We thank all reviewers for their time and understanding of the results, and we are glad that all reviewers appreciate the introduction of the first CIs that feature a double-descent behavior. We first address 2 points raised by several referees and then address comments by individual reviewers. All typos/writing issues will be fixed as suggested in the reviews. We believe the rebuttal addresses all actionable concerns, and we invite the referees to revisit their scores in light of this.

• **Why does (8) hold?** By (3), the weighted average in (8) is equal to -$\mathrm{trace}[\boldsymbol{T}_0(\boldsymbol{e}_j)]/\left\|\boldsymbol{y}-\boldsymbol{A}\widehat{\boldsymbol{\alpha}}\right\|_2$ in Theorem 1. The correctness of (8) then follows by Theorem 1. The term [additive correction] in (8) includes traces of $\boldsymbol{T}_1, \boldsymbol{T}_L$ and the first term in (11). We'll clarify in the camera-ready version. **This addresses a major concern from Reviewers 2, 3, 4.**
• **Why vector $\boldsymbol{\xi}$ is the argument of $\hat{f}$ in (3)?** The variable name $\boldsymbol{\xi}$ in (3) was picked to avoid confusion with the observed feature vectors $(\boldsymbol{x}_i)_{i\in[n]}$, since the function $\hat{f}$ itself in (3) depends implicitly on $(\boldsymbol{x}_i)_{i\in[n]}$ through $\widehat{\boldsymbol{\alpha}}$. It is thus important to avoid $\boldsymbol{x}_i$ as argument for the definition of the function $\hat{f}:\mathbb{R}^p\to\mathbb{R}$ in (3); for clarity we avoided $\boldsymbol{x}$ as well. Applied to $\boldsymbol{x}_i$, quantity $\hat{f}(\boldsymbol{x}_i)$ depends on $\boldsymbol{x}_i$ directly through the input given to $\hat{f}$ and implicitly through $\widehat{\boldsymbol{\alpha}}$ in the definition of $\hat{f}$. The implicit dependence of $\hat{f}$ on $\boldsymbol{x}_i$ through $\widehat{\boldsymbol{\alpha}}$ is the reason why classical CLT would not apply to $n^{-1/2}\sum_i \hat{f}(\boldsymbol{x}_i)$. Since $\hat{f}$ is the prediction function after training, $\hat{f}(\boldsymbol{\xi})$ can be thought of as the prediction at a point of interest such as an unlabeled future observation, though this is only a thought experiment and $\boldsymbol{\xi}$ is only used as the argument of $\hat{f}$ to set the notation for partial derivatives. The quantity ξ_L in Theorem 1 is unrelated to the vector $\boldsymbol{\xi}$ in (3); to avoid confusion we will clarify and use ζ_L instead. **This addresses a major concern from Reviewers 1, 2 and 3.**

Reviewer 1: *"the linear part of the target is the only thing that RF model can learn"* thanks–we'll add this comment with a reference as a motivation for the β_js being the targets of the CIs.*" (...) the CI is only established for the linear part of the target (...) In this case what would be the motivation of considering a nonlinear perturbation "* ⇒ The nonlinear perturbation adds noise correlated with the features, which is possibly more hostile than the independent additive noise $\boldsymbol{\varepsilon}$. This model of nonlinear perturbation wasn't introduced in the submission but in [Mei and Montanari 2019] to study the double-descent curves for the risk. It is not an ad hoc model to obtain CIs. *"I find the "double descent" phenomenon in the CI length to be interesting.(...) Can the author comment on the plausible mechanism of this observation?"* ⇒ A possible mechanism is the conjecture that the length of the CIs is an increasing function of the risk. Proving this appears highly non-trivial. Such phenomena were observed for M-estimators in [Celentano and Montanari 2019, Prop. 4.3(iii)]. *"The result in Section 2.4 (based on Mei and Montanari 2019) seems to be under the assumption of iid weight matrix W"* ⇒ Results in § 2.4 build upon results of Mei and Montanari. Hence, for that section only we require $\boldsymbol{W}$ with iid entries, but, e.g., Theorems 1 and 2 hold for bounded $\|\boldsymbol{W}\|_{op}$ with no iid assumption. *"(minor) A few related works(...)"* ⇒ Thanks–we'll add those. *"(minor) Does the characterization also hold for the ridgeless limit ($\lambda=0$)?"* ⇒ Not currently. A lead to study $\lambda=0$ is to first take $n,p\to+\infty$ (our result), then take $\lambda\to 0$ and study whether interchangeability of limits applies. This appears non-trivial. *"(minor) On Figure 2 Left, why is there a discrepancy between the predicted and simulated boxplot?"* ⇒ The predicted theory is the thick blue line only. The discrepancy is mild and expected for some boxplots when plotting that many boxplots. We'll increase number of runs.
Reviewer 2: *"Neural Networks, when used for classification, are known to suffer from poor calibration. This paper does not touch upon any of these issues prevalent in 2-layer neural networks."* ⇒ We understand the referee's concerns but it does not appear reasonable to solve, in a single 8-pages submission, both CIs for regression, classification and solve the mentioned calibration issues. *"The quantity nL^2 remains elusive (...)"* ⇒ In our results $L\asymp n^{-1/2}$, as for the length of CIs in classical statistics. nL^2 is of constant order after multiplication by n. That's why we study this quantity in figures. *" weighing in (8) seems to be a matter of analytic convenience (...) "* ⇒ This is a contribution of the paper to pinpoint this specific surprising weighting that leads to a pivotal quantity in Theorem 1. This allows the construction of CIs. *" It appears that the training loss appears everywhere. Is this consistent with (...)"* ⇒ Training loss, $\mathrm{trace}[\boldsymbol{H}]$ and population loss are linked through subtle nonlinear relationships studied [Mei and Montanari 2019] and our § 2.4. *"(...) overview into the proof techniques needs to be provided"* ⇒ A proof outline will be added (using the 9th page allowed).
Reviewer 3: *" (...)τ is chosen to be $\tau=(d/p)\lambda$, is this scaling (by d/p) technically important?(...) "* ⇒ No. Since d/p has a finite limit, parametrization of the tuning parameter can be performed through either τ or λ without changing the conclusions. *"Eq (2) and above, notations for transpose should be aligned (...) needs clarification. (...) "* ⇒ Thanks! We'll clarify the writing and fix the typos as suggested. *" (...) The connection between Eq (8) and Theorem 1 needs to be clarified (...) "* ⇒ cf. line 5-7 above. *" Figure 3:(...)or choosing max(x, 0.01x) could help? "* ⇒ Thx! We'll use it.
Reviewer 4:*"The main issue is that this CLT holds under a linear model(...) The coefficients within a linear model are not a natural quantity to be concerned with for the neural network(...) if we believed that a linear model held, we would fit a linear model, not a neural net"* ⇒ Thanks. Developing similar CIs where $\mathbb{E}[y|\boldsymbol{x}]$ is a nonlinear function of $\boldsymbol{x}$ is an interesting future direction. One reason for starting with a linear model is that the targets for the CIs are canonically defined (the unknown coefficients β_j), while for nonlinear models, it is unclear what to consider as canonical population targets (functionals of $\mathbb{E}[y|\boldsymbol{x}]$) for the CIs. Studying linear models has been fruitful to understand double descent curves in numerous recent works (cf. related work section or e.g., [Mei and Montanari]), and this submission is the first work to provide CIs within this line of research. *" (...)it is confusing to see that the results address the linear models and their betas.(...) Perhaps the authors can clear up this confusion, by answering how (8) exactly follows. "* ⇒ cf. line 5-7 above. The relationship between Theorem 1 and the linear coefficients β_js are addressed on line 187-188 around (14).