INSIGHTS ON GRADIENT-BASED ALGORITHMS IN HIGH-DIMENSIONAL NON-CONVEX OPTIMISATION

Lenka Zdeborová
(EPFL)

EPFL CIS - RIKEN seminar, 11. May 2022

UNDERSTANDING MACHINE LEARNING

- We know (among others):
  - Neural networks are universal approximators (Cybenko'89).
  - The optimisation problem is NP-hard (e.g. Blum, Rivest'89).
- We do not know (among others):

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

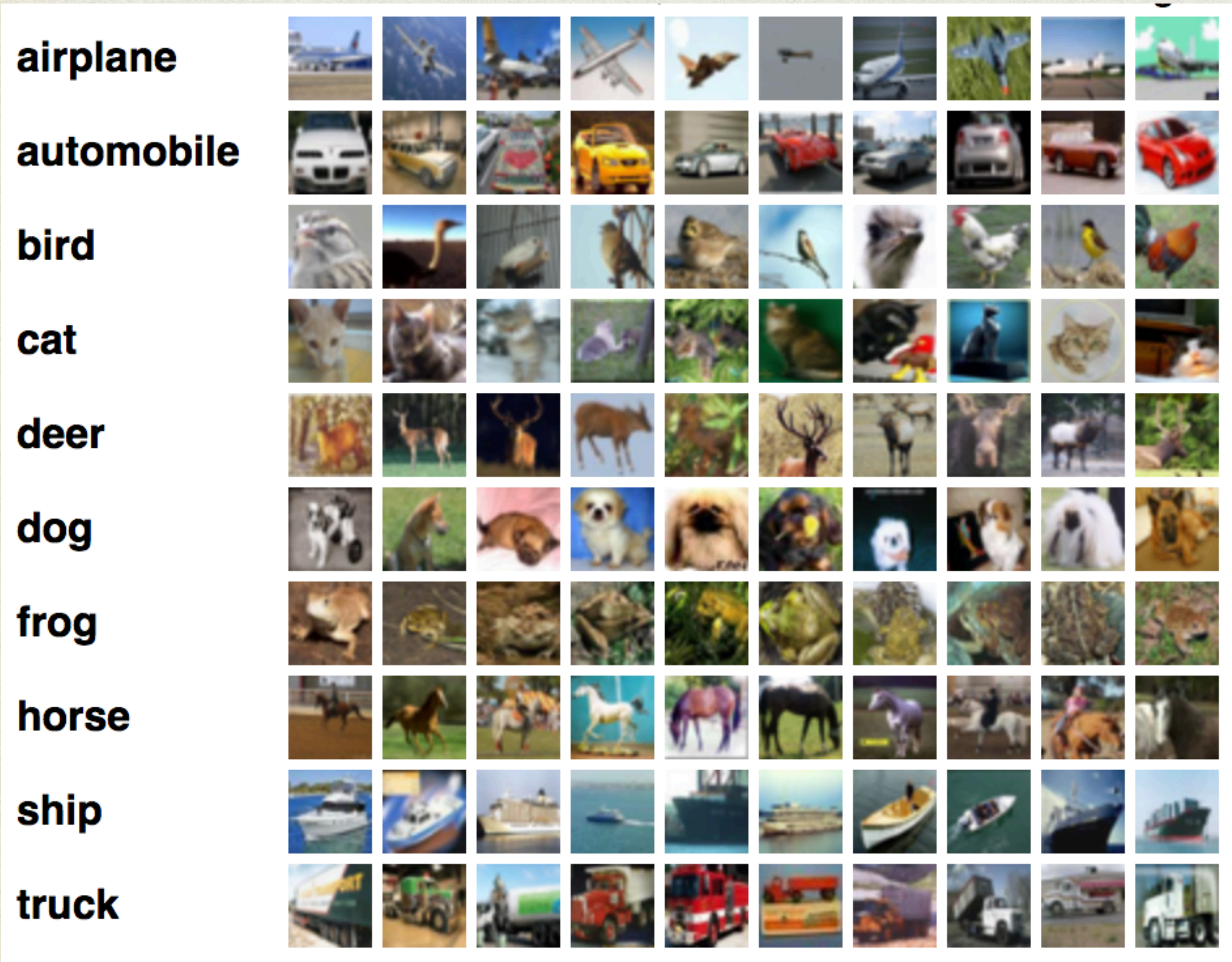
- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?

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

Still not answered!

SAMPLE COMPLEXITY

How many training samples are needed for a given task? Are we close to the minimum? If not, is it because of architectures or algorithms?



- Cifar10 - 50000 samples.
- How many samples are really needed?

UNDERSTANDING MACHINE LEARNING

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!

IN DEEP LEARNING

- Empirical observation: Global minima with bad generalisation error do exist.

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

- Literature: ~~No bad minima.~~ Implicit regularisation. Learning simple functions first. Etc.
  No really satisfactory answer yet.

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.

GOAL

Analyse generalisation properties of gradient-based algorithms in non-convex high-dimensional setting at low sample complexity.

Key points:

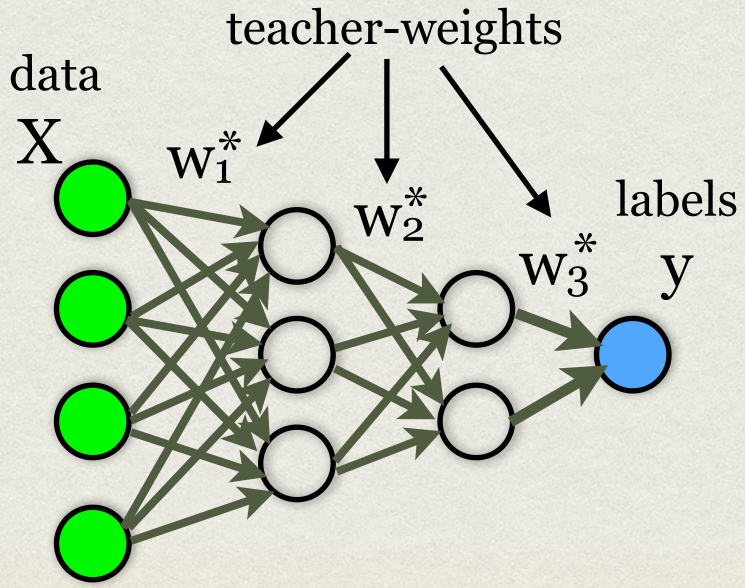
- non-convex
- high-dimensional
- low sample complexity

Next: A setting where this can be done.

TEACHER-STUDENT SETTING

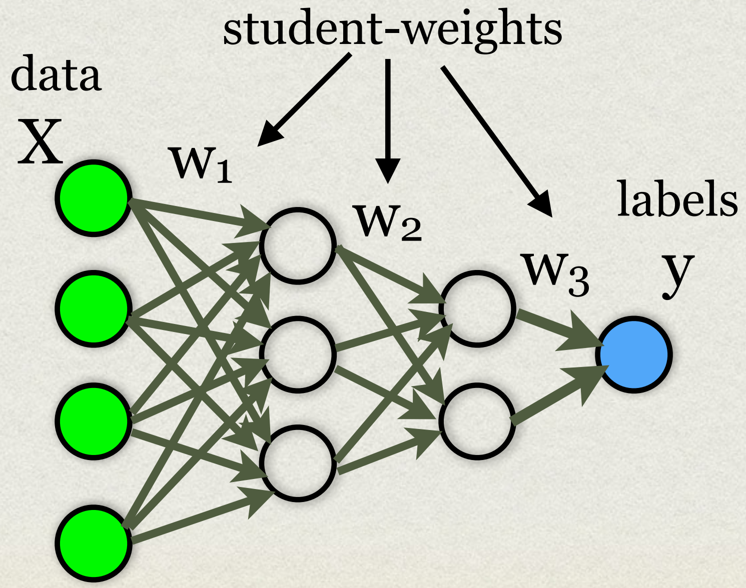
Teacher-network

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

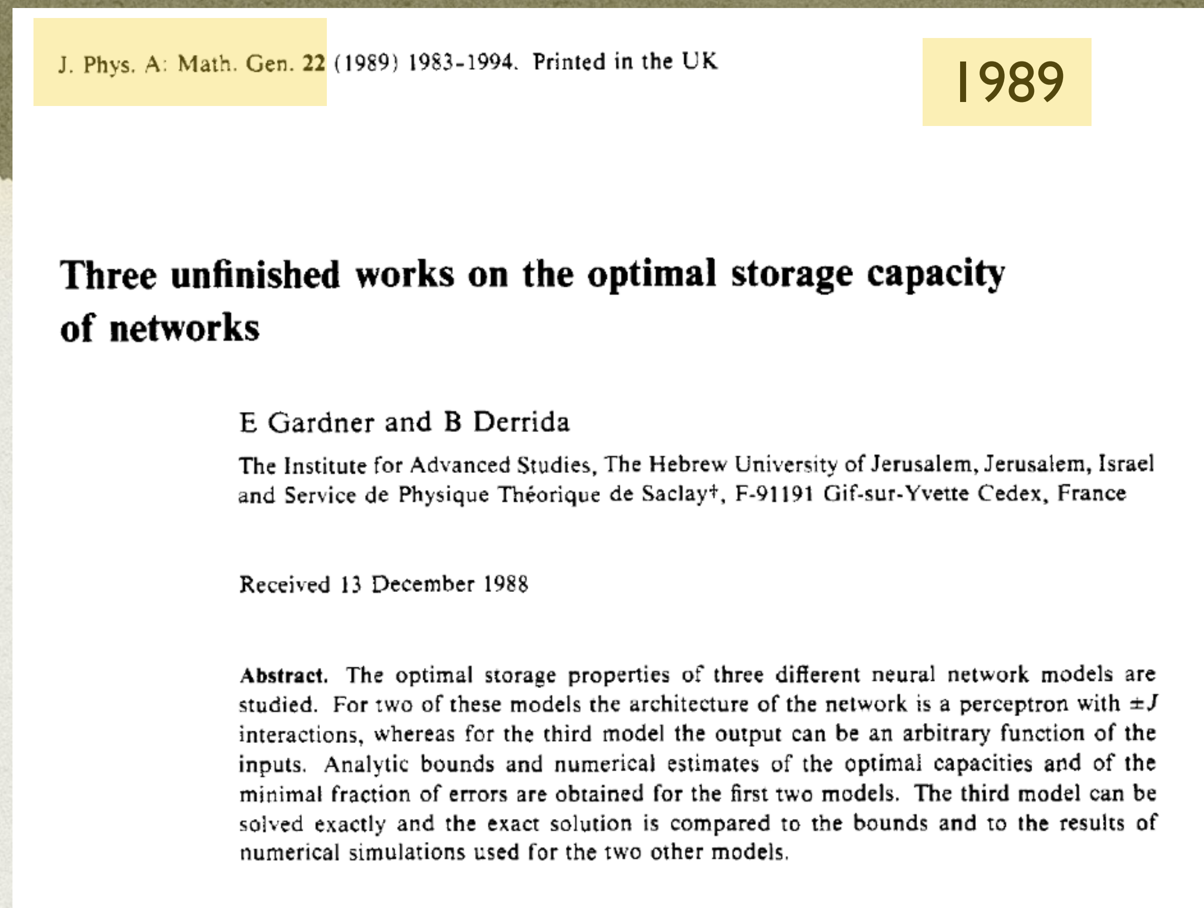


Student-network

- Observes X, y, the architecture of the network.
- How does the best achievable test 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

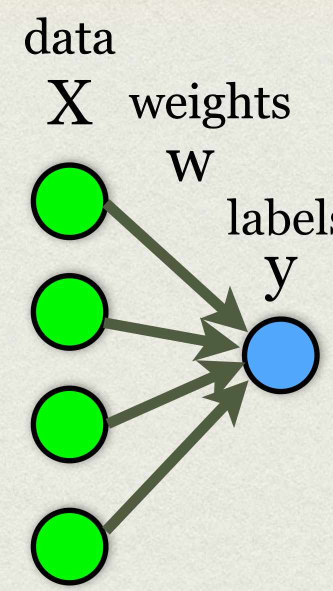
Three unfinished works on the optimal storage capacity of networks

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

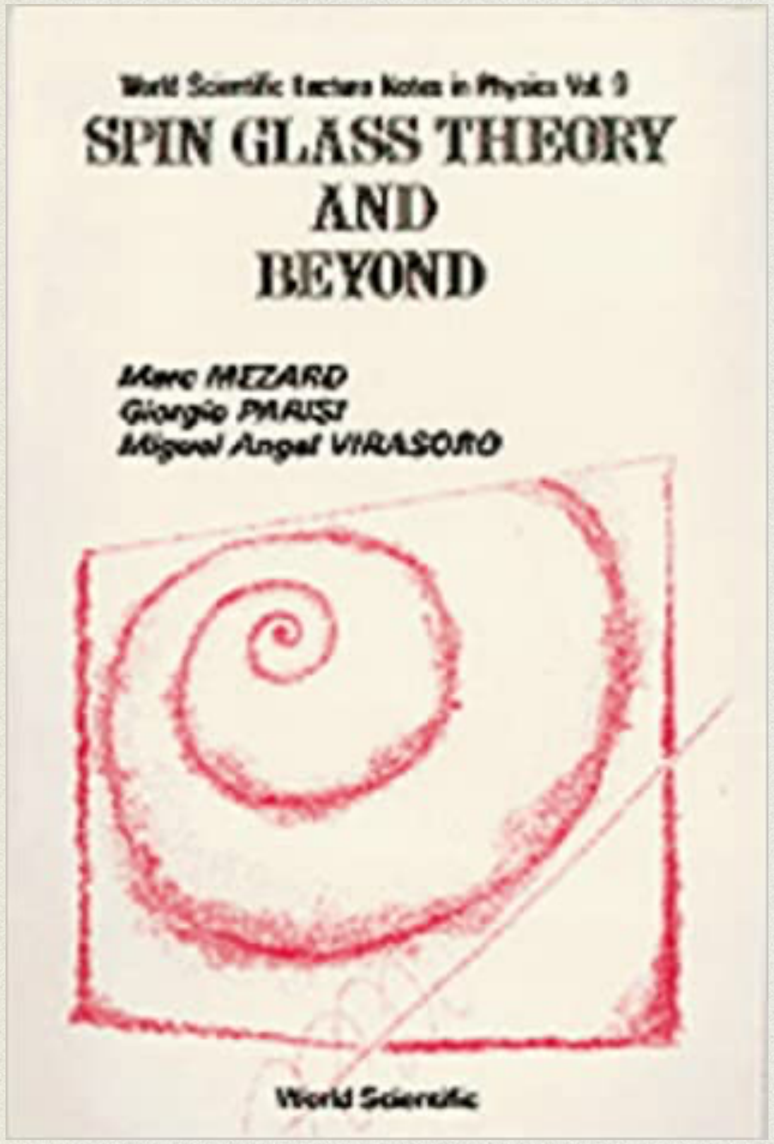
Received 13 December 1988

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.

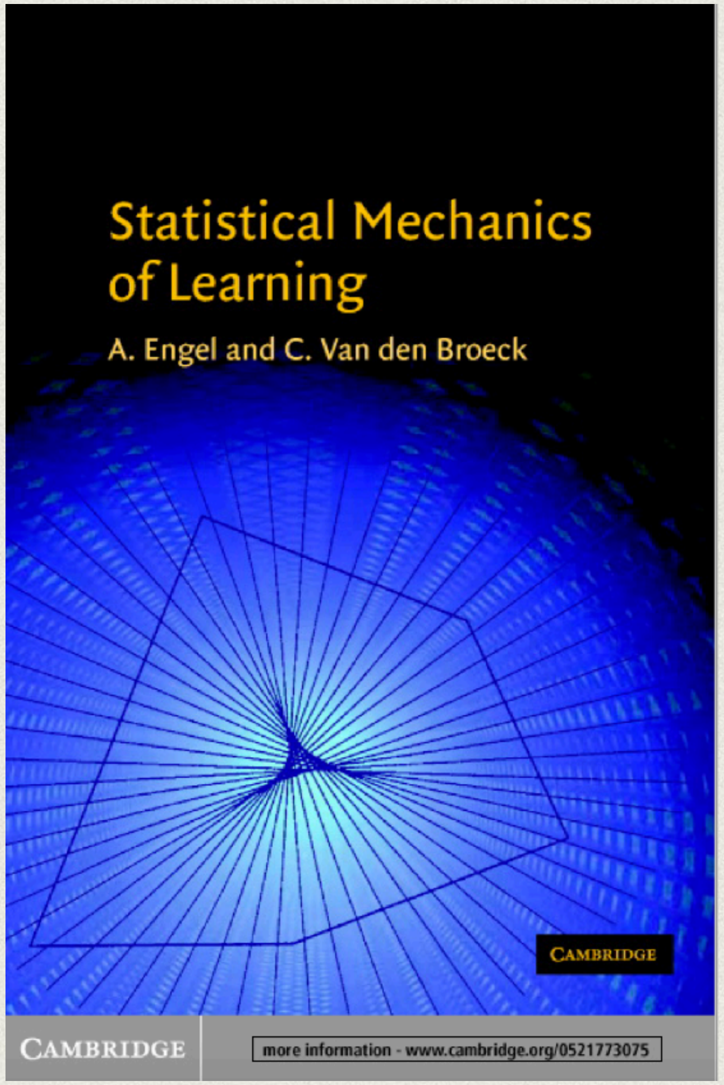


- Take random iid Gaussian $X_{\mu i}$, and random iid w_i^* from P_w.
- Create $y_\mu = \varphi\left(\sum_{i=1}^{d} X_{\mu i} w_i^*\right)$, e.g. $\varphi(z) = \text{sign}(z)$
- High-dimensional regime: $n \to \infty \quad d \to \infty$, $\alpha \equiv n/d = \Theta(1)$ — d dimensions, n samples

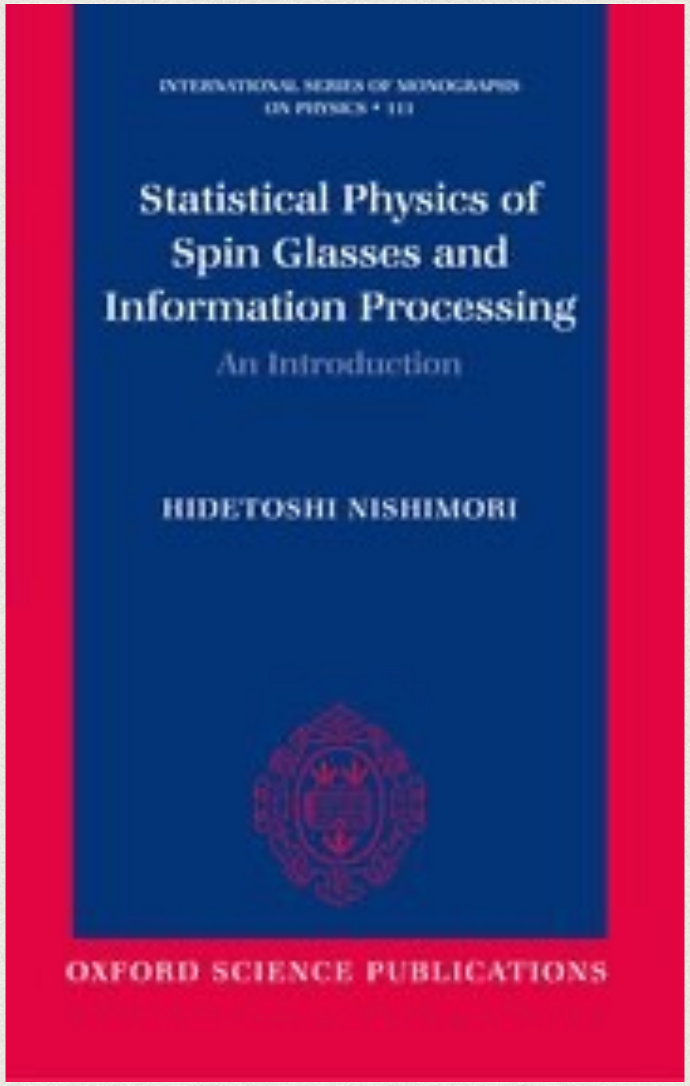
1987



2001



2001



BAYES-OPTIMAL PREDICTION

Posterior probability distribution:

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

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

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

BAYES VS RISK MINIMISATION

- Bayes-optimal estimation = marginals of the posterior:

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

- More common in ML: Empirical risk minimisation = minimisation of a loss function:

$$\min_w \Big[\sum_{\mu=1}^{n} \ell(y_\mu, X_\mu \cdot w) + \lambda \|w\|_2^2 \Big]$$

e.g. square loss $\ell(y, z) = (y - z)^2$, logistic loss $\ell(y, z) = \log_2(1 + e^{-yz})$

BAYES-OPTIMAL PERFORMANCE

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

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

Theorem 1:

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

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

where

$$\Phi_{P_w}(\hat{m}) \equiv \mathbb{E}_{z,w_0}\Big[\ln \mathbb{E}_w\Big(e^{\hat{m}ww_0 + \sqrt{\hat{m}}wz - \hat{m}w^2/2}\Big)\Big]$$

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

$w, w_0 \sim P_w \qquad z, v, \xi \sim \mathcal{N}(0,1) \qquad \rho = \mathbb{E}_{P_w}(w^2)$

BAYES-OPTIMAL PERFORMANCE

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

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

Theorem 1:

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

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

Theorem 2: Optimal generalisation error

$$\mathcal{E}_{\mathrm{test}} = \mathbb{E}_{v,\xi}\left[\varphi(\sqrt{\rho}v)^2\right] - \mathbb{E}_{v,z,\xi}\left[\varphi\left(\sqrt{m^*}v + \sqrt{\rho - m^*}z\right)\right]^2$$

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

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

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

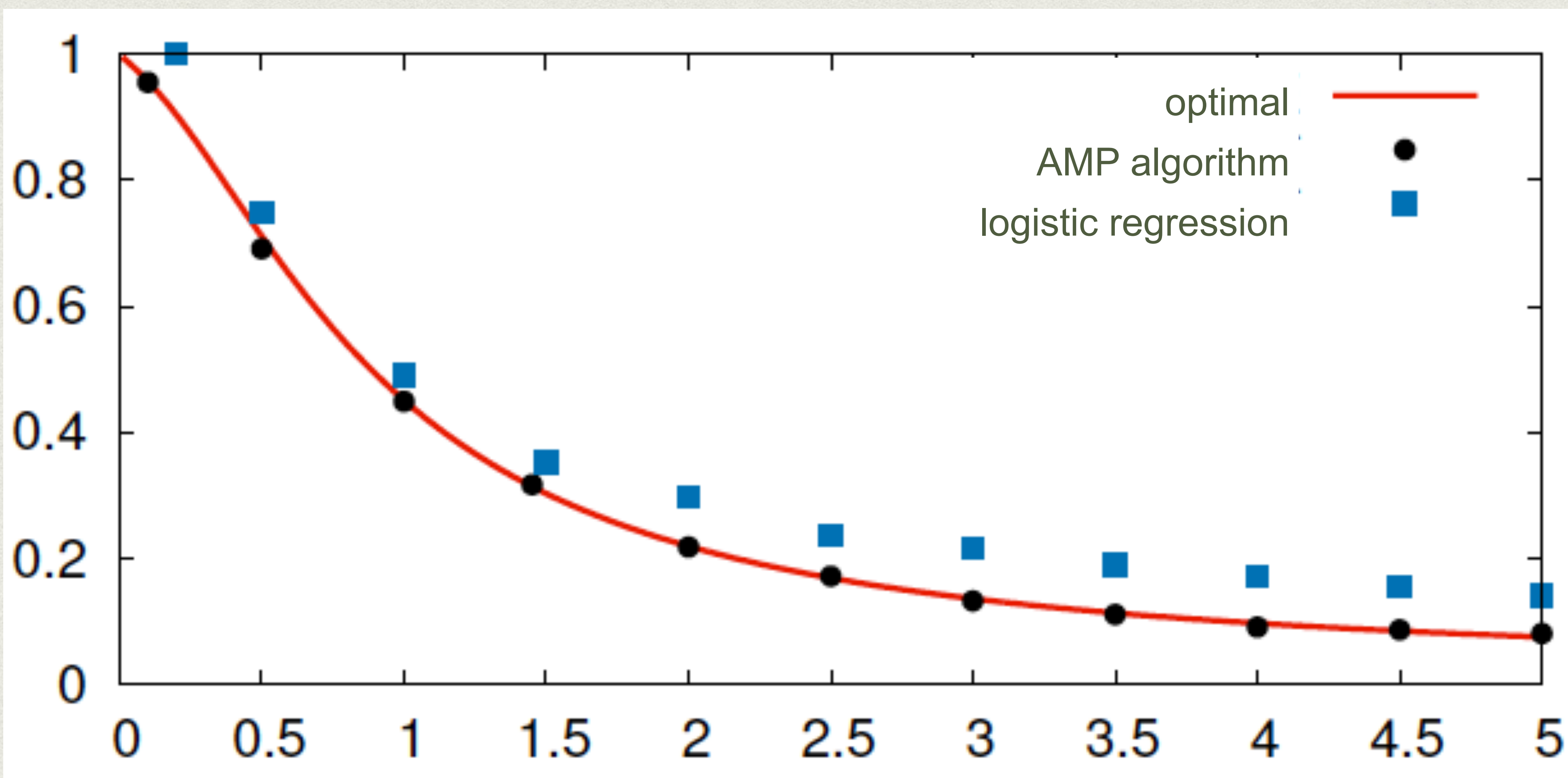
$\xi \sim P_\xi$

SPHERICAL PERCEPTRON

Data generated as: $X_{\mu i} \sim \mathcal{N}(0,1)$ iid

$$y_\mu = \mathrm{sign}\left(\sum_{i=1}^{d} X_{\mu i} w_i^*\right) \qquad P_{w^*} = \mathcal{N}(0,1)$$

test error



$n \to \infty$

$d \to \infty$

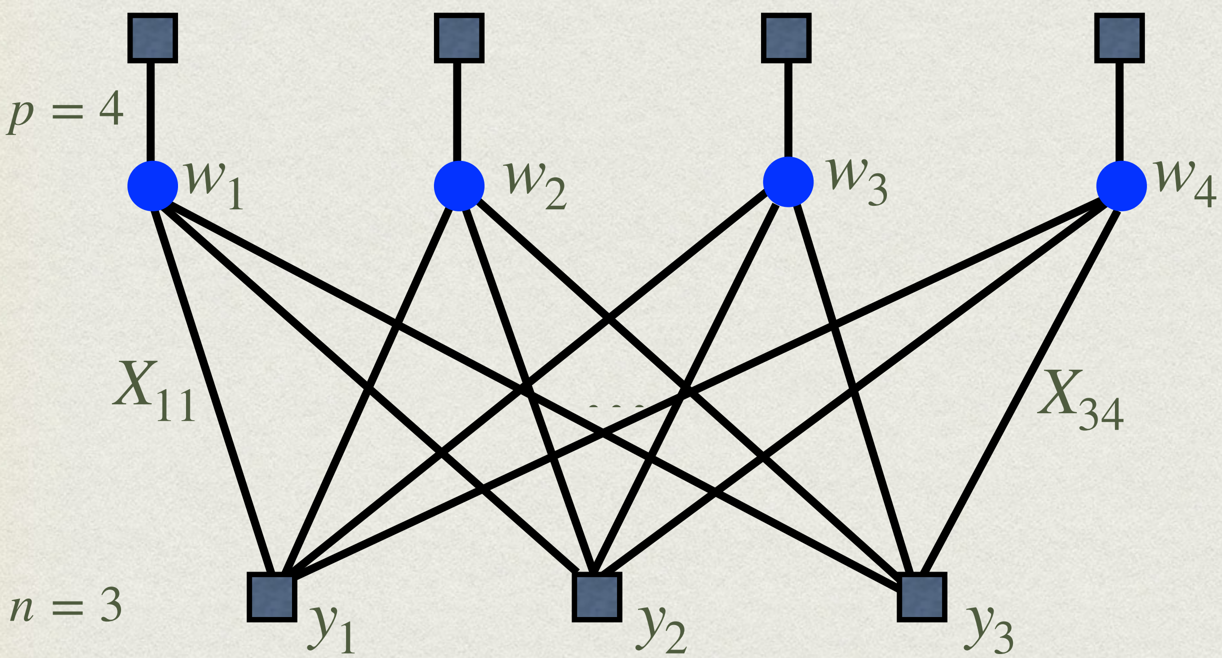
$n/d = \Theta(1)$

# of samples per dimension n/d

APPROXIMATE MESSAGE PASSING

Thouless-Anderson-Palmer'76, Mézard'89, Donoho, Maleki, Montanari'09, Rangan'10

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



Belief Propagation

$$m_{i\to\mu}(w_i) = \frac{1}{z_{i\to\mu}} P_w(w_i) \prod_{\gamma\neq\mu} m_{\gamma\to i}(w_i)$$

$$m_{\mu\to i}(w_i) = \frac{1}{z_{\mu\to i}} \int \prod_{j\neq i} [\mathrm{d}w_j m_{j\to\mu}(w_j)] P_{\text{out}}(y_\mu \,|\, \sum_l X_{\mu l} w_l)$$

The d-dimensional integral in BP is algorithmically intractable, but simplifies …

Algorithm 2 Generalized Approximate Message Passing (G-AMP)

Input: y

Initialize: $\mathbf{a}^0$,$\mathbf{v}^0$, t = 1 $g^0_{\text{out},\mu}$

repeat

AMP Update of ω_μ, V_μ

$$\boxed{\begin{aligned} V^t_\mu &\leftarrow \sum_i X^2_{\mu i} v^{t-1}_i \\ \omega^t_\mu &\leftarrow \sum_i X_{\mu i} a^{t-1}_i - V^t_\mu g^{t-1}_{\text{out}} \end{aligned}}$$

AMP Update of Σ_i, R_i and $g_{\text{out},\mu}$

$$\begin{aligned} \Sigma^t_i &\leftarrow \left[-\sum_\mu X^2_{\mu i} \partial_\omega g_{\text{out}}(\omega^t_\mu, y_\mu, V^t_\mu)\right]^{-1} \\ R^t_i &\leftarrow a^{t-1}_i + (\Sigma^{t+1}_i)^{-1} \sum_\mu X_{\mu i} g_{\text{out}}(\omega^t_\mu, y_\mu, V^t_\mu) \end{aligned}$$

AMP Update of the estimated marginals a_i, v_i

$$\begin{aligned} a^{t+1}_i &\leftarrow f_a(\Sigma_i, R^{t+1}_i,) \\ v^{t+1}_i &\leftarrow f_v(\Sigma_i, R^{t+1}_i) \end{aligned}$$

t ← t + 1

until Convergence on **a**,**v**

output: a,**v**.

Variances and means of the pre-activations

Simple to implement, only matrix multiplications, O(d²)

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

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

Algorithm 2 Generalized Approximate Message Passing (G-AMP)

Input: y

Initialize: $\mathbf{a}^0$,$\mathbf{v}^0$, $\mathrm{t}=1$ $g^0_{\mathrm{out},\mu}$

repeat

AMP Update of ω_μ, V_μ

$$V^t_\mu \leftarrow \sum_i X^2_{\mu i} v^{t-1}_i$$

$$\omega^t_\mu \leftarrow \sum_i X_{\mu i} a^{t-1}_i - V^t_\mu g^{t-1}_{\mathrm{out}}$$

AMP Update of Σ_i, R_i and $g_{\mathrm{out},\mu}$

$$\Sigma^t_i \leftarrow \left[-\sum_\mu X^2_{\mu i} \partial_\omega g_{\mathrm{out}}(\omega^t_\mu, y_\mu, V^t_\mu)\right]^{-1}$$

$$R^t_i \leftarrow a^{t-1}_i + (\Sigma^{t+1}_i)^{-1} \sum_\mu X_{\mu i} g_{\mathrm{out}}(\omega^t_\mu, y_\mu, V^t_\mu)$$

AMP Update of the estimated marginals a_i, v_i

$$a^{t+1}_i \leftarrow f_a(\Sigma_i, R^{t+1}_i,)$$

$$v^{t+1}_i \leftarrow f_v(\Sigma_i, R^{t+1}_i)$$

$\mathrm{t} \leftarrow \mathrm{t}+1$

until Convergence on **a**,**v**

output: a,**v**.

Variances and means of the pre-activations

Simple to implement, only matrix multiplications, O(d²)

Bayes-optimal prediction:

$$\boxed{\hat{y}^t_{\mathrm{new}} = \frac{1}{\sqrt{2\pi V^t}} \int \mathrm{d}z\, \mathrm{d}y\, y P_{\mathrm{out}}(y|z) e^{-\frac{1}{2V^t}\left(z - \sum_i F_{\mathrm{new},i} a^{t-1}_i\right)^2}}$$

STATE EVOLUTION

Bayati, Montanari'11, Bayati, Lelarge, Montanari'12, Javanmard, Montanari'13.

Define: $m^t \equiv \frac{1}{d}\sum_{i=1}^{d} w_i^* a_i^t$ then $\mathrm{MSE}(t) = \rho - m^t$

m^t in the AMP algorithm evolves as:

$$m^{t+1} = 2\partial_{\hat{m}}\Phi_{P_w}(\hat{m}^t)$$

$$\hat{m}^t = 2\alpha\partial_m\Phi_{P_{\mathrm{out}}}(m^t;\rho)$$

Recall the RS free energy

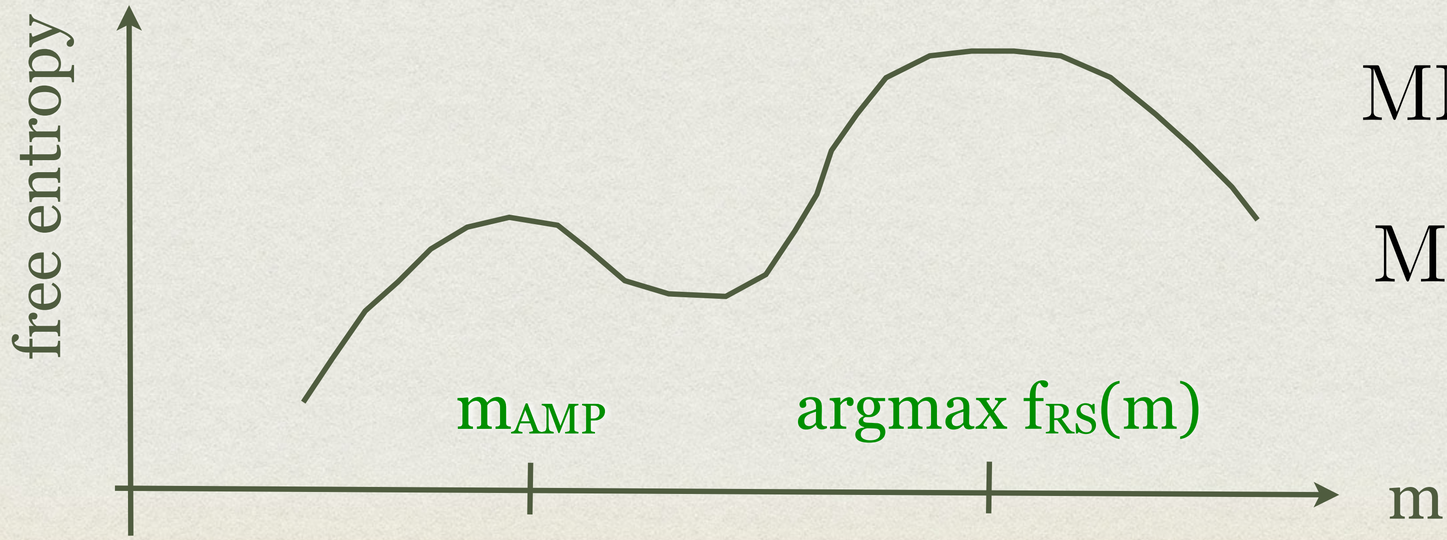
$$f_{\mathrm{RS}}(m,\hat{m}) = \Phi_{P_w}(\hat{m}) + \alpha\Phi_{P_{\mathrm{out}}}(m;\rho) - \frac{m\hat{m}}{2}$$

COROLLARY

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

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

- MMSE is given by the global maximum of the free entropy.
- AMP-MSE given by the local maximum of the free entropy reached starting from small m/large MSE.



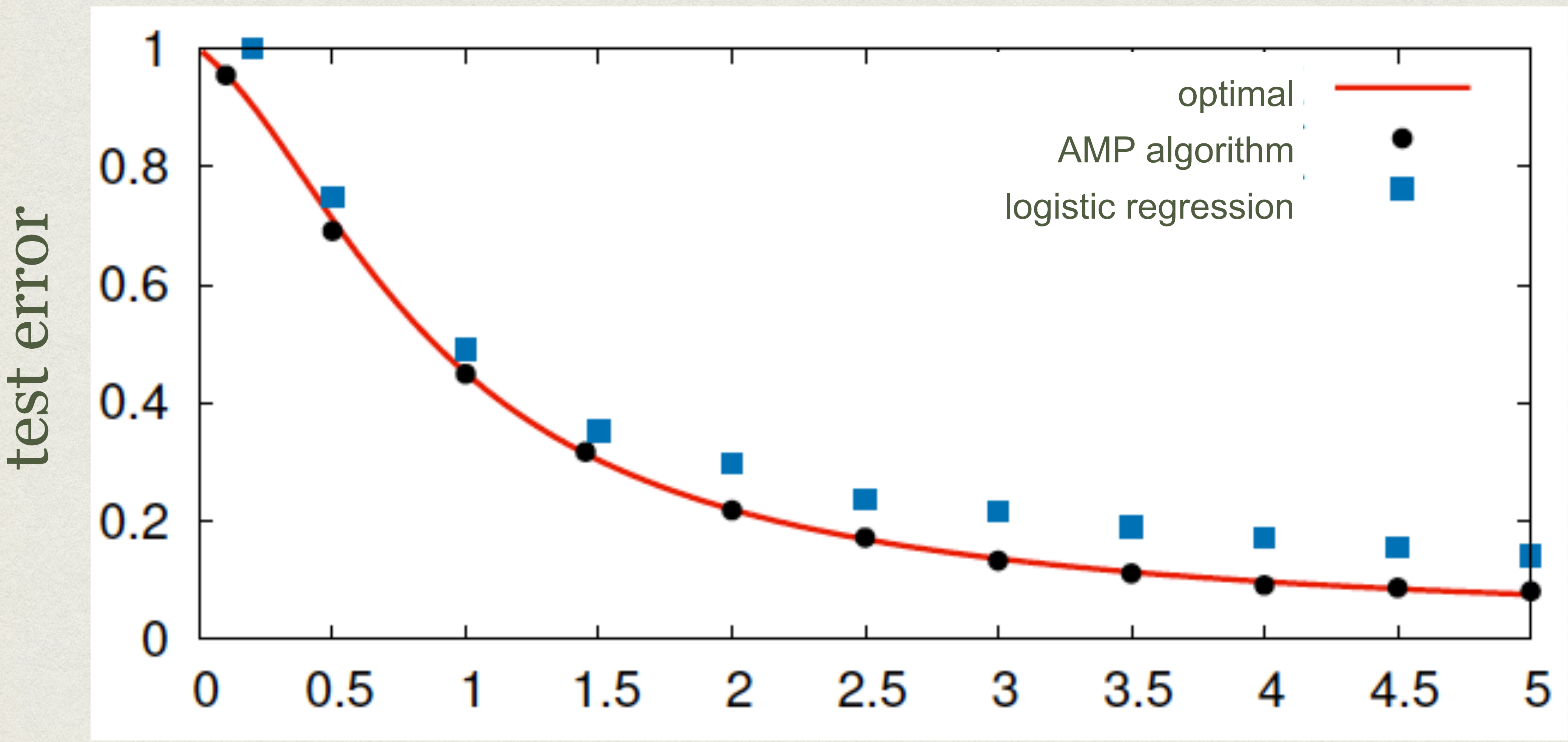
$$\mathrm{MMSE} = \rho - \mathrm{argmax} f_{RS}(m)$$

$$\mathrm{MSE}_{\mathrm{AMP}} = \rho - m_{\mathrm{AMP}}$$

SPHERICAL PERCEPTRON

Data generated as: $X_{\mu i} \sim \mathcal{N}(0,1)$ iid

$$y_\mu = \operatorname{sign}\left(\sum_{i=1}^{d} X_{\mu i} w_i^*\right) \qquad P_{w^*} = \mathcal{N}(0,1)$$



# of samples per dimension $\quad n/d$

$n \to \infty$

$d \to \infty$

$n/d = \Theta(1)$

BAYES VS LOGISTIC REGRESSION

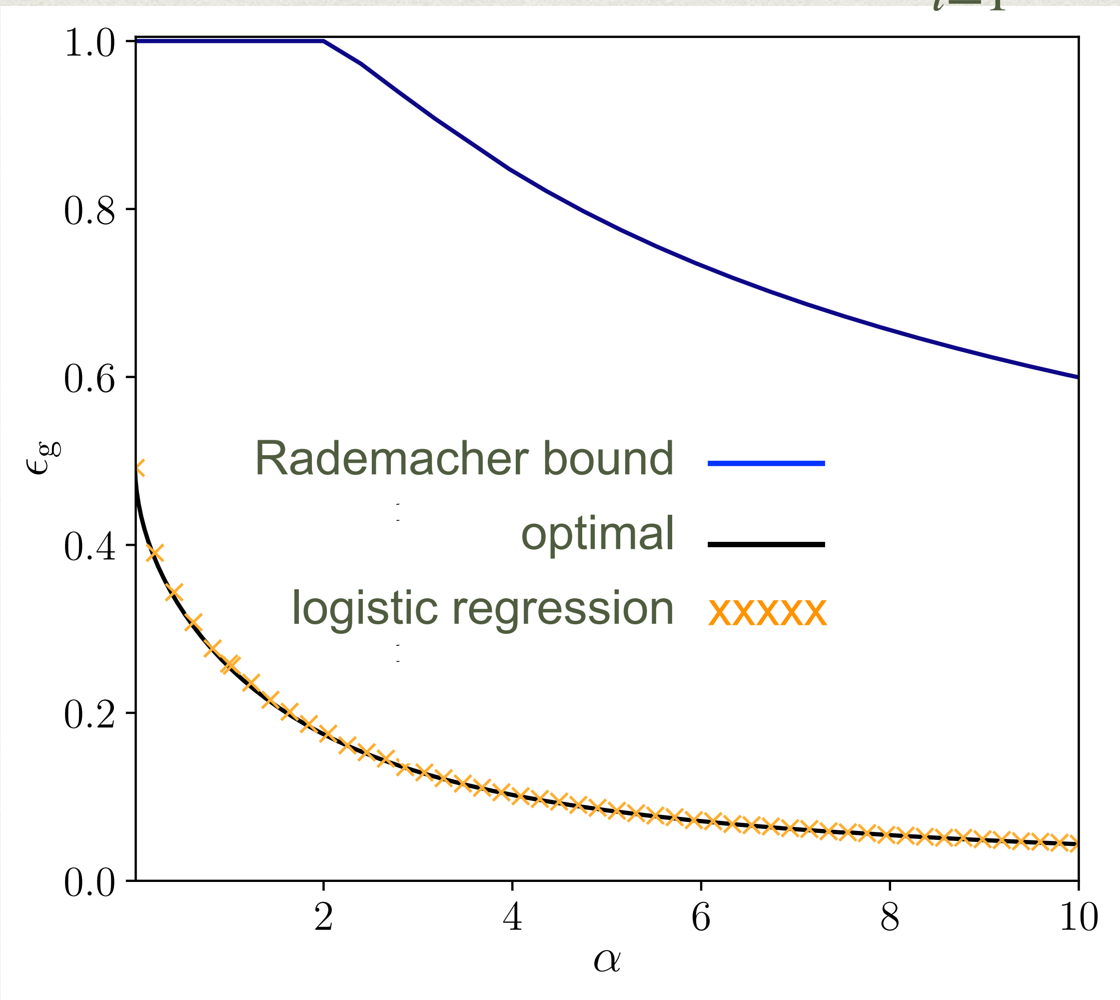
Aubin, Krzakala, Lu, LZ; NeurIPS'20, arXiv:2006.06560

Data generated as: $X_{\mu i} \sim \mathcal{N}(0,1)$ iid $\qquad P_{w^*} = \mathcal{N}(0,1)$

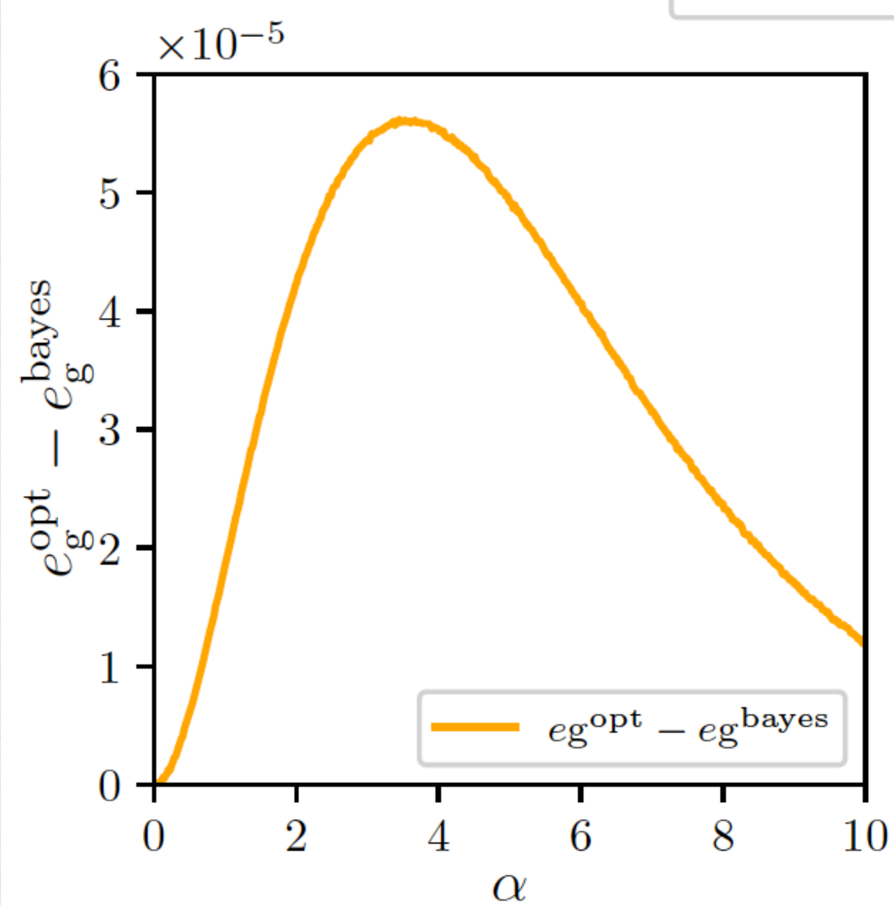
$$y_\mu = \mathrm{sign}\left(\sum_{i=1}^{d} X_{\mu i} w_i^*\right)$$

Here: Optimally regularized logistic regression essentially Bayes-optimal

test error



# of samples per dimension



ANOTHER EXAMPLE OF THE TEACHER-STUDENT SETTING (NON-CONVEX THIS TIME)

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.

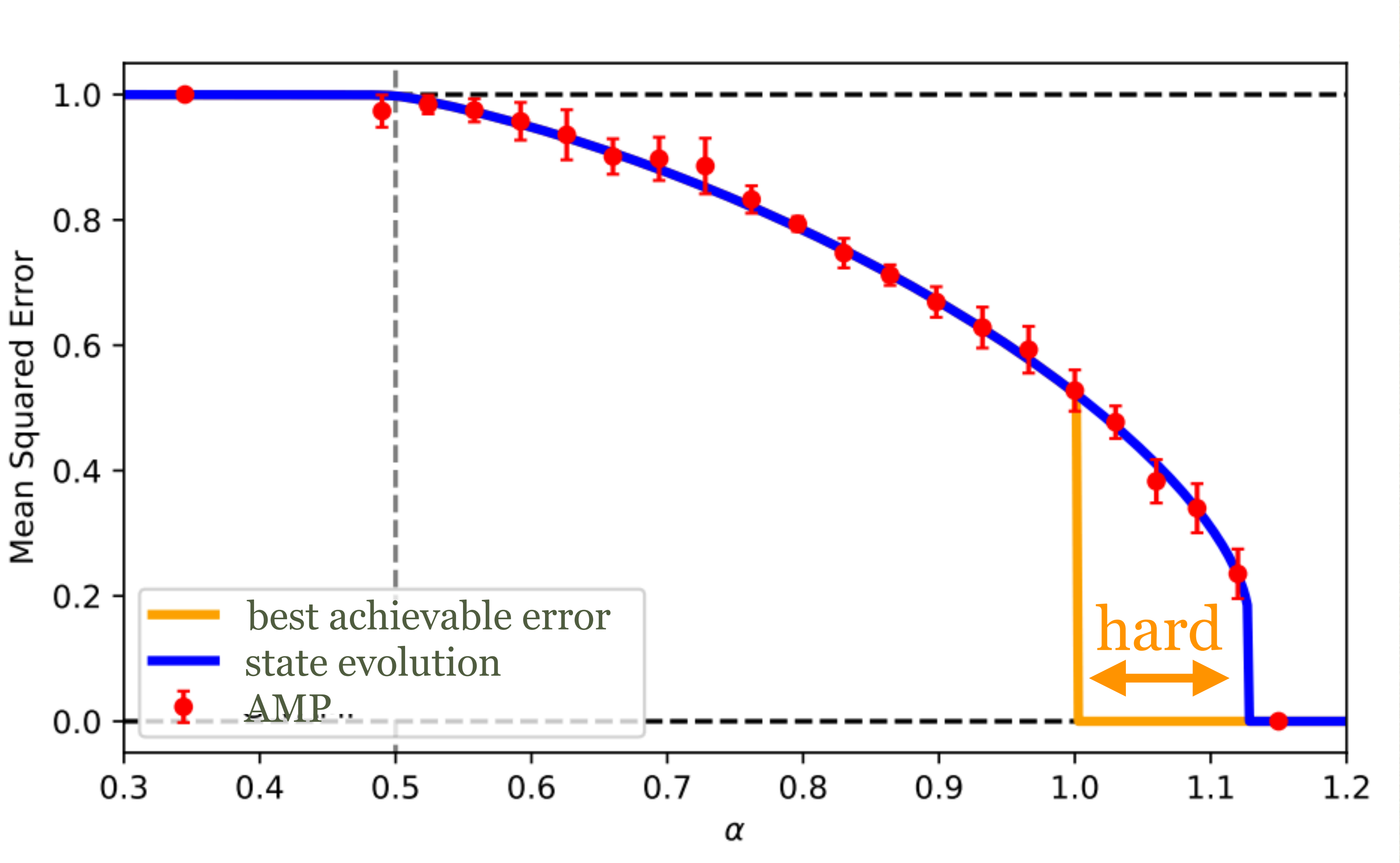
$$X_{\mu i} \sim \mathcal{N}(0,1/d) \qquad w_i^* \sim \mathcal{N}(0,1) \qquad \begin{aligned} \mu &= 1,\ldots,n \\ i &= 1,\ldots,d \end{aligned}$$

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

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

OPTIMAL PHASE RETRIEVAL

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



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

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

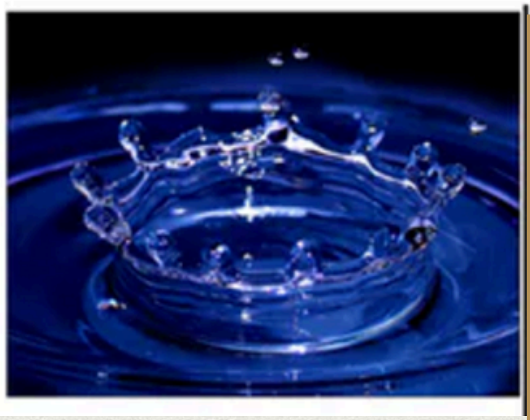
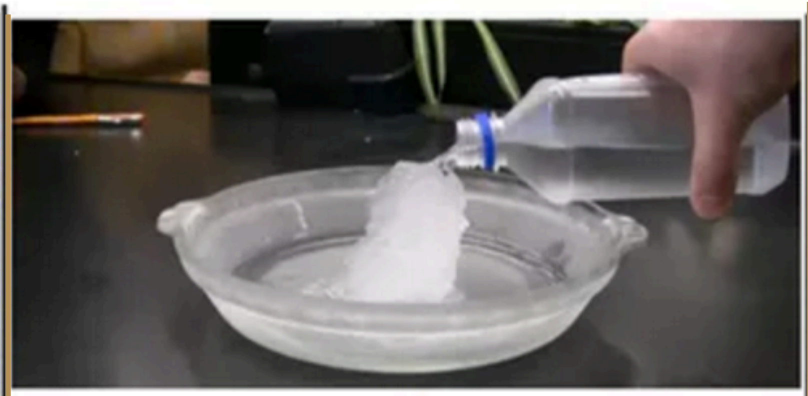
$$X_{\mu i}, w_i \in \mathbb{R}$$

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

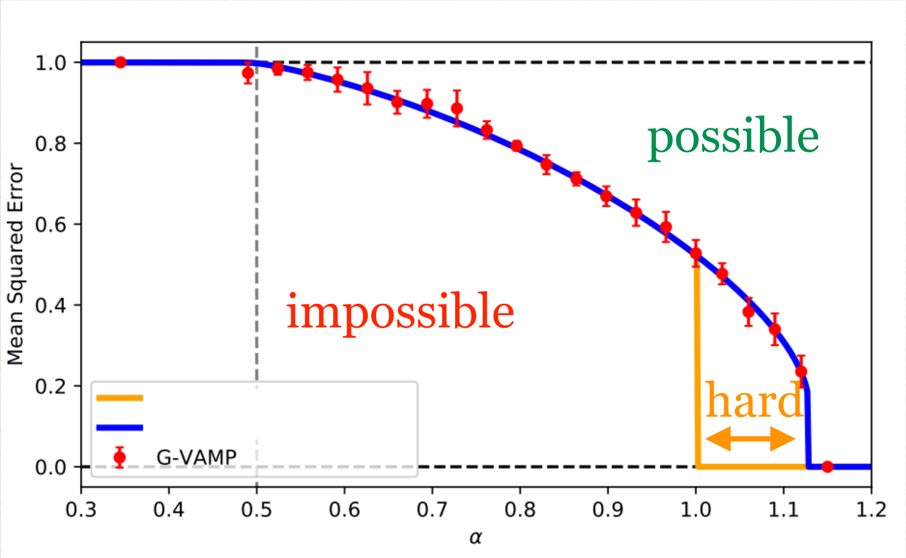
$\alpha_{\mathrm{IT}} = 1$ # of samples needed for perfect generalisation for any algorithm.

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

PHYSICS VS LEARNING


liquid	supercooled liquid	ice
impossible	computationally hard	possible



Presence of the hard phase signals hurdles for gradient-based algorithms including Langevin.

PHYSICAL REVIEW X

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

ABSTRACT

An algorithmically hard phase is described in a range of inference problems: Even if the signal can be reconstructed with a small error from an information-theoretic point of view, known algorithms fail unless the noise-to-signal ratio is sufficiently small. This *hard phase* is typically understood as a metastable branch of the dynamical evolution of message-passing algorithms. In this work, we study the metastable branch for a prototypical inference problem, the low-rank matrix factorization, that

EMPIRICAL RISK MINIMIZATION FOR PHASE RETRIEVAL

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

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

Gradient flow: $$\dot{w}_i(t) = -\,\partial_{w_i} \mathscr{L}\Big(\{ w_j(t) \}_{j=1}^{d} \Big) + \mu(t) w_i(t)$$

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

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

A non-convex optimisation problem.

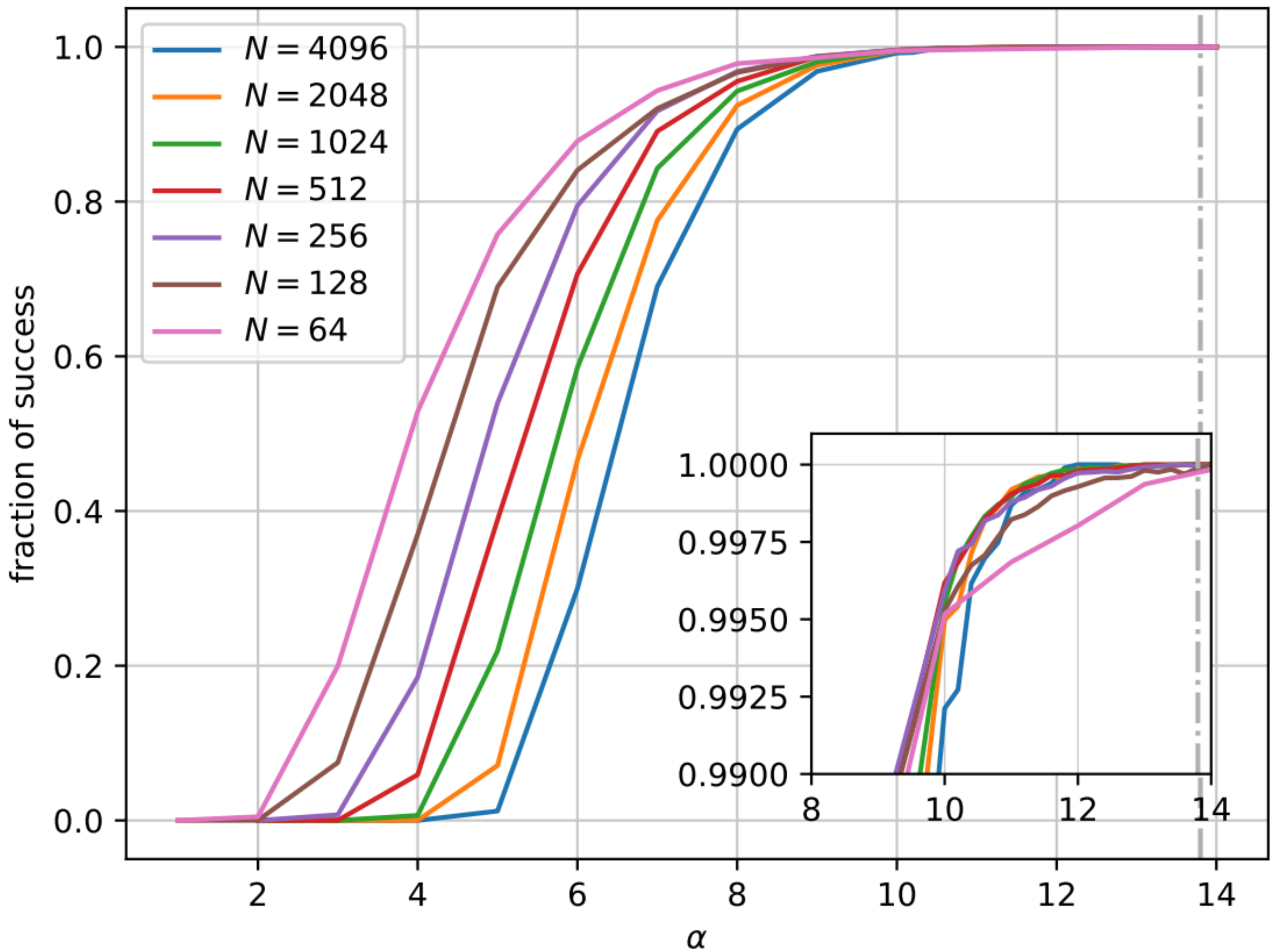
PERFORMANCE OF GRADIENT DESCENT

Number of samples to reach zero test error in phase retrieval:

GRADIENT DESCENT NUMERICALLY

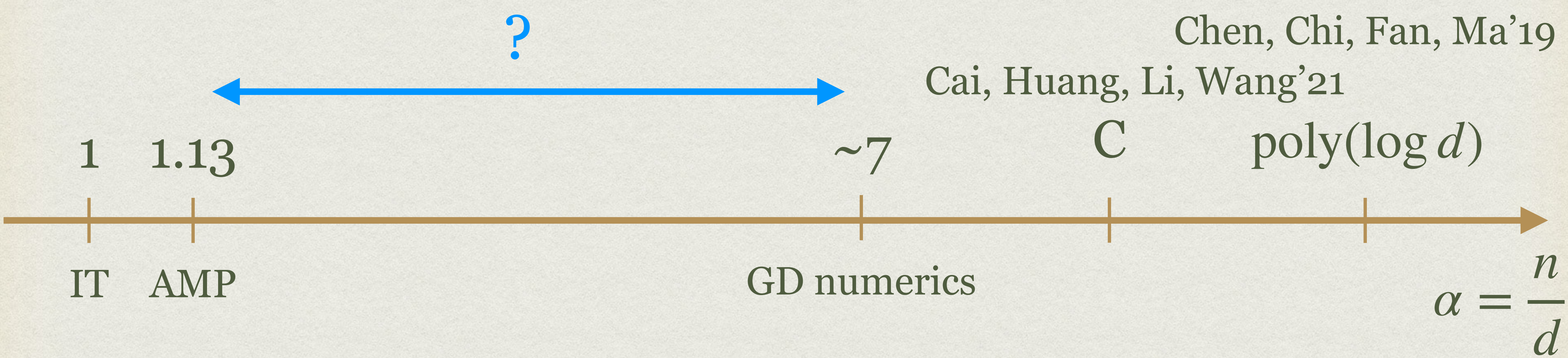
Sarao Mannelli, Biroli, Cammarota, Krzakala, LZ, NeurIPS 2020, 2006.06997.

$N = d$



PERFORMANCE OF GRADIENT DESCENT

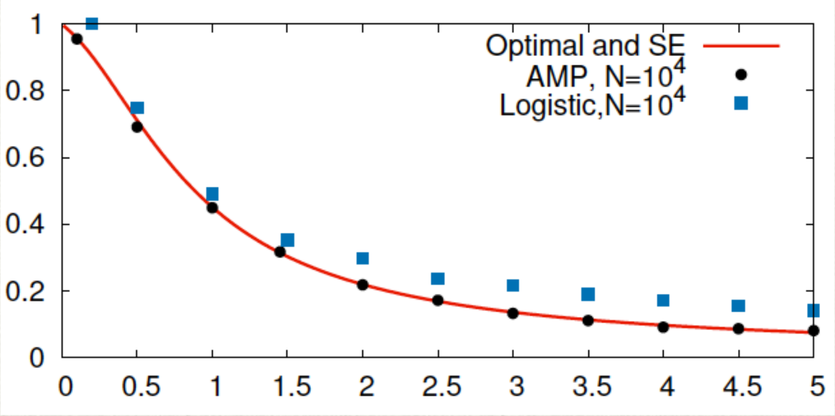
Closing the gap between GD and AMP?



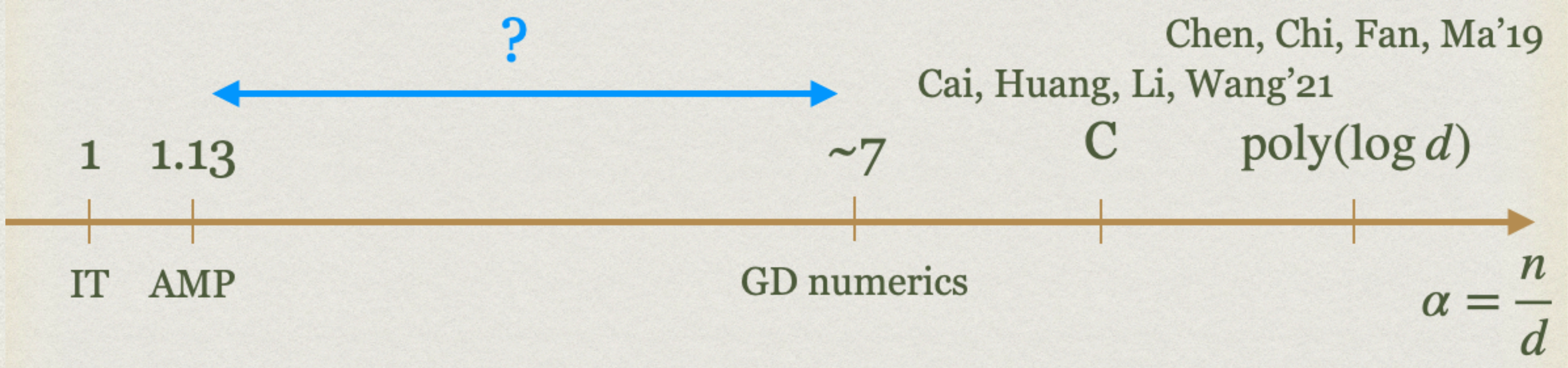
RECAP SO FAR

Two simple teacher-student examples:

- Spherical perceptron ERM close to Bayes-optimal.



- Phase retrieval ERM way worse than Bayes-optimal.

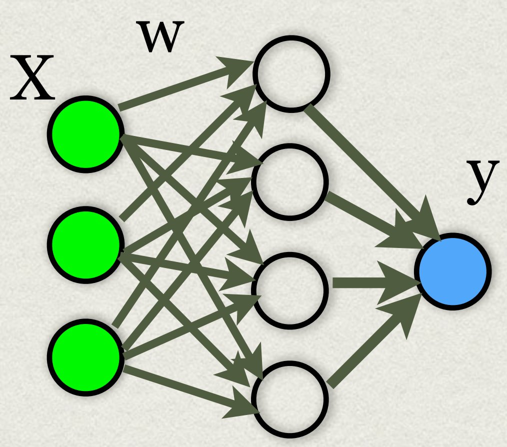


WHAT IS MISSING?

DEEP LEARNING IS OVER-PARAMETRIZED

OVER-PARAMETRIZED ERM FOR PHASE RETRIEVAL

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



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

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

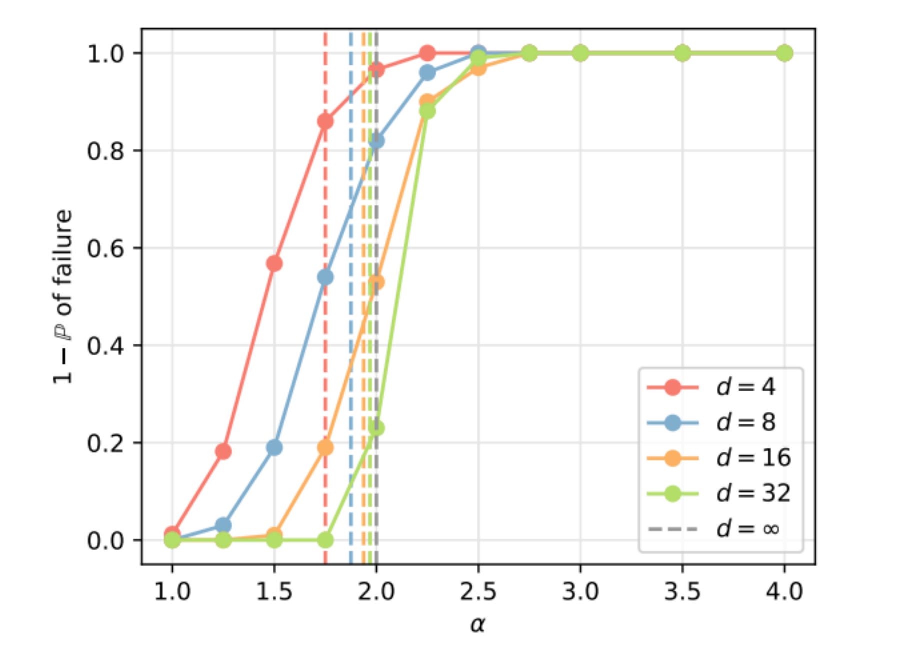
Sarao Mannelli, Vanden-Eijnden, LZ, NeurIPS'20, 2006.15459

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*

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

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

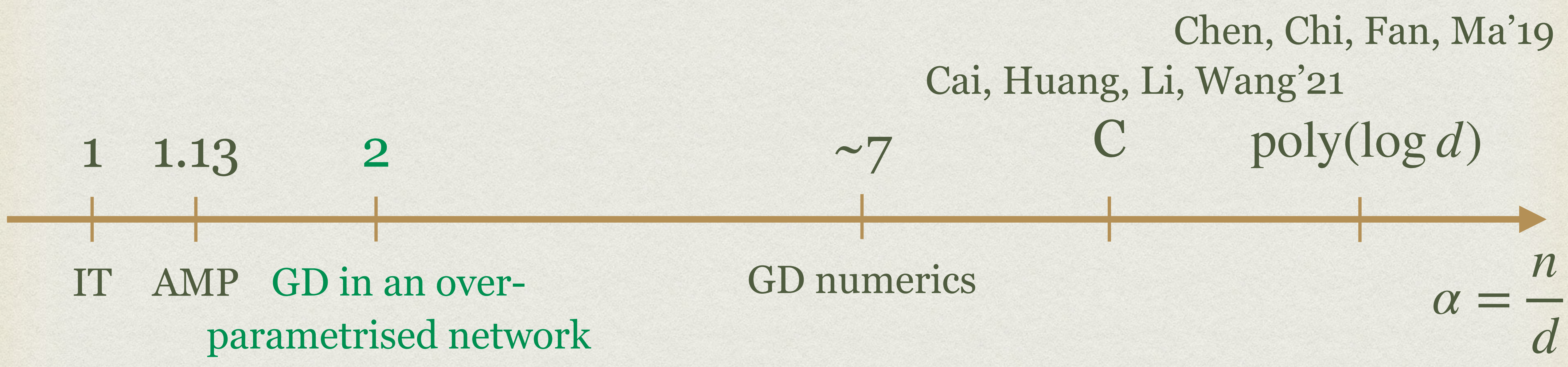
$$E_n(A_\infty) = 2L_n(\boldsymbol{w}_1^\infty, \ldots, \boldsymbol{w}_n^\infty) = 0. \tag{16}$$



PERFORMANCE OF GRADIENT DESCENT

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

Over-parametrised neural networks trained by gradient descent need fewer samples to learn phase retrieval



WHAT IS MISSING?

DEEP LEARNING USES STOCHASTIC GRADIENT DESCENT

PERSISTENT SGD

Mignaco, Urbani, Krzakala, LZ, NeurIPS 2020, 2006.06098

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

SGD	Persistent-SGD
$s_\mu(t) = \begin{cases} 1 & \text{w.p. } b \\ 0 & \text{w.p. } 1-b \end{cases}$	$\text{Prob}\left(s_\mu(t+\eta) = 1 \mid s_\mu(t) = 0\right) = \frac{\eta}{\tau}$ PERSISTENCE TIME of each sample $\text{Prob}\left(s_\mu(t+\eta) = 0 \mid s_\mu(t) = 1\right) = \frac{1-b}{b\tau}\eta$
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 generalized to the persistent stochastic GD and the planted model:

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

$\xrightarrow{\text{DMFT}}$

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

$$\partial_t h(t) = -\underbrace{\tilde{\nu}(t)}_{\text{eff. regularisation}} h(t) - \underbrace{s(t)\partial_1 v(\tilde{h}(t); h_0)}_{\text{stochastic noise}} + \int_0^t \underbrace{dt'\, M_R(t,t')h(t')}_{\text{memory}} + \underbrace{\chi(t)}_{\text{Gauss noise}}$$

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

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

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

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

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

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

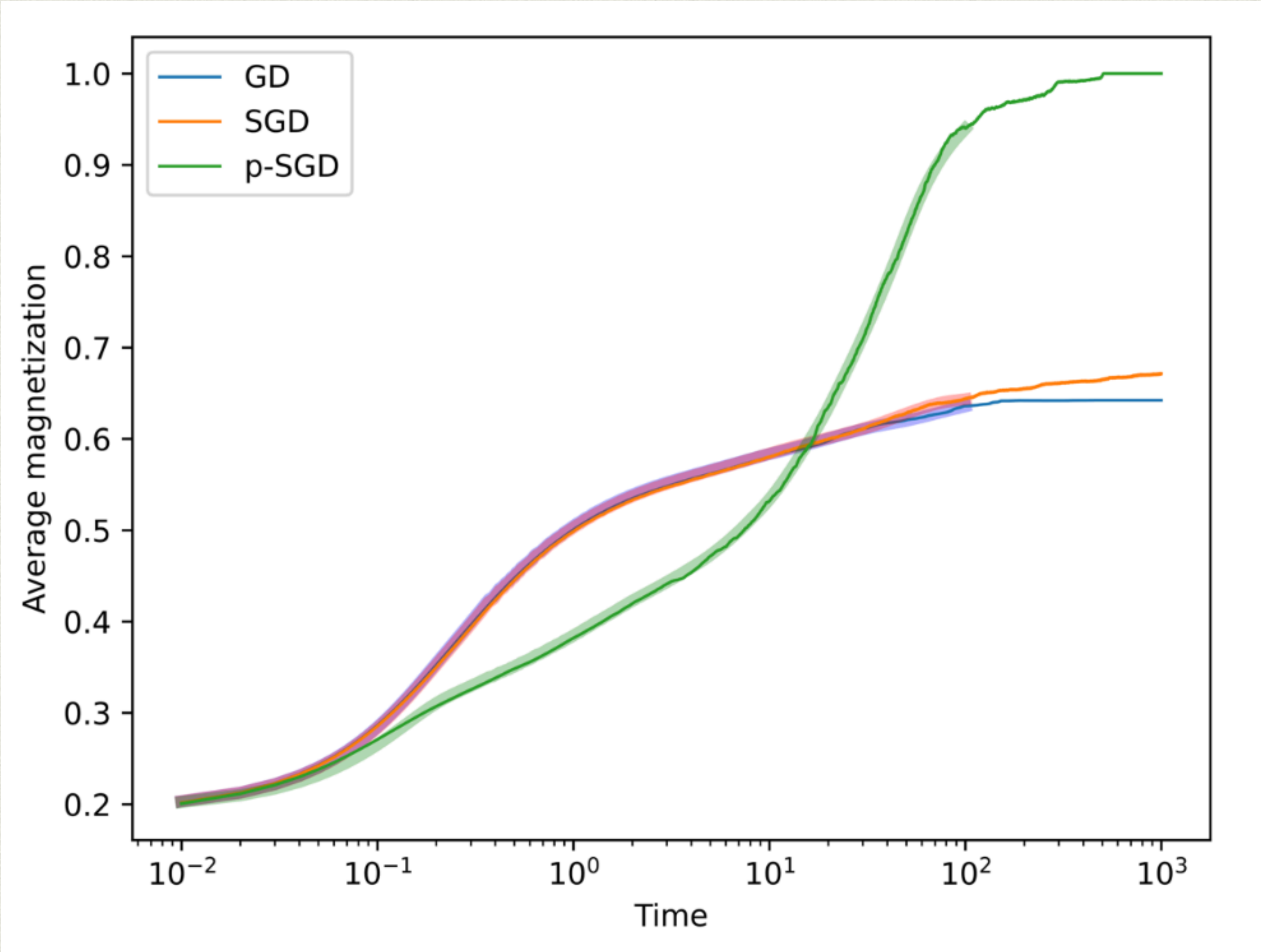
$$\delta\nu(t) = \frac{\alpha}{b}\langle s(t)\partial_1^2 v(\tilde{h}(t);h_0)\rangle \qquad \mu(t) = \frac{\alpha}{b}\langle s(t)h_0\partial_1 v(\tilde{h}(t);h_0)\rangle$$

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

Persistent-SGD better than GD or SGD

Mignacco, Urbani, LZ; MLST'21, 2103.04902.

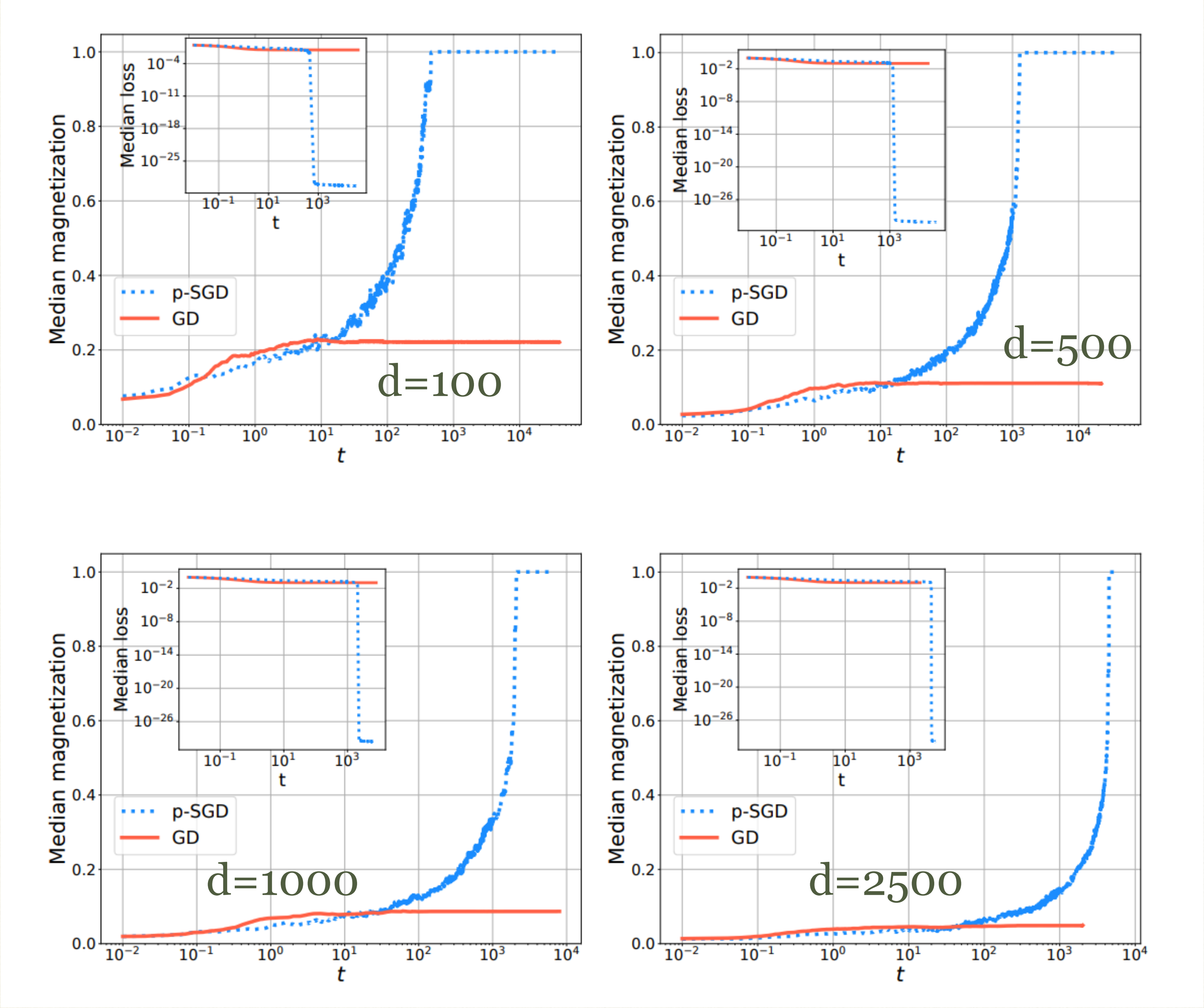
$\alpha = 3.0$
$\eta_{\text{SGD}} = 0.01$
$b = 0.5, \tau = 1$



From Mignaco, Urbani, LZ, 2103.04902; DMFT from E. Troiani master thesis.

P-SDG WITH RANDOM START

Mignacco, Urbani, LZ; MLST'21, 2103.04902.



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

$\alpha = 2.5$

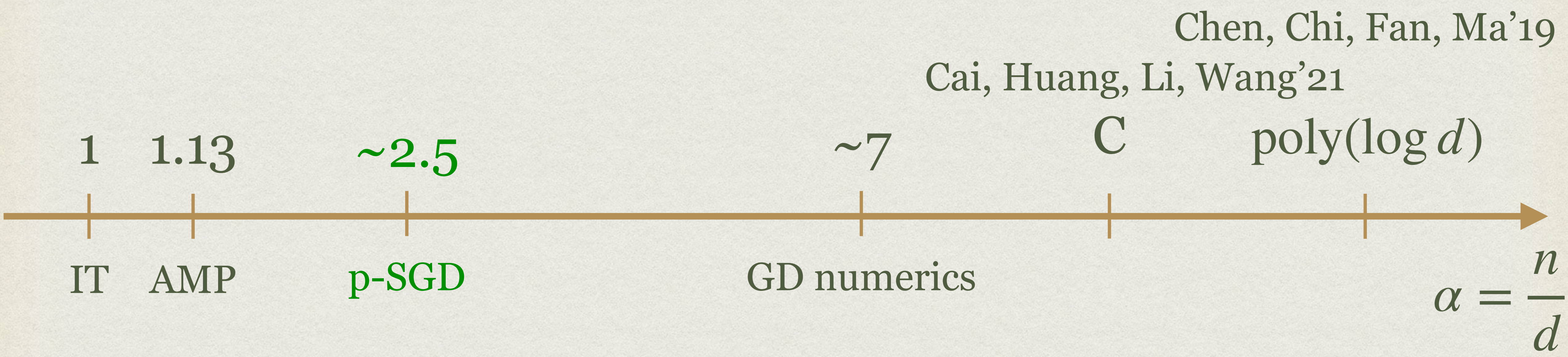
$\eta_{\rm SGD} = 0.01$

$b = 0.5, \tau = 2$

PERFORMANCE OF SGD IN PHASE RETRIEVAL

Mignacco, Urbani, LZ; MLST'21, 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 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 non-trivial flow limit, analysable by DMFT, performing better than SGD (without hidden units).

OPEN QUESTIONS

- 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.