We thank the reviewers for their careful consideration of our paper. Our analysis gives a new sample complexity threshold that shows how the generalization error in a teacher-student scenario depends on the number of teacher units m^* and the input dimension d, a result highlighted by **R1**. We also derive the convergence rate of GD as a function of m^* and d in the limit of many samples, a result that both **R1** and **R3** commended. We are glad that the reviewers found our proofs elegant and concise (**R4**), and appreciated that we validate these theoretical findings by extensive numerical experiments that are well laid out and clearly motivated (**R1**,**R4**), and also use the string method as new tool in the ML literature (**R2**). We are encouraged that **R4** thought that our paper clearly outlines its main contributions and reviews the relevant literature. We are also grateful for the reviewers specific suggestions to improve our results. One primary concern was that some of the arguments were unclear or incomplete. We agree. *In the revised version we have followed the reviewers recommendations and added proofs and new results accordingly*, as detailed next. We also thank **R1** and **R3** for pointing out typos and residuals of previous notation, i.e. $P \leftarrow n$ and $Qm^* \leftarrow A^*$.

R3: *the considered problem in this paper is quite over-idealized.* We respectfully disagree. In particular, questions related to the generalization error or the convergence rate of GD remain mostly open even for shallow networks–here we give precise answers to these questions in the context of networks with quadratic activation functions, which, as **R4** remarks, "are still in and of themselves an active area of interest for the ML community."

R3: *[They] considered only a little more general case than Ref. [24].* We don't think this assessment is fair. In Ref. [24] the authors give a nice and thorough analysis of the case when $m^* > d$, i.e. A^* is full rank and the threshold is at $n = d(d+1)/2$ by standard linear algebra results. The situation with $m^* < d$ that we focus on is relevant for phase retrieval and has a different sample complexity threshold that depends on m^* and d, the proof of which requires analysis techniques from random geometry that are different from the ones used in Ref. [24].

R1: *authors should discuss the limitations and applications of this work and future work*; **R2**: *It would better if this paper provides a more explicit exposition on the improvement over existing results.* We will expand on these points in the introduction of our paper, in particular stress that most of the results in the literature focus on the empirical loss, showing that over-parameterization helps to find good minima (in particular refs.[8,15,16]). Here we focus on generalization error and convergence rate of GD, which are less studied and not as well understood.

R2: *Theorem 4.2 only applies to the identity initialization. [..] I would like to recommend acceptance if this paper provides formal analysis to random initialization.* **R3**: *More details need for the phase that other terms in Eq. (17) take over rather than the quadratic term.* Our analysis of Eq. (14) was kept informal for readability, but it is not hard to make it rigorous and we will add a proof of Eqs. (20) and (23) in the Appendix. Also, we had focused on the case of $A(0) = \mathrm{Id}$ because, if the initial weights $w_i(0)$ are independent standard Gaussian vectors, $A(0) = m^{-1}\sum_{i=1}^m w_i^T(0)w_i(0)$ is close to the identity as long as $m \gg d$ by the Law of Large Numbers. However, the convergence analysis can be generalized to arbitrary initial $A(0)$ with full rank ($m \geq d$) and we can show that: (i) there exists a constant $C > 0$ such that, for all $t \geq 0$, $E(A(t)) \leq 1/(1+Ct)$ (including when $m^* < d$, i.e. when A^* is rank-deficient, which is the difficult case for analysis) and (ii) Eqs. (20) and (23) in the paper still hold as $t \to \infty$. We will add this result as a new theorem along with its proof based on the analysis of the dynamics of the off-diagonal term of $A(t)$: since $A(t) \to A^*$, these terms must eventually decay to zero at a rate that can be shown to be exponential, which is enough to establish the results (including the non-asymptotic convergence rate, using convexity of the loss).

R3: *The main results Theorem 3.1 only holds for a special case, i.e., $m^* = 1$. From the perspective of matrix analysis, it seems impossible to conclude the results rigorously without using the information from A^*.* The proof of Theorem 3.1 for $m^* = 1$ is based on a geometric rephrasing that uses the rank 1 structure of A^*. Extending this proof to A^* with rank $m^* > 1$ seems nontrivial, we agree, as the geometrical problem becomes more complex. Still our numerical experiments strongly suggest that the result from our heuristic argument is correct.

R4: *Lemmas 2.1 and 2.2 should explicitly be derived in the Appendix. Theorem 4.1 also needs a rigorous proof.* We will add these proofs: Lemma 2.1 follows by direct calculation from Eqs. (3) and (4); Lemma 2.2. follows from Lemma 2.1 by averaging over the Gaussian weights using Wick's theorem; and Theorem/Proposition 4.1 follows from the Stable Manifold Theorem which states that the GD flow reaches a local minimum of the empirical loss with probability one, and all these minima have zero loss by convexity of the loss in A – the result is stated under the assumption that $A(0)$ has full rank, for otherwise rank deficiency introduces a nontrivial constraint which may preclude the dynamics to reach the minimum; we note however that this assumption is sufficient but may not be necessary.

R4: *In line 33 of the Appendix, [..] In line 74 of the Appendix, [..]* When $m^* = 1$, A^* has only one nonzero eigenvalue and at most one eigenvalue of $A - A^*$ is negative. When $m^* > 1$, the number $\frac{1}{2}(d-m^*)(d-m^*-1)$ of constraints is obtained by requiring that $A^\perp$ be nonnegative definite in the $d - m^*$ subspace were it lives, which gives (A.13), but this argument is only heuristic for the reason stated in lines 78-80.

R4: *How the confidence intervals in Figure 1 are being calculated? What is the motivation behind the power law assumption?* The extrapolations consider a power law divergence at the transition [Cugliandolo, Kurchan'93] and the confidence interval comes assuming a t-student distribution of the samples.