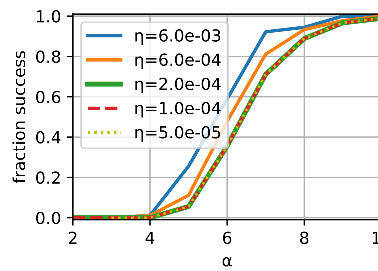
We kindly thank the reviewers (**R1**, **R2**, **R3**) for their comments. We are certain that they will improve the quality of the work. We also thank for pointing out typos (**R2**) and missing references (**R2**), and for giving suggestions to improve the flow of the text (**R2**) and the quality of the figures (**R1**). We will integrate these recommendations and the comments below in the final version. We discuss the main points below.

R1: **Why in Fig. 2 the Hessian is not marginal before the transition?** For the size used in Fig. 2 (N=2048) the system is already after the transition for $\alpha = 10$, see Fig. 3 and the discussion of finite size effects in Sec 2.2. According to our results, the dynamics converges to threshold states at long time only for α smaller than the transition point. In this case the spectrum of the hessian tends to a gapless one. However, the dynamics after the transition approaches only on short times and approximately the threshold states and then aims towards the signal, which is what shown in Fig. 2 We will add in the supplementary a version of Fig.2 for a value of α (and N) before the transition where the marginality clearly appears.

R1: **Mismatch between theory and simulations** As the reviewer remarked, the reasons behind the mismatch are probably due to the finite size effects or the 1RSB approximation. Note, however, that our theory holds on timescales not diverging with N—an additional mechanism that would make the system detect the signal on larger times (polynomial in N), could lead to a smaller value of α_c. We will emphasize this possibility in the conclusion.

R1: Learning rate in the experiments. As **R1** noted it may be possible that a high learning rate η shrinks the basin of attraction of minima, and makes weakly stable states and marginal ones inaccessible to the dynamics. Therefore this effect may shift to the right the numerical estimation of the transition. Following the advice of **R1**, we ran experiments for $N = 1024$ and several learning rates η. The figure here shows the fraction of successes for dynamics in 90 samples as a function of α. We observe only a slight shift to the right of the curve (with respect to the η used in the paper) before reaching a saturation. We will update the figures in the final version and add a brief discussion.



R2: **Missing proofs.** The methods that we use are not rigorous but, as **R1** mentioned, are widely accepted in statistical physics, and have been made rigorous in some specific cases. We are not aware of any proof technique for the setting that we discuss, the formal proof is left as an open problem.

R2: **Knowing the distribution** $P(\hat{y}, y)$**.** Our results indicate that the algorithmic transition happens when the threshold states undergo a BBP transition. We are left with the analysis of the Hessian of threshold states which is a random matrix with the same form of the one analyzed in reference [26] where it was shown that the BBP threshold depends only on $P(\hat{y}, y)$. This is reported in Theorem 1 of [26]. Therefore $P(\hat{y}, y)$ of threshold states can be used to determine the BBP transition and consequently the algorithmic transition, as indeed we do in Sec. 2. Reducing the problem of determining α_c to the knowledge of $P(\hat{y}, y)$ is a considerable advantage, since this distribution can be obtained by replica theory, allowing us to estimate the threshold analytically.

R2: **The use of** α **N rather then m** We adopted notations used in signal processing where the relevant transitions happen at finite α. We will add clarifications in the text when it can give rise to confusion.

R2: **Comment on line 61** The threshold $\alpha_{alg} = 1.13$ is the algorithmic threshold of approximate message passing. Above the threshold this algorithm can find the solution. The information theoretic transition instead occurs at $\alpha_{IT} = 1$.

R2: **Why threshold energy is higher for more samples?** The threshold states can be pictured as the states where the dynamics converge if the labels are shuffled or substituted by random variables. In this setting when $\alpha > 1$ with high probability there are no minima that have zero training error. Therefore the system has to satisfy αN constraint with N variables, as α increases above 1 it is harder and harder to satisfy the constraints and the dynamics end up on minima that violate more and more constraints. This implies that the energy grows with α. We will explain this point.

R3: **The results are limited to perceptrons.** In this work we indeed analyzed a specific perceptron (phase retrieval) and we expect that the same techniques could be applied to one-hidden layer neural networks with a finite number of hidden units where some replica approaches have been already applied (e.g. Aubin et al. NeurIPS 2018). Going toward yet more complex model is indeed an open problem.

R3: **There are no new phenomena compared to previous studies.** Previous work [8], that discussed very similar phenomenology, were on the spiked matrix-tensor model that is not a supervised learning problem. The phase retrieval we analyze is a genuine simple neural network and our work thus shows for the first time that this phenomenology extends to this more relevant setting. Although the underlying phenomenon is the same, we remark that the techniques used in previous works are difficult to extend to our case, and we had to develop an alternative approach.