UNDERSTANDING MACHINE LEARNING VIA EXACTLY SOLVABLE STATISTICAL PHYSICS MODELS

Lenka Zdeborová
(IPhT, CEA Saclay, France)

Workshop IV: Using Physical Insights for Machine Learning,
18.-22. 11. 2019, IPAM UCLA

MOTIVATION

- Deep learning brought unprecedented empirical/engineering progress into machine learning.
- 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

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.

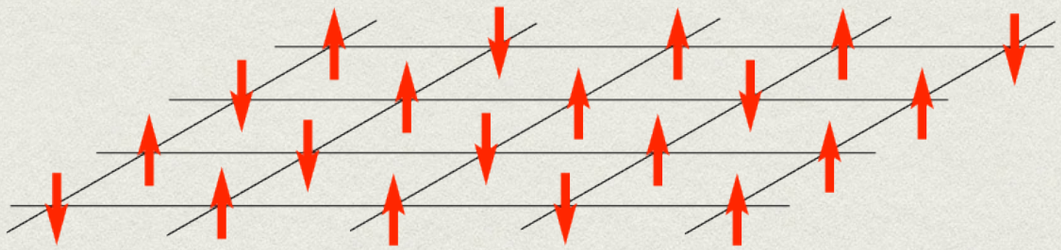
http://cargese.krzakala.org

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.

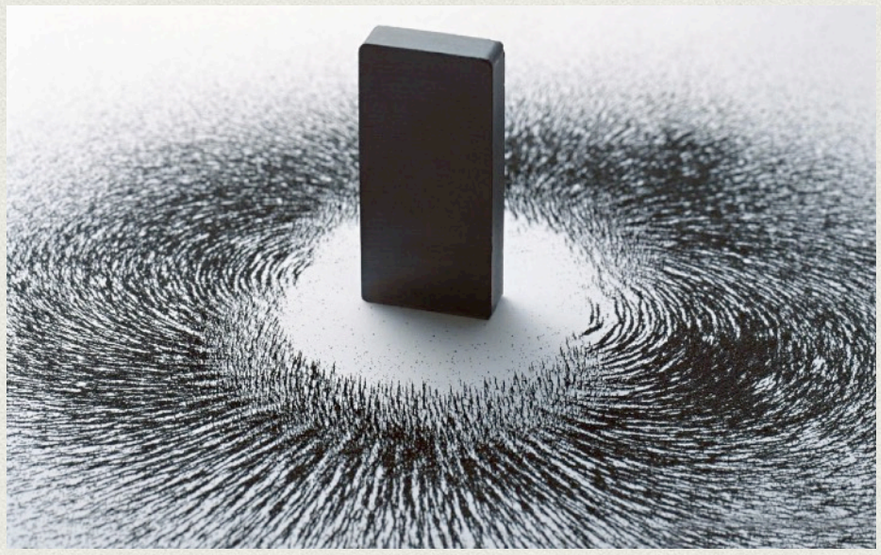


2–D Ising Model

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

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

magnetism of materials



WHAT TO MODEL IN DEEP LEARNING?

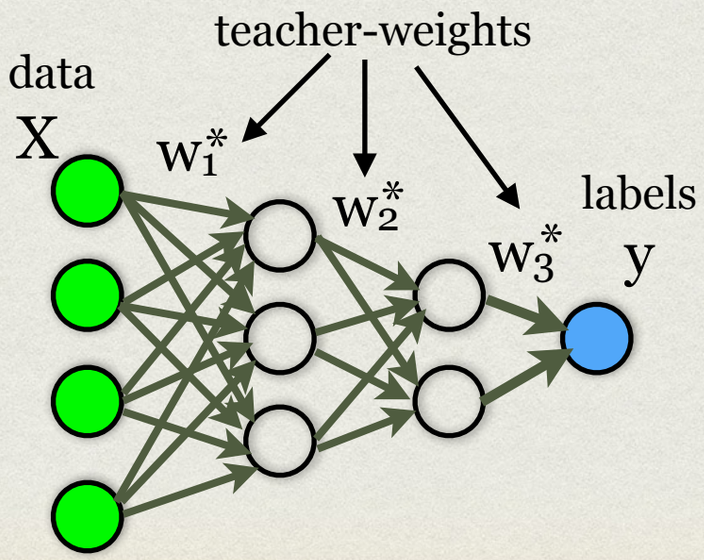
We aim to reproduce the salient behaviours of the real system.

Iterative process of improving the model.

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

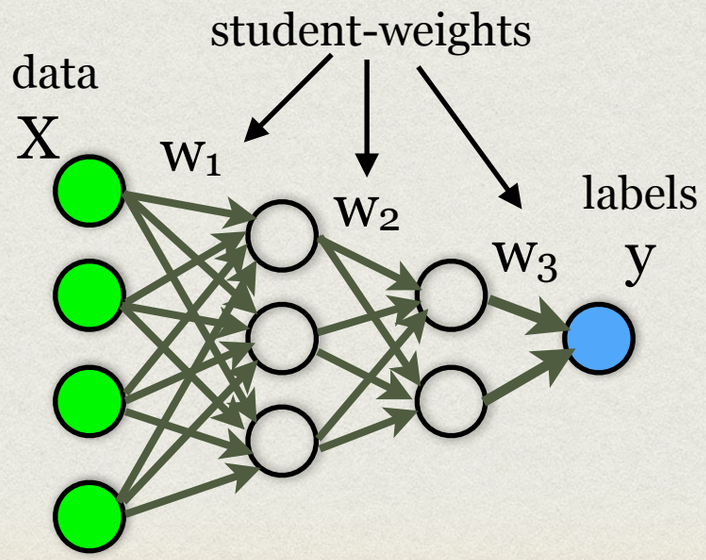
Teacher-network

- 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?



TEACHER-STUDENT PERCEPTRON

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

1989

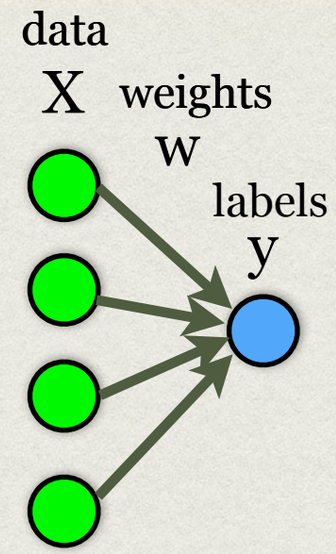
Three unfinished works on the optimal storage capacity of networks

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: $n \to \infty \quad p \to \infty$ p dimensions

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

Solved using the replica method in the high-dimensional limit

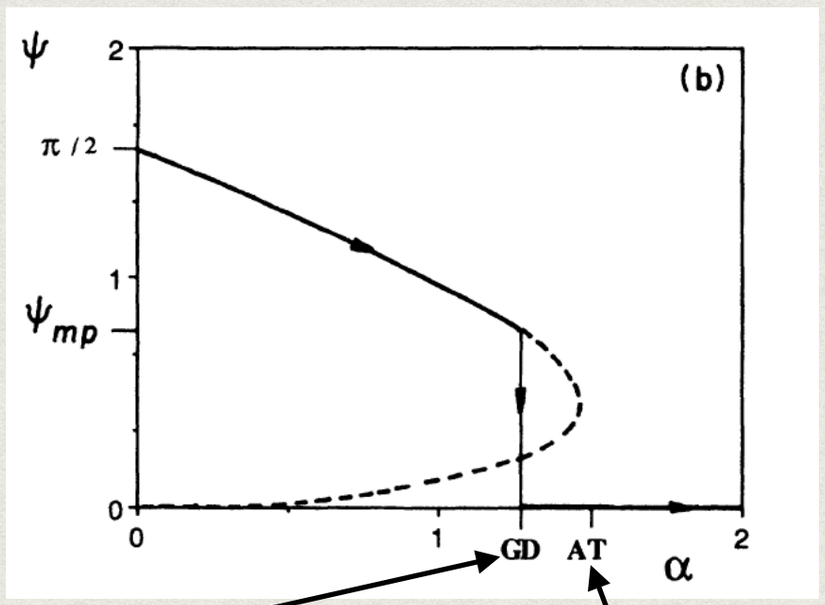
RAPID COMMUNICATIONS

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.



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

- Binary teacher-weights:

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

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

$$\alpha = n/p$$

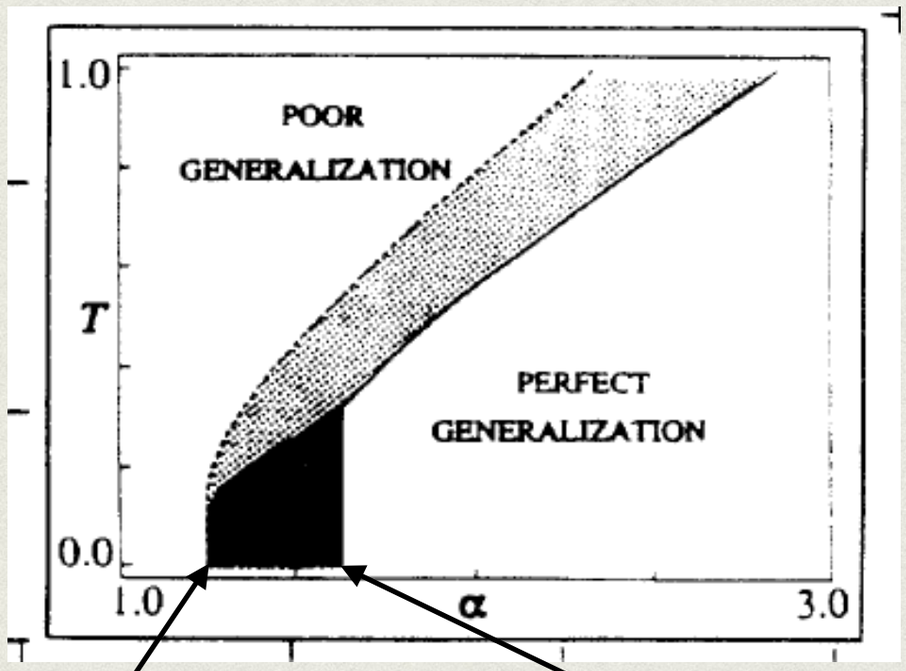
Learning from Examples in Large Neural Networks

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 $\alpha \rightarrow 1.24$. 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

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: $f = \lim_{p\to\infty} \frac{1}{p} \mathbb{E}_{y,X} \log Z(y, X)$ $\qquad \alpha = \frac{p}{n}$

Theorem 1:

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

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

where

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

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

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

CLOSED-FORM SOLUTION

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

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: 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 $\Sigma_i, R_i, g_{\text{out},\mu}$

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

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

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

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**.

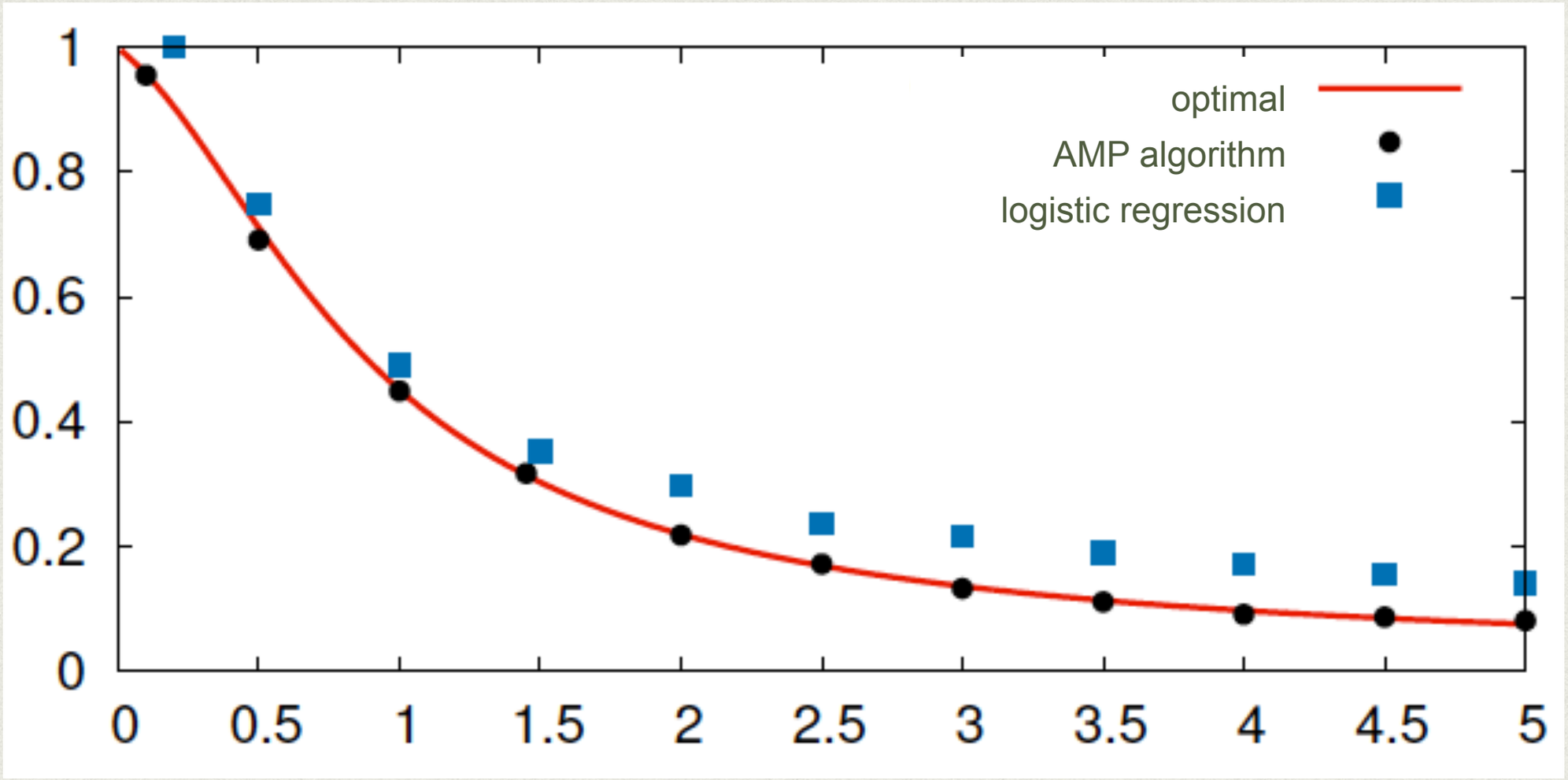
Simple to implement, only matrix multiplications, $O(N^2)$

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

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

SPHERICAL PERCEPTRON

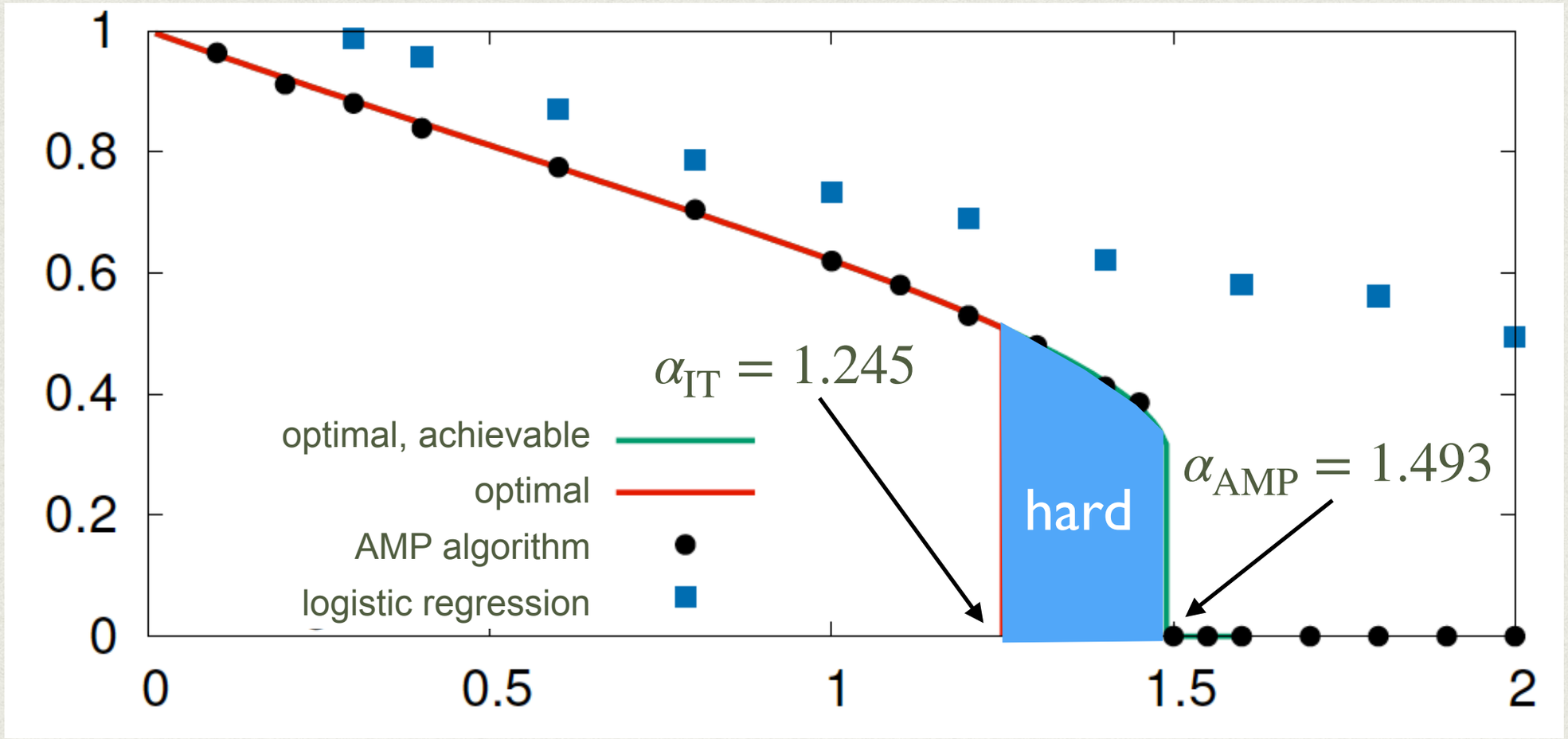
$\varphi(z) = \text{sign}(z)$ $\quad P_w = \mathcal{N}(0,1)$ $\quad n \to \infty,\ p \to \infty$ $\quad n/p = \Omega(1)$



BINARY PERCEPTRON

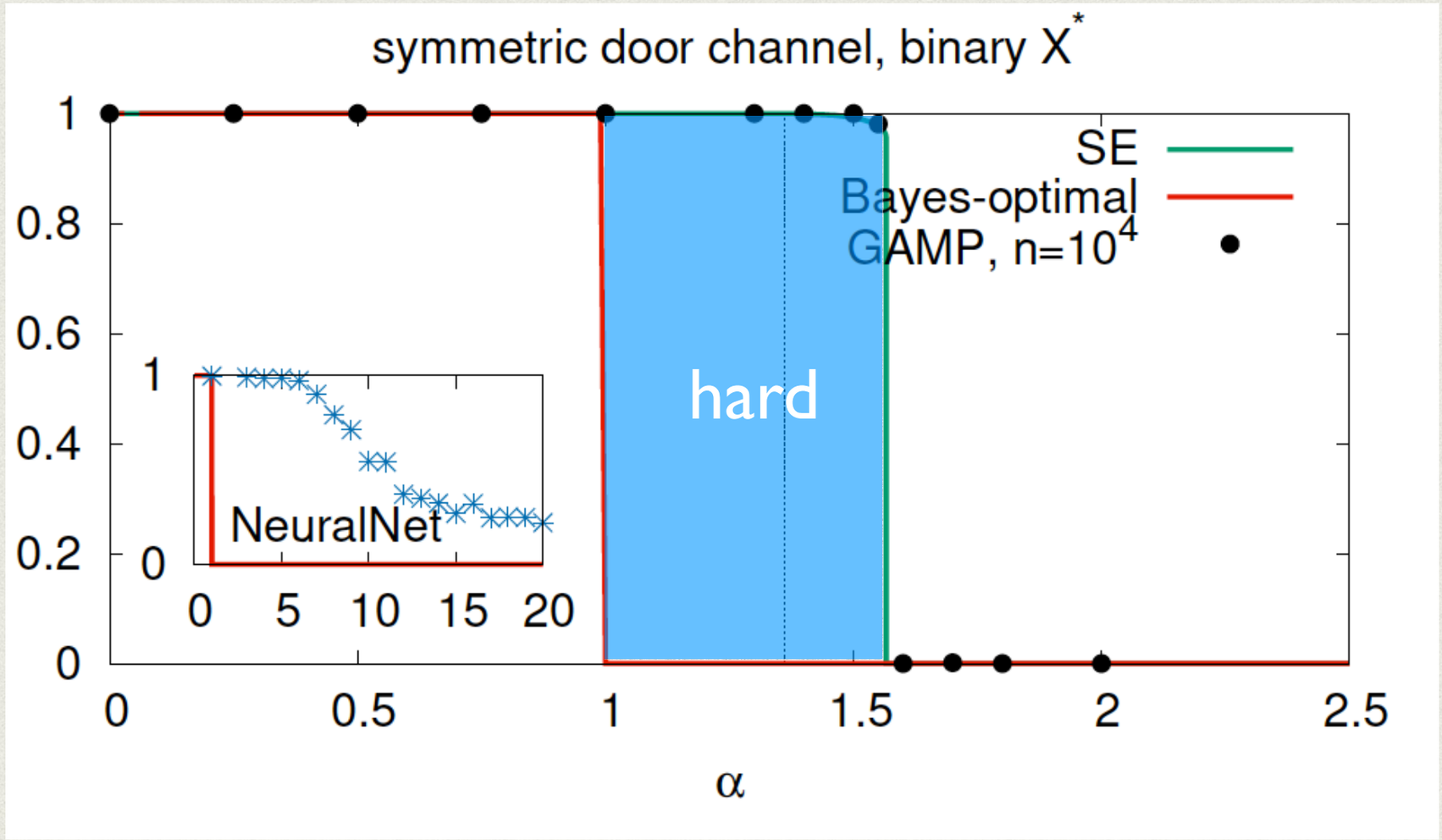
$$y_\mu = \mathrm{sign}\Big(\sum_{i=1}^{p} X_{\mu i} w_i\Big) \qquad w_i \in \{-1, +1\}$$

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



SYMMETRIC-DOOR PERCEPTRON

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



# of samples per dimension n/p

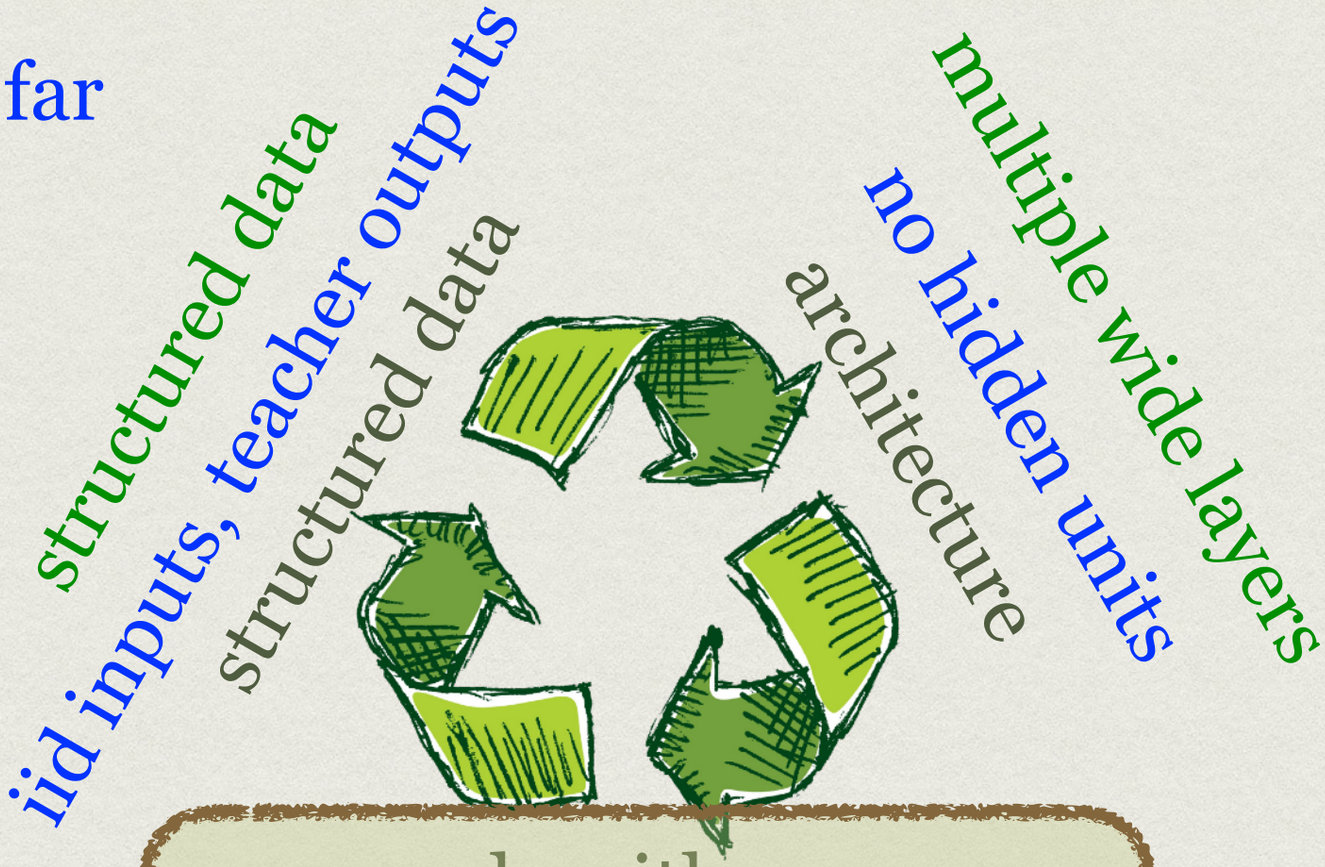
Is this bringing us towards the theory of deep learning?

TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed



algorithm

message passing

gradient-descent-based

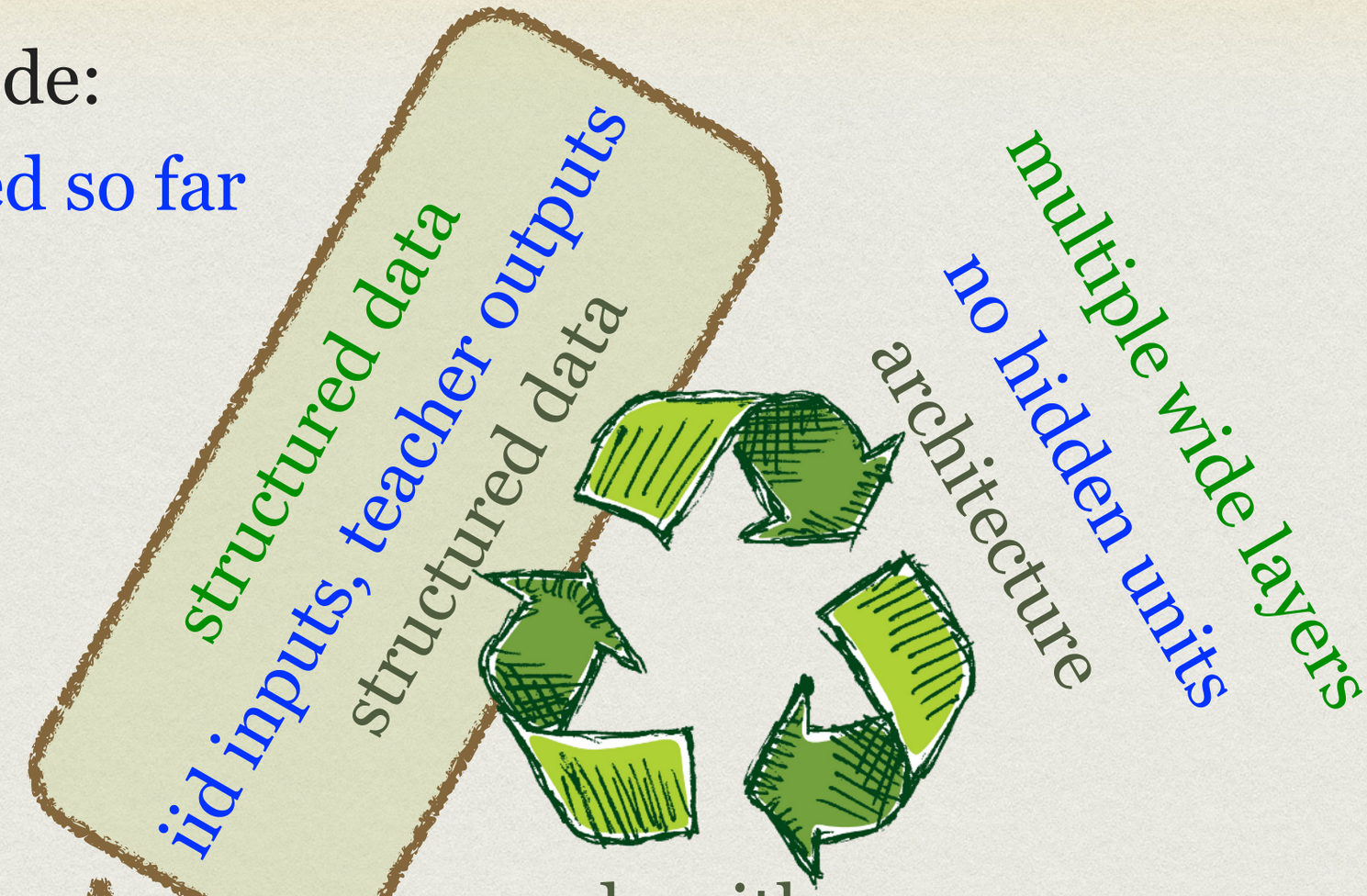
Gérard & Chiara tomorrow

TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed



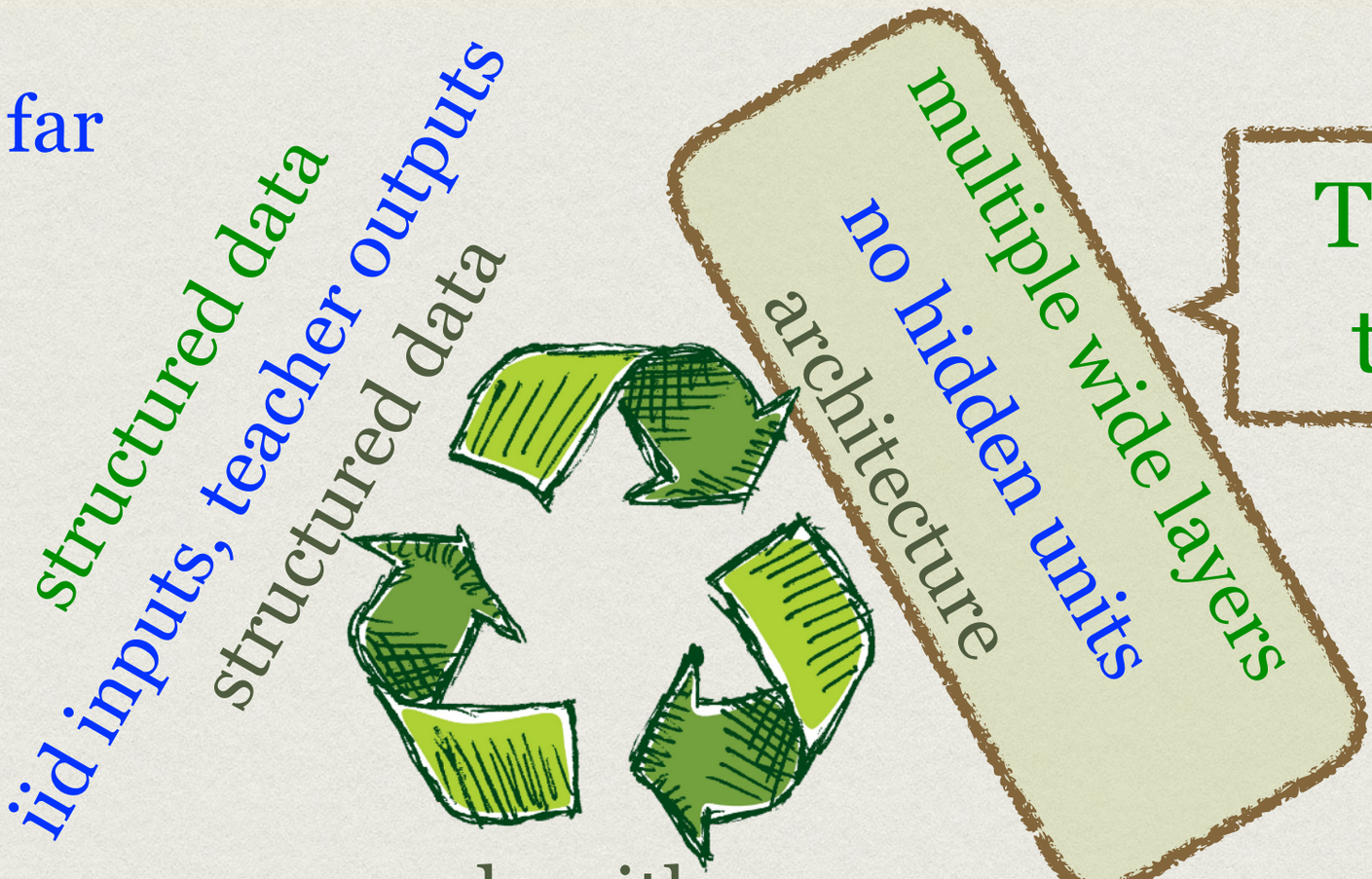
Marc
Thursday

TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed



The rest of this talk!

algorithm

message passing

gradient-descent-based

GOING MULTI-LAYER

Committee machine

Model from Schwarze'92.

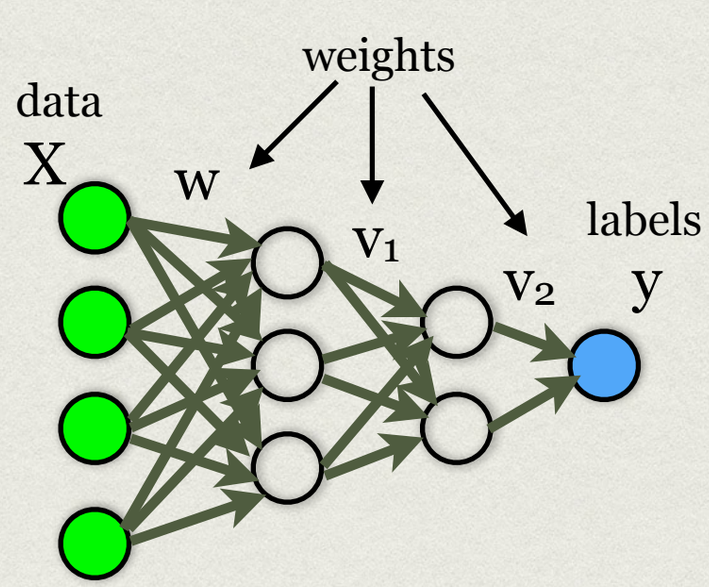
Proof of the replica formula, and approximate message passing Aubin, Maillard, Barbier, Macris, Krzakala, LZ, NeurIPS'18, arXiv:1806.05451.

p input units

K hidden units

output unit

n training samples



L=3 layers

w learned, v_1 & v_2 fixed

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

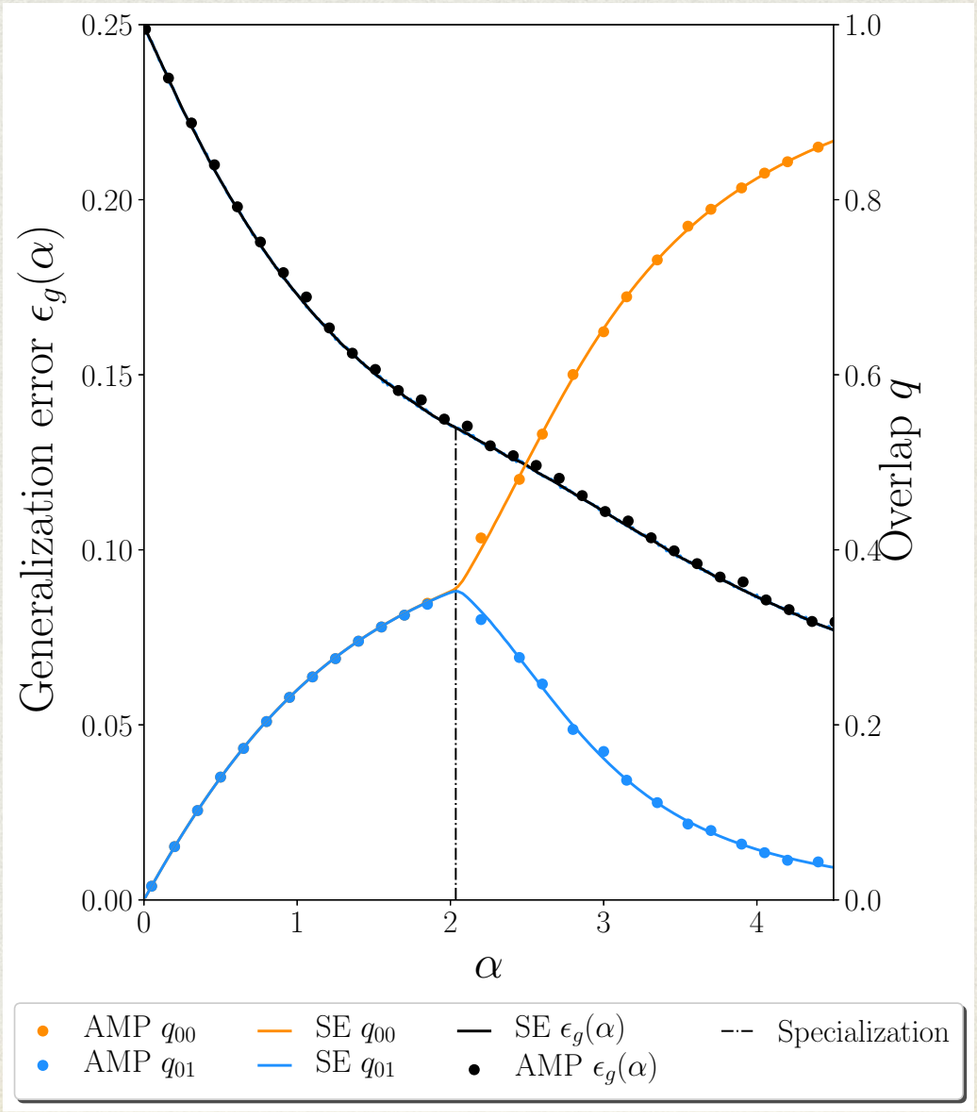
SPECIALISATION TRANSITION

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.



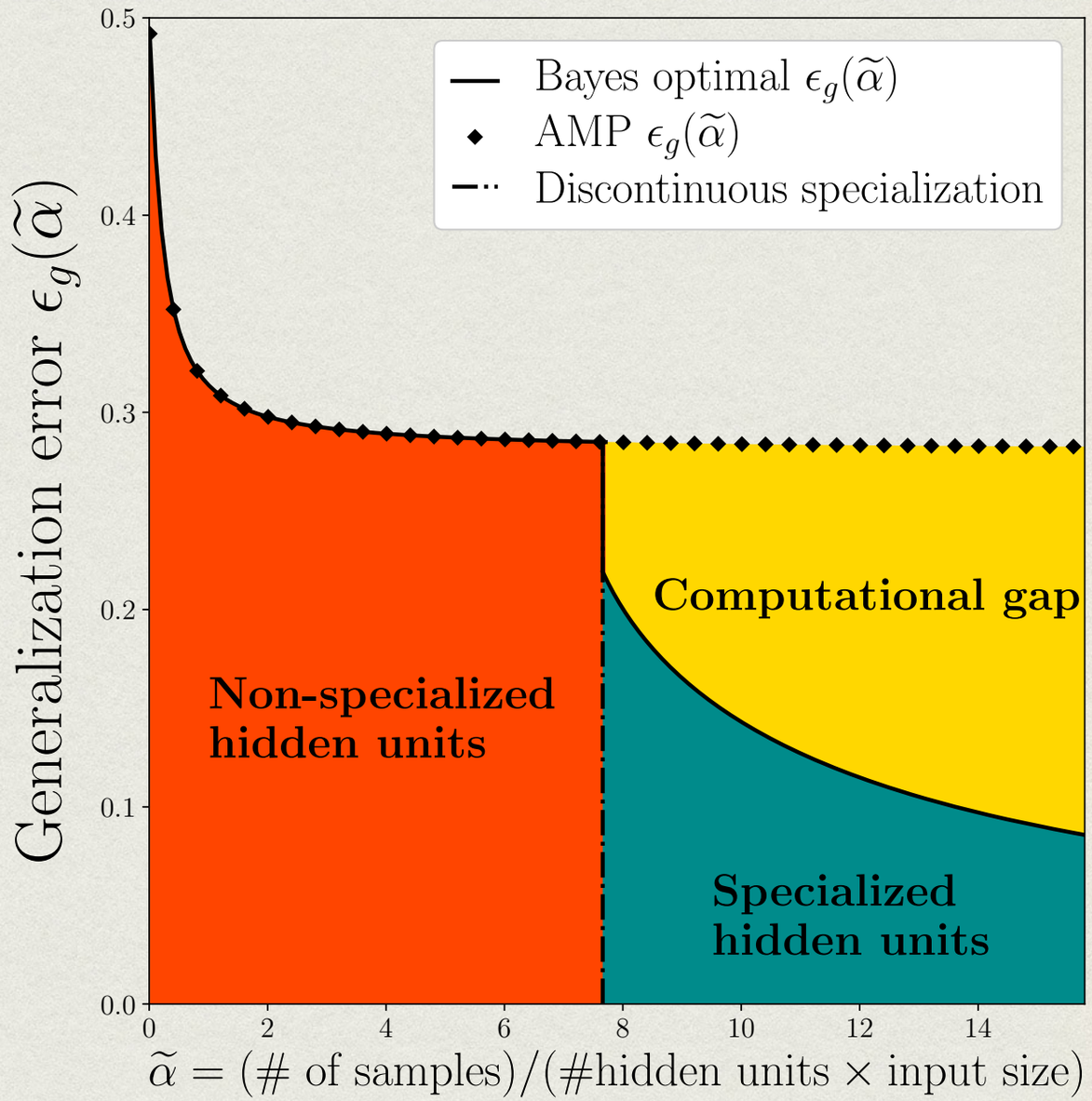
COMPUTATIONAL GAP

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

hidden units $K \gg 1$

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

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



SGD IN COMMITTEE MACHINE

- One pass SGD - i.e. putting the sample complexity aside.
- ODEs (Saad & Solla'94) for the SGD dynamics as $p \to \infty, K = \Omega(1)$
- Over-parametrised student network, i.e. wider than the teacher network.

➡ Formulas for the generalisation error as a function of which layer is learned, different activations, learning rate, label noise, initialisation, number of teacher and student hidden units.

Goldt, Advani, Saxe, Krzakala, LZ, NeurIPS'19 arXiv:1906.08632

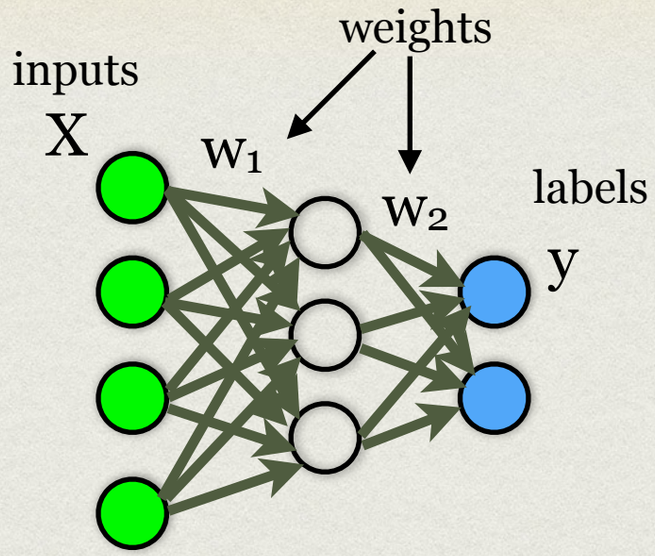
OPEN PROBLEM

p # input units

k # hidden units

m # output units

n training samples



2 layers
w_1 & w_2 learned

Limit: $n \to \infty$ $k \to \infty$ $n/p = \Omega(1)$
$p \to \infty$ $m \to \infty$ $k/p = \Omega(1)$
$m/p = \Omega(1)$

iid inputs X, iid teacher weights w_1^* and w_2^*, generate output y.

Open question: Optimal generalisation error of the student network?

No known (even heuristic) formula.

MULTI-LAYER GLM

$$y = \varphi^{(L)}(W^{(L)} \ldots \varphi^{(2)}(W^{(2)} \varphi^{(1)}(W^{(1)} x)))$$

- Compare: Neural network = learn $W^{(L)}, \ldots, W^{(2)}, W^{(1)}$, from n samples of (X, y). Analysis: Open problem.

- Def: Multi-layer GLM = known fixed $W^{(L)}, \ldots, W^{(2)}, W^{(1)}$, one sample of y, learn x. "Invert" the neural network.

  ➡ Replica formula for optimal solution. Multi-layer AMP algorithm (Manoel, Krzakala, Mezard, LZ, ISIT'17). Proof only for 2 layers (Gabrié, Luneau, Barbier, Macris, Krzakala, LZ, NeurIPS'18)

APPLICATIONS OF MULTI-LAYER GLM

- Scalable formula for mutual information between layers in learned neural network -> test the information bottleneck theory. (Gabrié, Luneau, Barbier, Macris, Krzakala, LZ, NeurIPS'18)

- Theory for estimation problems (e.g compressed sensing, sparse PCA) with generative neural-network prior. (Aubin, Loureiro, Maillard, Krzakala, LZ, NeurIPS'19)

RECOVER SIGNAL v FROM OBSERVATIONS y

$$y = \Gamma(v) \qquad v \in \mathbb{R}^p$$

observation — noisy transformation — signal

Simple examples

noise $\mathcal{N}(0,\Delta)$

- Denoising: $\Gamma(v) = v + \xi$
- Linear inverse problem: $\Gamma(v) = Av + \xi, \quad A \in \mathbb{R}^{n \times p}$
- Spiked matrix estimation: $\Gamma(v) = vv^T + \xi$

Basic paradigm of signal processing: Structure in v serves for recovery with better accuracy, larger noise, smaller n, etc.

GENERATIVE MODELS AS PRIORS

$$y = \Gamma(v)$$

Recover signal v from observations y, knowing that:

- Sparsity: v is k-sparse.
- A generative model learned from data: There exists $x \in \mathbb{R}^k$ such that

$$v = \varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}x))))$$

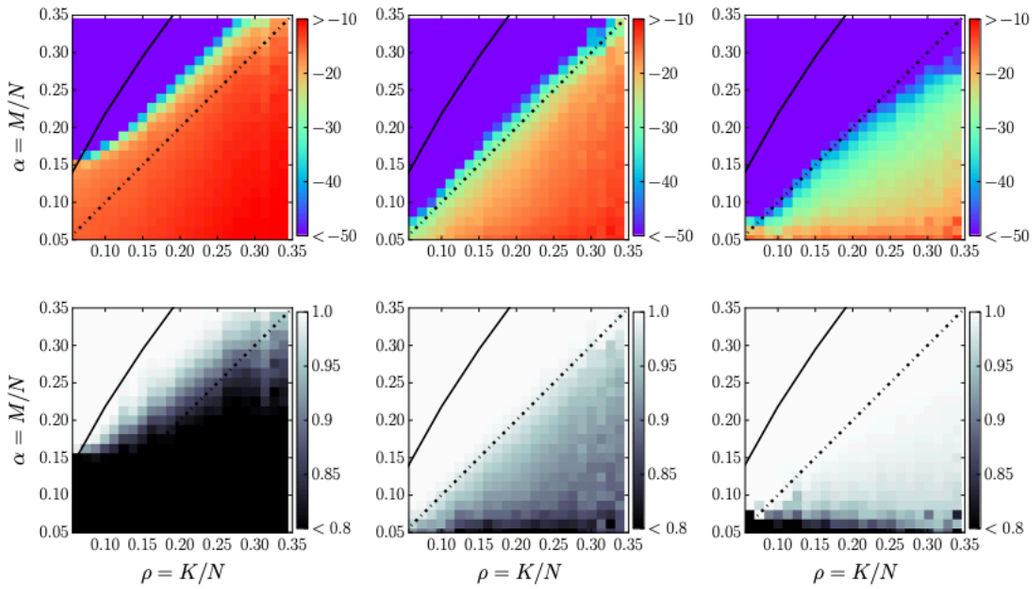
$$\varphi^{(i)}, W^{(i)}, i = 1,\ldots,L \quad \text{known, after training}$$

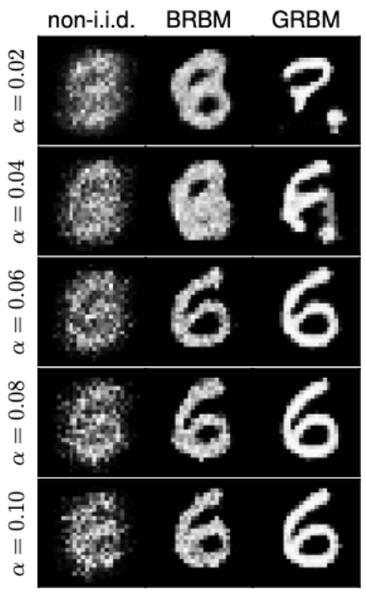
SELECTION OF EXISTING WORKS

Inferring Sparsity: Compressed Sensing using Generalized Restricted Boltzmann Machines

Eric W. Tramel[†], Andre Manoel[†], Francesco Caltagirone[‡], Marylou Gabrié[†] and Florent Krzakala[†§]
[†]Laboratoire de Physique Statistique (CNRS UMR-8550),
École Normale Supérieure, PSL Research University, 24 rue Lhomond, 75005 Paris, France
[§]Université Pierre et Marie Curie, Sorbonne Universités, 75005 Paris, France
[‡]INRIA Paris, 2 rue Simone Iff, 75012 Paris, France

arXiv:1606.03956





SELECTION OF EXISTING WORKS

Compressed Sensing using Generative Models

Ashish Bora* Ajil Jalal† Eric Price‡ Alexandros G. Dimakis§

arXiv:1703.03208

Abstract

The goal of compressed sensing is to estimate a vector from an underdetermined system of noisy linear measurements, by making use of prior knowledge on the structure of vectors in the relevant domain. For almost all results in this literature, the structure is represented by sparsity in a well-chosen basis. We show how to achieve guarantees similar to standard compressed sensing but without employing sparsity at all. Instead, we suppose that vectors lie near the range of a generative model $G : \mathbb{R}^k \to \mathbb{R}^n$. Our main theorem is that, if G is L-Lipschitz, then roughly $O(k \log L)$ random Gaussian measurements suffice for an ℓ_2/ℓ_2 recovery guarantee. We demonstrate our results using generative models from published variational autoencoder and generative adversarial networks. Our method can use 5-10x fewer measurements than Lasso for the same accuracy.

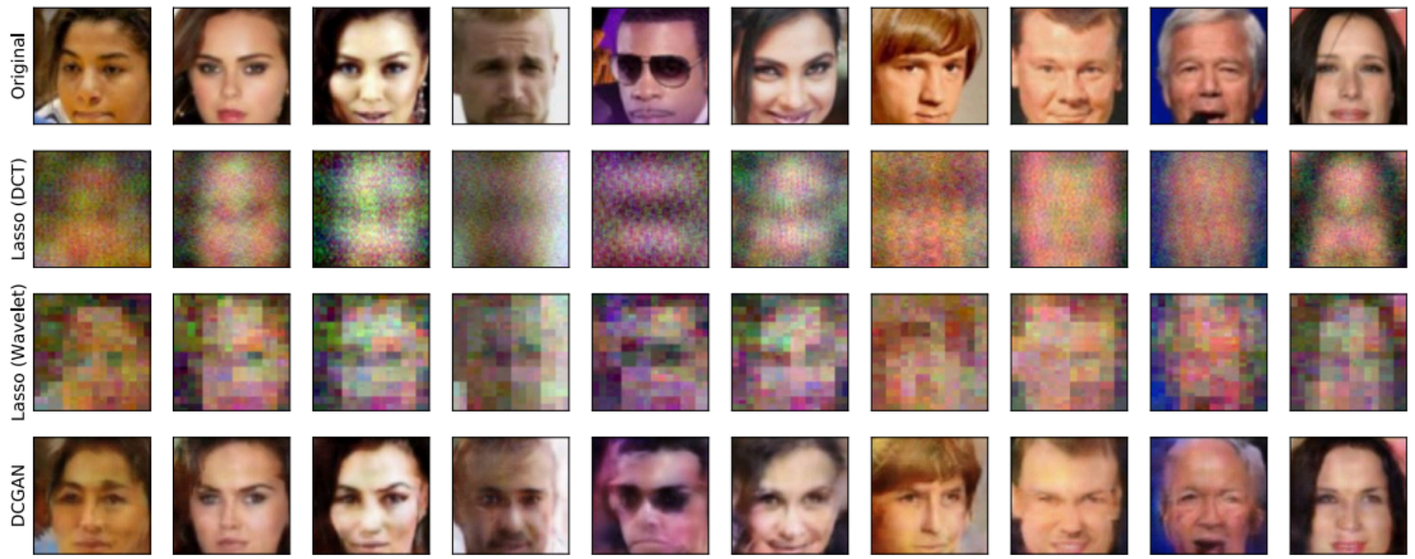
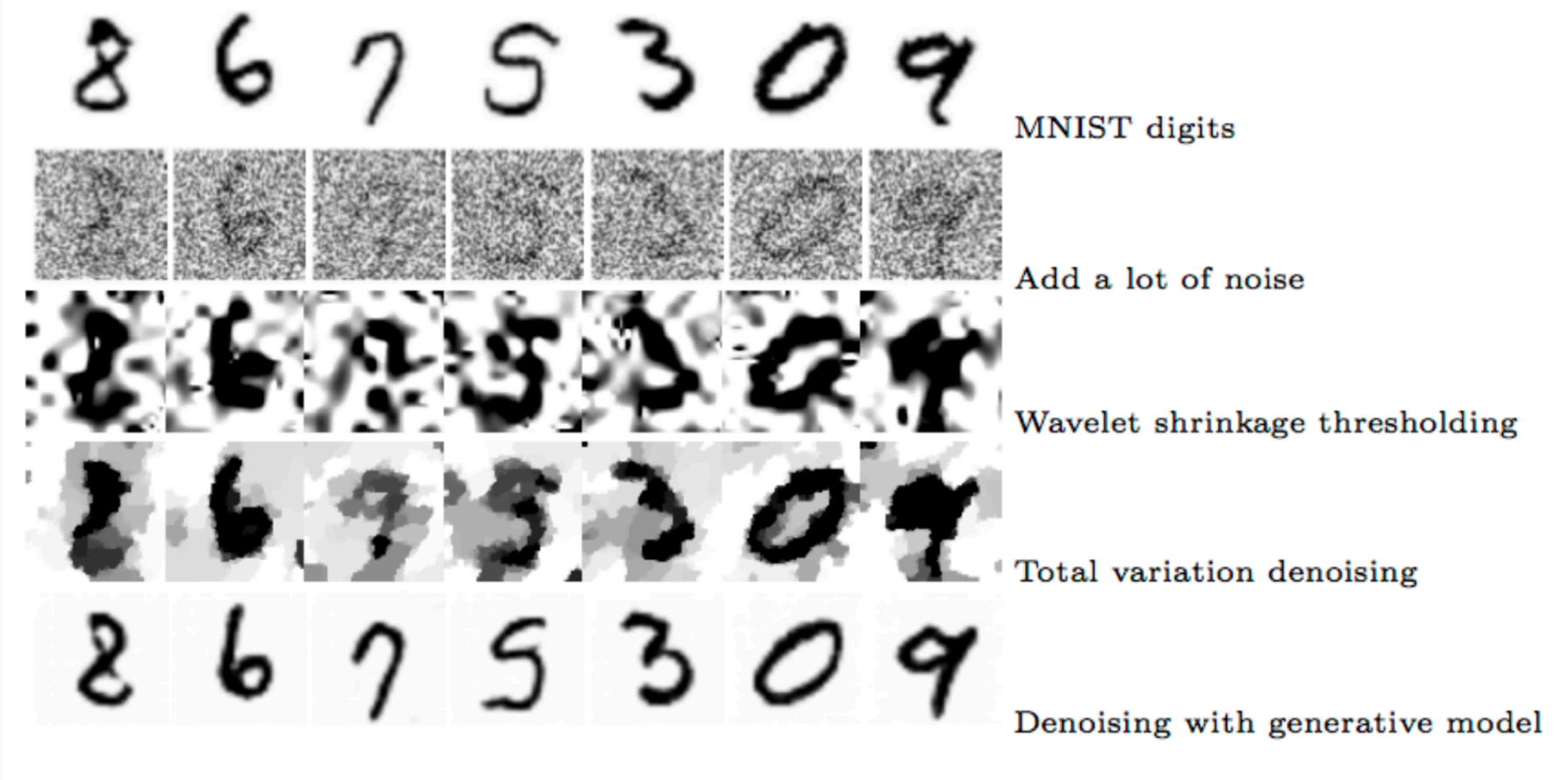


Figure 3: Reconstruction results on celebA with $m = 500$ measurements (of $n = 12288$ dimensional vector). We show original images (top row), and reconstructions by Lasso with DCT basis (second row), Lasso with wavelet basis (third row), and our algorithm (last row).

Generative models are the new sparsity?

Mar 28, 2018

Posted with : Data science, Data science, Data science



Blog by Soledad Villar, about arXiv:1803.09319
“SUNLayer: Stable denoising with generative networks”

FOCUS ON SPIKED MATRIX ESTIMATION

Simple examples

- Denoising: $\Gamma(v) = v + \xi$ — noise $\mathcal{N}(0,\Delta)$
- Compressed sensing: $\Gamma(v) = Av + \xi, \quad A \in \mathbb{R}^{n\times p}$
- Spiked matrix estimation: $\Gamma(v) = vv^T + \xi$

SPARSE PCA

$$Y = \frac{1}{\sqrt{p}} v^*(v^*)^T + \xi \qquad v^* \in \mathbb{R}^p \qquad \xi_{ij} \sim \mathcal{N}(0,\Delta)$$

$$P(v^*) = \prod_{i=1}^{p} \left[(1-\rho)\delta(v_i^*) + \rho\Phi(v_i^*) \right]$$

Result to recall:

(Rangan, Fletcher'12; Deshpande, Montanari'14; Lesieur, Krzakala, LZ'15-17, Krzakala, Xu, LZ'16; Barbier, Dia, Macris, Krzakala, Lesieur, LZ'16)

- At small $\rho = \Omega(1)$, large gap between information-theoretic and best-known-algorithmic performance.

SPIKED MATRIX MODEL WITH GENERATIVE PRIORS

Aubin, Loureiro, Maillard, Krzakala, LZ, aNeurIPS'19, arXiv:1905.12385

$$Y = \frac{1}{\sqrt{p}} v^*(v^*)^T + \xi \qquad v^* \in \mathbb{R}^p \qquad \xi_{ij} \sim \mathcal{N}(0,\Delta)$$

$$v^* = \varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}x^*)))) \qquad x^* \in \mathbb{R}^k$$

- Theory for $W^{(i)}, i = 1,\ldots,L$ with random iid components and independent from layer to layer.

BAYESIAN INFERENCE

$$P(v \mid Y) = \frac{1}{Z(Y,\Delta)} P_v(v) \prod_{i<j} e^{-\frac{1}{2\Delta}\left(Y_{ij} - v_i v_j/\sqrt{p}\right)^2}$$

Mutual information: $I(Y; v^*) = -\mathbb{E}_Y[\log Z(Y,\Delta)] + \frac{\rho_v p}{4\Delta} \qquad \rho_v \equiv \frac{1}{p}\mathbb{E}(v^T v)$

Main Theorem: $\lim_{p\to\infty} \frac{I(Y; v^\star)}{p} = \inf_{\rho_v \geq q_v \geq 0} i_{\mathrm{RS}}(\Delta, q_v)$

$$\mathrm{MMSE}_v = \rho_v - \operatorname{arginf} i_{\mathrm{RS}}(q_v)$$

where $i_{\mathrm{RS}}(\Delta, q_v) \equiv \frac{(\rho_v - q_v)^2}{4\Delta} + \frac{1}{p}\lim_{p\to\infty} I\left(v; v + \sqrt{\frac{\Delta}{q_v}}\xi\right)$

Proof: By Guerra interpolation from original to the denoising problem (Aubin, Loureiro, Maillard, Krzakala, LZ, arXiv:1905.12385).

PRIOR-MODEL DENOISING

$$v = \varphi^{(4)}(W^{(4)}\varphi^{(3)}(W^{(3)}\varphi^{(2)}(W^{(2)}\varphi^{(1)}(W^{(1)}x^*)))) + \frac{\Delta}{q_v}\xi$$

Multi-Layer Generalized Linear Estimation

Andre Manoel
Neurospin, CEA
Université Paris-Saclay

Florent Krzakala
LPS ENS, CNRS
PSL, UPMC & Sorbonne Univ.

Marc Mézard
Ecole Normale Supérieure
PSL Research University

Lenka Zdeborová
IPhT, CNRS, CEA
Université Paris-Saclay

ISIT'17

***Abstract*—We consider the problem of reconstructing a signal from multi-layered (possibly) non-linear measurements. Using non-rigorous but standard methods from statistical physics we present the Multi-Layer Approximate Message Passing (ML-AMP) algorithm for computing marginal probabilities of the corresponding estimation problem and derive the associated state evolution equations to analyze its performance. We also give the expression of the asymptotic free energy and the minimal information-theoretically achievable reconstruction error. Finally, we present some applications of this measurement model for compressed sensing and perceptron learning with structured matrices/patterns, and for a simple model of estimation of latent variables in an auto-encoder.**

components of each of these matrices are drawn independently at random, from a probability distribution $P_{W^{(\ell)}}$ having zero mean and variance $1/n_\ell$. We consider a signal $\boldsymbol{x} \in \mathbb{R}^{n_L}$ with elements x_i, $i = 1, \ldots, n_L$ sampled independently from a distribution $P_X(x_i)$. We then collect n_0 observations $\boldsymbol{y} \in \mathbb{R}^{n_0}$ of the signal $\boldsymbol{x}$ as

$$\boldsymbol{y} = f_{\xi^1}^{(1)}(W^{(1)} f_{\xi^2}^{(2)}(W^{(2)} \cdots f_{\xi^L}^{(L)}(W^{(L)}\boldsymbol{x}))), \qquad (1)$$

where the so-called *activation functions* $f_{\xi^\ell}^{(\ell)}$, $\ell = 1, \ldots, L$, are applied element-wise. These functions can be deterministic or stochastic and are, in general, non-linear. Assuming $f_\xi^{(\ell)}(z)$

PRIOR-MODEL DENOISING

Multi-Layer Generalized Linear Estimation

Andre Manoel
Neurospin, CEA
Université Paris-Saclay

Florent Krzakala
LPS ENS, CNRS
PSL, UPMC & Sorbonne Univ.

Marc Mézard
Ecole Normale Supérieure
PSL Research University

Lenka Zdeborová
IPhT, CNRS, CEA
Université Paris-Saclay

ISIT'17

***Abstract*—We consider the problem of reconstructing a signal from multi-layered (possibly) non-linear measurements. Using non-rigorous but standard methods from statistical physics we present the Multi-Layer Approximate Message Passing (ML-AMP) algorithm for computing marginal probabilities of the corresponding estimation problem and derive the associated state evolution equations to analyze its performance. We also give the expression of the asymptotic free energy and the minimal information-theoretically achievable reconstruction error. Finally, we present some applications of this measurement model for compressed sensing and perceptron learning with structured matrices/patterns, and for a simple model of estimation of latent variables in an auto-encoder.**

components of each of these matrices are drawn independently at random, from a probability distribution $P_{W^{(\ell)}}$ having zero mean and variance $1/n_\ell$. We consider a signal $\boldsymbol{x} \in \mathbb{R}^{n_L}$ with elements x_i, $i = 1, \ldots, n_L$ sampled independently from a distribution $P_X(x_i)$. We then collect n_0 observations $\boldsymbol{y} \in \mathbb{R}^{n_0}$ of the signal $\boldsymbol{x}$ as

$$\boldsymbol{y} = f^{(1)}_{\xi^1}(W^{(1)} f^{(2)}_{\xi^2}(W^{(2)} \cdots f^{(L)}_{\xi^L}(W^{(L)}\boldsymbol{x}))), \qquad (1)$$

where the so-called *activation functions* $f^{(\ell)}_{\xi^\ell}$, $\ell = 1, \ldots, L$, are applied element-wise. These functions can be deterministic or stochastic and are, in general, non-linear. Assuming $f^{(\ell)}_{\xi}(z)$

- Proof for single layer prior: Barbier, Krzakala, Macris, Miolane, Krzakala, LZ, COLT'18, PNAS'19
- Proof for two-layer prior: Gabrié, Manoel, Luneau, Macris, Krzakala, LZ, NeurIPS'18.

EXAMPLE OF A RESULT

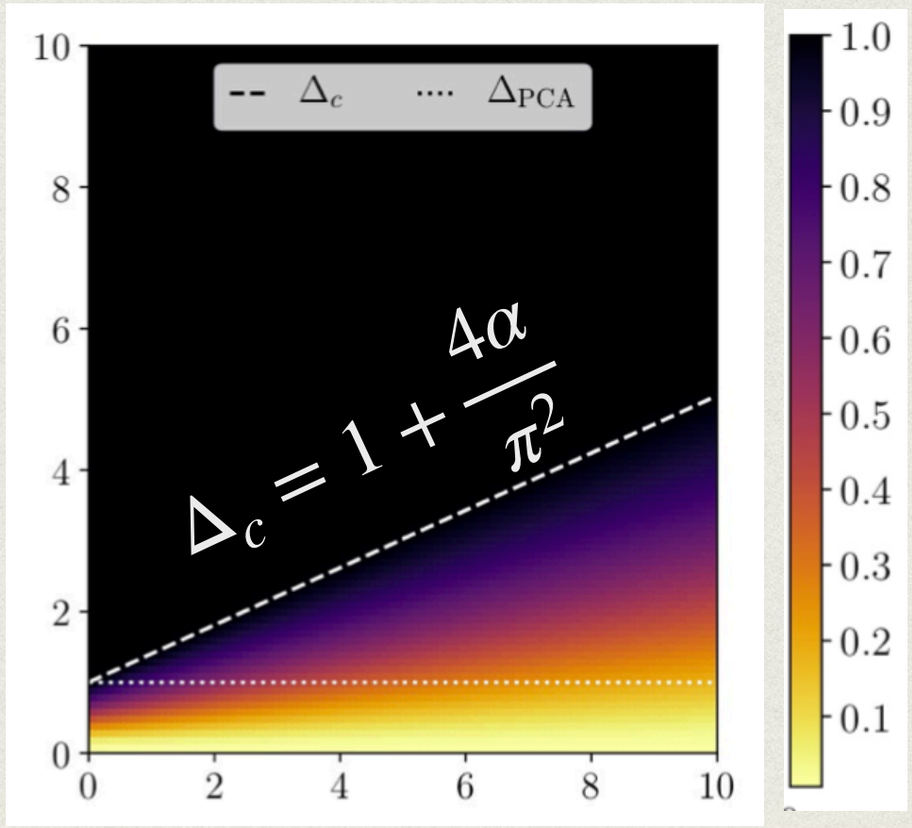
$$Y = \frac{1}{\sqrt{p}} v^*(v^*)^T + \xi$$

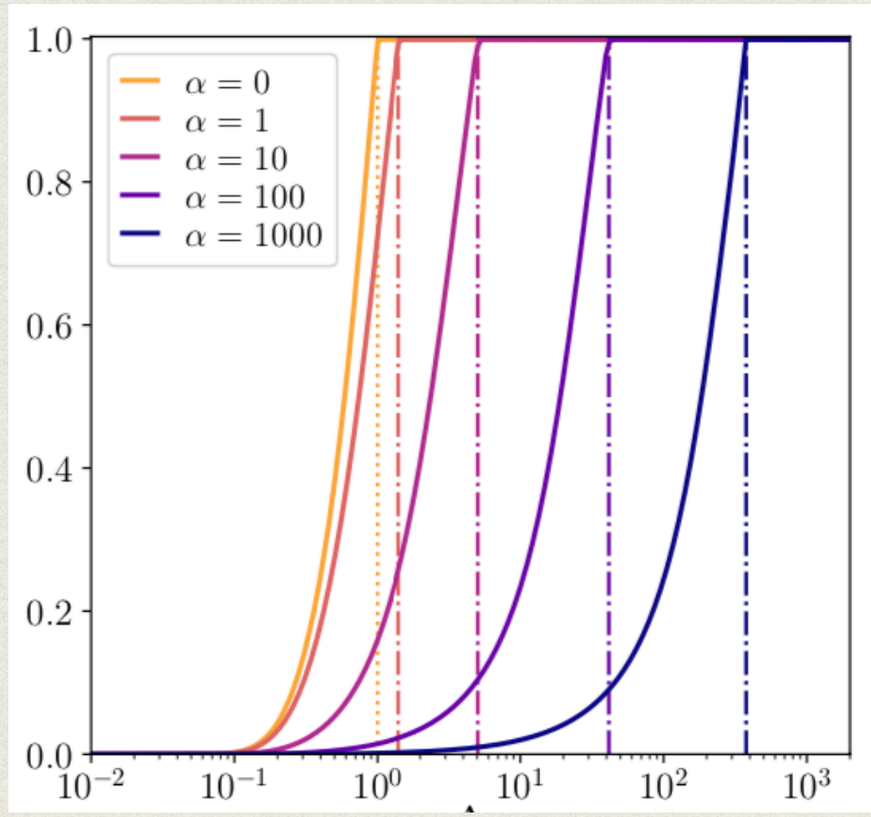
$$v^* = \mathrm{sign}(Wx^*)$$

$$v^* \in \mathbb{R}^p \qquad \xi_{ij} \sim \mathcal{N}(0,\Delta)$$

$$x^* \in \mathbb{R}^k \qquad x_i^* \sim \mathcal{N}(0,1)$$

$$W \in \mathbb{R}^{p\times k} \qquad W_{ij} \sim \mathcal{N}(0,1/p)$$





APPROXIMATE MESSAGE PASSING

Input: $Y \in \mathbb{R}^{p\times p}$ and $W \in \mathbb{R}^{p\times k}$:
Initialize to zero: $(\mathbf{g}, \hat{\mathbf{v}}, \mathbf{B}_v, A_v)^{t=0}$.
Initialize with: $\hat{\mathbf{v}}^{t=1} = \mathcal{N}(0, \sigma^2)$, $\hat{\mathbf{z}}^{t=1} = \mathcal{N}(0, \sigma^2)$, and $\hat{\mathbf{c}}_v^{t=1} = \mathbb{1}_p$, $\hat{\mathbf{c}}_z^{t=1} = \mathbb{1}_k$, $t = 1$.
repeat
Spiked layer:

$$\mathbf{B}_v^t = \frac{1}{\Delta}\frac{Y}{\sqrt{p}}\hat{\mathbf{v}}^t - \frac{1}{\Delta}\frac{\left(\mathbb{1}_p^\intercal \hat{\mathbf{c}}_v^t\right)}{p}\hat{\mathbf{v}}^{t-1} \quad \text{and} \quad A_v^t = \frac{1}{\Delta p}\|\hat{\mathbf{v}}^t\|_2^2 \mathrm{I}_p.$$

Generative layer:

$$V^t = \frac{1}{k}\left(\mathbb{1}_k^\intercal \hat{\mathbf{c}}_z^t\right)\mathrm{I}_p, \quad \boldsymbol{\omega}^t = \frac{1}{\sqrt{k}}W\hat{\mathbf{z}}^t - V^t\mathbf{g}^{t-1} \quad \text{and} \quad \mathbf{g}^t = f_{\text{out}}\left(\mathbf{B}_v^t, A_v^t, \boldsymbol{\omega}^t, V^t\right),$$

$$\Lambda^t = \frac{1}{k}\|\mathbf{g}^t\|_2^2 \mathrm{I}_k \quad \text{and} \quad \boldsymbol{\gamma}^t = \frac{1}{\sqrt{k}}W^\intercal \mathbf{g}^t + \Lambda^t\hat{\mathbf{z}}^t.$$

Update of the estimated marginals:

$$\hat{\mathbf{v}}^{t+1} = f_v(\mathbf{B}_v^t, A_v^t, \boldsymbol{\omega}^t, V^t) \qquad \text{and} \qquad \hat{\mathbf{c}}_v^{t+1} = \partial_B f_v(\mathbf{B}_v^t, A_v^t, \boldsymbol{\omega}^t, V^t),$$

$$\hat{\mathbf{z}}^{t+1} = f_z(\boldsymbol{\gamma}^t, \Lambda^t) \qquad \text{and} \qquad \hat{\mathbf{c}}_z^{t+1} = \partial_\gamma f_z(\boldsymbol{\gamma}^t, \Lambda^t),$$

$t = t + 1$.
until Convergence.
Output: $\hat{\mathbf{v}}, \hat{\mathbf{z}}$.

STATE EVOLUTION

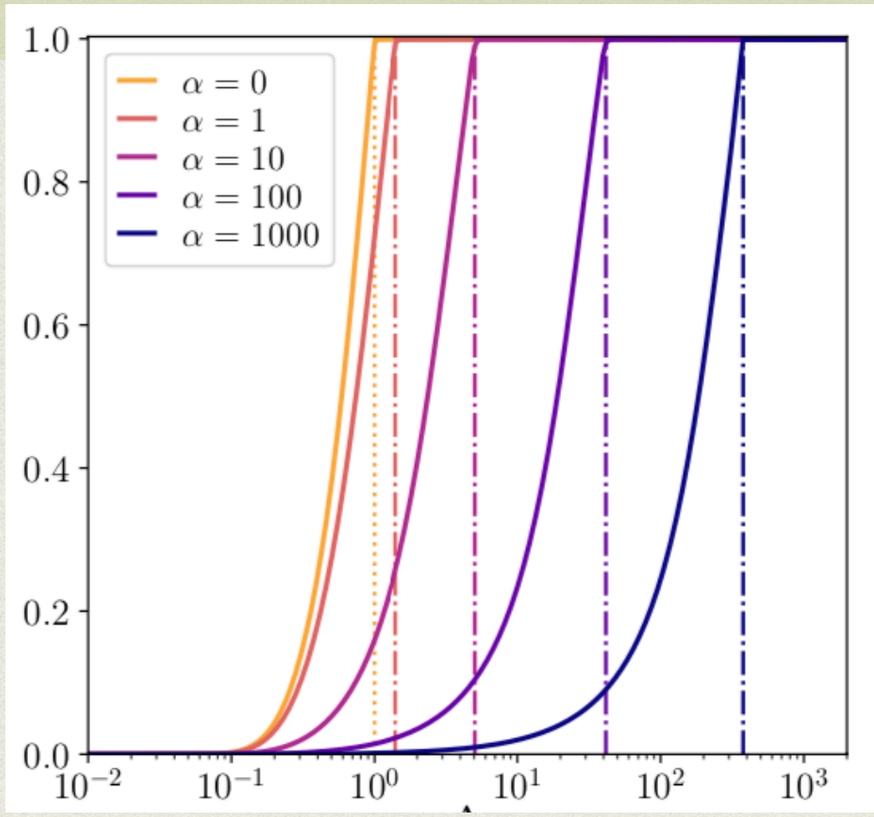
$$i_{\mathrm{RS}}(\Delta, q_v) \equiv \frac{(\rho_v - q_v)^2}{4\Delta} + \frac{1}{p} \lim_{p\to\infty} I\left(v; v + \sqrt{\frac{\Delta}{q_v}}\xi\right)$$

- As long as $i_{\mathrm{RS}}(q_v)$ has a unique minimiser, AMP matches the optimal performance as $p \to \infty$

$$Y = \frac{1}{\sqrt{p}} v^*(v^*)^T + \xi$$

$$v^* = \mathrm{sign}(Wx^*) \qquad \xi_{ij} \sim \mathcal{N}(0,\Delta)$$

$$W \in \mathbb{R}^{p\times k} \qquad \alpha = p/k$$



MESSAGE ABOUT GENERATIVE PRIORS

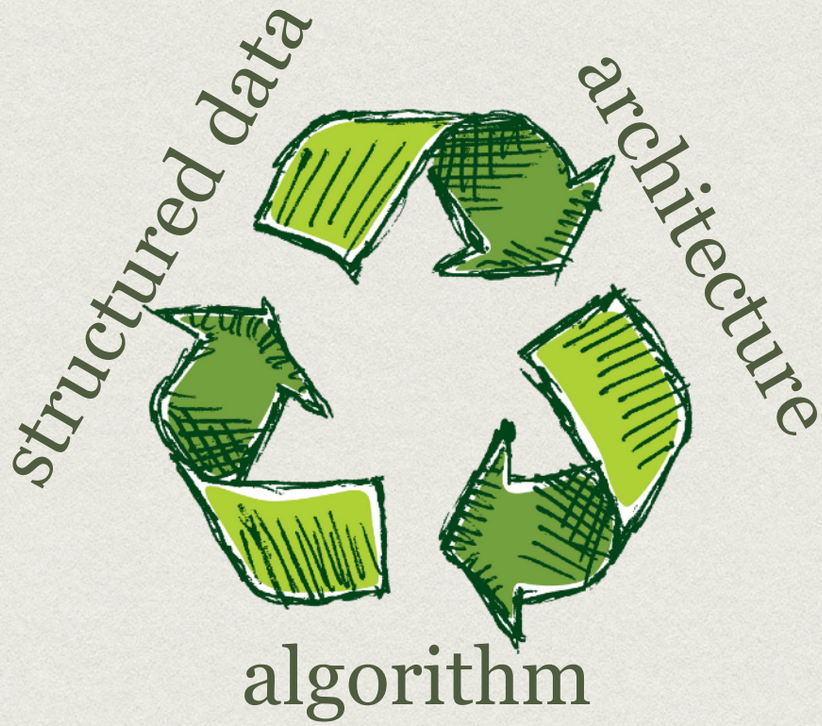
- Sparse prior: At small ρ, large gap between information-theoretic and best-known-algorithmic performance.

- Generative prior: No gap between information-theoretic and best-known-algorithmic performance.

Are generative priors better than sparsity?

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.
- Goldt, Advani, Saxe, Krzakala, LZ, *Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup*, NeurIPS'19 arXiv:1906.08632
- Manoel, Krzakala, Mezard, LZ, *Multi-layer generalized linear estimation*, ISIT'17, arXiv: 1701.06981.
- Gabrié, Luneau, Barbier, Macris, Krzakala, LZ, *Entropy and mutual information in models of deep neural networks*, NeurIPS'18, arXiv:1805.09785
- Aubin, Loureiro, Maillard, Krzakala, LZ, *The spiked matrix model with generative priors*, NeurIPS'19, arXiv:1905.12385

- Of independent interest: *Machine learning and the physical sciences*; Carleo, Cirac, Cranmer, Daudet, Schuld, Tishby, Vogt-Maranto, LZ; to appear Reviews of Modern Physics, arXiv:1903.10563

BONUS

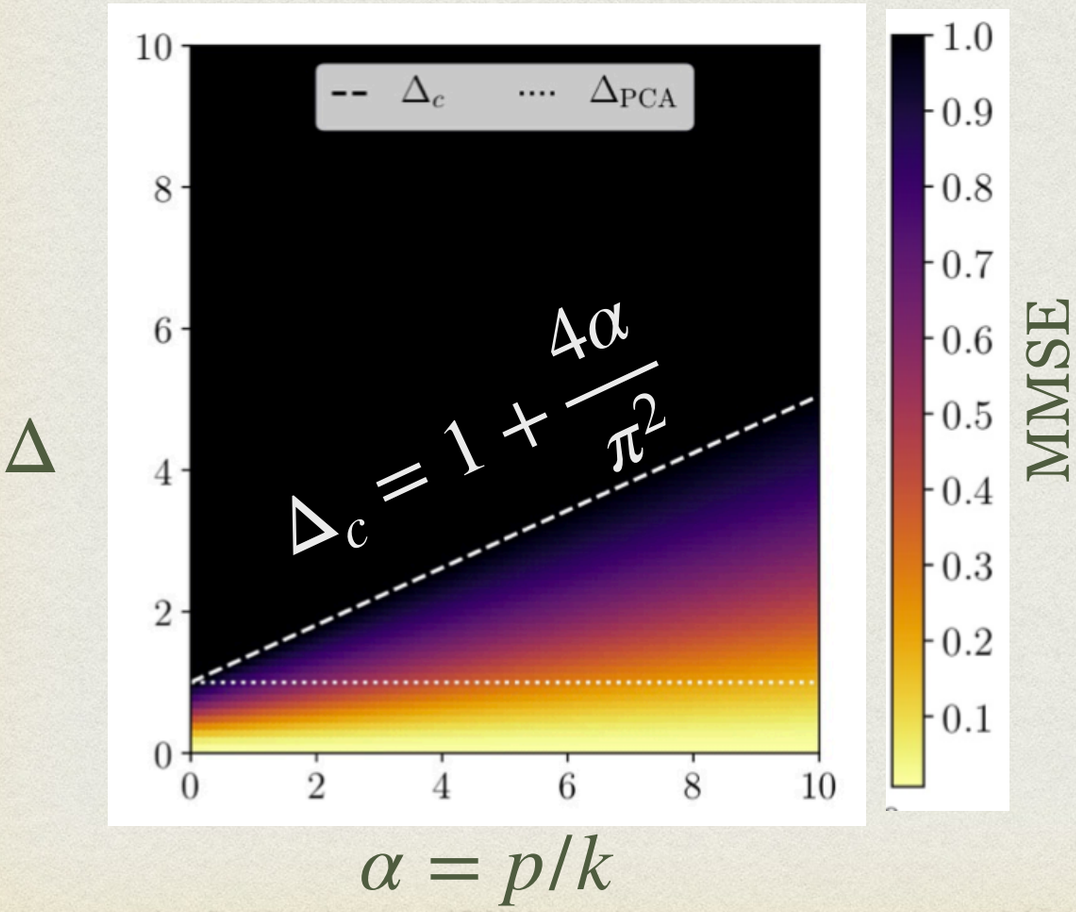
SPECTRAL ALGORITHMS

$$Y = \frac{1}{\sqrt{p}} v^*(v^*)^T + \xi$$

$$v^* = \text{sign}(Wx^*)$$

$$v^* \in \mathbb{R}^p \qquad x^* \in \mathbb{R}^k \qquad W \in \mathbb{R}^{p \times k}$$

$$\xi_{ij} \sim \mathcal{N}(0,\Delta) \qquad x_i^* \sim \mathcal{N}(0,1) \qquad W_{ij} \sim \mathcal{N}(0,1/p)$$

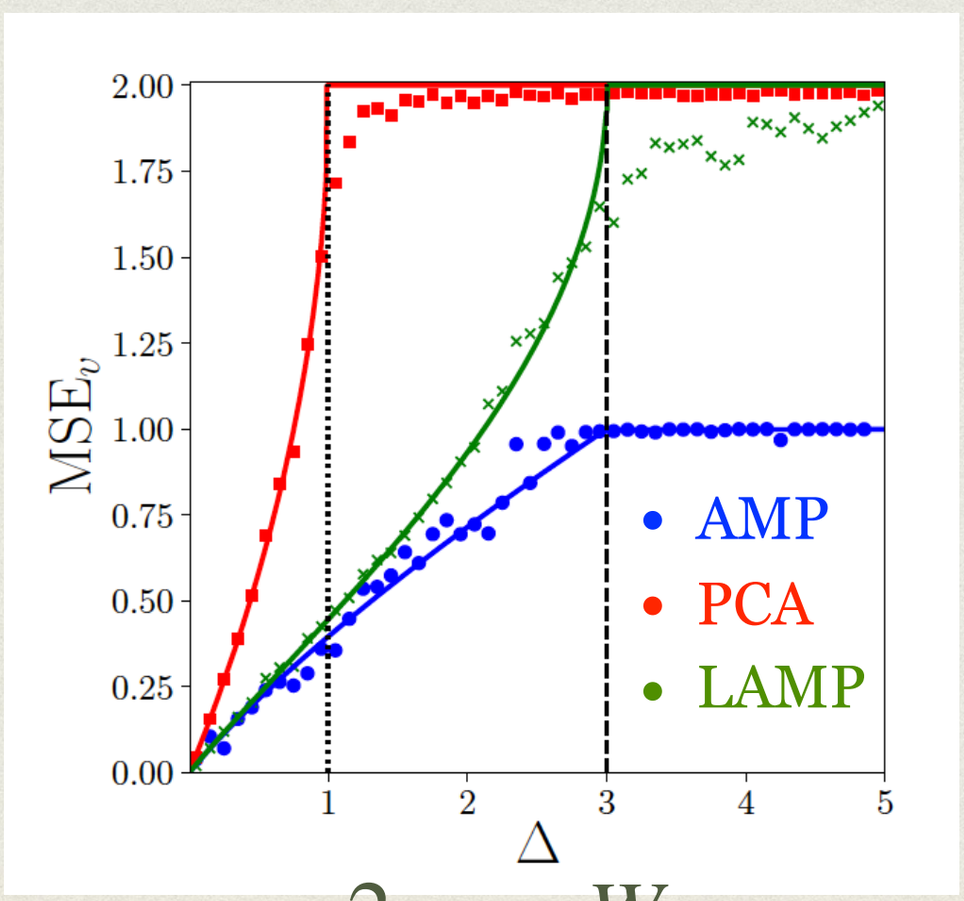


- AMP works for $\Delta < 1 + \frac{4\alpha}{\pi^2}$
- PCA works for $\Delta < 1$

Better spectral algorithms?

OPTIMAL AMONG SPECTRAL ALGORITHMS

- Strategy: Linearize approximate message passing or belief propagation (from Krzakala, Mossel, Moore, Neeman, Sly, LZ, Zhang, PNAS'13)
- Resulting conjecture: Optimal spectral algorithm LAMP uses



$\alpha = 2, v = Wx$

$$\Gamma = K_p \left[Y - \sqrt{p}\, \mathrm{I}_p \right]$$

$$K_p = \mathbb{E}(vv^T)$$

FOR RANDOM MATRIX THEORY LOVERS

$$Y = \frac{1}{\sqrt{p}} v^*(v^*)^T + \xi$$

$$v^* = Wx^*$$

$v^* \in \mathbb{R}^p$ $\quad$ $\xi_{ij} \sim \mathcal{N}(0,\Delta)$

$x^* \in \mathbb{R}^k$ $\quad$ $x_i^* \sim \mathcal{N}(0,1)$

$W \in \mathbb{R}^{p \times k}$ $\quad$ $W_{ij} \sim \mathcal{N}(0,1/p)$

$\alpha = p/k$

- Theorem: The leading eigenvector of Γ correlates with signal iff $\Delta < 1 + \alpha$

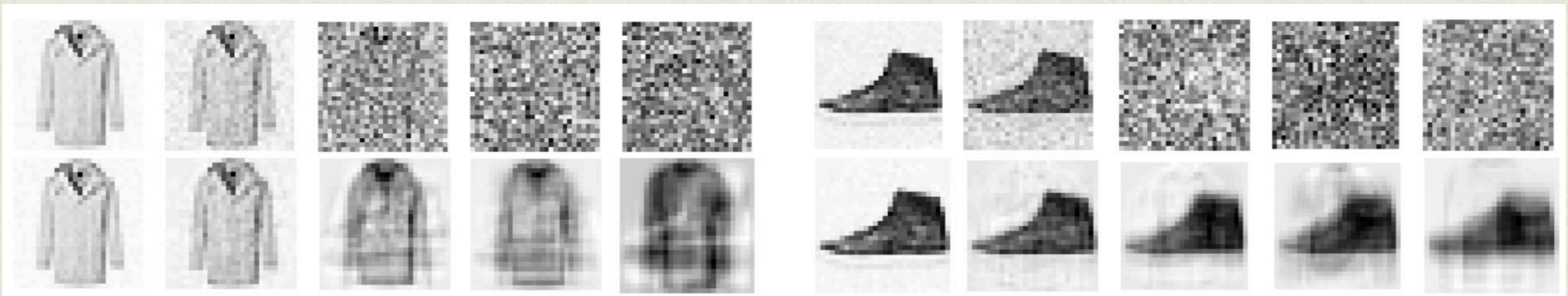
$$\Gamma = WW^T \left[Y - \sqrt{p}\, \mathrm{I}_p \right]$$

- Open problem for any other φ, with $v^* = \varphi(Wx^*)$

LAMP IMPROVES PCA WITHOUT TRAINING ON DATA

PCA (up) versus LAMP (bottom) on spiked matrix estimation

$\Delta = 0.01, 0.1, 1, 2, 10$



$$Y = \frac{1}{\sqrt{p}} v^*(v^*)^T + \xi \qquad \xi_{ij} \sim \mathcal{N}(0,\Delta)$$

LAMP: $\Gamma = K_p \left[Y - \sqrt{p}\, \mathrm{I}_p \right]$ K_p: empirical covariance

TAKE-HOME MESSAGE II

- Sparse prior: For $\rho = \Theta(1)$ no known algorithms with threshold better than PCA.

- Generative prior: spectral LAMP algorithm is better than PCA. Has the same threshold as AMP, conjectured optimal.

$$\Gamma = K_p \left[Y - \sqrt{p}\, \mathrm{I}_p \right]$$

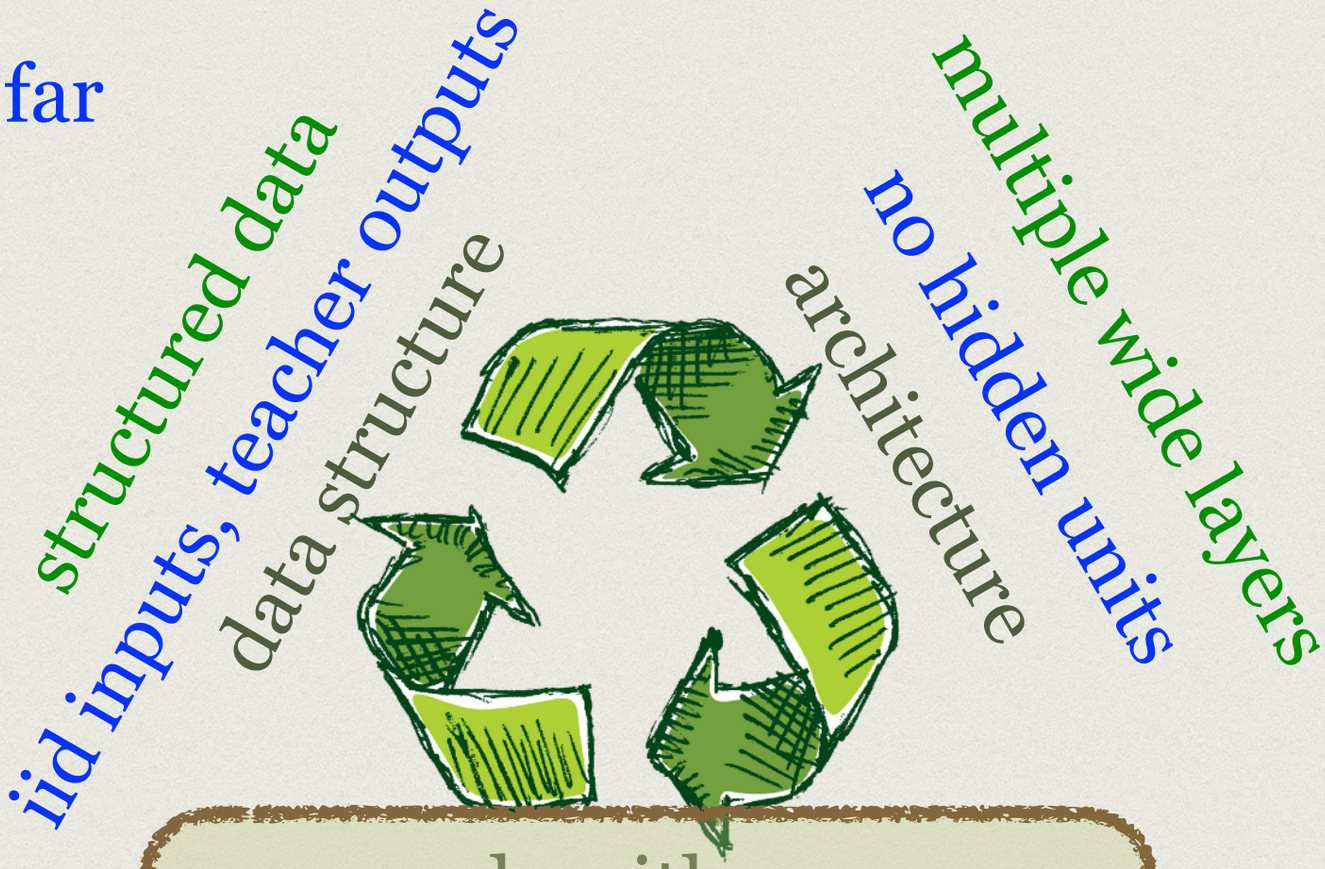
$$K_p = \mathbb{E}(vv^T)$$

TOWARDS THEORY OF DEEP LEARNING?

color-code:

described so far

needed



algorithm

message passing

gradient-descent-based

Chiara tomorrow

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!

Other progress recently: Linear networks Lazy training/NTK networks. Mean field limit. Implicit regularization in matrix factorizations.

GRADIENT-BASED ALGORITHMS

spherical constraint
(weight decay)

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

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

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

- High-dimensional. Non-convex loss.
- 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:

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

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

- Corresponding Hamiltonian (loss function, log-likelihood)

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

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

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.

OPTIMALITY AND AMP

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

- Approximate message passing (~TAP) algorithm provably matching the predicted performance. (Rangan, Fletcher'12, Matsushita, Tanaka'13, Deshpande, Montanari'14, Richard, Montanari'14 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; Miolane'17; Lesieur, Miolane, Lelarge, Krzakala, LZ'17)

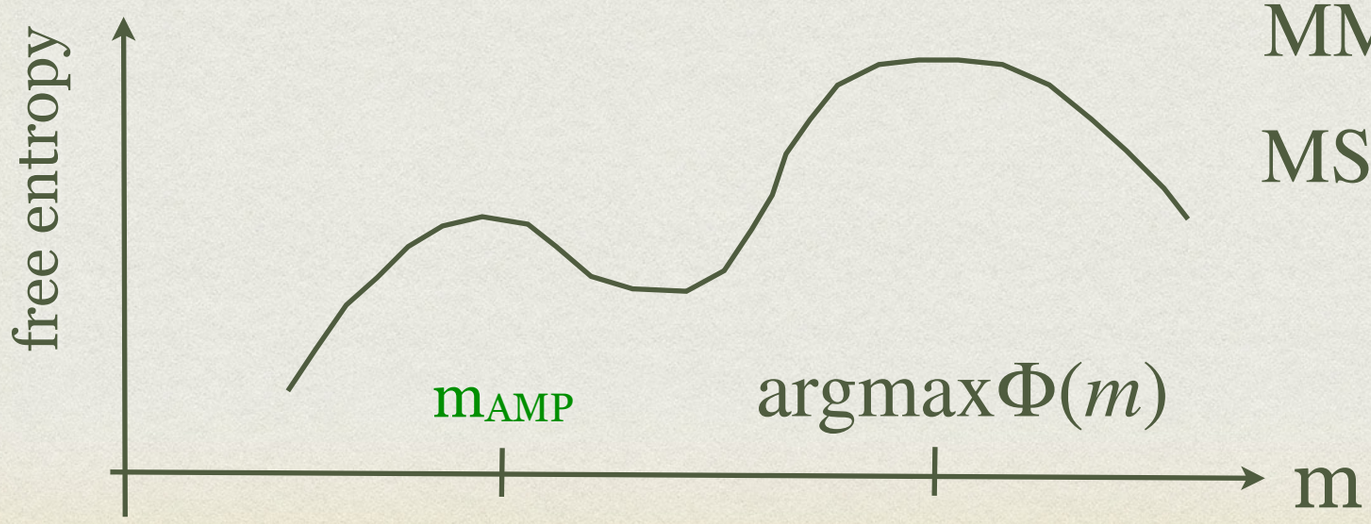
OPTIMALITY AND AMP

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

RS Free entropy: $$\Phi(m) = \frac{1}{2}\log(1-m) + \frac{m}{2} + \frac{m^2}{4\Delta_2} + \frac{m^p}{2p\Delta_p}$$

overlap with the planted configuration $m \in \mathbb{R}^+$

- Optimal MSE given by the global maximum of the free entropy.
- Mean-squared error (MSE) of AMP given by the local maximum of the free entropy, reached starting from small m/large MSE.



$$\text{MMSE} = 1 - \text{argmax}\,\Phi(m)$$

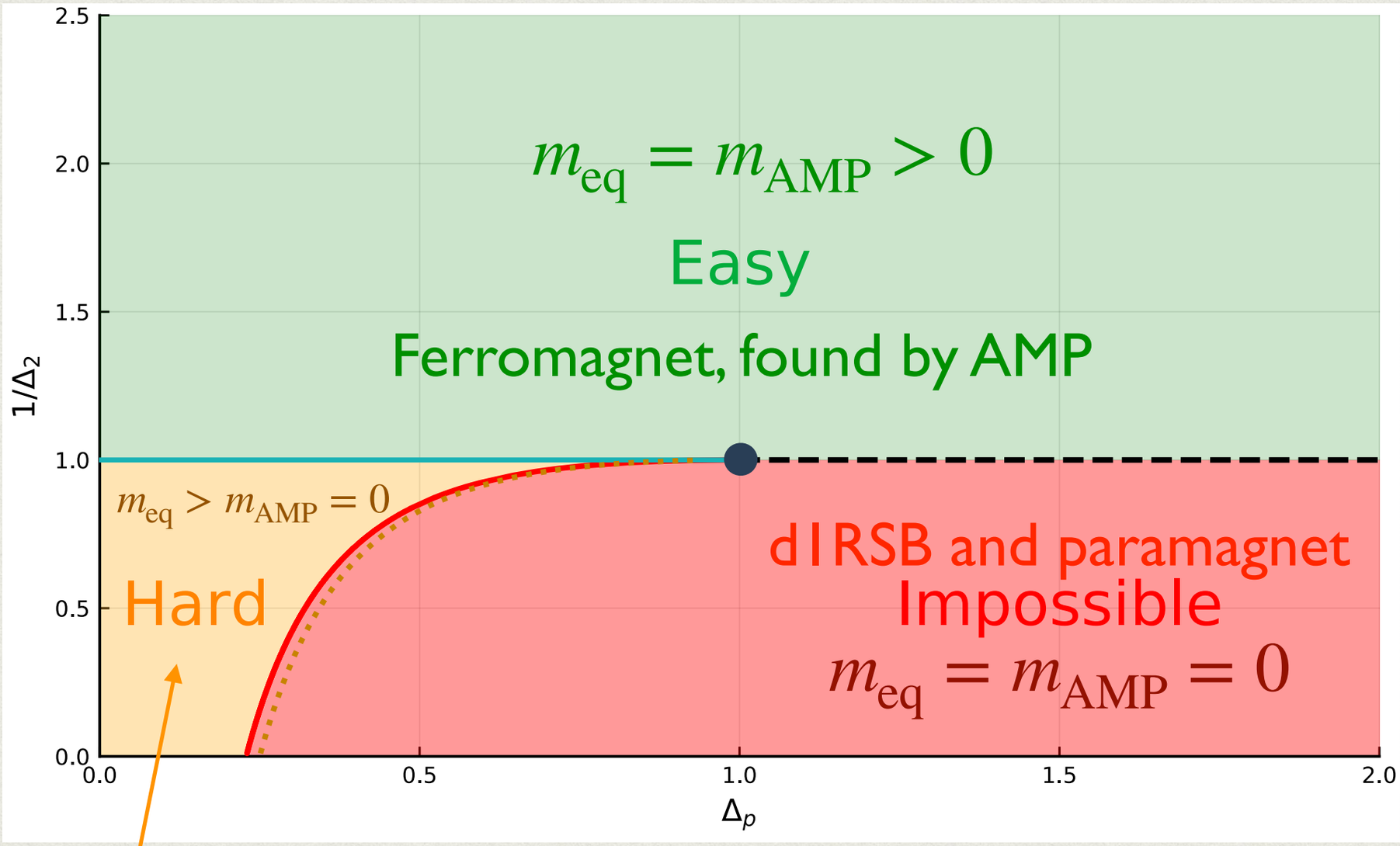
$$\text{MSE}_{\text{AMP}} = 1 - m_{\text{AMP}}$$

PHASE DIAGRAM

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

Bayes-optimal performance and AMP

p=3



Ferromagnet, not found by AMP

GRADIENT-BASED ALGORITHMS

spherical constraint
(weight decay)

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

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

gradient

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

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

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

$$N \to \infty$$

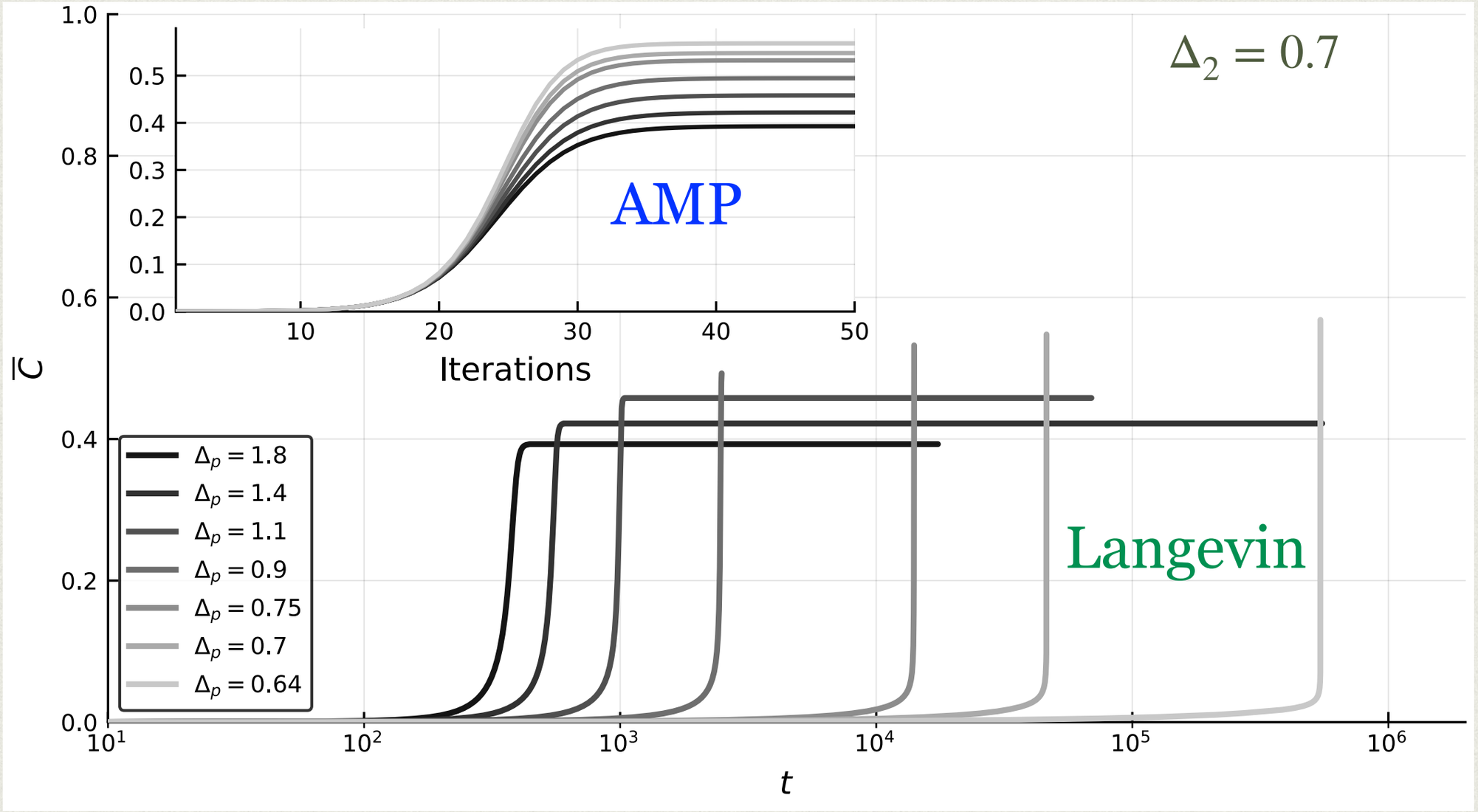
$$
\begin{aligned}
\frac{\partial}{\partial t}C(t,t') &= 2R(t',t) - \mu(t)C(t,t') + Q'(\overline{C}(t))\overline{C}(t') + \int_0^t dt'' R(t,t'')Q''(C(t,t''))C(t',t'') + \int_0^{t'} dt'' R(t',t'')Q'(C(t,t''))\,, \\
\frac{\partial}{\partial t}R(t,t') &= \delta(t-t') - \mu(t)R(t,t') + \int_{t'}^t dt'' R(t,t'')Q''(C(t,t''))R(t'',t')\,,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

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)

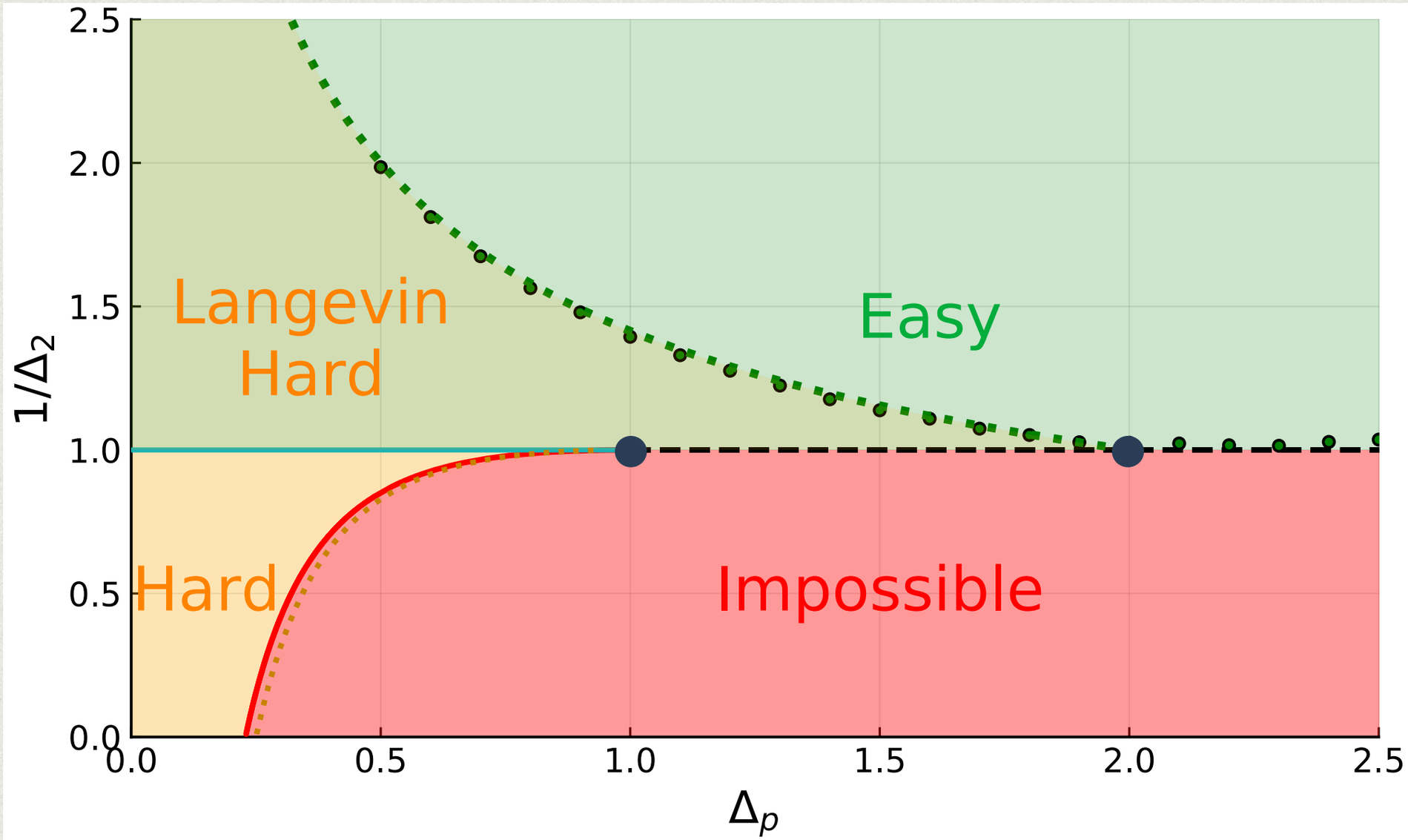
LANGEVIN STATE EVOLUTION (NUMERICAL SOLUTION)



github.com/sphinxteam/spiked_matrix-tensor

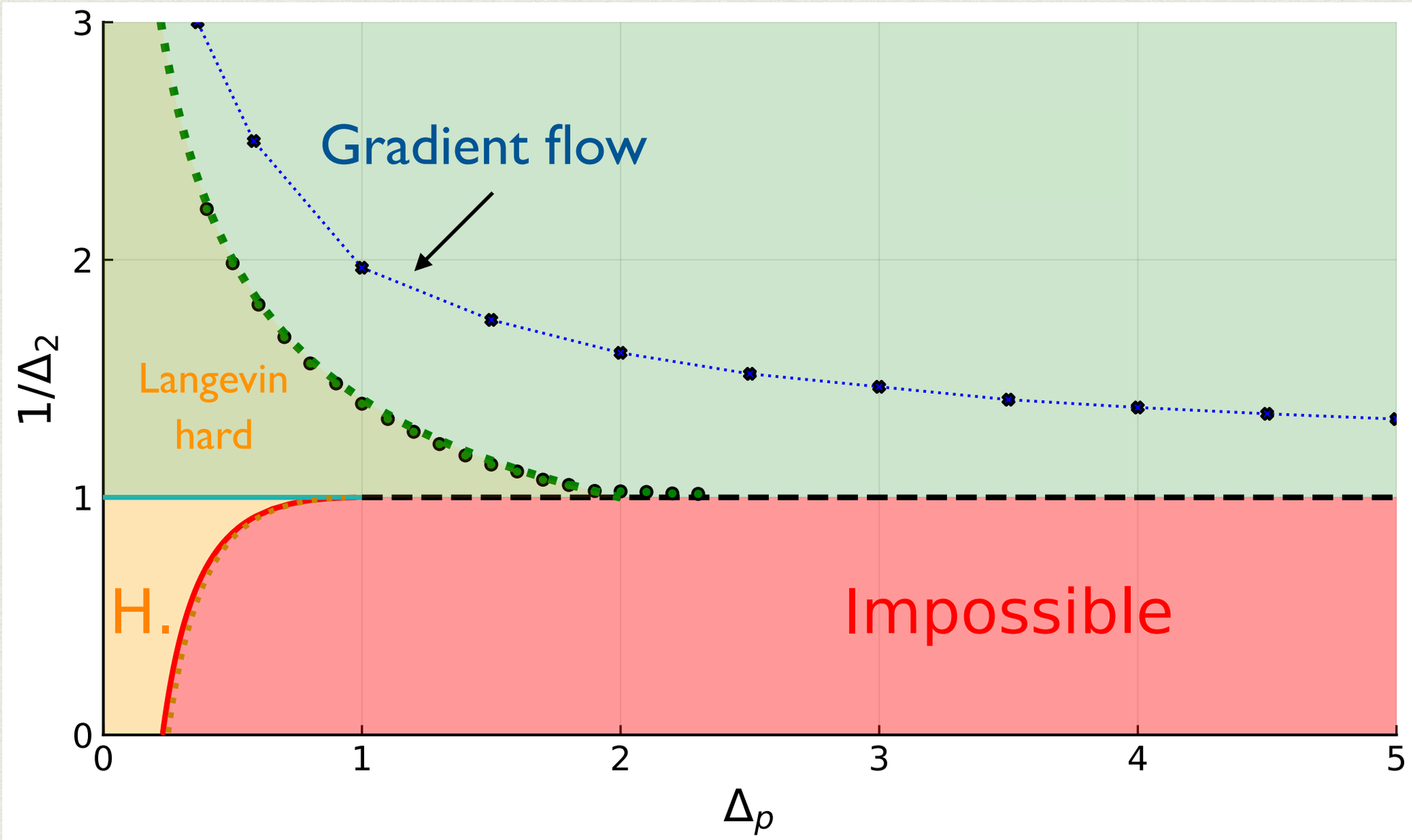
LANGEVIN PHASE DIAGRAM

p=3

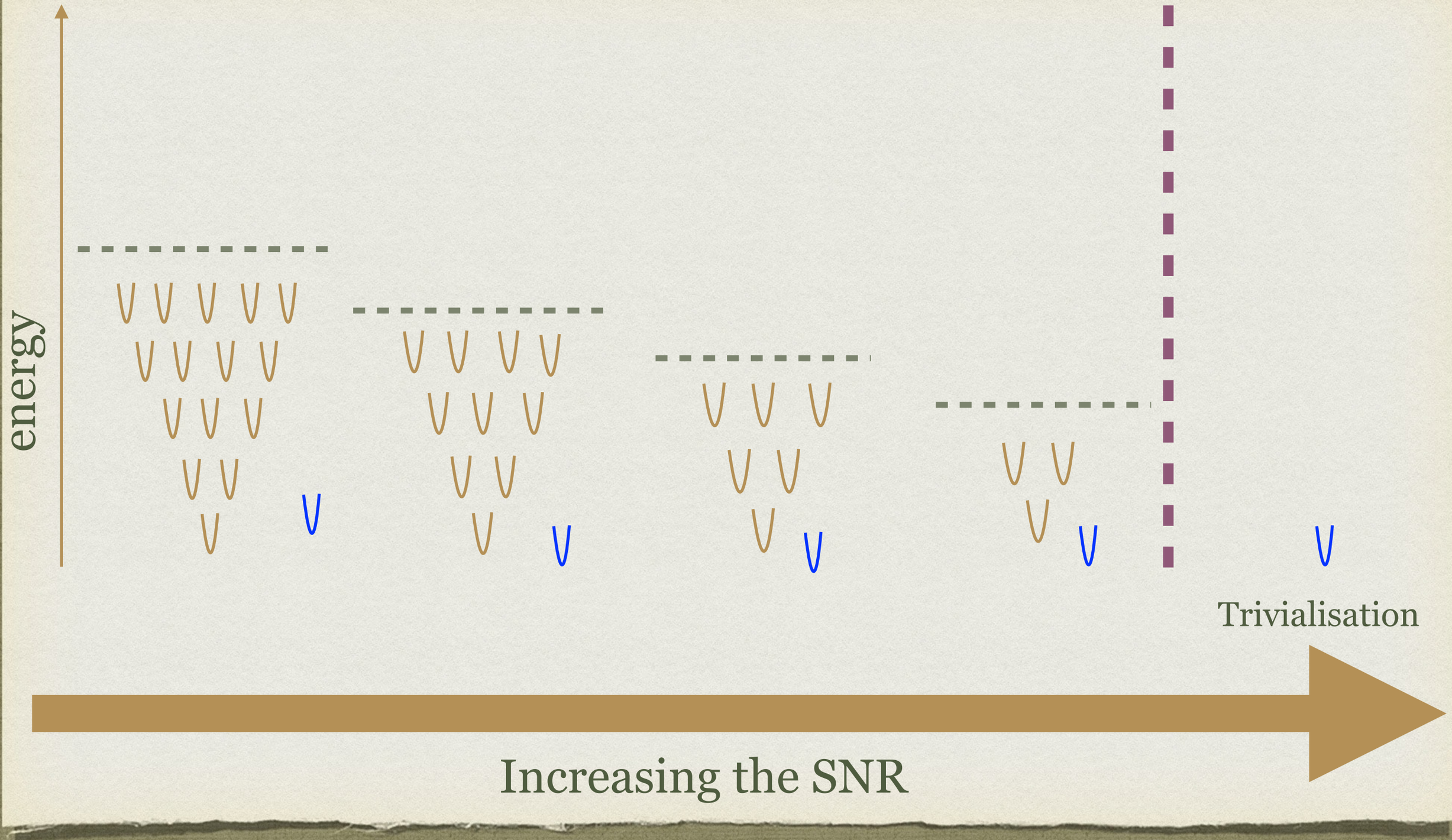


GRADIENT-FLOW PHASE DIAGRAM

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

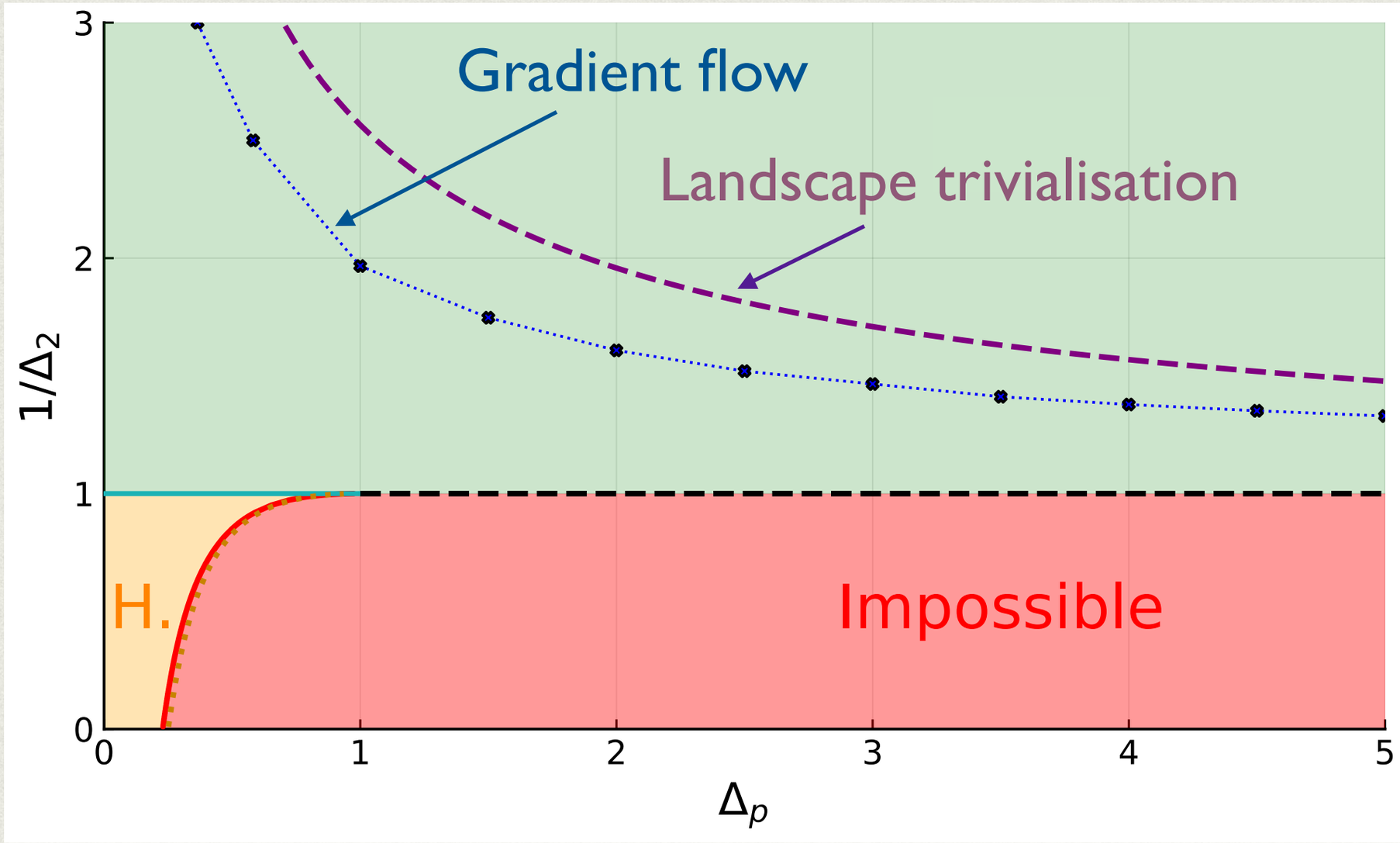
where:

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

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

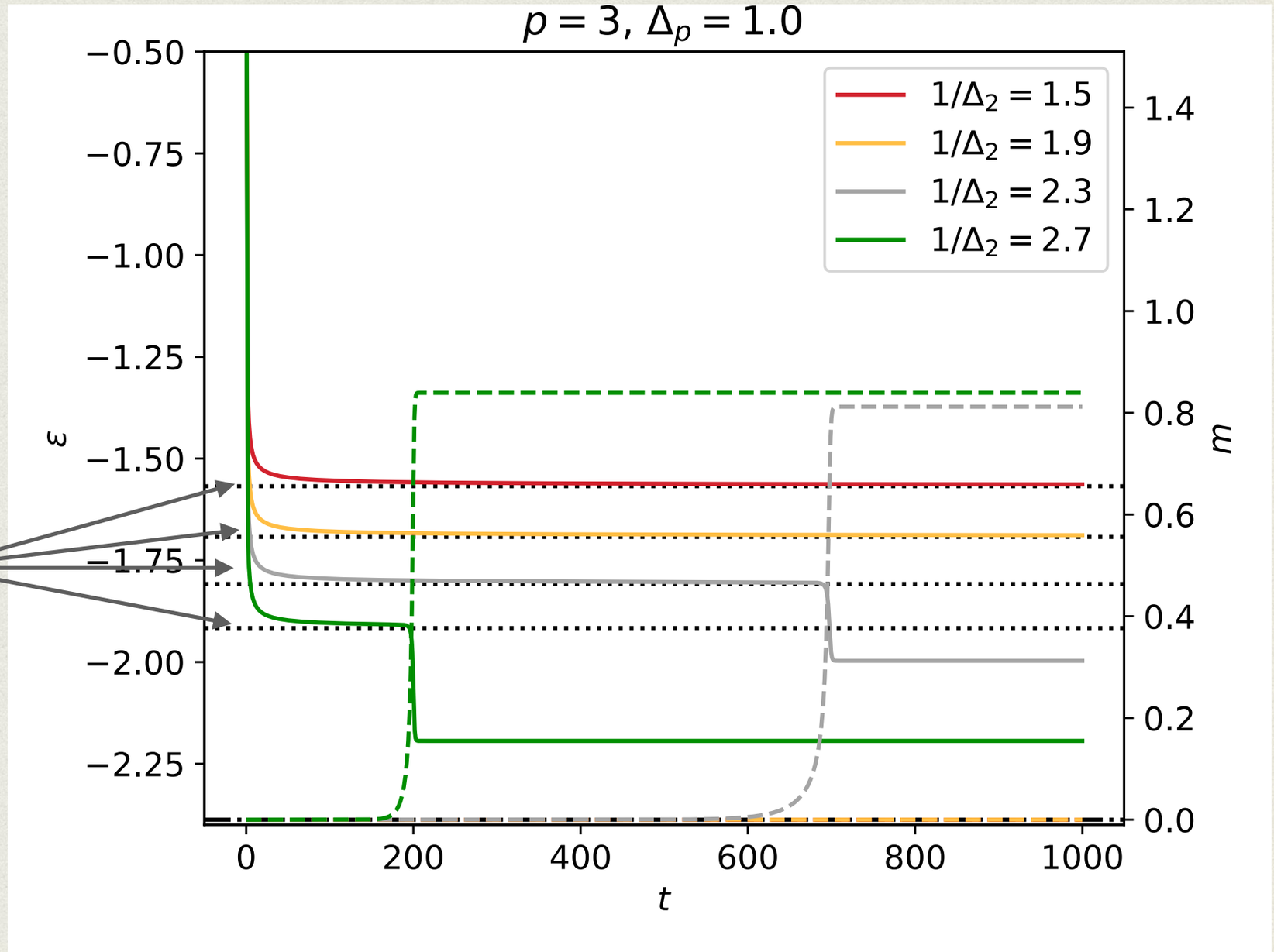
SPURIOUS MINIMA DO NOT NECESSARILY CAUSE GF TO FAIL

p=3



WHAT IS GOING ON?

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



TRANSITION RECIPE

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

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

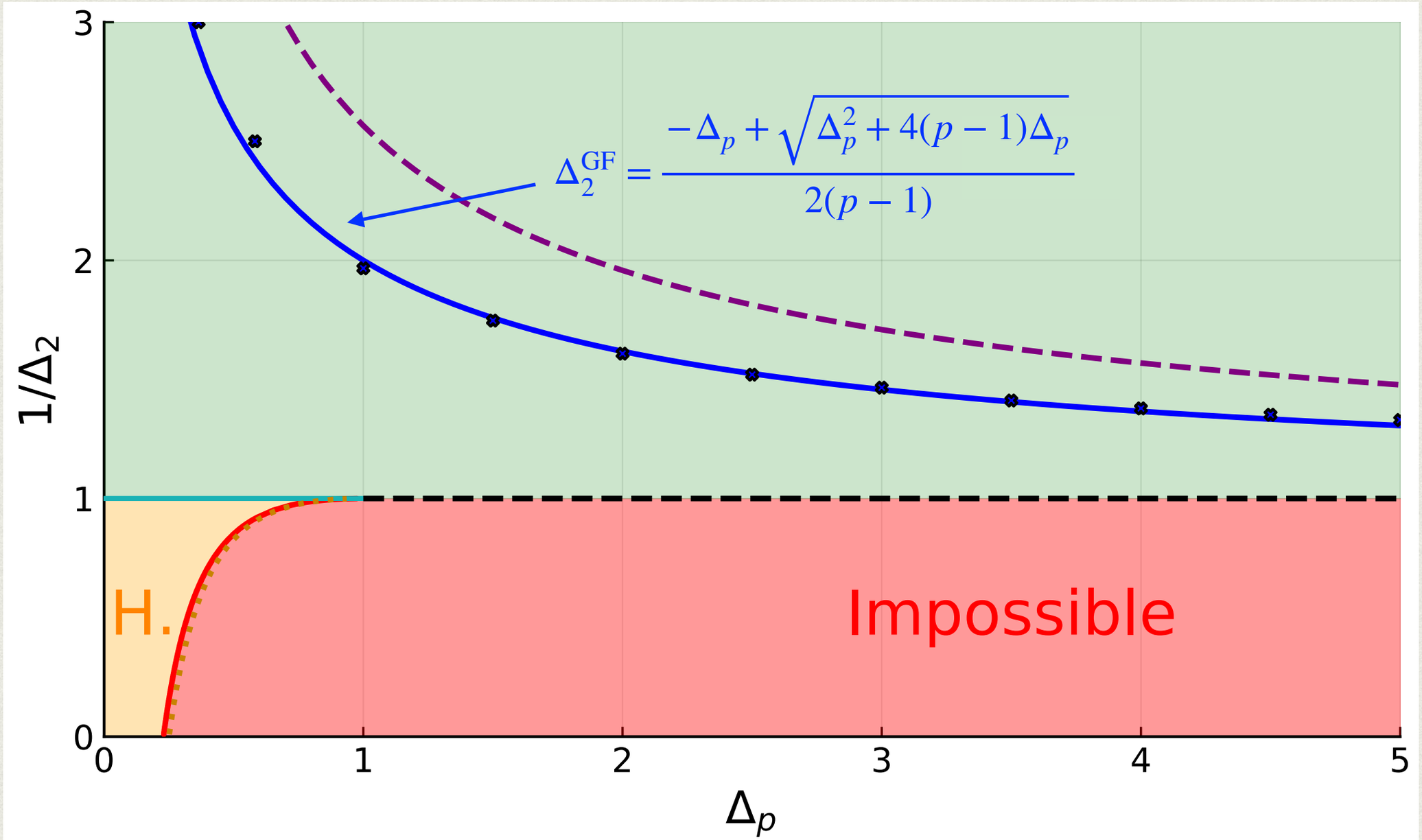
AMP state evolution at fixed q, determines stability of T=0:

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

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

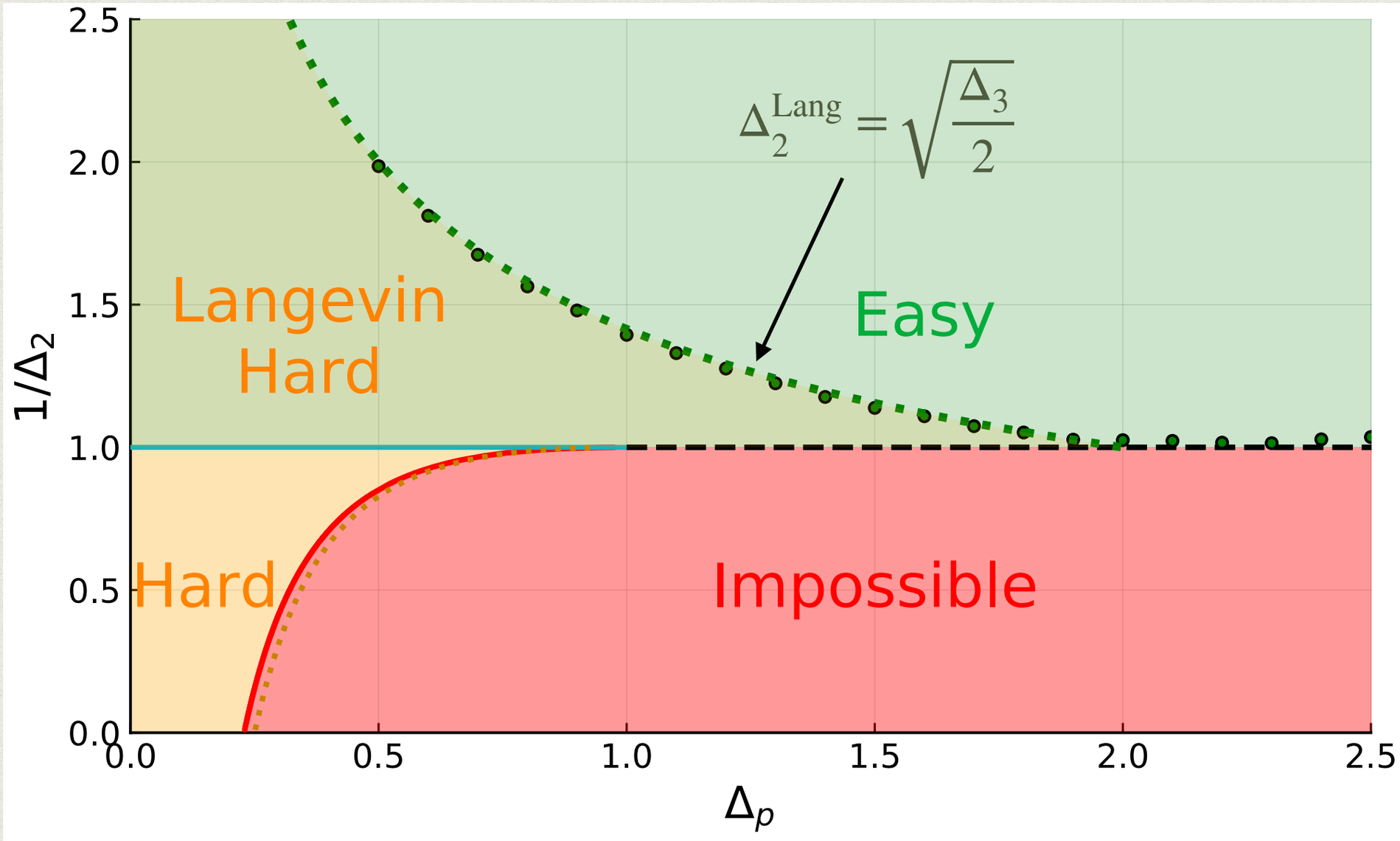
$$\frac{1}{\Delta_2^2} = (p-1)\frac{(1-T\Delta_2)^{p-2}}{\Delta_p} + \frac{1}{\Delta_2}$$

GRADIENT-FLOW PHASE DIAGRAM



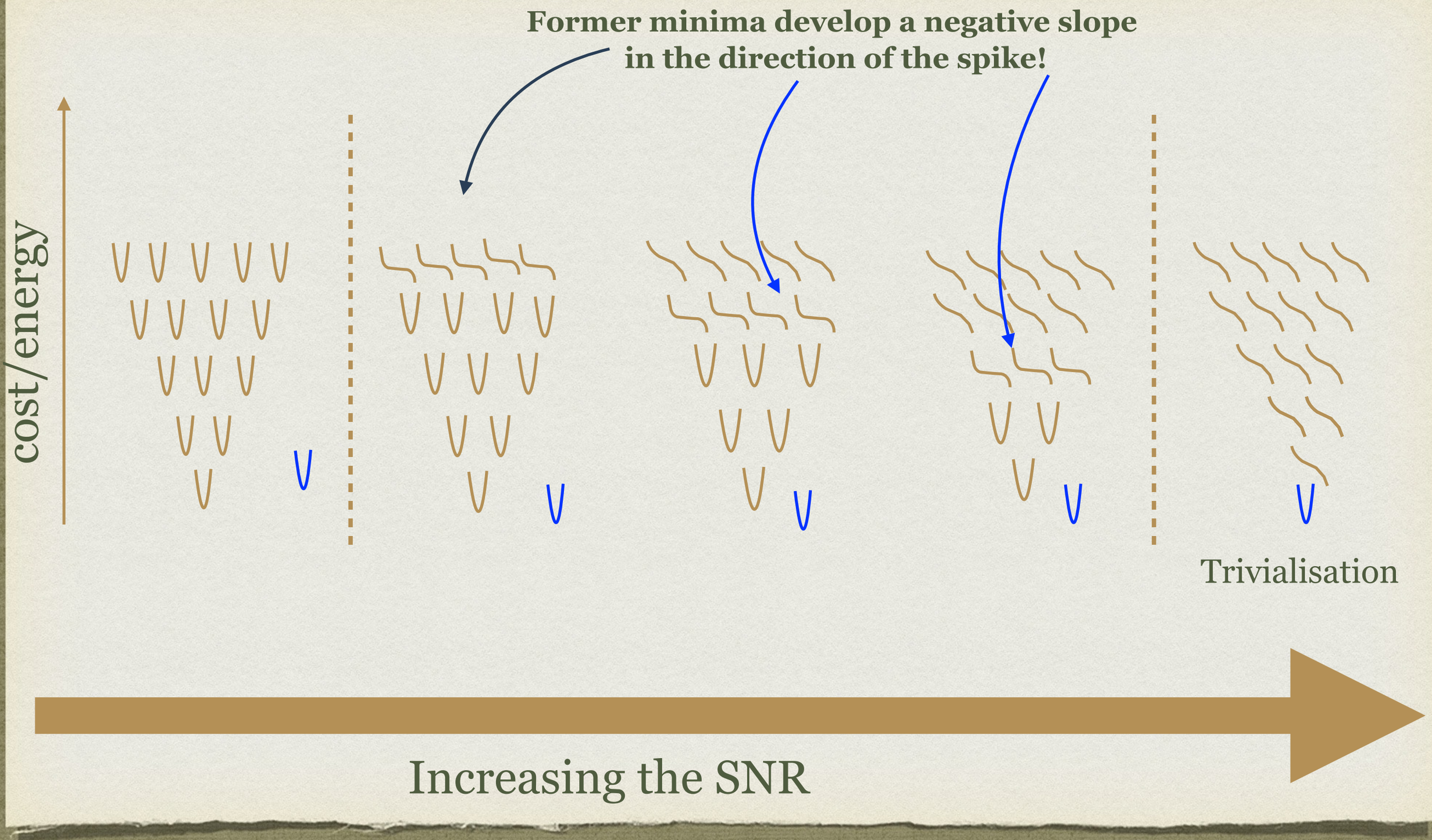
LANGEVIN PHASE DIAGRAM

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?
  Proof that our conjecture for the threshold is correct.

- 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

Marc
Thursday

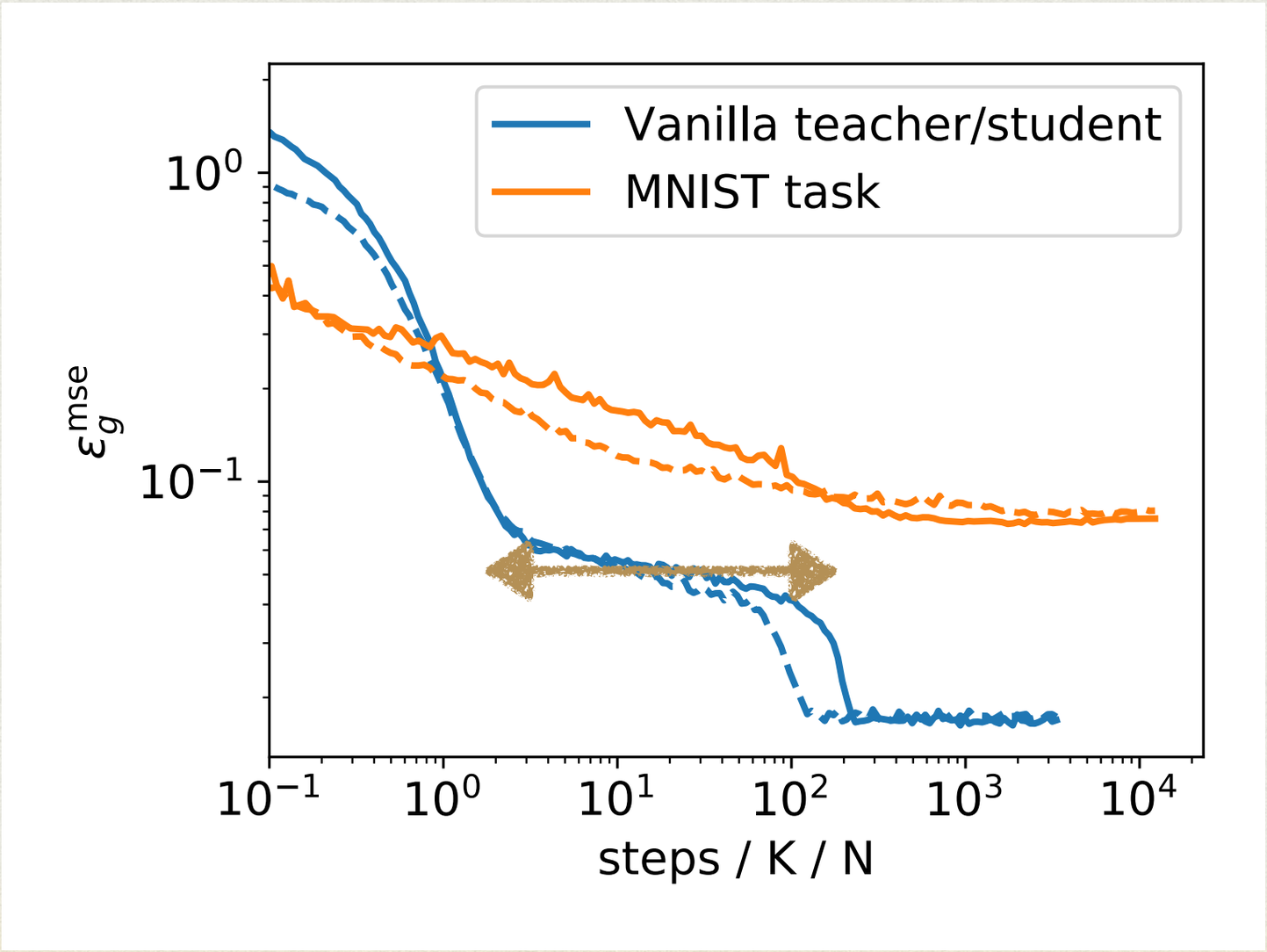
MNIST VS TEACHER/STUDENT

Goldt, Krzakala, Mézard, LZ; arXiv:1909.11500

Teacher/student:

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

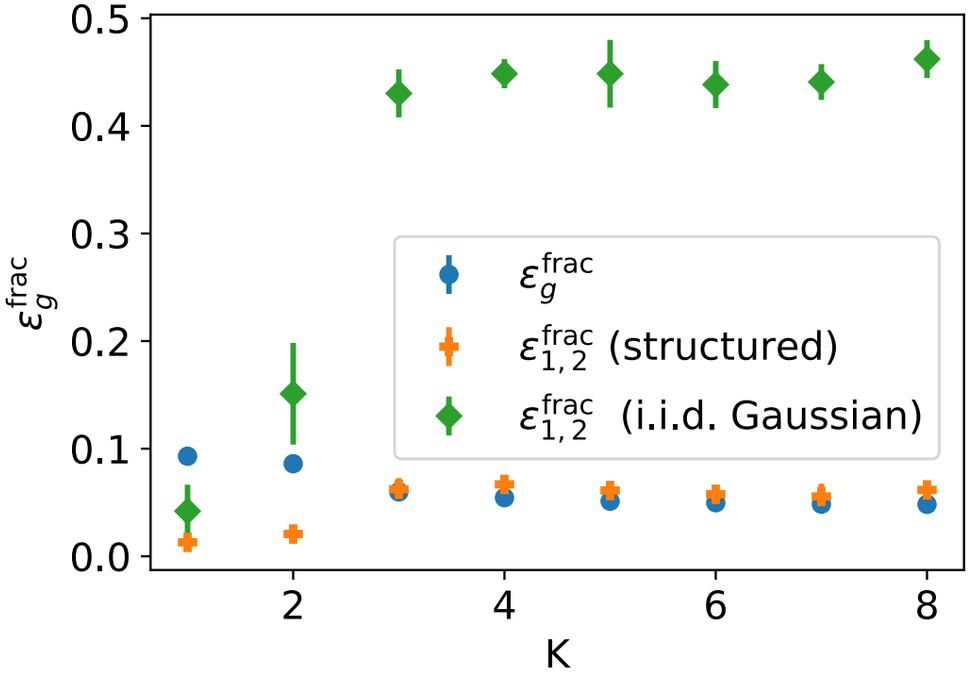
MNIST (even vs odd classification): No plateau ...



MNIST VS TEACHER/STUDENT

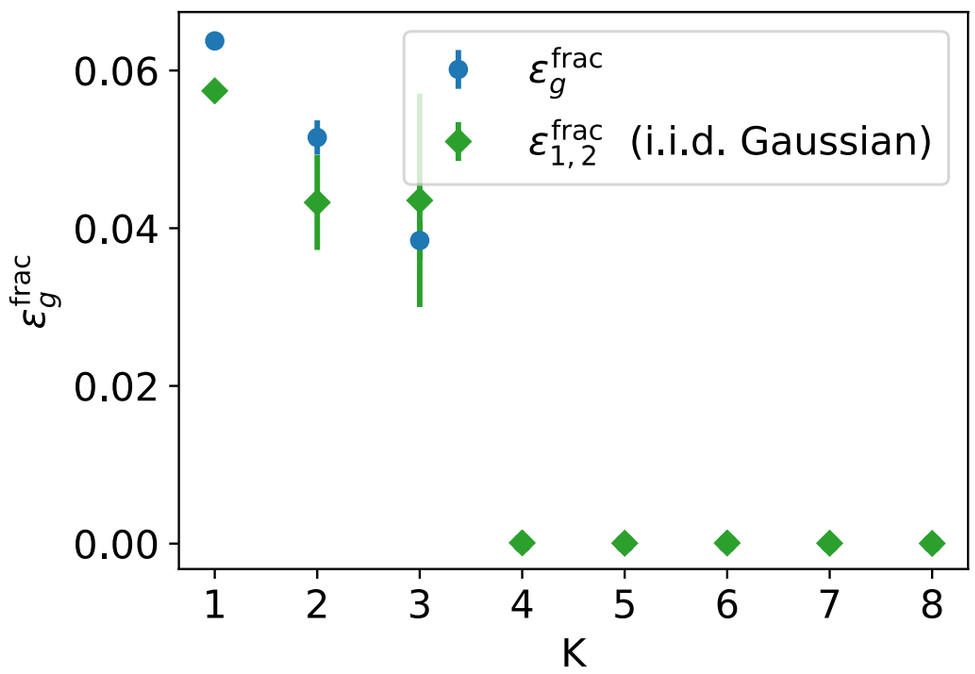
MNIST (odd vs even):

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

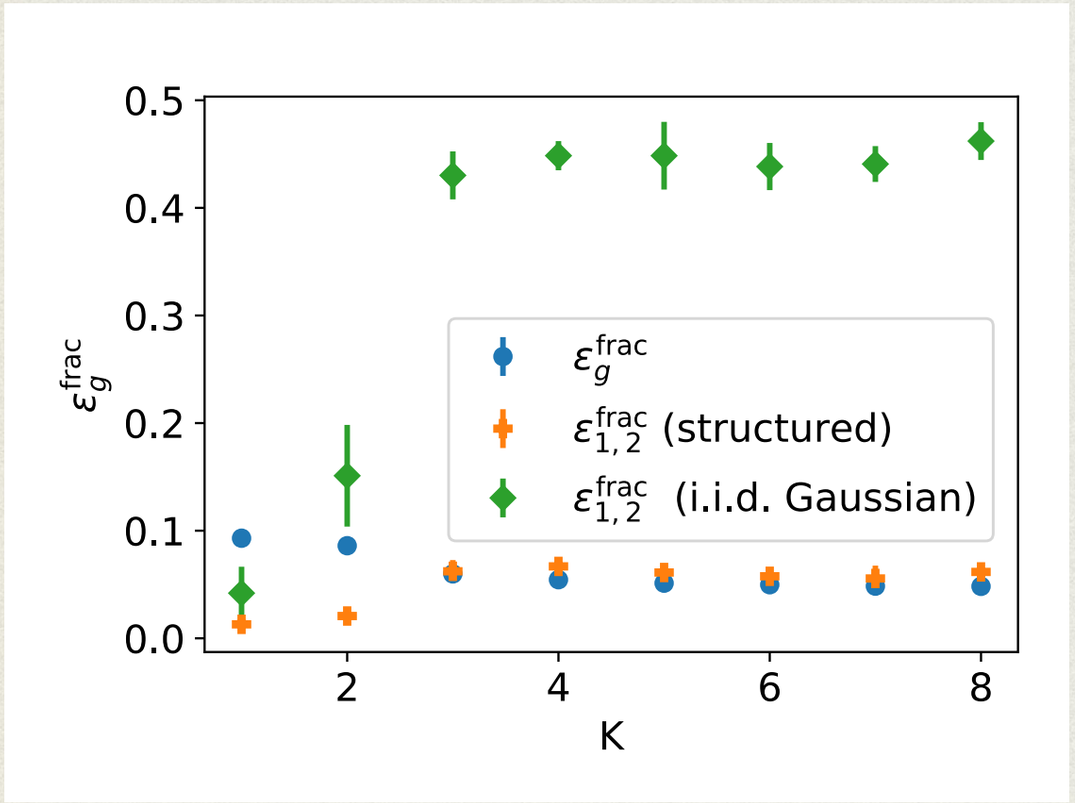
X is iid matrix

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

MNIST VS HIDDEN MANIFOLD

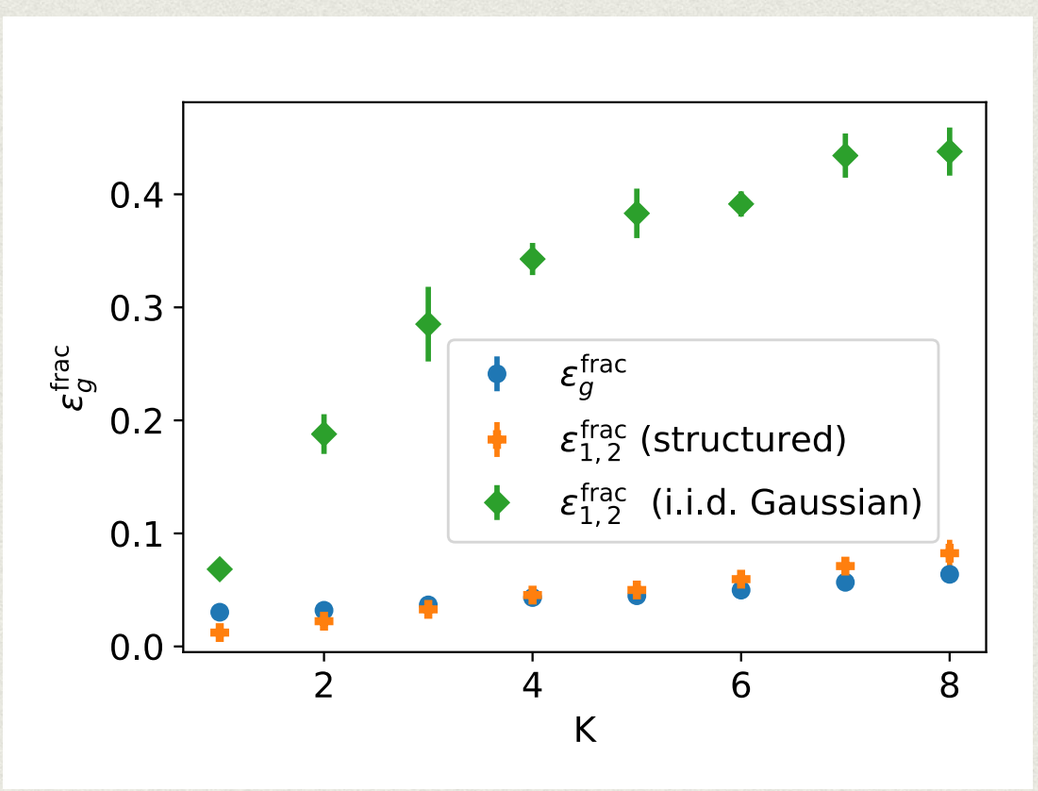
MNIST (odd vs even):

Two independent students do not learn the same function!



Hidden manifold (d=10)

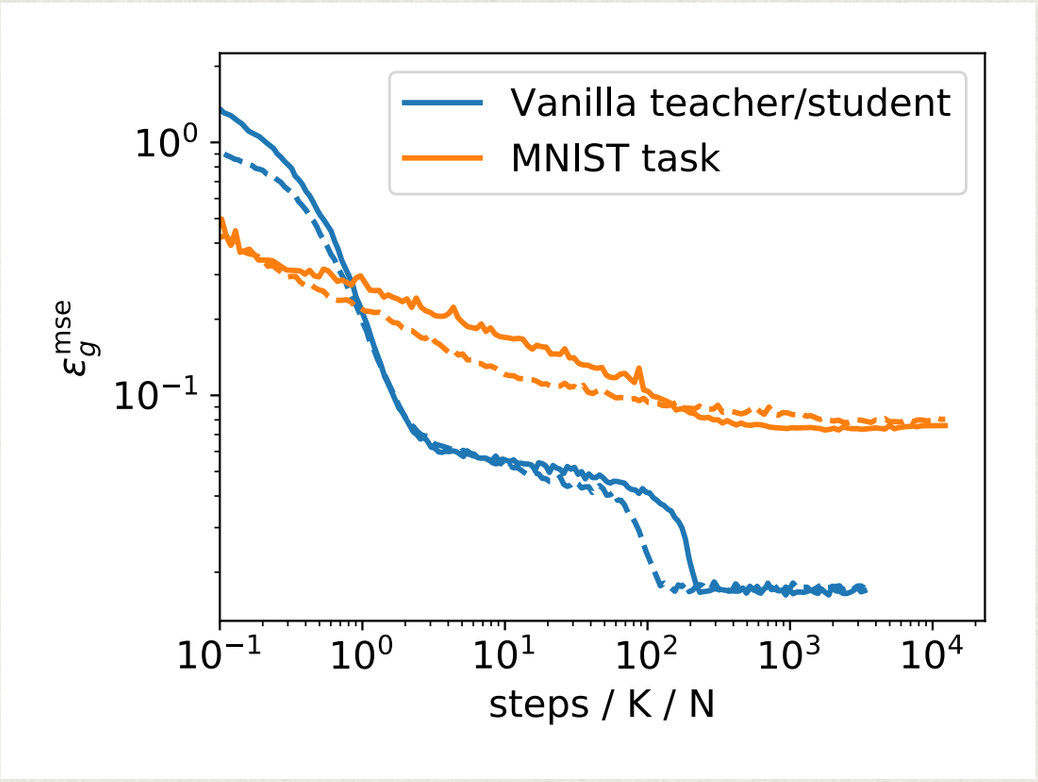
Two independent students do not learn the same function!

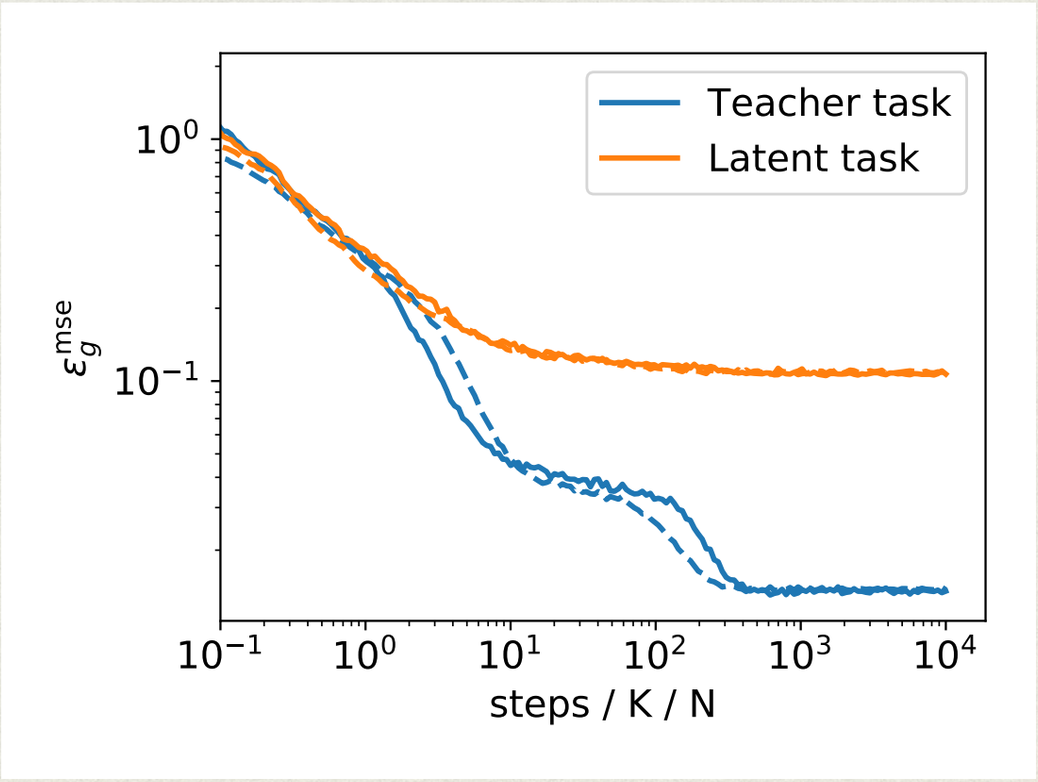


MNIST VS HIDDEN MANIFOLD

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.
- To appear: Solve analytically.
- TODO: Generalize to be able to demonstrate the advantage of depth.