We thank the reviewers for their helpful feedback. The reviews emphasized the significance of our results for the analysis of finite-width neural networks (**R1**, **R3**) and the strength of our technical contributions (**R1**, **R2**). One primary concern was that our results are only asymptotic. We address this concern and other questions below.

1. Theory

R2 & **R4**: *Being asymptotic in both n and t, Theorem 3.3 is not very interesting.* We agree that nonasymptotic bounds would be highly desirable, but such results are difficult and have only been established in limited settings that are not applicable to neural network training (e.g. weak interactions in [1]) beyond the *lazy* regime. Despite being asymptotic in n (just like the classical CLT), our results give insights on the evolution of the deviations from mean-field limit and set the groundwork for nonasymptotic results. In particular, they reveal a benign dependency on ambient dimension (see point below). Regarding the dependence on t, existing bounds on the fluctuations (cf. [2]) grow exponentially in t, whereas we show a finite bound as $t \to \infty$. We also discuss the decay of the fluctuations at finite t in the first remark under Sec. 2 below. Finally, we check the validity of our results in numerical experiments under finite n and t.

R1 & **R3**: *Assumption on $\hat{D}$.* We assume it is a *closed manifold* – a compact manifold without boundary – like a sphere.

R1: *Do the assumptions on activations hold for $\tanh$ and erf?* Yes. The only requirements are the Universal Approximation Theorem and our smoothness assumptions in Sec. 2.1 in the paper.

R2: *Is Prop. 2.1 tangential?* Not really - it shows by adding the regularization we can control the 2-norm of the loss minimizer (consistent with Col. 1 & 5, Row 2 in Fig. 1), on which the fluctuation bound in Thm. 3.3 crucially depends.

R2: *The bound in Theorem 3.3 is not truly dimension free.* Our bound depends on the norm of the target function in the variation-norm space introduced in Sec. 2.1. As the reviewer points out, the dependency on dimension is implicit through this norm, which is shown in [3] to depend polynomially in the dimension for functions with hidden dependency on low-dim structures.

R2: *Results under time discretization are lacking.* Indeed, we leave this for future work, though our continuous-time analysis already yields insights. Moreover, we see from Col. 3 of Fig. 1 that the loss evolutions under gradient descent nicely agree across different n, showing empirical consistency of the discretization scheme.

R3: *Are $\limsup_{t\to\infty}$ and $\lim_{n\to\infty}$ exchangeable in Thm. 3.3? How about changing n to n^a for $a \in (0, 1)$?* Not in our result (including in the latter case), unfortunately, as the $O(n^{-1})$ scaling of the fluctuations in n at finite time (thanks to CLT and the continuity of the flow) may not be preserved at the $t \to \infty$ limit. This is worthy of future investigations.

R3: *Concrete examples for assumptions (71) and (131).* We show in C.1.1 and C.2.1 that (71) and (131) can be satisfied if the flow $\mathbf{\Theta}_t(\boldsymbol{\theta})$ converges at a uniform rate of $O(t^{-\alpha})$ with $\alpha > 2$ and $\alpha > \frac{3}{2}$, respectively. Also, an alternative to (71) is $\int_0^\infty (\mathcal{L}(\mu_t))^{1/2} dt < \infty$, and so a sufficient condition is for the loss value to decrease faster than $O(t^{-2})$.

R3: *More explanations of $\boldsymbol{T}_t$.* We will add: $\forall \boldsymbol{\theta} \in D$, $\boldsymbol{T}_t(\boldsymbol{\theta})$ captures the deviation of the flow $\mathbf{\Theta}_t(\boldsymbol{\theta})$ due to the "initial deviation" ω_0, i.e. $\boldsymbol{T}_t = \lim_{n\to\infty} n^{1/2}(\mathbf{\Theta}_t^{(n)} - \mathbf{\Theta}_t)$. $\boldsymbol{T}_t$ satisfies an infinite-dim. linear ODE, (86). To control it, we show that 1) the asymptotic linear operator is PSD; 2) the source term lies in the range of the linear operator; and 3) finite time perturbations are controlled.

2. Experimental results

R1: *Why do the average fluctuations decay with t?* While Thm. 3.3 only speaks about the $t \to \infty$ limit, we can study the long-term behavior of $\mathbb{E}_0 \|g_t\|_{\hat{\nu}}$ by analyzing the t-asymptotic version of (25) or (86), $\dot{\boldsymbol{T}}_t = -(\mathcal{A}_\infty^{(K)} + \mathcal{A}_\infty^{(V)})\boldsymbol{T}_t + \boldsymbol{b}_\infty$. (Note that this also describes the exact dynamics of the fluctuations if we set $\mu_0 = \mu_\infty$.) In the unregularized case, we expect that f_∞ interpolates the data, and hence $\mathcal{A}_\infty^{(V)} = 0$. Thus, the solution to above is $\boldsymbol{T}_t = (1 - e^{-t\mathcal{A}_\infty^{(K)}})(\mathcal{A}_\infty^{(K)})^\dagger \boldsymbol{b}_\infty$. Also, in the ERM setting, $\mathcal{A}_t^{(K)}$ is a PSD operator with finitely many nonzero eigenvalues, and hence its nonzero eigenspaces are spanned by eigenfunctions $v_1, ..., v_k$ associated with eigenvalues $\lambda_1, ..., \lambda_k > 0$. By Lemma C.3, we can express $\boldsymbol{b}_\infty = \sum_{i=1}^k c_i v_i$. Using (26), we get $\|g_t\|_{\hat{\nu}}^2 = \|\bar{g}_\infty\|_{\hat{\nu}}^2 - \langle \boldsymbol{b}_\infty, (I - e^{-2t\mathcal{A}_\infty^{(K)}})(\mathcal{A}_\infty^{(K)})^\dagger \boldsymbol{b}_\infty \rangle = \|\bar{g}_\infty\|_{\hat{\nu}}^2 - \sum_{i=1}^k \lambda_j^{-1} \left(1 - e^{-2\lambda_j t}\right) c_i^2$, which decreases monotonically in t. This is consistent with the long-term behavior of the fluctuations in Col. 2, Rows 1 & 3 in Fig. 1.

R2: *Why do the fluctuations decrease faster for smaller n?* This is due to finite-n effects in the fluctuation dynamics, which are captured in (63) but not (25), and which decrease as n grows.

R2: *Why the alignment or lack thereof with teacher network's TV- and 2-norm in Fig. 1?* When (and only when) regularized, both the TV- and 2-norm of the student are controlled by those of the teacher, consistent with Prop. 2.1.

R2: *Why are some of the behaviors plotted in Fig. 1 non-monotonic?* Thm. 3.3 does not guarantee monotonic decay of the fluctuations during finite time, but only prescribes its behavior as $t \to \infty$. The calculations above of the fluctuation decay is also an asymptotic analysis. As for the norms, their non-monotonic evolution indicates that the regularization's effect become relatively stronger later in training, when the function reconstruction loss is low.

R1: *In Col. 3 of Fig. 1, why is the loss independent from width?* We first note that Col. 3 plots the training / population loss, as the study of generalization is beyond the scope of this paper (which is why we don't distinguish between ν and $\hat{\nu}$). The good agreement of loss evolution for different n validates empirically the convergence to a mean-fields solution.

R1: *How do we choose the teacher's neuron in Fig. 1?* We chose the teacher's neurons to have $c = 1$ and $\boldsymbol{z}$ randomly sampled on the hypersphere. For the plots on Col. 1, we chose one of the neurons as the "marker".

[1] Durmus et al. "An Elementary...". [2] Mei, Misiakiewicz, Montanari, "Mean-field...". [3] Bach, "Breaking...".