We thank all the reviewers for the time reading our paper! We especially thank **R4** for acknowledging that this is "a strong theory paper" and "the techniques developed can be very useful in the analysis of multi-layer ReLU nets."

We also thank **R6** for appreciating the presentation of this paper and shall continue to improve the writing to hopefully make **R5** satisfied as well. We will fix all the minor issues, and below we only address the main concerns.

- **R4+R6** has concerns about the polynomial network size not being applicable in practice

  Indeed, in practice, the step size is usually larger so we expect a different behavior. In this work, we give the first result proving that in the idealized setting, namely "large over-parameterization and small learning rate", neural network as complicated as RNNs can be trained to zero training error.

  We believe this can at least provide intuitions on the second phase of NN learning: where one decays the learning rate and the training error can go to zero. Our m dependency is the worst-case theoretical bound that holds for all possible inputs and labels. In practice, one usually expects a more benign training set and the bounds can be improved by a lot. (This is indeed the case when data is generated from some low-complexity concept class, where m no longer depends on a big polynomial over input size n, see follow up [1].)

- **R5:** Can the authors provide an easy-to-understand paragraph describing why GD/SGD will reach global minimum and how overparameterization helps?

  We will try. What we have provided in the current version is a 6-paged sketched proof (Sections 4+5 on pages 6-8 and Sections 7+8 on pages 13-16) so that the readers don't need to go to our appendix and can already understand our proofs and why over-parameterization helps. We're glad to see that Reviewer 6 has liked this structure. We can try to provide a half-page sketch for this 6 pages sketched. **Do you think that will help?**

- **R6** has pointed out 3 papers and asked us to compare with: [L] "On Lazy training..." arXiv 1812.07956, [G] "On the power ..." arXiv 1904.00687, and [J] Jacot et al. "Neural Tangent Kernel..."

  Note that [L] and [G] both appeared *after* our work (we appeared in October 2018), but we can cite all of them.

  In particular, [J] studies the infinite-width setting of neural network, which is worse than our *polynomial width* result. It's our fault to forget to cite it.

  As for the criticism raised by [L] and [G] about lazy training, here's our response. "Lazy training" has several magnitudes. Perhaps the "laziest" is when NN can be approximated by a linear model (like this paper). The less lazy one is to take into account interactions between layers (see follow up [2]), and the least lazy one is to take into account also sign changes (see follow up 1905.10337). As it goes less lazy, the power of NN becomes stronger. Perhaps surprisingly, in this laziest model, we can already prove that (R)NN memorizes data. (This is non-trivial! Since (1) why is RNN close to linear model in small learning rate regime? and (2) even so, why can linear model train to zero error?)

  Finally, we believe studying this laziest model is very meaningful. Not only it can give us intuitions about the second phase of NN learning (where one decays the learning rate), it also gives us technical tools for studying less lazy models (such as follow ups [2] and 1905.10337, both relying on this paper). Most significantly, the perturbation theorems proved in this paper is at the heart of all of those follow-ups.

- **R6:** it seems strange A, B can be set random, W moves little, and the network can produce zero training loss.

  There's no contradiction here. Recall when A, B, W are all at random initialization, (we have proved) the RNN outputs are of magnitude roughly constant. We claim by moving from W to $W + W'$, even when $\|W'\|_2 \ll \|W\|_2$, we can already change the output significantly (i.e., by more than a constant). There is no contradiction here:

  - W is random, so when interacting with A and B, there is a lot of cancellation and the output is small;
  - W' correlates with A, B, so when interacting with them, gives a much bigger output.

  Hence, W' need not be as big as W. As a toy example, when $a, w \in \mathbb{R}^m$ are random vectors with coordinates i.i.d. from $\mathcal{N}(0,1)$, then $|\langle a, w\rangle| \approx \sqrt{m}$ but $\|a\|_2 \approx \|w\|_2 \approx \sqrt{m}$. If we update $w + w' = w + \frac{1000}{\sqrt{m}} a$, then $\langle a, w + w'\rangle \approx 1000\sqrt{m}$ which is very different from $\langle a, w\rangle$ but $\|w'\| \ll \|w\|$.

  (There is practical evidence of this as well. One example is for 2-layer network on MNIST, see Fig. 5 on page 26 of prior work Li-Liang [30] of their NeurIPS camera ready version. There are also evidence in some follow-up works of the anonymous authors. We can consider adding some explanations of this in a next revision.)