ALGORITHMIC BARRIERS IN HIGH-DIMENSIONAL NON-CONVEX LANDSCAPES

Lenka Zdeborová

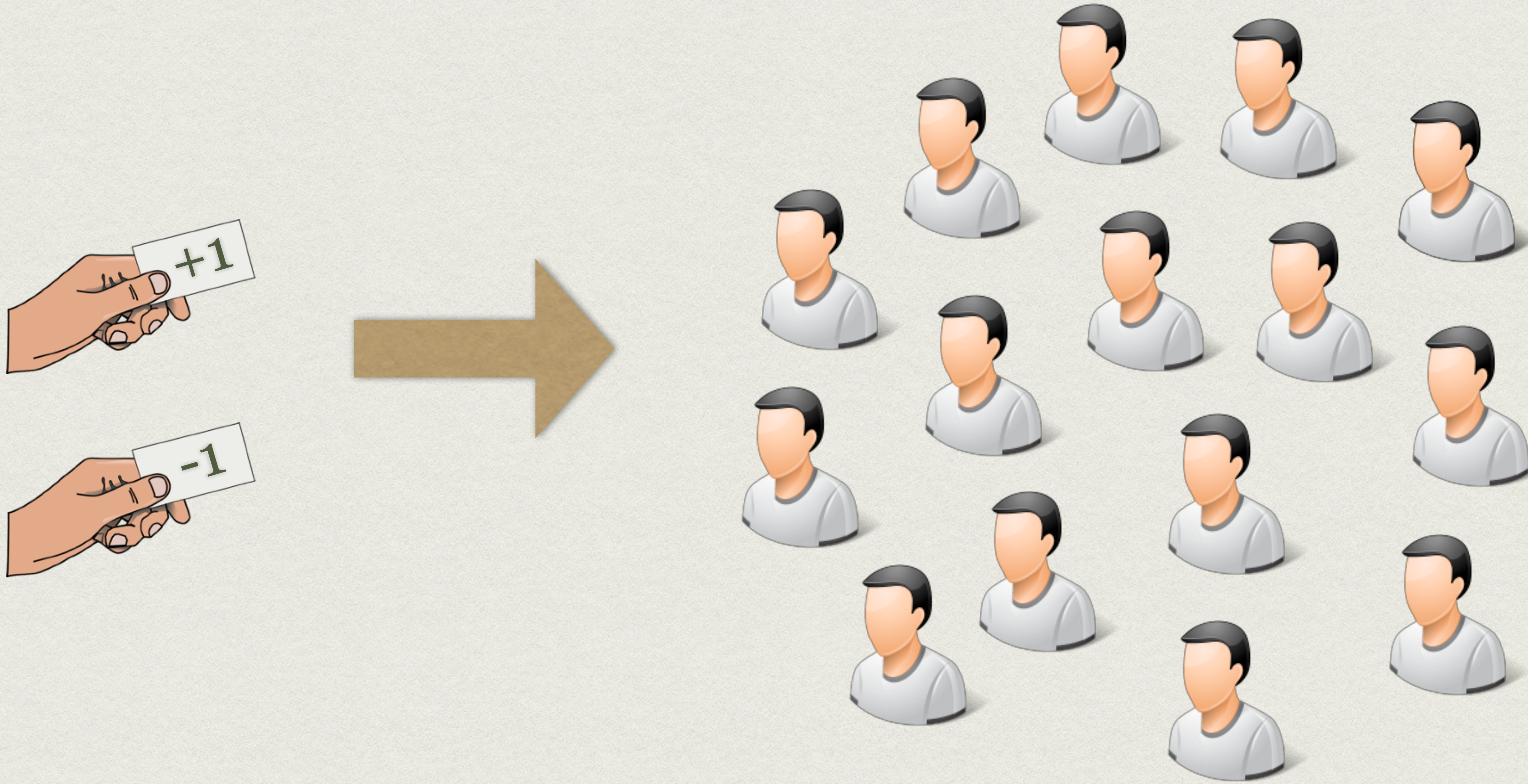
(IPhT, CEA Saclay, France)

with F. Antenucci, G. Biroli, C. Cammarota, S. Franz, T. Lesieur, F. Krzakala, S. Sarao Mannelli, P. Urbani.

Which high-dimensional inference problems (leading to non-convex objectives) are solvable (close to) optimally with tractable algorithms?

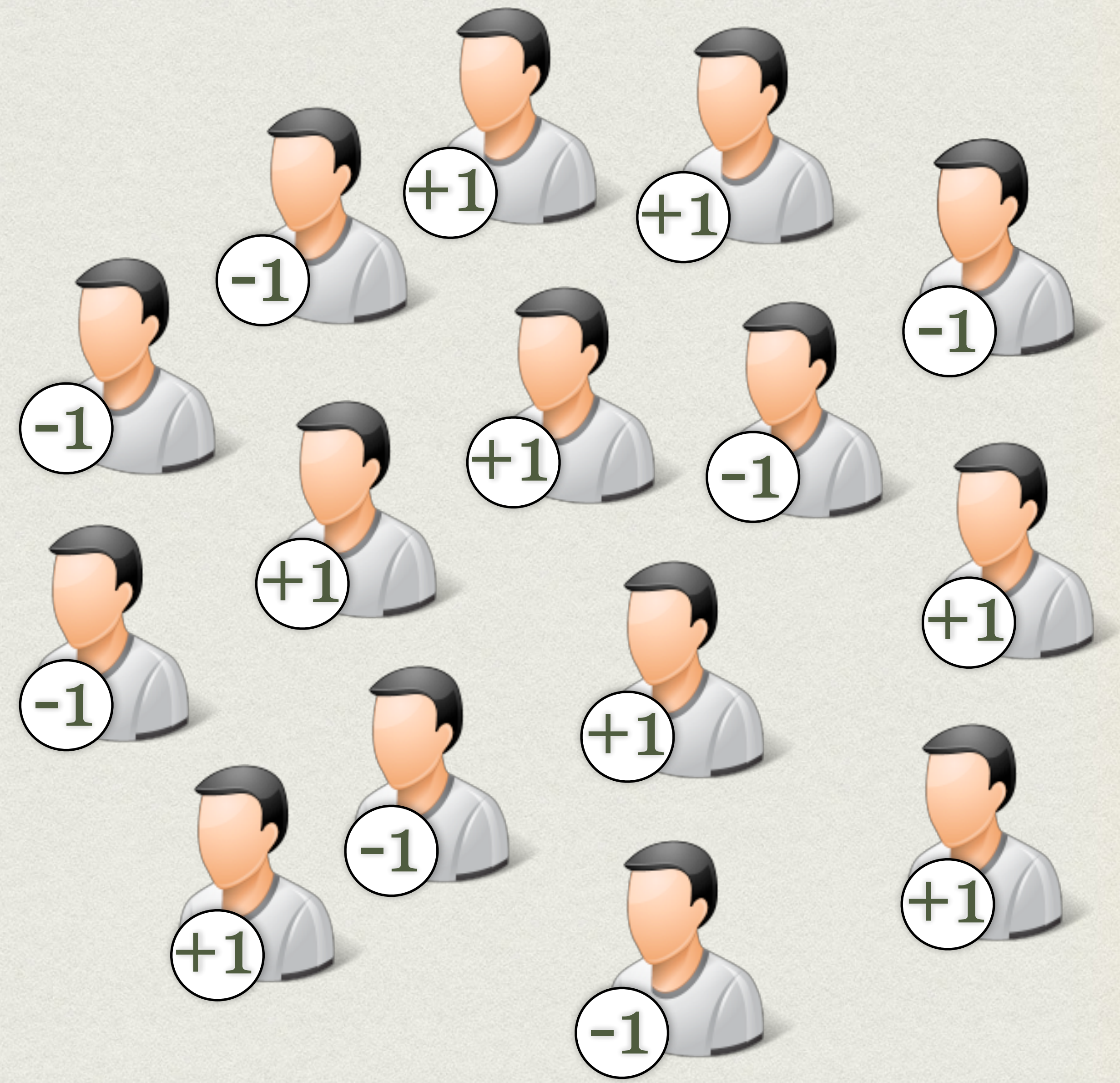
Which algorithms?

TYPICAL EXAMPLE

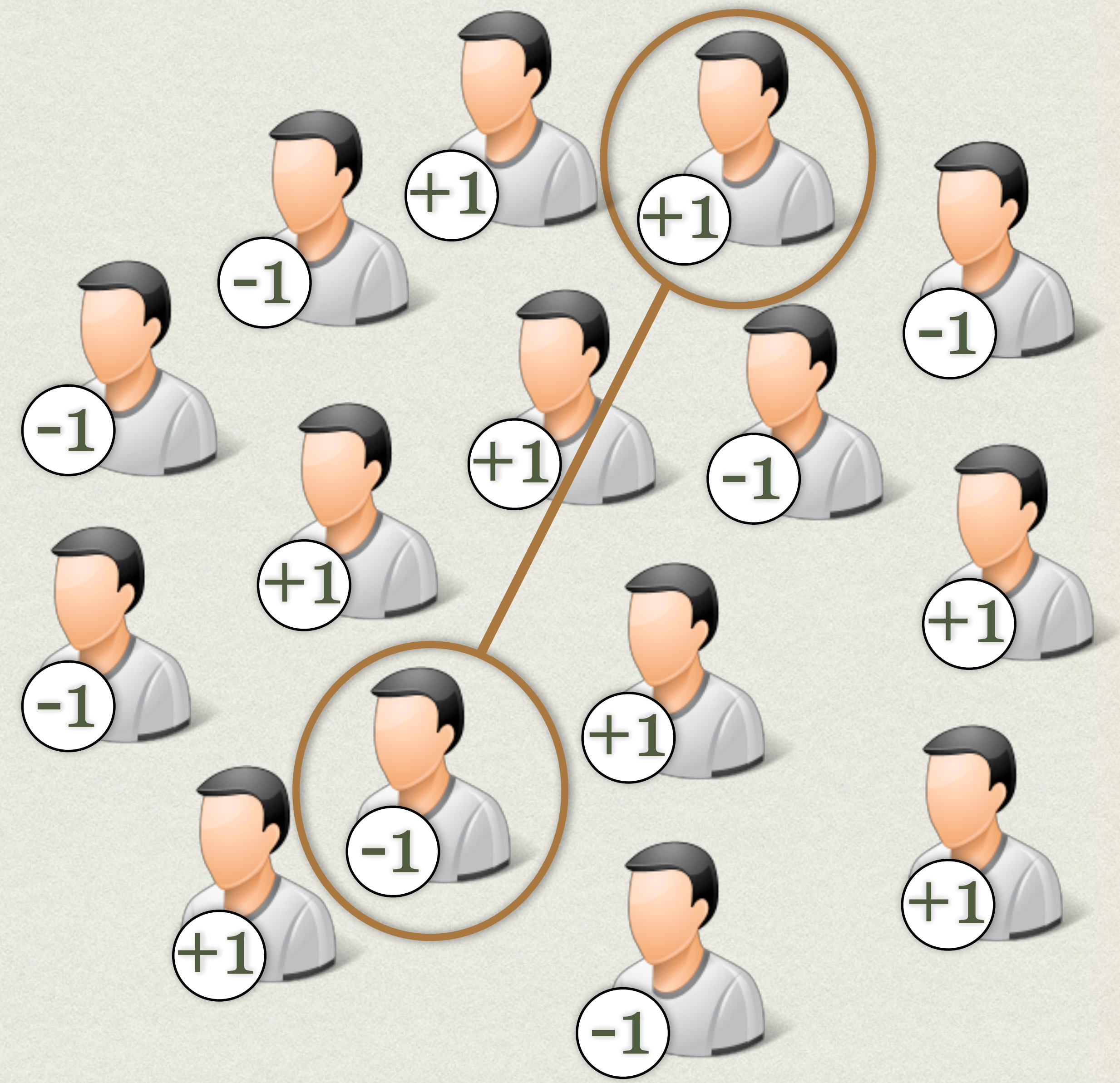


N=15 people

TYPICAL EXAMPLE



TYPICAL EXAMPLE



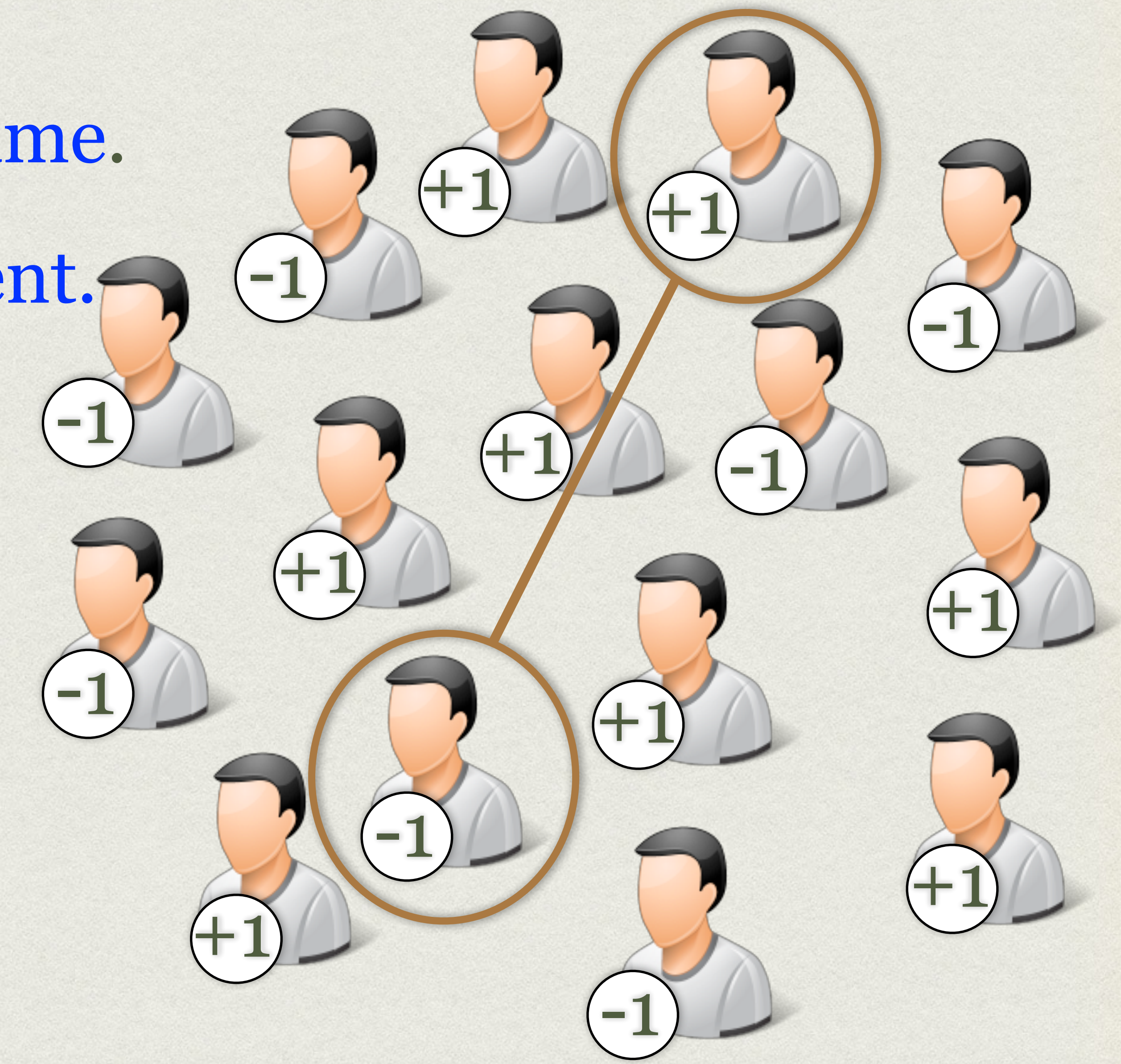
TYPICAL EXAMPLE

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

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

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

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



HOW TO SOLVE THIS?

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

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

$Z_{ij} \sim \mathcal{N}(0, \Delta)$ $\quad x_i^* \in \{-1, +1\}$

Eigen-decomposition of Y (aka PCA) minimises

$$\sum_{i<j} (Y_{ij} - \hat{Y}_{ij})^2 \quad \text{with } \mathrm{rank}(\hat{Y}) = 1$$

x_{PCA} (leading eigen-vector of Y) estimates x* (up to a sign).

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

$\Delta < 1 \qquad |x_{\text{PCA}} \cdot x^*| > 0$

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

BAYESIAN INFERENCE

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

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

Posterior distribution:

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

Bayes-optimal inference = computation of marginals
(argmax maximizes the number of correctly assigned values,
mean of marginals minimises the mean-squared error).

Physics: Sherrington-Kirkpatrick model with planted-disorder.

BAYESIAN INFERENCE

Values of cards: $x_i \sim P_X(x_i)$

Posterior distribution:

$$P(x \mid Y) = \frac{1}{Z(Y, \Delta)} \prod_{i=1}^{N} P_X(x_i) \prod_{i<j} e^{-\frac{1}{2\Delta}\left(Y_{ij} - x_i x_j/\sqrt{N}\right)^2}$$

Bayes-optimal inference = computation of marginals
(argmax maximizes the number of correctly assigned values, mean of marginals minimises the mean-squared error).

PROPERTIES OF THE BAYES-OPTIMAL ESTIMATOR

Theorem 1:

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

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

$m \in \mathbb{R}$

$x \sim P_X$

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

= replica symmetric free entropy

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

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

Proofs: +1/-1 Korada, Macris'10; generic: Krzakala, Xu, LZ, ITW'16, Barbier, Dia, Macris, Krzakala, Lesieur, LZ'16 & 18; simpler: Lelarge, Miolane'16; El-Alaoui, Krzakala'17

PROPERTIES OF THE BAYES-OPTIMAL ESTIMATOR

Theorem 1:

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

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

$m \in \mathbb{R}$

$x \sim P_X$

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

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

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

Proofs: +1/-1 Korada, Macris'10; generic: Krzakala, Xu, LZ, ITW'16, Barbier, Dia, Macris, Krzakala, Lesieur, LZ'16 & 18; simpler: Lelarge, Miolane'16; El-Alaoui, Krzakala'17

Can tractable algorithms achieve the Bayes-optimal error?

APPROXIMATE MESSAGE PASSING

AMP algorithm estimates means and variances of the marginals:

$$a_i^{t+1} = f(A^t, B_i^t)$$

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

$$A^t = \frac{1}{N\Delta} \sum_{l=1}^{N} \left(a_l^t\right)^2$$

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

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

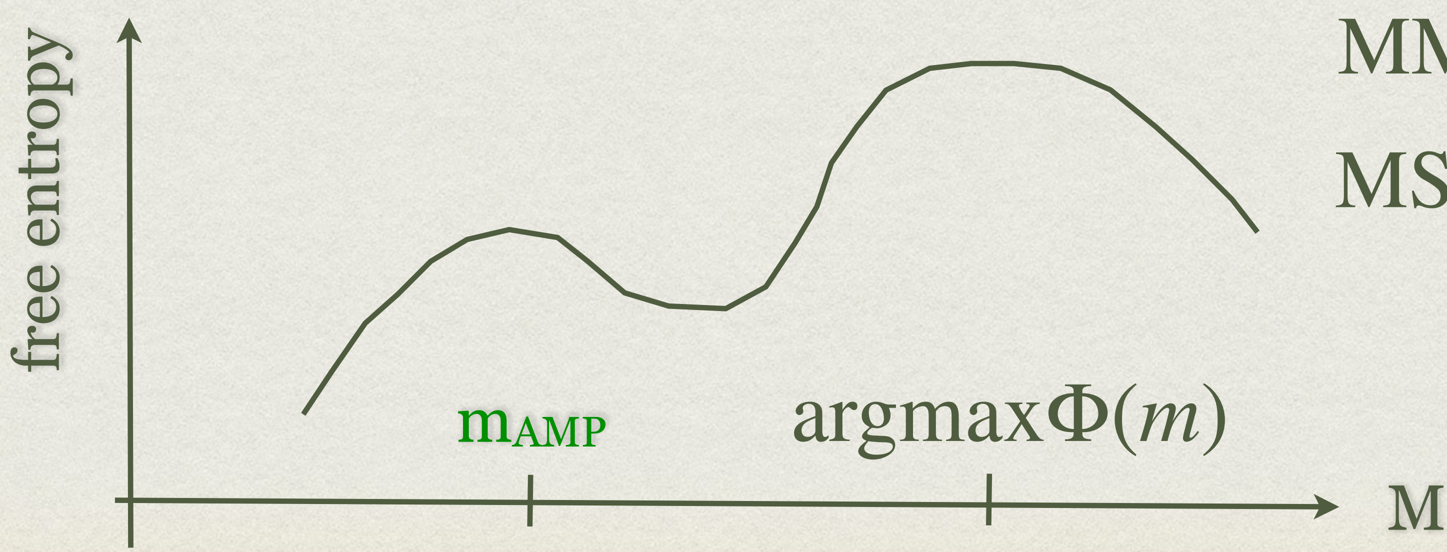
$$\mathscr{P}(x; A, B) = \frac{1}{\mathscr{Z}(A, B)} P_X(x) e^{Bx - Ax^2/2} \qquad f(A, B) = \mathbb{E}_{\mathscr{P}}(x)$$

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

STATE EVOLUTION

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

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



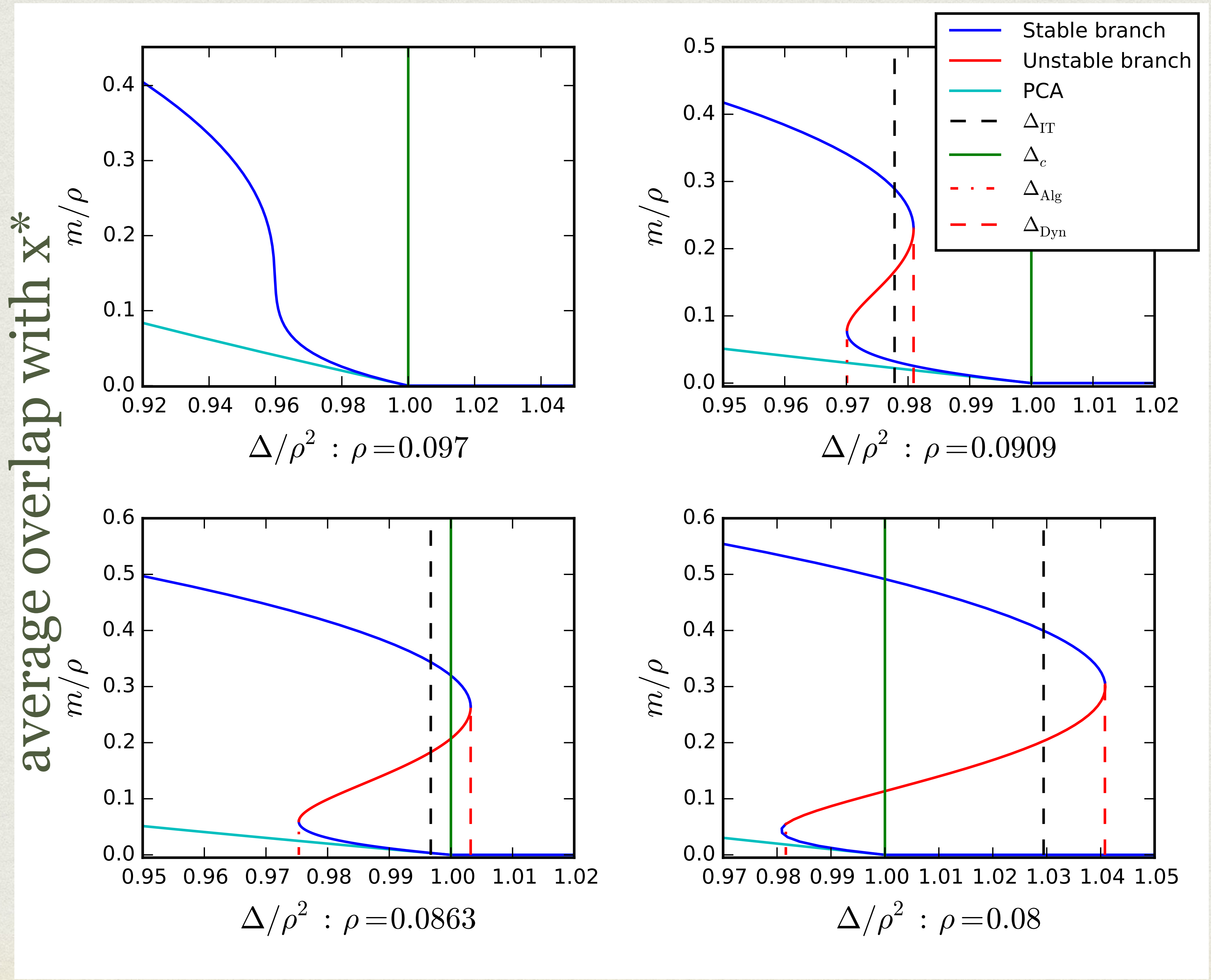
$$\text{MMSE} = \mathbb{E}_{P_X}(x^2) - \text{argmax}\,\Phi(m)$$

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

From fixed points to phase transitions:

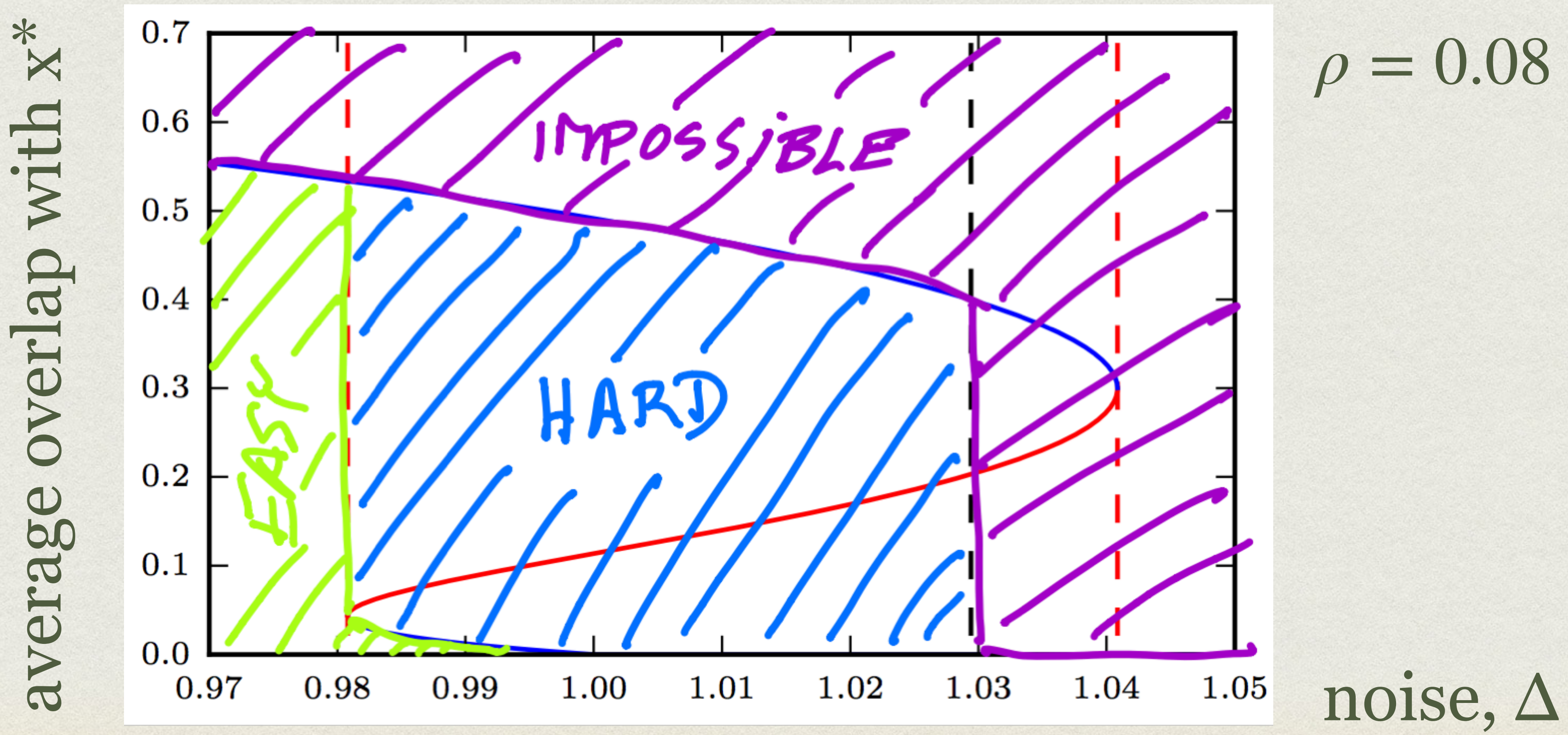
Lesieur, Krzakala, LZ'17

Sparse PCA: $P_X(x_i) = \frac{\rho}{2}\left[\delta(x_i - 1) + \delta(x_i + 1)\right] + (1-\rho)\delta(x_i)$



ALGORITHMIC INTERPRETATION

- Easy by approximate message passing.
- Impossible information theoretically.
- Hard phase: in presence of a *first order phase transition*.



$\rho = 0.08$

HARD PHASE

Hard phase = spinodal region of first order phase transitions.

Algorithmic threshold shared by spectral methods and SDPs.

Conjecture:

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

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

Hard phase identified in:

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

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

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

LANDSCAPE OF THE HARD PHASE

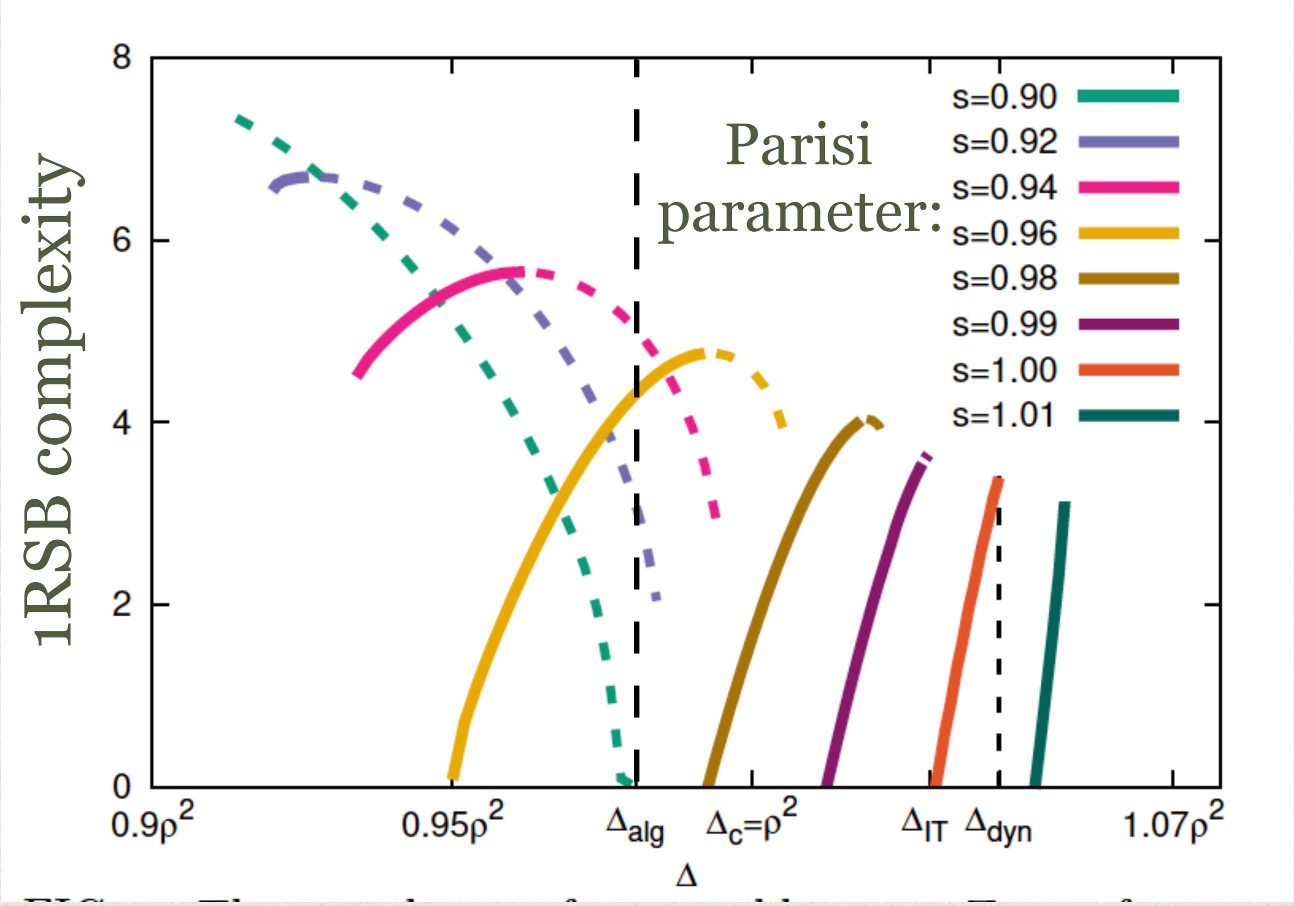
What are the properties of the Gibbs measure, in the hard phase and around, conditioned not to be close to the ground-truth x^*?

GLASSY NATURE OF THE HARD PHASE

Antenucci, Franz, Urbani, LZ, Phys. Rev. X'19

- Analyzed by 1-step replica symmetry breaking (see Mezard's tutorial).
  - The hard phase is glassy - many spurious local minima potentially blocking the dynamics.
  - The glassiness extends even below the algorithmic threshold.

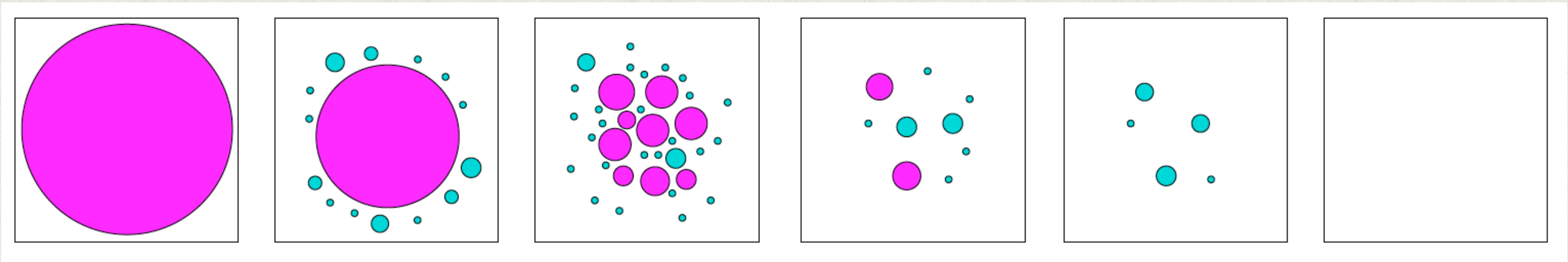
$\rho = 0.08$



SURVEY PROPAGATION

Mezard, Parisi, Zecchina, Science'02

as in Mezard's tutorial



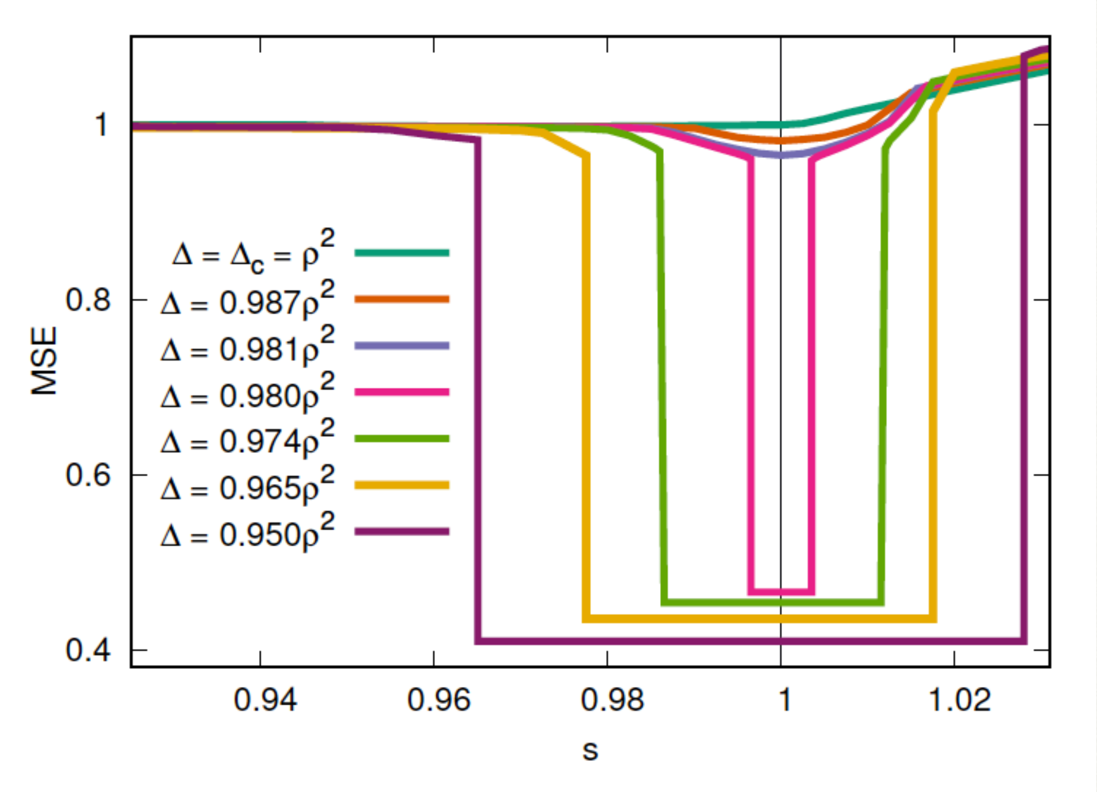
- Developed for the k-satisfiability problem
- Algorithm that takes into account the glassiness (1RSB structure).
- Provides large algorithmic improvement in K-SAT. State-of-the art on random K-SAT still today.
- ☆ Can this provide improvement in the inference-hard-phase?

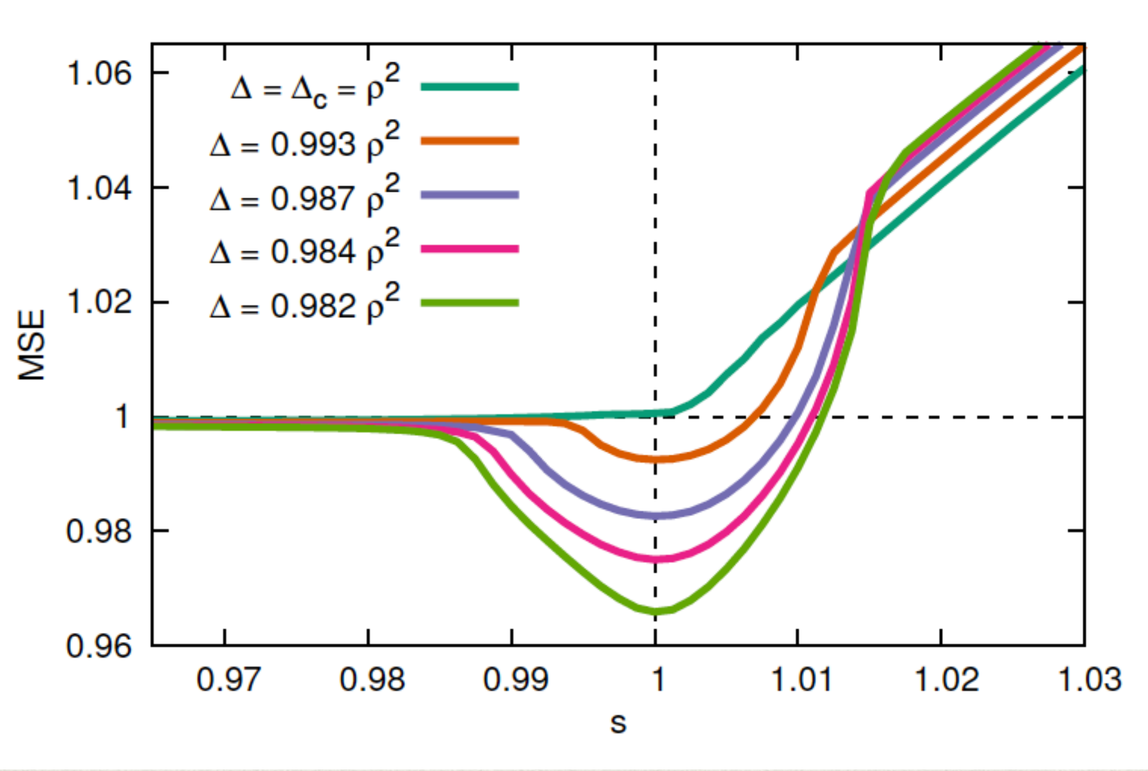
APPROXIMATE SURVEY PROPAGATION

Antenucci, Krzakala Urbani, LZ, arXiv:1807.01296

Message passing algorithm with state evolution = 1RSB fixed point equations for the Parisi parameter s.

$\rho = 0.08$





Result n. 1: ASP never better than Bayes-optimal-AMP.

Physically a mystery. Mathematically follows from proofs about optimality of AMP's denoising function (Deshpande, Montanari'13)

GLASSY NATURE OF THE HARD PHASE

Antenucci, Franz, Urbani, LZ, Phys. Rev. X'19

Result n.2: Residual glassiness below the algorithmic threshold. = Strong yet indirect indication of trouble for Gibbs-sampling or gradient based algorithms.

How to confirm this?

- Numerically — work in progress by Ricci-Tersenghi et al.
- Analytically — Gibbs samplers and gradient descents are much harder to analyse than message passing .... let's try anyway!

SEEKED INGREDIENTS OF THE MODEL

- Kind of spherical spin glass so that Langeving dynamics solvable via Crisanti-Horner-Sommers-Cugliandolo-Kurchan'93 equations (see Cugliandolo's tutorial on Thursday).
- Inference model with a AMP-hard phase.
- Model where AMP conjectured optimal, i.e. algorithmic threshold of the same order as the information theoretic (excludes spiked tensor model).

MIXED SPIKED MATRIX-TENSOR MODEL

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

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

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

- Bayes-optimal estimation = marginals for Hamiltonian

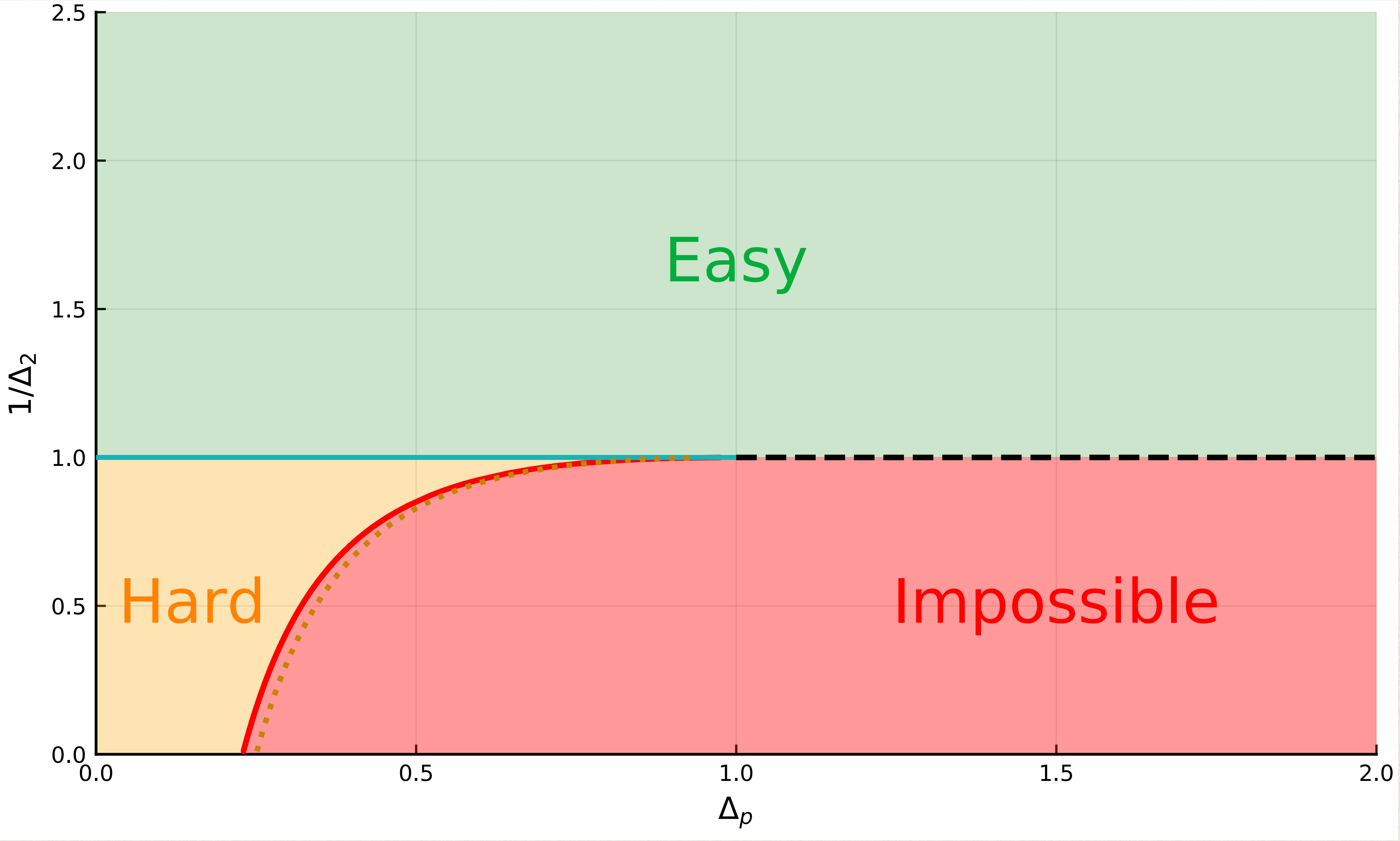
$$\mathcal{H}(x) = -\frac{1}{\Delta_2 \sqrt{N}} \sum_{i<j} Y_{ij} x_i x_j - \frac{\sqrt{(p-1)!}}{\Delta_p N^{(p-1)/2}} \sum_{i_1 < \dots < i_p} T_{i_1 \dots i_p} x_{i_1} \dots x_{i_p}$$

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

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

PHASE DIAGRAM P=3

Bayes-optimal performance and AMP



LANGEVIN ALGORITHM

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

spherical constraint

T=1 noise

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

gradient

At large time (exponentially) samples the posterior measure.

Where does it go in large constant time?

LANGEVIN STATE EVOLUTION

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

$$
\frac{\partial}{\partial t}C(t,t') = 2R(t',t) - \mu(t)C(t,t') + Q'(\overline{C}(t))\overline{C}(t') + \int_0^t dt'' R(t,t'')Q''(C(t,t''))C(t',t'') + \int_0^{t'} dt'' R(t',t'')Q'(C(t,t''))
$$

$$
\frac{\partial}{\partial t}R(t,t') = \delta(t-t') - \mu(t)R(t,t') + \int_{t'}^t dt'' R(t,t'')Q''(C(t,t''))R(t'',t')\,,
$$

$$
\frac{\partial}{\partial t}\overline{C}(t) = -\mu(t)\overline{C}(t) + Q'(\overline{C}(t)) + \int_0^t dt'' R(t,t'')\overline{C}(t'')Q(C(t,t''))\,,
$$

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

Generalization of the CHSCK equations to include the spike x*.

LANGEVIN STATE EVOLUTION

$$\frac{\partial}{\partial t}C(t,t') = 2R(t',t) - \mu(t)C(t,t') + Q'(\overline{C}(t))\overline{C}(t') + \int_0^t dt'' R(t,t'')Q''(C(t,t''))C(t',t'') + \int_0^{t'} dt'' R(t',t'')Q'(C(t,t''))$$

$$\frac{\partial}{\partial t}R(t,t') = \delta(t-t') - \mu(t)R(t,t') + \int_{t'}^t dt'' R(t,t'')Q''(C(t,t''))R(t'',t'),$$

$$\frac{\partial}{\partial t}\overline{C}(t) = -\mu(t)\overline{C}(t) + Q'(\overline{C}(t)) + \int_0^t dt'' R(t,t'')\overline{C}(t'')Q(C(t,t'')),$$

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

Generalization of the CHSCK equations to include the spike x*.

Without spike:

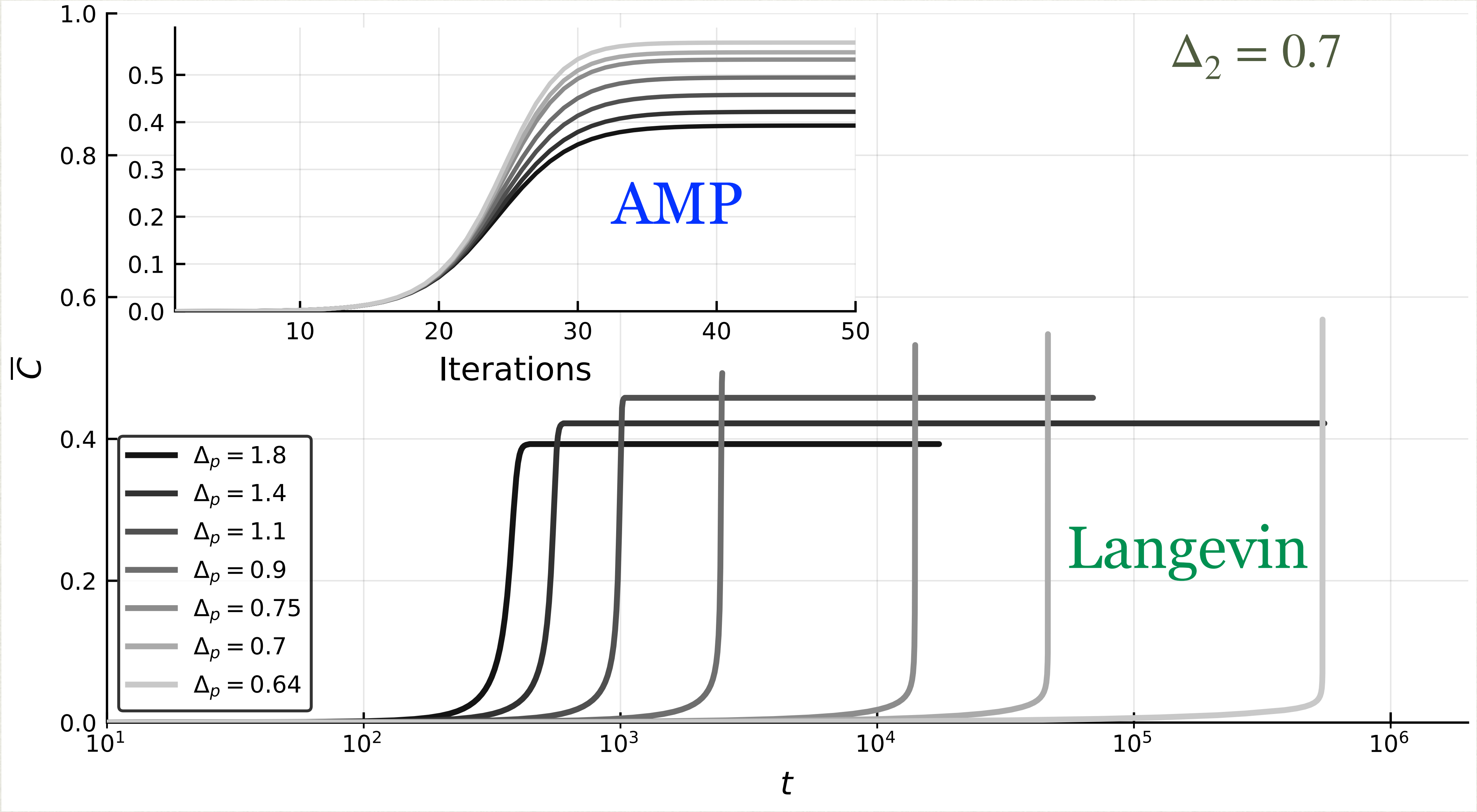
See Cugliandolo's tutorial & Ricci-Tersenghi's talk on Thursday!

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

(proof with spike: let's work on it?)

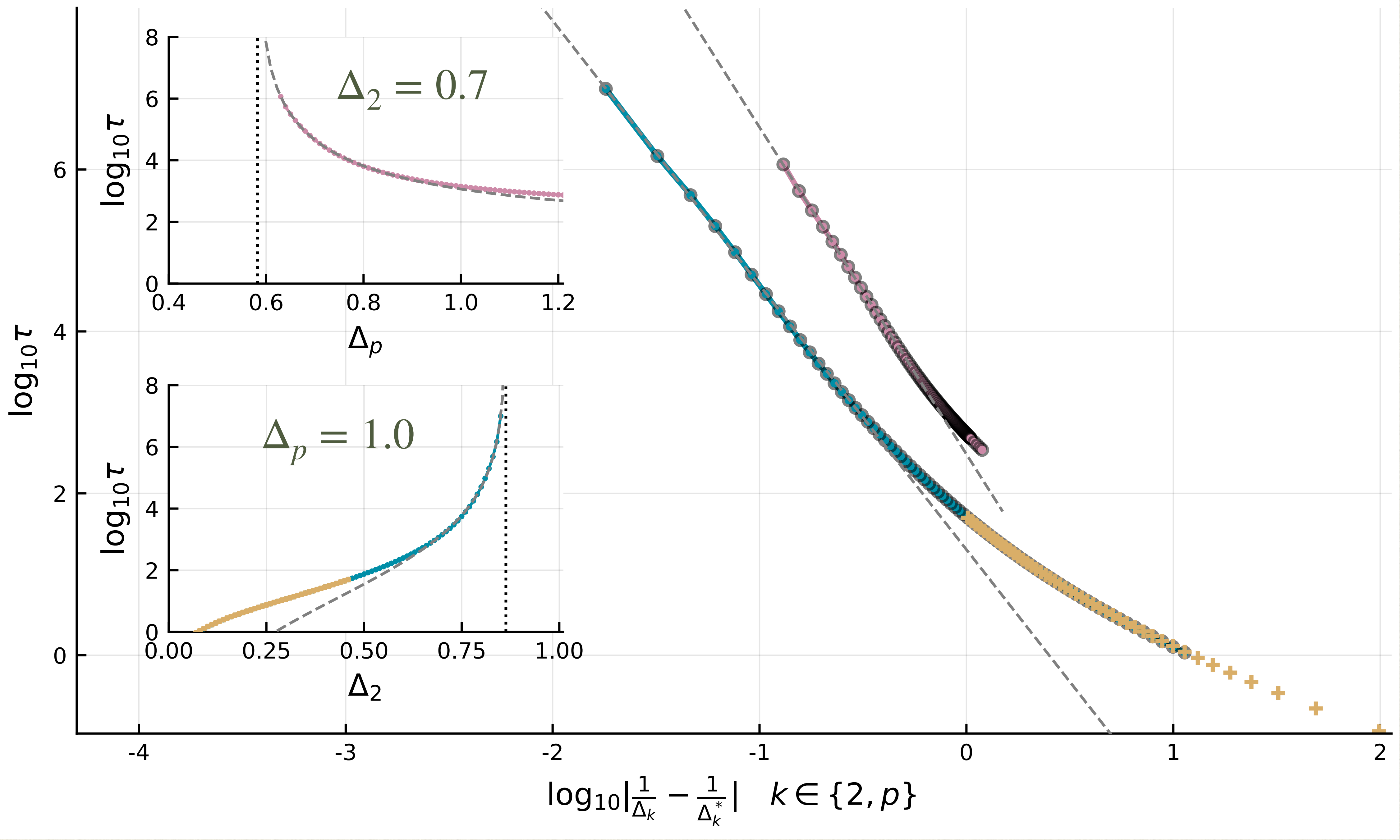
LANGEVIN STATE EVOLUTION (NUMERICAL SOLUTION)

correlation with ground truth

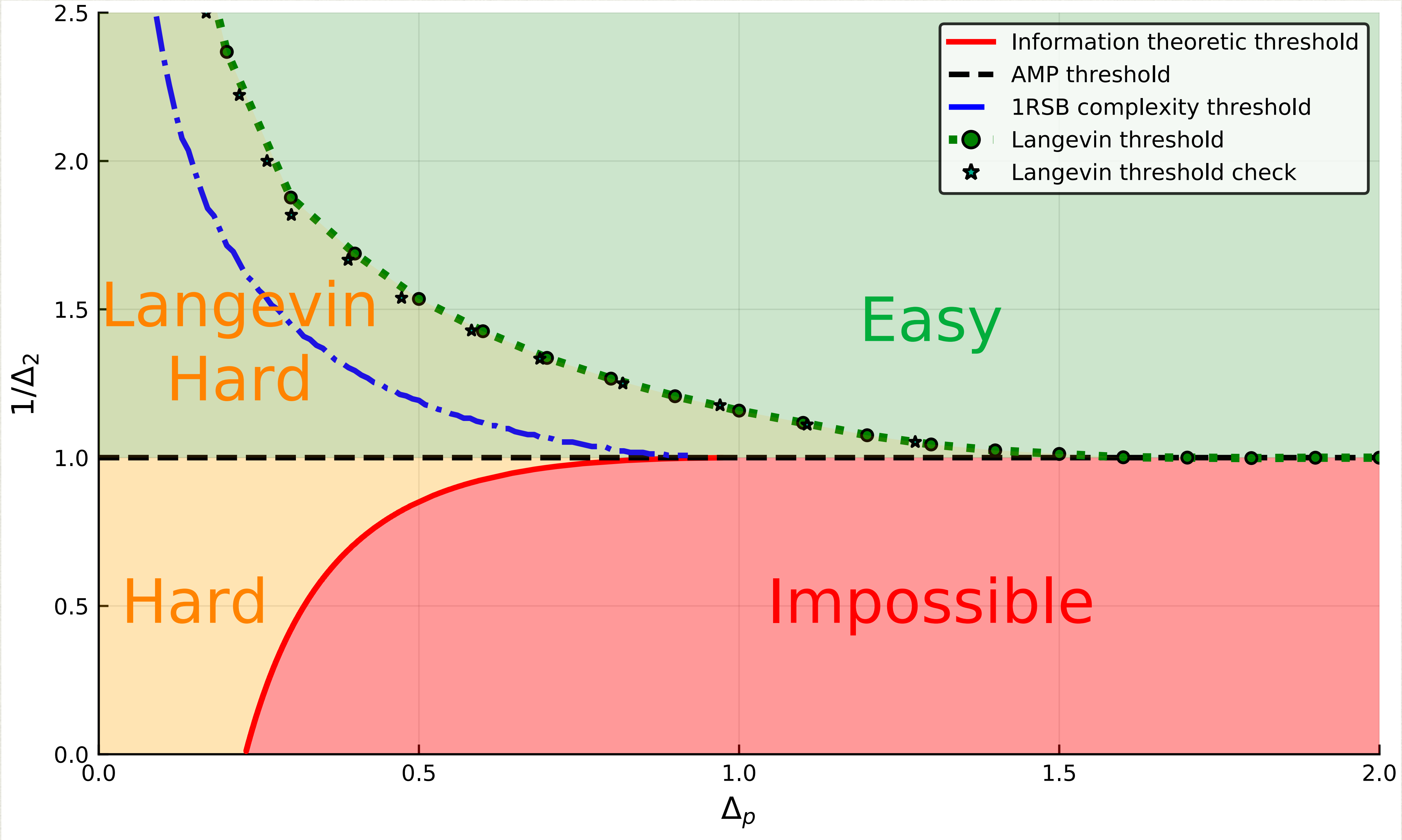


github.com/sphinxteam/spiked_matrix-tensor

EXTRAPOLATION OF LANGEVIN CONVERGENCE TIME



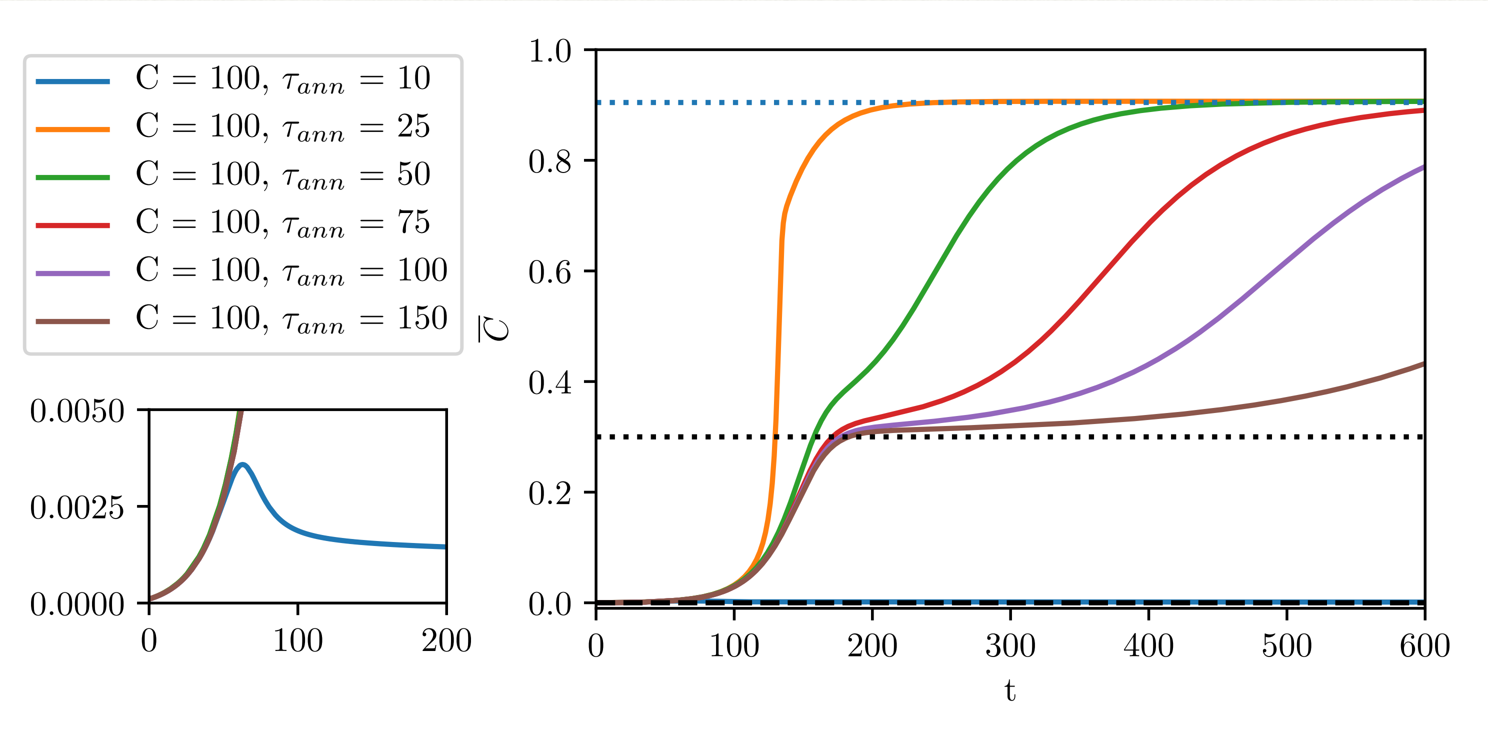
LANGEVIN PHASE DIAGRAM



MARVELS AND PITFALLS

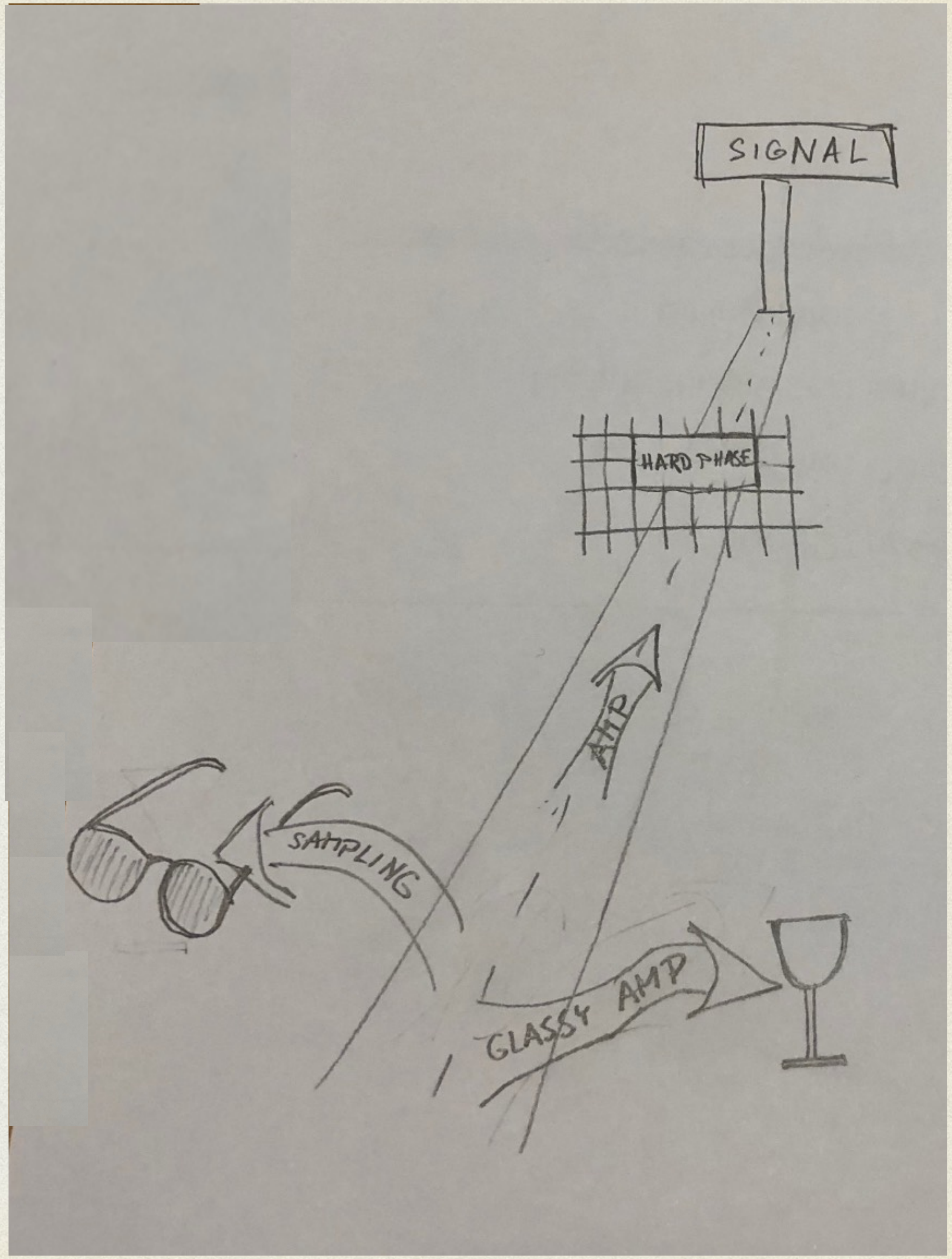
- Langevin fails because of residual glassiness. AMP ignores glassiness, optimises physically wrong objective, yet performs better. How is this physically possible?

- Can Langevin match AMP? Yes: anneal in Δ_p, but not Δ_2. Bayesian puzzle: It is more efficient to mismatch Δ_p from the true one. Ever observed before?



SUMMARY

- State evolution for Langevin.
- AMP superior by making physically wrong assumptions.
- Bayesian puzzle - wrong priors may bring computational advantage.
- Poster of S. Sarao: gradient descent, Kac-Rice annealed and quenched, and AMP at T=0.



We expect the same picture to hold in all problems having hard phase associated to the first order phase transition. (e.g. neural networks with hidden units ...) - work in progress.

TALK BASED ON

- Barbier, Dia, Macris, Krzakala, Lesieur, LZ *Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula*, NeurIPS'16, & arXiv:1812.02537
- Lesieur, Krzakala, LZ, *Constrained Low-rank Matrix Estimation: Phase Transitions, Approximate Message Passing and Applications*, J. Stat. Mech.'17
- Lesieur, Miolane, Lelarge, Krzakala, LZ, *Statistical and computational phase transitions in spiked tensor estimation*, ISIT'17

- Antenucci, Franz, Urbani, LZ, *On the glassy nature of the hard phase in inference problems,* Phys. Rev. X., arXiv:1805.05857
- Antenucci, Krzakala, Urbani, LZ, *Approximate Survey Propagation for Statistical Inference, arXiv:1807.01296*
- Sarao, Biroli, Cammarota, Krzakala, Urbani, LZ, *Marvels and Pitfalls of the Langevin Algorithm in Noisy High-dimensional Inference*, arXiv:1812.09066