We thank the reviewers for their valuable and constructive feedback. We are pleased that they generally appreciated our theoretical analysis in a simplified setting of phenomena such as slowing down of dynamics, rich and lazy learning, and over-fitting. We will implement all their suggestions which we think will greatly improve the quality of the text. We address specific points below.

Reviewer 1 The discussion about related contributions was focused on recent works addressing the mean-field limit of infinitely wide one-hidden layer networks, since this was the theoretical framework we address. We agree with the **R1** that previous works mainly from statistical physics on few hidden units and teacher-student settings provide an interesting and relevant context for comparison. We will add a discussion. Concerning the dataset, in a more complex case the average over the dataset is not independent of the parameters (Eq. 9 does not hold). From numerical experiments, and ongoing work, we find that this is related to specialization of nodes, whose dynamics is governed by subsets of the whole dataset. We will add a brief discussion in the conclusion, as well as references to previous cases where specialization was observed and analyzed. We will also specify in the abstract that the analytical treatment holds for the infinite dataset case, and that corrections are studied heuristically and numerically.

Reviewer 2: The derivation of our results is based on theoretical physics methods, which are within reach of probability theory, with extensions of methods used in [17-22] and developed to rigorously treat the hydrodynamic limit in physics, e.g. *C. Kipnis and C. Landim, Scaling limits of interacting particle systems, Springer (2013); S. Serfaty, Coulomb gases and Ginzburg-Landau vortices, Zurich Lect. in Adv. Math., EMS (2015)*. This makes us confident that all our results can be made rigorous, but we agree with **R2** that the non-rigorous steps should be stated and discussed more explicitly. We will do that when passing from eq. (1) to (2) [for which we assume the convergence of the empirical distribution to its average], and in the derivation of eq. (5) [for which we assume the convergence to the hydrodynamic limit of the dynamical process]. As for sec. 3.1, we will revise it to make it more clear. The parameters are normally distributed at initialization (see lines 71, 72), this implies that $a(0)$, $w^{\parallel}(0)$, and each component of $\mathbf{w}_{\perp}(0)$ in the space orthogonal to $\mathbf{w}^*$ are Gaussian distributed. We could have chosen different distributions, but we decided to focus on the Gaussian case as this is often used in practice.

Reviewer 3: We agree with the **R3** that considering the finiteness of the dataset is important, but it is also very challenging analytically. Our numerical experiments and heuristic arguments allow us to qualitatively understand the main new effects introduced by a finite dataset. An analytical treatment can be done in the case of a very large but finite dataset, in which one studies the Gaussian fluctuations of empirical averages around their means (these are small and of the order of (number of data)$^{-1/2}$), see eq. 18. A more complete treatment in which fluctuations of the empirical average are of the same order of the mean would be more conclusive indeed. However, this would require to control the entire distribution of the empirical averages, a quite difficult task in our case. One possibility to achieve this goal might be considering the simultaneous scaling limit $M \to \infty$ and $N \to \infty$ with a fixed ratio M/N, as done for example in Random Features models. This could provide a more conclusive picture but it is beyond the scope of the current work.

Reviewer 4: We chose a simple separable dataset to keep the model analytically tractable. We agree with the **R4** that realistic dataset are certainly more complex. However, we expect the dynamics of our simple model to share qualitative aspects with more realistic tasks, as in the early training a dataset can be roughly approximated as "two clusters with separate averages" and only later more complex features are learned. Several works have hinted at such progressive learning hierarchy *(e.g. A.M. Saxe, et al., Learning hierarchical categories in deep neural networks, (2013))*. Our numerical experiment was intended to illustrate our results for a realistic—yet still simple—dataset, since being limited to binary classification through a single hidden layer, one cannot reach good performance on real challenging tasks. Following the suggestion of **R4**, we report below (and will add to the SM) the results of experiments on CIFAR10 and ImageNet (analog to the ones of Fig. 4 on MNIST). The initial dynamics, while only poorly learning the tasks, does resemble the one we analyze in our simple model and found for MNIST (compare with Fig. 2 and Fig. 4). For ImageNet there is a difference in speed for the two classes, a feature that we could easily capture in our model by considering a different data distribution. Finally, we will add a broader discussion on theoretical studies on learning dynamics, in particular the ones on implicit regularization in which the dynamics of simplified models has also been analyzed.