LETTRES SCIENCES
ÉCOLE NORMALE SUPÉRIEURE

IUF

CFM
INSIGHT.DATA.CLARITY.

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

STATISTICAL PHYSICS OF NEURAL NETS

Florent Krzakala

The Kavli Institute for Theoretical Physics
University of California, Santa Barbara

At the Crossroad of Physics & Machine Learning

ÉCOLE NORMALE SUPÉRIEURE
IUF
CFM INSIGHT.DATA.CLARITY.
SORBONNE UNIVERSITÉ CRÉATEURS DE FUTURS DEPUIS 1257

STATISTICAL PHYSICS OF NEURAL NETS

OLD IDEAS FOR NEW PROBLEMS

Florent Krzakala

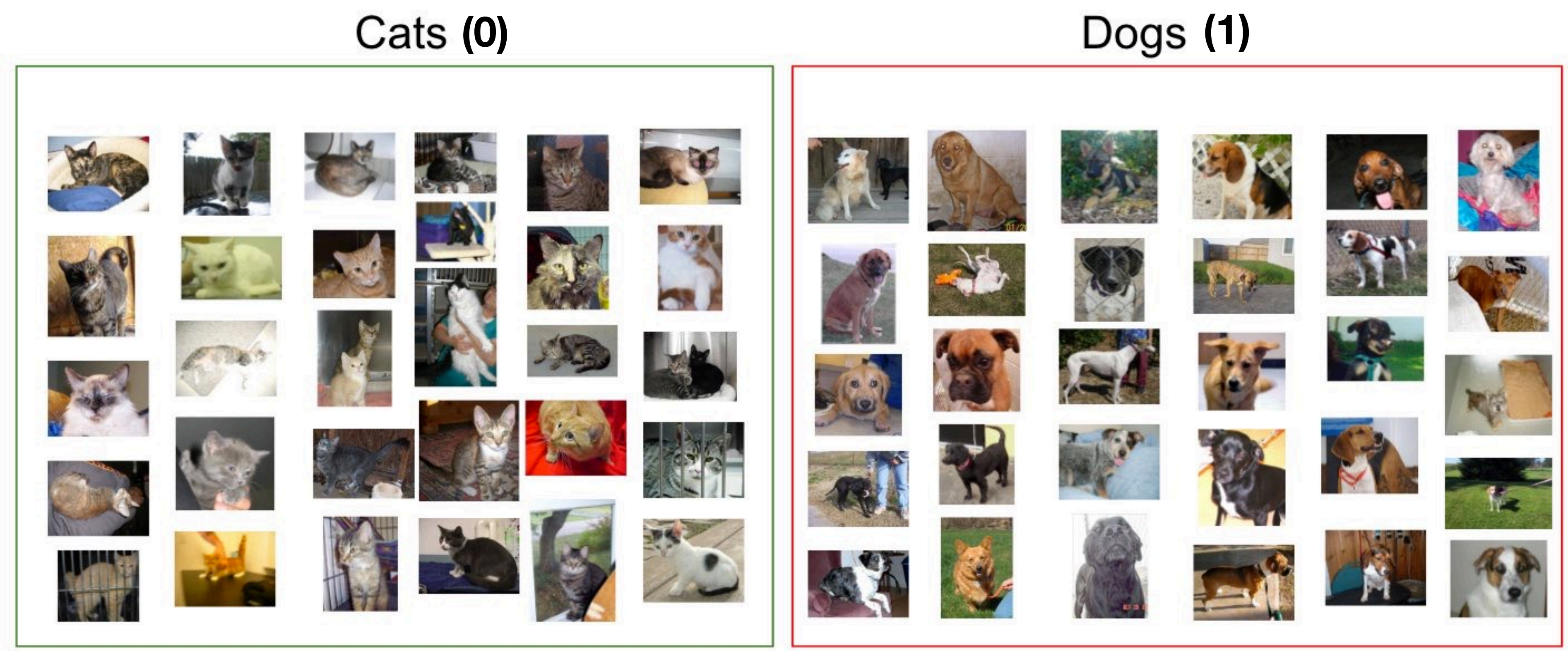
The Kavli Institute for Theoretical Physics
University of California, Santa Barbara

At the Crossroad of Physics
& Machine Learning

Supervised learning

<u>Training</u>: Find a function $f(\,.\,) \in \mathscr{F}$ **that predicts the right class** $y_i = f(\mathbf{x}_i)$ **in the dataset**

The fraction of mistakes is called the *<u>training error</u>*

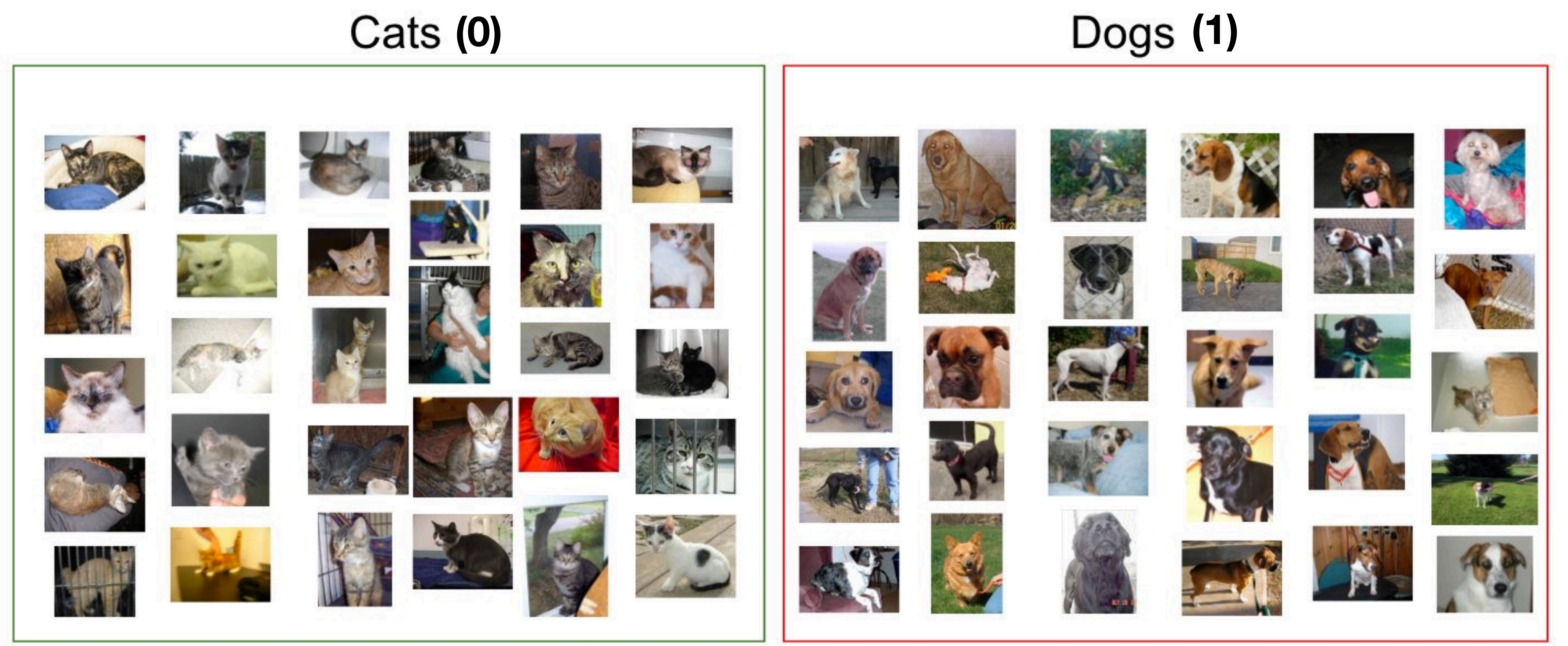


Sample of cats & dogs images from Kaggle Dataset

Supervised learning

Training: Find a function $f(\,.\,) \in \mathscr{F}$ that predicts the right class $y_i = f(\mathbf{x}_i)$ in the dataset

The fraction of mistakes is called the *training error*



Sample of cats & dogs images from Kaggle Dataset

Generalization: See how the function performs on new, unseen, images

$f(\mathbf{x}_{\text{new}})$ **?**

Supervised learning

<u>**Training**</u>**: Find a function** $f(\,.\,) \in \mathscr{F}$ **that predicts the right class** $y_i = f(\mathbf{x}_i)$ **in the dataset**

The fraction of mistakes is called the *training error*

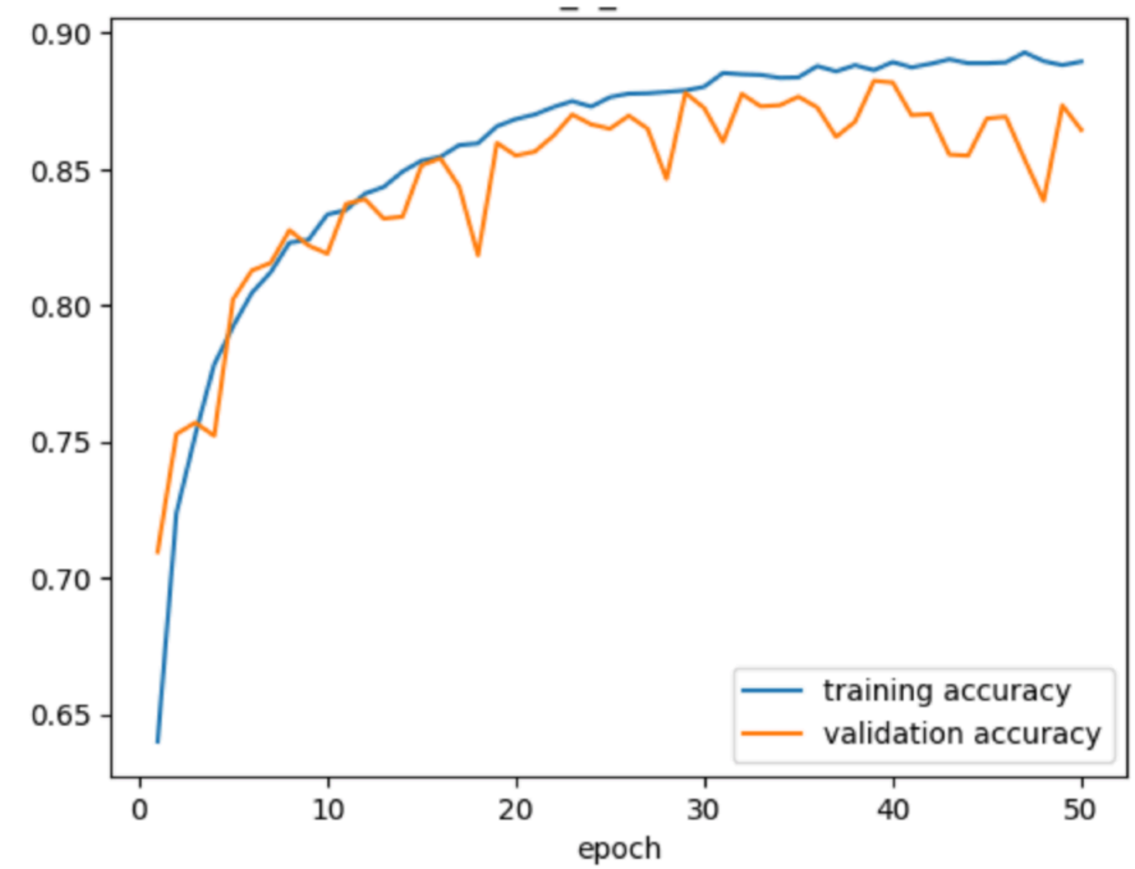
Cats **(0)** Dogs **(1)**

Sample of cats & dogs images from Kaggle Dataset

<u>**Generalization**</u>**: See how the function performs on new, unseen, images**

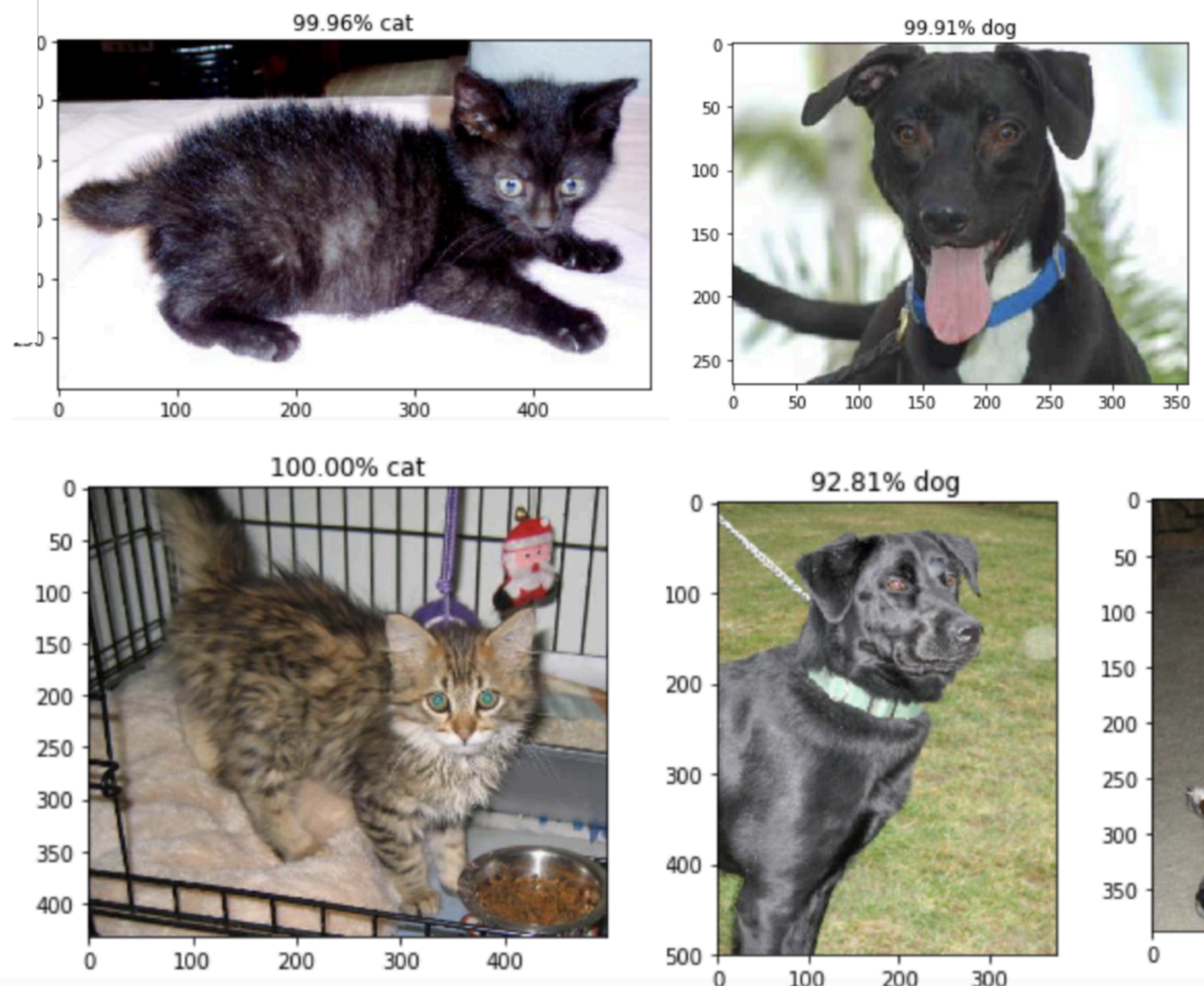
$f(\mathbf{x}_{\text{new}})$ **?**

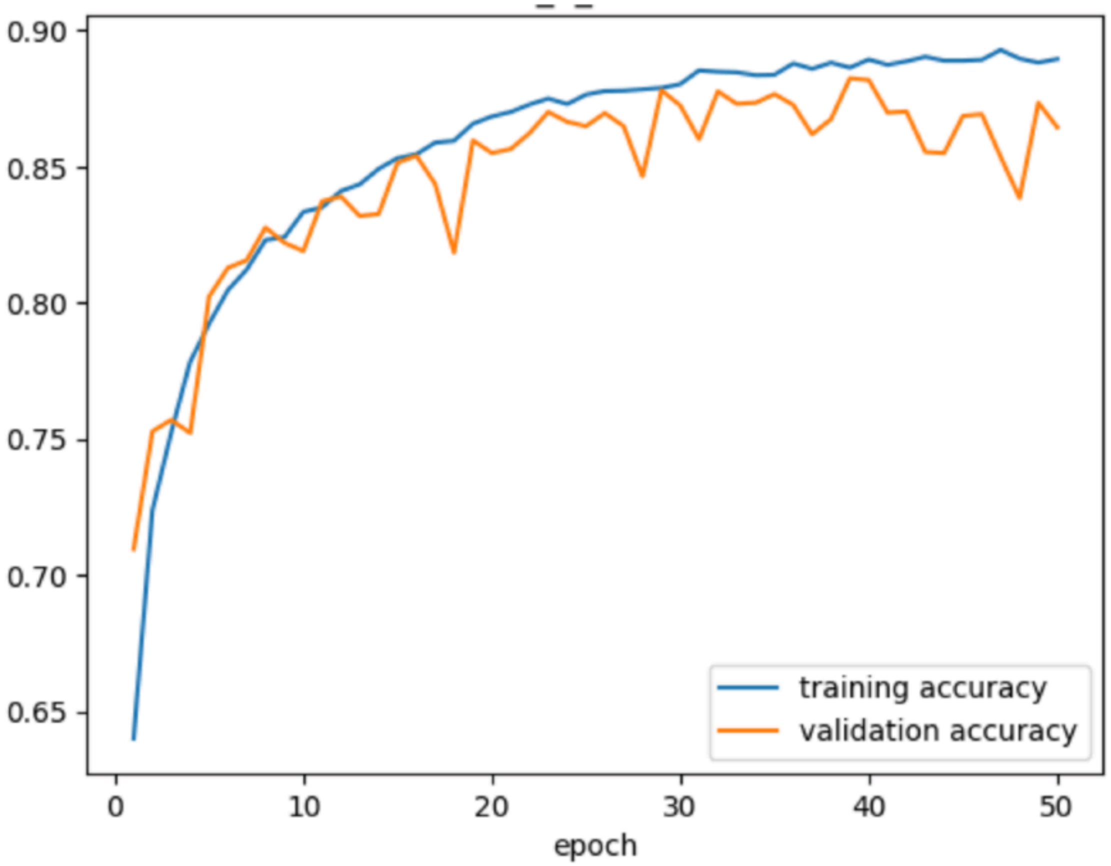
The fraction of mistakes on new images is called the *generalisation error*



Credit: https://towardsdatascience.com/

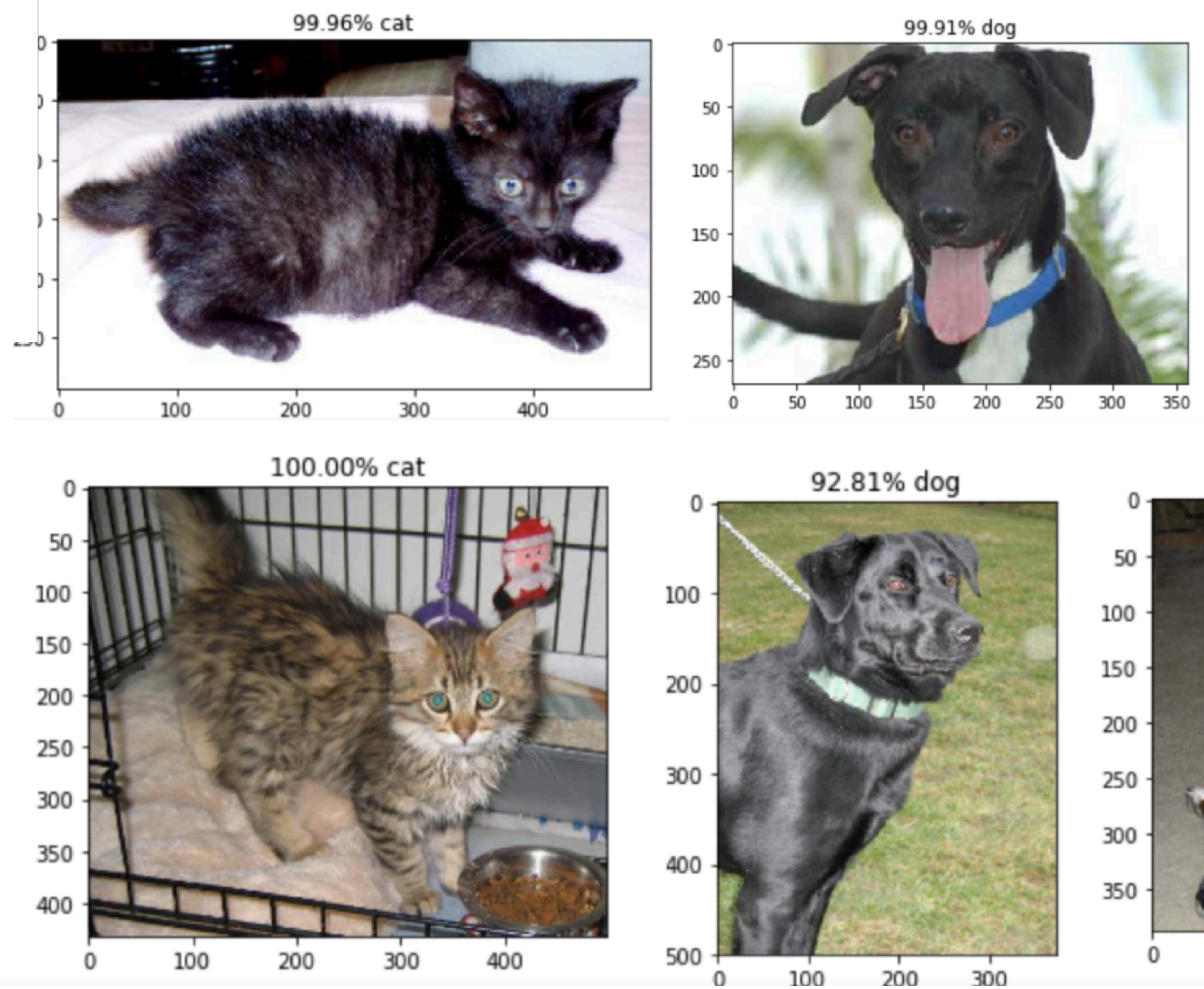
It works!





Credit: https://towardsdatascience.com/

It works!



Do we understand this?

The generalization crisis

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang[*]
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht[†]
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com

ABSTRACT

Despite their massive size, successful deep artificial neural networks can exhibit a remarkably small difference between training and test performance. Conventional wisdom attributes small generalization error either to properties of the model family, or to the regularization techniques used during training.

Through extensive systematic experiments, we show how these traditional approaches fail to explain why large neural networks generalize well in practice. Specifically, our experiments establish that state-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data. This phenomenon is qualitatively unaffected by explicit regularization, and occurs even if we replace the true images by completely unstructured random noise. We corroborate these experimental findings with a theoretical construction showing that simple depth two neural networks already have perfect finite sample expressivity as soon as the number of parameters exceeds the number of data points as it usually does in practice.

We interpret our experimental findings by comparison with traditional models.

The generalization crisis

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang[*]
Massachusetts Institute (
chiyuan@mit.edu

Benjamin Recht[†]
University of California,
brecht@berkeley.edu

Despite their mas
remarkably smal
wisdom attribute
ily, or to the regu
Through extensi
proaches fail to
Specifically, our
for image classif
dom labeling of
by explicit regul
pletely unstructu
with a theoretical
ready have perfe
exceeds the numl
We interpret our

To Understand Deep Learning We Need to Understand Kernel Learning

Mikhail Belkin, Siyuan Ma, Soumik Mandal
Department of Computer Science and Engineering
Ohio State University
{mbelkin, masi}@cse.ohio-state.edu, *mandal.32@osu.edu*

Abstract

Generalization performance of classifiers in deep learning has recently become a subject of intense study. Deep models, which are typically heavily over-parametrized, tend to fit the training data exactly. Despite this “overfitting", they perform well on test data, a phenomenon not yet fully understood.

The first point of our paper is that strong performance of overfitted classifiers is not a unique feature of deep learning. Using six real-world and two synthetic datasets, we establish experimentally that kernel machines trained to have zero classification error or near zero regression error (interpolation) perform very well on test data, even when the labels are corrupted with a high level of noise. We proceed to give a lower bound on the norm of zero loss solutions for smooth kernels, showing that they increase nearly exponentially with data size. We point out that this is difficult to reconcile with the existing generalization bounds.

The generalization crisis

UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION

Chiyuan Zhang*
Massachusetts Institute
chiyuan@mit.edu

Benjamin Recht†
University of California,
brecht@berkeley.edu

To Understand Deep Learning We Need to Understand Kernel Learning

Mikhail Belkin, Siyuan Ma, Soumik Mandal

Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior

Charles H. Martin* Michael W. Mahoney†

Abstract

We describe an approach to understand the peculiar and counterintuitive generalization properties of deep neural networks. The approach involves going beyond worst-case theoretical capacity control frameworks that have been popular in machine learning in recent years to revisit old ideas in the statistical mechanics of neural networks. Within this approach, we present a prototypical Very Simple Deep Learning (VSDL) model, whose behavior is controlled by two control parameters, one describing an effective amount of data, or load, on the

Rademacher and VC bounds

Given a space Z and a fixed distribution $D|_Z$, let $S = \{z_1, \ldots, z_m\}$ be a set of examples drawn i.i.d. from $D|_Z$. Furthermore, let $\mathcal{F}$ be a class of functions $f : Z \to \mathbb{R}$.

Definition. The *empirical Rademacher complexity* of $\mathcal{F}$ is defined to be

$$\hat{R}_m(\mathcal{F}) \;=\; \mathsf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right) \right]$$

where $\sigma_1, \ldots, \sigma_m$ are independent random variables uniformly chosen from $\{-1, 1\}$. We will refer to such random variables as *Rademacher variables*.

Definition. The *Rademacher complexity* of $\mathcal{F}$ is defined as

$$R_m(\mathcal{F}) \;=\; \mathsf{E}_D[\hat{R}_m(\mathcal{F})]$$

Theorem 2. *Fix distribution $D|_Z$ and parameter $\delta \in (0,1)$. If $\mathcal{F} \subseteq \{f : Z \to [a, a+1]\}$ and $S = \{z_1, \ldots, z_n\}$ is drawn i.i.d. from $D|_Z$ then with probability $\geq 1 - \delta$ over the draw of S, for every function $f \in \mathcal{F}$,*

$$\mathsf{E}_D[f(z)] \;\leq\; \hat{\mathsf{E}}_S[f(z)] + 2R_m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{m}}. \tag{1}$$

In addition, with probability $\geq 1 - \delta$, for every function $f \in \mathcal{F}$,

$$\mathsf{E}_D[f(z)] \;\leq\; \hat{\mathsf{E}}_S[f(z)] + 2\hat{R}_m(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{m}}. \tag{2}$$

By Sauer's Lemma, $\mathcal{H}[m] \leq m^d$ where d is the VC dimension of $\mathcal{H}$, so we can further simplify this result to

$$\hat{R}_m(\mathcal{H}) \leq \sqrt{\frac{2d \ln m}{m}}.$$

Credit: Machine Learning Theory, lecture notes @CMU

A back-of-the-envelope bound

$\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$

$f(\,.\,) \in \mathcal{F}$

$$\epsilon_{\text{generalization}} = \epsilon_{\text{training}} + \Delta$$

A back-of-the-envelope bound

$\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$

$f(\,.\,) \in \mathcal{F}$

$$\epsilon_{\text{generalization}} = \epsilon_{\text{training}} + \Delta$$

What is the worst that can happen ?

A back-of-the-envelope bound

$\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$

$f(\,.\,) \in \mathcal{F}$

$$\epsilon_{\text{generalization}} = \epsilon_{\text{training}} + \Delta$$

What is the worst that can happen ?

We are looking for a rule while there *is no rule* and the labels are actually random!

A back-of-the-envelope bound

$\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$

$f(\,.\,) \in \mathcal{F}$

$$\epsilon_{\text{generalization}} = \epsilon_{\text{training}} + \Delta$$

What is the worst that can happen ?

We are looking for a rule while there *is no rule* and the labels are actually random!

$\epsilon_{\text{training}}^{\text{random}}$ **can be optimised…**

A back-of-the-envelope bound

$\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$

$f(\,.\,) \in \mathcal{F}$

$$\epsilon_{\text{generalization}} = \epsilon_{\text{training}} + \Delta$$

What is the worst that can happen ?

We are looking for a rule while there *is no rule* and the labels are actually random!

$\epsilon_{\text{training}}^{\text{random}}$ **can be optimised…** **… but** $\epsilon_{\text{generalization}}^{\text{random}} = \frac{1}{2}$

A back-of-the-envelope bound

$\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$

$f(\,.\,) \in \mathscr{F}$

$$\epsilon_{\text{generalization}} = \epsilon_{\text{training}} + \Delta$$

What is the worst that can happen ?

We are looking for a rule while there *__is no rule__* and the labels are actually random!

$\epsilon_{\text{training}}^{\text{random}}$ **can be optimised…** **… but** $\epsilon_{\text{generalization}}^{\text{random}} = \dfrac{1}{2}$

So, in reality, we expect:

A back-of-the-envelope bound

$\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$

$f(\,.\,) \in \mathcal{F}$

$$\epsilon_{\text{generalization}} = \epsilon_{\text{training}} + \Delta$$

What is the worst that can happen ?

We are looking for a rule while there *<u>is no rule</u>* and the labels are actually random!

$\epsilon^{\text{random}}_{\text{training}}$ **can be optimised…** **… but** $\epsilon^{\text{random}}_{\text{generalization}} = \frac{1}{2}$

So, in reality, we expect:

$$\epsilon_{\text{generalization}} - \epsilon_{\text{training}} \leq \epsilon^{\text{random}}_{\text{generalization}} - \epsilon^{\text{random}}_{\text{training}}$$

A back-of-the-envelope bound

$\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$

$f(\,.\,) \in \mathcal{F}$

$$\epsilon_{\text{generalization}} = \epsilon_{\text{training}} + \Delta$$

What is the worst that can happen ?

We are looking for a rule while there *is no rule* and the labels are actually random!

$\epsilon^{\text{random}}_{\text{training}}$ **can be optimised…** **… but** $\epsilon^{\text{random}}_{\text{generalization}} = \frac{1}{2}$

So, in reality, we expect:

$$\epsilon_{\text{generalization}} - \epsilon_{\text{training}} \leq \epsilon^{\text{random}}_{\text{generalization}} - \epsilon^{\text{random}}_{\text{training}} = \frac{1}{2}(1 - 2\epsilon^{\text{random}}_{\text{training}})$$

A back-of-the-envelope bound

$\{(\mathbf{x}_i, y_i)\}_{i=1,\dots,n}$

$f(\,.\,) \in \mathcal{F}$

$$\epsilon_{\text{generalization}} = \epsilon_{\text{training}} + \Delta$$

What is the worst that can happen ?

We are looking for a rule while there *__is no rule__* and the labels are actually random!

$\epsilon_{\text{training}}^{\text{random}}$ **can be optimised…** **… but** $\epsilon_{\text{generalization}}^{\text{random}} = \frac{1}{2}$

So, in reality, we expect:

$$\epsilon_{\text{generalization}} - \epsilon_{\text{training}} \leq \epsilon_{\text{generalization}}^{\text{random}} - \epsilon_{\text{training}}^{\text{random}} = \frac{1}{2}(1 - 2\epsilon_{\text{training}}^{\text{random}}) = \frac{1}{2}\hat{\mathcal{R}}_n\{(\mathbf{x})\}$$

A back-of-the-envelope bound

$\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$
$f(\,.\,) \in \mathcal{F}$

$$\epsilon_{\text{generalization}} = \epsilon_{\text{training}} + \Delta$$

What is the worst that can happen ?

We are looking for a rule while there *__is no rule__* and the labels are actually random!

$\epsilon^{\text{random}}_{\text{training}}$ **can be optimised…** **… but** $\epsilon^{\text{random}}_{\text{generalization}} = \dfrac{1}{2}$

So, in reality, we expect:

$$\epsilon_{\text{generalization}} - \epsilon_{\text{training}} \leq \epsilon^{\text{random}}_{\text{generalization}} - \epsilon^{\text{random}}_{\text{training}} = \frac{1}{2}(1 - 2\epsilon^{\text{random}}_{\text{training}}) = \frac{1}{2}\hat{\mathcal{R}}_n\{(\mathbf{x})\}$$

$\mathcal{R}_n(\{\mathbf{x}\})$**is the empirical Rademacher complexity:**
it tells how well your method can fit random labels

A back-of-the-envelope bound

$\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$
$f(\,.\,) \in \mathcal{F}$

$$\epsilon_{\text{generalization}} = \epsilon_{\text{training}} + \Delta$$

What is the worst that can happen ?

We are looking for a rule while there *<u>is no rule</u>* and the labels are actually random!

$\epsilon^{\text{random}}_{\text{training}}$ **can be optimised…** **… but** $\epsilon^{\text{random}}_{\text{generalization}} = \frac{1}{2}$

So, in reality, we expect:

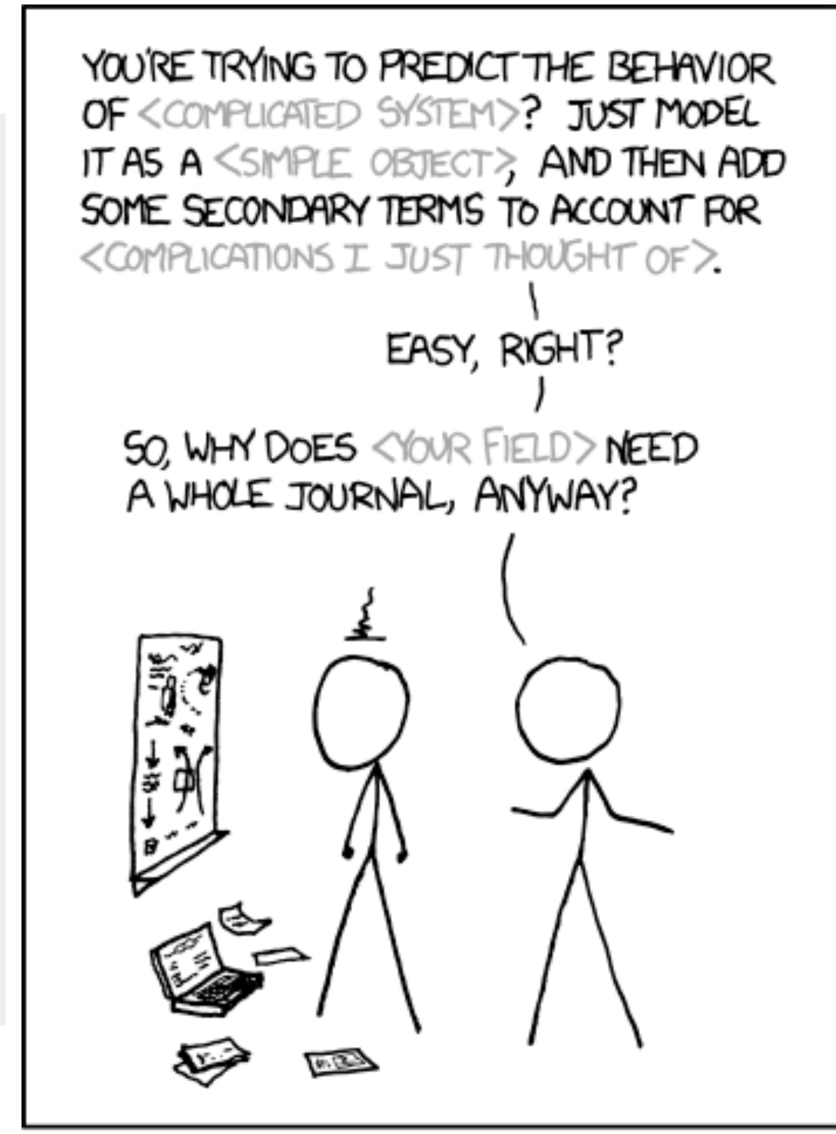
$$\epsilon_{\text{generalization}} - \epsilon_{\text{training}} \leq \epsilon^{\text{random}}_{\text{generalization}} - \epsilon^{\text{random}}_{\text{training}} = \frac{1}{2}(1 - 2\epsilon^{\text{random}}_{\text{training}}) = \frac{1}{2}\hat{\mathcal{R}}_n\{(\mathbf{x})\}$$

$\mathcal{R}_n(\{\mathbf{x}\})$ **is the empirical Rademacher complexity:**
it tells how well your method can fit random labels

$\epsilon_{\text{generalization}} - \epsilon_{\text{training}} \leq \frac{1}{2}\hat{\mathcal{R}}_n\{(\mathbf{x})\}$ **is a very <u>pessimistic</u> scenario!**

Physicists like Models

Spherical cow in vacuum



LIBERAL-ARTS MAJORS MAY BE ANNOYING SOMETIMES, BUT THERE'S *NOTHING* MORE OBNOXIOUS THAN A PHYSICIST FIRST ENCOUNTERING A NEW SUBJECT.

credit: XKCD

Physicists do not like worst case analysis, and instead study models of data

1

The Teacher-Student scenario

P. Carnevali & S. Patarnello (1987)
N. Tishby, E. Levin, & S. Solla (1989)
E. Gardner, B. Derrida (1989)
….

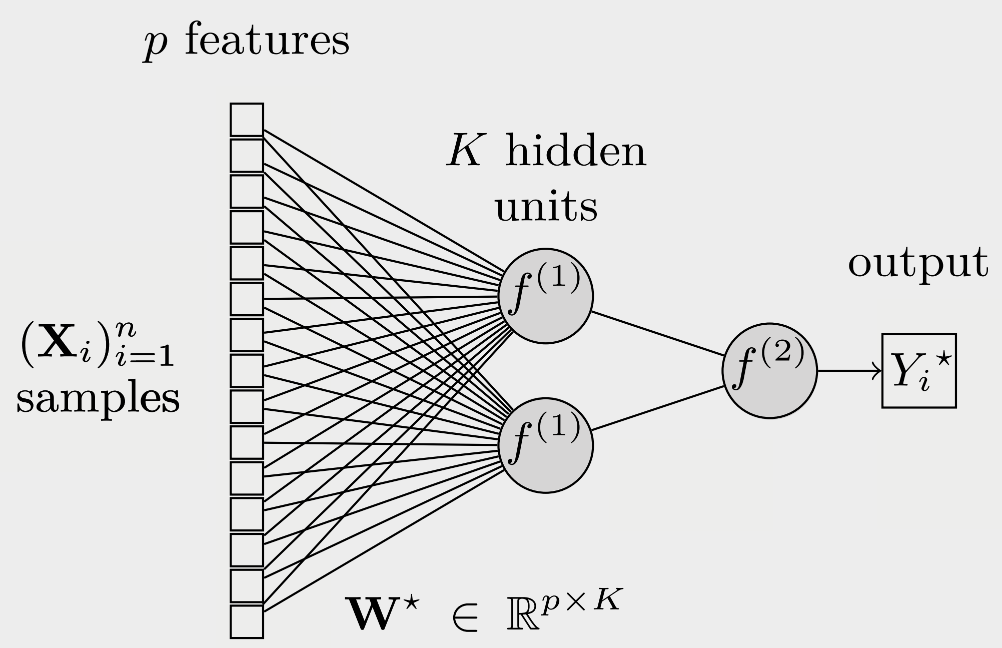
The Teacher-Student scenario: A tale of two networks

The Teacher-Student scenario: A tale of two networks

- **Teacher:**

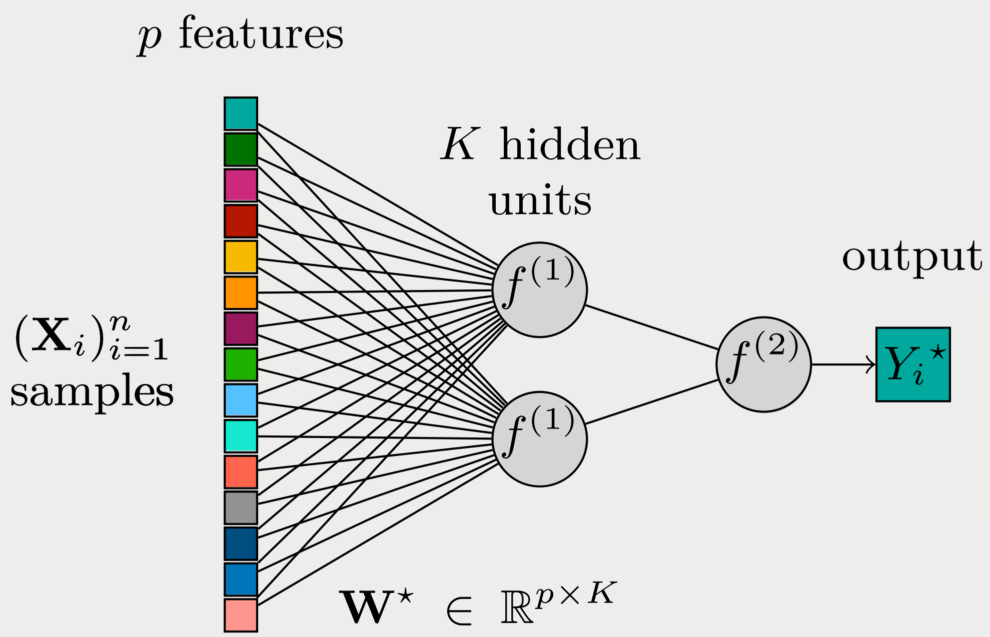
The Teacher-Student scenario: A tale of two networks

- **Teacher:**



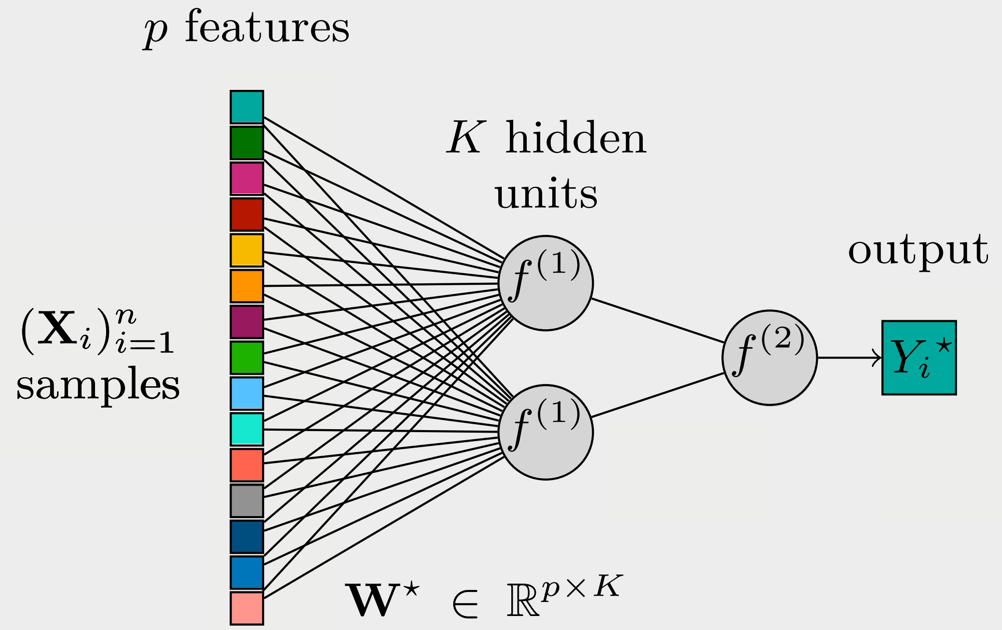
The Teacher-Student scenario: A tale of two networks

- **Teacher:**

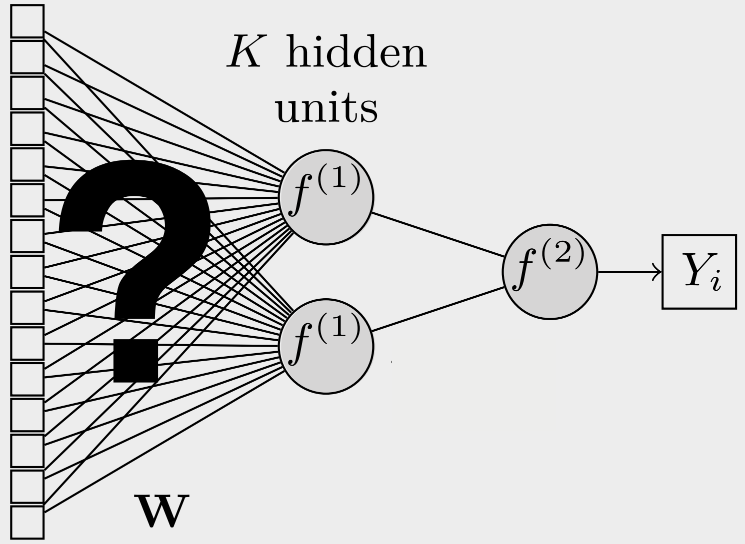


The Teacher-Student scenario: A tale of two networks

- **Teacher:**

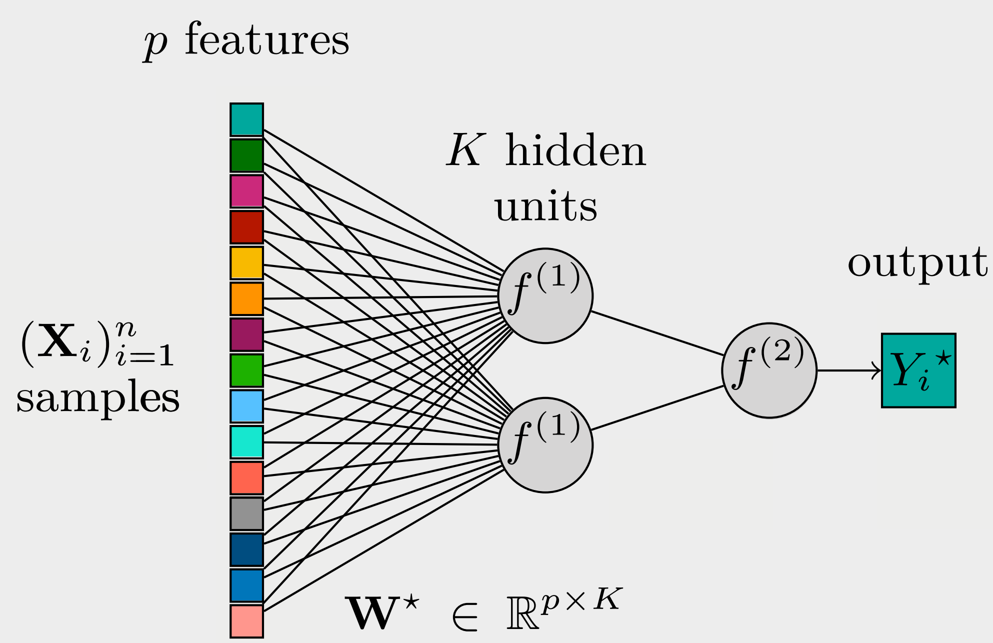


- **Student:**

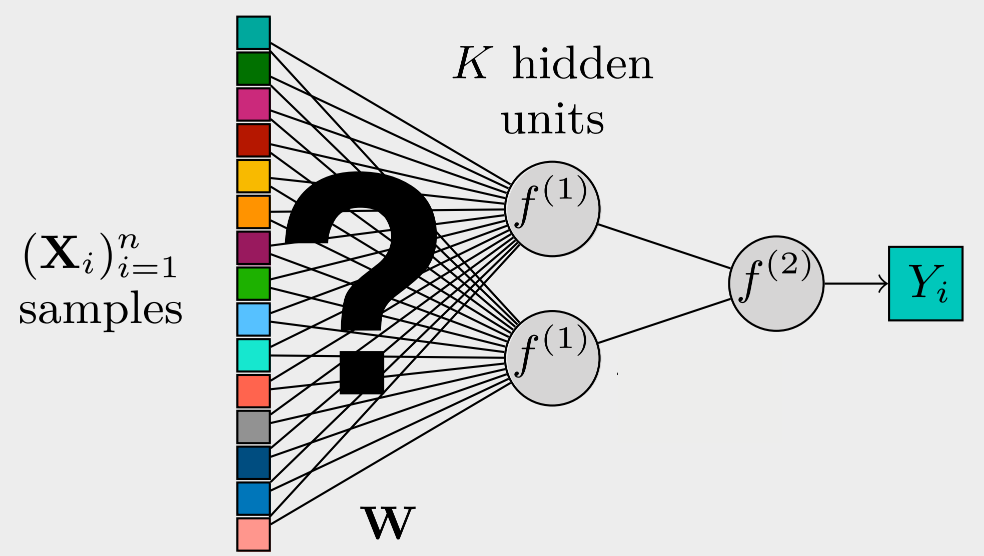


The Teacher-Student scenario: A tale of two networks

- **Teacher:**

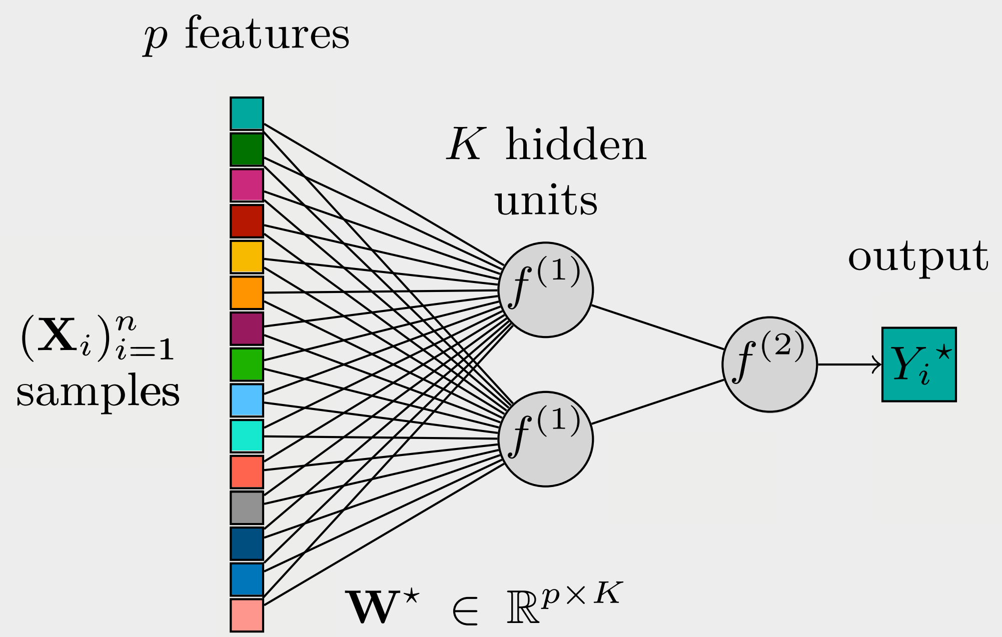


- **Student:**

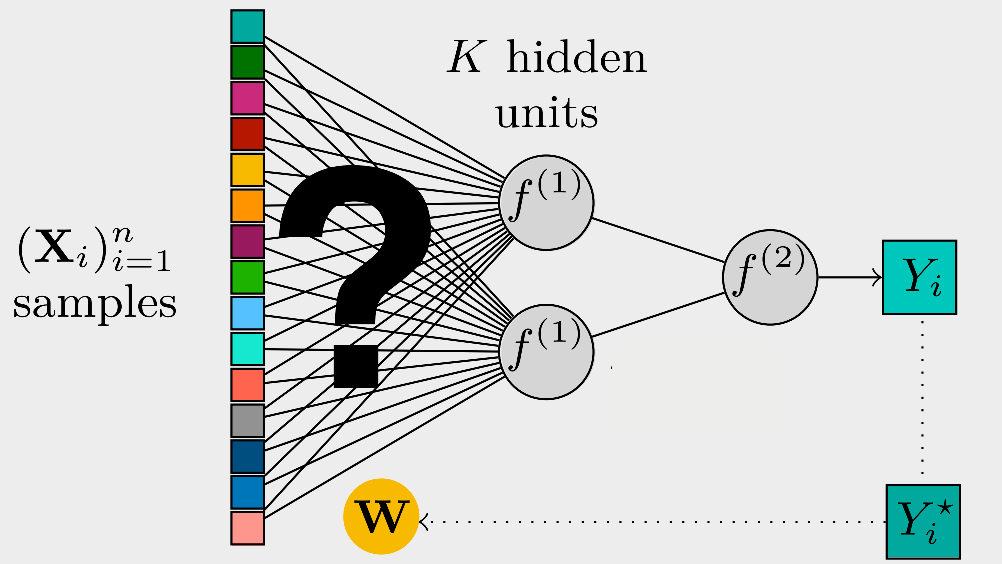


The Teacher-Student scenario: A tale of two networks

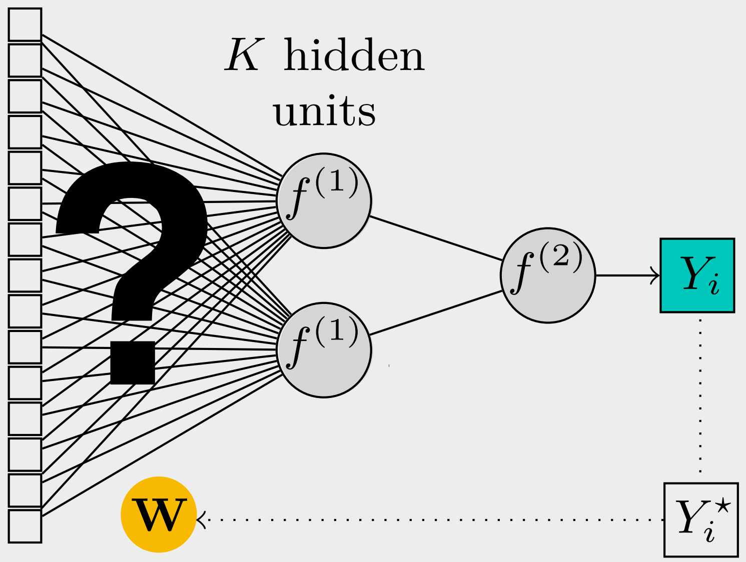
- **Teacher:**



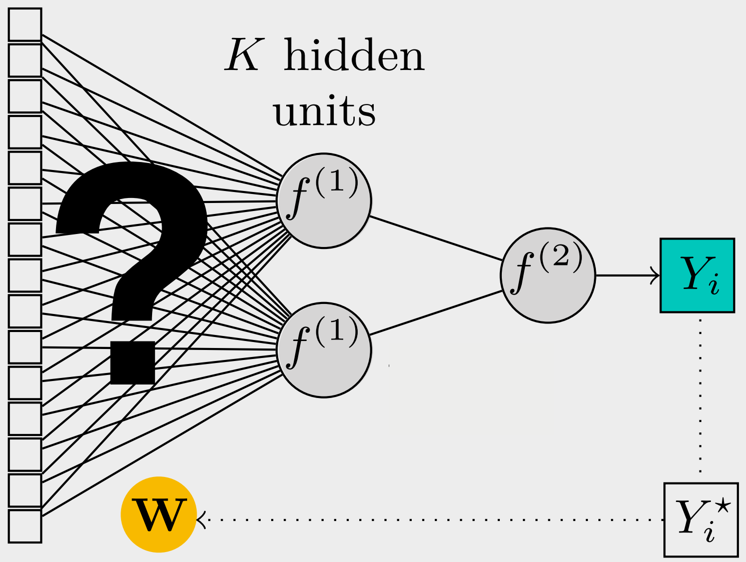
- **Student:**



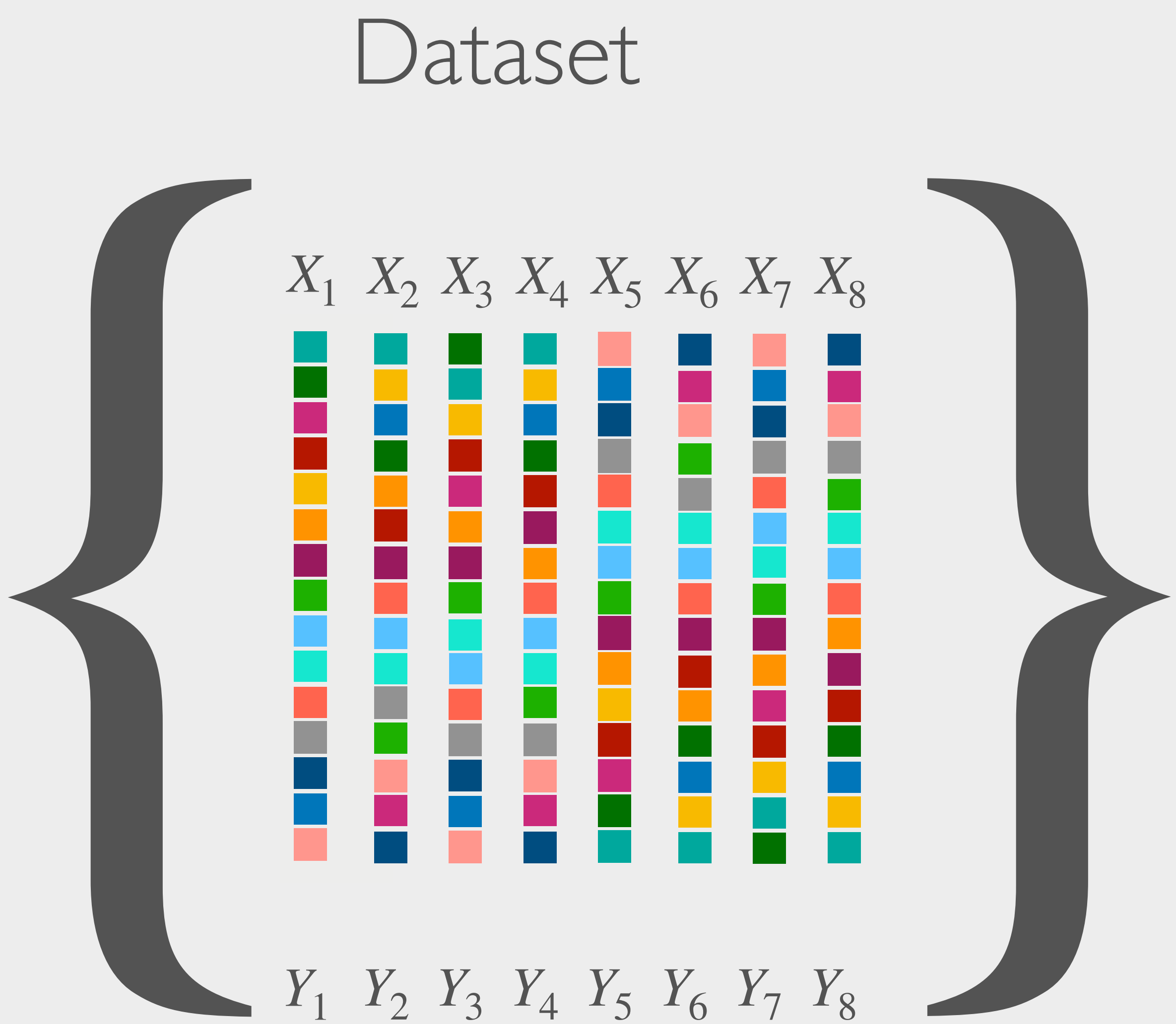
- **Student:**



- **Student:**

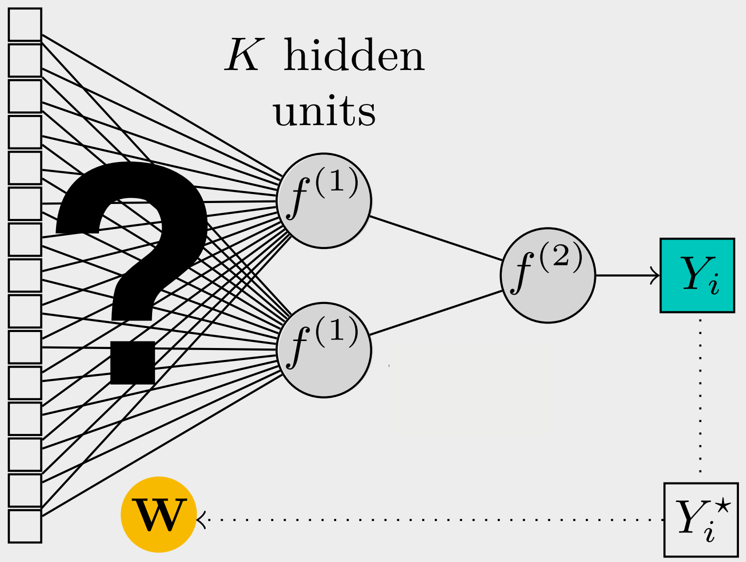


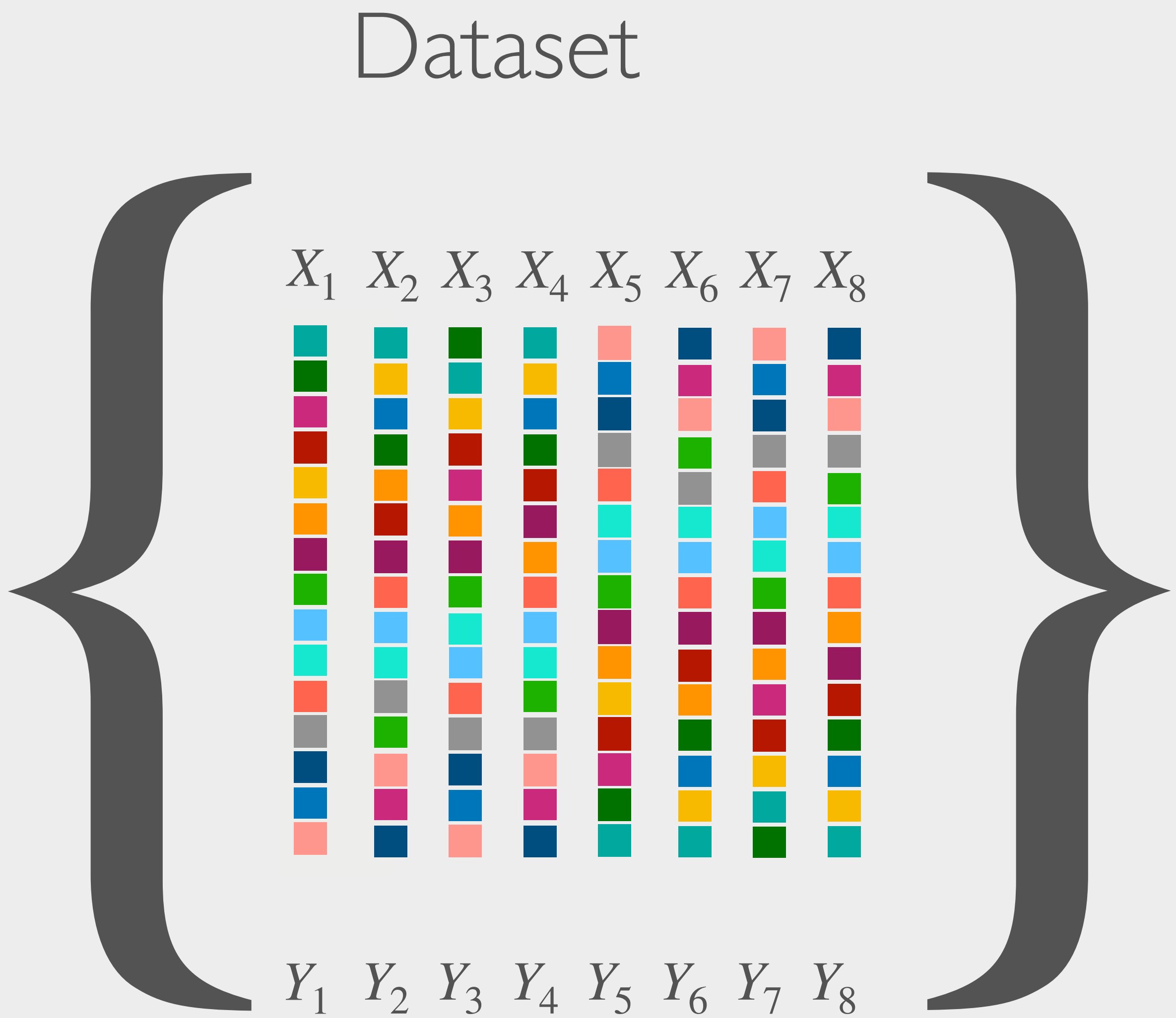
K hidden units
$f^{(1)}$
$f^{(1)}$
$f^{(2)}$
Y_i
$Y_i^\star$
W



Dataset
X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8
Y_1 Y_2 Y_3 Y_4 Y_5 Y_6 Y_7 Y_8

- **Student:**





Now that we have a model of data, we can compute anything….

Statistical Physics Setting

Hamiltonian == Cost function

Ground state energy == minimal error

Ex: Binary classification

$$\mathscr{H} = N - \sum_{i=1}^{N} \delta_{y_i, f(\overrightarrow{x}_i)} = N - \sum_{i=1}^{N} \frac{1 + y_i f(\overrightarrow{x}_i))}{2}$$

Statistical Physics Setting

Hamiltonian == Cost function

Ground state energy == minimal error

Ex: Binary classification

$$\mathscr{H} = N - \sum_{i=1}^{N} \delta_{y_i, f(\overrightarrow{x}_i)} = N - \sum_{i=1}^{N} \frac{1 + y_i f(\overrightarrow{x}_i))}{2}$$

Average over disorder == Average on data generated by the teacher

Spin-glass like model in statistical mechanics of disordered systems

$$F = -\beta \mathbb{E}_{\text{data}} \log Z(\text{data})$$

A BIT OF HISTORY

- Very active part of statistical physics in the 90s. An entire section of arxiv.org/cond-mat is devoted to Disordered Systems and Neural Networks. Hundreds of papers following these studies.
- Review articles and book:
  - Seung, Sompolinsky, Tishby. Statistical mechanics of learning from examples, Phys. Rev. A, 1992.
  - Watkin, Rau, Biehl. The statistical mechanics of learning a rule, Reviews of Modern Physics, 1993.
  - Engel, Van den Broeck. Statistical Mechanics of Learning, Cambridge University Press, 2001.
- Many questions left open, need to re-think many results (next slide).
- After 2000, not much activity on *_artificial_* neural networks among statistical physics community.
- Massive come-back in recent years as Deep Learning made his impact

(SOME) OPEN QUESTIONS

- Can one compute the worst-case Rademacher bound?
- Can one compute the optimal generalisation *rigorously* ?
- How does these two compare?
- Can optimal results be obtained by a tractable (i.e. polynomial) algorithms?
- How good is Stochastic Gradient Descent in this case?

(SOME) OPEN QUESTIONS

- Can one compute the worst-case Rademacher bound?
- Can one compute the optimal generalisation *rigorously* ?
- How does these two compare?
- Can optimal results be obtained by a tractable (i.e. polynomial) algorithms?
- How good is Stochastic Gradient Descent in this case?

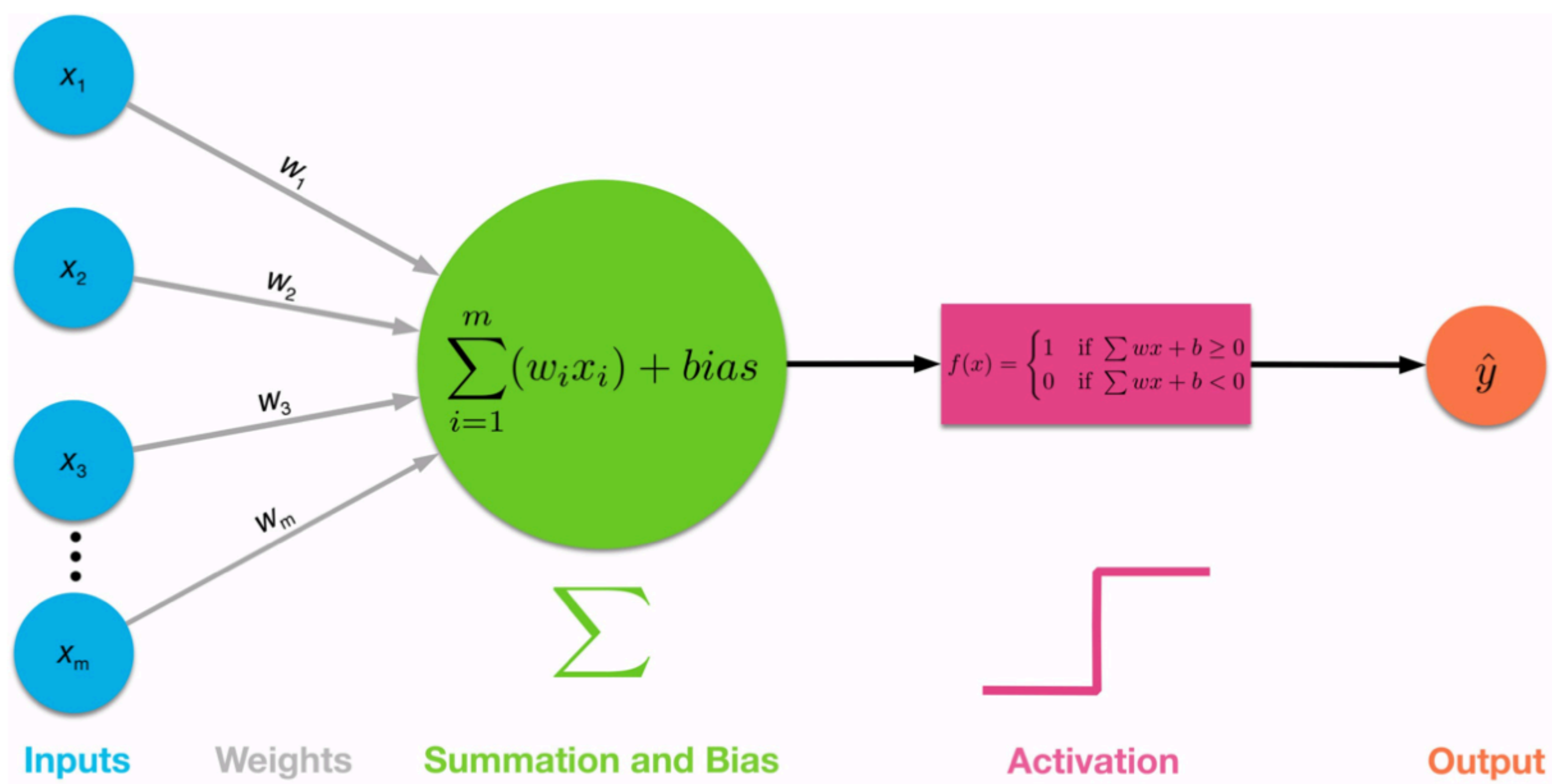
All answered in this talk.

2

Generalisation in single & multi-Layer teacher-student networks

Single Layer Neural Nets

Teacher is a SLNT, Student is a SLNT



$$y = \varphi_\xi(z) = \varphi_\xi(\mathbf{x} \cdot \mathbf{w})$$

$$P_{\text{out}}(y|z) = \mathbb{E}_{P_\xi}[\delta(y - \varphi_\xi(z))]$$

RESULT 1: BAYES OPTIMAL RESULT

Barbier, FK, Macris, Miolane, Zdeborova arXiv:1708.03395, COLT'18

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

Theorem 1 (replica free entropy, informally):

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

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

where

$$\Phi_{P_X}(\hat{m}) \equiv \mathbb{E}_{z,x_0}\left[\ln \mathbb{E}_x \left[e^{\hat{m}xx_0 + \sqrt{\hat{m}}xz - \hat{m}x^2/2}\right]\right]$$

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

$x, x_0 \sim P_X \qquad z, v, w \sim \mathcal{N}(0,1) \qquad \rho = \mathbb{E}_{P_X}(x^2)$

RESULT 1: BAYES OPTIMAL RESULT

Barbier, FK, Macris, Miolane, Zdeborova arXiv:1708.03395, COLT'18

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

Theorem 1 (replica free entropy, informally):

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

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

Theorem 2 (informally): Optimal generalisation error is

$$\mathbb{E}_{v,\xi}\left[\varphi_\xi(\sqrt{\rho}\, v)^2\right] - \mathbb{E}_v\left[\mathbb{E}_{w,\xi}\left[\varphi_\xi(\sqrt{m^*}\, v + \sqrt{\rho - m^*}\, w)\right]^2\right]$$

$\rho = \mathbb{E}_{P_X}(x^2)$
$v, w \sim \mathcal{N}(0,1)$

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

RESULT 1: BAYES OPTIMAL RESULT

Barbier, FK, Macris, Miolane, Zdeborova arXiv:1708.03395, COLT'18

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

Theorem 1 (replica free entropy, informally):

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

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

Theorem 2 (informally): Optimal generalisation error is

$$\mathbb{E}_{v,\xi}\left[\varphi_\xi(\sqrt{\rho}\, v)^2\right] - \mathbb{E}_v\left[\mathbb{E}_{w,\xi}\left[\varphi_\xi(\sqrt{m^*}\, v + \sqrt{\rho - m^*}\, w)\right]^2\right]$$

$\rho = \mathbb{E}_{P_X}(x^2)$
$v, w \sim \mathcal{N}(0,1)$

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

Generalization and rigorous proof of early results by Derrida, Gardner '89, Gyorgyi '90 & Sompolinsky, Tishby, Seung '92

RESULT 2: RADEMACHER COMPLEXITY

Aubin, FK, Zdeborova, in preparation

Rademacher complexity can be obtained with the replica method

Groundstate energy with random labels:

$$f^{\text{rand}}(T, \alpha) = \lim_{n,p\to\infty} \mathbb{E}_{\text{random label}} \frac{1}{p} \log\left(\int d\theta e^{-\beta\mathscr{H}}\right)$$

from replica method (1RSB level)

$$e_{\text{GS}} = -\partial_\beta f^{\text{rand}}(T, \alpha)$$

$$e^{\text{GS}} = \lim_{p\to\infty} \mathbb{E}_{\text{random label}} \frac{< \mathscr{H} >_{T\to 0}}{p}$$

Groundstate energy gives the Rademacher Complexity

$$\mathscr{R}_N = 1 - \frac{2e_{\text{GS}}(\alpha)}{\alpha}$$

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

RESULT 2: RADEMACHER COMPLEXITY

Aubin, FK, Zdeborova, in preparation

Rademacher complexity can be obtained with the replica method

Groundstate energy with random labels:

$$f^{\mathrm{rand}}(T,\alpha) = \lim_{n,p\to\infty} \mathbb{E}_{\mathrm{random\ label}} \frac{1}{p} \log\left(\int d\theta e^{-\beta\mathcal{H}}\right)$$

from replica method (1RSB level)

$$e_{\mathrm{GS}} = -\partial_\beta f^{\mathrm{rand}}(T,\alpha)$$

$$e^{\mathrm{GS}} = \lim_{p\to\infty} \mathbb{E}_{\mathrm{random\ label}} \frac{<\mathcal{H}>_{T\to 0}}{p}$$

Groundstate energy gives the Rademacher Complexity

$$\mathcal{R}_N = 1 - \frac{2e_{\mathrm{GS}}(\alpha)}{\alpha}$$

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

Generalization of early results by
Derrida and Gardner '89 & Mezard, Krauth '89

RESULT 2: RADEMACHER COMPLEXITY

Aubin, FK, Zdeborova, in preparation

Rademacher complexity can be obtained with the replica method

Groundstate energy with random labels:

$$f^{\text{rand}}(T,\alpha) = \lim_{n,p\to\infty} \mathbb{E}_{\text{random label}} \frac{1}{p} \log\left(\int d\theta e^{-\beta\mathscr{H}}\right)$$

from replica method (1RSB level)

$$e_{\text{GS}} = -\partial_\beta f^{\text{rand}}(T,\alpha)$$

$$e^{\text{GS}} = \lim_{p\to\infty} \mathbb{E}_{\text{random label}} \frac{<\mathscr{H}>_{T\to 0}}{p}$$

Groundstate energy gives the Rademacher Complexity

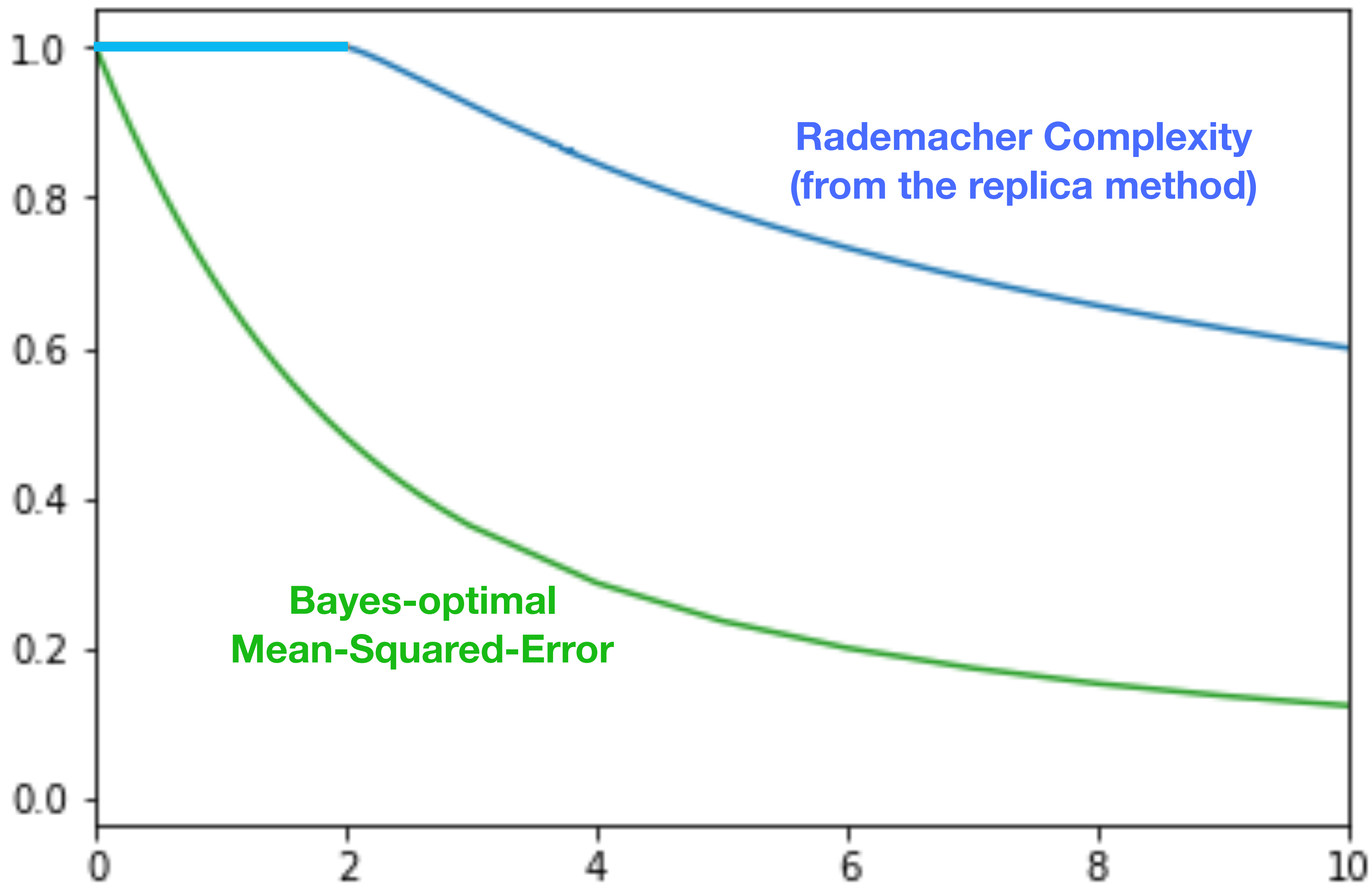
$$\mathscr{R}_N = 1 - \frac{2e_{\text{GS}}(\alpha)}{\alpha}$$

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

Generalization of early results by
Derrida and Gardner '89 & Mezard, Krauth '89
Mathematically open

Typical vs Worst case

Spherical Perceptron $\mathbf{W} \in \mathbb{R}^p; ||\mathbf{W}||_2^2 = 1$

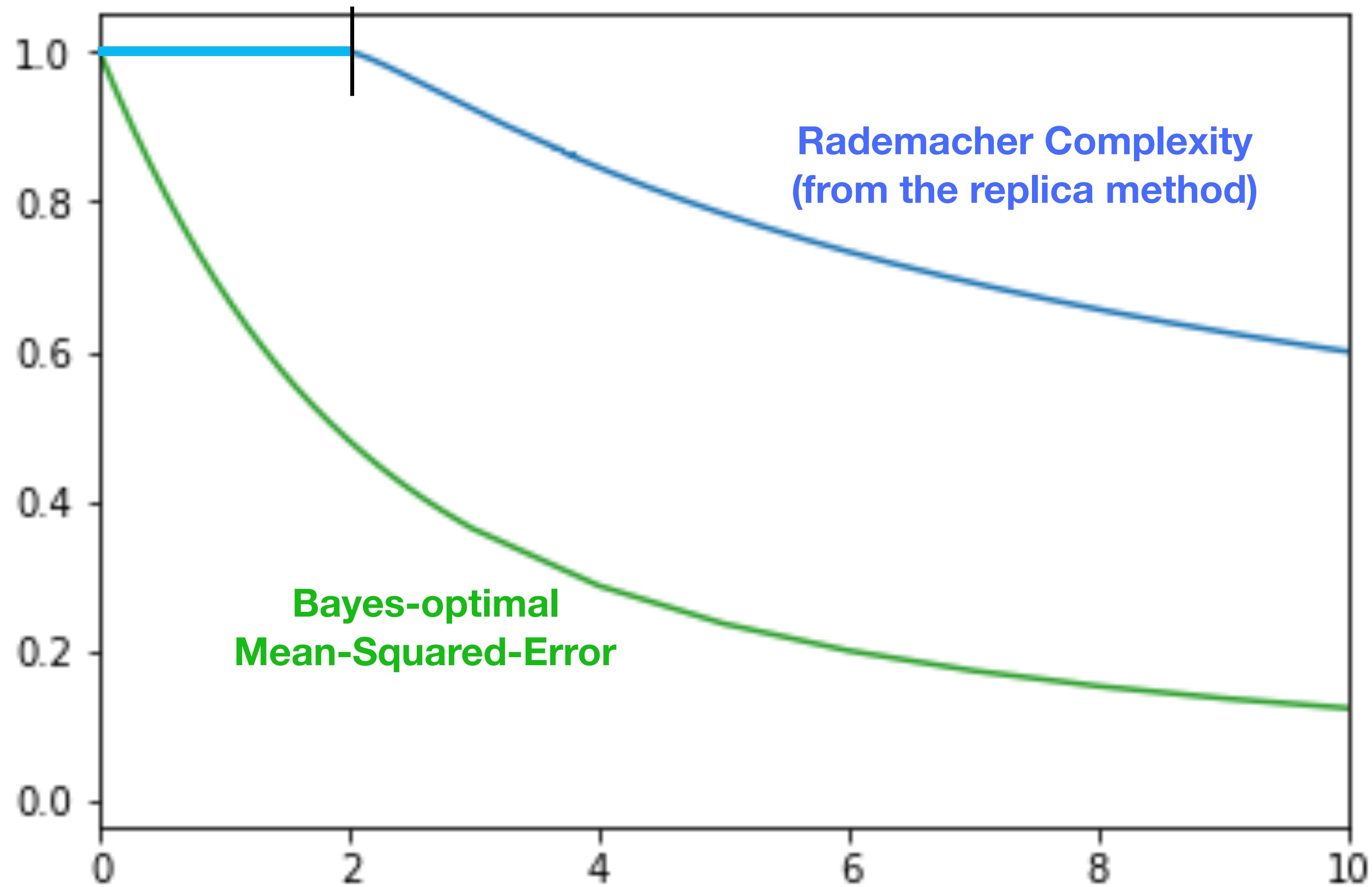


$$\alpha = \frac{N}{P}$$

Typical vs Worst case

Spherical Perceptron $\mathbf{W} \in \mathbb{R}^p; ||\mathbf{W}||_2^2 = 1$

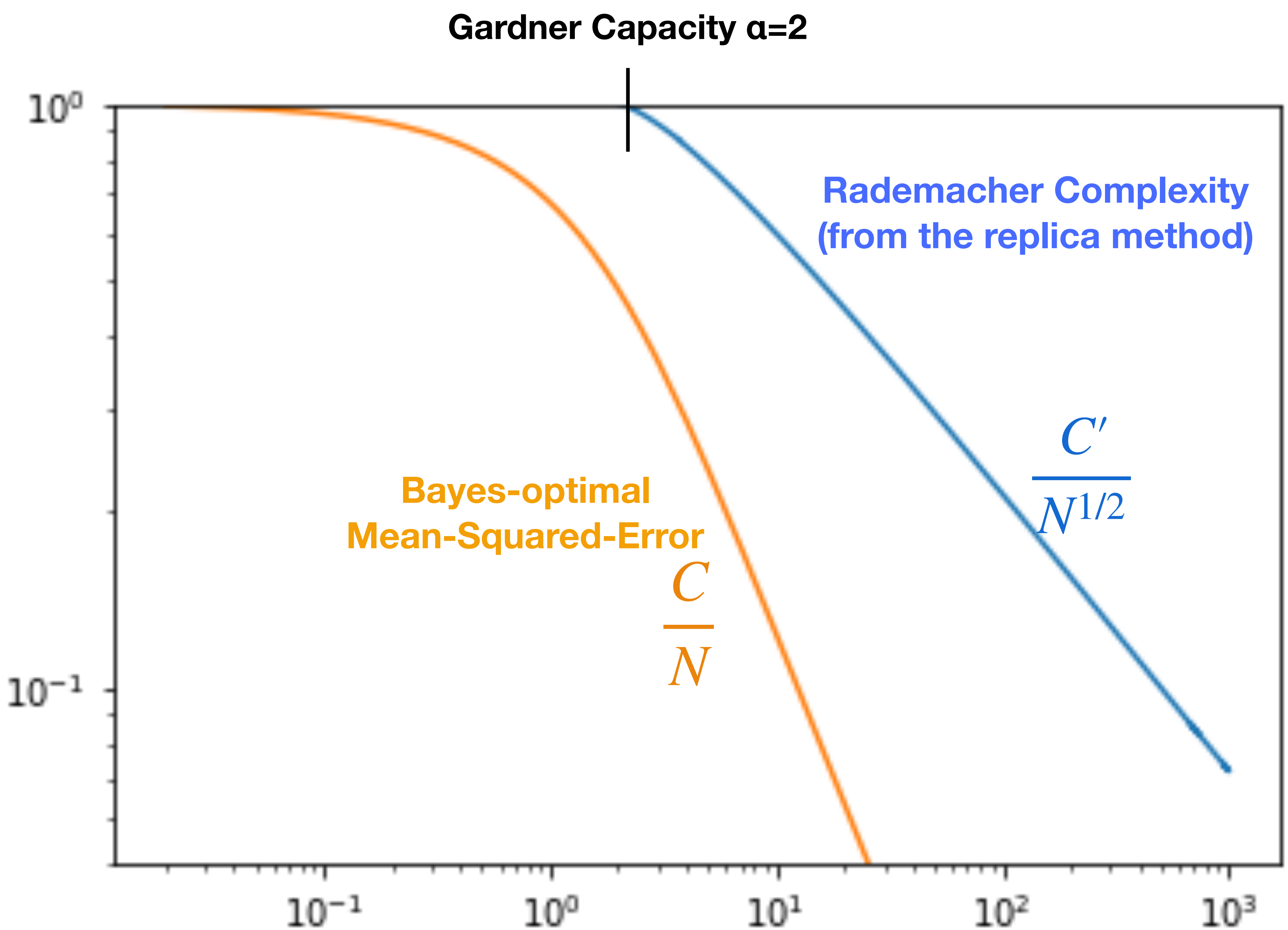
Gardner Capacity α=2 Cover (1965), Derrida-Gardner, 1988, 1989



$$\alpha = \frac{N}{P}$$

Typical vs Worst case

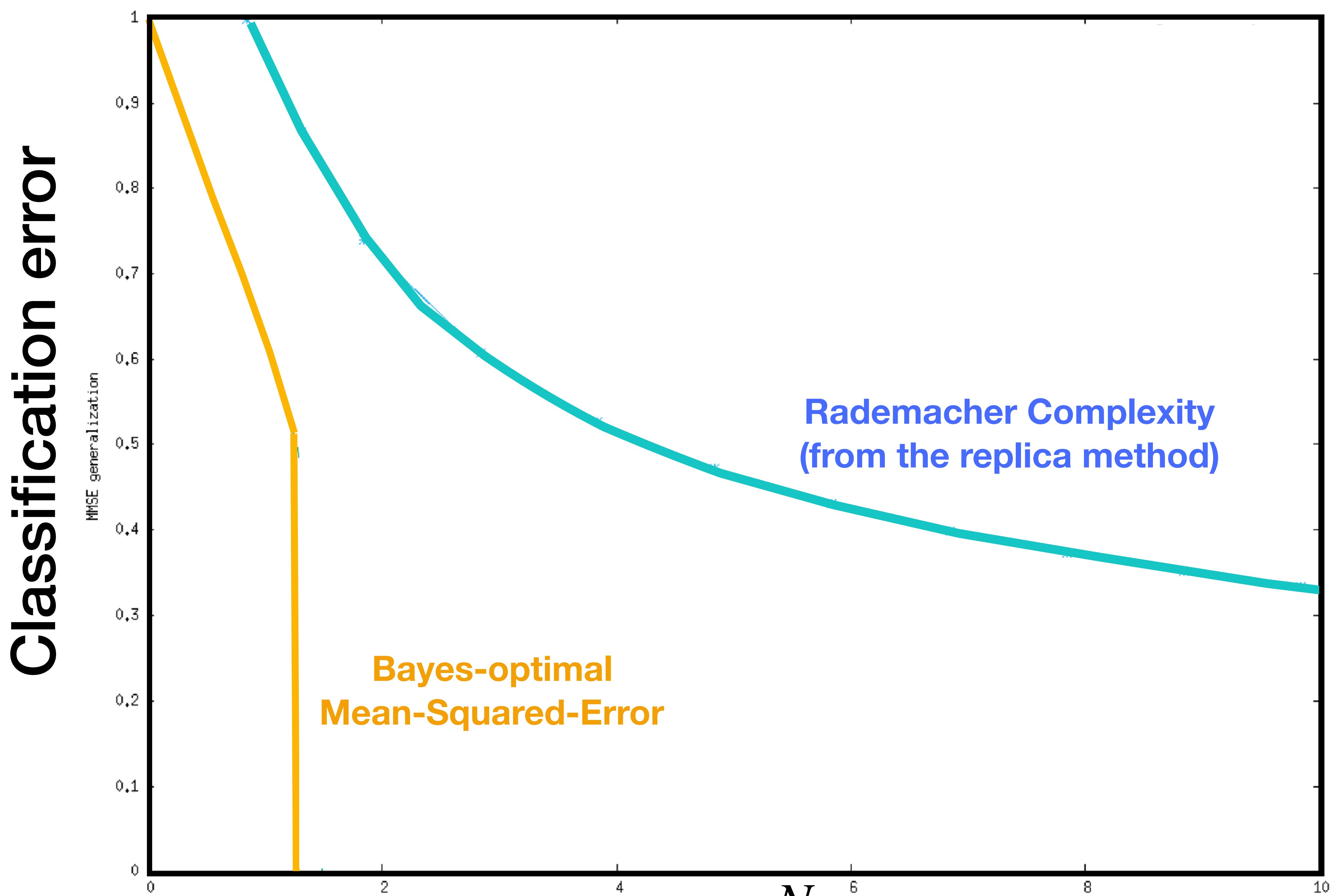
Spherical Perceptron $\mathbf{W} \in \mathbb{R}^p; ||\mathbf{W}||_2^2 = 1$



$$\alpha = \frac{N}{P}$$

Typical vs Worst case

Binary Perceptron $W_i = \pm 1$

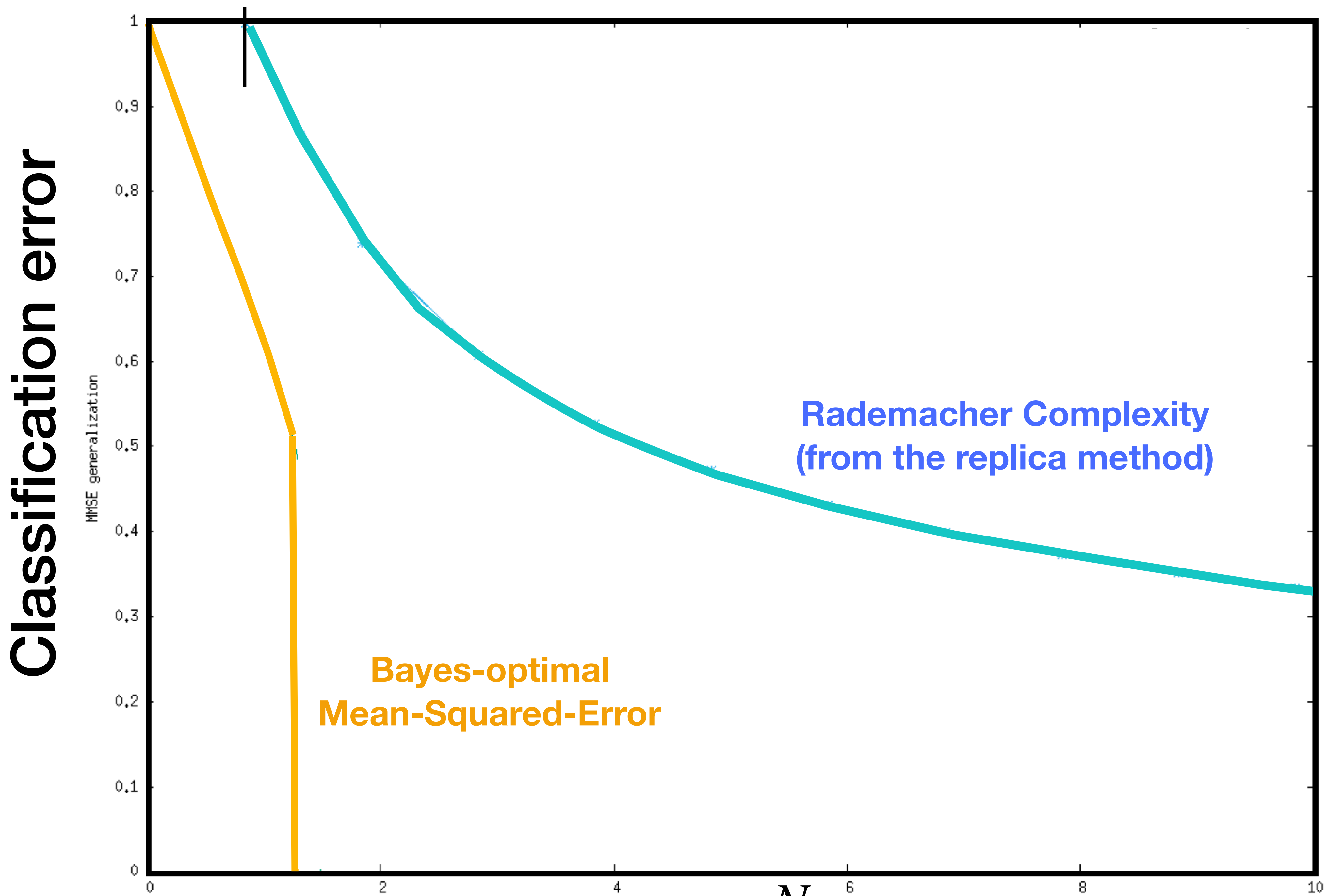


$$\alpha = \frac{N}{P}$$

Typical vs Worst case

Binary Perceptron $W_i = \pm 1$

Gardner Capacity α=0.8333 Mezard-Krauth '89

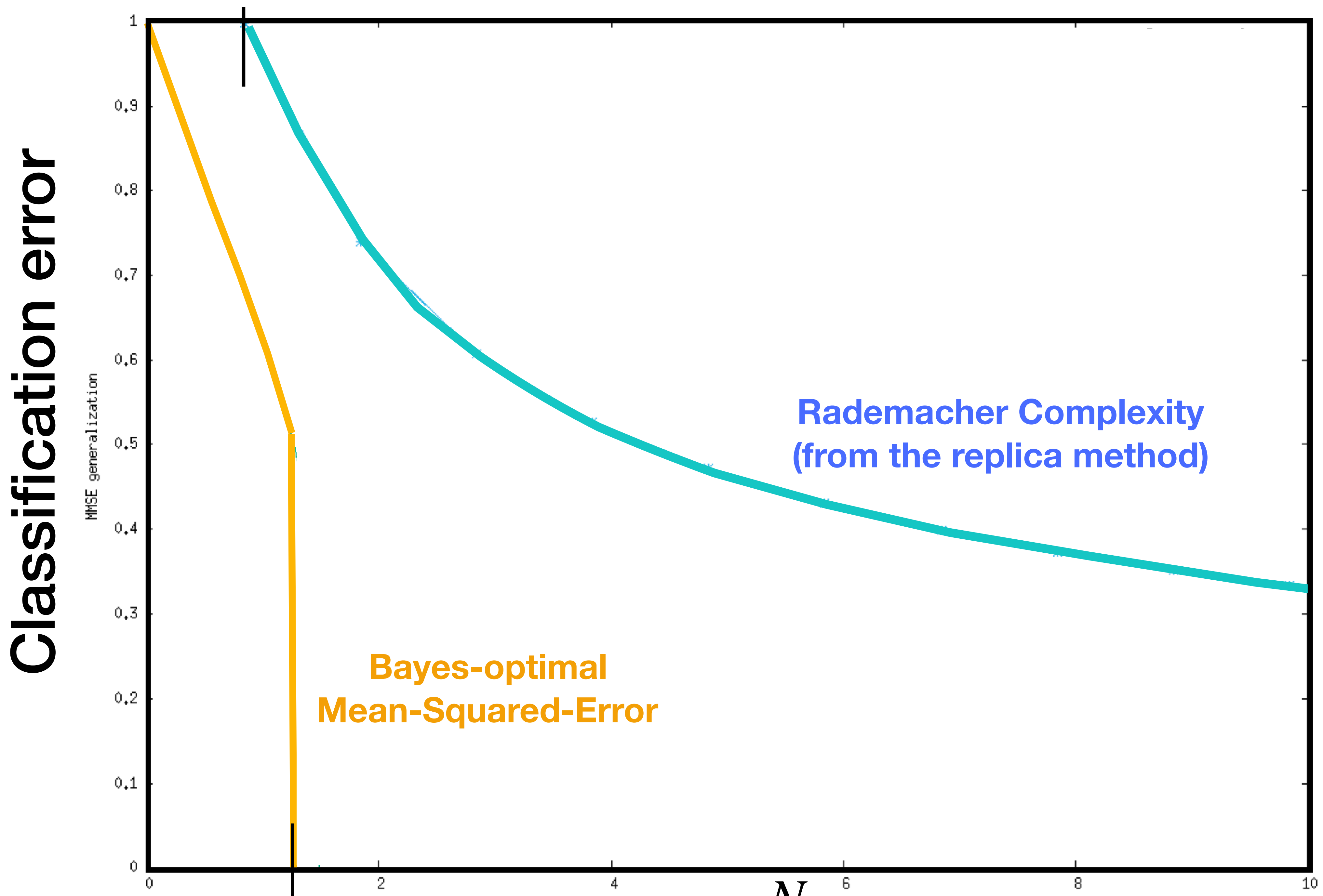


$$\alpha = \frac{N}{P}$$

Typical vs Worst case

Binary Perceptron $W_i = \pm 1$

Gardner Capacity α=0.8333 Mezard-Krauth '89

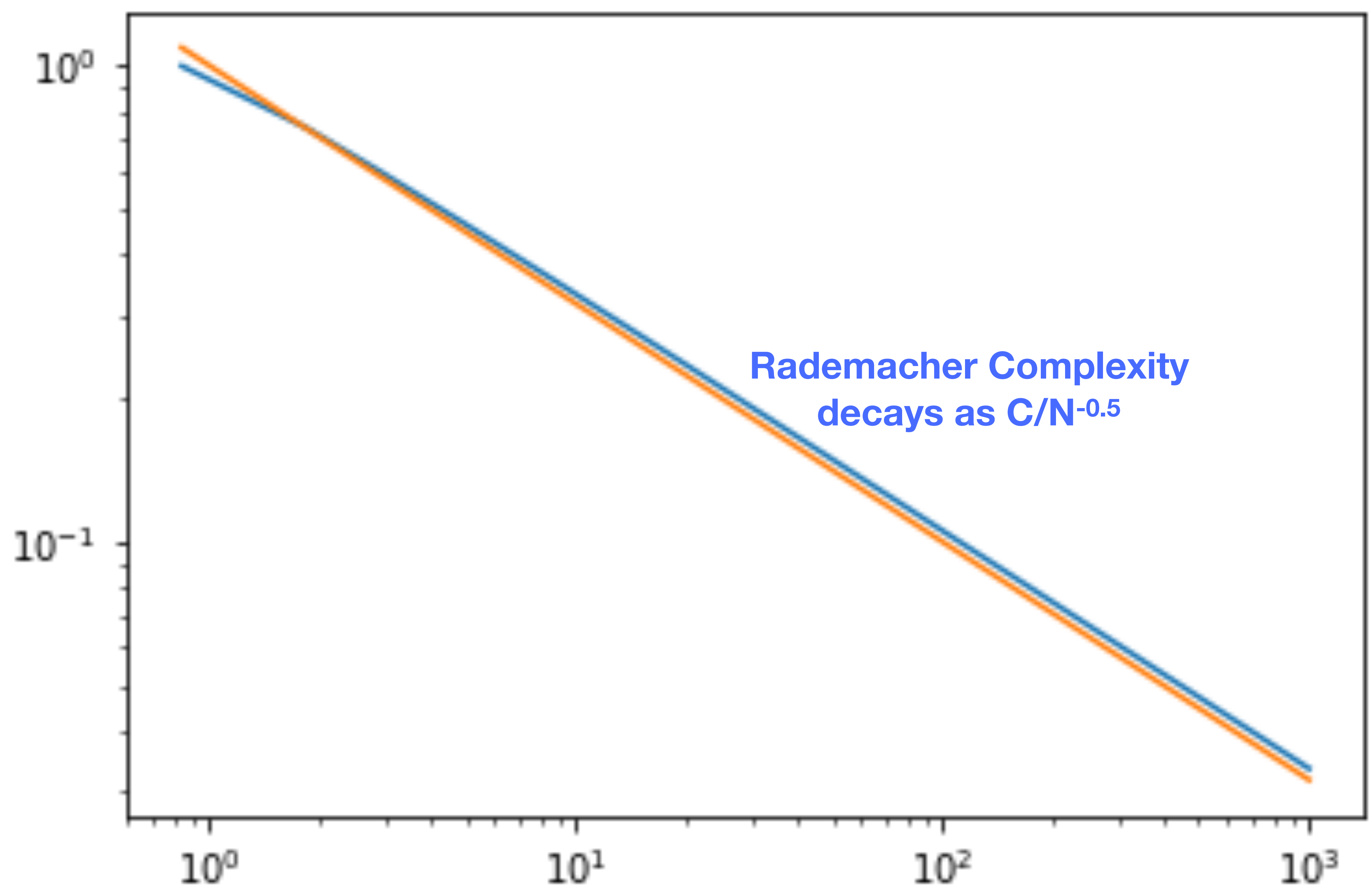


Perfect generalisation α=1.24…

$$\alpha = \frac{N}{P}$$

Typical vs Worst case

Binary Perceptron $W_i = \pm 1$

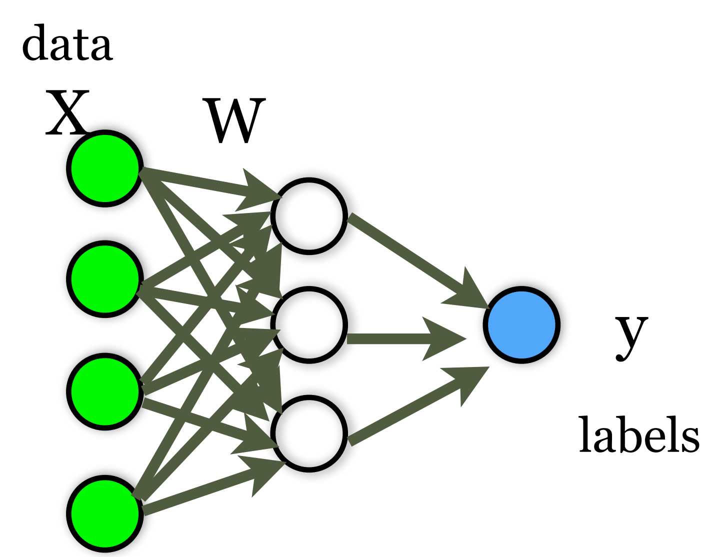


$$\alpha = \frac{N}{P}$$

Multi-Layer Neural Nets

The Committee machine

- P input units
- K hidden units
- output unit
- N training samples



$$K = O(1)$$

Limit: $\alpha = \frac{N}{P} = O(1)$ $\quad K = O(1)$ $\quad N, P \to \infty$

Rachemacher bound

Mitchison and Durbin [Biol. Cybern. 60, 345 (1989)],
Monasson-Zecchina ‘95

Gardner Capacity: $d_{\text{Gardner}} \approx PK\sqrt{\log K}$

Rachemacher bound

Mitchison and Durbin [Biol. Cybern. 60, 345 (1989)],
Monasson-Zecchina '95

Gardner Capacity: $d_{\text{Gardner}} \approx PK\sqrt{\log K}$

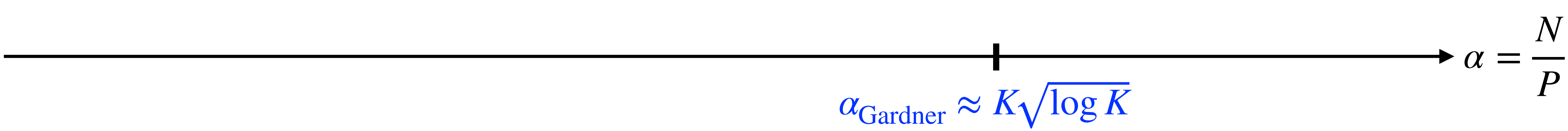
We thus expect a decay as $\dfrac{C}{\sqrt{N}}$ **shortly after** $N \gg PK\sqrt{\log K}$

Rachemacher bound

Mitchison and Durbin [Biol. Cybern. 60, 345 (1989)],
Monasson-Zecchina '95

Gardner Capacity: $d_{\text{Gardner}} \approx PK\sqrt{\log K}$

We thus expect a decay as $\dfrac{C}{\sqrt{N}}$ **shortly after** $N \gg PK\sqrt{\log K}$

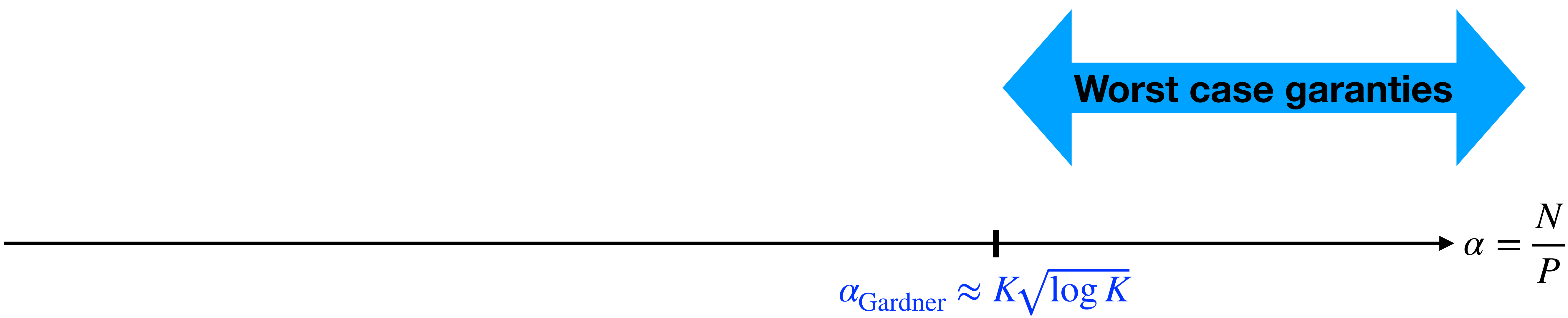


Rachemacher bound

Mitchison and Durbin [Biol. Cybern. 60, 345 (1989)],
Monasson-Zecchina ‘95

Gardner Capacity: $d_{\text{Gardner}} \approx PK\sqrt{\log K}$

We thus expect a decay as $\frac{C}{\sqrt{N}}$ **shortly after** $N \gg PK\sqrt{\log K}$



Optimal generalization

Optimal generalization

EUROPHYSICS LETTERS 15 October 1992

Europhys. Lett., **20** (4), pp. 375-380 (1992)

Generalization in a Large Committee Machine.

H. SCHWARZE and J. HERTZ

CONNECT, The Niels Bohr Institute and Nordita
Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark

(received 10 March 1992; accepted in final form 12 August 1992)

PACS. 87.10 – General, theoretical, and mathematical biophysics (inc. logic of biosystems, quantum biology and relevant aspects of thermodynamics, information theory, cybernetics, and bionics).
PACS. 02.50 – Probability theory, stochastic processes, and statistics.
PACS. 64.60C – Order-disorder and statistical mechanics of model systems.

Optimal generalization

EUROPHYSICS LETTERS 15 October 1992

Europhys. Lett., **20** (4), pp. 375-380 (1992)

Generalization in a Large Committee Machine.

J. Phys. A: Math. Gen. 26 (1993) 5781–5794. Printed in the UK

H. SCHWAR

CONNECT

Blegdamsve

(received 1

Learning a rule in a multilayer neural network

PACS. 87.10

H Schwarze

CONNECT, The Niels Bohr Institute, Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark

PACS. 02.50

PACS. 64.60

Received 4 June 1993

Abstract. The problem of learning from examples in multilayer networks is studied within the framework of statistical mechanics. Using the replica formalism we calculate the average generalization error of a fully connected committee machine in the limit of a large number of hidden units. If the number of training examples is proportional to the number of inputs in the network, the generalization error as a function of the training set size approaches a finite value. If the number of training examples is proportional to the number of weights in the network we find first-order phase transitions with a discontinuous drop in the generalization error for both binary and continuous weights.

Optimal generalization

EUROPHYSICS LETTERS 15 October 1992
Europhys. Lett., **20** (4), pp. 375-380 (1992)

Generalization in a Large Committee Machine.

H. SCHWAR
CONNECT
Blegdamsve

(received 1
PACS. 87.10
PACS. 02.50
PACS. 64.60

J. Phys. A: Math. Gen. 26 (1993) 5781–5794. Printed in the UK

Learning a rule in a multilayer n

H Schwarze
CONNECT, The Niels Bohr Institute, Ble

Received 4 June 1993

Abstract. The problem of learning fron
the framework of statistical mechanics. I
generalization error of a fully connected
hidden units. If the number of training ex
network, the generalization error as a func
If the number of training examples is pro
find first-order phase transitions with a di
binary and continuous weights.

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[⋆]

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.

NeurIPS 2018

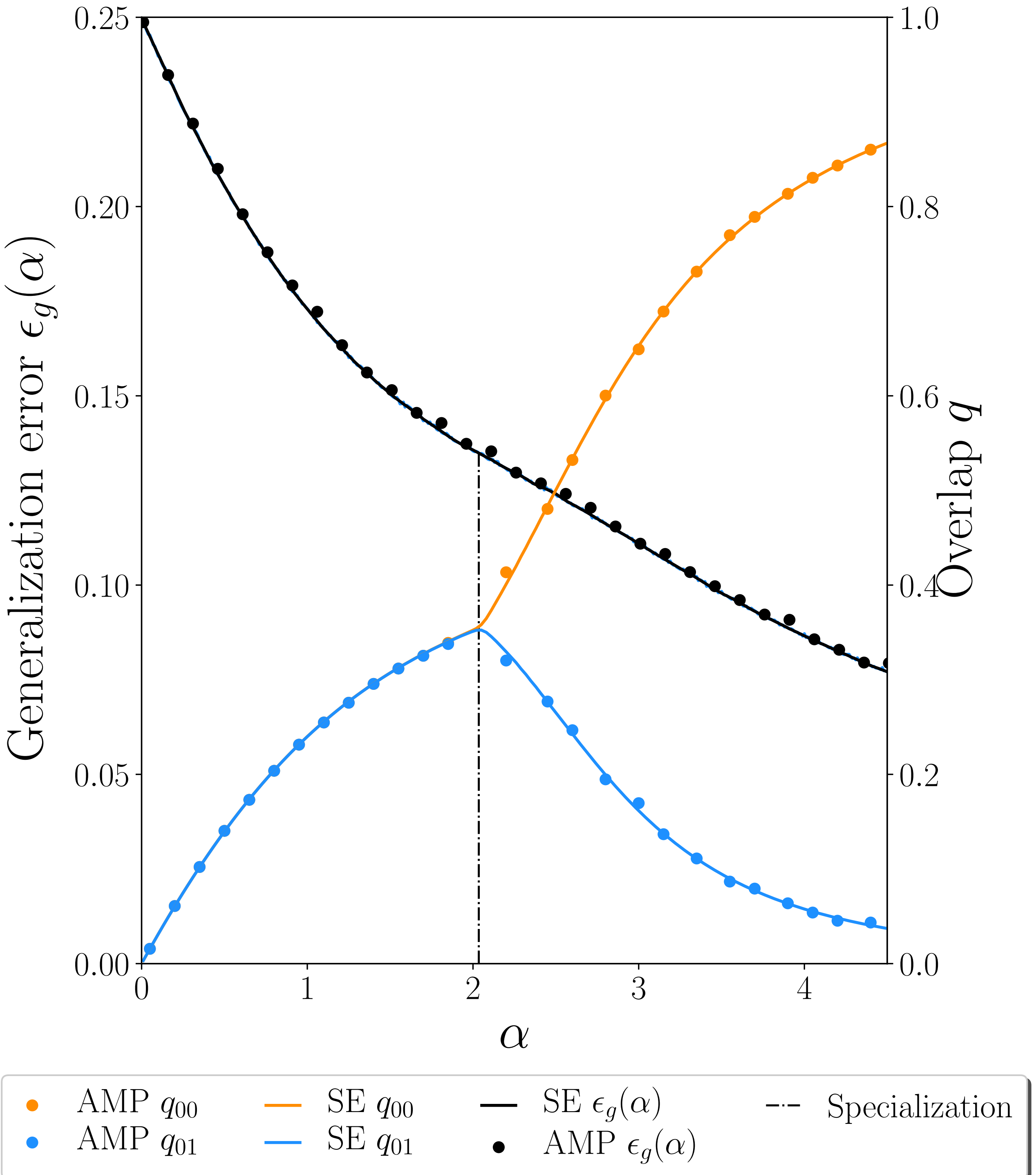
Optimal generalization

K=2

$$y_\mu = \text{sign}\left[\text{sign}\left(\sum_i F_{\mu,i} x_{i,1}\right) + \text{sign}\sum_i \left(F_{\mu,i} x_{i,2}\right)\right]$$

$$\text{sign}(0) = 0$$

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



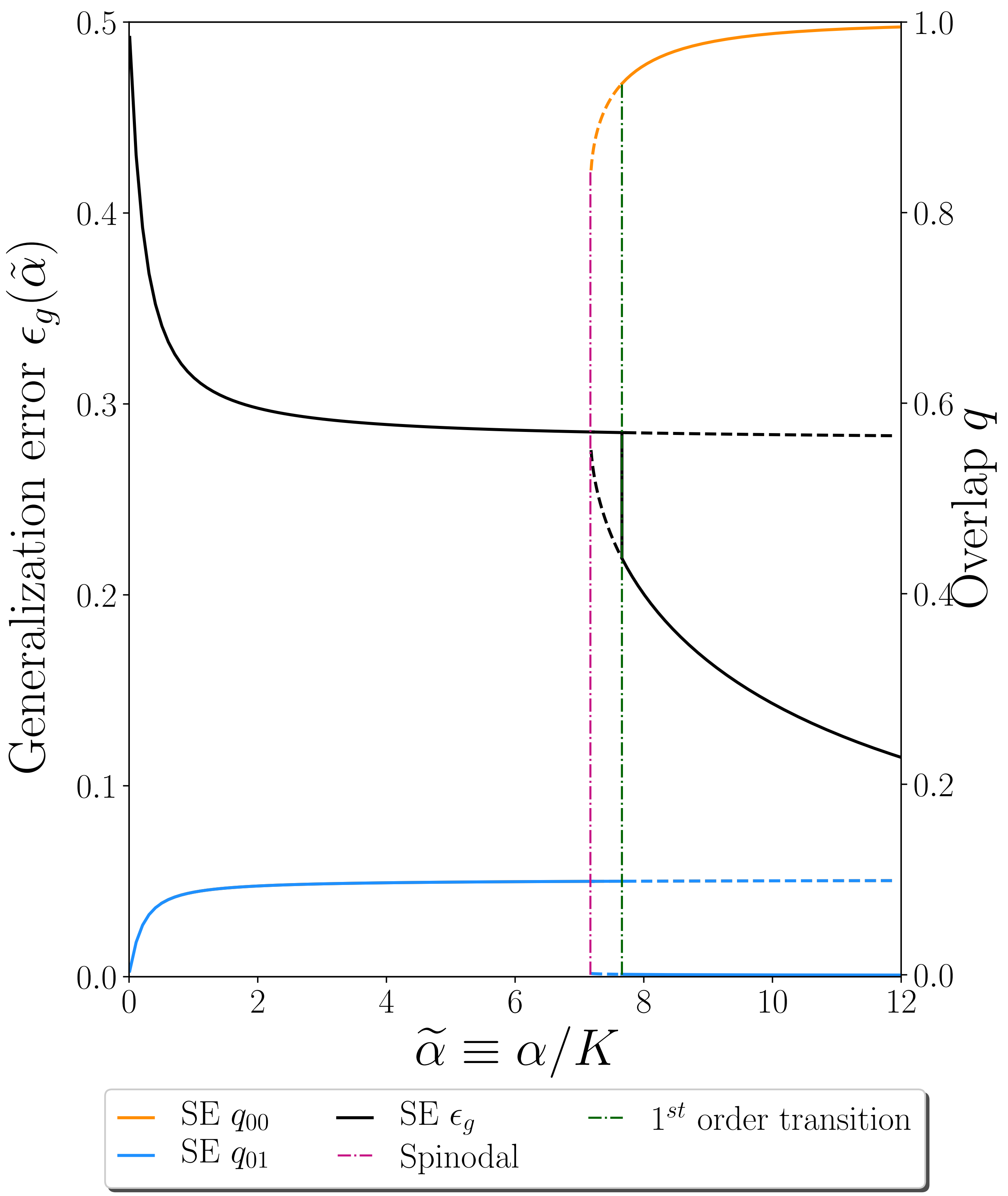
Optimal generalization

$N \gg K \gg 1$

$$y_\mu = \text{sign}\Big[\sum_{l=1}^{K} \text{sign}\big(\sum_{i=1}^{p} F_{\mu i} x^*_{il}\big)\Big]$$

- Specialization phase transition
- First-order threshold:

$$N > 7.65KP$$



Optimal generalization

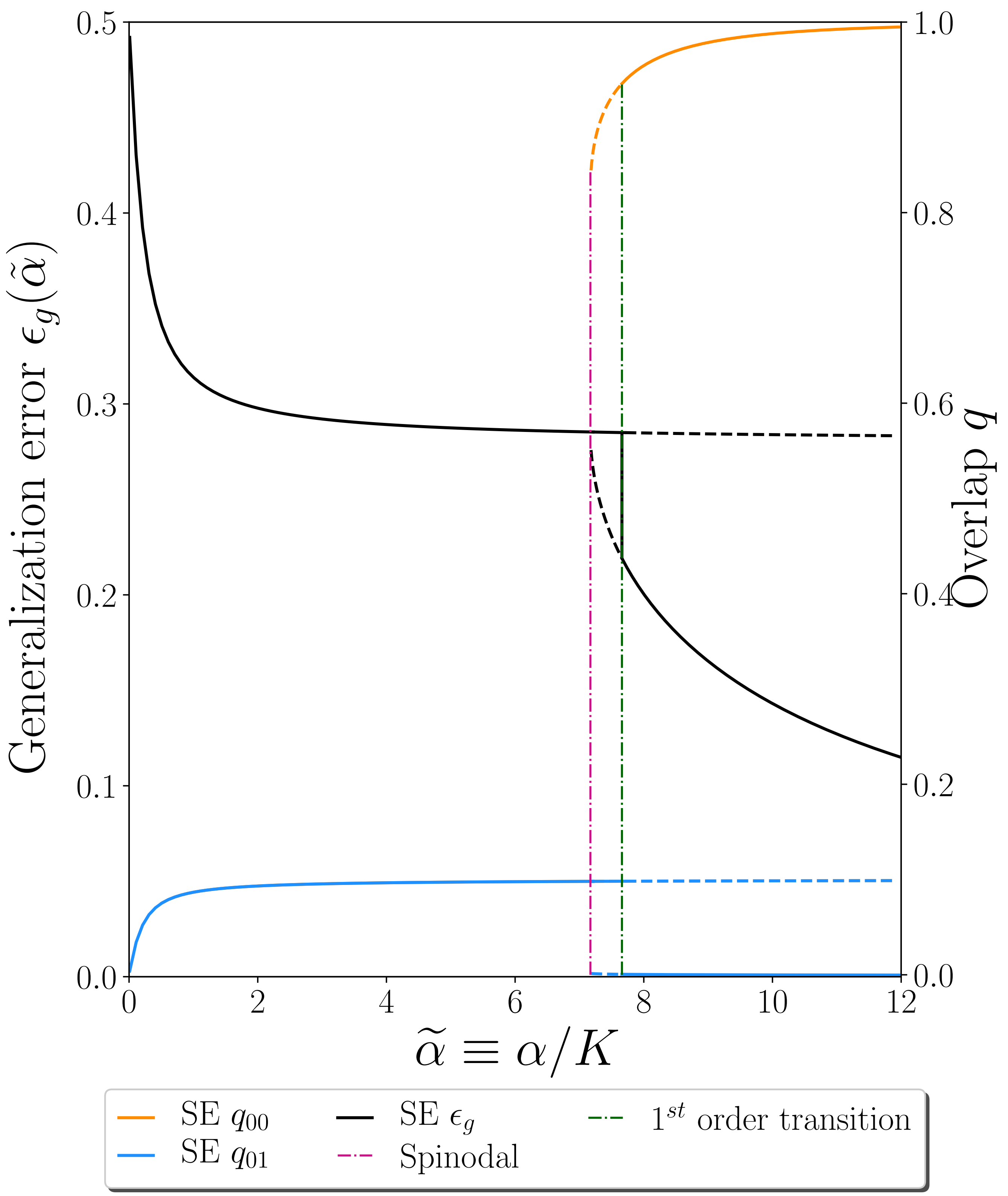
$N \gg K \gg 1$

$$y_\mu = \mathrm{sign}\Big[\sum_{l=1}^{K} \mathrm{sign}\big(\sum_{i=1}^{p} F_{\mu i} x^*_{il}\big)\Big]$$

- Specialization phase transition
- First-order threshold:

$$N > 7.65KP$$

Capacity: $d_{\mathrm{Gardner}} \geq CstPK\sqrt{\log K}$



Committee machine

Very large gap between typical and worst case!

No good learning

Good “typical” performances

Good “worst case” performances

$\alpha = \frac{N}{D}$

$\alpha_{\text{IT}} = O(PK)$

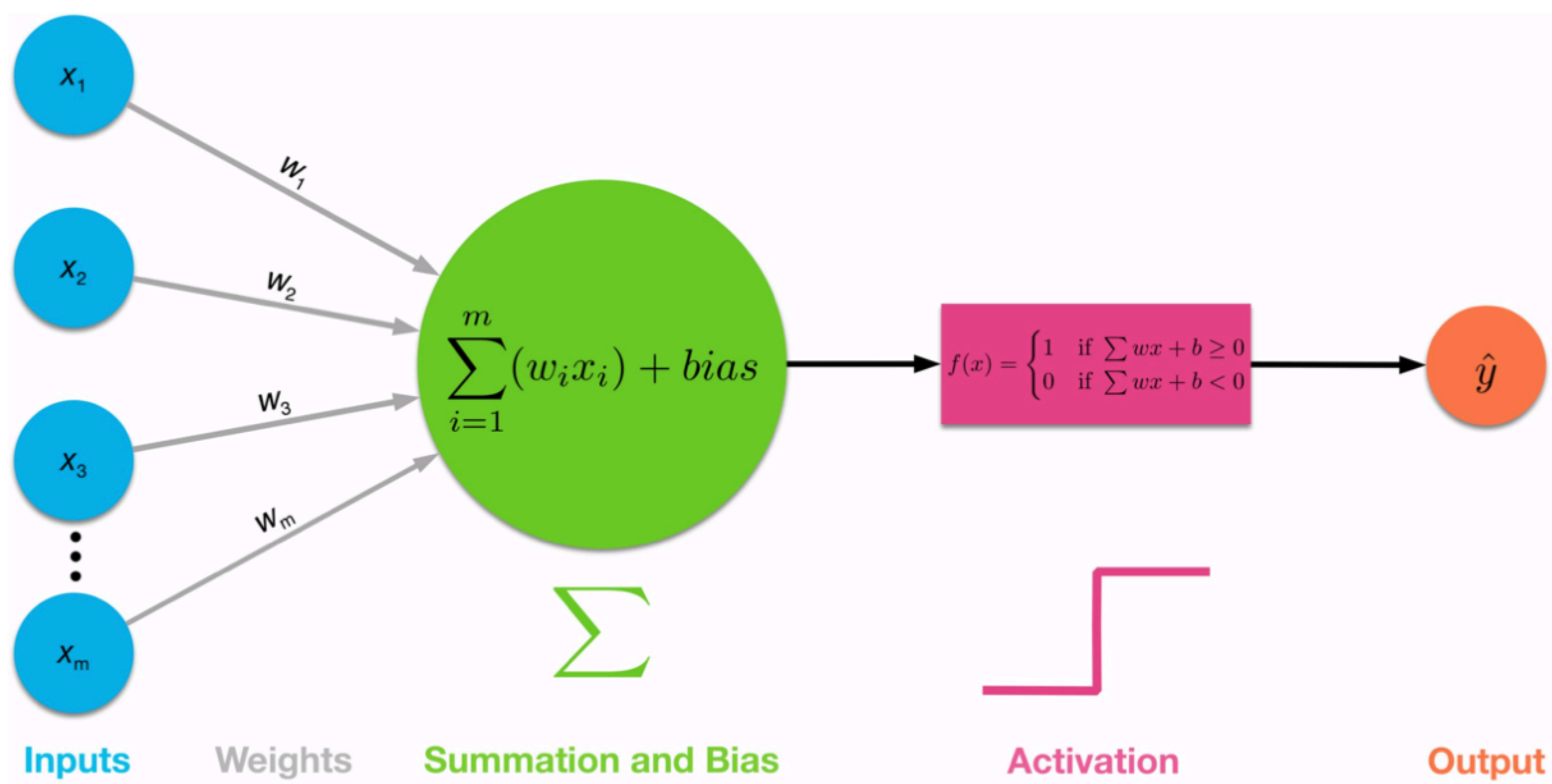
$\alpha_{\text{IT}} = O(K\sqrt{\log(K)})$

3

Efficient Optimal Algorithm

Single Layer Neural Nets

Teacher is a SLNT, Student is a SLNT



$$y = \varphi_\xi(z) = \varphi_\xi(\mathbf{x} \cdot \mathbf{w})$$

$$P_{\text{out}}(y|z) = \mathbb{E}_{P_\xi}[\delta(y - \varphi_\xi(z))]$$

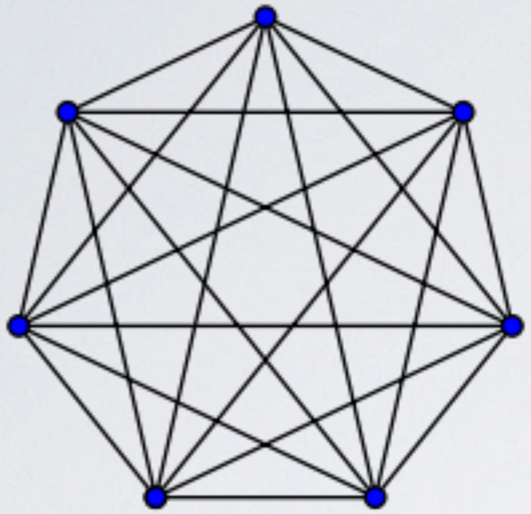
APPROXIMATE MESSAGE PASSING (AMP)

- No optimal and efficient algorithm during the classic period of stat-mech of neural nets (e.g. State of the art: MCMC for binary perceptron when $P<50$)
- Spectacular recent progress on AMP, a mean-field method "*on steroid*":
  - Thouless-Anderson-Palmer '76 (TAP): improved mean-field equations
  - George-Yeddida '91:TAP is a correction to standard mean-field
  - Applying TAP to various problems: Neural networks Mezard '89, Hopfield model Sompolinsky '92, Error-correction Tanaka '02
  - TAP becomes an *iterative* algorithm "AMP": Donoho, Maleki, Montanari'09 for compressed sensing and linear estimation, Rangan'10 generic output for linear estimation
  - Rigorous results on AMP: Bolthausen '09, Bayati, Montanari'10,

AMP IN A NUTSHELL

(

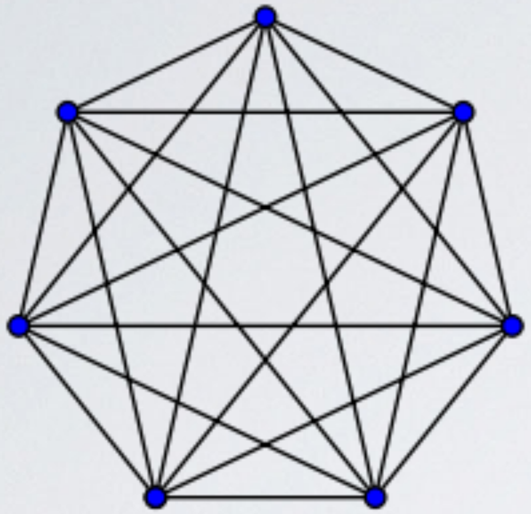
THE SHERRINGTON-KIRKPATRICK MODEL



$$H_N(\sigma, J) = -\frac{1}{\sqrt{N}} \sum_{(i,j)} J_{ij}\sigma_i\sigma_j - h\sum_i \sigma_i,$$

Ising model with disorder

THE SHERRINGTON-KIRKPATRICK MODEL



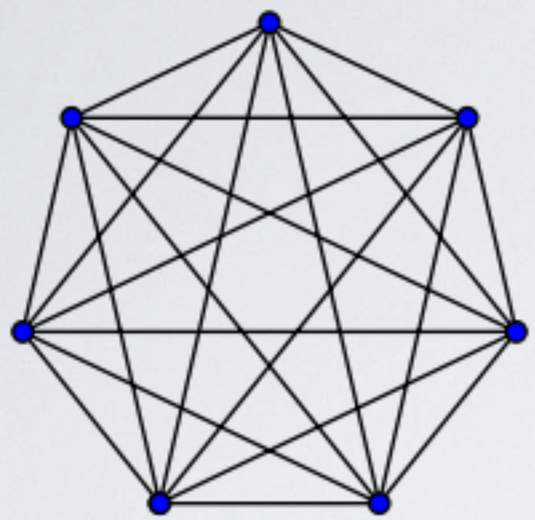
$$H_N(\sigma, J) = -\frac{1}{\sqrt{N}} \sum_{(i,j)} J_{ij}\sigma_i\sigma_j - h\sum_i \sigma_i,$$

Ising model with disorder

Naive mean-field

$$m_i = \tanh\left[h_j + \beta \sum_j J_{ij} m_j\right]$$

THE SHERRINGTON-KIRKPATRICK MODEL



$$H_N(\sigma, J) = -\frac{1}{\sqrt{N}} \sum_{(i,j)} J_{ij}\sigma_i\sigma_j - h\sum_i \sigma_i,$$

Ising model with disorder

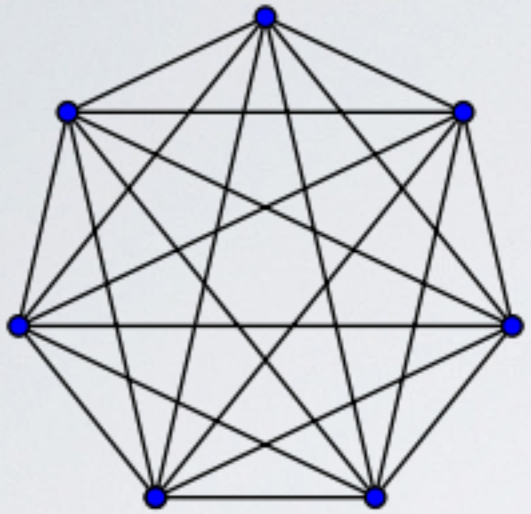
Naive mean-field

$$m_i = \tanh\left[h_j + \beta \sum_j J_{ij} m_j\right]$$

Thouless-Anderson-Palmer

$$m_i = \tanh\left[h + \sum_j \beta J_{ij} m_j - \beta^2 \sum_j J_{ij}^2 (1 - m_j^2) m_i\right]$$

THE SHERRINGTON-KIRKPATRICK MODEL



$$H_N(\sigma, J) = -\frac{1}{\sqrt{N}} \sum_{(i,j)} J_{ij}\sigma_i\sigma_j - h \sum_i \sigma_i,$$

Ising model with disorder

Naive mean-field

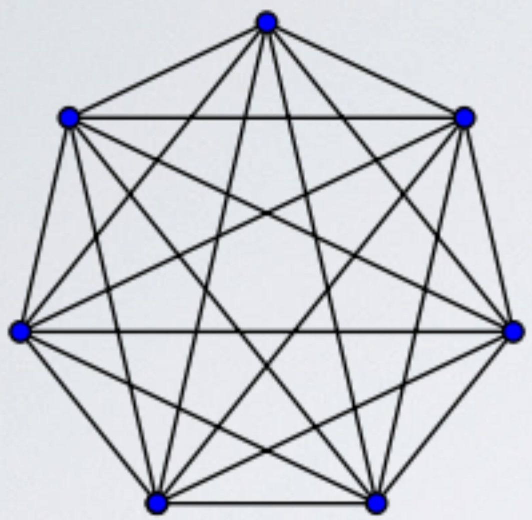
$$m_i = \tanh\left[h_j + \beta \sum_j J_{ij} m_j\right]$$

Thouless-Anderson-Palmer

$$m_i = \tanh\left[h + \sum_j \beta J_{ij} m_j - \underbrace{\beta^2 \sum_j J_{ij}^2 (1 - m_j^2) m_i}_{\text{Onsager term}}\right]$$

Onsager term

THE SHERRINGTON-KIRKPATRICK MODEL



$$H_N(\sigma, J) = -\frac{1}{\sqrt{N}} \sum_{(i,j)} J_{ij}\sigma_i\sigma_j - h\sum_i \sigma_i,$$

Ising model with disorder

Naive mean-field

$$m_i = \tanh\left[h_j + \beta \sum_j J_{ij} m_j\right]$$

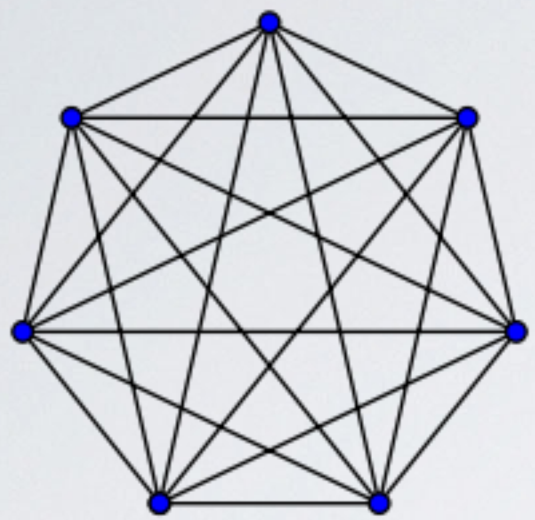
Thouless-Anderson-Palmer

$$m_i = \tanh\left[h + \sum_j \beta J_{ij} m_j - \beta^2 \sum_j J_{ij}^2 (1 - m_j^2) m_i\right]$$

Approximate Message Passing

$$m_i = \tanh\left[h + \sum_j \beta J_{ij} m_j - \beta^2 \sum_j J_{ij}^2 (1 - m_j^2) m_i\right]$$

THE SHERRINGTON-KIRKPATRICK MODEL



$$H_N(\sigma, J) = -\frac{1}{\sqrt{N}} \sum_{(i,j)} J_{ij}\sigma_i\sigma_j - h \sum_i \sigma_i,$$

Ising model with disorder

Naive mean-field

$$m_i = \tanh\left[h_j + \beta \sum_j J_{ij} m_j\right]$$

Thouless-Anderson-Palmer

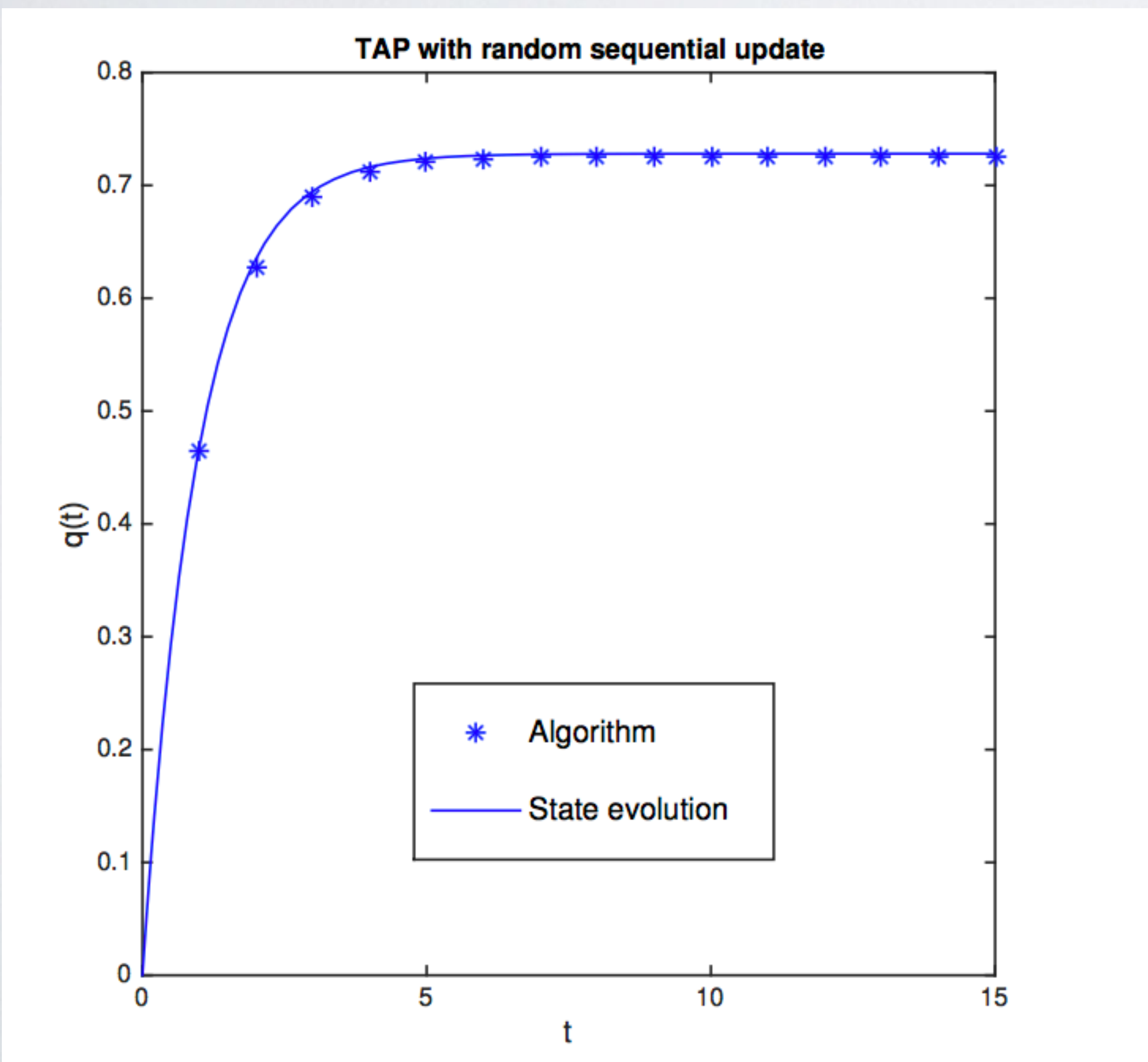
$$m_i = \tanh\left[h + \sum_j \beta J_{ij} m_j - \beta^2 \sum_j J_{ij}^2 (1 - m_j^2) m_i\right]$$

Approximate Message Passing

$$m_i^{t+1} = \tanh\left[h + \sum_j \beta J_{ij} m_j^{t} - \beta^2 \sum_j J_{ij}^2 (1 - (m_j^{t})^2) m_i^{t-1}\right]$$

AMP FOLLOWS THE REPLICA FREE ENERGY

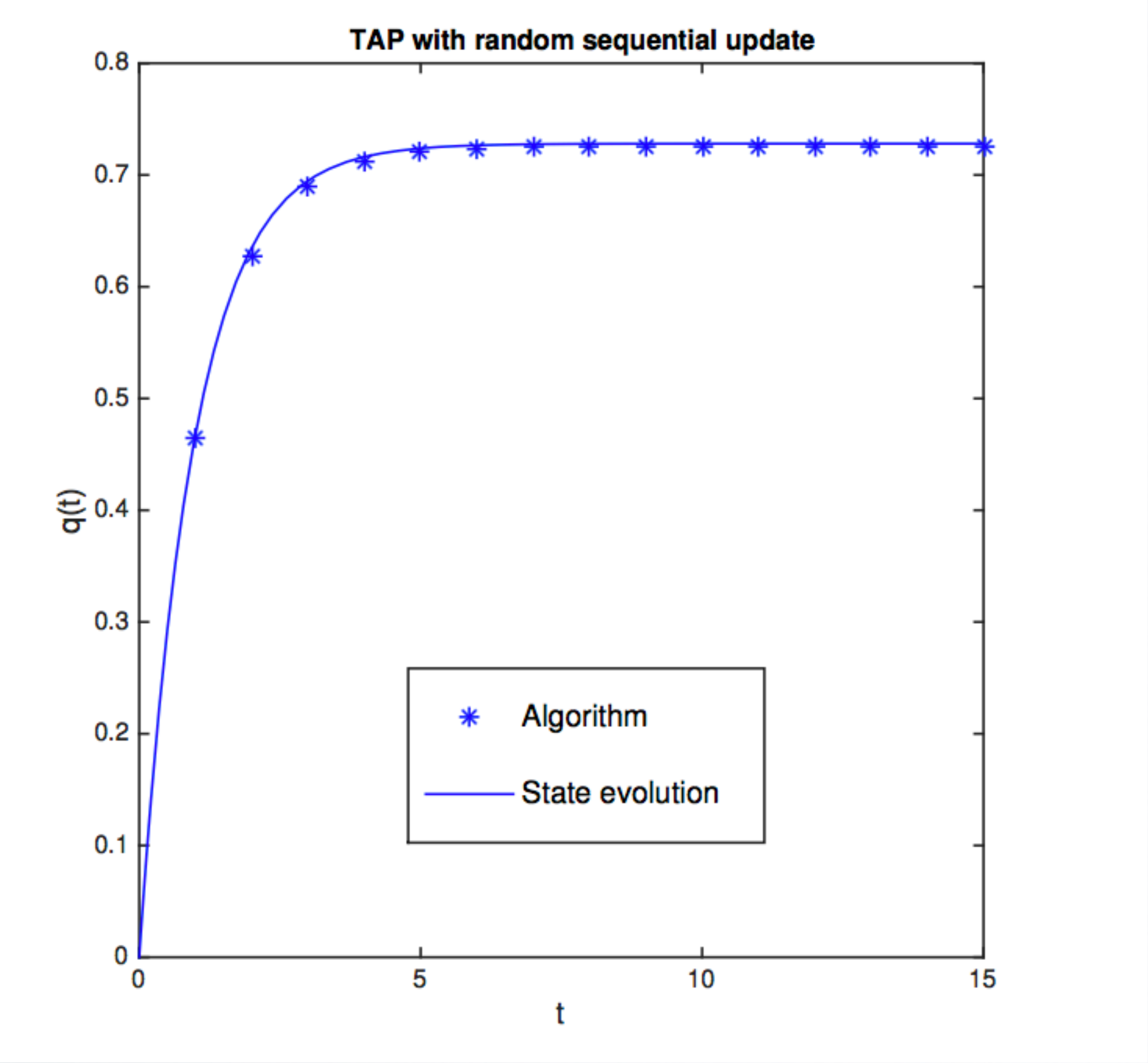
$$m_i^{t+1} = \tanh\left[h + \sum_j \beta J_{ij} m_j^{t} - \beta^2 \sum_j J_{ij}^2 (1 - (m_j^{t})^2) m_i^{t-1}\right]$$



AMP FOLLOWS THE REPLICA FREE ENERGY

$$m_i^{t+1} = \tanh\left[h + \sum_j \beta J_{ij} m_j^{t} - \beta^2 \sum_j J_{ij}^2 (1 - m_j^{2\,t}) m_i^{t-1}\right]$$

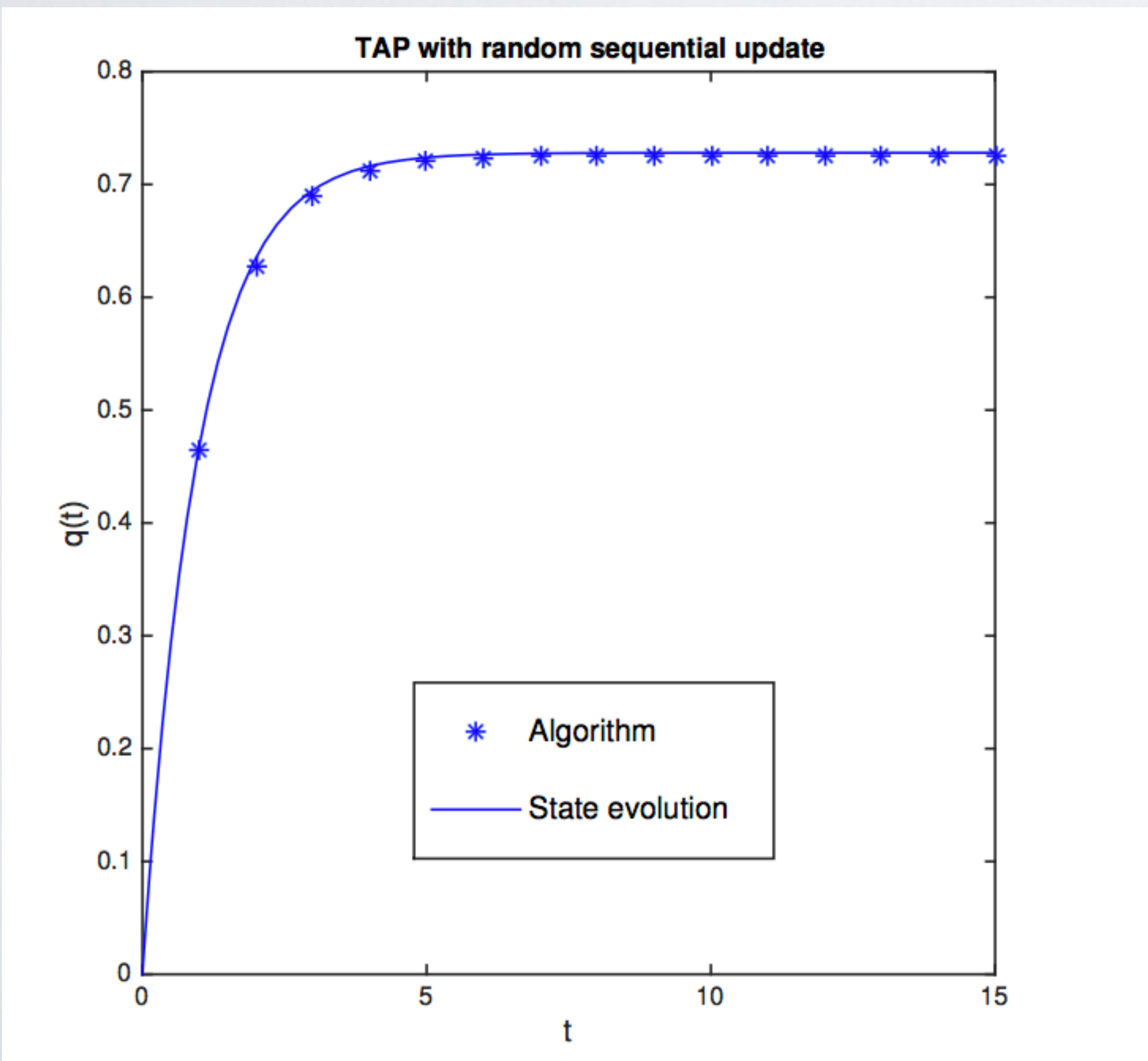
$$f = \min f_{\mathrm{SK}}(q) \qquad \beta f_{\mathrm{SK}} = -\frac{(\beta J)^2}{4(1-q)^2} - \frac{1}{\sqrt{2\pi}} \int dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} \log 2\cosh\left(\beta z J \sqrt{q} + \beta h\right)$$



AMP FOLLOWS THE REPLICA FREE ENERGY

$$m_i^{t+1} = \tanh\left[h + \sum_j \beta J_{ij} m_j^{t} - \beta^2 \sum_j J_{ij}^2 (1 - (m_j^{t})^2) m_i^{t-1}\right]$$

$$f = \min f_{\mathrm{SK}}(q) \qquad \beta f_{\mathrm{SK}} = -\frac{(\beta J)^2}{4(1-q)^2} - \frac{1}{\sqrt{2\pi}} \int dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} \log 2\cosh\left(\beta z J \sqrt{q} + \beta h\right)$$



$$q^{t+1} = \int dz \frac{e^{-z^2/2}}{\sqrt{2\pi}} \tanh\left(\beta z J \sqrt{q^t} + \beta h\right)$$

THE SHERRINGTON-KIRKPATRICK MODEL

)

AMP FOR TEACHER-STUDENT

Distribution available on github.com/sphinxteam/GLMStructuredInput

Algorithm 2 Generalized Approximate Message Passing (G-AMP)

Input: $\mathbf{y}$

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

repeat

AMP Update of ω_μ, V_μ

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

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

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

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

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

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

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

Onsager terms

Simple to implement, only matrix multiplications, $O(N^2)$

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

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

WHY DO WE ♥ AMP? STATE EVOLUTION

$N, M \to \infty, \alpha \equiv M/N = O(1)$

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

m^t in the AMP algorithm evolves as:

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

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

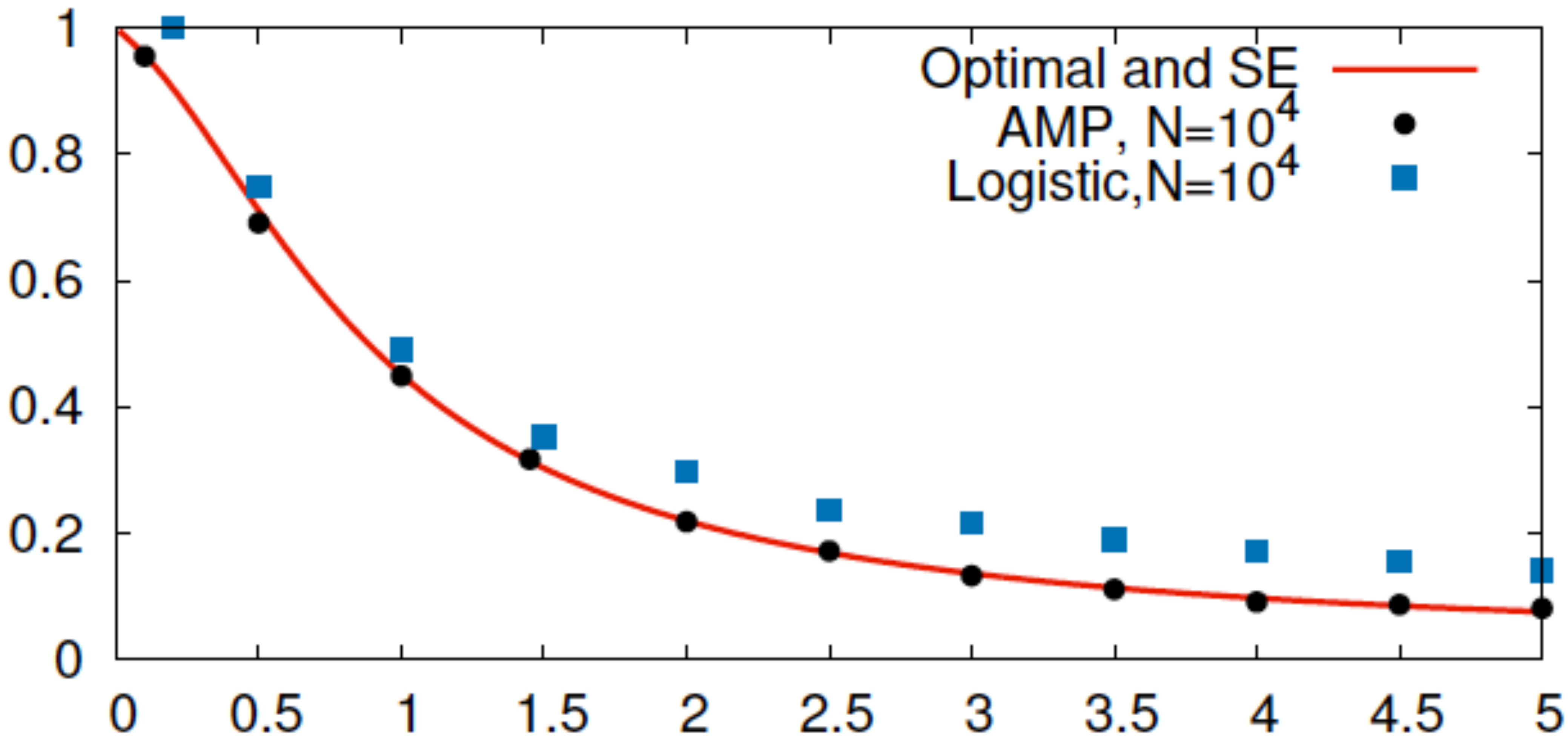
Recall the RS free energy we proved few slides ago?

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

→ AMP is doing a "gradient" descent in the replica free energy

AMP vs Optimal learning

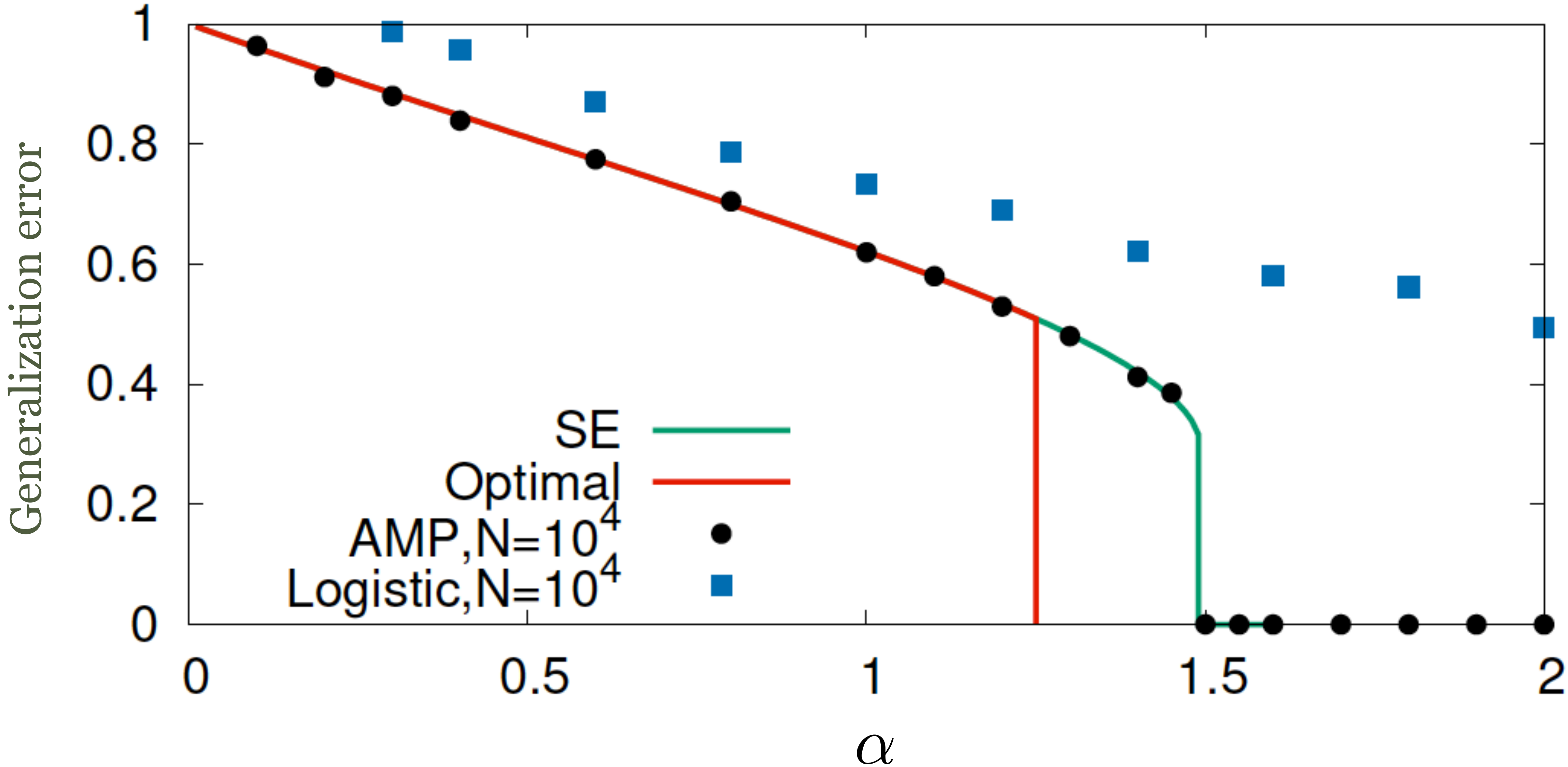
Real value case



α

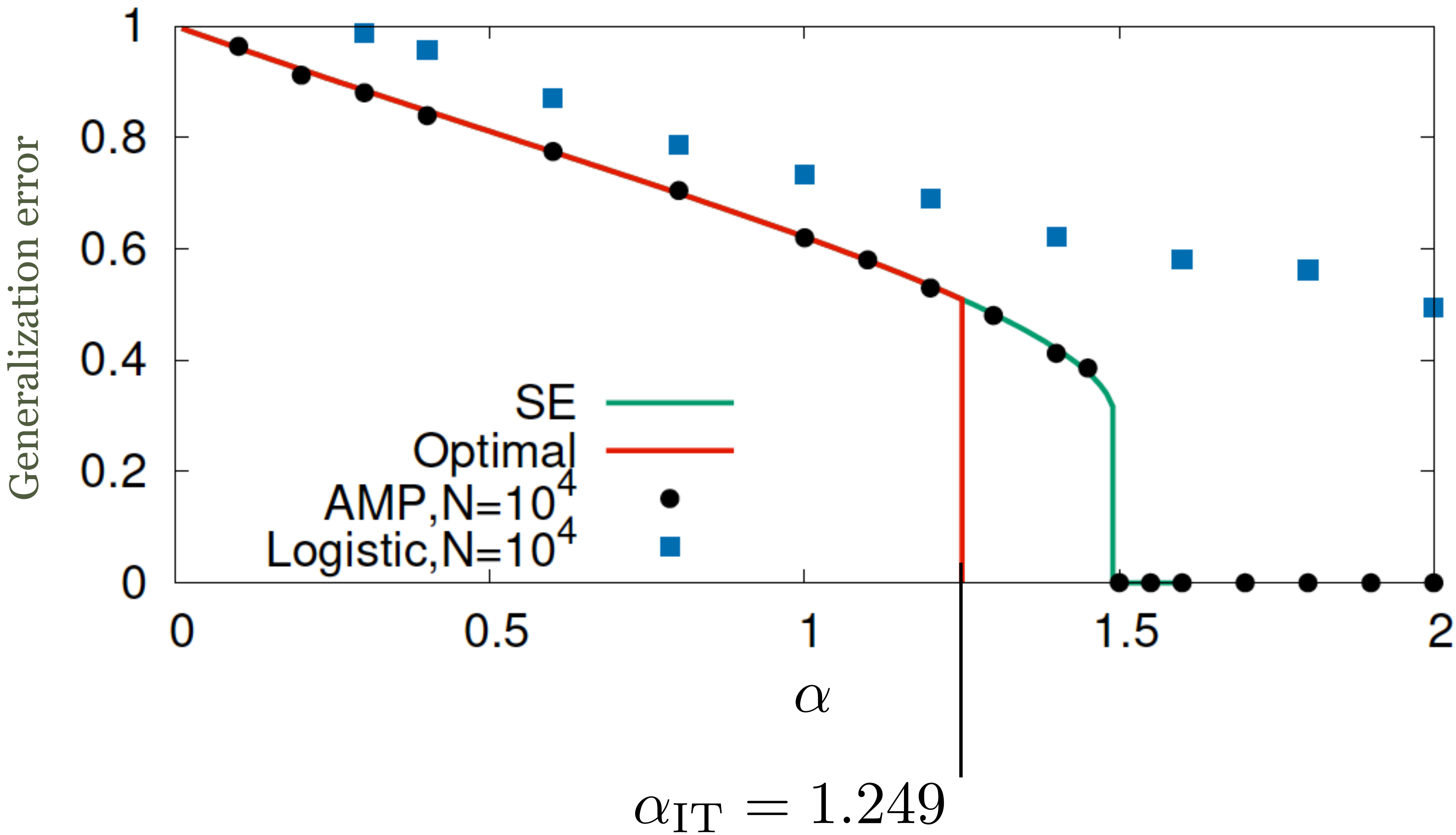
AMP vs Optimal learning

Binary case



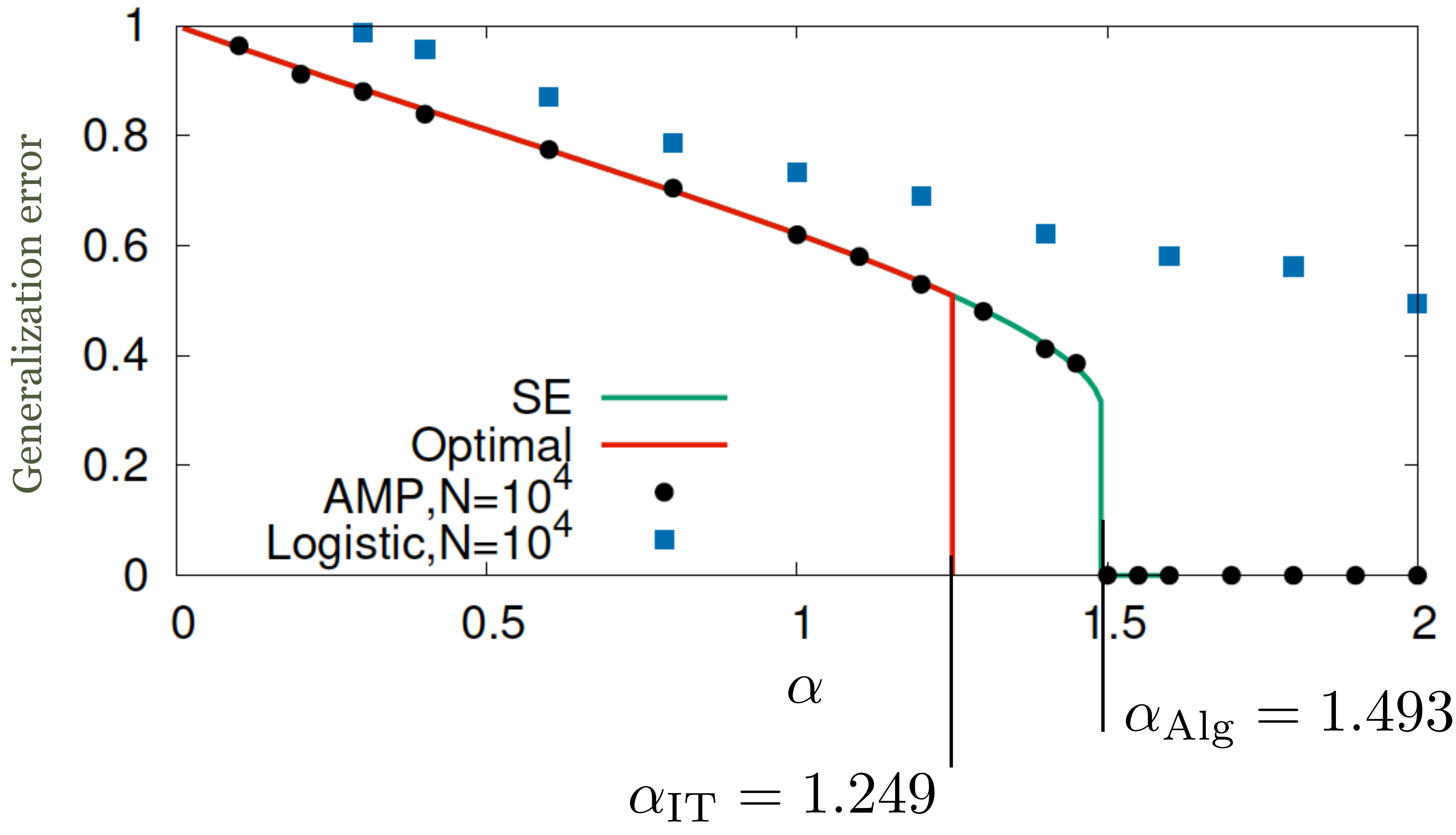
AMP vs Optimal learning

Binary case



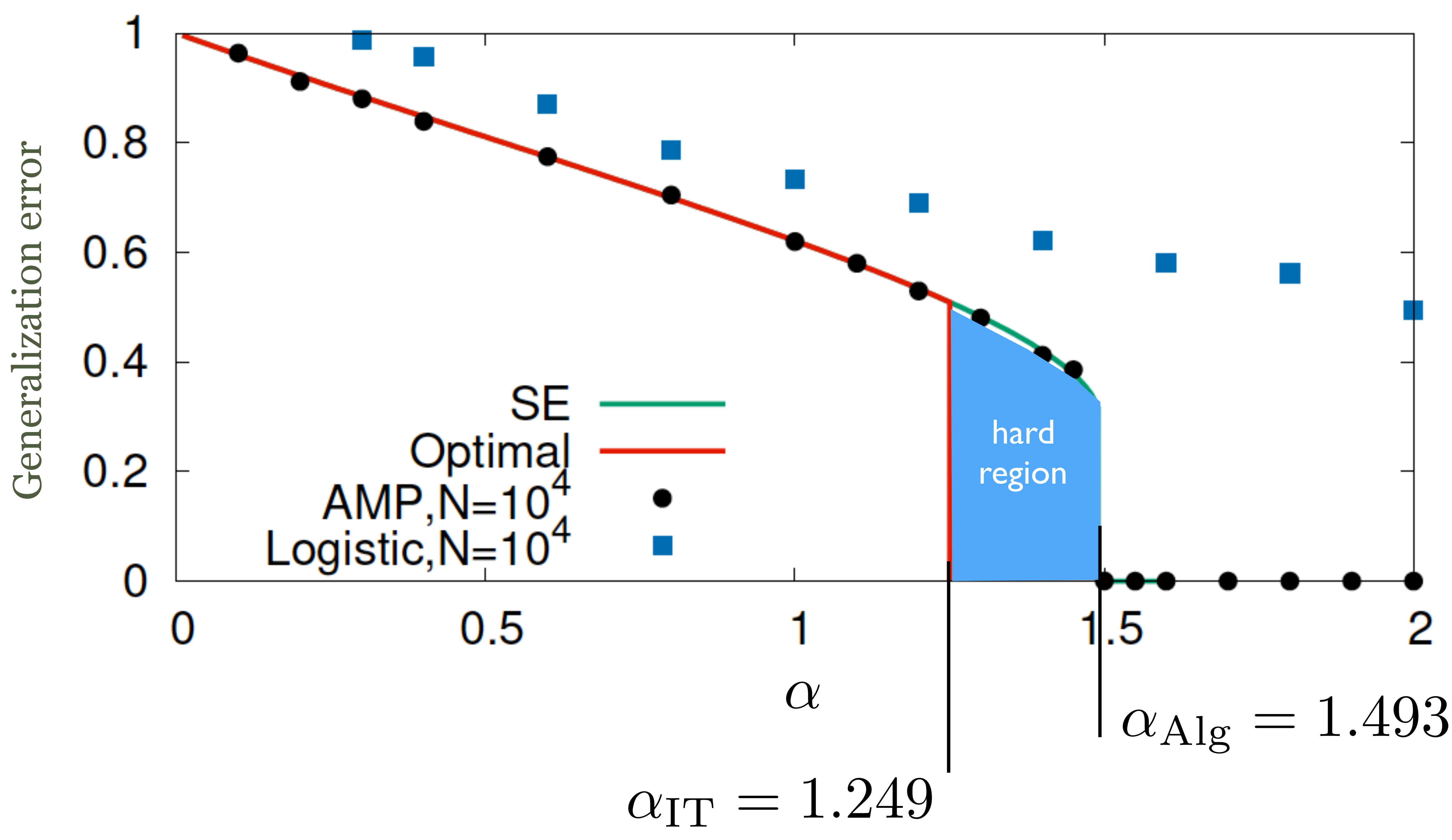
AMP vs Optimal learning

Binary case



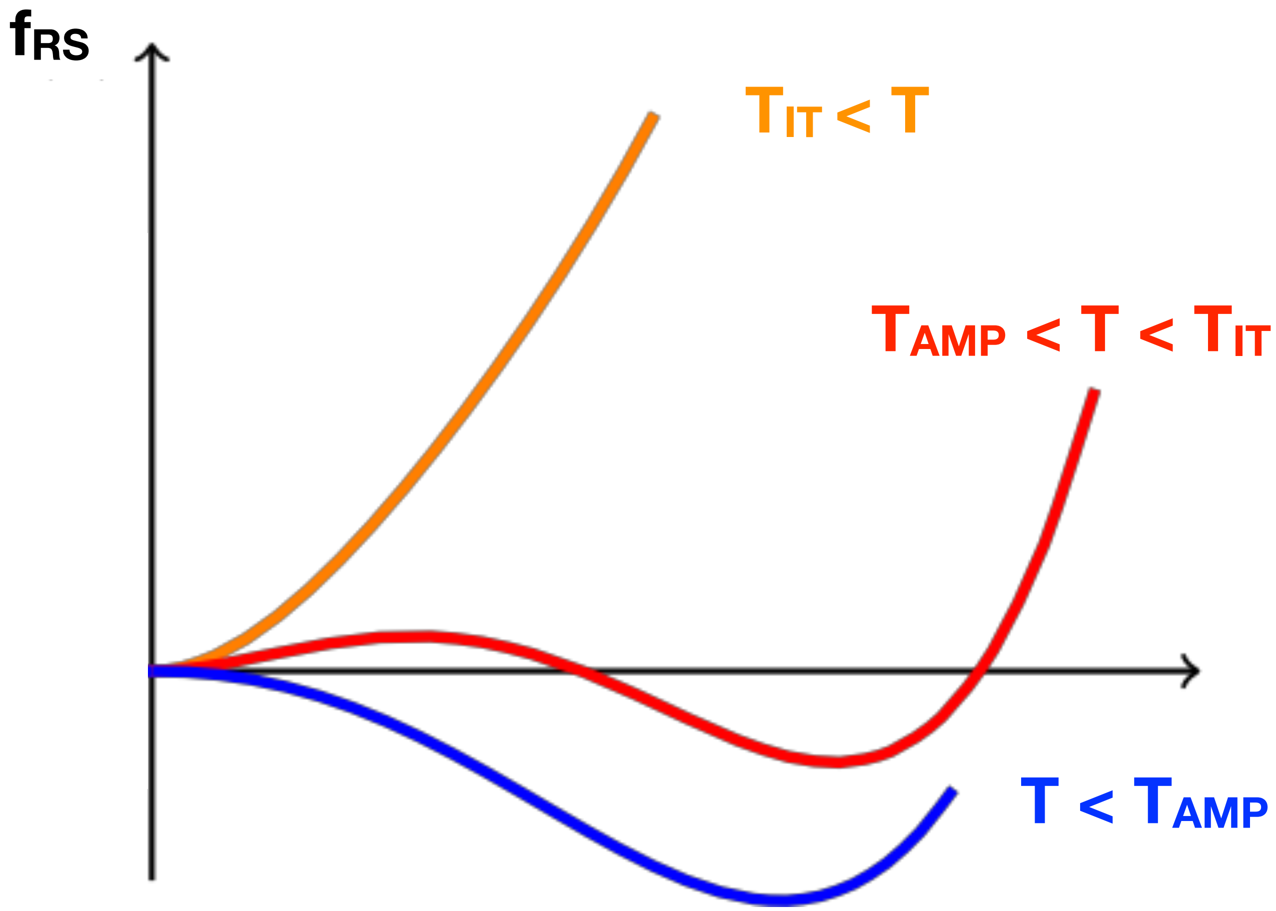
AMP vs Optimal learning

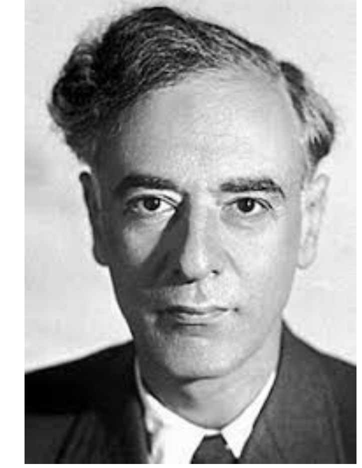
Binary case



FIRST ORDER AND THE HARD PHASE!

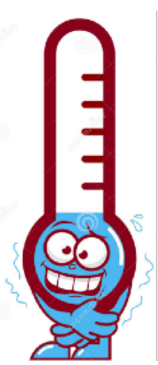
Johannes Diderik van der Waals





Lev Landeau

EASY ***HARD*** ***IMPOSSIBLE***



low temperature
more data

high temperature
Less data

Hard phases everywhere!

Identified in probabilistic models for:

- stochastic block model
- dense planted sub-matrix;
- low-rank tensor completion;
- compressed sensing;
- planted constraint satisfaction;
- Gaussian mixture clustering;
- low-density parity check error correcting codes;
- sparse principal component analysis;
- generalised linear regression;
- dictionary learning;
- blind source separation;
- learning in binary perceptron;
- phase retrieval; ...

Journal
Advances in Physics ›
Volume 65, 2016 - Issue 5

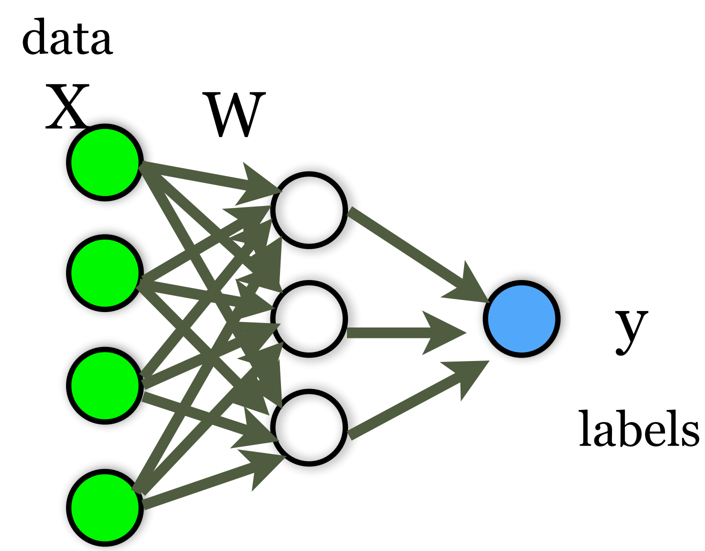
Statistical physics of inference: thresholds and algorithms

Lenka Zdeborová & Florent Krzakala

Multi-Layer Neural Nets

The Committee machine

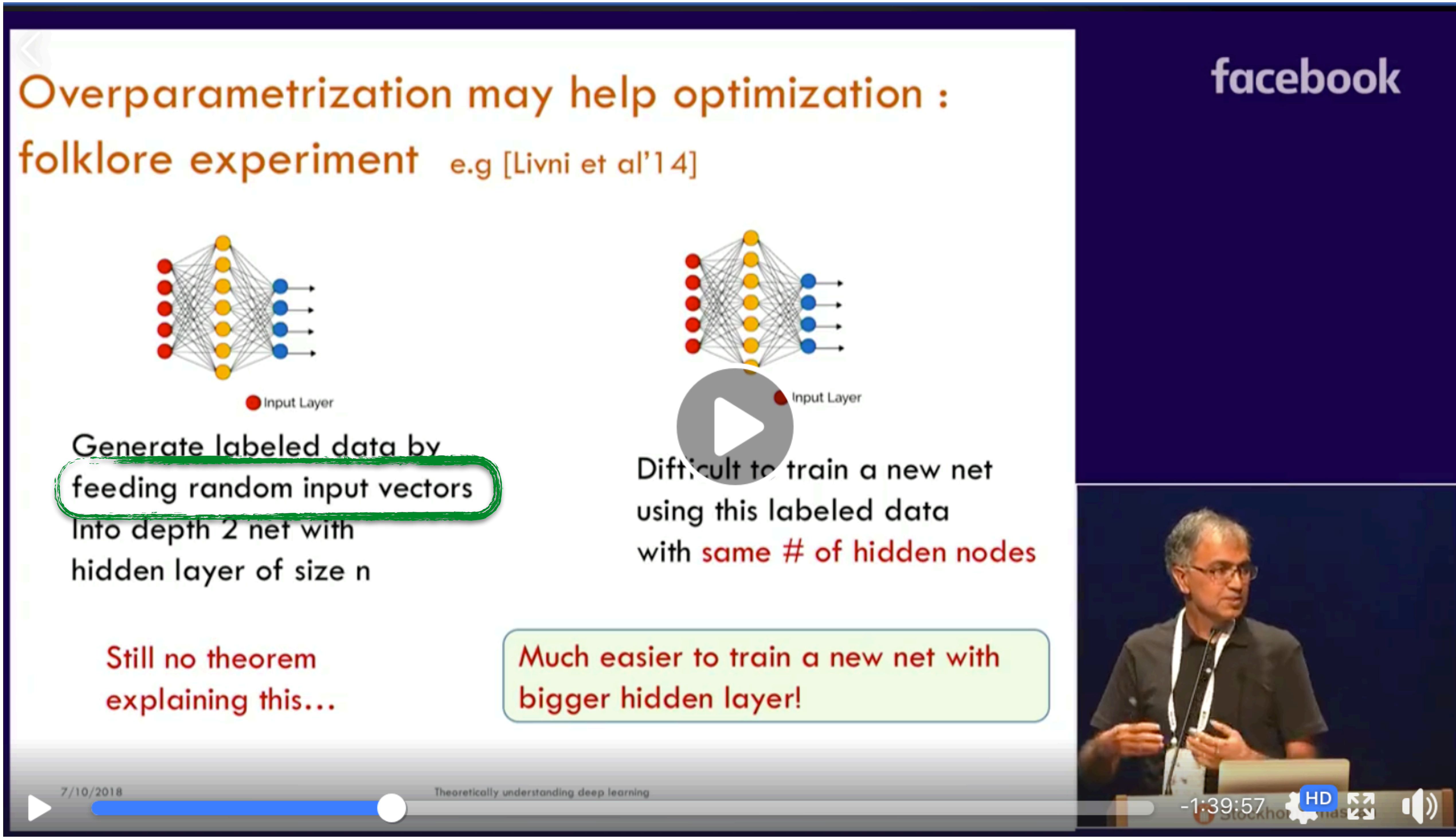
- P input units
- K hidden units
- output unit
- N training samples



$$K = O(1)$$

Limit: $\alpha = \frac{N}{P} = O(1)$ $\qquad K = O(1) \qquad N, P \to \infty$

Sanjeev Arora at ICML'18: Tutorial on theory of deep learning.



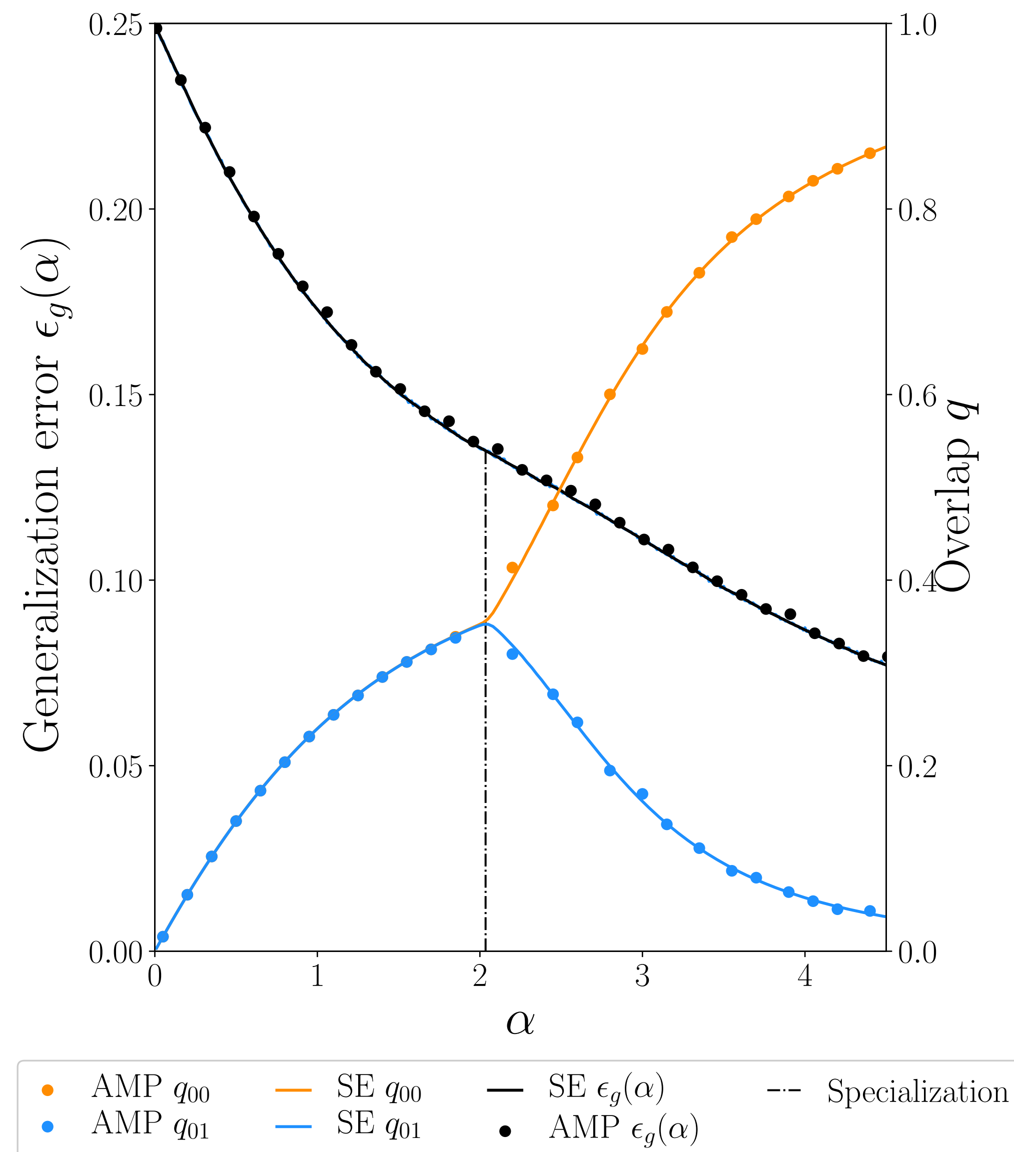
Optimal generalization

K=2

$$y_\mu = \mathrm{sign}\Big[\mathrm{sign}\big(\sum_i F_{\mu,i} x_{i,1}\big) + \mathrm{sign}\sum_i \big(F_{\mu,i} x_{i,2}\big)\Big]$$

$$\mathrm{sign}(0) = 0$$

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



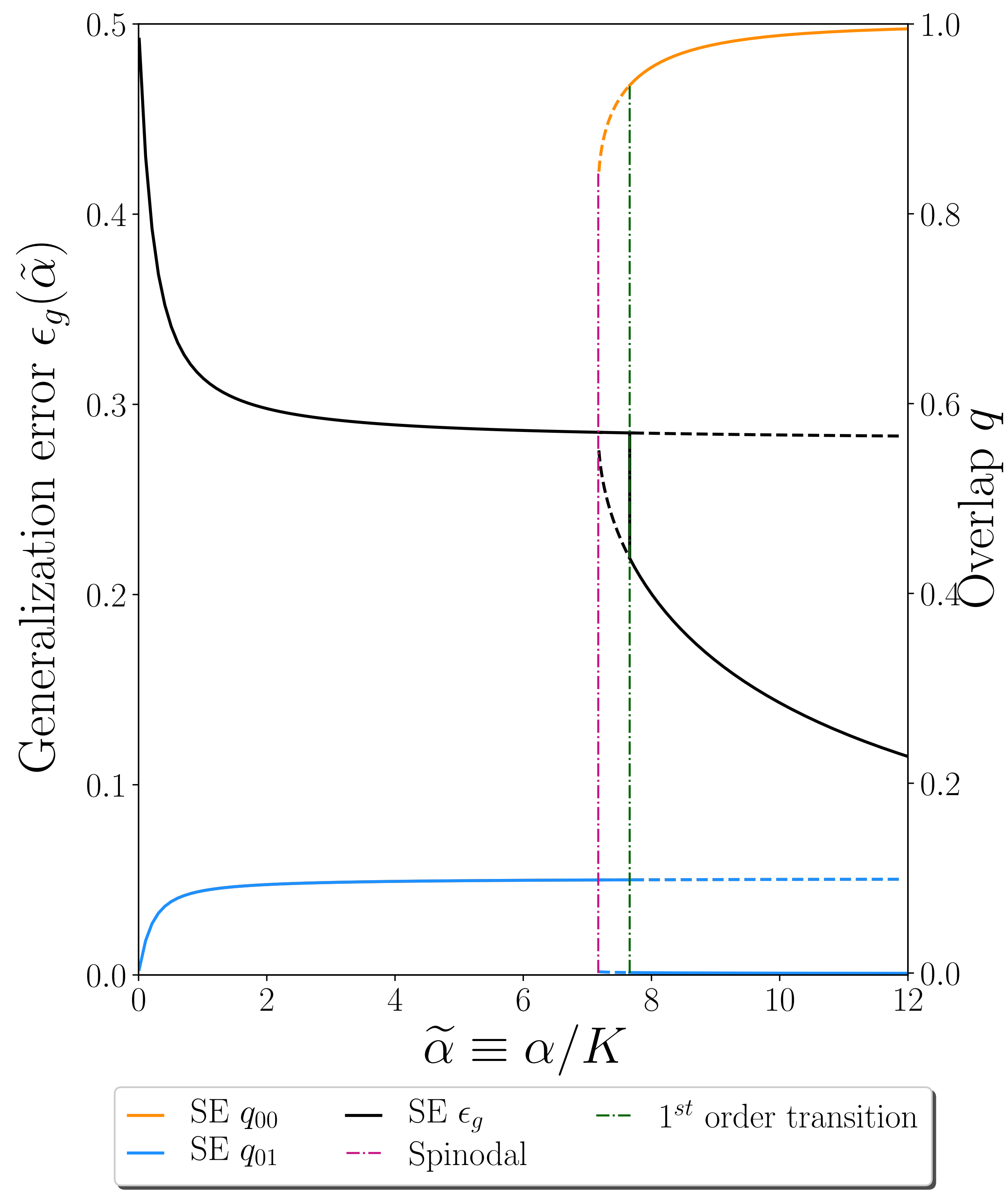
Optimal generalization

$N \gg K \gg 1$

$$y_\mu = \mathrm{sign}\Big[\sum_{l=1}^{K} \mathrm{sign}\big(\sum_{i=1}^{p} F_{\mu i} x^*_{il}\big)\Big]$$

- Specialization phase transition
- First-order threshold:

$$N > 7.65KP$$



Optimal generalization

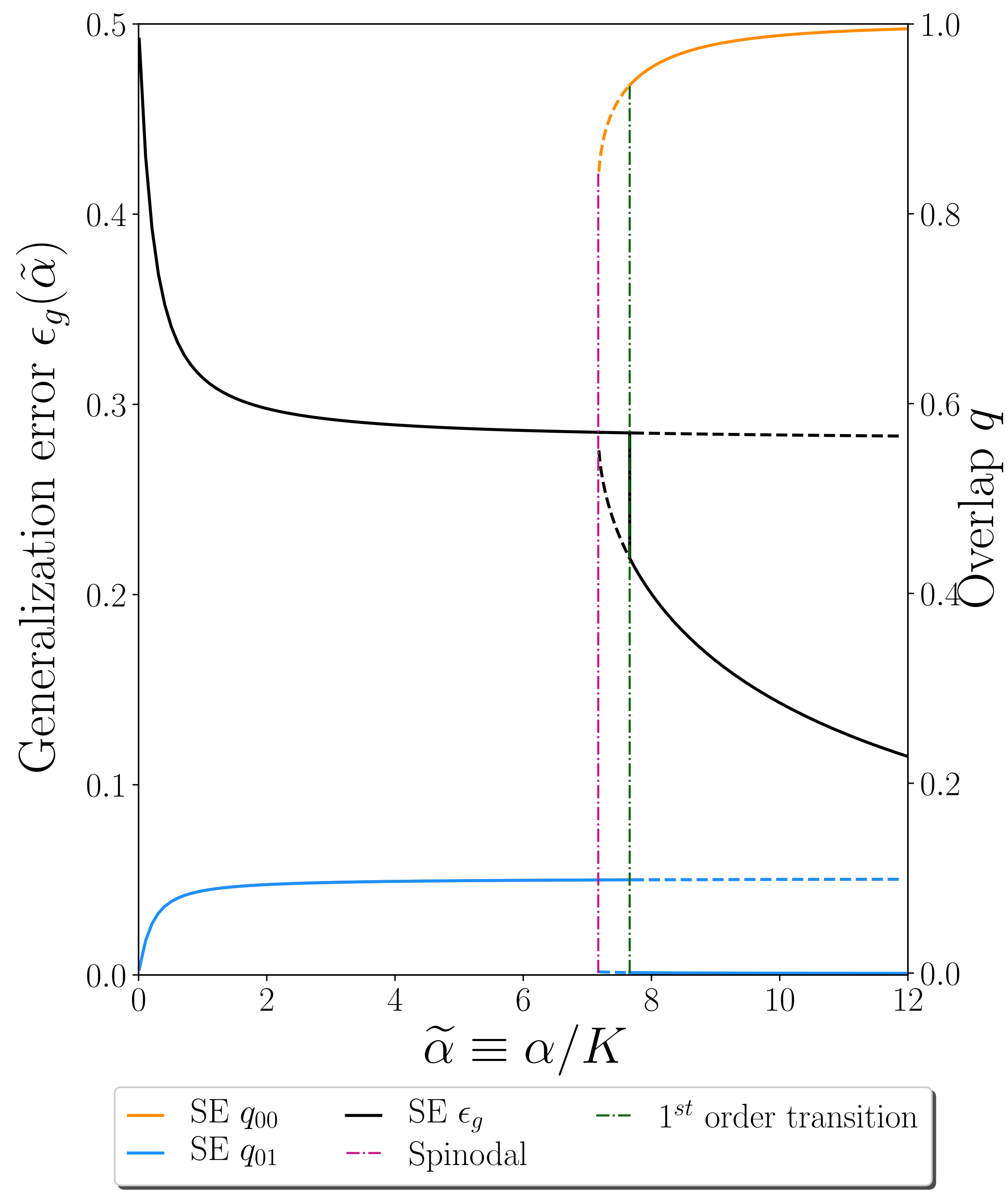
$N \gg K \gg 1$

$$y_\mu = \mathrm{sign}\Big[\sum_{l=1}^{K} \mathrm{sign}\big(\sum_{i=1}^{p} F_{\mu i} x^*_{il}\big)\Big]$$

- Specialization phase transition
- First-order threshold:

$$N > 7.65KP$$

Capacity: $d_{\mathrm{Gardner}} \geq CstPK\sqrt{\log K}$

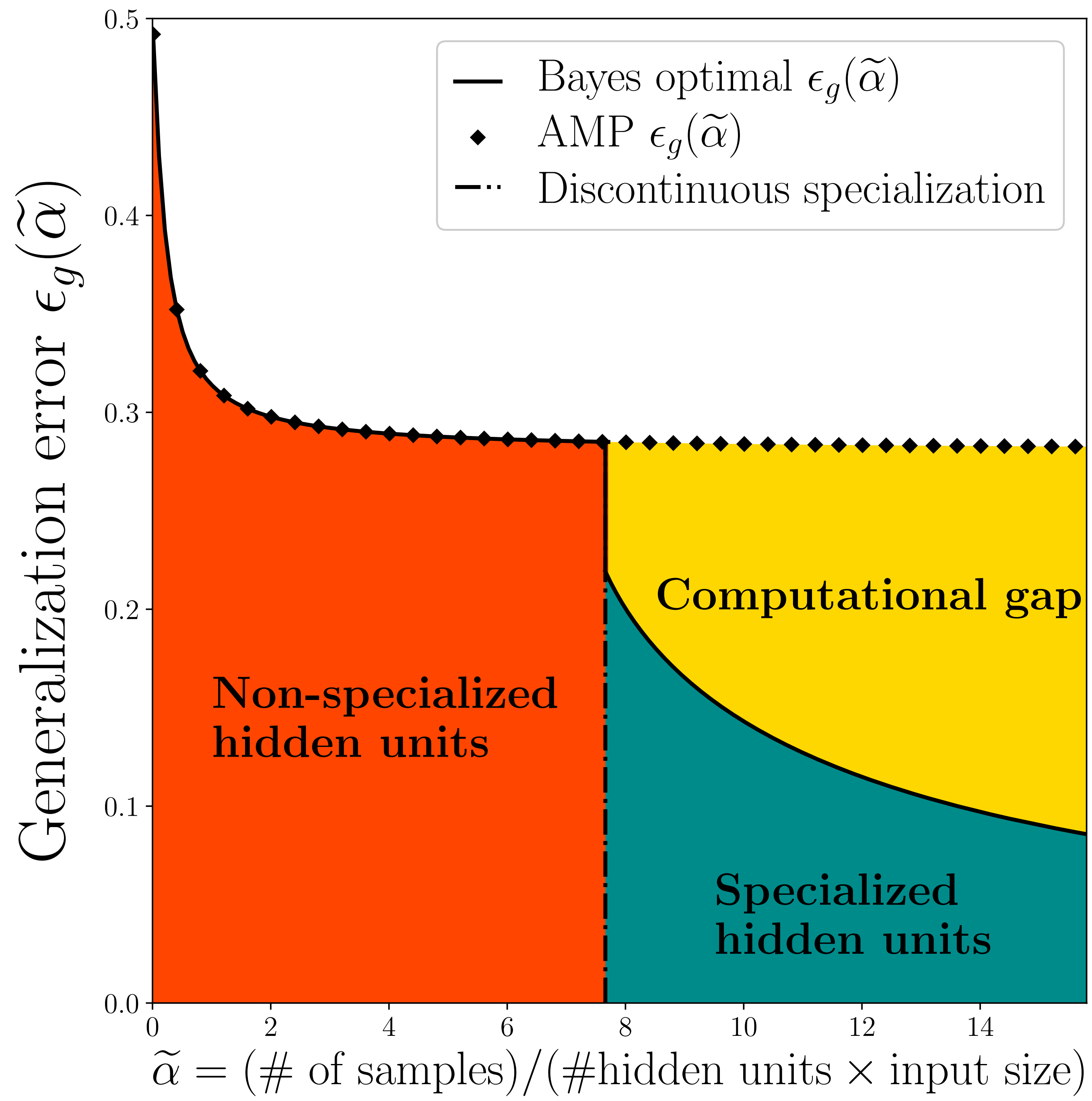


Optimal generalization

$K \gg 1$

$$y_\mu = \text{sign}\Big[\sum_{l=1}^{K} \text{sign}\big(\sum_{i=1}^{p} F_{\mu i} x^*_{il}\big)\Big]$$

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



Committee machine

Very large gap between typical and worst case!

No good learning

Good “typical” performances

Good “worst case” performances

$\alpha = \frac{N}{P}$

$\alpha_{\mathrm{IT}} = O(PK)$

$\alpha_{\mathrm{IT}} = O(K\sqrt{\log(K)})$

Committee machine

Very large gap between typical and worst case!

Online learning with SGD

4

Generalisation dynamics of online learning in over-parameterised neural networks

Sebastian Goldt[1], Madhu S. Advani[2], Andrew M. Saxe[3], Florent Krzakala[4] and Lenka Zdeborová[1]

arxiv:1901.09085

Gradient-Descent, one sample at a time…

At each time, minimize: $$E(\{W\}, \mathbf{x}_i) = \frac{1}{2}(\phi(W, \mathbf{x}_i) - y_i)^2$$

Weight decay

Learning rate

$$W_k^{t+1} = W_k^t - \frac{\kappa}{P} W_k^t - \frac{\eta}{\sqrt{P}} \nabla E(W) \big|_{(\mathbf{x}^t, y^t)}$$

Stochastic Gradient

Gradient-Descent, one sample at a time…

$$W_k^{t+1} = W_k^t - \frac{\kappa}{P} W_k^t - \frac{\eta}{\sqrt{P}} \nabla E(W) \big|_{(\mathbf{x}^t, y^t)}$$

Can be analysed efficiently in teacher/student setting
by a ordinary differential equations in the teacher/student case

Single Layer

W. Kinzel and P. Rujan ’90
C.W.H.Mace & A.C.C.Coolen ‘98
E. Oja and J. Karhunen ’85
…
Wang & Lu ‘16

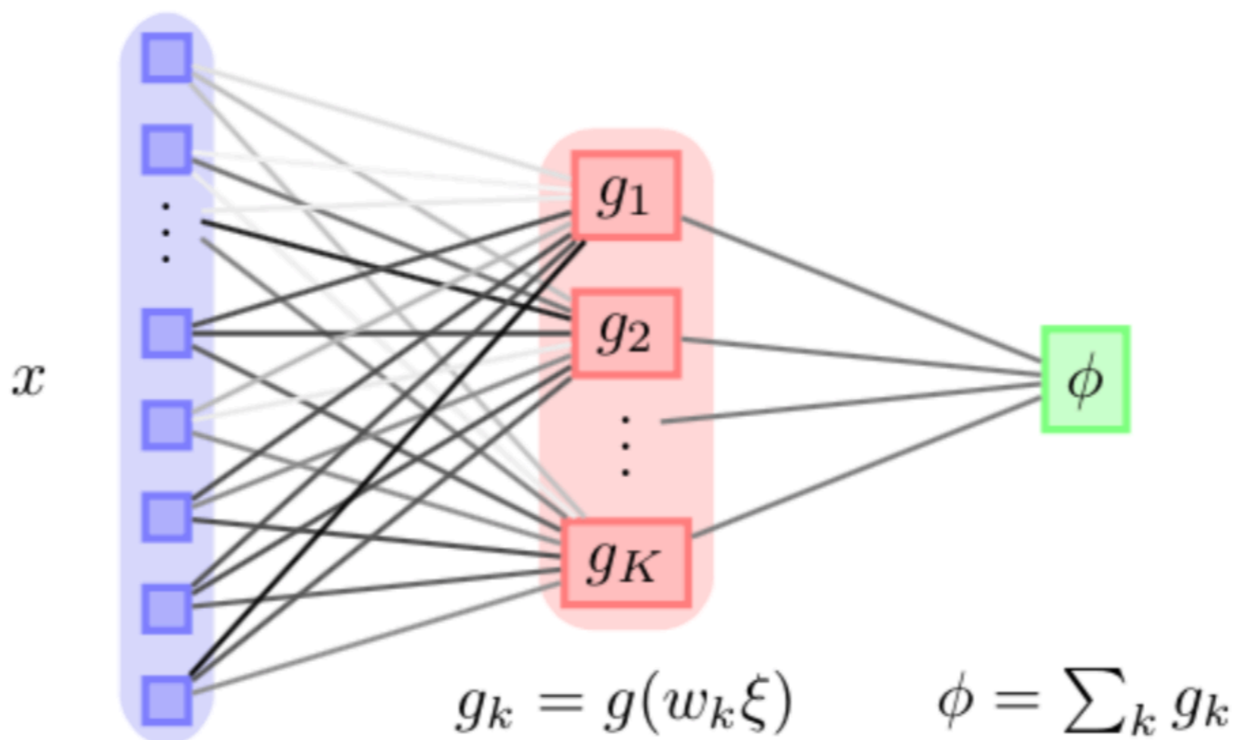
Multi-Layer

M. Biehl and H. Schwarze '95
Saad and S.A. Solla ’95
…
Last talk by Andrea (Different setting…)

The committee machine

Teacher: $\phi(B,\mathbf{x}) = \sum_{m=1}^{K} g\left(\frac{\mathbf{B_m x}}{\sqrt{P}}\right) + \sqrt{\sigma}\xi$

***Student*:** $\phi(B,\mathbf{x}) = \sum_{k=1}^{M} g\left(\frac{\mathbf{W_k x}}{\sqrt{P}}\right)$



$$w_k^{\mu+1} = w_k^{\mu} - \frac{\kappa}{P} w_k^{\mu} - \frac{\eta}{\sqrt{P}} x^{\mu} \delta_k^{\mu}$$

where

$$\delta_k^{\mu} \equiv g'(\lambda_k^{\mu}) \left[\phi(w, x^{\mu}) - y_B^{\mu}\right]$$

and we have defined $\lambda_k^{\mu} \equiv w_k x^{\mu} / \sqrt{P}$.

$$\epsilon_g = \frac{1}{2} \mathbb{E}_{\mathbf{x}_{\text{new}}} \left(\sum_{m=1}^{M} g\left(\frac{\mathbf{B_m x}_{\text{new}}}{\sqrt{P}}\right) - \sum_{k=1}^{K} g\left(\frac{\mathbf{W_k x}_{\text{new}}}{\sqrt{P}}\right) \right)^2$$

The committee machine

Teacher: $\phi(B, \mathbf{x}) = \sum_{m=1}^{K} g\left(\frac{\mathbf{B_m x}}{\sqrt{P}}\right) + \sqrt{\sigma}\xi$

Student: $\phi(B, \mathbf{x}) = \sum_{k=1}^{M} g\left(\frac{\mathbf{W_k x}}{\sqrt{P}}\right)$

$$\frac{\mathrm{d}R_{in}}{\mathrm{d}\alpha} = -\kappa R_{in} + \eta\langle\delta_i \nu_n\rangle$$

$$\frac{\mathrm{d}Q_{ik}}{\mathrm{d}\alpha} = -2\kappa Q_{ik} + \eta\langle\delta_i \lambda_k\rangle + \eta\langle\delta_k \lambda_i\rangle + \eta^2\langle\delta_i\delta_k\rangle + \eta^2\sigma^2\langle g'(\lambda_i)g'(\lambda_k)\rangle$$

$$\nu_m^\mu = \frac{B_m x^\mu}{\sqrt{P}}$$

$$\lambda_k^\mu = \frac{W_k x^\mu}{\sqrt{P}}$$

$$R_{km} = \frac{W_k B_m}{N}$$

$$Q_{kl} = \frac{W_k W_l}{N}$$

The committee machine

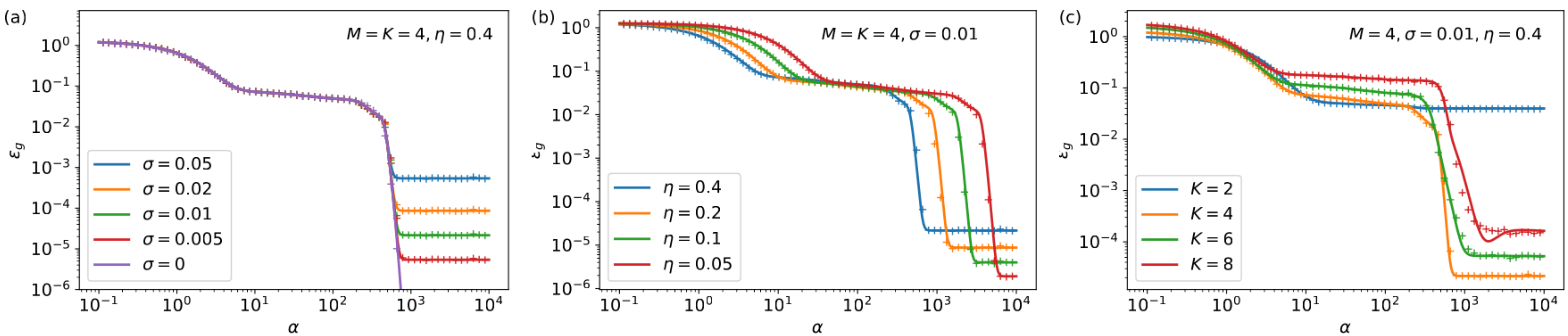


Figure 2. **The analytical description of the generalisation dynamics of sigmoidal networks (solid) matches simulations (crosses).** We show learning curves $\epsilon_g(\alpha)$ obtained by integration of the ODEs (12) (solid). From left to the right, we vary the variance of the teacher's output noise σ, the learning rate η, and the number of hidden units in the student K. For each combination of parameters shown in the plots, we ran a single simulation of a network with $N = 784$ and plot the generalisation observed (crosses). $\kappa = 0$ in all cases.

https://github.com/sgoldt/pyscm

Asymptotic generalisation error

$$\epsilon_g^* = \frac{\sigma^2 \eta}{2\pi}\left(L + \frac{M}{\sqrt{3}}\right) + \mathcal{O}(\eta^2)$$

Derived for Sigmoidal networks

Asymptotic generalisation error

$$\epsilon_g^* = \frac{\sigma^2 \eta}{2\pi} \left(L + \frac{M}{\sqrt{3}} \right) + \mathcal{O}(\eta^2)$$

Derived for Sigmoidal networks

Very robust scaling!

Asymptotic generalisation error

$$\epsilon_g^* = \frac{\sigma^2 \eta}{2\pi}\left(L + \frac{M}{\sqrt{3}}\right) + \mathcal{O}(\eta^2)$$

Derived for Sigmoidal networks

Very robust scaling!

Linear model $\epsilon_g^* = \frac{1}{4}\eta\sigma^2(L+M) + \mathcal{O}\left(\eta^2\right)$

Asymptotic generalisation error

$$\epsilon_g^* = \frac{\sigma^2 \eta}{2\pi}\left(L + \frac{M}{\sqrt{3}}\right) + \mathcal{O}(\eta^2)$$

Derived for Sigmoidal networks

Very robust scaling!

Relu models

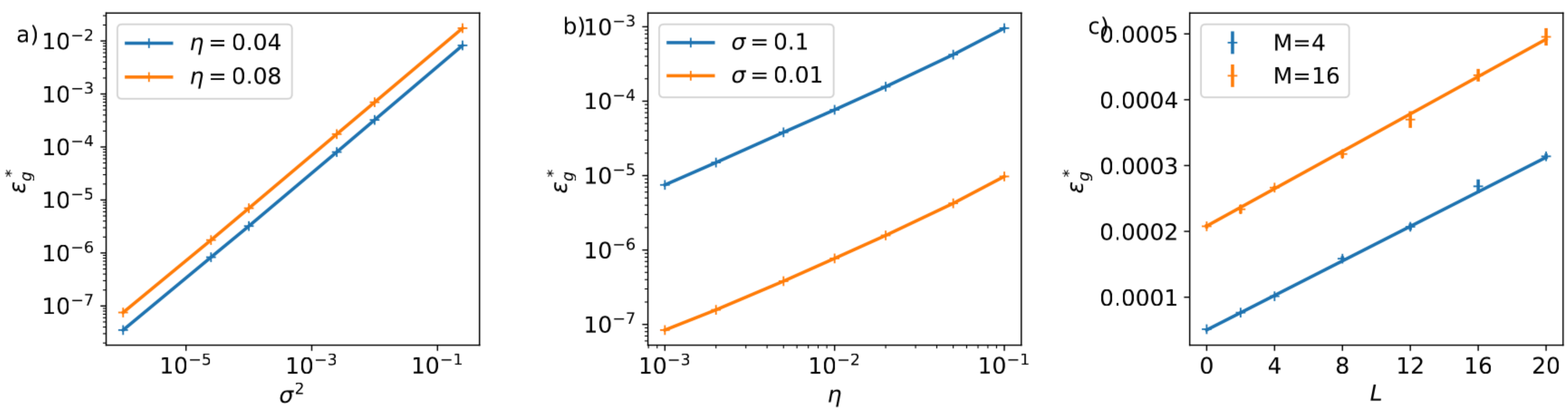


Figure 5. **The final generalisation error of over-parametrised ReLU networks scales as $\epsilon_g^* \sim \eta\sigma^2 L$.** Simulations confirm that the asymptotic generalisation error ϵ_g^* of a ReLU student learning from a ReLU teacher scales with the learning rate η, the variance of the

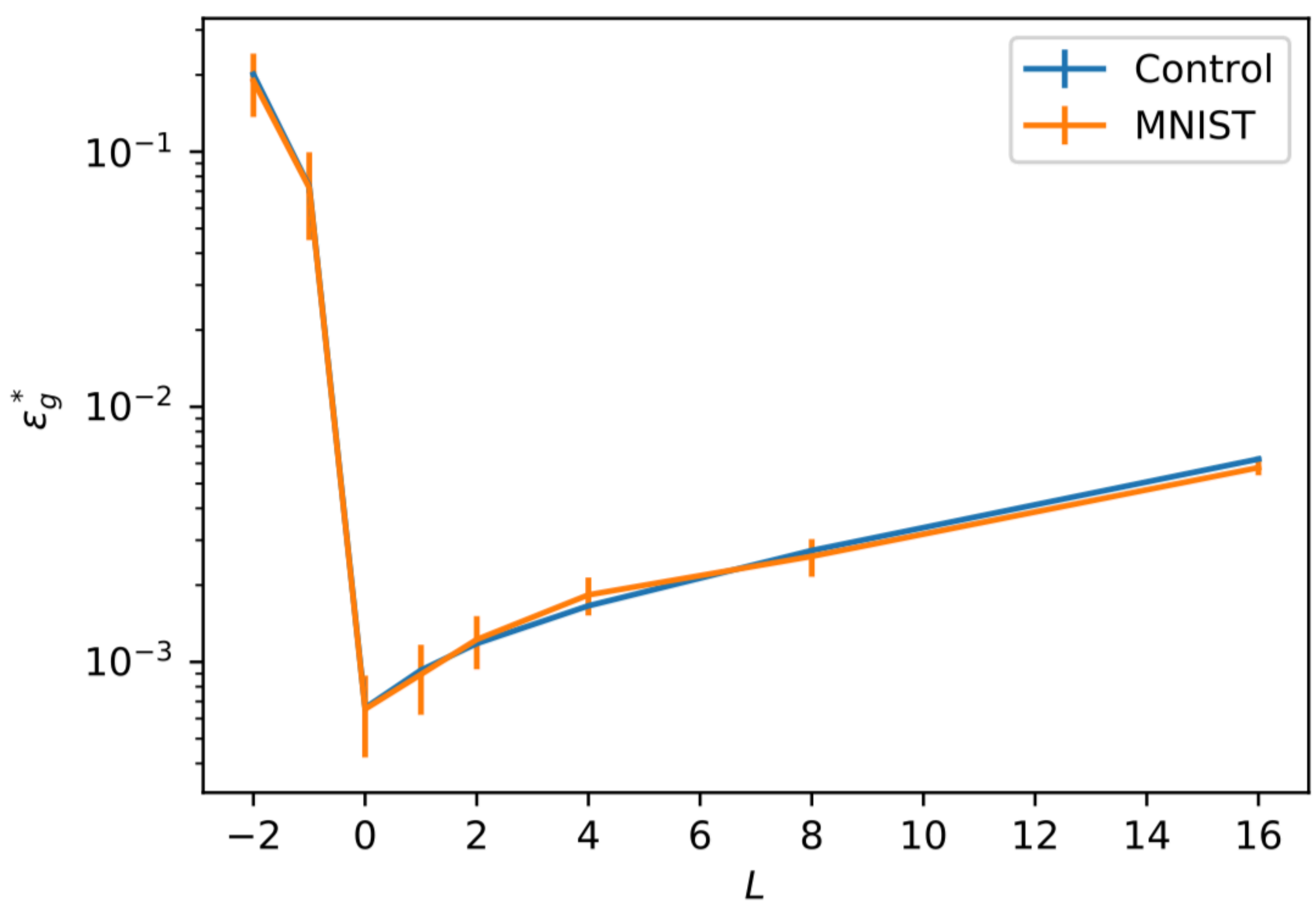
Asymptotic generalisation error

$$\epsilon_g^* = \frac{\sigma^2 \eta}{2\pi} \left(L + \frac{M}{\sqrt{3}} \right) + \mathcal{O}(\eta^2)$$

Derived for Sigmoidal networks

Very robust scaling!

Structured patterns (i.e. MNIST)
In random teacher-student



Asymptotic generalisation error

$$\epsilon_g^* = \frac{\sigma^2 \eta}{2\pi} \left(L + \frac{M}{\sqrt{3}} \right) + \mathcal{O}(\eta^2)$$

Online learning works and generalize well…

… even in the overparametrized regime!

Asymptotic generalisation error

$$\epsilon_g^* = \frac{\sigma^2 \eta}{2\pi} \left(L + \frac{M}{\sqrt{3}} \right) + \mathcal{O}(\eta^2)$$

Online learning works and generalize well…

… even in the overparametrized regime!

However !

Asymptotic generalisation error

$$\epsilon_g^* = \frac{\sigma^2 \eta}{2\pi}\left(L + \frac{M}{\sqrt{3}}\right) + \mathcal{O}(\eta^2)$$

Online learning works and generalize well…

… even in the overparametrized regime!

However ! Notice the scaling $\epsilon_g^* \propto \eta L$

Asymptotic generalisation error

$$\epsilon_g^* = \frac{\sigma^2\eta}{2\pi}\left(L + \frac{M}{\sqrt{3}}\right) + \mathcal{O}(\eta^2)$$

Online learning works and generalize well…

… even in the overparametrized regime!

However !

Notice the scaling $\epsilon_g^* \propto \eta L$

Time is samples:

Asymptotic generalisation error

$$\epsilon_g^* = \frac{\sigma^2 \eta}{2\pi} \left(L + \frac{M}{\sqrt{3}} \right) + \mathcal{O}(\eta^2)$$

Online learning works and generalize well…

… even in the overparametrized regime!

However ! Notice the scaling $\epsilon_g^* \propto \eta L$

Time is samples:

Overparametrized students require larger training set!

Asymptotic generalisation error

$$\epsilon_g^* = \frac{\sigma^2 \eta}{2\pi}\left(L + \frac{M}{\sqrt{3}}\right) + \mathcal{O}(\eta^2)$$

Online learning works and generalize well…

… even in the overparametrized regime!

However ! Notice the scaling $\epsilon_g^* \propto \eta L$

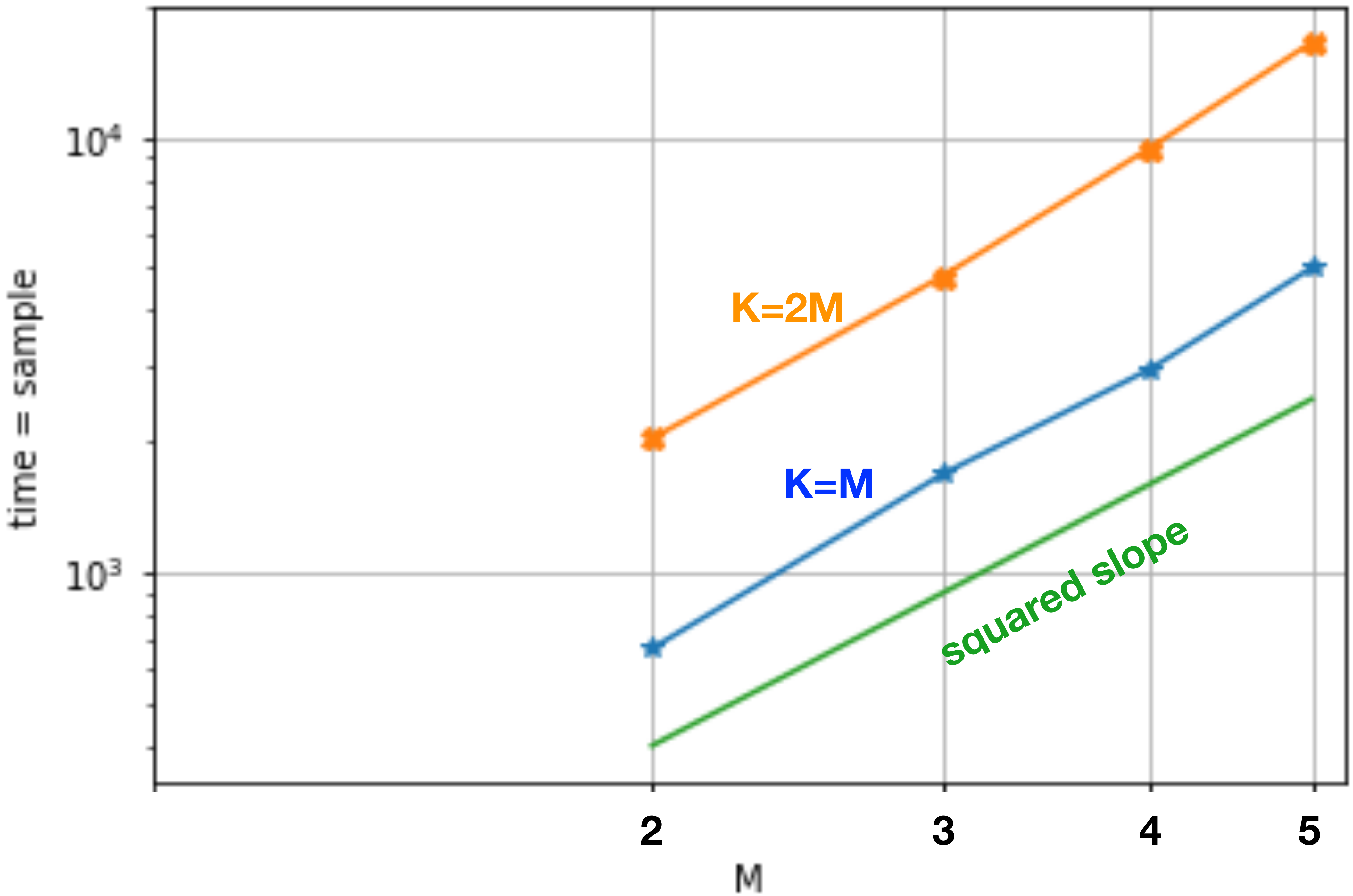
Time is samples:

Overparametrized students require larger training set!

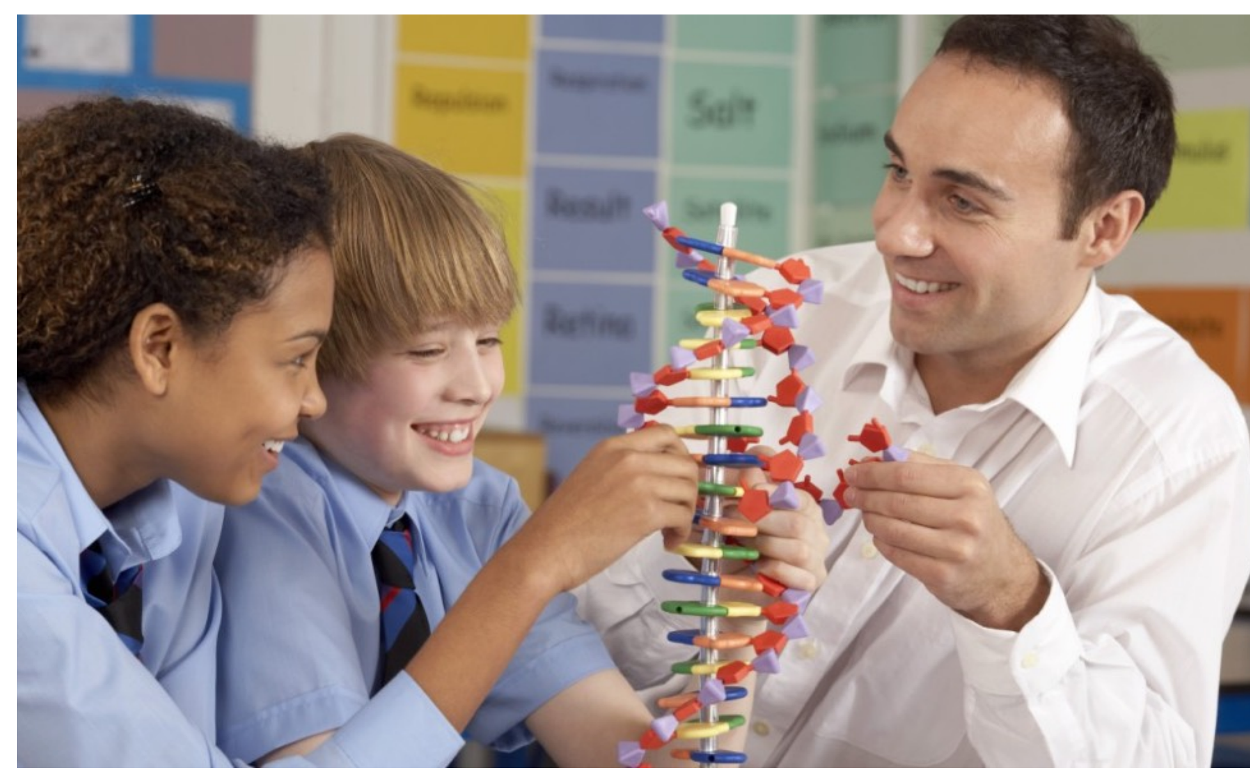
Online SGD does not perform magic

The hard phase is still hard!

time = samples needed to converge grows as PK^2, just as for AMP
Over-parametrization and SGD do not perform magic in the hard region



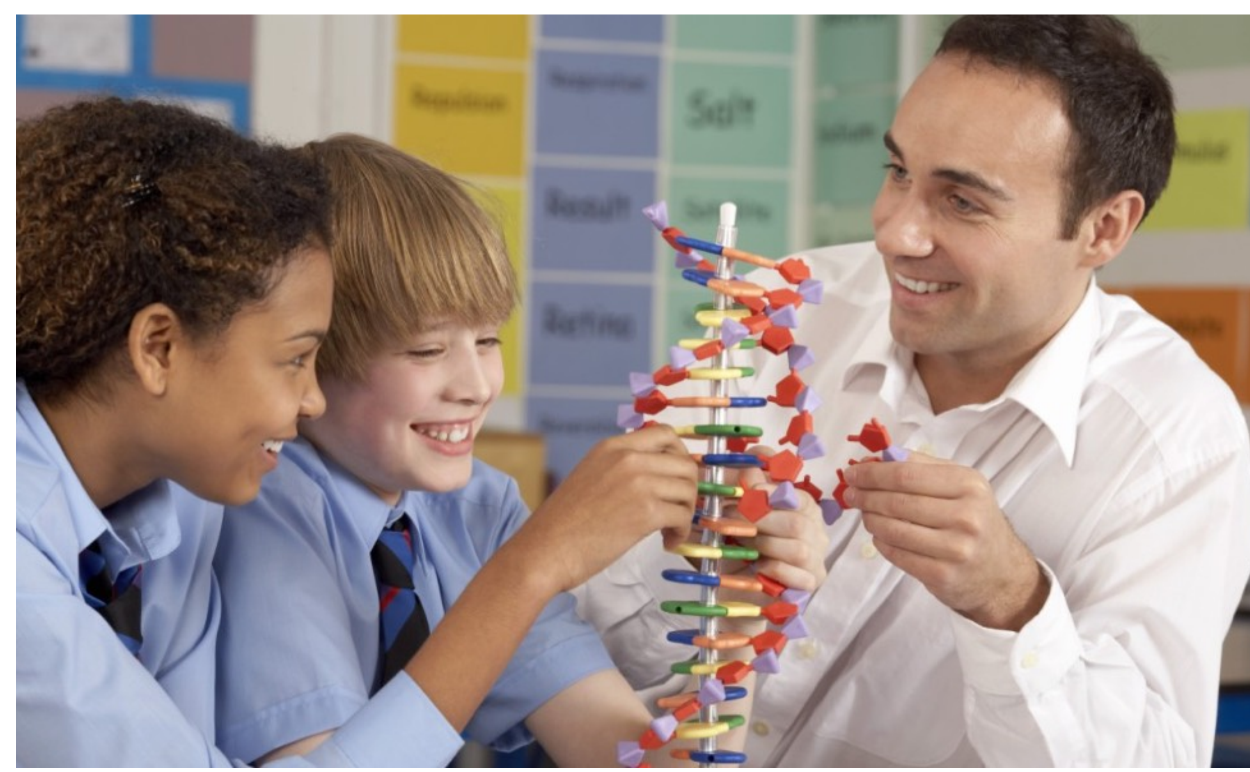
Teacher-Student Scenario



Teacher-Student Scenario

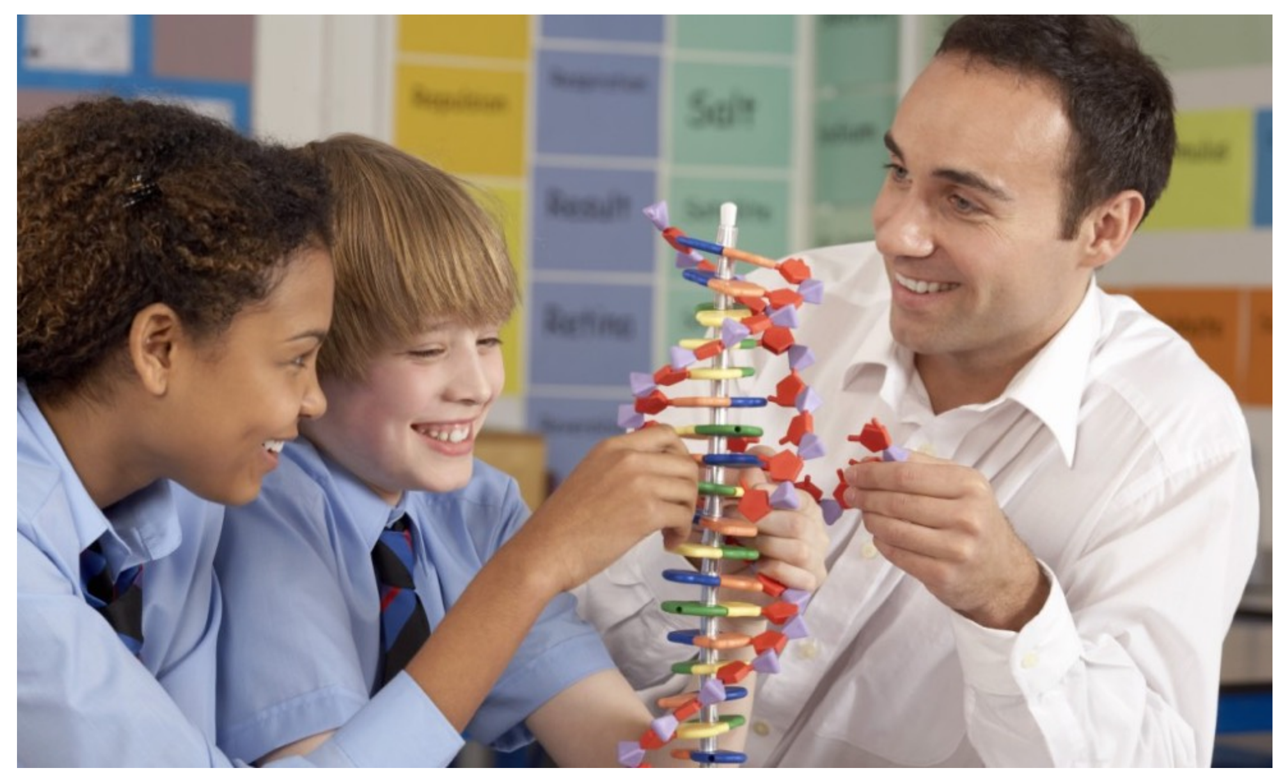
Teacher-Student Scenario

★ **Allows for a detailed analytical description (& some mathematically rigorous statements)**



Teacher-Student Scenario

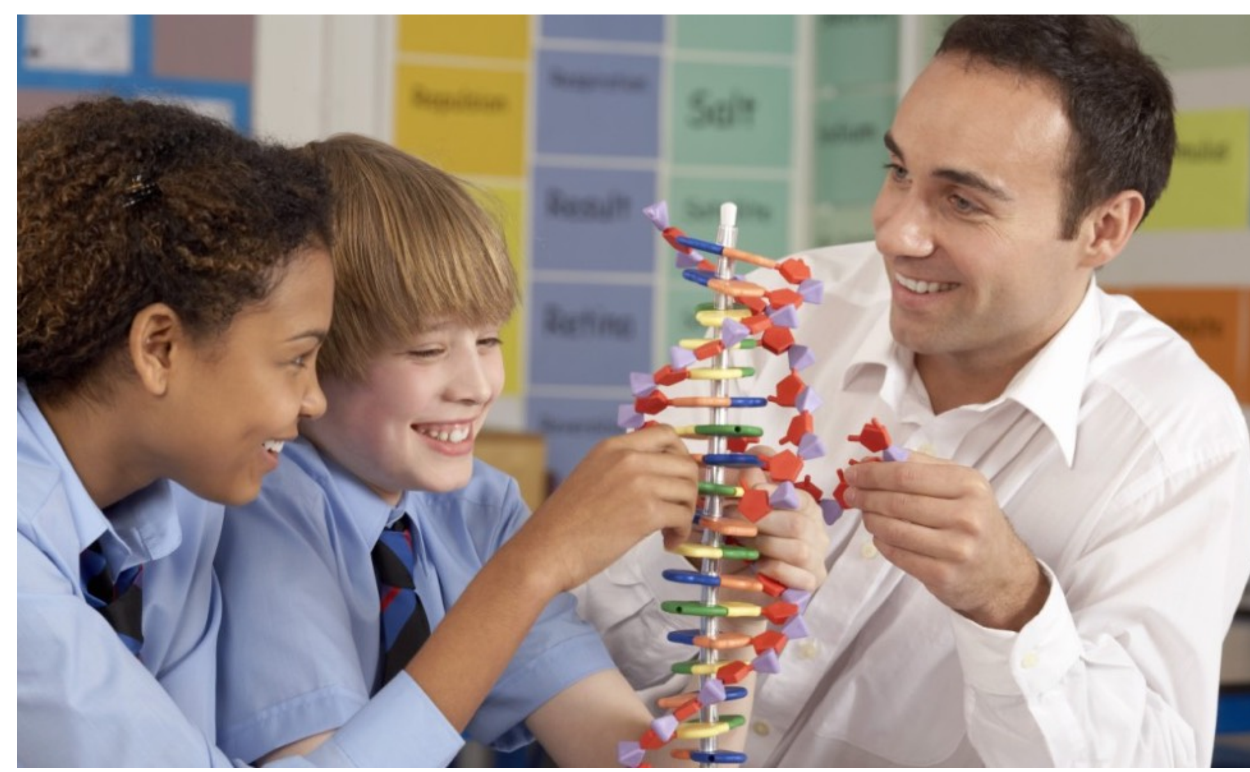
★ Allows for a detailed analytical description (& some mathematically rigorous statements)



Teacher-Student Scenario

★ Allows for a detailed analytical description (& some mathematically rigorous statements)

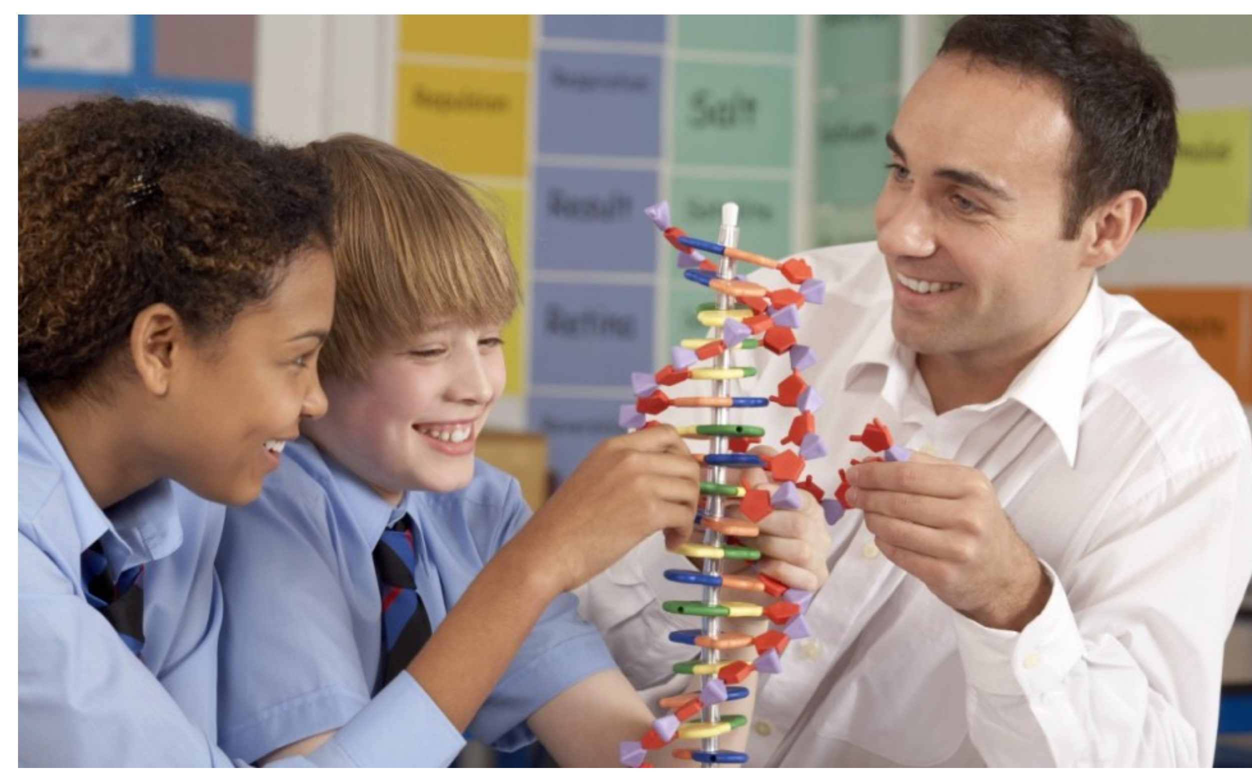
★ Rich picture for optimal generalization, Rademacher bounds, various algorithms, etc…



Teacher-Student Scenario

★Allows for a detailed analytical description (& some mathematically rigorous statements)

★Rich picture for optimal generalization, Rademacher bounds, various algorithms, etc…

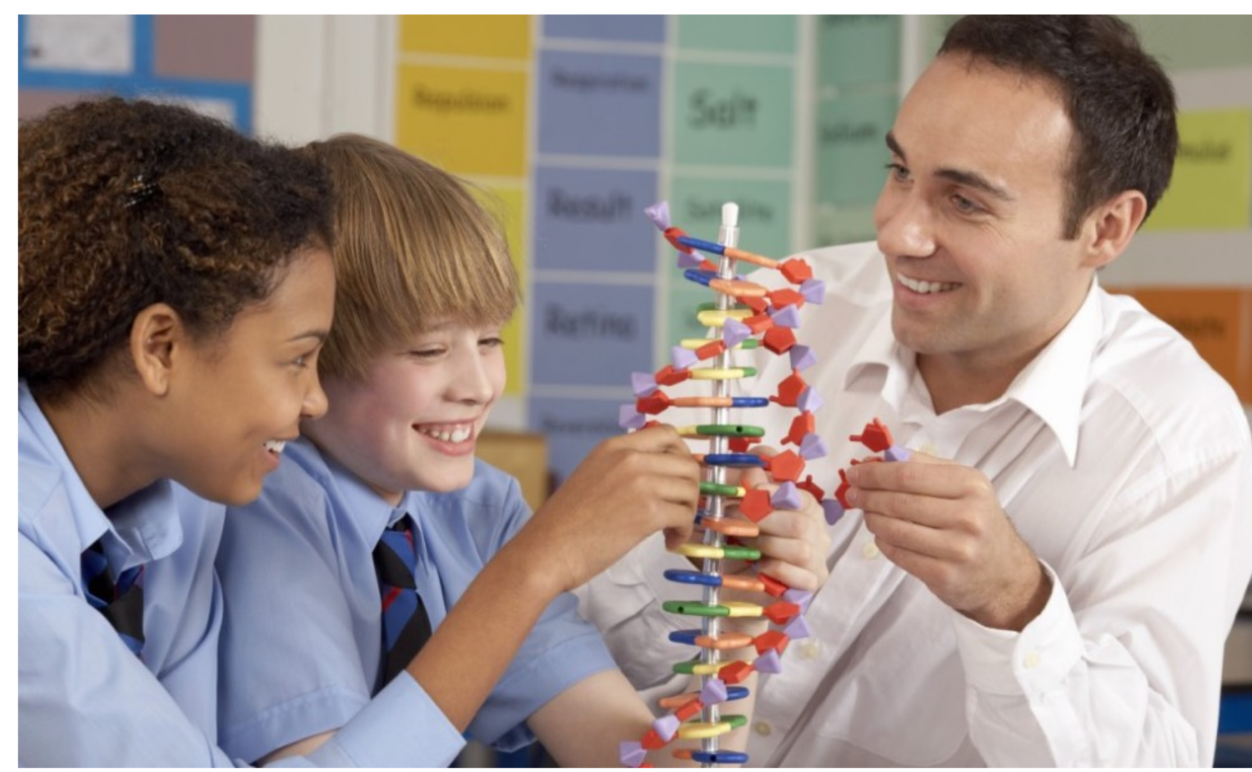


Teacher-Student Scenario

★Allows for a detailed analytical description (& some mathematically rigorous statements)

★Rich picture for optimal generalization, Rademacher bounds, various algorithms, etc…

★**Old topic, but still much to do and many recent works in this setting**
(*e.g. Zecchina's group, see also Marylou Gabrie's poster on mutual information*)

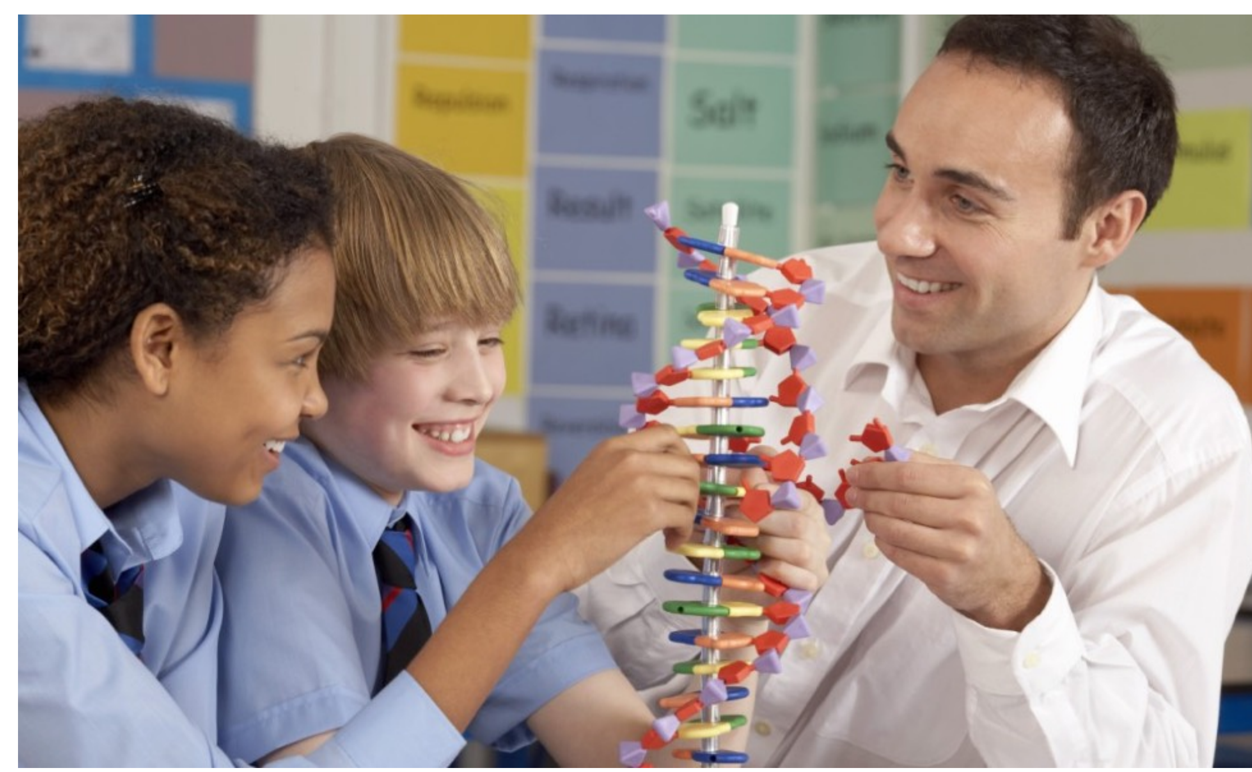


Teacher-Student Scenario

★Allows for a detailed analytical description (& some mathematically rigorous statements)

★Rich picture for optimal generalization, Rademacher bounds, various algorithms, etc…

★Old topic, but still much to do and many recent works in this setting
(*e.g. Zecchina's group, see also Marylou Gabrie's poster on mutual information*)



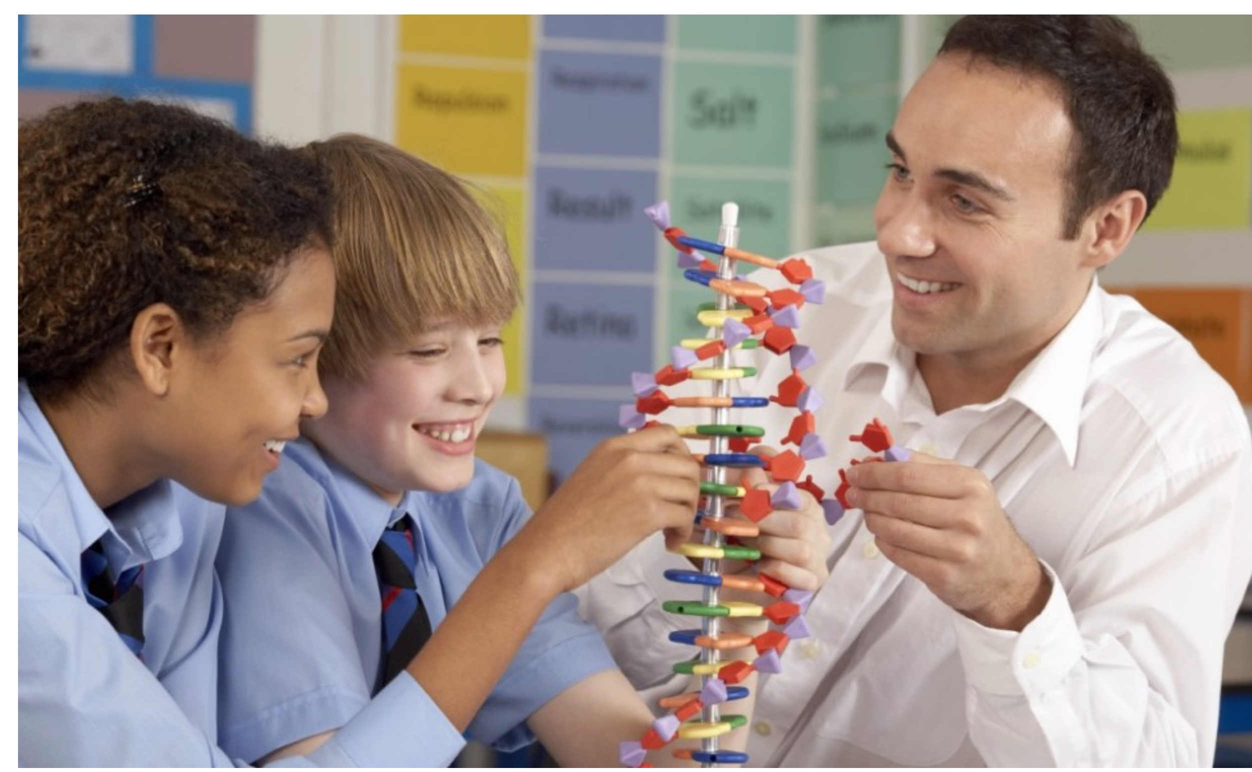
Teacher-Student Scenario

★ Allows for a detailed analytical description (& some mathematically rigorous statements)

★ Rich picture for optimal generalization, Rademacher bounds, various algorithms, etc…

★ Old topic, but still much to do and many recent works in this setting
(*e.g. Zecchina's group, see also Marylou Gabrie's poster on mutual information*)

★ **What is needed now ?**



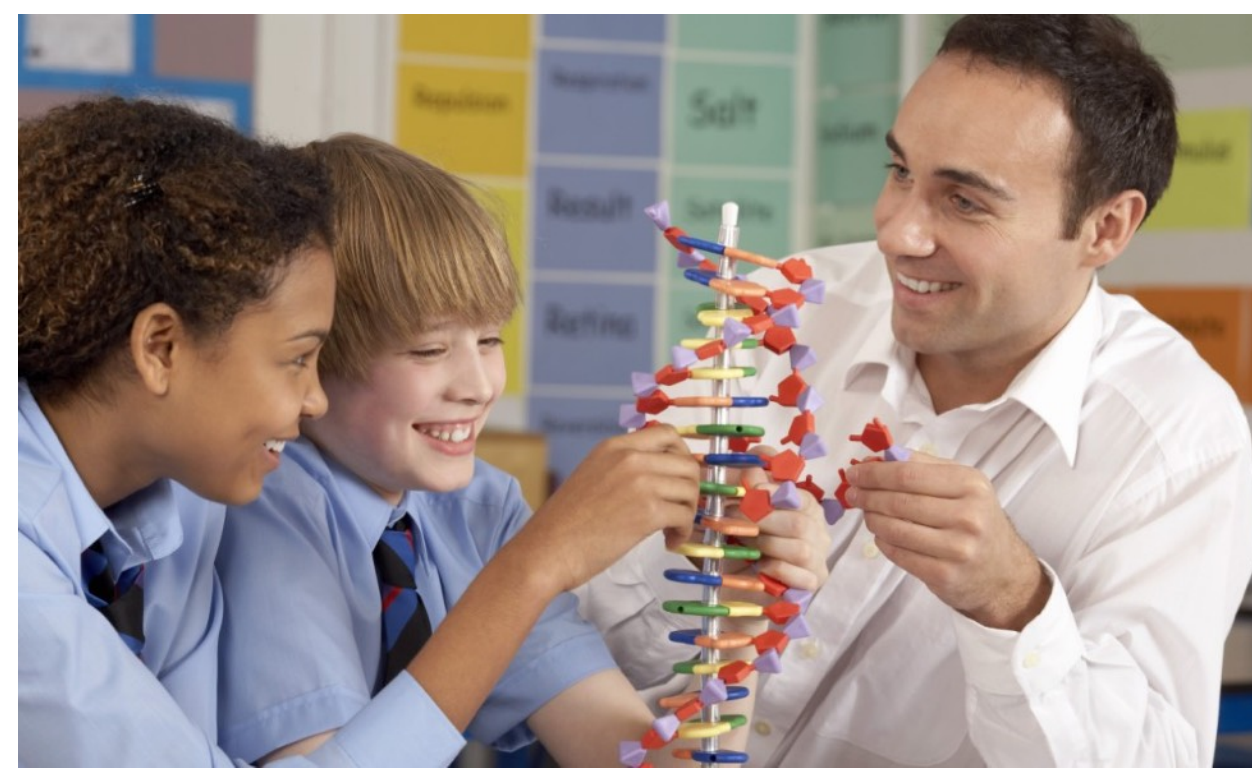
Teacher-Student Scenario

★Allows for a detailed analytical description (& some mathematically rigorous statements)

★Rich picture for optimal generalization, Rademacher bounds, various algorithms, etc…

★Old topic, but still much to do and many recent works in this setting
(*e.g. Zecchina's group, see also Marylou Gabrie's poster on mutual information*)

★ **What is needed now ?** **(a) Realistic teacher, with *structured* and *correlated* data**



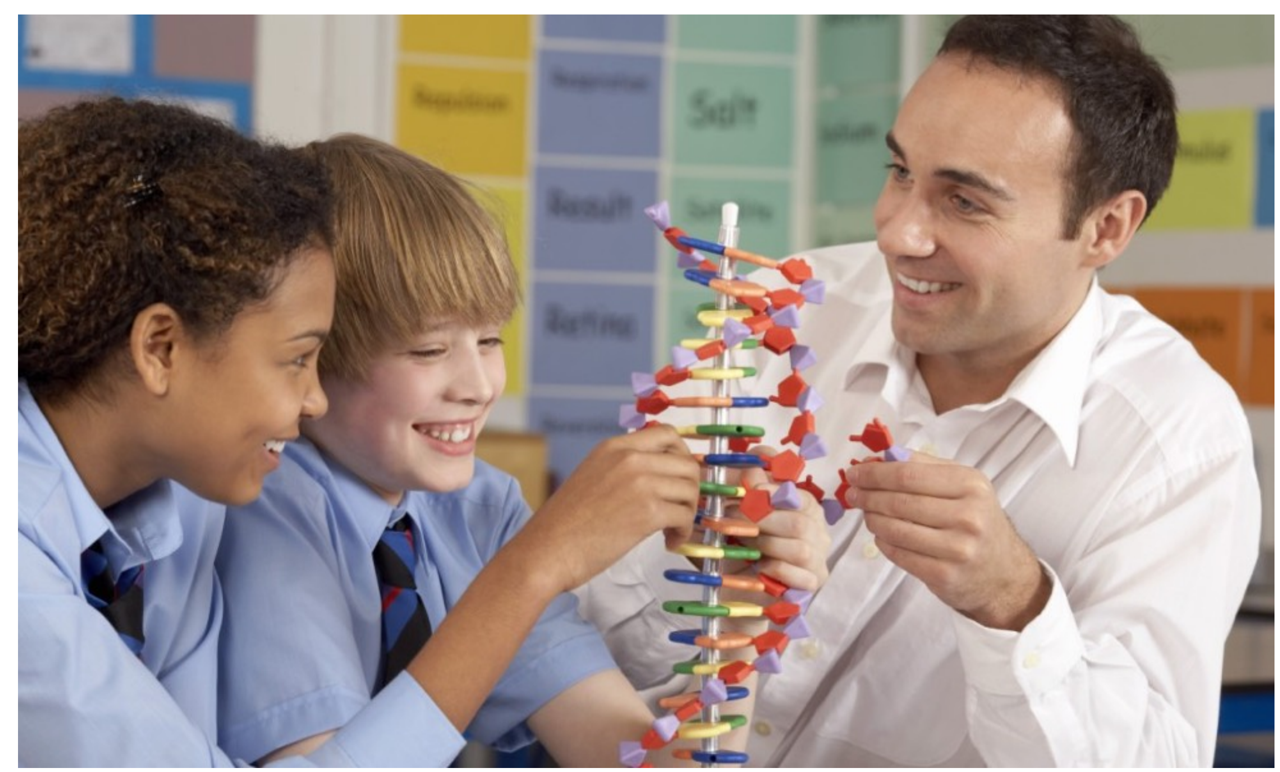
Teacher-Student Scenario

★ Allows for a detailed analytical description (& some mathematically rigorous statements)

★ Rich picture for optimal generalization, Rademacher bounds, various algorithms, etc…

★ Old topic, but still much to do and many recent works in this setting
(*e.g. Zecchina's group, see also Marylou Gabrie's poster on mutual information*)

★ **What is needed now ?** **(a) Realistic teacher, with *structured* and *correlated* data**
(b) More studies on *over-parametrized* models



Teacher-Student Scenario

★Allows for a detailed analytical description (& some mathematically rigorous statements)

★Rich picture for optimal generalization, Rademacher bounds, various algorithms, etc…

★Old topic, but still much to do and many recent works in this setting
(*e.g. Zecchina's group, see also Marylou Gabrie's poster on mutual information*)

★ **What is needed now ?**

(a) Realistic teacher, with *structured* and *correlated* data
(b) More studies on *over-parametrized* models
(c) More studies on practical algorithms