STATISTICAL PHYSICS OF COMPUTATIONAL PROBLEMS

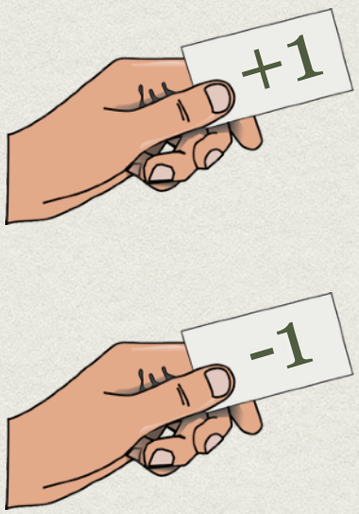
Lenka Zdeborová
(IPhT CEA Saclay)

Bezout working group, May 14th, 2019

CO-RESPONSIBLE

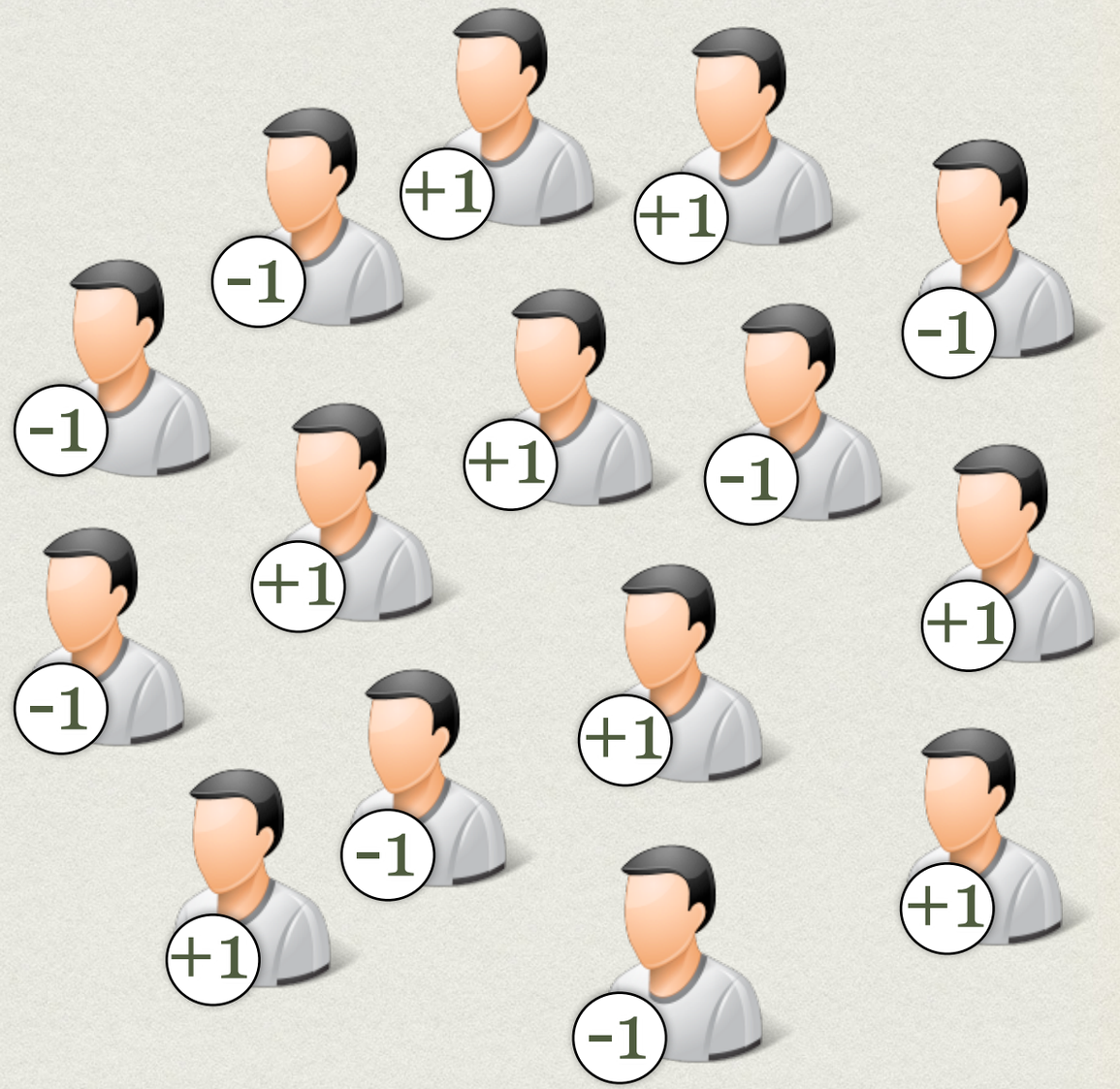
Madhu Advani, Ahmed El Alaoui, Fabrizio Antenucci, Maria-Chiara Angelini, John Ardelius, Benjamin Aubin, Jess Banks, Jean Barbier, Giulio Biroli, Alfredo Braunstein, Francesco Caltagirone, Chiara Cammarota, Michele Castellana, Michael Chertkov, Andrea Crisanti, Amin Coja-Oghlan, Luca Dall'Asta, Varsha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

LET'S PLAY A GAME

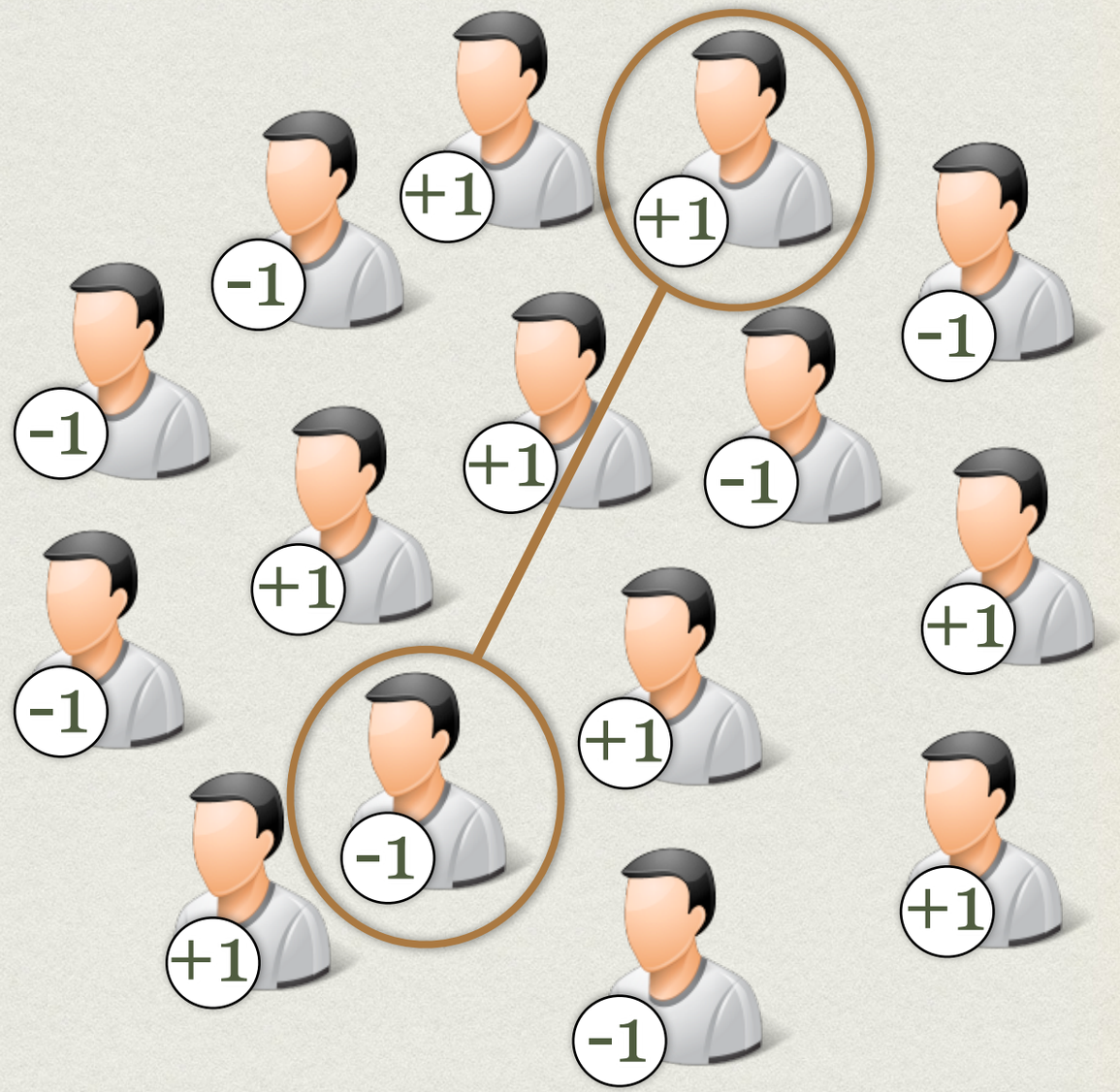


N=15 people

LET'S PLAY A GAME



LET'S PLAY A GAME



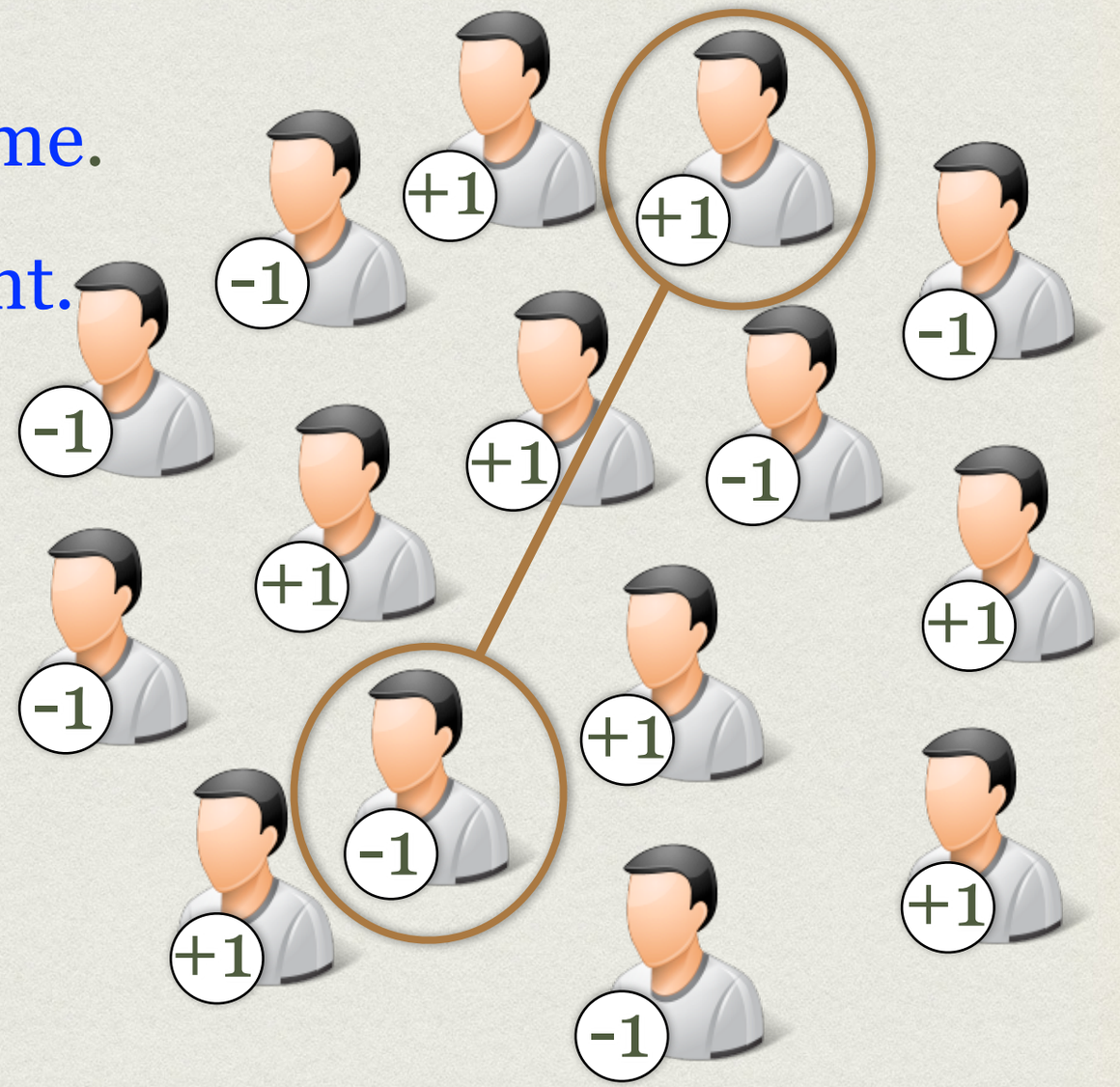
LET'S PLAY A GAME

- Each pair reports:
  - $Y_{ij}=Z_{ij}+1/\sqrt{N}$ if cards the same.
  - $Y_{ij}=Z_{ij}-1/\sqrt{N}$ if cards different.

$$Z_{ij} \sim \mathcal{N}(0,\Delta^*)$$

Collect Y_{ij} for every pair (ij).

Goal: Recover cards (up to symmetry) purely from the knowledge of $\mathbf{Y} = \{Y_{ij}\}_{i<j}$



CHALLENGE

- High-dimensional (non-convex) problem.
- No statistical consistency as $N \to \infty$.

Outside the box of "traditional" statistical learning.

HOW TO SOLVE THIS?

$$Y_{ij} = \frac{1}{\sqrt{N}} x_i^* x_j^* + Z_{ij}$$

true values of cards: $x^* \in \{-1, +1\}^N$

$x_i^* \in \{-1, +1\}$

$$Z_{ij} \sim \mathcal{N}(0, \Delta^*)$$

x_{PCA} = leading eigenvector of Y estimates x* (up to a sign).

BBP phase transition: $\Delta^* > 1$ $\quad x_{\mathrm{PCA}} \cdot x^* \approx 0$

Watkin, Nadal'94
Baik, BenArous, Pechet'04 $\quad \Delta^* < 1$ $\quad |x_{\mathrm{PCA}} \cdot x^*| > 0$

PCA: not optimal error value (does not maximise the number of correctly assigned cards)

BAYESIAN INFERENCE

$$P(x|Y) = \frac{P(x)P(Y|x)}{P(Y)}$$

Values of cards: $x \in \{-1, +1\}^N$

$x_i \in \{-1, +1\}$

Posterior distribution:

$$P(x|Y) = \frac{1}{Z(Y,\Delta)} \prod_{i=1}^{N} [\delta(x_i + 1) + \delta(x_i - 1)] \prod_{i<j} e^{-\frac{(Y_{ij} - x_i x_j/\sqrt{N})^2}{2\Delta}}$$

Bayes-optimal inference = computation of marginals

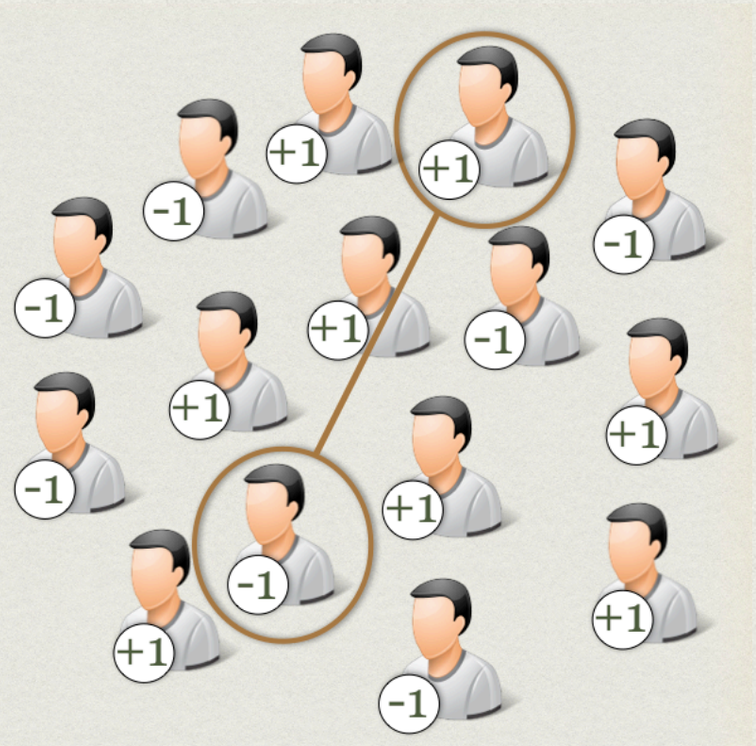
Computationally expensive in general (#P-hard)

How do we compute the Bayes-optimal performance?

Map to a spin glass?

$$Y \rightarrow J$$

$$x_i \rightarrow S_i$$



BACK TO THE CARD GAME

$$P(S|J) = \frac{1}{Z(J,\Delta)} \prod_{i<j} e^{-\frac{1}{2\Delta}(J_{ij} - S_i S_j/\sqrt{N})^2}$$

$$S_i \in \{-1, +1\}$$

Hamiltonian

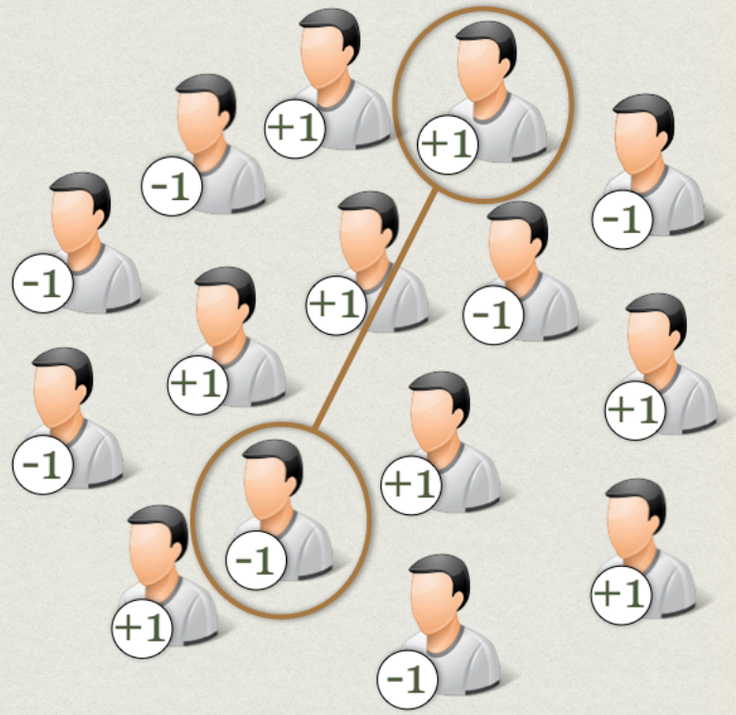
Boltzmann distribution

$$P(S|J) = \frac{1}{\tilde{Z}(J,\Delta)} e^{-\frac{1}{\Delta\sqrt{N}} \sum_{i<j} J_{ij} S_i S_j}$$

Partition function

Mean-field Ising spin glass
(Sherrington-Kirkpatrick'75 model)

J_{ij} conditioned on S_i^*: *planted* disorder



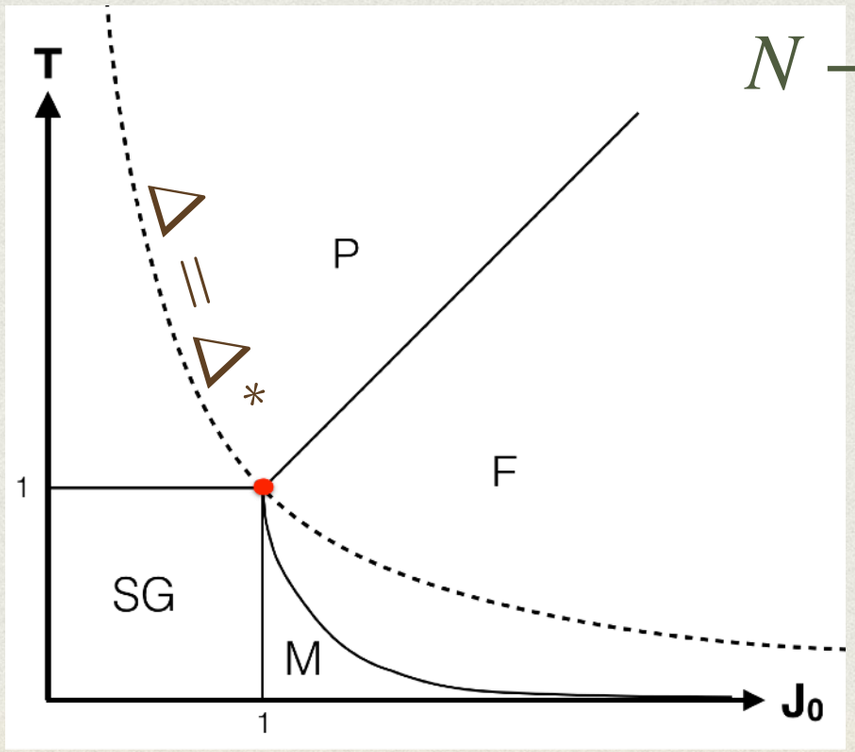
MEAN-FIELD SPIN GLASS

- Mean-field spin glass models are exactly solvable using replica method / cavity method. (Mezard, Parisi, Nishimori, Watkin, Nadal, Sompolinsky, many many others 70s-80s.)
- For Ising spins, under Gauge transform $S_i \to \tilde{S}_i S_i^*, J_{ij} \to \tilde{J}_{ij} S_i^* S_j^*$ planting is equivalent to ferromagnetic bias $J_0 = 1/\sqrt{\Delta^*}$ at $T = \sqrt{\Delta}$

Ill: N. Elmehed. © Nobel Media 2016
David J. Thouless

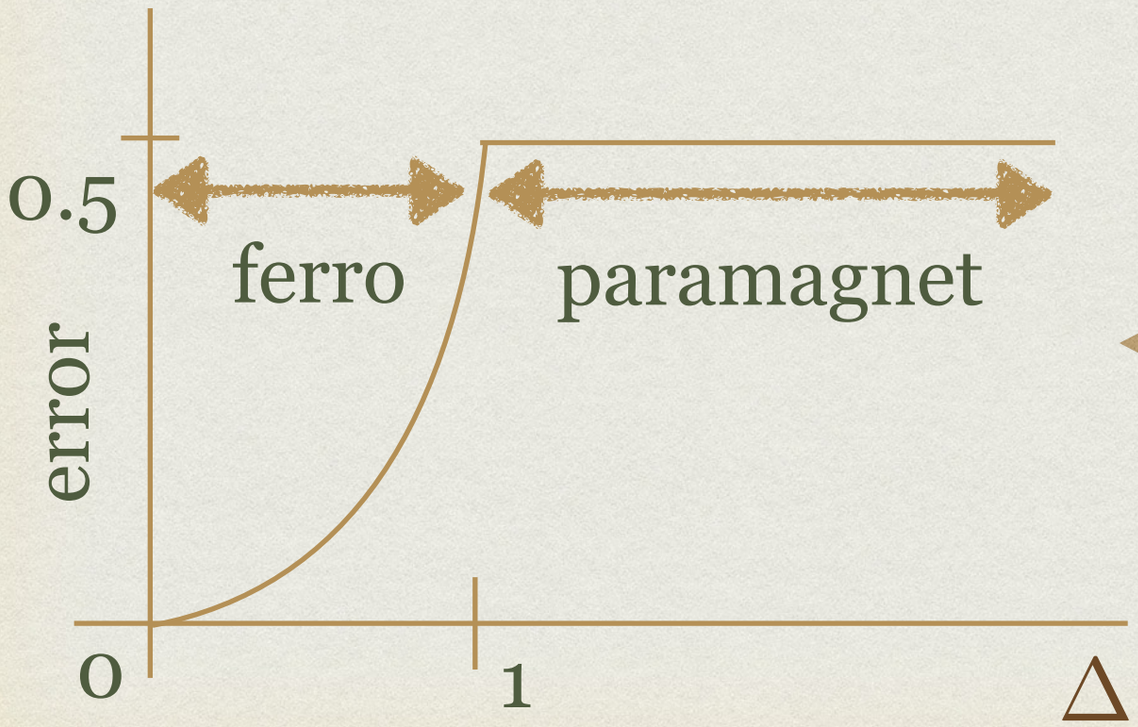
De Almeida; Thouless'78:

$\sqrt{\Delta}$

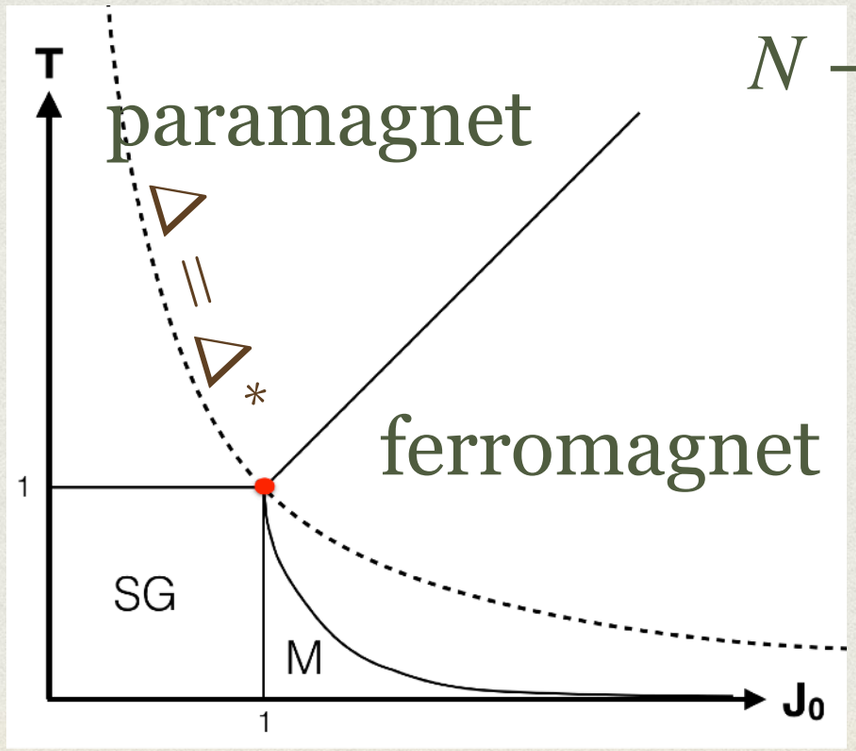


MEAN-FIELD SPIN GLASS

- Mean-field spin glass models are exactly solvable using replica method / cavity method. (Mezard, Parisi, Nishimori, Watkin, Nadal, Sompolinsky, many many others 70s-80s.)
- For Ising spins, under Gauge transform $S_i \to \tilde{S}_i S_i^*, J_{ij} \to \tilde{J}_{ij} S_i^* S_j^*$ planting is equivalent to ferromagnetic bias $J_0 = 1/\sqrt{\Delta^*}$ at $T = \sqrt{\Delta}$



$\sqrt{\Delta}$



$N \to \infty$

$\frac{1}{\sqrt{\Delta^*}}$

LET'S JUMP ~40 YEARS FORWARD:

RECENT TECHNICAL PROGRESS

(by my group and colleagues)

- Solution of low-rank matrix estimation for any noise distribution, any (separable) prior and rank. (Lesieur, Krzakala, LZ'15-17)

- Rigorous proof that the replica solution for Bayes-optimal inference is correct. (Krzakala, Xu, LZ'16 and Barbier, Dia, Macris, Krzakala, Lesieur, LZ'16)

- Approximate message passing algorithm matching the predicted performance. (Rangan, Fletcher'12, Matsushita, Tanaka'13, Deshpande, Montanari'14, Lesieur, Krzakala, LZ'15-17)

LOW-RANK MATRIX (TENSOR) ESTIMATION
"GENERALISED GAME"

Bayes-optimal inference for generic prior, output, and rank

$$P(x|Y) = \frac{1}{Z(Y)} \prod_{i=1}^{N} P_X(x_i) \prod_{i<j} P_{\text{out}}(Y_{ij}|x_i^T x_j/\sqrt{N}) \qquad x_i \in \mathbb{R}^r$$

or

$$P(u,v|Y) = \frac{1}{Z(Y)} \prod_{i=1}^{N} P_U(u_i) \prod_{j=1}^{M} P_V(v_j) \prod_{i,j} P_{\text{out}}(Y_{ij}|u_i^T v_j/\sqrt{N})$$

or

$$P(x|Y) = \frac{1}{Z(Y)} \prod_{i=1}^{N} P_X(x_i) \prod_{i_1<\ldots<i_p} P_{\text{out}}(Y_{i_1\ldots i_p}|\frac{\sqrt{(p-1)!}}{N^{(p-1)/2}} x_{i_1}\ldots x_{i_p})$$

Generate ground-truth x_i^* from P_X. Generate Y_{ij} from P_{out}.

Goal: Infer x^* from Y.

LOW-RANK MATRIX ESTIMATION

- Symmetric
  - Stochastic Block Model
  - Matrix completion.
  - Submatrix localization.
  - Z2 synchronization.
  - Spiked Wigner models.
- Tensor
  - Spiked tensor model
  - Hyper-graph clustering
  - Tensor completion.
  - Sub-tensor localisation
- Non-symmetric
  - Gaussian mixture clustering.
  - Biclustering.
  - Dawid-Skene model for crowdsourcing.
  - Johnstone's spiked covariance model.
  - Restricted Boltzmann machine with random weights.

RECENT TECHNICAL PROGRESS

(by my group and colleagues)

- Solution of low-rank matrix estimation for any noise distribution, any (separable) prior and rank. (Lesieur, Krzakala, LZ'15-17)

- Rigorous proof that the replica solution for Bayes-optimal inference is correct. (Krzakala, Xu, LZ'16 and Barbier, Dia, Macris, Krzakala, Lesieur, LZ'16)

- Approximate message passing algorithm matching the predicted performance. (Rangan, Fletcher'12, Matsushita, Tanaka'13, Deshpande, Montanari'14, Lesieur, Krzakala, LZ'15-17)

PROPERTIES OF THE BAYES-OPTIMAL ESTIMATOR

Theorem 1:

$\frac{1}{N}\log Z(Y,\Delta)$ concentrates around maximum of $\Phi(m)$

$$\Phi(m) = \mathbb{E}_{x,w}\left[\log \mathcal{Z}\left(\frac{m}{\Delta}, \frac{m}{\Delta}x + \sqrt{\frac{m}{\Delta}}w\right)\right] - \frac{m^2}{4\Delta} \qquad \begin{array}{l} m \in \mathbb{R} \\ x \sim P_X \\ w \sim \mathcal{N}(0,1) \end{array}$$

= replica symmetric free entropy

$\mathcal{Z}(A,B)$ auxiliary function defined by:

$$\mathcal{P}(x;A,B) = \frac{1}{\mathcal{Z}(A,B)} P_X(x) e^{Bx - Ax^2/2}$$

Proofs: Krzakala, Xu, LZ, ITW'16, Barbier, Dia, Macris, Krzakala, Lesieur, LZ'16 & 18; simpler: Lelarge, Miolane'16; El-Alaoui, Krzakala'17

PROPERTIES OF THE BAYES-OPTIMAL ESTIMATOR

Theorem 1:

$\frac{1}{N}\log Z(Y,\Delta)$ concentrates around maximum of $\Phi(m)$

$$\Phi(m) = \mathbb{E}_{x,w}\left[\log \mathcal{Z}\left(\frac{m}{\Delta}, \frac{m}{\Delta}x + \sqrt{\frac{m}{\Delta}}w\right)\right] - \frac{m^2}{4\Delta}$$

$m \in \mathbb{R}$

$x \sim P_X$

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

Theorem 2: mean-squared-error of the Bayes-optimal estimator

$$\text{MMSE} = \mathbb{E}_{P_X}(x^2) - \operatorname{argmax} \Phi(m)$$

Proofs: Krzakala, Xu, LZ, ITW'16, Barbier, Dia, Macris, Krzakala, Lesieur, LZ'16 & 18; simpler: Lelarge, Miolane'16; El-Alaoui, Krzakala'17

RECENT TECHNICAL PROGRESS

(by my group and colleagues)

- Solution of low-rank matrix estimation for any noise distribution, any (separable) prior and rank. (Lesieur, Krzakala, LZ'15-17)

- Rigorous proof that the replica solution for Bayes-optimal inference is correct. (Krzakala, Xu, LZ'16 and Barbier, Dia, Macris, Krzakala, Lesieur, LZ'16)

- Approximate message passing algorithm matching the predicted performance. (Rangan, Fletcher'12, Matsushita, Tanaka'13, Deshpande, Montanari'14, Lesieur, Krzakala, LZ'15-17)

APPROXIMATE MESSAGE PASSING

AMP algorithm estimates means and variances of the marginals:

$$a_i^{t+1} = f(A^t, B_i^t) \qquad v_i^{t+1} = \partial_B f(A^t, B_i^t)$$

$$B_i^t = \frac{1}{\Delta\sqrt{N}} \sum_{l=1}^{N} Y_{il} a_l^t - \frac{1}{\Delta}\left(\frac{1}{N}\sum_{l=1}^{N} v_l^t\right) a_i^{t-1} \qquad A^t = \frac{1}{N\Delta}\sum_{l=1}^{N} \left(a_l^t\right)^2$$

$f(A, B)$ auxiliary function defined by:

$$\mathscr{P}(x; A, B) = \frac{1}{\mathscr{Z}(A, B)} P_X(x) e^{Bx - Ax^2/2} \qquad f(A, B) = \mathbb{E}_{\mathscr{P}}(x)$$

Derived in: Rangan, Fletcher'12; Matsushita, Tanaka'13; Javanmard, Montanari'13; Deshpande, Montanari'14; Lesieur, Krzakala, LZ'15

Traces back to: Thouless, Anderson, Palmer'76

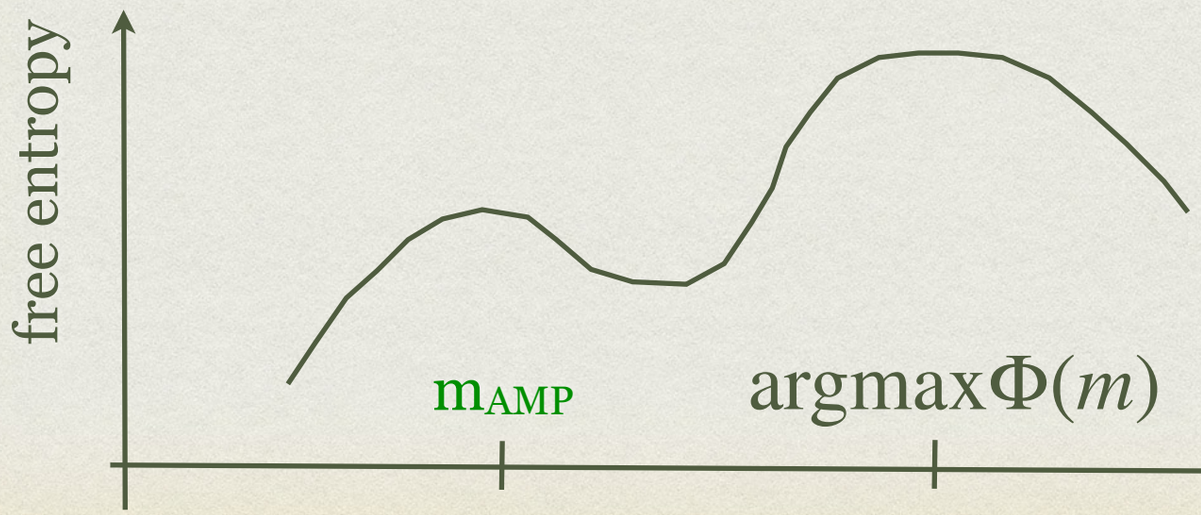
STATE EVOLUTION

$$\Phi(m) = \mathbb{E}_{x,w}\left[\log \mathcal{Z}\left(\frac{m}{\Delta}, \frac{m}{\Delta}x + \sqrt{\frac{m}{\Delta}}w\right)\right] - \frac{m^2}{4\Delta}$$

As $N \to \infty$:

- AMP-MSE given by the local maximum of the free entropy reached ascent starting from small m/large MSE. (Proofs: Rangan, Fletcher'12, Javanmard, Montanari'12, Deshpande, Montanari'14)

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



$$\text{MMSE} = \mathbb{E}_{P_X}(x^2) - \text{argmax}\,\Phi(m)$$

$$\text{MSE}_{\text{AMP}} = \mathbb{E}_{P_X}(x^2) - m_{\text{AMP}}$$

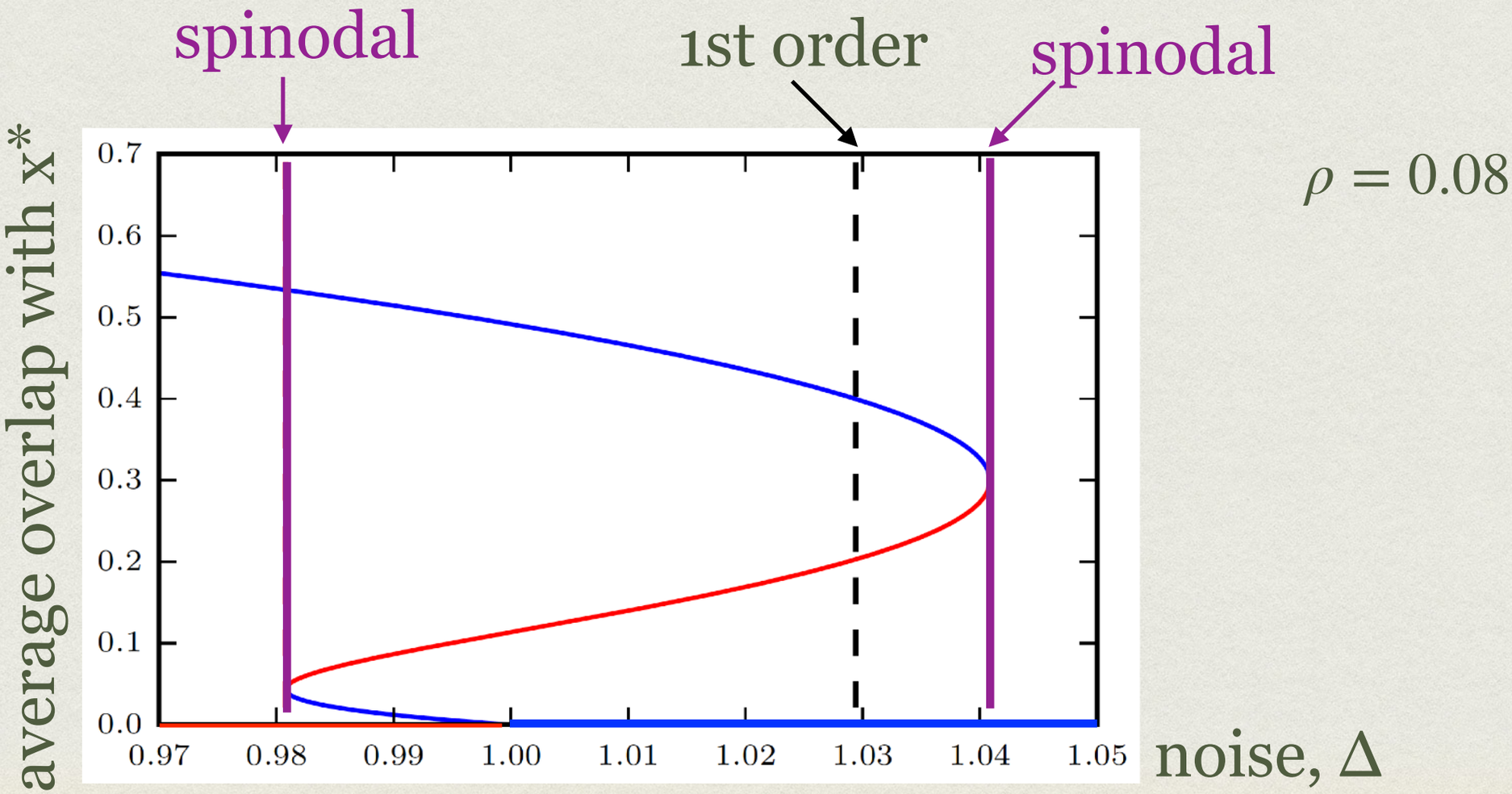
1ST ORDER PHASE TRANSITIONS

1ST ORDER PHASE TRANSITION

Slight change of the rules of the game:

$$P_X(x) = \rho[\delta(x-1) + \delta(x+1)]/2 + (1-\rho)\delta(x)$$

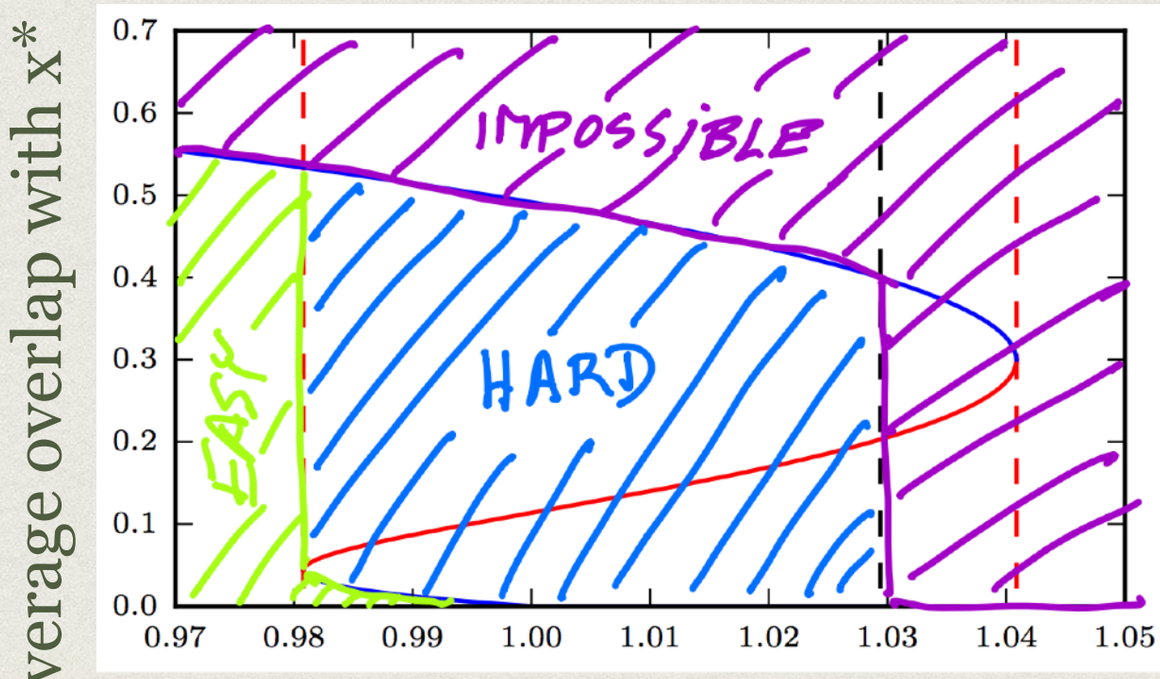
(sparse PCA - relevant in data-science applications to learn relevant dimensions)



$\rho = 0.08$

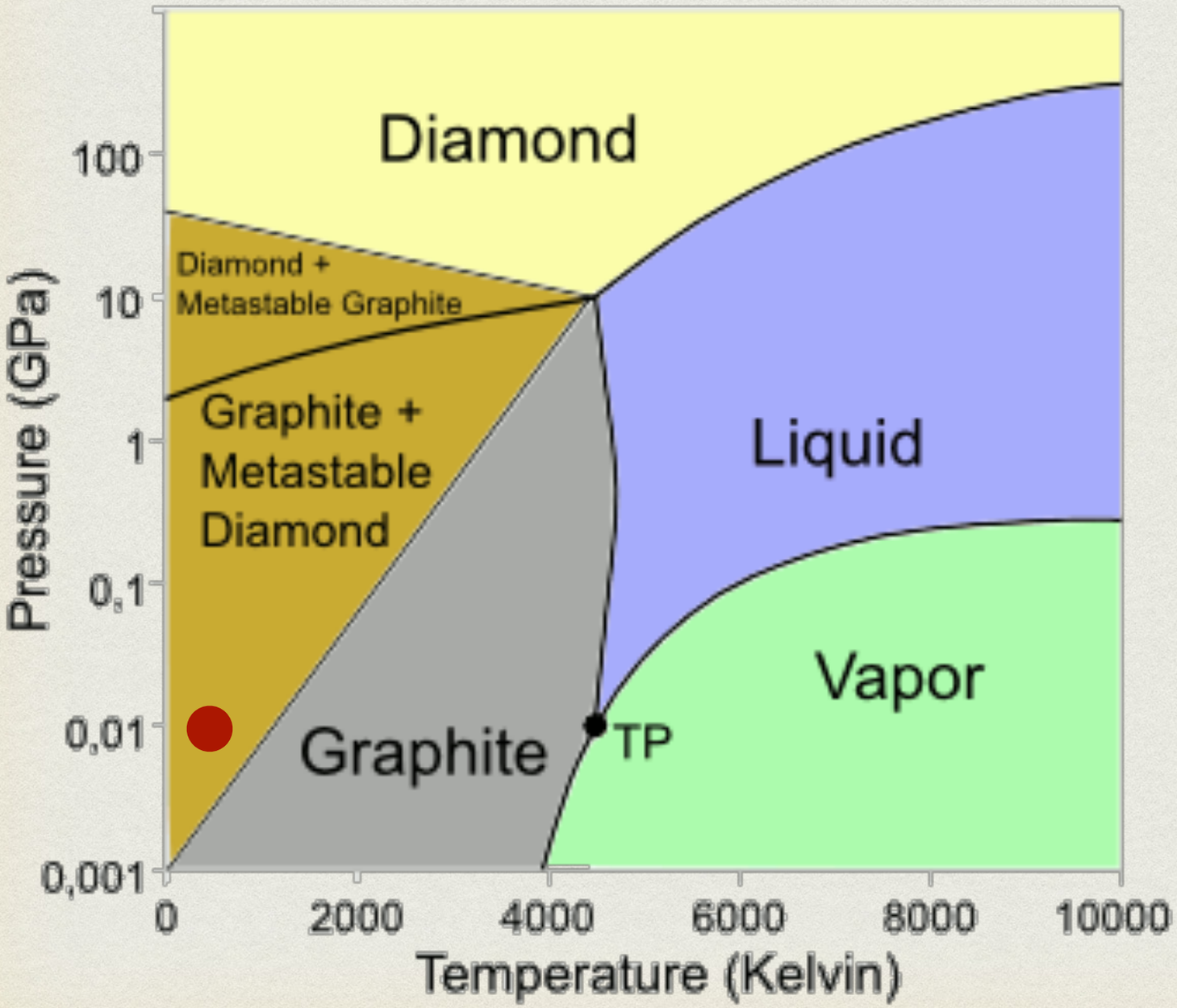
ALGORITHMIC INTERPRETATION

- Easy by approximate message passing algorithms.
- Impossible information theoretically.
- Hard phase: AMP algorithm does not achieve optimal performance.



HARD PHASE

Hard phase: Algorithms “stuck” at low accuracy for exponential time.



Metastable diamond
= low accuracy.

Equilibrium graphite
= high accuracy.

HARD PHASE

Hard phase = spinodal region of first order phase transitions.

Algorithmic threshold shared by spectral methods and SDPs.

Conjecture:

AMP achieves (in the large N limit) the lowest error among all polynomial algorithms.

Deshpande, Montanari'13: AMP optimal within a large class of related algorithms.

Hard phase identified in:

- dense planted sub-matrix;
- sparse principal component analysis;
- Gaussian mixture clustering;
- low-rank tensor completion;

- stochastic block model
- planted constraint satisfaction;
- low-density parity check error correcting codes;

- generalised linear regression;
- compressed sensing;
- learning in binary perceptron;
- phase retrieval;
- committee machine; ...

Computational Threshold Phenomena for Average-Case Problems in Statistics, Machine Learning, and Combinatorial Optimization

STOC 2018 Workshop. June 29, 2018. Los Angeles, CA.

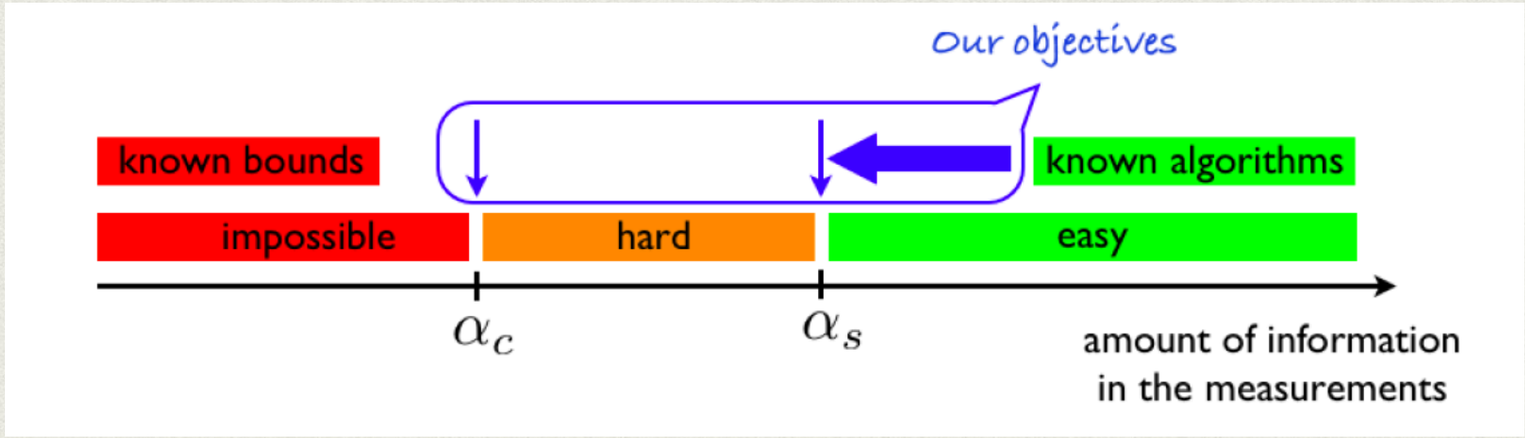
STOC = Symposium of the theory of computing

(Leading conference in computational complexity.)

HOW CAN ALGORITHMIC PHASE TRANSITIONS BE USEFUL?

AIMING TO REACH THRESHOLDS GIVES NEW ALGORITHMS

- Examples (in my works):
  - Spatially coupled measurements for compressed sensing.
    (Krzakala, Mezard, Sausset, Sun, LZ, PRX'12)
  - Non-backtracking spectral methods for sparse data.
    (Krzakala, Mossel, Moore, Neeman, Sly, Zhang, LZ, PNAS'13)



What does this type of analysis imply for neural networks?

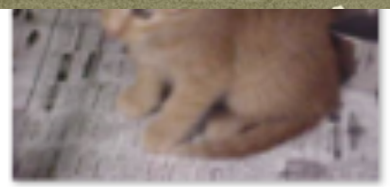
Hard phase identified in:

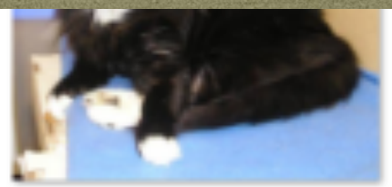
- dense planted sub-matrix;
- sparse principal component analysis;
- Gaussian mixture clustering;
- low-rank tensor completion;

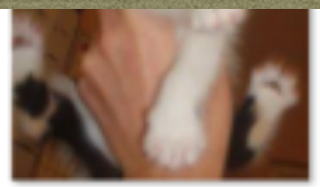
- stochastic block model
- planted constraint satisfaction;
- low-density parity check error correcting codes;

- generalised linear regression;
- compressed sensing;
- learning in binary perceptron;
- phase retrieval;
- committee machine; ...

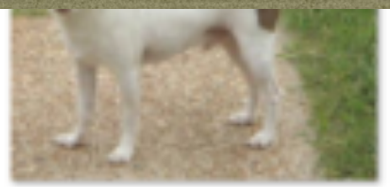
LEARNING A RULE


701.jpg

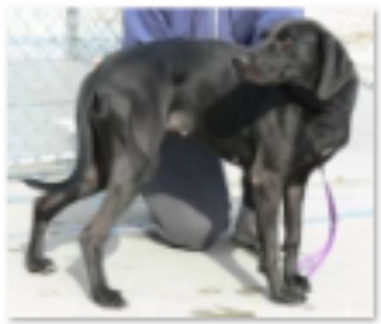

702.jpg


703.jpg


704.jpg


705.jpg


706.jpg


707.jpg


708.jpg


709.jpg


710.jpg


711.jpg


712.jpg


713.jpg

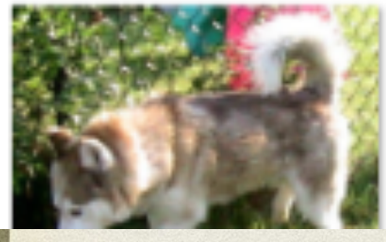

714.jpg

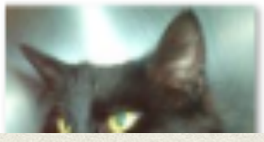

715.jpg


716.jpg


717.jpg


718.jpg





LEARNING A RULE

$= X_\mu =$ (01001010 01110011 10001100 01001011 01110000 10001100 ..... all the pixels ....)

Goal: Find a function f so that

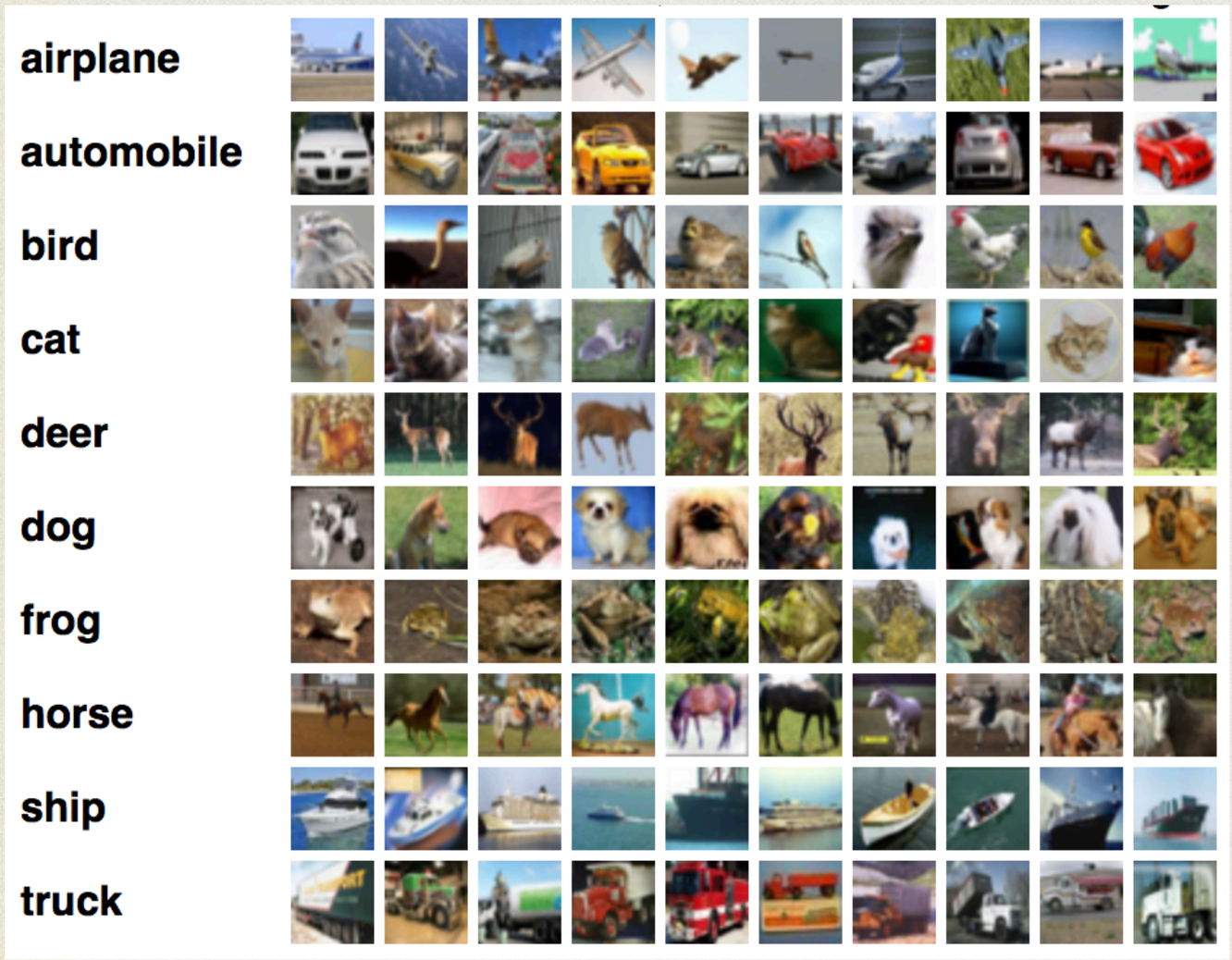
$$f(X_\mu) = +1 \qquad \text{for a picture of a cat.}$$

$$f(X_\mu) = -1 \qquad \text{for a picture of a dog.}$$

Is routinely done with convolutional deep neural networks.

SAMPLE COMPLEXITY

How low is the optimal sample complexity? Are we achieving it?
If not, is it because of architectures or algorithms?



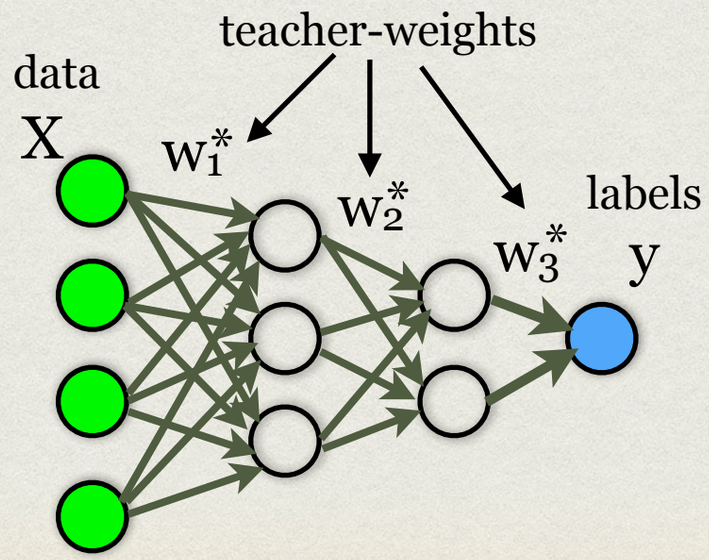
- Cifar10 - 50000 samples.

- How many samples are really needed?

OUR QUESTION: CAN A NEURAL NETWORK LEARN A TEACHER-NEURAL NETWORK?

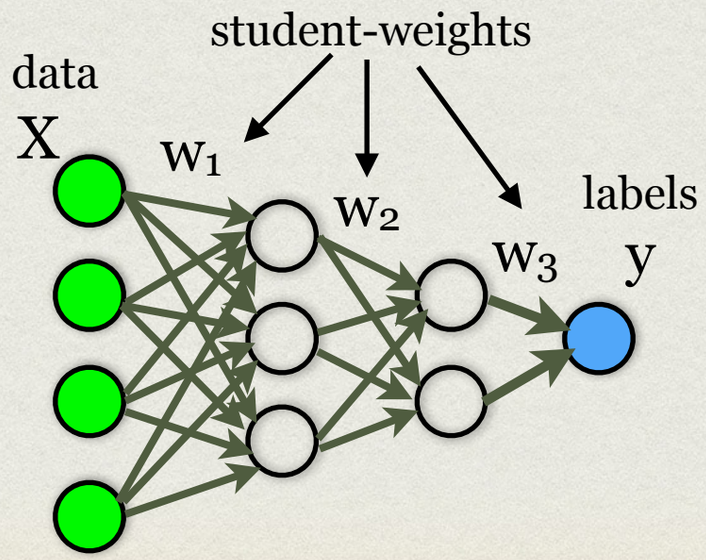
Teacher-network

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



Student-network

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



TEACHER-STUDENT PERCEPTRON

Gardner, Derrida'89, Gyorgyi'90

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

1989

Model B in:

Three unfinished works on the optimal storage capacity of networks

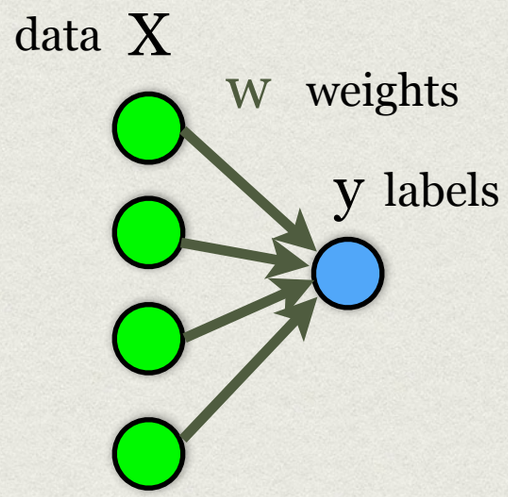
E Gardner and B Derrida

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.

Single layer neural network



$$y_\mu = \text{sign}\left(\sum_{i=1}^{p} X_{\mu i} w_i\right)$$

p dimensions

n samples

high-dimensional limit

$$n \to \infty \quad p \to \infty$$

$$n/p = \alpha = \Omega(1)$$

Solved using the replica method in the high-dimensional limit

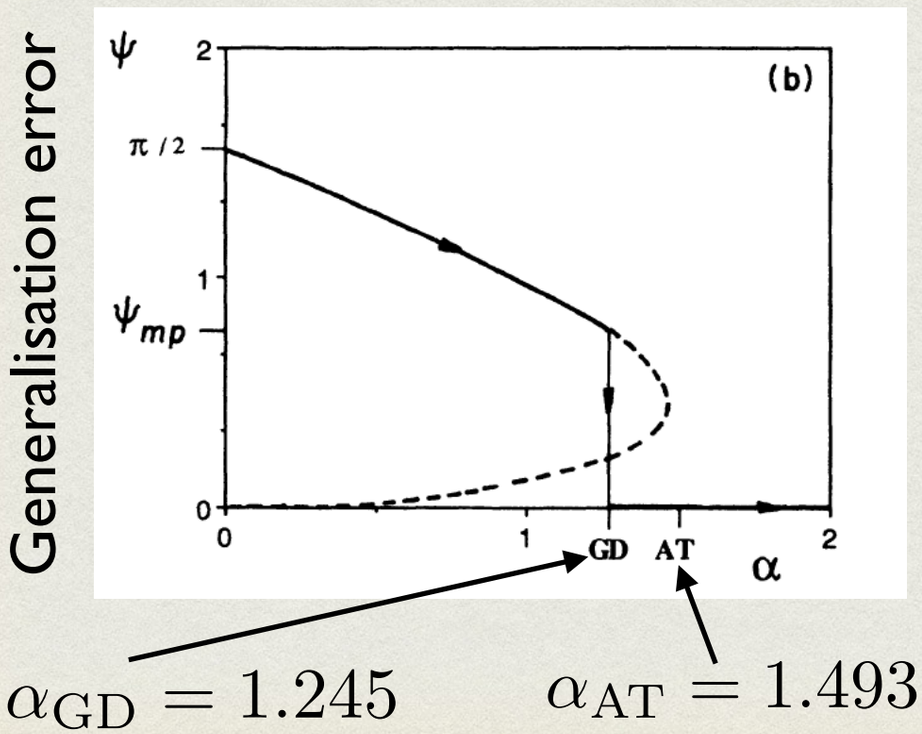
RAPID COMMUNICATIONS

PHYSICAL REVIEW A VOLUME 41, NUMBER 12 15 JUNE 1990

First-order transition to perfect generalization in a neural network with binary synapses

Géza Györgyi*
School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332-0430
(Received 9 February 1990)

Learning from examples by a perceptron with binary synaptic parameters is studied. The examples are given by a reference (teacher) perceptron. It is shown that as the number of examples increases, the network undergoes a first-order transition, where it freezes into the state of the reference perceptron. When the transition point is approached from below, the generalization error reaches a minimal positive value, while above that point the error is constantly zero. The transition is found to occur at $\alpha_{GD}=1.245$ examples per coupling.



- Binary teacher-weights:

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

- "The dashed lines represent non-physical segments of the curves." (Gyorgyi'90)

RECENT PROGRESS

- Solution of generalised linear regression for any (noisy) activation function and any (separable) prior.

- Rigorous proof that the replica solution for the teacher-student model is correct.

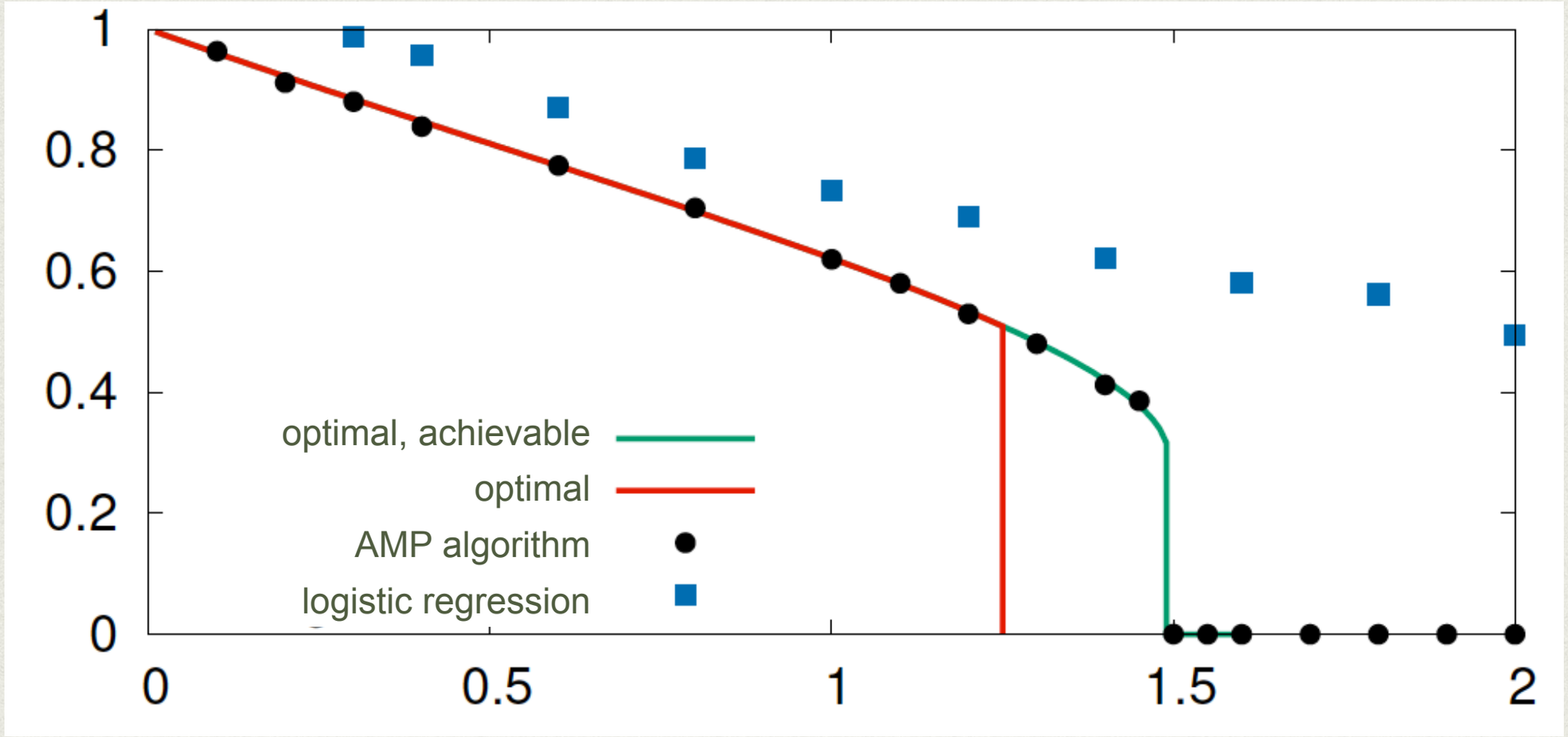
- Regions of optimality of approximate message passing algorithm.

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

PHASE TRANSITIONS

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

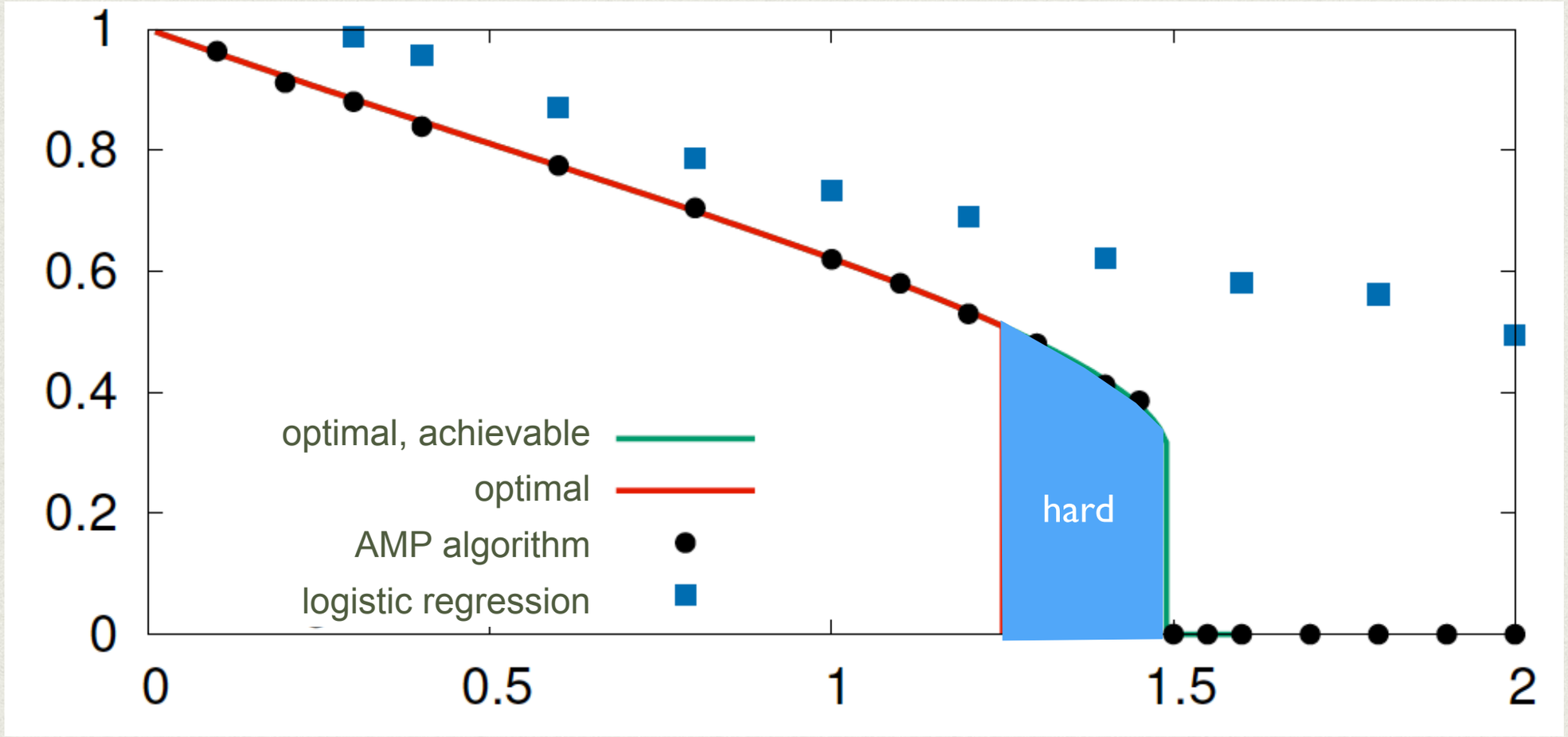
$$\begin{matrix} n \to \infty \\ p \to \infty \end{matrix} \qquad n/p = \Omega(1)$$



PHASE TRANSITIONS

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

$$\begin{matrix} n \to \infty \\ p \to \infty \end{matrix} \qquad n/p = \Omega(1)$$



GOING MULTI-LAYER

Committee machine

Model from Schwarze'92.

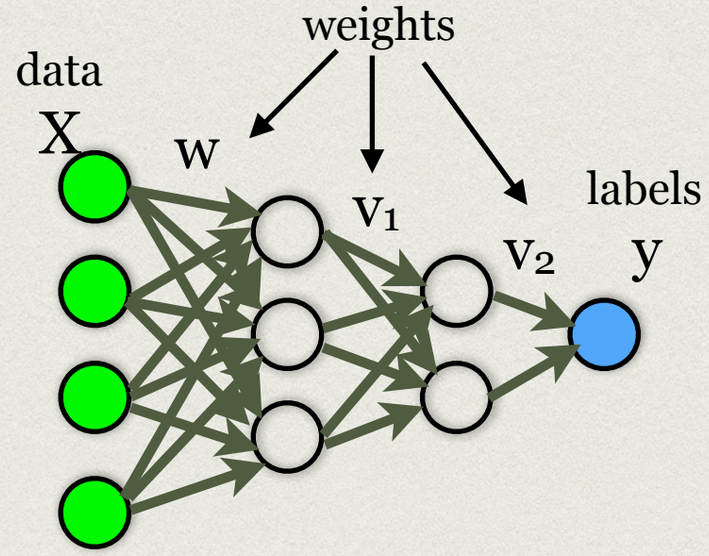
Proof of the replica formula, and approximate message passing Aubin, Maillard, Barbier, Macris, Krzakala, LZ'18, spotlight at NeurIPS'18.

p input units

K hidden units

output unit

n training samples



L=3 layers

w learned, v_1 & v_2 fixed

Limit: $n \to \infty$, $p \to \infty$ $\quad \alpha = n/p = \Omega(1) \qquad K = O(1)$

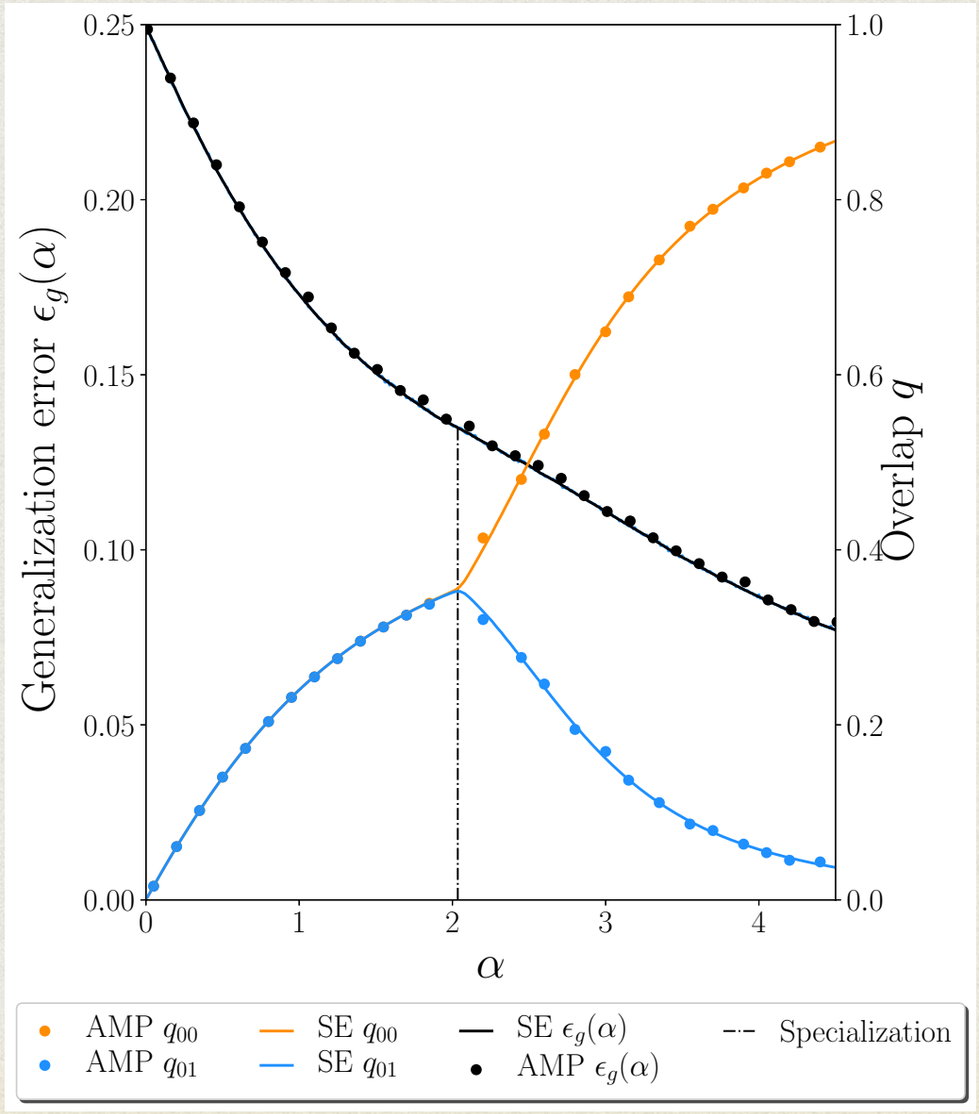
PHASE TRANSITIONS

K=2

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

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

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



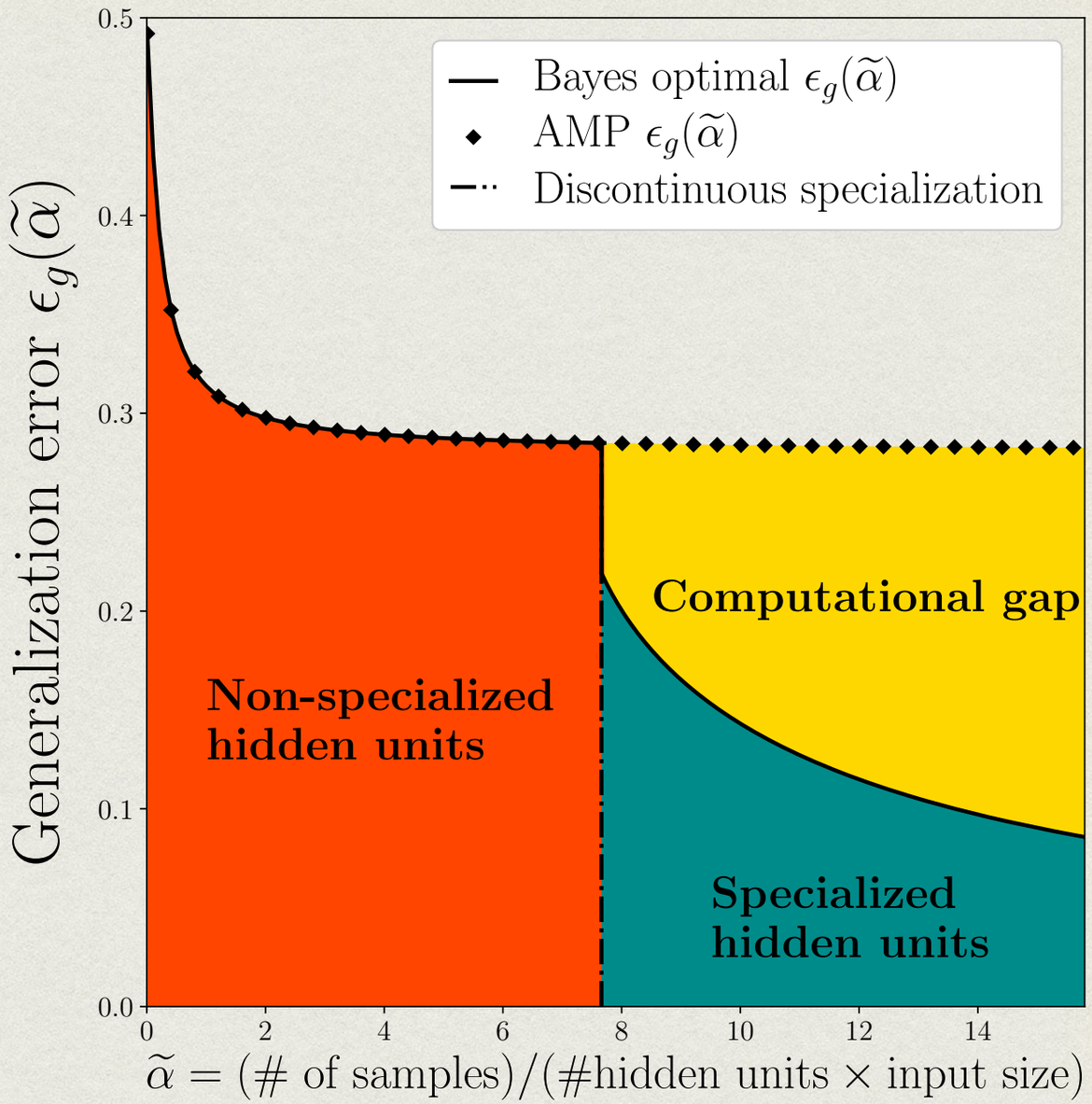
PHASE TRANSITIONS

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

$K \gg 1$

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

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



TOWARDS THEORY OF DEEP LEARNING?

algorithm

message passing

(stochastic) gradient descent

GRADIENT FLOW

spherical constraint
(weight decay)

$$\dot{x}_i(t) = -\mu(t)x_i(t) - \frac{\partial \mathcal{H}}{\partial x_i}$$

gradient

Where does it go in large constant time?

CHALLENGE

Analysis of randomly initialised gradient flow dynamics
in non-convex, high-dimensional, strongly correlated
models with limited number of samples
and notion of the generalisation error.

MIXED SPIKED MATRIX-TENSOR MODEL

Sarao, Biroli, Cammarota, Krzakala, Urbani, LZ, arXiv:1812.09066

- On the same signal x* observe a matrix Y and tensor T as:

$$Y_{ij} = \frac{1}{\sqrt{N}} x_i^* x_j^* + \xi_{ij} \qquad \xi_{ij} \sim \mathcal{N}(0, \Delta_2)$$

$$T_{i_1 \dots i_p} = \frac{\sqrt{(p-1)!}}{N^{(p-1)/2}} x_{i_1}^* \dots x_{i_p}^* + \xi_{i_1 \dots i_p} \qquad \xi_{i_1, \dots, i_p} \sim \mathcal{N}(0, \Delta_p)$$

- Maximise likelihood = minimize Hamiltonian

$$\mathscr{H}(x) = -\frac{1}{\Delta_2 \sqrt{N}} \sum_{i<j} Y_{ij} x_i x_j - \frac{\sqrt{(p-1)!}}{\Delta_p N^{(p-1)/2}} \sum_{i_1 < \dots < i_p} T_{i_1 \dots i_p} x_{i_1} \dots x_{i_p}$$

spherical constraint: $\sum_{i=1}^{N} x_i^2 = N$

Spiked version of the mixed 2+p spherical spin glass model.

LANGEVIN STATE EVOLUTION

Sarao, Biroli, Cammarota, Krzakala, Urbani, LZ, arXiv:1812.09066

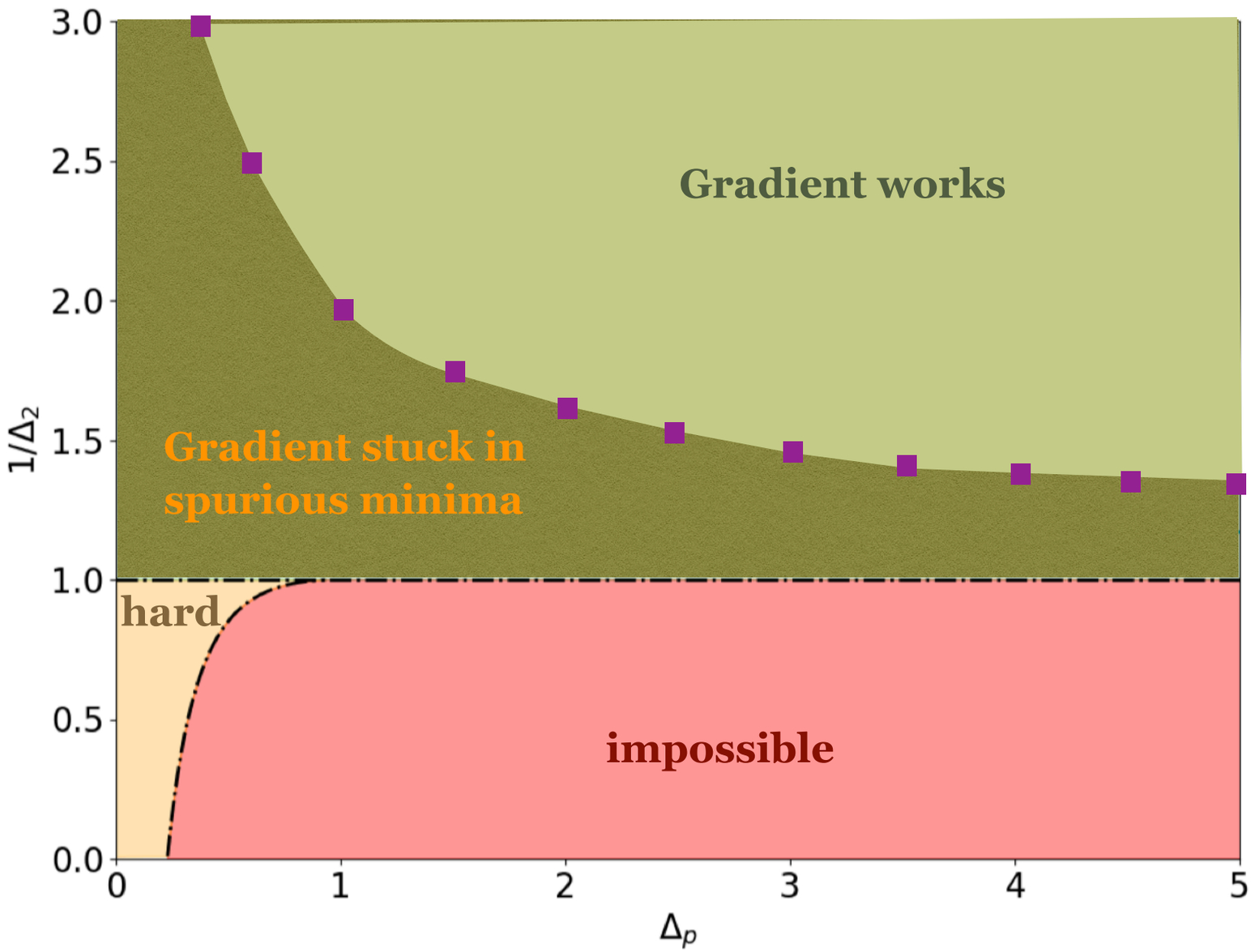
Generalization of the Crisanti-Horner-Sommers-Cugliandolo-Kurchan'93 equations to include the spike x*.

$$
\begin{aligned}
C_N(t,t') &\equiv \tfrac{1}{N}\textstyle\sum_{i=1}^{N} x_i(t)x_i(t')\,, \\
\overline{C}_N(t) &\equiv \tfrac{1}{N}\textstyle\sum_{i=1}^{N} x_i(t)x_i^*\,, \\
R_N(t,t') &\equiv \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \partial x_i(t)/\partial h_i(t')|_{h_i=0}\,,
\end{aligned}
\qquad N\to\infty
$$

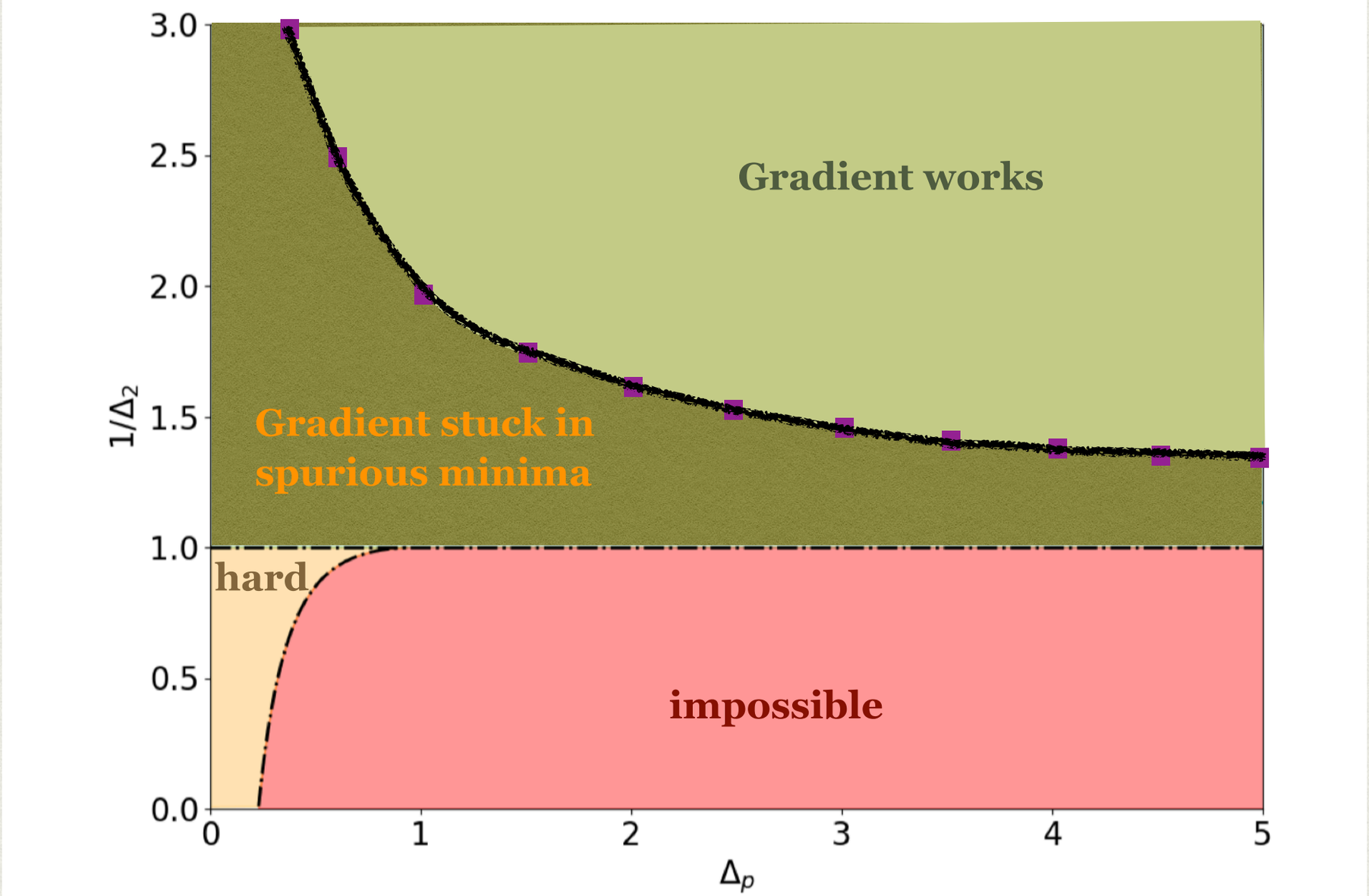
$$
\begin{aligned}
\frac{\partial}{\partial t}C(t,t') &= -\tilde{\mu}(t)C(t,t') + Q'(m(t))m(t') + \int_0^t dt'' R(t,t'')Q''(C(t,t''))C(t',t'') \\
&\quad + \int_0^{t'} dt'' R(t',t'')Q'(C(t,t''))\,, \\
\frac{\partial}{\partial t}R(t,t') &= -\tilde{\mu}(t)R(t,t') + \int_{t'}^t dt'' R(t,t'')Q''(C(t,t''))R(t'',t')\,, \\
\frac{\partial}{\partial t}m(t) &= -\tilde{\mu}(t)m(t) + Q'(m(t)) + \int_0^t dt'' R(t,t'')m(t'')Q(C(t,t''))\,,
\end{aligned}
$$

$$Q(x) = x^2/(2\Delta_2) + x^p/(p\Delta_p).$$

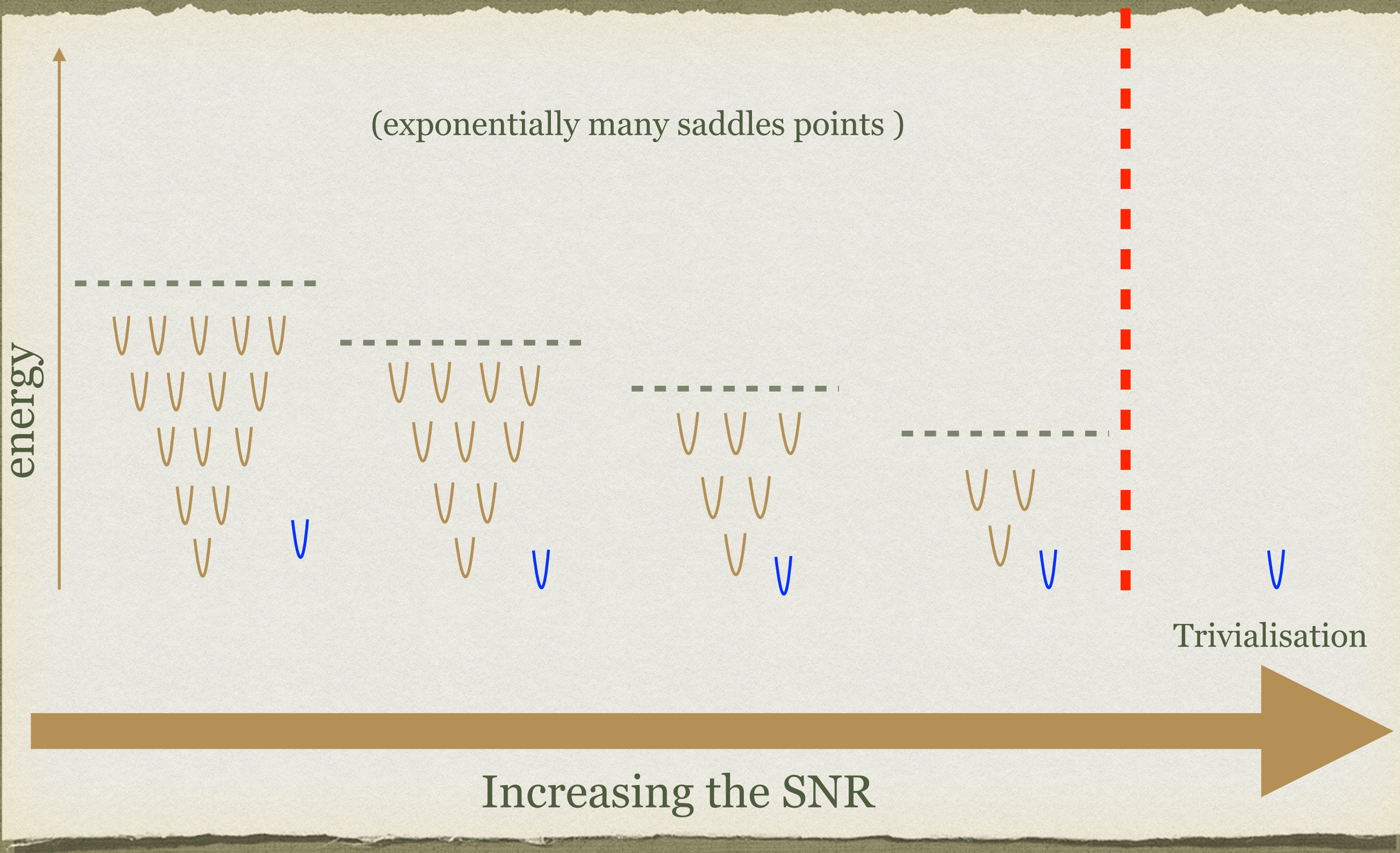
NUMERICAL INTEGRATION



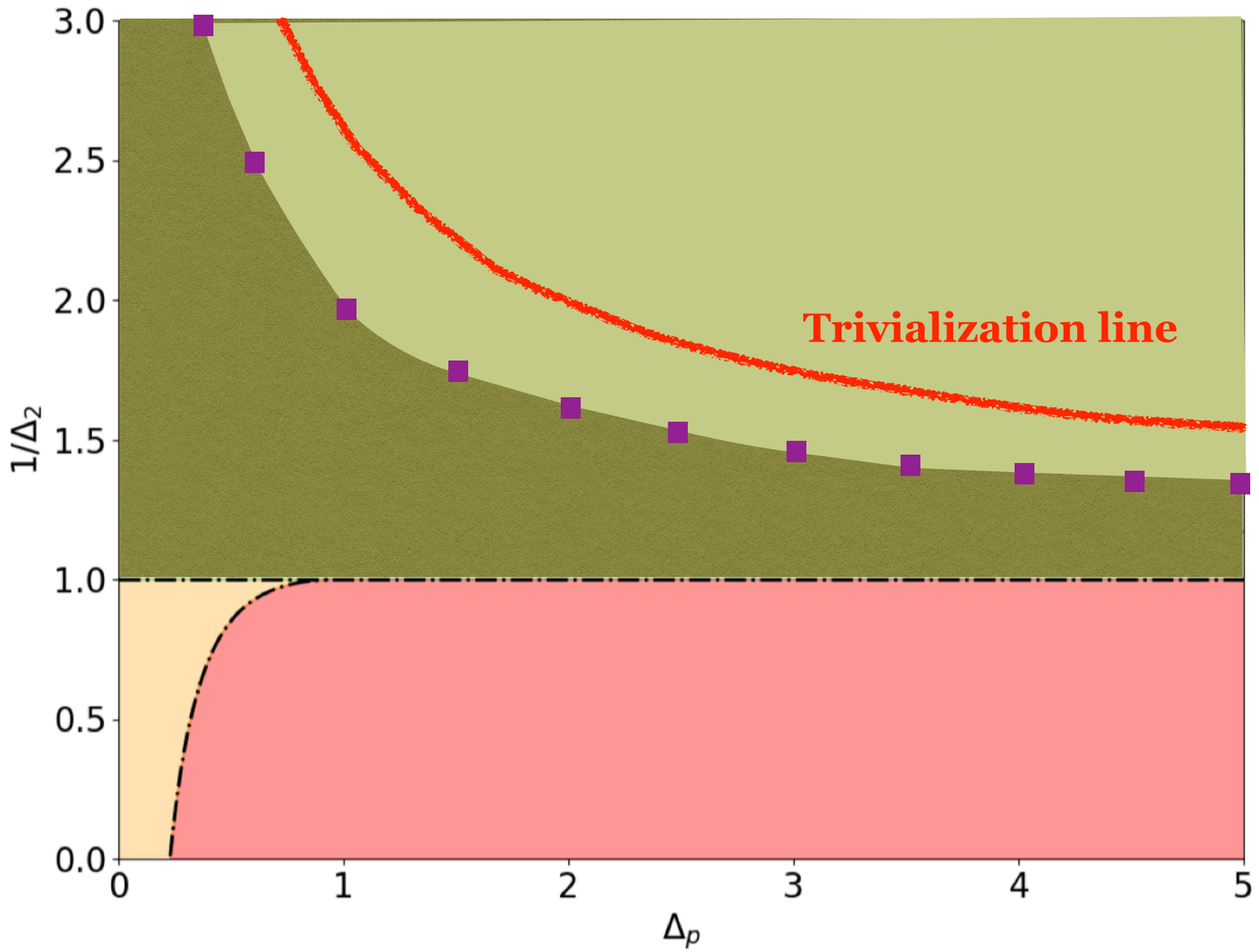
ANALYTICAL SOLUTION



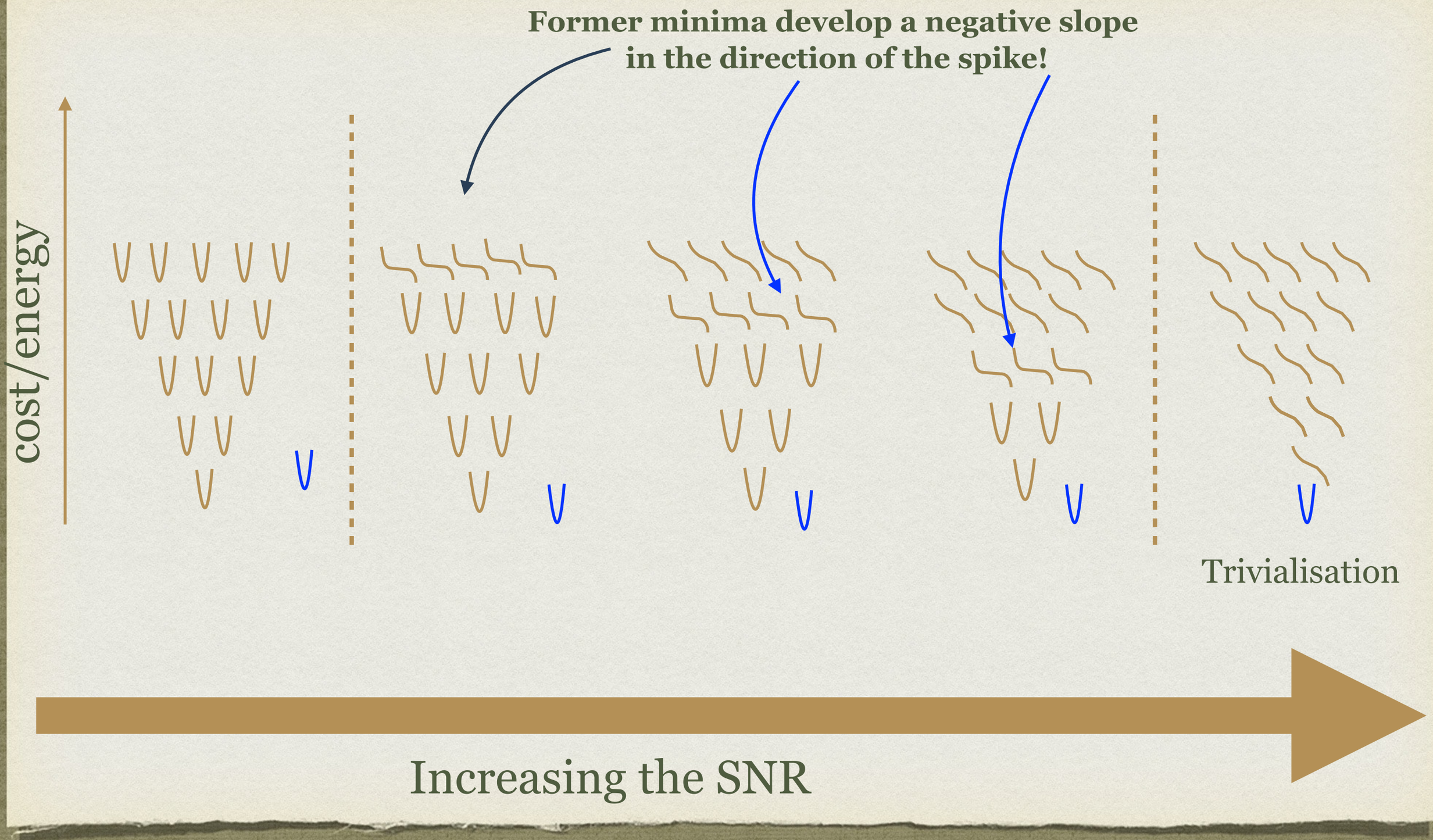
LANDSCAPE ANALYSIS
WITH KAC-RICE APPROACH



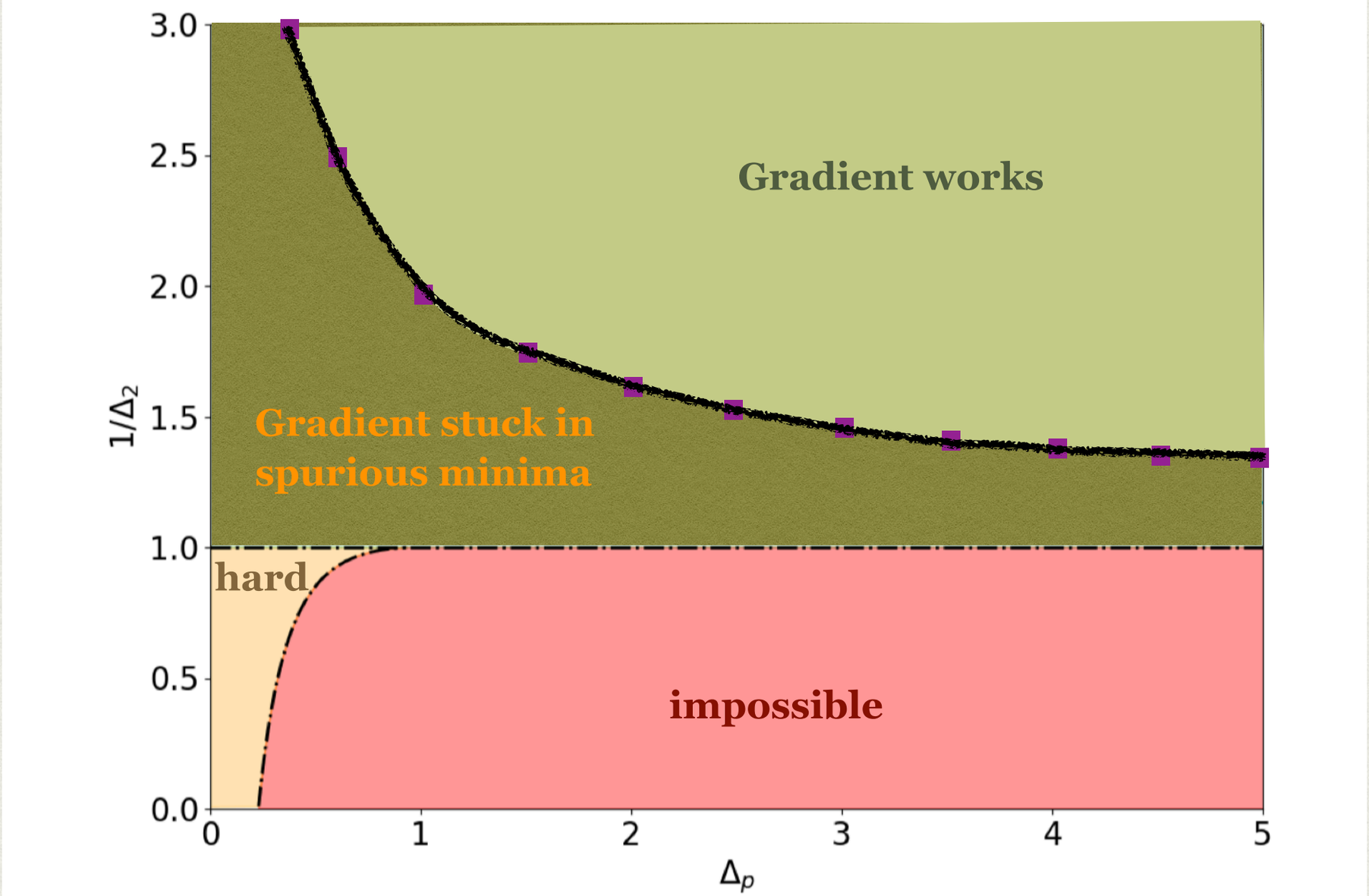
WHAT IS GOING ON?



LANDSCAPE ANALYSIS

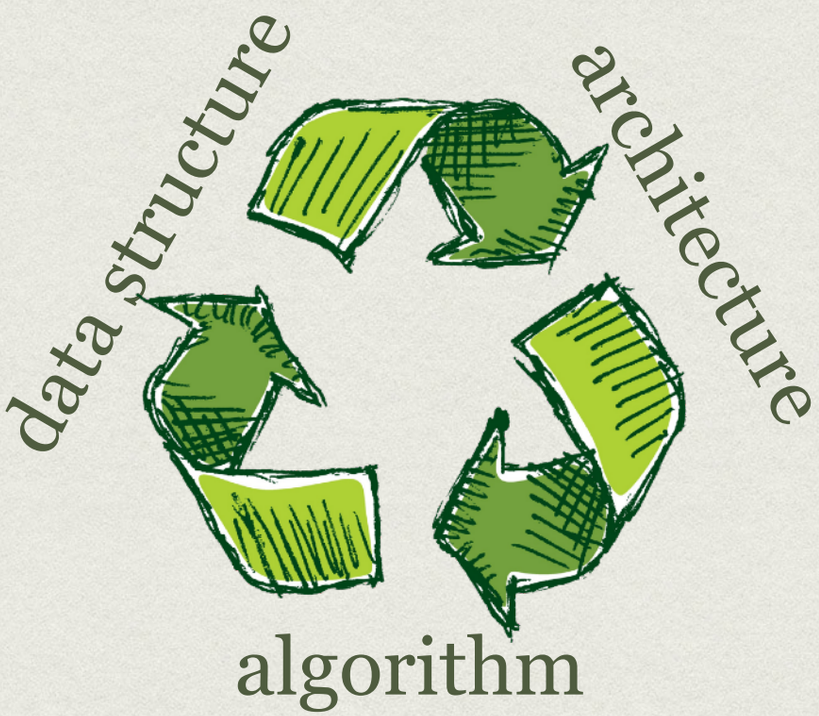


ANALYTICAL SOLUTION



WHERE ARE WE GOING?

Building the toolbox to understand learning from data.



- Sample complexity.
- Generalization error.
- Setting hyper parameters.
- Better algorithms/architectures.

MORE SPECIFICALLY ...

- Priors from another graphical model. Multi-layer compositions of graphical models. "Feynman diagrams" of composed graphical models. Preliminary: Manoel, Krzakala, Mézard, LZ, 1701.06981 ; Rangan, Fletcher 1706.09549; Reeves 1710.04580.
- Over-parametrized networks. Teacher network mismatching the student network. Goldt, Advani, Saxe, Krzakala, LZ, 1901.09085
- Structured data. Input data coming from a generative random network. Labels from teacher networks. Gabrie, Manoel, Barbier, Luneau, Macris, Krzakala, LZ, NeurIPS'18, 1805.09785
- Extensive hidden layers. Resisting the replica analysis - technical challenge.
- Analyzing gradient-based algorithms. Extensions of solvable dynamical mean-field-theory beyond spherical p-spin. Agoritas, Biroli, Urbani, Zamponi, 1710.04894

CONCLUSIONS

- Many problems studied in math/statistics/signal processing/machine learning are physics models of disordered systems.
- 1st order phase transitions imply sharply-defined algorithmically hard regions.
- Identifying threshold and aiming to reach them leads to better algorithms.
- Learning with multi-layer neural networks is within the reach of this methodology.