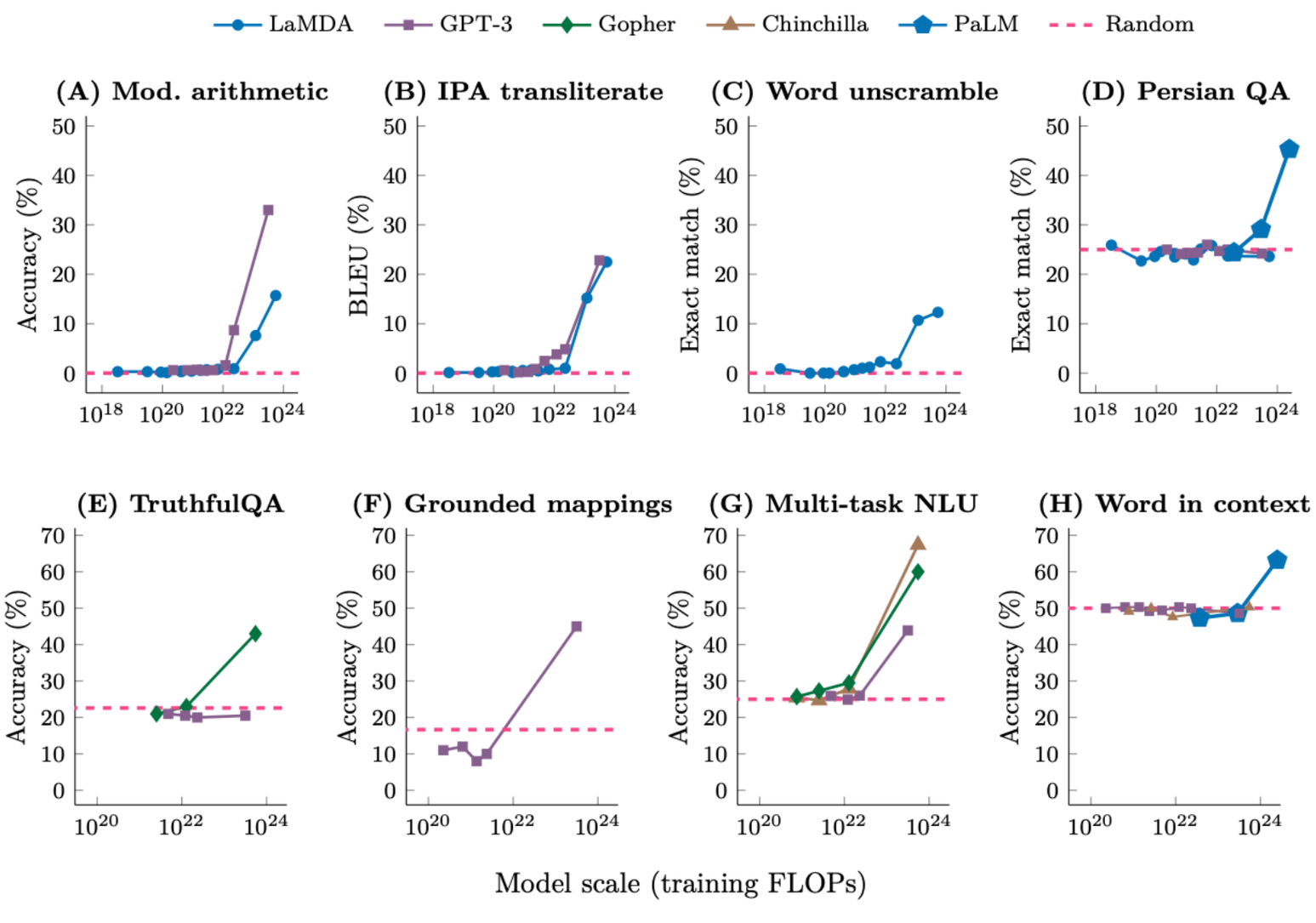
Keep in mind: The goal is low test error, not low loss.

PHASE TRANSITIONS IN LEARNING WITH NEURAL NETWORKS

Emergent Abilities of Large Language Models

Jason Wei [1] *jasonwei@google.com*
Yi Tay [1] *yitay@google.com*
Rishi Bommasani [2] *nlprishi@stanford.edu*
Colin Raffel [3] *craffel@gmail.com*
Barret Zoph [1]
Sebastian Borgeaud [4]
Dani Yogatama [4]
Maarten Bosma [1]
Denny Zhou [1]
Donald Metzler [1]
Ed H. Chi [1]
Tatsunori Hashimoto [2]
Oriol Vinyals [4]
Percy Liang [2]
Jeff Dean [1]
William Fedus [1]

[1] *Google Research* [2] *Stanford University* [3] *UNC C*



RAPID COMMUNICATIONS

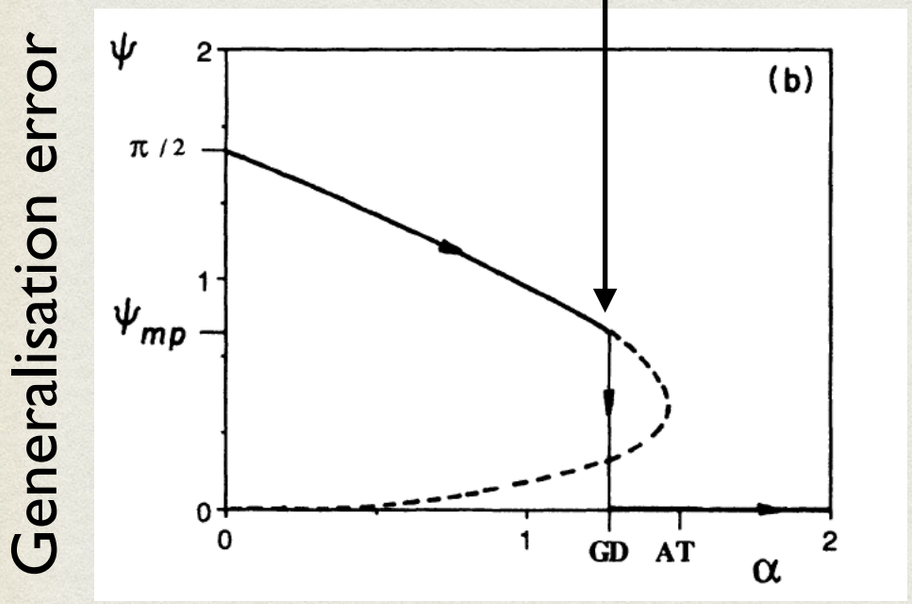
PHYSICAL REVIEW A VOLUME 41, NUMBER 12 15 JUNE 1990

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.

$\alpha_{GD} = 1.245$



samples / # input dimensions

VOLUME 65, NUMBER 13 PHYSICAL REVIEW LETTERS 24 SEPTEMBER 1990

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)

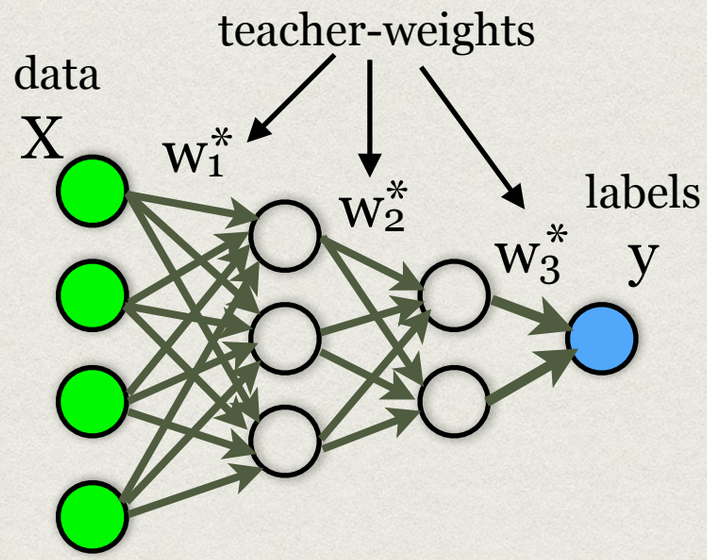
A statistical mechanical theory of learning from examples in layered networks at finite temperature is studied. When the training error is a smooth function of continuously varying weights the generalization error falls off asymptotically as the inverse number of examples. By analytical and numerical 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.

PACS numbers: 87.10.+e, 02.50.+s, 05.20.−y

TEACHER-STUDENT SETTING

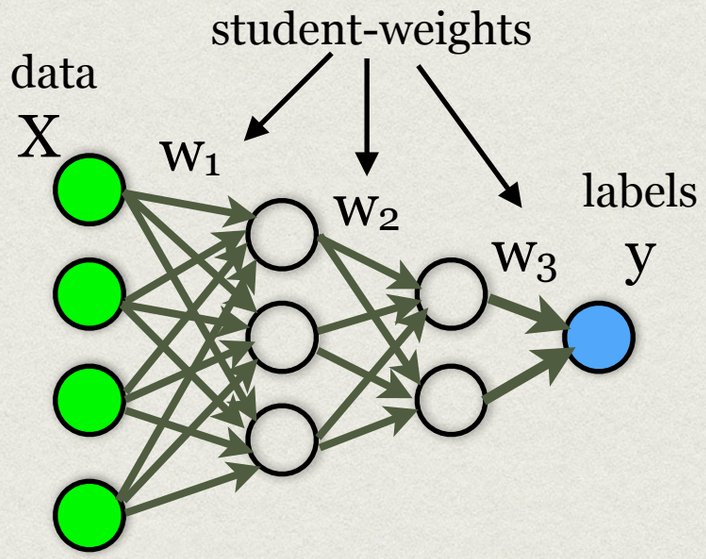
Teacher-network

- Generate data X, n samples of d dimensional data, iid Gaussian.
- Generate random weights w*.
- Generate labels y using:



Student-network

- Observes X, y, the architecture of the network.
- Goal: Learn the same function as used by the teacher (have a good test error).



GARDNER PROGRAM

J. Phys. A: Math. Gen. **22** (1989) 1983-1994. Printed in the UK

1989

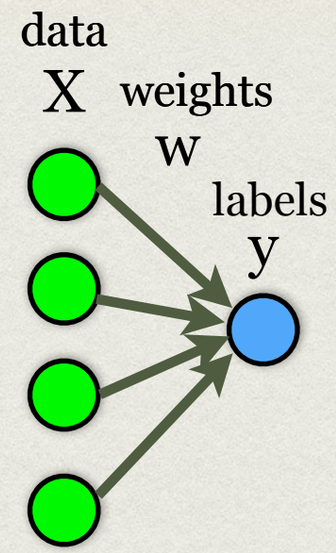
Three unfinished works on the optimal storage capacity of networks

E Gardner and B Derrida

The Institute for Advanced Studies, The Hebrew University of Jerusalem, Jerusalem, Israel and Service de Physique Théorique de Saclay†, F-91191 Gif-sur-Yvette Cedex, France

Received 13 December 1988

Abstract. The optimal storage properties of three different neural network models are studied. For two of these models the architecture of the network is a perceptron with $\pm J$ interactions, whereas for the third model the output can be an arbitrary function of the inputs. Analytic bounds and numerical estimates of the optimal capacities and of the minimal fraction of errors are obtained for the first two models. The third model can be solved exactly and the exact solution is compared to the bounds and to the results of numerical simulations used for the two other models.



- Take random iid Gaussian $X_{\mu i}$, and random iid w_i^* from P_w
- Create $y_\mu = \sigma\left(\sum_{i=1}^{d} X_{\mu i} w_i^* \right)$
- High-dimensional regime: $n \to \infty \quad d \to \infty$ p dimensions
  $\alpha \equiv n/d = \Theta(1)$ n samples

QUESTIONS OF INTEREST:

For a given task and a number of data samples:

- What is the best information-theoretically achievable test error?
- What is the best efficiently achievable test error?

BAYES-OPTIMAL GENERALIZATION IN THE TEACHER-STUDENT SETTING

Posterior probability distribution:

$$P(W|y,X) = \frac{1}{Z(y,X)} P_W(W) \prod_{\mu=1}^{n} P_{\text{out}}(y_\mu | X_\mu, W)$$

where $P_{\text{out}}(y_\mu | X_\mu, W) = \delta(y_\mu - f_W(X_\mu))$

$W = \{w_1, w_2, \ldots, w_L\}$ $\quad$ $X_\mu \in \mathbb{R}^d$

- A new sample X_{new} is given. Bayes-optimal prediction of a new label: $\hat{y}_{\text{new}} = \mathbb{E}_{P(W|y,X)} \left[f_W(X_{\text{new}}) \right]$

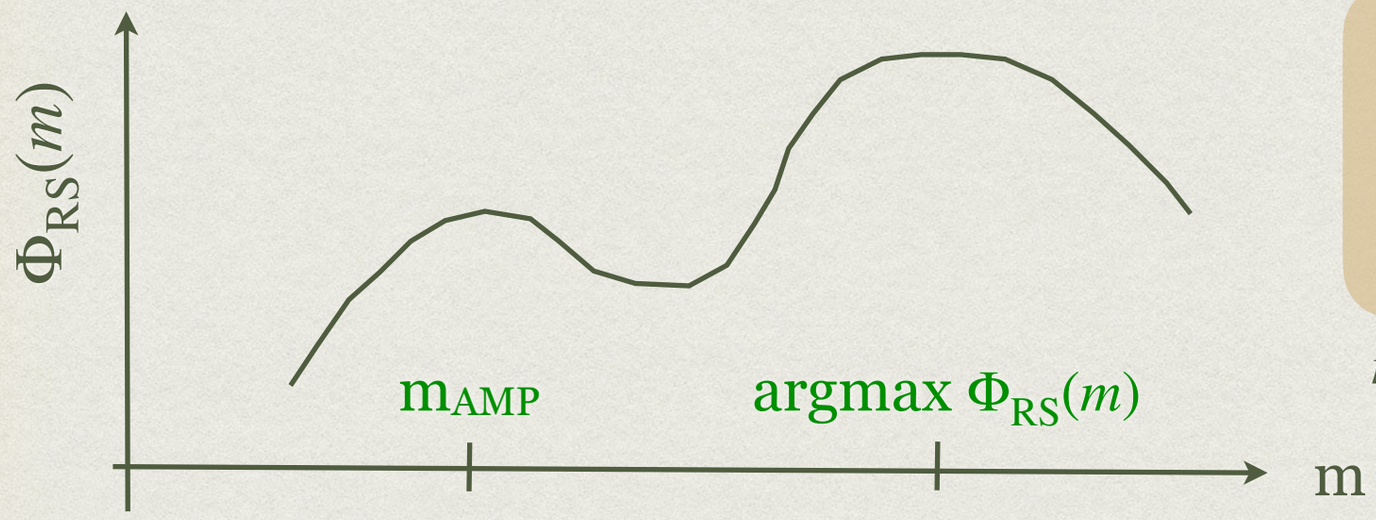
RESULT FOR THE OPTIMAL ERROR

Barbier, Krzakala, Macris, Miolane, LZ; arXiv:1708.03395, COLT'18, PNAS'19

- Assumptions: $X_{\mu i} \sim \mathcal{N}(0,1), w_i^* \sim P_w(w_i^*), y_\mu = \sigma\left(\sum_{i=1}^{d} X_{\mu i} w_i^*\right).$
- Main theorem in the limit $d \to \infty, n \to \infty, \alpha = n/d = \Theta(1)$:

A formula for so-called free entropy $\Phi_{\rm RS}(m)$ that implies both the test MMSE & test $\mathrm{MSE}_{\rm AMP}$ as:

$m \in \mathbb{R}$



$$\mathrm{MMSE} = \rho - \mathrm{argmax}\Phi_{\rm RS}(m)$$

$$\mathrm{MSE}_{\rm AMP} = \rho - m_{\rm AMP}$$

$m_{\rm AMP} \equiv$ local max of $\Phi_{\rm RS}(m)$ with lowest m

$\rho \equiv \mathbb{E}_{P_w}(w^2)$

RESULTS FOR THE OPTIMAL ERROR

Barbier, Krzakala, Macris, Miolane, LZ; arXiv:1708.03395, COLT'18, PNAS'19

Def. free entropy: $$\Phi = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}_{y,X} \log Z(y, X) \qquad \alpha = \frac{n}{d}$$

Theorem:

$$\Phi = \sup_{m} \inf_{\hat{m}} \Phi_{\mathrm{RS}}(m, \hat{m}) \qquad \Phi_{\mathrm{RS}}(m) = \inf_{\hat{m}} \Phi_{\mathrm{RS}}(m, \hat{m})$$

$$\Phi_{\mathrm{RS}}(m, \hat{m}) = \Phi_{P_w}(\hat{m}) + \alpha\Phi_{P_{\mathrm{out}}}(m; \rho) - \frac{m\hat{m}}{2}$$

where

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

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

$$w, w_0 \sim P_w \qquad z, v, \xi \sim \mathcal{N}(0,1) \qquad \rho = \mathbb{E}_{P_w}(w^2)$$

RESULTS FOR THE OPTIMAL ERROR

Barbier, Krzakala, Macris, Miolane, LZ; arXiv:1708.03395, COLT'18, PNAS'19

Def. free entropy: $$\Phi = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}_{y,X} \log Z(y, X) \qquad \alpha = \frac{n}{d}$$

Theorem:

$$\Phi = \sup_{m} \inf_{\hat{m}} \Phi_{\mathrm{RS}}(m, \hat{m}) \qquad \Phi_{\mathrm{RS}}(m) = \inf_{\hat{m}} \Phi_{\mathrm{RS}}(m, \hat{m})$$

$$\Phi_{\mathrm{RS}}(m, \hat{m}) = \Phi_{P_w}(\hat{m}) + \alpha\Phi_{P_{\mathrm{out}}}(m; \rho) - \frac{m\hat{m}}{2}$$

Corollary: Optimal test error

$$\mathscr{E}_{\mathrm{test}} = \mathbb{E}_{v,\xi}\left[\varphi(\sqrt{\rho}v)^2\right] - \mathbb{E}_{v,z,\xi}\left[\varphi\left(\sqrt{m^*}v + \sqrt{\rho - m^*}z\right)\right]^2$$

where m* is the extremizer of $\Phi_{\mathrm{RS}}(\mathrm{m})$.

$$\rho = \mathbb{E}_{P_w}(w^2)$$
$$v, z \sim \mathcal{N}(0,1)$$
$$\xi \sim P_\xi$$

Algorithm 2 Generalized Approximate Message Passing (G-AMP)

Input: y

Initialize: $\mathbf{a}^0$,$\mathbf{v}^0$, t = 1 $g^0_{\text{out},\mu}$

repeat

AMP Update of ω_μ, V_μ

$$
\begin{aligned}
V_\mu^t &\leftarrow \sum_i X_{\mu i}^2 v_i^{t-1} \\
\omega_\mu^t &\leftarrow \sum_i X_{\mu i} a_i^{t-1} - V_\mu^t g_{\text{out}}^{t-1}
\end{aligned}
$$

Variances and means of the pre-activations

AMP Update of Σ_i, R_i and $g_{\text{out},\mu}$

$$
\begin{aligned}
\Sigma_i^t &\leftarrow \left[-\sum_\mu X_{\mu i}^2 \partial_\omega g_{\text{out}}(\omega_\mu^t, y_\mu, V_\mu^t) \right]^{-1} \\
R_i^t &\leftarrow a_i^{t-1} + (\Sigma_i^{t+1})^{-1} \sum_\mu X_{\mu i} g_{\text{out}}(\omega_\mu^t, y_\mu, V_\mu^t)
\end{aligned}
$$

AMP Update of the estimated marginals a_i, v_i

$$
\begin{aligned}
a_i^{t+1} &\leftarrow f_a(\Sigma_i, R_i^{t+1},) \\
v_i^{t+1} &\leftarrow f_v(\Sigma_i, R_i^{t+1})
\end{aligned}
$$

Variances and means of the weights.

t $\leftarrow$ t + 1

until Convergence on **a**,**v**

output: a,**v**.

$$
f_a(\Sigma, R) = \frac{\int \mathrm{d}x\, x\, P_w(x)\, e^{-\frac{(x-R)^2}{2\Sigma}}}{\int \mathrm{d}x\, P_w(x)\, e^{-\frac{(x-R)^2}{2\Sigma}}}, \qquad f_v(\Sigma, R) = \Sigma \partial_R f_a(\Sigma, R)\,.
$$

$$
g_{\text{out}}(\omega, y, V) \equiv \frac{\int \mathrm{d}z P_{\text{out}}(y|z)\,(z-\omega)\, e^{-\frac{(z-\omega)^2}{2V}}}{V \int \mathrm{d}z P_{\text{out}}(y|z) e^{-\frac{(z-\omega)^2}{2V}}}\,.
$$

STATE EVOLUTION

Bolthausen'09; Bayati, Montanari'11; Javanmard, Montanari'13.

Define: $m^t \equiv \frac{1}{d}\sum_{i=1}^{d} w_i^* a_i^t$ then $\mathrm{MSE}(t) = \rho - m^t$

m^t in the AMP algorithm evolves as:

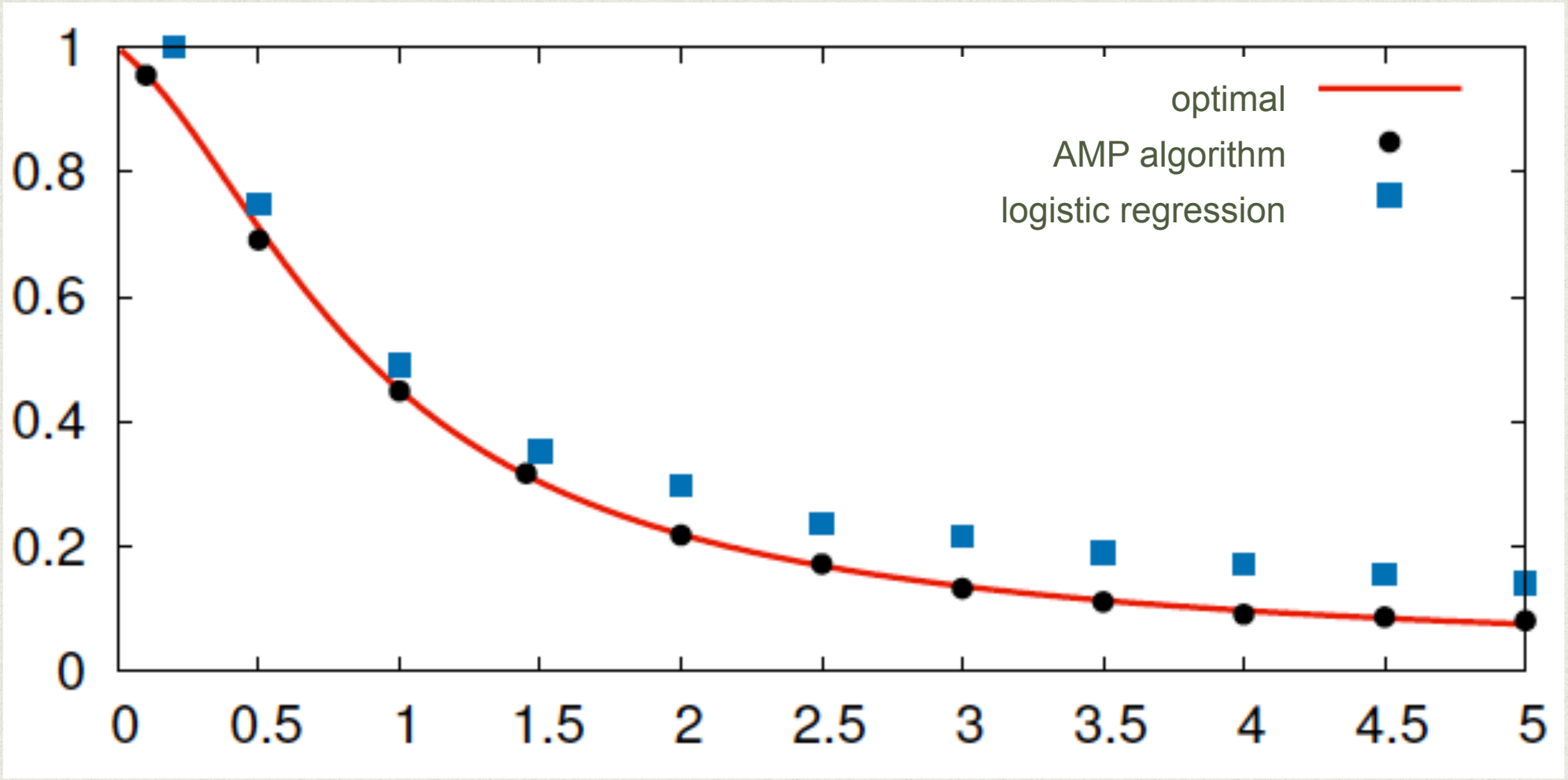
$$m^{t+1} = 2\partial_{\hat{m}}\Phi_{P_w}(\hat{m}^t) \qquad \hat{m}^t = 2\alpha\partial_m\Phi_{P_{\text{out}}}(m^t;\rho)$$

Recall the RS free entropy

$$\Phi_{\text{RS}}(m,\hat{m}) = \Phi_{P_w}(\hat{m}) + \alpha\Phi_{P_{\text{out}}}(m;\rho) - \frac{m\hat{m}}{2}$$

LEARNING CURVES

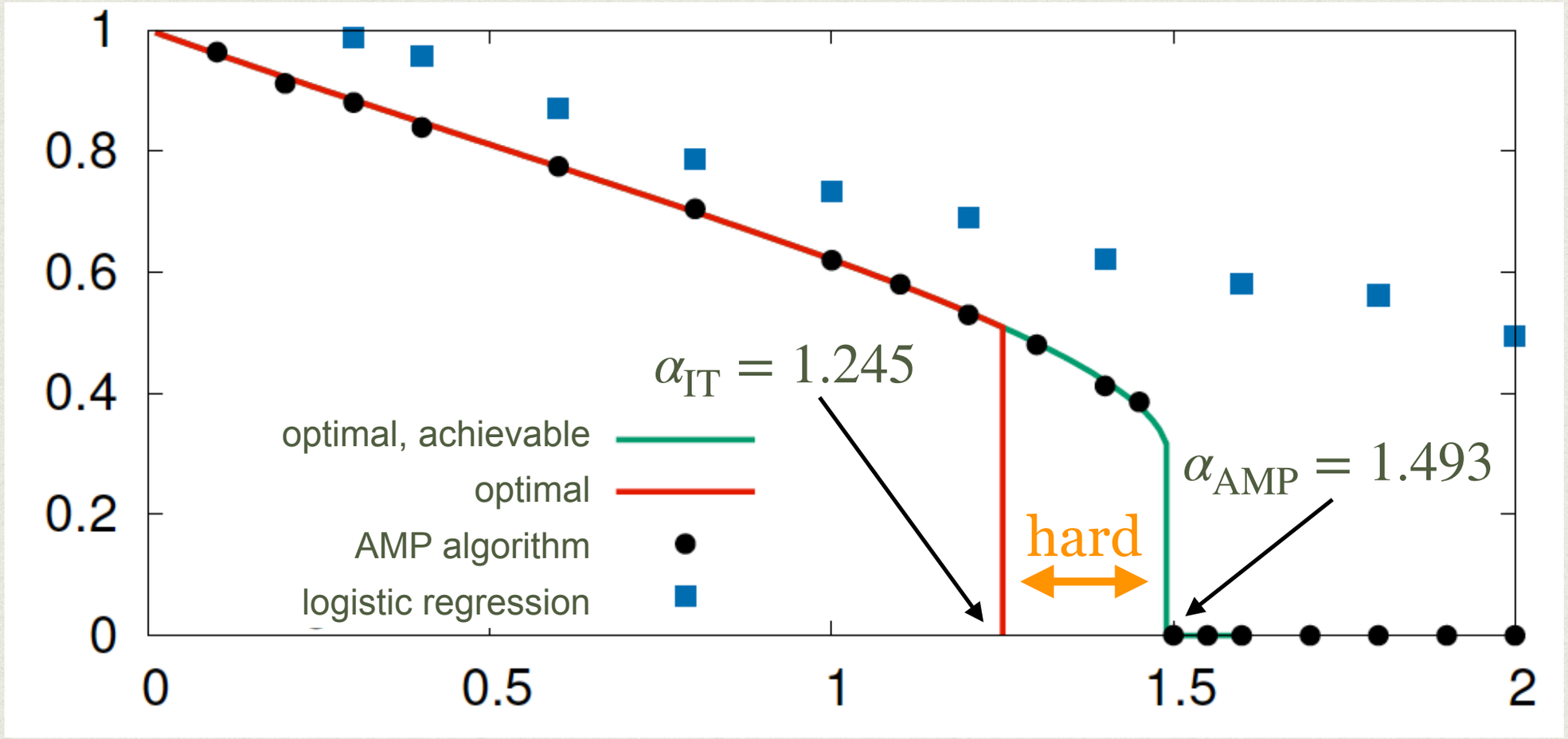
$\varphi(z) = \text{sign}(z)$ $\quad P_w = \mathcal{N}(0,1)$ $\quad n \to \infty,\ p \to \infty \quad n/p = \Omega(1)$



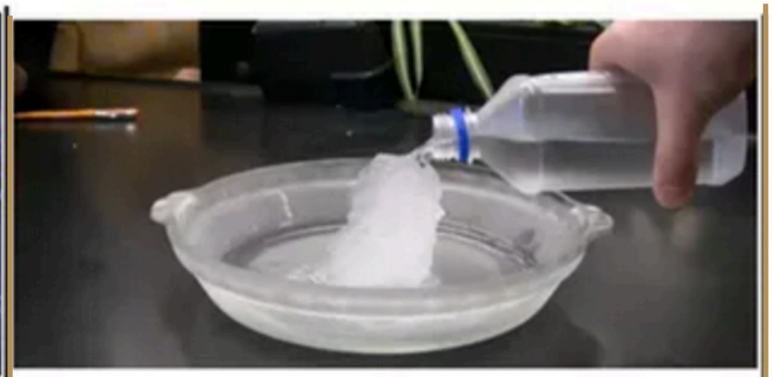
BINARY PERCEPTRON

Barbier, Krzakala, Macris, Miolane, LZ; arXiv:1708.03395, COLT'18, PNAS'19

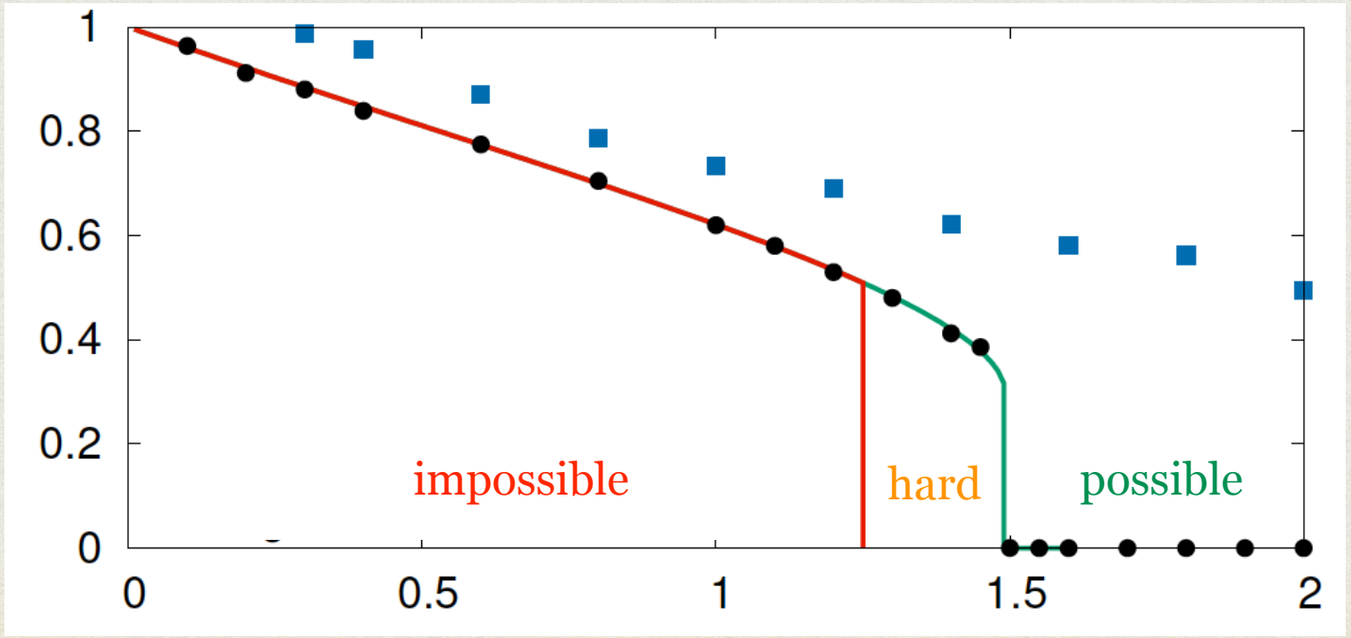
$$y_\mu = \mathrm{sign}\Big(\sum_{i=1}^{d} X_{\mu i} w_i^*\Big) \qquad X_{\mu i} \sim \mathcal{N}(0,1) \qquad w_i^* \in \{-1,+1\} \qquad n \to \infty \qquad d \to \infty \qquad n/d = \Theta(1)$$



SUDDEN EMERGENCE OF PERFECT LEARNING (= A PHASE TRANSITION)

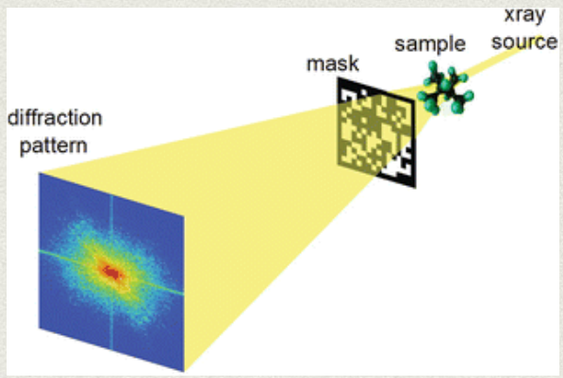


liquid	supercooled liquid	ice
impossible	computationally hard	possible



PHASE RETRIEVAL

- Broad range of applications in signal processing and imaging.



- Teacher-student setting with teacher having no hidden units, teacher's activation function is the absolute value.

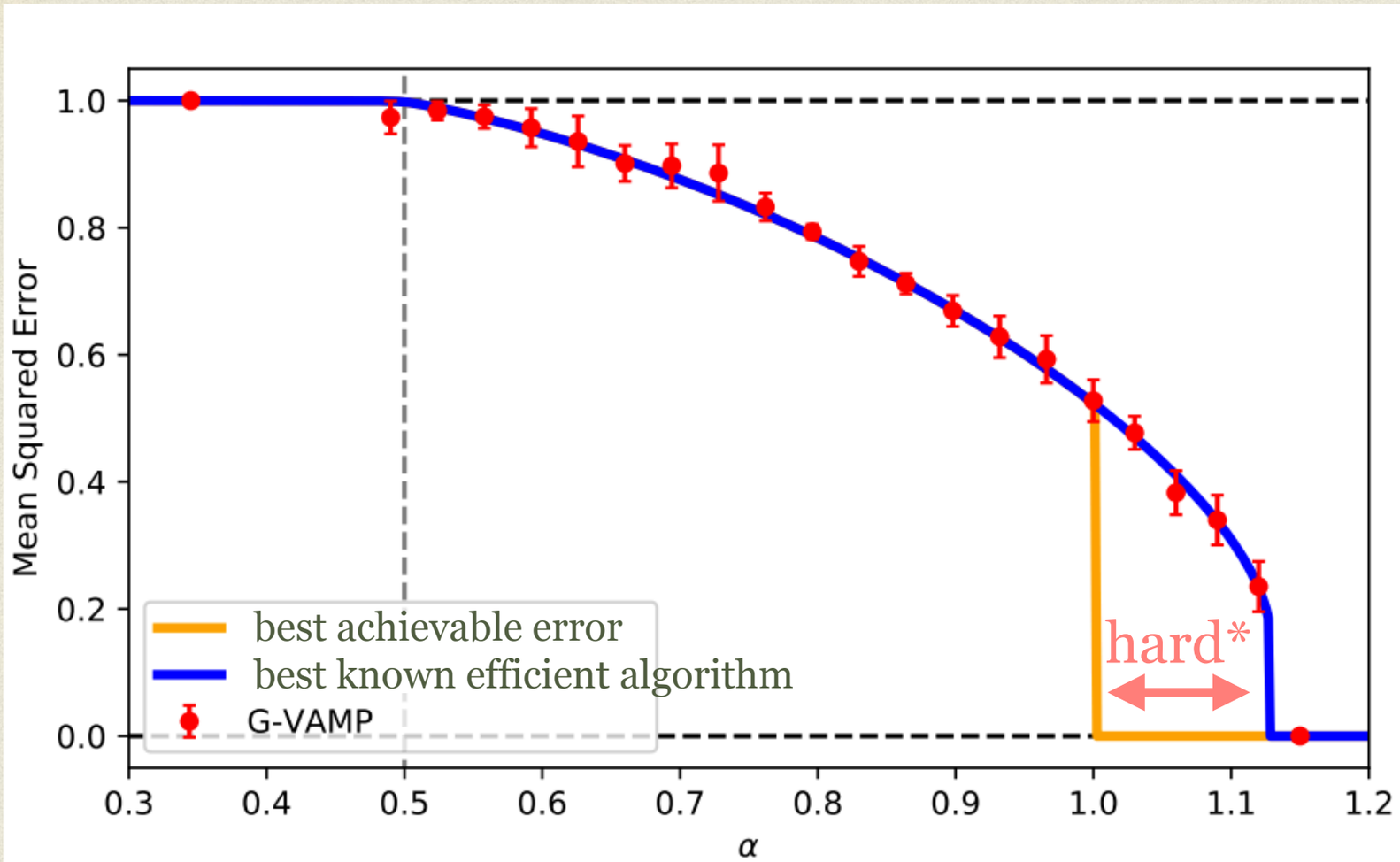
$$X_{\mu i} \sim \mathcal{N}(0,1/d) \qquad w_i^* \sim \mathcal{N}(0,1) \qquad \mu = 1,\ldots,n \qquad i = 1,\ldots,d$$

$$y_\mu = \left| \sum_{i=1}^{d} X_{\mu i} w_i^* \right|$$

Phase retrieval: Regression from training data $\{\mathbf{X}_\mu, y_\mu\}_{\mu=1}^{n}$

PHASE (SIGN) RETRIEVAL

Maillard, Loureiro, Krzakala, **LZ**, arXiv:2006.05288, NeurIPS'20.



$$y_\mu = \left| \sum_{i=1}^{d} X_{\mu i} w_i^* \right|$$

$$w_i^* \sim \mathcal{N}(0,1)$$

$$X_{\mu i} \sim \mathcal{N}(0,1/d)$$

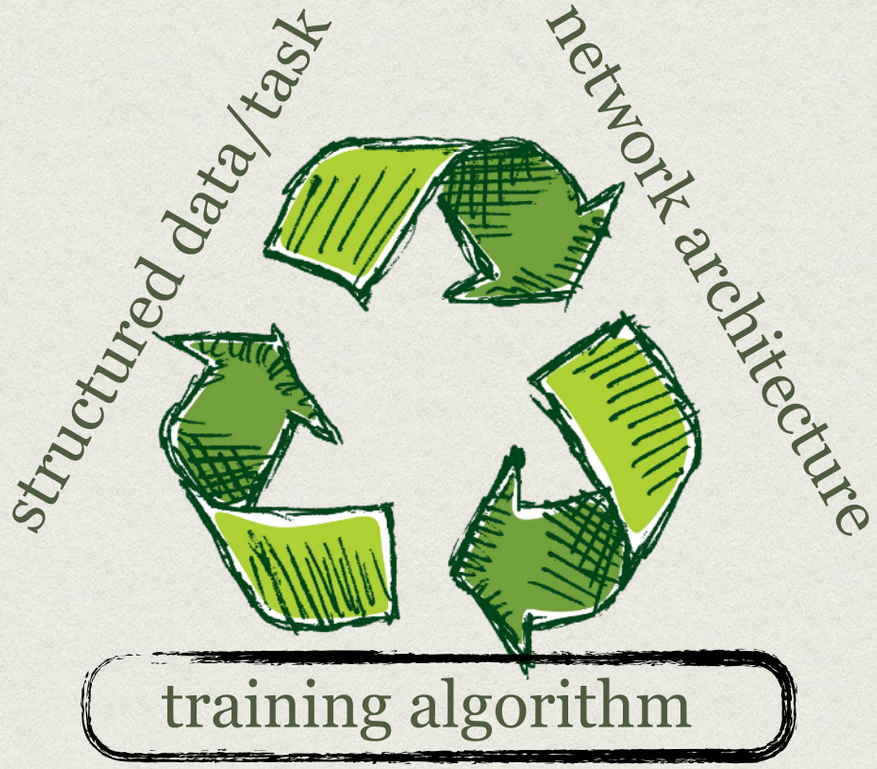
$$\alpha = \frac{n}{d}, n \to \infty$$

$\alpha_{\text{WR}} = 1/2$ weak-recovery, # of samples for generalisation better than random.

$\alpha_{\text{IT}} = 1$ # of samples for perfect generalisation for any algorithm.

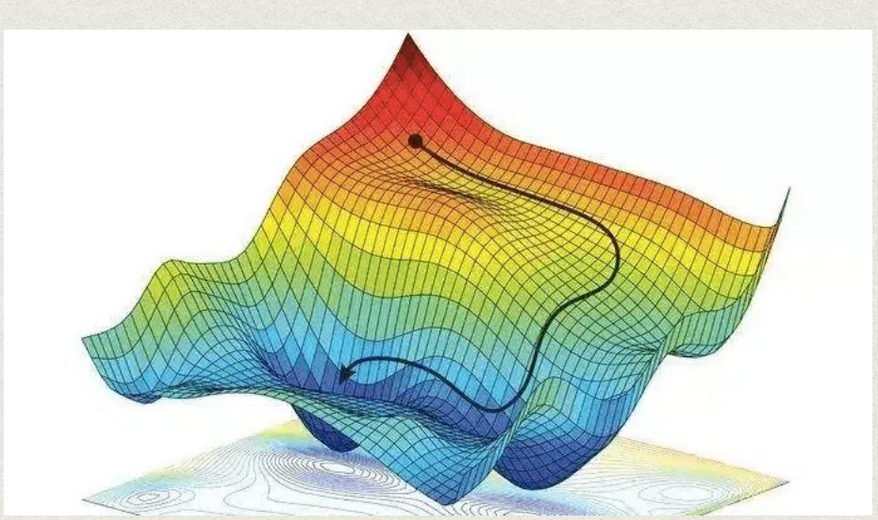
$\alpha_{\text{AMP}} = 1.13$ # of samples needed for perfect generalisation for AMP algorithms.

* LLL lattice-based reduction algorithm work in poly time for $\alpha > 1$ (Song, Zadik, Bruna'21)


structured data/task
network architecture
training algorithm

DEEP LEARNING USES GRADIENT DESCENT

(NOT MESSAGE PASSING)



GRADIENT DESCENT FOR PHASE RETRIEVAL

Loss function:

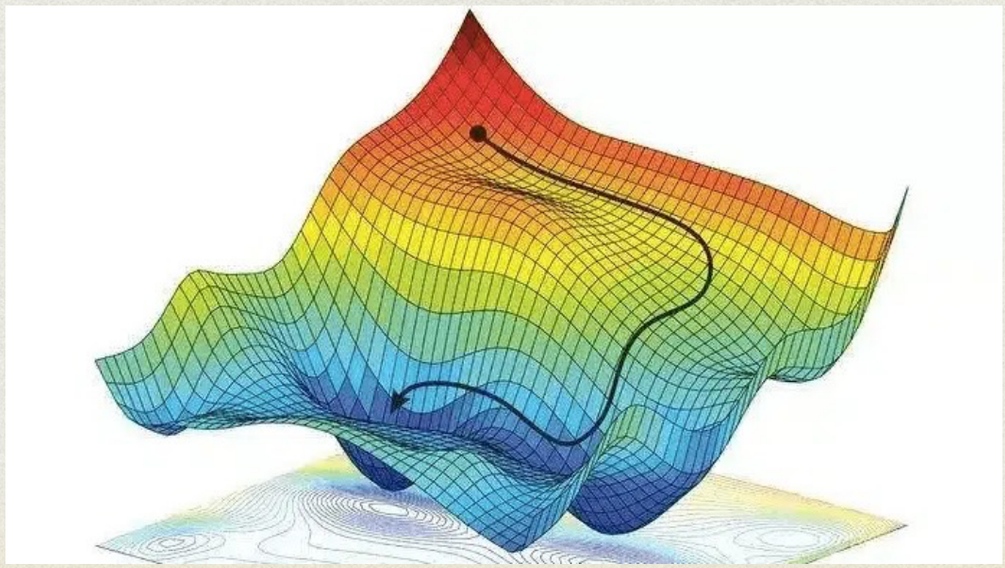
$$\mathscr{L}(\{w_i\}_{i=1}^p) = \sum_{\mu=1}^{n} \left[y_\mu^2 - \left(\sum_{i=1}^{d} X_{\mu i} w_i \right)^2 \right]^2$$

$$\text{where} \quad y_\mu = \left| \sum_{i=1}^{d} X_{\mu i} w_i^* \right|$$

$$X_{\mu i} \sim \mathcal{N}(0, 1/d) \qquad w_i^* \sim \mathcal{N}(0,1)$$

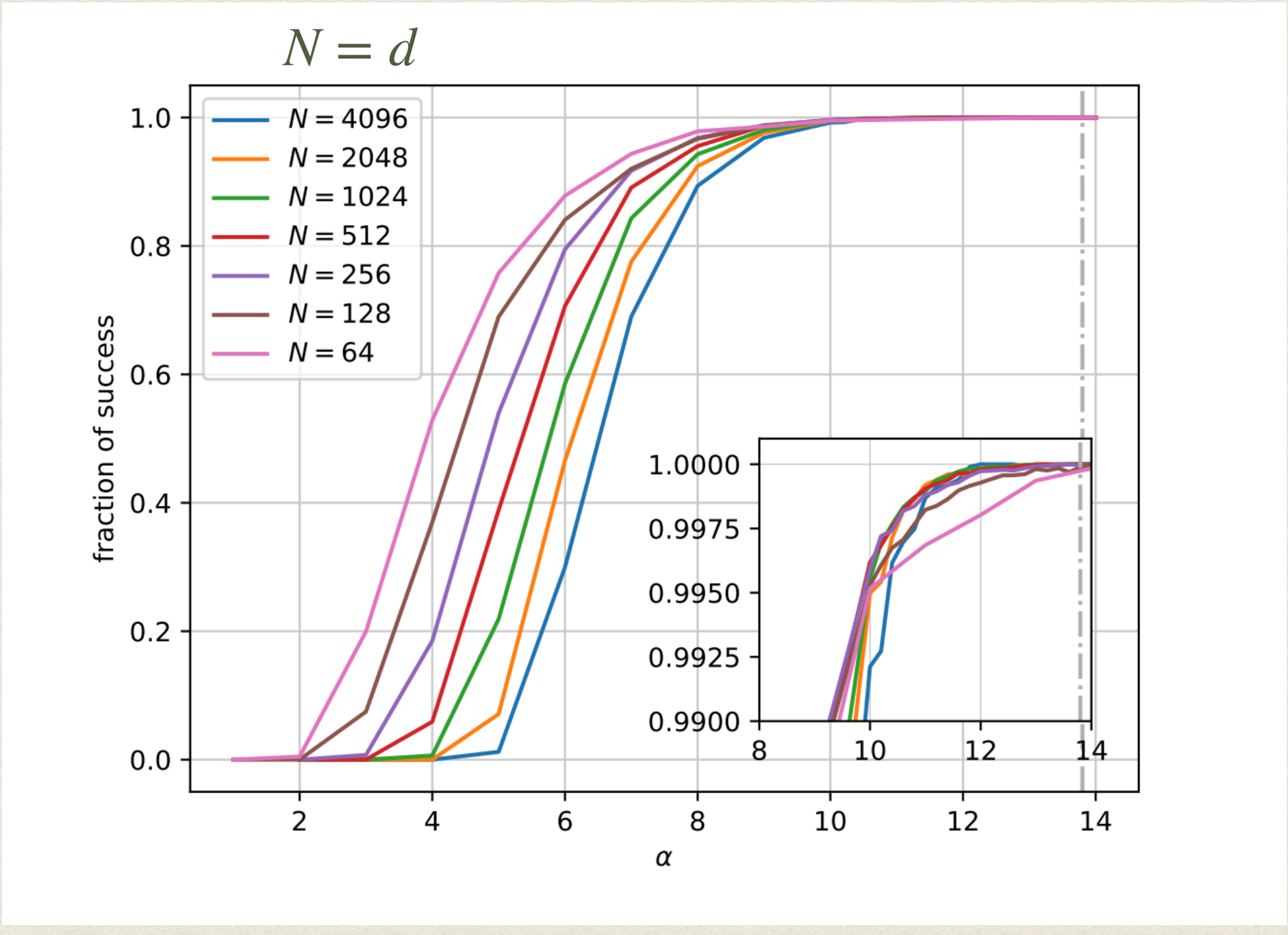
Gradient flow: $\dot{w}_i(t) = -\partial_{w_i} \mathscr{L}\left(\{w_j(t)\}_{j=1}^d \right)$

Initialisation: $w_i(0) \sim \mathcal{N}(0,1)$



GD IN PHASE RETRIEVAL

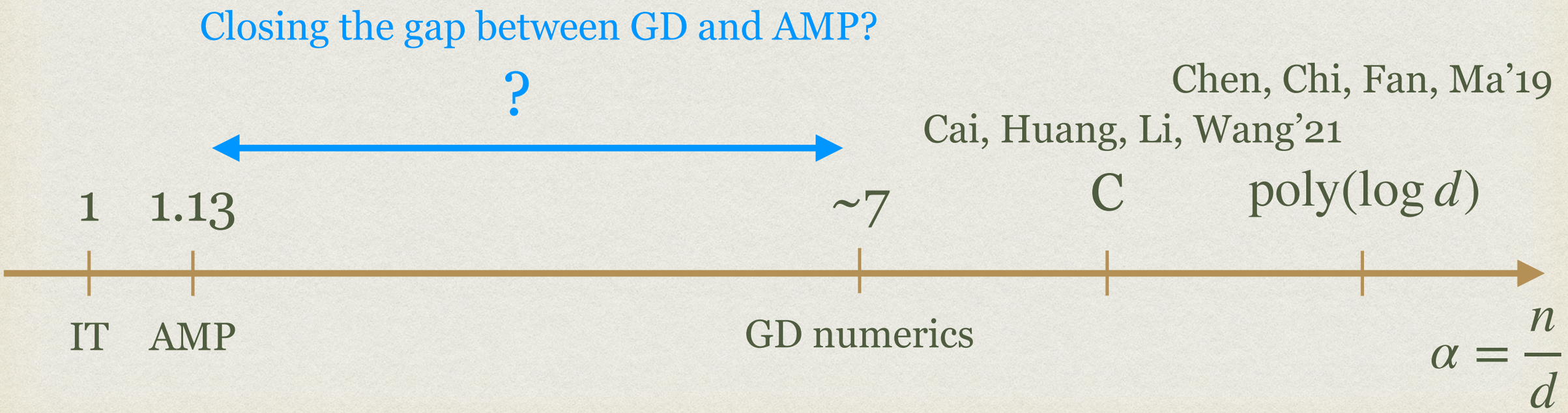
Sarao Mannelli, Biroli, Cammarota, Krzakala, Urbani, LZ; 2006.06997, NeurIPS'20.



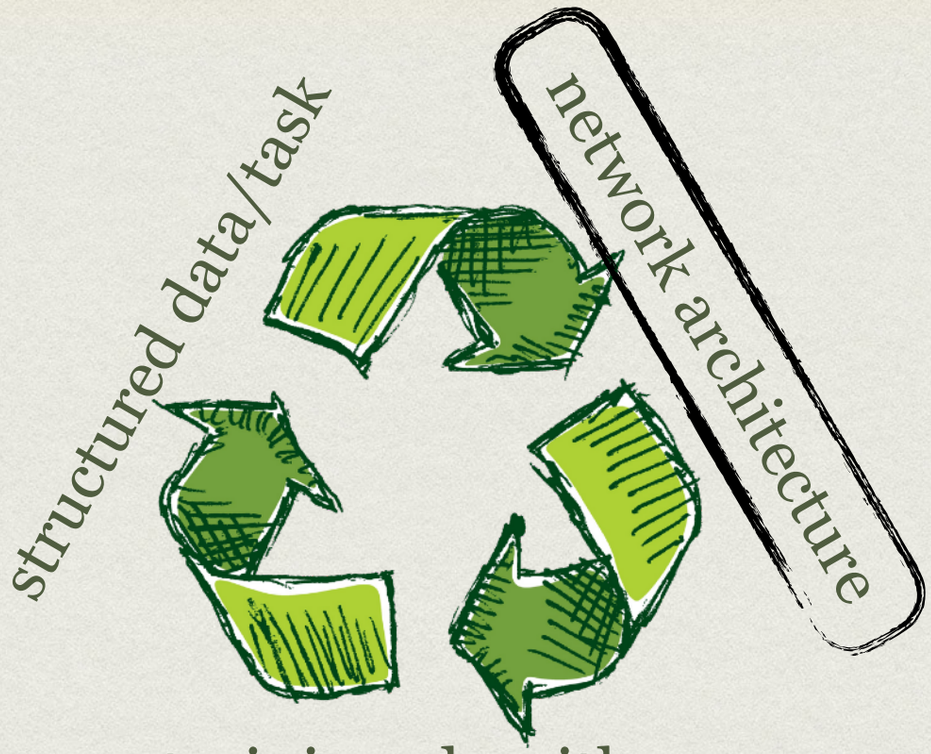
GRADIENT DESCENT IN PHASE RETRIEVAL

Sarao Mannelli, Biroli, Cammarota, Krzakala, Urbani, LZ; arxiv: 2006.06997, NeurIPS'20.

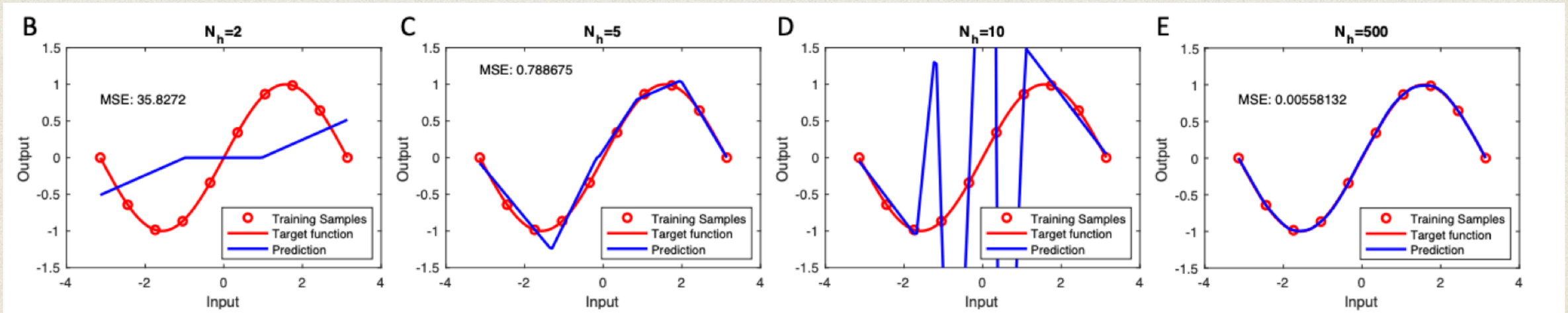
Sample complexity needed for exact recovery



Note: Kernel methods need $\alpha = O(d)$ to solve phase retrieval.


network architecture
structured data/task
training algorithm

DEEP LEARNING IS OVER-PARAMETRIZED



Saxe, Advani'17

DOUBLE-DESCENT

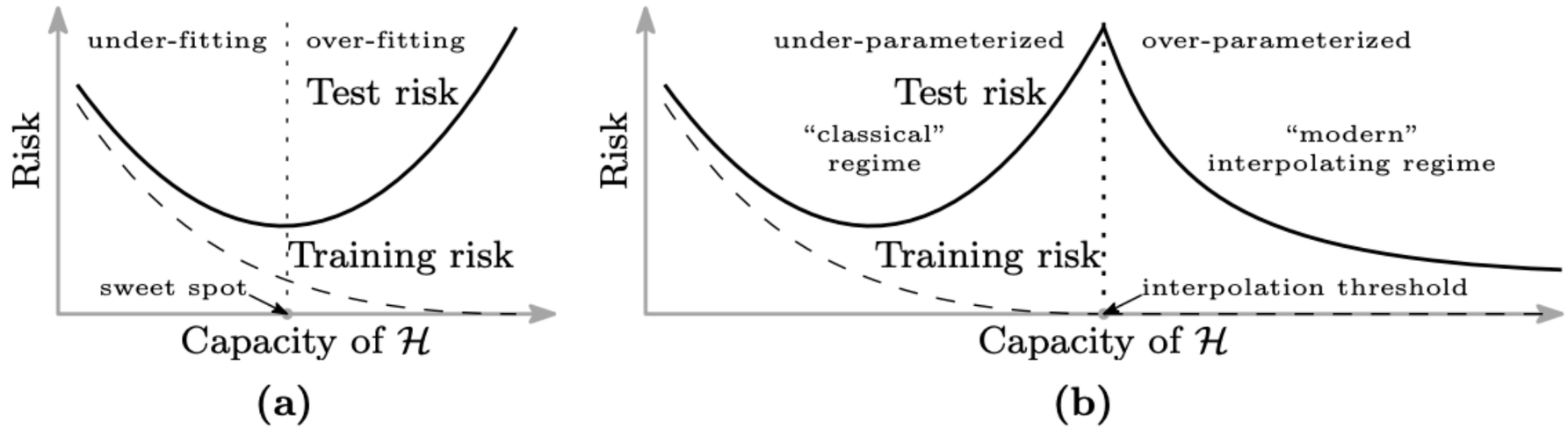


Figure 1: **Curves for training risk (dashed line) and test risk (solid line).** (**a**) The classical *U-shaped risk curve* arising from the bias-variance trade-off. (**b**) The *double descent risk curve*, which incorporates the U-shaped risk curve (i.e., the "classical" regime) together with the observed behavior from using high capacity function classes (i.e., the "modern" interpolating regime), separated by the interpolation threshold. The predictors to the right of the interpolation threshold have zero training risk.

Belkin et al. 2019

MORE ON DOUBLE DESCENT

- Generalisation error in learning with random features and the hidden manifold model, Gerace, Loureiro, Krzakala, Mézard, LZ, ICML'20.

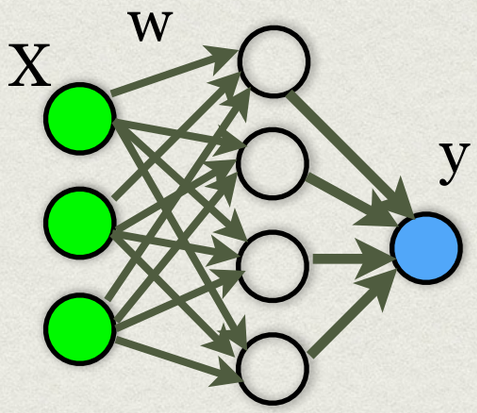
Paper n. 1 for flipped class.

OVER-PARAMETRIZED PHASE RETRIEVAL

Loss function: $$\mathscr{L}(\{w_{ia}\}_{i,a=1}^{d,m}) = \sum_{\mu=1}^{n}\left[y_\mu^2 - \frac{1}{m}\sum_{a=1}^{m}\left(\sum_{i=1}^{d} X_{\mu i} w_{ia}\right)^2\right]^2$$

where $$y_\mu = \left|\sum_{i=1}^{d} X_{\mu i} w_i^*\right|$$

$$X_{\mu i} \sim \mathcal{N}(0,1/d) \qquad w_i^* \sim \mathcal{N}(0,1)$$



Wide (m>d) over-parametrised two-layer neural network

Gradient flow: $$\dot{w}_{ia}(t) = -\,\partial_{w_{ia}}\mathscr{L}\left(\{w_{jb}(t)\}_{j,b=1}^{d,m}\right)$$

Initialisation: $$w_{ia}(0) \sim \mathcal{N}(0,1)$$

OVER-PARAMETRISED LANDSPACE

Sarao Mannelli, Vanden-Eijnden, LZ, arxiv:2006.15459, NeurIPS'20.

Theorem 3.1 (Single unit teacher). *Consider a teacher with $m^* = 1$ and a student with $m \geq d$ hidden units respectively, so that A^* has rank 1 and A has full rank. Given a data set $\{\boldsymbol{x}_k\}_{k=1}^n$ with each $\boldsymbol{x}_k \in \mathbb{R}^d$ drawn independently from a standard Gaussian, denote by $\mathcal{M}_{n,d}$ the set of minimizer of the empirical loss constructed with $\{\boldsymbol{x}_k\}_{k=1}^n$ over symmetric positive semidefinite matrices A, i.e.*

$$\mathcal{M}_{n,d} = \left\{ A = A^T, \text{ positive semidefinite such that } E_n(A) = 0 \right\}. \tag{10}$$

Set $n = \lfloor \alpha d \rfloor$ for $\alpha \geq 1$ and let $d \to \infty$. Then

$$\lim_{d\to\infty} \mathbb{P}\left(\mathcal{M}_{\lfloor \alpha d \rfloor, d} \neq \{A^*\} \right) = 1 \qquad \text{if } \alpha \in [0, 2] \tag{11}$$

whereas

$$\lim_{d\to\infty} \mathbb{P}\left(\mathcal{M}_{\lfloor \alpha d \rfloor, d} = \{A^*\} \right) > 0 \qquad \text{if } \alpha \in (2, \infty). \tag{12}$$

$$A(t) = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{w}_i(t) \boldsymbol{w}_i^T(t), \quad A^* = \frac{1}{m^*} \sum_{i=1}^{m^*} \boldsymbol{w}_i^* (\boldsymbol{w}_i^*)^T,$$

OVER-PARAMETRIZED PHASE RETRIEVAL

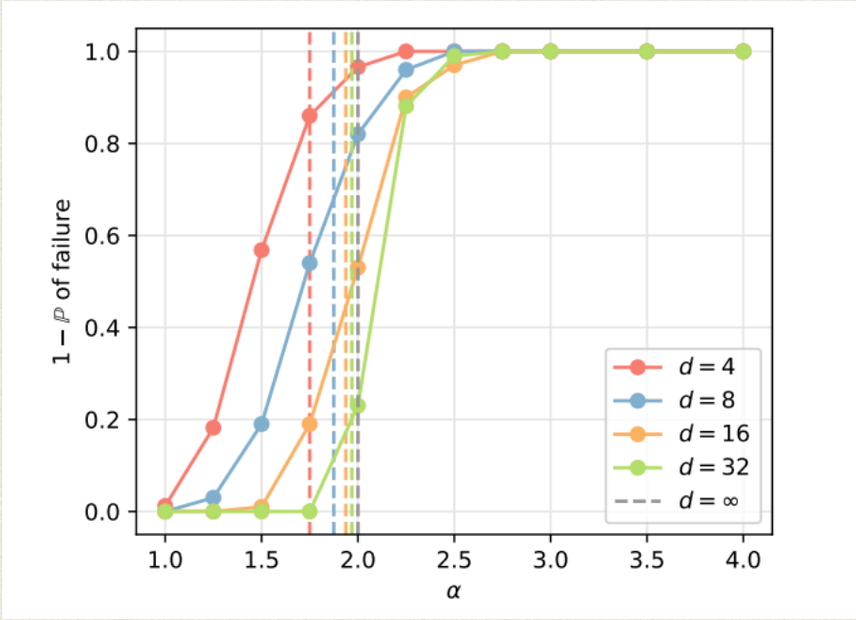
Sarao Mannelli, Vanden-Eijnden, LZ, NeurIPS'20, 2006.15459

Theorem 4.1. *Let $\{\boldsymbol{w}_i(t)\}_{i=1}^m$ be the solution to (3) for the initial data $\{\boldsymbol{w}_i(0)\}_{i=1}^m$. Assume that $m \geq d$ and each $\boldsymbol{w}_i(0)$ is drawn independently from a distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. Then*

$$A = \frac{1}{m}\sum_{i=1}^m \boldsymbol{w}_i(t)\boldsymbol{w}_i^T(t) \to A_\infty = \frac{1}{m}\sum_{i=1}^m \boldsymbol{w}_i^\infty(\boldsymbol{w}_i^\infty)^T \quad as \;\; t\to\infty \tag{15}$$

and A_∞ is a global minimizer of the empirical loss, i.e.

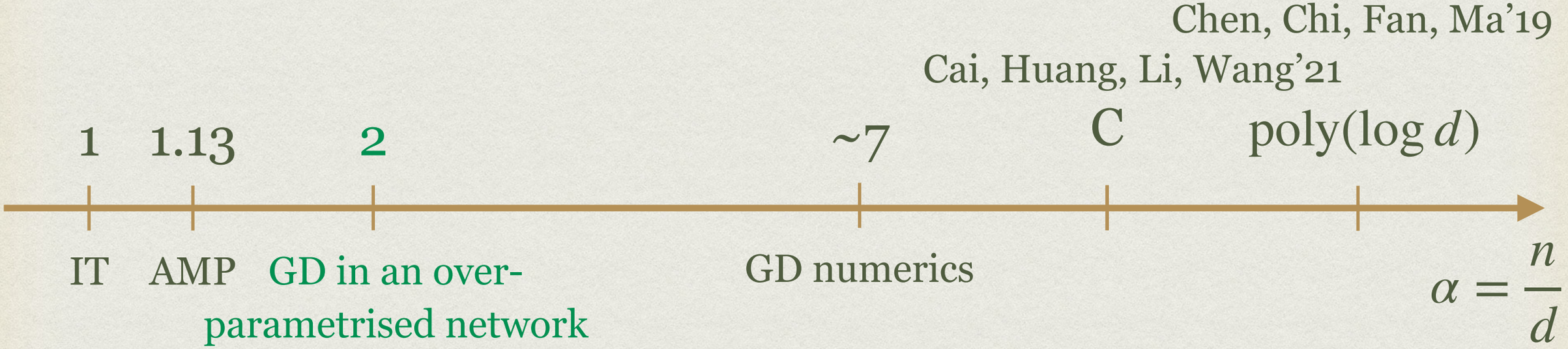
$$E_n(A_\infty) = 2L_n(\boldsymbol{w}_1^\infty, \dots, \boldsymbol{w}_n^\infty) = 0. \tag{16}$$

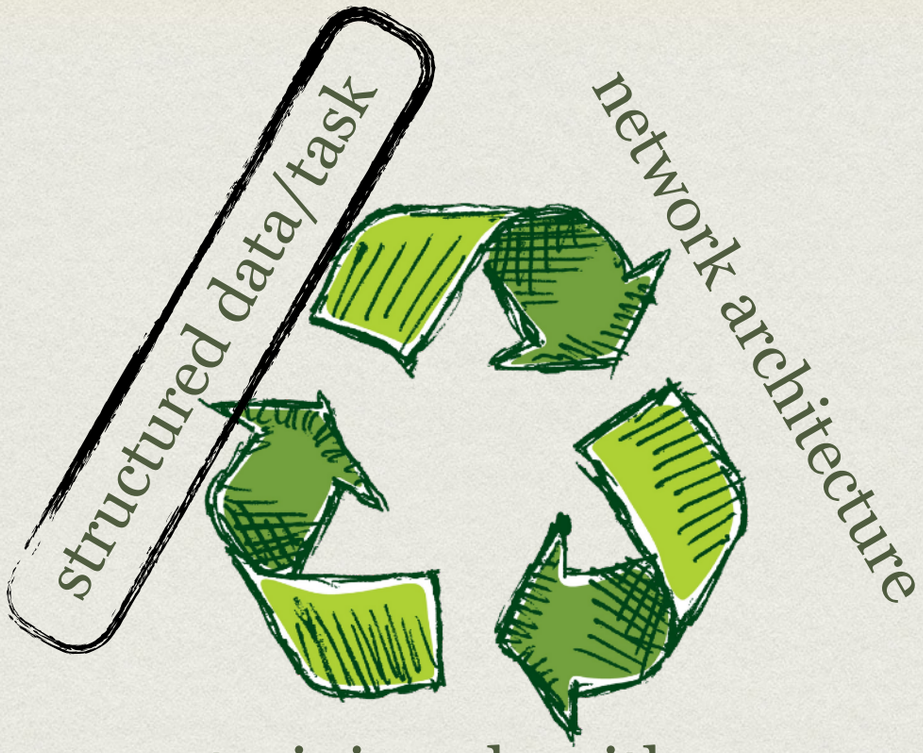


OVER-PARAMETRIZED PHASE RETRIEVAL

Sarao Mannelli, Vanden-Eijnden, LZ, arxiv:2006.15459, NeurIPS'20.

Over-parametrised neural network trained by gradient descent needs fewer samples to solve phase retrieval




structured data/task
network architecture
training algorithm

Teacher/target functions considered:

- So far: no hidden units, $d \to \infty, n \to \infty, \alpha = n/d = \Theta(1)$.
  Generalised linear model, single index model.

- $\Theta(1)$ hidden units, $d \to \infty, n \to \infty, \alpha = n/d = \Theta(1)$.
  Committee machine, multi-index model.

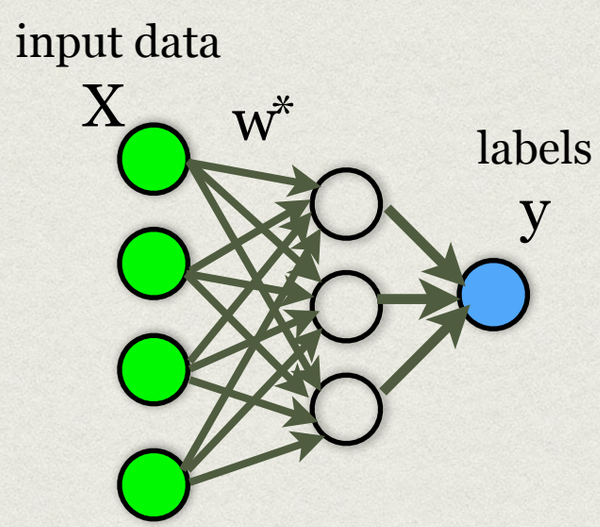
- $\Theta(d)$ hidden units, $d \to \infty, n \to \infty, \alpha = n/d^2 = \Theta(1)$.
  Extensive width.

EXTENSIVE WIDTH PHASE RETRIEVAL

Maillard, Troiani, Simon, Krzakala, **LZ**. arXiv:2408.03733

Teacher/target function with quadratic activation:

$$y_\mu = \frac{1}{m}\sum_{a=1}^{m}\left(\sum_{i=1}^{d} X_{\mu i} w^*_{ia}\right)^2 \qquad w^*_{ia} \sim \mathcal{N}(0,1) \qquad X_{\mu i} \sim \mathcal{N}(0,1/d)$$

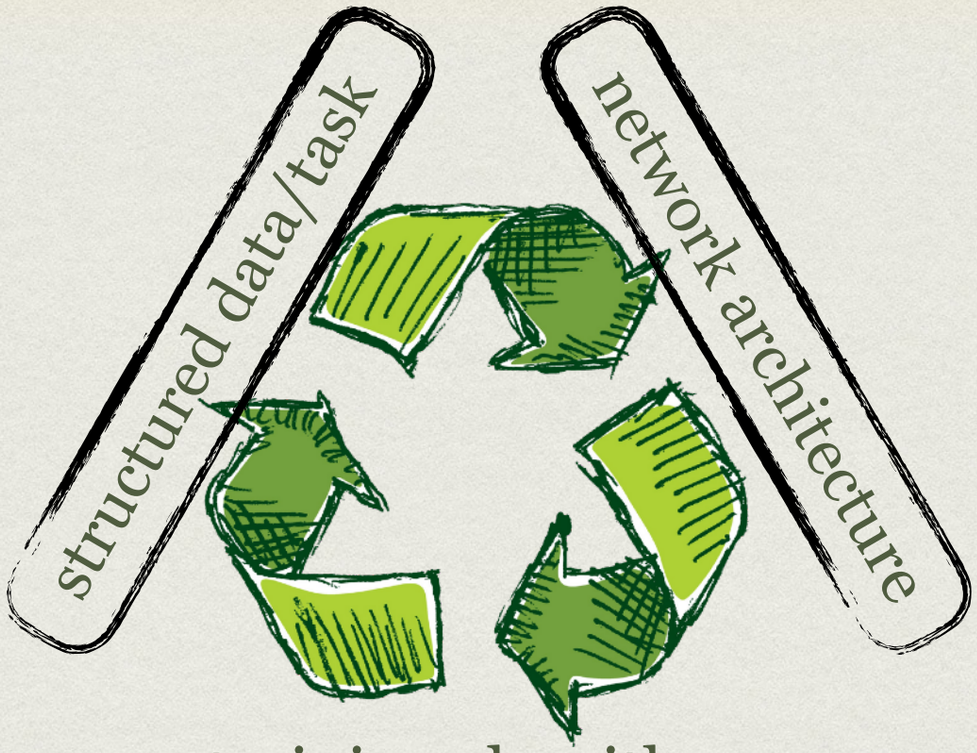


Paper n. 2 for flipped class.

Solvable in the limit:

$$\frac{m}{d} = \kappa, \frac{n}{d^2} = \alpha, d \to \infty, \kappa, \alpha = \Theta(1)$$

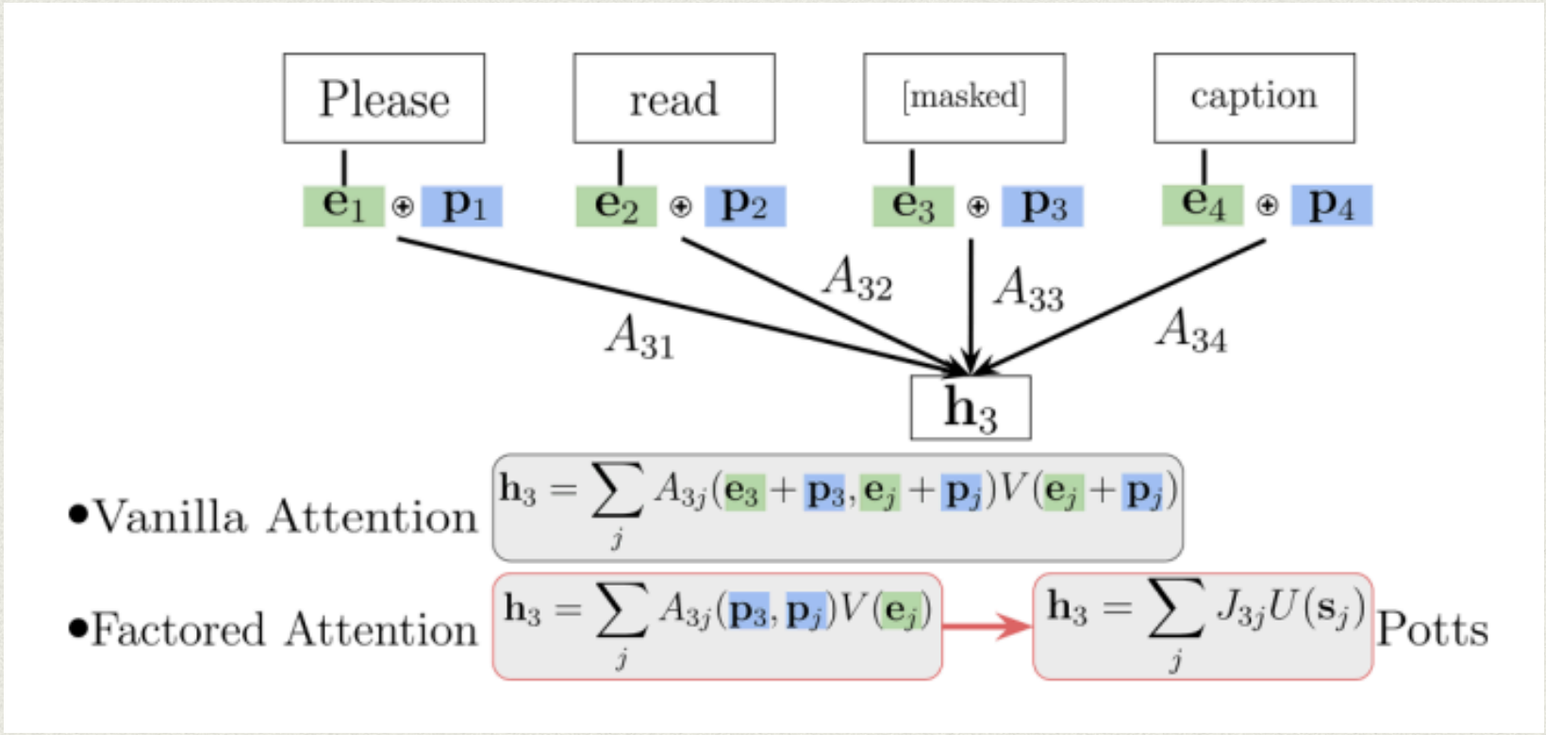
AND WHAT ABOUT LEARNING FROM SEQUENCES (AS LLM/TRANSFORMERS DO)?


structured data/task
network architecture
training algorithm

PHYSICAL REVIEW RESEARCH **6**, 023057 (2024)

Mapping of attention mechanisms to a generalized Potts model

Riccardo Rende, Federica Gerace, Alessandro Laio, and Sebastian Goldt*
Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy



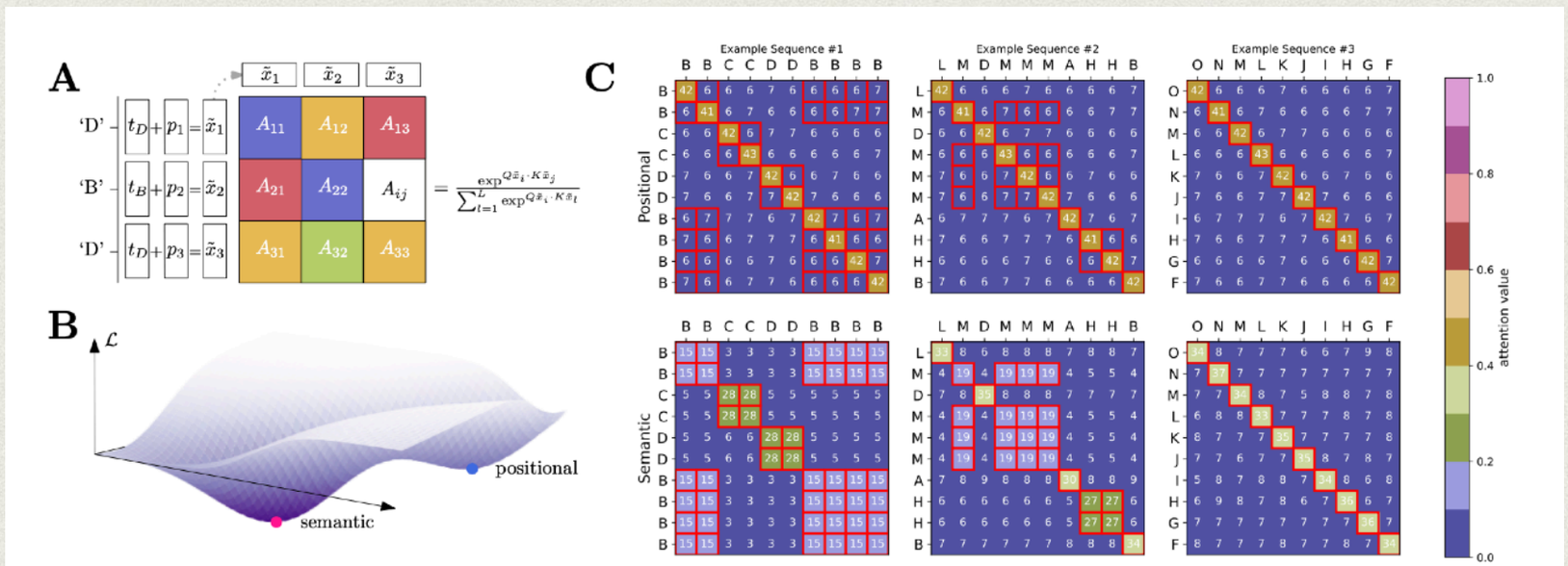
Paper n. 3 for flipped class.

A Phase Transition between Positional and Semantic Learning in a Solvable Model of Dot-Product Attention

Hugo Cui[1], Freya Behrens[1], Florent Krzakala[2], and Lenka Zdeborová[1]

[1]Statistical Physics Of Computation laboratory, EPFL, Switzerland
[2]Information Learning & Physics laboratory, EPFL, Switzerland

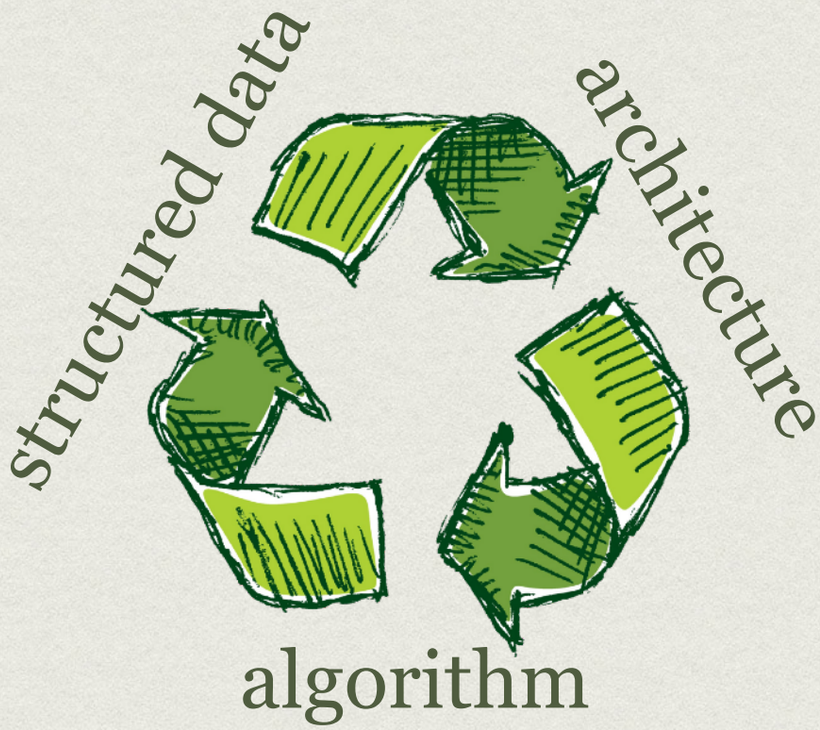
arxiv:2402.03902



Paper n. 4 for flipped class.

CONCLUSION

Physics has many useful tools applicable to understand machine learning / deep learning / AI.



BONUS/OLD

CAN WE HELP WITH PHYSICS?

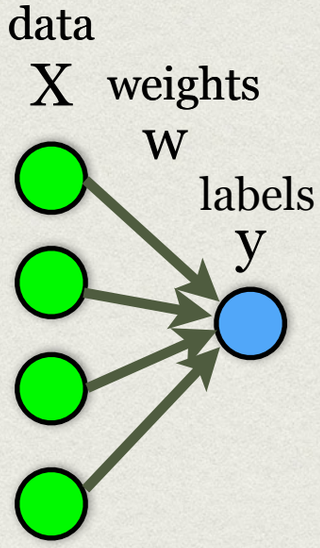
What is a good model to understand deep learning?

We aim to reproduce the salient behaviours of the real system.

Iterative process of improving the model.

SIMPLEST NEURAL NETWORK

Single layer neural network = perceptron = generalized linear regression.



Given (X,y) find w such that

$$y_\mu = \varphi\left(\sum_{i=1}^{p} X_{\mu i} w_i\right) \qquad \begin{matrix} i = 1,\ldots,p \\ \mu = 1,\ldots,n \end{matrix}$$

(noisy) activation function

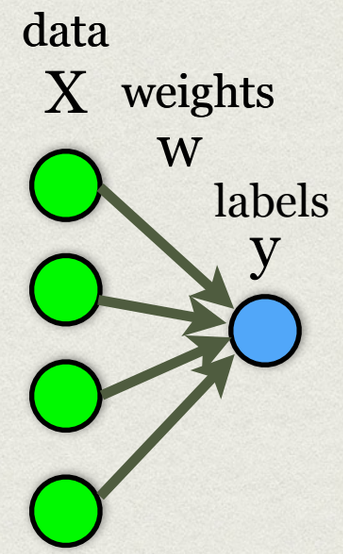
p dimensions.

n samples/ data points

TEACHER-STUDENT MODEL

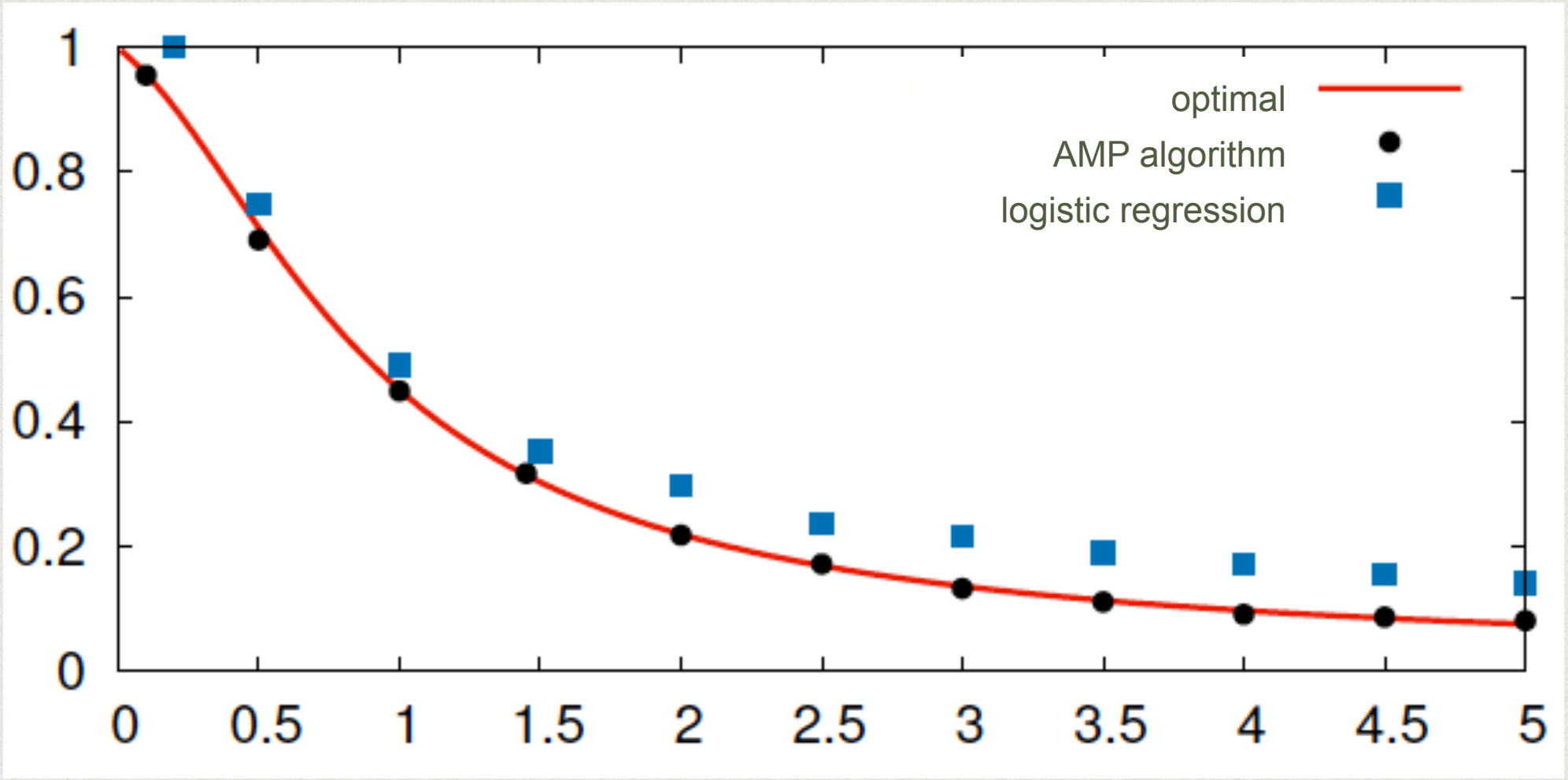
Gardner, Derrida'89, Gyorgyi'90

- Take random iid Gaussian $X_{\mu i}$ and random iid w_i^* from P_w
- Create $$y_\mu = \varphi\left(\sum_{i=1}^{p} X_{\mu i} w_i^*\right)$$
- Goal: Compute the best possible generalisation error achievable with n samples of dimension p.
- High-dimensional regime: $$n \to \infty$$ $$p \to \infty$$ $$n/p = O(1)$$



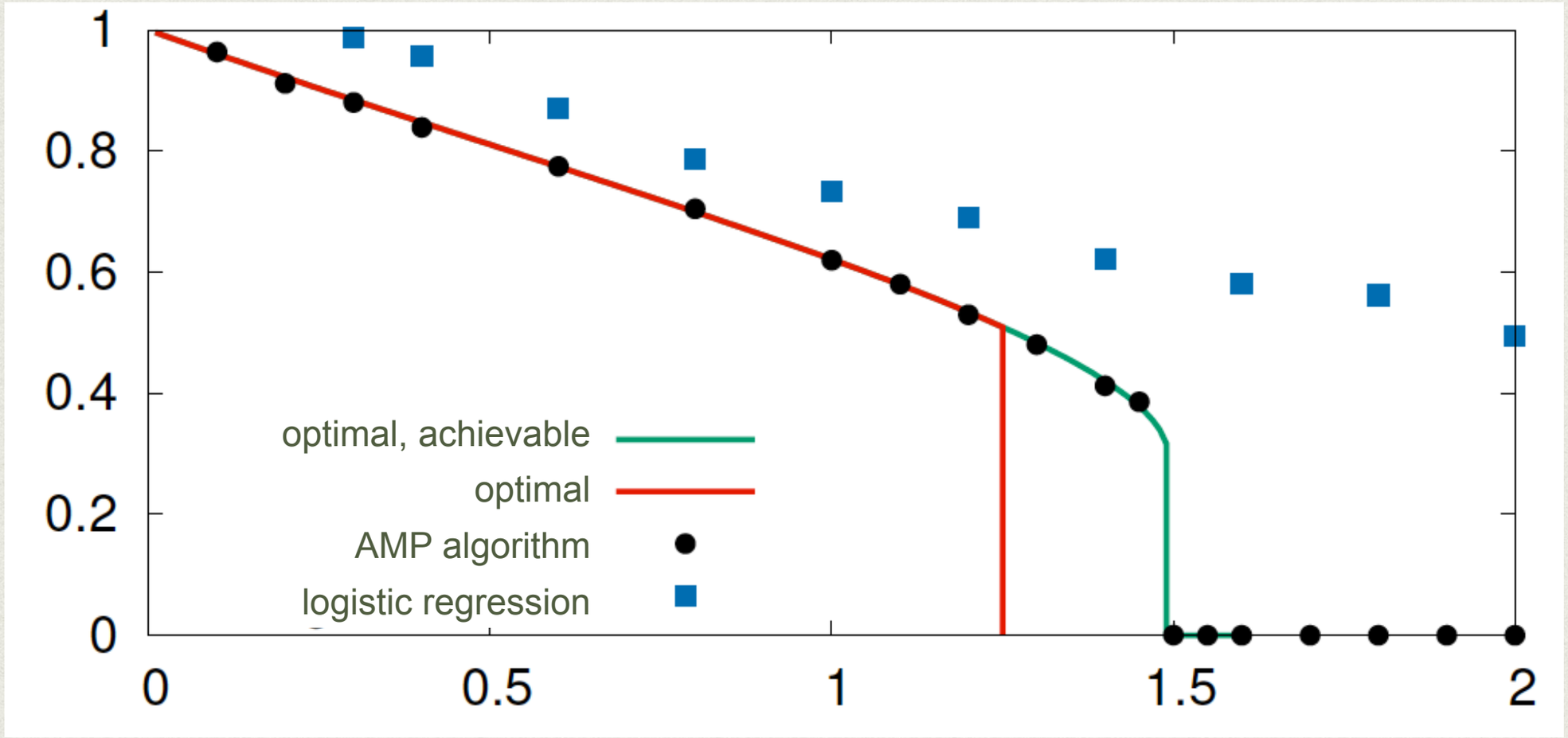
LEARNING CURVES

$\varphi(z) = \text{sign}(z)$ $\quad P_w = \mathcal{N}(0,1)$ $\quad n \to \infty,\ p \to \infty \quad n/p = \Omega(1)$



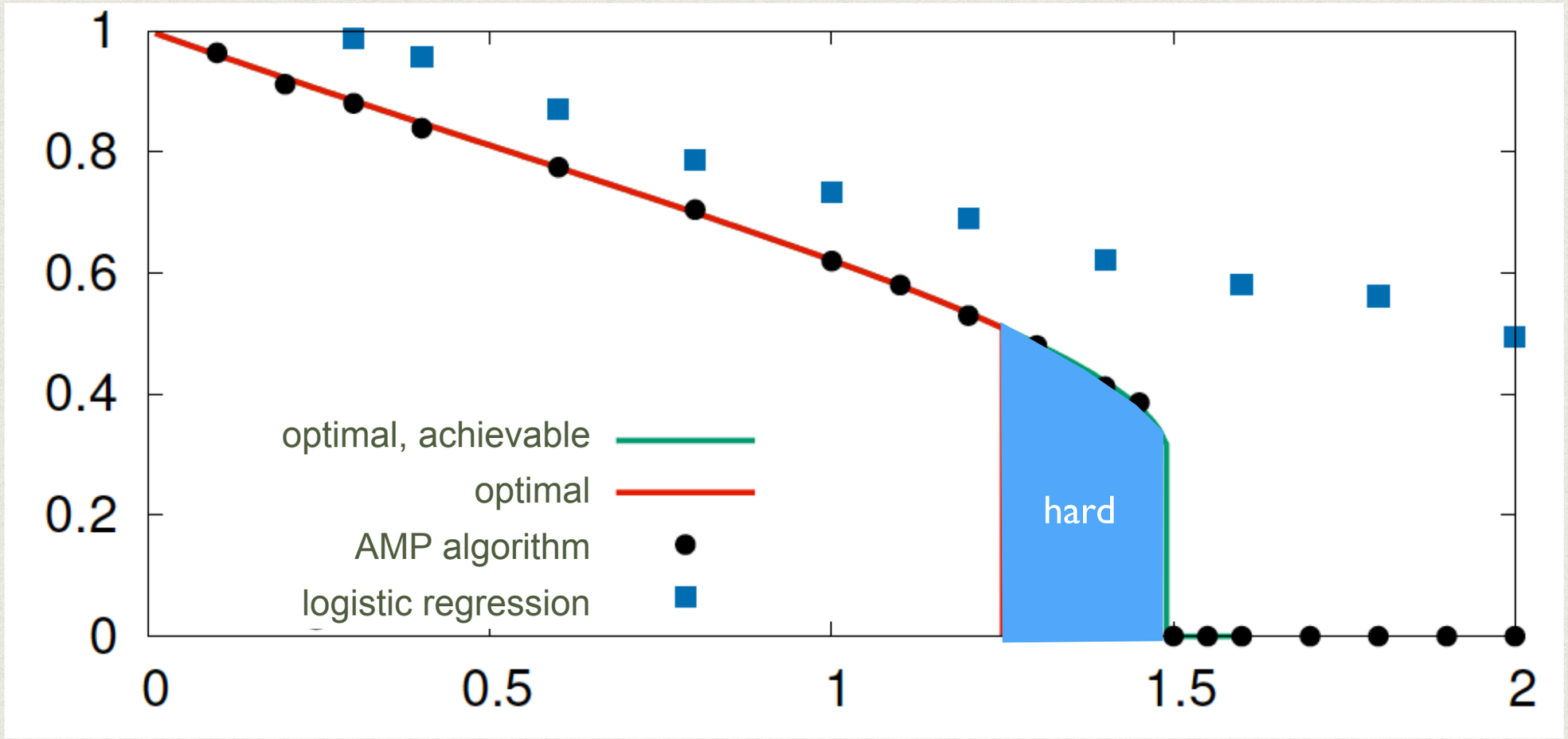
PHASE TRANSITIONS

$\varphi(z) = \text{sign}(z)$ $\quad w_i \in \{-1, +1\}$ $\quad \begin{matrix} n \to \infty \\ p \to \infty \end{matrix}$ $\quad n/p = \Omega(1)$



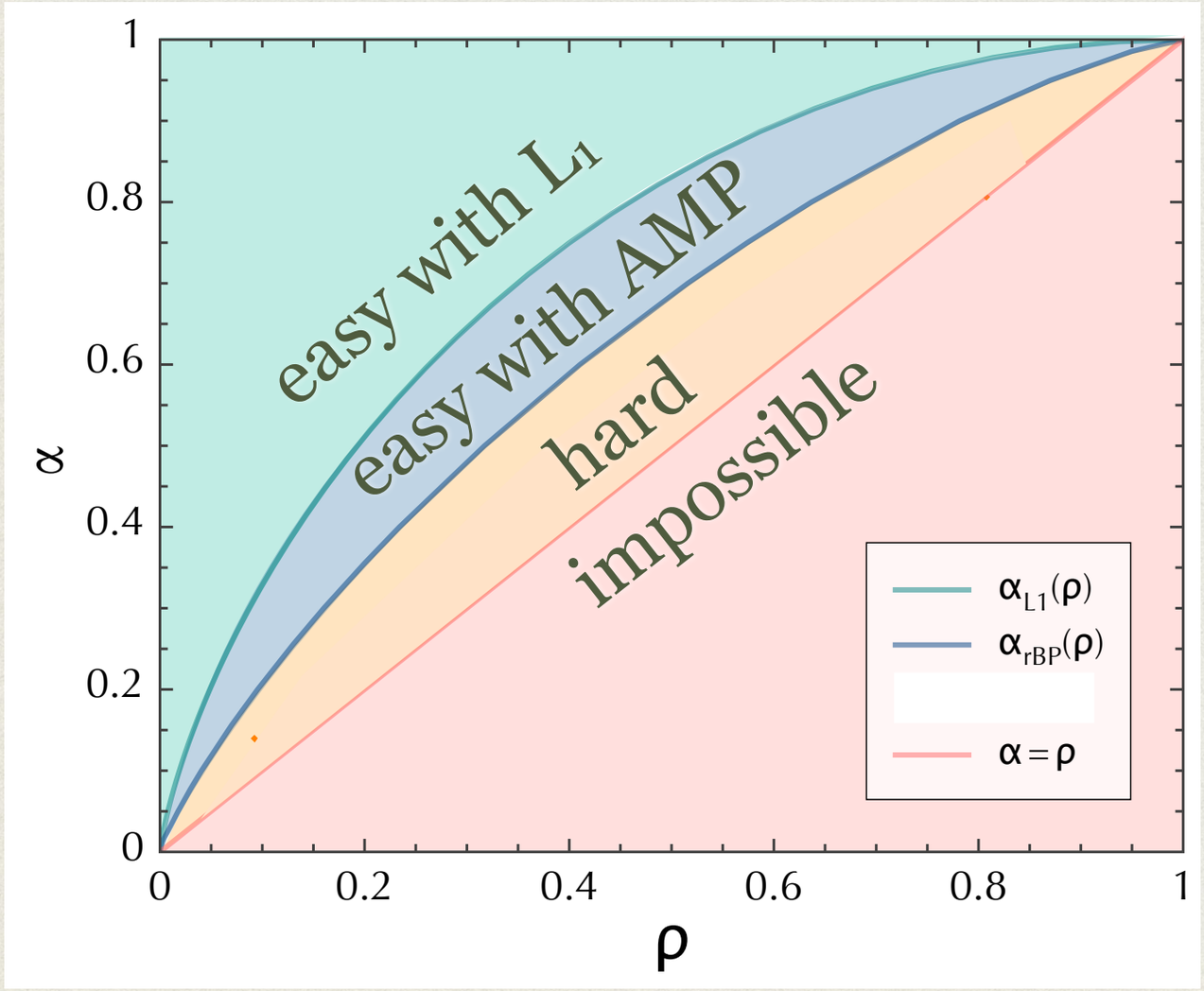
PHASE TRANSITIONS

$\varphi(z) = \text{sign}(z)$ $\quad w_i \in \{-1, +1\}$ $\quad \begin{matrix} n \to \infty \\ p \to \infty \end{matrix}$ $\quad n/p = \Omega(1)$



COMPRESSED SENSING

$\varphi(z) = z$ $\quad P_w(w_i) = (1-\rho)\delta(w_i) + \rho\mathcal{N}(w_i;0,1)$ $\quad \begin{matrix} n \to \infty \\ p \to \infty \end{matrix} \quad n/p = \Omega(1)$



Easy/hard threshold = spinodal of a 1st order phase transition.

Freedom in the design of X.

Spatial coupling.

TEACHER-STUDENT GLM

Paper n. 1 for flipped class.

Optimal errors and phase transitions in high-dimensional generalized linear models,

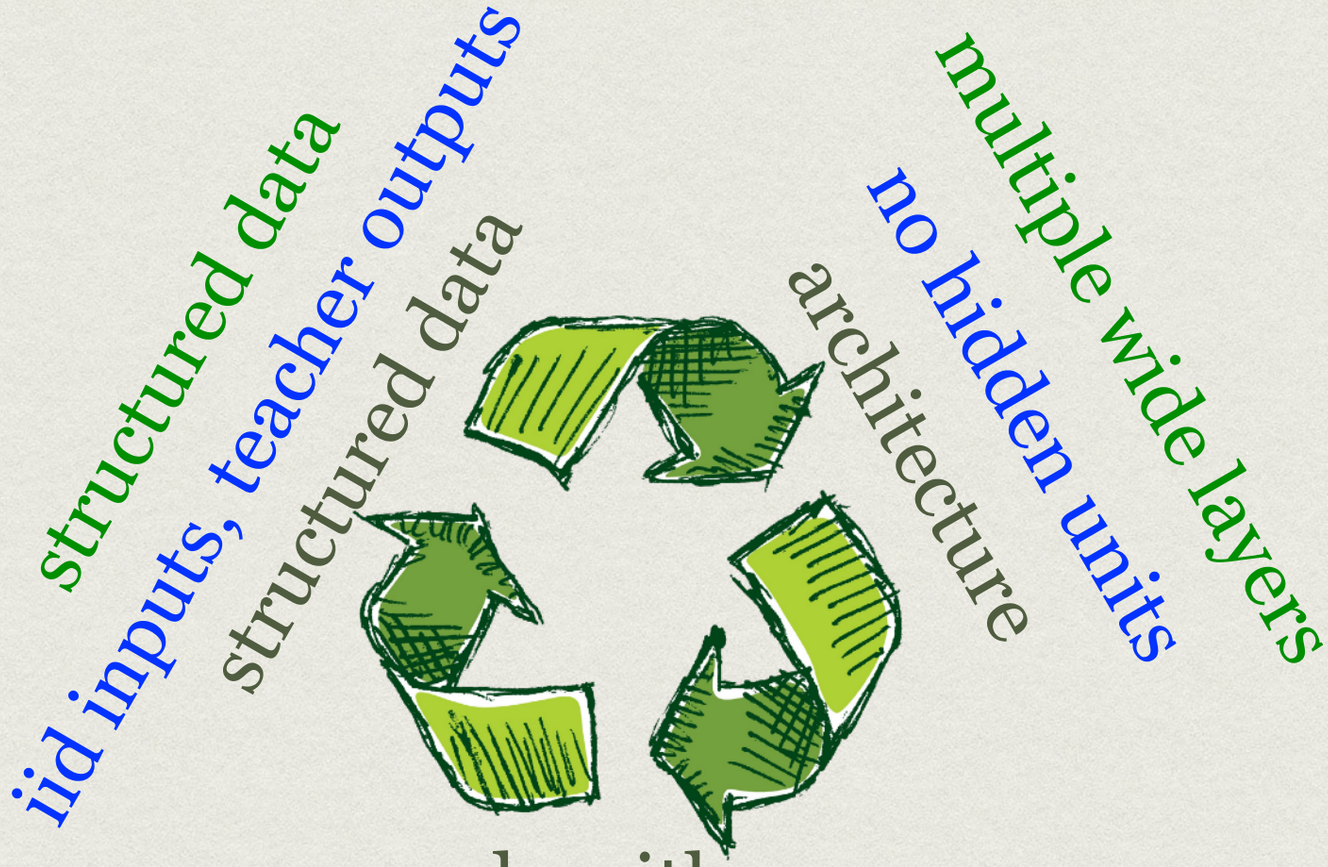
Barbier, Krzakala, Macris, Miolane, LZ, arXiv:1708.03395, COLT'18, PNAS'19

TOWARDS THEORY OF DEEP LEARNING?

color-code:

paper n. 1

needed



algorithm

message passing

gradient-descent-based

TOWARDS THEORY OF DEEP LEARNING?

- Gradient-based dynamics: Marvels and Pitfalls of the Langevin Algorithm in Noisy High-dimensional Inference, Mannelli, Biroli, Cammarota, Krzakala, Urbani, LZ, PRX'20.

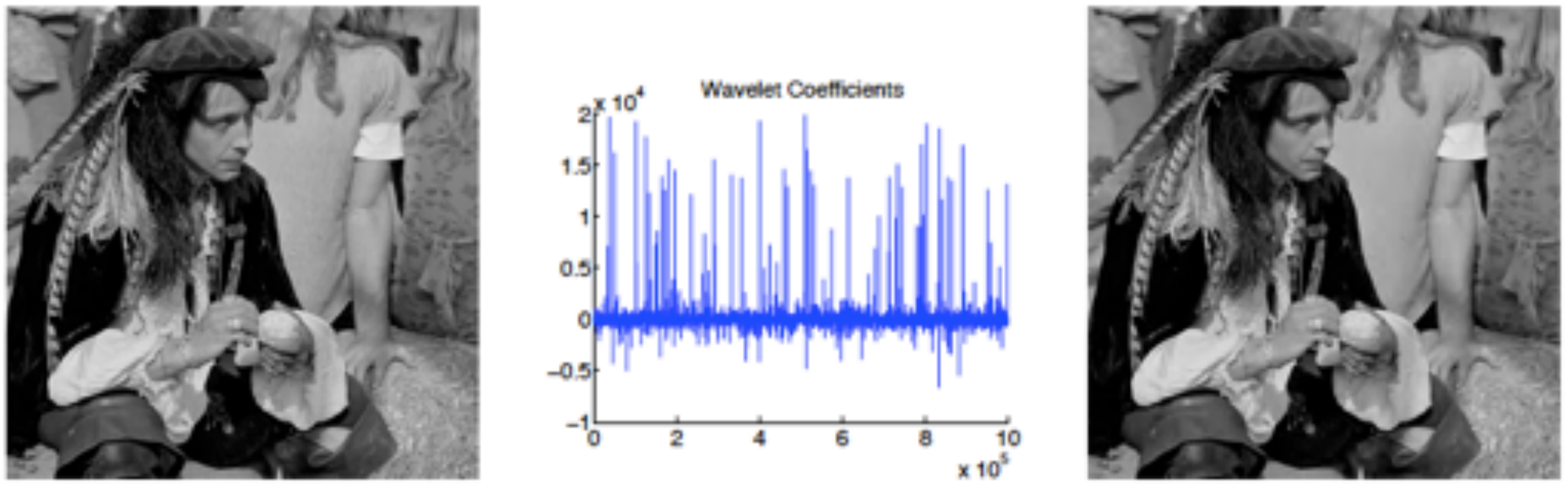
Paper n. 2 for flipped class.

- Structured data & architectures with hidden layers: Generalisation error in learning with random features and the hidden manifold model, Gerace, Loureiro, Krzakala, Mézard, LZ, ICML'20.

Paper n. 3 for flipped class.

How to make the hardness go away?

COMPRESSED SENSING



From 10^6 wavelet coefficients, keep only 25k.

Most signals of interest are sparse in an appropriate basis.

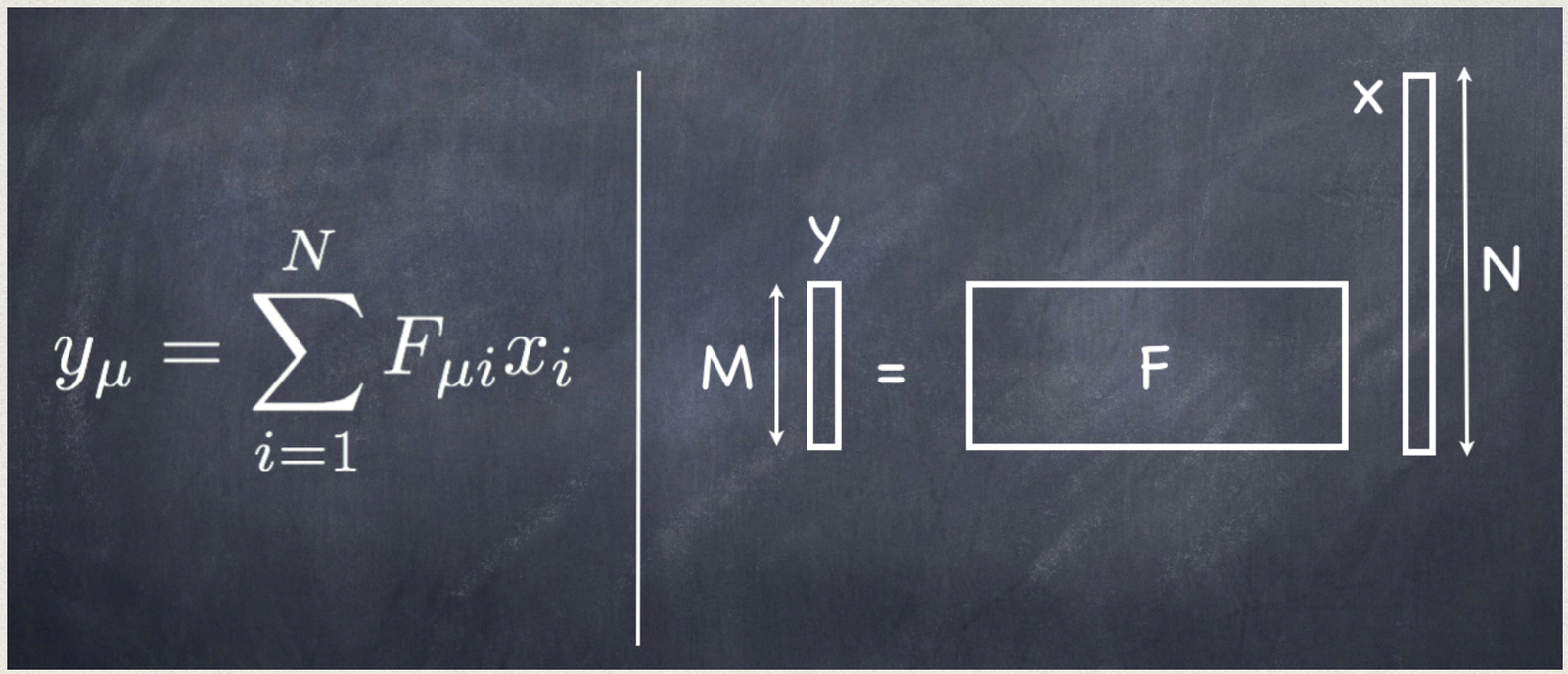
(Exploited everywhere for data compression. Jpeg2000.)

We record the full data and then compress to keep only few bits.

Idea: Can we record directly only the relevant bits. How?

MATHEMATICAL SETTING

Design the matrix F such that sparse signal x can be reconstructed efficiently from measurements y.

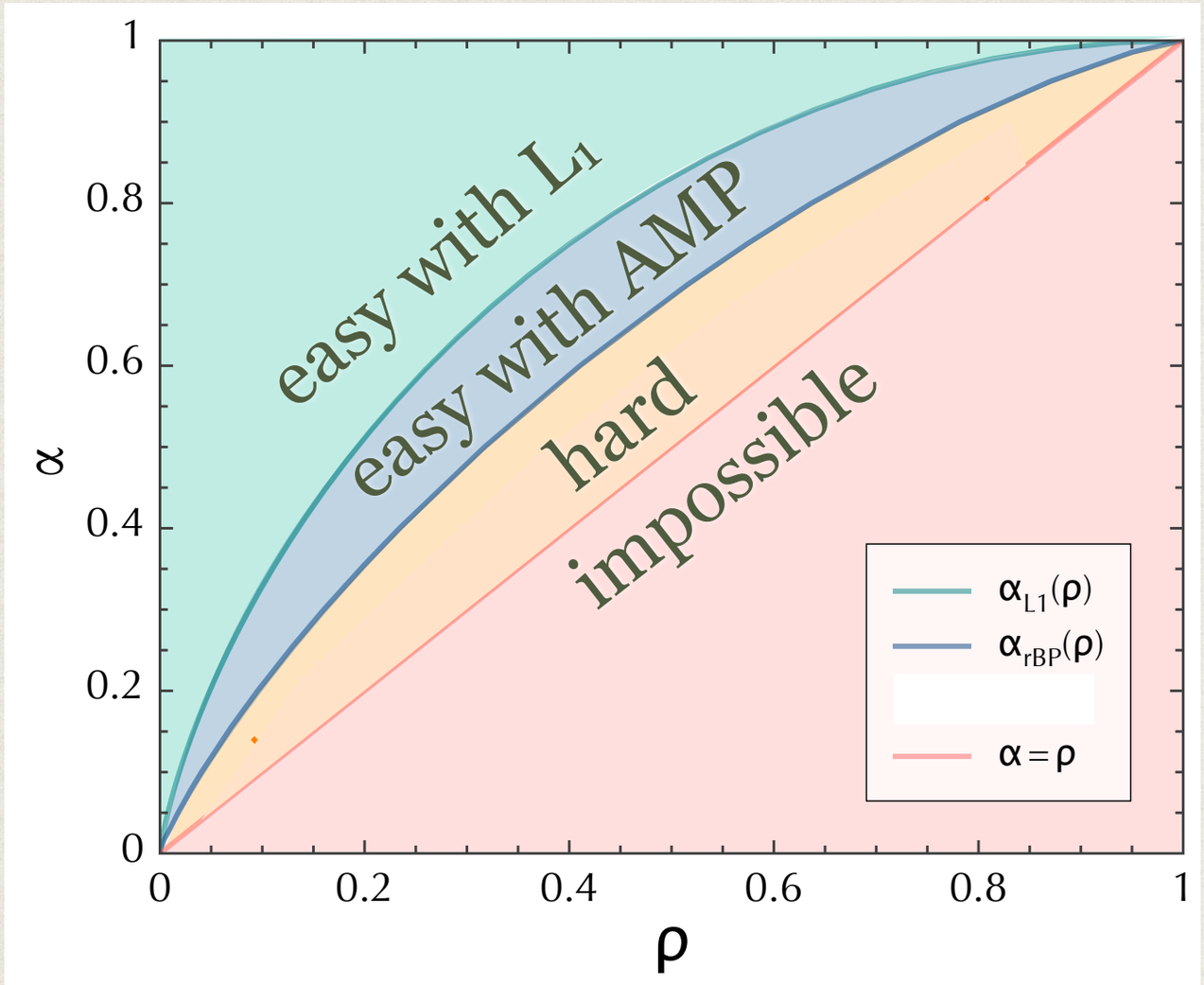


Vector x is sparse, i.e. only ρN elements are non-zero.
The linear problem has many solutions, only is one sparse.

PHASE DIAGRAM

(Dohono, Maleki, Montanari'09, Krzakala, Mézard, Sausset, Sun, Zdeborová'12)

Bayes-optimal compressed sensing, random iid F:



Sparse prior:

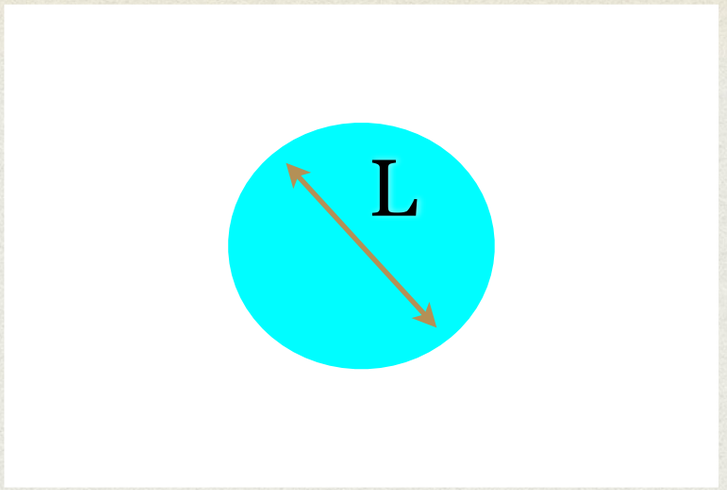
$$P_X(x_i) = (1-\rho)\delta(x_i) + \rho\mathcal{N}(x_i;0,1)$$

Easy/hard threshold = spinodal of a 1st order phase transition.

Freedom in the design of F.

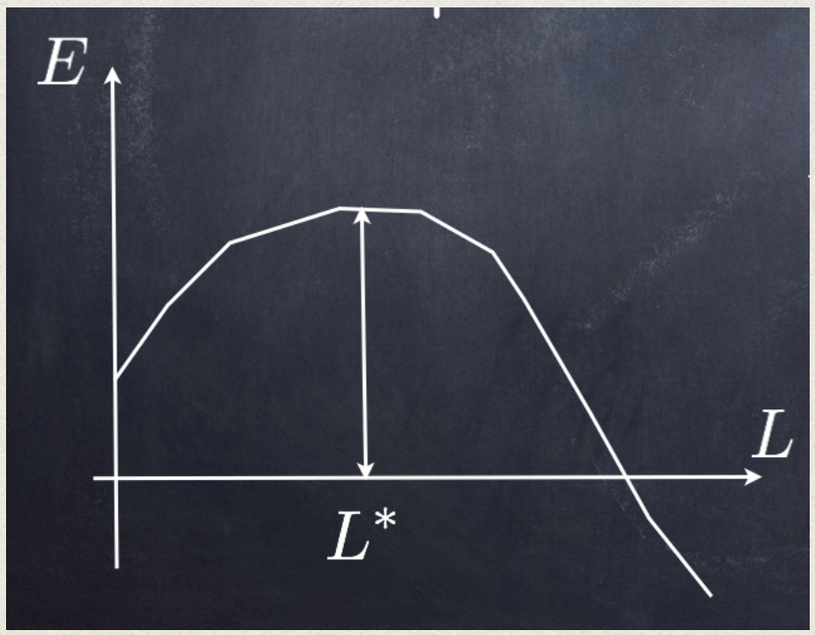
NUCLEATION IN PHYSICS

- Infinitely (exp N) living metastable states exist only in mean field systems (when surface as large as volume).



- Nucleation in finite dimension

Cost to flip a metastable droplet: $E = \Gamma L^{d-1} - \Delta f L^d$



$$L^* = \frac{\Gamma}{\Delta f}\frac{d-1}{d}$$

finite in N !!

$L \ll L^*$ surface wins, droplet shrinks

$L \gg L^*$ volume wins, droplet grows

NUCLEATION FOR OPTIMALITY

Thanks to: UCGP 2008, Kyoto, Japan

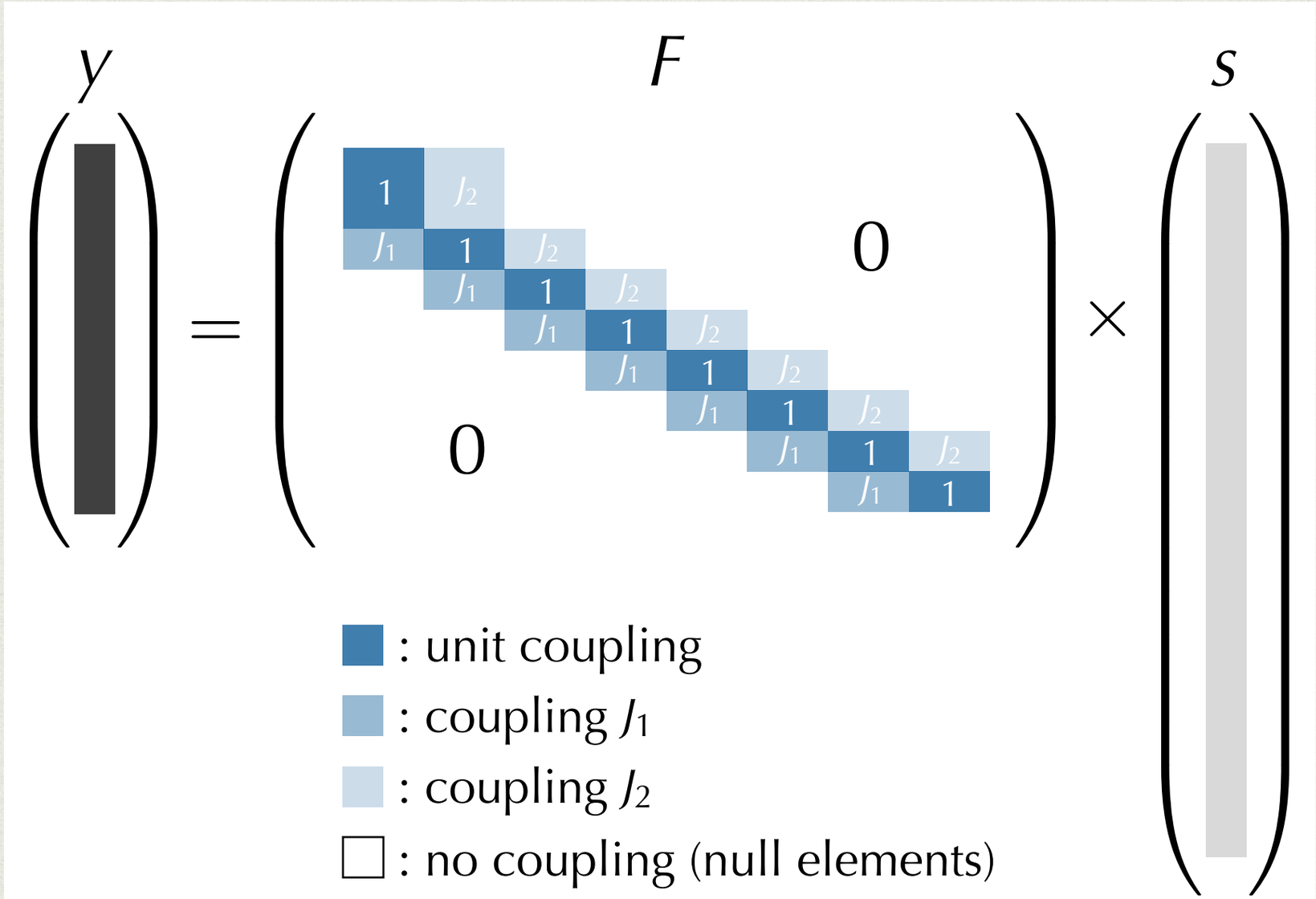
Heating pad or hand warmer:

sodium acetate

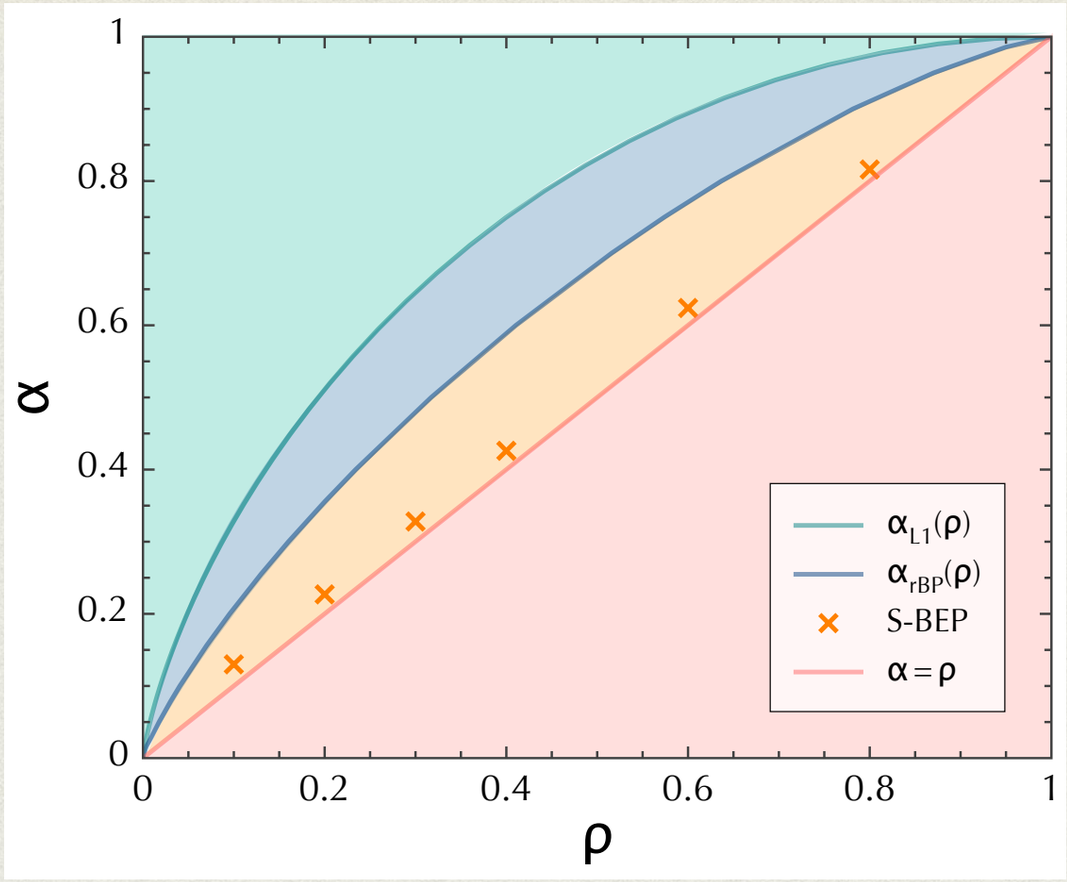
melts at 58 C


AURELE

INDUCING NUCLEATION IN COMPRESSED SENSING



METASTABILITY VANISHES!

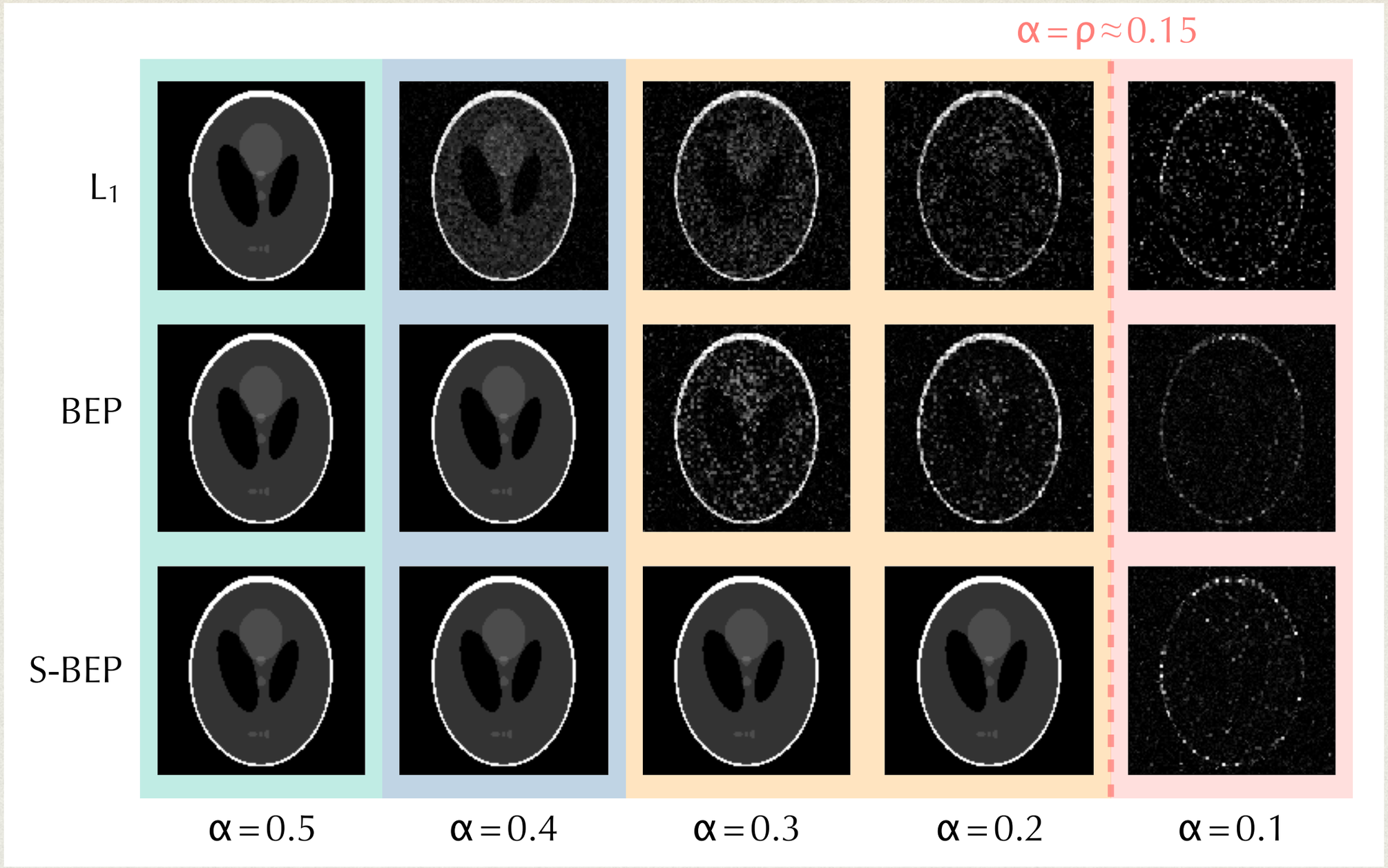


Thanks to induced nucleation compressed sensing is computationally tractable down the information theoretic limit!

Krzakala, Mezard, Sausset, Sun, Zdeborova, Phys. Rev. X 2012.

Proof: Donoho, Javanmard, Montanari, ISIT 2012.

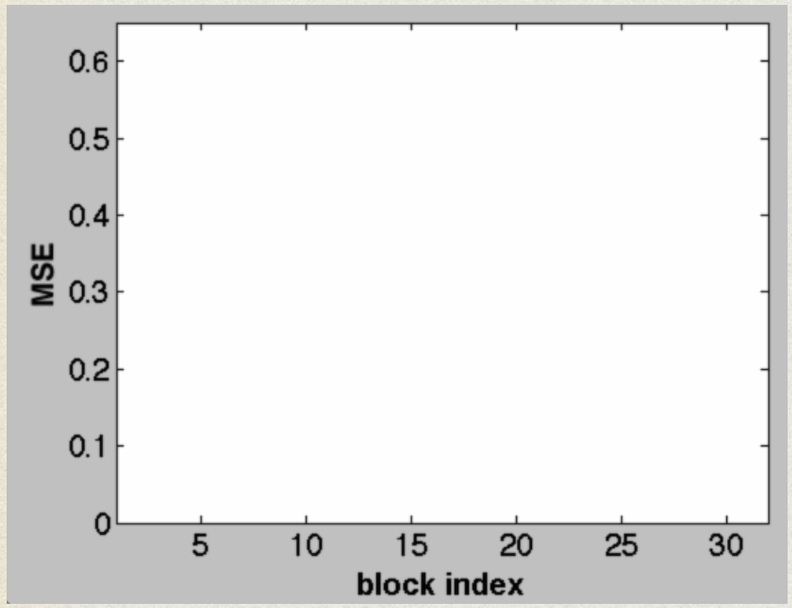
EXAMPLE FOR BENCHMARK DATA

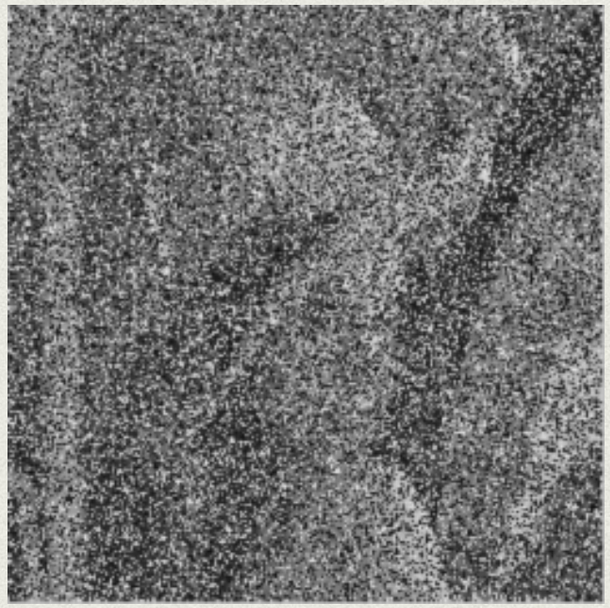


Shepp-Logan phantom,
sparse in the Haar-wavelet representation

EXAMPLE FOR BENCHMARK DATA

Decoding with sparse superposition codes.





Fisher-KPP type of
wave-front propagation

- from: **J. Barbier et al.** Threshold Saturation of Spatially Coupled Sparse Superposition Codes for All Memoryless Channels. IEEE Trans. Inf. Th.'16, ITW'16.