WHAT HAS STATISTICAL PHYSICS TO SAY ABOUT MACHINE LEARNING?

Lenka Zdeborová
(IPhT, CEA Saclay, France)

PAISS, 3.-5. 10. 2019, Paris

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, Bertrand Thirion, Eric W. Tramel, Pierfrancesco Urbani, Gaël Varoquaux, Massimo Vergassola, Yingying Xu, Jiaming Xu, Sun Yifan, Riccardo Zecchina, Pan Zhang, Hai-jun Zhou.

ENGINEERING & SCIENCE

- ML mostly developed by engineering design process: Define an objective (e.g. to reach the best accuracy on ImageNet). Create a tool that reaches the objective.

Rank	Method	Top 1 Accuracy	Top 5 Accuracy	Number of params	Extra Training Data	Paper Title	Year	Paper	Code
1	**FixResNeXt-101 32x48d**	86.4%	98.0%	829M	✓	Fixing the train-test resolution discrepancy	2019		

➡ Deep learning is a revolutionary engineering progress.

- Science aims to understand behaviour of existing world. Do we understand why FixResNeXt-101 works?

➡ Science/understanding of deep learning is in its infancy.

MOTIVATION

- Do we need understanding? Isn't engineering enough, simply because "it works"?
- Some open questions:

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?

From "Reflections after refereeing papers for NIPS", Leo Breiman, 1995.

Still not answered!

TOWARDS THEORY OF DEEP LEARNING?

Inter-play of three ingredients

information theory
statistics
signal processing
data structure

learning theory
approximation theory
architecture

algorithm
computer science
optimisation theory

See also: E. Mossel, Deep Learning Boot Camp in Simons Institute, Berkeley (June 2019).

LONG-LASTING FRIENDSHIP BETWEEN MACHINE LEARNING AND STATISTICAL PHYSICS

STATISTICAL PHYSICS AND MACHINE LEARNING

ÉCOLE DE PHYSIQUE DES HOUCHES

Yann LeCun is with Levent Sagun and 3 others.
August 30

Stéphane Mallat's tutorial at the "Statistical Physics and Machine Learning back Together" summer school in Cargese, Corsica.

There is a long history of theoretical physicists (particularly condensed matter physicists) bringing ideas and mathematical methods to machine learning, neural networks, probabilistic inference, SAT problems, etc.

In fact, the wave of interest in neural networks in the 1980s and early 1990s was in part caused by the connection between spin glasses and recurrent nets popularized by John Hopfield. While this caused some physicists to morph into neuroscientists and machine learners, most of them left the field when interest in neural networks wanted in the late 1990s

With the prevalence of deep learning and all the theoretical questions that surround it, physicists are coming back!

Many young physicists (and mathaticians) are now working on trying to explain why deep learning works so well. This summer school is for them.

We need to find ways to connect this emerging community with the ML/AI community. It's not easy because (1) papers submitted by physicists to ML conferences rarely make it because of a lack of qualified reviewers; (2) conference papers don't count in a physicist's CV.

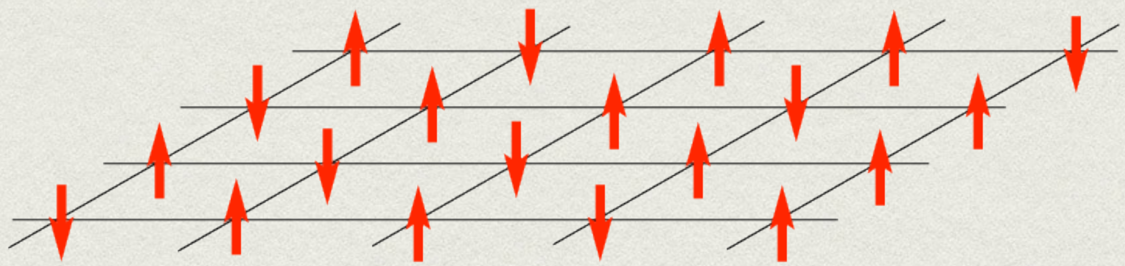
http://cargese.krzakala.org

Disordered Systems and Biological Organization

MODELS

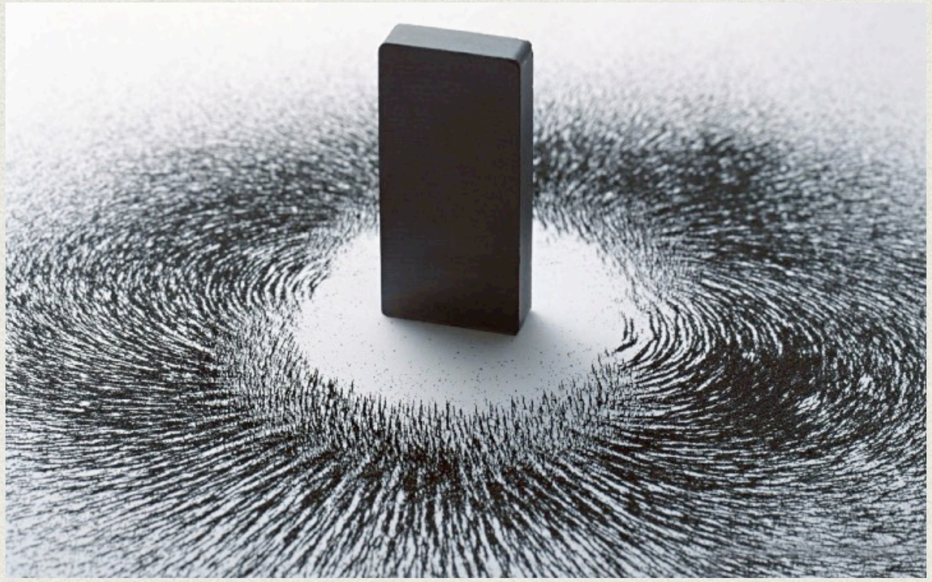
- In data science, models are used to fit the data (e.g. linear regression: Best straight line that captures the dependence of y on x?). In physics we could call those an “ansatz”.

- In physics, models are the main tool for understanding.



2–D Ising Model

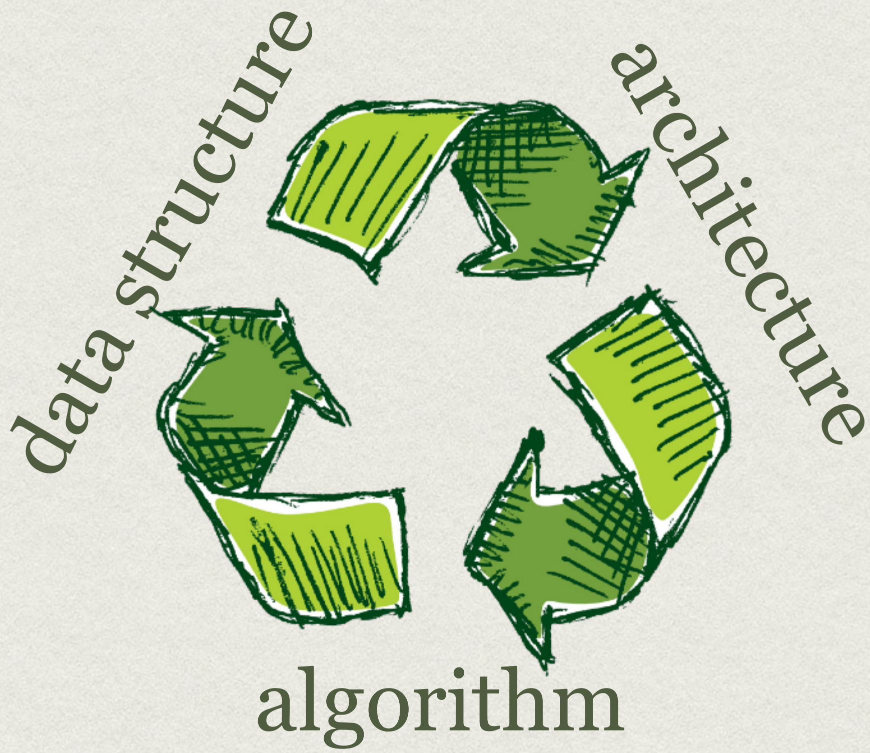
magnetism of materials



$$P(\{S_i\}_{i=1,\dots,N}) = \frac{e^{-\beta\mathcal{H}}}{Z}$$

$$\mathcal{H} = -J \sum_{(ij)\in\mathbb{E}} S_i S_j$$

WHAT TO MODEL IN DEEP LEARNING?



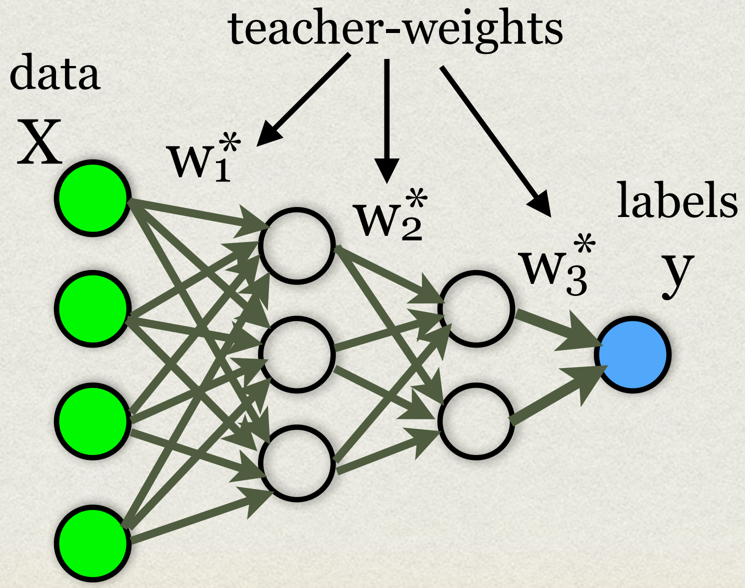
We aim to reproduce the salient behaviours of the real system.

Iterative process of improving the model.

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

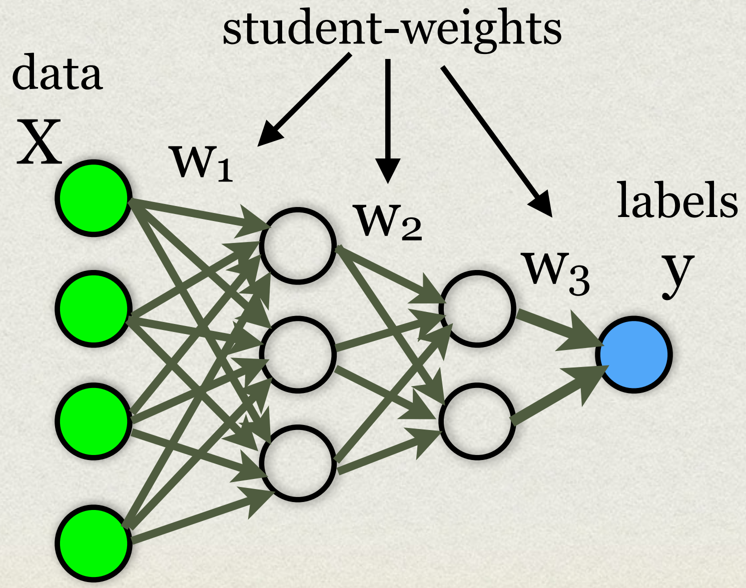
Teacher-network

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

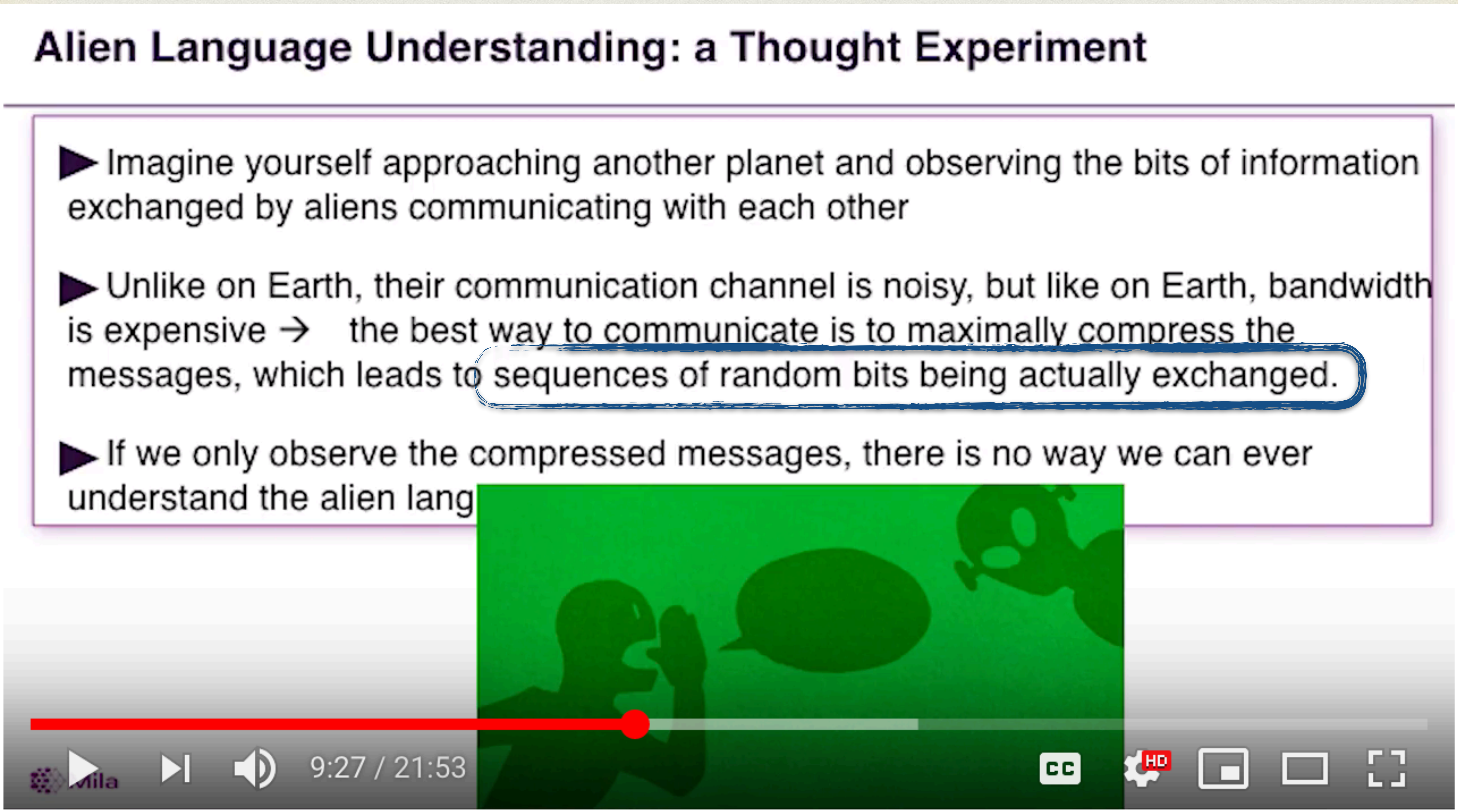


Student-network

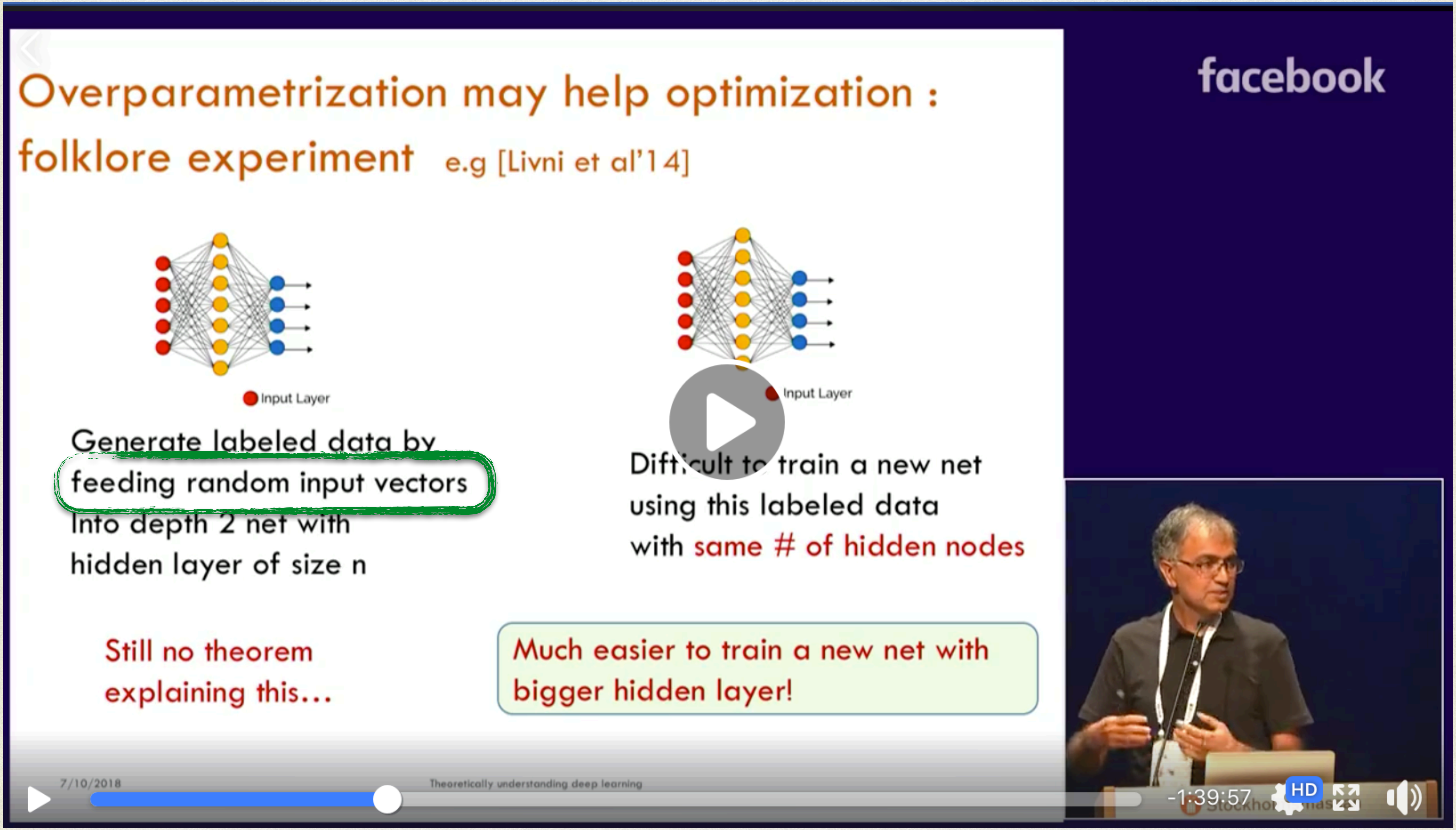
- Observes X, y, the architecture of the network.
- How does the best achievable generalisation error depend on the number of samples n?



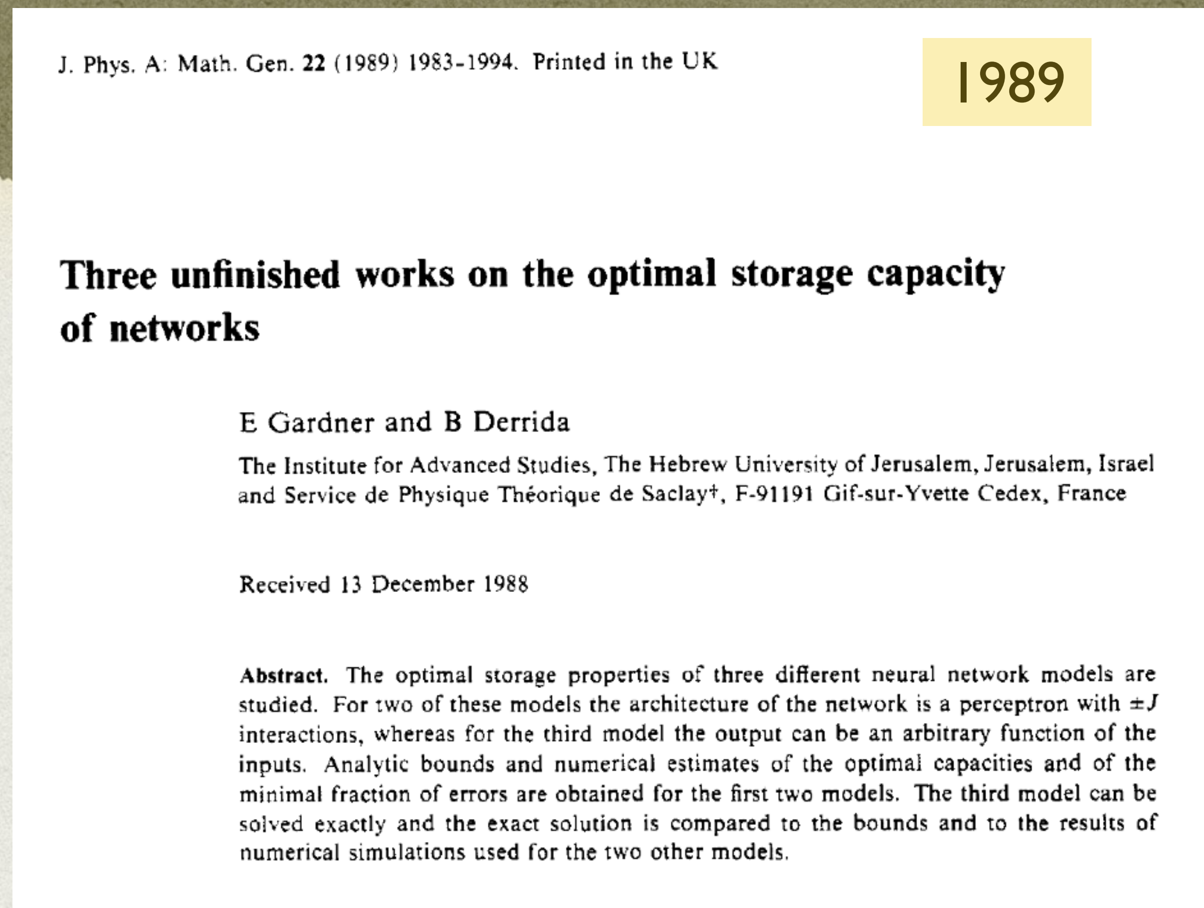
Yoshua Bengio at France in AI'18: On challenges of deep learning towards AI.


Alien Language Understanding: a Thought Experiment
Imagine yourself approaching another planet and observing the bits of information exchanged by aliens communicating with each other
Unlike on Earth, their communication channel is noisy, but like on Earth, bandwidth is expensive → the best way to communicate is to maximally compress the messages, which leads to sequences of random bits being actually exchanged.
If we only observe the compressed messages, there is no way we can ever understand the alien lang
9:27 / 21:53
CC
HD
Mila

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



TEACHER-STUDENT PERCEPTRON



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

1989

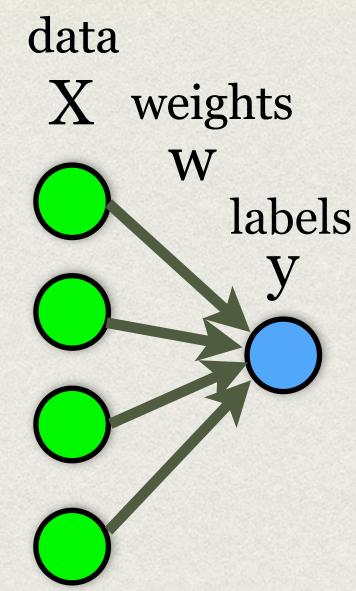
Three unfinished works on the optimal storage capacity of networks

E Gardner and B Derrida

The Institute for Advanced Studies, The Hebrew University of Jerusalem, Jerusalem, Israel and Service de Physique Théorique de Saclay†, F-91191 Gif-sur-Yvette Cedex, France

Received 13 December 1988

Abstract. The optimal storage properties of three different neural network models are studied. For two of these models the architecture of the network is a perceptron with $\pm J$ interactions, whereas for the third model the output can be an arbitrary function of the inputs. Analytic bounds and numerical estimates of the optimal capacities and of the minimal fraction of errors are obtained for the first two models. The third model can be solved exactly and the exact solution is compared to the bounds and to the results of numerical simulations used for the two other models.



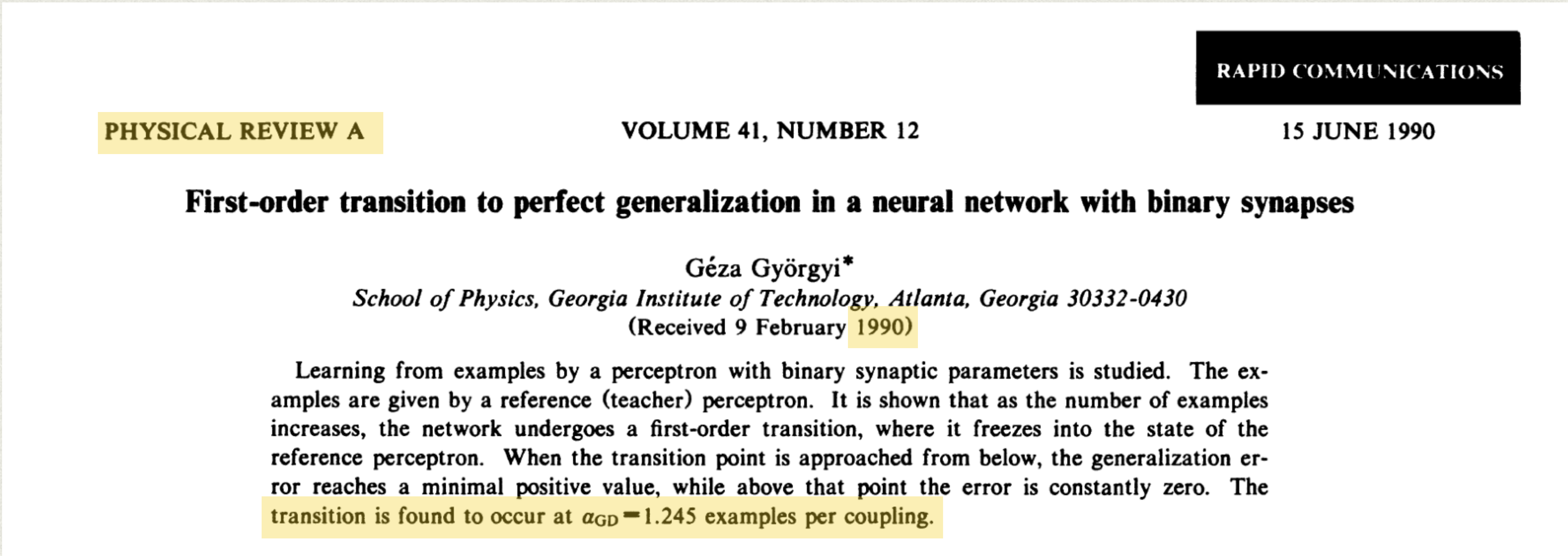
- Take random iid Gaussian $X_{\mu i}$ and random iid w_i^* from P_w
- Create $y_\mu = \varphi\left(\sum_{i=1}^{p} X_{\mu i} w_i^*\right)$
- High-dimensional regime: $n \to \infty \quad p \to \infty$

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

p dimensions

n samples

Solved using the replica method in the high-dimensional limit



RAPID COMMUNICATIONS

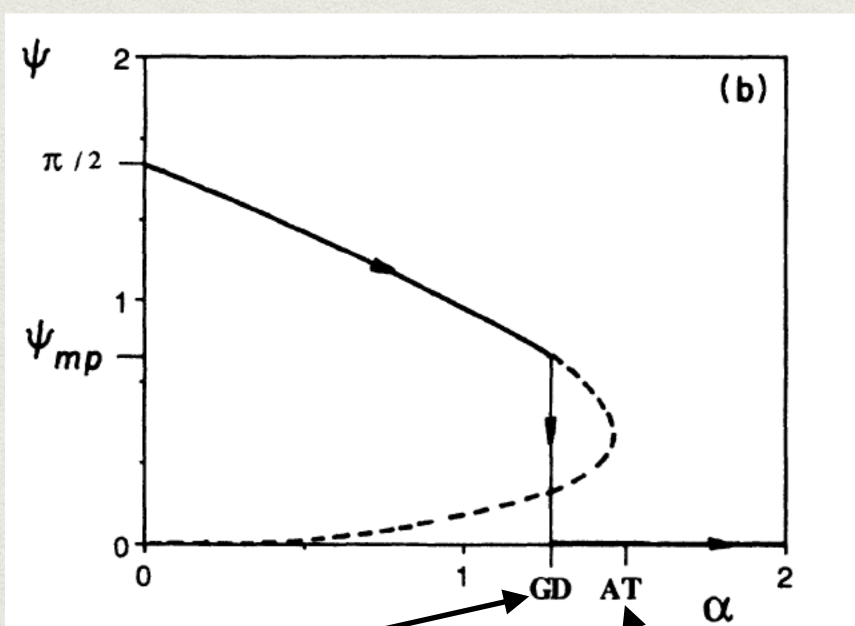
PHYSICAL REVIEW A VOLUME 41, NUMBER 12 15 JUNE 1990

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

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

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

Generalisation error



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

- Binary teacher-weights:

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

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

$\alpha = n/p$

RECENT PROGRESS

- Solution for any activation function, general class of priors on weights.
- Regions of optimality of approximate message passing (=TAP) algorithm.
- Rigorous proof that the replica solution for the teacher-student model is correct.

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

CLOSED-FORM SOLUTION

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

Theorem 1:

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

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

where

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

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

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

CLOSED-FORM SOLUTION

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

Theorem 1:

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

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

Theorem 2: Optimal generalisation error

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

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

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

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

$$\xi \sim P_\xi$$

Algorithm 2 Generalized Approximate Message Passing (G-AMP)

Input: $\mathbf{y}$

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**.

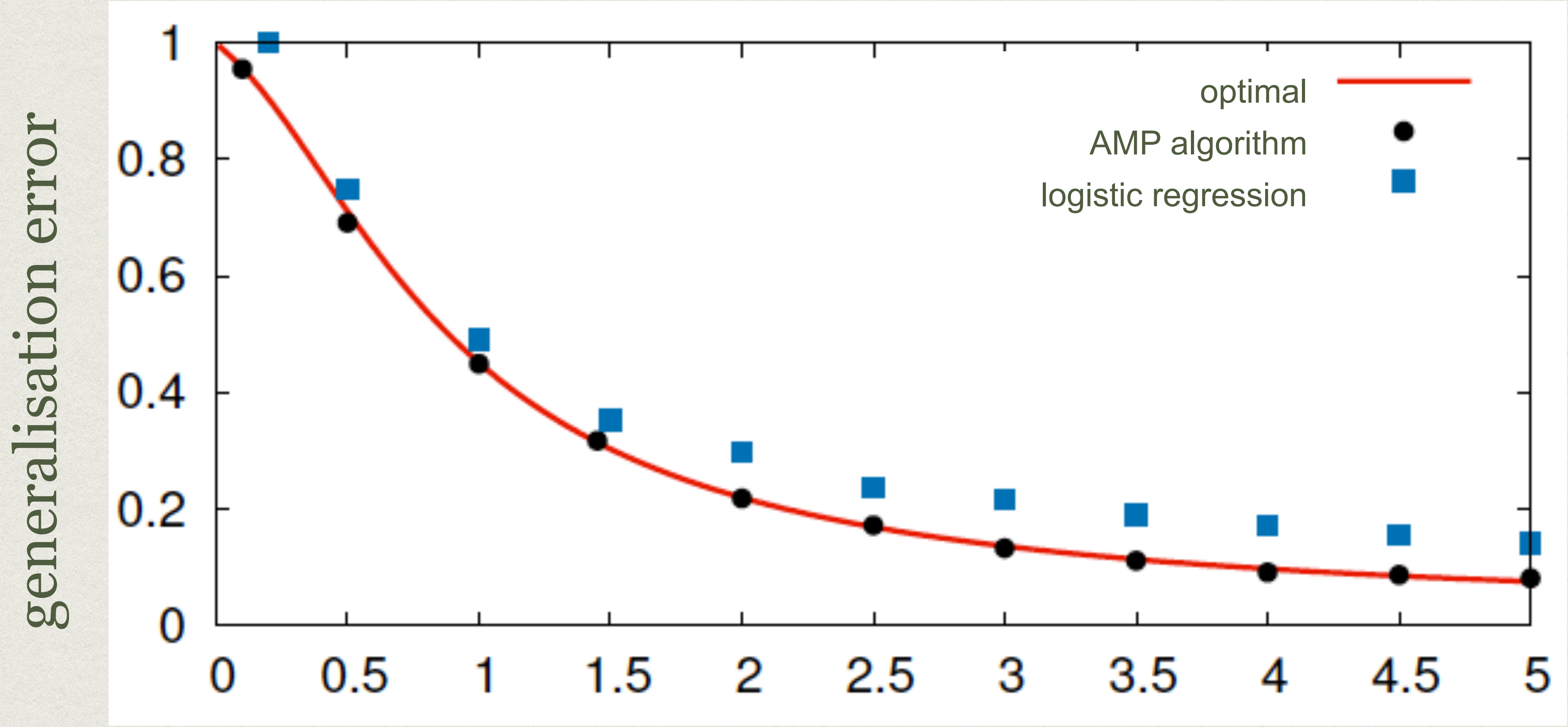
Simple to implement, only matrix multiplications, O(N²)

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

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

SPHERICAL PERCEPTRON

$\varphi(z) = \text{sign}(z)$ $P_w = \mathcal{N}(0,1)$ $n \to \infty$, $p \to \infty$ $n/p = \Omega(1)$

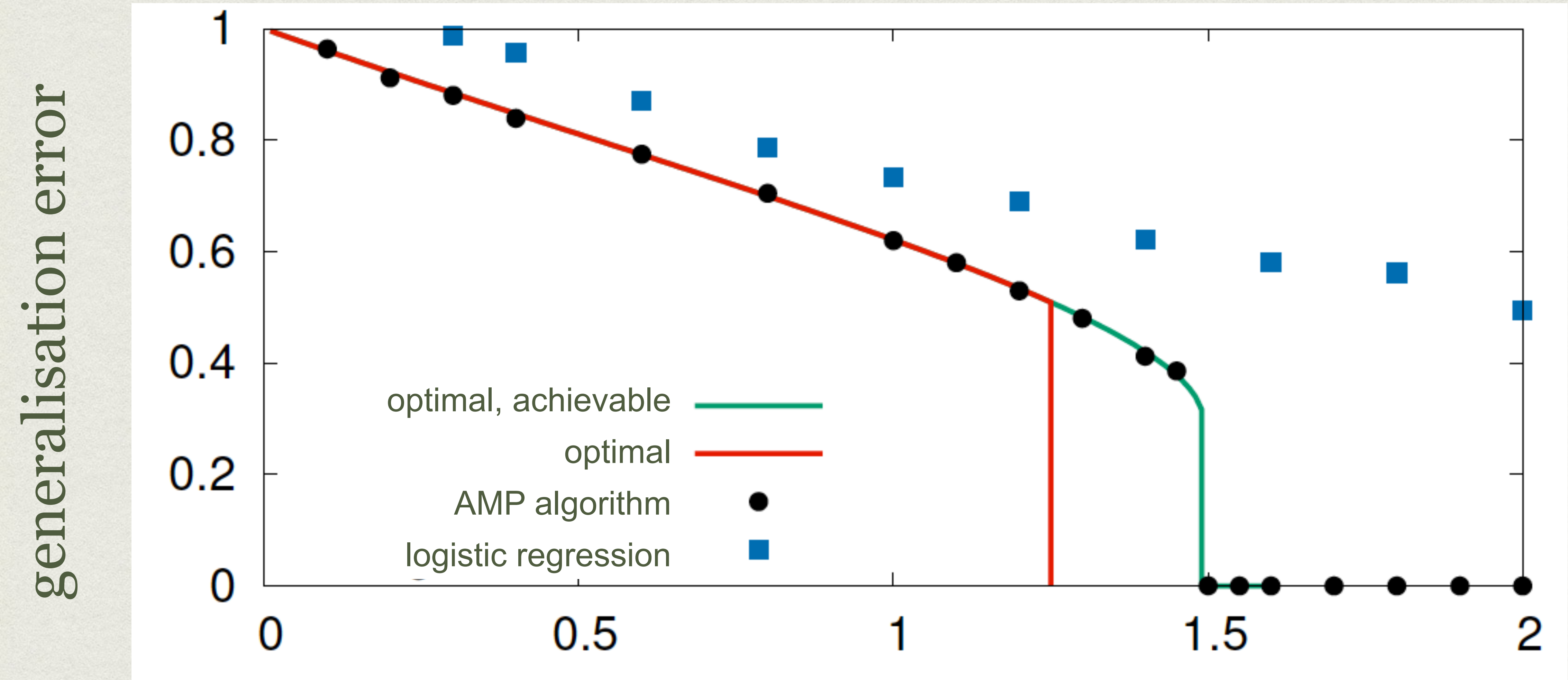


# of samples per dimension n/p

BINARY PERCEPTRON

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

$$n \to \infty \quad p \to \infty \qquad n/p = \Omega(1)$$



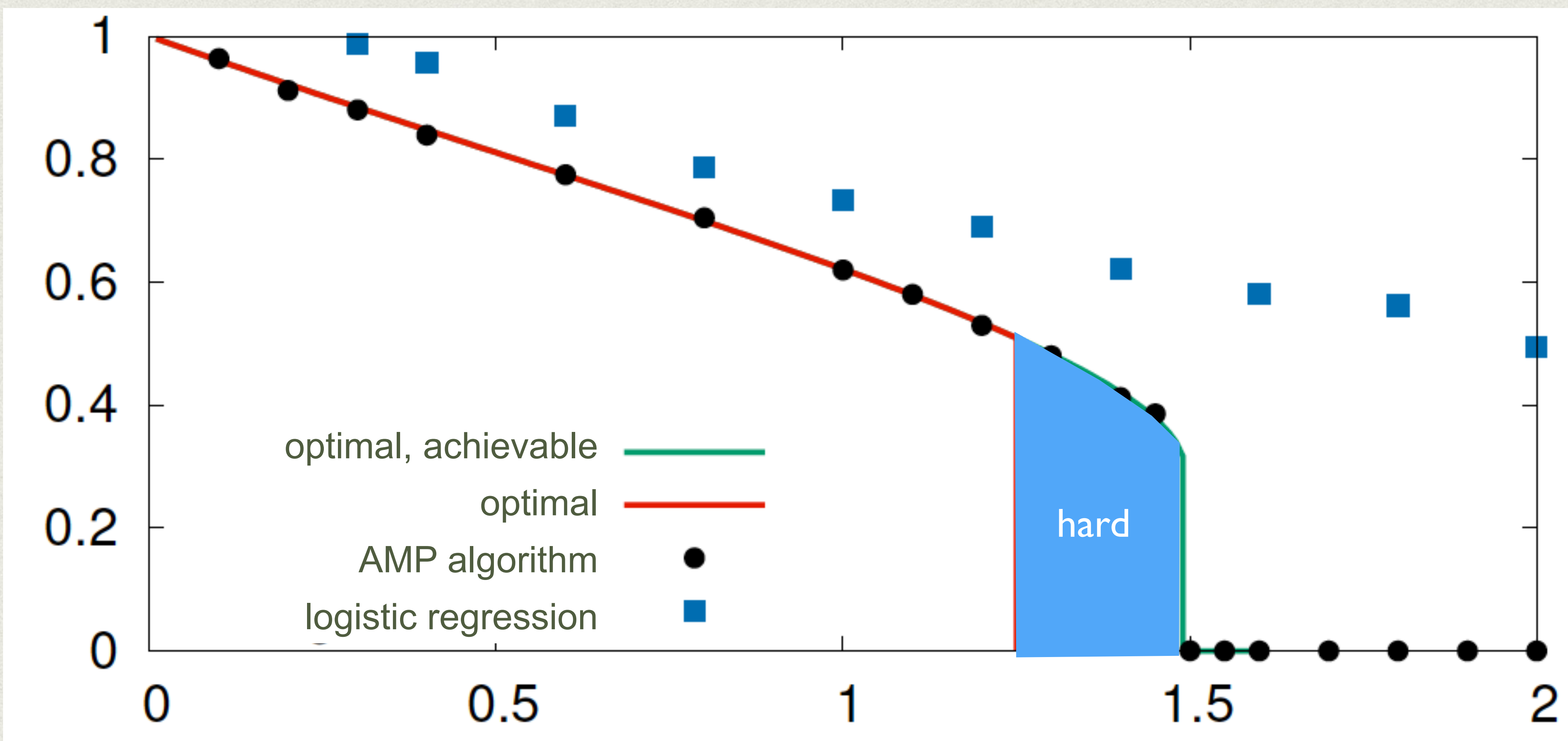
# of samples per dimension n/p

BINARY PERCEPTRON

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

$$n \to \infty \qquad n/p = \Omega(1)$$
$$p \to \infty$$

generalisation error

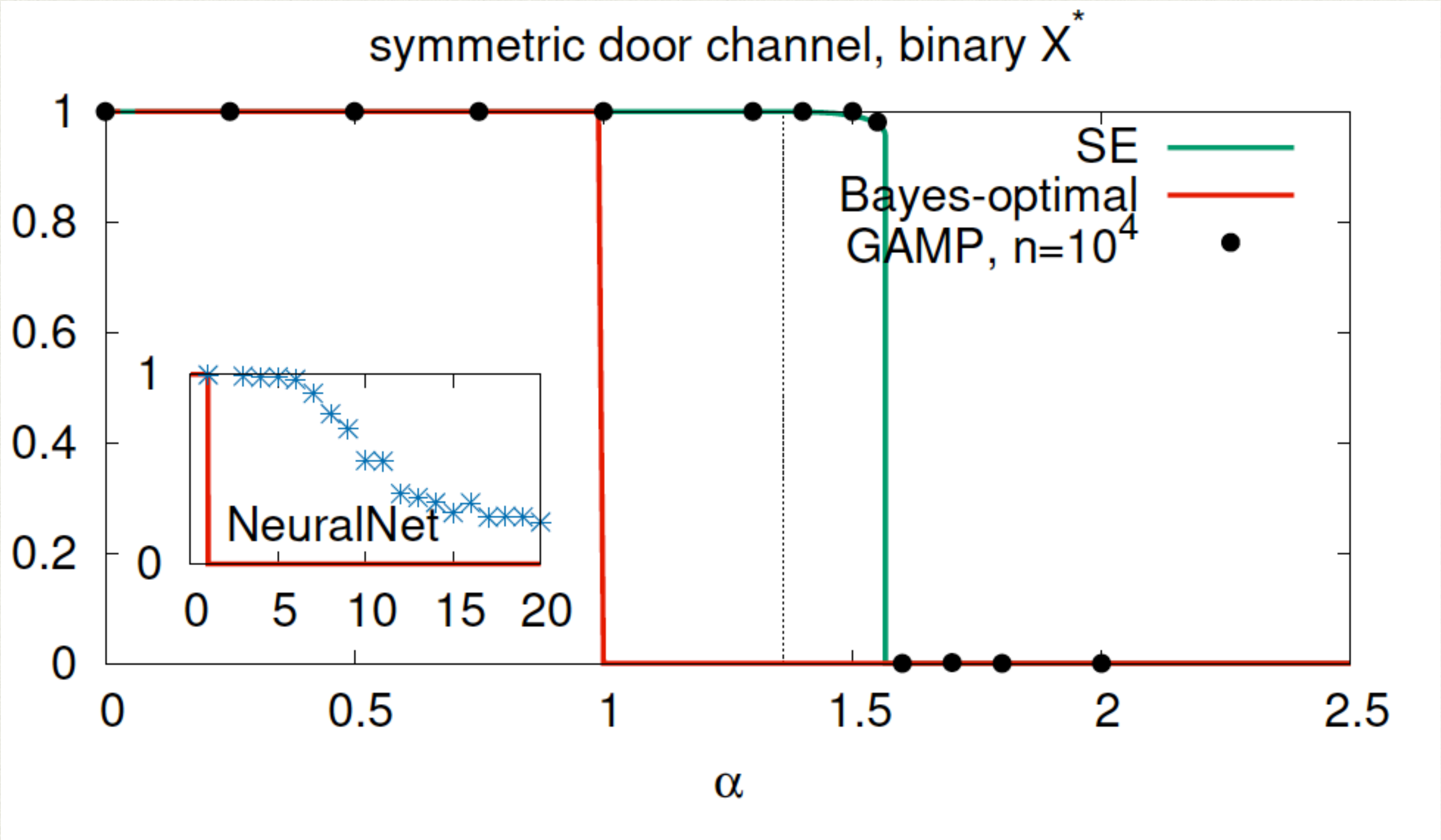


# of samples per dimension n/p

SYMMETRIC-DOOR PERCEPTRON

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

generalisation error



# of samples per dimension n/p

Is this bringing us towards the theory of deep learning?

TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed

structured data

iid inputs, teacher outputs

data structure

multiple wide layers

no hidden units

architecture

algorithm

message passing

gradient-descent-based

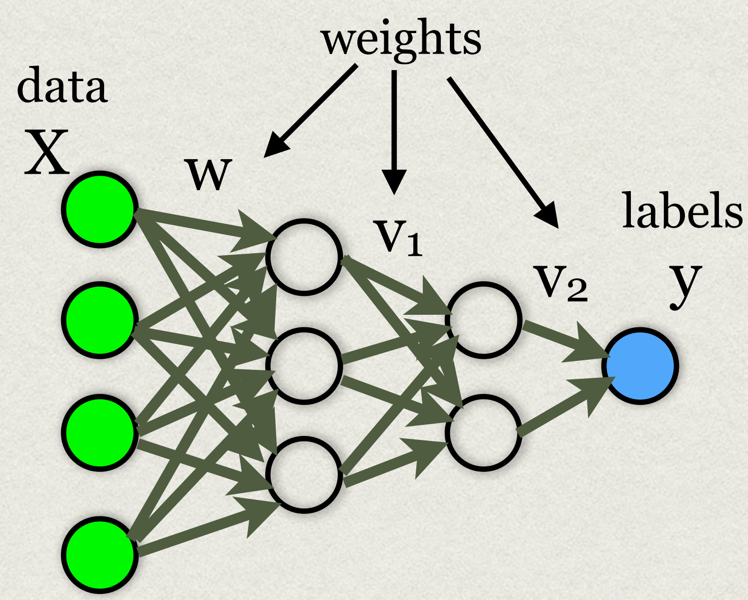
GOING MULTI-LAYER

Committee machine

Model from Schwarze'92.

Proof of the replica formula, and approximate message passing Aubin, Maillard, Barbier, Macris, Krzakala, LZ, NeurIPS'18, arXiv:1806.05451.

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



L=3 layers

w learned, v_1 & v_2 fixed

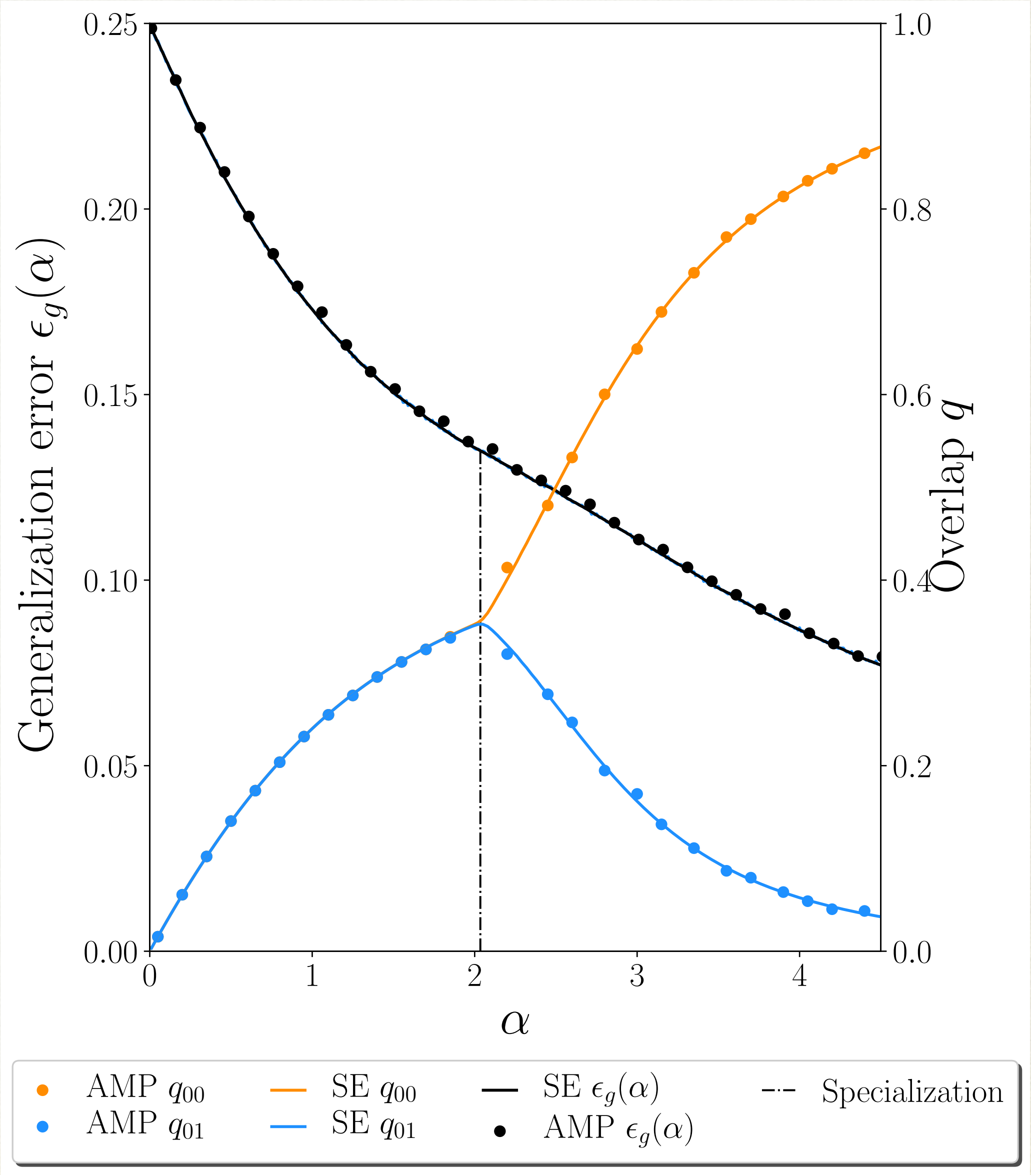
Limit: $\begin{matrix} n \to \infty \\ p \to \infty \end{matrix}$ $\quad \alpha = n/p = \Omega(1) \quad K = O(1)$

SPECIALISATION TRANSITION

hidden units
K=2

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

- Specialization phase transition = hidden units specialise to correlate with specific features.
- Consequence: Sharp threshold for number of samples below which linear regression is the best thing to do.

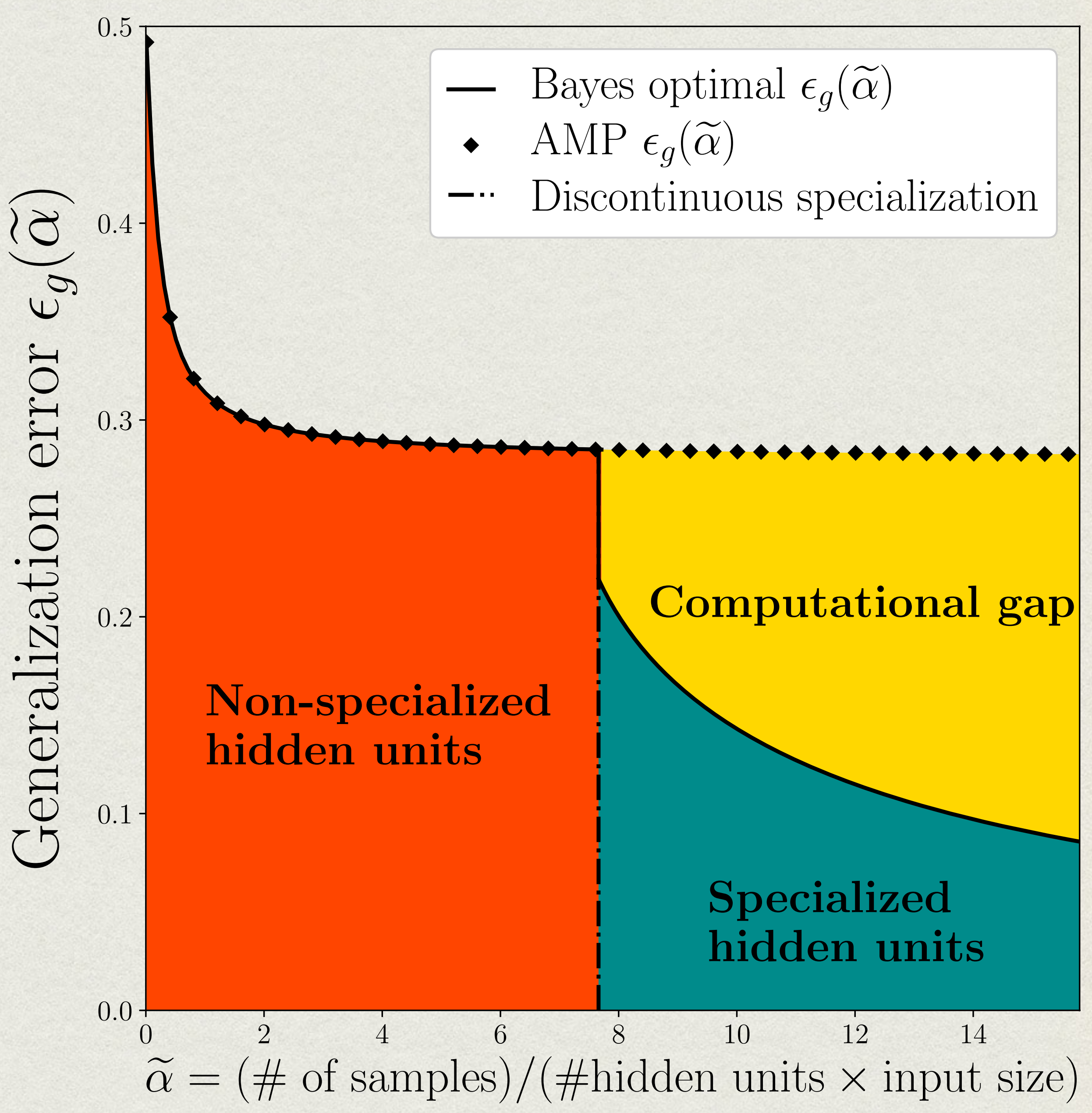


COMPUTATIONAL GAP

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

hidden units $K \gg 1$

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

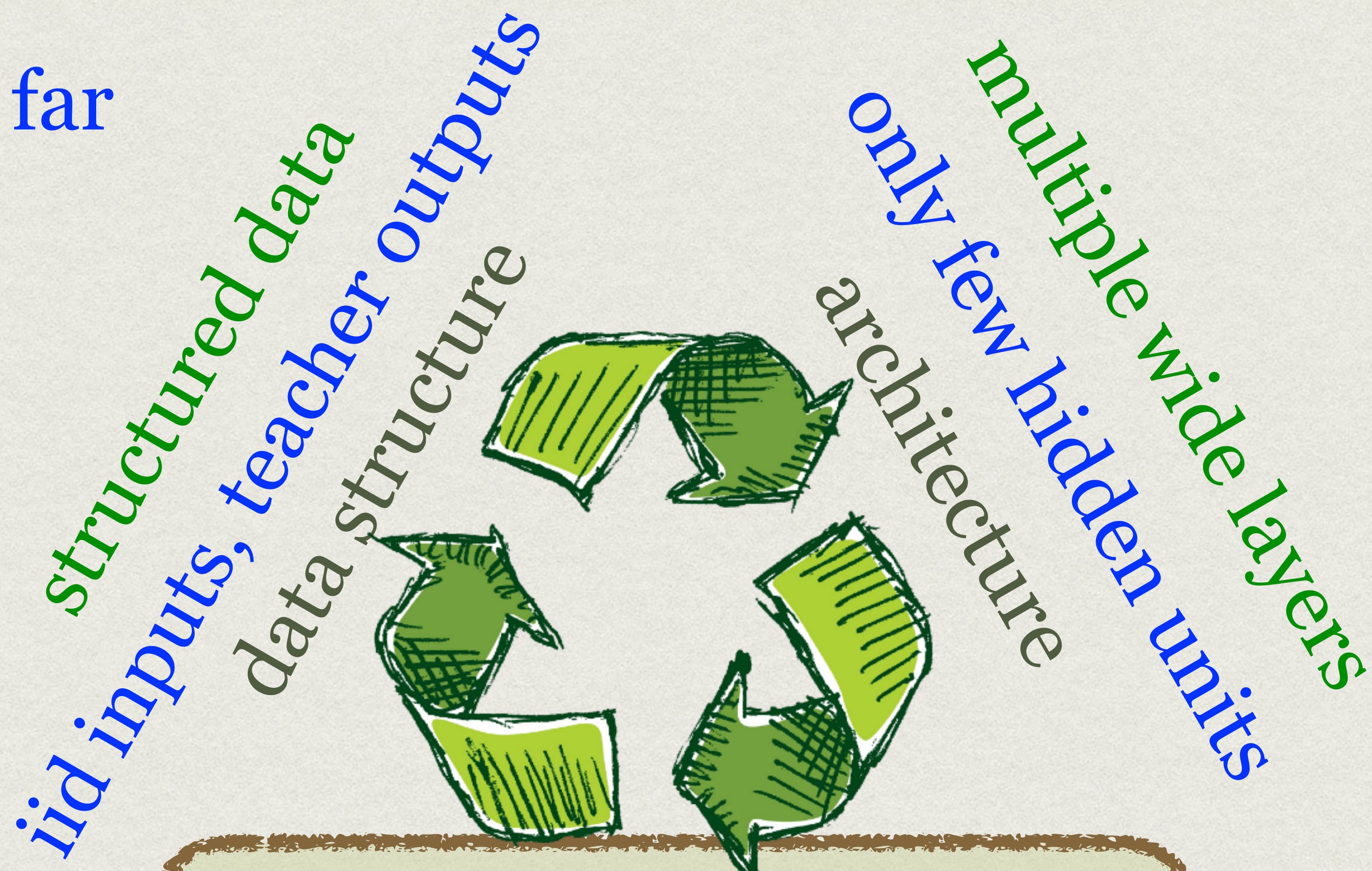


TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed



algorithm

message passing

gradient-descent-based

OPEN QUESTION

How do the gradient-descent-based algorithms compare to the performance of approximate message passing?

Deep learning is fuelled by gradient descent.

Understanding is needed!

Progress recently: Linear networks (trivial fixed point). Lazy training (NTK) networks close to initialization. Infinitely wide single layer networks.

GRADIENT-BASED ALGORITHMS

spherical constraint
(weight decay)

$\langle \eta_i(t)\eta_j(t') \rangle = 2T\delta_{ij}\delta(t-t')$
noise

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

gradient

- T=1 Langevin algorithm: At large time (exponentially) samples the posterior measure.
- T=0 Gradient flow.

Where do they go in large constant time?

MODEL INGREDIENTS

WANTED

- High-dimensional. Non-convex loss. Random perceptron?
- Notion of a "good" configuration (~ generalisation error) beyond lowest-loss configuration. Teacher-student perceptron? Hard to analyze (Agoritsas, Biroli, Urbani, Zamponi'18)
- Solvability: Error of gradient flow and Langevin algorithm follows a closed-form tractable equation. Spiked tensor model?
- Have (hopefully) behaviour that has a large universality class.

MIXED SPIKED MATRIX-TENSOR MODEL

Sarao, Biroli, Cammarota, Krzakala, Urbani, LZ'18

- Signal x* on a sphere, observe a matrix Y and tensor T as:

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

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

- Corresponding Hamiltonian (loss function, log-likelihood)

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

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

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

ESTIMATORS

Bayes-optimal inference = computation of marginals/local magnetization of the Boltzmann measure at T=1.

➡ Langevin algorithm.

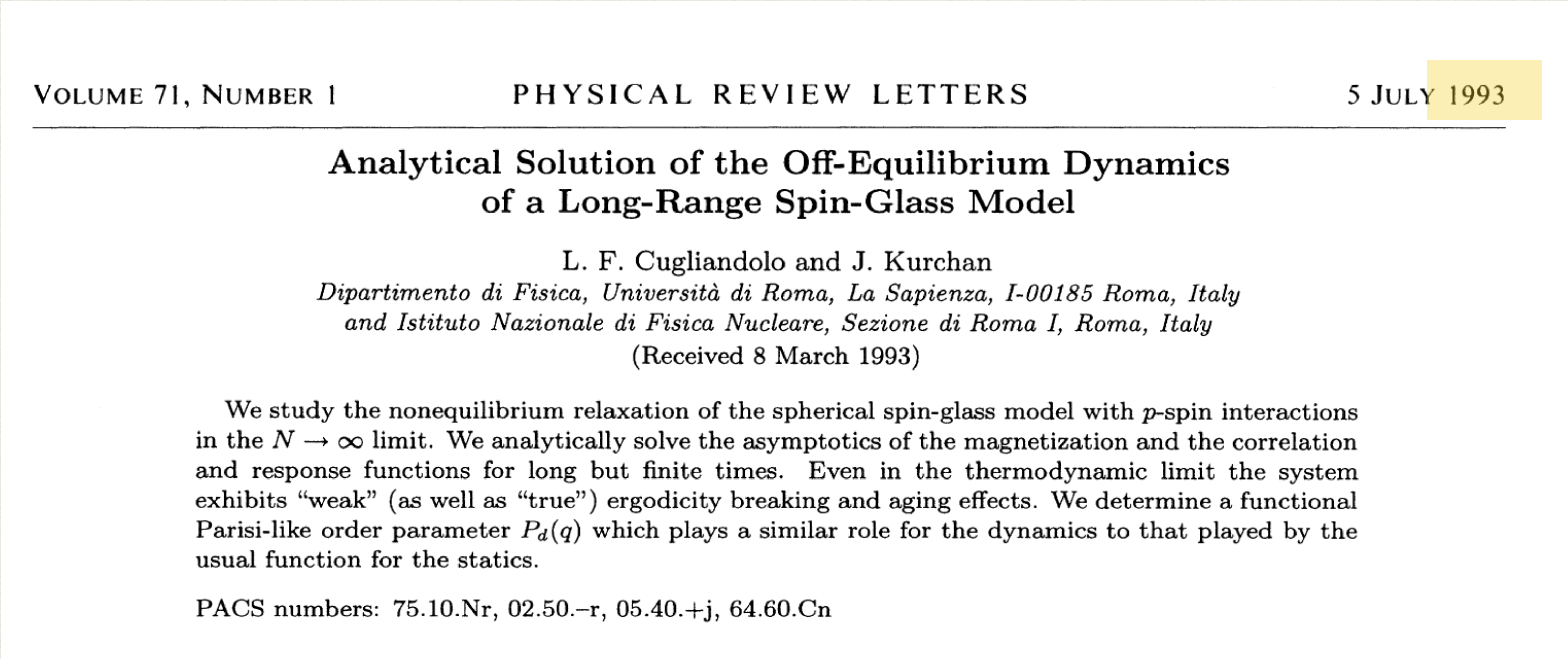
Maximum likelihood inference = computing the ground state.

➡ Gradient flow.

DYNAMICAL MEAN FIELD THEORY

The same model without spike: mixed spherical p-spin glass

Mean field theory of glassy dynamics:


VOLUME 71, NUMBER 1 PHYSICAL REVIEW LETTERS 5 JULY 1993

Analytical Solution of the Off-Equilibrium Dynamics of a Long-Range Spin-Glass Model

L. F. Cugliandolo and J. Kurchan
Dipartimento di Fisica, Università di Roma, La Sapienza, I-00185 Roma, Italy
and Istituto Nazionale di Fisica Nucleare, Sezione di Roma I, Roma, Italy
(Received 8 March 1993)

We study the nonequilibrium relaxation of the spherical spin-glass model with p-spin interactions in the $N \to \infty$ limit. We analytically solve the asymptotics of the magnetization and the correlation and response functions for long but finite times. Even in the thermodynamic limit the system exhibits "weak" (as well as "true") ergodicity breaking and aging effects. We determine a functional Parisi-like order parameter $P_d(q)$ which plays a similar role for the dynamics to that played by the usual function for the statics.

PACS numbers: 75.10.Nr, 02.50.–r, 05.40.+j, 64.60.Cn

Proof of this without spike: BenArous, Dembo, Guionnet'06.

LANGEVIN STATE EVOLUTION

Sarao, Biroli, Cammarota, Krzakala, Urbani, LZ'18&19

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

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

$$N \to \infty$$

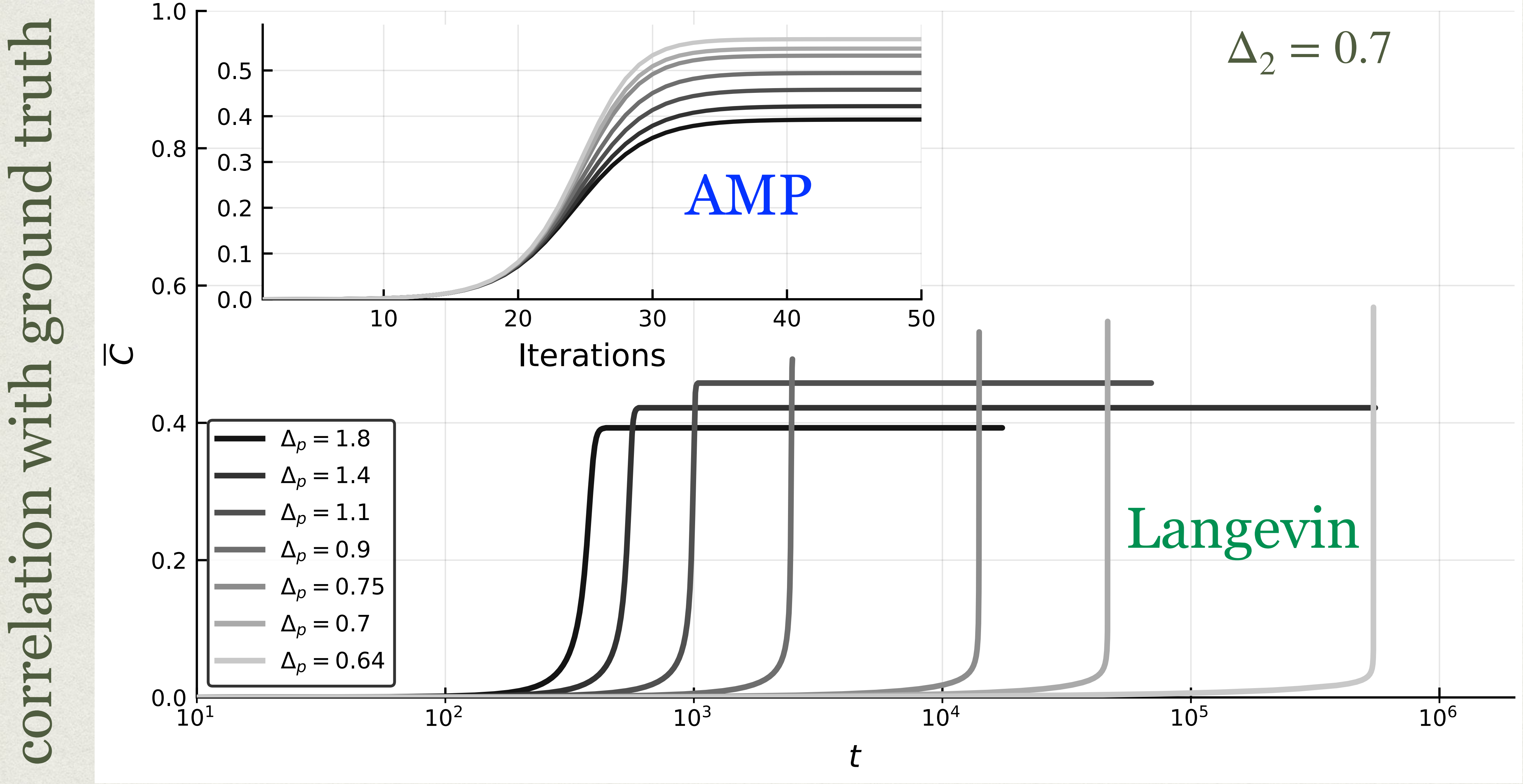
$$
\begin{aligned}
\frac{\partial}{\partial t}C(t,t') &= 2R(t',t) - \mu(t)C(t,t') + Q'(\overline{C}(t))\overline{C}(t') + \int_0^t dt'' R(t,t'')Q''(C(t,t''))C(t',t'') + \int_0^{t'} dt'' R(t',t'')Q'(C(t,t''))\,,\\
\frac{\partial}{\partial t}R(t,t') &= \delta(t-t') - \mu(t)R(t,t') + \int_{t'}^t dt'' R(t,t'')Q''(C(t,t''))R(t'',t')\,,\\
\frac{\partial}{\partial t}\overline{C}(t) &= -\mu(t)\overline{C}(t) + Q'(\overline{C}(t)) + \int_0^t dt'' R(t,t'')\overline{C}(t'')Q''(C(t,t''))\,,
\end{aligned}
$$

Langevin algorithm (T=1)

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

Gradient flow (T=0)

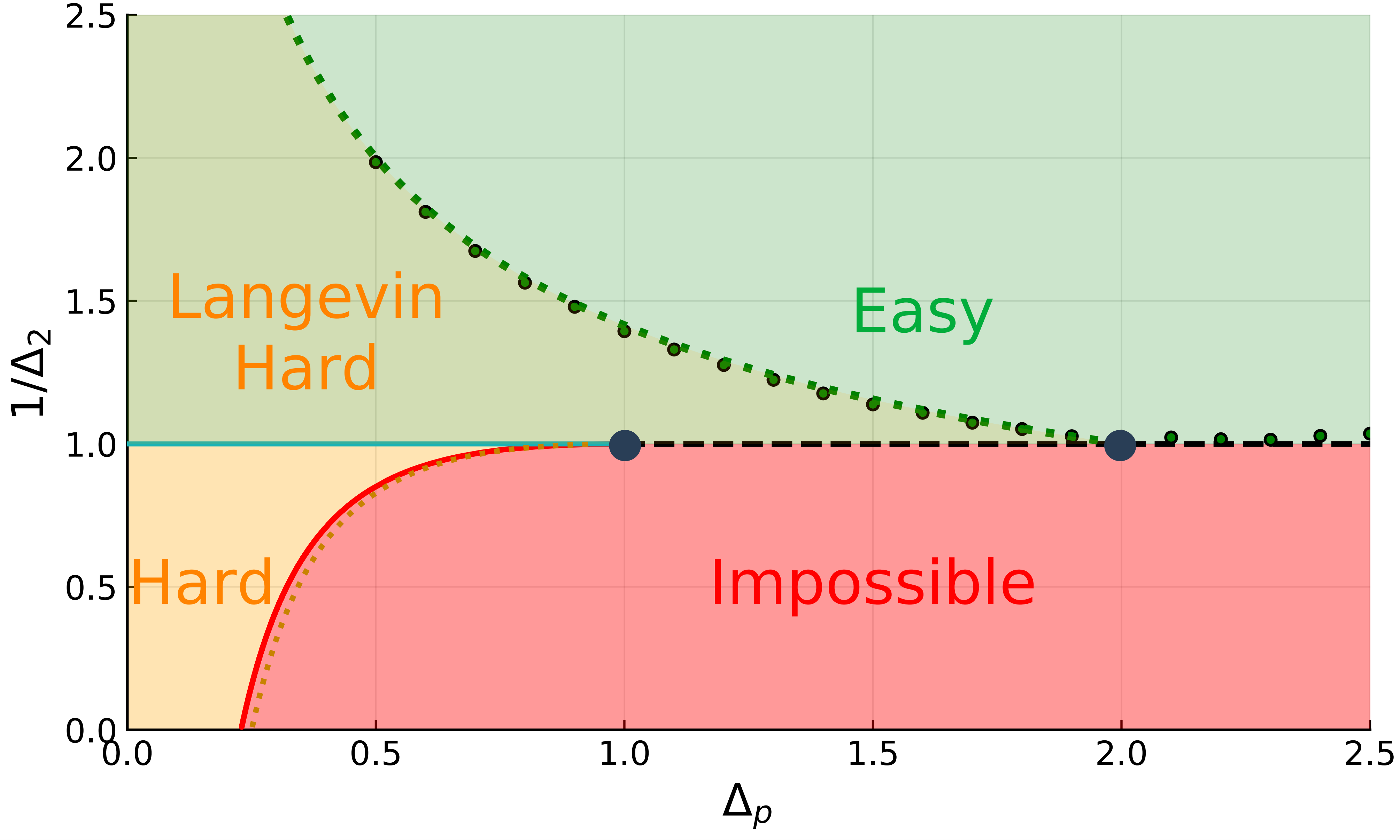
LANGEVIN STATE EVOLUTION (NUMERICAL SOLUTION)



github.com/sphinxteam/spiked_matrix-tensor

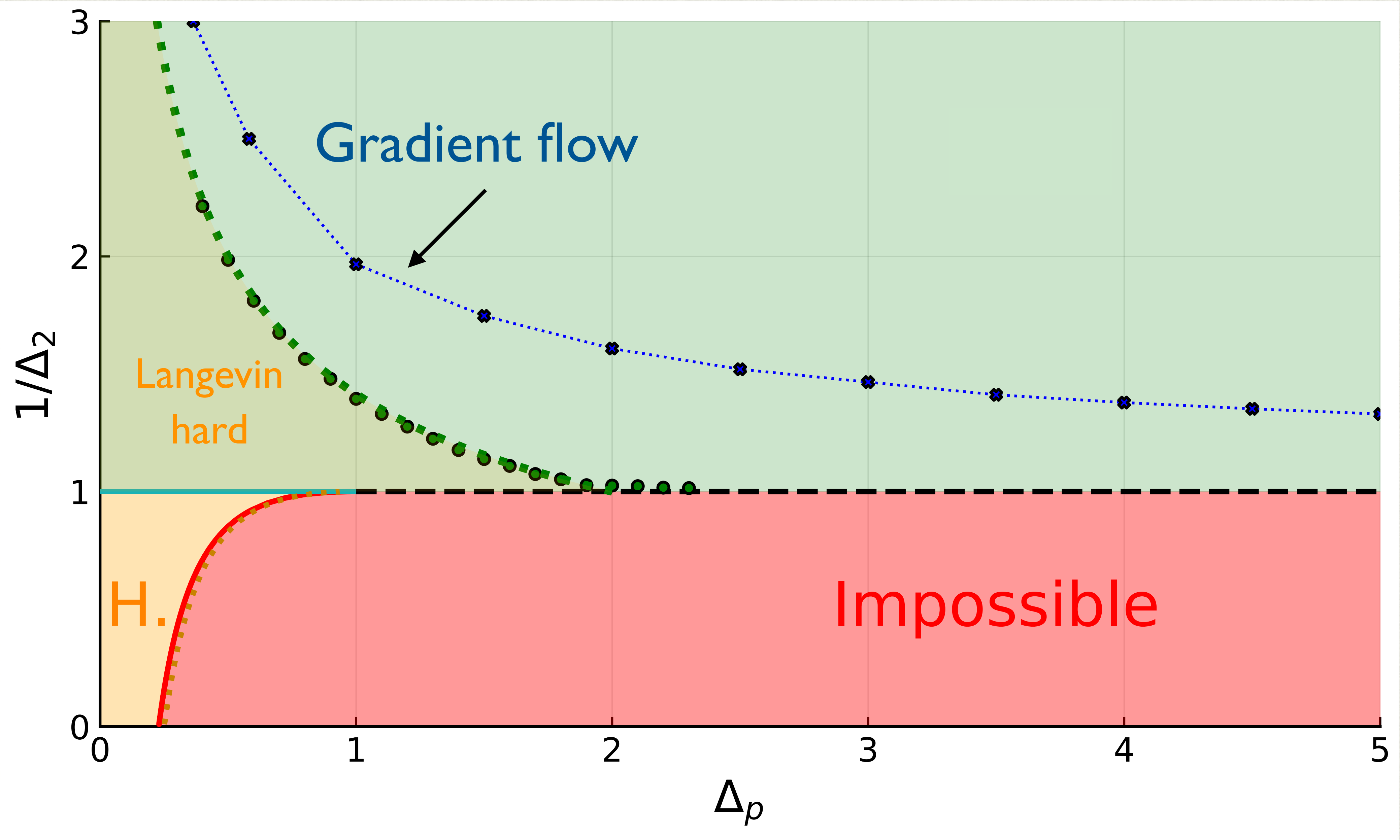
LANGEVIN PHASE DIAGRAM

p=3

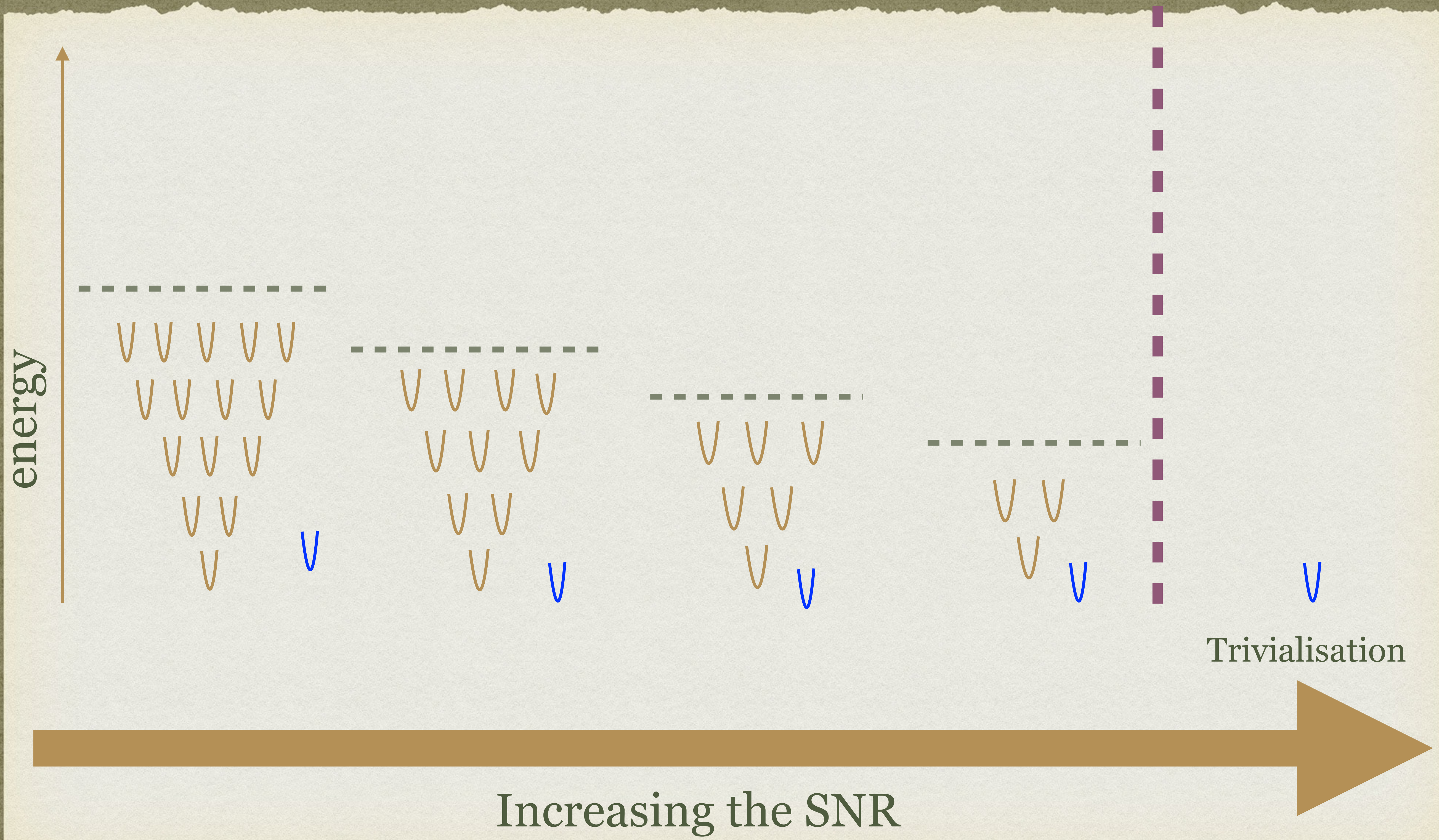


GRADIENT-FLOW PHASE DIAGRAM

p=3



POPULAR "EXPLANATION"



COUNTING MINIMA: KAC-RICE

Sarao, Biroli, Cammarota, Krzakala, Urbani, LZ'19

Annealed entropy of local minima (at m=0 also quenched):

$$\tilde{\Sigma}_{\Delta_2,\Delta_p}(m,\epsilon_2,\epsilon_p) = \frac{1}{2}\log\frac{\frac{p-1}{\Delta_p}+\frac{1}{\Delta_2}}{\frac{1}{\Delta_p}+\frac{1}{\Delta_2}} + \frac{1}{2}\log(1-m^2)$$
$$-\frac{1}{2}\frac{\left(\frac{m^{p-1}}{\Delta_p}+\frac{m}{\Delta_2}\right)^2}{\frac{1}{\Delta_p}+\frac{1}{\Delta_2}}(1-m^2) - \frac{p\Delta_p}{2}\left(\epsilon_p+\frac{m^p}{p\Delta_p}\right)^2$$
$$-\Delta_2\left(\epsilon_2+\frac{m^2}{2\Delta_2}\right)^2 + \Phi(t) - L(\theta,t),$$

Similar to Ben Arous, Mei, Song, Montanari, Nica'17; Ros, Ben Arous, Biroli, Cammarota'18 for spiked tensor model

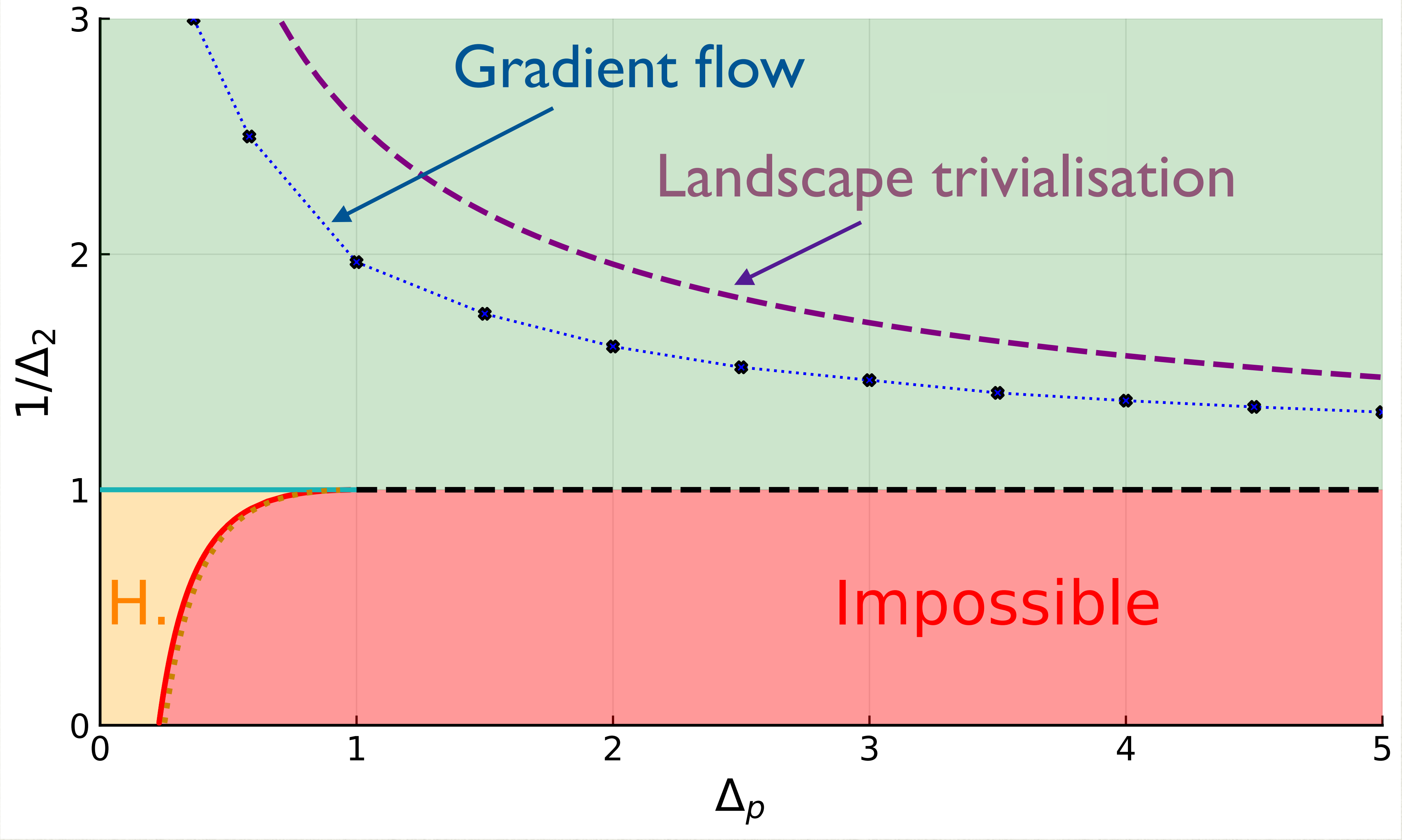
where:

$$\Phi(t) = \frac{t^2}{4} + \mathbb{1}_{|t|>2}\left[\log\left(\sqrt{\frac{t^2}{4}-1}+\frac{|t|}{2}\right) - \frac{|t|}{4}\sqrt{t^2-4}\right]$$

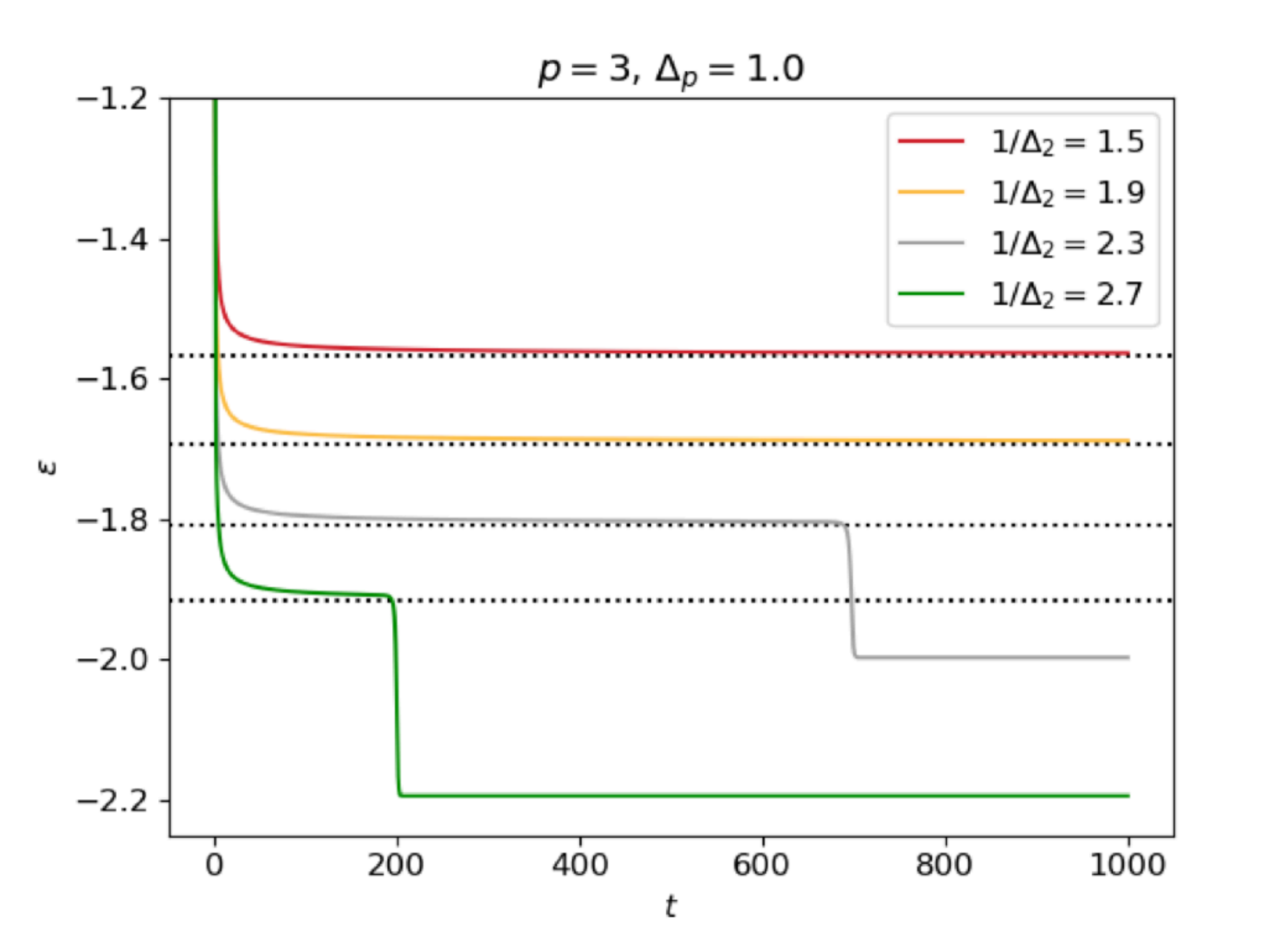
$$L(\theta,t) = \begin{cases} \frac{1}{4}\int_{\theta+\frac{1}{\theta}}^{t}\sqrt{y^2-4}\,dy - \frac{\theta}{2}\left(t-\left(\theta+\frac{1}{\theta}\right)\right) \\ \quad + \frac{t^2-\left(\theta+\frac{1}{\theta}\right)^2}{8} & \theta>1,\ 2\le t<\frac{\theta^2+1}{\theta} \\ \infty & t<2 \\ 0 & \text{otherwise.} \end{cases}$$

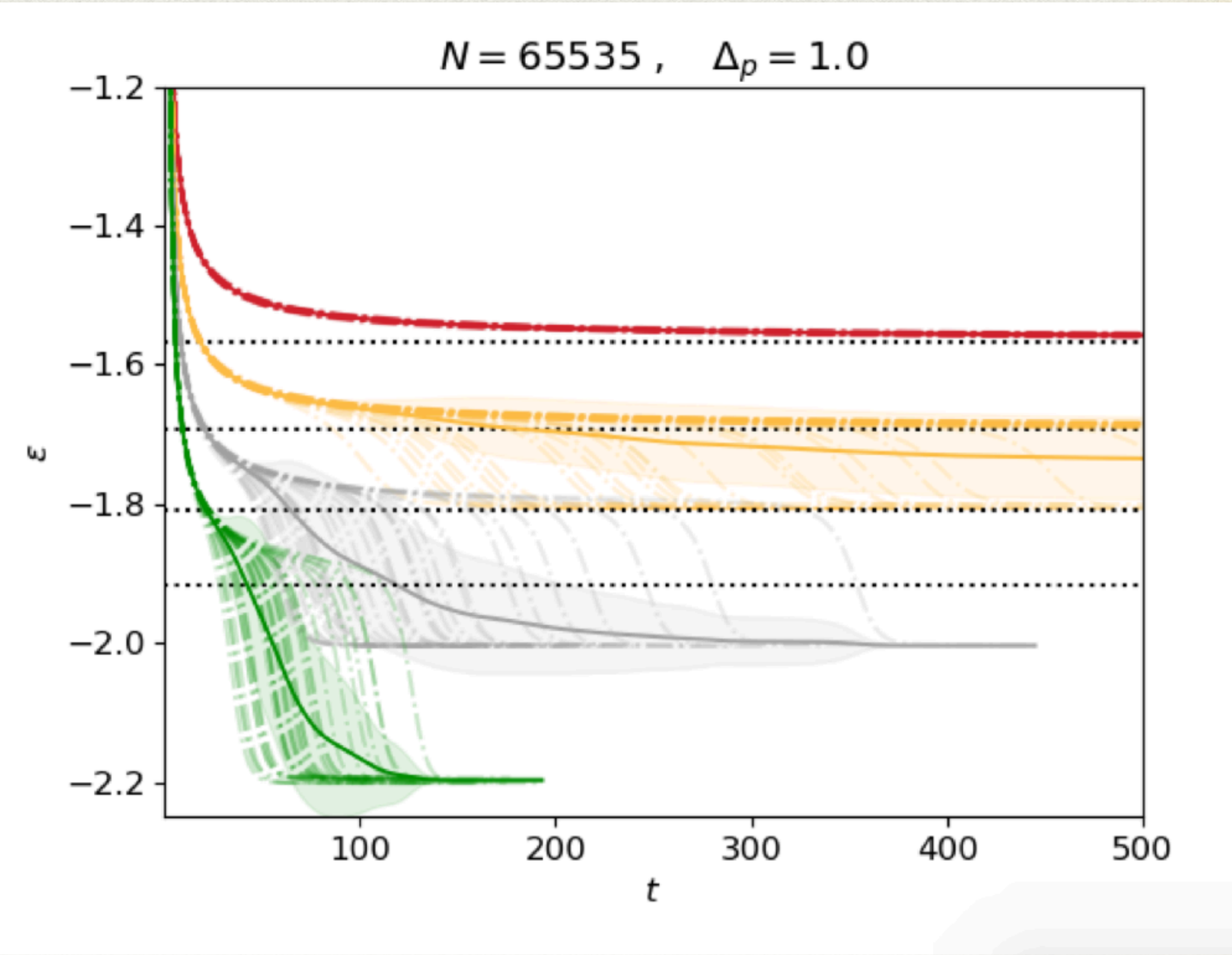
SPURIOUS MINIMA DO NOT NECESSARILY CAUSE GF TO FAIL

p=3



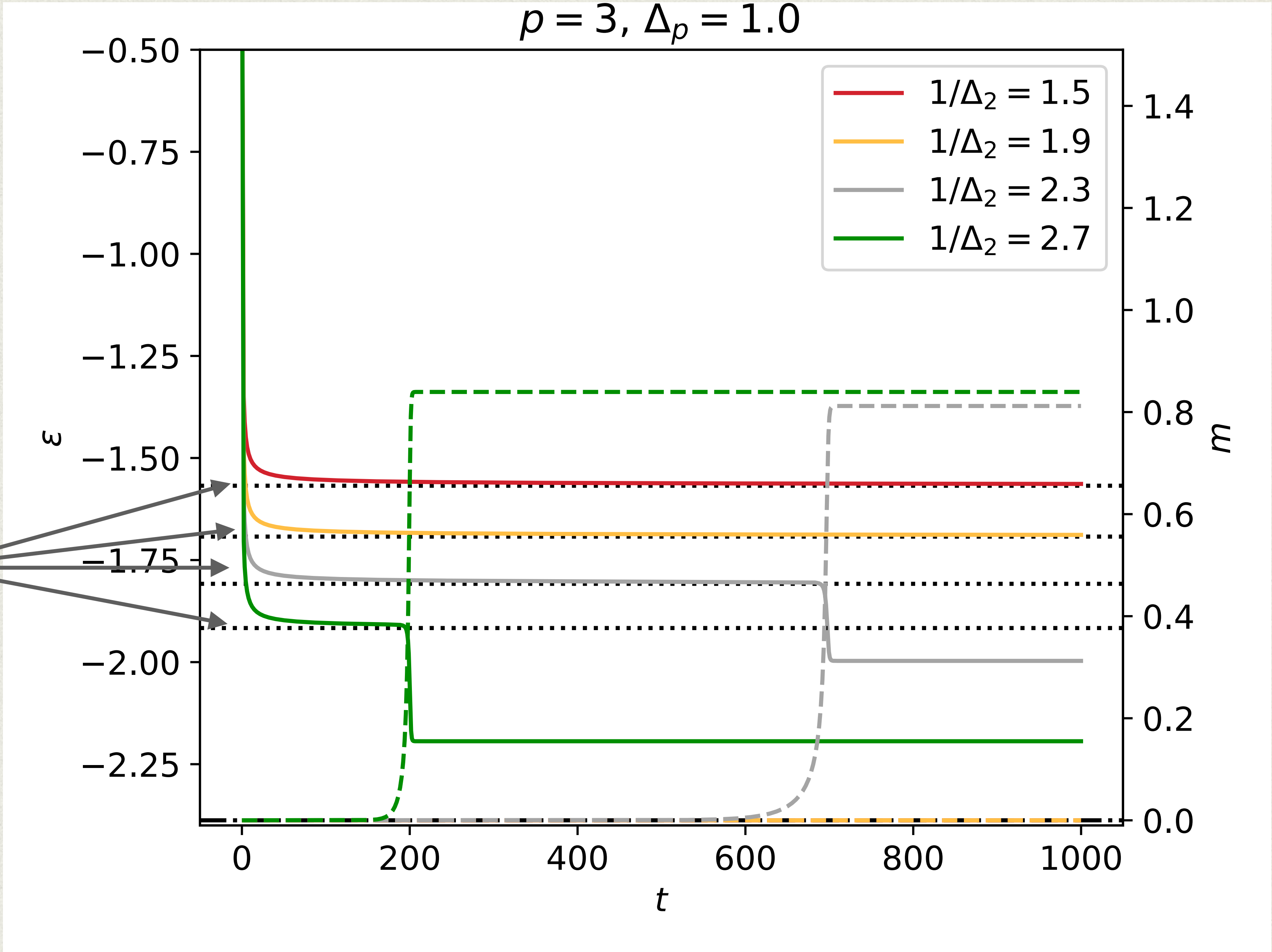
WHAT IS GOING ON?





WHAT IS GOING ON?

Threshold energy in the non-planted model (m=0)



TRANSITION RECIPE

Dynamics first goes to the threshold states (replicon condition):

$$\frac{T^2}{(1-q^{\text{th}})^2} = (p-1)\frac{(q^{\text{th}})^{p-2}}{\Delta_p} + \frac{1}{\Delta_2}$$

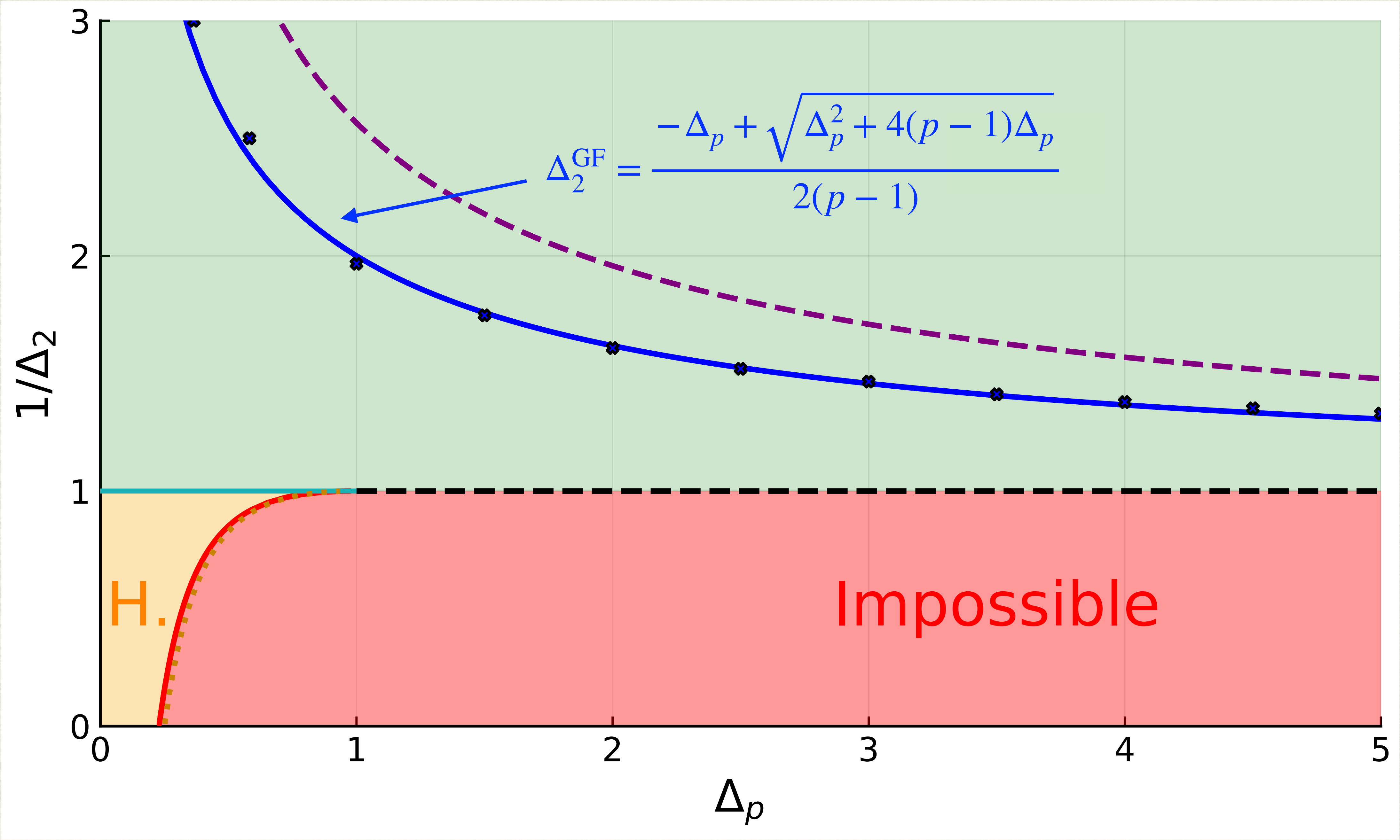
AMP state evolution at fixed q, determines stability of T=0:

$$m^{t+1} = \frac{1-q}{T}\left(\frac{m^t}{\Delta_2} + \frac{(m^t)^{p-2}}{\Delta_p}\right)$$

Leads to the Langevin/gradient-flow transition (conjecture):

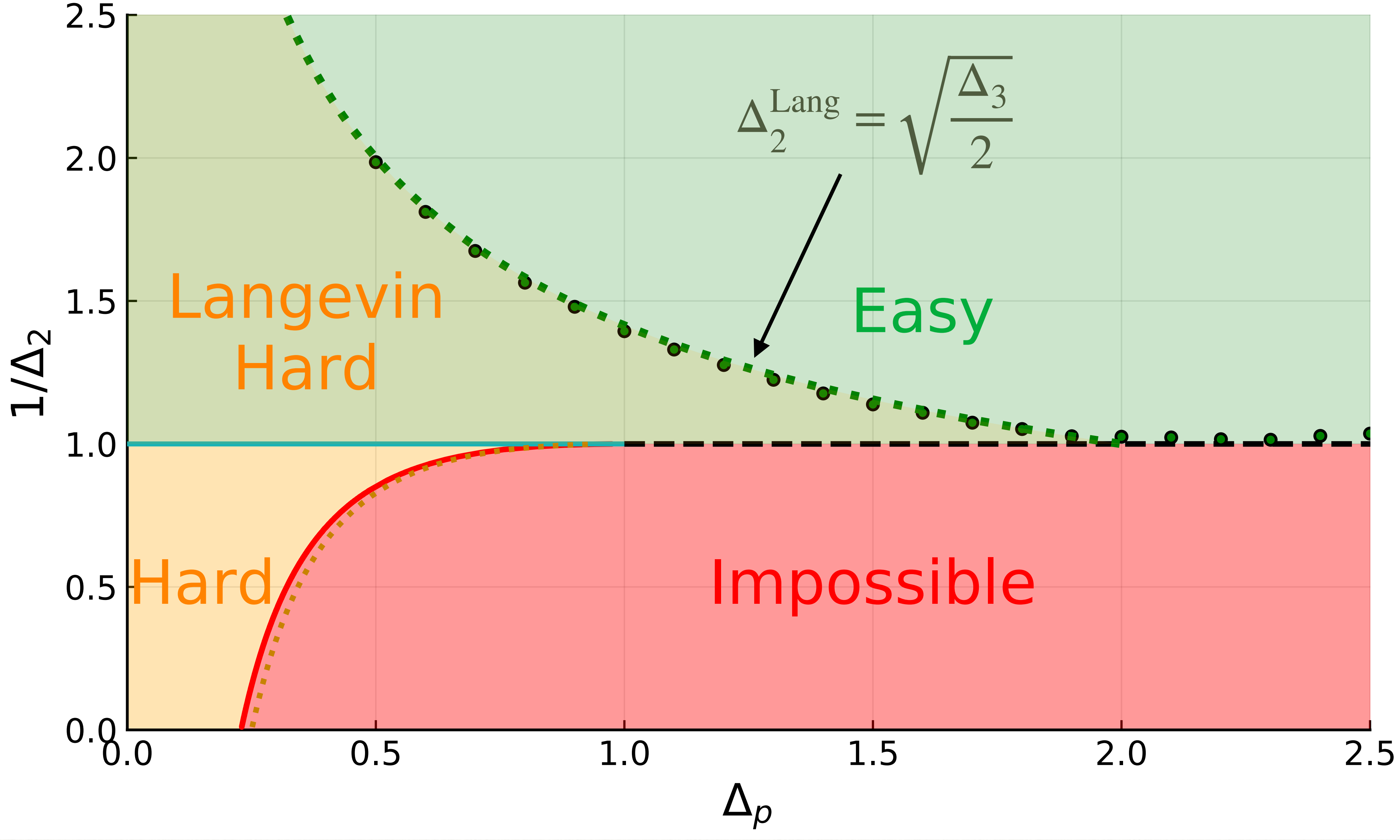
$$\frac{1}{\Delta_2^2} = (p-1)\frac{(1-T\Delta_2)^{p-2}}{\Delta_p} + \frac{1}{\Delta_2}$$

GRADIENT-FLOW PHASE DIAGRAM



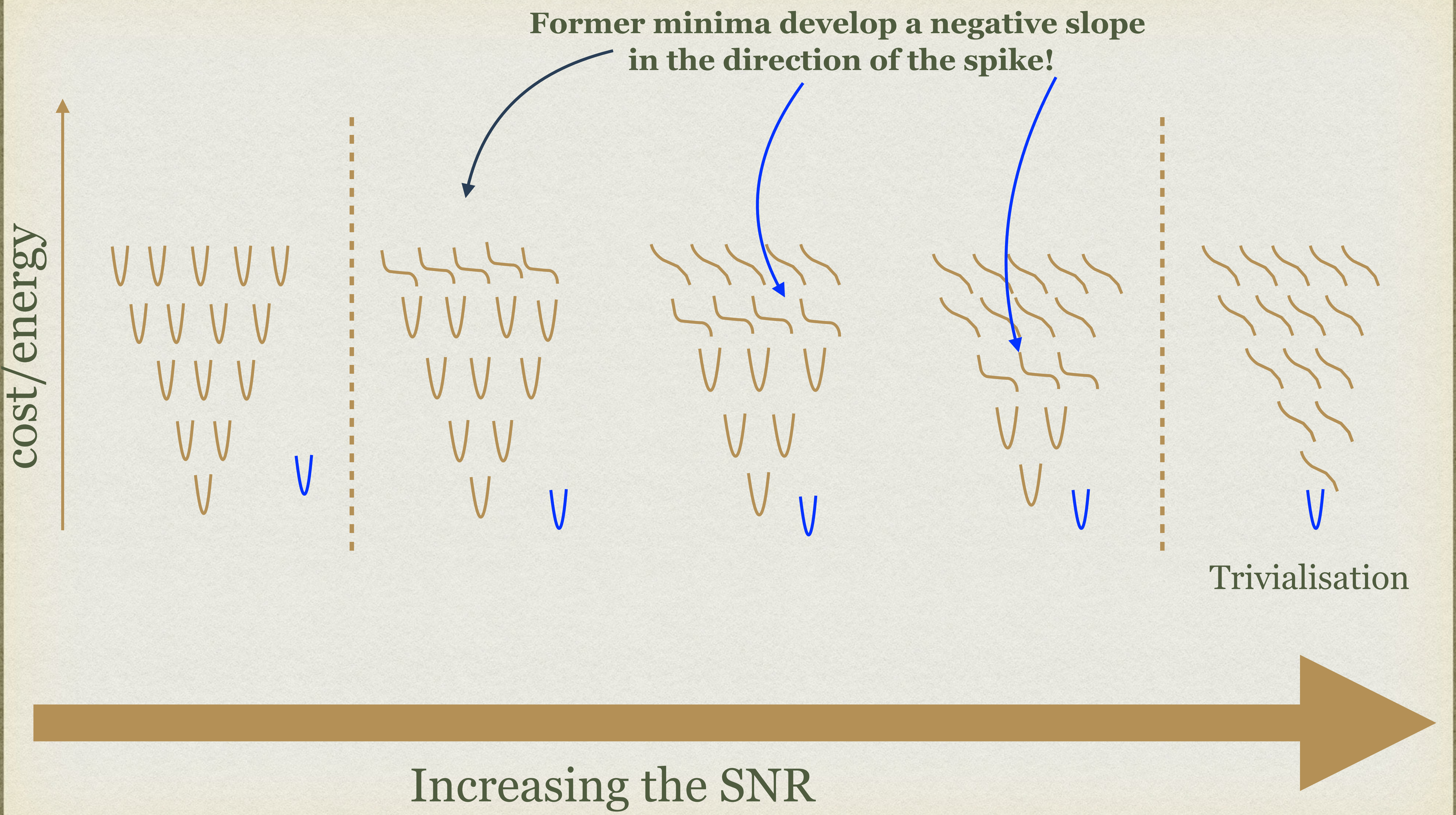
LANGEVIN PHASE DIAGRAM

p=3



LANDSCAPE ANALYSIS

Sarao, Biroli, Cammarota, Krzakala, LZ'19



CONCLUSION ON GRADIENT-ALGORITHMS

- Gradient flow worse than Langevin algorithm, both worse than AMP. How can GF & LA be improved to reach the AMP threshold?
- Gradient flow (sometimes) works even when spurious local minima are present. Quantified with the Kac-Rice approach.
- First time we have a closed-form conjecture for the threshold of gradient-based algorithms including constants. Applicable beyond the present model?

TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed

data structure

iid inputs, teacher outputs

structured data

architecture

only few hidden units

multiple wide layers

algorithm

message passing

gradient-descent-based

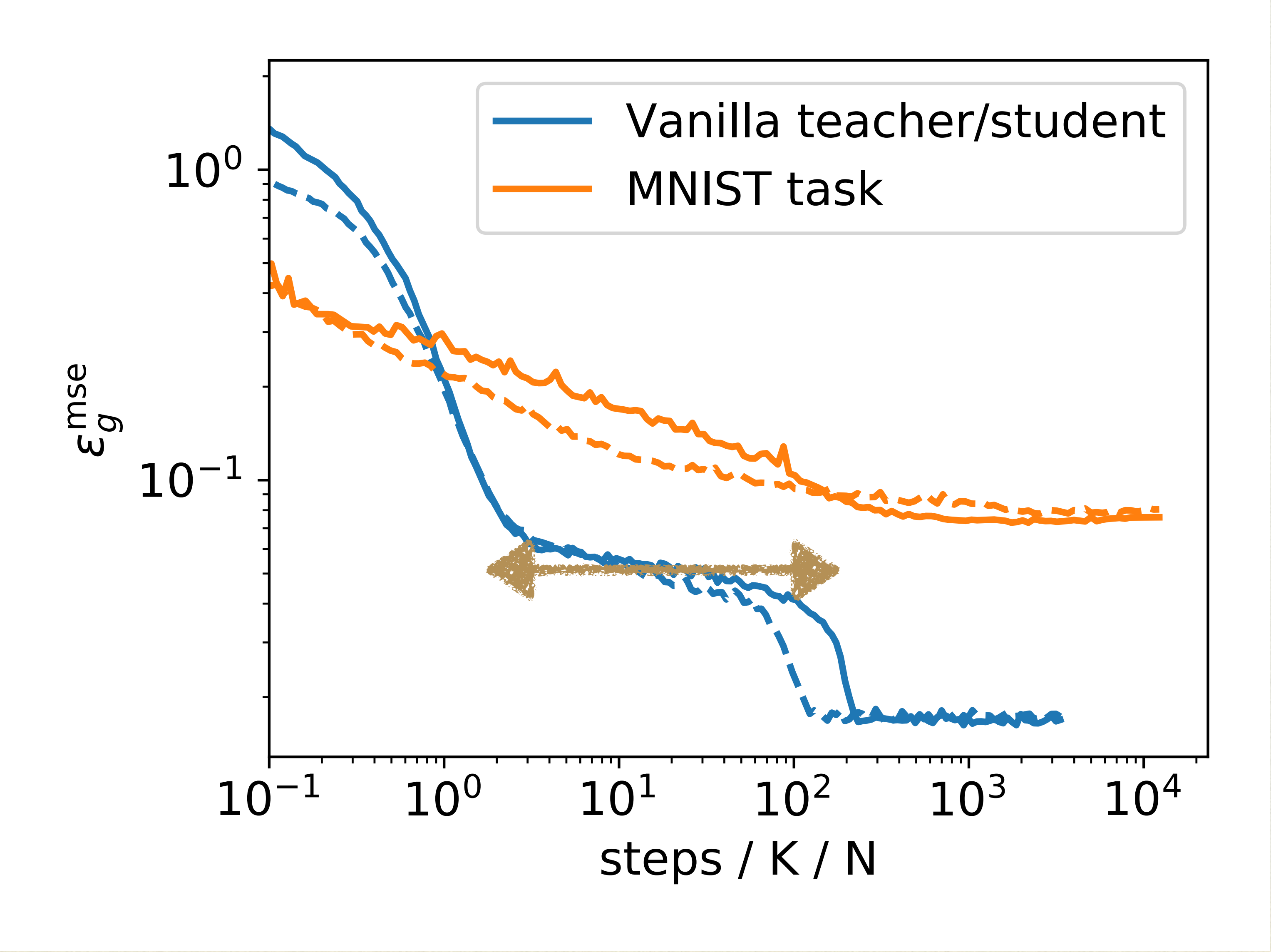
MNIST VS TEACHER/STUDENT

Goldt, Krzakala, Mézard, LZ; arXiv:1909.11500

Teacher/student:

Plateau in learning dynamics, due to specialisation (Saad, Solla'95).

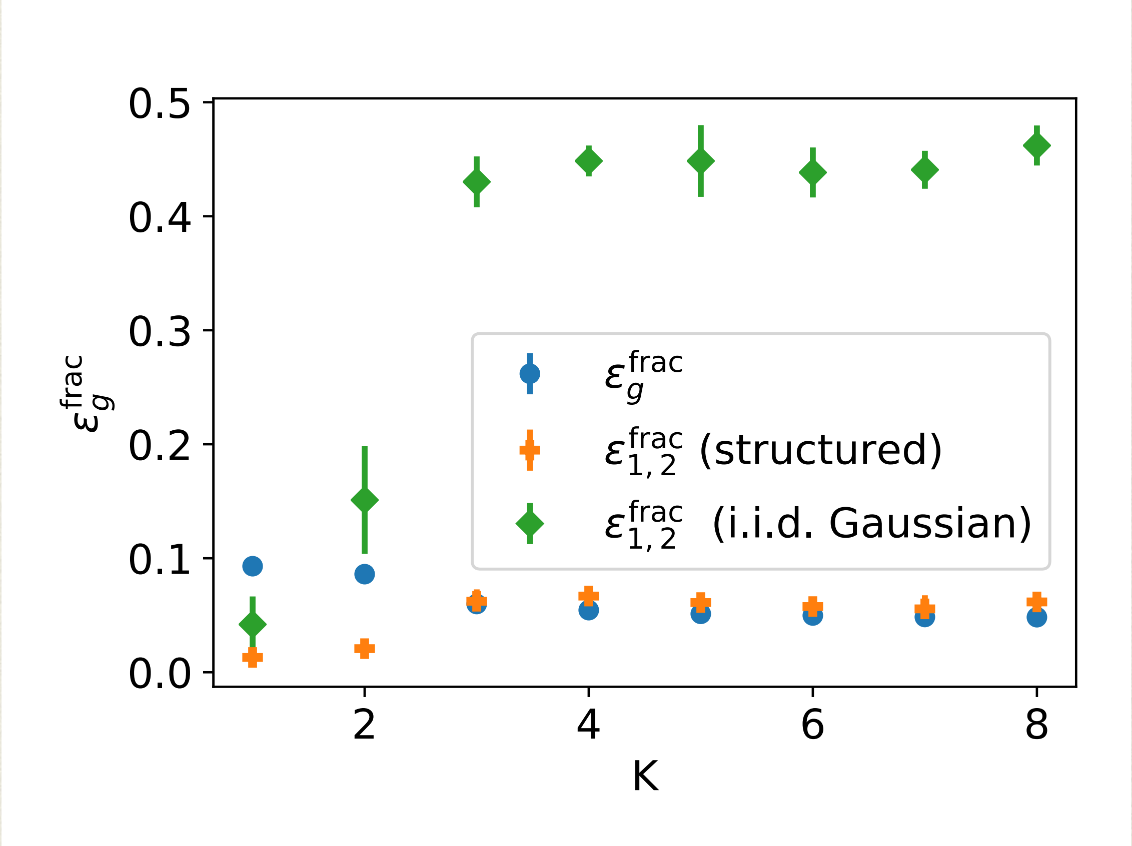
MNIST (even vs odd classification): No plateau …



MNIST VS TEACHER/STUDENT

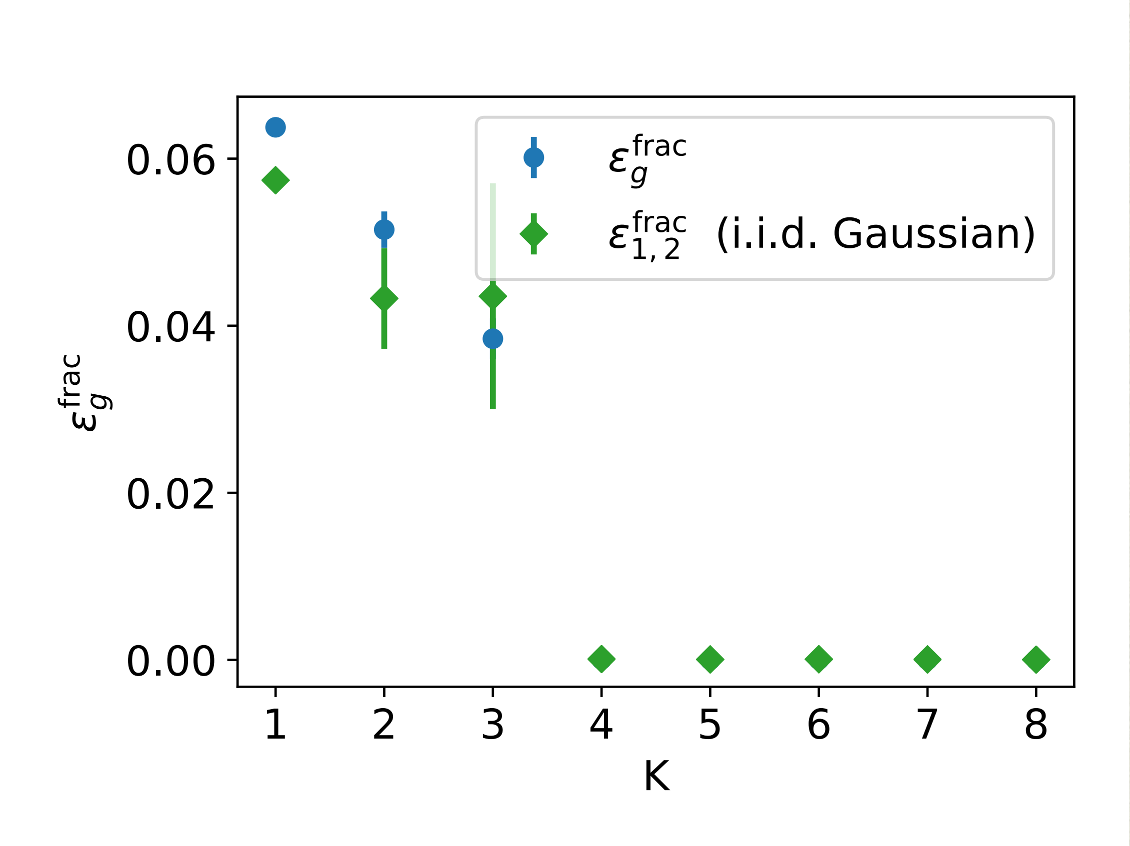
MNIST (odd vs even):

Two independent students do not learn the same function!



Teacher/student:

Two independent students learn the same function!



HIDDEN MANIFOLD MODEL

Input data: $X \in \mathbb{R}^{n \times p}$ $C \in \mathbb{R}^{n \times d}$ $F \in \mathbb{R}^{d \times p}$

n samples, p input & d latent dimension.

Input on low-dimensional manifold.

$$X = f(CF)$$

C, F iid matrices.

True labels:

Depend on the latent coordinates C.

$$\tilde{y}_\mu = \sum_{m=1}^{M} \tilde{v}_m \, g\left(\langle \tilde{\mathbf{w}}_m, \mathbf{C}_\mu \rangle\right)$$

Vanilla teacher/student

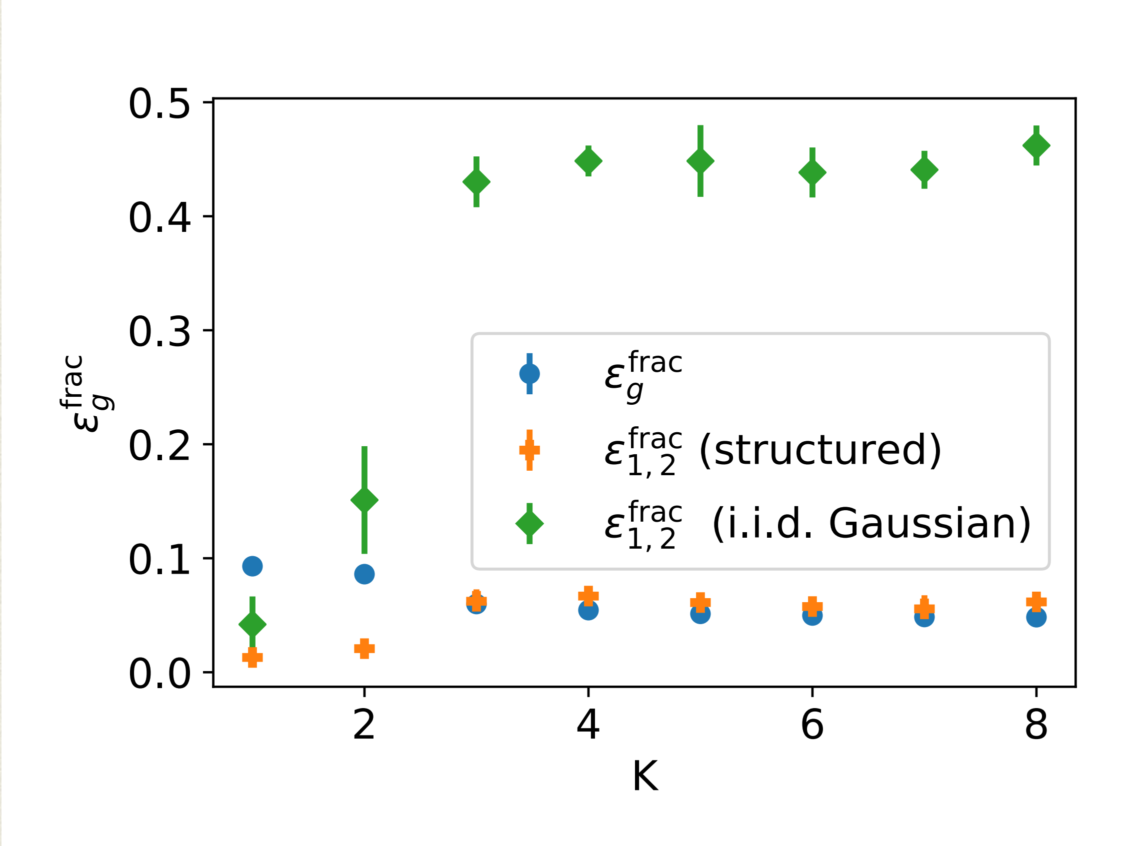
X is iid matrix

$$y_\mu = \sum_{m=1}^{M} v_m \, g\left(\langle \mathbf{w}_m, \mathbf{X}_\mu \rangle\right)$$

MNIST VS HIDDEN MANIFOLD

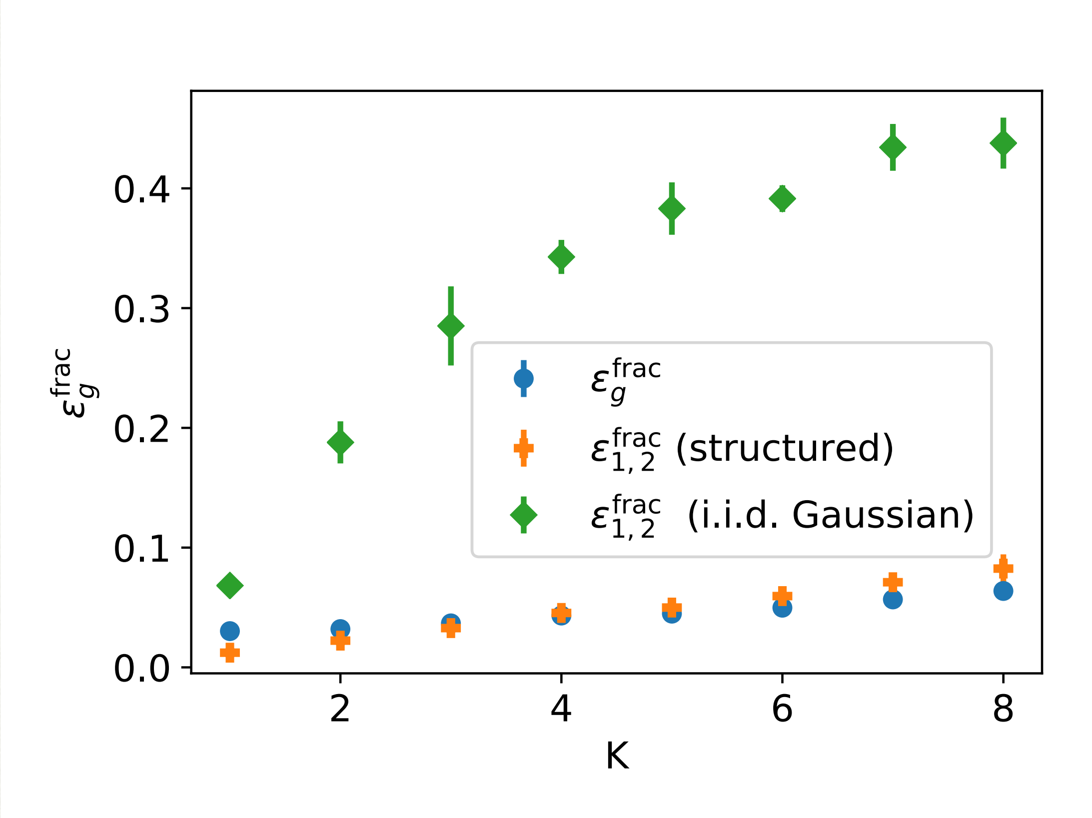
MNIST (odd vs even):

Two independent students do not learn the same function!



Hidden manifold (d=10)

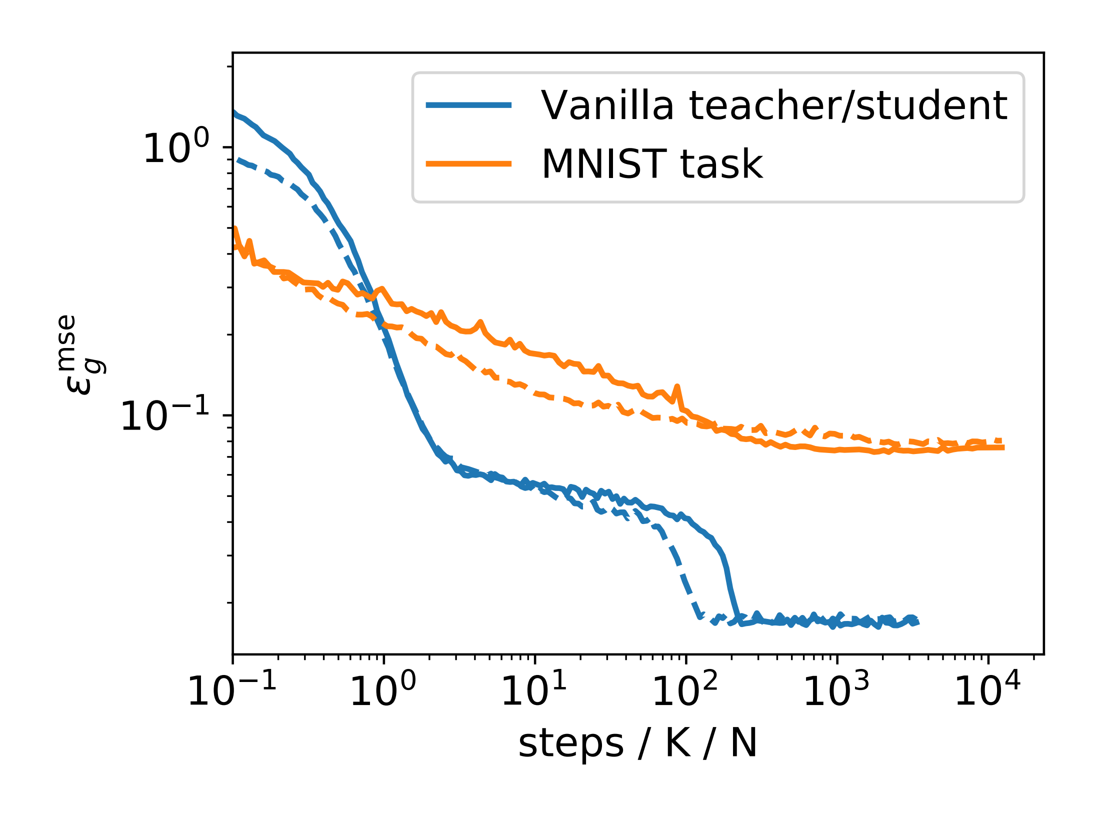
Two independent students do not learn the same function!

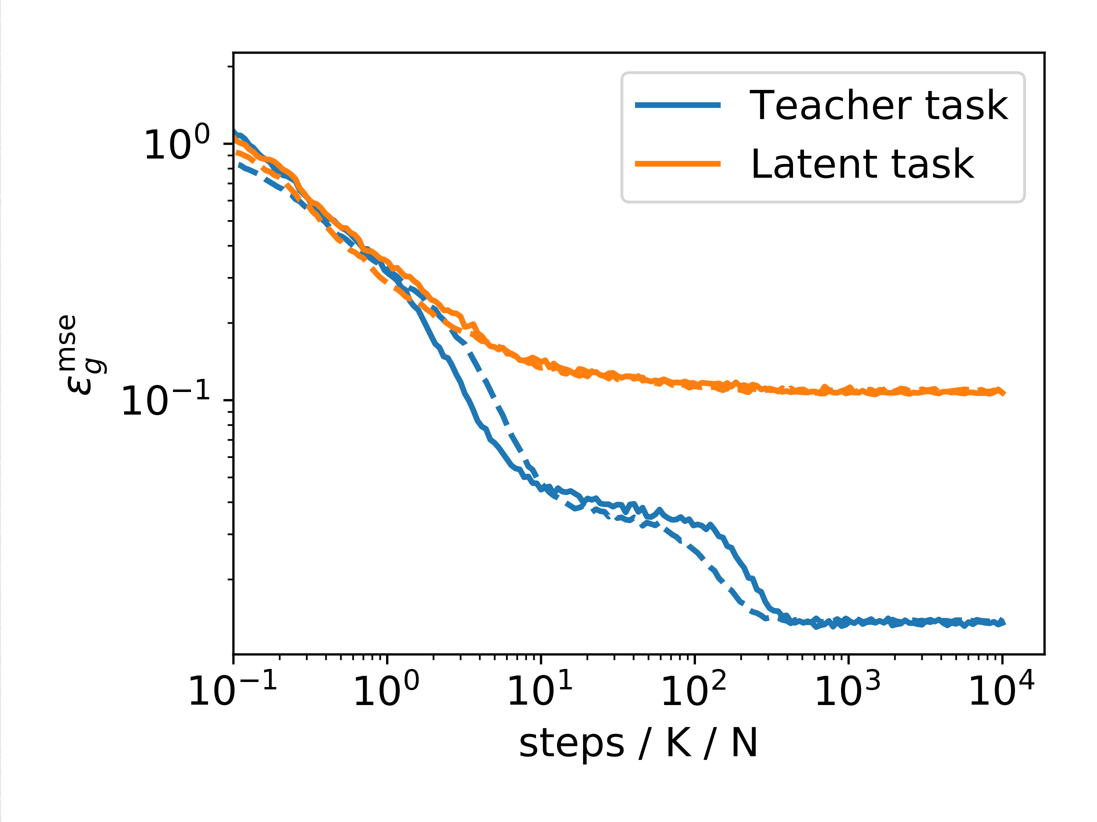


MNIST VS HIDDEN MANIFOLD

Teacher acting on X: Plateau in learning dynamics

MNIST & hidden manifold: No plateau ...





CONCLUSION ON HIDDEN MANIFOLD

- The hidden manifold model reproduces/captures behaviour of learning-dynamics on MNIST.
- Both (i) low-dimensional structure of input, and (ii) labels depending on the latent representation are needed.
- TODO: Solve analytically.
- TODO: Generalize to be able to demonstrate the advantage of depth.

TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed

hidden manifold data

iid inputs, teacher outputs

data structure

multiple wide layers

only few hidden units

architecture

algorithm

message passing

gradient-descent-based

CONCLUSION

Physics has many useful tools applicable in high-dimensional inference and learning.

REFERENCES FOR THIS TALK

- Barbier, Krzakala, Macris, Miolane, LZ; *Optimal errors and phase transitions in high-dimensional generalized linear models;* COLT'18, PNAS'19, arXiv:1708.03395.
- Aubin, Maillard, Barbier, Macris, Krzakala, LZ; *The committee machine: Computational to statistical gaps in learning a two-layers neural network*, spotlight at NeurIPS'18, arXiv:1806.05451.
- Sarao, Biroli, Cammarota, Krzakala, Urbani, LZ; *Marvels and Pitfalls of the Langevin Algorithm in Noisy High-dimensional Inference*, arXiv:1812.09066.
- Sarao, Krzakala, Urbani, LZ; *Passed & Spurious: Descent Algorithms and Local Minima in Spiked Matrix-Tensor Models;* ICML'19, arXiv:1902.00139.
- Sarao, Biroli, Cammarota, Krzakala, Urbani, LZ; *Who is Afraid of Big Bad Minima? Analysis of Gradient-Flow in a Spiked Matrix-Tensor Model*; spotlight at NeurIPS'19, arXiv:1907.08226.
- Goldt, Krzakala, Mézard, LZ; *Modelling the influence of data structure on learning in neural networks;* arXiv:1909.11500.
- Of independent interest: *Machine learning and the physical sciences*; Carleo, Cirac, Cranmer, Daudet, Schuld, Tishby, Vogt-Maranto, LZ; arXiv:1903.10563