We thank all the reviewers for their diligent reading of our paper, we address their comments in order of appearance:

Reviewer #1 **Presentation of the theorem** We thank the reviewer for the suggestions concerning the presentation of the theorem, which we will implement in the final version. **Lipschitz** The function $f(m)$ is Lipschitz-continuous and we take our initial conditions to be deterministic by assumption A3, so the uniqueness of the solution is guaranteed. **Implications of Eq. (10)** This result only applies to networks where the first layer is trained, and we report different behaviours when we train both layers in Sec. 3. We used SGD for training in both cases, so the (indirect) implications of Eq. (10) and (15) are that firstly, the (implicit) regularisation observed when training both layers is not a property of just the algorithm, and thus secondly, for general neural networks, we still have to find the precise mechanism allowing bigger networks to generalise better. **Analytical results** We cannot solve the ODE in closed form. We obtain our analytic results by linearising the equations in the limit of small noise σ around the fixed point at zero noise. **Line 82** We mean that there exist weights which are fixed points of SGD dynamics, in the sense that the generalisation error will stay stationary at a value $\hat{\epsilon}_g$ if SGD is started with these weights. However, starting from random initialisation, SGD finds weights with a generalisation error that is higher than $\hat{\epsilon}_g$. **Line 118** The teacher-student overlaps $R^\mu = [R^\mu_{in}]$ capture the *similarity* between the weights of the ith student node and the nth teacher node. **Line 124** By "a closed set of ODEs", we mean a set of coupled ODEs where the variables that appear in them are governed by an ODE in that set. Our ODEs do not have a known closed-form solution. **Line 147** We mean that running SGD will yield networks where the generalisation error (="performance") improves or worsens with over-parameterisation (="drastically different").

Reviewer #2 **Message for practitioners:** We have to be careful with the widely cited claim that bigger networks are better. It is not always true and we still have to understand the range of the cases in which it is. **Eqs. S26-S30** The reviewer is right that Eqs. S26-S30 are only valid for sigmoidal activation; we will emphasise this in the revised manuscript. **Additional experiments and relation to mean field limit**: We numerically checked the behaviour of ϵ_g when the number of student hidden nodes becomes much larger, see Fig. 1. This is the same plot as Fig. 4 of the main paper, only that we extended the range of K to 1000. We plot the final generalisation error after convergence. The plot demonstrates that the trend observed in the paper persists: the normalised network, where we train only the first layer and divide the network output by the number of hidden units, beats the performance of a two-layer network where we train both layers. We think our results connect with the cited mean-field analyses in that they suggest that the "distributional" fixed points corresponding to the mean-field analysis persist even down to relatively small sizes of the hidden layer. However, we also found other fixed points which are not captured by the mean field analysis, such as the one leading to the increasing generalisation error. Elaborating this connection and pinning down the differences more precisely is a very interesting direction for future research.

Reviewer #3 **Validity of the expansion** We found the results of our expansion in good agreement with numerics up to a noise with $\sigma \simeq 0.3$. **Details on the numerical experiments:** We will collect all the parameters needed to reproduce the experiments, including the reference to codes, in one section of the appendix, instead of scattering them in various places of the text. The stopping criterion is a fixed number of steps chosen in each experiment manually to be large enough to reach a stationary point of the generalisation error, usually on the order of $10^6 N$. **Context of the MNIST experiments:** We intended to verify the qualitative validity of our result in Eq. (10) for the final test error of networks where only the first layer is trained, in a setting which violates two key assumptions of our theoretical treatment: (1) that all inputs are i.i.d. draws from the multi-normal distribution and (2) that at every step, we use a previously unseen sample (x^μ, y^μ). Crucially, it is still another teacher that generates the labels for the images. Indeed, the plots in Fig. S7 show that substituting MNIST inputs (orange curve) for Gausssian inputs (blue) does not change the shape of the $\epsilon_g^* \sim L$ curve, with the minimal final test error occurring for $K = M$. **Proof precisions.** Regarding the reviewer's comments on the proof (we will clarify all points in the final version): (i) The expectation $\mathbb{E}_\mu$ is the conditional expectation conditioned on the state of the Markov chain at step μ, m^μ. (ii) We use q for any but only *one* of the time-dependent order parameters and m to denote the set of all order parameters, resp. (iii) We were indeed not precise about the step from S11 to S12; we crucially also used assumption (A3) by which the initial macroscopic state is deterministic and therefore the average $\mathbb{E}$ in that line is just an average over the first sample shown during training. (iv) We kept the discussion of the coupling trick more compact than the other parts of the proof because it is not original, but due to Wang et al., Ref. 46. We will expand this section in the final version to make the paper more self-contained. The stochastic process b^μ lives in the same space as m^μ, and similarly for the deterministic process d^μ. They do not require additional assumptions.

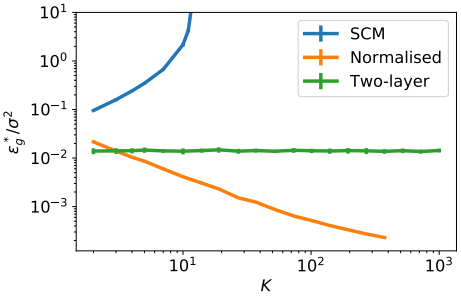

Figure 1: Asymptotic generalisation error of linear networks as a function of the number of student nodes K. Parameters: $N = 100, M = 4, v^* = 4, \eta = 0.01, \sigma = 0.01$.