# INSIGHTS ON GRADIENT-BASED ALGORITHMS IN HIGH-DIMENSIONAL LEARNING

Lenka Zdeborová
(EPFL)

Co-authors: G. Biroli, C. Cammarota, F. Krzakala, S. Mannelli Sarao, F. Mignacco, P. Urbani, E. Vanden-Eijnden.

Physics ∩ ML webinar, 9. 9. 2020

# REFERENCES FOR THIS TALK


- Sarao Mannelli, Biroli, Cammarota, Krzakala, Urbani, LZ; *Marvels and Pitfalls of the Langevin Algorithm in Noisy High-dimensional Inference;* Phys. Rev. X'20, arXiv:1812.09066
- Sarao Mannelli, Krzakala, Urbani, LZ; *Passed & Spurious: Descent Algorithms and Local Minima in Spiked Matrix-Tensor Models;* ICML'19, arXiv:1902.00139.
- Sarao Mannelli, Biroli, Cammarota, Krzakala, LZ; *Who is Afraid of Big Bad Minima? Analysis of Gradient-Flow in a Spiked Matrix-Tensor Model*; NeurIPS'19, arXiv:1907.08226.

- Sarao Mannelli, Biroli, Cammarota, Krzakala, Urbani, LZ; *Complex Dynamics and Simple Neural Networks: Understanding Gradient Flow in Phase Retrieval,* arXiv:2006.06997.
- Sarao Mannelli, Vanden-Eijnden, LZ; *Landscape and Dynamics in Over-Parametrized Neural Networks with Quadratic Activation*; arXiv:2006.15459.

- Mignacco, Urbani, Krzakala, LZ; *Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification*; arXiv:2006.06098.

# WORKHORSE OF MACHINE LEARNING

Stochastic Gradient Descent

Machine Learning

# IN DEEP LEARNING

- Empirical observation: Local (even global) minima with bad generalisation error do exist.

**Bad Global Minima Exist and SGD Can Reach Them**

**Shengchao Liu, Dimitris Papailiopoulos**
University of Wisconsin–Madison

**Dimitris Achlioptas**
University of California, Santa Cruz

**Abstract**

Several recent works have aimed to explain why severely overparameterized models, generalize well when trained by Stochastic Gradient Descent (SGD). The emergent consensus explanation has two parts: the first is that there are "no bad local minima", while the second is that SGD performs implicit regularization by having a bias towards low complexity models. We revisit both of these ideas in the context of image classification with common deep neural network architectures. Our first finding is that there exist bad *global* minima, *i.e.*, models that fit the training set perfectly, yet have poor generalization. Our second finding is that given only *unlabeled* training data, we can easily construct initializations that will cause SGD to quickly converge to such bad global minima. For example, on CIFAR, CINIC10, and (Restricted) ImageNet, this can be achieved by starting SGD at a model derived by fitting random labels on the training data: while subsequent SGD training (with the correct labels) will reach zero training error, the resulting model will exhibit a test accuracy degradation of up to 40% compared to training from a random initialization. Finally, we show that regularization seems to provide SGD with an escape route: once heuristics such as data augmentation are used, starting from a complex model (adversarial initialization) has no effect on the test accuracy.

- Question: How do gradient-based algorithms manage to avoid bad minima with limited number of samples?

# STRATEGY

- Goal: We need to understand the whole trajectory of gradient-based algorithms in non-convex high-dimensional problems.

- In practice: Number of samples is limited & constants matter.

- Simplify: Work with synthetic model data as a first step to get insight on the behaviour of algorithms.

# SPIKED MATRIX-TENSOR MODEL

Loss: $$\mathscr{L}(x) = \|xx^\top - Y\|_2^2 + \|x^{\otimes p} - T\|_2^2$$

where: $$Y = x^*(x^*)^\top + \mathcal{N}(0, \tilde{\Delta}_2)$$
$$T = (x^*)^{\otimes p} + \mathcal{N}(0, \tilde{\Delta}_p)$$

$$x, x^* \in \mathbb{S}^{N-1} \qquad N \to \infty$$

Goal: Find back a vector close to x* by gradient-descent on the loss.

# SPIKED MATRIX-TENSOR MODEL

- 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.

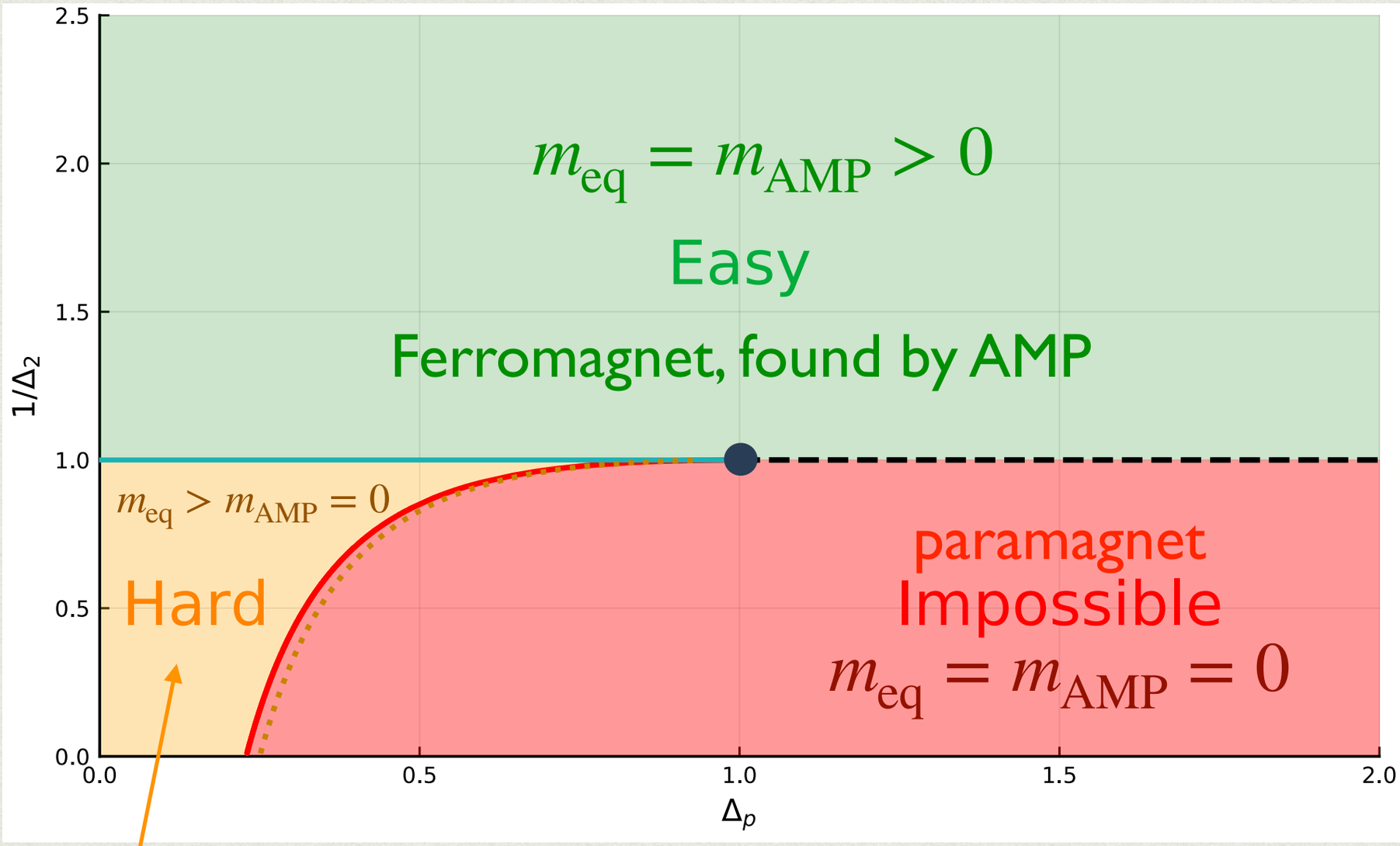
Maximum likelihood inference = computing the ground state.

➡ Gradient flow.

# PHASE DIAGRAM

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.

What happens at large constant time?

# 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.

# DYNAMICAL MEAN FIELD THEORY

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

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

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

$$N \to \infty$$

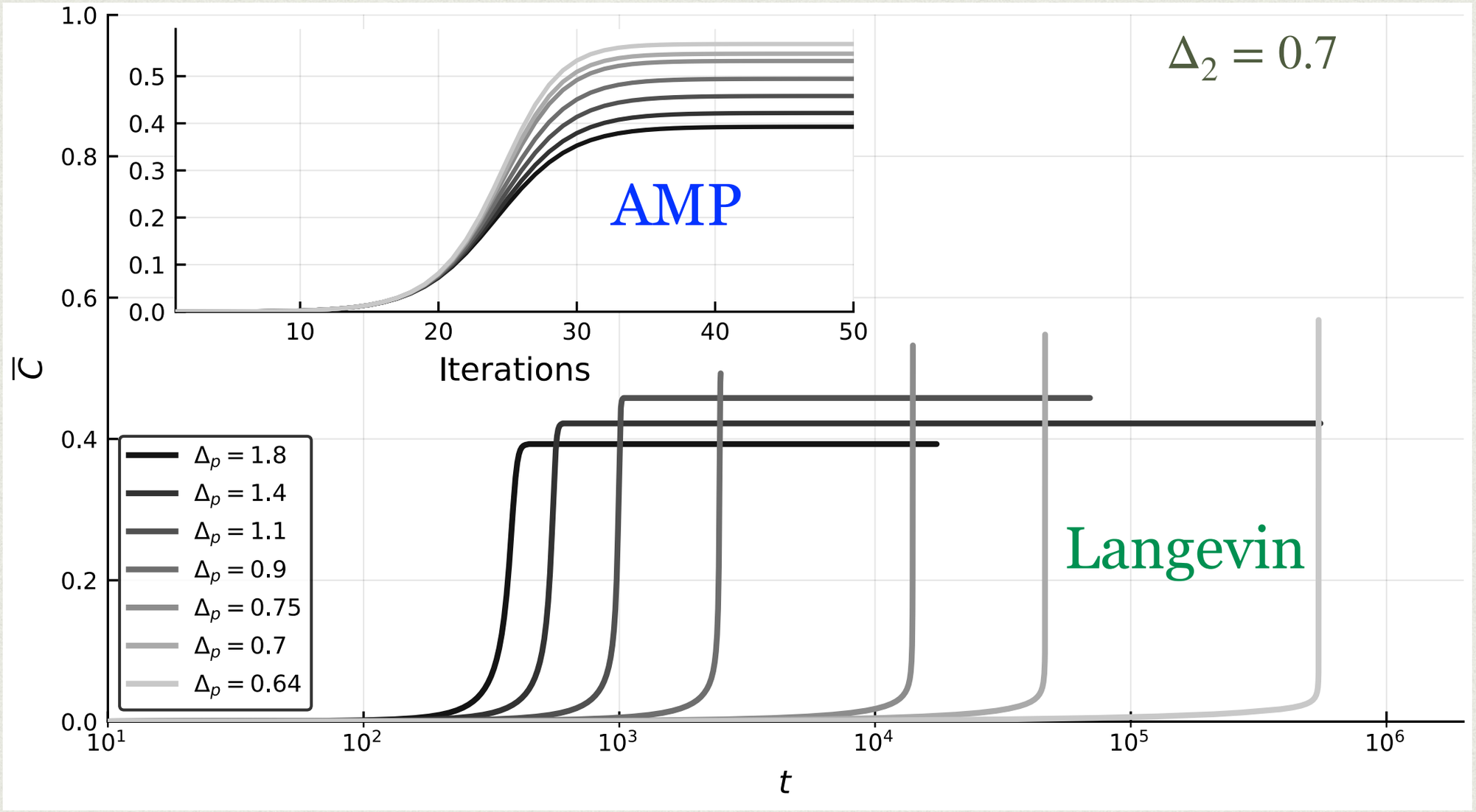
$$
\begin{aligned}
\frac{\partial}{\partial t}C(t,t') &= 2R(t',t) - \mu(t)C(t,t') + Q'(\overline{C}(t))\overline{C}(t') + \int_0^t dt'' R(t,t'')Q''(C(t,t''))C(t',t'') + \int_0^{t'} dt'' R(t',t'')Q'(C(t,t''))\,, \\
\frac{\partial}{\partial t}R(t,t') &= \delta(t-t') - \mu(t)R(t,t') + \int_{t'}^t dt'' R(t,t'')Q''(C(t,t''))R(t'',t')\,,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

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)

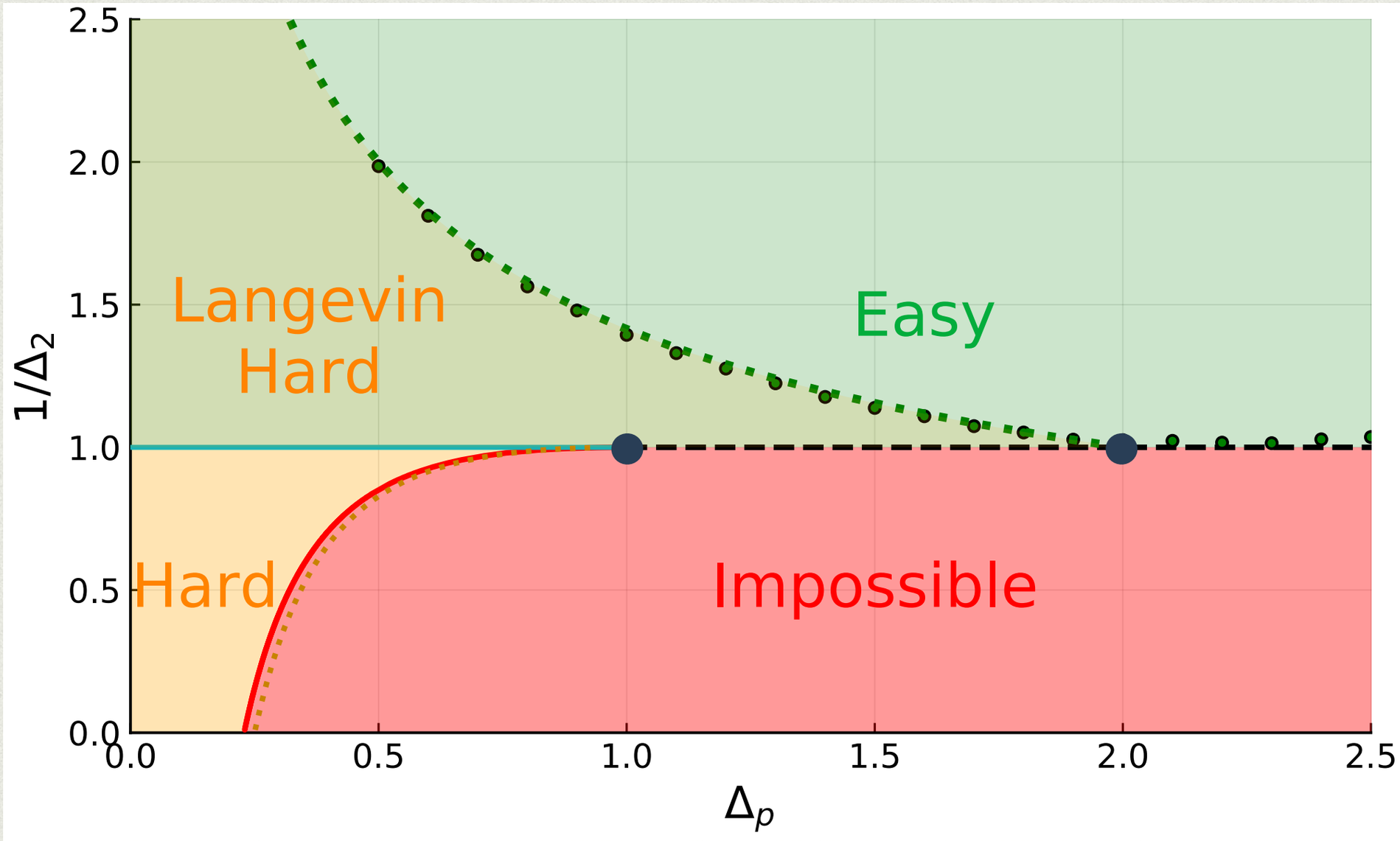
# LANGEVIN STATE EVOLUTION (NUMERICAL SOLUTION)




github.com/sphinxteam/spiked_matrix-tensor

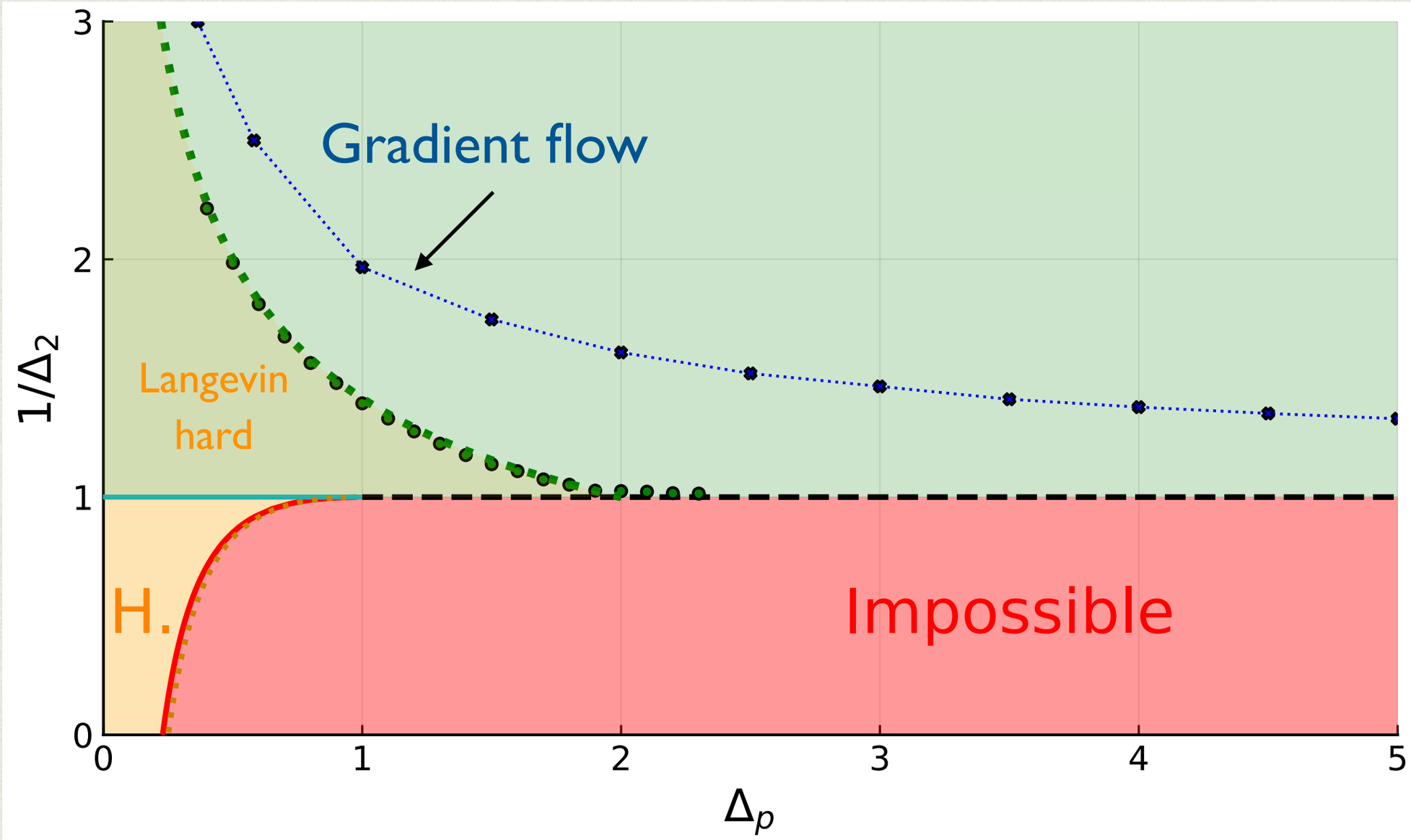
# LANGEVIN PHASE DIAGRAM

p=3

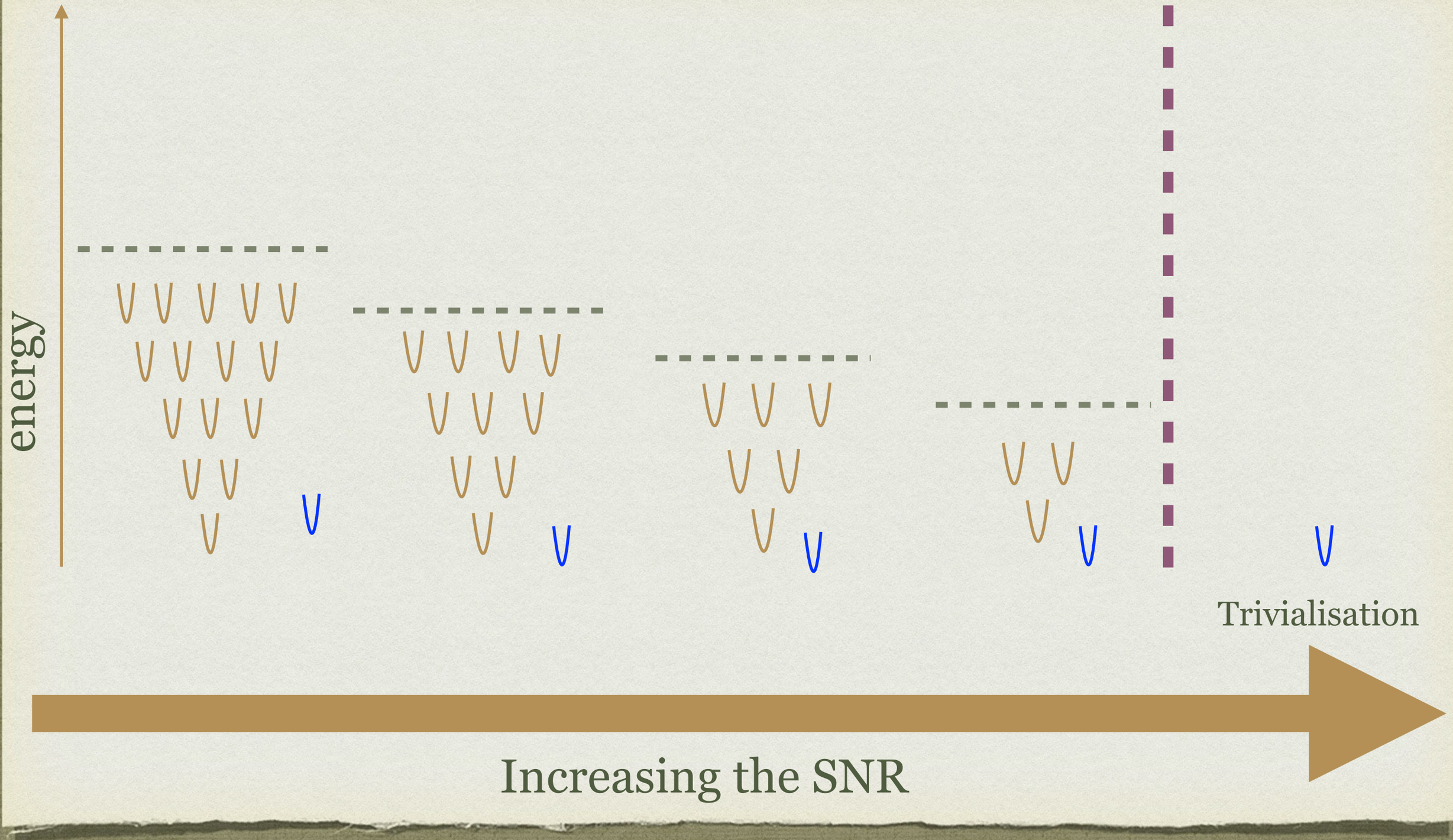


# GRADIENT-FLOW PHASE DIAGRAM

p=3



# POPULAR "EXPLANATION"



# COUNTING MINIMA: KAC-RICE

Sarao, Krzakala, Urbani, LZ, ICML'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

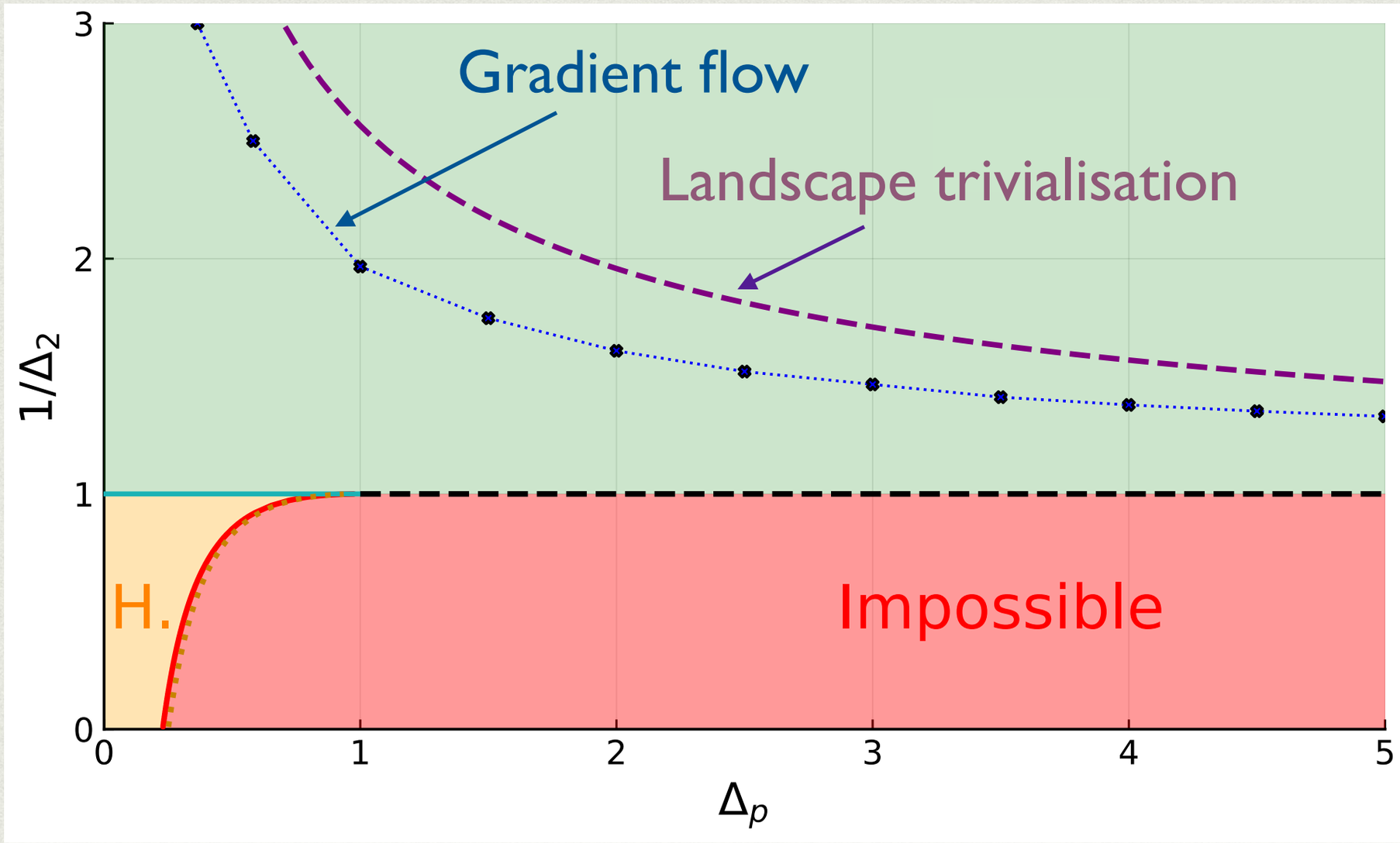
where:

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

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

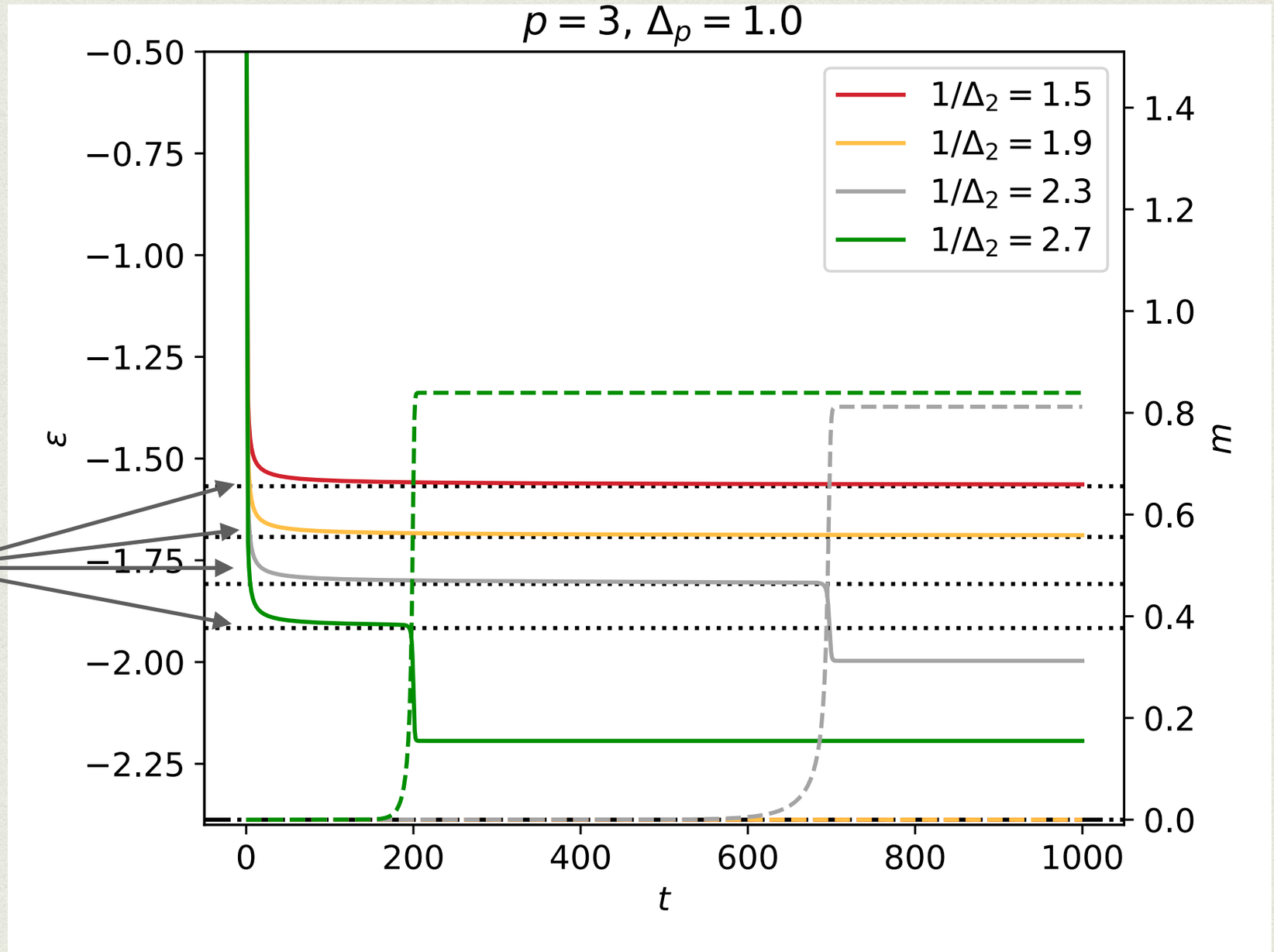
# SPURIOUS MINIMA DO NOT NECESSARILY CAUSE GF TO FAIL

p=3



# WHAT IS GOING ON?

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



# TRANSITION RECIPE

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

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

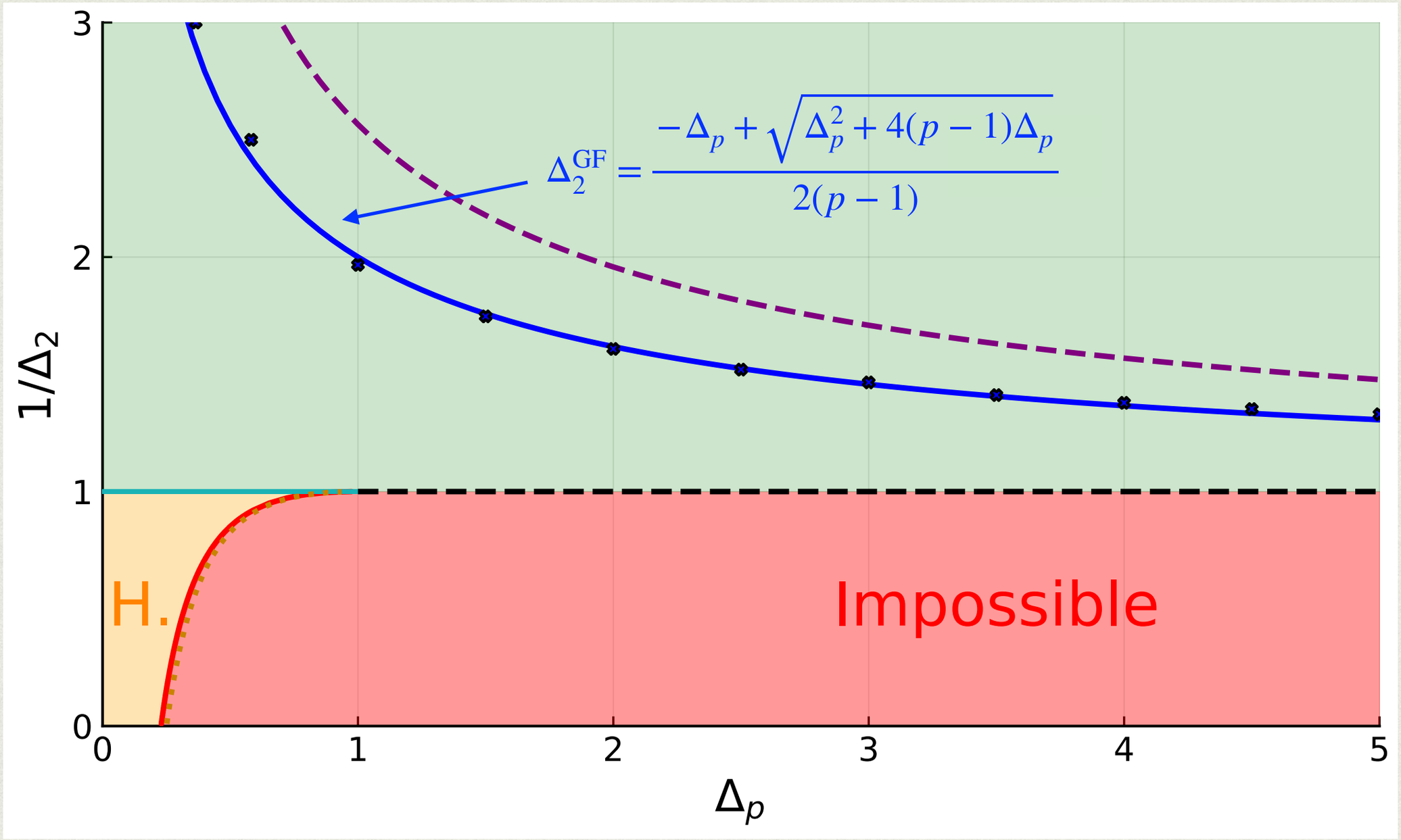
Condition for instability toward the solution at fixed q:
(derived from both Kac-Rice, and DMFT)

$$T\Delta_2 = 1 - q$$

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

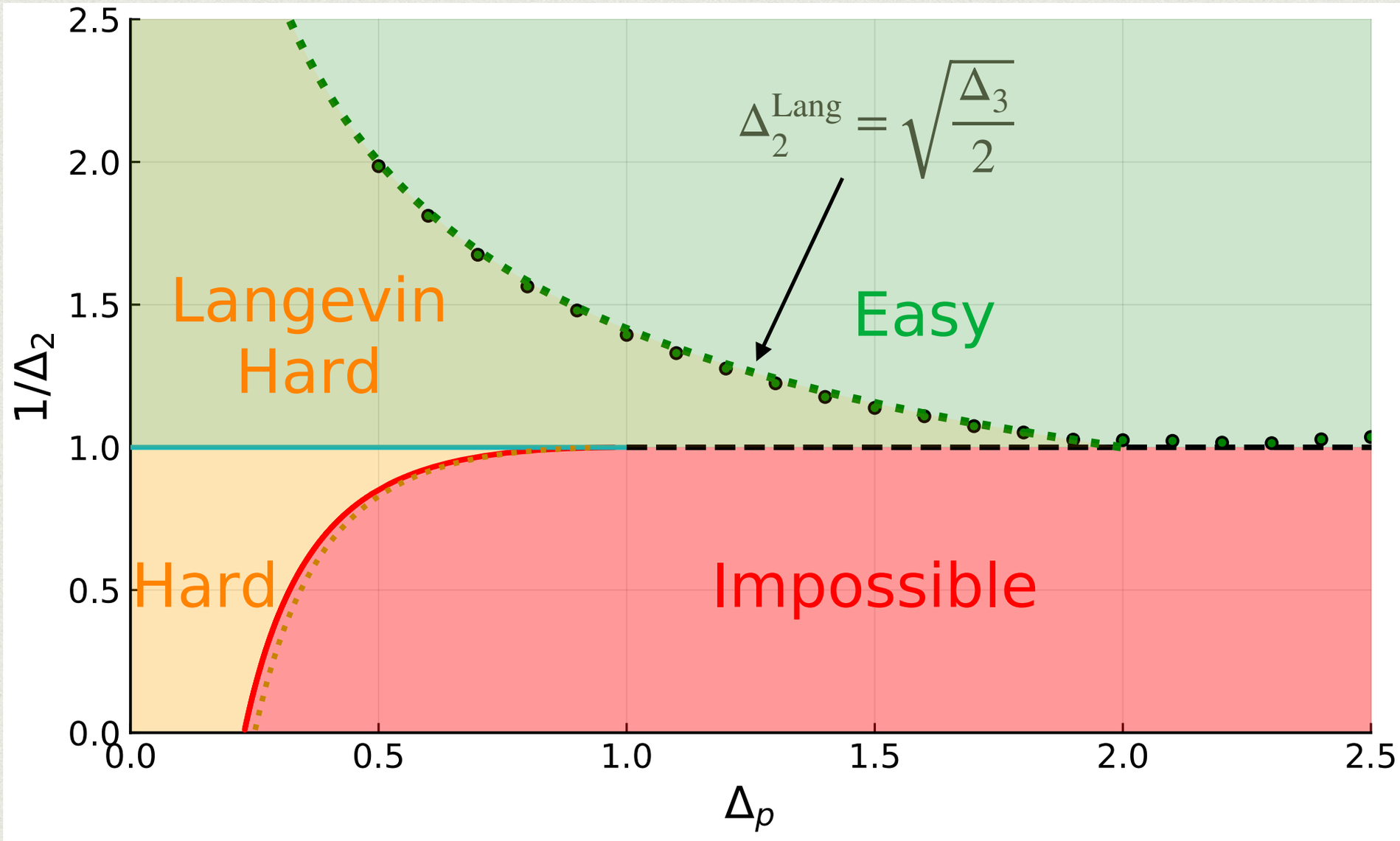
$$\frac{1}{\Delta_2^2} = (p-1)\frac{(1-T\Delta_2)^{p-2}}{\Delta_p} + \frac{1}{\Delta_2}$$

# GRADIENT-FLOW PHASE DIAGRAM



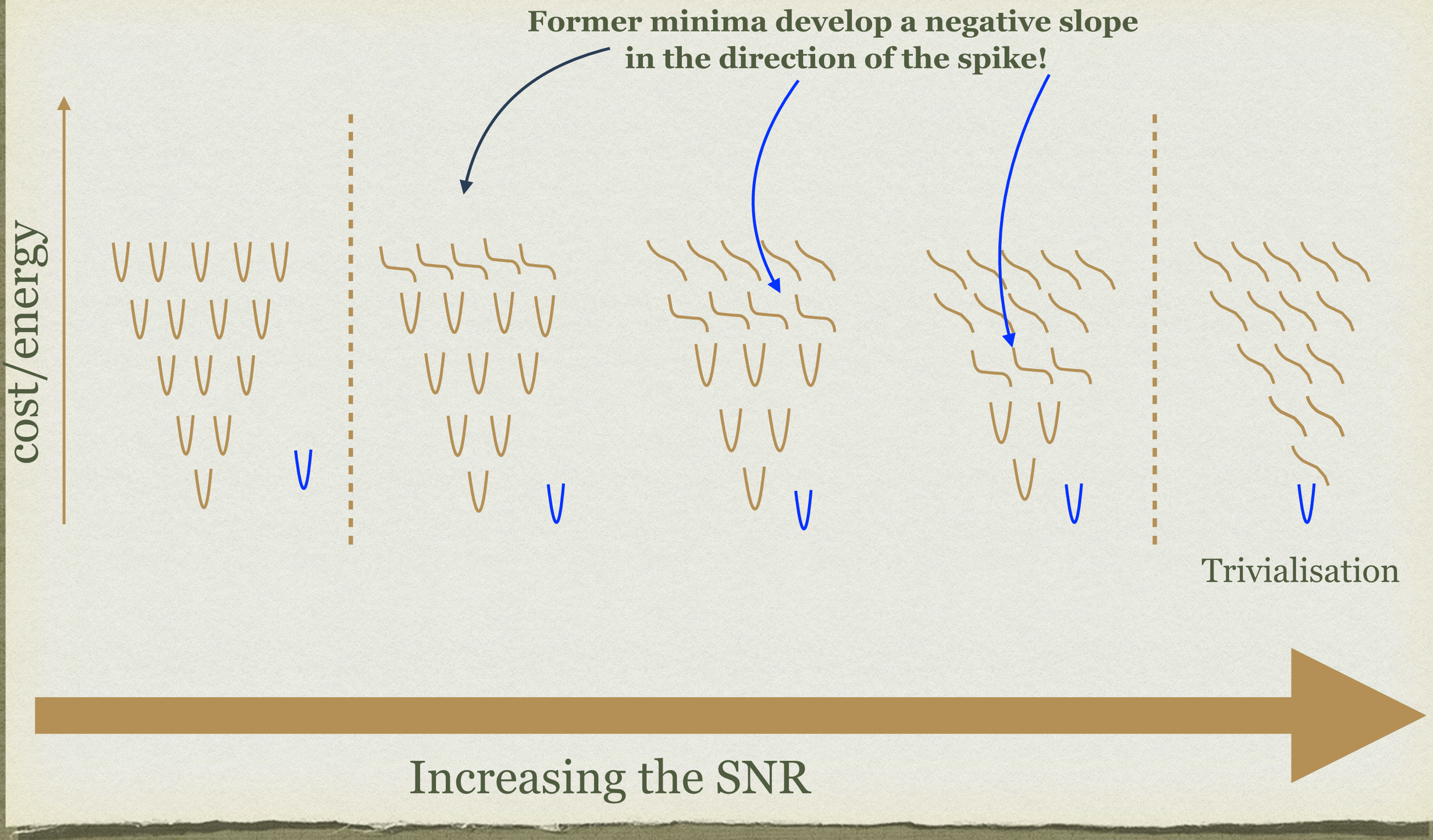
# LANGEVIN PHASE DIAGRAM

p=3



# LANDSCAPE ANALYSIS

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



# CONCLUSION ON SPIKED MATRIX-TENSOR MODEL

- First time we have a closed-form conjecture for the threshold of gradient-based algorithms including constants.
  Applicable for (simple) neural networks?

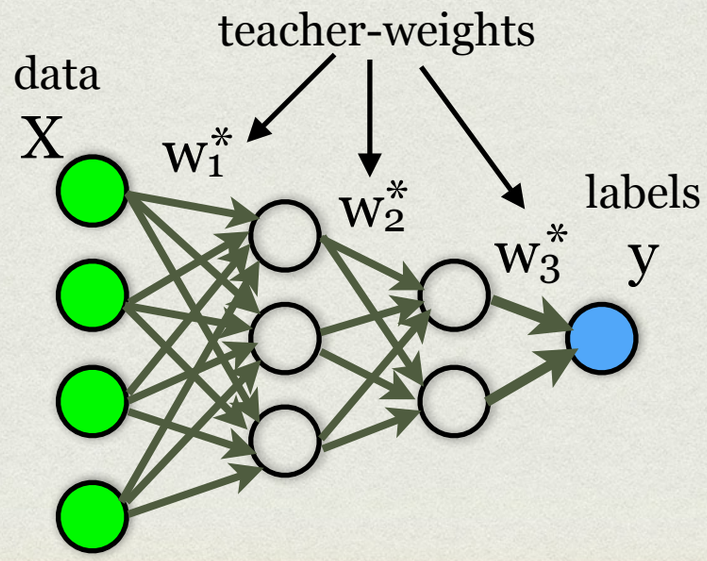
- Gradient flow worse than Langevin algorithm, both worse than AMP.
  How can GF & LA be improved to approach the AMP threshold?

- Gradient flow (sometimes) works even when spurious local minima are present. Quantified with the Kac-Rice approach.
  What about stochastic gradient descent?

# TEACHER-NEURAL SETTING

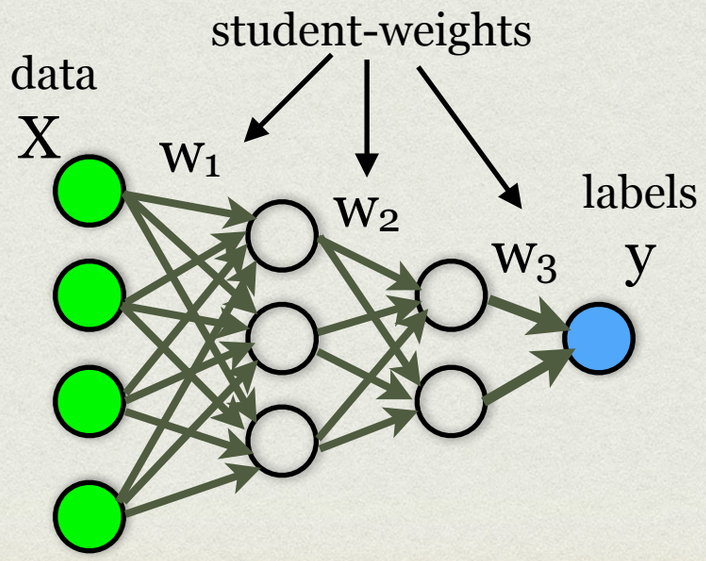
## Teacher-network

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




## Student-network

- Observes X, y.
- The architecture of the network is the same as the teacher or different.
- How does the 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

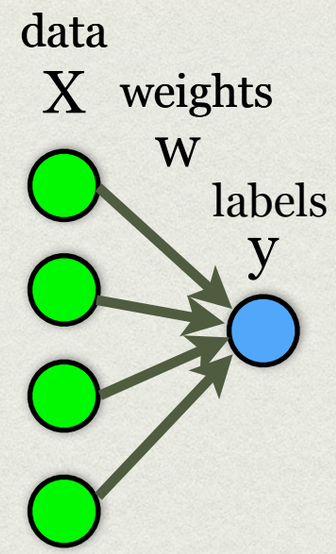
## Three unfinished works on the optimal storage capacity of networks

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 $y_\mu = \text{sign}\left( \sum_{i=1}^{p} X_{\mu i} w_i^* \right)$
- High-dimensional regime: $n \to \infty \quad p \to \infty$ p dimensions

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

# PHASE RETRIEVAL

- Broad range of applications in signal processing and imaging.
- Teacher-student setting with teacher having no hidden units, teacher's activation function is absolute value.

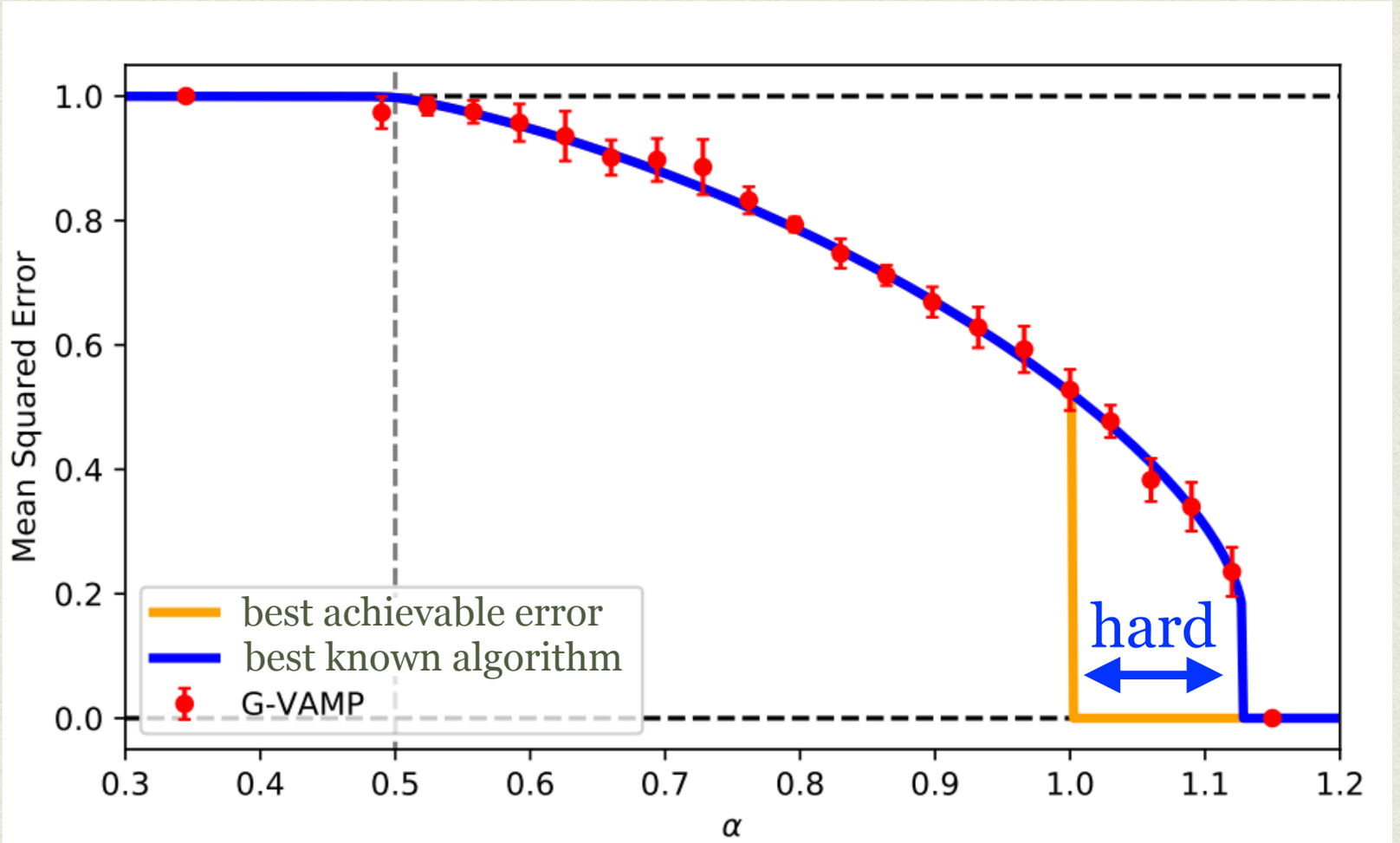
$$X_{\mu i} \sim \mathcal{N}(0,1/p) \qquad w_i^* \sim \mathcal{N}(0,1) \qquad \begin{matrix} \mu = 1,\ldots,n \\ i = 1,\ldots,p \end{matrix}$$

$$y_\mu = \left| \sum_{i=1}^{p} X_{\mu i} w_i^* \right|$$

Phase retrieval: Regression from training data $\{\mathbf{X}_\mu, y_\mu\}_{\mu=1}^{n}$

# PHASE RETRIEVAL: OPTIMAL SOLUTION

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




$$\alpha = \frac{n}{p}$$

$\alpha_{\mathrm{IT}} = 1$ # of samples need for perfect generalisation for any algorithm.

$\alpha_{\mathrm{AMP}} = 1.13$ # of samples need for perfect generalisation for approximate message passing algorithm (conjectured optimal among polynomial ones).

# GRADIENT DESCENT FOR PHASE RETRIEVAL

Loss function:

$$\mathscr{L}(\{w_i\}_{i=1}^p) = \sum_{\mu=1}^{n} \left[ y_\mu^2 - \left( \sum_{i=1}^{p} X_{\mu i} w_i \right)^2 \right]^2$$

where $y_\mu = \left| \sum_{i=1}^{p} X_{\mu i} w_i^* \right|$

Gradient flow:

$$\dot{w}_i(t) = -\partial_{w_i} \mathscr{L}\left( \{w_j(t)\}_{j=1}^p \right) + \mu(t) w_i(t)$$

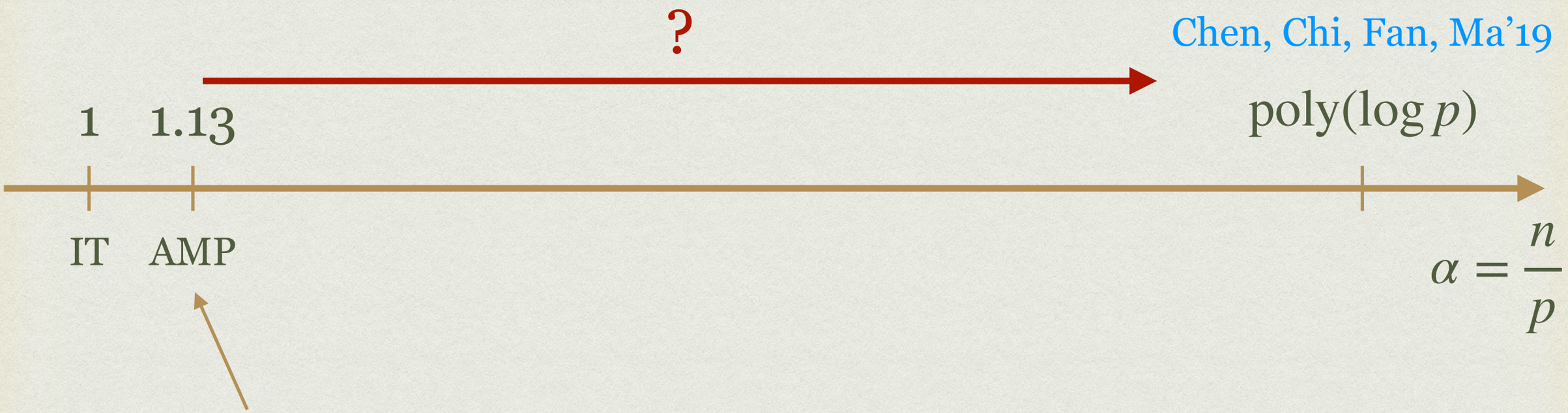
ensuring $|w|_2^2 = p$

Initialisation:

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

# PERFORMANCE OF GRADIENT DESCENT

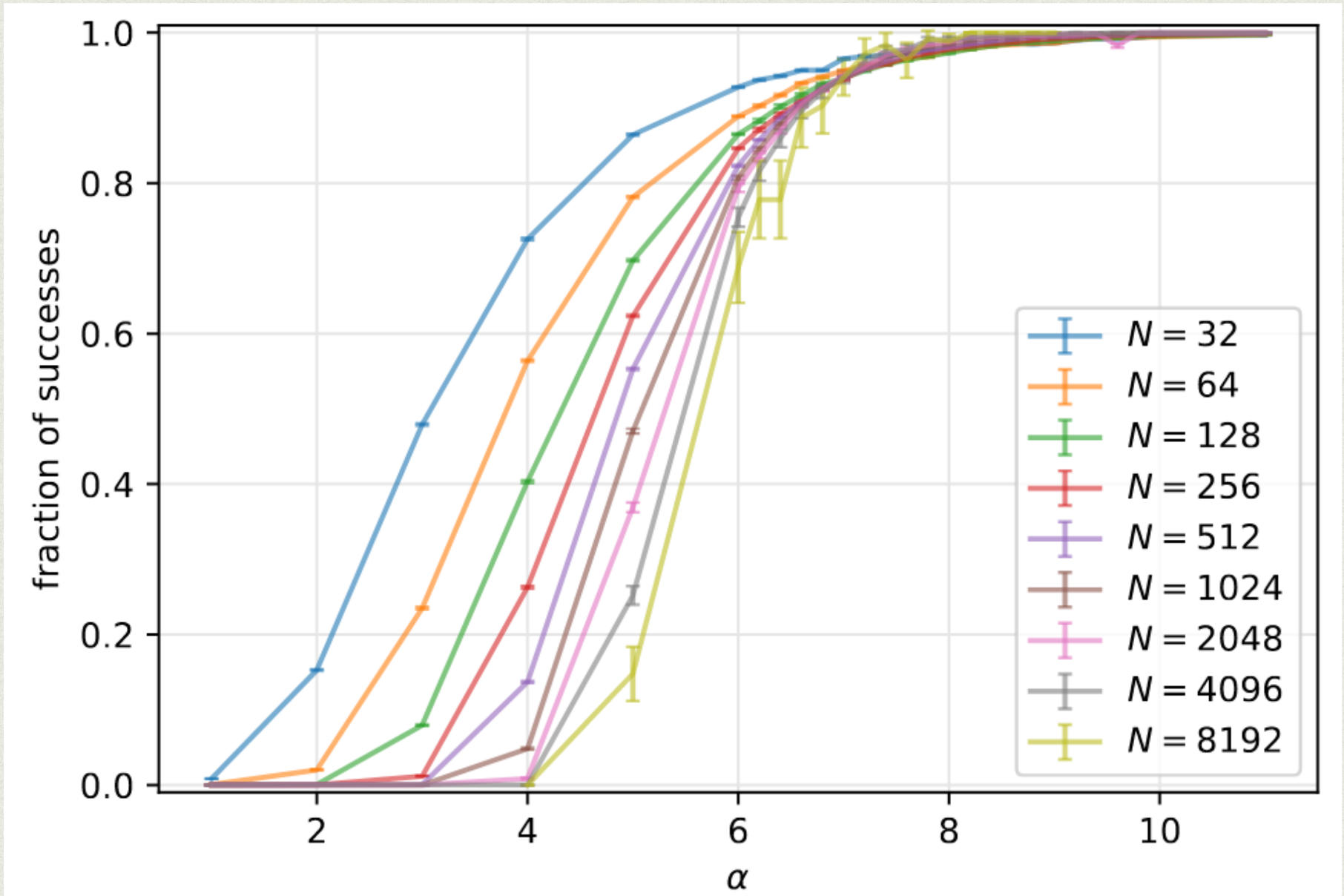
How many samples needed for perfect generalization?




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

# GRADIENT DESCENT NUMERICALLY

Sarao Mannelli, Biroli, Cammarota, Krzakala, LZ, 2006.06997.



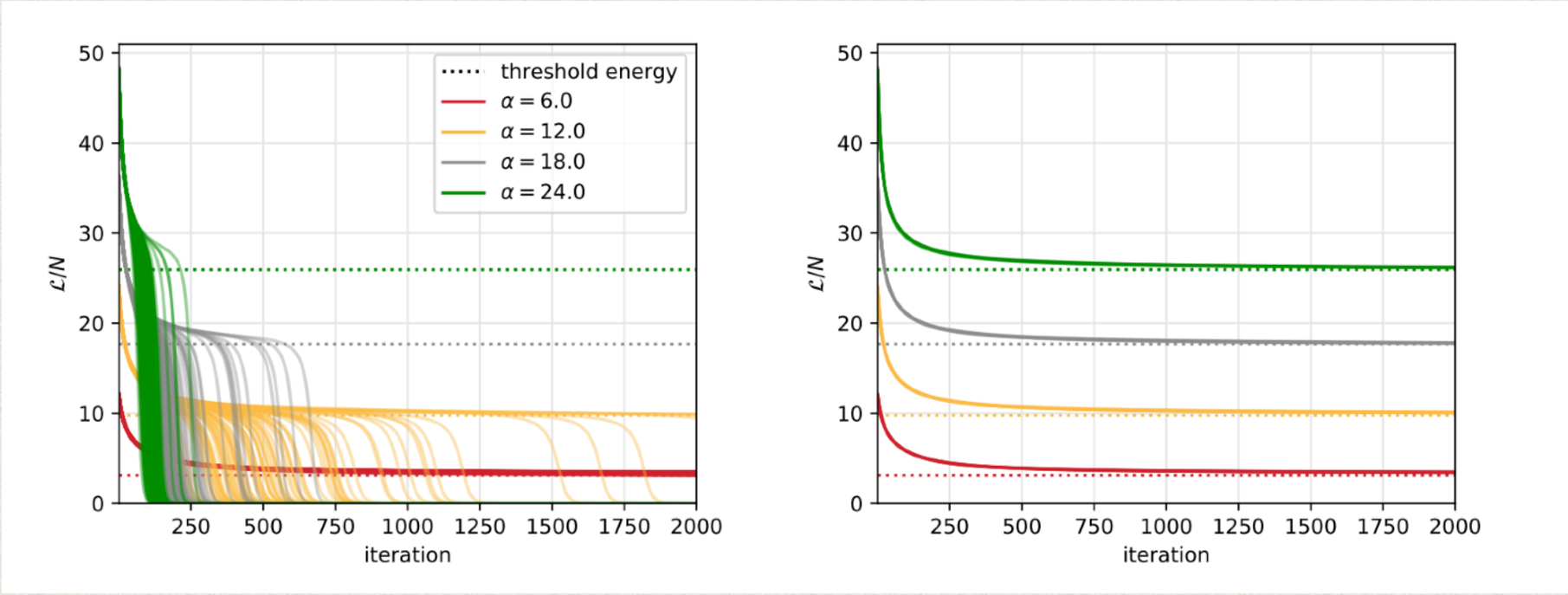

N=p

# TOWARDS A THEORY

Sarao Mannelli, Biroli, Cammarota, Krzakala, LZ, 2006.06997.

- Lesson from the spiked 2+p spin model: GF first goes to the threshold states and a BBP-like transition in the Hessian then drives success v.s. failure.

- True also in phase retrieval.

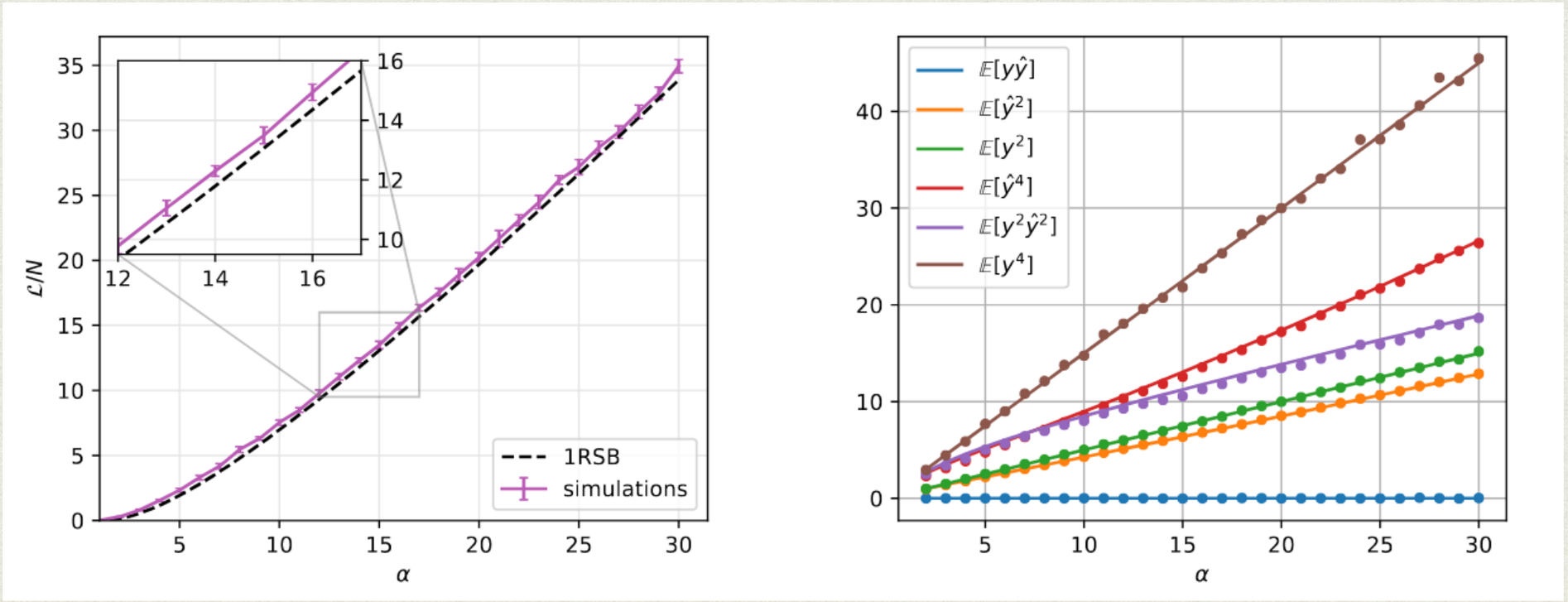


# TOWARDS A THEORY

Sarao Mannelli, Biroli, Cammarota, Krzakala, LZ, 2006.06997.

- Random matrix results (BBP-like) results from (Lu, Li'19) + marginality of threshold states = expression for $\alpha_{\rm GD}\left[P(\hat{y}, y)\right]$

- One-step replica symmetry breaking theory provides an approximation of $P_{\rm 1RSB}(\hat{y}, y)$



# TOWARDS A THEORY

Sarao Mannelli, Biroli, Cammarota, Krzakala, LZ, 2006.06997.

- Random matrix results (BBP-like) results from (Lu, Li'19) + marginality of threshold states = expression for $\alpha_{\rm GD}\left[P(\hat{y}, y)\right]$

- One-step replica symmetry breaking theory provides an approximation of $P_{\rm 1RSB}(\hat{y}, y)$

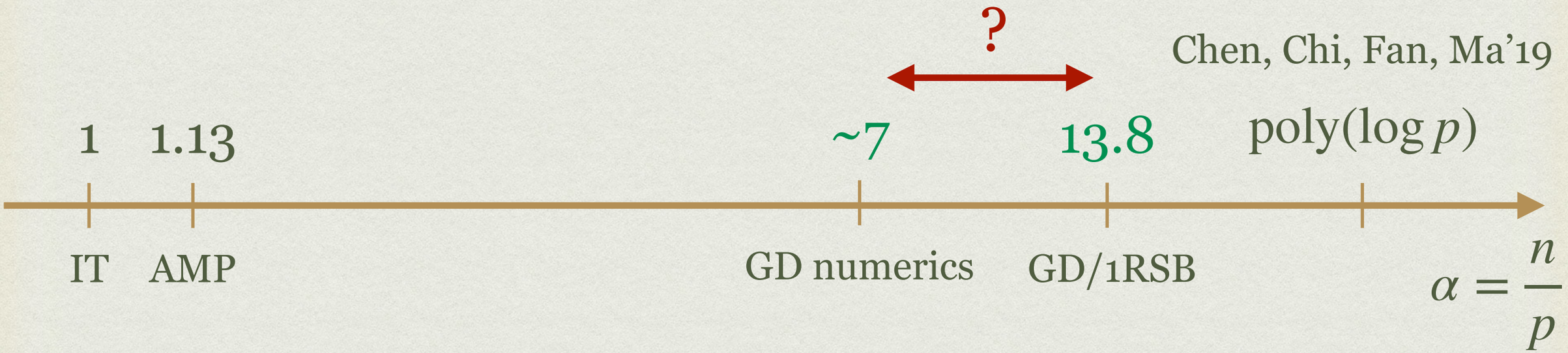
- Leading to

$$\alpha_{\rm GD}^{\rm 1RSB} \approx 13.8$$

# PERFORMANCE OF GRADIENT DESCENT

Sarao Mannelli, Biroli, Cammarota, Krzakala, LZ, 2006.06997.

How many samples needed for perfect generalization?




Only $O(p)$ samples seem to be needed. Precise constant?

# PERFORMANCE OF GRADIENT DESCENT

Closing the gap between GD and AMP?

?

Chen, Chi, Fan, Ma'19

1 1.13 ~7 13.8 poly(log $p$)

IT AMP GD numerics GD/1RSB

$\alpha = \frac{n}{p}$

# CONCLUSION ON SPIKED MATRIX-TENSOR MODEL

- First time we have a closed-form conjecture for the threshold of gradient-based algorithms including constants.
  Applicable for (simple) neural networks?

- Gradient flow worse than Langevin algorithm, both worse than AMP.
  How can GF & LA be improved to approach the AMP threshold?

- Gradient flow (sometimes) works even when spurious local minima are present. Quantified with the Kac-Rice approach.
  What about stochastic gradient descent?

# OVER-PARAMETRISATION & GRADIENT DESCENT

# PHASE RETRIEVAL

- Teacher-student setting with teacher having no hidden units, teacher's activation function is absolute value.

$$X_{\mu i} \sim \mathcal{N}(0,1/p) \qquad w_i^* \sim \mathcal{N}(0,1) \qquad \begin{matrix}\mu = 1,\ldots,n \\ i = 1,\ldots,p\end{matrix}$$

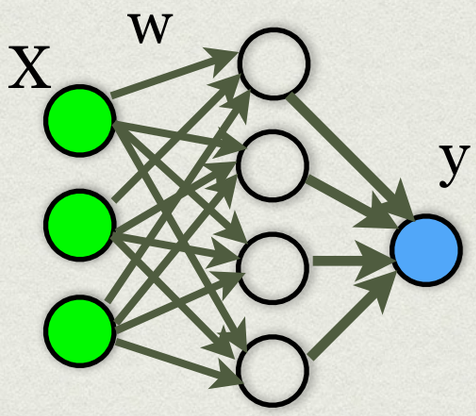
$$y_\mu = \left| \sum_{i=1}^{p} X_{\mu i} w_i^* \right|$$

Phase retrieval: Regression from training data $\{\mathbf{X}_\mu, y_\mu\}_{\mu=1}^{n}$

# GRADIENT DESCENT FOR PHASE RETRIEVAL

Loss function: $$\mathscr{L}(\{w_{ia}\}_{i,a=1}^{p,m}) = \sum_{\mu=1}^{n}\left[y_\mu^2 - \frac{1}{m}\sum_{a=1}^{m}\left(\sum_{i=1}^{p} X_{\mu i} w_{ia}\right)^2\right]^2$$

where $$y_\mu = \left|\sum_{i=1}^{p} X_{\mu i} w_i^*\right|$$




Wide (m>p) over-parametrised two-layer neural network

Gradient flow: $$\dot{w}_{ia}(t) = -\,\partial_{w_{ia}}\mathscr{L}\left(\{w_{jb}(t)\}_{j,b=1}^{p,m}\right)$$

Initialisation: $$w_{ia}(0) \sim \mathcal{N}(0,1)$$

# OVER-PARAMETRISED LANDSPACE

Sarao Mannelli, Vanden-Eijnden, LZ, 2006.15459

**Theorem 3.1** (Single unit teacher). *Consider a teacher with $m^* = 1$ and a student with $m \geq d$ hidden units respectively, so that $A^*$ has rank 1 and $A$ has full rank. Given a data set $\{\boldsymbol{x}_k\}_{k=1}^n$ with each $\boldsymbol{x}_k \in \mathbb{R}^d$ drawn independently from a standard Gaussian, denote by $\mathcal{M}_{n,d}$ the set of minimizer of the empirical loss constructed with $\{\boldsymbol{x}_k\}_{k=1}^n$ over symmetric positive semidefinite matrices $A$, i.e.*

$$\mathcal{M}_{n,d} = \left\{ A = A^T, \text{ positive semidefinite such that } E_n(A) = 0 \right\}. \tag{10}$$

*Set $n = \lfloor \alpha d \rfloor$ for $\alpha \geq 1$ and let $d \to \infty$. Then*

$$\lim_{d\to\infty} \mathbb{P}\left( \mathcal{M}_{\lfloor \alpha d \rfloor, d} \neq \{A^*\} \right) = 1 \qquad \text{if } \alpha \in [0, 2] \tag{11}$$

*whereas*

$$\lim_{d\to\infty} \mathbb{P}\left( \mathcal{M}_{\lfloor \alpha d \rfloor, d} = \{A^*\} \right) > 0 \qquad \text{if } \alpha \in (2, \infty). \tag{12}$$

$$A(t) = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{w}_i(t) \boldsymbol{w}_i^T(t), \quad A^* = \frac{1}{m^*} \sum_{i=1}^{m^*} \boldsymbol{w}_i^* (\boldsymbol{w}_i^*)^T,$$

p=d

# GD FOR OVER-PARAMETRISED PHASE RETRIEVAL

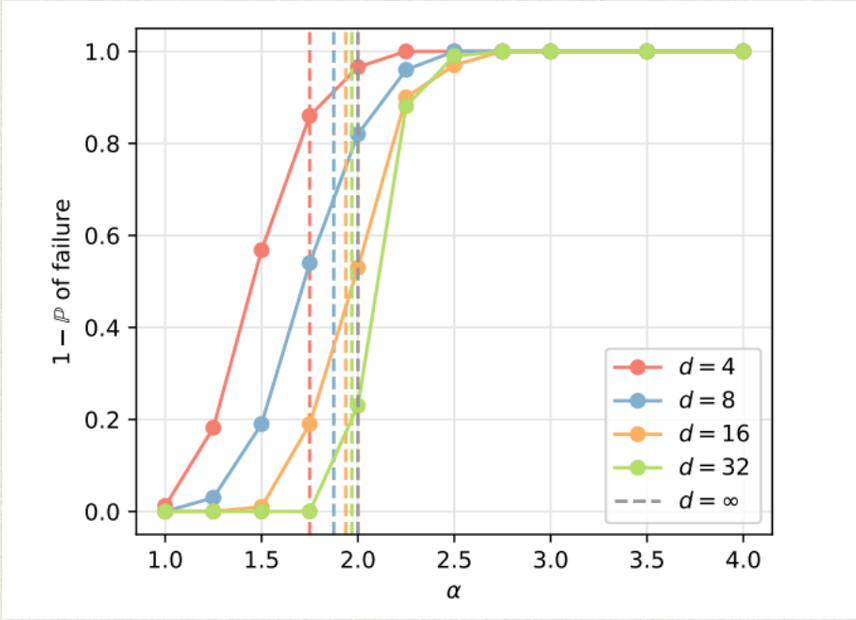
Sarao Mannelli, Vanden-Eijnden, LZ, 2006.15459

**Theorem 4.1.** *Let $\{\boldsymbol{w}_i(t)\}_{i=1}^m$ be the solution to (3) for the initial data $\{\boldsymbol{w}_i(0)\}_{i=1}^m$. Assume that $m \geq d$ and each $\boldsymbol{w}_i(0)$ is drawn independently from a distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. Then*

$$A = \frac{1}{m}\sum_{i=1}^m \boldsymbol{w}_i(t)\boldsymbol{w}_i^T(t) \to A_\infty = \frac{1}{m}\sum_{i=1}^m \boldsymbol{w}_i^\infty(\boldsymbol{w}_i^\infty)^T \quad as \;\; t\to\infty \tag{15}$$

*and $A_\infty$ is a global minimizer of the empirical loss, i.e.*

$$E_n(A_\infty) = 2L_n(\boldsymbol{w}_1^\infty, \dots, \boldsymbol{w}_n^\infty) = 0. \tag{16}$$

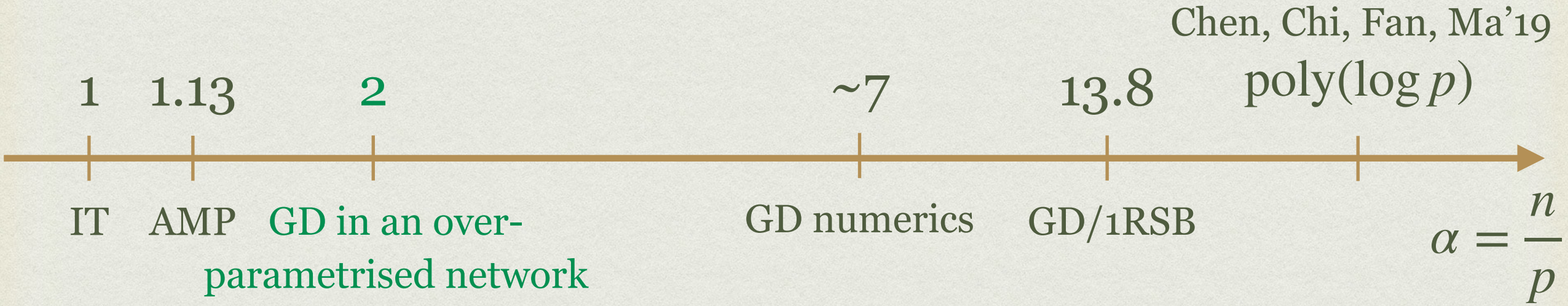



p=d

# PERFORMANCE OF GRADIENT DESCENT

Sarao Mannelli, Vanden-Eijnden, LZ, 2006.15459

Over-parametrised neural network needs fewer samples to learn



# CONCLUSION ON SPIKED MATRIX-TENSOR MODEL

- First time we have a closed-form conjecture for the threshold of gradient-based algorithms including constants.
  Applicable for (simple) neural networks?

- Gradient flow worse than Langevin algorithm, both worse than AMP.
  How can GF & LA be improved to approach the AMP threshold?

- Gradient flow (sometimes) works even when spurious local minima are present. Quantified with the Kac-Rice approach.
  What about stochastic gradient descent?

# STOCHASTIC GRADIENT DESCENT

$$\mathscr{L}(\{w_i\}_{i=1}^p) = \sum_{\mu=1}^{n} \ell(y_\mu, \mathbf{X}_\mu, \{w_i\}_{i=1}^p) + \lambda \|\mathbf{w}\|_2^2$$

$$w_j(t+\eta) = w_j(t) - \eta\left[\lambda w_j(t) + \partial_{w_j} \ell(y_\mu, \mathbf{X}_\mu, w(t))\right]$$

- Online SGD = each iteration uses samples never used before. Minimises directly the population loss, no notion of generalisation gap, i.e. train and test error are the same (in physics: Saad, Solla'95; Saad'09; Goldt, Advani, Saxe, Krzakala, Zdeborová'19)

- In practice: multi-pass SGD, reuses each sample many times. Much less existing theory ...

# CONTINUOUS TIME LIMIT?

$$w_j\left(t+\eta\right) = w_j(t) - \eta\left[\lambda w_j(t) + \sum_{\mu=1}^{n} s_\mu(t)\,\partial_{w_j}\ell(y_\mu, X_\mu, w(t))\right]$$

## SGD

$$s_\mu(t) = \begin{cases} 1 & \text{w.p. } b \\ 0 & \text{w.p. } 1-b \end{cases}$$

DISCRETE-TIME STOCHASTIC PROCESS

## Persistent-SGD

$$\text{Prob}\left(s_\mu\left(t+\eta\right)=1\,|\,s_\mu(t)=0\right) = \frac{\eta}{\tau}$$

PERSISTENCE TIME of each sample

$$\text{Prob}\left(s_\mu\left(t+\eta\right)=0\,|\,s_\mu(t)=1\right) = \frac{1-b}{b\,\tau}\eta$$

WELL-DEFINED CONTINUOUS LIMIT

stochastic gradient flow, $\eta \to 0$

$$\dot{w}_j(t) = -\,\eta\left[\lambda w_j(t) + \sum_{\mu=1}^{n} s_\mu(t)\,\partial_{w_j}\ell(y_\mu, X_\mu, w(t))\right] \qquad w_j(0) \sim \mathcal{N}(0,R)$$

$p, n \to \infty$ at fixed $\alpha = n/p, b, \tau$ batch size: $bn, 0 \le b \le 1$

# MODEL FOR DATA

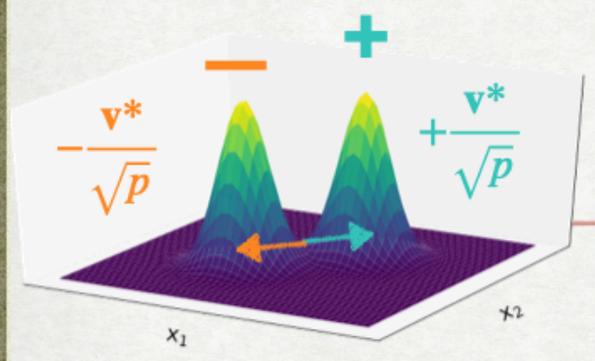
Binary classification of a Gaussian mixture:

Data $\mathbf{x}_\mu = c_\mu \dfrac{\mathbf{v}^*}{\sqrt{p}} + \sqrt{\Delta}\,\mathbf{z}_\mu \in \mathbb{R}^p \;,\; \mu = 1,...n$

$\mathbf{v}^* = (1,1,...1) \in \mathbb{R}^p,\; \mathbf{z}_\mu \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_p\right)$

Labels $\mathbf{y} = (y_1, \dots y_n)^\top \in \{-1, +1\}^n$

$\alpha = n/d$




## 2 clusters

$$c_\mu = \begin{cases} -1 & \text{w.p. } 1/2 \\ +1 & \text{w.p. } 1/2 \end{cases}$$

$$y_\mu = c_\mu$$

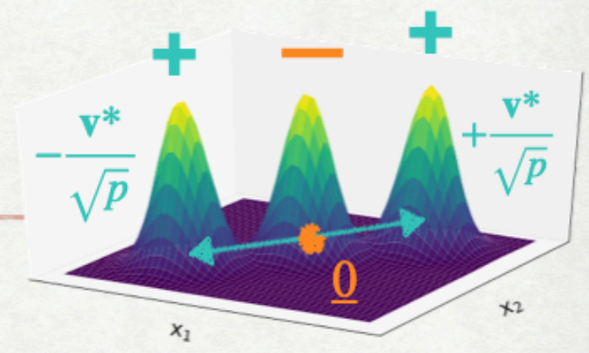
Regression

$$\hat{y}_\mu(\mathbf{w}) = \operatorname{sgn}\left[\frac{1}{\sqrt{p}}\mathbf{w}^\top \mathbf{x}_\mu\right]$$

## 3 clusters

$$c_\mu = \begin{cases} 0 & \text{w.p. } 1/2 \\ +1 & \text{w.p. } 1/4 \\ -1 & \text{w.p. } 1/4 \end{cases}$$

$$y_\mu = \begin{cases} -1 & \text{if } c_\mu = 0 \\ +1 & \text{if } c_\mu = \pm 1 \end{cases}$$




Regression

$$\hat{y}_\mu(\mathbf{w}) = \operatorname{sgn}\left[\left(\frac{1}{\sqrt{p}}\mathbf{w}^\top \mathbf{x}_\mu\right)^2 - L^2\right]$$

# DYNAMICAL MEAN-FIELD THEORY

(Mézard, Parisi, Virasoro, '87, Georges, Kotliar, Krauth, Rozenberg, '96)

**IOP** Publishing

Journal of Physics A: Mathematical and Theoretical

J. Phys. A: Math. Theor. **51** (2018) 085002 (36pp)

https://doi.org/10.1088/1751-8121/aaa68d

## Out-of-equilibrium dynamical mean-field equations for the perceptron model

**Elisabeth Agoritsas[1], Giulio Biroli[1,2], Pierfrancesco Urbani[2] and Francesco Zamponi[1]**

We generalize to the stochastic GD and data model with tests error well defined.

Markovian dynamics of a strongly coupled system of $p \to \infty$ degrees of freedom

DMFT →

Non-Markovian dynamics of one degree of freedom with memory

# DYNAMICAL MEAN-FIELD THEORY

## Effective (scalar) stochastic process for a typical "gap"

$h(t)$ related to $\mathbf{w}(t)^\top \mathbf{z}_\mu \big/ \sqrt{p}$

$$\partial_t h(t) = -\left(\lambda + \hat{\lambda}(t)\right) h(t) - \sqrt{\Delta}\, s(t)\, \Lambda'\left(y(c), r(t)\right) + \int_0^t \mathrm{d}t'\, M_R(t,t') h(t') + \xi(t)$$

$$\partial_t m(t) = -\lambda m(t) - \mu(t), \qquad m(0) = 0^+$$

→ deterministic equation for the "magnetisation"

$$m(t) = \mathbf{w}(t)^\top \mathbf{v}^* \big/ p$$

## Memory kernels & auxiliary functions

$$\mu(t) = \alpha \left\langle s(t)\left(c + \sqrt{\Delta} h_0\right) \Lambda'\left(y(c), r(t)\right) \right\rangle,$$

$$\hat{\lambda}(t) = \alpha\Delta \left\langle s(t) \Lambda''\left(y(c), r(t)\right) \right\rangle,$$

$$M_C(t,t') = \alpha\Delta \left\langle s(t) s(t') \Lambda'\left(y(c), r(t)\right) \Lambda'\left(y(c), r(t')\right) \right\rangle,$$

$$M_R(t,t') = \alpha\Delta \frac{\delta}{\delta Y(t')} \left\langle s(t) \Lambda'\left(y(c), r(t)\right) \right\rangle \Big|_{Y=0}$$

$$r(t) = \sqrt{\Delta}\, h(t) + m(t)\left(c + \sqrt{\Delta}\, h_0\right)$$

$$\langle \xi(t) \rangle = 0, \quad \langle \xi(t)\, \xi(t') \rangle = M_C(t,t')$$

The stochastic process must be solved self-consistently (Eissfeller, Opper '92)

# DYNAMICAL MEAN-FIELD THEORY

Correlation $C(t,t') = \frac{1}{p}\mathbf{w}(t)^\top \mathbf{w}(t')$ and response $R(t,t') = \frac{1}{p}\sum_{i=1}^{p} \frac{\delta \mathrm{w}_i(t)}{\delta H_i(t')}$ functions:

$$\partial_t C(t',t) = -\left(\lambda + \hat{\lambda}(t)\right) C(t,t') + \int_0^t \mathrm{d}s\, M_R(t,s) C(t',s) + \int_0^{t'} \mathrm{d}s\, M_C(t,s) R(t',s) - m(t')\left(\int_0^t \mathrm{d}s\, M_R(t,s) m(s) + \mu(t) - \hat{\lambda}(t) m(t)\right) \quad \text{if } t \neq t',$$

$$\frac{1}{2}\partial_t C(t,t) = -\left(\lambda + \hat{\lambda}(t)\right) C(t,t) + \int_0^t \mathrm{d}s\, M_R(t,s) C(t,s) + \int_0^t \mathrm{d}s\, M_C(t,s) R(t,s) - m(t)\left(\int_0^t \mathrm{d}s\, M_R(t,s) m(s) + \mu(t) - \hat{\lambda}(t) m(t)\right),$$

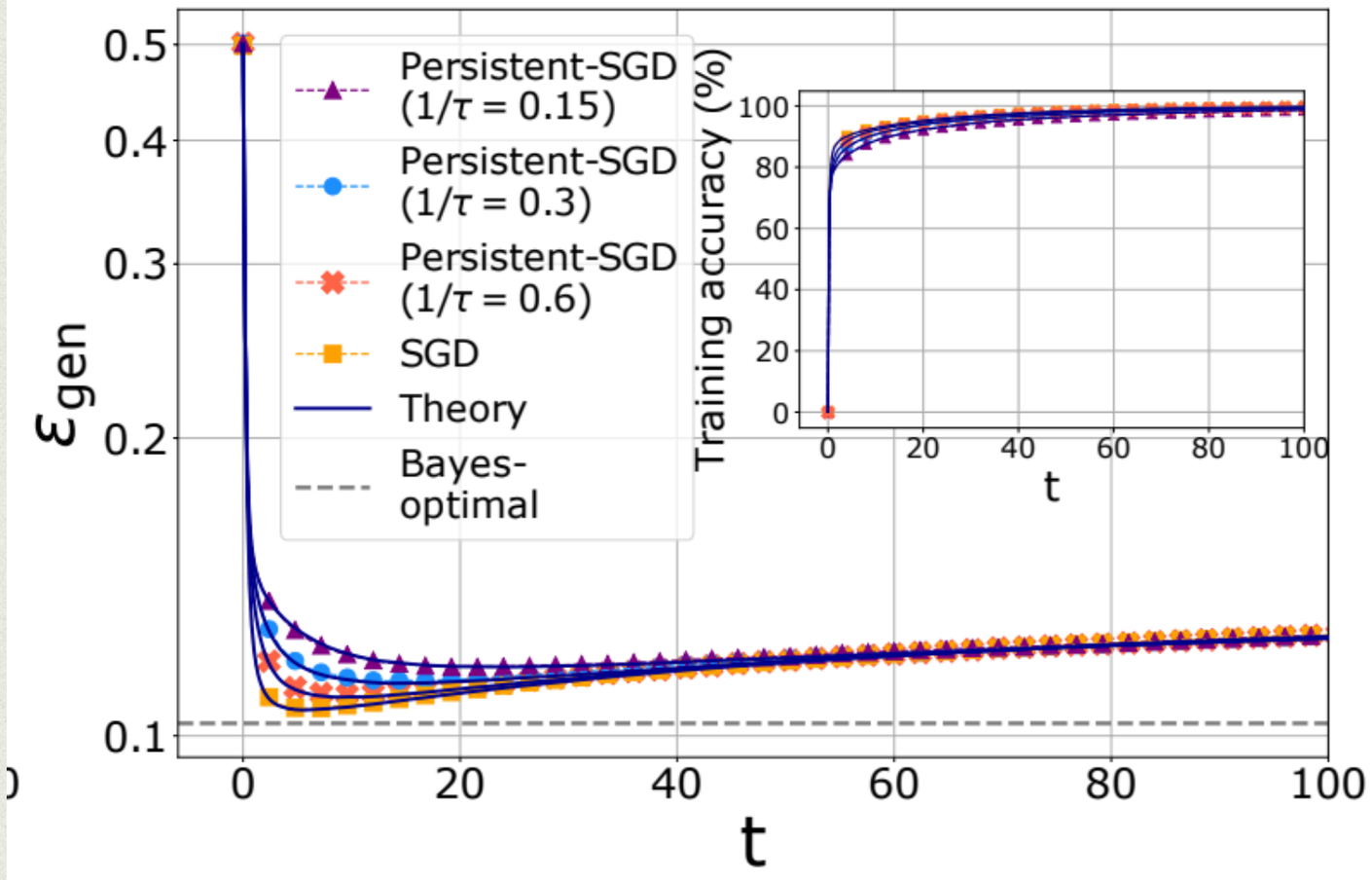
$$\partial_t R(t',t) = -\left(\lambda + \hat{\lambda}(t)\right) R(t,t') + \delta(t-t') + \int_{t'}^t \mathrm{d}s\, M_R(t,s) R(s,t').$$

Training loss: $e(t) = \alpha \left\langle \ell\left(y\Phi\left(r(t)\right)\right)\right\rangle$ Training accuracy: $a(t) = 1 - \left\langle \theta\left(-y\Phi\left(r(t)\right)\right)\right\rangle$

Generalisation error: $\varepsilon_{\text{gen}}(t) \equiv \frac{1}{4}\mathbb{E}_{\mathbf{X},\mathbf{y},\mathbf{x}_{\text{new}},y_{\text{new}}}\left[\left(\hat{y}_{\text{new}} - y_{\text{new}}\right)^2\right] = \begin{cases} \frac{1}{2}\operatorname{erfc}\left(\frac{m(t)}{\sqrt{2\Delta\, C(t,t)}}\right) \\ \frac{1}{2}\operatorname{erfc}\left(\frac{L}{\sqrt{2\Delta\, C(t,t)}}\right) + \frac{1}{4}\left(\operatorname{erf}\left(\frac{L-m(t)}{\sqrt{2\Delta\, C(t,t)}}\right) + \operatorname{erf}\left(\frac{L+m(t)}{\sqrt{2\Delta\, C(t,t)}}\right)\right) \end{cases}$

# DMFT FOLLOWS THE WHOLE TRAJECTORY

## 2-clusters



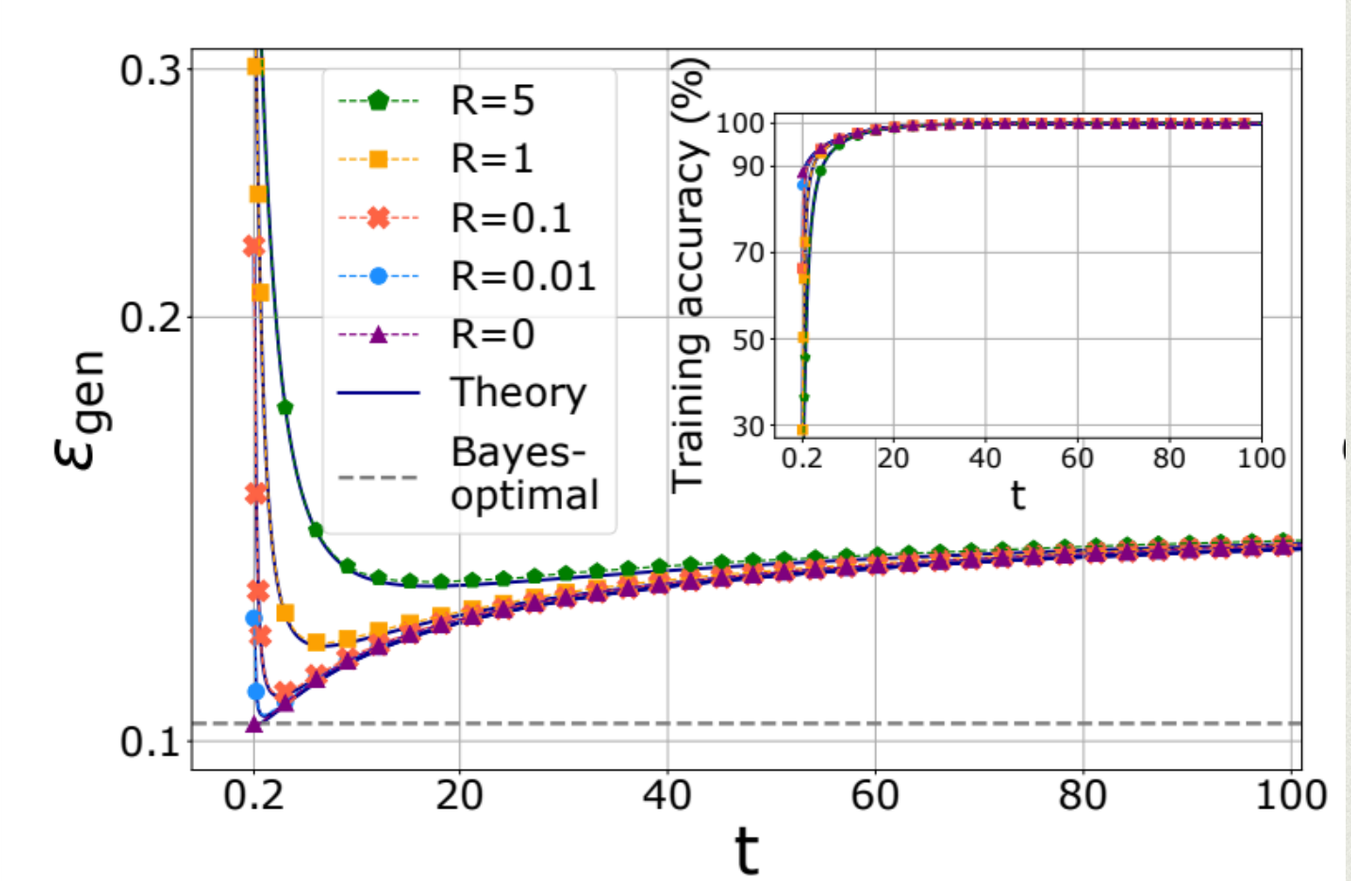

Small persistence time → SGD

SGD-inspired discretisation is ad hoc, yet agrees with simulations

$b = 0.3, \alpha = 2.0, \Delta = 0.5, \lambda = 0, \eta = 0.2, R = 0.01$

# DMFT FOLLOWS THE WHOLE TRAJECTORY

## 2-clusters, full-batch




R = variance at initialization

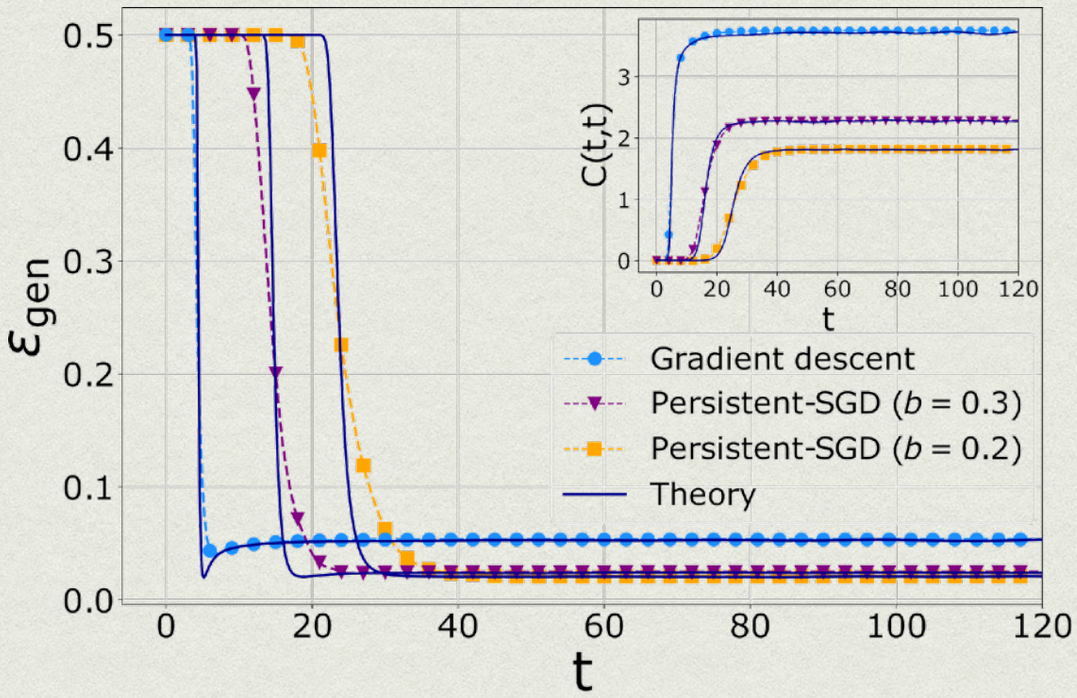
For 2-clusters R=0 is optimal after one iteration.

$b = 1, \alpha = 2.0, \Delta = 0.5, \lambda = 0, \eta = 0.2$

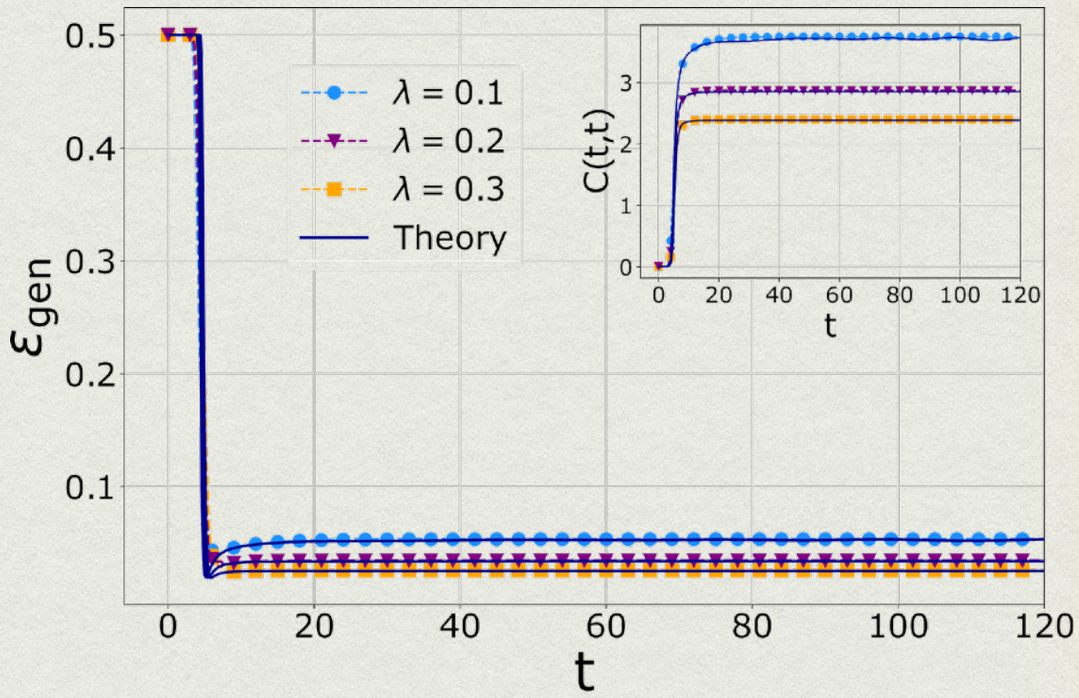
# DMFT FOLLOWS THE WHOLE TRAJECTORY

## 3-clusters

$\lambda = 0.1$




$b = 1$




$\alpha = 3.0, \Delta = 0.05, L = 0.7, R = 0.01, \eta = 0.2$

The finite batch size acts as an effective regularisation.

# CONCLUSION

- DMFT tracks the trajectory of GD/SGD for a range of data models.
- Extensions:
  - Deduce more insights from the DMFT equations (optimal hyper-parameter setting, nature of noise …)
  - Other data models, networks with hidden units, variants of GD/SGD.
  - Rigorous proof of the equations/thresholds.

# REFERENCES FOR THIS TALK



- Sarao Mannelli, Biroli, Cammarota, Krzakala, Urbani, LZ; *Marvels and Pitfalls of the Langevin Algorithm in Noisy High-dimensional Inference;* Phys. Rev. X'20, arXiv:1812.09066
- Sarao Mannelli, Krzakala, Urbani, LZ; *Passed & Spurious: Descent Algorithms and Local Minima in Spiked Matrix-Tensor Models;* ICML'19, arXiv:1902.00139.
- Sarao Mannelli, Biroli, Cammarota, Krzakala, LZ; *Who is Afraid of Big Bad Minima? Analysis of Gradient-Flow in a Spiked Matrix-Tensor Model*; NeurIPS'19, arXiv:1907.08226.

- Sarao Mannelli, Biroli, Cammarota, Krzakala, Urbani, LZ; *Complex Dynamics and Simple Neural Networks: Understanding Gradient Flow in Phase Retrieval,* arXiv:2006.06997.
- Sarao Mannelli, Vanden-Eijnden, LZ; *Landscape and Dynamics in Over-Parametrized Neural Networks with Quadratic Activation*; arXiv:2006.15459.

- Mignacco, Urbani, Krzakala, LZ; *Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification*; arXiv:2006.06098.