Complex Dynamics in Simple Neural Networks: Understanding Gradient Flow in Phase Retrieval

Sarao Mannelli Stefano, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Pierfrancesco Urbani, Lenka Zdeborová

Complex Dynamics in Simple Neural Networks: Understanding Gradient Flow in Phase Retrieval

Stefano Sarao Mannelli[1], **Giulio Biroli**[2], **Chiara Cammarota**[3,4], **Florent Krzakala**[5], **Pierfrancesco Urbani**[1], **and Lenka Zdeborová**[6]

Abstract

Despite the widespread use of gradient-based algorithms for optimizing high-dimensional non-convex functions, understanding their ability of finding good minima instead of being trapped in spurious ones remains to a large extent an open problem. Here we focus on gradient flow dynamics for phase retrieval from random measurements. When the ratio of the number of measurements over the input dimension is small the dynamics remains trapped in spurious minima with large basins of attraction. We find analytically that above a critical ratio those critical points become unstable developing a negative direction toward the signal. By numerical experiments we show that in this regime the gradient flow algorithm is not trapped; it drifts away from the spurious critical points along the unstable direction and succeeds in finding the global minimum. Using tools from statistical physics we characterize this phenomenon, which is related to a BBP-type transition in the Hessian of the spurious minima.

1 Introduction

In many machine learning applications one optimizes a non-convex loss function; this is often achieved using simple descending algorithms such as gradient descent or its stochastic variations. The positive results obtained in practice are often hard to justify from the theoretical point of view, and this apparent contradiction between non-convex landscapes and good performance of simple algorithms is a recurrent problem in machine learning.

A successful line of research has studied the geometrical properties of the loss landscape, distinguishing between good minima - that lead to good generalization error - and spurious minima - associated with bad generalization error. The results showed that in some regimes, for several problems from matrix completion [1] to wide neural networks [2, 3], spurious minima disappear and consequently under weak assumptions [4] gradient descent will converge to good minima. However, these results do not justify numerous other results showing that good and spurious minima are present, but systematically gradient descent works [5, 6]. In [7] it was theoretically shown that in a toy model - the spiked matrix-tensor model - it is possible to find good minima with high probability in a regime where exponentially many spurious minima are provably present. In [8] it was shown that this is due to the presence of the so-called threshold states in the landscape, that play a key role in the

1. Université Paris-Saclay, CNRS, CEA, Institut de physique théorique, 91191, Gif-sur-Yvette,France.
2. Laboratoire de Physique de l'Ecole normale supérieure ENS, Université PSL, CNRS, Sorbonne Université, Université Paris-Diderot, Sorbonne Paris Cité Paris, France
3. Dipartimento di Fisica, Sapienza Università di Roma, P.le A. Moro 5, 00185 Rome, Italy
4. Department of Mathematics, King's College London, Strand London WC2R 2LS, UK
5. IdePHICS laboratory, EPFL, Switzerland
6. SPOC laboratory, EPFL, Switzerland

Preprint. Under review.

dynamics of the gradient flow [9, 10]: at first attracting it, and successively triggering the converge towards lower minima under certain conditions [11, 12]. However, the spiked matrix-tensor model is an unsupervised learning model and it remained open whether the picture put forward in [7, 8] happens also in learning with neural networks.

In this work we thus study learning with a simple single-layer neural network on data stemming from the well-known phase retrieval problem - that consists of reconstructing a hidden vector having access only to the absolute value of its projection onto random directions. The problem emerges naturally in a variety of imaging applications where the intensity is easier to access than the phase [13–16] but it appears also in acoustics [17] and quantum mechanics [18]. The phase retrieval problem considered here leads to a high-dimensional and non-convex optimization problem defined as follows.

Phase retrieval. Consider αN N-dimensional sensing vectors $\boldsymbol{x}_m$ with unitary norm and generated according to a centered Gaussian distribution, and the true labels y_m^2 with $y_m = \langle \boldsymbol{x}_m, \boldsymbol{W}^* \rangle$ and $\boldsymbol{W}^*$ an unknown teacher-vector (the signal in the phase retrieval literature) from $\mathbb{S}^{N-1}(\sqrt{N})$. Given the dataset, the goal is to build an estimator $\boldsymbol{W}$ of $\boldsymbol{W}^*$ by minimizing the loss function

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2}\sum_{m=1}^{\alpha N} \left(\langle \boldsymbol{x}_m, \boldsymbol{W} \rangle^2 - \langle \boldsymbol{x}_m, \boldsymbol{W}^* \rangle^2 \right)^2 = \frac{1}{2}\sum_{m=1}^{\alpha N} \ell(\hat{y}_m, y_m), \tag{1}$$

with $\hat{y}_m = \langle \boldsymbol{x}_m, \boldsymbol{W} \rangle$ and $\ell(\hat{y}, y) = (\hat{y}^2 - y^2)^2$ is a modified square loss commonly used in the literature [19, 20] that ensures a smoother landscape compared to the square loss with the absolute values. The loss function, Eq. (1), is minimized using gradient-descent flow on the sphere starting from random initialization

$$\begin{aligned} \dot{\boldsymbol{W}}(t) &= -\nabla_{\boldsymbol{W}} \mathcal{L}(t) + \mu(t) \boldsymbol{W}(t), \\ W_i(t=0) &\sim \mathcal{N}(0,1) \quad \forall\, i, \end{aligned} \tag{2}$$

with $\mu(t)$ the Lagrange multiplier that enforces the spherical constraint during the dynamics. The value of $\mu(t)$ can be readily evaluated by taking the scalar product of the gradient of the loss with $\boldsymbol{W}$ and dividing by N

$$\mu(t) = \frac{1}{2N}\sum_{m=1}^{\alpha N} \frac{\partial}{\partial \hat{y}_m} \ell(\hat{y}_m(t), y_m) \hat{y}_m(t). \tag{3}$$

We can finally define the Hessian of the problem, denoting $\delta_{i,j}$ the Kronecker delta, it reads

$$\mathcal{H}_{i,j}(\boldsymbol{W}) = \frac{1}{2}\sum_{m=1}^{\alpha N} \frac{\partial^2}{\partial \hat{y}_m^2} \ell(\hat{y}_m, y_m) x_{m,i} x_{m,j} - \mu \delta_{i,j}. \tag{4}$$

Related work and our main contributions.

Numerous algorithms can be applied to achieve a good reconstruction of the hidden signal (teacher vector) [19, 21]. For random i.i.d. data and labels generated by a teacher the information-theoretically optimal generalization error has been analyzed in [22], showing that for $\alpha < \alpha_{\rm WR} = 0.5$ in the limit of $N \to \infty$ no estimator is able to obtain a generalization error better than a random guess. On the other hand for $\alpha > \alpha_{\rm IT} = 1$ algorithms (not necessarily polynomial ones) exist that are able to achieve zero generalization error. While the weak-recovery threshold is achievable with efficient algorithms [23], the information-theoretic threshold is conjectured not to be and an algorithmic gap has been conjectured to exist, with perfect recovery achievable with the approximate message passing algorithm only for $\alpha > \alpha_{\rm alg} = 1.13$ [22].

In the present paper we are interested in the algorithmic performance of the gradient descent algorithm. The motivation of the present work is not to compare vanilla gradient descent to other algorithms but rather use this well studied phase-retrieval problem to get insights on the properties of the gradient descent algorithm in high-dimensional non-convex optimization. The spurious minima are known to disappear in phase retrieval if the number of samples in the dataset scales as $O(N \log^3 N)$ with N the input dimension [24]. Later [20] showed that randomly initialized plain gradient descent will solve the phase retrieval problem with $O(N\, \mathrm{poly}(\log N))$ samples. An open theoretical question is whether randomly initialized gradient descent is able to solve the phase retrieval problem with $O(N)$ samples. We explore this question in the present work.

Our main contributions can be summarized in the following points:

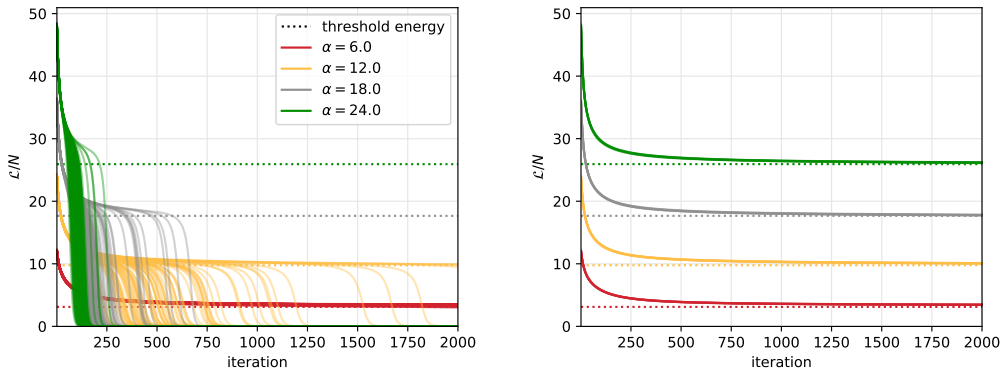

Figure 1: Evolution of the training loss for systems of size $N = 16384$ with number of samples αN. The left panel has labels created using a teacher, while the right panel has random labels constructed using a Gaussian distribution with variance matching the teacher-labels. The left figure shows that also the simulations with a teacher approach the threshold energy before transiting to the global minimum, if α is large enough.

- We show empirically that in the phase retrieval problem, randomly initialized gradient flow is attracted by the so-called threshold states. We show that as the number of samples increases, the geometry of the threshold states changes from minima to saddles with the slope pointing toward the teacher-vector. This transition is akin to the BBP phase transition known in random matrix theory [25]. This transition affects gradient descent that, following the slope, achieves zero generalization error.
- We obtain a close formula for the number of samples per dimension α_c at which this transition happens. This depends on the joint probability distribution of the labels of the student and the teacher at the threshold.
- Using the replica theory from statistical physics as a non-rigorous proxy we characterize the approximate distribution of the labels of the student and the teacher leading to a approximate prediction for threshold $\alpha_c = 13.8$, notably suggesting that the additional logarithmic factors in the previous works might not be needed.

2 BBP on the threshold states

In Fig. 1 (left) we show the loss as a function of iteration time for the phase retrieval problem, defined above, with varying number of samples to dimension ratio α in many different runs (full lines). In the right hand side of the figure we show the loss but this time for labels that do not come from a teacher network, but that are randomly reshuffled. We see that in that case the loss converges after a very long time towards a value marked by the dotted line (reproduced also in the left part), that we defined to be the so-called threshold energy. We see that for small α, e.g. $\alpha = 6$ the train loss on the phase retrieval problem does not decrease to zero (nor the test one), while for the larger values $\alpha = 18$ and 24 it does very rapidly. The value $\alpha = 12$ is close to a critical regime where some realization find perfect generalization and other do not, with the dynamics staying for a long time close to the threshold energy.

In a different model, the spiked matrix-tensor, Ref. [8] described exactly this phenomenology and showed that gradient flow starting from random initial conditions has a transition when the Hessian of the spurious minima that trap the dynamics, the so-called threshold states, display a BBP transition [25]. This leads to the emergence of a descending direction toward the informative minimum which is correlated with the ground truth signal $\boldsymbol{W}^*$. In Fig. 2 we argue that the same mechanism is at play in the phase retrieval problem and based on these insights we derive an analytic equation for the corresponding threshold in Sec. 2.1.

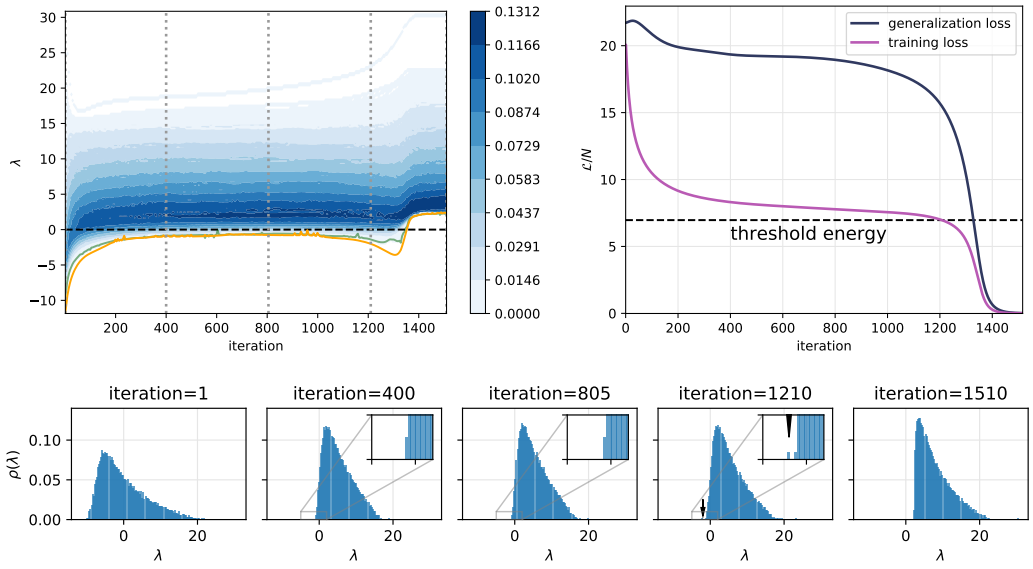

Figure 2: Properties of the Hessian for phase retrieval of a system of size $N = 2048$ at $\alpha = 10$. On the left figure, we show the evolution of the density of the bulk of the eigenvalues, from zero density in white to high density in blue, the smallest eigenvalue in orange, and the second smallest in green. The picture shows that a BBP transition occurs when the training loss approaches the threshold energy. The right panel depicts the evolution of the training loss (purple) and the generalization loss (dark blue) in time. The training loss rapidly approaches a plateau at the level of the threshold states (black dashed line) and converges towards the teacher after the BBP transition. Namely, when the smallest eigenvalue detaches from the rest of the bulk. The bottom panels show the distribution of the eigenvalues at five different instants in the dynamics. At iteration ≈ 1210 we cross the threshold energy and we observe an isolated eigenvalue detaching from the bulk.

2.1 Theory for the BBP threshold

Based on the numerical results just presented, we aim to obtain an equation determining the value of threshold α_c such that for $\alpha > \alpha_c$ the BBP transition occurs, whereas for $\alpha < \alpha_c$ it does not. For $\alpha < \alpha_c$ the system is at long times trapped in the threshold states and not able to recover, even weakly, the signal. We define $P(\hat{y}, y)$ the long-time limit of the distribution of the estimated labels and the true labels; $P(\hat{y}, y)$ allows us to study the Hessian of the threshold states, which is the random matrix defined in (4) with $\hat{y}_\mu$ and y_μ distributed following the law $P(\hat{y}, y)$.
A type of random matrix $\mathcal{M}_{i,j}$ with similar structure as the contribution to the Hessian coming from the Loss function, $\mathcal{H}_{i,j} = -\mathcal{M}_{i,j} - \mu\delta_{i,j}$, has been studied recently in [26] and the convergence in probability for large N of the largest and the second largest eigenvalues of such matrix was proven. We applied the results of the Theorem 1 of [26] to determine the behavior of the smallest and second smallest eigenvalue of the Hessian. Call

$$\Psi_\alpha(\lambda) = \lambda\left[\frac{1}{\alpha} - \mathbb{E}_{\hat{y},y}\left(\frac{\alpha\partial^2_{\hat{y}}\ell(\hat{y},y)}{2\lambda + \alpha\partial^2_{\hat{y}}\ell(\hat{y},y)}\right)\right], \qquad \Phi(\lambda) = -\lambda\mathbb{E}_{\hat{y},y}\left(\frac{\alpha\partial^2_{\hat{y}}\ell(\hat{y},y)y^2}{2\lambda + \alpha\partial^2_{\hat{y}}\ell(\hat{y},y)}\right); \quad (5)$$

let $\bar{\lambda} = \arg\min \Psi_\alpha(\lambda)$ and $\xi_\alpha(\lambda) = \Psi_\alpha(\max\{\lambda, \bar{\lambda}\})$ then the largest eigenvalue of $\mathcal{M}_{i,j}$ is $\lambda_1 \to_{\mathbb{P}} \xi_\alpha(\lambda^*_\alpha)$ with λ^*_α being the solution of $\xi_\alpha(\lambda) = \Phi(\lambda)$. The second largest is $\lambda_2 \to_{\mathbb{P}} \Psi_\alpha(\bar{\lambda}_\alpha)$.

A BBP transition occurs at $\alpha_{\rm BBP}$, the largest α such that $\lambda_1 = \lambda_2$. Following [26] this leads to an equation on $\bar{\lambda}$ and $\alpha_{\rm BBP}$ which reads:

$$\frac{1}{\alpha_{\rm BBP}} = \mathbb{E}_{\hat{y},y}\left(\frac{\alpha_{\rm BBP}\partial^2_{\hat{y}}\ell(\hat{y},y)}{2\bar{\lambda} + \alpha_{\rm BBP}\partial^2_{\hat{y}}\ell(\hat{y},y)}\right)^2. \quad (6)$$

We now use an additional assumption, which comes from studies of gradient descent dynamics of mean-field spin-glasses [27], that the threshold states are marginal, i.e. the smallest eigenvalue of their

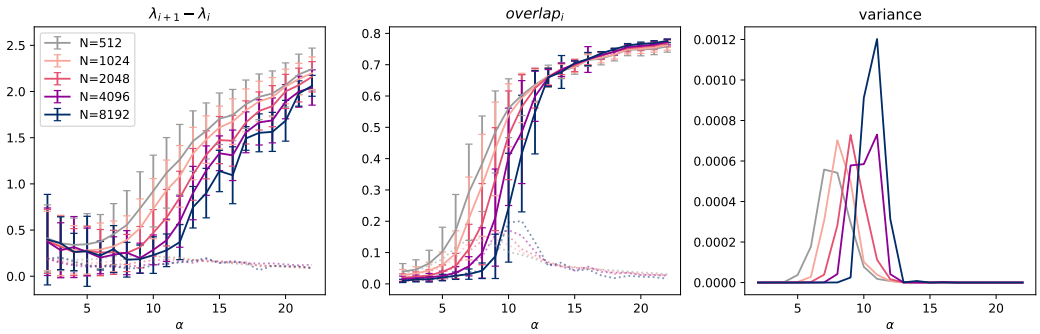

Figure 3: The three images show the occurring of a BBP transition at the moment when the training loss crosses the threshold energy. The images investigate the BBP as the input dimension is increased form $N = 512$ to $N = 8192$. At the BBP transition the smallest eigenvalue pops out of the bulk of eigenvalues and the associated eigenvector contains information on the signal. The left figure is the difference of the smallest eigenvalues (the second with the first in solid line, the third with the second in dotted line). The central image shows the overlap of the first eigenvector (full) and second eigenvector (dotted) with the signal. The transition appears to shift as in the left image. Finally on the right the fluctuations of the overlap first eigenvector with the teacher are shown, the peak corresponds to the transition.

Hessian is null. As the smallest eigenvalue of the Hessian is determined by the largest eigenvalue of $\mathcal{M}_{i,j}$, this imposes the additional condition $\lambda_1 = \lambda_2 = -\mu$. Using the second Eq. in (5) to enforce this last condition and the definition of μ in Eq. (3) one finds (see SM for more details):

$$\mu = \bar{\lambda}\mathbb{E}_{\hat{y},y}\left(\frac{\alpha_{\rm BBP}\partial^2_{\hat{y}}\ell(\hat{y},y)y^2}{2\bar{\lambda}+\alpha_{\rm BBP}\partial^2_{\hat{y}}\ell(\hat{y},y)}\right), \qquad \mu = \frac{\alpha_{\rm BBP}}{2}\mathbb{E}_{\hat{y},y}\left(\frac{\partial}{\partial\hat{y}}\ell(\hat{y},y)\hat{y}\right). \tag{7}$$

Together the three eqs. (6), and the two in (7) allow to determine the three unknown $\alpha_{\rm BBP}, \bar{\lambda}, \mu$. Assuming that $\alpha_c = \alpha_{\rm BBP}$, they therefore provide an equation for the algorithmic threshold once $P(\hat{y}, y)$ at threshold states is known.

2.2 Further numerical justifications

We now illustrate and test the theory by numerical simulations. Fig. 2 shows that the training loss slowly tends to the threshold energy, before departing in direction of the global minimum. According to the theory described in the previous section, the phenomenon occurring is a BBP transition. In order to confirm this we characterize the spectrum of the Hessian and focus on the smallest eigenvalues. On left figure of Fig. 2 we show the density of eigenvalues in blue and we show separately the smallest eigenvalue in orange and the second smallest eigenvalue in green. During the dynamics the spectrum moves compactly forming a bulk - except for the largest eigenvalues that do not play role in the transition. Approaching the threshold states, for $\alpha > \alpha_c$ the dynamics feels the presence of a descending direction associated with the smallest eigenvalue. This becomes increasingly strong and when the negative eigenvalue pops out of the bulk, the dynamics will follow and will converge to the global minimum. In the lower panel of Fig. 2 we show 5 snapshots of the spectral density during the dynamics.

In Fig. 3 we study the spectral properties of the Hessian at the time when the training loss hits the threshold energy. In the first panel from the left, we show the difference between the second smallest eigenvalue and the smallest eigenvalue (solid line) and the difference between the third smallest and the second smallest eigenvalue (dotted). The two quantities are close, and very small, until a certain value value of α - that depends on the size and is reported in the caption - after which the former increases linearly whereas the latter remains small. The second property that we investigate, central figure, is the overlap of the first and second eigenvectors with the teacher-vector $\boldsymbol{W}^*$ (respectively solid and dotted lines). This overlap is zero before the transition, then as $\lambda_2 - \lambda_1$, it suddenly increases and finally saturates. The first two panels are provide a strong evidences that a BBP transition is taking place. To further corroborate this findings, in the right panel we consider the fluctuation of the

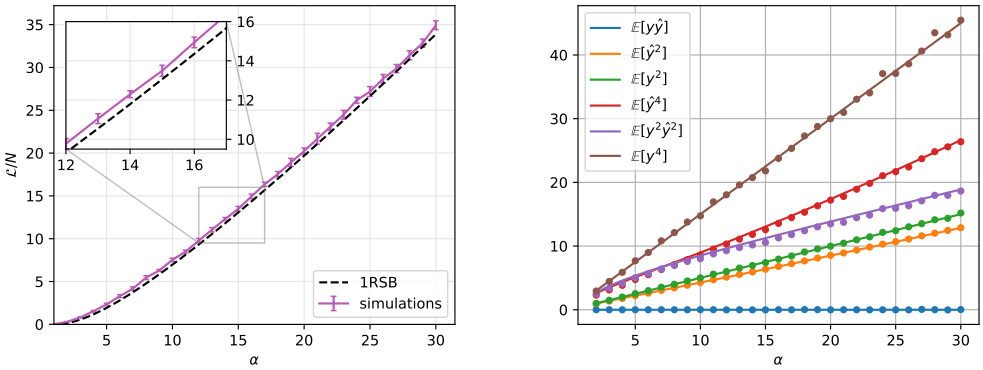

Figure 4: On the left panel is the loss of the threshold states from the simulations and from the analysis with the 1RSB replica method, evaluated for different values of α. The errorbars given by mean and standard deviation of 1000 simulations with input dimension $N = 2048$ and shuffled labels. As expected increasing the number of datapoints the number of violated constraints increases and so does the threshold energies. On the right we compare the moments of the label distribution Eq. (17) (solid lines) with the moments obtained in the simulations (circles).

overlap (Fig. 3 right panel). In statistical physics terms, the overlap plays the role of order parameter, and its fluctuations diverge at the BBP transition. We indeed find that at the value of α at which the BBP transition seems to take place fluctuations peak (the more so the larger is N as expected for a phase transition).

Finally, we also note the presence of strong finite size effects that shift the effective value of α at which the transition (cross-over for finite N) takes place.

3 Characterization of threshold states

The pivotal point of our analysis is that the gradient-flow dynamics is that the spurious minima trapping the dynamics and hindering weak recovery are the threshold states. The study of threshold states has been started in statistical physics of disordered systems, where it has been conjectured and verified that they play a prominent role in the gradient descent dynamics of such random systems [9]. according to this theory, threshold states are the most numerous minima and the Hessian associated to the threshold states has a spectrum that is positive semi-definite and gap-less.

In the previous section we have obtained a closed set of equations that allow us to obtain $\alpha_{\rm BBP}$ once the distribution $\mathbb{P}(\hat{y}, y)$ is known. Such distribution can be derived using tools from statistical physics, in particular replica theory [10, 28, 29]. The main quantity of interest is the partition function, which is defined as the normalization constant of the Gibbs distribution associated with the loss Eq. (1),

$$\mathcal{Z} = \int_{\mathbb{S}^{\alpha(N-1)}(\sqrt{N})} e^{-\beta \mathcal{L}(\boldsymbol{W})} d\boldsymbol{W}, \tag{8}$$

where β is a parameter associated to the inverse temperature in the physical analog of the problem. We consider the $\beta \to \infty$ limit to study the minima of $\mathcal{L}$ that are relevant in gradient-flow dynamics.

The disordered systems approach focuses on the average properties of the systems. In order to ensure concentration properties as the input dimension goes to infinity, the quantity to average is the logarithm of the partition function. This quantity in general is hard to compute and we resort to replica theory. The idea is to apply replica method, described below and in more detail in the SM, to move from the moments of the partition function to average of its logarithmic value. The analysis leads to the free energy density, Eq. (11), that we use to compute the distribution of the labels.

We start writing the moments of Eq. (8)

$$\overline{\mathcal{Z}^n} = \int_{\mathbb{S}^{\alpha(N-1)}(\sqrt{N})} \mathbb{E}_{(\boldsymbol{x},y)} e^{-\beta n \mathcal{L}(\boldsymbol{W})} d\boldsymbol{W}, \tag{9}$$

where we represent the average over the possible datasets with the overbar. The high-dimensional integral in Eq. (9) can be written in term of the overlap matrix $Q_{a,b} = \langle \boldsymbol{W}_a \cdot \boldsymbol{W}_b \rangle / N$ that encodes the similarity between two configurations extracted with the Gibbs measure, of the labels $\hat{y}_\mu$ and y_μ, $\overline{\mathcal{Z}^n} = \int e^{NnS(\boldsymbol{Q})} \prod_{a \geq b=0}^{n} dQ_{a,b}$ with

$$S(\boldsymbol{Q}) = \frac{1}{2} \log\det \boldsymbol{Q} + \alpha \log \exp \left[\frac{1}{2} \sum_{a,b=0}^{n} Q_{a,b} \frac{\partial^2}{\partial h_a \partial h_b} \right] \exp \left[-\frac{\beta}{2} \sum_a \ell(h_a, h_0) \right] \Bigg|_{\{h_a = 0 \ \forall a\}} . \tag{10}$$

We perform a so-called *1-step replica symmetry breaking* (1RSB) analysis [30] that allows us to obtain an explicit expression for the distribution of $(\hat{y}, y)$. Using the 1RSB ansatz on this problem we reduce the number of parameters in Eq. (10) to χ and z: where χ is related to amplitude of the fluctuations in minimum when this is perturbed, and z is a Lagrange multiplier that we use to enforce that the minima in consideration have a Hessian with gapless spectrum, i.e. they are threshold states. The general prescription for finding the global minimum of this kind of problem would require to optimize over χ and z, however, the same formalism can be used to characterize the threshold states by imposing a value for z and optimizing only over χ [28, 31]. We evaluate the integral of the partition function Eq. (8) at the threshold using Laplace's method on Eq. (10). Finally we consider the replica method, by taking the analytic continuation of the moments of the partition function and the limit $n \to 0^+$. With this considerations we obtain the free energy density in low temperature limit $f(\chi|\alpha, z) = -\lim_{\beta\to\infty} \lim_{N\to\infty} \frac{1}{\beta N} \overline{\log \mathcal{Z}}/(\beta N)$ that reads

$$f(\chi|\alpha, z) = -\frac{1}{2z} \log \frac{\chi + z}{\chi} - \frac{\alpha}{z} \int_{\mathbb{R}} dy \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \log \gamma_1 \star e^{\frac{-z}{2} V(\hat{y}, y|\chi)} \Bigg|_{\hat{y}=0} , \tag{11}$$

where $V(\hat{y}, y|\chi) \doteq \min_{\tilde{y}} \frac{(\hat{y}-\tilde{y})^2}{\chi} + \ell(\tilde{y}, y)$, the $\star$ symbol represent a convolution and we write γ_1 as a compact notation to represent a centered Gaussian distribution with unit variance. The arguments χ and z are given via implicit equations that respectively impose that the Laplace's approximation on f and that the free energy describes the threshold states (further details are in the SM):

$$\frac{1}{\chi(\chi + z)} = \frac{\alpha}{4} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \frac{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\hat{y}^2}{2}} e^{-\frac{z}{2} V(\hat{y}, y|\chi)} \left(\partial_{\hat{y}} V(\hat{y}, y|\chi)\right)^2 d\hat{y}}{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\hat{y}^2}{2}} e^{-\frac{z}{2} V(\hat{y}, y|\chi)} d\hat{y}} dy; \tag{12}$$

$$1 = \frac{\alpha}{4} \chi^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \frac{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\hat{y}^2}{2}} e^{-\frac{z}{2} V(\hat{y}, y|\chi)} \left(\partial^2_{\hat{y}} V(\hat{y}, y|\chi)\right)^2 d\hat{y}}{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\hat{y}^2}{2}} e^{-\frac{z}{2} V(\hat{y}, y|\chi)} d\hat{y}} dy \, . \tag{13}$$

In order to obtain $\mathbb{P}(\hat{y}, y)$ we follow the strategy introduced in [29]. The partition function in Eq. (8) can be written as a functional integral over the empirical labels $\rho_{\text{quen}}(\hat{y}, y) = \frac{1}{\alpha N} \sum_{m=1}^{\alpha N} \delta(y - y_m)\delta(\hat{y} - \hat{y}_m)$:

$$\mathcal{Z} = \int \mathcal{D}[\rho_{\text{quen}}(\hat{y}, y)] \, e^{-\frac{\beta \alpha N}{2} \int d\hat{y} dy \rho_{\text{quen}}(\hat{y}, y) \ell(\hat{y}, y) + N S[\rho_{\text{quen}}(\hat{y}, y)]} , \tag{14}$$

where $S[\rho_{\text{quen}}(\hat{y}, y)]$ is the entropic factor (evaluated on configurations corresponding to the threshold states). This makes clear that the free-energy can also be obtained in terms of a large-deviation principle:

$$f = \text{Min}_\rho \left(\frac{\alpha}{2} \int d\hat{y} dy \rho_{\text{quen}}(\hat{y}, y) \ell(\hat{y}, y) - S[\rho_{\text{quen}}(\hat{y}, y)] \right) \tag{15}$$

The minimizer of this variational problem corresponds by definition to $\mathbb{P}(\hat{y}, y)$. By taking the functional derivative of f with respect to $\ell(h, h_0)$ one obtains:

$$\frac{\delta f}{\delta \ell(\hat{y}, y)} = \frac{\alpha}{2} \mathbb{P}(\hat{y}, y). \tag{16}$$

Since we have an explicit expression of f in terms of ℓ, see eq. (11), we can obtain this distribution directly. In taking the functional derivative of Eq. (11) we must be careful in considering the implicit

dependence of $V(\hat{y}, y|\chi)$ from $\ell(\hat{y}, y)$. Finally inverting Eq. 9160 we obtain the label distribution

$$\mathbb{P}_{\rm 1RSB}(\hat{y}, y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \frac{\exp\left[-\frac{\hat{y}^2}{2} - \frac{z}{2} V(\hat{y}, y|\chi)\right]}{\int_{\mathbb{R}} \exp\left[-\frac{\tilde{y}^2}{2} - \frac{z}{2} V(\tilde{y}, y|\chi)\right] d\tilde{y}}. \tag{17}$$

The index 1RSB denotes that the free-energy has been obtained within the 1RSB scheme.

We run numerical experiments to characterize the threshold states in terms of their energy and their moments with respect to the label probability distribution. In Fig. 4 (left) the 1RSB threshold energy is plotted together with the value of the plateau obtained from the simulations in Fig. 1. The 1RSB ansatz appears to be a good approximation of the empirically obtained energy. In the inset we highlight the discrepancy with the empirical line, which may be due to both finite size effects as well as to the fact that the 1RSB scheme is an approximation and more involved schemes must be employed to obtain the exact distribution [30]. Finally in Fig. 4 (right) we show the first 4 moments of the label distribution from the 1RSB analysis and we compare them with the numerical results. The 1RSB moments give a nice agreement in relation to the empirical ones.

Encouraged by the reasonable, though not perfect, accuracy of the 1RSB approximation, we use the expression (17) as input for the three eqs. in (6,7). Their solutions leads to a finite threshold at $\alpha_{\rm BBP}^{\rm 1RSB} \approx 13.8$. This value could be compatible with the numerical results presented in Fig. 3 if the finite size shift saturates for yet larger sizes.

4 Discussion

In Fig. 5 we present the fraction of runs of gradient descent that converge to zero test error in the phase retrieval problem under consideration. We see that a large fraction of simulations achieves convergence before the $\alpha_{\rm BBP}^{\rm 1RSB}$ threshold, in general before $\alpha \approx 11$. The mechanism behind this difference is not clear to us, as well as it is not clear whether it will hold as the input dimension tends to infinity. The numerical results are affected by strong finite size effects and may be both consistent with a transition taking place at finite alpha in the infinite dimensional limit and with curves that shift in alpha as a logarithm in the input dimension. Empirically we do not observe any unsuccess above $\alpha_{\rm BBP}^{\rm 1RSB}$ which is consistent with our theoretical scenario for any $N > 64$. In [7] the authors showed that large finite size effects in the initialization affect the location of the BBP transition. Whether this happens also in this case is unclear and deserves further investigations.

Let us conclude by commenting on the limitations and possible generalizations of the results presented in this paper. A key element of the phase retrieval model is the existence of a phase at small α in which the best achievable generalization error is not better than random guess. This arises in models that present a symmetry, such as the ± 1 symmetry due to the absolute value in the phase retrieval problem. The picture shown in Fig. 1 where the threshold states are characterized using the gradient-flow dynamics in a variant of the model with randomized labels is enabled by this symmetry and the rest of our analysis relies on the simplifications stemming from this. We expect that the BBP picture presented in this work generalizes to all learning where at small sample complexity error is not better than random guess. Of course working out the details, even for two layer neural networks, is an open problem of interest for future work. In most learning problems a naive linear regression often achieves a better than random guess performance, and thus an extension of our theory for problems of that type (lacking a symmetry) would be an interesting direction for future work.

Acknowledgments and Disclosure of Funding

We thank Carlo Lucibello for precious discussions and Federico Ricci-Tersenghi for interesting comments on the manuscript. We acknowledge funding from the ERC under theEuropean Union's Horizon 2020 Research and Innovation Programme Grant Agreement 714608-SMiLe;from the French government under management of Agence Nationale de la Recherche (ANR) grant PAIL,and grant "Investissements d'Avenir" LabEx PALM (ANR-10-LABX-0039-PALM) (StatPhysDisSys) and(ANR-19-P3IA-0001) (PRAIRIE 3IA Institute); and from the Simons Foundation collaboration Crackingthe Glass Problem (#454935, Giulio Biroli).

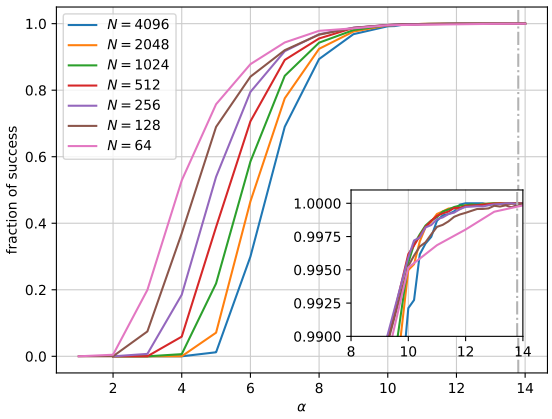

Figure 5: Fraction of simulations with gradient descent that achieve zero test error for various input dimensions N. The number of simulations increases with the smaller input dimension in order the account for the larger fluctuations. The number varies from 10000 simulations for $N = 4192$ to 100000 simulations for $N = 64$. The learning rate is 0.0002 and the simulation runs until either the loss goes below 10^{-8} or the number of steps exceeds $1500 \log_2 N$ iterations. The logarithmic term in the time accounts for the fact that the average initial overlap with the ground truth is $\sim 1/\sqrt{N}$. In the inset we zoom in the region where all the simulations converge showing the indication of a crossing with large fluctuations due to finite size effects.

Broader Impact

Our work is theoretical in nature, and as such the potential societal consequence are difficult to foresee. We anticipate that deeper theoretical understanding of the functioning of machine learning systems will lead to their improvement in the long term.

References

[1] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[2] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

[3] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

[4] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016.

[5] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. *arXiv preprint arXiv:1712.08968*, 2017.

[6] Shengchao Liu, Dimitris Papailiopoulos, and Dimitris Achlioptas. Bad global minima exist and sgd can reach them. *arXiv preprint arXiv:1906.02613*, 2019.

[7] Stefano Sarao Mannelli, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborova. Passed & spurious: analysing descent algorithms and local minima in spiked matrix-tensor model. *ICML*, 2019.

[8] Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, and Lenka Zdeborová. Who is afraid of big bad minima? analysis of gradient-flow in spiked matrix-tensor models. In *Advances in Neural Information Processing Systems*, pages 8676–8686, 2019.

[9] Leticia F Cugliandolo and Jorge Kurchan. Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model. *Physical Review Letters*, 71(1):173, 1993.

[10] Fabrizio Antenucci, Silvio Franz, Pierfrancesco Urbani, and Lenka Zdeborová. Glassy nature of the hard phase in inference problems. *Physical Review X*, 9(1):011020, 2019.

[11] TLH Watkin and J-P Nadal. Optimal unsupervised learning. *Journal of Physics A: Mathematical and General*, 27(6):1899, 1994.

[12] Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Marvels and pitfalls of the langevin algorithm in noisy high-dimensional inference. *Physical Review X*, 10(1):011057, 2020.

[13] Adriaan Walther. The question of phase retrieval in optics. *Optica Acta: International Journal of Optics*, 10(1):41–49, 1963.

[14] Rick P Millane. Phase retrieval in crystallography and optics. *JOSA A*, 7(3):394–411, 1990.

[15] Robert W Harrison. Phase problem in crystallography. *JOSA a*, 10(5):1046–1055, 1993.

[16] Oliver Bunk, Ana Diaz, Franz Pfeiffer, Christian David, Bernd Schmitt, Dillip K Satapathy, and J Friso Van Der Veen. Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels. *Acta Crystallographica Section A: Foundations of Crystallography*, 63(4):306–314, 2007.

[17] Radu Balan, Pete Casazza, and Dan Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, 2006.

[18] John V Corbett. The pauli problem, state reconstruction and quantum-real numbers. *Reports on Mathematical Physics*, 1(57):53–68, 2006.

[19] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[20] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1-2):5–37, 2019.

[21] Philip Schniter and Sundeep Rangan. Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing*, 63(4):1043–1055, 2014.

[22] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[23] Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. *Foundations of Computational Mathematics*, 19(3):703–773, 2019.

[24] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

[25] Jinho Baik, Gérard Ben Arous, Sandrine Péché, et al. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[26] Yue M Lu and Gen Li. Phase transitions of spectral initialization for high-dimensional nonconvex estimation. *Information and Inference: A Journal of the IMA*, 2019.

[27] Leticia F Cugliandolo. Dynamics of glassy systems. *arXiv preprint cond-mat/0210312*, 2002.

[28] Rémi Monasson. Structural glass transition and the entropy of the metastable states. *Physical review letters*, 75(15):2847, 1995.

[29] Silvio Franz, Giorgio Parisi, Maxime Sevelev, Pierfrancesco Urbani, and Francesco Zamponi. Universality of the sat-unsat (jamming) threshold in non-convex continuous constraint satisfaction problems. *SciPost Phys*, 2(3):019, 2017.

[30] Marc Mézard, Giorgio Parisi, and Miguel-Angel Virasoro. *Spin glass theory and beyond.* World Scientific Publishing, 1987.

[31] Francesco Zamponi. Mean field theory of spin glasses. *arXiv preprint arXiv:1008.4844*, 2010.

[32] Ker-Chau Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992.

[33] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*, volume 111. Clarendon Press, 2001.

[34] Florent Krzakala and Lenka Zdeborová. Hiding quiet solutions in random constraint satisfaction problems. *Physical review letters*, 102(23):238701, 2009.

[35] Jorge Kurchan, Giorgio Parisi, Pierfrancesco Urbani, and Francesco Zamponi. Exact theory of dense amorphous hard spheres in high dimension. ii. the high density regime and the gardner transition. *The Journal of Physical Chemistry B*, 117(42):12979–12994, 2013.

A BBP transition and phase transition of spectra initialization

In non-convex estimation problems, such as Phase Retrieval, big advantages follow from the development of well tailored spectral methods to be used as initialization step. Recently the outcomes of one such spectral method widely used in Phase Retrieval and based on the construction of a data matrix

$$\mathcal{M}_{i,j} = \frac{1}{\alpha N}\sum_{m=1}^{\alpha N} \mathcal{T}(Y_m) u_{m,i} u_{m,j} = \frac{1}{\alpha}\sum_{m=1}^{\alpha N} \mathcal{T}(Y_m) x_{m,i} x_{m,j} \tag{18}$$

from sensing vectors $\boldsymbol{u}_m$, with elements of order one (or sensing vectors $\boldsymbol{x}_m$ on the unitary sphere, according to our definition) and measurements Y_m has been exactly derived [26]. The method involves a pre processing function $\mathcal{T}(Y_m)$ that can be optimized to further improve the results. Once the data matrix is constructed the eigenvector $\boldsymbol{v_1}$ corresponding to the largest eigenvalue λ_1 can be used as an estimator of the signal $\boldsymbol{W^*}$.
To obtain the performances of this kind of spectral initialization it is assumed [26] that the measurements Y_m are independently drawn according to a density function conditional on $y_m = \langle \boldsymbol{x}_m \boldsymbol{W^*} \rangle$ associated to the particular acquisition process, and it is recalled that y_m are themselves Gaussian random variables, due to the definition of the problem. Finally in the large N limit the empirical average used to construct the data matrix can be replaced by the expected value $\mathbb{E}_{Y,y}$ over these two distributions.
The result [26] goes as follows. Given two functions defined as

$$\Psi_\alpha(\lambda) \equiv \lambda \left[\frac{1}{\alpha} + \mathbb{E}_{Y,y}\left(\frac{\mathcal{T}(Y)}{\lambda - \mathcal{T}(Y)} \right) \right] \tag{19}$$

and

$$\Phi(\lambda) = \lambda \mathbb{E}_{Y,y}\left(\frac{\mathcal{T}(Y) y^2}{\lambda - \mathcal{T}(Y)} \right) \tag{20}$$

with $\lambda > \mathcal{T}(Y)$, and given

$$\xi_\alpha(\lambda) = \Psi_\alpha(\max\{\lambda, \bar{\lambda}\}) \tag{21}$$

with

$$\bar{\lambda} = \arg\min \Psi_\alpha(\lambda) \, , \tag{22}$$

the two largest eigenvalues of $\mathcal{M}$, λ_1 and λ_2, are such that

$$\lambda_1 \to_{\mathbb{P}} \xi_\alpha(\lambda_\alpha^*) \, , \tag{23}$$

with λ_α^* the solution of $\xi_\alpha(\lambda) = \Phi(\lambda)$, and

$$\lambda_2 \to_{\mathbb{P}} \Psi_\alpha(\bar{\lambda}). \tag{24}$$

A phase transition occurs at the largest α such that $\lambda_1 = \lambda_2$, which can be evaluated by imposing $\Psi'_\alpha(\lambda_\alpha^*) = 0$ or equivalently

$$\frac{1}{\alpha} = \mathbb{E}_{Y,y}\left(\frac{\mathcal{T}(Y)^2}{(\bar{\lambda} - \mathcal{T}(Y))^2} \right) \, , \tag{25}$$

which corresponds to $\Psi'_\alpha(\bar{\lambda}) = 0$ as at that point $\lambda_\alpha^* = \bar{\lambda}$. At larger α the largest eigenvalue pops out from the spectrum bulk ($\lambda_1 \neq \lambda_2$) and the corresponding eigenvector develops a finite correlation with the signal, in a phenomenon called BBP transition [25], hence the definition of α_{BBP} when this occurs.

It is interesting to note that the structure of the data matrix is closely reminiscent of the structure of the first term in the Hessian of our problem (see Eq. (4)). Indeed incidentally the original idea of this spectral method for initialization can be traced back to the study of Hessian's principal directions [32]. In particular we observe that

$$-\mathcal{H}_{i,j}(\boldsymbol{W}) = -\frac{1}{2}\sum_{m=1}^{\alpha N} \frac{\partial^2}{\partial \hat{y}_m^2} \ell(\hat{y}_m, y_m) x_{m,i} x_{m,j} + \mu \delta_{i,j} = \mathcal{M}_{i,j} + \mu \delta_{i,j} \, , \tag{26}$$

provided that the pre processing function is

$$\mathcal{T}(Y) = -\frac{\alpha}{2}\frac{\partial^2}{\partial \hat{y}^2}\ell(\hat{y}, y) = -\frac{\alpha}{2}\partial_{\hat{y}}^2 \ell(\hat{y}, y) \, . \tag{27}$$

Note that we consider a case for which measurements Y have a one to one correspondence with y (*i.e.* $Y = y^2$). Moreover in the problem discussed empirical averages involve not only measurements Y but also estimated $\hat{y} = \langle \boldsymbol{x}\boldsymbol{W} \rangle$ in correspondence of some $\boldsymbol{W}$ of interest. Therefore the expected value, $\mathbb{E}_{\hat{y},y}$, should be taken over the relative joint probability distribution function $P(\hat{y}, y)$. In conclusion, the results just mentioned tell immediately what is the largest α (*i.e.* α_{BBP}):

$$\frac{1}{\alpha_{\rm BBP}} = \mathbb{E}_{\hat{y},y} \left(\frac{\alpha_{\rm BBP} \partial^2_{\hat{y}} \ell(\hat{y}, y)}{2\bar{\lambda} + \alpha_{\rm BBP} \partial^2_{\hat{y}} \ell(\hat{y}, y)} \right)^2 , \tag{28}$$

before the smallest eigenvalue of the Hessian pops out from the spectrum bulk, being also associated to an eigenvector with finite projection on the signal to be detected.

In this case we are focusing on performance of a gradient descent dynamics, which for mean-field spin-glasses naturally gets stuck on what are called threshold states [27]. We argue that the gradient descent dynamics applied to Phase Retrieval, when retrieval fails, will also approach threshold states which are mainly characterized by their property of being marginal, *i.e.* the smallest eigenvalue of their Hessian is null. This qualifies the relevant $\boldsymbol{W}$ as a typical configuration belonging to threshold states and $P(\hat{y}, y)$ as the joint probability distribution at threshold states. Moreover it allows to introduce the marginal condition $\lambda_2 = \lambda_1 = -\mu$, which can be re-expressed by equating $\lambda_2 = \Psi_\alpha(\bar{\lambda})$ to $-\mu$:

$$\mu = \bar{\lambda} \mathbb{E}_{\hat{y},y} \left(\frac{\alpha_{\rm BBP} \partial^2_{\hat{y}} \ell(\hat{y}, y) y^2}{2\bar{\lambda} + \alpha_{\rm BBP} \partial^2_{\hat{y}} \ell(\hat{y}, y)} \right) . \tag{29}$$

Finally the definition of the spherical parameter (3) in the main text must be considered to close the system of equations

$$\mu = \frac{\alpha_{\rm BBP}}{2} \mathbb{E}_{\hat{y},y} \left(\partial_{\hat{y}} \ell(\hat{y}, y) \hat{y} \right) \tag{30}$$

to be used to determine $\alpha_{\rm BBP}, \bar{\lambda}, \mu$ in correspondence of $P(\hat{y}, y)$ from threshold states.

The resulting picture is as follows. In the small α regime, gradient descent will systematically approach threshold states and remain stuck there. However starting from $\alpha_{\rm BBP}$ in the Hessian of these states, that is otherwise marginal, an isolated eigenvalue pops out from the bulk immediately becoming negative. Moreover the eigenvector associated to such negative direction, naturally followed by gradient descent, has a finite overlap with the signal. Therefore, we argue, it is from that point on that the signal should be easily retrieved.

B Replica Analysis

The field of physics of disordered systems has developed numerous tools to deal with random systems [30, 33]. At an abstract level of thinking, using those tools means that we identify the inference problem with a physical systems that subject to a certain potential. The randomness comes from the having a dataset made of random projections. The estimator is mapped into a spherical spin and the loss function becomes the energy - or *Hamiltonian*. Finally the system the temperature in which the system lives is sent to zero and the system tend to the lowest energy, herefore minimizing the loss. The ground truth in the inference problem becomes equivalent to a minimum planted in the energetic landscape [34]. For example this formulation is equivalent to a physical system that before the experiment is liquid, but as we cool it down it can become either a crystal - an ordered solid - or a glass - an amorphous solid. Finding the crystal means reconstructing the signal.

B.1 Partition function

Moving to the mathematical aspects of the problem. We define a Gibbs distribution associated with the problem and evaluate its normalization constant - the *partition function* $\mathcal{Z}$. The partition function, and in particular its logarithm divided by the temperature - the free energy -, contains all the information we aim to understand in the problem. By taking the proper derivatives, and possibly add an external field, we can compute relevant macroscopic properties such as: average overlap with the ground truth, average loss achieved. In disordered system we have to consider the additional complication given by the randomness. Therefor, we need to consider the average free

energy that is the average of the logarithm of a high-dimensional integral, that can be done by the simple observation:

$$\log \mathcal{Z} = \lim_{n\to 0} \frac{\mathcal{Z}^n - 1}{n}. \tag{31}$$

Which for arbitrary n is not simpler than computing the logarithm, but it is much simpler if $n \in \mathbb{N}$ and we perform an analytic continuation to $n \in \mathbb{R}$. Under this - replica trick - the average of the logarithm is equivalent to compute the average of the nth moment of the partition function and take the limit. Formally the nth moment correspond to the partition function of n identical - replicated - system that do not interact but with the same realization of the disorder.

The problem is now computing the moments of the partition function which in general is prohibitive and we have to use an ansatz on the specific form of the solution, in particular we use the so called *replica symmetry breaking ansatz* [30]. This largely reduces the number of parameters. Finally the average free energy can be evaluated by set of saddle point equations.

We can now move to the analysis. The partition function already defined in the main text is

$$\mathcal{Z} = \int_{\mathbb{S}^{N-1}(\sqrt{N})} \mathcal{D}\boldsymbol{W}_S \, \exp\left\{ -\frac{\beta}{2} \sum_{\mu} \ell(\langle \boldsymbol{x}_\mu, \boldsymbol{W}_S\rangle, \langle \boldsymbol{x}_\mu, \boldsymbol{W}_T\rangle) \right\} \tag{32}$$

and consider its nth moment - the *replicated partition function* -

$$\overline{\mathcal{Z}^n} = \mathbb{E}_{\{\boldsymbol{x}_\mu\}} \int_{\mathbb{S}^{n(N-1)}(\sqrt{N})} \prod_{a=1}^{n} \mathcal{D}\boldsymbol{W}_a \, \exp\left\{ -\frac{\beta}{2} \sum_{\mu} \ell(|\langle \boldsymbol{x}_\mu, \boldsymbol{W}_S\rangle|, |\langle \boldsymbol{x}_\mu, \boldsymbol{W}_T\rangle|) \right\}. \tag{33}$$

This is formally equivalent to have n independent systems. Introduce the overlaps with the projector $r_\mu^{(a)} = \langle \boldsymbol{x}_\mu, \boldsymbol{W}_a\rangle$ with indices $a = 0, \dots, n$ where $a = 0$ is the overlap with the ground truth and the others are the overlaps with the estimators of the n systems. Introduce those quantities in the replicated partition function via Dirac's deltas using their Fourier transform

$$\begin{aligned}\overline{\mathcal{Z}^n} &= \mathbb{E}_{\{\boldsymbol{x}_\mu\}} \int \prod_a \mathcal{D}\boldsymbol{W}_a \int \mathcal{D}(r,\hat{r}) \exp\left\{ -\frac{\beta}{2} \sum_{a,\mu} \ell(r_\mu^{(a)}, r_\mu^{(0)}) + i \sum_{a,\mu} \hat{r}_\mu^{(a)} r_\mu^{(a)} - i \sum_{a,\mu} \langle \boldsymbol{x}_\mu, \hat{r}_\mu^{(a)} \boldsymbol{W}_a\rangle \right\} = \\ &= \int \prod_a \mathcal{D}\boldsymbol{W}_a \int \mathcal{D}(r,\hat{r}) \exp\left\{ -\frac{\beta}{2} \sum_{a,\mu} \ell(r_\mu^{(a)}, r_\mu^{(0)}) + i \sum_{a,\mu} \hat{r}_\mu^{(a)} r_\mu^{(a)} - \frac{1}{2N} \sum_{\mu} \sum_{a,b=0}^{n} \hat{r}_\mu^{(a)} \hat{r}_\mu^{(b)} \langle \boldsymbol{W}_a, \boldsymbol{W}_b\rangle \right\}.\end{aligned} \tag{34}$$

Where $\mathcal{D}$ contains the normalization factor of the Fourier transform. We introduce the matrix of overlaps between estimators and ground ground truth $\boldsymbol{Q}$, $Q_{ab} = \frac{1}{N}\langle \boldsymbol{W}_a, \boldsymbol{W}_b\rangle$. This is done using the same idea of introducing delta function and gives a contribution $\frac{N}{2}\log\det \boldsymbol{Q}$ to the action [31, 35]. The equation can now be factorized in μ, so we can drop the μs from r and $\hat{r}$. Observing that

$$e^{-\frac{1}{2}\sum_{ab} \hat{r}^{(a)}\hat{r}^{(b)} Q_{ab}} = e^{\frac{1}{2}\sum_{ab} Q_{ab} \frac{\partial^2}{\partial h_a \partial h_b}} e^{-i \, h_a \hat{r}^{(a)}}\big|_{\{h_a=0\}_a}, \tag{35}$$

we can integrate over r and $\hat{r}$, and write a simplified replicated partition function $\overline{\mathcal{Z}^n} = \int \prod_{a\geq b=0}^{n} dQ_{ab} e^{NS(\boldsymbol{Q})}$ with action

$$S(\boldsymbol{Q}) = \frac{1}{2}\log\det \boldsymbol{Q} + \alpha \log \exp\left[\frac{1}{2} \sum_{a,b=0}^{n} Q_{a,b} \frac{\partial^2}{\partial h_a \partial h_b} \right] \exp\left[-\frac{\beta}{2} \sum_a \ell(h_0, h_a) \right]\Big|_{\{h_a=0\}_a}. \tag{36}$$

where the first term is an entropic term that accounts for the degeneracy of the matrix $\boldsymbol{Q}$ in the space of symmetric matrices. And the second term is an energetic term that accounts for the potential acting on the system.

Observe that so far we did not make any ansatz on the structure of the overlap matrix $\boldsymbol{Q}$. In the next subsections we will consider the 1-step replica symmetry breaking ansatz (1RSB).

B.2 1 step replica symmetric breaking

The 1RSB scheme consists in making an ansatz on the structure of the overlap matrix $\boldsymbol{Q}$ [30]. The assumption is that not all the replicated systems will have the same overlap - which correspond to the replica symmetric ansatz - but the systems are clustered. Systems inside the same cluster will have a larger overlap, systems outside the cluster will have a smaller overlap. This translates into the following parameters: q_1 overlaps inside the same cluster, q_0 overlaps in different clusters, x dimension of the clusters, finally m the overlap with the signal. Schematically we have

$$\boldsymbol{Q} = \begin{pmatrix} 1 & m & & m & \\ m & \begin{pmatrix} 1 & q_1 & q_1 \\ q_1 & 1 & q_1 \\ q_1 & q_1 & 1 \end{pmatrix} & & q_0 & \\ m & & q_0 & & \begin{pmatrix} 1 & q_1 & q_1 \\ q_1 & 1 & q_1 \\ q_1 & q_1 & 1 \end{pmatrix} \end{pmatrix} \tag{37}$$

with $\boldsymbol{Q}$ of dimension $(n+1)\times(n+1)$, and the inner matrices of dimension $x \times x$.

The analysis proceed as in a standard way: we derive the 1RSB free energy with the associated saddle point equations and move to the zero temperature limit [30, 33], we impose the marginal stability - corresponding to the threshold states - [28], finally we derive the label distribution [29].

For notational convenience we call $\gamma_a(x)$ the probability density function of a Gaussian with zero mean and (co-)variance a, and we use the symbol $\star$ to indicate the convolutions, i.e. $f \star g(x) = \int_{\mathbb{R}} f(x-t)g(t)dt$.

We can plug Eq. (37) into Eq. (36) and obtain 1RSB formulation of the action.

$$\begin{aligned} \frac{1}{n} S_{\text{1RSB}}(\boldsymbol{Q}) = \frac{1}{2}\log(1-q_1) + \frac{1}{2x}\log\frac{1-q_1+x(q_1-q_0)}{1-q_1} + \frac{1}{2}\frac{q_0-m^2}{1-q_1+x(q_1-q_0)} + \\ + \alpha \int_{\mathbb{R}^2} d\hat{y}dy\gamma_{\boldsymbol{\Sigma}}(\hat{y},y)\frac{1}{x}\log\left[\gamma_{q_1-q_0}\star\left(\gamma_{1-q_1}\star e^{-\frac{\beta}{2}\ell(\hat{y},y)}\right)^x\right]. \end{aligned} \tag{38}$$

and using Eq. (31) the free energy is $-\frac{1}{n\beta}S_{\text{1RSB}}$.

We can now take derivatives to find the saddle point equations.

$$\begin{aligned} \frac{1}{x}\left(\frac{1}{1-q_1} - \frac{1}{1-q_1+x\,(q_0-q_1)}\right) + \frac{q_0-m^2}{[1-q_1+x\,(q_0-q_1)]^2} = \\ = \alpha\int_{\mathbb{R}^2} d\hat{y}dy\gamma_\Sigma(\hat{y},y)e^{-x\,f(x,\hat{y},y)}\gamma_{q_1-q_0}\star\left[e^{x\,f(1,\hat{y},y)}\left(\frac{d}{d\hat{y}}f(1,\hat{y},y)\right)^2\right]; \end{aligned} \tag{39}$$

$$\frac{q_0-m^2}{[1-q_1+x\,(q_0-q_1)]^2} = \alpha\int_{\mathbb{R}^2} d\hat{y}dy\gamma_\Sigma(\hat{y},y)\left(\frac{d}{dh}f(x,\hat{y},y)\right)^2; \tag{40}$$

$$\frac{m}{1-q_1+x\,(q_0-q_1)} = \alpha\int_{\mathbb{R}^2} d\hat{y}dy\gamma_\Sigma(\hat{y},y)\frac{d^2}{d\hat{y}dy}f(x,\hat{y},\hat{y}); \tag{41}$$

where we define

$$\begin{aligned} f(1,\hat{y},y) &= \log\left[\gamma_{1-q_1}\star e^{-\frac{\beta}{2}\ell(\hat{y},y)}\right]; \\ f(x,\hat{y},y) &= \frac{1}{x}\log\left[\gamma_{q_1-q_0}\star\left(\gamma_{1-q_1}\star e^{-\frac{\beta}{2}\ell(\hat{y},y)}\right)^x\right] = \frac{1}{x}\log\left[\gamma_{q_1-q_0}\star e^{xf(1,\hat{y},y)}\right]; \\ \ell(\hat{y},y) &= (\hat{y}^2-y^2)^2; \end{aligned}$$

and Σ is a 2×2 covariance matrix with entries $\Sigma_{11}=1$, $\Sigma_{12}=m$ and $\Sigma_{22}=q_0$.

Zero temperature and parameter ansatz As the zero temperature goes to zero the order parameters q_1 and x needs to be rescaled with the temperature as $q_1 \approx 1-\chi T$ and $x \approx zT$ with $\chi, z \sim O(1)$. Instead m and q_0 will not be affected by the limit. The reason why some parameters need to be rescaled is that as long as there is a positive temperature the replicated systems exploit this thermal

energy to fluctuate in the basin of the minimum, therefor their overlap q_1 is given by average overlap in the basin of attraction. When the temperature drops to zero the thermal goes to zero and all the system shrink to a point. χ represent the fluctuation that systems can have when they receive an infinitesimal amount of thermal energy. With the same physical reason, also the cluster itself shrinks to a point. The rate of convergence to the point is given by z.

Under those observation we obtain the 1RSB free energy

$$-f(\boldsymbol{Q}) \approx_{\beta\gg 1} \frac{1}{2z}\log\frac{\chi+z(1-q_0)}{\chi}+\frac{1}{2}\frac{q_0}{\chi+z(1-q_0)}+ \\ +\frac{\alpha}{y}\int_{\mathbb{R}} dy\gamma_{\boldsymbol{\Sigma}}(\hat{y},y)\log\gamma_{1-q_0}\star e^{\frac{-y}{2}V(\hat{y},y|\chi)}\Big|_{\hat{y}=0}. \tag{42}$$

and the corresponding saddle point equations from Eqs. (39-41).

However, the solution of those equation will lead to the global minimum of the loss, while we are interested in the threshold states. We follow the idea of [28] where z is used as a Lagrange multiplier that selects the threshold states. z is fixed by imposing the marginal stability condition, i.e. that the spectrum of the Hessian of the minima in consideration touches zero, following [29] this is given by

$$1=\frac{\alpha}{4}\chi^2\int_{\mathbb{R}} dy\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}\frac{\gamma_1\star\left[V''(\hat{y},y|\chi)^2e^{-\frac{z}{2}V(\hat{y},y|\chi)}\right]}{\gamma_1\star e^{-\frac{z}{2}V(\hat{y},y|\chi)}}\Bigg|_{\hat{y}=0}. \tag{43}$$

At the threshold the overlap with the signal is zero, $m=0$, and for symmetries of the problem also $q_0=0$. In fact $m=0$ is always a solution of Eq. (41), and if $m=0$ then $q_0=0$ is also solution of the second saddle point equation, Eq. (40), as in the RHS the Gaussian becomes degenerate in h and $d_h f(x,h,h_0)|_{h=0}$ being an odd function in h. Therefore we can restrict the equations to just the one for χ, Eq. (39) that becomes

$$\frac{\beta^2}{z}\left(\frac{1}{\chi}-\frac{1}{\chi+z}\right)=\alpha\int_{\mathbb{R}} dy\gamma_1(y)e^{-x\ f(x,0,y)}\int_{\mathbb{R}} d\hat{y}\gamma_1(\hat{y})\left[e^{x\ f(1,\hat{y},y)}\left(\frac{d}{d\hat{y}}f(1,\hat{y},y)\right)^2\right]. \tag{44}$$

Rewriting these equation expliciting the convolutions we obtain the equations presented in the main text. With those element we can obtain the label distribution presented in the main text.

C Additional details on the numerical experiments

C.1 Dynamics with shuffled labels

In the main text we showed that the dynamics is always attracted by the threshold states before (possibly) finding the good direction that leads to the optimal solution. To visually complement the explanation, in Fig. 6 we consider the dynamics in the same setting of Fig. 2, same dataset and same initialization, but using shuffled labels. The upper figure shows the density of eigenvalues during the evolution while the lower figure considers the histogram for five cuts at iteration 1, 453, 906, 1208 and 1510. We remark that in the initial stages of their evolution the two simulations follow the same dynamics (iteration 1, 453, and 906). Only after a transient the simulation with the correct labels shows the presence of the BBP transition (iteration 1208) and converge to the solution (iteration 1510). Instead, the simulation with the shuffled labels keeps approaching the threshold states for the rest of the simulation time tending to the marginal condition. We remark that in the upper figure the smallest (dashed orange line) and second smallest eigenvalues (solid green line) that are always very close, as it is expect since in this case the distribution of the eigenvalues always forms a bulk.

C.2 Numerical threshold

On the choice of the learning rate. The dynamics considered in the main text is gradient flow, while numerically we rely on the discretized version of the algorithm, namely gradient descent. In order to have agreement between the theoretical analysis and the simulations we must consider a learning rate sufficiently small to reduce the discrepancy between the two algorithms. We consider different input dimensions and tested different learning rates, in Fig. 7 we report this process for $N=1024$. In the

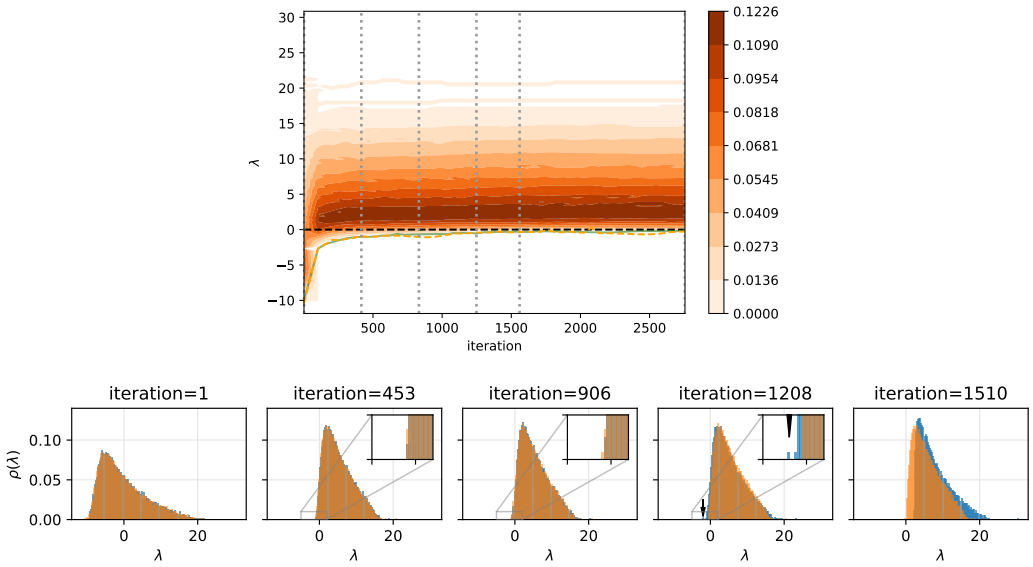

Figure 6: Comparison of the properties of the Hessian for phase retrieval between the problem with the correct labels and the problem with shuffled labels. The system simulated has size $N = 2048$ and $\alpha = 10$. The database considered in the simulation in the upper figure is the one used for Fig. 2 with the shuffled labels. The figures represent the same quantities plotted in Fig. 2 and are intended to make the comparison between the evolution in the two settings. The figure below represents the distribution of the eigenvalues during the dynamics for the problem with correct and shuffled labels, respectively in blue and orange.

figure on the left we can observe that the dynamics of 10 simulations at $\alpha = 7$ testing the learning rates : $\eta = 2 \times 10^{-4}$ (solid lines), $\eta = 1 \times 10^{-4}$ (dashed lines), $\eta = 5 \times 10^{-5}$ (dotted lines). We can notice that the lines are nicely overlapping, this give us the learning rate $\eta = 2 \times 10^{-4}$ that we use in our simulations. In the right figure we show the fraction of successful simulations, where success is defined as the fact that the overlap $\boldsymbol{W} \cdot \boldsymbol{W}^*/N > 0.99$. Notice that the line is shifting until we reach the learning rate $\eta = 2 \times 10^{-4}$.

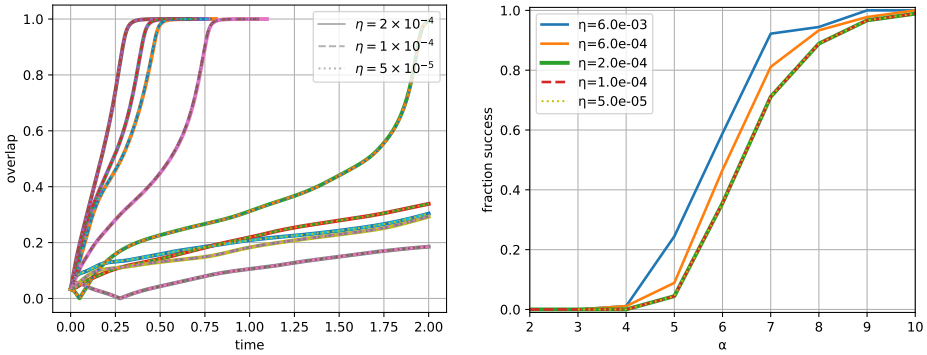

Figure 7: Choice of the learning rate for the phase retrieval problem with input dimension $N = 1024$. On the left we show the overlap $\boldsymbol{W} \cdot \boldsymbol{W}^*/N$ in time for 10 different simulations at $\alpha = 7$. The different line styles refer to different learning rates. On the right we show the fraction of success, as reported in Fig. 5, as we change the learning rate. In the simulations we adapted the stopping criterion to the learning rate so that the curve get to the same simulation time, more specifically for the left panel the stopping times are : $1000 \log_2 N$ for $\eta = 2 \times 10^{-4}$, $2000 \log_2 N$ for $\eta = 1 \times 10^{-4}$, and $4000 \log_2 N$ for $\eta = 5 \times 10^{-5}$.

On the nature of the transition. Our analysis assumes a first order transition between: the easy phase, where gradient flow with high probability solves the problem; and the hard phase, where gradient flow with high probability does not find the optimal solution and this solution is information theoretically achievable. According to this hypothesis threshold identifies a jump in the overlap between estimator and ground truth, which defines the first order transition. We bring as additional support on the matter Fig. 8, where we show the final overlaps reached by the simulations for different size of the dataset α. As α increases, the figures show that densities become bimodal and finally concentrate on 1 meaning that the problem is easy. This is an indicator of a first order transition. Contrary to the second order (continuous) transition, where the distribution of the overlaps should be unimodal.

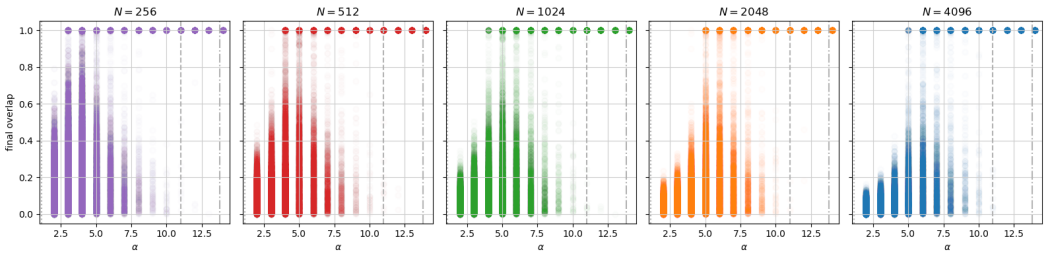

Figure 8: The five figures show the final overlaps between estimator and teacher for a large number of simulations at different values of α. The figure refer to different input dimensions, from left to right : 256, 512, 1024, 2048, and 4096.