We thank the reviewers for their positive opinions and constructive feedback. Responses are given below. All the issues discussed (and the typos) will be fixed in the final version of the paper. In addition, since submission, we found that our main results can be straightforwardly extended to provide an upper bound on the timescale/lengthscale of signal propagation in quantized RNNs/CNN. This is done by a combination of our results with previous mean-field results on RNNs/CNNs. We plan to add this practically relevant extension to the main paper as a short paragraph. We respond to specific comments below:

R1: "...the authors refer to [25] to justify that the convergence to a fixed point (Q^*, C^*) of M is slowest in the C^* direction ... it seems that there is only empirical rather than analytic evidence for this..." » Our original intention was to reference the analytic claim in [25], which appears their on the second paragraph of page 5. However, upon closer inspection, we now see that their analysis is inapplicable for quantized activations (as it is based on a Taylor-expansion of the activation function, done in appendix 7.1 there). We will therefore clarify that this claim is currently only empirical. We will also add our own empirical evidence, which supports this claim for quantized activations. It might be possible also to get an analytical argument for quantized activations, but we are not sure yet.

R1: "Throughout the article, the authors consider the possibility only of a C^* that lies in [0,1]. It is not a priori clear to me that there is no stable fixed point in [-1,0]..." » Indeed, this should have been better clarified. We will add a short description to address the region [-1, 0). In general, our equations do hold for this region. When there is bias ($\sigma_b^2 > 0$) there will be no fixed point in [-1,0), because the bias makes the hidden states more positively correlated while the rest of the operations bring the correlation closer to zero. In the case of $\sigma_b^2 = 0$, assuming an anti-symmetric activation function (as was used for activation with constant/linear spacing), the entire network becomes anti-symmetric upon initialization and $C = -1$ becomes an infinitely unstable fixed point as well.

R1: (point 3)"... But when the authors treat the annealed stochastic rounding version of the NTK, and so it would make sense to say what the authors do prove about the corresponding J_{STE}'s" » As the reviewer points out, in the stochastic rounding setting one could study the moments of the spectrum of J_{STE} and obtain dynamical isometry conditions. Indeed, this is an interesting topic for future work. We will clarify that we have not explored it in the present work.

R1: "On line 160, I do not agree with the statement that χ is bounded above by $\frac{2}{\pi}$. It seems like it is bounded below (!) by $\frac{2}{\pi}$ when $\sigma_b = 0$..." » Please see the proof in appendix K.1 which fully covers this issue: it shows that increasing σ_b will, in fact, result in a smaller slope χ at the fixed point. If this was not sufficiently clear from the main text, we can also give the reader some intuition about this: by adding the approximated equation for the fixed point slope using a Taylor expansion $\left(C^* \simeq 1 - \left(\frac{8}{\pi^2}\right)\left(\frac{\sigma_w^2}{\sigma_w^2+\sigma_b^2}\right)^2\right)$, and plugging it into equation (11): $\chi = \frac{2\sigma_w^2}{\pi\left(\sigma_w^2+\sigma_b^2\right)\sqrt{1-(C^*)^2}}$, we obtain an expression decreasing in σ_b^2. We ultimately dropped this argument from the final version, after we derived the proof in appendix K.1. However, we can add this argument back if it is helpful.

R1: "On lines 157-158, I am confused the reasoning for why M cannot have a fixed point..." » Please note the additional condition, regarding $\mathcal{M}(C)$ diverging at $C = 1$, forcing $\mathcal{M}(1-\epsilon) < 1-\epsilon$ for some $\epsilon > 0$. With $\mathcal{M}(0) \geq 0$ and $\mathcal{M}(C)$ being convex, the fixed point slope can not exceed the slope of the linear function between those points: $\frac{\mathcal{M}(1-\epsilon)-\mathcal{M}(0)}{1-\epsilon} < \frac{1-\epsilon}{1-\epsilon} = 1$.

R2: "This point [about the test accuracy dynamics from a viewpoint of Neural tangent kernel (NTK)] has, however, been already discussed in arXiv:1711.00165 titled "Deep neural networks as gaussian processes" and more detailed numerical experiments have been given there" » The covariance map studied in this part of our work is indeed identical to the NNGP kernel in arXiv:1711.00165 which can be used for inference. However the authors of that work do not consider inference with the NTK or the dynamics of the generalization error of neural networks in the linear regime. Their argument in section 3 is indeed similar to ours, but applies to NNGP inference (rather than wide neural networks trained with gradient descent). We thank the reviewer for pointing this out, and we will add a reference and discussion in the revision.

R2: "... For the experiment described in lines 208-217, I cannot find the result" » The relevant figure for this experiment is figure 1, but the reference is missing. Will fix.

R2: "Lines 621-623: I could not understand the manipulations here. Could you explain them in more detail?" » γ is chosen as a smooth extension of $\widehat{\alpha}$ in order to enable one to take derivatives in the lines below. The second order terms drop from symmetry. This calculation is a recapitulation of the one in [13]. We will clarify this further.