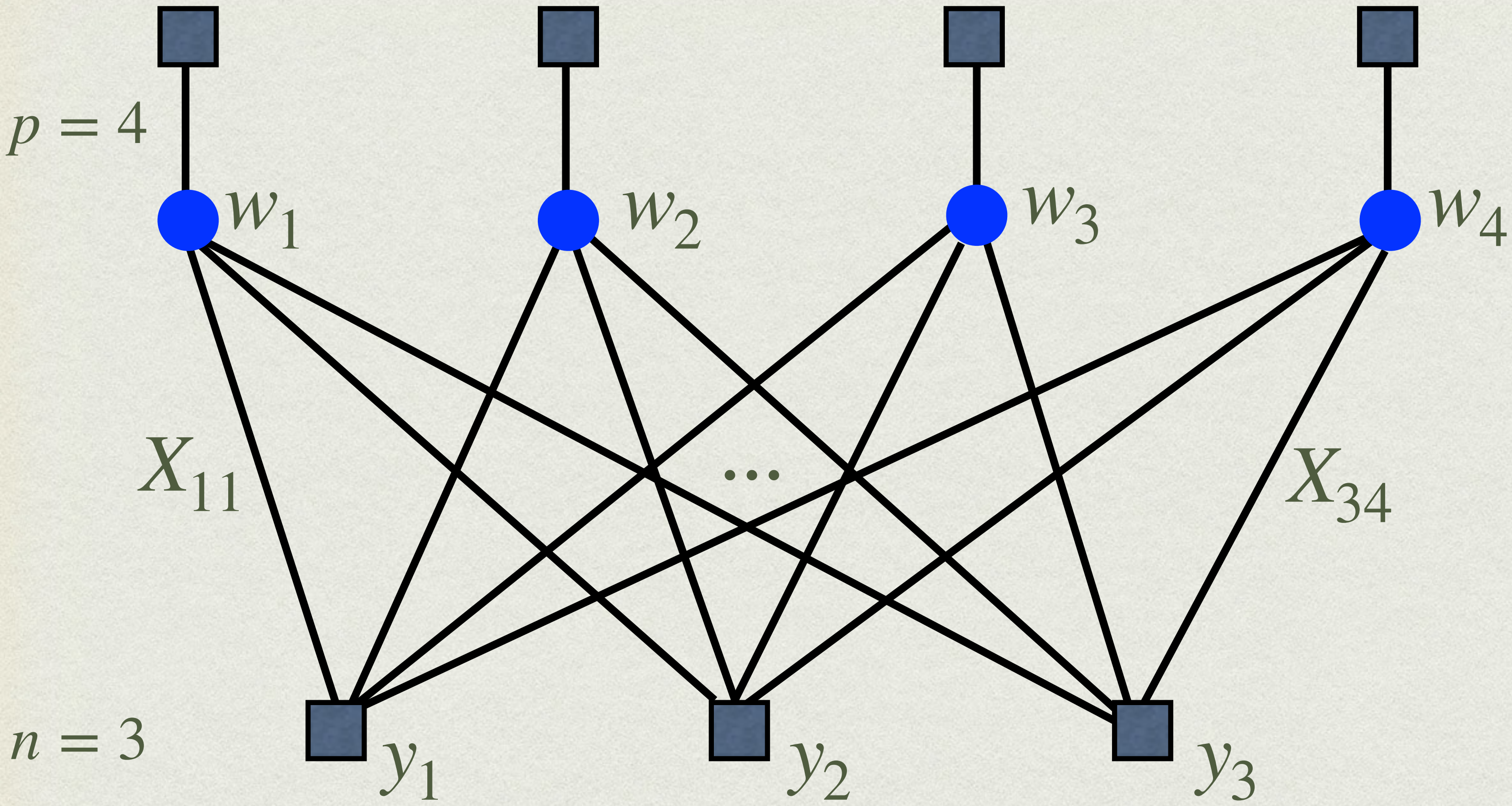
STATISTICAL PHYSICS AND COMPUTATION IN HIGH DIMENSION
LECTURE III

Lenka Zdeborová & Florent Krzakala
(CNRS & CEA Saclay, ENS Paris, EPFL)

Probability, Geometry, and Computation in High Dimensions Boot Camp
Simons institute for Theory of Computing, 19.-28. 8. 2020

GRAPHICAL MODEL



High-dimensional limit:

$p \to \infty, n \to \infty$

$\alpha \equiv n/p = \Theta(1)$

$\mathbf{w} \in \mathbb{R}^p, w_i \in \mathbb{R}$

$\mathbf{y} \in \mathbb{R}^n, y_\mu \in \mathbb{R}$

$X \in \mathbb{R}^{n \times p}, \mathbf{X}_\mu \in \mathbb{R}^p, X_{\mu i} \in \mathbb{R}$

Probability distribution:

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

Solvable for some $P_{\text{data}}(y_\mu, \mathbf{X}_\mu)$, examples follow.

EXAMPLE 1

PERCEPTRON STORAGE CAPACITY

J. Phys. A: Math. Gen. **21** (1988) 271–284. Printed in the UK

Optimal storage properties of neural network models

E Gardner† and B Derrida‡

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

Received 29 May 1987

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

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

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.

PERCEPTRON STORAGE CAPACITY

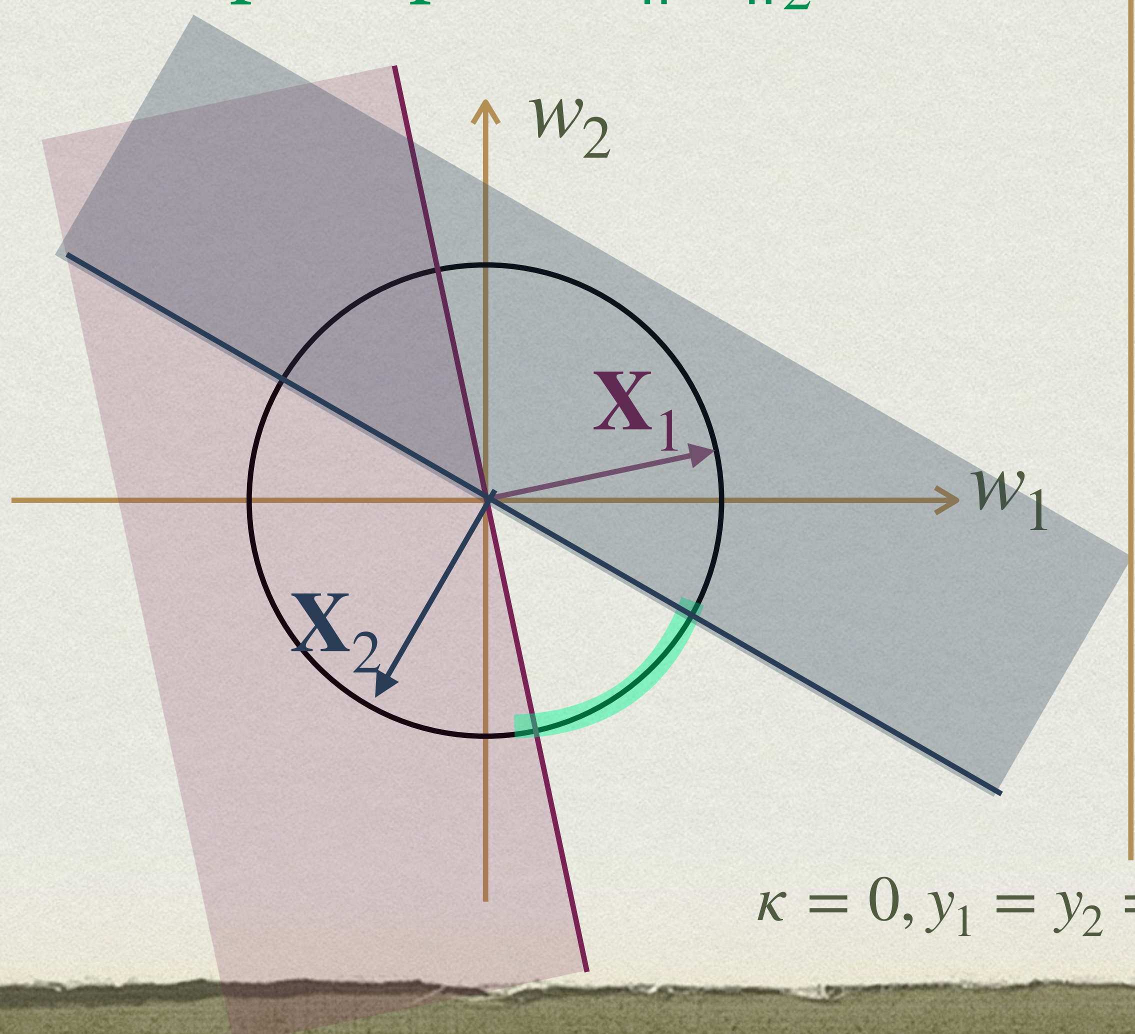
input data (patterns): random iid $X_{\mu i} \sim \mathcal{N}(0,1/p)$

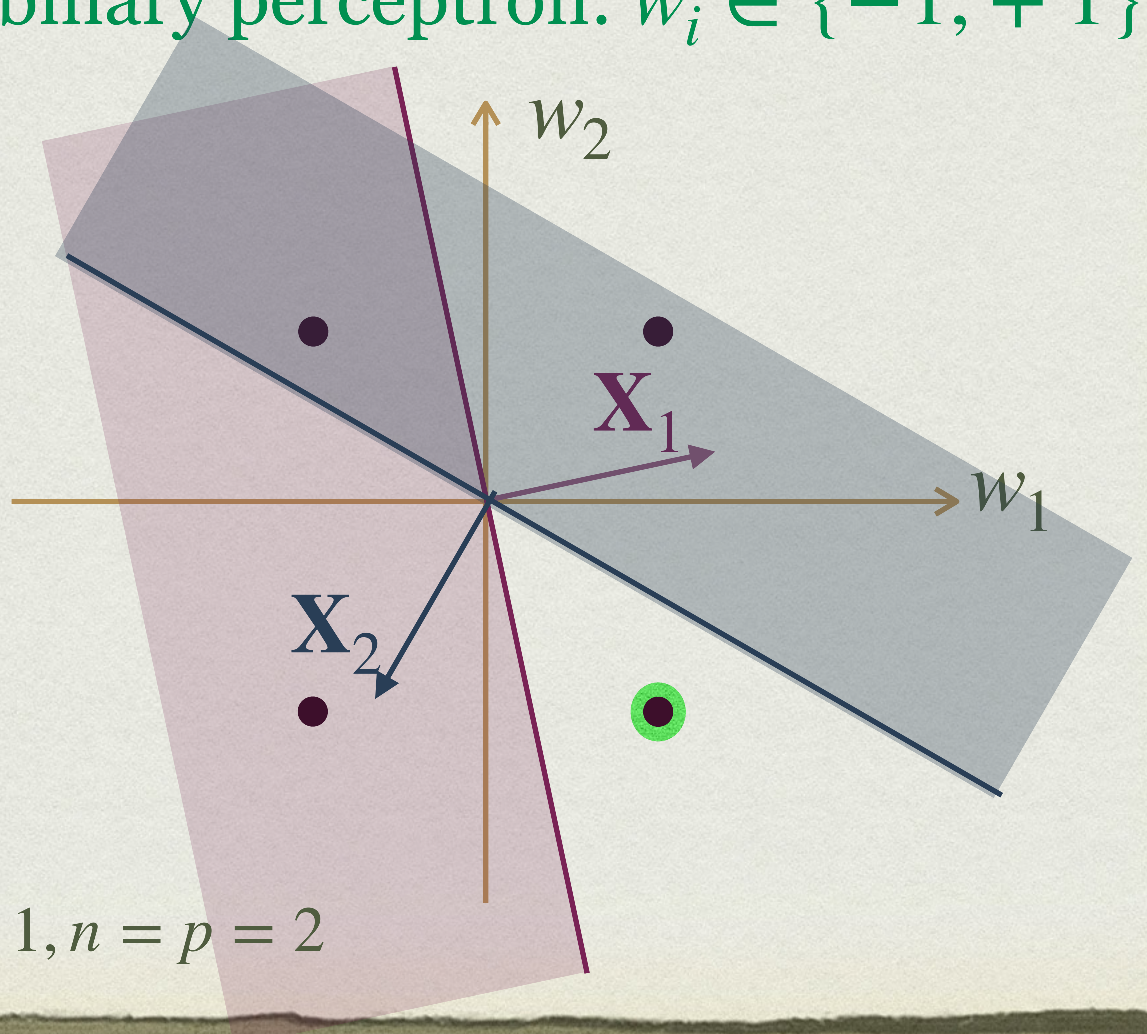
labels: random iid Rademacher $y_\mu \sim \delta(y_\mu + 1)/2 + \delta(y_\mu - 1)/2$

constraints: $P_{\text{out}}(y_\mu, \mathbf{X}_\mu \cdot \mathbf{w}) = \mathbb{I}(y_\mu \mathbf{X}_\mu \cdot \mathbf{w} > \kappa)$

spherical perceptron: $\|\mathbf{w}\|_2 = 1$

binary perceptron: $w_i \in \{-1, +1\}$





$\kappa = 0, y_1 = y_2 = 1, n = p = 2$

PERCEPTRON STORAGE CAPACITY

input data (patterns): random iid $X_{\mu i} \sim \mathcal{N}(0,1/p)$

labels: random iid Rademacher $y_\mu \sim \delta(y_\mu + 1)/2 + \delta(y_\mu - 1)/2$

constraints: $P_{\text{out}}(y_\mu, \mathbf{X}_\mu \cdot \mathbf{w}) = \mathbb{I}(y_\mu \mathbf{X}_\mu \cdot \mathbf{w} > \kappa)$

spherical perceptron: $\|\mathbf{w}\|_2 = 1$

binary perceptron: $w_i \in \{-1, +1\}$

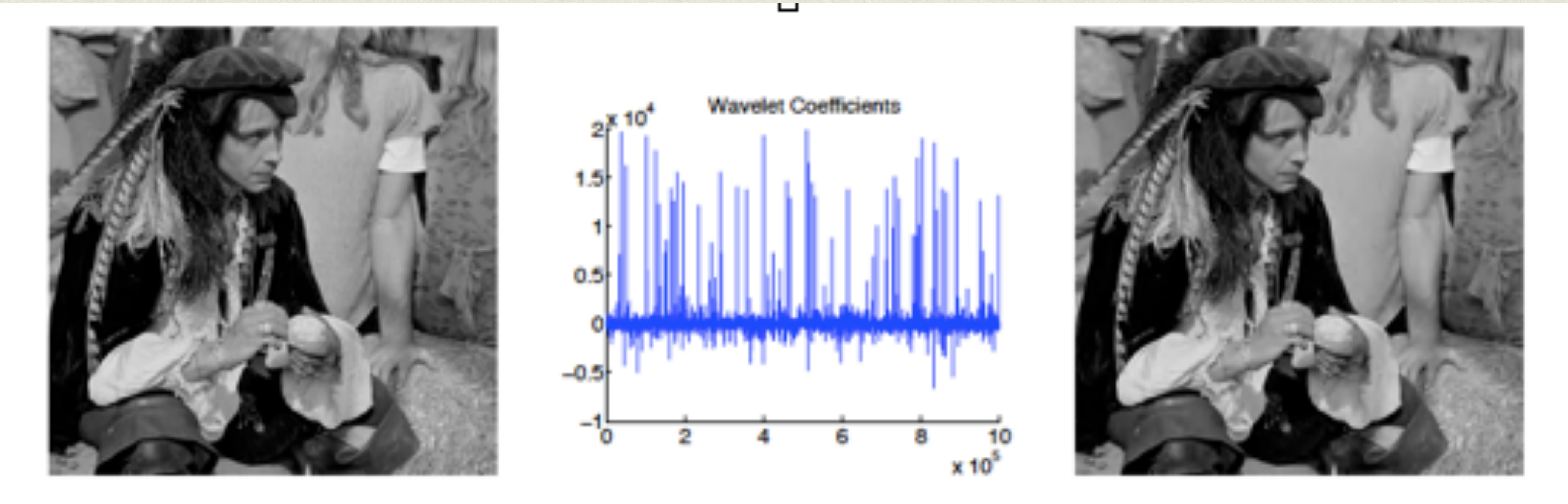
Def: storage capacity as the largest $\alpha_c(\kappa) = n/p$ such that with high probability (as $p \to \infty$)

$$\exists \mathbf{w} \in \mathbb{R}^p : y_\mu \sum_{i=1}^{p} X_{\mu i} w_i > \kappa \quad \forall \mu = 1, \dots, n$$

Def: ground state energy as the smallest possible (over choices of w) number of unsatisfied constraints.

EXAMPLE 2

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?

COMPRESSED SENSING

e.g. Donoho'06

$$\mathbf{y} = G\mathbf{s}^* + \xi \qquad \mathbf{w}^* = \Phi\mathbf{s}^* \qquad X = G\Phi^{-1}$$

- G: measurement matrix of the apparatus (x-ray, NMR).
- Φ: Transform that makes the signal sparse.

$$\mathbf{y} = X\mathbf{w}^* + \xi$$

  - Random iid X is good in "conserving the information".
  - w* has many zeros, $P_w(w_i) = (1-\rho)\delta(w_i) + \rho\mathcal{N}(0,1)$
- Goal: Recover w* from as few measurements as possible.

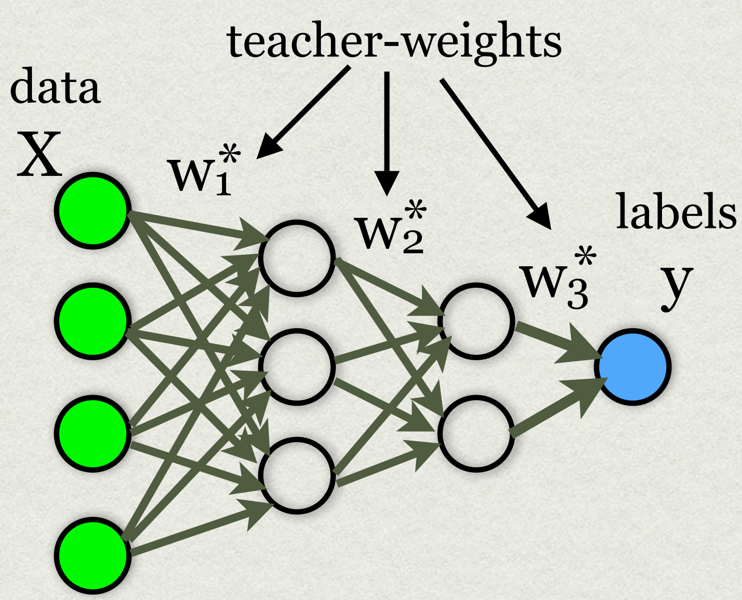
EXAMPLE 3

TEACHER-STUDENT NEURAL NETWORK

(k=1, perceptron) Gardner, Derrida'88, (k>1, committee machine) Schwarze'92

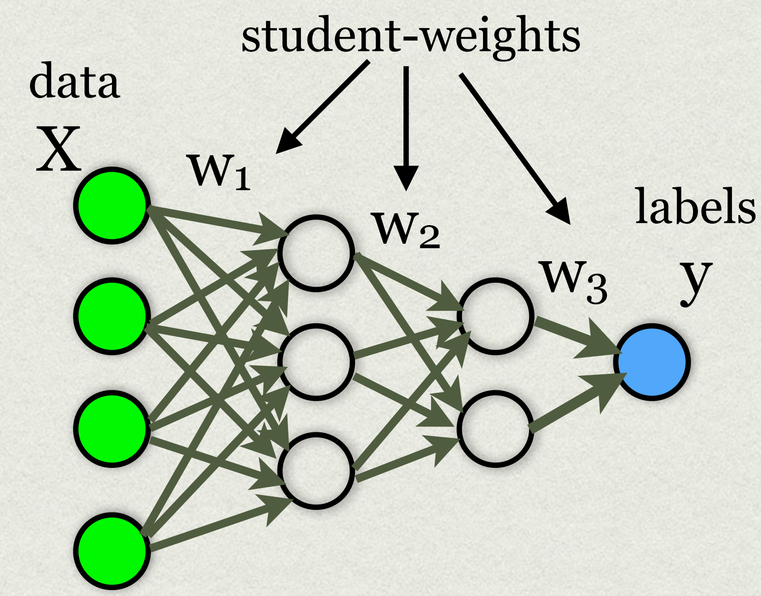
Teacher-network

- Generates data X, n samples of p dimensional data, e.g. iid Gaussian.
- 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?



$p \to \infty, n \to \infty \qquad \alpha \equiv n/p = \Theta(1)$

# of hidden units $\quad k = \Theta(1)$

EXAMPLE 4

with non-separable prior P_w

GENERATIVE PRIORS

e.g. Bora, Jalal, Price, Dimakis'17;

$$\mathbf{y} = \sigma(G\mathbf{s}^*)$$

- G: measurement matrix of the apparatus (x-ray, NMR).
- s^* signal from a range of generative neural network with small input dimension k, $\mathbf{z}^* \in \mathbb{R}^k$

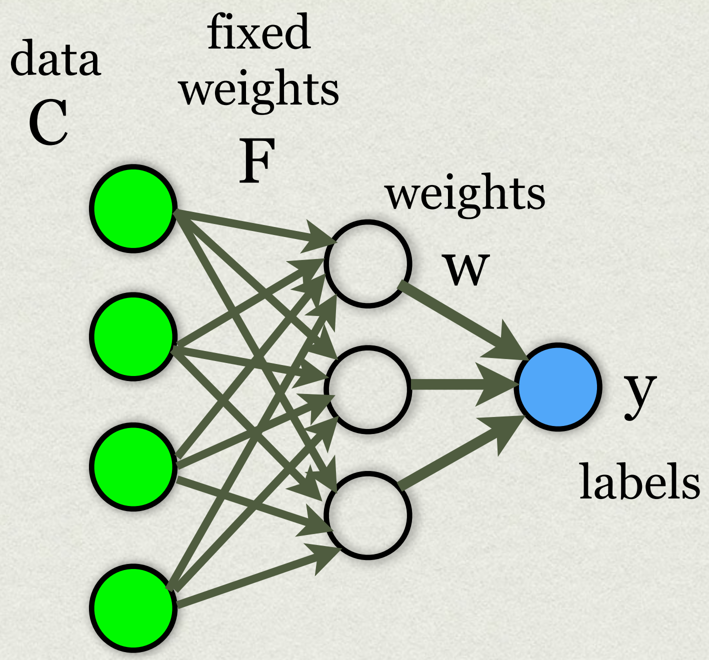
$$\mathbf{s}^* = \varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}\mathbf{z}^*))))$$

Signal comes from a generative neural network

EXAMPLE 5

with non-iid matrix X

RANDOM FEATURE LEARNING



iid Gaussian data, $C \in \mathbb{R}^{n\times d}$

1st layer fixed weights, $F \in \mathbb{R}^{d\times p}$

post-activations, $X = \sigma(CF)$

teacher labels, $y_\mu = \tilde{\sigma}(\mathbf{C}_\mu \cdot \tilde{w}^*)$

Close relation to kernel machines (Rahimi, Recht'08)

Solvable limit $n, p, d \to \infty, n/p = \Theta(1), d/p = \Theta(1)$

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.

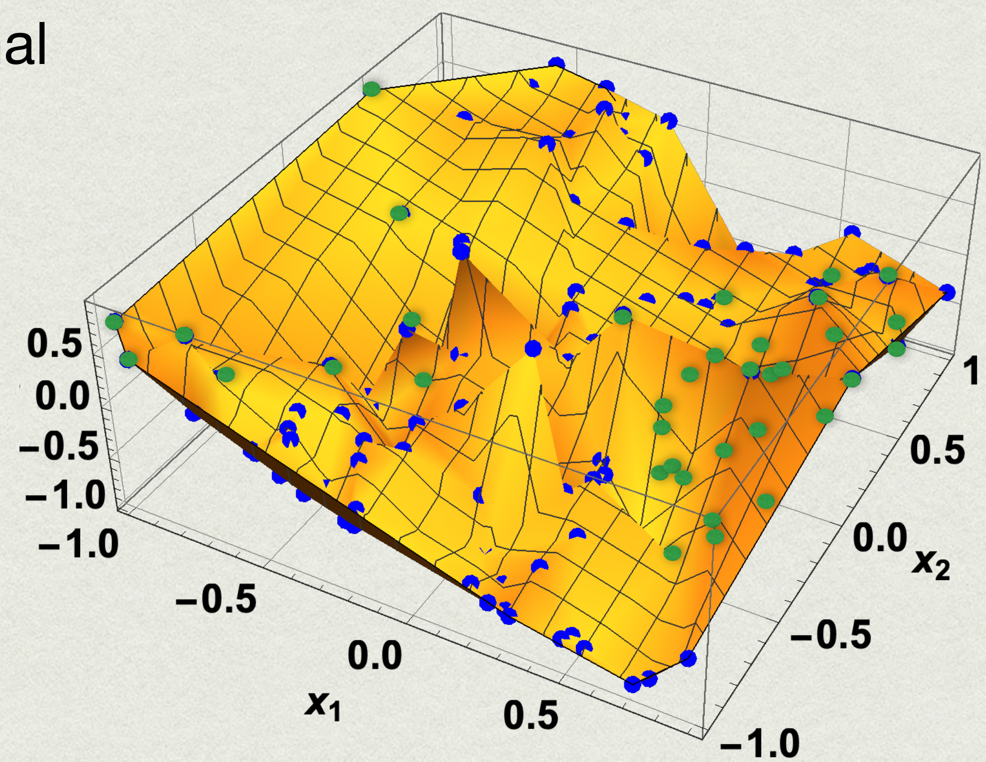
point coordinates in sub-space (dimension d).

low-dimensional sub-space

$$X = \sigma(CF)$$

Data points (dimension p)

Non-linear function



$$y_\mu = \tilde{\sigma}(\mathbf{C}_\mu \cdot \tilde{w}^*)$$

$\boxed{\text{X comes from a generative neural network}}$

LET'S GET INTO MORE DETAILS

BACK TO EXAMPLE 1
STORAGE CAPACITY

PERCEPTRON STORAGE CAPACITY

input data (patterns): random iid $X_{\mu i} \sim \mathcal{N}(0, 1/p)$

labels: random iid Rademacher $y_\mu \sim \delta(y_\mu + 1)/2 + \delta(y_\mu - 1)/2$

constraints: $P_{\text{out}}(y_\mu, \mathbf{X}_\mu \cdot \mathbf{w}) = \mathbb{I}(y_\mu \mathbf{X}_\mu \cdot \mathbf{w} > \kappa)$

spherical perceptron: $\|\mathbf{w}\|_2 = 1$

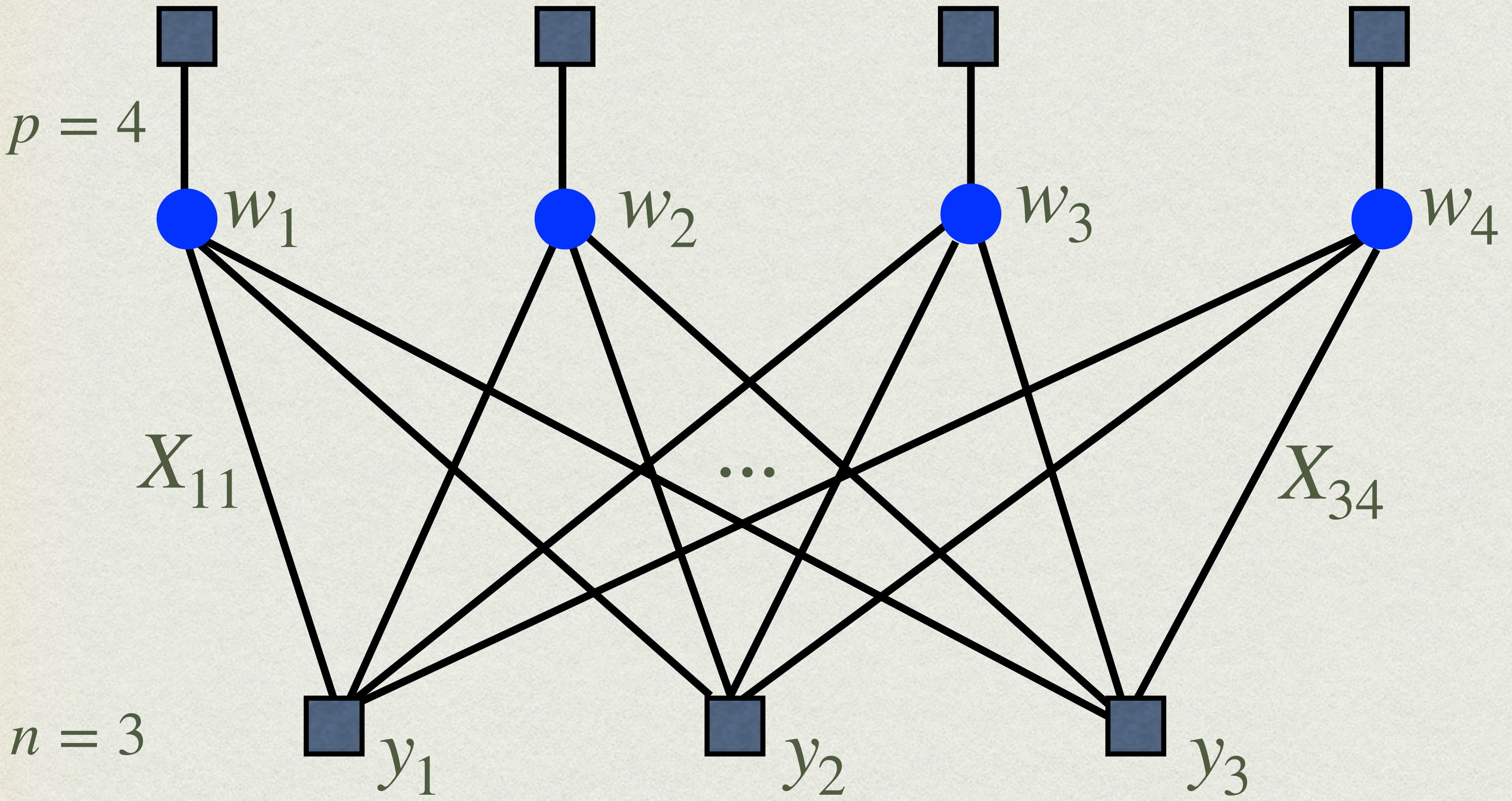
binary perceptron: $w_i \in \{-1, +1\}$

Define storage (Gardner) capacity as the largest $\alpha_G(\kappa) = n/p$ such that with high probability (as $p \to \infty$)

$$\exists w \in \mathbb{R}^p : y_\mu \sum_{i=1}^{p} X_{\mu i} w_i > \kappa \quad \forall \mu = 1, \dots, n$$

For $\kappa = 0$, storage capacity = linear separability threshold.

GRAPHICAL MODEL



Limit:

$p \to \infty, n \to \infty$

$\alpha \equiv n/p = \Theta(1)$

$\mathbf{w} \in \mathbb{R}^p, w_i \in \mathbb{R}$

$\mathbf{y} \in \mathbb{R}^n, y_\mu \in \mathbb{R}$

$X \in \mathbb{R}^{n\times p}, \mathbf{X}_\mu \in \mathbb{R}^p, X_{\mu i} \in \mathbb{R}$

Probability distribution:

step function

$$P(\mathbf{w}\,|\,\mathbf{y}, X) = \frac{1}{Z(\mathbf{y}, X)} \prod_{i=1}^{p} P_w(w_i) \prod_{\mu=1}^{n} \theta(y_\mu \mathbf{X}_\mu \cdot \mathbf{w} - \kappa)$$

SPHERICAL PERCEPTRON

spherical perceptron: $\|\mathbf{w}\|_2 = 1$

- $\kappa = 0$: $\alpha_G = 2$, Cover'65.
- $\kappa \geq 0$: Conjecture from replica method by Derrida, Gardner'88. Proof - Shcherbina, Tirozzi'03.
- $\kappa < 0$: open problem, replica symmetry breaking present (Franz, Parisi'16; Franz, Parisi, Sevelev, Urbani, and Zamponi'17; Mihailo Stojnic, arXiv:1306.3980)

BINARY PERCEPTRON

binary perceptron: $w_i \in \{-1, +1\}$

- Krauth, Mézard'89 conjecture from replica method:

$$q_0, \hat{q}_0 \geq 0$$

$$\phi_{\mathrm{RS}}(q_0, \hat{q}_0) = \frac{1}{2}\left(q_0 - 1\right)\hat{q}_0 + \int Dt \log\left[2\cosh\left(t\sqrt{\hat{q}_0}\right)\right] + \alpha \int Dt \log\left[\int_{\frac{K - t\sqrt{q_0}}{\sqrt{1-q_0}}}^{\infty} Du\right]$$

Saddle point $q_0^*, \hat{q}_0^*$

$\alpha < \alpha_G : \phi_{\mathrm{RS}}(q_0^*, \hat{q}_0^*) > 0$ $\quad \phi_{\mathrm{RS}}(q_0^*, \hat{q}_0^*) = \lim_{n.p \to \infty} \frac{1}{p} \mathbb{E}_{\mathbf{y},X} \log(Z(\mathbf{y}, X))$

$\alpha > \alpha_G : \phi_{\mathrm{RS}}(q_0^*, \hat{q}_0^*) < 0$ $\quad \alpha_G(K=0) = 0.833...$

BINARY PERCEPTRON

- What is known rigorously?
- Kim, Roche'98: $0.005 < \alpha_G(K=0) < 0.9973$
- Ding, Sun'18: tight lower bound, proof technique inspired by the physics result.
- Xu'19, sharpness of the threshold.

Perhaps the most basic open problem on artificial neural networks, with a simple explicit conjecture.

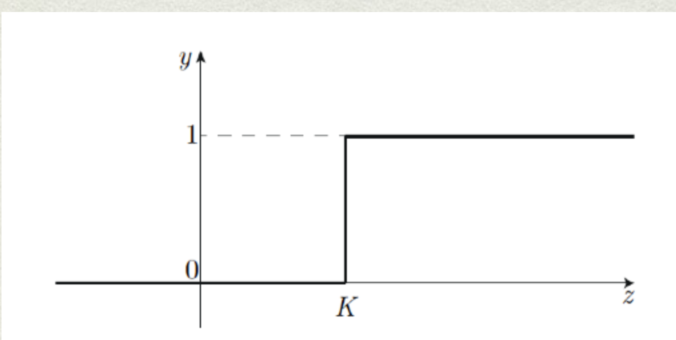
$\alpha_G(K=0) = 0.833...$

SYMMETRIC BINARY PERCEPTRONS

described so far

$z_\mu > \kappa$

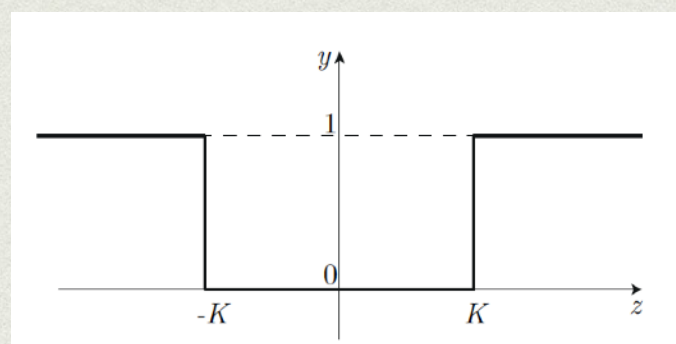
step



Aubin, Perkins, LZ, 1901.00314

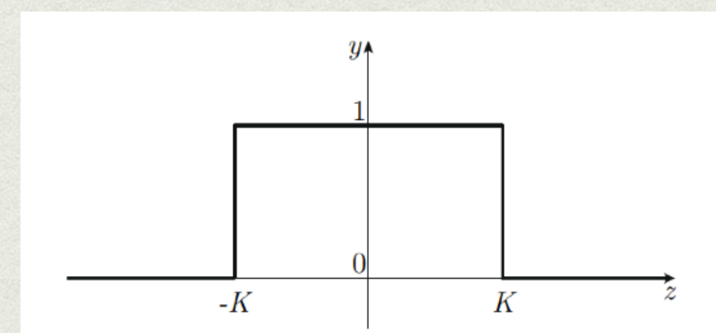
$|z_\mu| > \kappa$

u-shape



$|z_\mu| < \kappa$

rectangle



$$P(\mathbf{w}) = \frac{1}{Z(X)} \prod_{\mu=1}^{n} \varphi\left(z_\mu, \kappa\right)$$

$$z_\mu \equiv \sum_{i=1}^{p} X_{\mu i} w_i$$

$w_i \in \{-1,1\}$ $X_{\mu i}$ iid $\alpha = n/p$ $p \to \infty$

PERCEPTRON STORAGE CAPACITY

$$\alpha_G = -\frac{\log 2}{\log p_\kappa^{s,u,r}}$$

where $p_\kappa^{s,u,r}$ is the probability that a Gaussian random variable of zero mean and unit variance satisfies the step/u-shape/rectangle constraint.

$z_\mu > \kappa$	$\lvert z_\mu \rvert > \kappa$	$\lvert z_\mu \rvert < \kappa$
step	u-function	rectangle
never correct	$\forall \kappa < \kappa^* \sim 0.817$	$\forall \kappa \in \mathbb{R}^+$

Aubin, Perkins, LZ, 1901.00314

CAN SOLUTIONS BE FOUND EFFICIENTLY?

- Statistical physics:
  - For any $\alpha > 0$ almost all solutions in a frozen-1RSB structure, i.e. vanishing entropy blobs separated by extensive distance (Krauth, Mézard'89, Huang, Wong, Kabashima'13).
  - Frozen-1RSB solutions are conjectured algorithmically hard to find with efficient algorithms.
  - Rare solutions in a large wide cluster easy to find for $\alpha \lesssim 0.75$ (Baldassi, Ingrosso, Lucibello, Saglietti, Zecchina'15).
- Rigorously: Close to nothing is known.
  - Open problem 1: Algorithmically constructive α>0.005 lower bound for binary perceptron (symmetric, if simpler).

Are perceptrons with random labels relevant for learning with neural networks?

- No, because generalisation is ill posed.
  But see teacher-student setting (starting in 2 slides).

- Yes, because of relation to

  (a) the VC dimension: $d_{\mathrm{VC}} \geq \dfrac{\alpha_G}{2} p$

  (b) The Rademacher complexity (next slide).

RADEMACHER COMPLEXITY

Def: Given a function class f_w, and random iid $y_\mu \in \{\pm 1\}$, the Rademacher complexity is $\mathscr{R}_n = \mathbb{E}_{y,X} \sup_w \frac{1}{n} \sum_{\mu=1}^{n} y_\mu f_w(X_\mu)$.

Theorem: With high probability $R_{\text{emp}} - R_{\text{pop}} \leq \mathscr{R}_n + o(1)$.

"If you are bad at fitting random labels, you must generalize well."

Note: For $f_w(X_\mu) = \text{sign}(\mathrm{X}_\mu \cdot \mathrm{w})$ (the perceptron)
$\mathscr{E}_{\text{GS}} = \frac{\alpha}{2}(1 - \mathscr{R})$ where $\mathscr{E}_{\text{GS}} = \frac{1}{p} \inf_w \sum_{\mu=1}^{n} \mathbb{I}[y_\mu \neq f_w(X_\mu)]$ is the ground state energy of the perceptron problem.

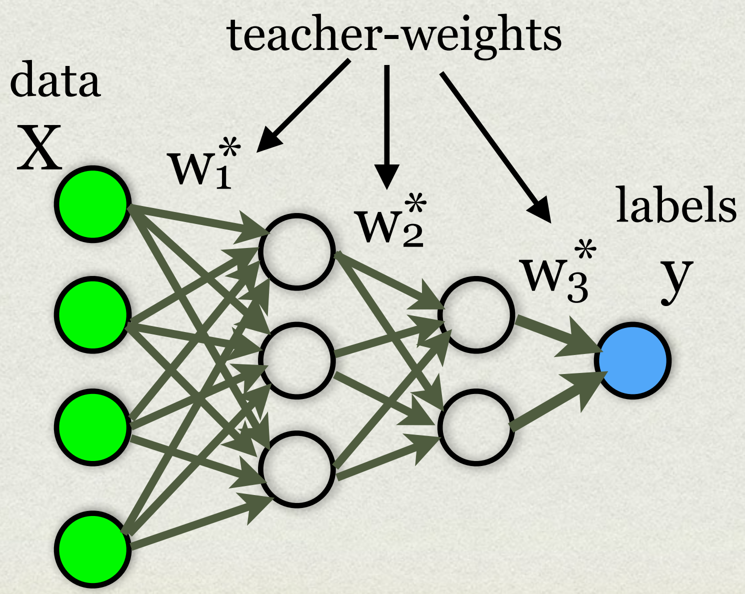
Abbara, Aubin, FK, LZ, 1912.02729

EXAMPLE 2&3
TEACHER-STUDENT
GENERALISED LINEAR MODEL

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

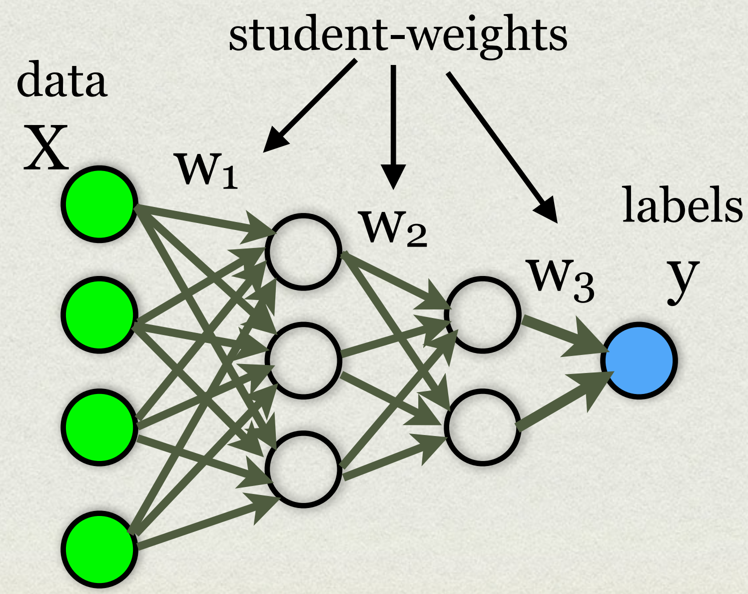
Teacher-network

- Generates data X, n samples of p 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

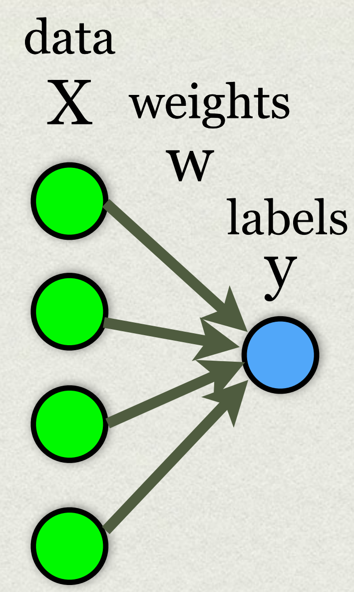
1989

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 $y_\mu = \mathrm{sign}\left(\sum_{i=1}^{p} X_{\mu i} w_i^*\right)$
- High-dimensional regime: $n \to \infty \quad p \to \infty$ p dimensions
  $\alpha \equiv n/p = \Theta(1)$ 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

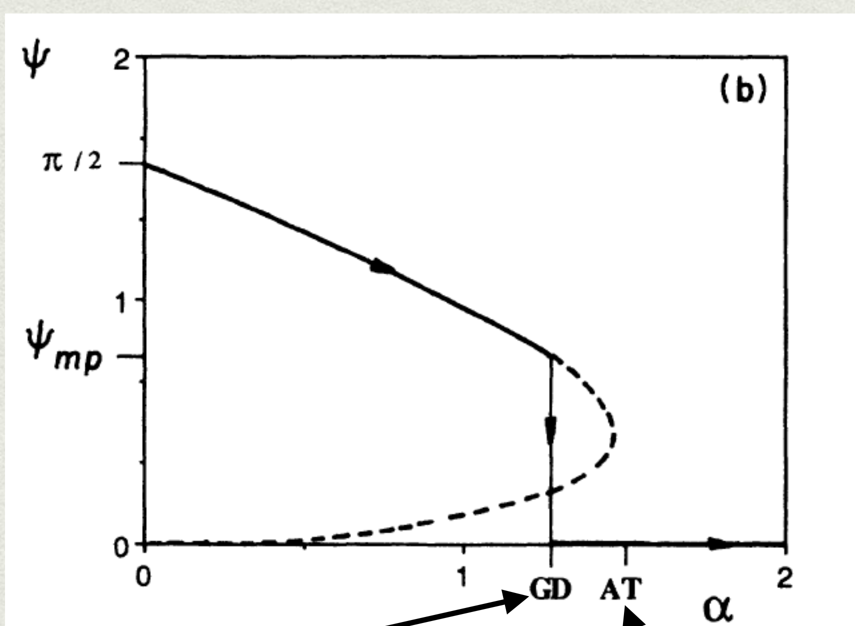
Géza Györgyi*

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.

Generalisation error



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

- Binary teacher-weights:

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

- Phase transition in the generalization error's dependence on the sample complexity.

$$\alpha = n/p$$

STATE-OR-THE-ART

- 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, FK, Macris, Miolane, LZ, arXiv:1708.03395, COLT'18, PNAS'19

BAYES-OPTIMAL GENERALIZATION

Posterior probability distribution:

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

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

(noisy) activation function

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

$\neq$ minimization of a loss function (empirical risk minimization)

REPLICA METHOD SOLUTION

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

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

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

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

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

REPLICA METHOD SOLUTION

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

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

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

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

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

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

$\xi \sim P_\xi$

PROOF IDEA

Barbier, FK, Macris, Miolane, LZ arXiv:1708.03395

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

$\Phi_{P_w}(\hat{m})$ is the free energy of a scalar denoising problem

$$y' = \sqrt{\hat{m}}\, w^* + \xi$$

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

$w^* \sim P_w$

$\Phi_{P_{\mathrm{out}}}(m;\rho)$ is the free energy of a scalar denoising problem

$$\tilde{y} \sim P_{\mathrm{out}}(\tilde{y} | \sqrt{m}\, v + \sqrt{\rho - m}\, z^*)$$

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

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

PROOF IDEA

Barbier, FK, Macris, Miolane, LZ arXiv:1708.03395

Adaptive interpolation between the original posterior and p + n independent scalar denoising problems.

Interpolating Hamiltonian (=log-likelihood):

$$\mathcal{H}_t = -\sum_{\mu=1}^{n} \ln P_{\rm out}(y_\mu \,|\, s_{t,\mu}) + \frac{1}{2}\sum_{i=1}^{p} [\sqrt{t\hat{m}}(w_i^* - w_i) + \xi_i]^2$$

$$s_{t,\mu} = \sqrt{1-t}[Xw]_\mu + \sqrt{\int_0^t m(t')dt'}\, v_\mu + \sqrt{\int_0^t (\rho - m(t'))dt'}\, z_\mu$$

PROOF IDEA

Barbier, FK, Macris, Miolane, LZ arXiv:1708.03395

Interpolating free energy:

$$f_p(t=0) = f_p - \frac{1}{2}$$

$$f_p(t=1) = -\frac{1+m\hat{m}}{2} + \Phi_{P_w}(\hat{m}) + n/p\Phi_{P_{\rm out}}(\int_0^1 m(t)dt;\rho)$$

Main aim: Choose interpolation path m(t) so that $f_p(t)$ effectively does not depend on t!

Key property for this to work (Nishimori): Under expectations ground truth w* is exchangeable for a sample from P(w|y,X).

$$\mathbb{E}_{w^*,y}\mathbb{E}_{P(w|y)}[g(y,w^{(1)},w^*)] = \mathbb{E}_y\mathbb{E}_{P(w|y)}[g(y,w^{(1)},w^{(2)})]$$

PROOF IDEA

Barbier, FK, Macris, Miolane, LZ arXiv:1708.03395

Interpolating free energy:

$$f_p(t=0) = f_p - \frac{1}{2}$$

$$f_p(t=1) = -\frac{1+m\hat{m}}{2} + \Phi_{P_w}(\hat{m}) + n/p\,\Phi_{P_{\text{out}}}\left(\int_0^1 m(t)dt;\rho\right)$$

Main aim: Choose interpolation path m(t) so that $f_p(t)$ effectively does not depend on t!

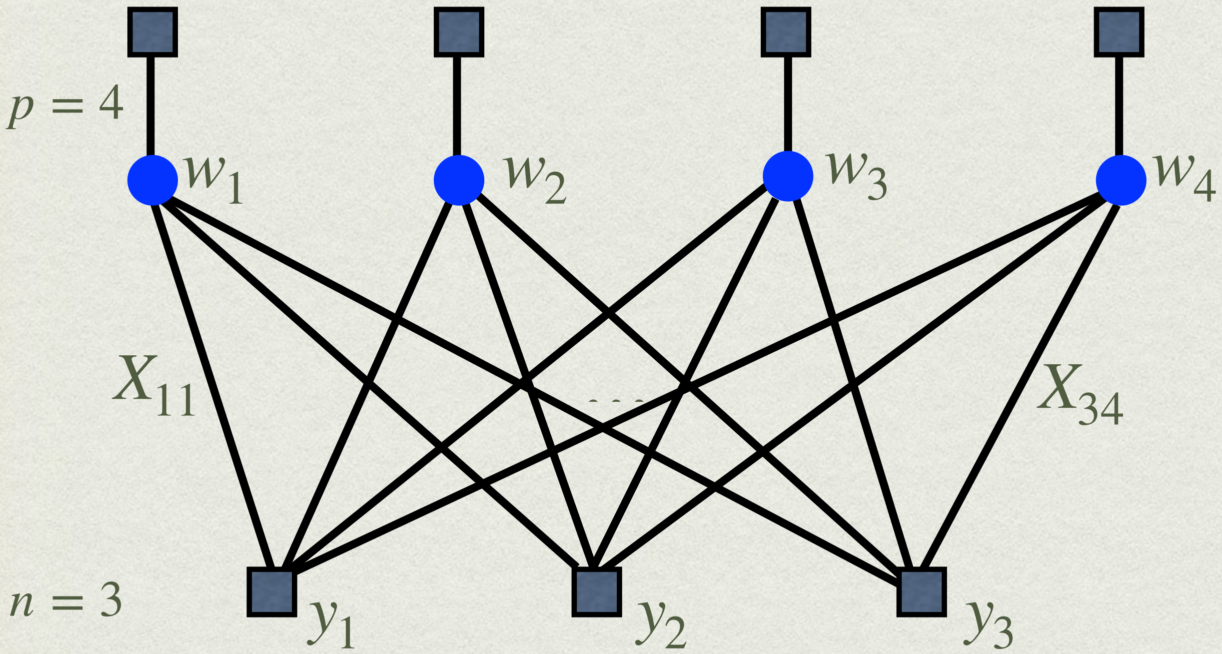
Work hard and get at the end:

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

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

APPROXIMATE MESSAGE PASSING

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



Belief Propagation

$$m_{i\to\mu}(w_i) = \frac{1}{z_{i\to\mu}} P_w(w_i) \prod_{\gamma\neq\mu} m_{\gamma\to i}(w_i)$$

$$m_{\mu\to i}(w_i) = \frac{1}{z_{\mu\to i}} \int \prod_{j\neq i} [\mathrm{d}w_j m_{j\to\mu}(w_j)] P_{\text{out}}(y_\mu\,|\,\sum_l X_{\mu l} w_l)$$

The p-dimensional integral in BP is algorithmically intractable ...

APPROXIMATE MESSAGE PASSING

$$m_{i\to\mu}(w_i) = \frac{1}{z_{i\to\mu}} P_w(w_i) \prod_{\gamma\neq\mu} m_{\gamma\to i}(w_i)$$

$$m_{\mu\to i}(w_i) = \frac{1}{z_{\mu\to i}} \int \prod_{j\neq i} [\mathrm{d}w_j m_{j\to\mu}(w_j)] P_{\text{out}}(y_\mu \mid \sum_l X_{\mu l} w_l)$$

The p-dimensional integral in BP is algorithmically intractable …

BP assumes incoming messages are independent. And there are many of them. Central limit theorem implies that we can close the equations on only means and variances of the messages.

Moreover, all the messages depend only weakly on the “target” node, expand about the point-estimations and collect terms that matter into the so-called Onsager terms.

Algorithm 2 Generalized Approximate Message Passing (G-AMP)

Input: $\mathbf{y}$

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

repeat

AMP Update of ω_μ, V_μ

$$V^t_\mu \leftarrow \sum_i F^2_{\mu i} v^{t-1}_i$$

$$\omega^t_\mu \leftarrow \sum_i F_{\mu i} a^{t-1}_i - V^t_\mu g^{t-1}_{\text{out},\mu}$$

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

$$a^t_i \leftarrow f_a(\Sigma^t_i, R^t_i)$$

$$v^t_i \leftarrow f_v(\Sigma^t_i, R^t_i)$$

$\mathrm{t} \leftarrow \mathrm{t} + 1$

until Convergence on $\mathbf{a}$,$\mathbf{v}$

output: $\mathbf{a}$,$\mathbf{v}$.

$X \to F$

$P_w \to P_X$

Simple to implement, only matrix multiplications, O(p²)

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

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

https://github.com/sphinxteam/GeneralizedLinearModel2017

Algorithm 2 Generalized Approximate Message Passing (G-AMP)

Input: $\mathbf{y}$

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

repeat

AMP Update of ω_μ, V_μ

$$V^t_\mu \leftarrow \sum_i F^2_{\mu i} v^{t-1}_i$$

$$\omega^t_\mu \leftarrow \sum_i F_{\mu i} a^{t-1}_i - \boxed{V^t_\mu g^{t-1}_{\text{out},\mu}}$$

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

$$g^t_{\text{out},\mu} \leftarrow g_{\text{out}}(\omega^t_\mu, y_\mu, V^t_\mu)$$

$$\Sigma^t_i \leftarrow \left[-\sum_\mu F^2_{\mu i}\partial_\omega g_{\text{out}}(\omega^t_\mu, y_\mu, V^t_\mu)\right]^{-1}$$

$$R^t_i \leftarrow \boxed{a^{t-1}_i} + \Sigma^t_i \sum_\mu F_{\mu i} g^t_{\text{out},\mu}$$

AMP Update of the estimated marginals a_i, v_i

$$a^t_i \leftarrow f_a(\Sigma^t_i, R^t_i)$$
$$v^t_i \leftarrow f_v(\Sigma^t_i, R^t_i)$$

$\mathrm{t} \leftarrow \mathrm{t}+1$

until Convergence on $\mathbf{a}$,$\mathbf{v}$

output: $\mathbf{a}$,$\mathbf{v}$.

$X \rightarrow F$

$P_w \rightarrow P_X$

Onsager terms

Simple to implement, only matrix multiplications, O(p²)

GAMP for prediction: $$\hat{y}^t_{\text{new}} = \frac{1}{\sqrt{2\pi V^t}} \int \mathrm{d}z\, \mathrm{d}y\, y P_{\text{out}}(y|z) e^{-\frac{1}{2V^t}\left(z - \sum_i F_{\text{new},i} a^{t-1}_i\right)^2}$$

https://github.com/sphinxteam/GeneralizedLinearModel2017

STATE EVOLUTION

Define: $m^t \equiv \frac{1}{p}\sum_{i=1}^{p} w_i^* a_i^t$ 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

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

SELECTED RELATED WORK

AMP is closely related to the Thouless-Anderson-Palmer'76 equations for the Sherrington-Kirkpatrick spin glass. For perceptron written by Mezard'89 as a way to derive the replica result without replicas, not used as an actual algorithm.

TAP had a problem with time-indices and hence with convergence (only Bolthausen fixed the issue in ~2008, and later AMP).

AMP for general prior written by Donoho, Maleki, Montanari in 2009.

G-AMP derived by Rangan'10, but also appeared earlier in Kabashima'03 (as a way to unify perceptron and CDMA).

State evolution proven by Bayati, Montanari'11 for Gaussian matrices and output, by Bayati, Lelarge, Montanari'12 for general iid matrices, and Gaussian output. General output and Gaussian matrices in Javanmard, Montanari'13.

BOTTOM LINE

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

- w* is generated from P_w, y from P_{out}. X is random iid.
- The analysis gave us the free energy $f_{\text{RS}}(m)$

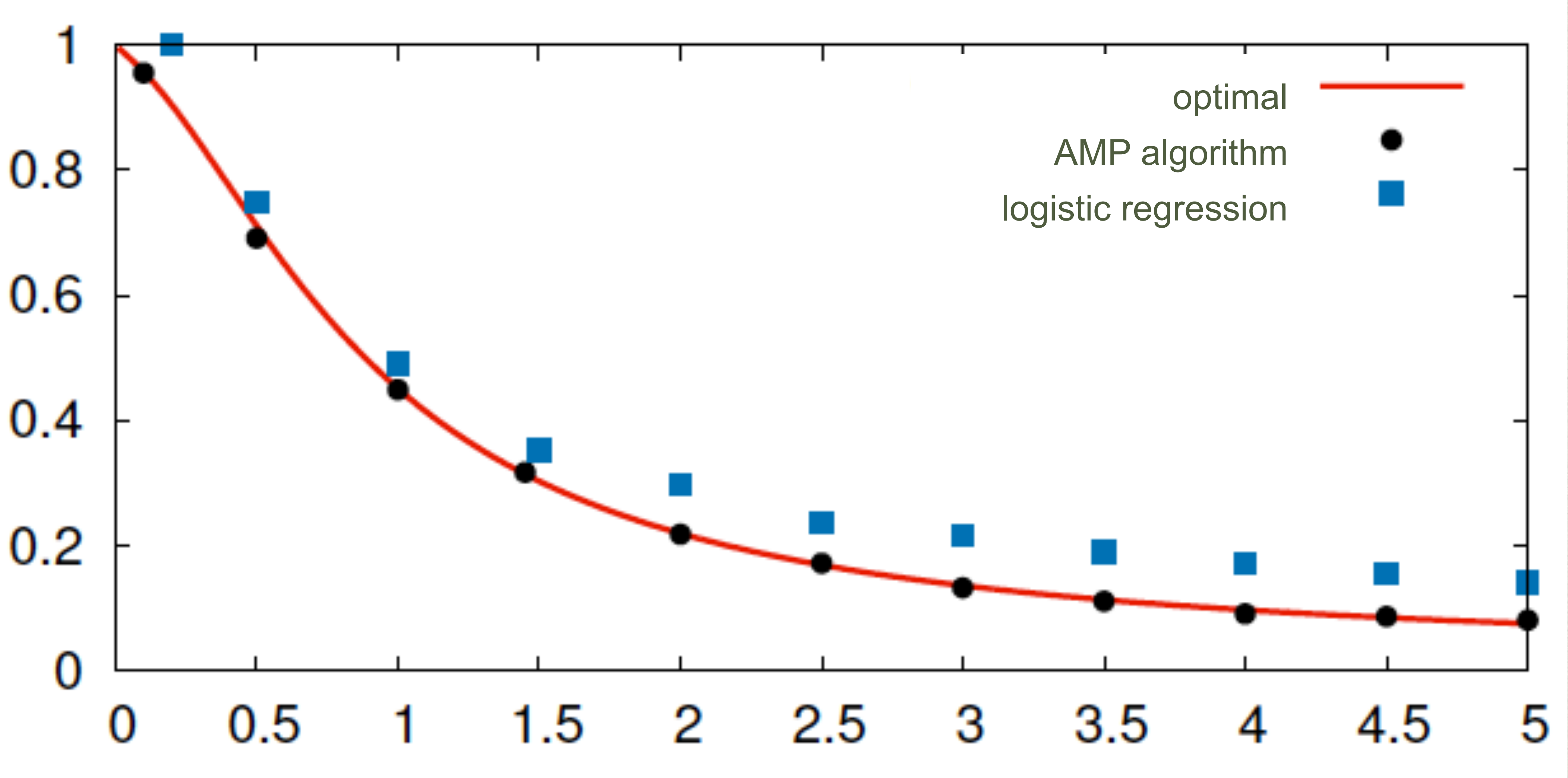
$$\text{MMSE} = \rho - \text{argmax} f_{\text{RS}}(m)$$

MSE_{AMP} = local extremum of $f_{\text{RS}}(m)$, reached from un-informed initialisation of state evolution.

SPHERICAL PERCEPTRON

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

generalisation error



# of samples per dimension n/p

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:

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

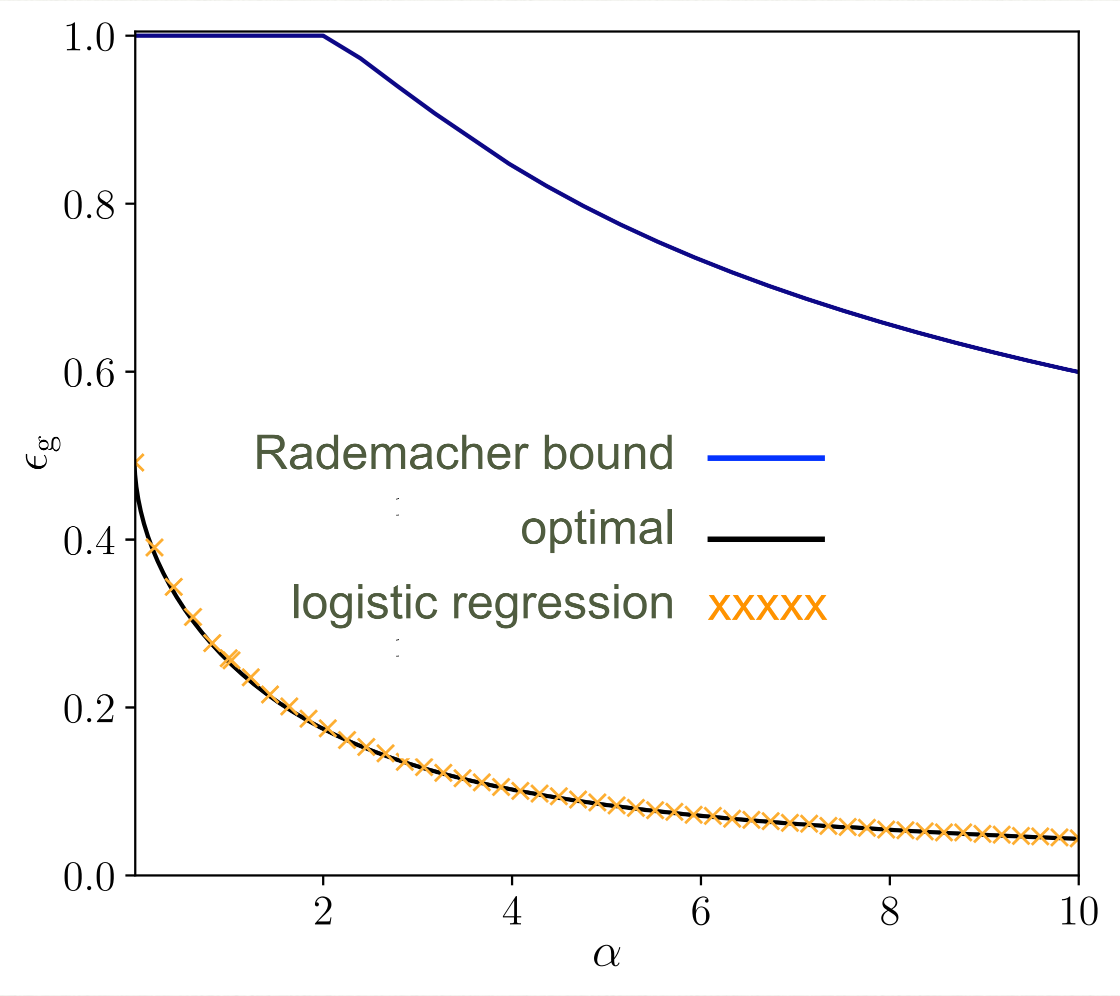
e.g. square loss $\ell(y,z) = (y-z)^2$, logistic loss $\ell(y,z) = \log_2(1+e^{-yz})$

BAYES VS RISK MINIMISATION

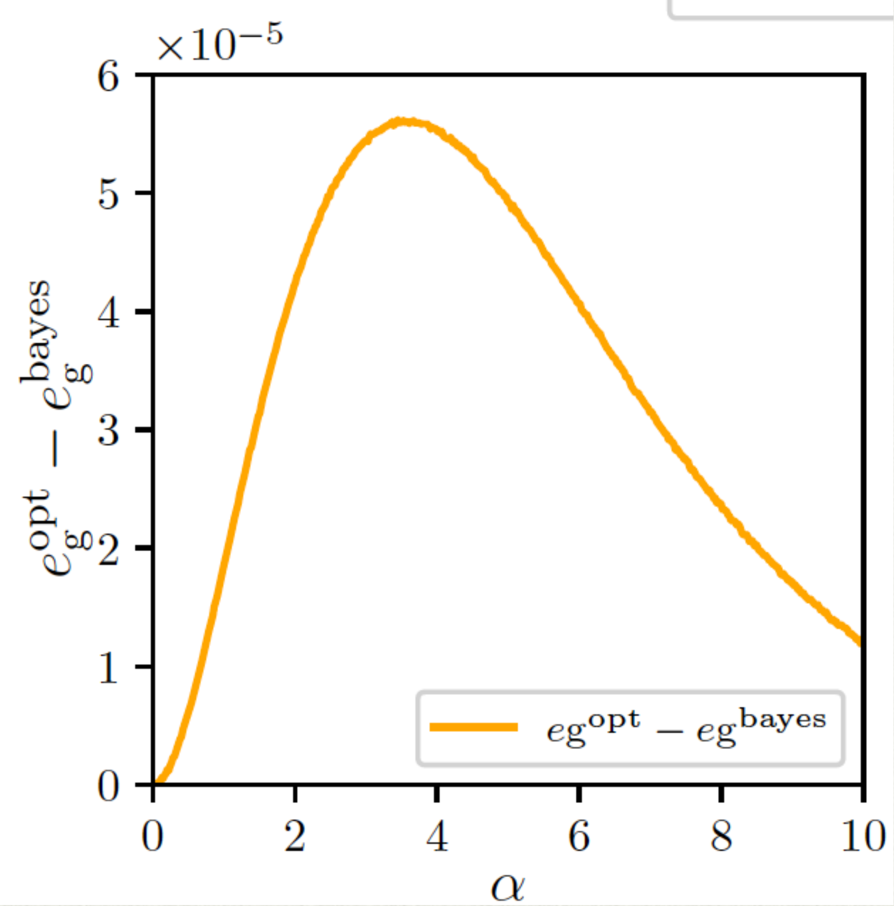
$$y_\mu = \text{sign}\Big(\sum_{i=1}^{p} X_{\mu i} w_i\Big) \qquad P_w = \mathcal{N}(0,1)$$

Optimally regularized logistic regression essentially Bayes-optimal

generalisation error



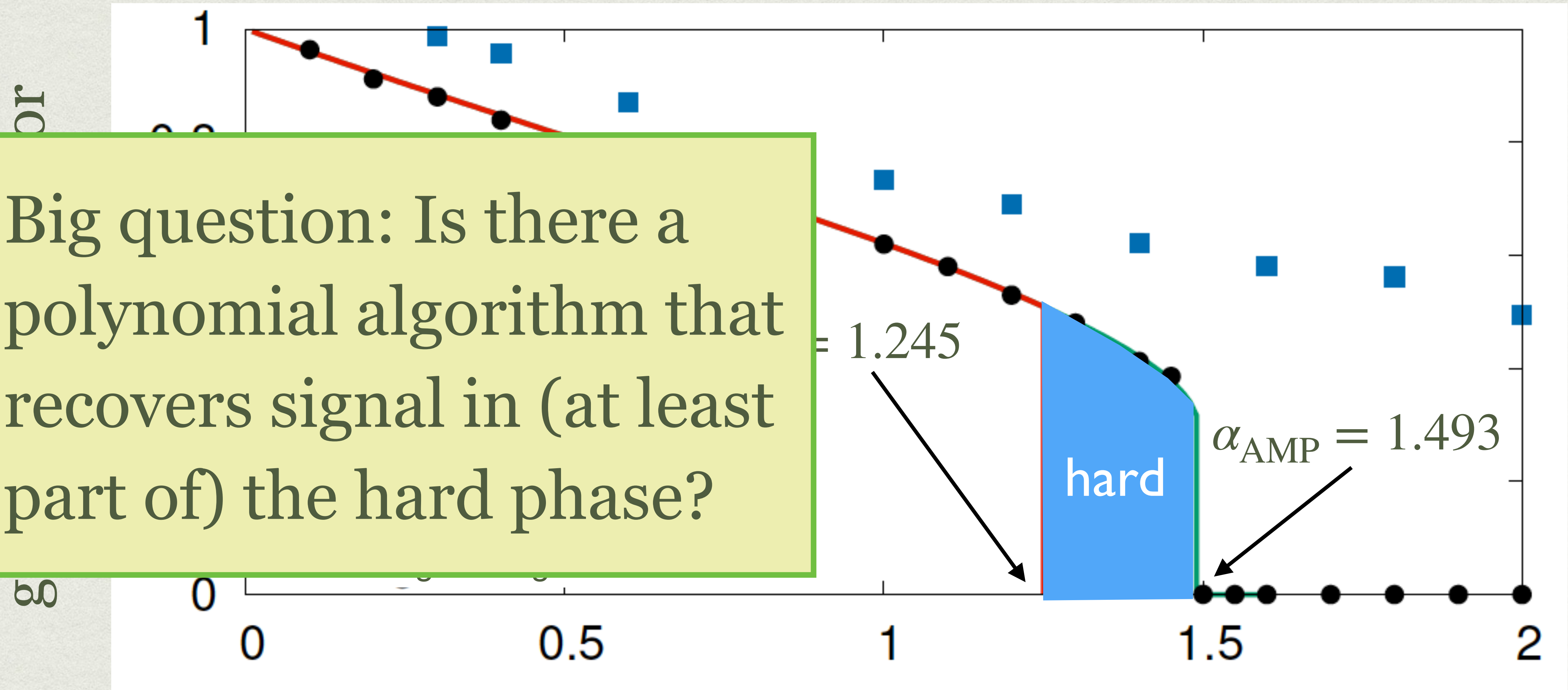
# of samples per dimension



Aubin, Lu, FK, LZ, 2006.06560

BINARY PERCEPTRON

$$y_\mu = \mathrm{sign}\left(\sum_{i=1}^{p} X_{\mu i} w_i\right) \qquad w_i \in \{-1, +1\} \qquad \begin{matrix} n \to \infty \\ p \to \infty \end{matrix} \qquad n/p = \Theta(1)$$

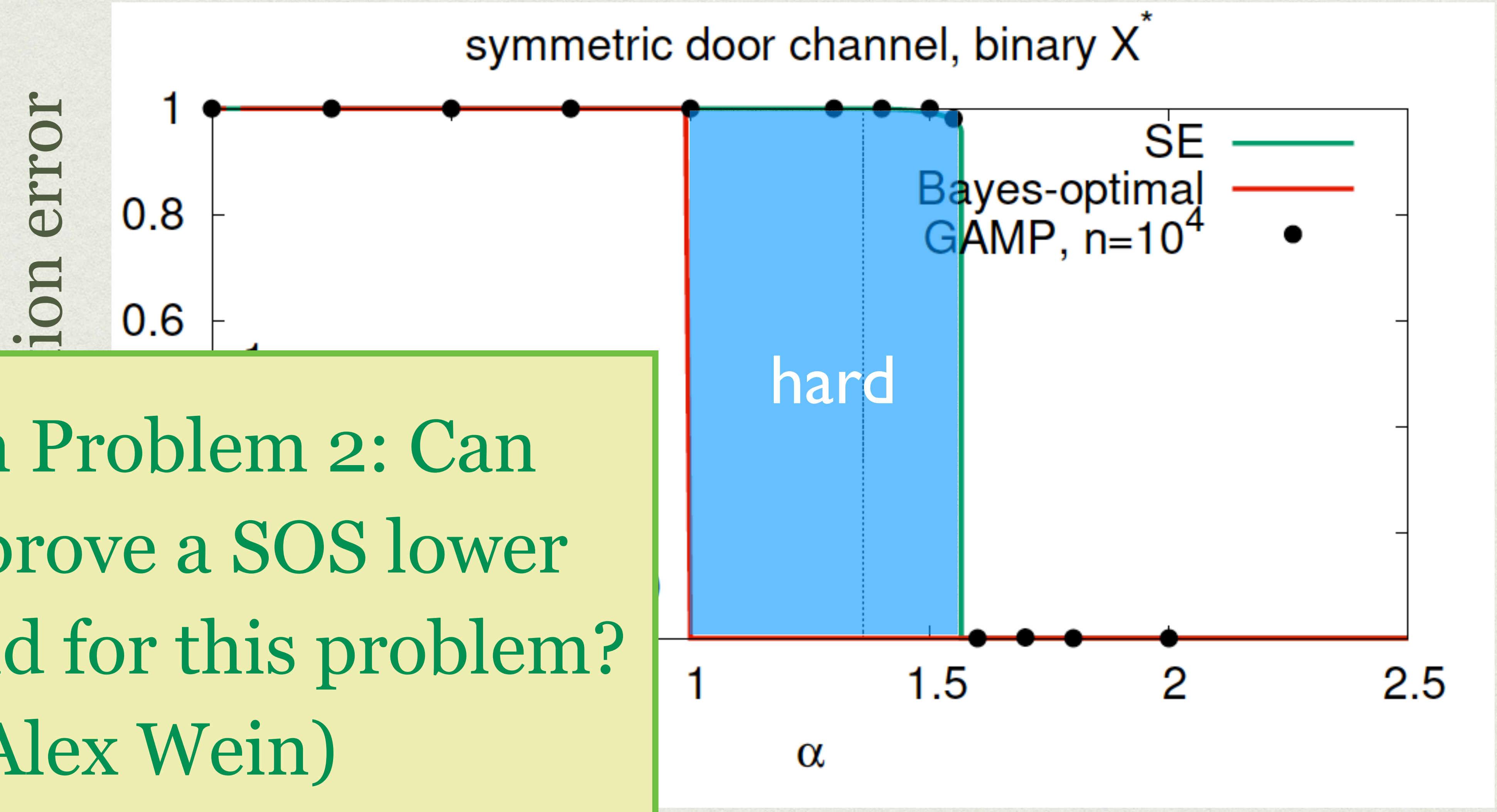


Big question: Is there a polynomial algorithm that recovers signal in (at least part of) the hard phase?

\# of samples per dimension n/p

SYMMETRIC-DOOR PERCEPTRON

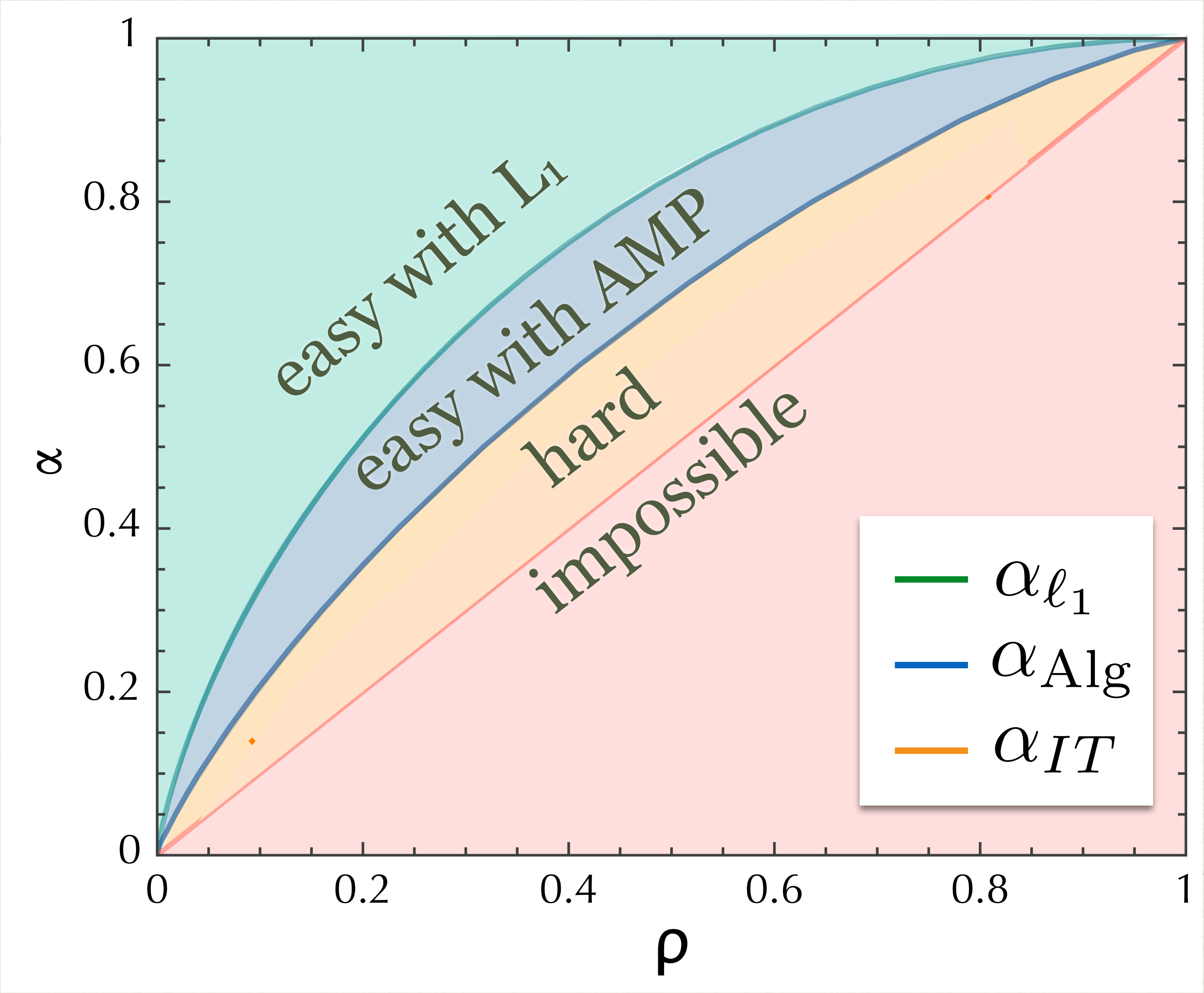
$$y_\mu = \mathrm{sign}\Big(\big| \sum_{i=1}^{p} X_{\mu i} w_i \big| - K \Big) \qquad w_i \in \{-1, +1\} \qquad \begin{matrix} n \to \infty \\ p \to \infty \end{matrix} \qquad n/p = \Theta(1)$$



Open Problem 2: Can one prove a SOS lower bound for this problem? (for Alex Wein)

\# of samples per dimension n/p

PHASE DIAGRAM OF (NOISELESS) SPARSE LINEAR ESTIMATION

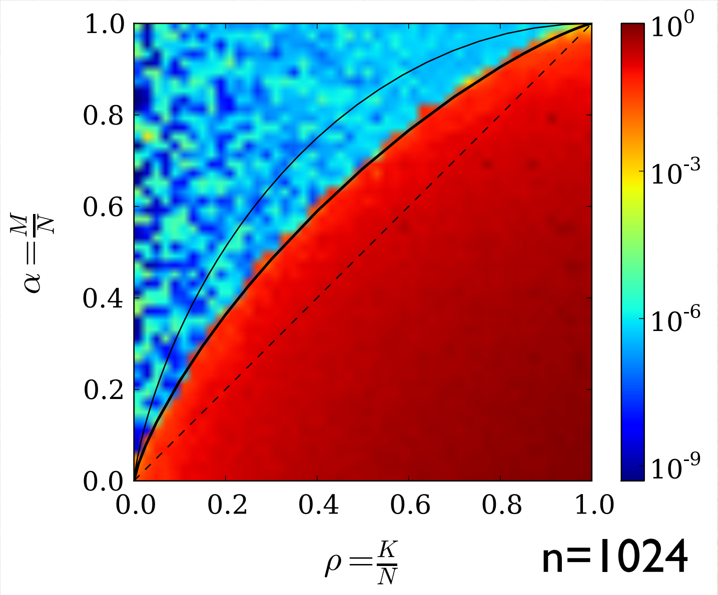


P_w Gauss-Bernoulli(ρ)

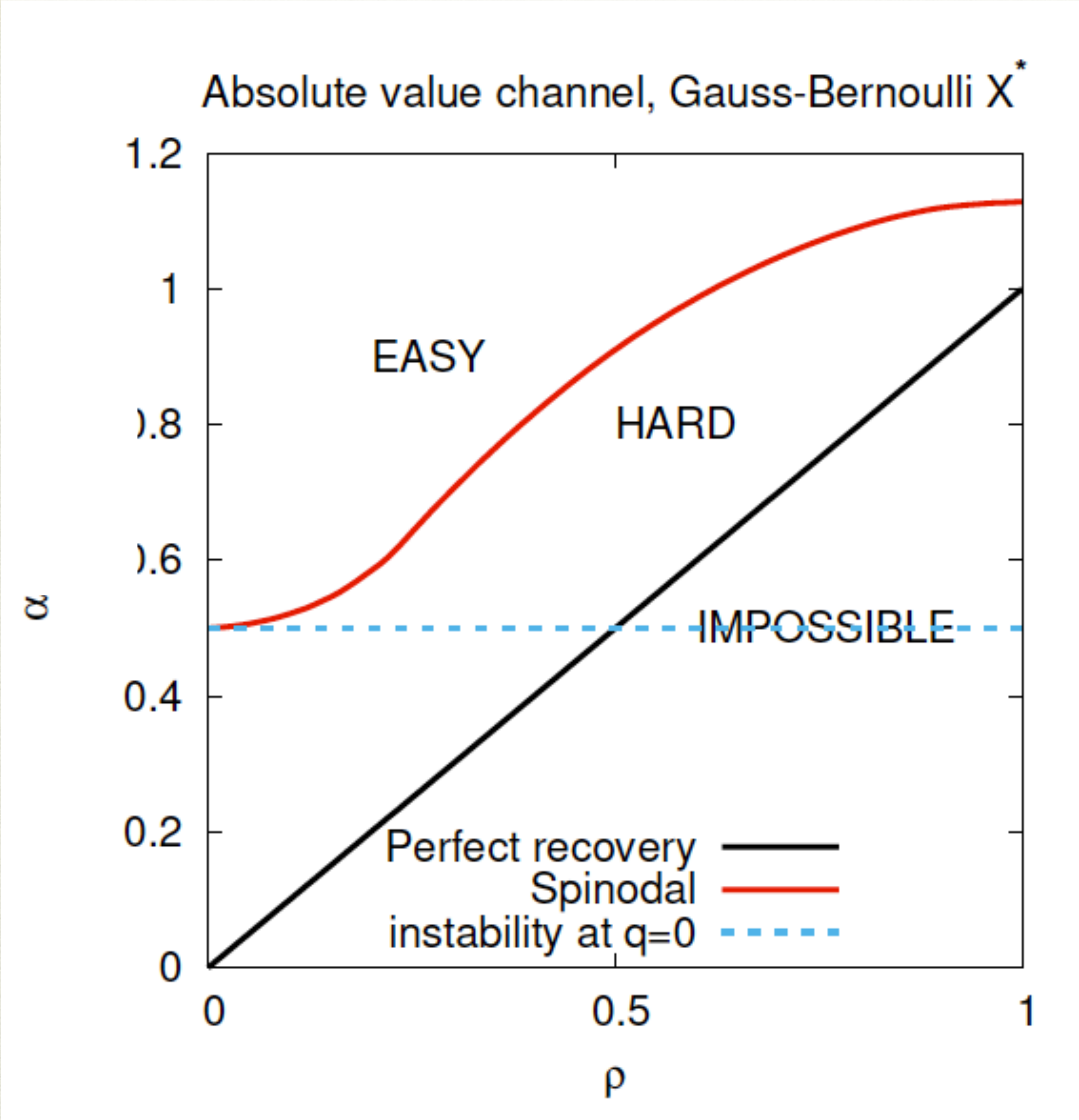
AMP = approximate message passing

Hard for all known algorithms:

$$\alpha_{IT} < \alpha < \alpha_{\mathrm{Alg}}$$



COMPRESSED PHASE RETRIEVAL

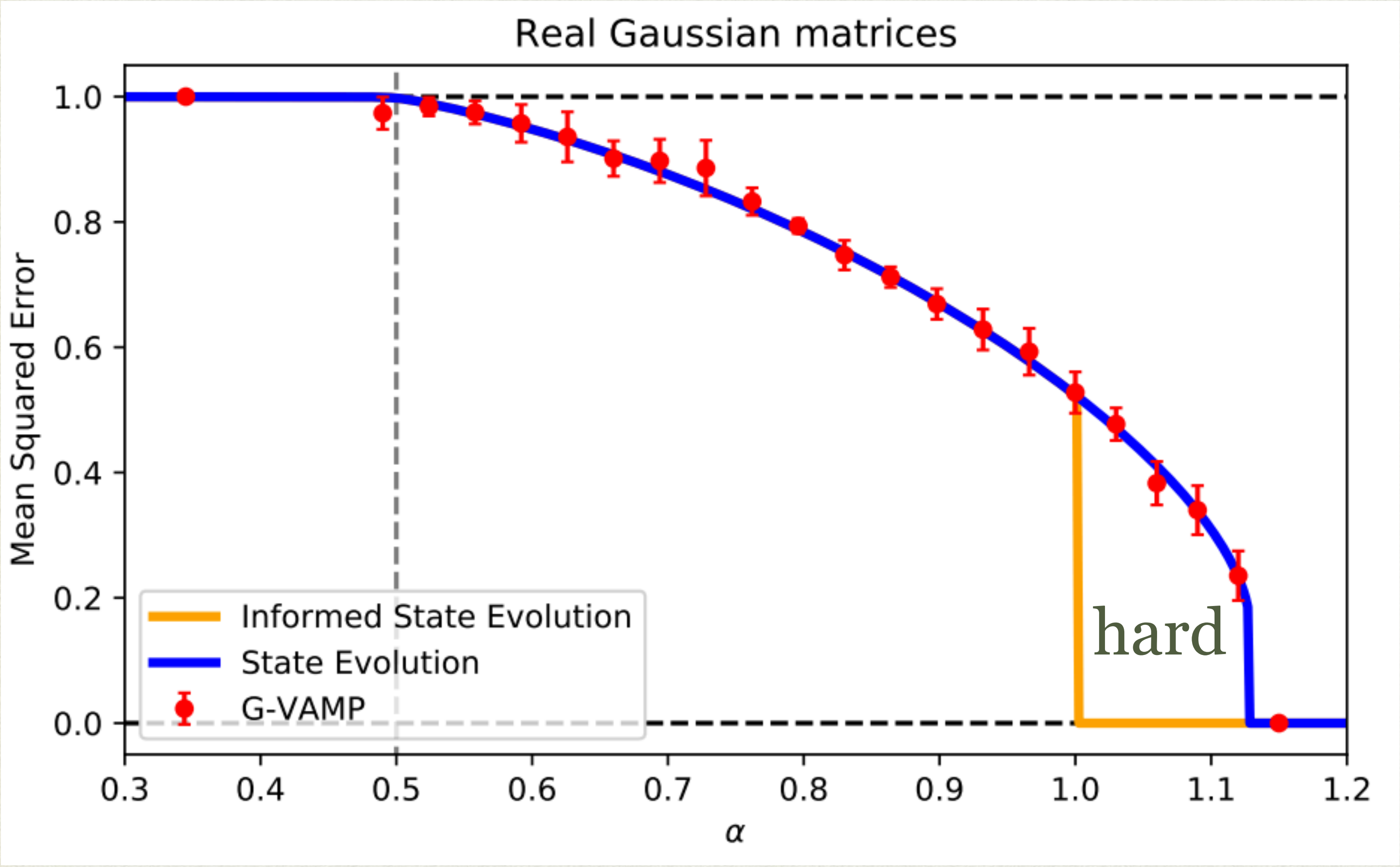


$$y = |Xw^*|$$

P_w Gauss-Bernoulli(ρ)

You cannot sense compressively if you lost the signs!

PHASE RETRIEVAL



$$y = |Xw^*|$$

P_w Gaussian

Hard phase exists even for continuous, non-sparse weights.

SOTA FOR PROOFS

For separable priors P_w, and Gaussian iid inputs X

- Replica theory gives predictions for generic GLM teacher-GLM students. (In non-convex non-Bayes-optimal case RSB is possible.)

- Rigorously proven for (a) Bayes-optimal estimation using adaptive interpolation. (b) ERM for convex losses using Gordon mini-max theory (Gaussian comparison).

- Open problem 3: Prove replica formula (even the RS one) for any non-convex & non-Bayes-optimal case.

STATISTICAL PHYSICS AND COMPUTATION IN HIGH DIMENSION
LECTURE IV

Lenka Zdeborová & Florent Krzakala
(CNRS & CEA Saclay, ENS Paris, EPFL)

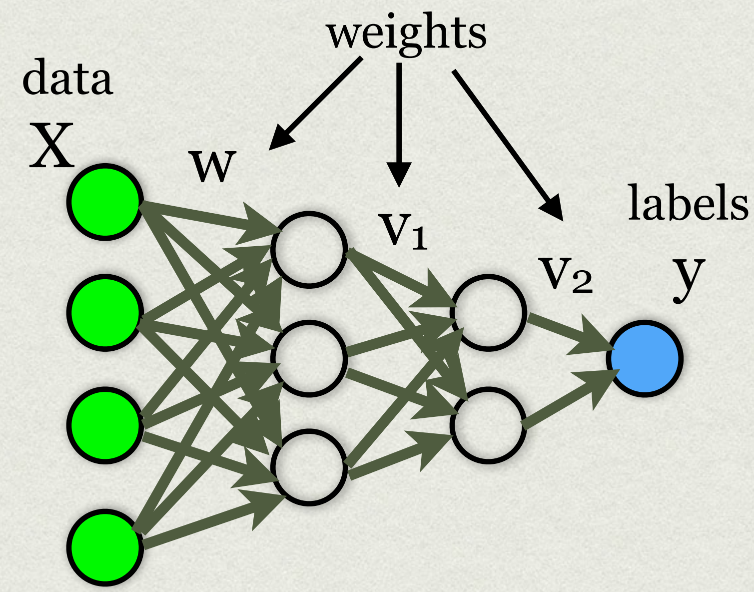
Probability, Geometry, and Computation in High Dimensions Boot Camp
Simons institute for Theory of Computing, 19.-28. 8. 2020

ADDING LAYERS OF HIDDEN VARIABLES

ADDING HIDDEN UNITS

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

Committee machine



- p input units
- M hidden units
- output unit
- n training samples

L=3 layers

w learned, v_1 & v_2 fixed

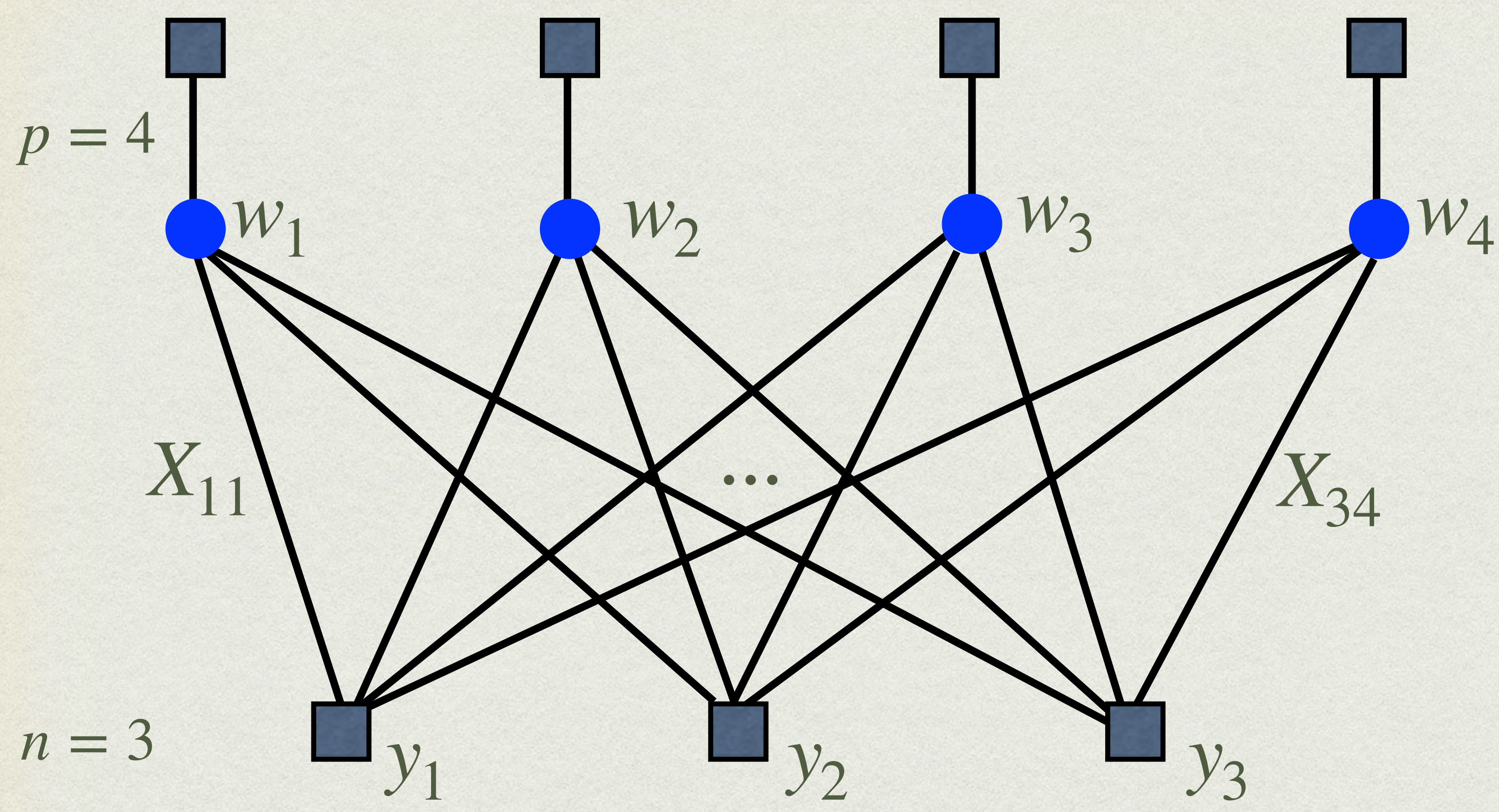
$K \leq M$ hidden units in the 1st layer

Limit: $n \to \infty$, $p \to \infty$ $\quad \alpha = n/p = \Theta(1) \quad M = \Theta(1)$

$P_{\text{data}}(y_\mu, \mathbf{X}_\mu)$: X Gaussian i.i.d., y from a teacher.

Replica solution by Schwarze'92.

GRAPHICAL MODEL



High-dimensional limit:

$p \to \infty, n \to \infty$

$\alpha \equiv n/p = \Theta(1)$

$K = \Theta(1)$

$w \in \mathbb{R}^{p \times K}, \mathbf{w}_i \in \mathbb{R}^K$

$\mathbf{y} \in \mathbb{R}^n, y_\mu \in \mathbb{R}$

$X \in \mathbb{R}^{n \times p}, \mathbf{X}_\mu \in \mathbb{R}^p, X_{\mu i} \in \mathbb{R}$

Probability distribution:

$$P(w \,|\, \mathbf{y}, X) = \frac{1}{Z(\mathbf{y}, X)} \prod_{i=1}^{p} P_w(\mathbf{w}_i) \prod_{\mu=1}^{n} P_{\text{out}}(y_\mu, \sum_i X_{\mu i} w_{ik}, \mathbf{v})$$

Example: $P_{\text{out}}(y_\mu, \mathbf{X}_\mu \cdot \mathbf{w}_k) = \mathbb{1}\left[y_\mu = \text{sign}(\sum_i X_{\mu i} w_{i1}) + \text{sign}(\sum_i X_{\mu i} w_{i2})\right]$

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

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

2018

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.

Essentially GLM with K-dimensional vectors, order parameters K x K matrices.

Theorem 2.1 (Replica formula) *Suppose (H1): The prior P_0 has bounded support in $\mathbb{R}^K$; (H2): The activation $\varphi_{\rm out} : \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}$ 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:*

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

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

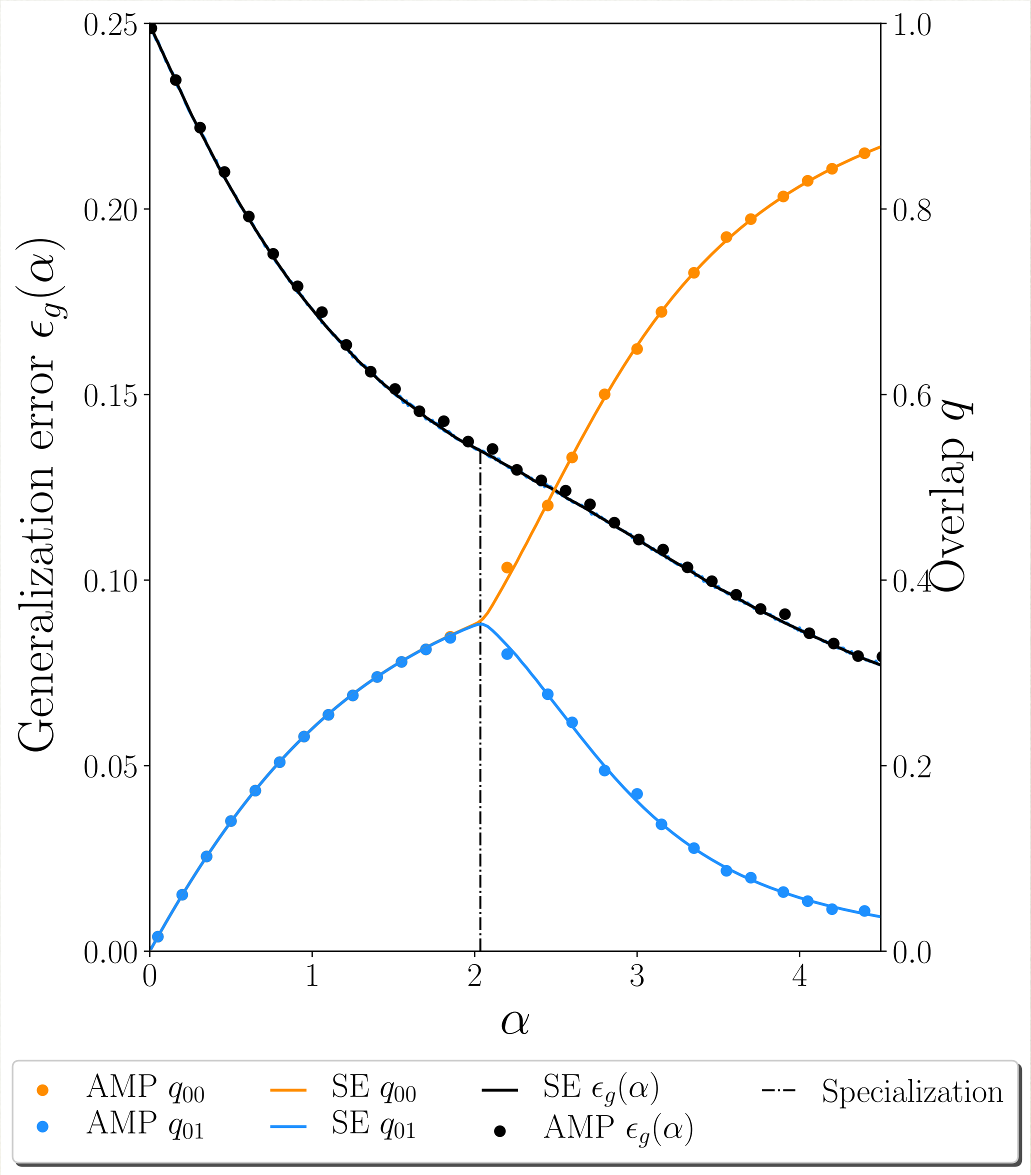
SPECIALISATION TRANSITION

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

hidden units
K=2

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

- 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.



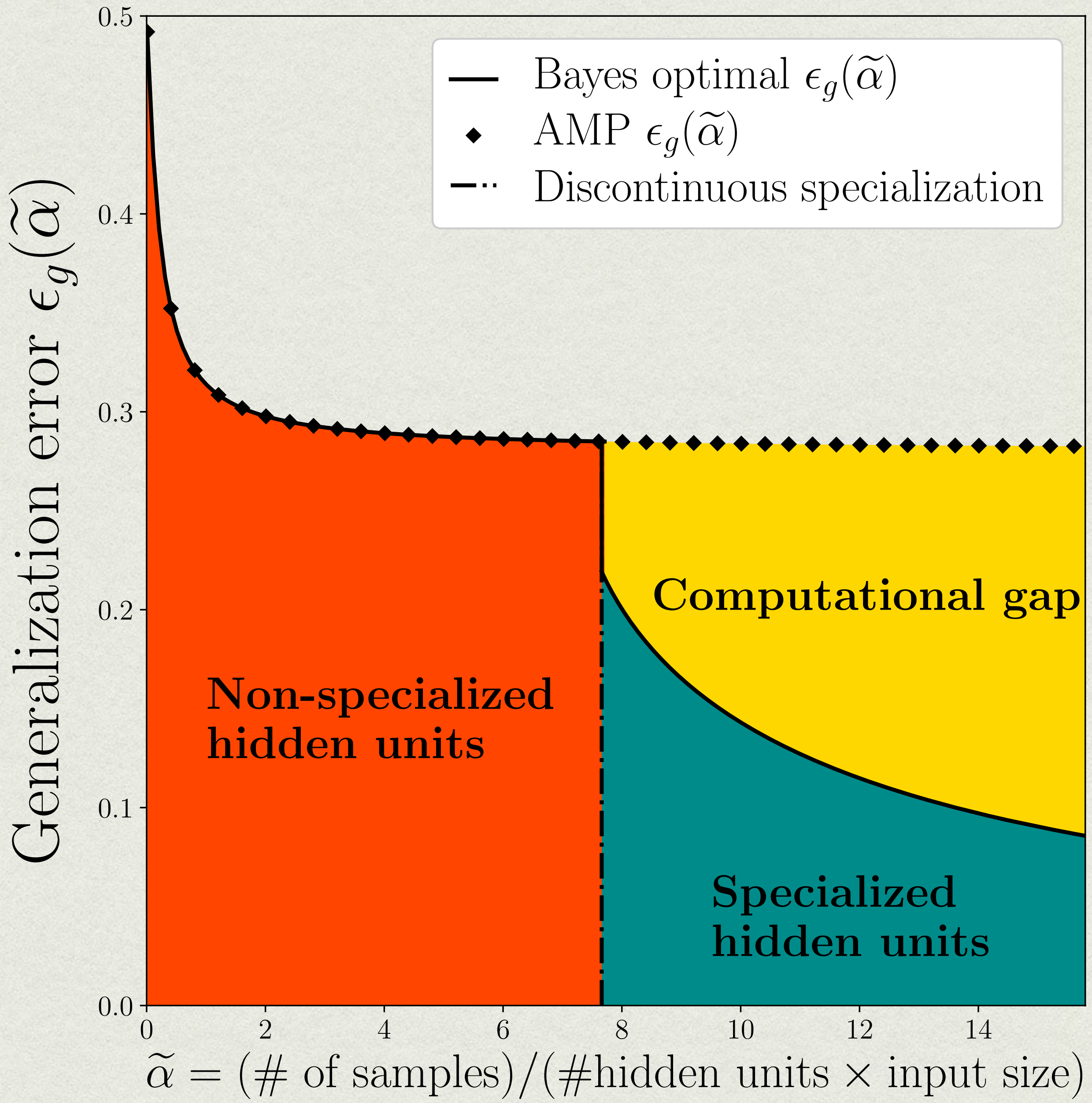
COMPUTATIONAL GAP

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

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

hidden units $K \gg 1$ (after taking $n, p \to \infty$)

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



MORE HIDDEN UNITS?

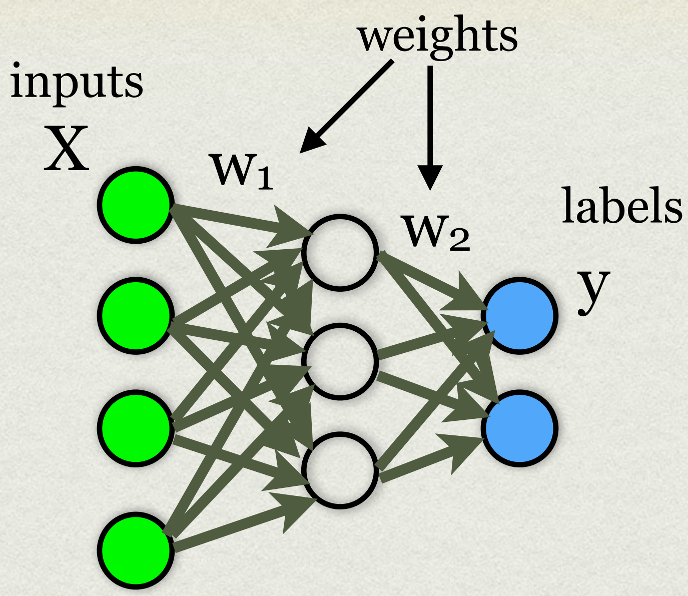
TWO-(EXTENSIVE)LAYERS PERCEPTRON

p # input units

k # hidden units

m # output units

n training samples



2 layers

w_1 & w_2 learned

Limit: $n \to \infty$ $\quad k \to \infty \quad n/p = \Theta(1)$

$p \to \infty \quad m \to \infty \quad k/p = \Theta(1)$

$m/p = \Theta(1)$

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

Optimal generalisation error of the student network?

No known closed-form (not even heuristic replica) formula.

GOING DEEP (MULTI-LAYER)

- Learning multiple (more than one) layers entirely open even for a single (extensive) hidden layer.
  - O(1) hidden layer = committee machine. Linear networks - not expressive. NTK - no feature learning. Single hidden layer much larger than dimension = mean field limit - no closed high-d formula.
- Deep generative priors for the vector w. (e.g. Manoel, FK, Mezard, LZ, ISIT, 1701.06981; Gabrié, Luneau, Barbier, Macris, FK, LZ, NeurIPS, 1805.09785; Aubin, Loureiro, Baker, FK, LZ, MSML, 1912.02008)
- Data samples coming from learned (deep) generative neural networks. (Goldt, Mezard, FK, LZ, 1909.11500; Gerace, Loureiro, FK, Mezard, LZ, ICML, 2002.09339; Goldt, Reeves, Mezard, FK, LZ, 2006.14709)

GENERATIVE PRIORS

e.g. Bora, Jalal, Price, Dimakis'17;

$$\mathbf{y} = \sigma(F\mathbf{s}^*)$$

- G: known measurement matrix of the apparatus.
- s^* signal from a range of generative neural network learned from data. There exists $\mathbf{x}^* \in \mathbb{R}^k$, $k \ll p$ such that

$$\mathbf{s}^* = \varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}\mathbf{x}^*))))$$

$\varphi^{(i)}, W^{(i)}, i = 1,\ldots,L$ known, after training

Signal comes from a generative neural network

GENERATIVE PRIORS

$$\mathbf{y} = \sigma(F\varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}\mathbf{x}*)))))$$

- G: known measurement matrix of the apparatus.

- $s*$ signal from a range of generative neural network learned from data. There exists $\mathbf{x}* \in \mathbb{R}^k$, $k \ll p$ such that

$$\mathbf{s}* = \varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}\mathbf{x}*))))$$

$\varphi^{(i)}, W^{(i)}, i = 1,\ldots,L$ known, after training

Signal comes from a generative neural network

GENERATIVE PRIORS

$$\mathbf{y} = \varphi^{(5)}(W^{(5)}\varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}\mathbf{x}^*)))))$$

- G: known measurement matrix of the apparatus.

- s^* signal from a range of generative neural network learned from data. There exists $\mathbf{x}^* \in \mathbb{R}^k$, $k \ll p$ such that

$$\mathbf{s}^* = \varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}\mathbf{x}^*))))$$

$\varphi^{(i)}, W^{(i)}, i = 1,\ldots,L$ known, after training

Signal comes from a generative neural network

SOLVABLE CASE

$$\mathbf{y} = \varphi^{(5)}(W^{(5)}\varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}\mathbf{x}^*)))))$$

- $W^{(i)}, i = 1,\ldots,L$ random iid (or random rotationally invariant). $\forall i$ the aspect ratios of $W^{(i)}$ are $\Theta(1)$.
- Latent variables $\mathbf{x}^* \in \mathbb{R}^k$ generated iid from a prior P_X. $\mathbf{y}$ generated by a teacher.
- Goal: From the knowledge of $\mathbf{y}$, $W^{(i)}, i = 1,\ldots,L$ estimate back the $\mathbf{x}^*$

KEY OBSERVATION ABOUT G-AMP

$$P(x\,|\,y, W) = \frac{1}{Z(y,F)} \prod_{\mu=1}^{n} P_{\text{out}}(y_\mu\,|\,z_\mu \equiv W_\mu \cdot x) \prod_{i=1}^{p} P_X(x_i)$$

Marginals of x and z:

$$\boxed{\begin{aligned} \mu_z(z_\mu) &= \frac{1}{Z_Z} P(y_\mu|z_\mu)\mathcal{N}(z_\mu;\omega_\mu, V_\mu) \\ \mu_x(x_i) &= \frac{1}{Z_X} P_X(x_i) e^{-\frac{1}{2}A_i x_i^2 + B_i x_i} \end{aligned}}$$

GAMP update:

$$V_\mu(t) = \sum_i \left[W_{\mu i}\right]^2 \sigma_i(t),$$

$$\omega_\mu(t) = \sum_i W_{\mu i} \hat{h}_i(t) - V_\mu(t) g_\mu(t-1),$$

$$A_i(t) = -\sum_\mu \left[W_{\mu i}\right]^2 \partial_\omega g_\mu(t),$$

$$B_i(t) = \sum_\mu W_{\mu i} g_\mu(t) + A_i(t)\hat{x}_I(t).$$

$$\hat{x}_i = \mathbb{E}_{\mu_x}(x_i)$$

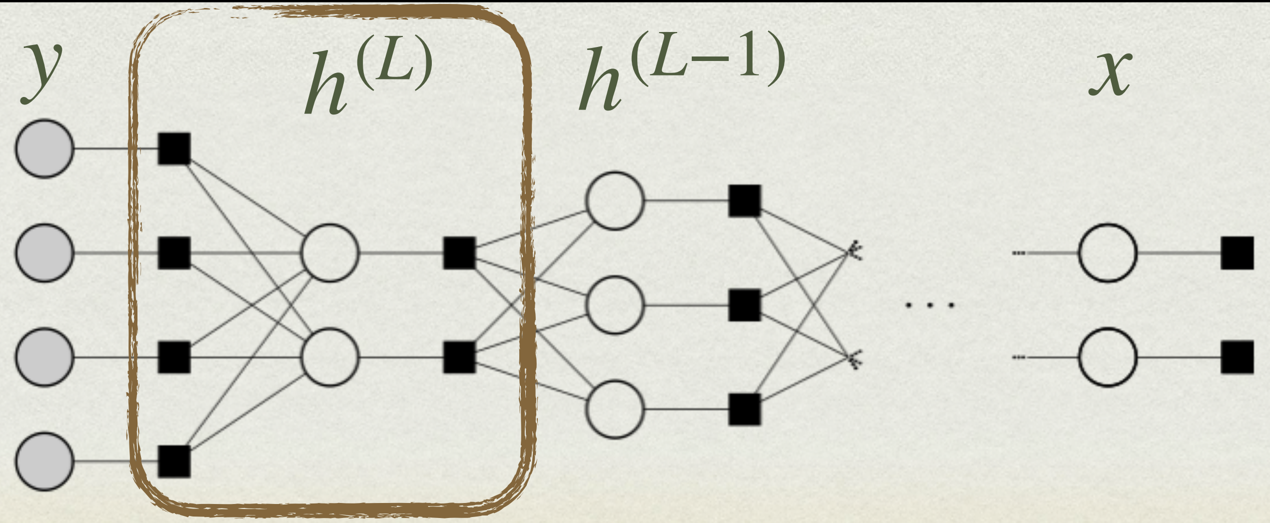
$$g_\mu = \mathbb{E}_{\mu_z} \frac{z_\mu - \omega_\mu}{V_\mu}$$

MULTI-LAYER GENERALISED LINEAR ESTIMATION

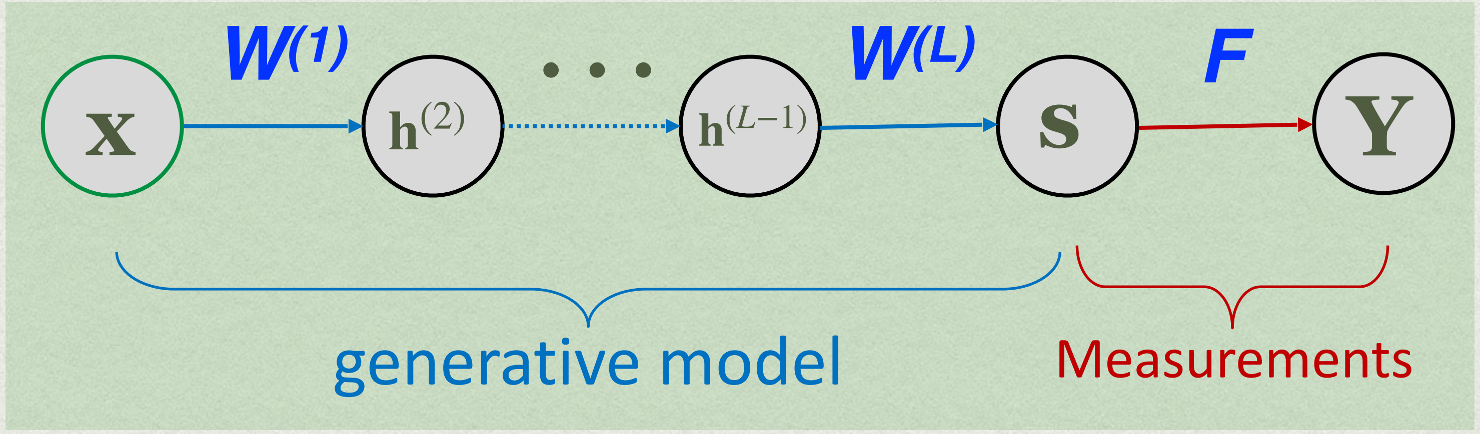
$$\mathbf{y} = \varphi^{(5)}(W^{(5)}\varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}\mathbf{x}^*)))))$$

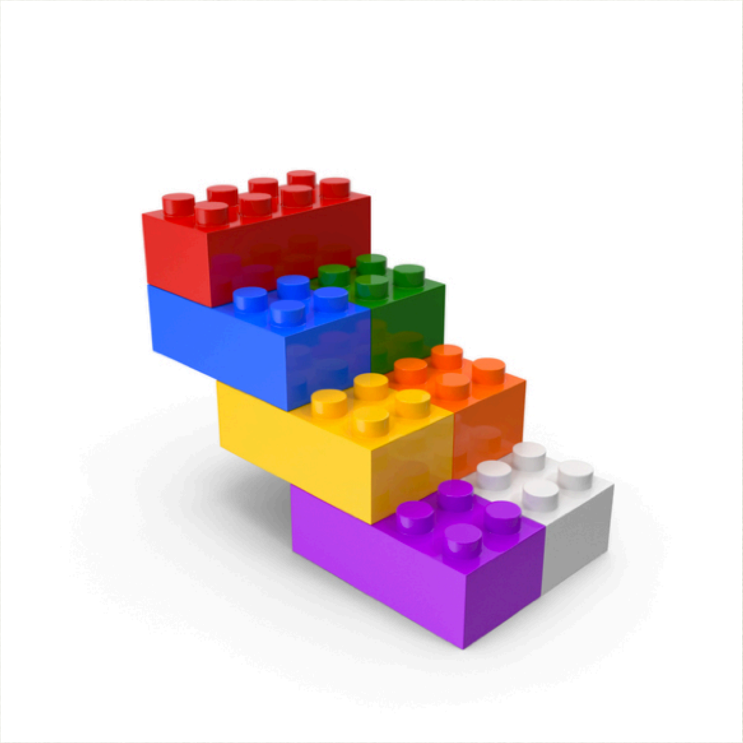
Introduce auxiliary variables h:

$$\begin{aligned} y_\mu &\sim P_{\mathrm{out}}^{(L)}\left(y_\mu \,\middle|\, \textstyle\sum_{i=1}^{n_L} W_{\mu i}^{(L)} h_i^{(L)}\right) \\ h_\mu^{(L)} &\sim P_{\mathrm{out}}^{(L-1)}\left(h_\mu^{(L)} \,\middle|\, \textstyle\sum_{i=1}^{n_{L-1}} W_{\mu i}^{(L-1)} h_i^{(L-1)}\right) \\ &\vdots \\ h_\mu^{(2)} &\sim P_{\mathrm{out}}^{(1)}\left(h_\mu^{(2)} \,\middle|\, \textstyle\sum_{i=1}^{n_1} W_{\mu i}^{(1)} x_i\right), \\ x_\mu &\sim P_X(x_\mu) \end{aligned}$$



A "LEGO" PRINCIPLE





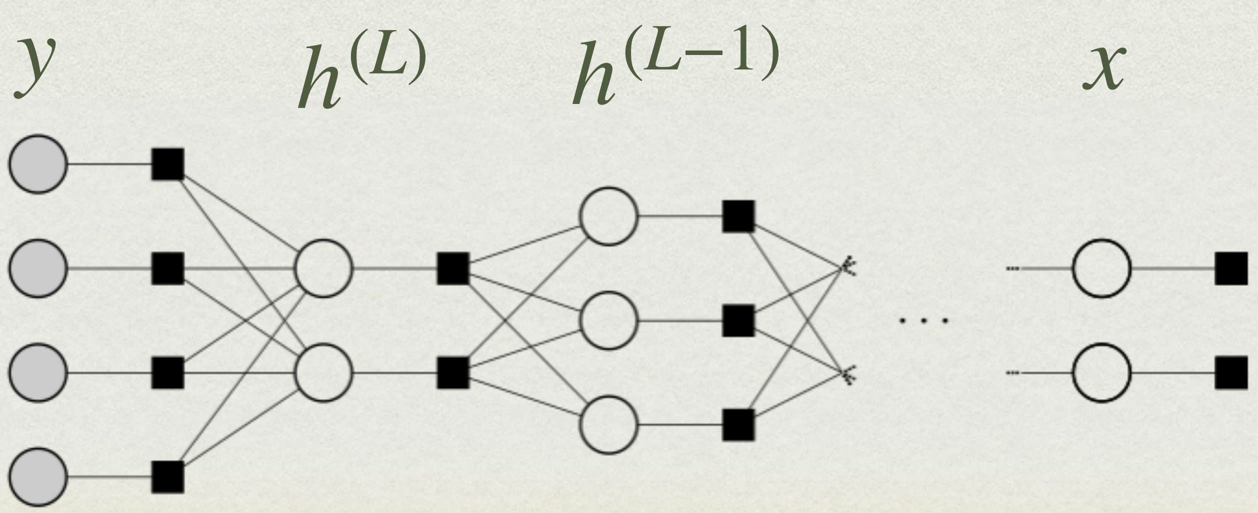
Free energy (and estimation/generalization error) of chains (trees) of solvable graphical models follow by recursively combining individual blocks.

MULTI-LAYER AMP

Each layer is G-AMP with an effective prior and an effective output channel:

$$P_X^{\mathrm{eff}}(h^{(\ell)}\,|\,V^{(\ell-1)},\omega^{(\ell-1)}) = \int \mathrm{d}z\, P_{\mathrm{out}}^{(\ell-1)}(h^{(\ell)}\,|\,z)\frac{e^{-\frac{(z-\omega^{(\ell-1)})^2}{2V^{(\ell-1)}}}}{\sqrt{2\pi V^{(\ell-1)}}}$$

$$P_{\mathrm{out}}^{\mathrm{eff}}(z^{(\ell)}\,|\,A^{(\ell+1)}B^{(\ell+1)}) = \int \mathrm{d}h\, P_{\mathrm{out}}^{(\ell)}(h\,|\,z^{(\ell)})\, e^{-\frac{1}{2}A^{(\ell+1)}h^2+B^{(\ell+1)}h}$$



MULTI-LAYER AMP

Update:

$$V_\mu^{(\ell)}(t) = \sum_i \left[W_{\mu i}^{(\ell)}\right]^2 \sigma_i^{(\ell)}(t),$$

$$\omega_\mu^{(\ell)}(t) = \sum_i W_{\mu i}^{(\ell)} \hat{h}_i^{(\ell)}(t) - V_\mu^{(\ell)}(t) g_\mu^{(\ell)}(t-1),$$

$$A_i^{(\ell)}(t) = -\sum_\mu \left[W_{\mu i}^{(\ell)}\right]^2 \partial_\omega g_\mu^{(\ell)}(t),$$

$$B_i^{(\ell)}(t) = \sum_\mu W_{\mu i}^{(\ell)} g_\mu^{(\ell)}(t) + A_i^{(\ell)}(t) \hat{h}_i^{(\ell)}(t).$$

G-AMP

---

where:

$$g_\mu^{(\ell)}(t) = \partial_\omega \log \mathcal{Z}^{(\ell)}(A_\mu^{(\ell+1)}, B_\mu^{(\ell+1)}, V_\mu^{(\ell)}, \omega_\mu^{(\ell)}),$$

$$\hat{h}_i^{(\ell)}(t+1) = \partial_B \log \mathcal{Z}^{(\ell-1)}(A_i^{(\ell)}, B_i^{(\ell)}, V_i^{(\ell-1)}, \omega_i^{(\ell-1)}),$$

$$\mathcal{Z}^{(\ell)}(A^{(\ell+1)}, B^{(\ell+1)}, V^{(\ell)}, \omega^{(\ell)}) \equiv \frac{1}{\sqrt{2\pi V^{(\ell)}}} \int \mathrm{d}h\, \mathrm{d}z\, P_{\text{out}}^{(\ell)}(h|z)\, e^{-\frac{1}{2}A^{(\ell+1)}h^2 + B^{(\ell+1)}h}\, e^{-\frac{(z-\omega^{(\ell)})^2}{2V^{(\ell)}}}$$

MULTI-LAYER GENERALISED LINEAR ESTIMATION

$$\mathbf{y} = \varphi^{(L)}(W^{(L)} \dots \varphi^{(1)}(W^{(1)}\mathbf{x}^*))$$

Generalizing single layer results (Manoel, FK, Mezard, LZ, ISIT, 1701.06981; Gabrié, Luneau, Barbier, Macris, FK, LZ, NeurIPS, 1805.09785)

- Asymptotically exact mutual information/free energy.
- MMSE of Bayes-optimal inference.
- State evolution for asymptotic performance of ML-AMP.
- Regions where ML-AMP asymptotically optimal.
- Proof for the Bayes-optimal case (so far only for 2 layers)

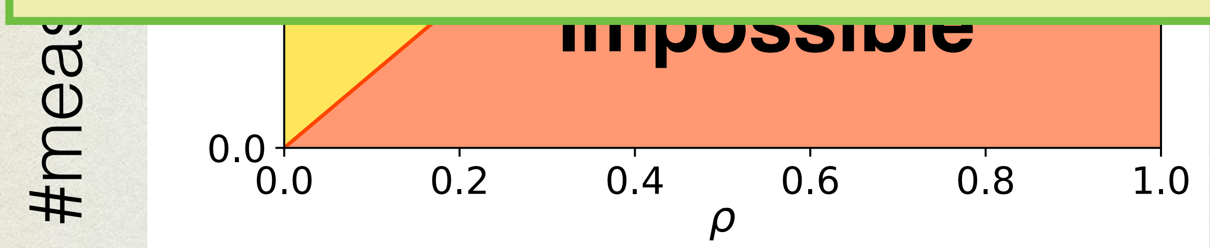
EX: PHASE RETRIEVAL $y = |Fs|$

Aubin, Loureiro, Baker, FK, LZ, 1912.02008

Sparse prior

With generative priors, compressive phase retrieval is possible. Hard phase shrinks/disappears when using generative priors

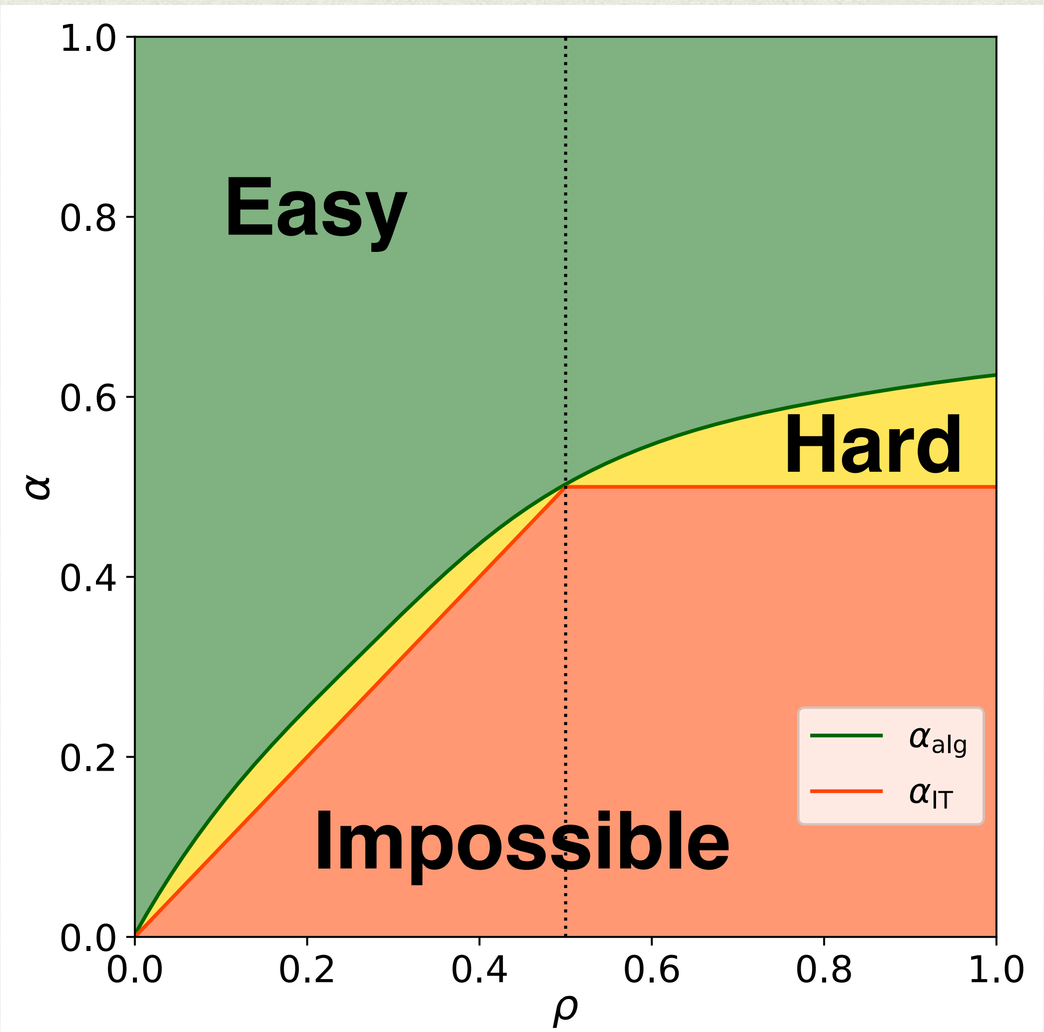
(also Hand, Leong, Voroninski'18)



$$\rho_S = \frac{\#\text{non}-\text{zero components}}{\text{dimension of signal}}$$

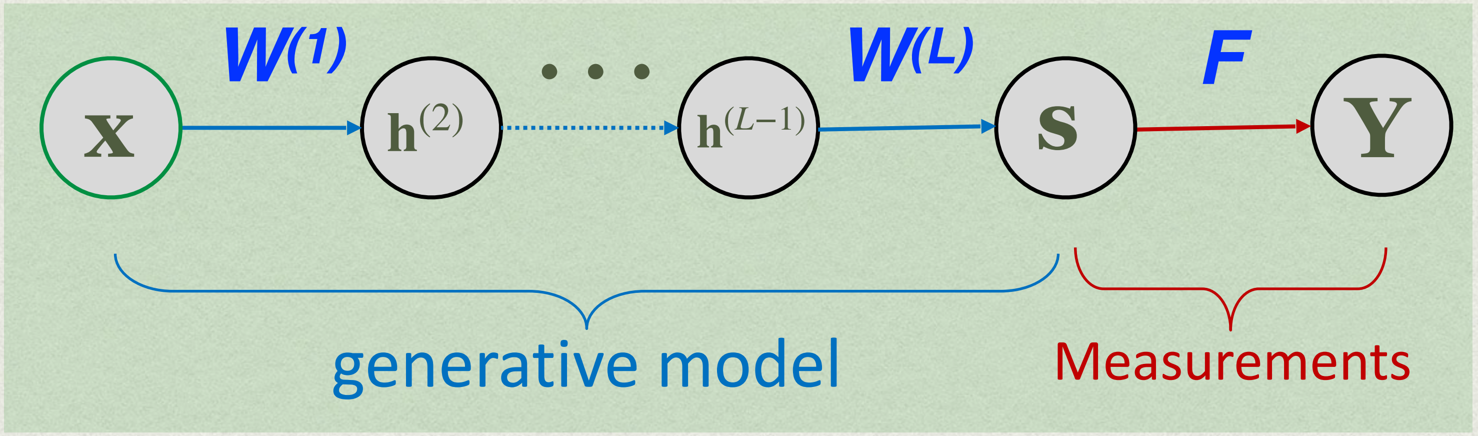
Generative prior

$s = \mathrm{Relu}(Wx)$

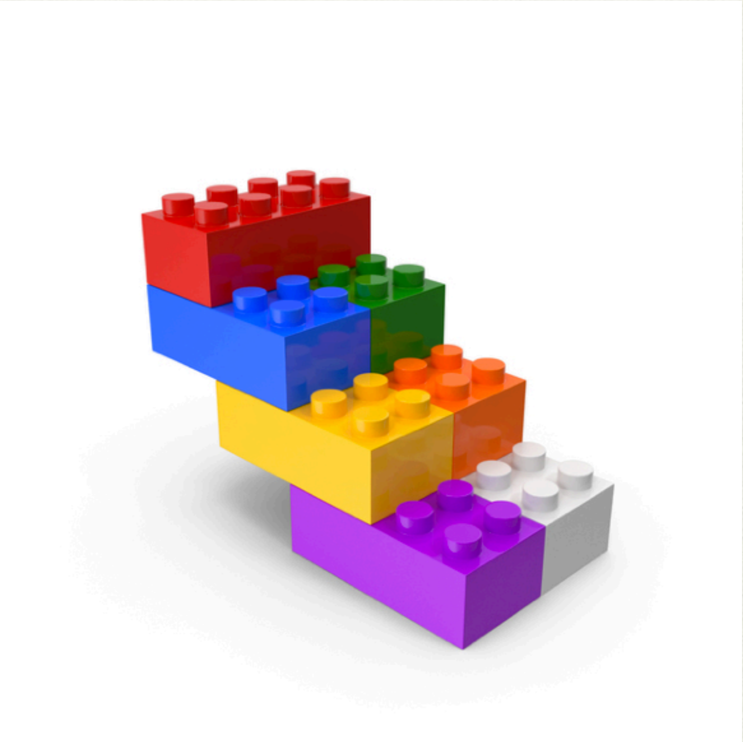


$$\rho_G = \frac{\#\text{latent variables}}{\text{dimension of signal}}$$

A "LEGO" PRINCIPLE



Free energy of chains of solvable graphical models is solvable.



Modular implementation of AMP for any tree-like probabilistic graphical model.

Baker, Aubin, FK, LZ, 2004.01571

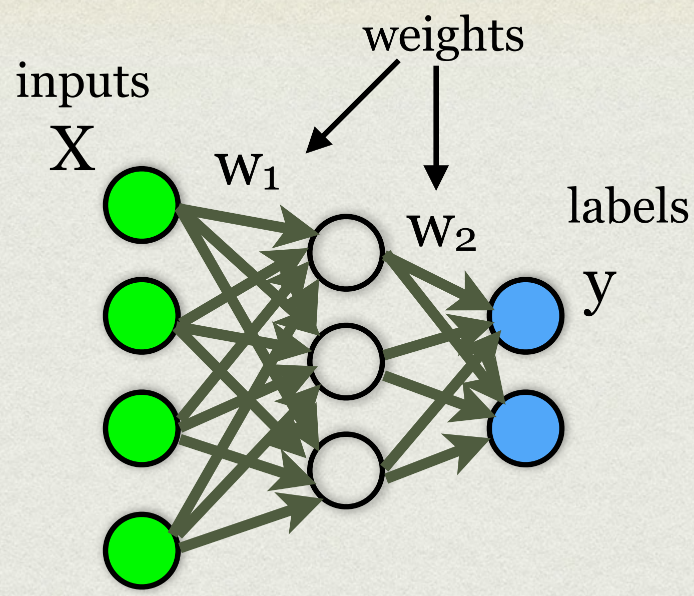
TWO-(EXTENSIVE)LAYERS PERCEPTRON

p # input units

k # hidden units

m # output units

n training samples



2 layers
w_1 & w_2 learned

Limit: $n \to \infty$ $\quad k \to \infty$ $\quad n/p = \Theta(1)$

$p \to \infty$ $\quad m \to \infty$ $\quad k/p = \Theta(1)$

$m/p = \Theta(1)$

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

Optimal generalisation error of the student network?

No known closed-form (not even heuristic replica) formula.

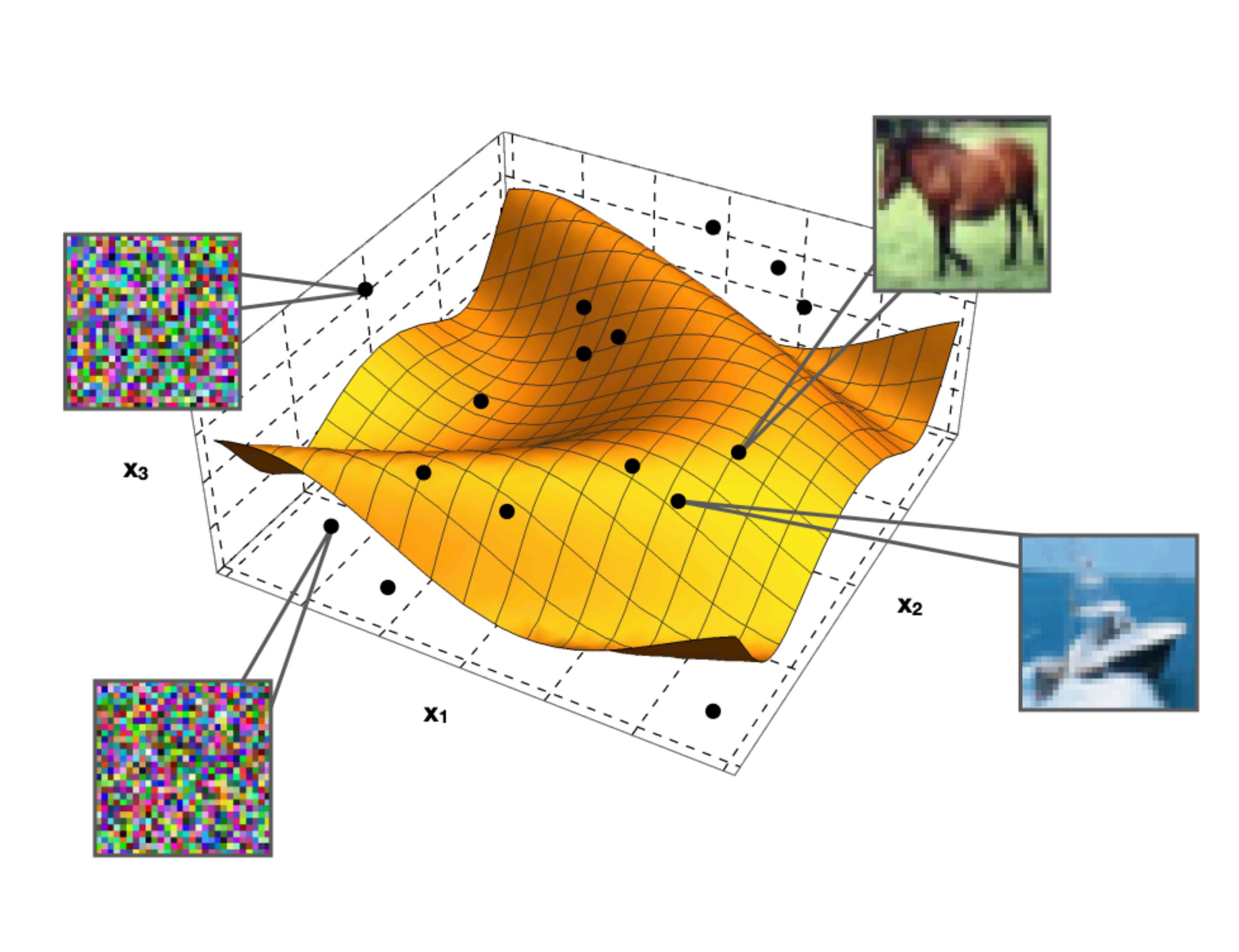
GOING DEEP (MULTI-LAYER)

- Learning multiple (more than one) layers entirely open even for a single (extensive) hidden layer.
  - O(1) hidden layer = committee machine. Linear networks - not expressive. NTK - no feature learning. Single hidden layer much larger than dimension = mean field limit - no closed high-d formula.
- Deep generative priors for the vector w. (e.g. Manoel, FK, Mezard, LZ, ISIT, 1701.06981; Gabrié, Luneau, Barbier, Macris, FK, LZ, NeurIPS, 1805.09785; Aubin, Loureiro, Baker, FK, LZ, 1912.02008)
- $\boxed{\text{Data samples coming from learned (deep) generative neural networks.}}$ (Goldt, Mezard, FK, LZ, 1909.11500; Gerace, Loureiro, FK, Mezard, LZ, ICML, 2002.09339; Goldt, Reeves, Mezard, FK, LZ, 2006.14709)

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).

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

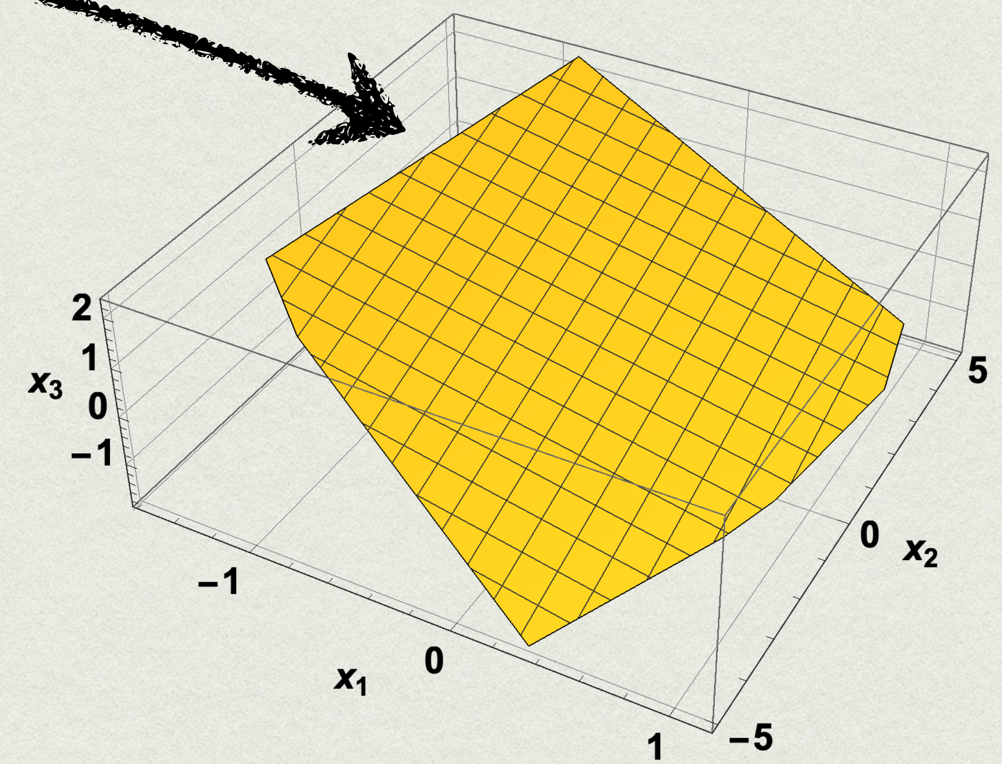
$X_\mu \in \mathbb{R}^p \quad C_\mu \in \mathbb{R}^d \quad F \in \mathbb{R}^{p\times d}$

p input & d latent dimension, p>d.

Hidden manifold model

low-dimensional
sub-space

F

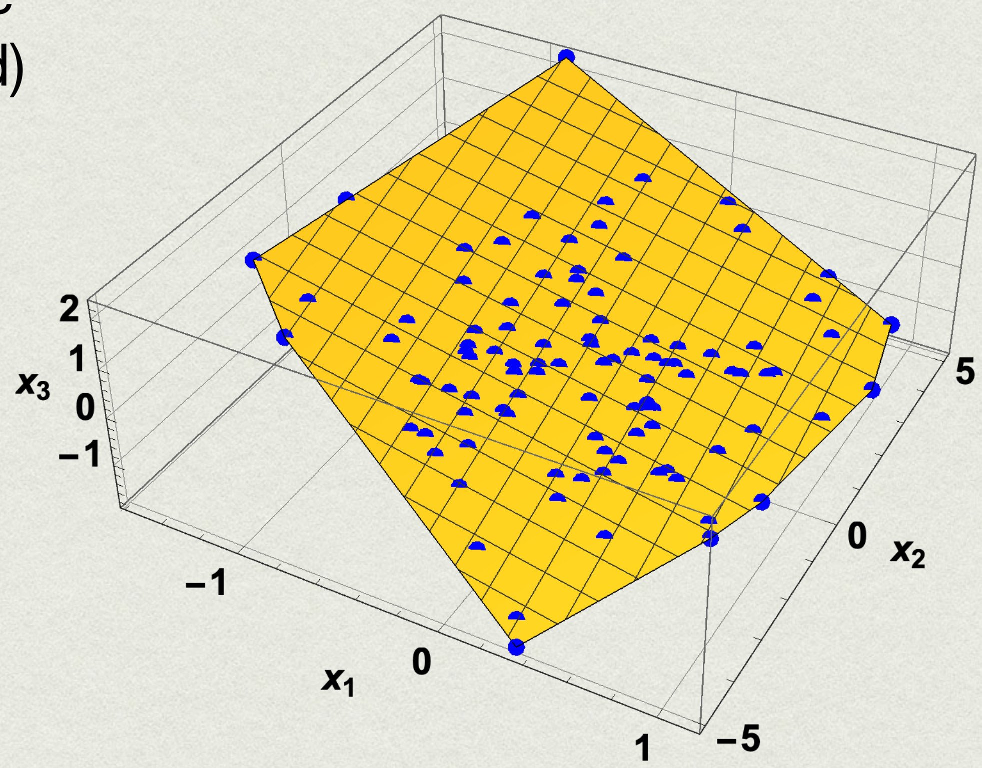


Hidden manifold model

low-dimensional
sub-space

point coordinates
in sub-space
(dimension d)

FC

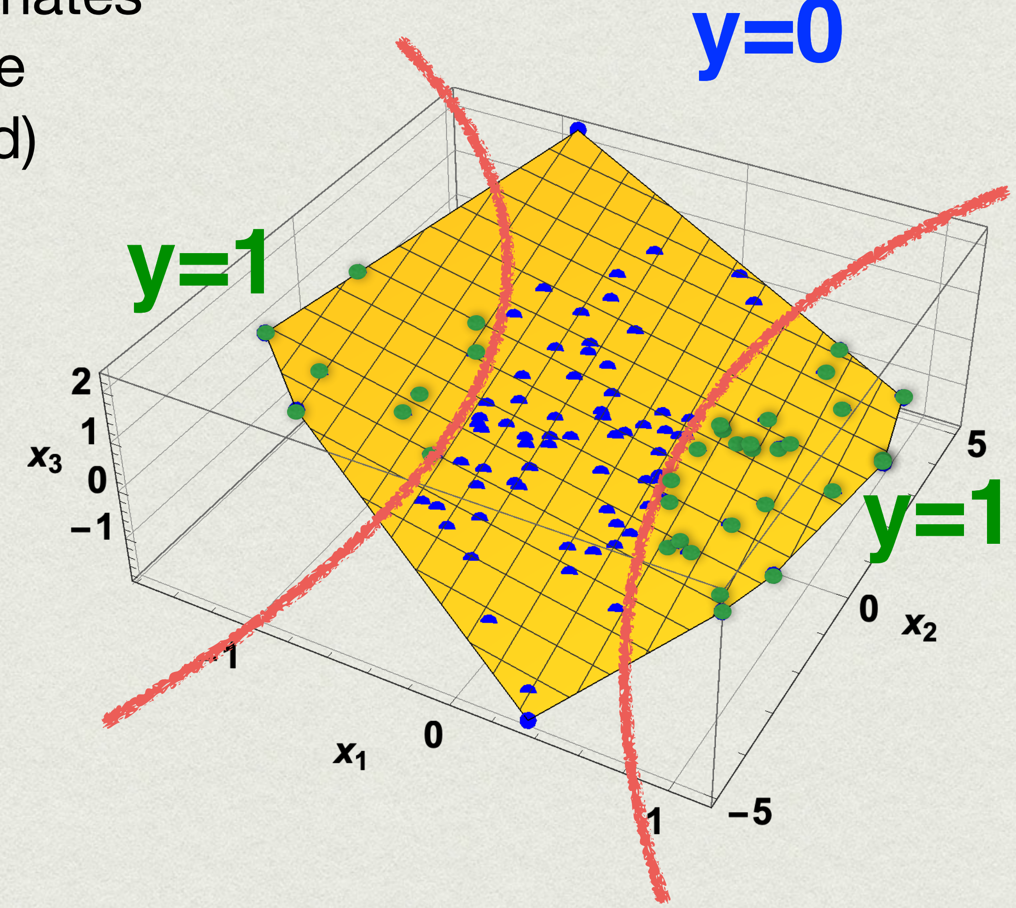


Hidden manifold model

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

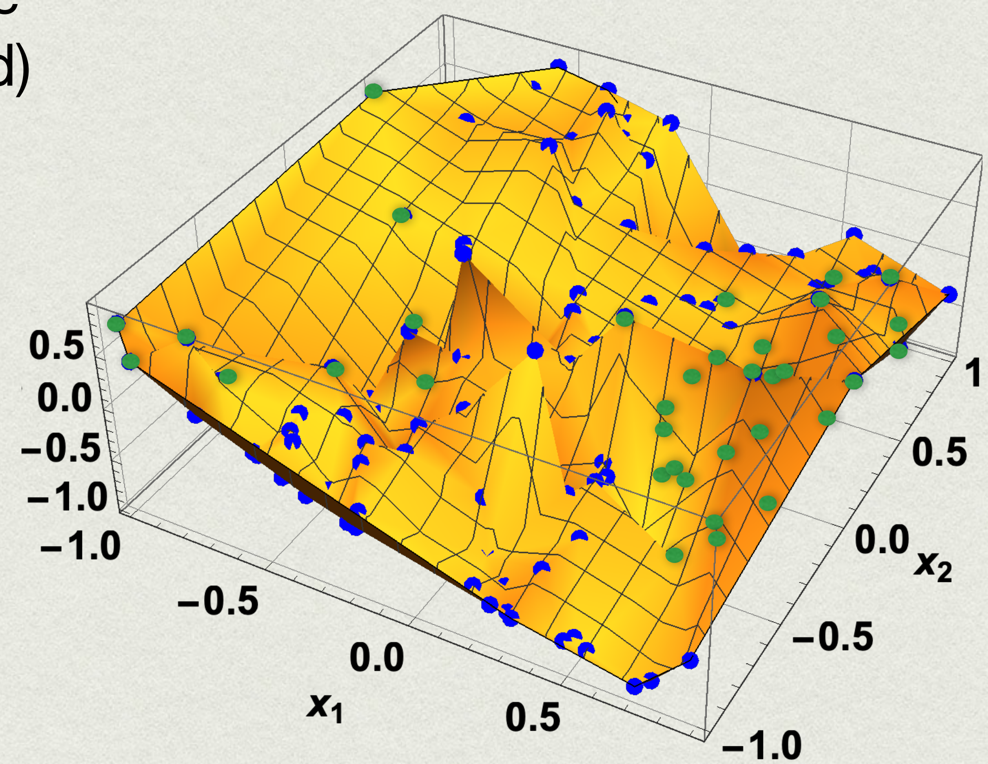
low-dimensional sub-space

point coordinates in sub-space (dimension d)

$$X = f(FC)$$

Data points (dimension p)

Non-linear function



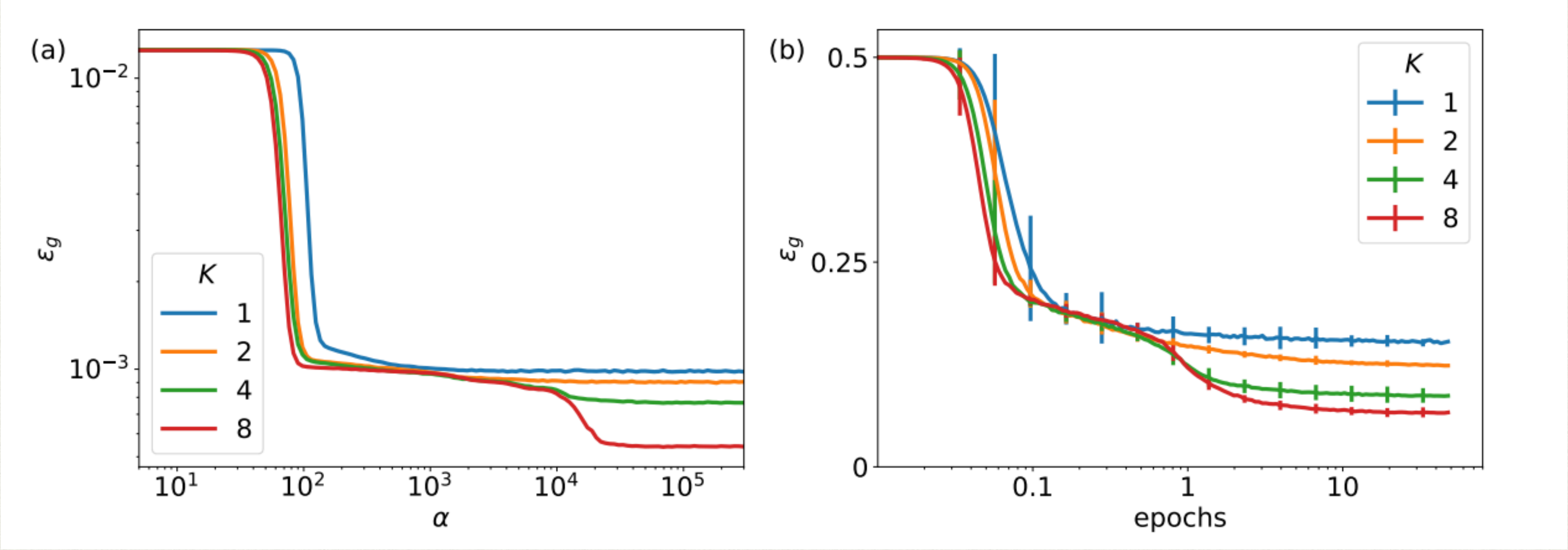
$$Y = g(C)$$

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

HIDDEN MANIFOLD VS MNIST

Hidden manifold (d=10) MNIST (odd vs even):

Neural network: single hidden layer, sigmoidal activation, K hidden units.

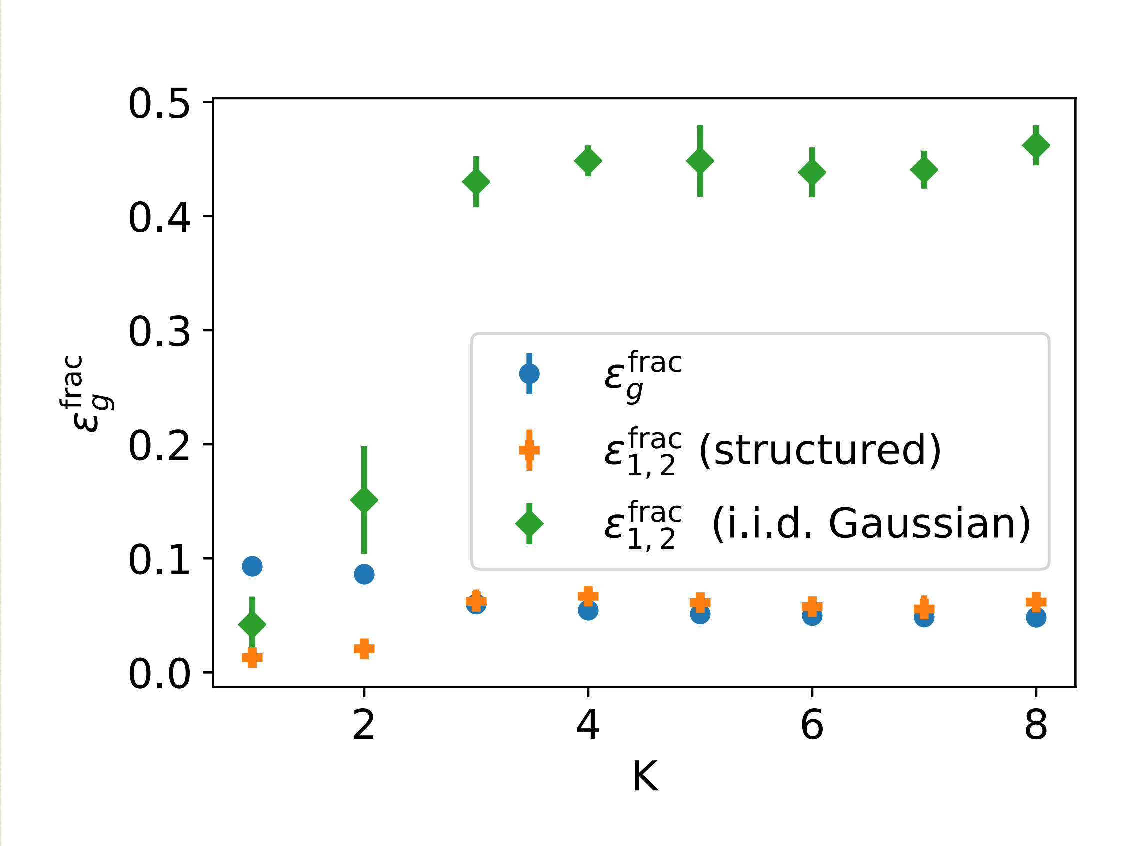


The neural network learns a simpler function first.

MNIST VS HIDDEN MANIFOLD

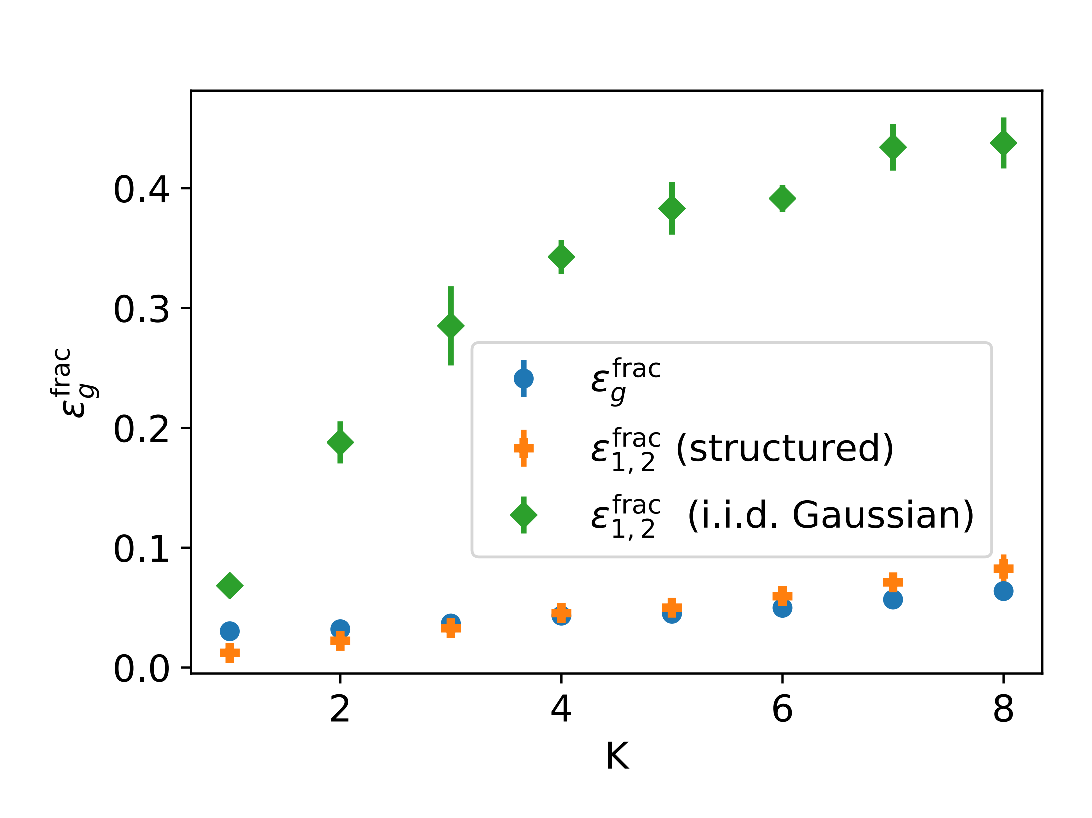
MNIST (odd vs even):

Two independent students do not learn the same function!



Hidden manifold (d=10)

Two independent students do not learn the same function!



SOLVING THE HMM

- With random F, least-square regression, and max-margin: Mei, Montanari'19, Montanari, Ruan, Sohn, Yan'19.
- Generic F, committee machine on (X,y) from HMM with online SGD algorithm. Goldt, FK, Mézard, LZ; arXiv:1909.11500
- Generic F, generalized linear regression on (X,y) from HMM. Gerace, Loureiro, FK, Mézard, LZ, ICML, 2002.09339

Open problem 4: Prove that result (for convex losses).

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

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

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

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

Replica solution

Solution:

Consider the unique fixed point of the following system of equations

$$\begin{cases} \hat{V}_s = \frac{\alpha}{\gamma}\kappa_1^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\partial_\omega \eta(y,\omega_1)}{V}\right], \\ \hat{q}_s = \frac{\alpha}{\gamma}\kappa_1^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\left(\eta(y,\omega_1)-\omega_1\right)^2}{V^2}\right], \\ \hat{m}_s = \frac{\alpha}{\gamma}\kappa_1 \mathbb{E}_{\xi,y}\left[\partial_\omega \mathcal{Z}\left(y,\omega_0\right)\frac{(\eta(y,\omega_1)-\omega_1)}{V}\right], \\ \hat{V}_w = \alpha\kappa_\star^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\partial_\omega \eta(y,\omega_1)}{V}\right], \\ \hat{q}_w = \alpha\kappa_\star^2 \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0\right)\frac{\left(\eta(y,\omega_1)-\omega_1\right)^2}{V^2}\right], \end{cases} \quad \begin{cases} V_s = \frac{1}{\hat{V}_s}\left(1 - z\, g_\mu(-z)\right), \\ q_s = \frac{\hat{m}_s^2+\hat{q}_s}{\hat{V}_s}\left[1 - 2z g_\mu(-z) + z^2 g_\mu'(-z)\right] \\ \qquad -\frac{\hat{q}_w}{(\lambda+\hat{V}_w)\hat{V}_s}\left[-z g_\mu(-z) + z^2 g_\mu'(-z)\right], \\ m_s = \frac{\hat{m}_s}{\hat{V}_s}\left(1 - z\, g_\mu(-z)\right), \\ V_w = \frac{\gamma}{\lambda+\hat{V}_w}\left[\frac{1}{\gamma} - 1 + z g_\mu(-z)\right], \\ q_w = \gamma\frac{\hat{q}_w}{(\lambda+\hat{V}_w)^2}\left[\frac{1}{\gamma} - 1 + z^2 g_\mu'(-z)\right], \\ \qquad +\frac{\hat{m}_s^2+\hat{q}_s}{(\lambda+\hat{V}_w)\hat{V}_s}\left[-z g_\mu(-z) + z^2 g_\mu'(-z)\right], \end{cases}$$

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

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

$\boxed{\kappa_0 = \mathbb{E}\left[\sigma(z)\right], \kappa_1 \equiv \mathbb{E}\left[z\sigma(z)\right], \kappa_\star \equiv \mathbb{E}\left[\sigma(z)^2\right] - \kappa_0^2 - \kappa_1^2}$, 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 $(\nu,\lambda) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\rho & M^\star\\ M^\star & Q^\star\end{pmatrix}\right)$

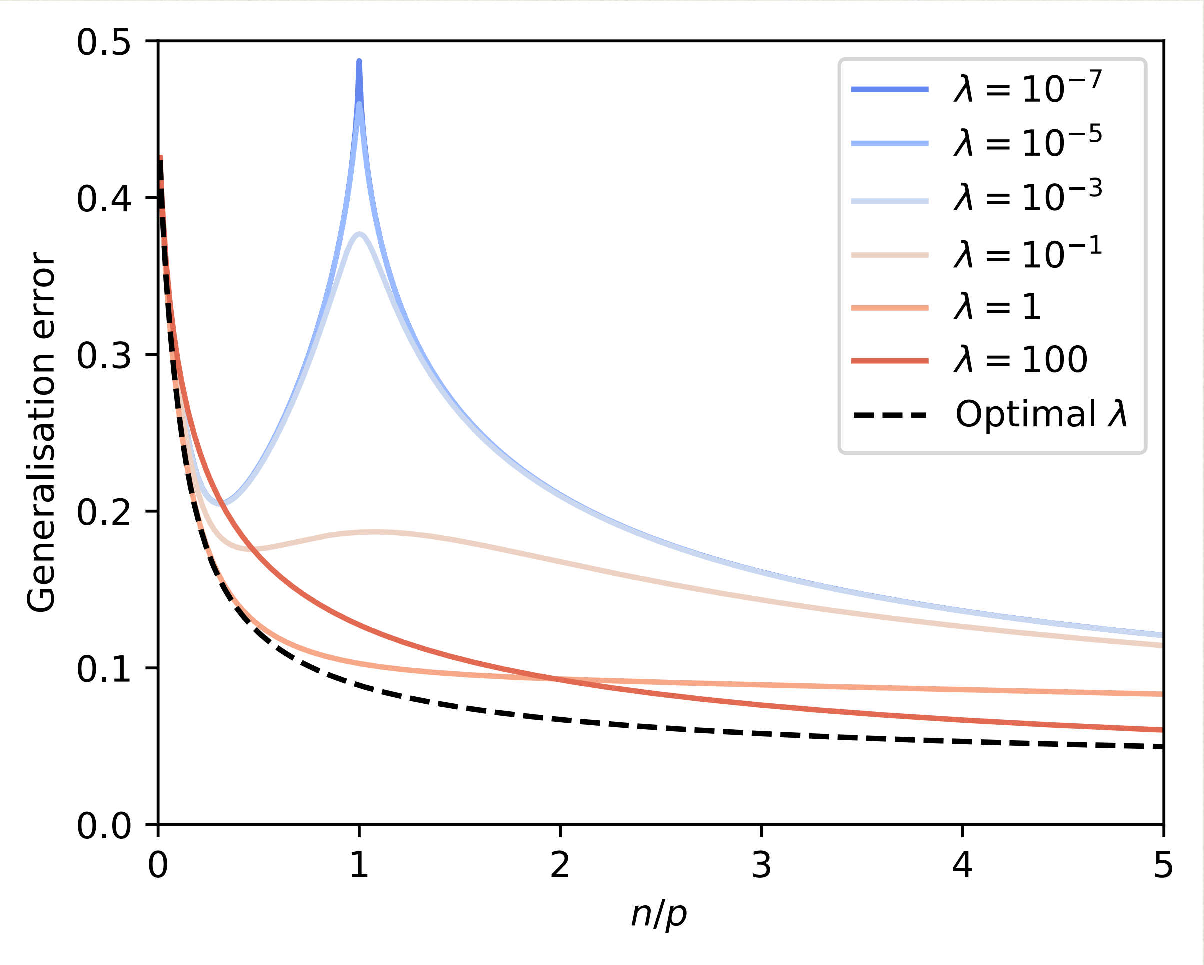
$$\mathcal{L}_{\text{training}} = \frac{\lambda}{2\alpha}q_w^\star + \mathbb{E}_{\xi,y}\left[\mathcal{Z}\left(y,\omega_0^\star\right)\ell\left(y,\eta(y,\omega_1^\star)\right)\right]$$

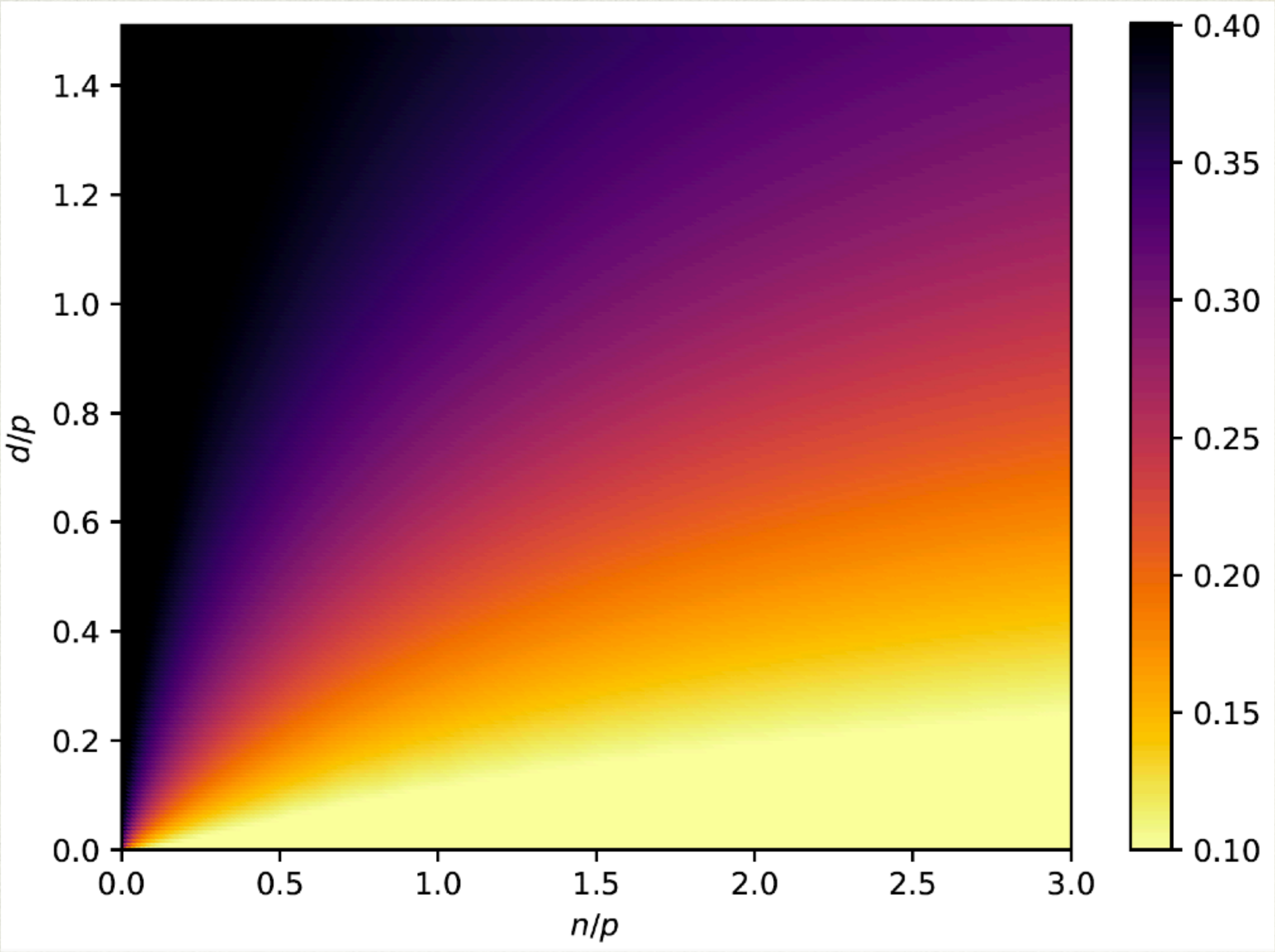
with $\omega_0^\star = M^\star/\sqrt{Q^\star}\xi$, $\omega_1^\star = \sqrt{Q^\star}\xi$

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

PHASE DIAGRAM

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





d/p=0.1

RANDOM FEATURES

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

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

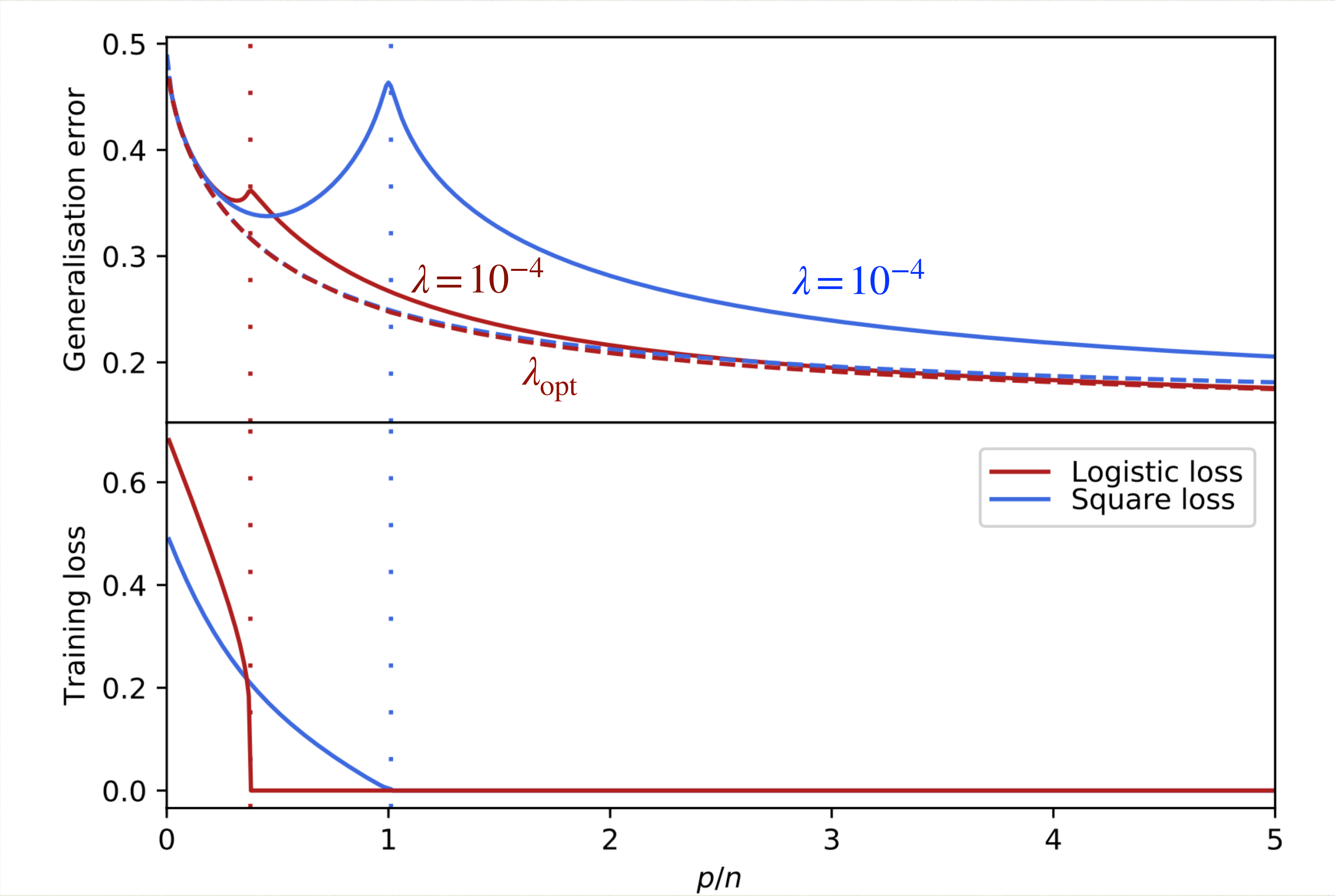
$C_\mu \in \mathbb{R}^d$	input data
$F \in \mathbb{R}^{p\times d}$	features
$X_\mu \in \mathbb{R}^p$	projections

n = # samples, d= input dimension, p = # features (width)

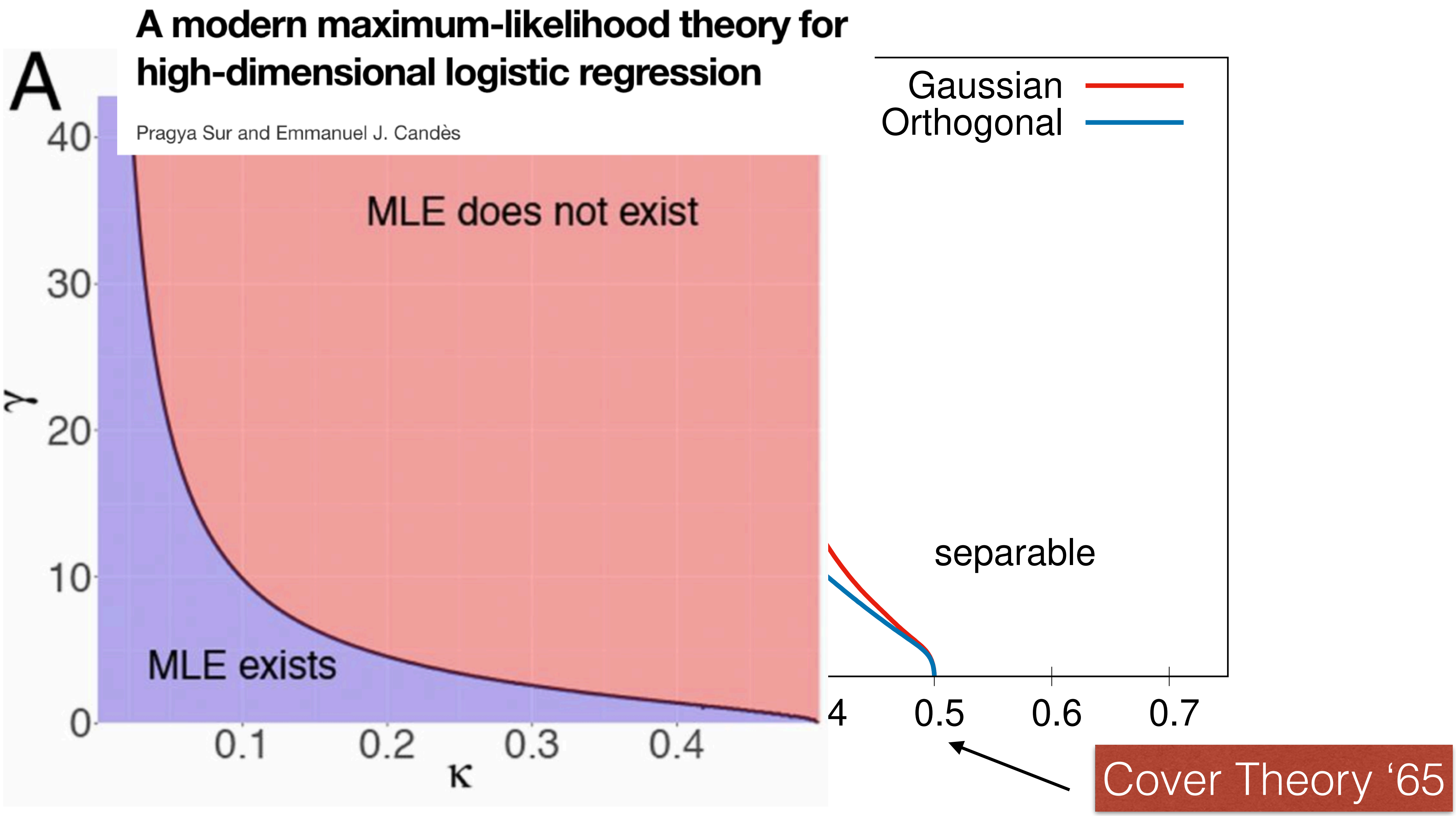
Over-parametrization

n/d=3

logistic loss, square loss



Phase transition of perfect separability

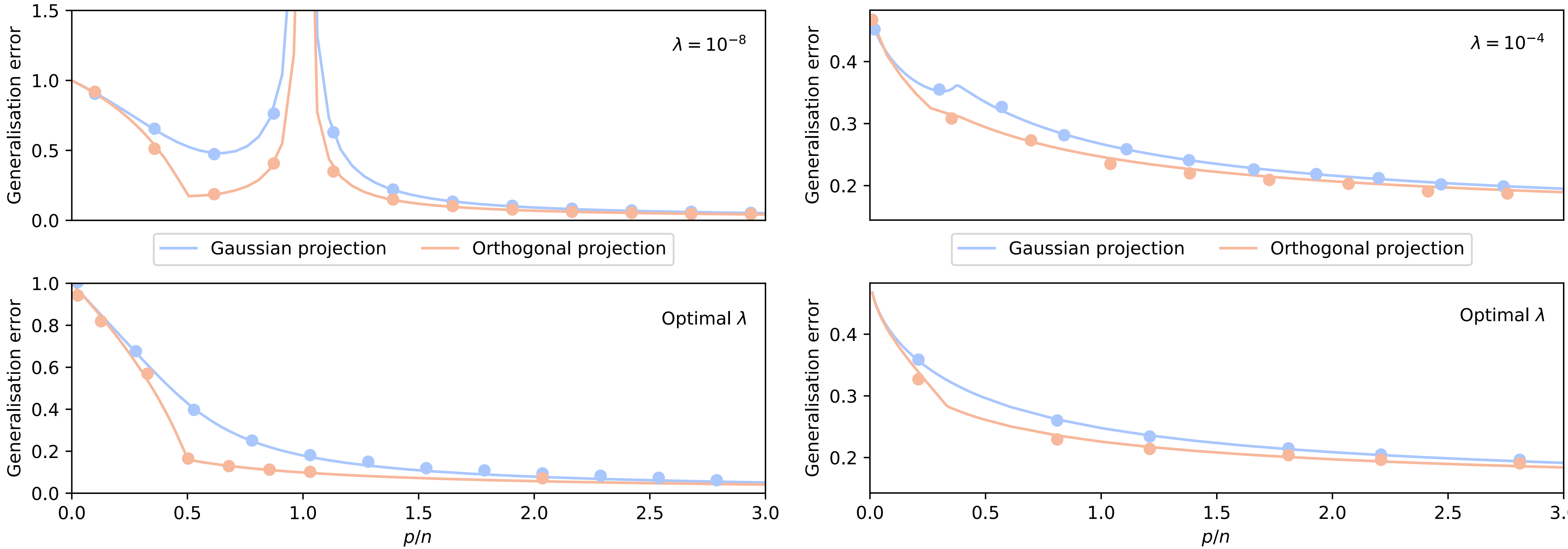


Generalizes the storage capacity phase transition
[Cover '65; Gardner '87; Sur & Candes, '18]

Asymptotics accurate even at d=200!

Regression task — ℓ_2 loss

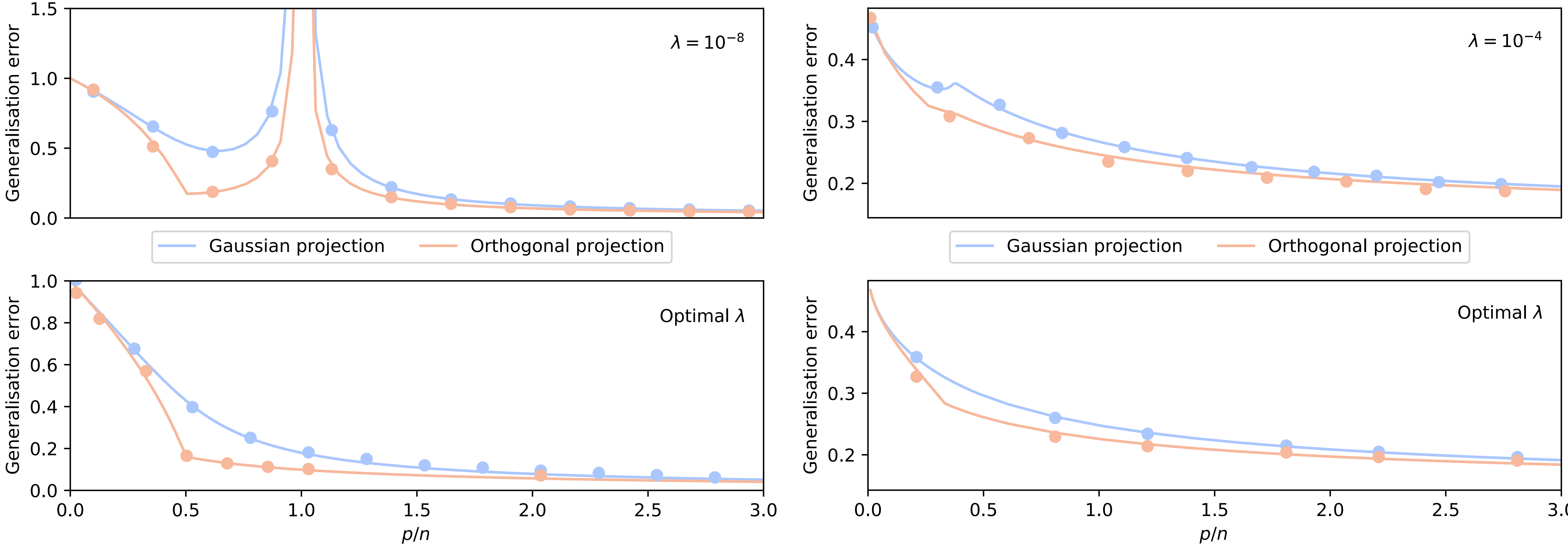
Classification task — logistic loss



First layer: random Gaussian Matrix

First layer: subsampled Fourier matrix

Gaussian v.s. orthogonal features



The Unreasonable Effectiveness of Structured Random Orthogonal Embeddings

Nips '17

Krzysztof Choromanski *
Google Brain Robotics
kchoro@google.com

Mark Rowland *
University of Cambridge
mr504@cam.ac.uk

Adrian Weller
University of Cambridge and Alan Turing Institute
aw665@cam.ac.uk

Does the analysis work when the generative model is deep?

DEEP GENERATIVE MODELS

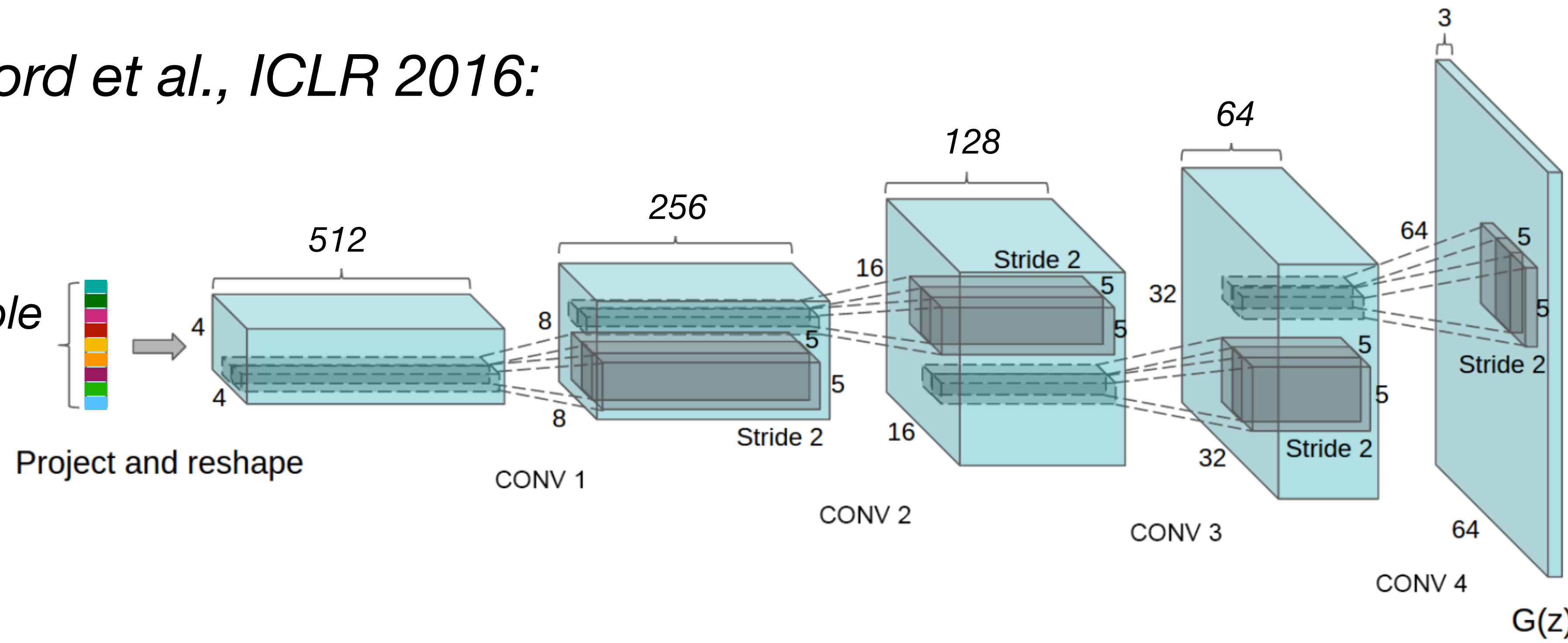
Goldt, FK, Mézard, Reeves LZ; arXiv:2006.14709

- Data model: Inputs generated by multi-layer neural networks. Teacher acting on the latent space.
- Result: Closed-formula for online SGD on one (small) hidden layer neural networks. Generalization of Saad, Sola'95 ODEs.
- Theoretically justified for random and independent weight matrices, works great even for learned generators.

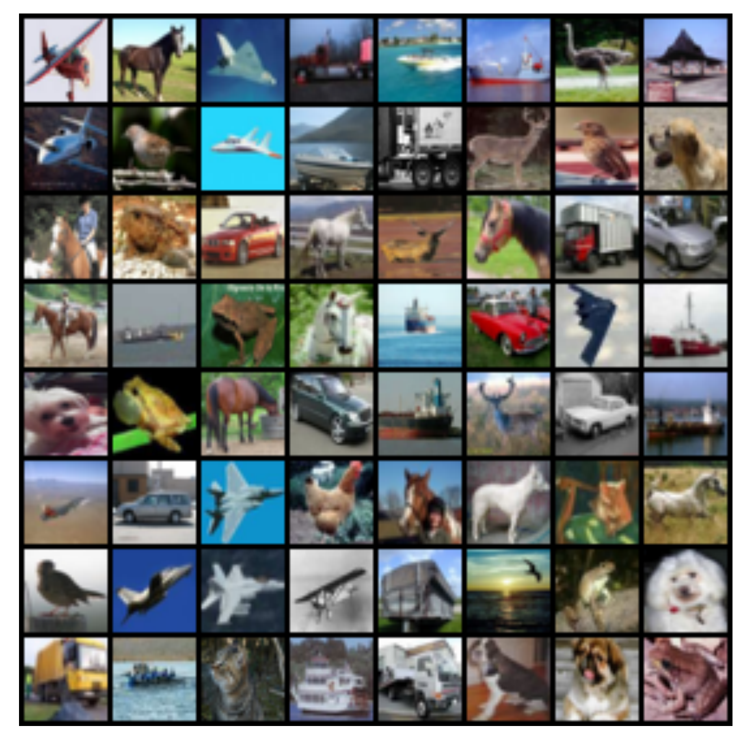
Deep convolutional GAN with random weights

Radford et al., ICLR 2016:

Latent variable
$c \in \mathbb{R}^D$



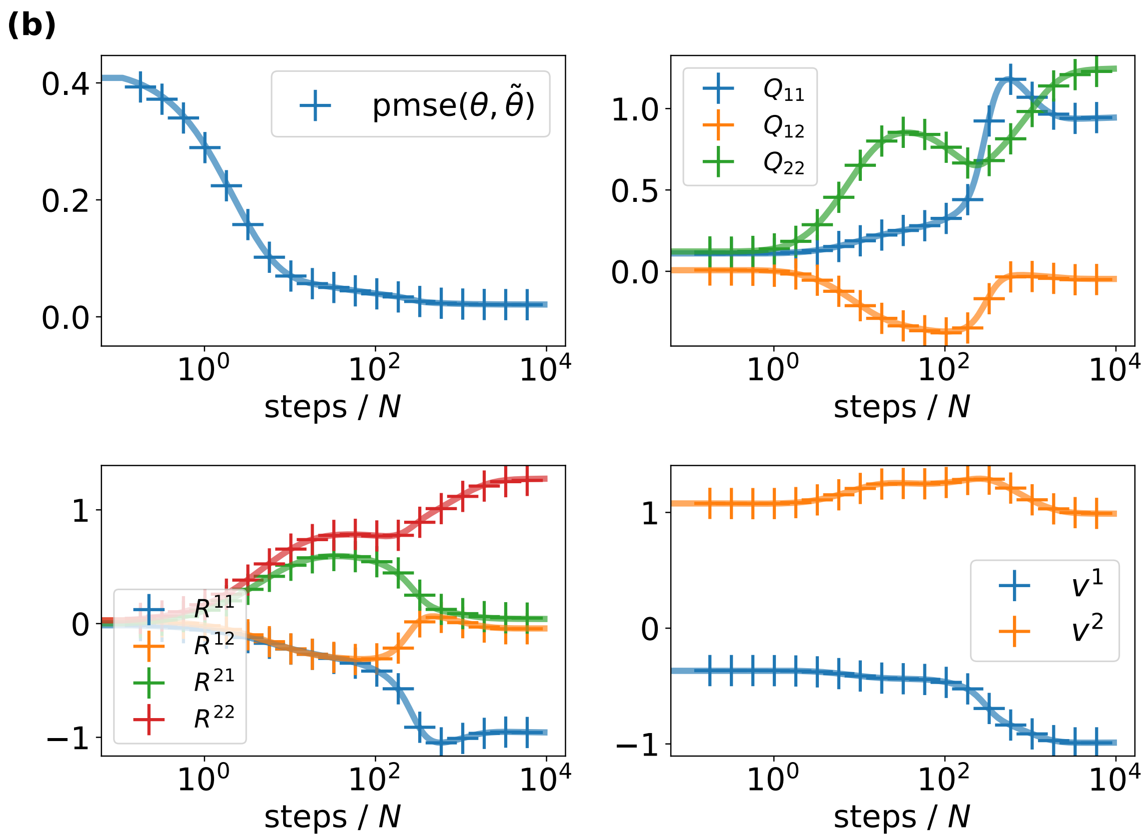
- Just five layers of 2D convolutions
- No pooling layer, no fully-connected layers
- ReLU activation after each layer, Tanh at the end



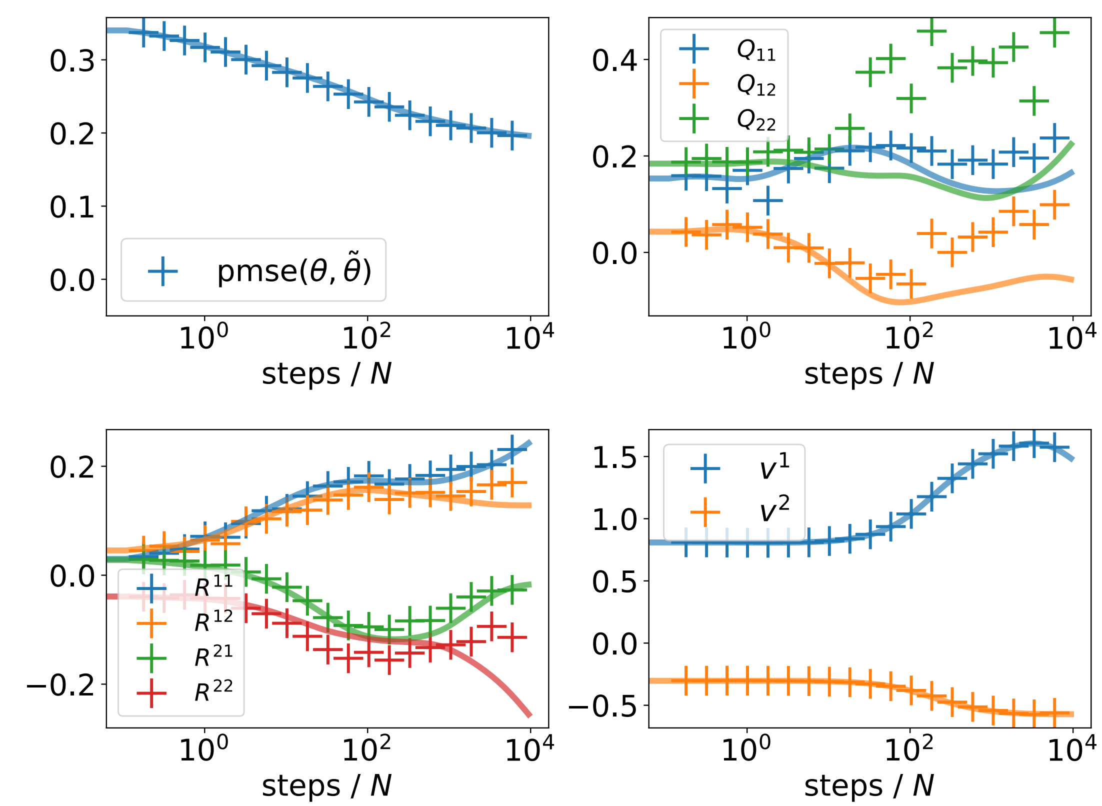
Images generated by a DCGAN trained on CIFAR10

Deep convolutional GAN: ODE vs simulation

Random weights



Pre-trained weights (CIFAR10)



Both experiments: $g(x) = \mathrm{erf}(x/\sqrt{2})$, *M=K=2,* $\eta = 0.2$, *D=100, N=3072*

- Great agreement for random weights.
- Reasonable agreement for learned weights.

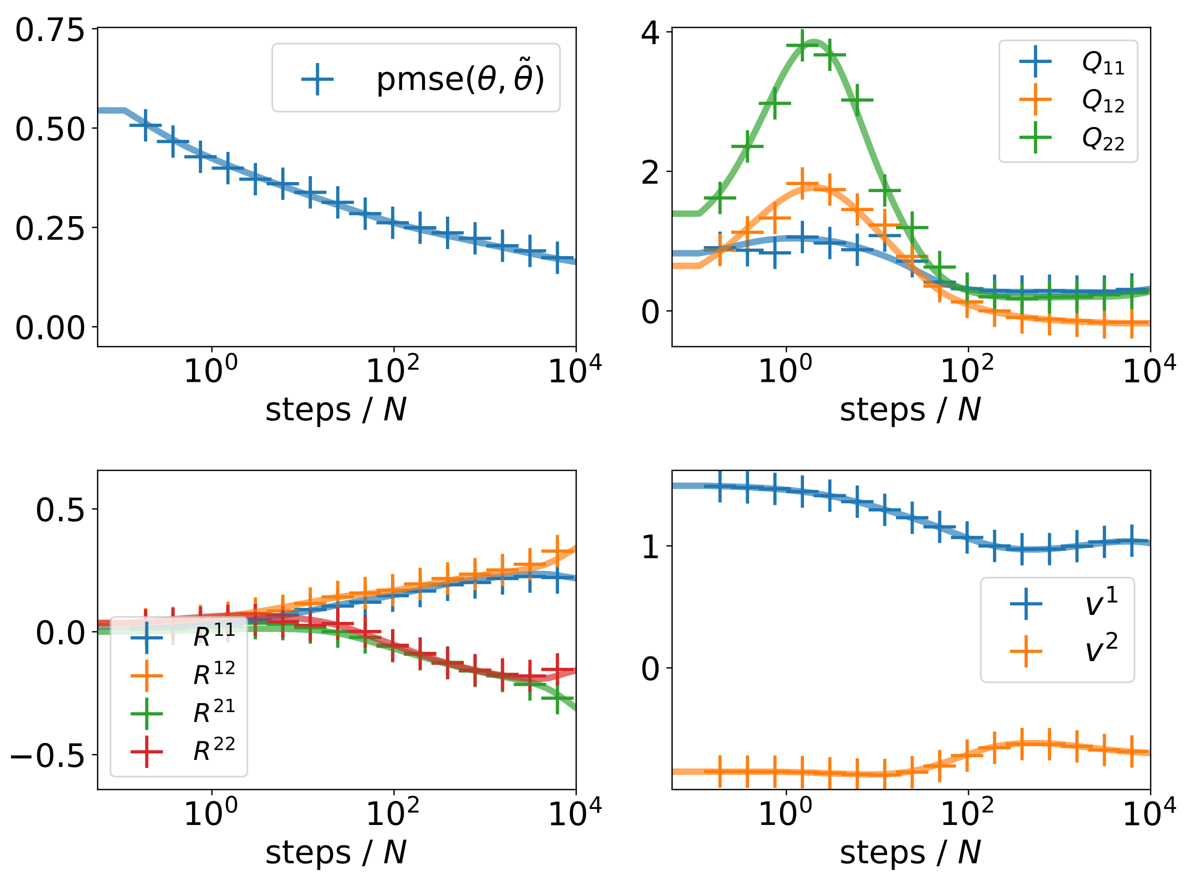
The realNVP: normalising flows

Dinh, Sohl-Dickstein, Bengio (ICLR 2017)

- Normalising flows generate inputs using a series of invertible transformations.

- Very good agreement between ODE and simulation for a pre-trained realNVP

Top half: CIFAR10 images
Bottom half: Samples from realNVP trained on CIFAR10



M=K=2, $\eta = 0.2$, D=3072, N=3072

THANK YOU!