EPFL

UNDERSTANDING MACHINE LEARNING VIA EXACTLY SOLVABLE STATISTICAL PHYSICS MODELS

Lenka Zdeborová
(EPFL)

9.6. 2021, Institute for Physics of Intelligence, University of Tokyo

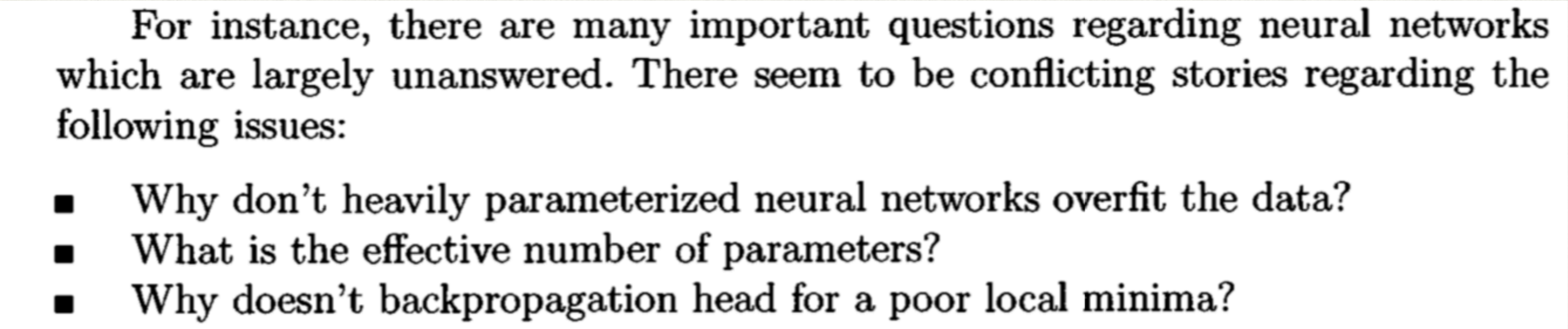
CO-RESPONSIBLE

Alia Abbara, Madhu Advani, Ahmed El Alaoui, Fabrizio Antenucci, Maria-Chiara Angelini, John Ardelius, Benjamin Aubin, Jess Banks, Antoine Baker, Jean Barbier, Giulio Biroli, Alfredo Braunstein, Francesco Caltagirone, Chiara Cammarota, Michele Castellana, Michael Chertkov, Andrea Crisanti, Amin Coja-Oghlan, Hugo Cui, Luca Dall'Asta, Varsha Dani, Mohamad Dia, Aurelien Decelle, Laura Foini, Silvio Franz, Marylou Gabrié, Federica Gerace, Cedric Gerbelot, Sebastian Goldt, Emmanuelle Gouillart, Nils-Eric Guenther, Václav Janiš, Michael I Jordan, Yoshyiuki Kabashima, Brian Karrer, Lukas Kroc, Florent Krzakala, Alejandro Lage Castellanos, Marc Lelarge, Thibault Lesieur, Luca Leuzzi, Martin Loebl, Bruno Loureiro, Yue M. Lu, Clément Luneau, Nicolas Macris, Antoine Maillard, Andre Manoel, Yoshiki Matsuda, Marc Mézard, Léo Miolane, Francesca Mignacco, Andrea Montanari, Cristopher Moore, Richard G. Morris, Elchanan Mossel, Joe Neeman, Mark Newman, Hidetoshi Nishimori, Will Perkins, Henry D Pfister, Sundeep Rangan, Aaditya Ramdas, Abolfazl Ramezanpour, Joerg Reichardt, Federico Ricci-Tersenghi, Alaa Saade, Luca Saglietti, Stefano Mannelli Sarao, Ayaka Sakata, Francois Sausset, Andrew Saxe, Christian Schmidt, Christophe Schulke, Guilhem Semerjian, Cosma R. Shalizi, David Sherrington, Allan Sly, Phil Schniter, Bertrand Thirion, Eric W. Tramel, Pierfrancesco Urbani, Eric Vanden-Eijnden, Gaël Varoquaux, Massimo Vergassola, Yingying Xu, Jiaming Xu, Sun Yifan, Riccardo Zecchina, Pan Zhang, Hai-jun Zhou.

MOTIVATION

- Deep learning brought unprecedented empirical/engineering progress into many applications, including physics.

- Some open questions:



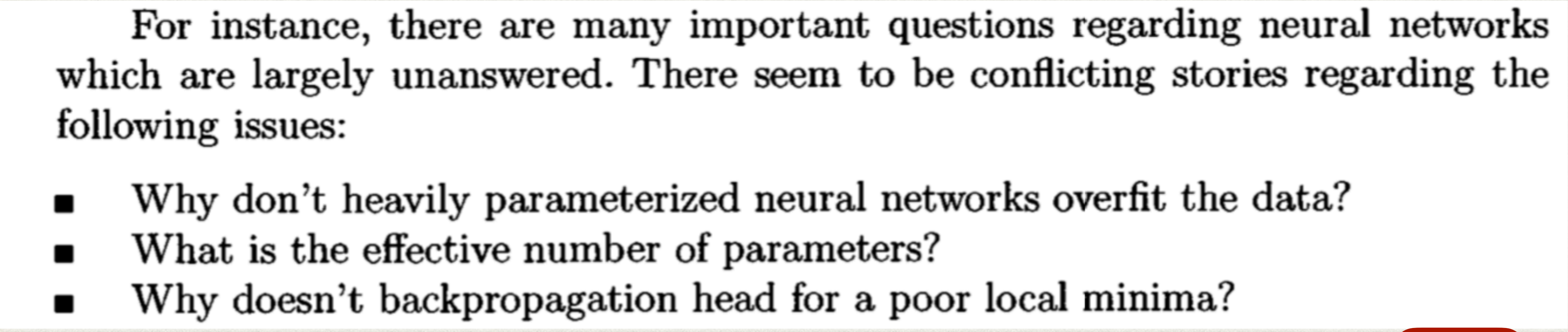
For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- 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?

MOTIVATION

- Deep learning brought unprecedented empirical/engineering progress into many applications, including physics.

- Some open questions:



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

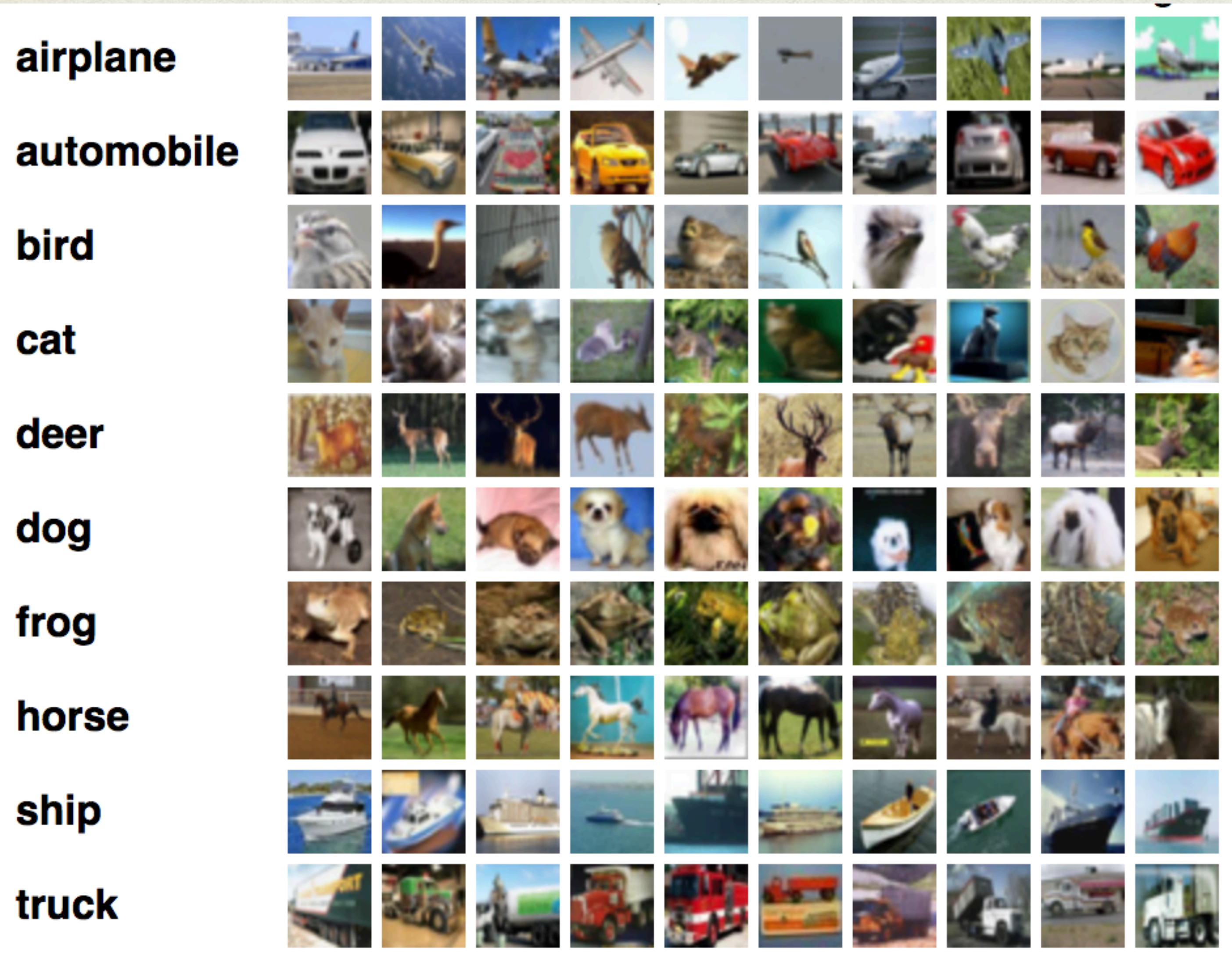
- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?

From "Reflections after refereeing papers for NIPS", Leo Breiman, 1995.

Still not answered!

SAMPLE COMPLEXITY

How many training samples are needed for a given task? Are we close to the minimum? If not, is it because of architectures or algorithms?

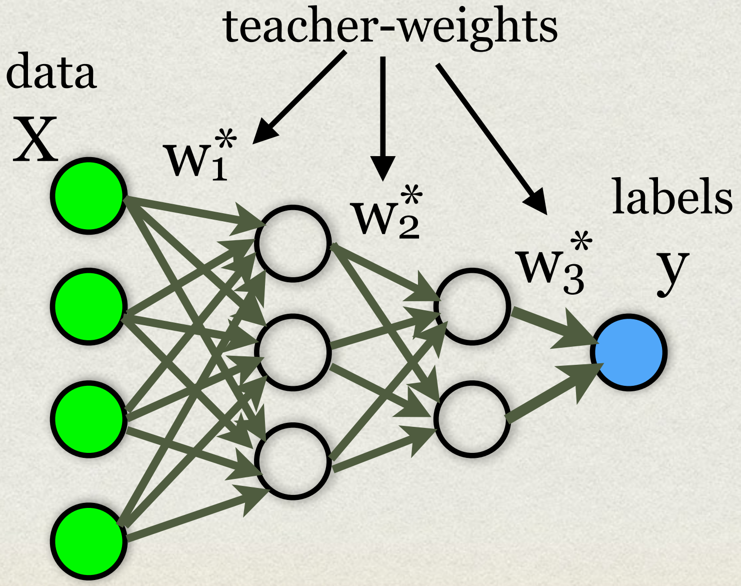


- Cifar10 - 50000 samples.
- How many samples are really needed?

WHEN CAN A NEURAL NETWORK LEARN A TEACHER-NEURAL NETWORK?

Teacher-network

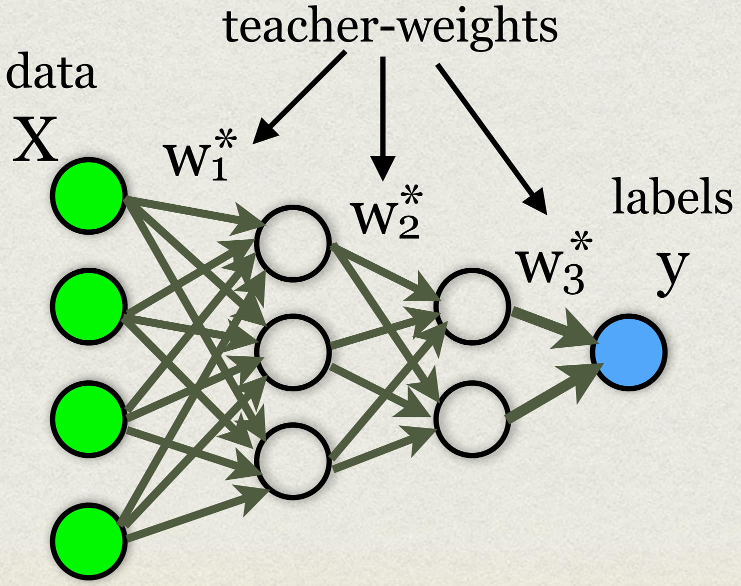
- Generates data X, n samples of d dimensional data, e.g. random input vectors.
- Generates weights w*, e.g. iid random.
- Generates labels y.



WHEN CAN A NEURAL NETWORK LEARN A TEACHER-NEURAL NETWORK?

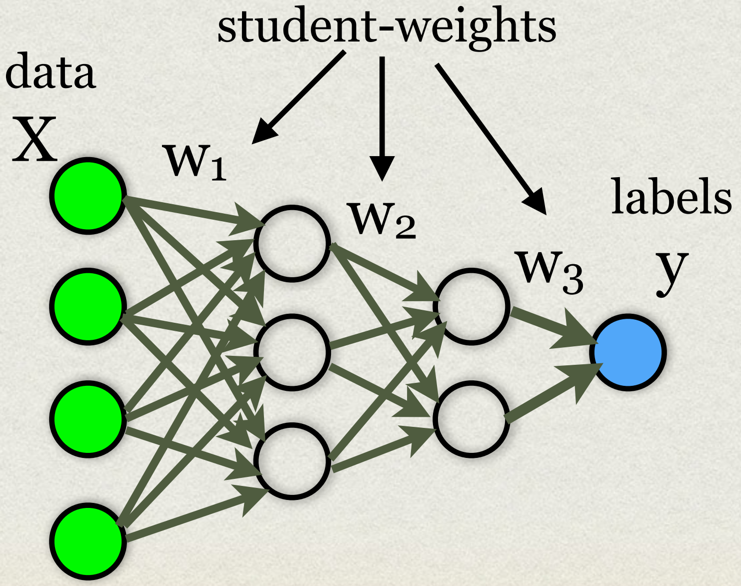
Teacher-network

- Generates data X, n samples of d dimensional data, e.g. random input vectors.
- Generates weights w*, e.g. iid random.
- Generates labels y.

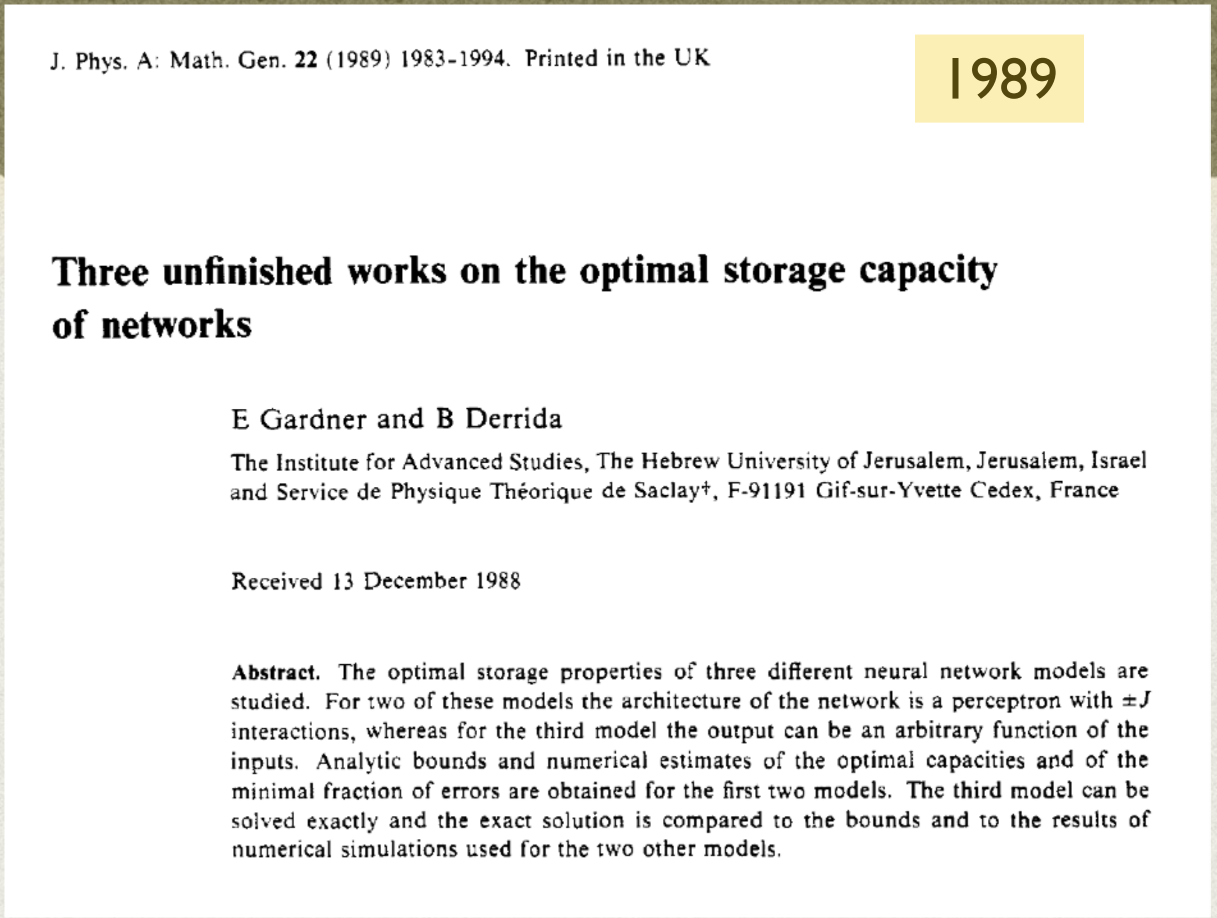


Student-network

- Observes X, y, the architecture of the network.
- How does the best achievable generalisation error depend on the number of samples n?



TEACHER-STUDENT PERCEPTRON



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

1989

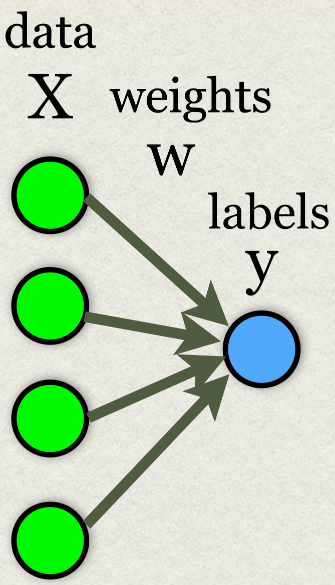
Three unfinished works on the optimal storage capacity of networks

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.

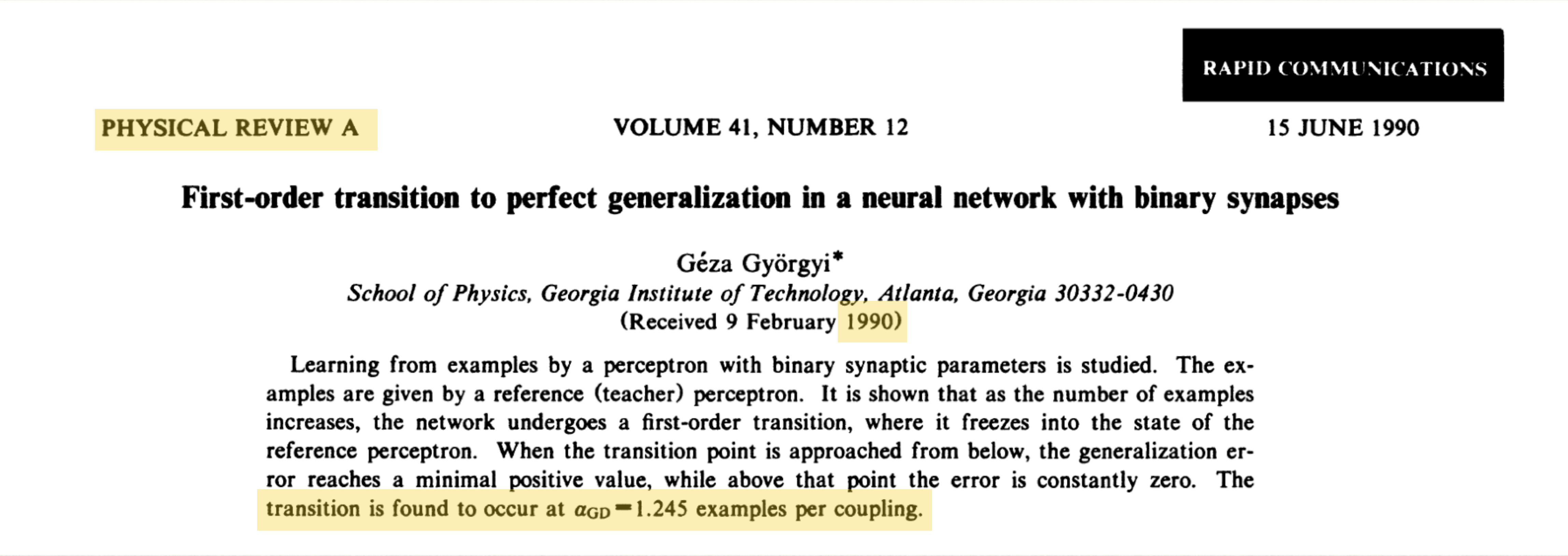


- Take random iid Gaussian $X_{\mu i}$, and random iid w_i^* from P_w
- Create $y_\mu = \text{sign}\left(\sum_{i=1}^{d} X_{\mu i} w_i^*\right)$
- High-dimensional regime: $n \to \infty \quad d \to \infty$ $\alpha \equiv n/d = \Theta(1)$

p dimensions

n samples

Solved using the replica method in the high-dimensional limit



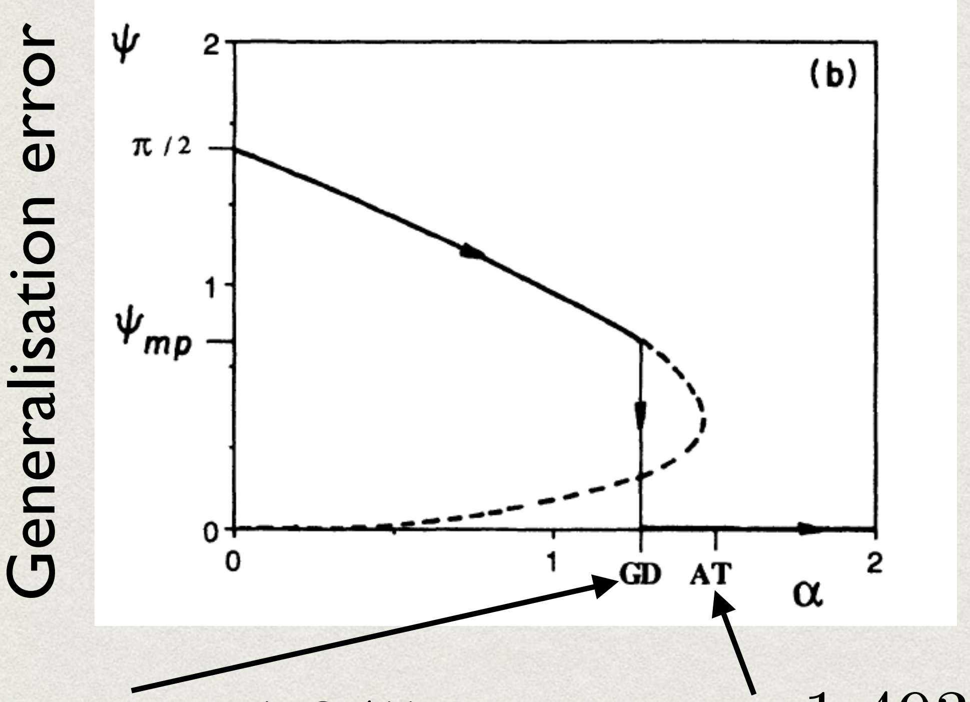
RAPID COMMUNICATIONS

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.



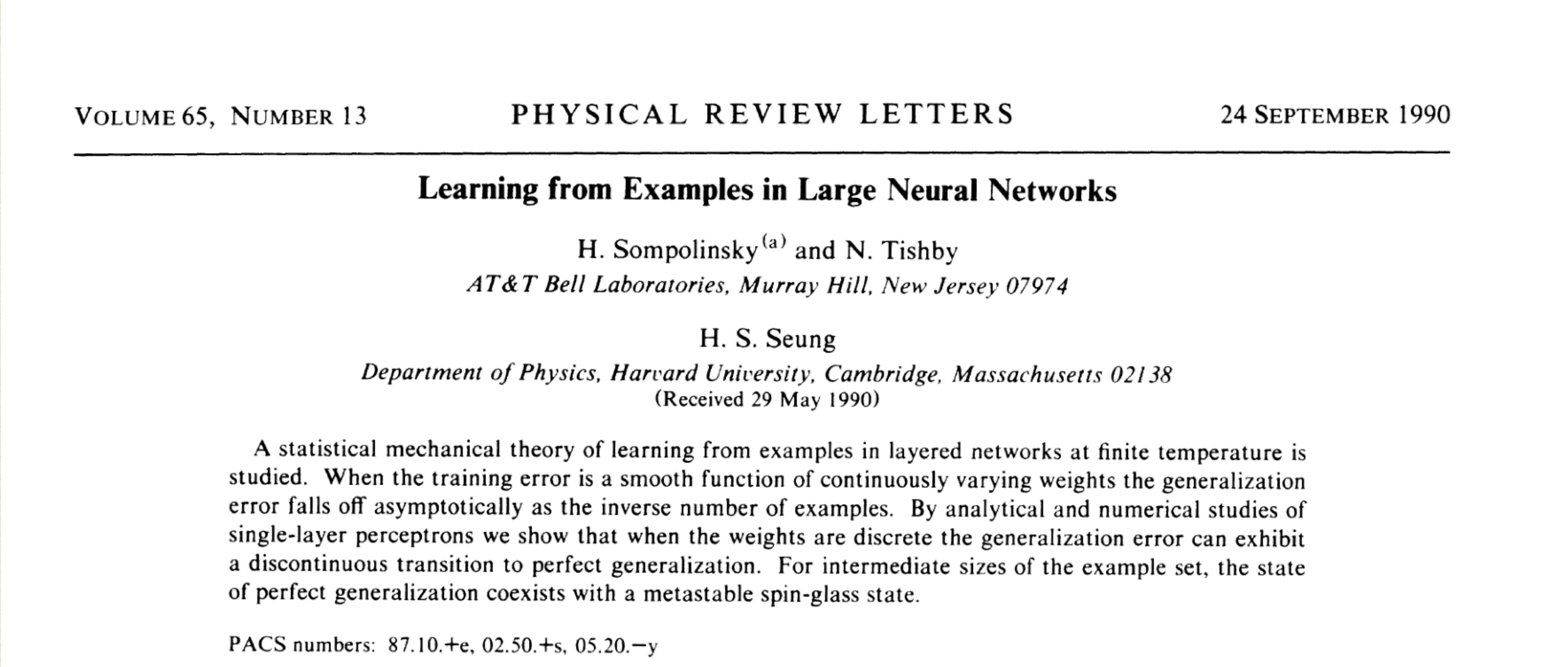
$\alpha_{\mathrm{GD}} = 1.245$ $\alpha_{\mathrm{AT}} = 1.493$

- Binary teacher-weights:

$$w^* \in \{-1,1\}^d$$

- 1st order phase transition in the learning curve.

$\alpha = n/d$



VOLUME 65, NUMBER 13 PHYSICAL REVIEW LETTERS 24 SEPTEMBER 1990

Learning from Examples in Large Neural Networks

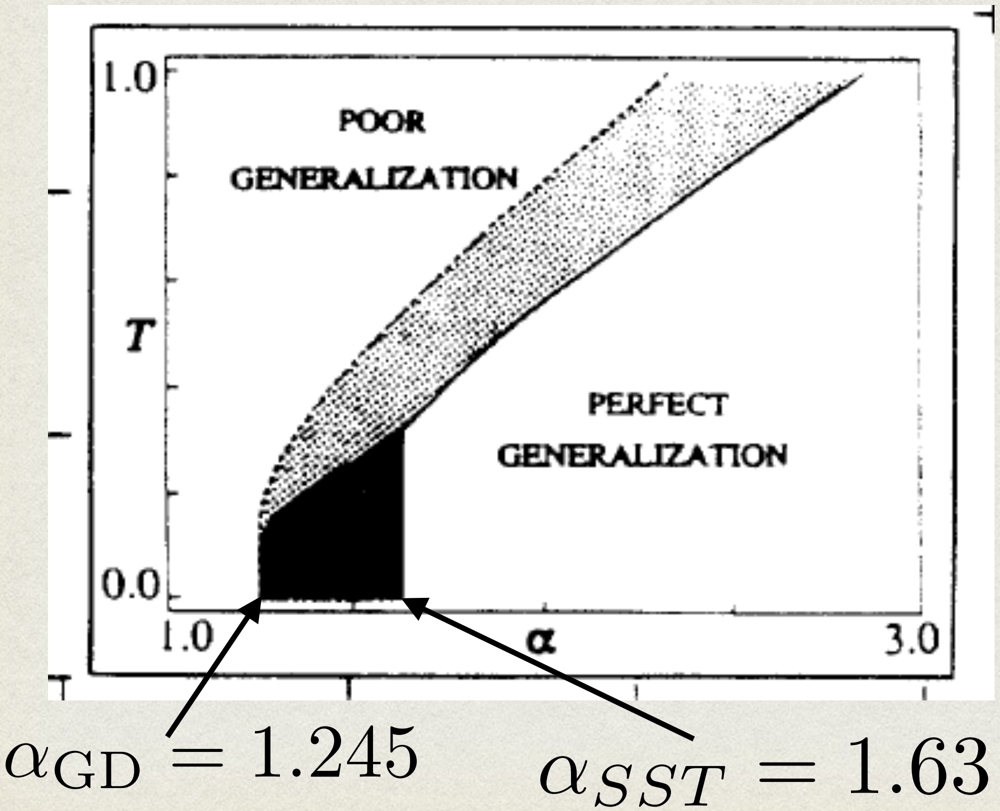
H. Sompolinsky[a] and N. Tishby
AT&T Bell Laboratories, Murray Hill, New Jersey 07974

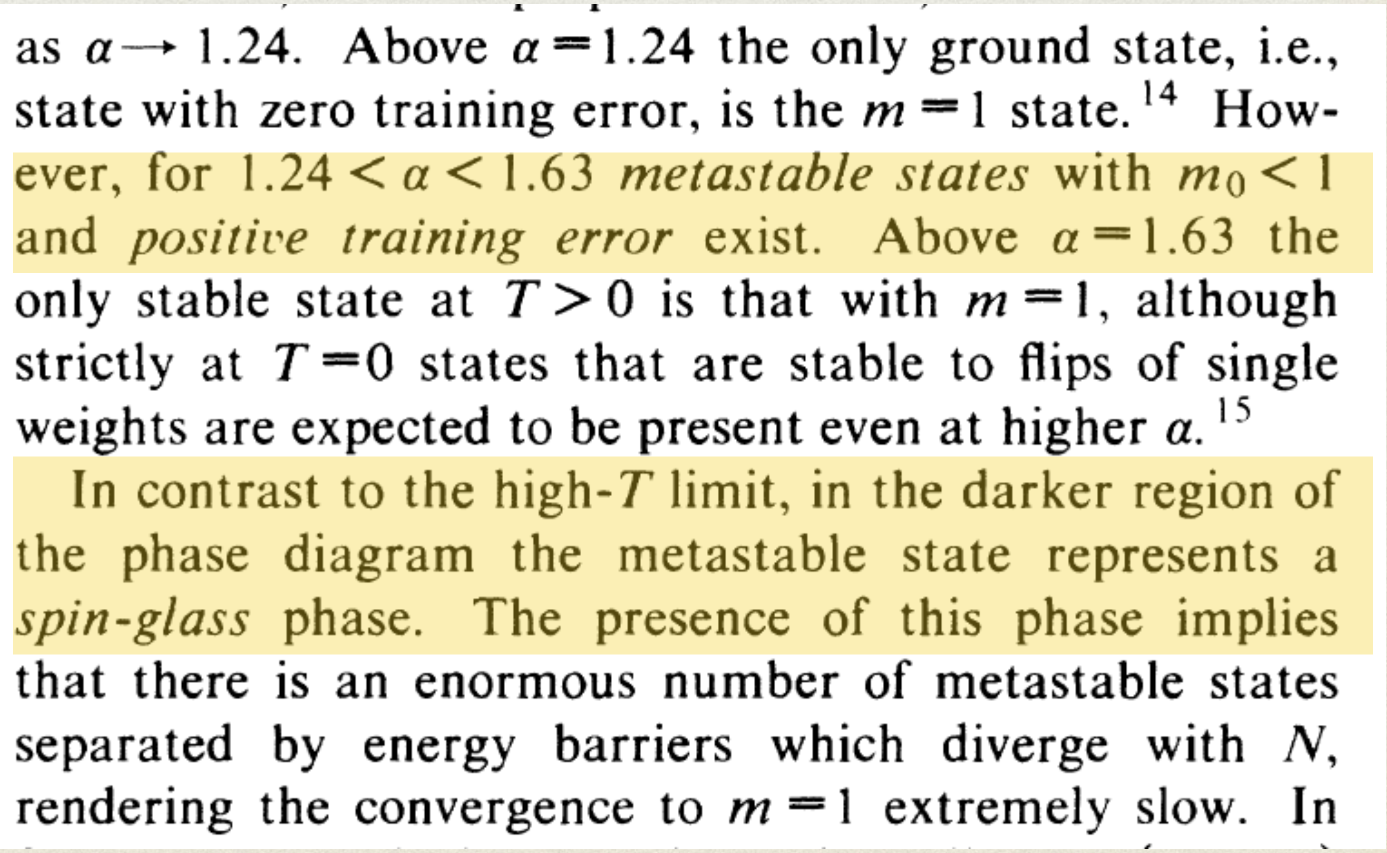
H. S. Seung
Department of Physics, Harvard University, Cambridge, Massachusetts 02138

(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





as $\alpha \rightarrow 1.24$. Above $\alpha = 1.24$ the only ground state, i.e., state with zero training error, is the $m = 1$ state.[14] However, for $1.24 < \alpha < 1.63$ *metastable states* with $m_0 < 1$ and *positive training error* exist. Above $\alpha = 1.63$ the only stable state at $T > 0$ is that with $m = 1$, although strictly at $T = 0$ states that are stable to flips of single weights are expected to be present even at higher α.[15]

In contrast to the high-T limit, in the darker region of the phase diagram the metastable state represents a *spin-glass* phase. The presence of this phase implies that there is an enormous number of metastable states separated by energy barriers which diverge with N, rendering the convergence to $m = 1$ extremely slow. In

STATE-OF-THE-ART GENERALIZED LINEAR MODEL

- Best achievable generalisation error for the single-layer teacher-student model for any activation function, any prior on weights.
- Regions of optimality of approximate message passing algorithm.
- Rigorous proof that the replica solution for the teacher-student model is correct.

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

BAYES-OPTIMAL GENERALIZATION

Posterior probability distribution:

$$P(w \,|\, y, X) = \frac{1}{Z(y, X)} \prod_{i=1}^{d} P_w(w_i) \prod_{\mu=1}^{n} P_{\text{out}}(y_\mu \,|\, X_\mu \cdot w)$$

where $\quad P_{\text{out}}(y_\mu \,|\, X_\mu \cdot w) = \delta(y_\mu - \varphi(X_\mu \cdot w))$

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

$\neq$ empirical risk minimization

REPLICA METHOD SOLUTION

Def. “quenched” free energy: $f = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}_{y,X} \log Z(y,X)$ $\alpha = \frac{d}{n}$

Theorem 1:

$$f = \sup_{m} \inf_{\hat{m}} f_{RS}(m,\hat{m})$$

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

where

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

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

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

REPLICA METHOD SOLUTION

Def. "quenched" free energy: $f = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}_{y,X}\log Z(y,X)$

$$\alpha = \frac{d}{n}$$

Theorem 1:

$$f = \sup_{m}\inf_{\hat{m}} f_{RS}(m,\hat{m})$$

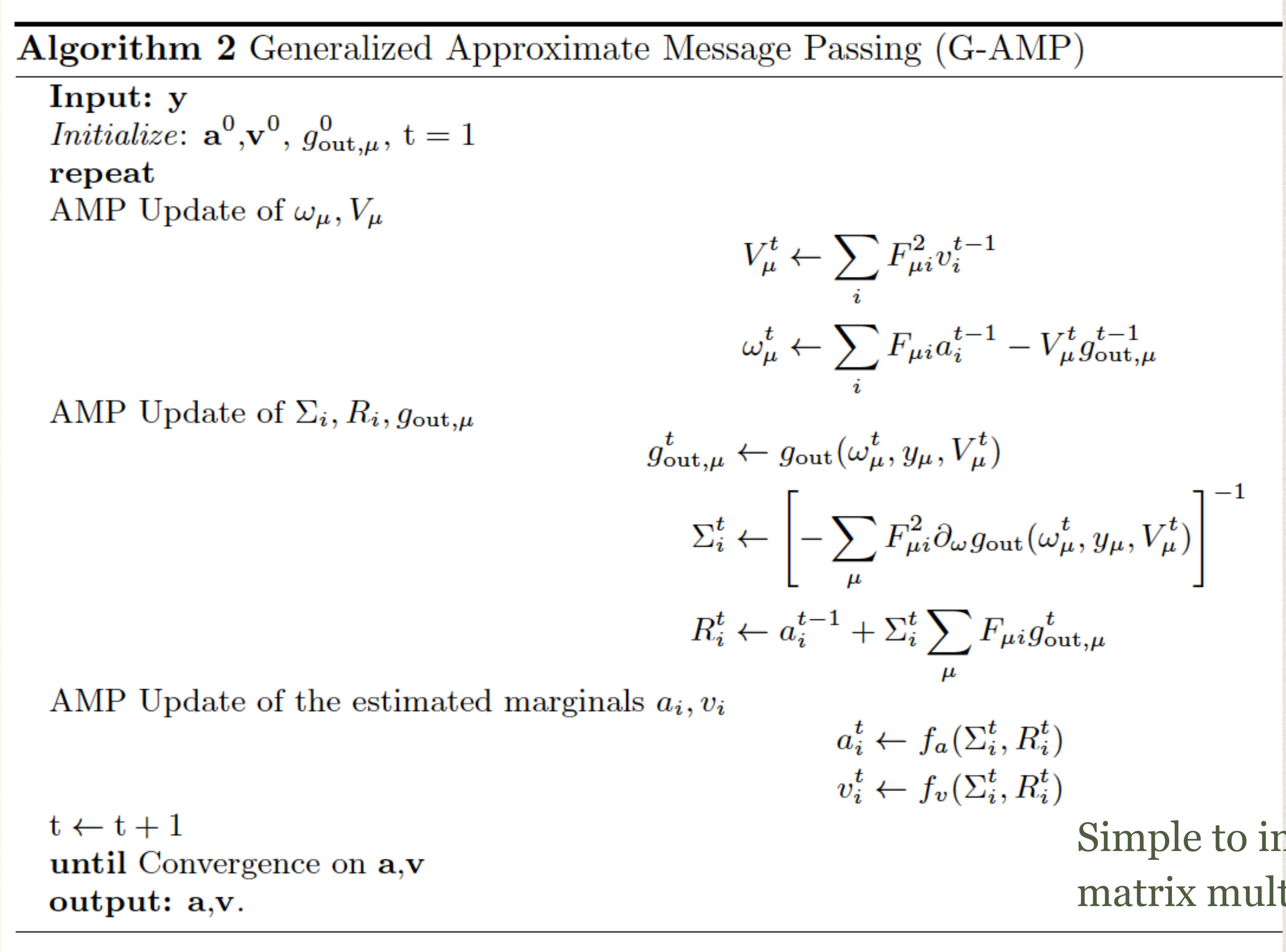
$$f_{\mathrm{RS}}(m,\hat{m}) = \Phi_{P_w}(\hat{m}) + \alpha\Phi_{P_{\mathrm{out}}}(m;\rho) - \frac{m\hat{m}}{2}$$

Theorem 2: Optimal generalisation error

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

where m* is the extremizer of f_{RS}.

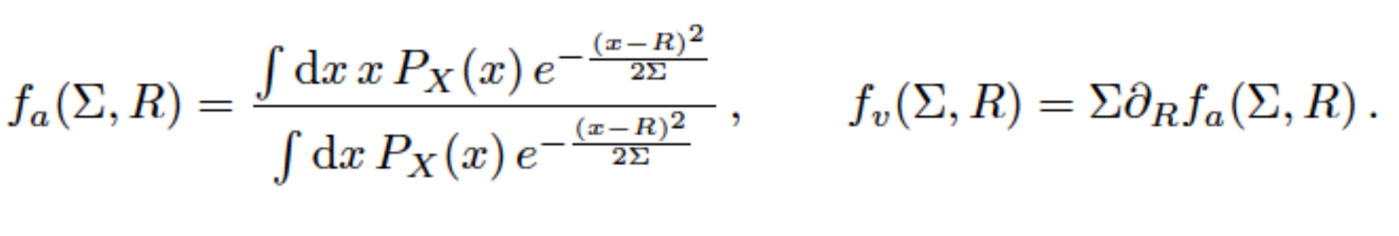
$$\rho = \mathbb{E}_{P_w}(w^2)$$
$$v, z \sim \mathcal{N}(0,1)$$
$$\xi \sim P_\xi$$



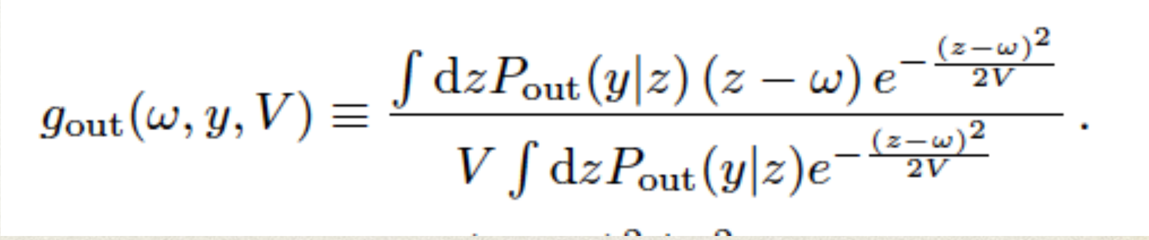
Algorithm 2 Generalized Approximate Message Passing (G-AMP)

```
Input: y
Initialize: $\mathbf{a}^0$,$\mathbf{v}^0$, $g^0_{\text{out},\mu}$, t = 1
repeat
AMP Update of $\omega_\mu, V_\mu$
    $V^t_\mu \leftarrow \sum_i F^2_{\mu i} v^{t-1}_i$
    $\omega^t_\mu \leftarrow \sum_i F_{\mu i} a^{t-1}_i - V^t_\mu g^{t-1}_{\text{out},\mu}$
AMP Update of $\Sigma_i, R_i, g_{\text{out},\mu}$
    $g^t_{\text{out},\mu} \leftarrow g_{\text{out}}(\omega^t_\mu, y_\mu, V^t_\mu)$
    $\Sigma^t_i \leftarrow \left[-\sum_\mu F^2_{\mu i} \partial_\omega g_{\text{out}}(\omega^t_\mu, y_\mu, V^t_\mu)\right]^{-1}$
    $R^t_i \leftarrow a^{t-1}_i + \Sigma^t_i \sum_\mu F_{\mu i} g^t_{\text{out},\mu}$
AMP Update of the estimated marginals $a_i, v_i$
    $a^t_i \leftarrow f_a(\Sigma^t_i, R^t_i)$
    $v^t_i \leftarrow f_v(\Sigma^t_i, R^t_i)$
t ← t + 1
until Convergence on a,v
output: a,v.
```

Simple to implement, only matrix multiplications, O(N²)



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



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

STATE EVOLUTION

Bayati, Montanari'11, Bayati, Lelarge, Montanari'12, 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:

$$m^{t+1} = 2\partial_{\hat{m}}\Phi_{P_w}(\hat{m}^t)$$

$$\hat{m}^t = 2\alpha\partial_m\Phi_{P_{\mathrm{out}}}(m^t;\rho)$$

Recall the RS free energy

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

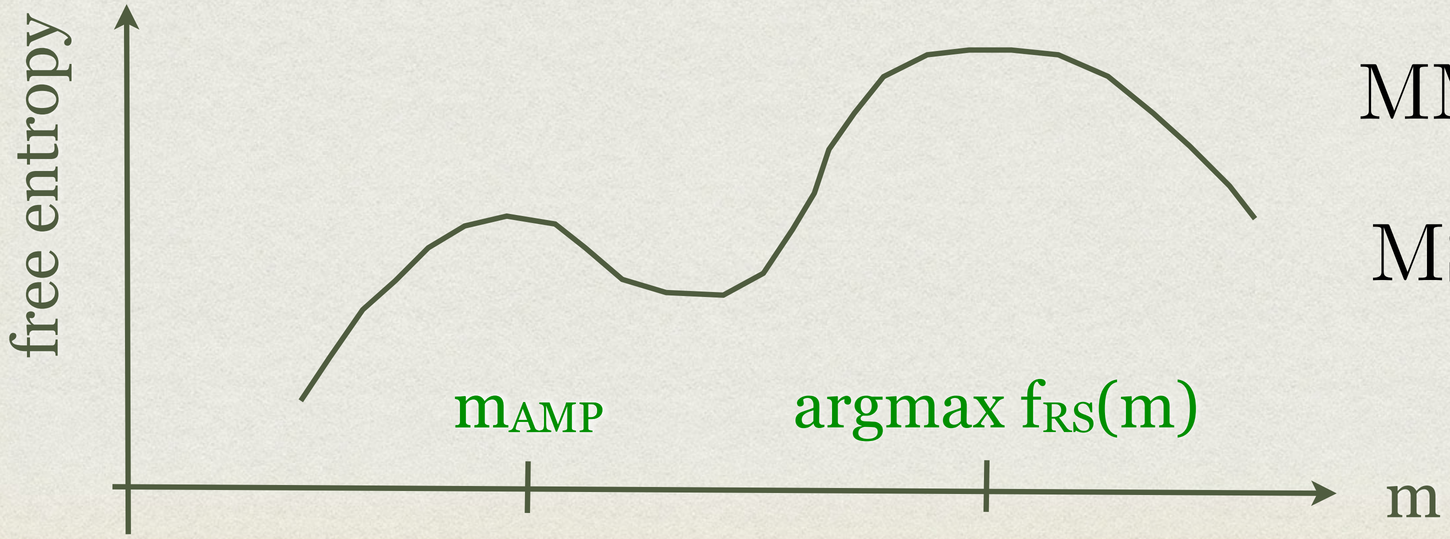
BOTTOM LINE

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

$$f_{RS}(m) = \inf_{\hat{m}} f_{RS}(m, \hat{m})$$

- AMP-MSE given by the local maximum of the free entropy reached starting from small m/large MSE.

- MMSE is given by the global maximum of the free entropy.

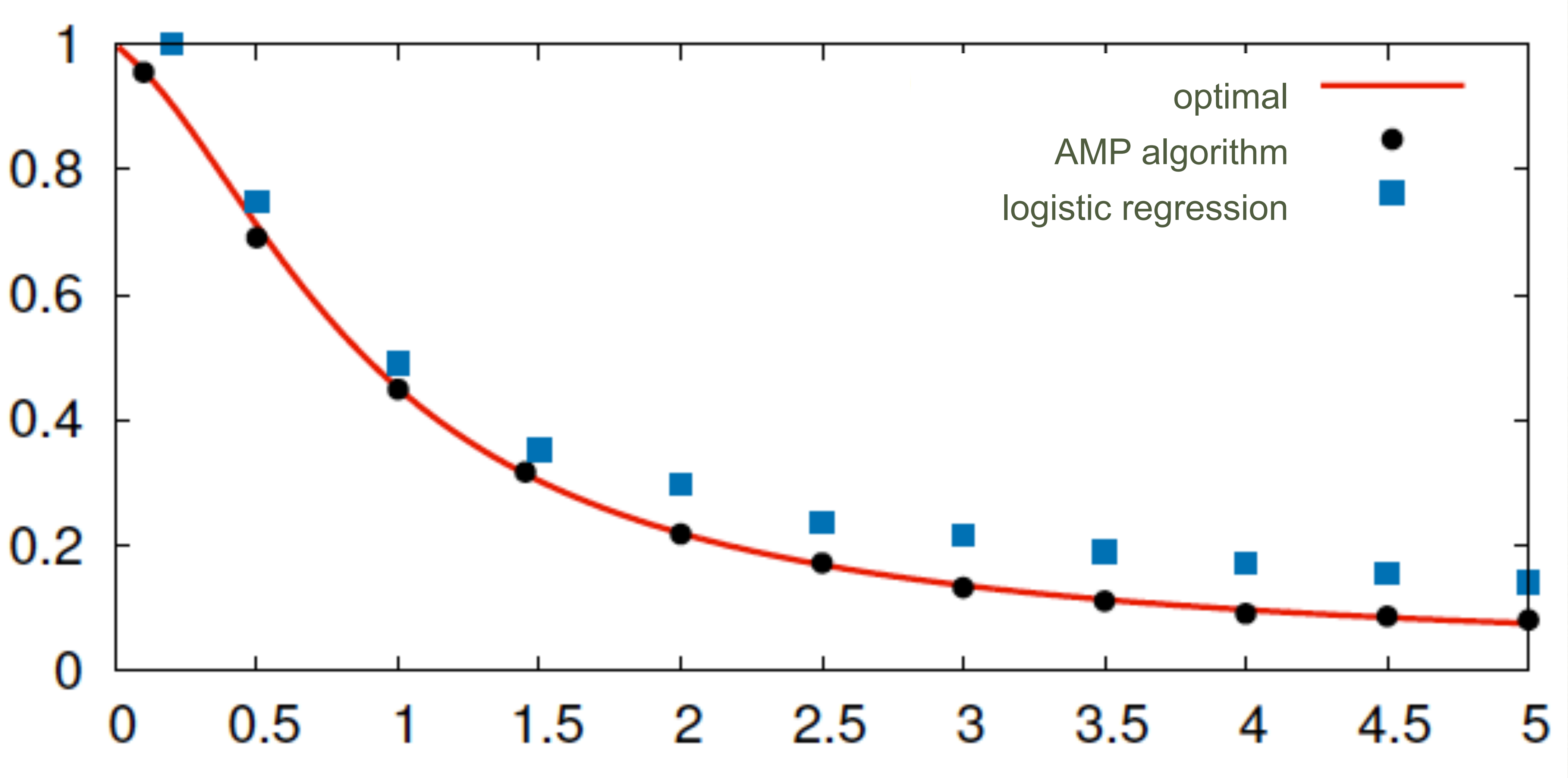


$$\mathrm{MMSE} = \rho - \mathrm{argmax} f_{RS}(m)$$

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

SPHERICAL PERCEPTRON

$y_\mu = \mathrm{sign}\left(\sum_{i=1}^{d} X_{\mu i} w_i\right)$ $P_w = \mathcal{N}(0,1)$ $n \to \infty$ $d \to \infty$ $n/d = \Theta(1)$



test error

# of samples per dimension n/d

BAYES VS RISK MINIMISATION

- So far: Bayes-optimal estimation = marginals of the posterior:

$$P(w\,|\,y,X) = \frac{1}{Z(y,X)}\prod_{i=1}^{p} P_w(w_i)\prod_{\mu=1}^{n} P_{\text{out}}(y_\mu\,|\,X_\mu\cdot w)$$

- More common: Empirical risk minimisation = minimisation of a loss function:

$$\min_w\Big[\sum_{\mu=1}^{n}\ell(y_\mu,\mathbf{X}_\mu\cdot\mathbf{w})+\lambda\|w\|_2^2\Big]$$

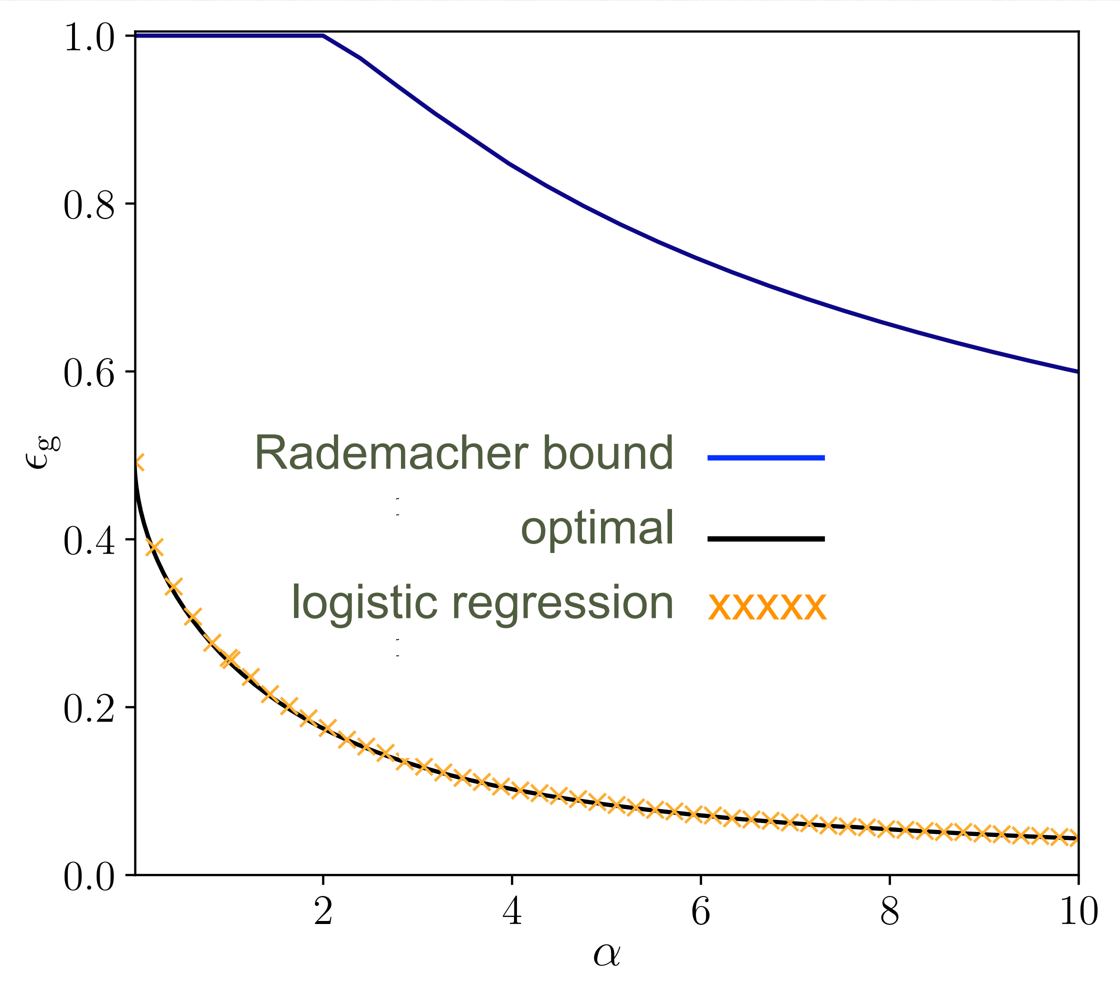
e.g. square loss $\ell(y,z)=(y-z)^2$, logistic loss $\ell(y,z)=\log_2(1+e^{-yz})$

BAYES VS RISK MINIMISATION

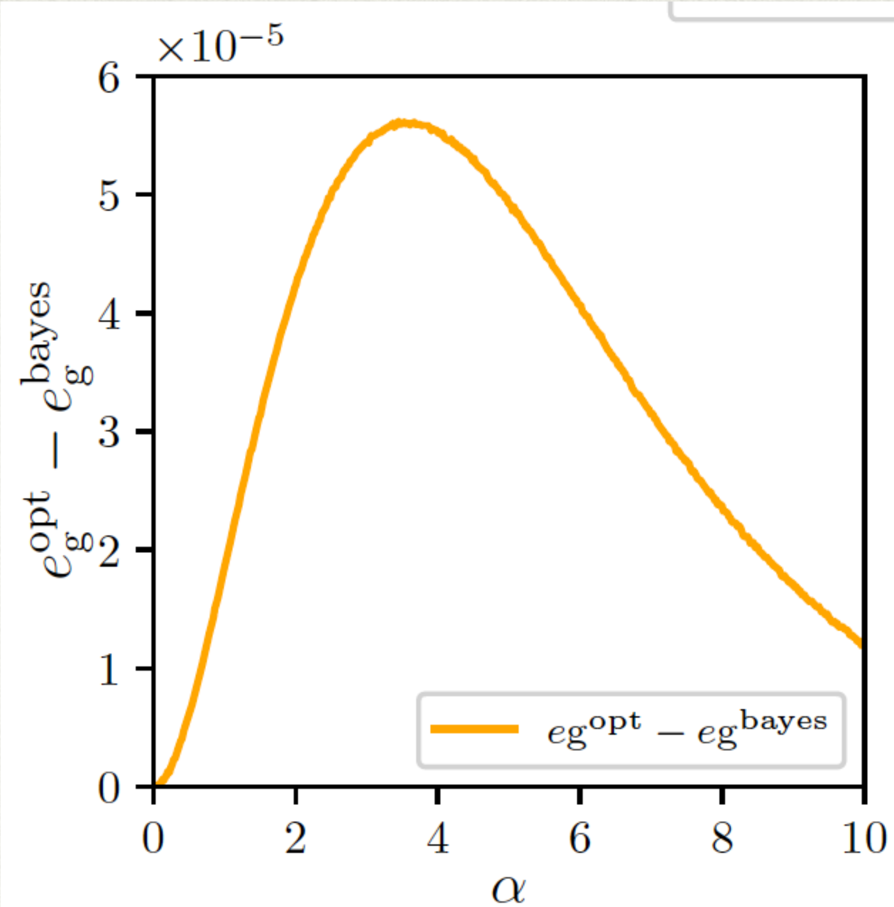
$$y_\mu = \mathrm{sign}\left(\sum_{i=1}^{d} X_{\mu i} w_i\right) \qquad P_w = \mathcal{N}(0,1)$$

Optimally regularized logistic regression essentially Bayes-optimal

generalisation error



# of samples per dimension



Aubin, Lu, FK, LZ, 2006.06560

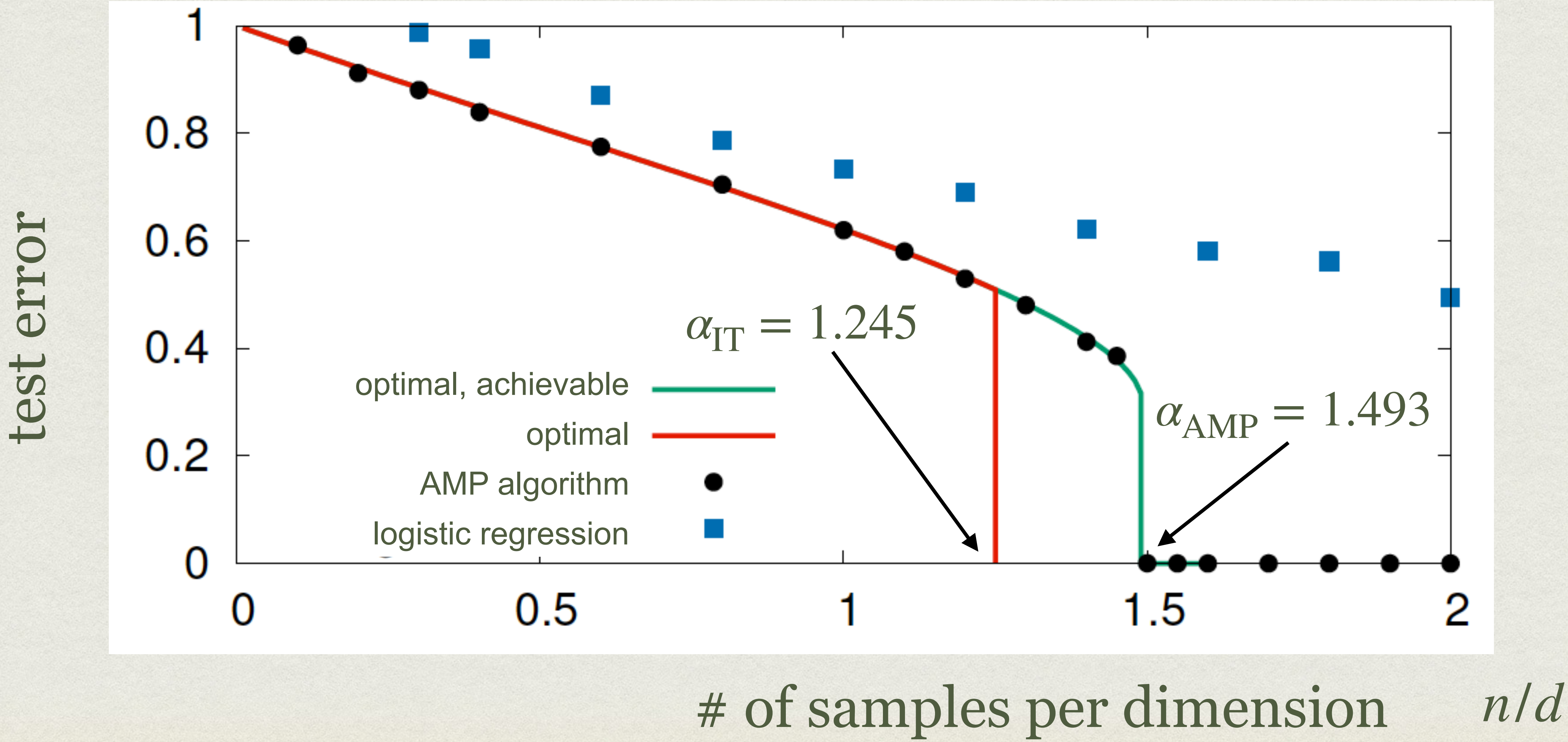
BINARY PERCEPTRON

$$y_\mu = \mathrm{sign}\Big(\sum_{i=1}^{d} X_{\mu i} w_i\Big) \qquad w_i \in \{-1, +1\}$$

$n \to \infty$

$d \to \infty$

$n/d = \Theta(1)$



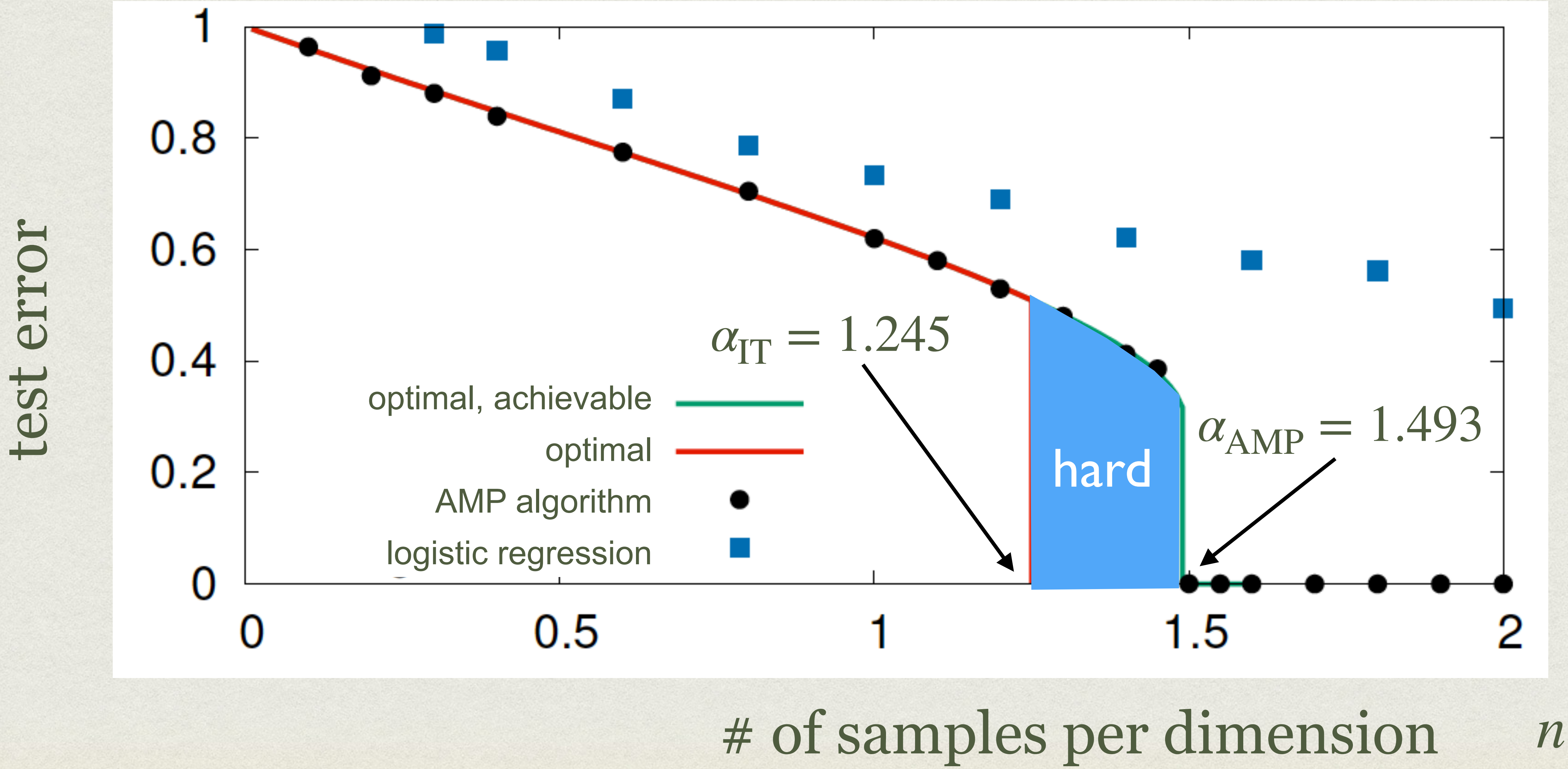
BINARY PERCEPTRON

$$y_\mu = \mathrm{sign}\left(\sum_{i=1}^{d} X_{\mu i} w_i\right) \qquad w_i \in \{-1, +1\}$$

$n \to \infty$

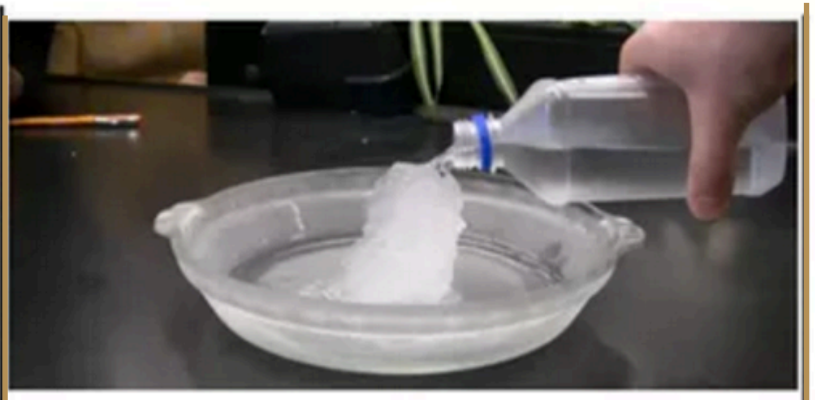
$d \to \infty$

$n/d = \Theta(1)$

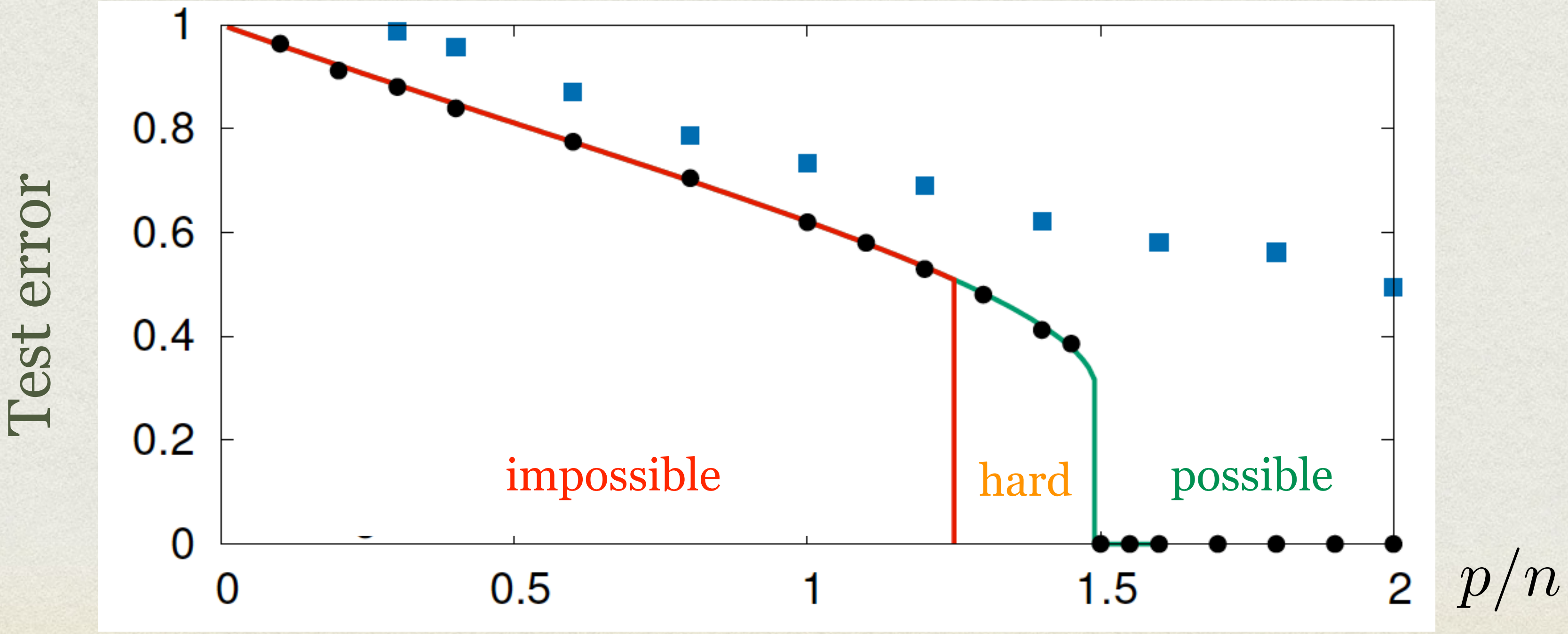


# of samples per dimension n/d

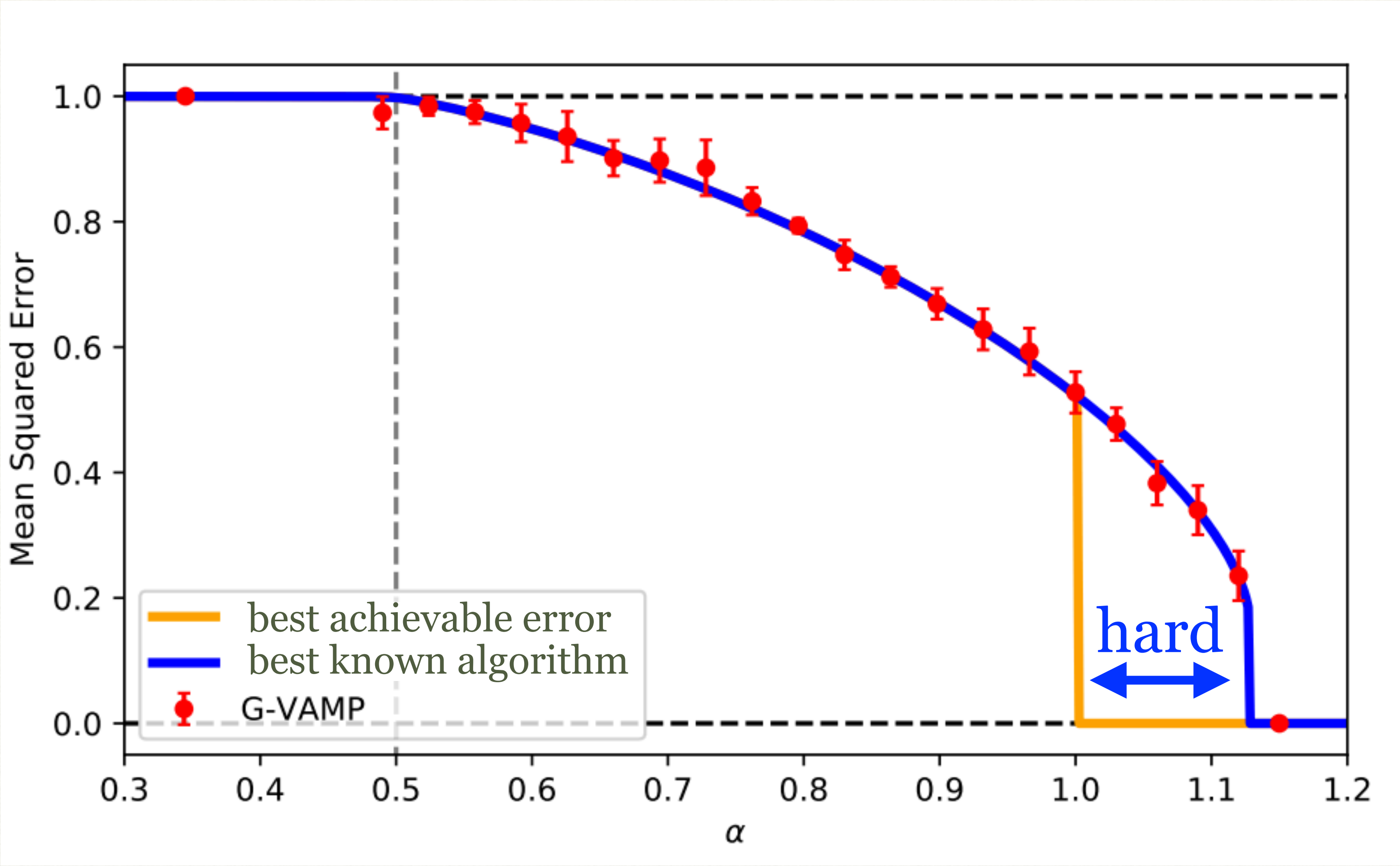
PHYSICS VS LEARNING



liquid	supercooled liquid	ice
impossible	computationally hard	possible



PHASE RETRIEVAL



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

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

$$X_{\mu i}, w_i \in \mathbb{R}$$

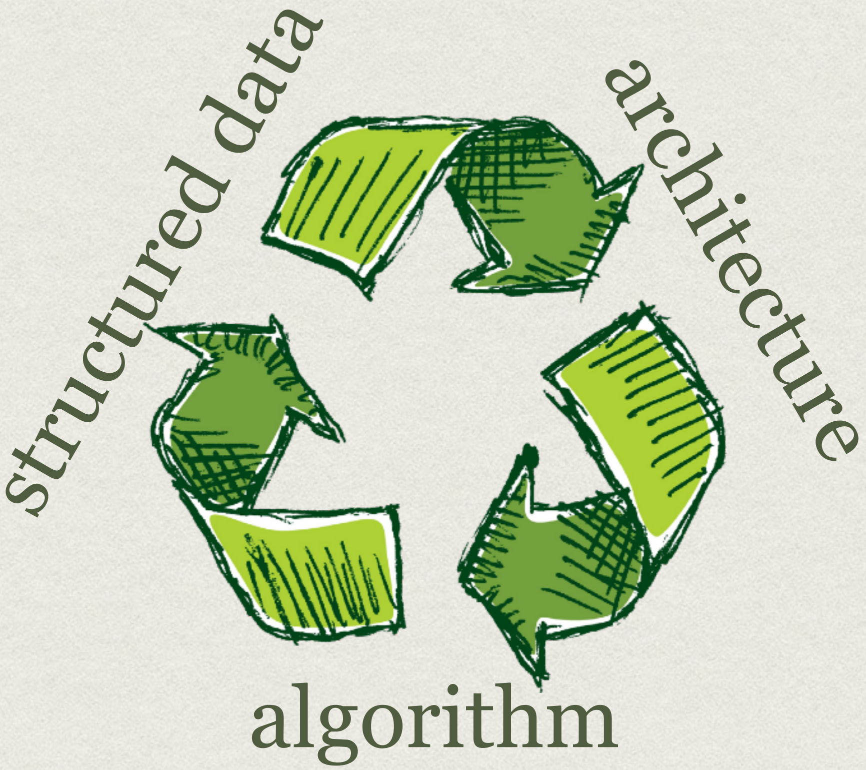
$$\alpha = \frac{n}{d}$$

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

$\alpha_{\text{AMP}} = 1.13$ # of samples needed for perfect generalisation for approximate message passing algorithm (conjectured optimal among polynomial ones).

Is this bringing us towards the theory of deep learning?

TOWARDS THEORY OF DEEP LEARNING?



structured data

architecture

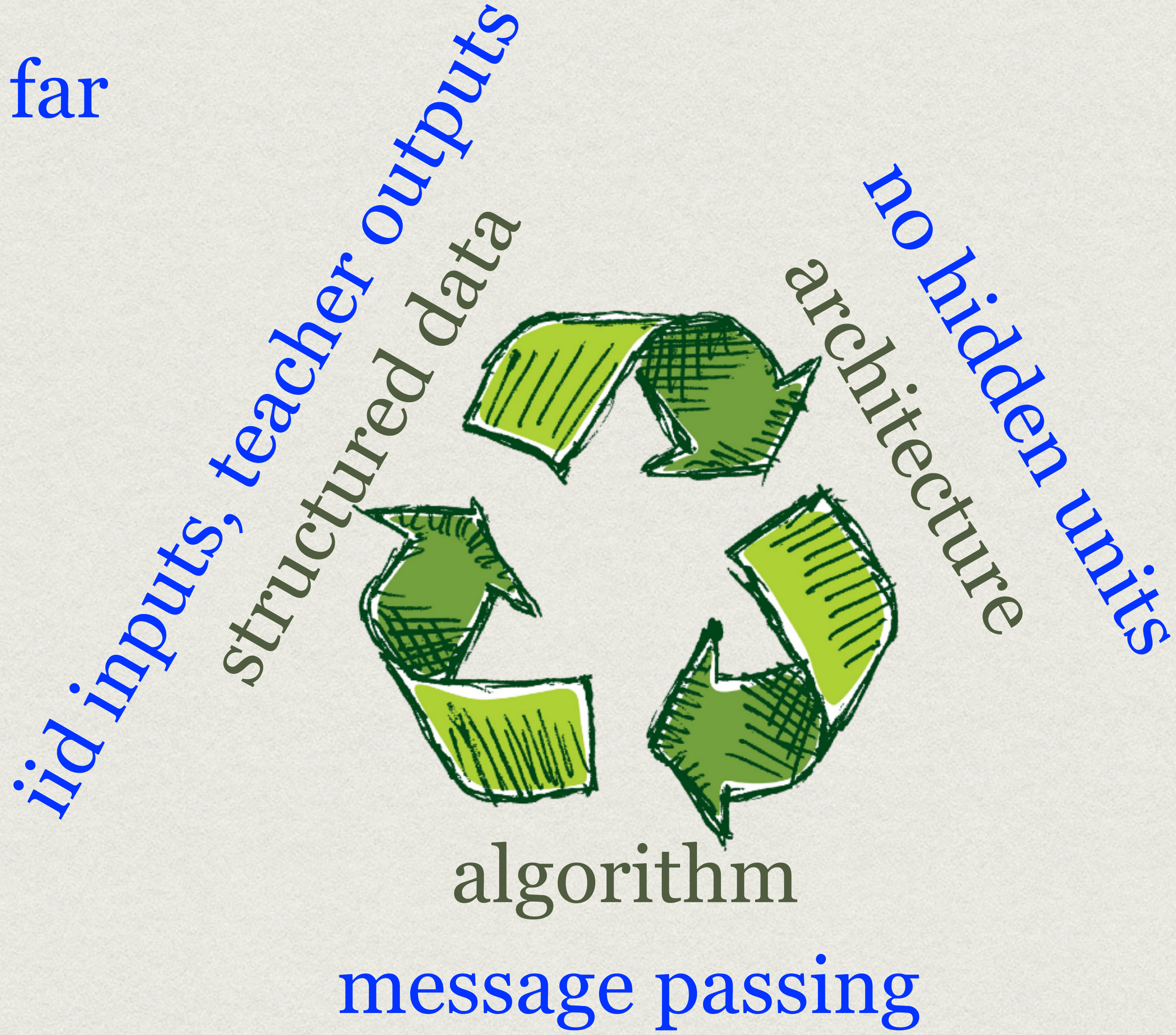
algorithm

TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed



iid inputs, teacher outputs

structured data

no hidden units

architecture

algorithm

message passing

TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed

structured data

iid inputs, teacher outputs

realistically structured data

architecture

no hidden units

multiple wide layers

algorithm

message passing

gradient-descent-based

TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed

realistically structured data

iid inputs, teacher outputs

structured data

multiple wide layers

no hidden units

architecture

algorithm

message passing

gradient-descent-based

ADDING HIDDEN UNITS

Aubin, Maillard, Barbier, Macris, Krzakala, LZ, NeurIPS'18, arXiv:1806.05451.

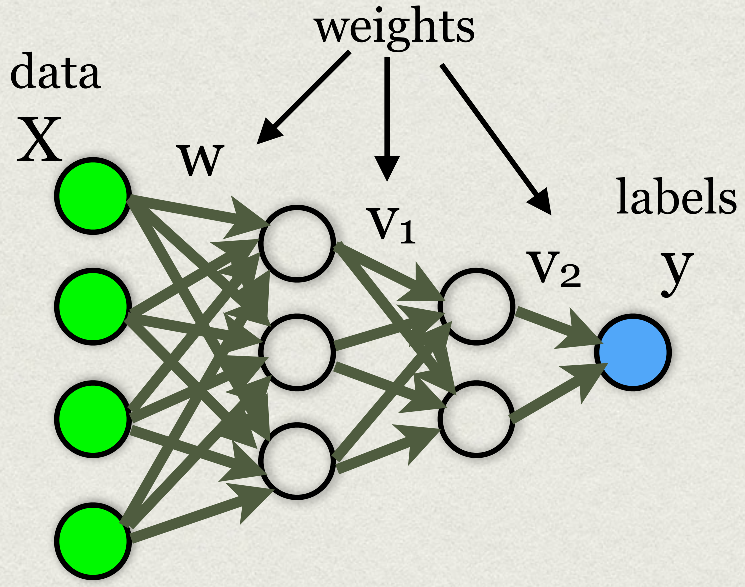
Committee machine

🟢 p input units

○ K hidden units

🔵 output unit

n training samples

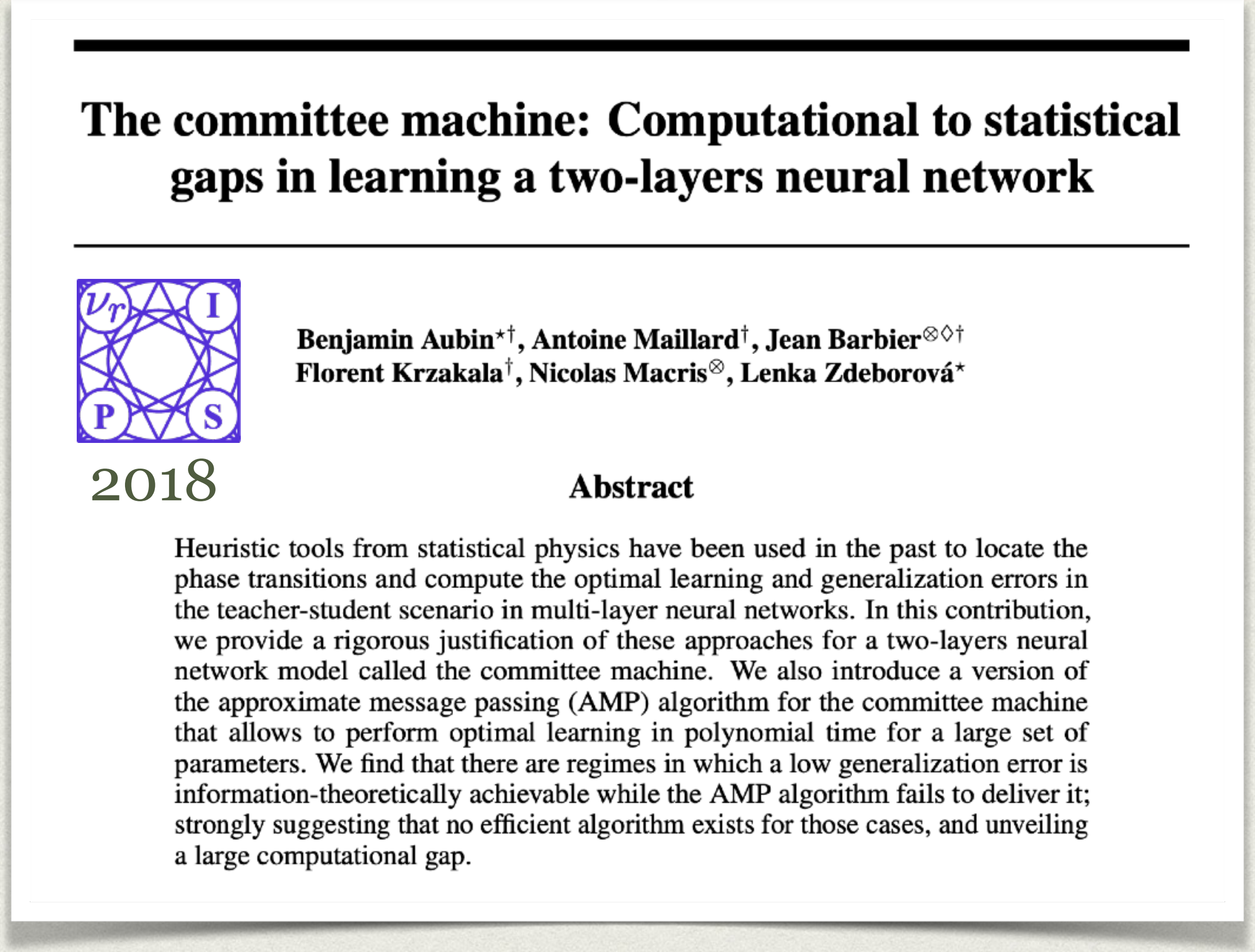


L=3 layers

w learned, v_1 & v_2 fixed

Limit: $n \to \infty$ $d \to \infty$ $\alpha = n/d = \Theta(1)$ $K = \Theta(1)$

Replica solution in Schwarze'92.



The committee machine: Computational to statistical gaps in learning a two-layers neural network

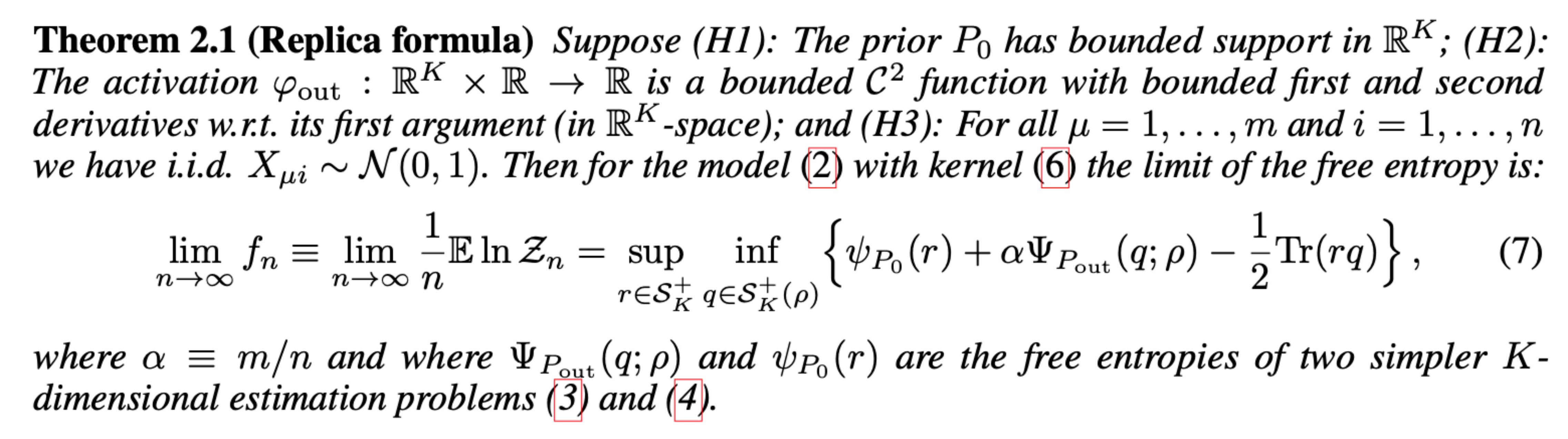
2018

Benjamin Aubin[⋆†], Antoine Maillard[†], Jean Barbier[⊗◇†]
Florent Krzakala[†], Nicolas Macris[⊗], Lenka Zdeborová[⋆]

Abstract

Heuristic tools from statistical physics have been used in the past to locate the phase transitions and compute the optimal learning and generalization errors in the teacher-student scenario in multi-layer neural networks. In this contribution, we provide a rigorous justification of these approaches for a two-layers neural network model called the committee machine. We also introduce a version of the approximate message passing (AMP) algorithm for the committee machine that allows to perform optimal learning in polynomial time for a large set of parameters. We find that there are regimes in which a low generalization error is information-theoretically achievable while the AMP algorithm fails to deliver it; strongly suggesting that no efficient algorithm exists for those cases, and unveiling a large computational gap.

Technical contribution: Approximate message passing and proof of the replica formula.



Theorem 2.1 (Replica formula) *Suppose (H1): The prior P_0 has bounded support in $\mathbb{R}^K$; (H2): The activation $\varphi_{\text{out}} : \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}$ is a bounded $\mathcal{C}^2$ function with bounded first and second derivatives w.r.t. its first argument (in $\mathbb{R}^K$-space); and (H3): For all $\mu = 1, \ldots, m$ and $i = 1, \ldots, n$ we have i.i.d. $X_{\mu i} \sim \mathcal{N}(0,1)$. Then for the model (2) with kernel (6) the limit of the free entropy is:*

$$\lim_{n\to\infty} f_n \equiv \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\ln \mathcal{Z}_n = \sup_{r\in\mathcal{S}_K^+} \inf_{q\in\mathcal{S}_K^+(\rho)} \left\{ \psi_{P_0}(r) + \alpha\Psi_{P_{\text{out}}}(q;\rho) - \frac{1}{2}\text{Tr}(rq) \right\}, \tag{7}$$

where $\alpha \equiv m/n$ and where $\Psi_{P_{\text{out}}}(q;\rho)$ and $\psi_{P_0}(r)$ are the free entropies of two simpler K-dimensional estimation problems (3) and (4).

SPECIALISATION TRANSITION

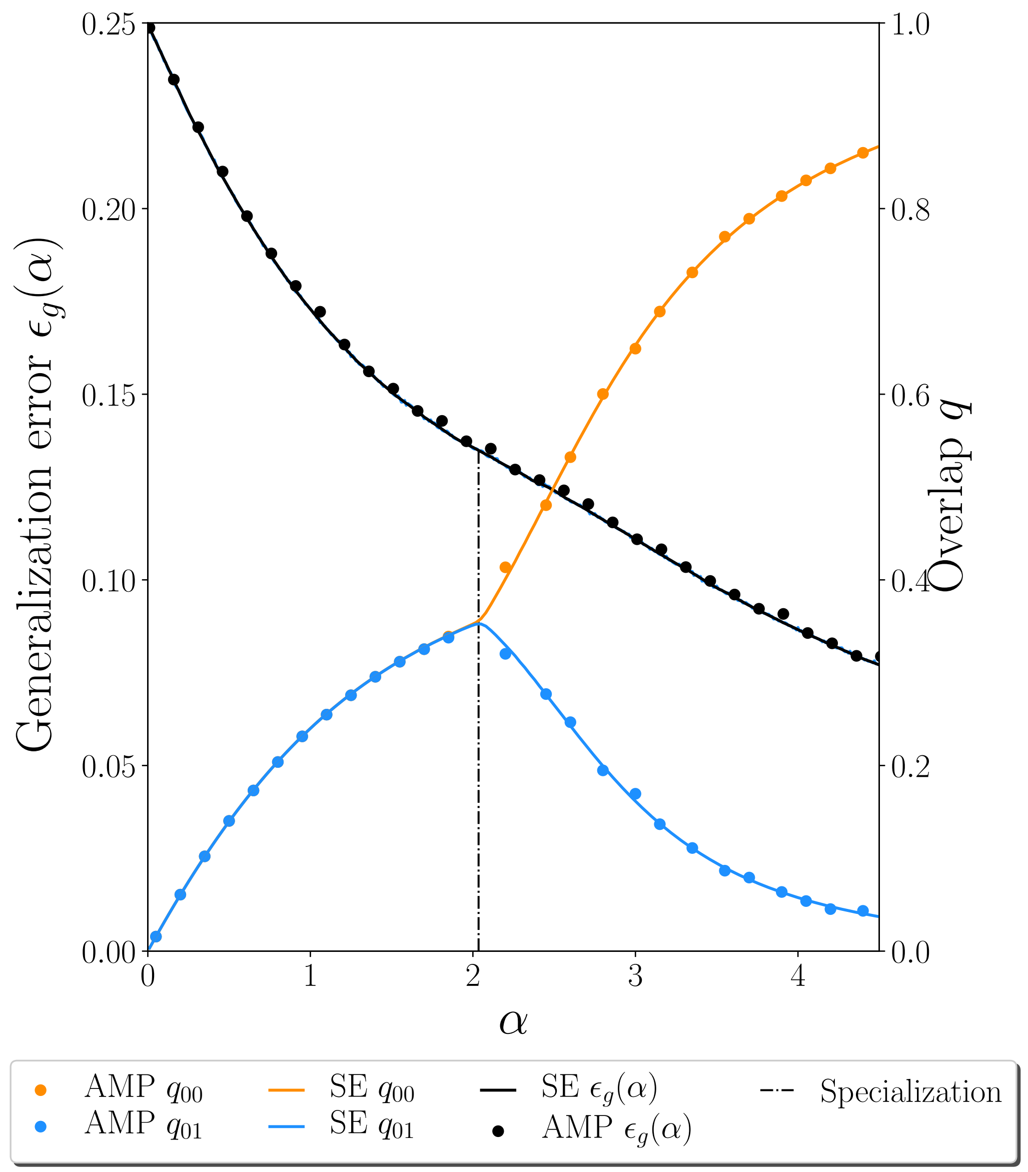
Aubin, Maillard, Barbier, Macris, Krzakala, LZ, NeurIPS'18, arXiv:1806.05451.

hidden units
K=2

$$y_\mu = \text{sign}\Big[\text{sign}\Big(\sum_i X_{\mu,i} w_{i,1}\Big) + \text{sign} \sum_i \big(X_{\mu,i} w_{i,2}\big)\Big]$$

- Specialization phase transition = hidden units specialise to correlate with specific features.

- Consequence: Sharp threshold for number of samples below which linear regression is the best thing to do.



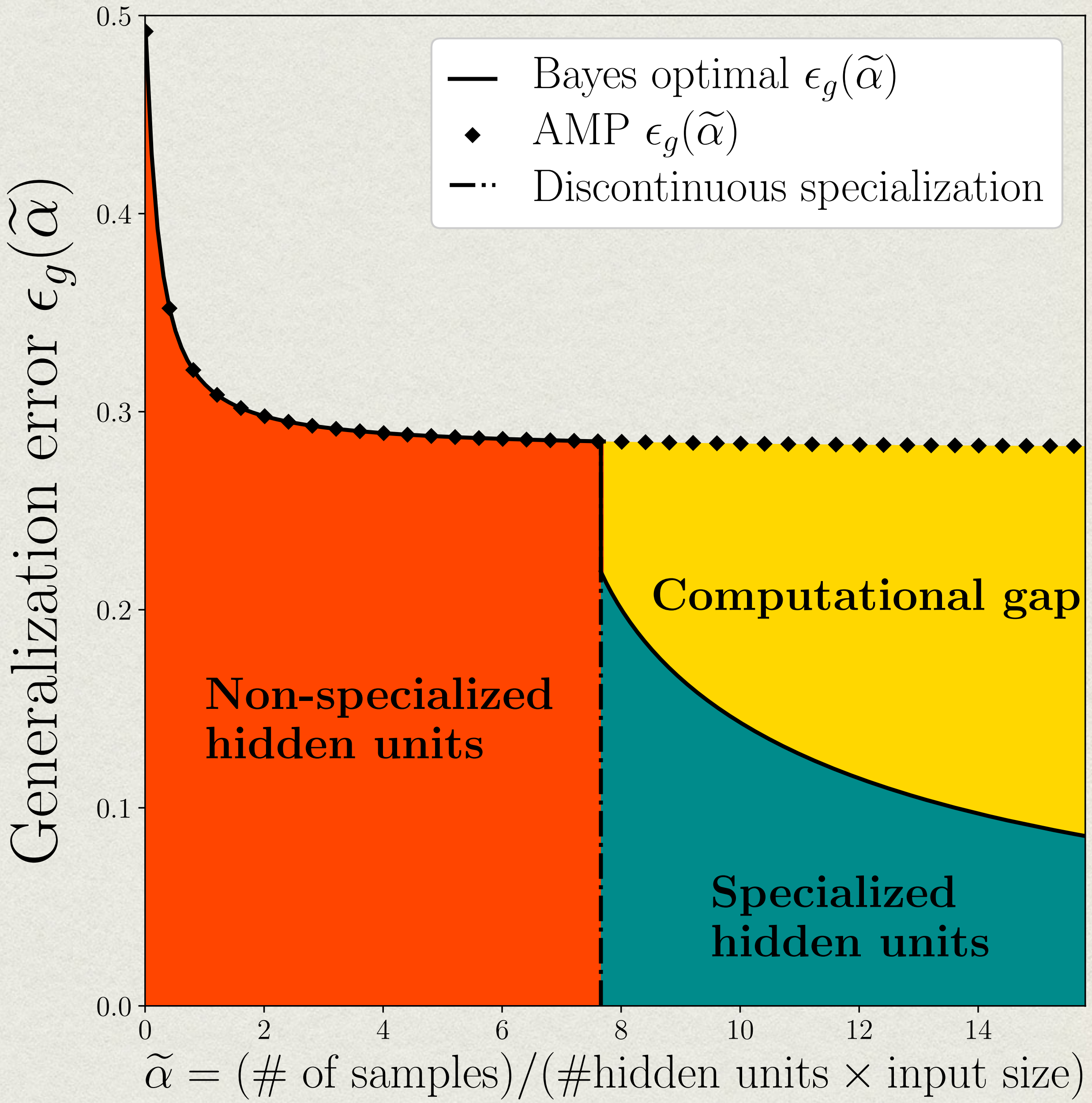
COMPUTATIONAL GAP

Aubin, Maillard, Barbier, Macris, Krzakala, LZ, NeurIPS'18, arXiv:1806.05451.

$$y_\mu = \mathrm{sign}\left[\sum_{a=1}^{K} \mathrm{sign}\left(\sum_i X_{\mu,i} w_{i,a}\right)\right]$$

hidden units $K \gg 1$

- Large algorithmic gap:
  - IT threshold: $n > 7.65Kd$
  - Algorithmic threshold $n > \mathrm{const}\,.\,K^2 d$



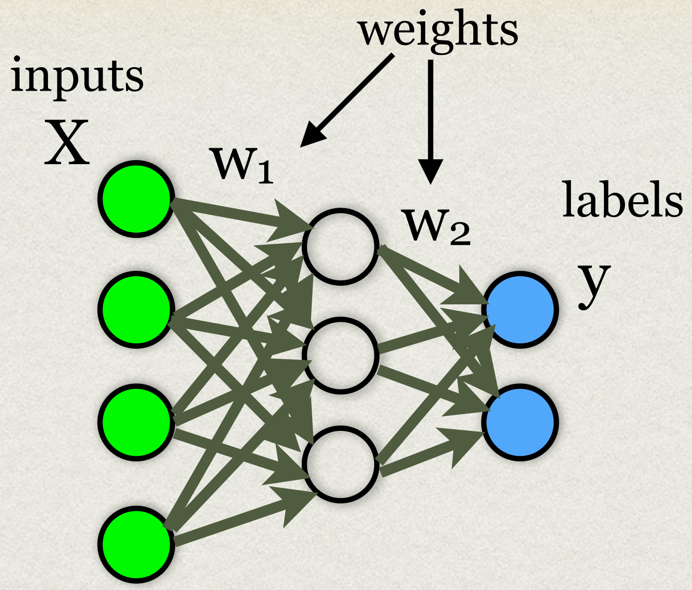
OPEN PROBLEM

🟢 p # input units

○ k # hidden units

🔵 m # output units

n training samples



2 layers
w_1 & w_2 learned

Limit: $n \to \infty$ $\quad p \to \infty$ $\quad k \to \infty$ $\quad m \to \infty$

$$n/p = \Theta(1)$$
$$k/p = \Theta(1)$$
$$m/p = \Theta(1)$$

iid inputs X, iid teacher weights w_1^* and w_2^*, generate output y.

Open question: Optimal generalisation error of the student network?

No known (even heuristic) formula.

TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed

structured data

iid inputs, teacher outputs

realistically structured data

architecture

O(1) hidden units*

multiple wide layers

algorithm

message passing

gradient-descent based

* see also: deep linear networks, and infinitely wide ones.

TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed

realistically structured data

iid inputs, teacher outputs

structured data

architecture

O(1) hidden units*

multiple wide layers

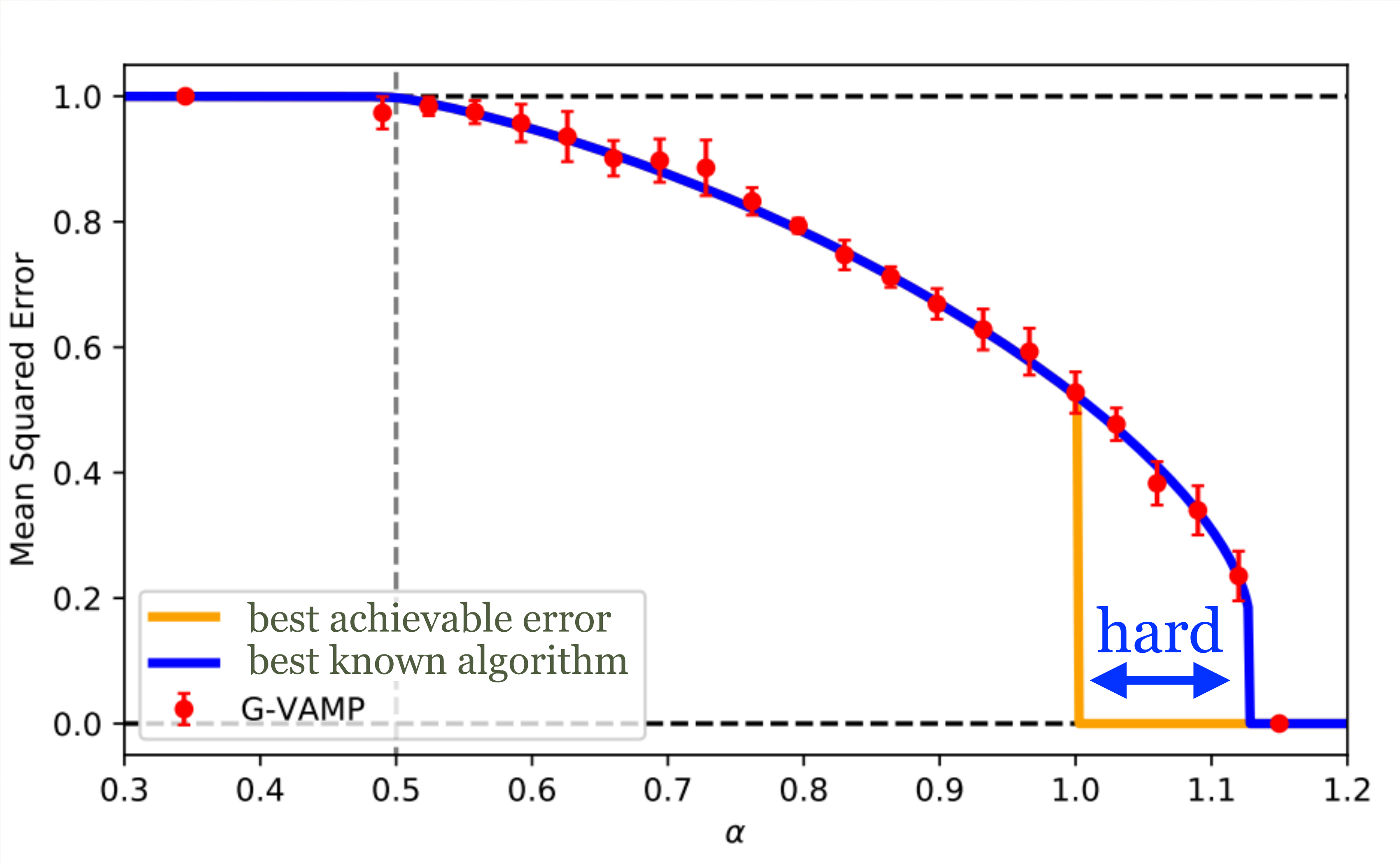
algorithm

message passing

gradient-descent based

* see also: deep linear networks, and infinitely wide ones.

PHASE RETRIEVAL



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

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

$$\alpha = \frac{n}{d}$$

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

$\alpha_{\text{AMP}} = 1.13$ # of samples needed for perfect generalisation for approximate message passing algorithm (conjectured optimal among polynomial ones).

GRADIENT DESCENT FOR PHASE RETRIEVAL

Loss function:

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

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

Gradient flow:

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

ensuring $\|w\|_2^2 = d$

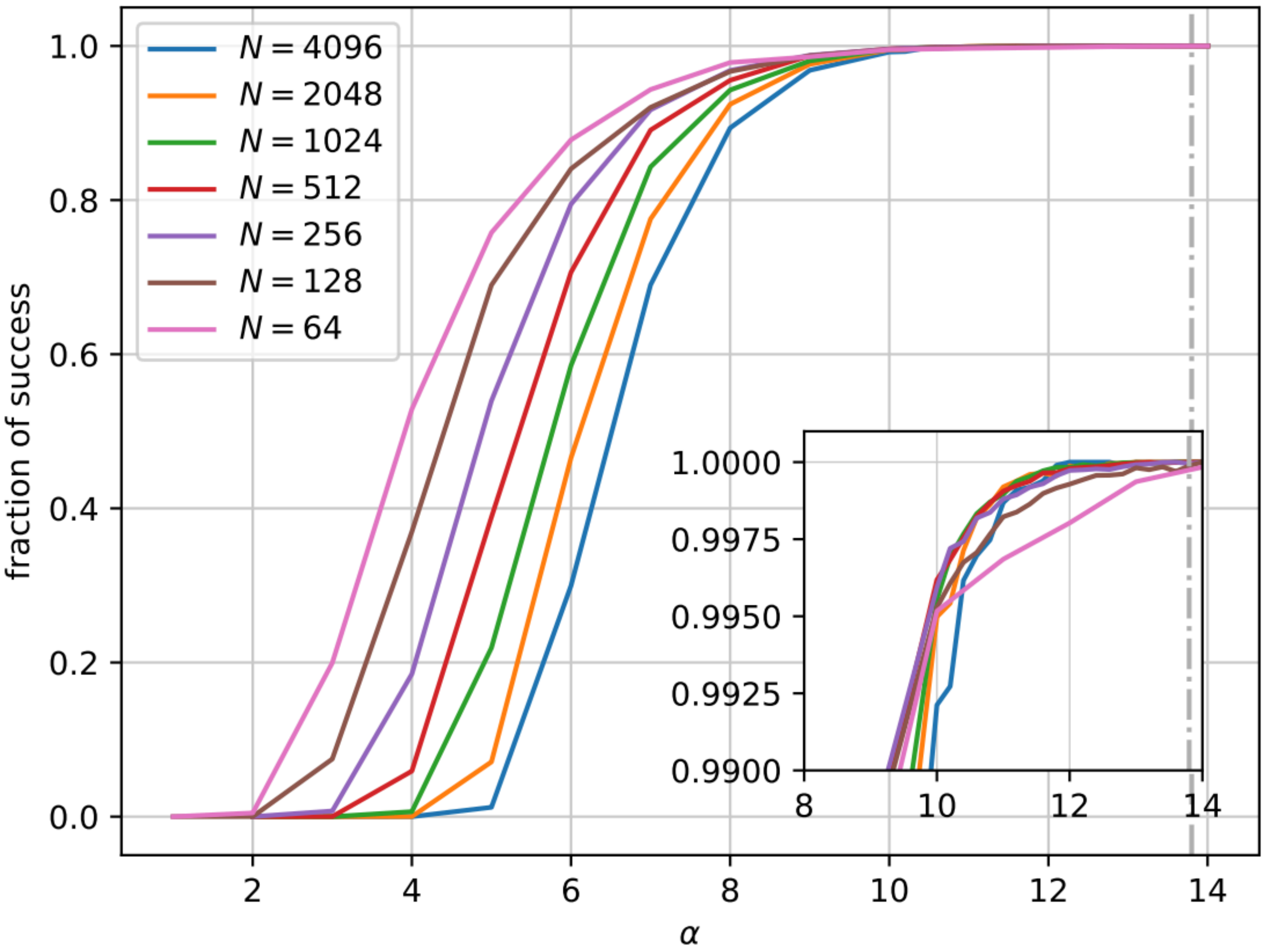
Initialisation:

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

GRADIENT DESCENT NUMERICALLY

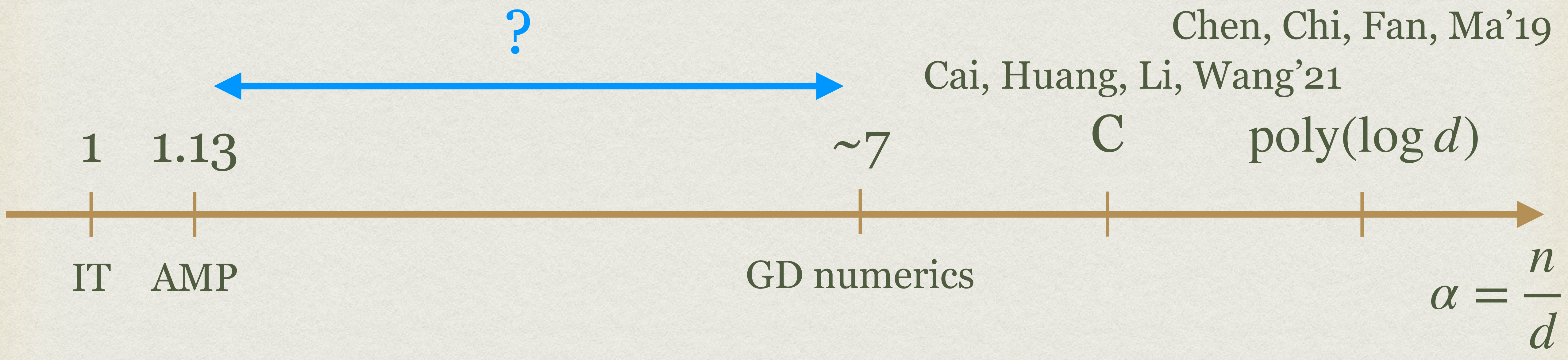
Sarao Mannelli, Biroli, Cammarota, Krzakala, LZ, 2006.06997.

$N = d$



PERFORMANCE OF GRADIENT DESCENT

Closing the gap between GD and AMP?



?

Chen, Chi, Fan, Ma’19

Cai, Huang, Li, Wang’21

1 1.13 ~7 C poly(log d)

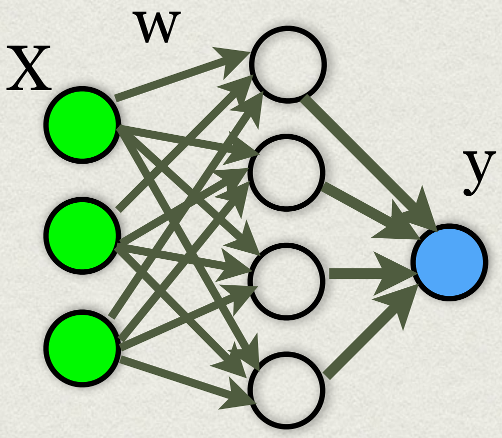
IT AMP GD numerics

$\alpha = \frac{n}{d}$

OVER-PARAMETRISATION & GRADIENT DESCENT

GRADIENT DESCENT FOR PHASE RETRIEVAL

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



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

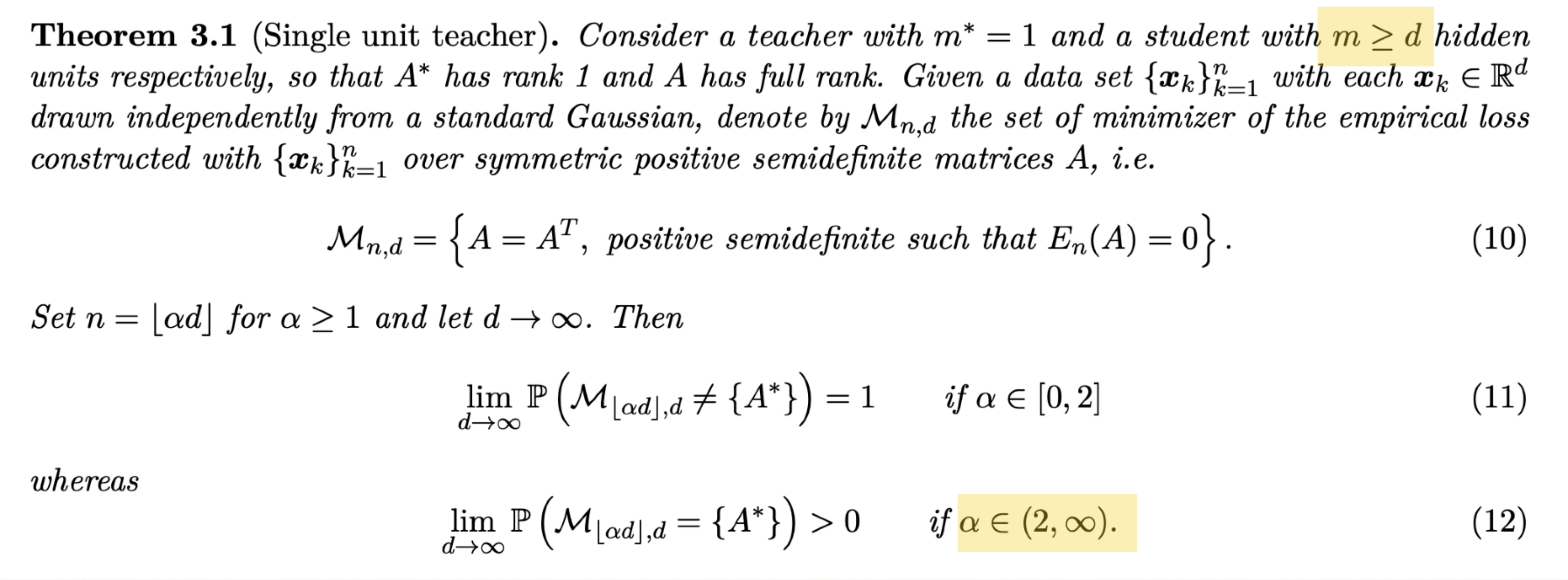
Wide (m>d) over-parametrised two-layer neural network

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

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

OVER-PARAMETRISED LANDSPACE

Sarao Mannelli, Vanden-Eijnden, LZ, 2006.15459



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.*

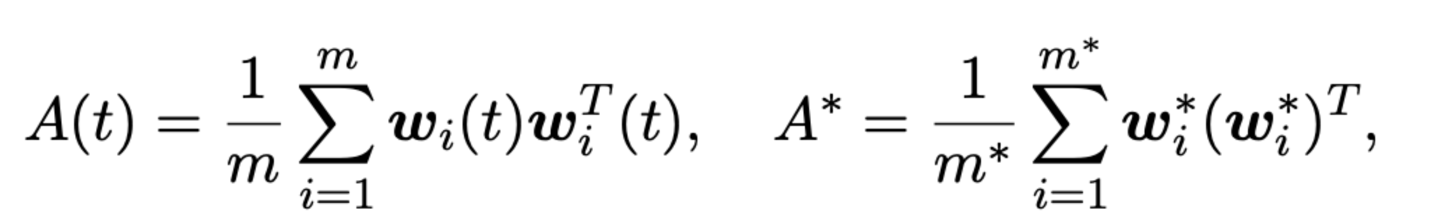
$$\mathcal{M}_{n,d} = \left\{ A = A^T, \textit{ positive semidefinite such that } E_n(A) = 0 \right\}. \tag{10}$$

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

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

whereas

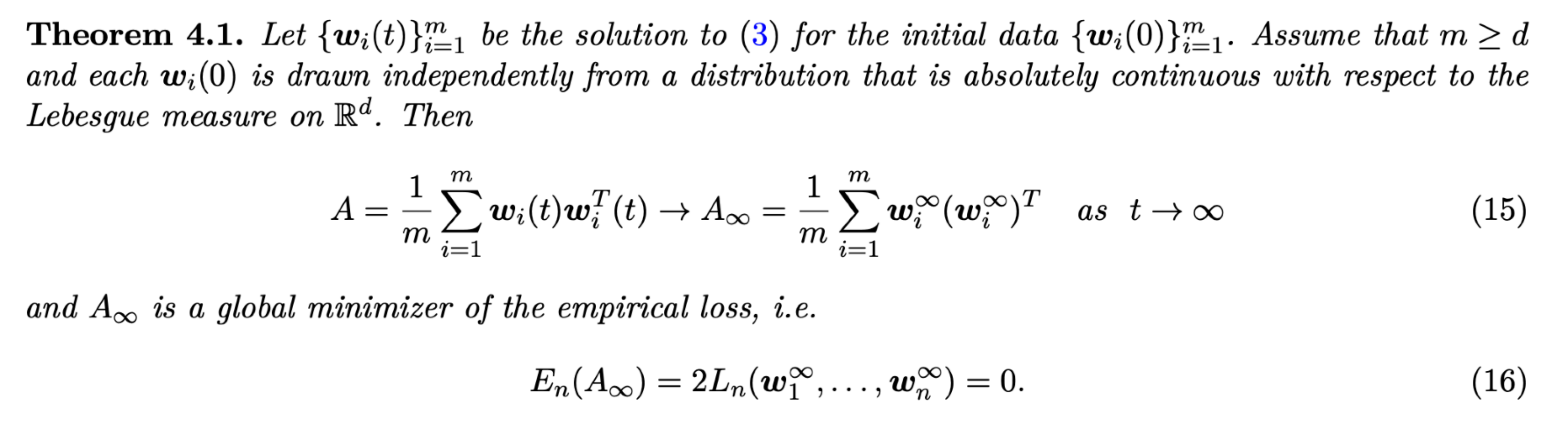
$$\lim_{d\to\infty} \mathbb{P}\left(\mathcal{M}_{\lfloor \alpha d \rfloor, d} = \{A^*\} \right) > 0 \qquad \textit{if } \alpha \in (2, \infty). \tag{12}$$



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

GD FOR OVER-PARAMETRISED PHASE RETRIEVAL

Sarao Mannelli, Vanden-Eijnden, LZ, 2006.15459

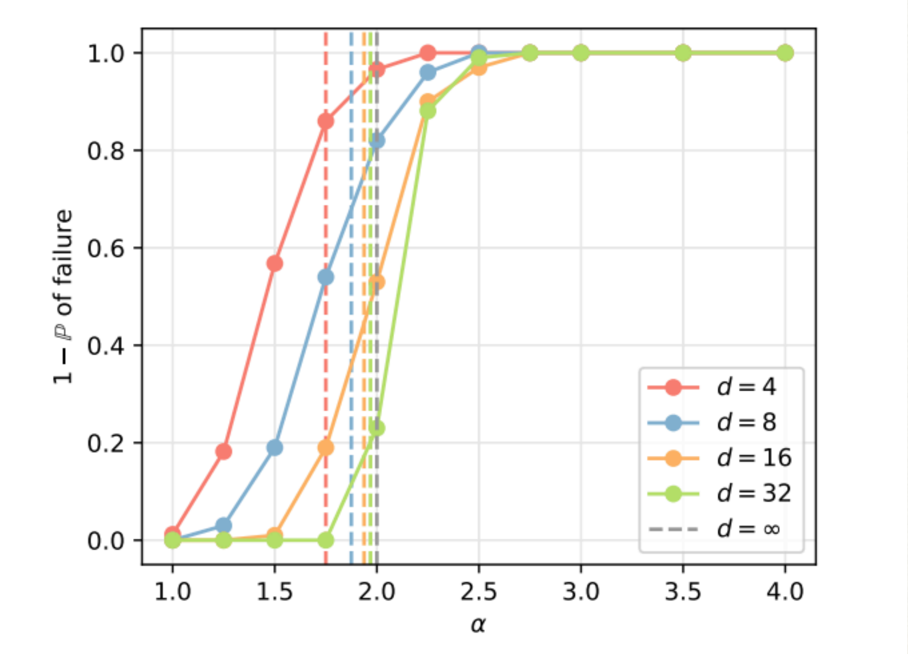


Theorem 4.1. *Let $\{\boldsymbol{w}_i(t)\}_{i=1}^m$ be the solution to (3) for the initial data $\{\boldsymbol{w}_i(0)\}_{i=1}^m$. Assume that $m \geq d$ and each $\boldsymbol{w}_i(0)$ is drawn independently from a distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$.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

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

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

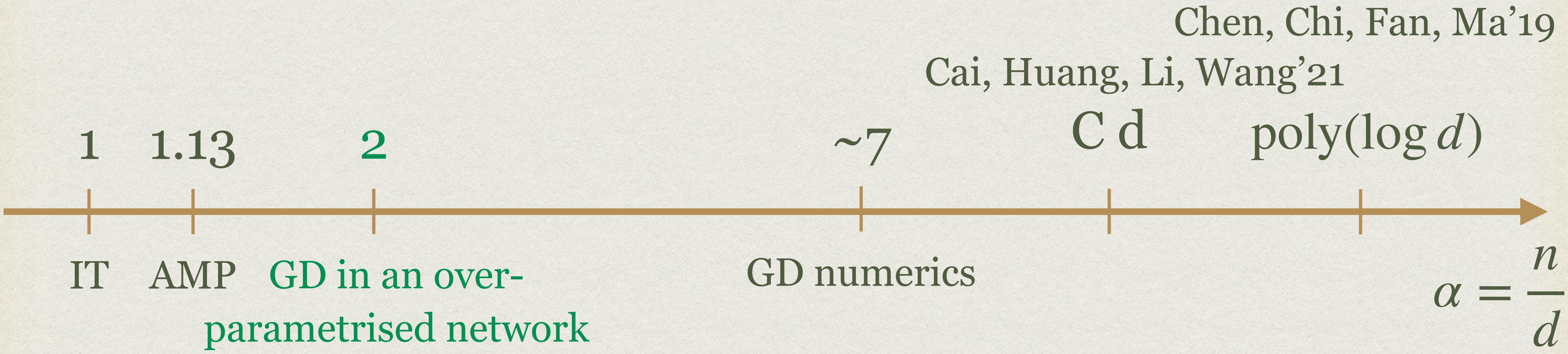
$$E_n(A_\infty) = 2L_n(\boldsymbol{w}_1^\infty, \ldots, \boldsymbol{w}_n^\infty) = 0. \tag{16}$$



PERFORMANCE OF GRADIENT DESCENT

Sarao Mannelli, Vanden-Eijnden, LZ, 2006.15459

Over-parametrised neural networks need fewer samples to learn



Chen, Chi, Fan, Ma'19

Cai, Huang, Li, Wang'21

1 — IT

1.13 — AMP

2 — GD in an over-parametrised network

~7 — GD numerics

C d

poly(log d)

$\alpha = \frac{n}{d}$

ANALYSIS OF GRADIENT-BASED ALGORITHM IN NON-CONVEX HIGH-DIMENSIONAL PROBLEMS

- Sarao Mannelli, Biroli, Cammarota, Krzakala, Urbani, LZ; *Marvels and Pitfalls of the Langevin Algorithm in Noisy High-dimensional Inference;* Phys. Rev. X'20, arXiv:1812.09066.
- Sarao Mannelli, Krzakala, Urbani, LZ; *Passed & Spurious: Descent Algorithms and Local Minima in Spiked Matrix-Tensor Models;* ICML'19, arXiv:1902.00139.
- Sarao Mannelli, Biroli, Cammarota, Krzakala, LZ; *Who is Afraid of Big Bad Minima? Analysis of Gradient-Flow in a Spiked Matrix-Tensor Model*; NeurIPS'19, arXiv:1907.08226.

- Mignacco, Urbani, Krzakala, LZ; *Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification*; NeurIPS'20, arXiv:2006.06098.

- Sarao Mannelli, Biroli, Cammarota, Krzakala, Urbani, LZ; *Complex Dynamics and Simple Neural Networks: Understanding Gradient Flow in Phase Retrieval,* NeurIPS'20, arXiv:2006.06997.
- Mignacco, Urbani, LZ; *Stochasticity helps to navigate rough landscapes: comparing gradient-descent-based algorithms in the phase retrieval problem, MLST,* arXiv:2103.04902.

TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed

structured data

iid inputs, teacher outputs

realistically structured data

architecture

O(1) hidden units*

multiple wide layers

algorithm

gradient-descent based

* see also: deep linear networks, and infinitely wide ones.

TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed

realistically structured data

iid inputs, teacher outputs

structured data

multiple wide layers

O(1) hidden units*

architecture

algorithm

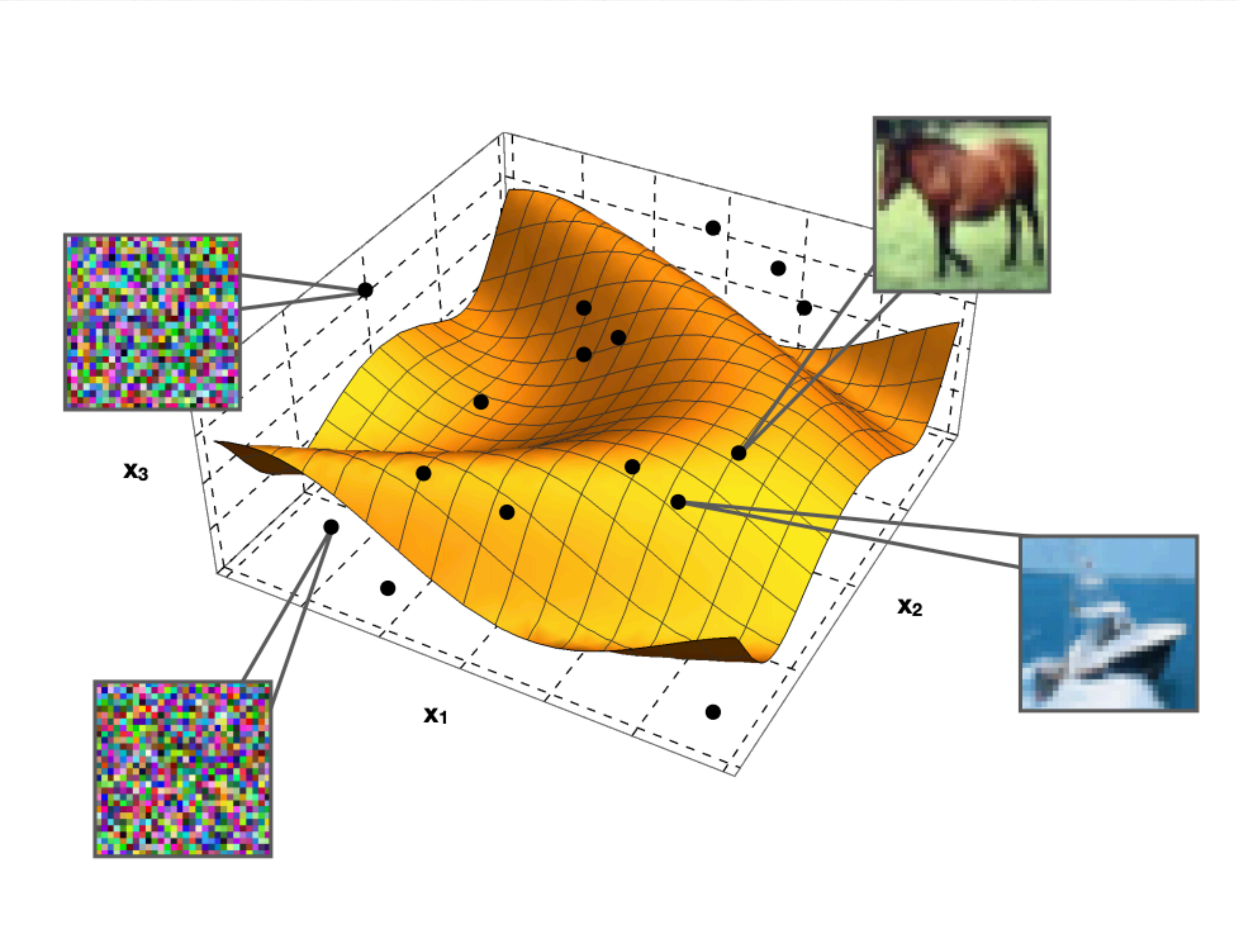
gradient-descent based

* see also: deep linear networks, and infinitely wide ones.

GANs generated photos of people.

DATA ON MANIFOLDS

- Real input data lie of low-dimensional manifolds; they can be generated by GANs and VAEs with small input dimension.



HIDDEN MANIFOLD MODEL

Goldt, FK, Mézard, LZ; arXiv:1909.11500

- Real input data lie of low-dimensional manifolds; they can be generated by GANs and VAEs with small input dimension.

- Hidden manifold model (C random iid matrix, F generic).

$$X_\mu = f(FC_\mu) \qquad y_\mu = g(C_\mu)$$

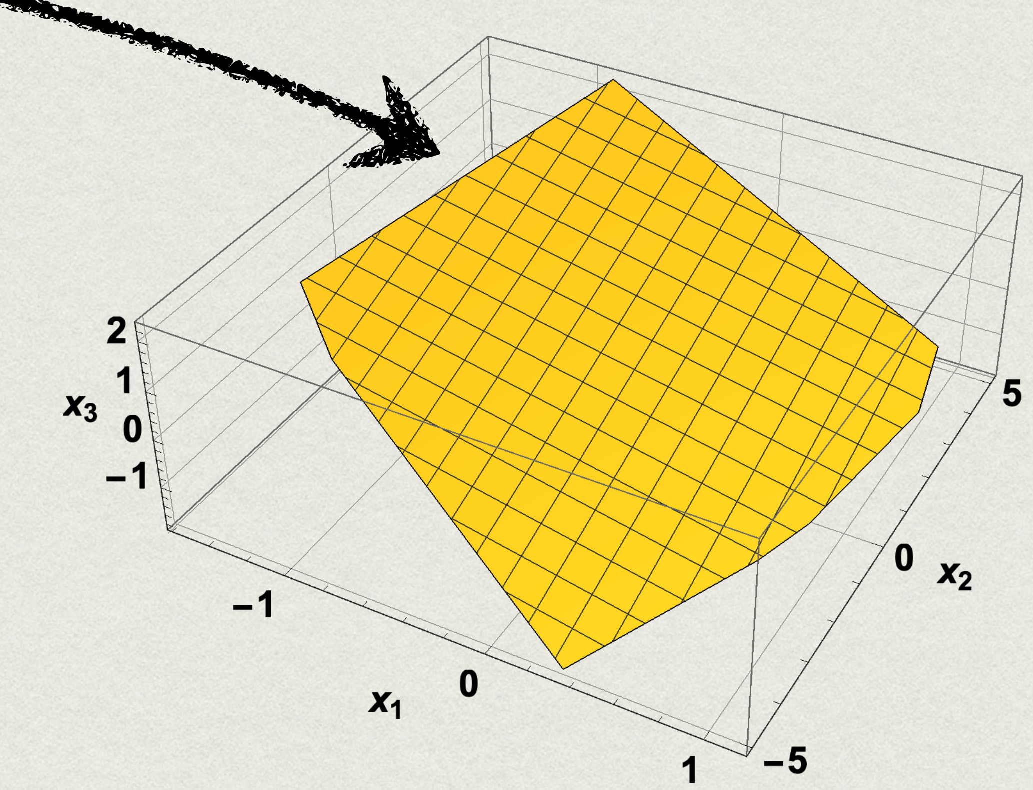
$X_\mu \in \mathbb{R}^p \quad C_\mu \in \mathbb{R}^d \quad F \in \mathbb{R}^{p\times d}$

p input & d latent dimension, p>d.

Hidden manifold model

low-dimensional
sub-space

F

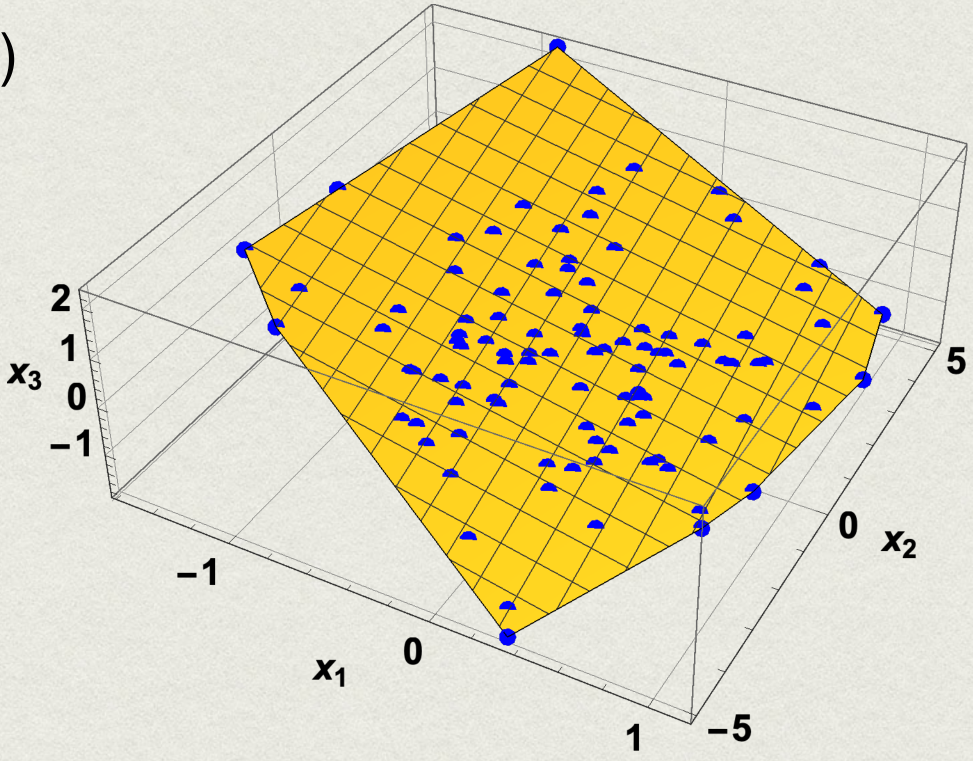


Hidden manifold model

low-dimensional
sub-space

point coordinates
in sub-space
(dimension d)

FC

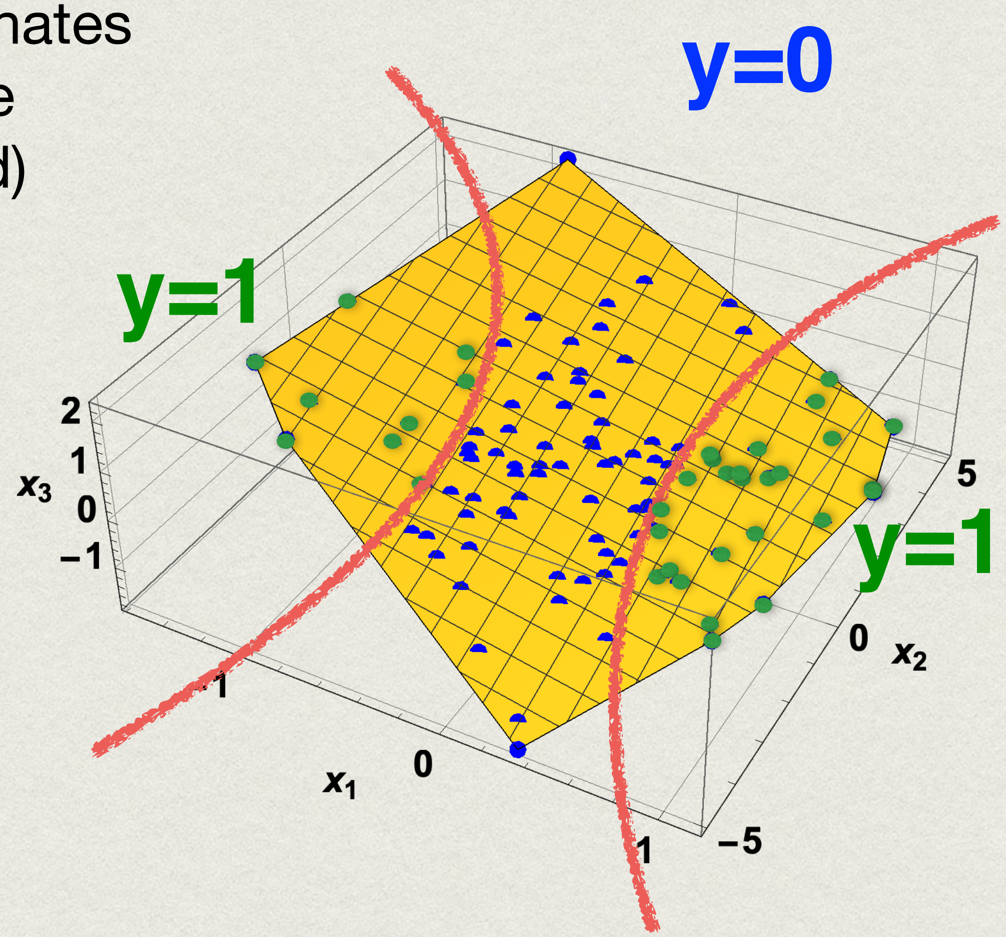


Hidden manifold model

low-dimensional
sub-space

point coordinates
in sub-space
(dimension d)

FC



$$Y = g(C)$$

Key: The true labels depend **only** on the latent representation of the point!

Hidden manifold model

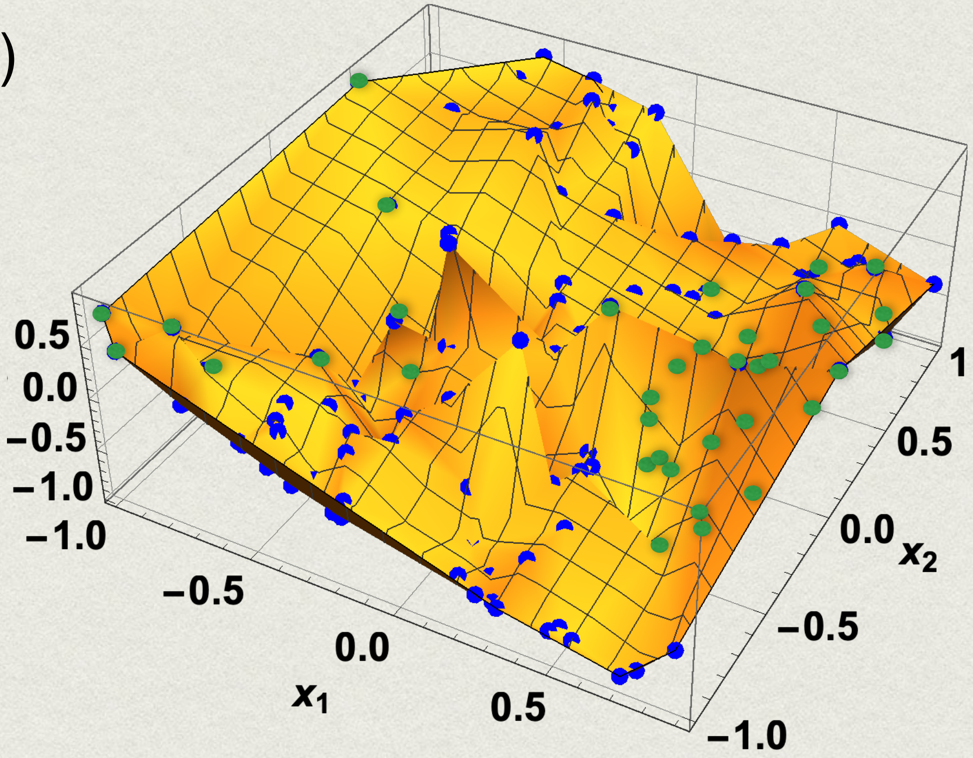
low-dimensional sub-space

point coordinates in sub-space (dimension d)

$$X = f(FC)$$

Data points (dimension p)

Non-linear function



$$Y = g(C)$$

Key: The true labels depend **only** on the latent representation of the point!

GAUSSIAN EQUIVALENCE

In the limit $p, n, d \to \infty$, while $n/p = \Theta(1)$ and $d/p = \Theta(1)$, generalisation error of the committee machine for

$$X_\mu = f(FC_\mu) \qquad y_\mu = g(C_\mu) \qquad X_\mu \in \mathbb{R}^p \quad C_\mu \in \mathbb{R}^d \quad F \in \mathbb{R}^{p\times d}$$

is the same as the one of

$$X_\mu = \kappa_1 F C_\mu + \kappa_* \mathcal{N}(0, \mathbb{I}_p) + \kappa_0 \mathbb{I}_p \qquad y_\mu = g(C_\mu)$$

$$\kappa_0 = \mathbb{E}\left[f(z)\right], \kappa_1 \equiv \mathbb{E}\left[zf(z)\right], \kappa_\star \equiv \mathbb{E}\left[f(z)^2\right] - \kappa_0^2 - \kappa_1^2$$

Formally: Goldt, FK, Mézard, Reeves, LZ, arXiv:2006.14709

Replica solution

Solution:

Consider the unique fixed point of the following system of equations

$$\begin{cases} \hat{V}_s = \frac{\alpha}{\gamma}\kappa_1^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\partial_\omega \eta(y,\omega_1)}{V}\right], \\ \hat{q}_s = \frac{\alpha}{\gamma}\kappa_1^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\left(\eta(y,\omega_1)-\omega_1\right)^2}{V^2}\right], \\ \hat{m}_s = \frac{\alpha}{\gamma}\kappa_1 \mathbb{E}_{\xi,y}\left[\partial_\omega\mathcal{Z}\left(y,\omega_0\right)\frac{\left(\eta(y,\omega_1)-\omega_1\right)}{V}\right], \\ \hat{V}_w = \alpha\kappa_\star^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\partial_\omega \eta(y,\omega_1)}{V}\right], \\ \hat{q}_w = \alpha\kappa_\star^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\left(\eta(y,\omega_1)-\omega_1\right)^2}{V^2}\right], \end{cases}$$

$$\begin{cases} V_s = \frac{1}{\hat{V}_s}\left(1 - z\ g_\mu(-z)\right), \\ q_s = \frac{\hat{m}_s^2+\hat{q}_s}{\hat{V}_s}\left[1-2zg_\mu(-z)+z^2g'_\mu(-z)\right] \\ \qquad -\frac{\hat{q}_w}{(\lambda+\hat{V}_w)\hat{V}_s}\left[-zg_\mu(-z)+z^2g'_\mu(-z)\right], \\ m_s = \frac{\hat{m}_s}{\hat{V}_s}\left(1 - z\ g_\mu(-z)\right), \\ V_w = \frac{\gamma}{\lambda+\hat{V}_w}\left[\frac{1}{\gamma}-1+zg_\mu(-z)\right], \\ q_w = \gamma\frac{\hat{q}_w}{(\lambda+\hat{V}_w)^2}\left[\frac{1}{\gamma}-1+z^2g'_\mu(-z)\right], \\ \qquad +\frac{\hat{m}_s^2+\hat{q}_s}{(\lambda+\hat{V}_w)\hat{V}_s}\left[-zg_\mu(-z)+z^2g'_\mu(-z)\right], \end{cases}$$

$$\begin{cases} \eta(y,\omega) = \underset{x\in\mathbb{R}}{\mathrm{argmin}}\left[\frac{(x-\omega)^2}{2V}+\ell(y,x)\right] \\ \mathcal{Z}(y,\omega) = \int\frac{dx}{\sqrt{2\pi V^0}}e^{-\frac{1}{2V^0}(x-\omega)^2}\delta\left(y-f^0(x)\right) \end{cases}$$

where $V=\kappa_1^2V_s+\kappa_\star^2V_w$, $V^0=\rho-\frac{M^2}{Q}$, $Q=\kappa_1^2q_s+\kappa_\star^2q_w$, $M=\kappa_1m_s$, $\omega_0=M/\sqrt{Q}\xi$, $\omega_1=\sqrt{Q}\xi$ and $g_\mu$is the Stieltjes transform of FF^T

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

Then in the high-dimensional limit:

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

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

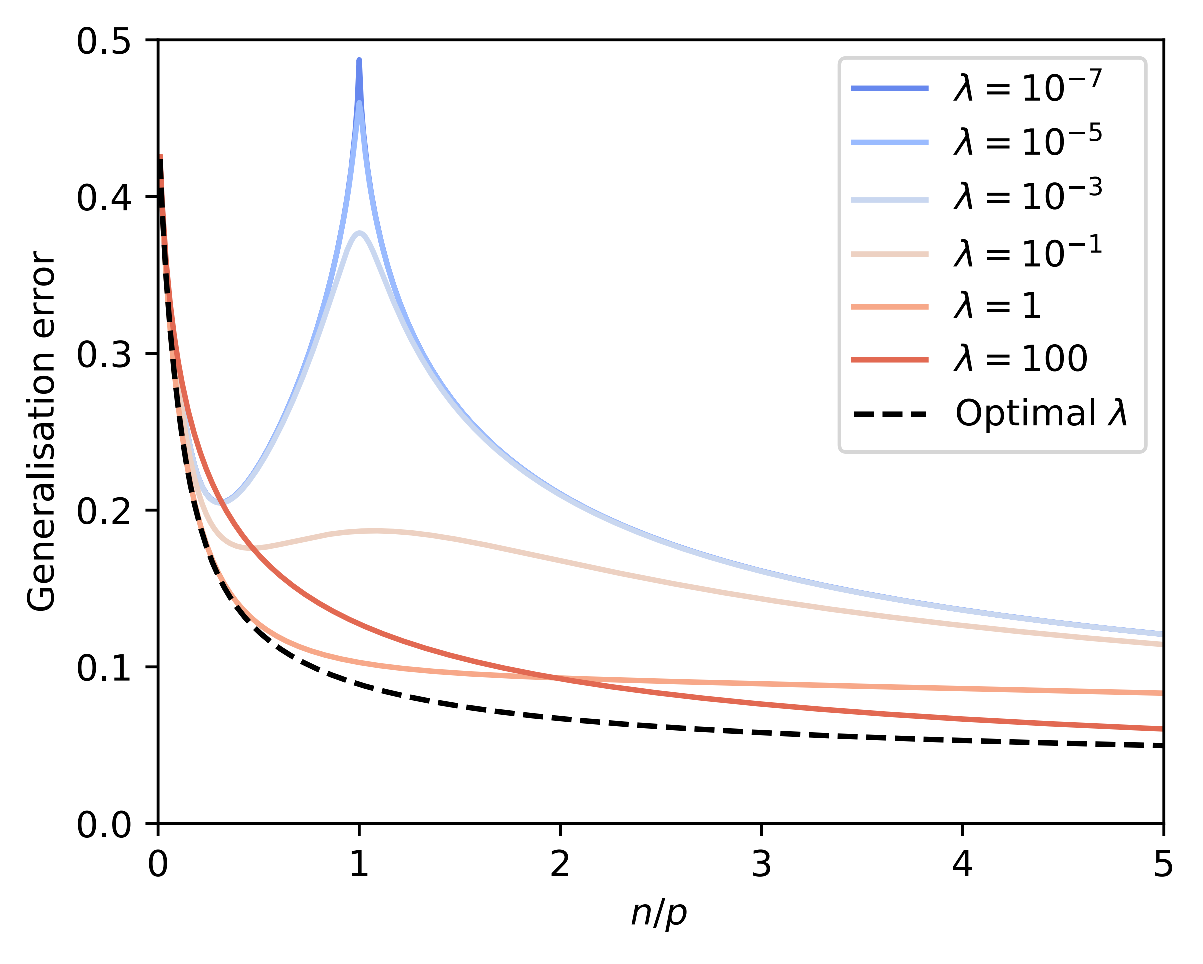
$$\mathcal{L}_{\text{training}}=\frac{\lambda}{2\alpha}q_w^\star+\mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0^\star\right)\ell\left(y,\eta(y,\omega_1^\star)\right)\right]$$

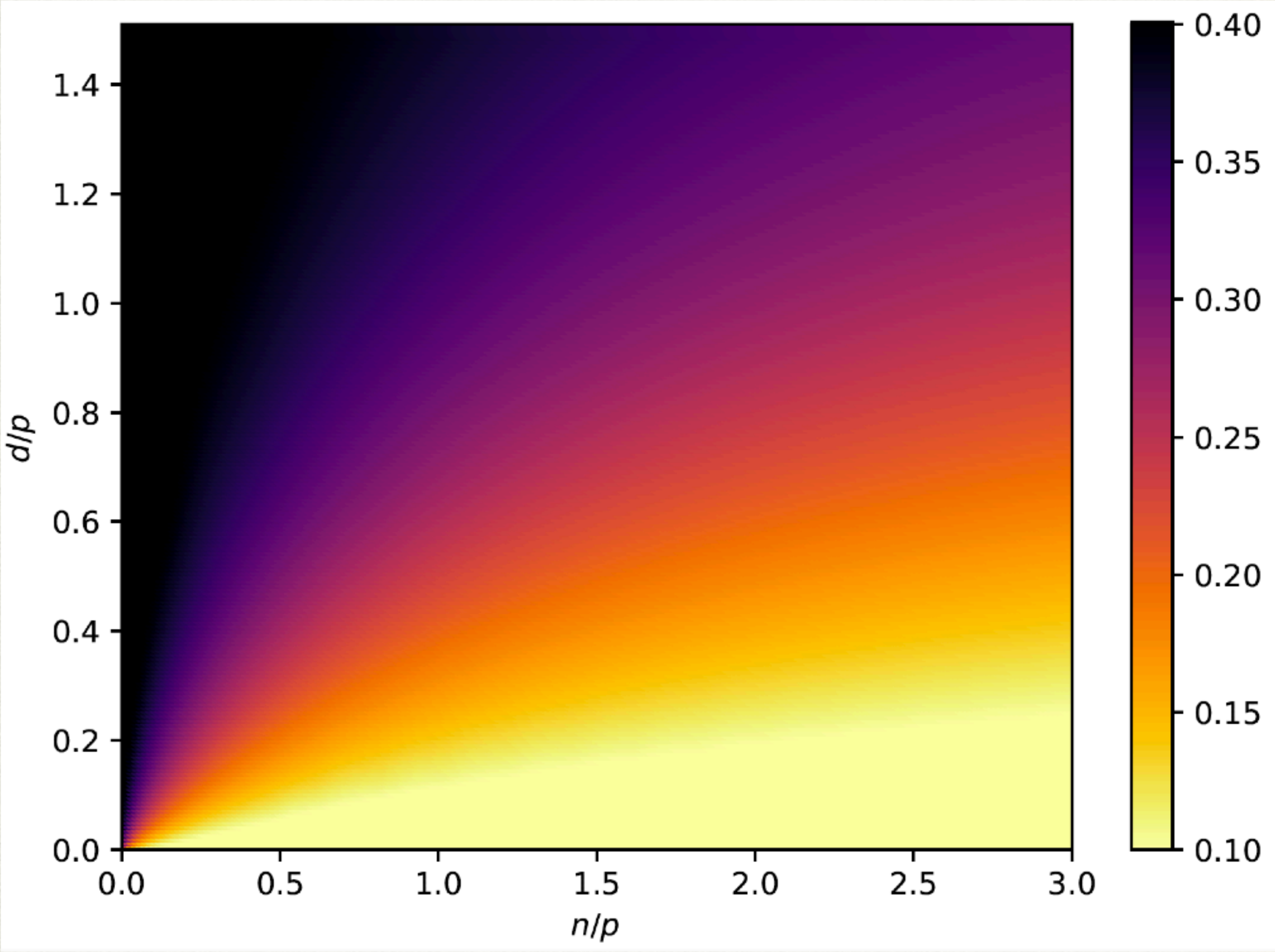
with $\omega_0^\star=M^\star/\sqrt{Q^\star}\xi, \omega_1^\star=\sqrt{Q^\star}\xi$

[Gerace, Loureiro, FK, Mezard, LZ, ICML, 2002.09339],

PHASE DIAGRAM

$X_\mu = \mathrm{erf}(FC_\mu)$ $y_\mu = \mathrm{sign}(C_\mu \cdot w^0)$ classification, least-squares loss

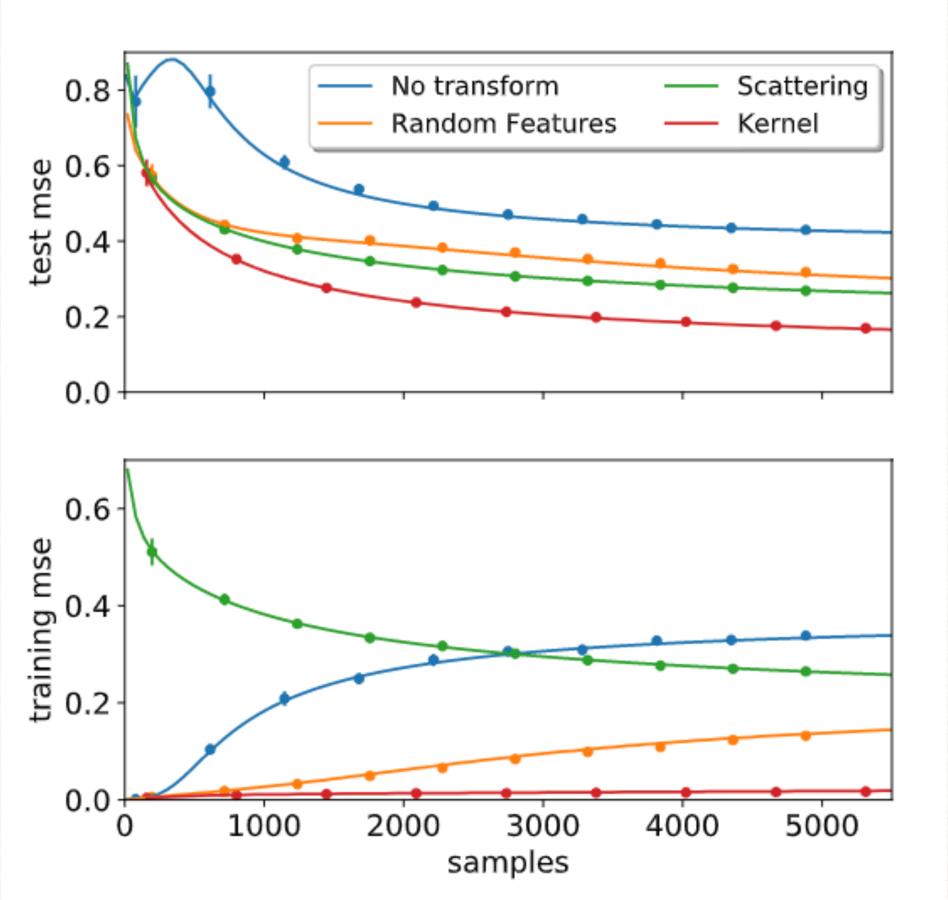




d/p=0.1

CAPTURING LEARNING CURVES OF REAL DATA

- Loureiro, Sicuro, Gerbelot, Pacco, Krzakala, LZ, *Learning curves of generic features maps for realistic datasets with a teacher-student model;* arXiv:2102.08127.
- Loureiro, Gerbelot, Cui, Goldt, Krzakala, Mézard, LZ, *Learning Gaussian Mixtures with Generalised Linear Models;* arXiv:2106.03791.



CONCLUSIONS

realistically structured data

structured data

architecture

O(1) hidden units*

multiple wide layers

algorithm

gradient-descent based

CONCLUSIONS

structured data

realistically structured data

architecture

O(1) hidden units*

multiple wide layers

algorithm

gradient-descent based