HOW CLOSE TO OPTIMAL SAMPLE COMPLEXITY DOES GRADIENT DESCENT GET IN PHASE RETRIEVAL?

Lenka Zdeborová

(EPFL)

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

Computational Complexity of Statistical Inference, Simons institute, 12. 10. 2021

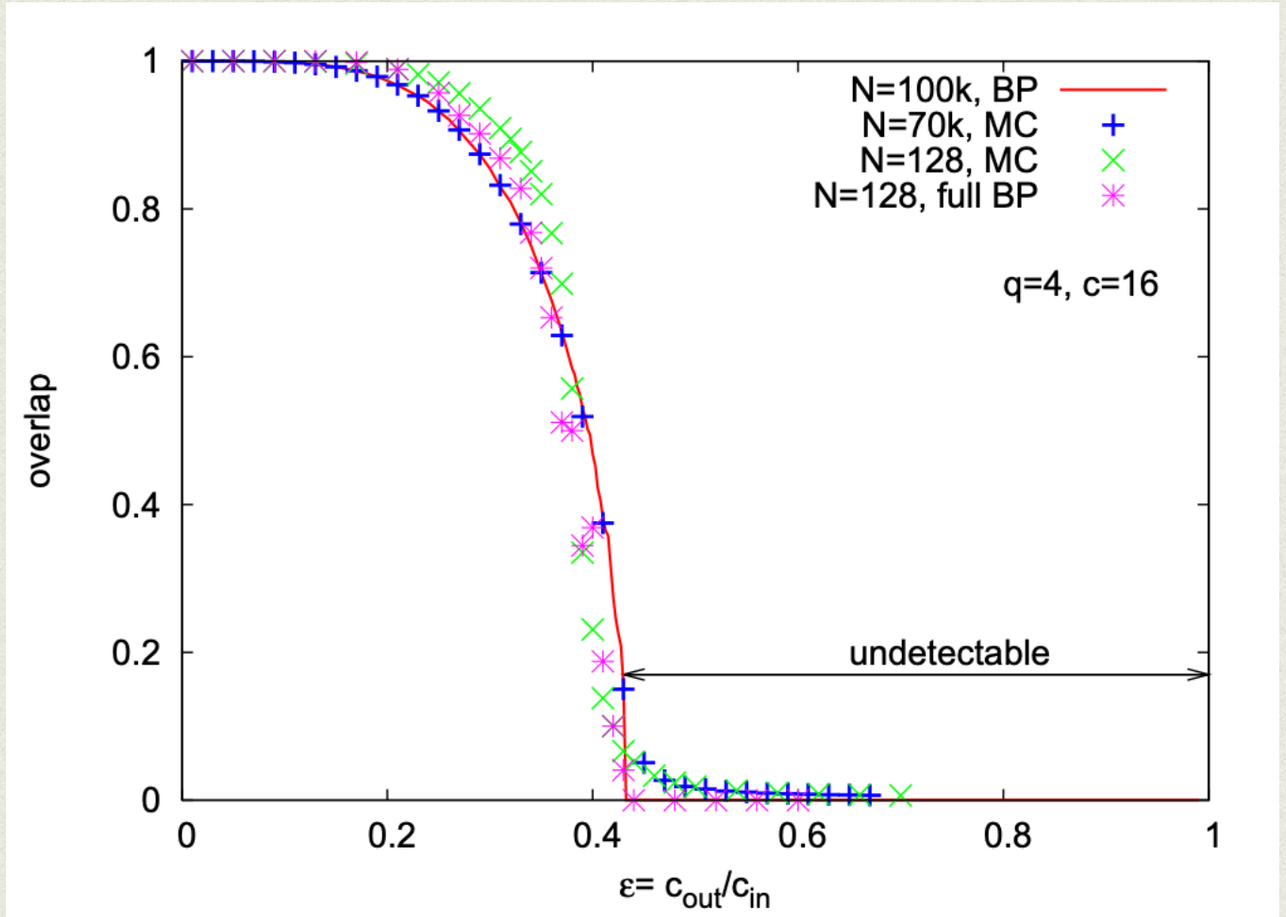
Parisi's discoveries make it possible to understand and describe many different and apparently entirely random complex materials and phenomena, not only in physics but also in other, very different areas, such as mathematics, biology, neuroscience and machine learning.

COMPARING

- Message passing algorithms: Belief propagation, Approximate Message Passing

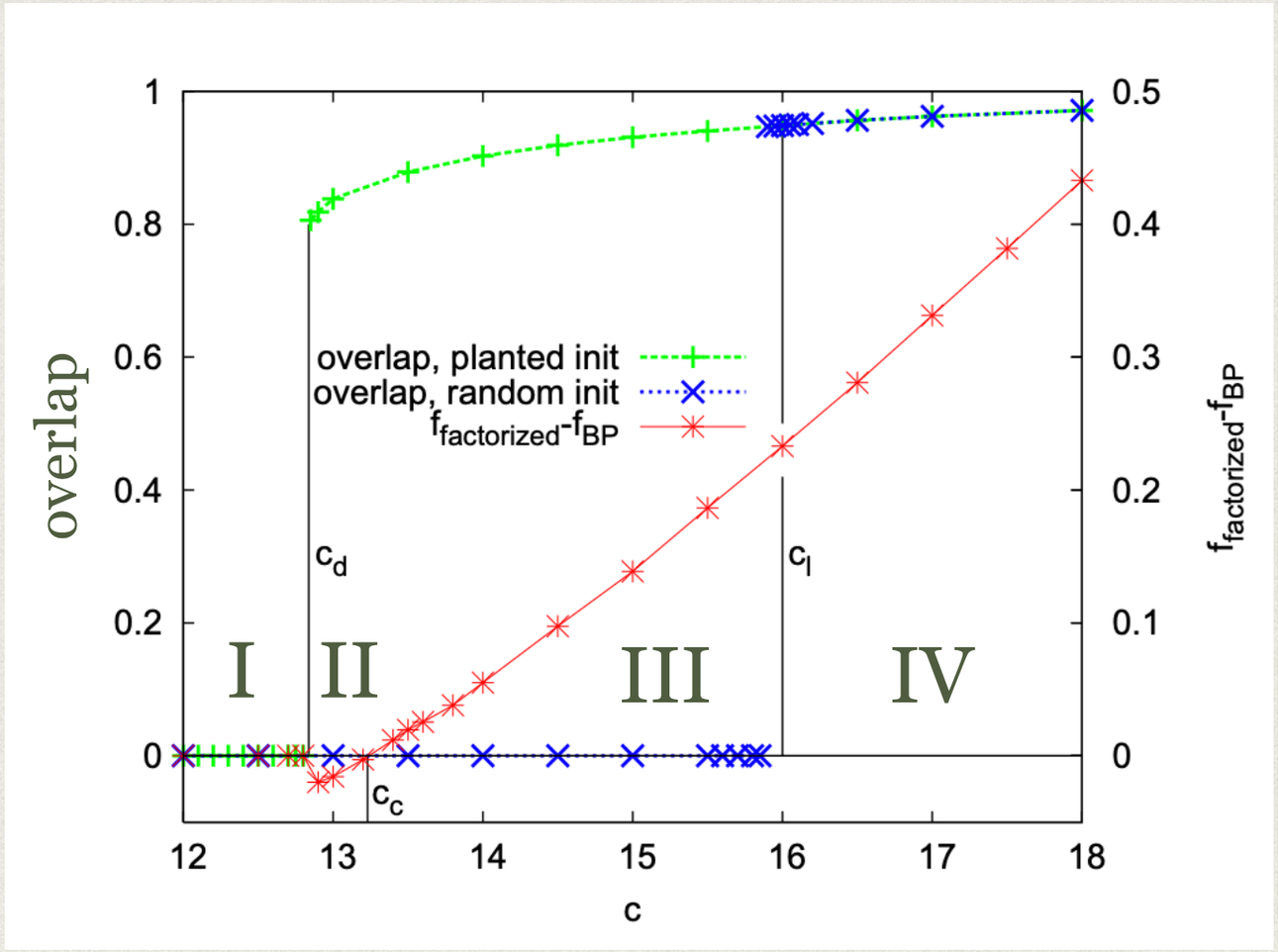
- Gradient & sampling based algorithms: Metropolis Monte Carlo, Gibbs Sampling, Langevin Algorithm, Gradient Descent

STOCHASTIC BLOCK MODEL



Decelle, Krzakala, Moore, LZ'11: Numerical evidence that BP and MCMC both reach the detectability threshold.

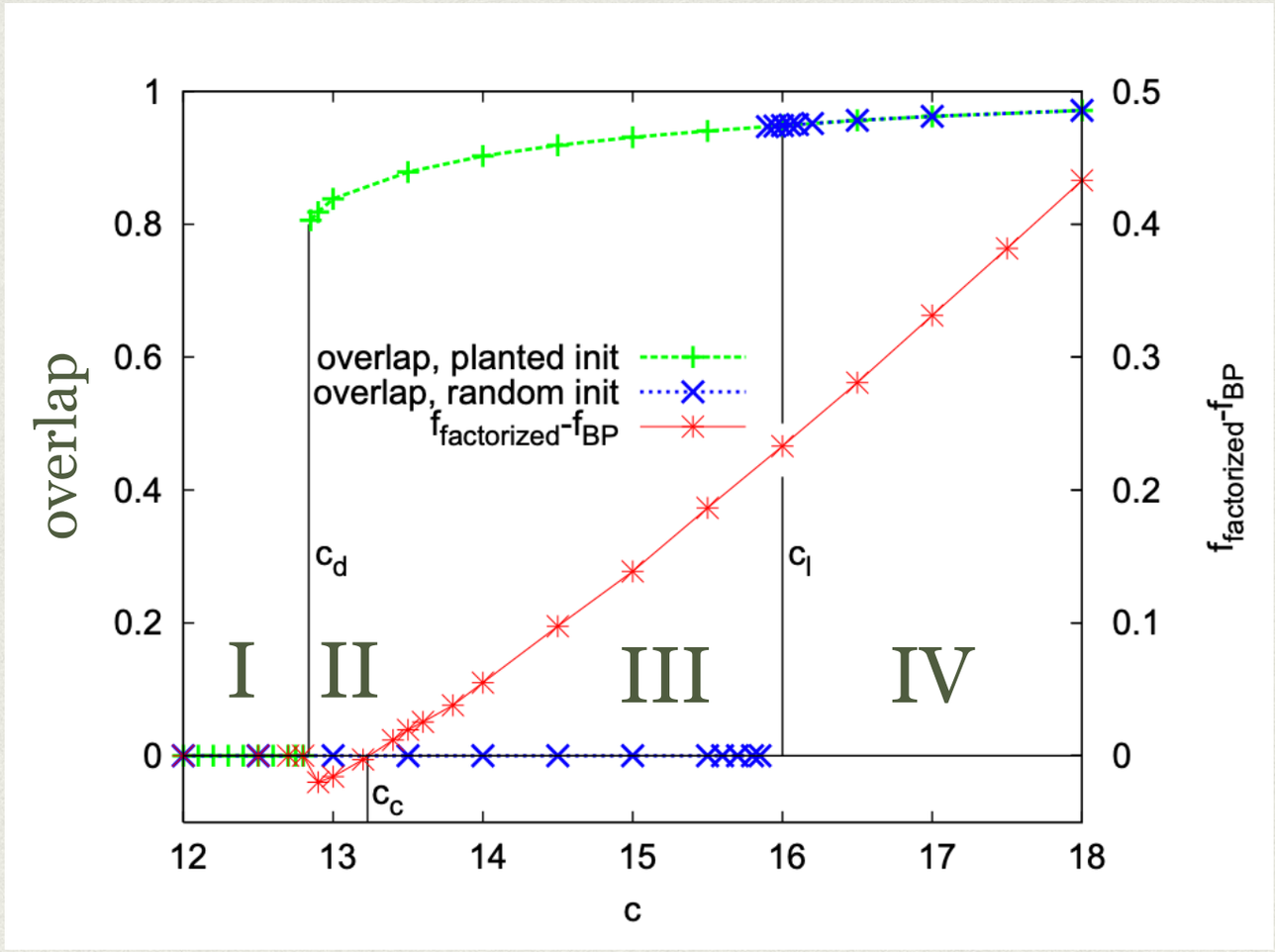
Does MCMC match BP even in the presence of a hard phase?



From Decelle, Krzakala, Moore, LZ'11:

We also investigated the case $q = 5$, $c_{in} = 0$, illustrated in Fig. 3, with Gibbs sampling, i.e., the Markov chain Monte Carlo algorithm. For the planted initialization, its performance is generally similar to BP. For the random initialization, MCMC agrees with BP only in phases (I) and (IV). It follows from results on glassy systems [42] that in phases (II) and (III), the equilibration time of MCMC is exponentially large as a function of N, and that its performance in linear time, i.e., CN for any constant C, does not yield any information about the original group assignment.

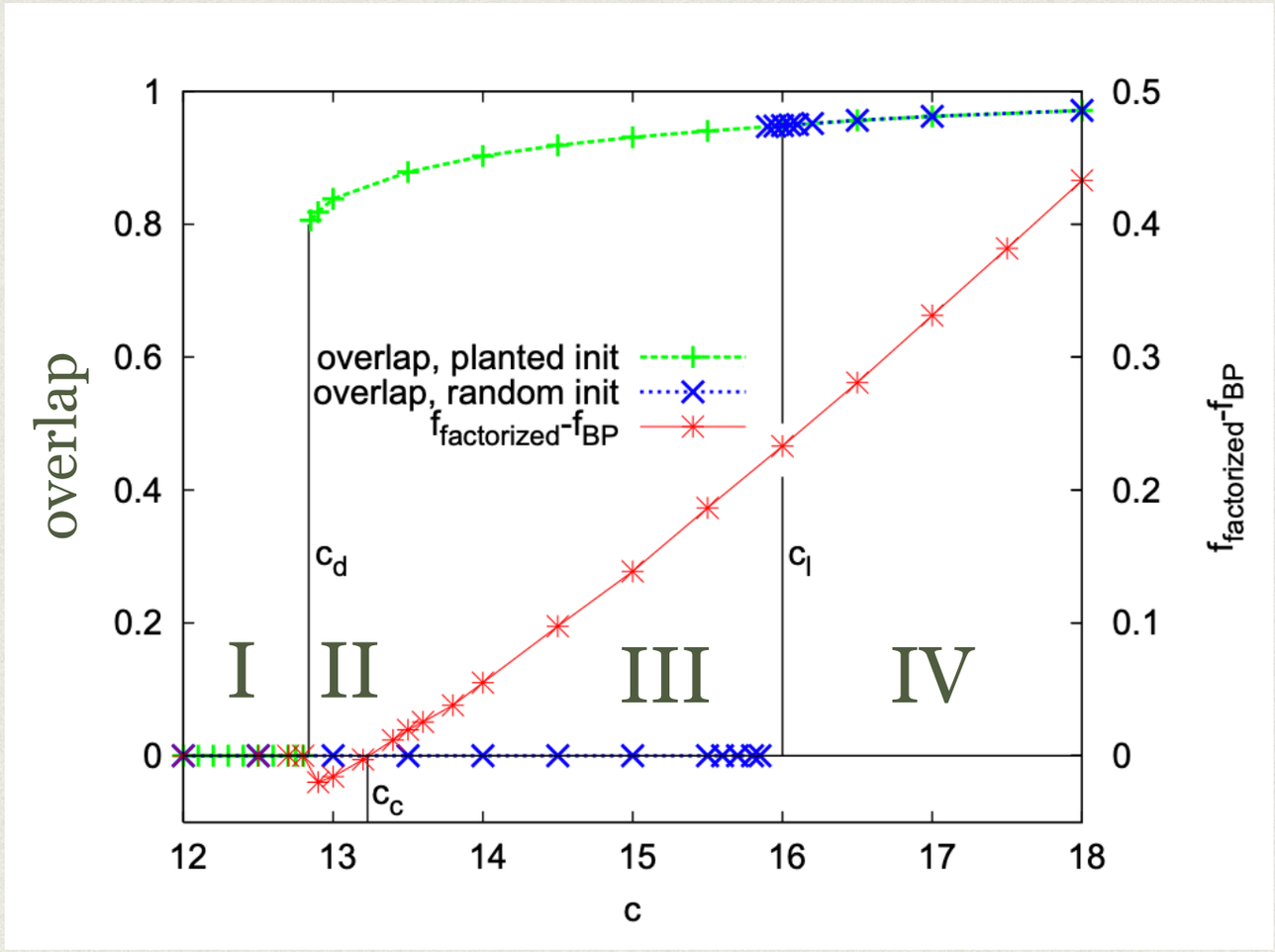
Does MCMC match BP even in the presence of a hard phase?



From Decelle, Krzakala, Moore, LZ'11:

We also investigated the case $q = 5$, $c_{in} = 0$, illustrated in Fig. 3, with Gibbs sampling, i.e., the Markov chain Monte Carlo algorithm. For the planted initialization, its performance is generally similar to BP. For the random initialization, ~~MCMC agrees with BP only in phases (I) and (IV)~~. It follows from results on glassy systems [42] that in phases (II) and (III), the equilibration time of MCMC is exponentially large as a function of N, and that its performance in linear time, i.e., CN for any constant C, does not yield any information about the original group assignment.

Does MCMC match BP even in the presence of a hard phase?



From Decelle, Krzakala, Moore, LZ'11:

We also investigated the case $q = 5$, $c_{in} = 0$, illustrated in Fig. 3, with Gibbs sampling, i.e., the Markov chain Monte Carlo algorithm. For the planted initialization, its performance is generally similar to BP. For the random initialization, ~~MCMC agrees with BP only in phases (I) and (IV)~~. It follows from results on glassy systems [42] that in phases (II) and (III), the equilibration time of MCMC is exponentially large as a function of N, and that its performance in linear time, i.e., CN for any constant C, does not yield any information about the original group assignment.

In this talk we argue that generically close to 1st order phase transitions MCMC is worse than BP!

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*?

PHYSICAL REVIEW X

Highlights Recent Subjects Accepted Collections Authors Referees Search

Open Access

Glassy Nature of the Hard Phase in Inference Problems

Fabrizio Antenucci, Silvio Franz, Pierfrancesco Urbani, and Lenka Zdeborová
Phys. Rev. X **9**, 011020 – Published 31 January 2019

Model — sparse rank-one low-rank estimation:

$$Y_{ij} = x_i^* x_j^* / \sqrt{N} + \xi_{ij}$$

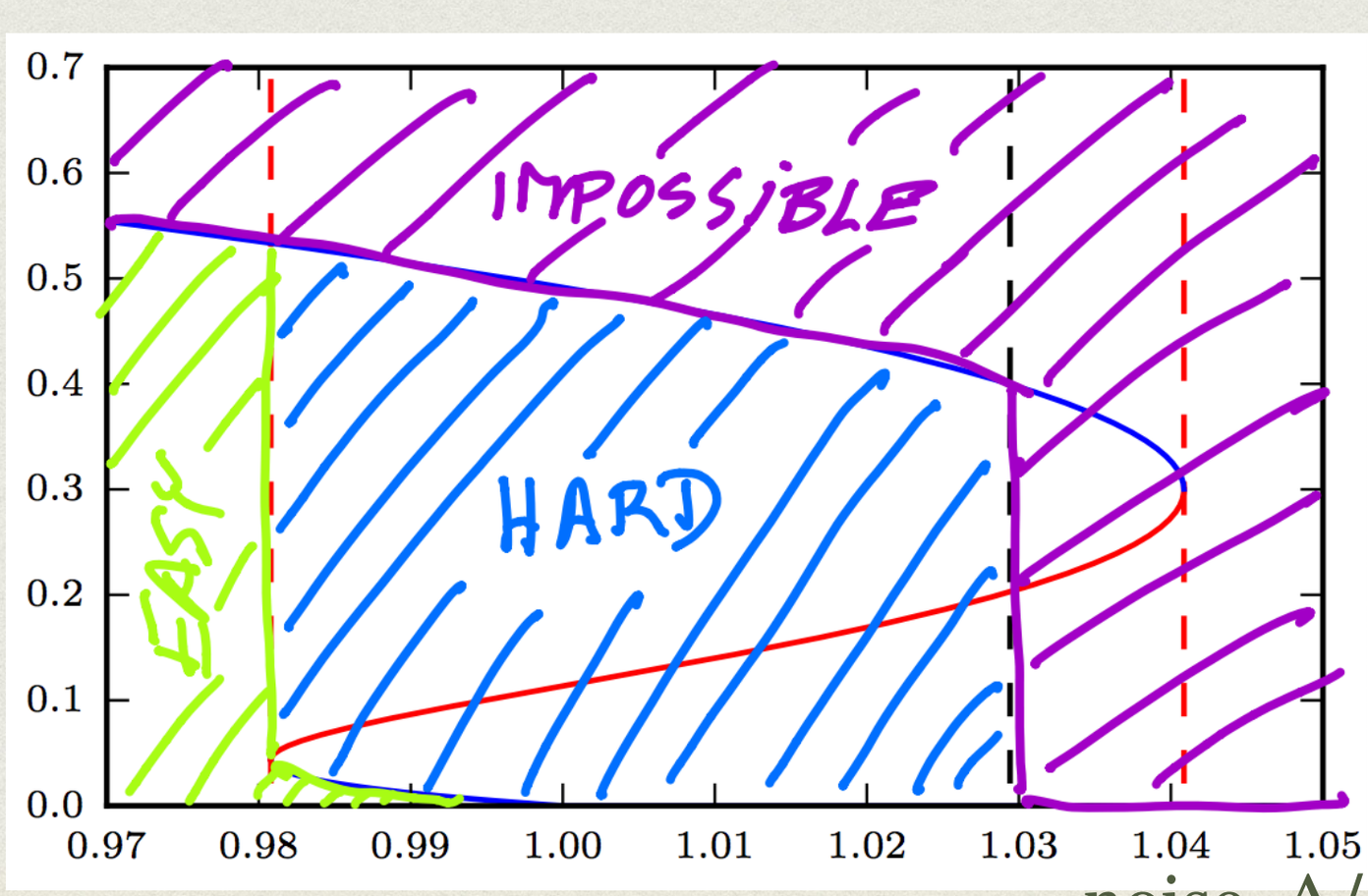
$$P_x(x_i^*) = (1-\rho)\delta(x_i^*) + \rho[\delta(x_i^* - 1) + \delta(x_i^* + 1)]/2$$

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

BAYES-OPTIMAL & AMP PHASE DIAGRAM

Lesieur, Krzakala, LZ, J. Stat. Mech,'17

- Easy by approximate message passing algorithms.
- Impossible information theoretically.
- Hard phase conjecture: No efficient algorithm works.

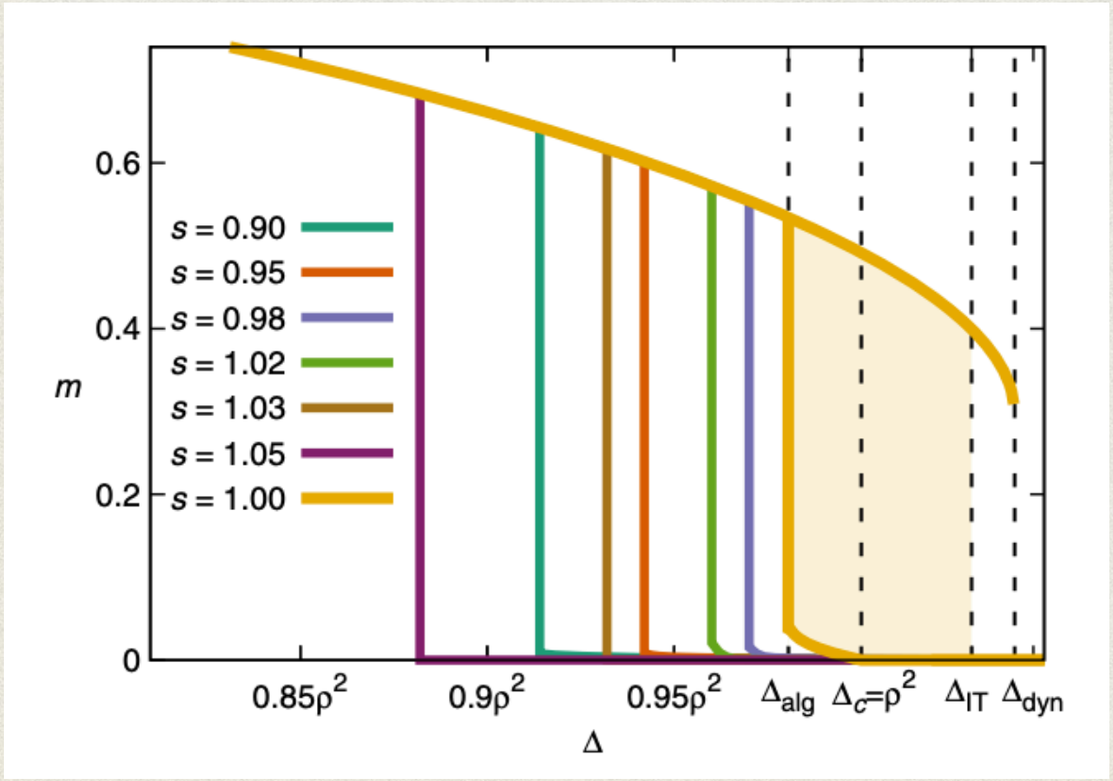


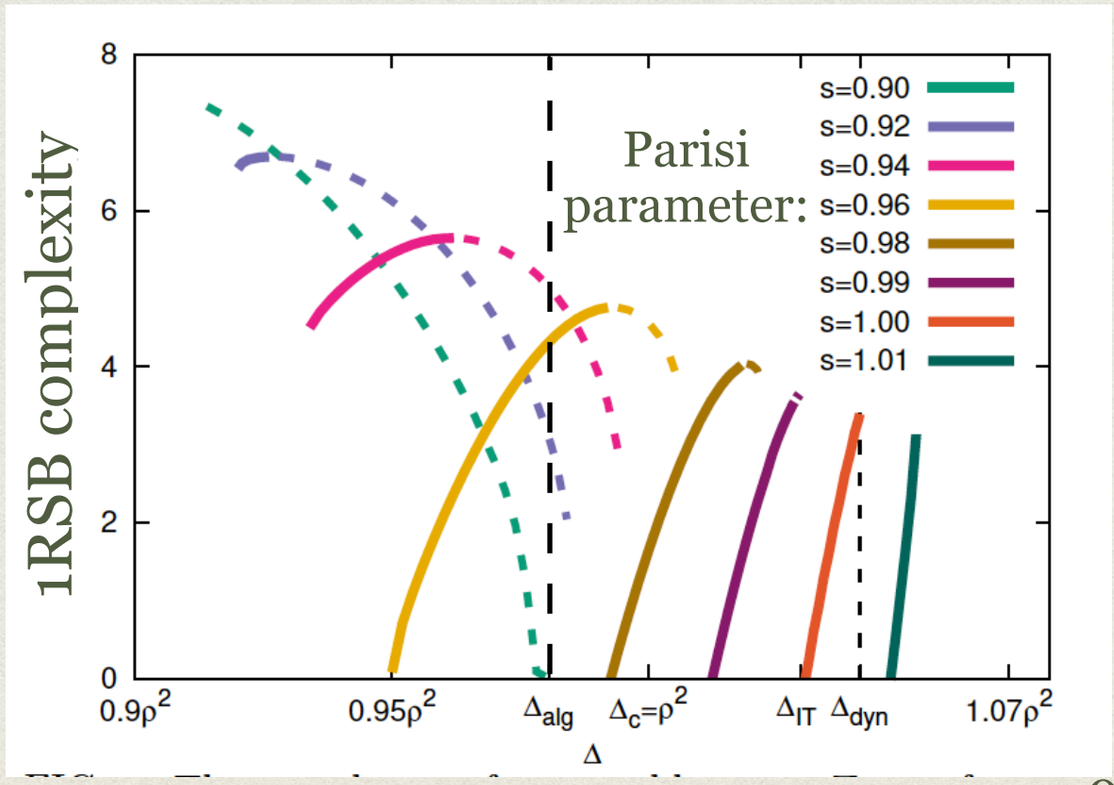
$\rho = 0.08$

GLASSY NATURE OF THE HARD PHASE

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

- Analyzed by 1-step replica symmetry breaking.
- The hard phase is glassy - many spurious local minima potentially blocking the (MCMC, GD, Langevin …) dynamics.
- The glassiness extends well inside the AMP-easy phase.





$\rho = 0.08$

GLASSY NATURE OF THE HARD PHASE

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

Residual glassiness below the algorithmic threshold. =>

Strong yet indirect indication of algorithmic troubles for Gibbs-sampling or gradient based algorithms.

GLASSY NATURE OF THE HARD PHASE

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

Residual glassiness below the algorithmic threshold. =>

Strong yet indirect indication of algorithmic troubles for Gibbs-sampling or gradient based algorithms.

How to confirm this?

- Analytically — Gibbs samplers and gradient descents are harder to analyse than message passing .... let's try anyway!

SPIKED MATRIX-TENSOR MODEL

Sarao, Biroli, Cammarota, Krzakala, Urbani, LZ, PRX'20

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

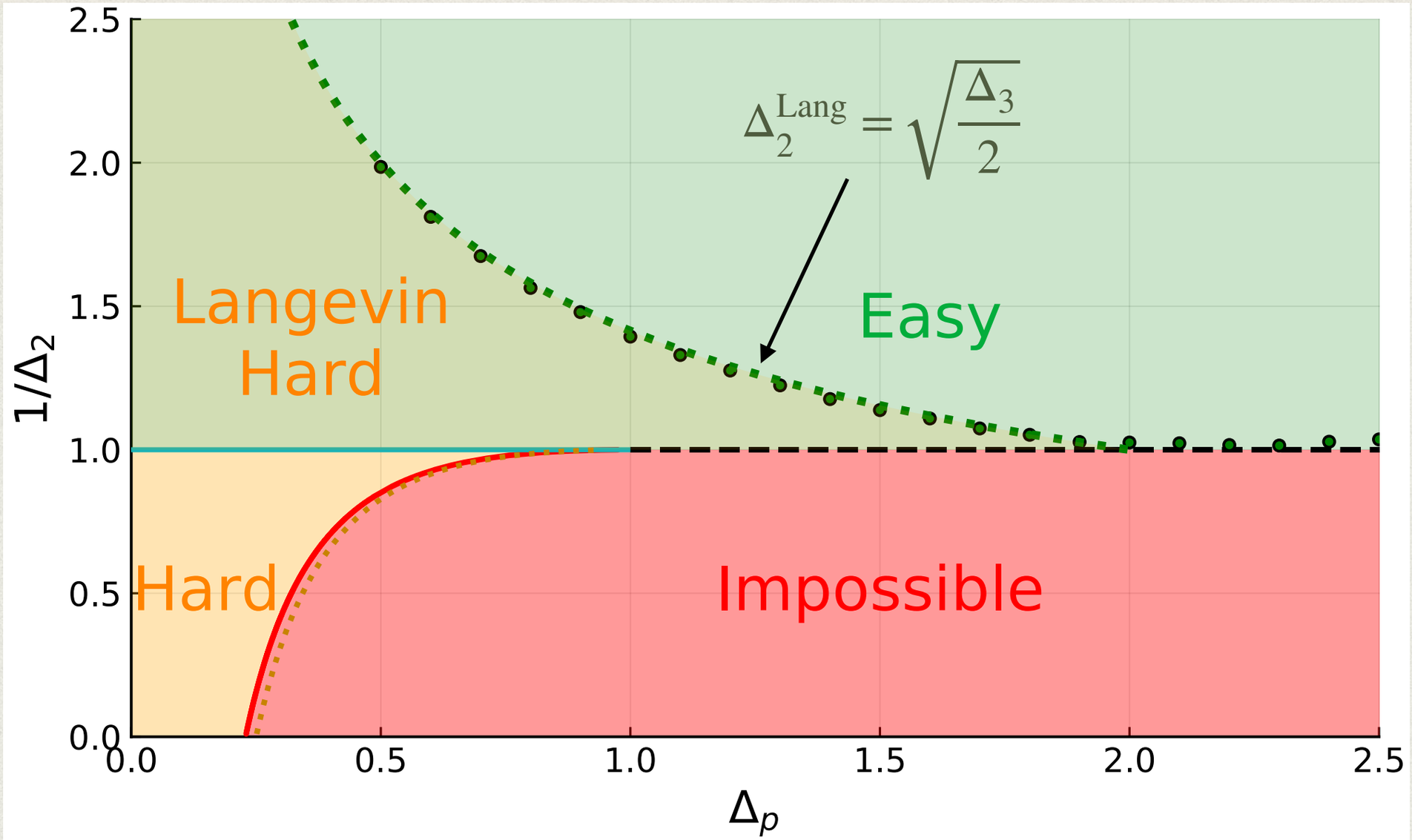
where: $$Y = x^*(x^*)^\top + \mathcal{N}(0, \tilde{\Delta}_2)$$
$$T = (x^*)^{\otimes p} + \mathcal{N}(0, \tilde{\Delta}_p)$$

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

Goal: Estimate x* by Langevin algorithm set to sample the posterior.

PHASE DIAGRAM

Sarao, Biroli, Cammarota, Krzakala, Urbani, LZ, PRX'20



$$\mathscr{L}(x) = \|xx^\top - Y\|_2^2 + \|x^{\otimes p} - T\|_2^2 \qquad \text{p=3}$$

PHASE RETRIEVAL

WHY PHASE RETRIEVAL?

- Phase retrieval is a simple neural network, gradient-descent based algorithm used for learning in practice.
- Non-convex, high-dimensional, limited sample complexity. -> challenging regime for computational learning theory.
- Behavious we observe akin to some aspects of deep neural networks.

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 the absolute value, w* are teacher weights.

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

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

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

BAYES-OPTIMAL GENERALIZATION

Posterior probability distribution:

$$P(w \mid y, X) = \frac{1}{Z(y, X)} \prod_{i=1}^{d} P_w(w_i) \prod_{\mu=1}^{n} P_{\text{out}}(y_\mu \mid X_\mu \cdot w)$$

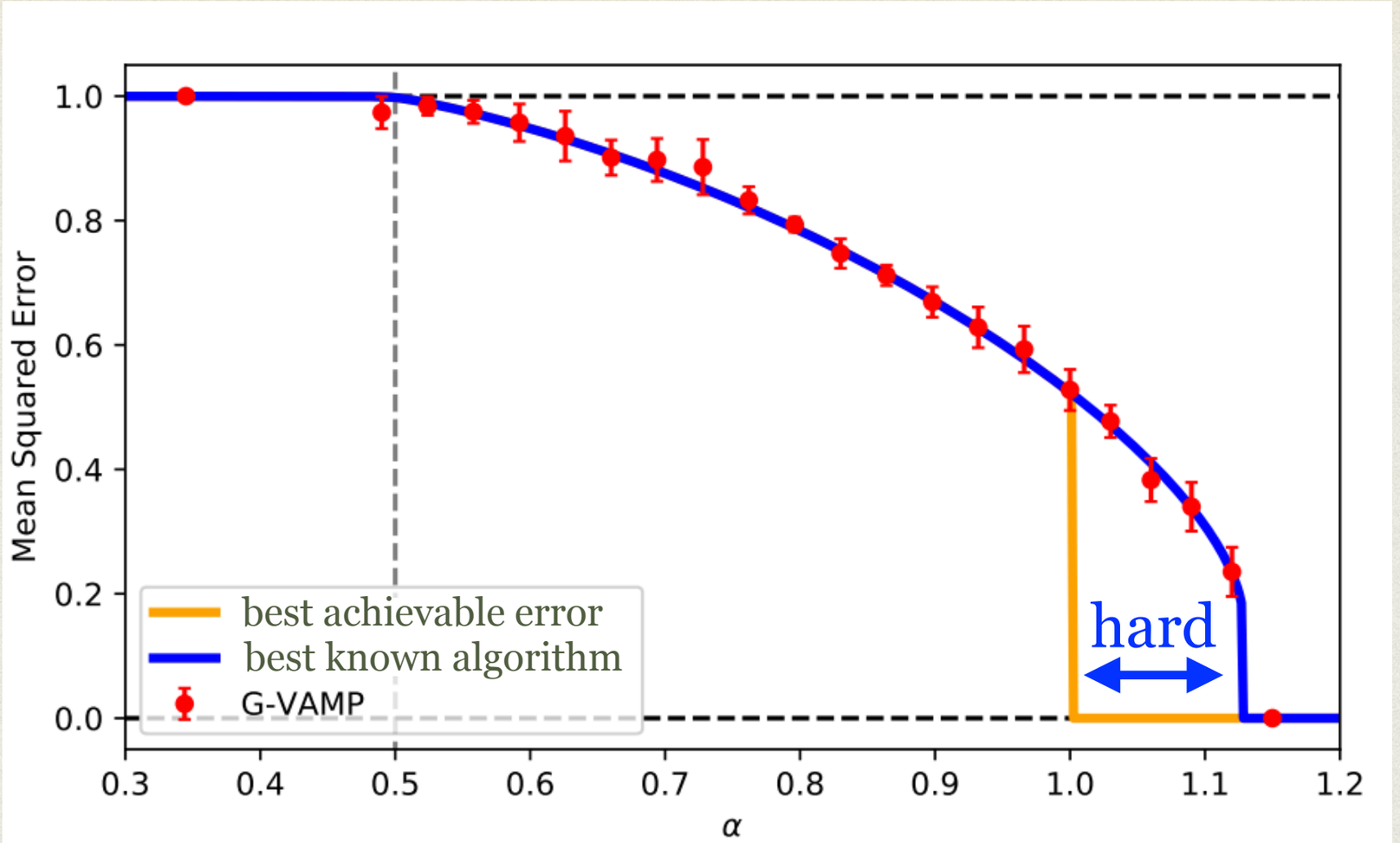
where $P_{\text{out}}(y_\mu \mid X_\mu \cdot w) = \delta(y_\mu - \varphi(X_\mu \cdot w))$

- A new sample X_{new} is given. Bayes-optimal prediction of a new label: $\hat{y}_{\text{new}} = \mathbb{E}_{P(w|y,X)} \left[\varphi(X_{\text{new}} \cdot w) \right]$

$\neq$ empirical risk minimization

BAYES-OPTIMAL ERROR

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



$n \to \infty$

$d \to \infty$

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

$\alpha_{\rm IT} = 1$ # of samples needed for perfect generalisation for any algorithm, achieved by LLL-based algorithm in absence of noise (Song, Zadik, Bruna'21).

$\alpha_{\rm AMP} = 1.13$ # of samples needed for perfect generalisation with approximate message passing algorithm (conjectured optimal among noise-robust ones).

DEEP LEARNING USES EMPIRICAL RISK MINIMISATION

(NOT BAYESIAN ESTIMATION)

ERM & GRADIENT DESCENT

Loss function:

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

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

Gradient flow:

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

ensuring $\|w\|_2^2 = d$

Initialisation:

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

PERFORMANCE OF GRADIENT DESCENT

Chen, Chi, Fan, Ma'19

Cai, Huang, Li, Wang'21

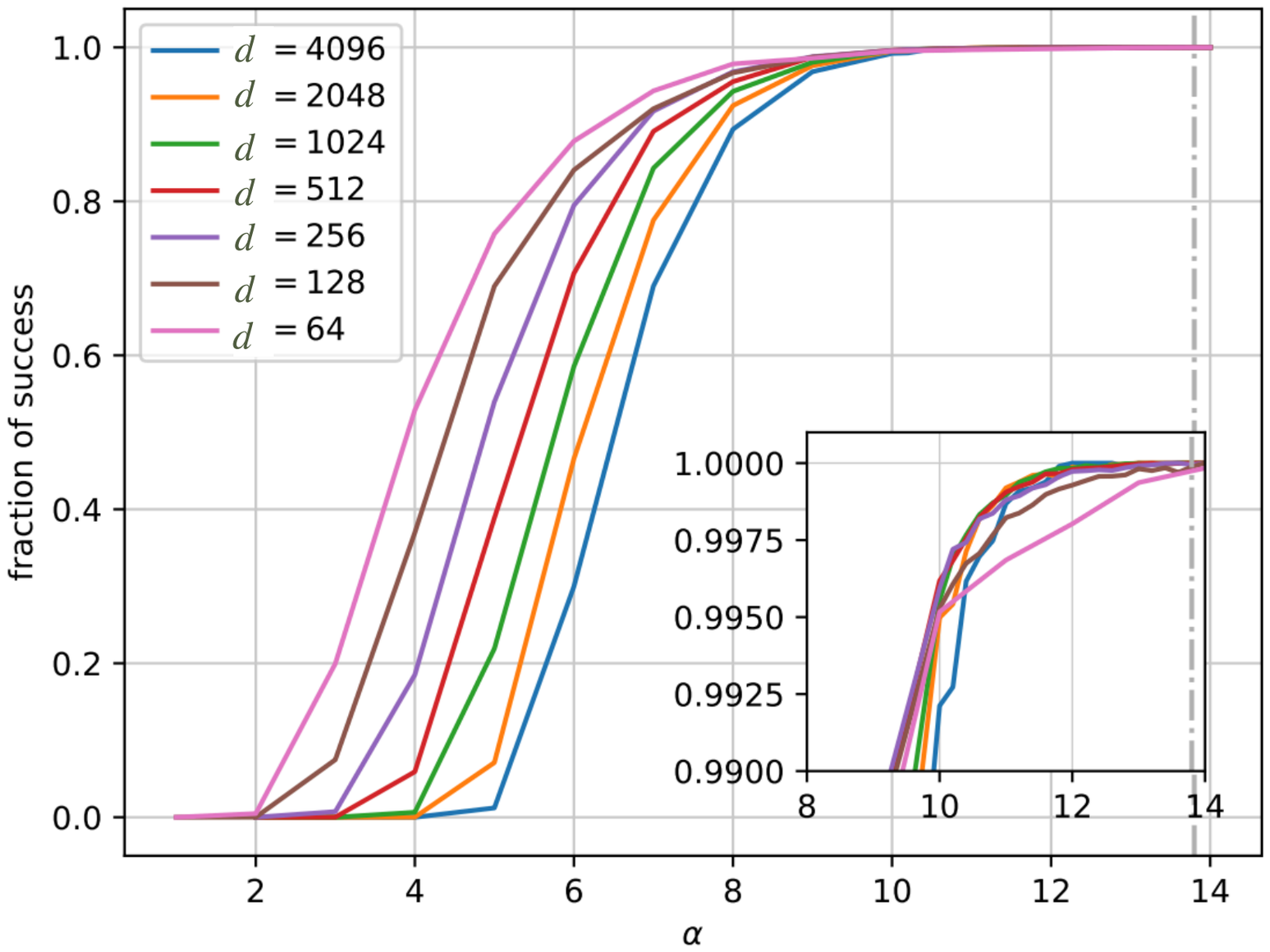
1 1.13 C poly(log d)

IT AMP

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

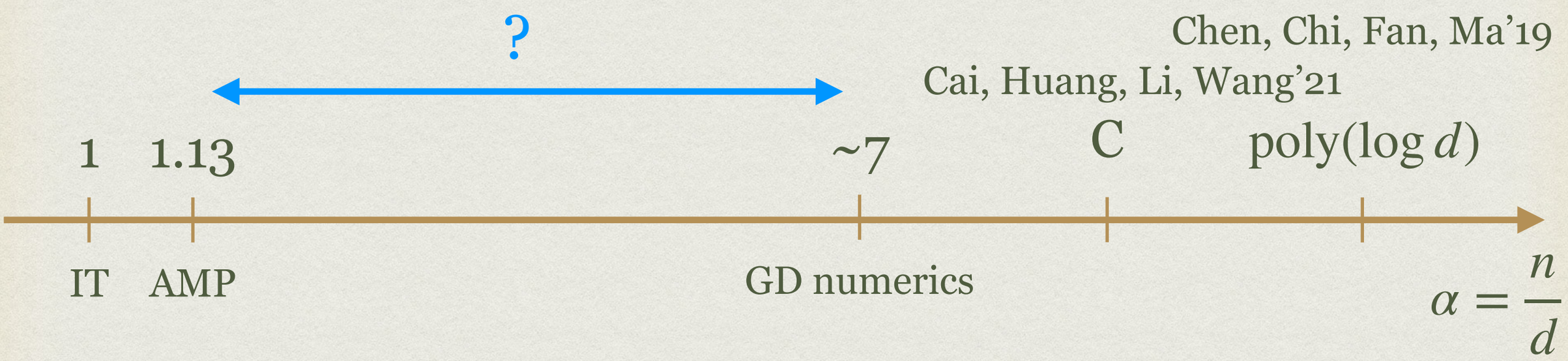
GRADIENT DESCENT NUMERICALLY

Sarao Mannelli, Biroli, Cammarota, Krzakala, LZ, NeurIPS'20, 2006.06997.



PERFORMANCE OF GRADIENT DESCENT

Closing the gap between GD and AMP?



Chen, Chi, Fan, Ma'19

Cai, Huang, Li, Wang'21

DEEP LEARNING IS OVER-PARAMETRIZED

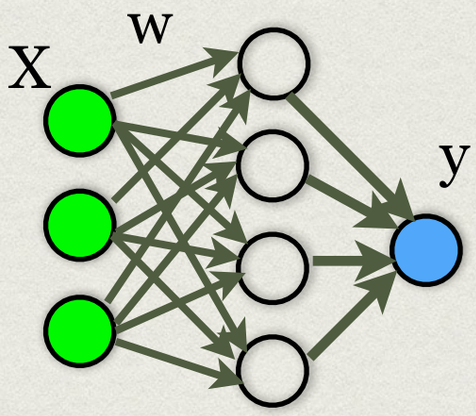
OVER-PARAMETRIZED PHASE RETRIEVAL

- Kernel regression, random features, NTK: Need $O(d^2)$ samples to solve phase retrieval. As opposed to $O(d)$ with AMP/GD.

- Needed instead: Over-parametrized, with feature learning & linear sample complexity.

GRADIENT DESCENT FOR PHASE RETRIEVAL

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



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

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

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

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

OVER-PARAMETRISED LANDSPACE

Sarao Mannelli, Vanden-Eijnden, LZ, NeurIPS'20, 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

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

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

GD FOR OVER-PARAMETRISED PHASE RETRIEVAL

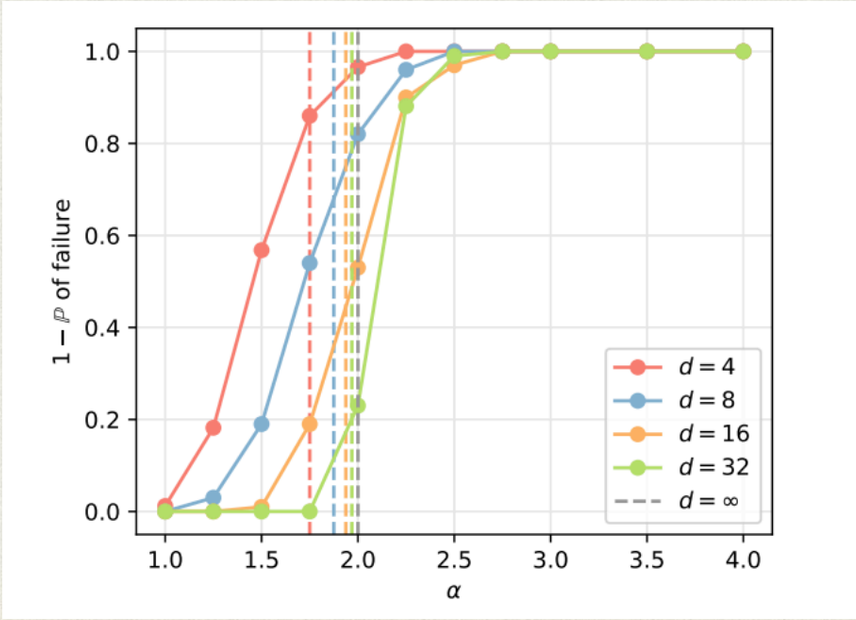
Sarao Mannelli, Vanden-Eijnden, LZ, NeurIPS'20, 2006.15459

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_∞ is a global minimizer of the empirical loss, i.e.

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



PERFORMANCE OF GRADIENT DESCENT

Sarao Mannelli, Vanden-Eijnden, LZ, NeurIPS'20, 2006.15459

Over-parametrised neural network needs fewer samples to learn

Chen, Chi, Fan, Ma'19

Cai, Huang, Li, Wang'21

1 1.13 2 ~7 C $\mathrm{poly}(\log d)$

IT AMP GD in an over-parametrised network GD numerics

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

DEEP LEARNING USES STOCHASTIC GRADIENT DESCENT

PERSISTENT SGD

Mignaco, Urbani, Krzakala, LZ, NeurIPS'20, 2006.06098

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

$$\ell(y_\mu, X_\mu, w) = \left[y_\mu^2 - \left(\sum_{i=1}^{d} X_{\mu i} w_i\right)^2\right]^2$$

SGD	Persistent-SGD
$s_\mu(t) = \begin{cases} 1 & \text{w.p. } b \\ 0 & \text{w.p. } 1-b \end{cases}$	$\text{Prob}\left(s_\mu\left(t+\eta\right)=1 \,\middle|\, s_\mu(t)=0\right) = \frac{\eta}{\tau}$ PERSISTENCE TIME of each sample $\text{Prob}\left(s_\mu\left(t+\eta\right)=0 \,\middle|\, s_\mu(t)=1\right) = \frac{1-b}{b\,\tau}\eta$
DISCRETE-TIME STOCHASTIC PROCESS	WELL-DEFINED CONTINUOUS LIMIT

stochastic gradient flow, $\eta \to 0$

$$\dot{w}_j(t) = -\hat{\nu}(t) w_j(t) - \sum_{\mu=1}^{n} s_\mu(t)\, \partial_{w_j} \ell(y_\mu, X_\mu, w(t))$$

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

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 planted model

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

Mignaco, Urbani, Krzakala, LZ, NeurIPS'20, 2006.06098

Lectures by Urbani to watch at http://leshouches2020.krzakala.org/

Effective scalar stochastic process

eff. regularisation — stochastic noise — memory — Gauss noise

$$\partial_t h(t) = -\tilde{\nu}(t)h(t) - s(t)\partial_1 v(\tilde{h}(t); h_0) + \int_0^t dt'\, M_R(t,t')h(t') + \chi(t)$$

$$h_0 \sim \mathcal{N}(0,1)$$

$$\tilde{h}(t) \equiv h(t) + h_0 m(t)$$

Gaussian effective noise:

$$\langle \chi(t) \rangle = 0, \qquad \langle \chi(t)\chi(t') \rangle = 2T\delta(t-t') + M_C(t,t')$$

MEMORY KERNELS AND OTHER VARIABLES

$$\partial_t m(t) = -\hat{\nu}(t)m(t) - \mu(t) \qquad m(0) = m_0$$

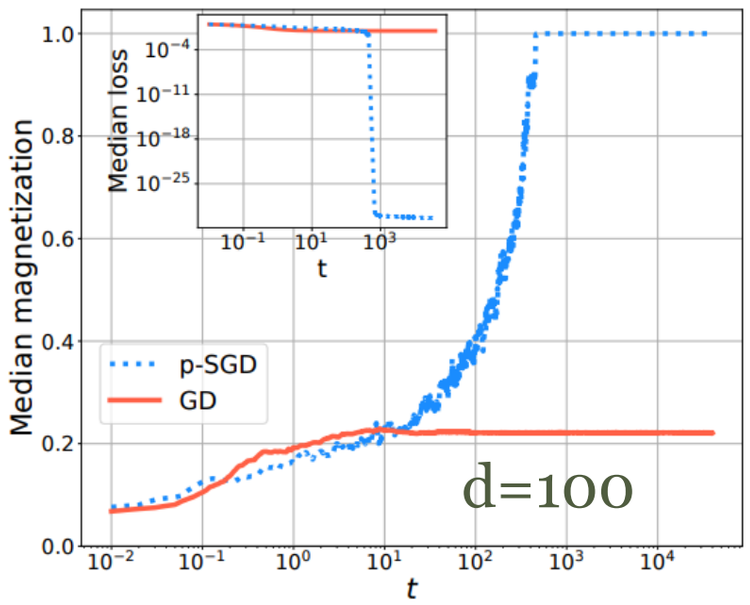
$$M_C(t,t') = \frac{\alpha}{b^2}\langle s(t)s(t')\partial_1 v(\tilde{h}(t);h_0)\partial_1 v(\tilde{h}(t');h_0))\rangle$$

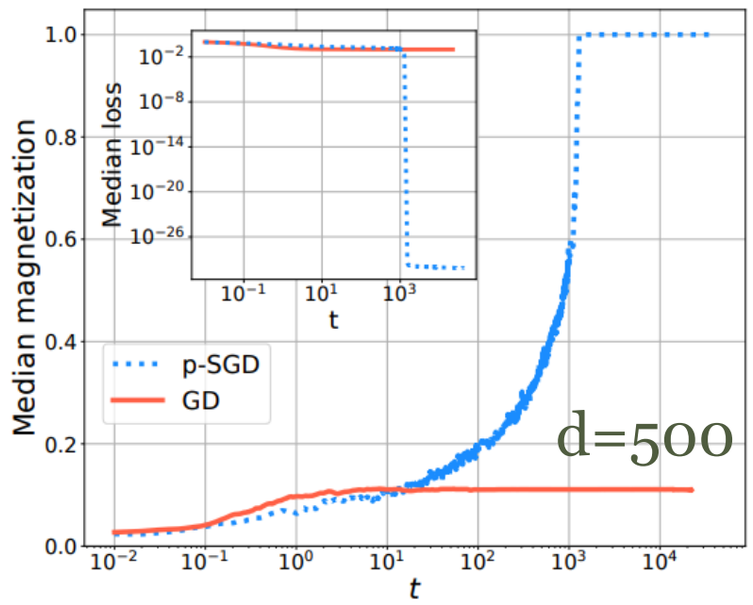
$$M_R(t,t') = \frac{\alpha}{b^2}\frac{\delta}{\delta P(t')}\langle s(t)\partial_1 v(\tilde{h}(t);h_0)\rangle\Big|_{P=0}$$

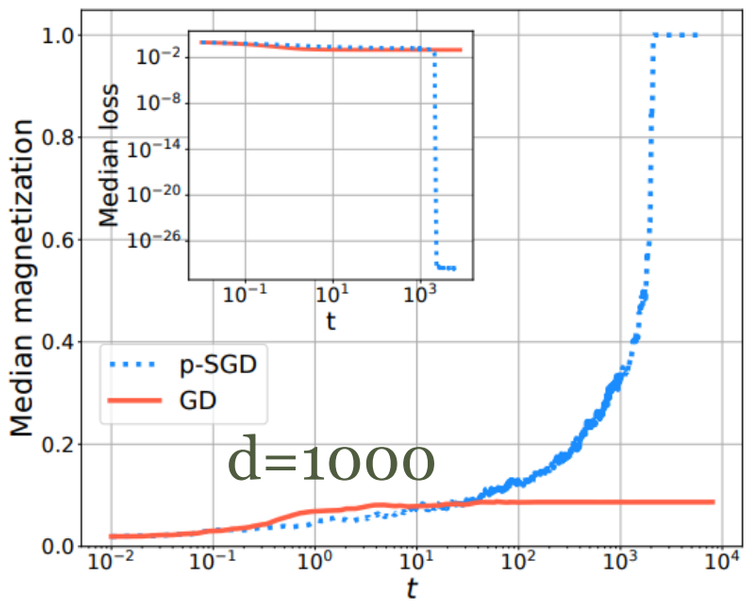
$$\delta\nu(t) = \frac{\alpha}{b}\langle s(t)\partial_1^2 v(\tilde{h}(t);h_0)\rangle \qquad \mu(t) = \frac{\alpha}{b}\langle s(t)h_0\partial_1 v(\tilde{h}(t);h_0)\rangle$$

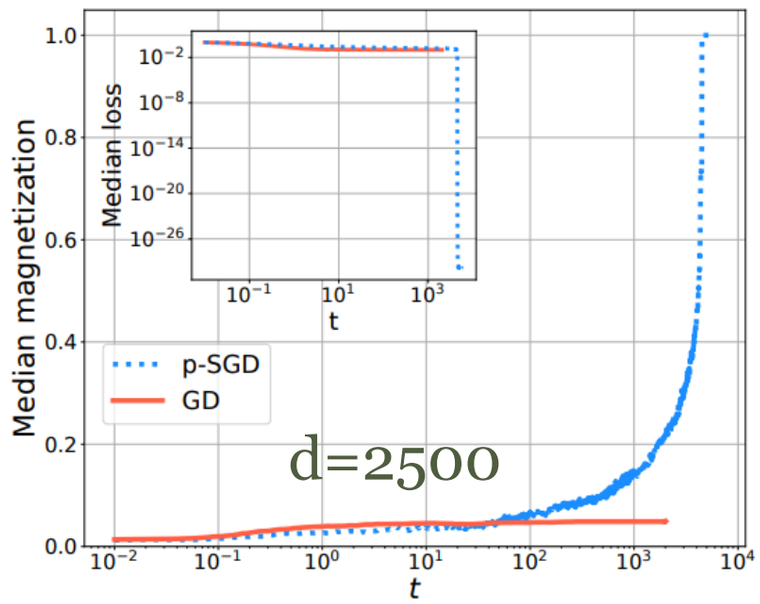
$$\hat{\nu}(t) = -\frac{\alpha}{b}\langle s(t)\tilde{h}(t)\partial_1 v(\tilde{h}(t);h_0)\rangle \qquad \tilde{\nu}(t) = \hat{\nu}(t) + \delta\nu(t)$$

P-SGD WITH RANDOM START









GD/p-SGD in phase retrieval, random start.

$\alpha = 2.5$

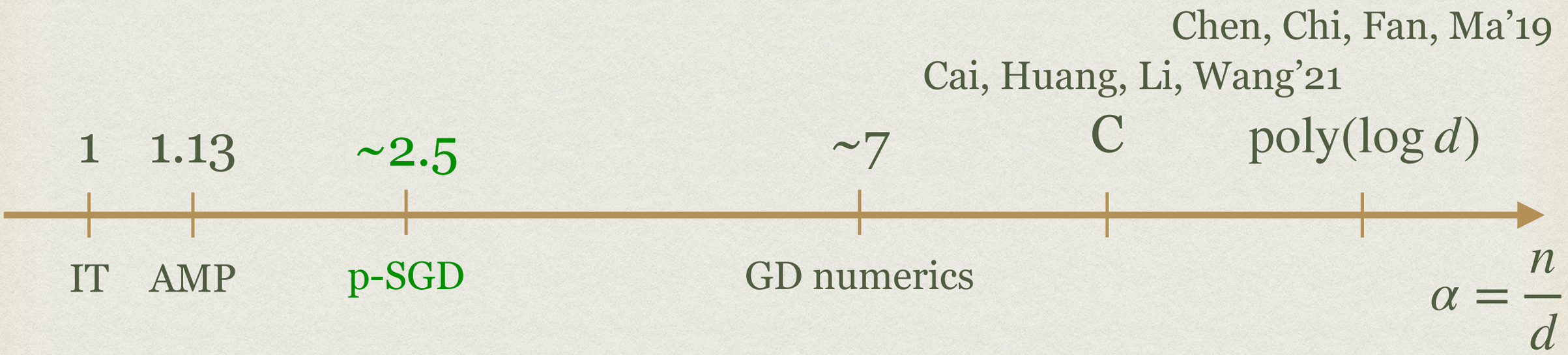
$\eta_{\mathrm{SGD}} = 0.01$

$b = 0.5, \tau = 2$

PERFORMANCE OF P-SGD

Mignacco, Urbani, LZ; MLST, 2103.04902.

p-SGD needs fewer samples to learn phase retrieval



SUMMARY

Phase-retrieval (high-d, real-valued teacher-student setting, Gaussian input data, Gaussian teacher weights) is a neat model to study dynamics of learning with neural networks.

- Sample complexity of gradient-based algorithms can be improved with over-parametrization or with p-SGD.
- Solvable case of feature learning in high-d over-parametrized setting.
- Persistent gradient descent - a variant of SGD with a nice flow limit, analysable by DMFT, performing better than SGD (without hidden units).

OPEN QUESTIONS

Phase-retrieval (high-d, real-valued teacher-student setting, Gaussian input data, Gaussian teacher weights):

- Sample complexity of GD and how does it depend on the loss, initialisation, learning rate?
- Architectures for which GD/SGD needs smaller sample complexity than $\alpha = 2$?
- Sample complexity of GD with number of hidden units 1<m<d?
- Sample complexity of SGD for over-parametrized networks m>1?
- etc.

OPEN QUESTIONS

Phase-retrieval (high-d, real-valued teacher-student setting, Gaussian input data, Gaussian teacher weights):

- Sample complexity of GD and how does it depend on the loss, initialisation, learning rate?
- Architectures for which GD/SGD needs smaller sample complexity than $\alpha = 2$?
- Sample complexity of GD with number of hidden units 1<m<d?
- Sample complexity of SGD for over-parametrized networks m>1?
- etc.