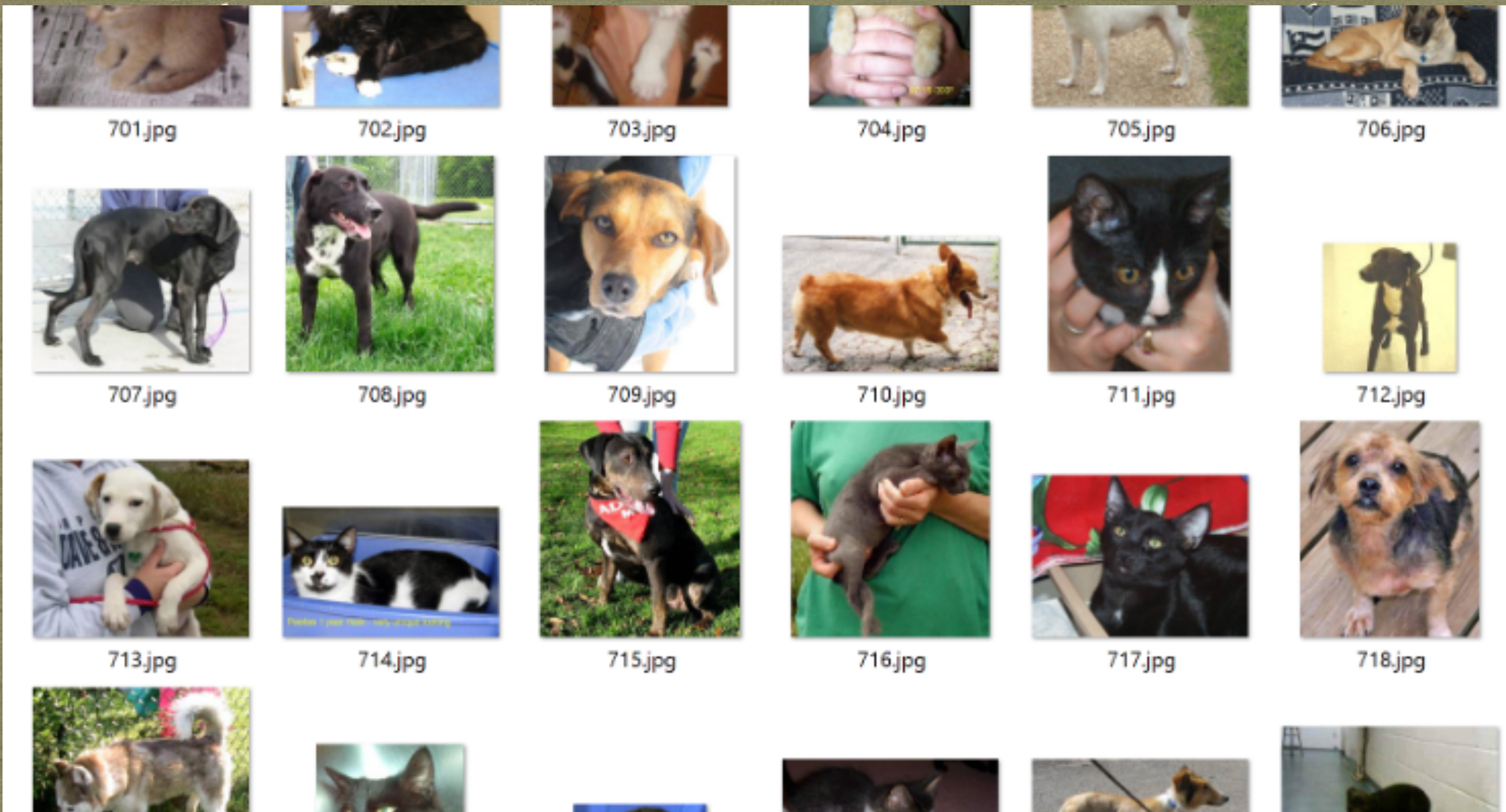
STATISTICAL PHYSICS OF LEARNING A RULE: DECADES OLD STORY CONTINUED

Lenka Zdeborová
(IPhT, CEA Saclay, France)

With: F. Krzakala, M. Mezard, N. Macris, J. Barbier, F. Caltagirone, A. Manoel, L. Miolane, F. Sausset, C. Schulke, Y. Sun, E. Tramel,

Memory Formation Conference, KITP, February 12-16, 2018.

LEARNING A RULE



LEARNING A RULE

 = F_μ = (01001010 01110011 10001100 01001011 01110000 10001100 ..... all the pixels ....)

Goal: Find a function f so that

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

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

Today this can be done with deep neural networks.

(causing excitement in many areas of interest, in science, in business ...).

STATISTICAL PHYSICS OF LEARNING A RULE

M "pictures" $F_\mu \in \mathbb{R}^N \quad \mu = 1, \ldots, M$

A rule: $f : F_\mu \to y_\mu \in \{+1, -1\}$

Model B in Gardner, Derrida'88: (teacher-student perceptron)

Elements of F (matrix) generated as iid random Gaussians.

Rule/teacher x* so that

$$y_\mu = \mathrm{sign}\left(\sum_{i=1}^{N} F_{\mu i} x_i^*\right)$$

Goal: Learn x* from M samples/examples of (F_μ,y_μ).

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

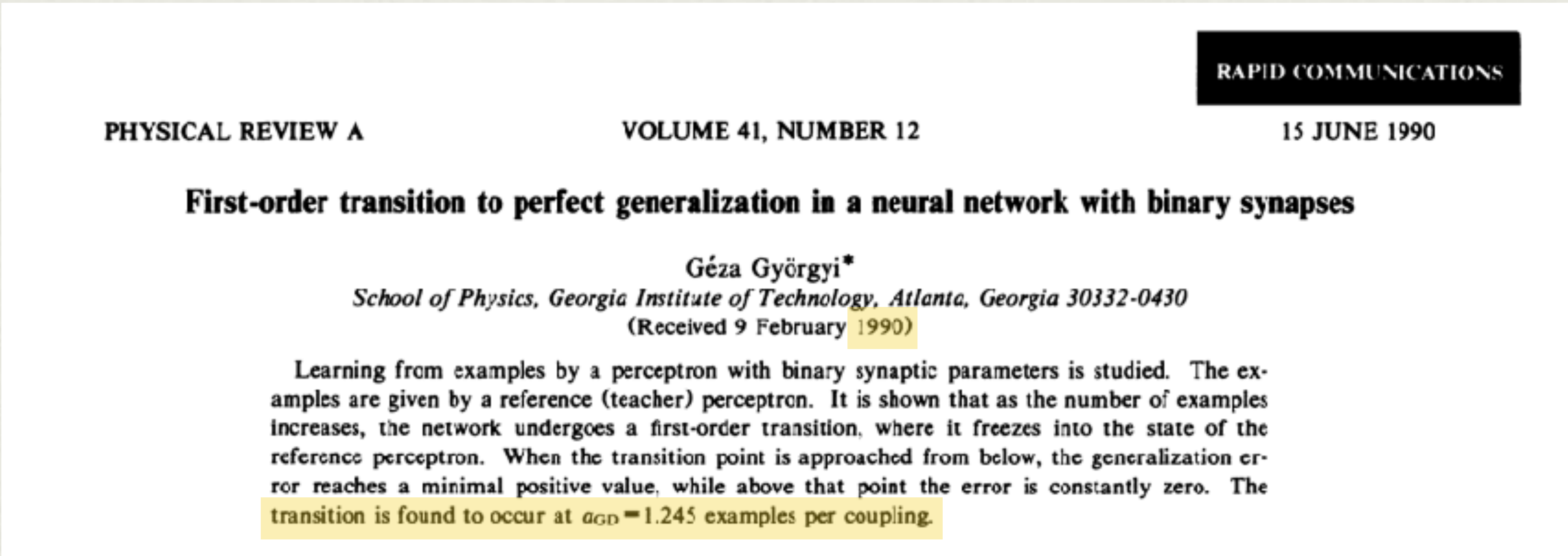
Received 13 December 1988

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 ±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.

STORAGE CAPACITY

- Main focus of Gardner&Derrida was storage capacity:
  - y iid random, F iid random, (no x*).
  - Is there an x so that for all μ=1,...,M: $y_\mu = \text{sign}(\sum_{i=1}^{N} F_{\mu i} x_i)$
- Interesting mathematically (constraint satisfaction problem), but no notion of generalisation error (=when we get a new picture the rule should be able to tell a dog from a cat).
- Back to the learning the rule setting where $y_\mu = \text{sign}(\sum_{i=1}^{N} F_{\mu i} x_i^*)$ and we need to find x* back from (y,F).

Solved using the replica method in the limit $N \to \infty \quad \alpha = M/N$



RAPID COMMUNICATIONS

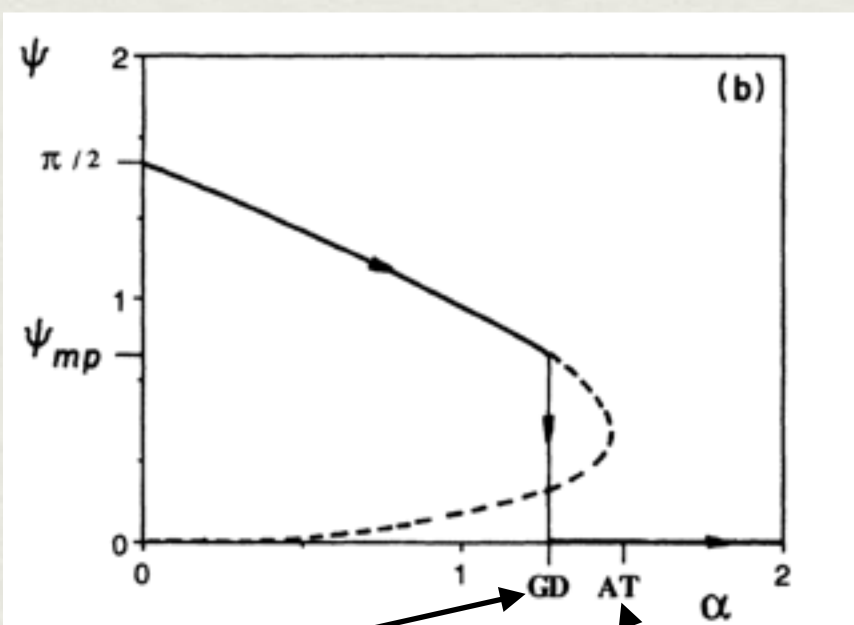
PHYSICAL REVIEW A VOLUME 41, NUMBER 12 15 JUNE 1990

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



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

- Binary weights/synapses:

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

- “The dashed lines represent non-physical segments of the curves.” (Gyorgyi’90)

VOLUME 65, NUMBER 13 PHYSICAL REVIEW LETTERS 24 SEPTEMBER 1990

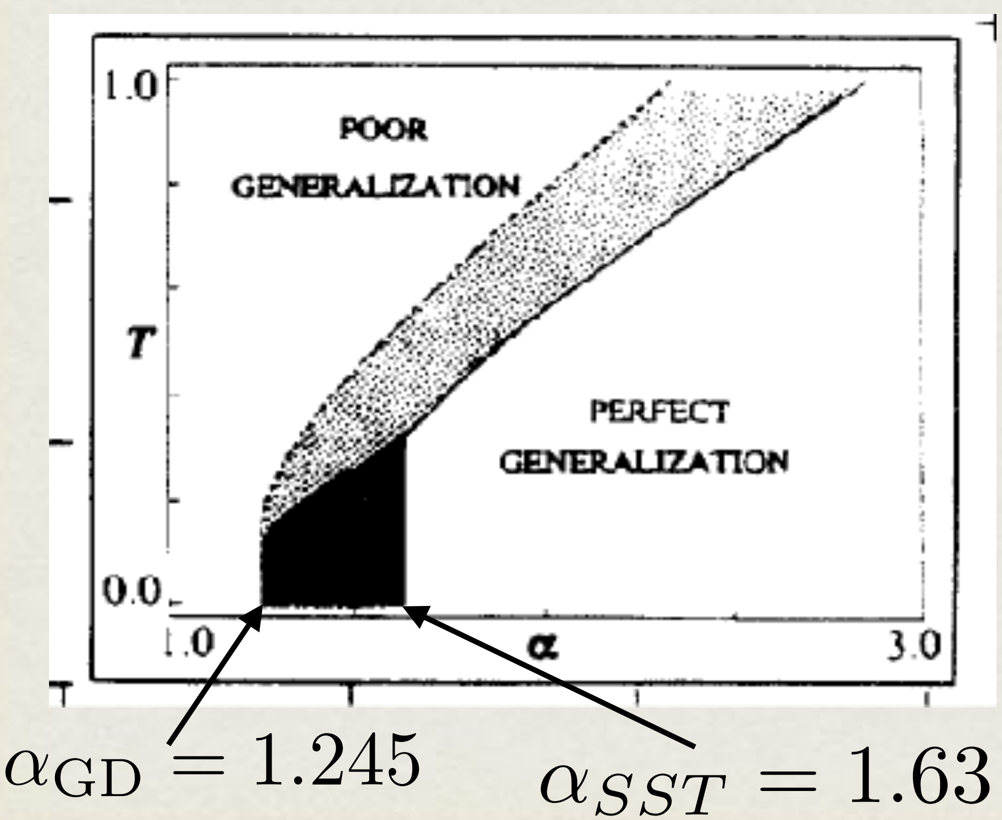
Learning from Examples in Large Neural Networks

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

H. S. Seung
Department of Physics, Harvard University, Cambridge, Massachusetts 02138
(Received 29 May 1990)

A statistical mechanical theory of learning from examples in layered networks at finite temperature is studied. When the training error is a smooth function of continuously varying weights the generalization error falls off asymptotically as the inverse number of examples. By analytical and numerical studies of single-layer perceptrons we show that when the weights are discrete the generalization error can exhibit a discontinuous transition to perfect generalization. For intermediate sizes of the example set, the state of perfect generalization coexists with a metastable spin-glass state.

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



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

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

VOLUME 76, NUMBER 11 PHYSICAL REVIEW LETTERS 11 MARCH 1996

Mean Field Approach to Bayes Learning in Feed-Forward Neural Networks

Manfred Opper
Institut für Theoretische Physik, Julius-Maximilians-Universität, Am Hubland, D-97074 Würzburg, Germany

Ole Winther
CONNECT, The Niels Bohr Institute, Blegdamsvej 17, 2100 Copenhagen Ø, Denmark

(Received 6 October 1995)

We propose an algorithm to realize Bayes optimal predictions for feed-forward networks which is based on the Thouless-Anderson-Palmer mean field method developed for the statistical mechanics of disordered systems. We conjecture that our approach will be exact in the thermodynamic limit. The algorithm results in a simple built-in leave-one-out cross validation of the predictions. Simulations for the case of the simple perceptron and the committee machine are in excellent agreement with the results of replica theory.

PACS numbers: 87.10.+e, 64.60.Cn

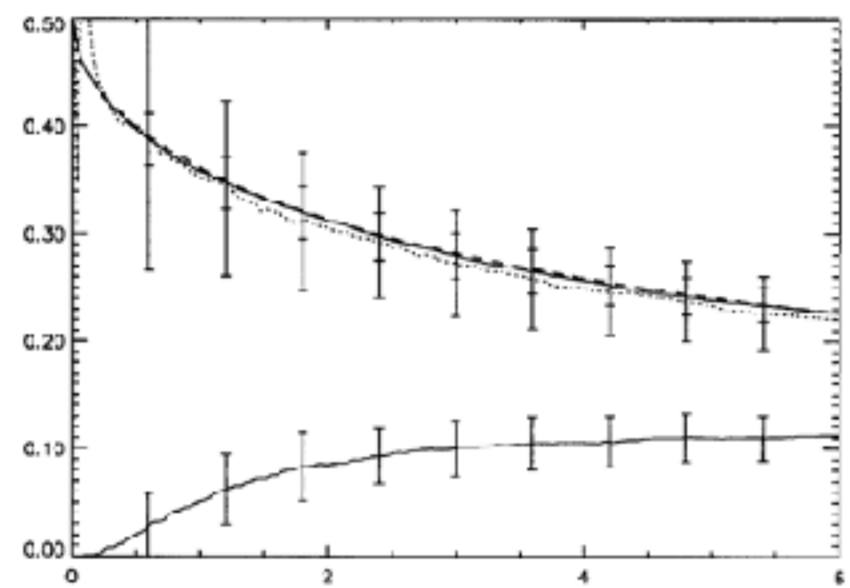


FIG. 1. The Bayes learning curve for the simple perceptron with output noise $\beta = 0.5$ and $N = 50$ averaged over 200 runs. The full lines are the simulation results (upper curve shows prediction error and the lower curve shows training error). The dashed line is the theoretical prediction. The dotted line with larger error bars is the moving control estimate.

- Spherical weights/synapses $\sum_i x_i^2 = N$

- But: TAP do not converge for large N.

- But: Conjecture false for binary weights.

A BIT OF HISTORY

- Very active part of statistical physics in the 90s. Whole section of arxiv.org/cond-mat/ 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 (next slide).

- After 2000, not much activity on *artificial* neural networks among statistical physics community.
- Massive come-back in recent years when Deep Learning became widely known and used.

OPEN QUESTIONS

- If the optimal generalization error by Gyorgyi & Sompolinsky, Tishby, Seung is correct, can we prove it mathematically rigorously?
- What is the smallest α reachable with tractable algorithms?
- What if the activation function was different (e.g. relu instead of sign)?
- What if the weights were different (e.g. sparse instead of binary)?

All answered in this talk.

GENERALIZED LINEAR REGRESSION

component-wise function

$$y = f_\xi(Fx^*)$$

e.g.: $f_\xi(z) = \mathrm{sign}(z)$
$x_i^* \in \{\pm 1\}$

labels: $y \in \mathbb{R}^M$
data matrix: $F \in \mathbb{R}^{M \times N}$
ground truth weights: $x^* \in \mathbb{R}^N$
noise: $\xi \in \mathbb{R}^M$

- Goal: Estimate x from examples (F_μ,y_μ).
- Special cases: Signal reconstruction in computed tomography, magnetic resonance imaging, phase retrieval, compressed sensing, LASSO, superposition error correcting codes, code-division multiple-access problem, group testing, logistic regression, ...

Google Scholar

generalized linear regression models

Articles

About 2,620,000 results (**0.29** sec)

Any time
Since 2017
Since 2016
Since 2013
Custom range...

Sort by relevance
Sort by date

include patents
include citations

Create alert

Generalized linear models
P McCullagh - European Journal of Operational Research, 1984 - Elsevier
... where S,cs and S,~ are the **regression** and residual 290 P. McCullagh / **Generalized linear models** sum of ... A natural **generalization** corresponding to canonical **regression** would be to write 11, = flo + (flVx,)e, (27) but the above **model** is no longer of the gener- alized **linear** type ...
Cited by 32787 Related articles All 15 versions

Longitudinal data analysis using generalized linear models
KY Liang, SL Zeger - Biometrika, 1986 - academic.oup.com
... With a single observation for each subject (nt =1), a **generalized linear model** (McCullagh & Nelder, 1983) can be applied to obtain such a description ... This paper presents an extension of **generalized linear models** to the analysis of longitudinal data when **regression** is the ...
Cited by 15100 Related articles All 19 versions

[BOOK] **Generalized linear models**
JA Nelder, RJ Baker - 1972 - Wiley Online Library
A **statistical model** is the specification of a probability distribution. For example, the **model** implicit in much of **regression** analysis is that the observations have a normal distribution*, the means being **linearly** related to the covariate values. Similarly, a log-**linear model** for
Cited by 6260 Related articles All 10 versions

[BOOK] **Generalized additive models**
T Hastie, R Tibshirani - 1990 - Wiley Online Library
... Linearity always remains a special case, and thus simple **linear** relationships can be easily ... Friedman (5) proposed a **generalization** of additive **modeling** that finds interactions among prognostic factors ... Software for fitting **generalized** additive **models** is available as part of the S ...
Cited by 14911 Related articles All 37 versions

GENERALIZED LINEAR REGRESSION

component-wise function

$$y = f_\xi(Fx^*)$$

e.g.: $f_\xi(z) = \mathrm{sign}(z)$
$x_i^* \in \{\pm 1\}$

labels: $y \in \mathbb{R}^M$
data matrix: $F \in \mathbb{R}^{M\times N}$
ground truth weights: $x^* \in \mathbb{R}^N$
noise: $\xi \in \mathbb{R}^M$

- Goal: Estimate x from examples (F_μ,y_μ).
- Model considered in this talk:
  - F iid of zero mean and variance 1/N;
  - x* iid random from P_x (e.g. sparse, binary);
  - High-dimensional limit: $N, M \to \infty, \alpha \equiv M/N = O(1)$

BAYES-OPTIMAL ESTIMATION

$$P(x|y,F) = \frac{1}{Z(y,F)} \prod_{\mu=1}^{M} P_{\text{out}}(y_\mu|z_\mu) \prod_{i=1}^{N} P_X(x_i) \qquad z_\mu = \sum_{i=1}^{N} F_{\mu i} x_i$$

- x* ~ P_X; y generated from $P_{\text{out}}(y|z) = \mathbb{E}_{P_\xi}[\delta(y - f_\xi(z))]$
- Estimate x* from (F, y). For a new row, F_{new}, predict the label y_{new}.
- Bayes-optimal inference $\hat{x}_i$ = marginal mean of x_i in P(x|y,F). Optimal because it minimizes $\text{MSE} = \frac{1}{N}\sum_{i-1}^{N}(\hat{x}_i - x_i^*)^2$
- Bayes-optimal prediction/generalization:

$$\hat{y}_{\text{new}} = \mathbb{E}_{P(x|y,F),P_\xi}\left[f_\xi(F_{\text{new}}x)\right]$$

No over-fitting! No other procedure can be better.

CLOSING 27 YEARS OLD CONJECTURE

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

Def. “quenched” free energy: $f \equiv \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{y,F} \log Z(y,F)$ $\qquad \alpha = \frac{M}{N}$

Theorem 1 (informally): The replica free energy is correct.

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

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

where

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

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

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

CLOSING 27 YEARS OLD CONJECTURE

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

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

Theorem 1 (informally): The replica free energy is correct.

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

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

Theorem 2: Optimal error on estimation of x* is:

$$\mathrm{MMSE} = \rho - m^*$$

where m* is the extremizer of f_{RS}. $\qquad \rho = \mathbb{E}_{P_X}(x^2)$

CLOSING 27 YEARS OLD CONJECTURE

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

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

Theorem 1 (informally): The replica free energy is correct.

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

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

Theorem 3: Optimal generalisation error is

$$\mathcal{E}_{\text{gen}} = \mathbb{E}_{v,\xi}\left[f_\xi(\sqrt{\rho}\,v)^2\right] - \mathbb{E}_{v}\,\mathbb{E}_{w,\xi}\left[f_\xi(\sqrt{m^*}\,v + \sqrt{\rho - m^*}\,w)\right]^2$$

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

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

ONE SLIDE ON THE PROOF

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

Guerra-Toninelli interpolation between the original posterior and N + M independent scalar denoising problems.

Novel (more powerful) variant where the interpolation parameter depends is a suitably chosen function of the interpolation "time".

Key property for the proof to work (Nishimori): Under expectations ground truth x^* is exchangeable for a sample from $P(x|y,F)$.

$$\text{Physics} \underset{\text{working algorithm}}{\overset{\text{free energy computation}}{\rightleftarrows}} \text{Math}$$

IS THE OPTIMAL ERROR REACHABLE WITH EFFICIENT ALGORITHMS?

Algorithm 2 Generalized Approximate Message Passing (G-AMP)

Rangan'10

Input: y

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

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

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

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

$t \leftarrow t + 1$

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

output: a,**v**.

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

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

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

Algorithm 2 Generalized Approximate Message Passing (G-AMP)

Rangan'10

Input: y

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

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

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

$t \leftarrow t + 1$

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

output: a,**v**.

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

GAMP for prediction:

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

THE STORY OF GAMP

GAMP 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 was used as an iterative algorithm, but had wrong iteration-indices and consequently did not convergence.

Bolthausen fixed the issue in ~2008 and proved state evolution for the corrected TAP equations.

Ma come è possibile che un matematico ci abbia sorpassato!!!! Fb x

Giorgio Parisi <giorgio.parisi@roma1.infn.it> 9/26/15

to Andrea, Irene, Federico, Enzo, Miguel, Francesco, jorge, me, Florent

Italian > English View translated message Always translate: Italian

Guardate lo schema iterativo per la TAP, pagina 4
in http://arxiv.org/pdf/1201.2891v1.pdf (i.e Bolthausen's paper)

si non
k+1, k e $k-1$!!!!

Converge!!!!!

da non crederci......

L'ho letto nella tesi di Lenka.... (i.e. LZ, Krzakala, Advances in Physics'16)

Da non crederci.

...

Giorgio Parisi <giorgio.parisi@roma1.infn.it> 9/26/15

to Andrea, Irene, Federico, Enzo, Miguel, Francesco, jorge, me, Florent

Italian > English Translate message Turn off for: Italian ×

converge nella fase RS, se capisco bene, ma le iterazioni normali non
convergono nemmeno in quella fase.

Il 26 settembre 2015 19:56, Giorgio Parisi
<giorgio.parisi@roma1.infn.it> ha scritto:

Ma come è possibile che un matematico ci abbia sorpassato!!! (But how is it that a mathematician has passed us !!!!)

Flo x

Giorgio Parisi <giorgio.parisi@roma1.infn.it> 9/26/15

to Andrea, Irene, Federico, Enzo, Miguel, Francesco, jorge, me, Florent

Italian > English View original message Always translate: Italian

Look iterative scheme for the TAP, page 4
in http://arxiv.org/pdf/1201.2891v1.pdf (i.e. Bolthausen's paper)

you do not
k + 1, k $ k-1 $!!!!!

Converge !!!!!!

unbelievable......

I read the thesis Lenka .... (i.e. LZ, Krzakala, Advances in Physics'16)

Unbelievable.

-
Department of Physics, University of Rome `` La Sapienza "
Piazzale Aldo Moro 2, 00185, Rome, Italy
tel. +390649914311 Fax + 390649694323

Giorgio Parisi <giorgio.parisi@roma1.infn.it> 9/26/15

to Andrea, Irene, Federico, Enzo, Miguel, Francesco, jorge, me, Florent

Italian > English View original message Always translate: Italian

converges in the RS phase, if I understand well, but not the normal iterations
converge even at that stage.

THE STORY OF GAMP

GAMP 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 was used as an iterative algorithm, but had wrong iteration-indices and consequently did not convergence.

Bolthausen fixed the issue in ~2008 and proved state evolution for the corrected TAP equations.

For GAMP state evolution proven by Bayati, Montanari'11, Bayati, Lelarge, Montanari'12, Javanmard, Montanari'13.

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:

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

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

Recall the RS free energy

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

No “state evolution” for naive mean-field, nor MCMC, nor Langevin dynamics (except spherical p-spin, much more complex int.-diff. equations).

BOTTOMLINE

$$P(x|y,F) = \frac{1}{Z(y,F)} \prod_{\mu=1}^{M} P_{\text{out}}(y_\mu | \sum_{i=1}^{N} F_{\mu i} x_i) \prod_{i=1}^{N} P_X(x_i)$$

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

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

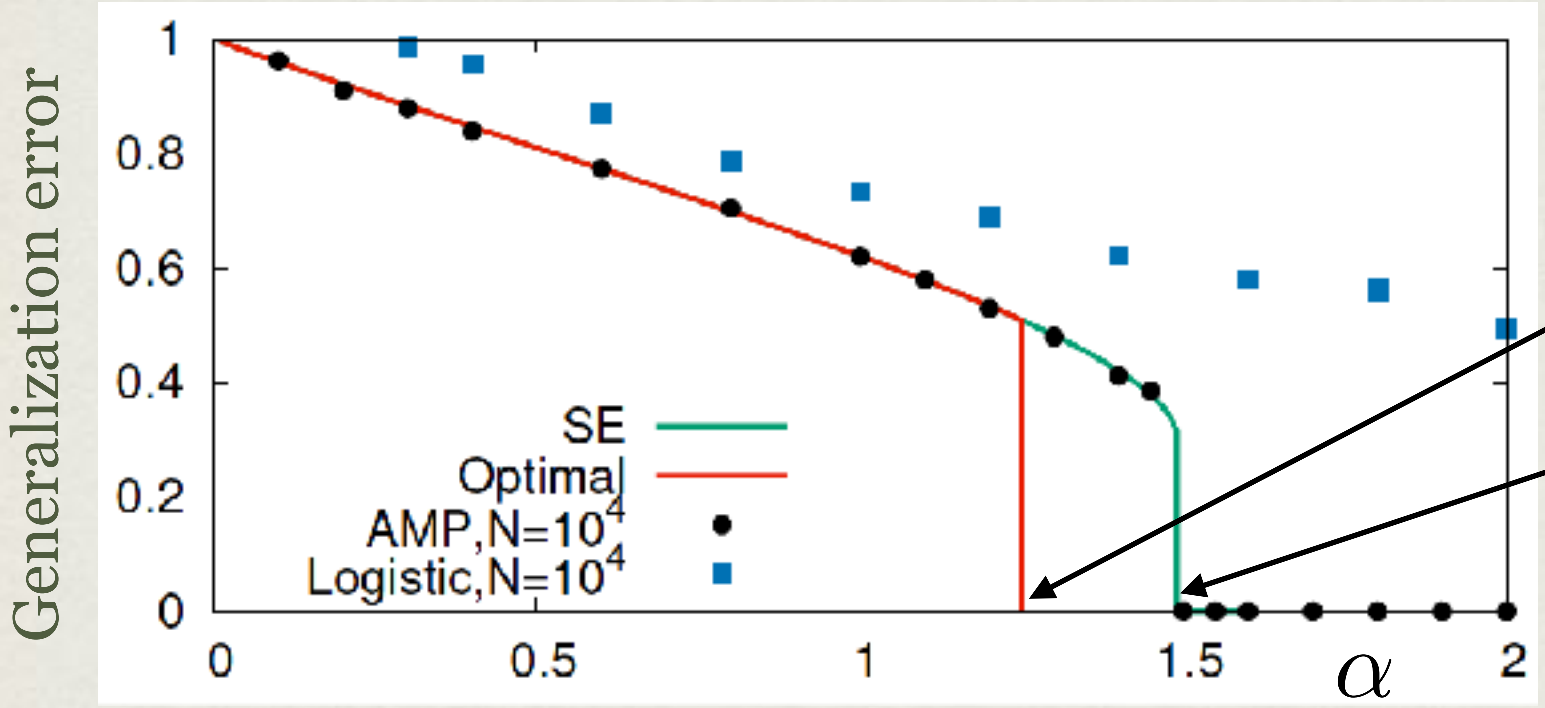
MSE_{AMP} = local extremum of $f_{\text{RS}}(m)$, reached from un-informed initialisation of state evolution.

RESULTS

BINARY PERCEPTRON

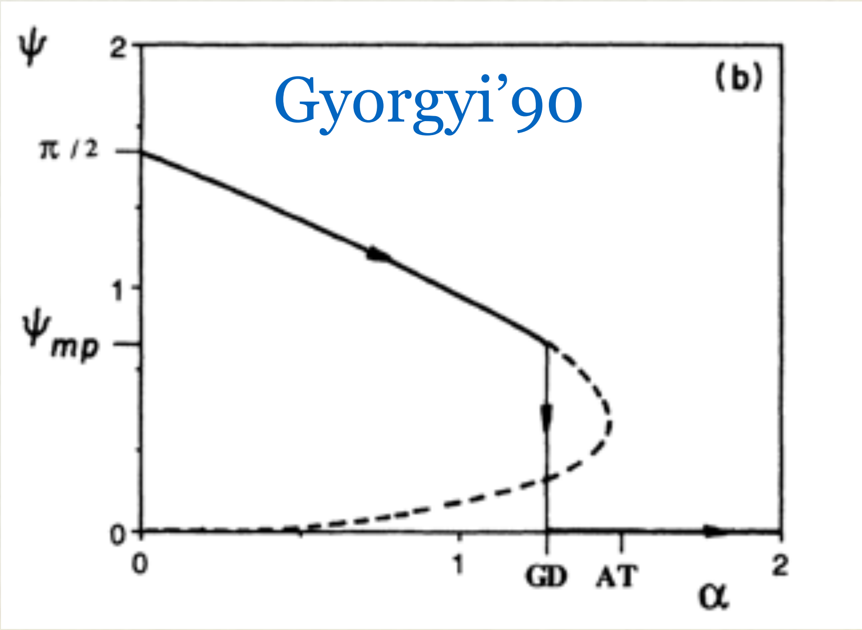
Gardner, Derrida'89, Gyorgyi'90, Sompolinsky, Tishby, Seung'90

$$y = \mathrm{sign}(Fx^*) \qquad P_X(x) = \frac{1}{2}[\delta(x-1) + \delta(x+1)]$$



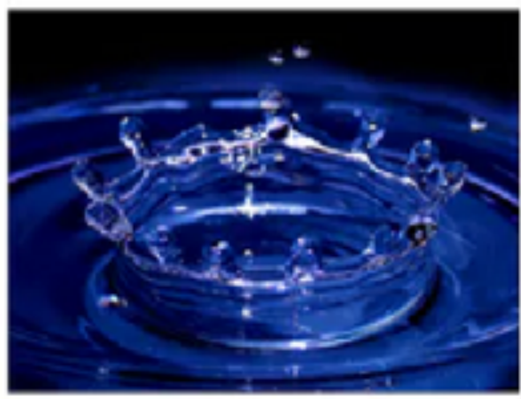
$\alpha_{\mathrm{IT}} = 1.249$

$\alpha_{\mathrm{Alg}} = 1.493$

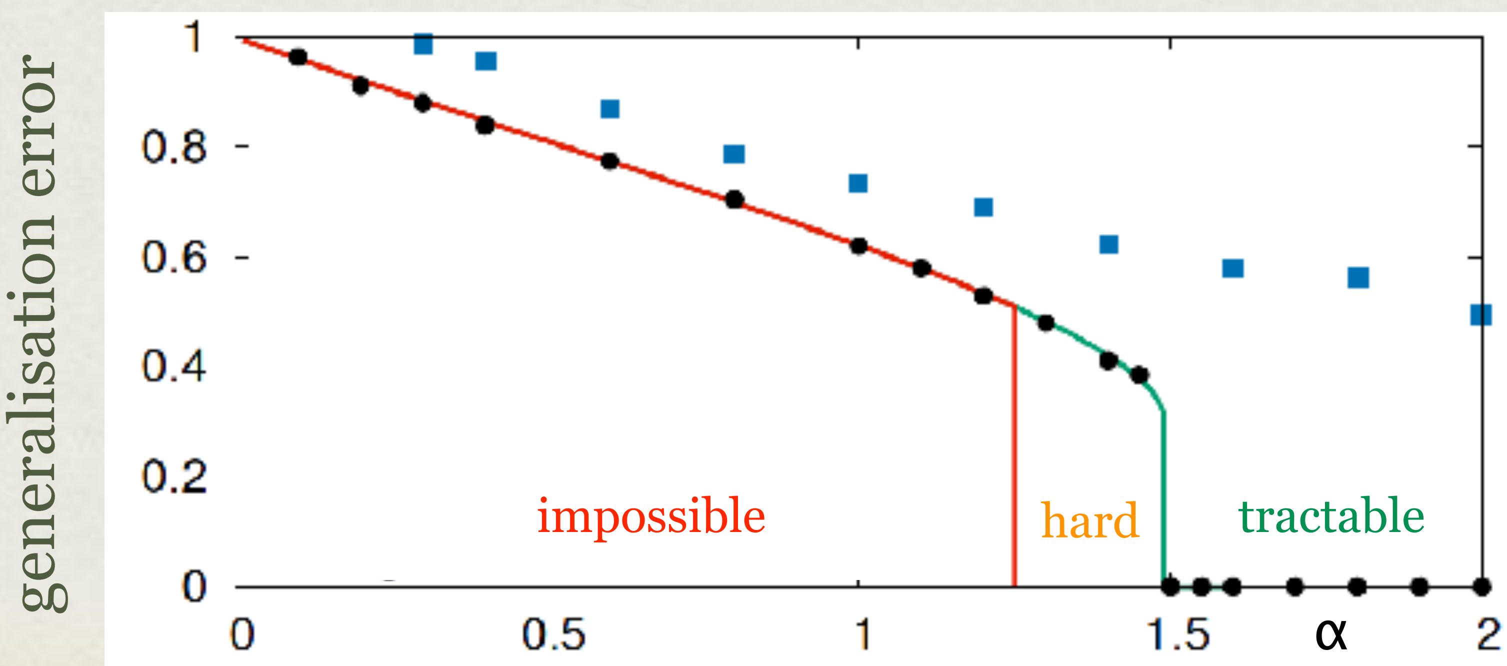


- GAMP is optimal starting from α_{Alg}.
- Redemption of the "un-physical" branch.

PHYSICS VS LEARNING



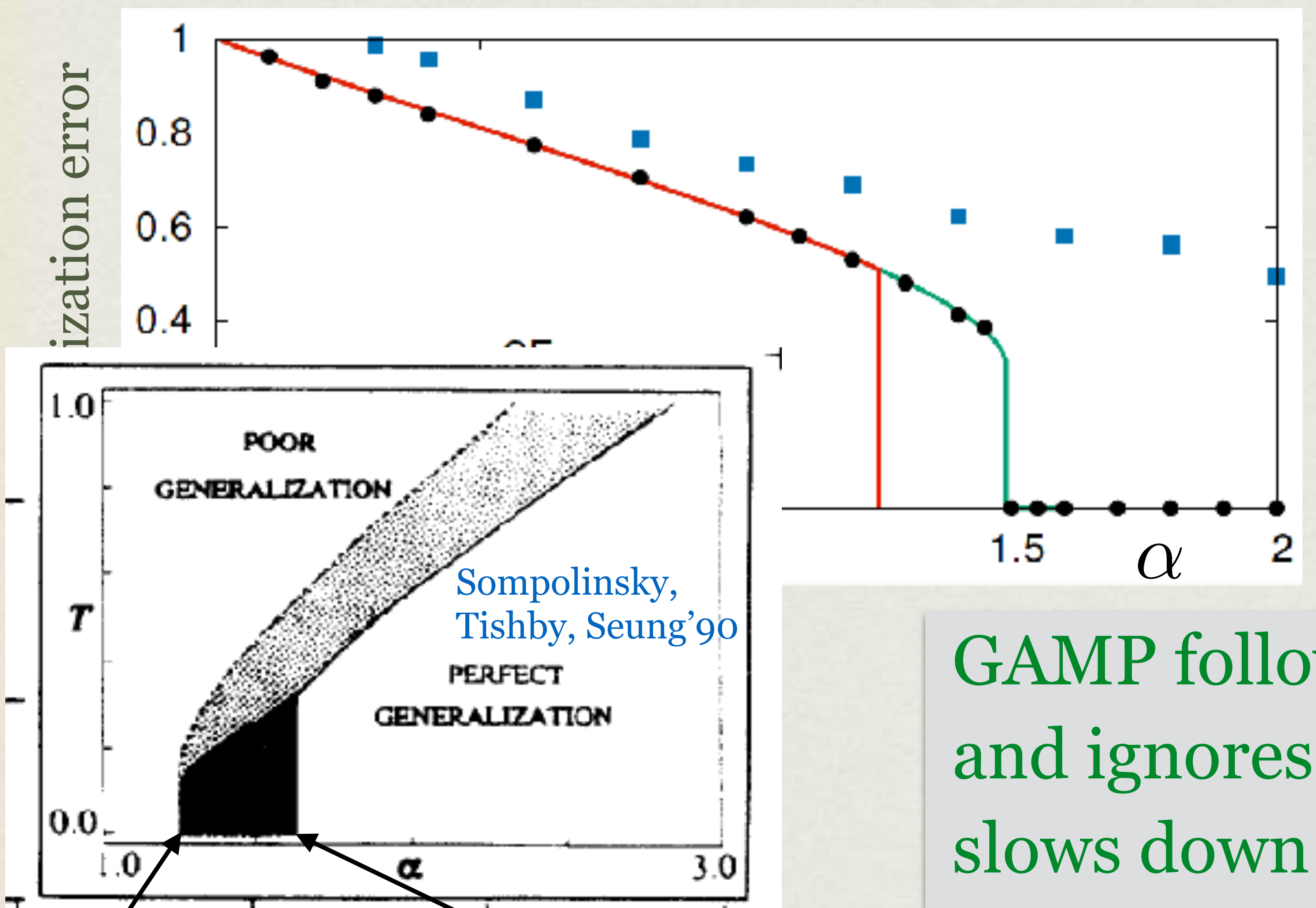
liquid	metastable liquid	crystal
impossible	hard	tractable



BINARY PERCEPTRON

Gardner, Derrida'89, Gyorgyi'90, Sompolinsky, Tishby, Seung'90

$$y = \mathrm{sign}(Fx^*) \qquad P_X(x) = \frac{1}{2}[\delta(x-1) + \delta(x+1)]$$



$\alpha_{\mathrm{GD}} = 1.245$ $\alpha_{SST} = 1.63$

$$\alpha_{\mathrm{IT}} = 1.249$$

$$\alpha_{\mathrm{Alg}} = 1.493$$

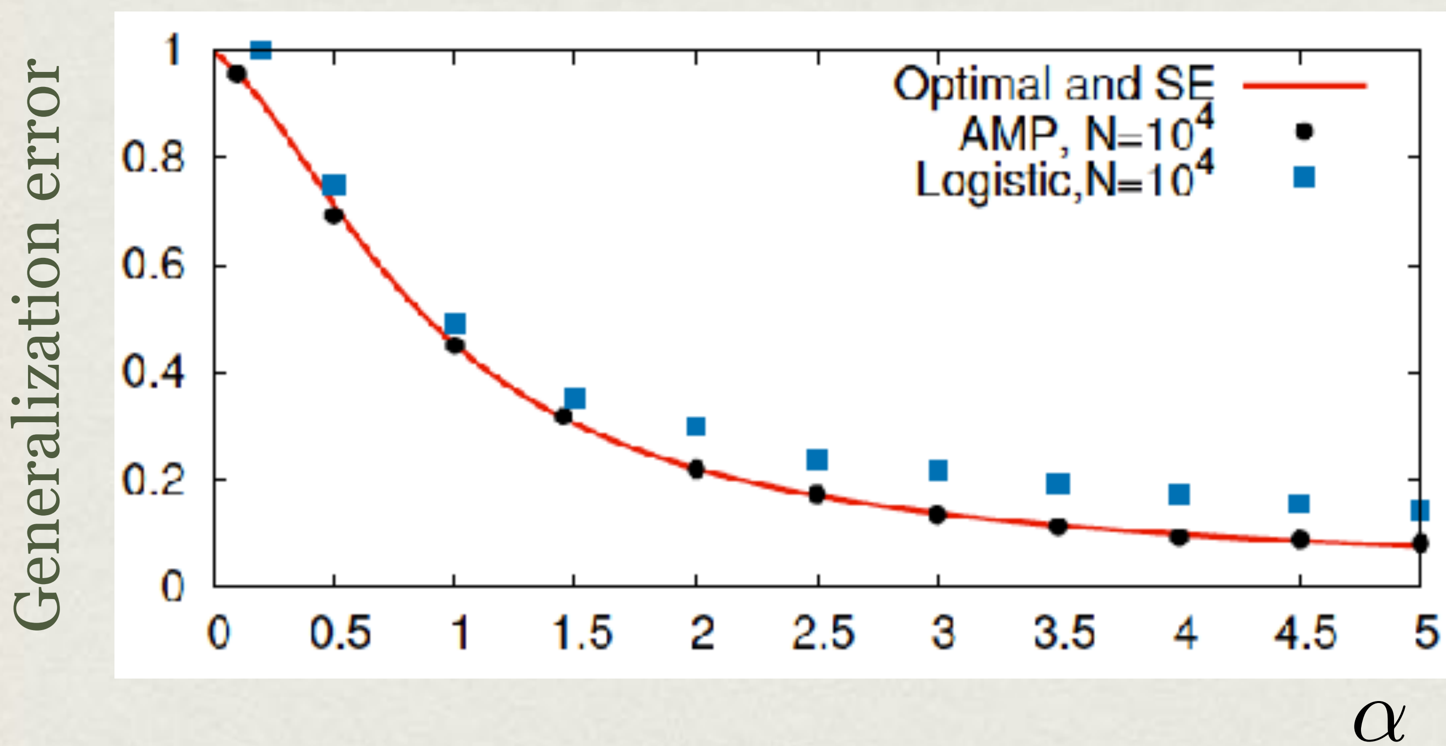
GAMP follows the liquid-spinodal, and ignores the glassiness that slows down MCMC. Can other algorithms match GAMP?

GAUSS-BERNOULLI PERCEPTRON

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

$$y = \mathrm{sign}(Fx^*)$$

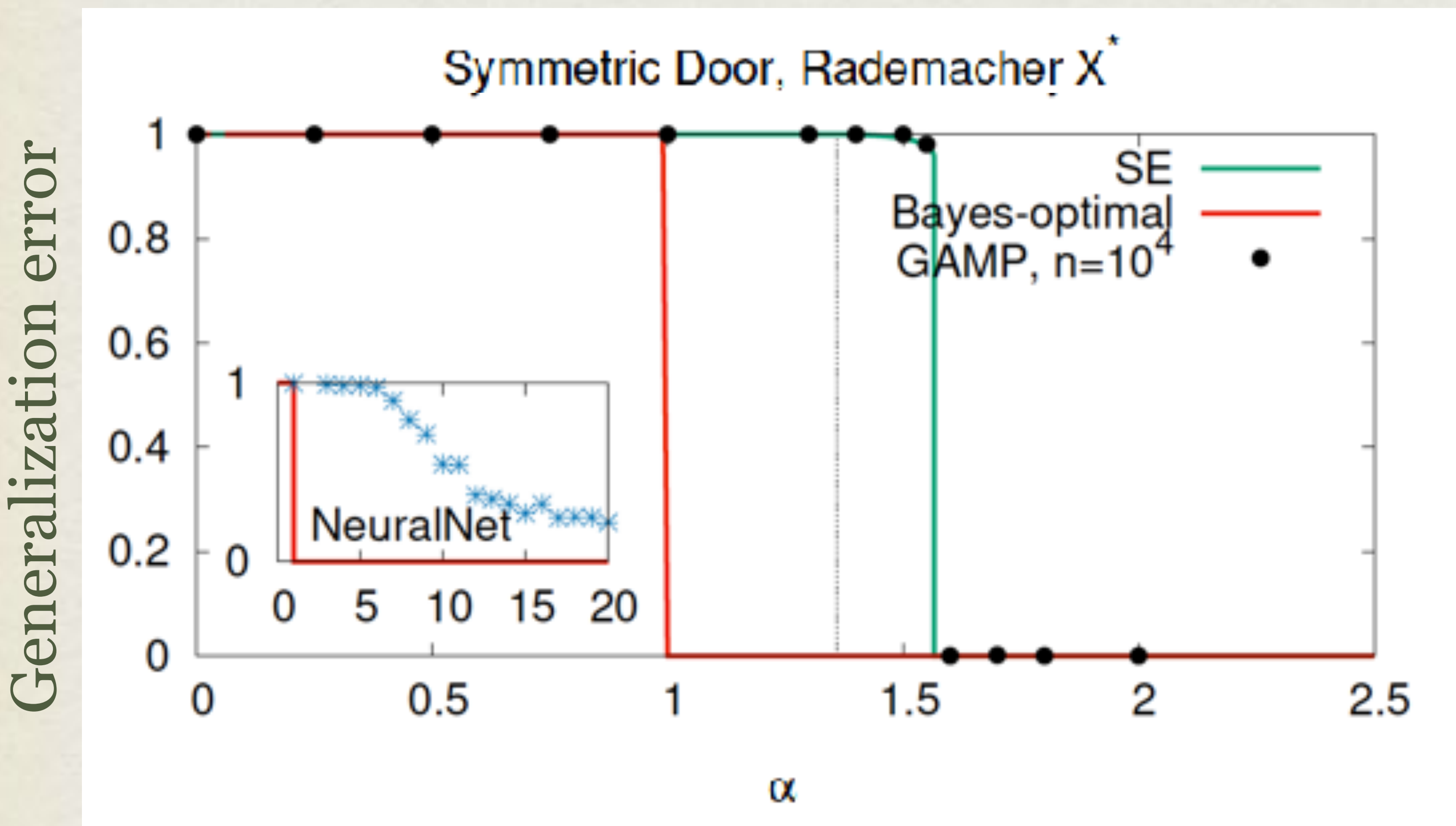
$$\rho = 0.2$$



SYMMETRIC BINARY PERCEPTRON

$$y = \mathrm{sign}(|Fx^*| - K)$$

$$P_X(x) = \frac{1}{2}[\delta(x-1) + \delta(x+1)]$$



$$\alpha_{\mathrm{IT}} = 1$$

$$\alpha_{\mathrm{Alg}} = 1.566$$

Very simple yet very hard benchmark for classification!

K chosen so that P(y=1)=0.5

from: Barbier, Krzakala, Macris, Miolane, LZ arXiv:1708.03395

NEW WITH RESPECT TO 1990

- Generic P_X and P_{out}, plug-and-go formula/algorithm (Rangan'10; LZ, Krzakala'16)
- Proof of the optimal error. (Barbier, Krzakala, Macris, Miolane, LZ'17)
- GAMP with correct time indices (Kabashima'03; Bolthausen'08; Donoho, Montanari, Maleki'09) follows the state evolution. (Bayati, Montanari'11; et al.)
- GAMP ignores glassiness, it follows the "unphysical" spinodal.
- Conjecture: GAMP optimal among tractable algorithms. (challenge for future work ...)

ONGOING WORK

- Generalized linear model as an interesting benchmarks for generic-purpose algorithms. How many samples does a deep network need to learn these simple rules?

- Beyond random iid matrices in order to study structured data.

- Beyond separable priors, extending to multiple-layers. With fixed weights (Krzakala, Manoel, Mezard, LZ'17).

- Learning of weights in multiple layers - a case where we still have to find the right decoupling to make the replica method work.

REFERENCES

F. Krzakala, M. Mézard, F. Sausset, Y. Sun, LZ, Statistical physics-based reconstruction in compressed sensing, Phys. Rev. X (2012), arXiv:1109.4424

F. Krzakala, M. Mézard, F. Sausset, Y. Sun, LZ, Probabilistic Reconstruction in Compressed Sensing: Algorithms, Phase Diagrams, and Threshold Achieving Matrices, J. Stat. Mech. (2012), arXiv:1206.3953.

LZ, F. Krzakala, Statistical Physics of Inference: Threshold and Algorithms, Advances of Physics (2016), arXiv:1511.02476.

J. Barbier, N. Macris, L. Miolane, F. Krzakala, LZ, Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models, arXiv:1708.03395.

Thank you for your attention!