GENERALISED LINEAR REGRESSION OPTIMAL ERRORS AND ALGORITHMS FOR RANDOM INSTANCES.

Lenka Zdeborová

(IPhT, CEA Saclay, France)

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

JDS2018, June 28, 2018.

WHICH PROBLEMS CAN WE SOLVE WITH COMPUTERS?

	Worst-case data	Probabilistic data model
Unlimited computation	x	**Information Theory, Statistics**
Limited computational resources	**Computational Complexity Theory**	?

- Non-convex, NP-hard problems are solved every day (neural networks, optimization, scheduling, …).
- Data is far from worst case (biology, physics, statistics, machine learning).

GENERALIZED LINEAR REGRESSION

one-layer fully-connected feedforward neural network

dependent variables: $y \in \mathbb{R}^n$

data matrix: $F \in \mathbb{R}^{n \times p}$

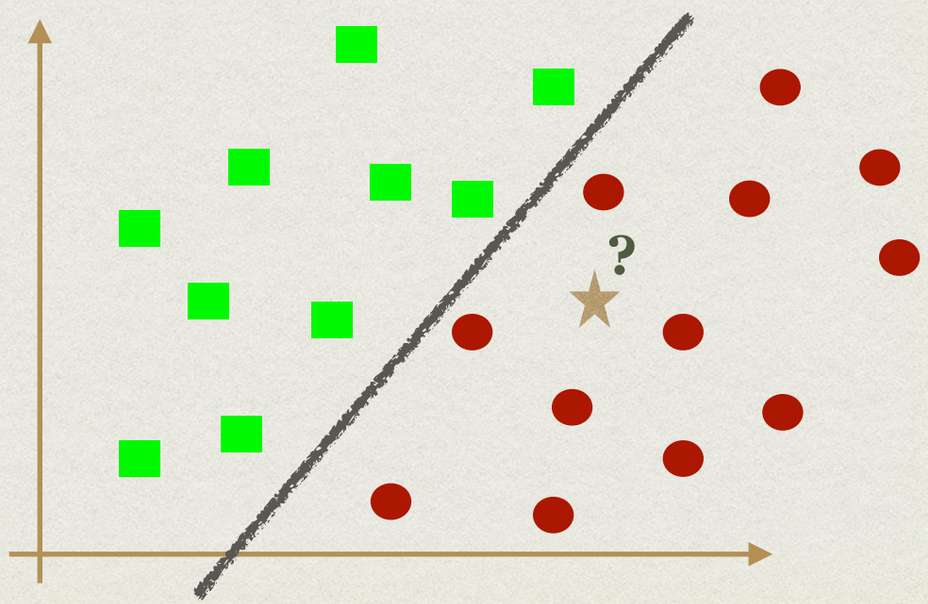
unknown fitting parameters: $x \in \mathbb{R}^p$

$$y = \varphi(Fx)$$

component-wise function

Goal: Minimize prediction/generalization error.

Example: $\varphi(z) = \operatorname{sign}(z)$



BASIC QUESTIONS:

- Given (y,F) and φ
- High-dimensional limit: α = n/p = Θ(1) while $n, p \to \infty$

What is the minimal achievable prediction error?

What is the minimal efficiently achievable prediction error?

Answer: Depends on (y,F) …

PROBABILISTIC MODEL

Gardner, Derrida'89; Mézard'89; ..... many more papers in statistical physics.

- Teacher:
  - F iid components, $F_{\mu i} \sim \mathcal{N}(0, 1/p)$ $F \in \mathbb{R}^{n \times p}$
  - x* iid components, $x_i^* \sim P_X$
  - Generates: $y = \varphi_\xi(Fx^*)$ noise: ξ
- Student:
  - Aims to estimate x from (y,F). Knows P_X, and φ.

Teacher-student setting

WHY TEACHER/STUDENT SETTING?

- Tractable theory (including constants), complementary to the worst-case theories.

- Sometimes realistic (e.g. compressed sensing, super-position error correcting codes, code multiple-division access problem).

- Interesting: phase transitions and algorithmic gaps.

- Challenging: Formula for optimal prediction error conjectured by Gardner, Derrida'89; Gyorgyi'90. Partial results in Talagrand'03: "Spin Glasses: A Challenge for Mathematicians". Proof in Barbier, Krzakala, Macris, Miolane, LZ, arXiv:1708.03395, COLT'18.

BAYES-OPTIMAL ESTIMATION

for the teacher-student setting

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

- x* is generated from P_X, y is generated from

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

- x* is then "forgotten" and has to be recovered from F and y.

- Bayes optimal estimator $\hat{x}_i$ = posterior mean of x_i.
  Optimal because it minimises

$$\text{MSE} = \frac{1}{p} \sum_{i=1}^{p} (x_i^* - \hat{x}_i)^2$$

🙂

BAYES-OPTIMAL GENERALIZATION

for the teacher-student setting

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

- x* is generated from P_X, y is generated from

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

- New row, F_{new}, is given; predict the label that x* would generate, without knowing x*.

- Bayes optimal prediction:

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

COMPUTATIONAL CAVEAT

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

- High-dimensional limit: $\alpha = n/p = \Theta(1)$ while $p, n \to \infty$
- Computing posterior averages:
  - Easy for Gaussian P_X and P_{out} (ridge regression).
  - NP-hard for general P_X and P_{out}. 😕

PERCEPTRON

highlight of the results

TEACHER-STUDENT PERCEPTRON

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

- introduce the model -

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.

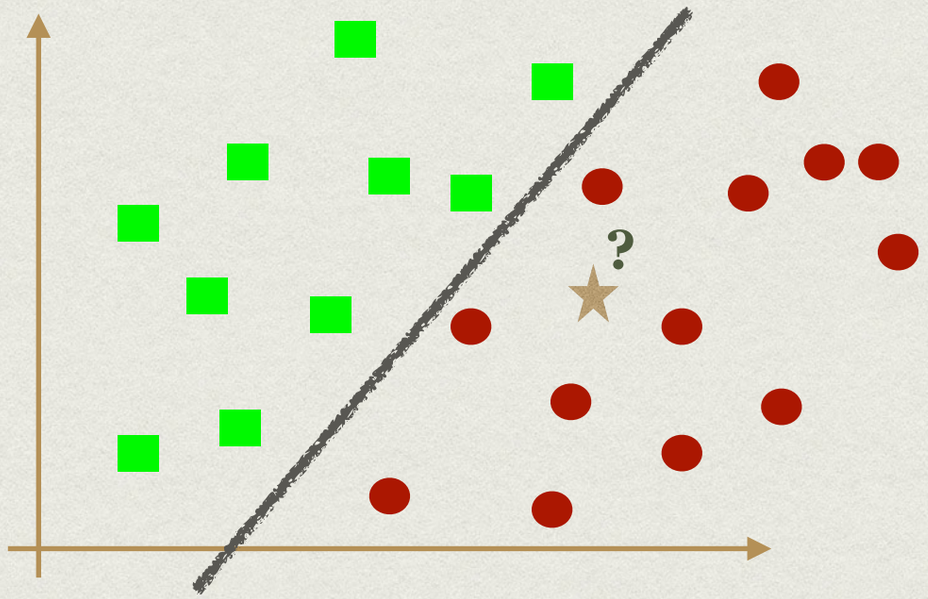
coordinates of points: $F_{\mu i} \sim \mathcal{N}(0, 1/p)$

hyper-plane parameters: $x_i^* \sim P_X$

labels ■ ● $y = \mathrm{sign}(Fx^*)$

Gyorgyi'90: conjectures formula for optimal generalisation error using replica method from disordered systems.

$$\alpha = \frac{n}{p} = \Theta(1) \qquad n, p \to \infty$$

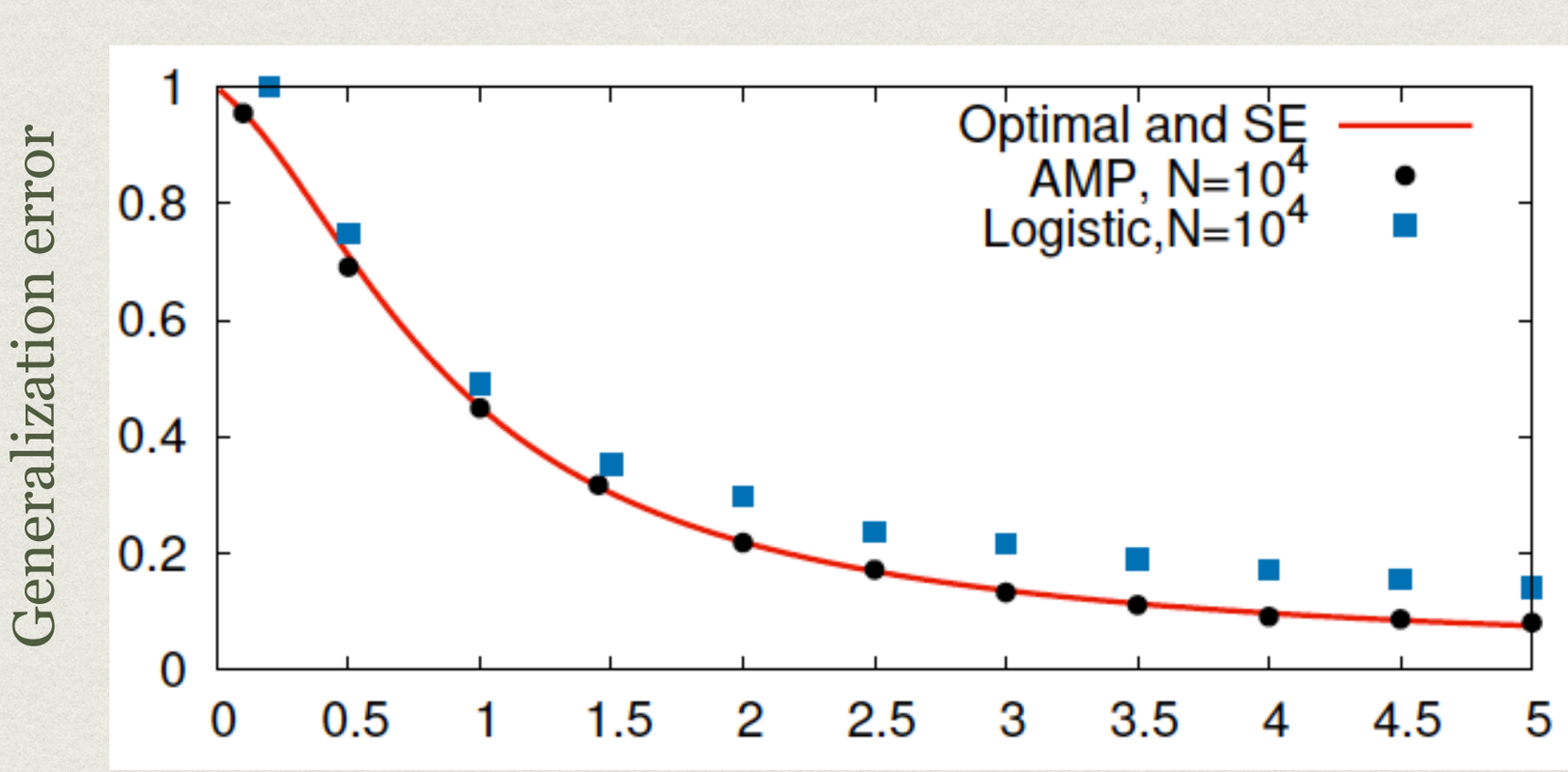


p=2, n=22, α=n/p=11

GAUSS-BERNOULLI PERCEPTRON

$$P_X(x) = \rho\mathcal{N}(0,1) + (1-\rho)\delta(x)$$
$$\rho = 0.2$$

$$n, p \to \infty$$
$$\alpha = \frac{n}{p} = \Theta(1)$$

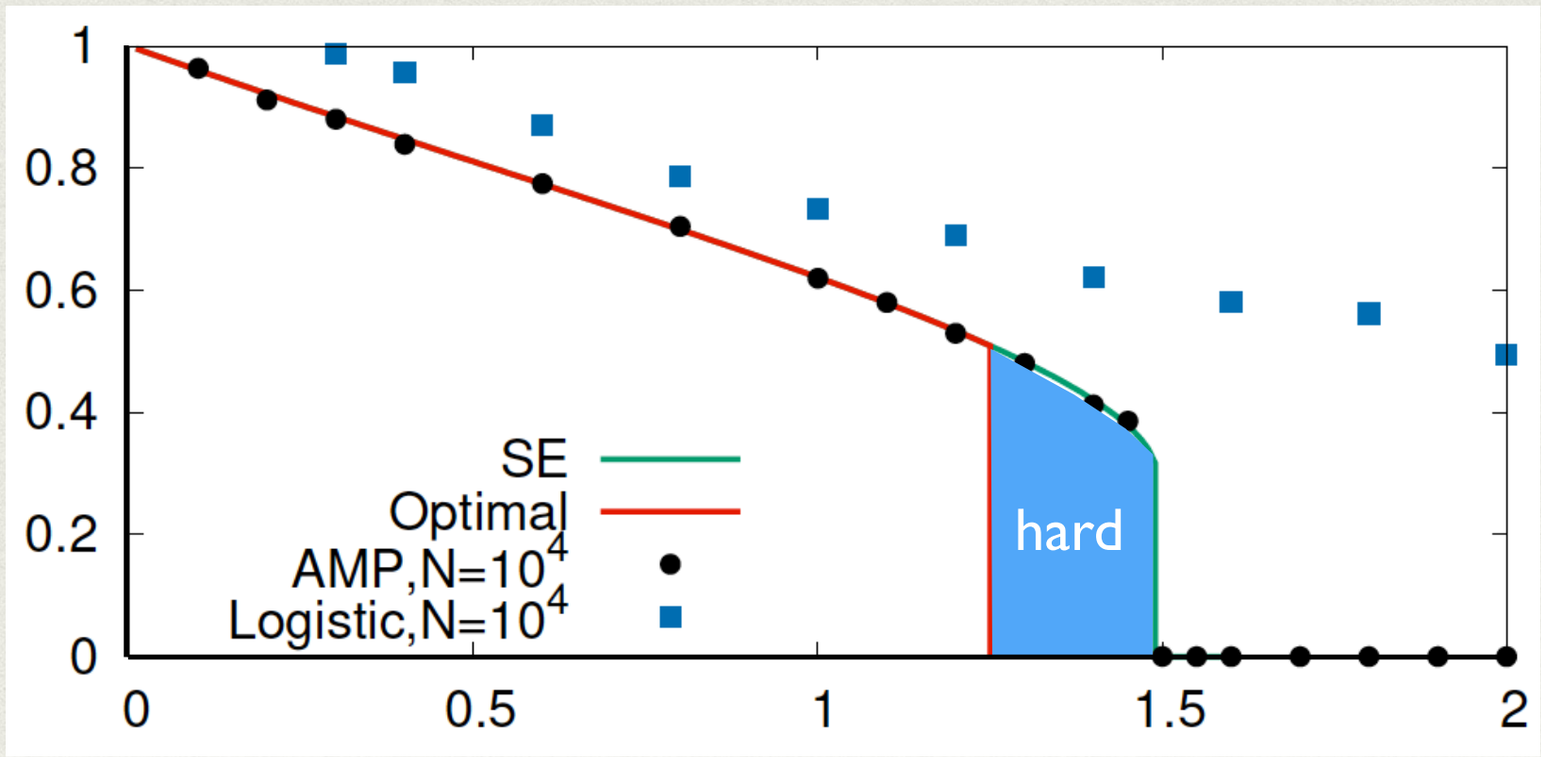


BINARY PERCEPTRON

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

$$n, p \to \infty$$

$$\alpha = \frac{n}{p} = \Theta(1)$$



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

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

OUTLINE OF THE REST

- Formula for the optimal error (+ proof idea).
- Approximate message passing (AMP) algorithms reaching optimality (out of a hard phase).
- Hard phase conjecture: AMP gets lowest error from all polynomial algorithms.

REMINDER OF THE SETTING

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

- F random iid, $F_{\mu i} \sim \mathcal{N}(0, 1/p)$
- x* is generated from P_X, y is generated from $P_{\text{out}}(y|Fx^*)$
- High-dimensional limit: $\alpha = n/p = \Theta(1)$ while $p, n \to \infty$

MAIN THEOREMS

Barbier, Krzakala, Macris, Miolane, LZ 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):

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

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

where

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

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

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

MAIN THEOREMS

Barbier, Krzakala, Macris, Miolane, LZ 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):

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

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

Theorem 2 (optimal estimation error, informally):

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

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

MAIN THEOREMS

Barbier, Krzakala, Macris, Miolane, LZ 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):

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

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

Theorem 3 (optimal generalisation error, informally):

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

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

$\rho = \mathbb{E}_{P_X}(x^2)$

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

$\xi \sim P_\xi$

SELECTED RELATED WORK I

Replica method of statistical physics leads to the same free entropy and expression for optimal error.

> Gardner, Derrida'89, Gyorgyi'90, Sompolinsky, Tishby, Seung'90 free entropy conjectured for the teacher-student setting of binary perceptron. Followed by large activity of statistical physicist on neural networks, reviewed Engel, Broeck'01.
>
> In information and communication theory the replica formula known as the Tanaka'02 formula for the CDMA problem.
>
> In compressed sensing Guo, Baron, Shamai'09, Wu, Verdu'11, Krzakala, Mezard, Sausset, Sun, LZ'12.

Previous rigorous proof only for a special case of the linear output with Gaussian noise: Barbier, Dia, Macris, Krzakala'16, and independently Reeves, Pfister'16.

PROOF IDEA

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

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

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

$$y' = \sqrt{\hat{m}}x^* + \xi \qquad \xi \sim \mathcal{N}(0,1) \qquad x^* \sim P_X$$

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

$$\tilde{y} \sim P_{\rm out}(\tilde{y}|\sqrt{m}\,v + \sqrt{\rho - m}\,z^*) \qquad v, z^* \sim \mathcal{N}(0,1) \qquad \rho = \mathbb{E}_{P_X}(x^2)$$

PROOF IDEA

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

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

Interpolating Hamiltonian (=log-likelihood):

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

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

PROOF IDEA

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

Interpolating free entropy:

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

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

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

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

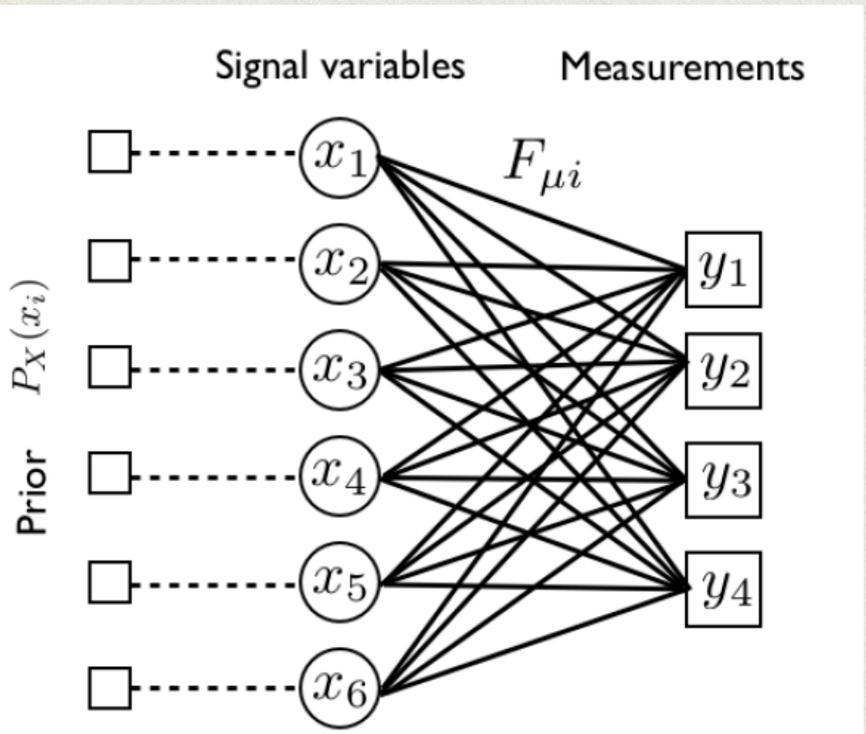
$$\mathbb{E}_{x^*,y}\mathbb{E}_{P(x|y)}[g(y,x^{(1)},x^*)] = \mathbb{E}_y\mathbb{E}_{P(x|y)}[g(y,x^{(1)},x^{(2)})]$$

□

IS THE OPTIMAL ERROR REACHABLE WITH EFFICIENT ALGORITHMS?

APPROXIMATE MESSAGE PASSING

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



Belief propagation (BP):

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

$$m_{\mu\to i}(x_i) = \frac{1}{z_{\mu\to i}} \int \prod_{j\neq i} [\mathrm{d}x_j m_{j\to\mu}(x_j)] \, P_{\text{out}}(y_\mu | \sum_l F_{\mu l} x_l)$$

The p-dimensional integral in BP is algorithmically intractable … but when $p \to \infty$ we can simplify into approximate message passing.

Algorithm 2 Generalized Approximate Message Passing (G-AMP)

Input: y

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

repeat

AMP Update of ω_μ, V_μ

$$V_\mu^t \leftarrow \sum_i F_{\mu i}^2 v_i^{t-1}$$

$$\omega_\mu^t \leftarrow \sum_i F_{\mu i} a_i^{t-1} - V_\mu^t g_{\text{out},\mu}^{t-1}$$

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

$$g_{\text{out},\mu}^t \leftarrow g_{\text{out}}(\omega_\mu^t, y_\mu, V_\mu^t)$$

$$\Sigma_i^t \leftarrow \left[-\sum_\mu F_{\mu i}^2 \partial_\omega g_{\text{out}}(\omega_\mu^t, y_\mu, V_\mu^t) \right]^{-1}$$

$$R_i^t \leftarrow a_i^{t-1} + \Sigma_i^t \sum_\mu F_{\mu i} g_{\text{out},\mu}^t$$

Onsager terms

AMP Update of the estimated marginals a_i, v_i

$$a_i^t \leftarrow f_a(\Sigma_i^t, R_i^t)$$

$$v_i^t \leftarrow f_v(\Sigma_i^t, R_i^t)$$

t ← t + 1

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

output: a,**v**.

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

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

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

Algorithm 2 Generalized Approximate Message Passing (G-AMP)

Input: y

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

repeat

AMP Update of ω_μ, V_μ

$$V_\mu^t \leftarrow \sum_i F_{\mu i}^2 v_i^{t-1}$$

$$\omega_\mu^t \leftarrow \sum_i F_{\mu i} a_i^{t-1} - V_\mu^t g_{\text{out},\mu}^{t-1}$$

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

$$g_{\text{out},\mu}^t \leftarrow g_{\text{out}}(\omega_\mu^t, y_\mu, V_\mu^t)$$

$$\Sigma_i^t \leftarrow \left[-\sum_\mu F_{\mu i}^2 \partial_\omega g_{\text{out}}(\omega_\mu^t, y_\mu, V_\mu^t) \right]^{-1}$$

$$R_i^t \leftarrow a_i^{t-1} + \Sigma_i^t \sum_\mu F_{\mu i} g_{\text{out},\mu}^t$$

AMP Update of the estimated marginals a_i, v_i

$$a_i^t \leftarrow f_a(\Sigma_i^t, R_i^t)$$

$$v_i^t \leftarrow f_v(\Sigma_i^t, R_i^t)$$

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

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

output: a,**v**.

Onsager terms

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

GAMP for prediction:

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

STATE EVOLUTION

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

m^t in the AMP algorithm ($n, p \to \infty, \alpha = \Theta(1)$) evolves as:

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

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

Recall the RS free entropy

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

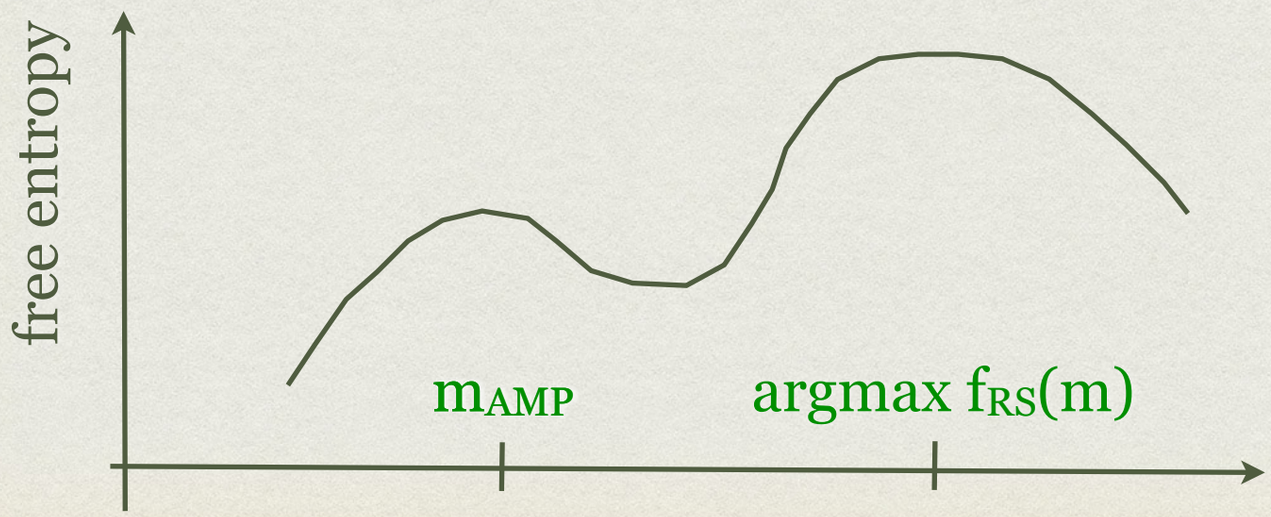
BOTTOM LINE

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

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

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

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



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

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

SELECTED RELATED WORK II

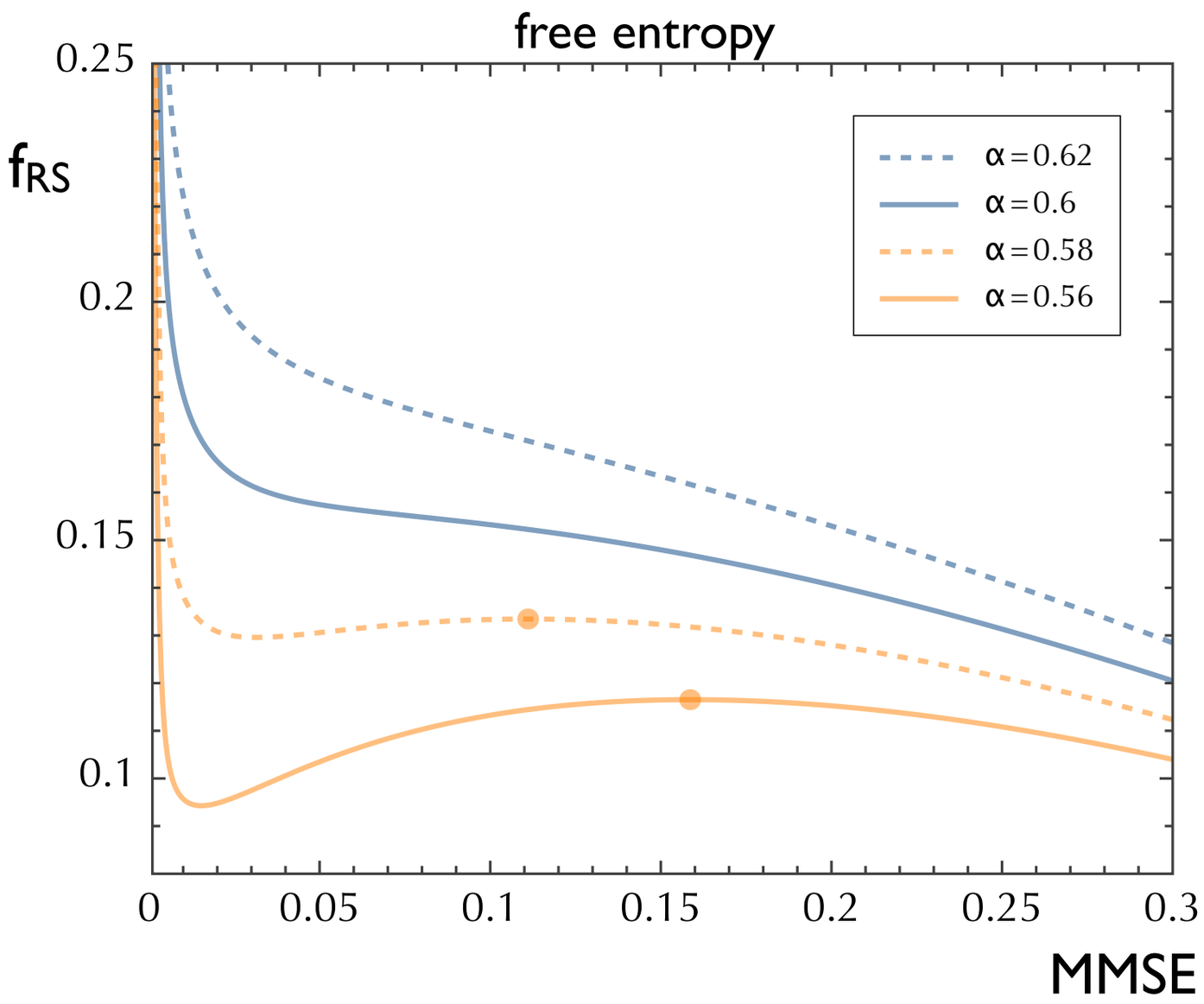
Idea from Thouless-Anderson-Palmer'76 but problem with time-indices and hence with convergence. Resolved by Bolthausen in ~2008.

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 in Bolthausen'08 for another model, regression by Bayati, Montanari'11 for Gaussian matrices and output, and by Bayati, Lelarge, Montanari'12 for general iid matrices, and Gaussian output. General output and Gaussian matrices in Javanmard, Montanari'13.

RESULTS

PHASE TRANSITIONS



Compressed sensing:

P_x Gauss-Bernoulli(ρ)

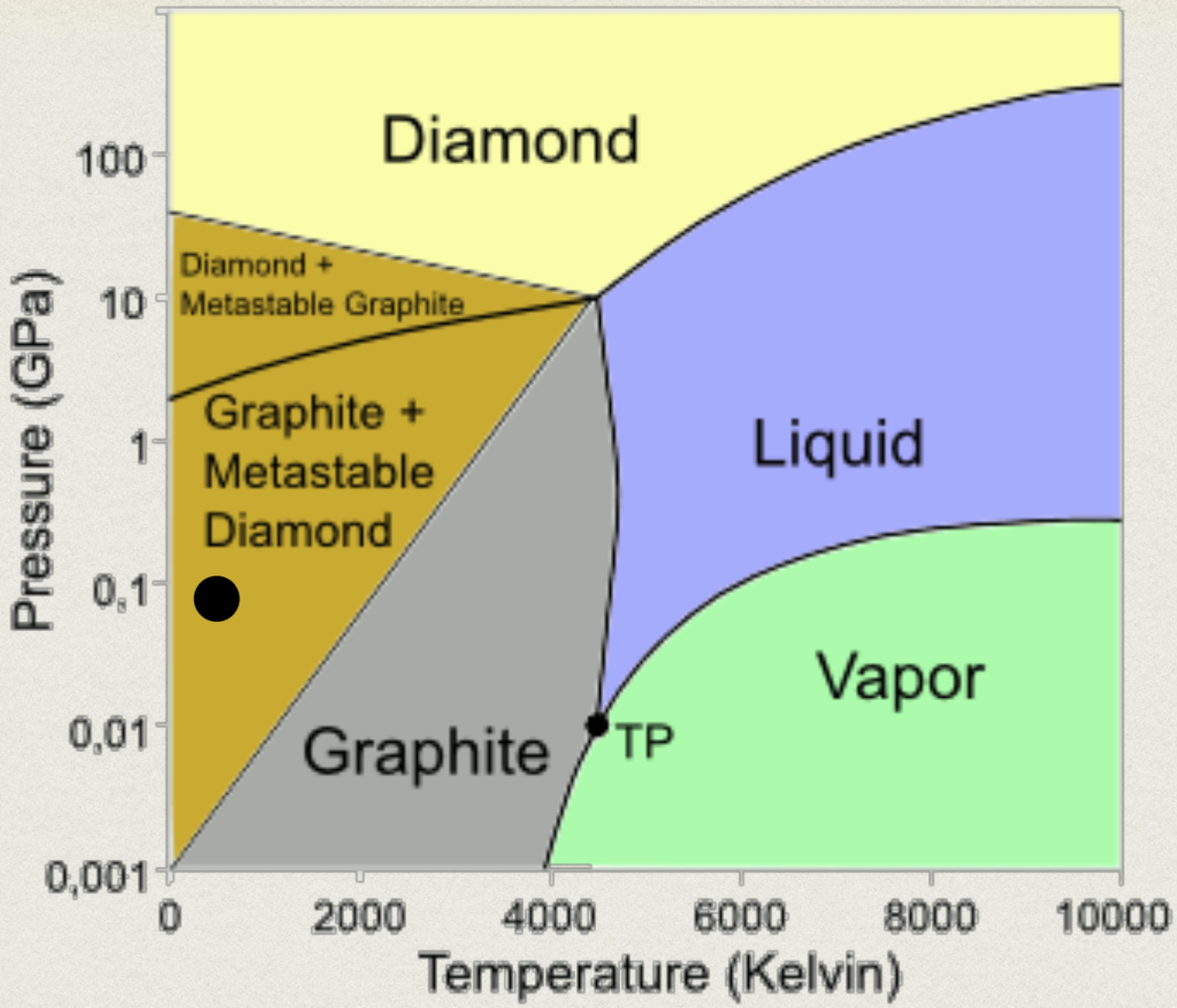
Noiseless linear output.

Fraction of non-zeros in x.

$\rho = 0.4$

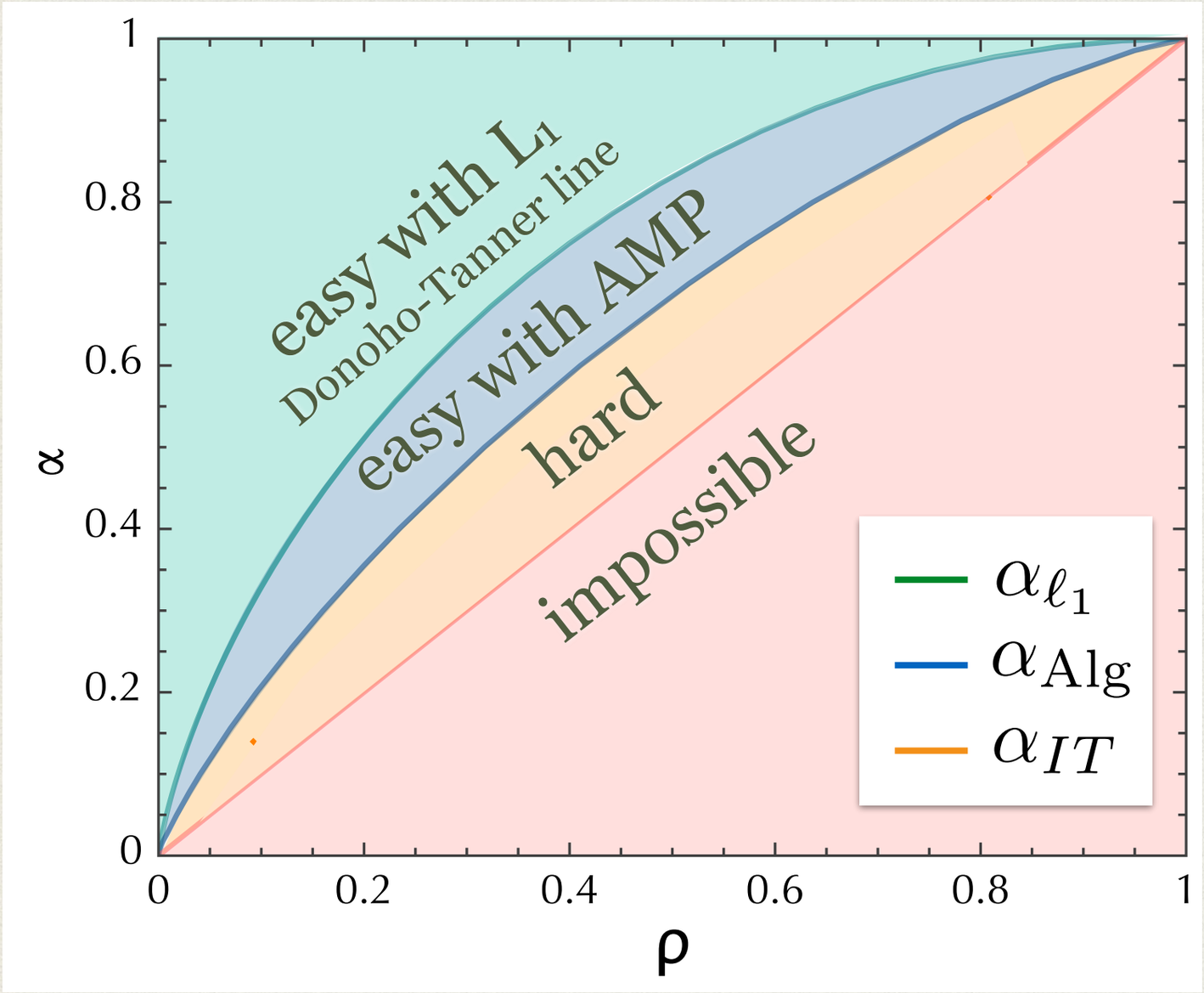
Non-zeros are Gaussians(0,1)

HARD PHASE IN NATURE



Metastable diamond = high error. Equilibrium graphite = low error. Algorithms are stuck at high error for exponential time.

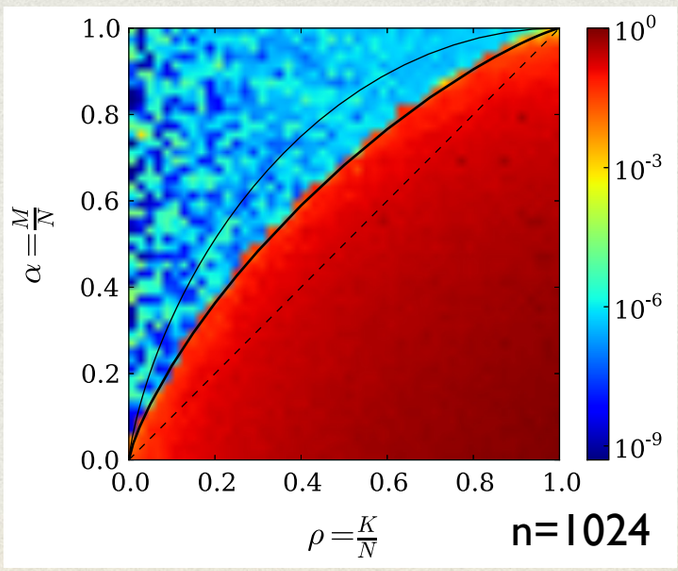
PHASE DIAGRAM OF (NOISELESS) SPARSE LINEAR ESTIMATION



P_x Gauss-Bernoulli(ρ)

AMP = approximate message passing

Hard for all known polynomial algorithms: $\alpha_{IT} < \alpha < \alpha_{\mathrm{Alg}}$



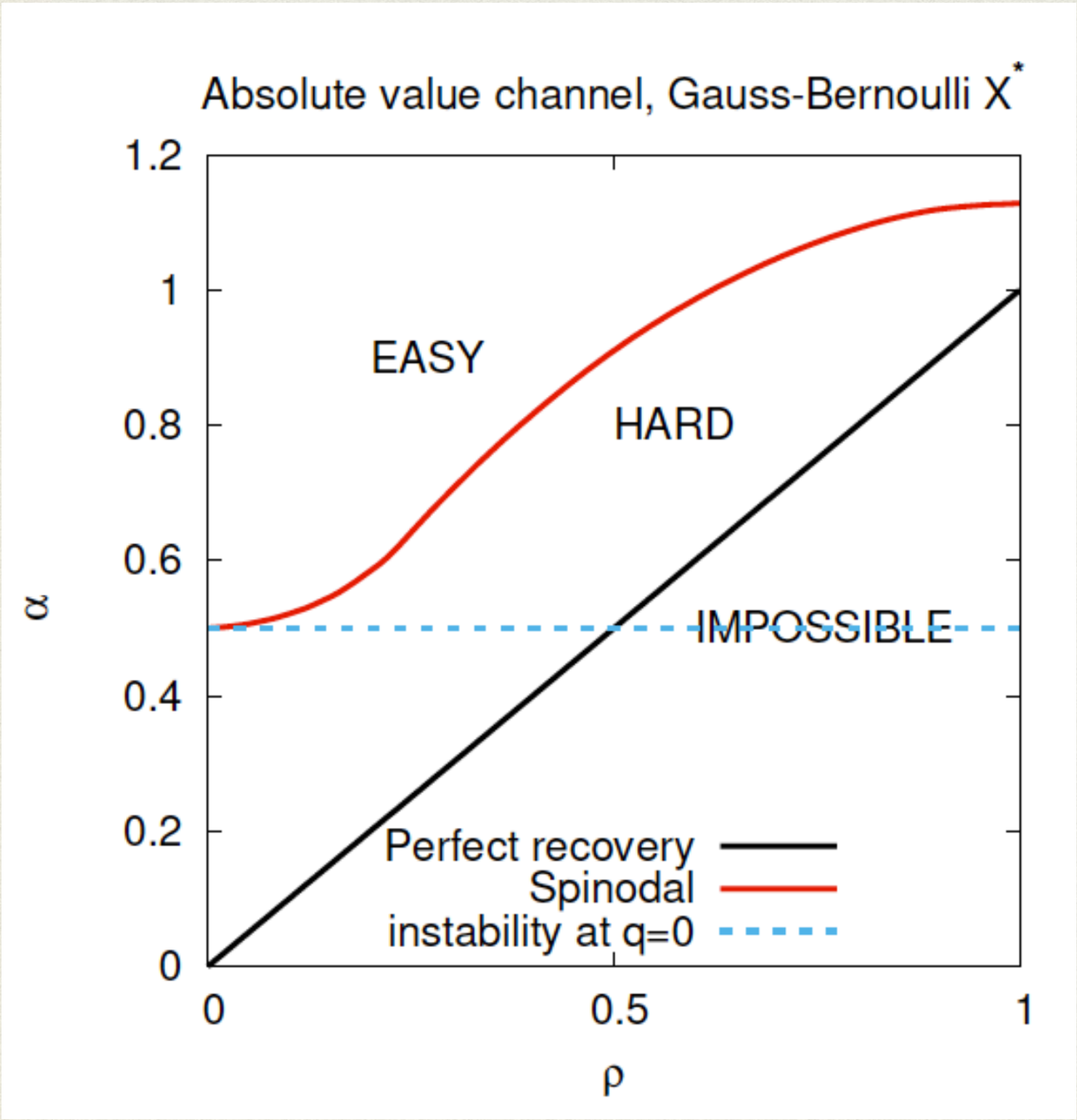
HARD PHASE EVERYWHERE!

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; ...

SIGN-LESS COMPRESSED SENSING

real-valued phase retrieval

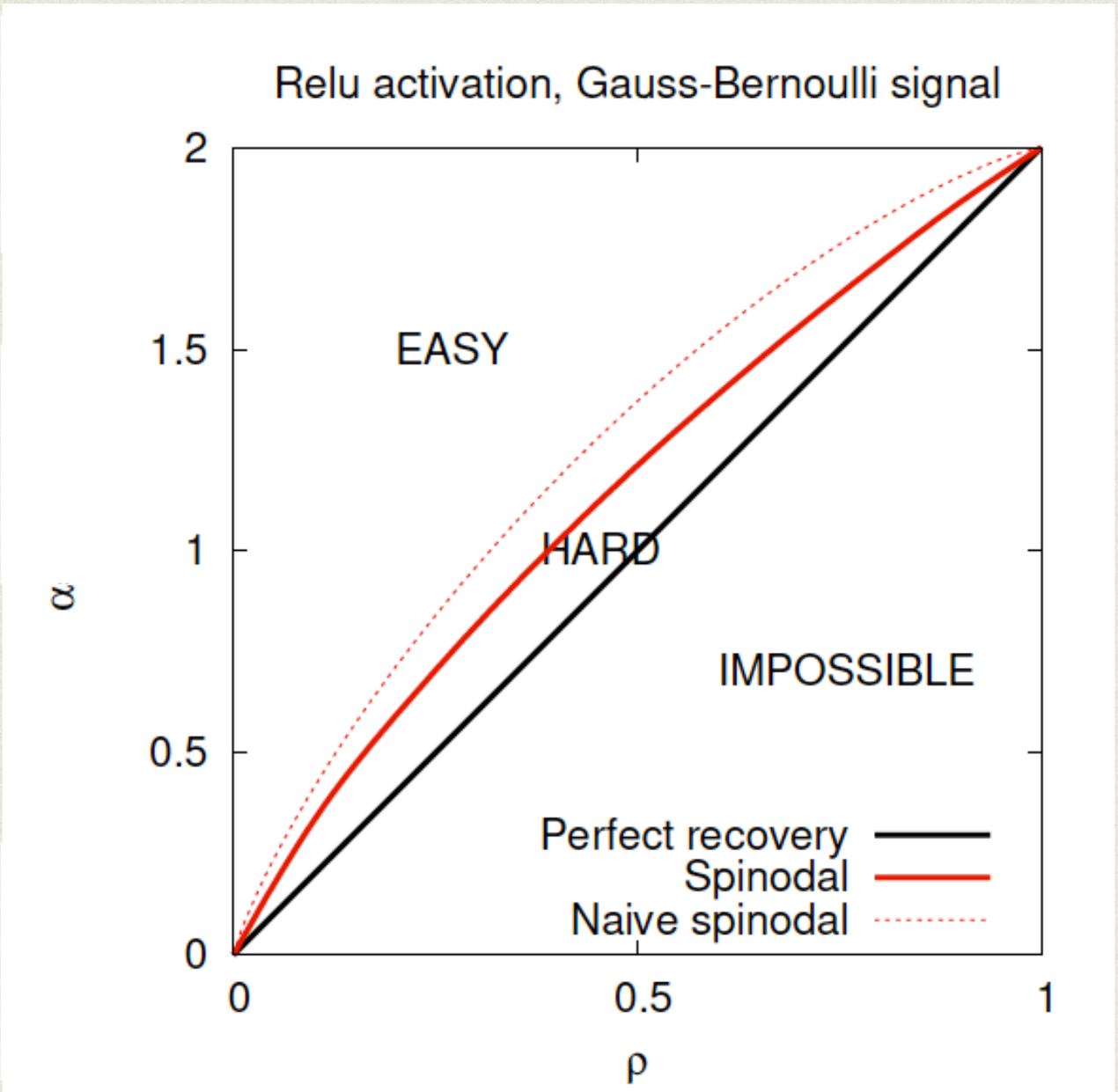


$$y = |Fx^*|$$

P_X Gauss-Bernoulli(ρ)

You cannot sense compressively if you lost the signs!

RELU COMPRESSED SENSING



$$y = \max(0, Fx^*)$$

P_x Gauss-Bernoulli(ρ)

Zeros in relu are not helpful information-theoretically, but algorithmically they are.

PERCEPTRON

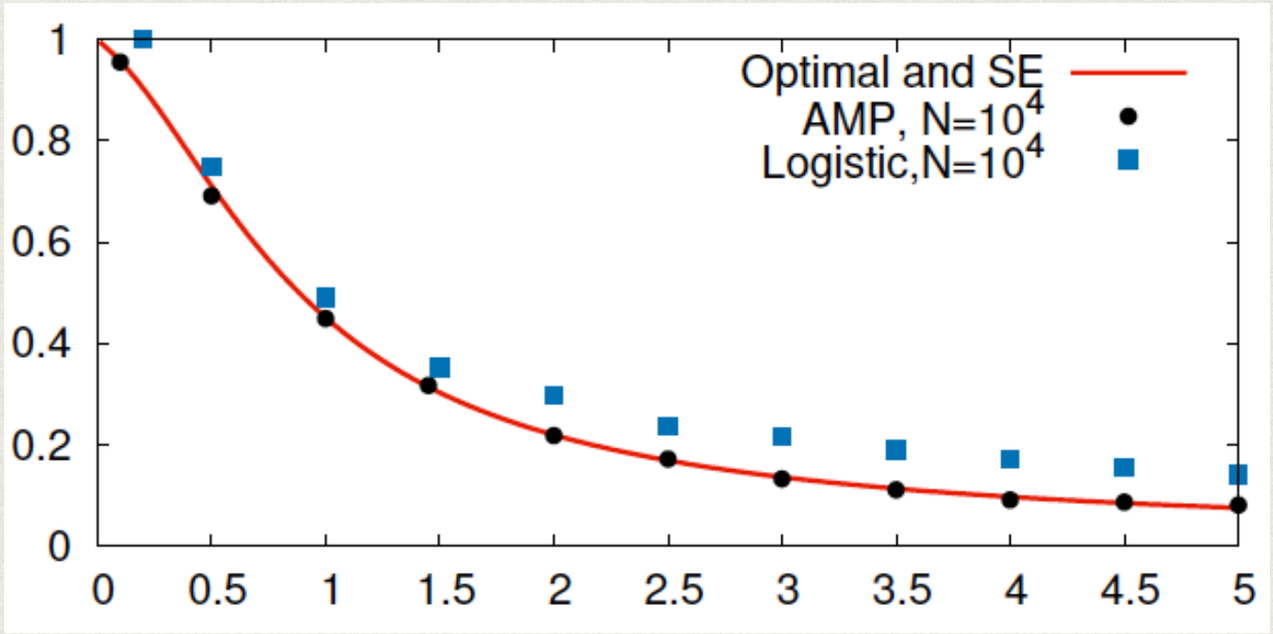
GAUSS-BERNOULLI PERCEPTRON

OR 1-BIT COMPRESSED SENSING

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

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

$$\rho = 0.2$$

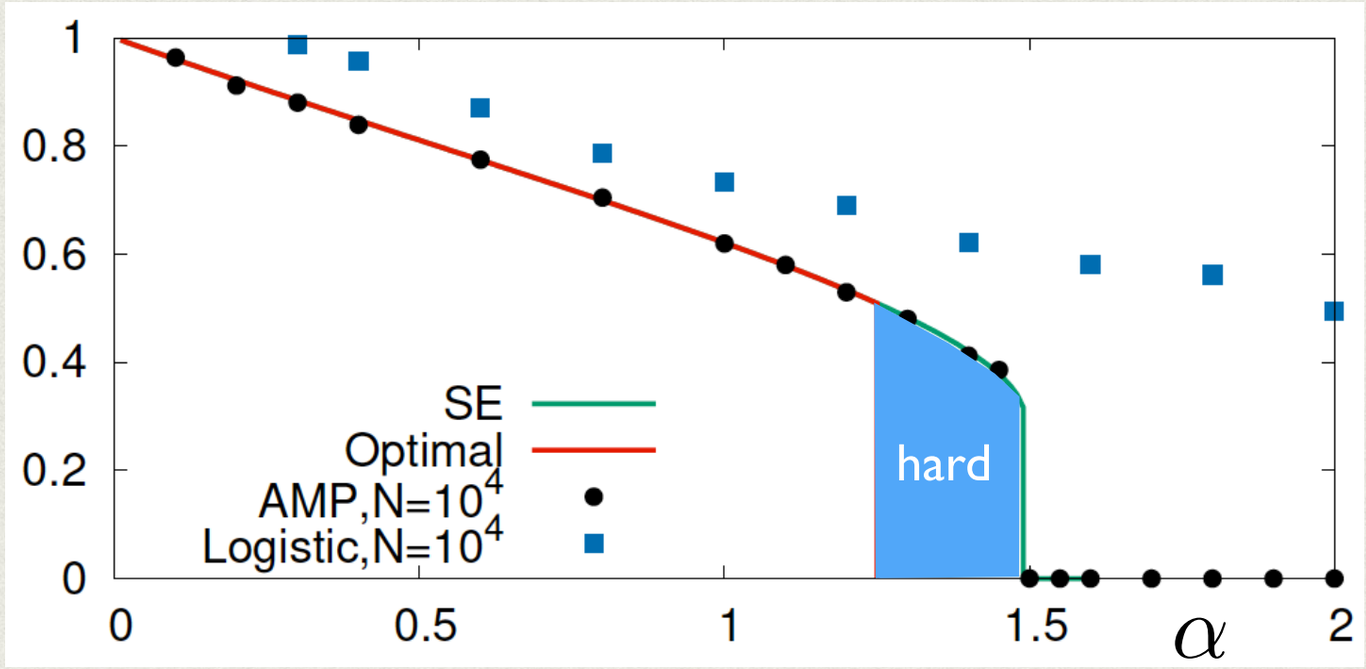


BINARY PERCEPTRON

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

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

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



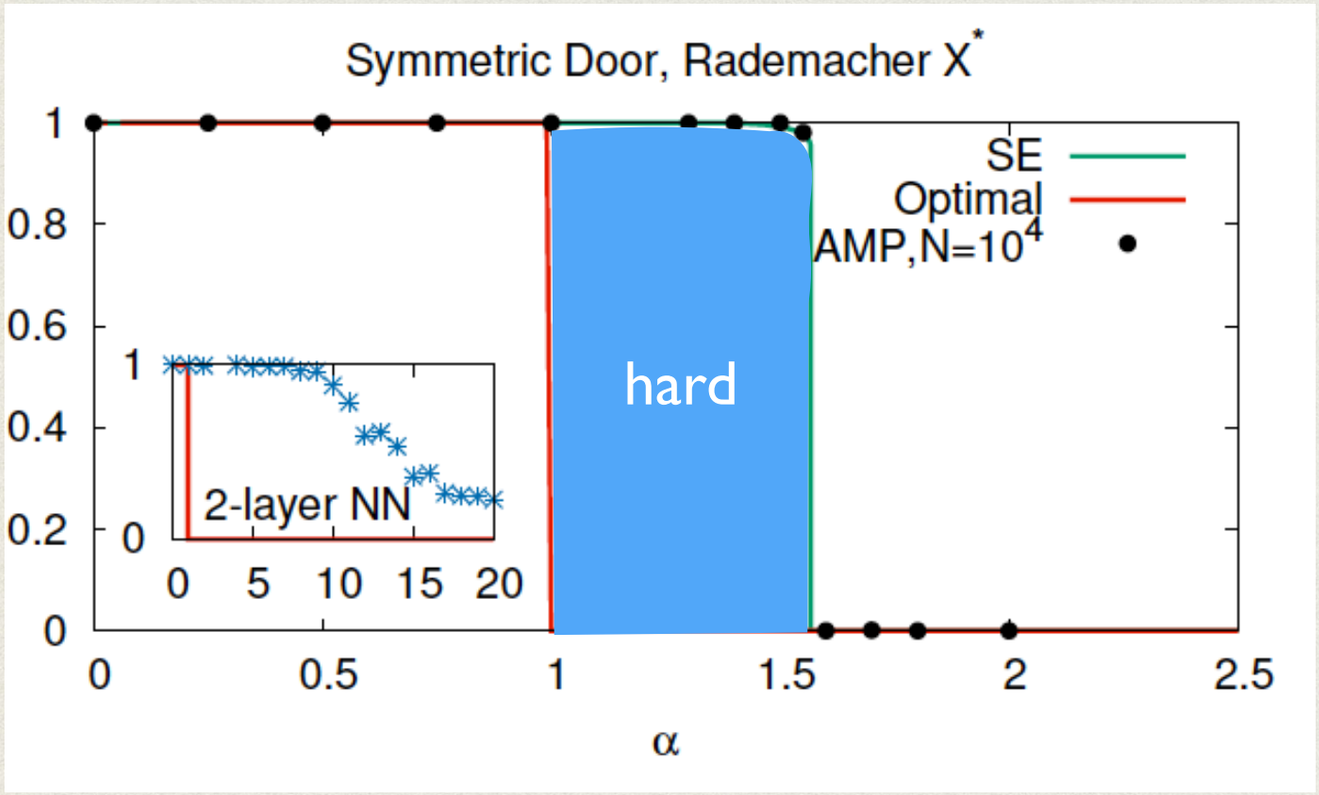
$$\alpha_{\mathrm{IT}} = 1.249$$

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

SYMMETRIC BINARY PERCEPTRON

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

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



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

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

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

Very simple yet very hard benchmark for classification!

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

INCLUDING HIDDEN VARIABLES

Committee machine

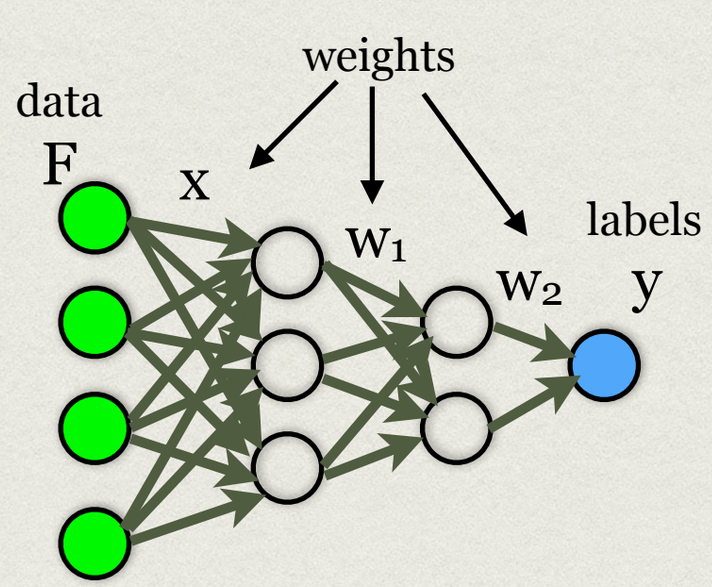
Teacher-student model studied in Schwarze, Herz'92. Proof and approximate message passing (Aubin, Maillard, Barbier, Macris, Krzakala, LZ'19), submitted to NIPS.

p input units

K hidden units

output unit

n training samples



L=3 layers

x learned, w fixed

Limit: $\alpha = \dfrac{n}{p} = \Theta(1)$ $\quad K = \Theta(1)$ $\quad n, p \to \infty$

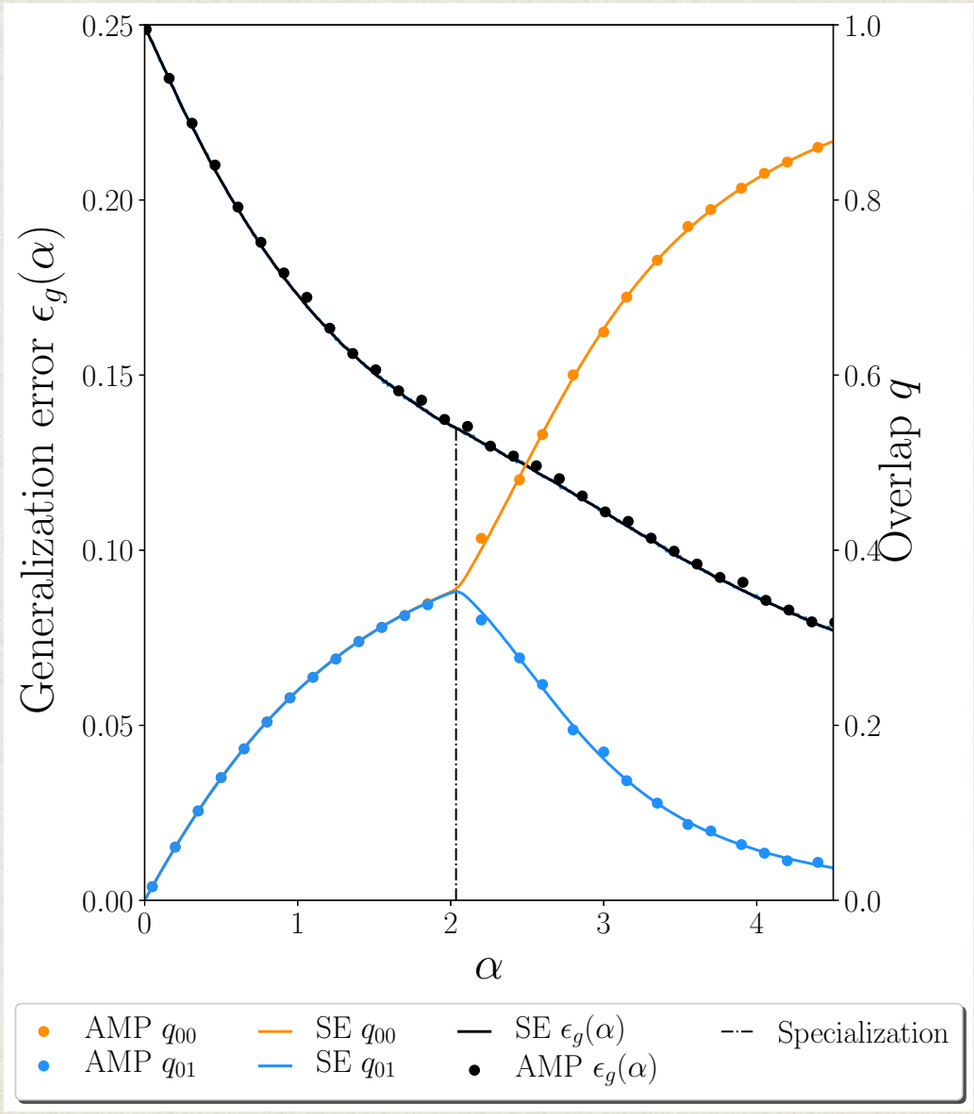
COMMITTEE MACHINE

K=2

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

$$y_\mu = \text{sign}[\text{sign}(\sum_{i=1}^{p} F_{\mu i} x^*_{i1}) + \text{sign}(\sum_{i=1}^{p} F_{\mu i} x^*_{i2})]$$

- Specialization phase transition



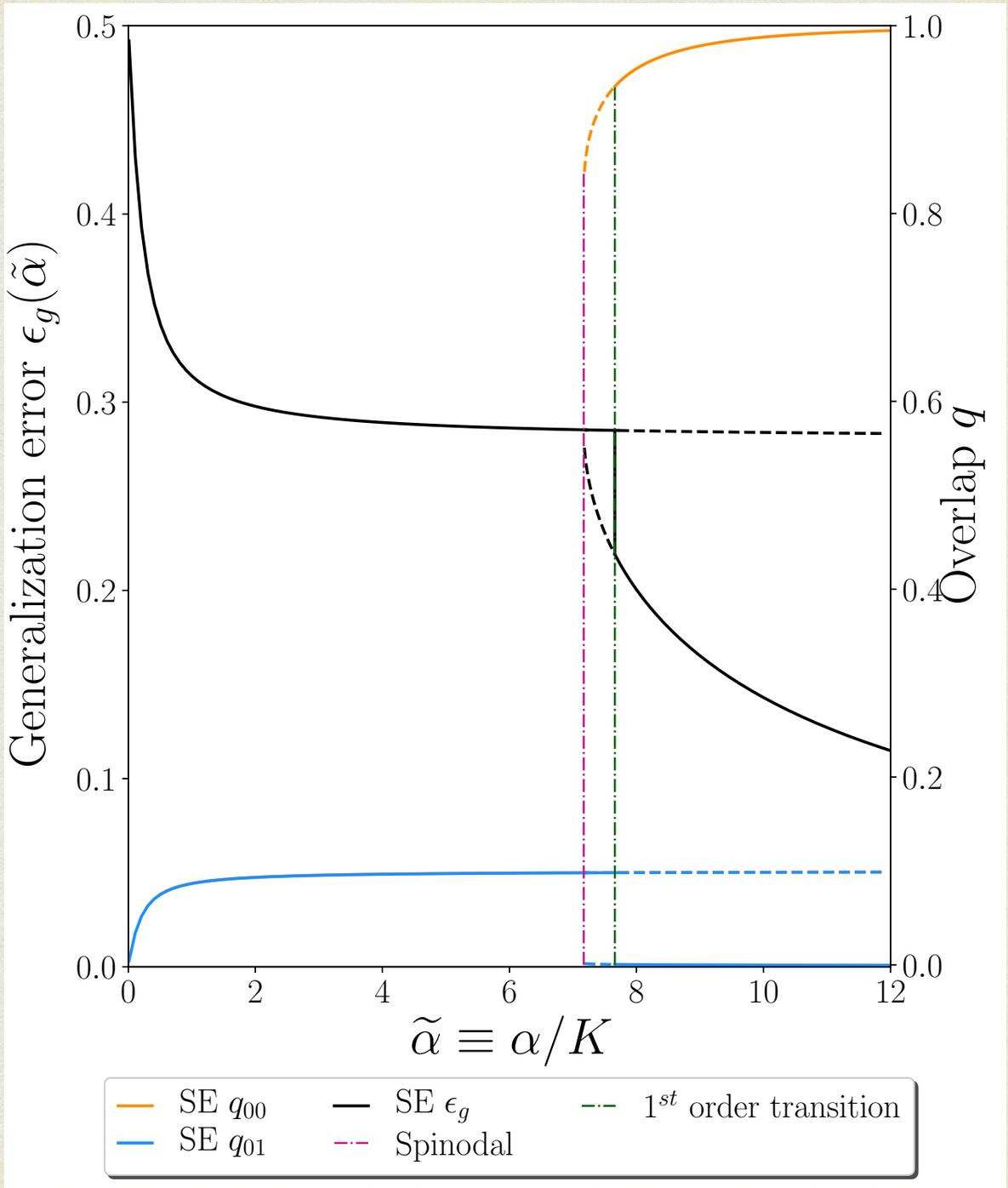
COMMITTEE MACHINE

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

$K \gg 1$

- Specialization phase transition
- Large algorithmic gap:
  - IT threshold: $n > 7.65Kp$
  - Algorithmic threshold

$$n > \mathrm{const.}K^2 p$$



OTHER EXTENSIONS

- Teacher's model mismatching the student's model.

  Replica predictions known (maybe replica symmetry breaking), no proof yet.

- Beyond random iid data F.

  So far random orthogonally invariants matrices solved (Kabashima, arxiv: 0808.3900) and partially proven (Gabrie, Manoel, Barbier, Luneau, Macris, Krzakala, LZ, arxiv:1805.09785).

- Beyond separable priors P_X.

  Priors coming from another GLM solved (Manoel, Krzakala, Mezard, LZ arxiv: 1701.06981 ; Rangan, Fletcher arxiv: 1706.09549; Reeves arxiv:1710.04580).

CONCLUSION

- Teacher student setting for the Generalized linear model.
- Analysis of the Bayes-optimal estimator for random iid matrices, and separable output and prior.
- Determination of the phase transition and their algorithmic meaning, approximate message passing.
- Hardness of computational tasks in probabilistic setting (still a lot to be done).

Thank you for your attention!