Response to Reviewer #1

- *“how come leverage score initialization gives no improvement on the bounds of Theorem 3.6 and 3.7..."*

RE: Thank you for the comment. The reason is that both bounds in Theorem 3.6, 3.7 come from two parts: 1) initialization phase and 2) training phase. Part 1) requires the width to be large enough, so that the initialized dynamic kernels $H(0)$ and $\hat{H}(0)$ are close enough to the neural tangent kernel (NTK) H^{cts}. Part 2) requires the width to be large enough, so that the dynamic kernels $H(t)$ and $\overline{H}(t)$ (defined in Definition D.3 in supplementary material) are close enough to the NTK H^{cts} during the training process. Leverage score initialization optimizes the initialization bound in part 1) while keeping the bound for part 2) the same. The current state-of-art analysis gives a tighter bound in part 2), so the final bound for width is the same for both cases. If analysis for part 2) can be improved and part 1) dominates, then initializing using leverage score will be beneficial in terms of the width needed. We will address these discussion in the final version.

- *“presentation need to improve..more discussion..."* **RE:** Thank you for the advice. We will elaborate our results and address the above discussion in detail in the final version.

- *“undefined notations and typos"* **RE:** Thank you for pointing out. We will address these issues in the final version.

Response to Reviewer #2

- *“incremental upon the work of Avron et. al...limited discussion on differences between this paper..."*

RE: Thanks for your comment. We respectfully disagree. We emphasize the major contributions of our work includes 1) *proving the convergence result for training neural network with polynomial width*, and 2) rigorously *building the connection between the leverage score sampling theory and the theoretical understanding of training deep neural networks*. Both our theory of understanding training regularized neural networks and training neural networks with leverage score initialization do not exist in the literature and is not mentioned in the work of Avron et. al.. We will add more discussion about this in the final version.

- *“no experiments are provided in this paper..."*

RE: Thank you for the comment. This work mainly focuses on the theoretical understanding of the connection between leverage score sampling and training regularized neural networks.

Response to Reviewer #4

- *“equivalence between NN initialization and random feature...is known since the mid 1990s[1][2]..."*

RE: We respectfully disagree. We emphasize that the work in [1][2] only argues the equivalence when the hidden units in the neural network go to infinity, while our results give out specific polynomial bounds for the network width for such equivalence to hold, which is much harder. We will address this important reference in the final version.

- *“...relies on the fact that the dynamic kernel during training stays close to the kernel at initialisation which in turn is a random feature approximation of neural tangent kernel. This assumption is very strong and unrealistic..."*

RE: Thank you for the comment. We point out the fact that *the dynamic kernel during training stays close to the kernel at initialization* is not an assumption but a rigorously proved statement in our work. It is derived from the over-parametrization property of the neural networks, which enables the variable to converge to a point near its initialization when the network is sufficiently wide. Similar idea has shown up in previous literature [3][4] for the unregularized case.

- *“do not find any interesting novel results..."*

RE: Thanks for your comment. In this work, our contribution is three-folded. Apart from the generalization of leverage score sampling theory you have mentioned, we rigorously characterize the equivalence between training regularized NN and KRR by giving out a polynomial bound for the network width and number of training steps. Based upon that, we novelly apply the idea of leverage score sampling to initialize neural networks, which gives better network width bound on approximating neural tangent kernel in the initialization phase.

Response to Reviewer #5

- *“not convinced...derive reasonable rates of convergence"*

RE: This is a very good comment. The key point we claim is that the bounds we obtained in Theorem 3.6,3.7 and 3.9 on the network widths are all polynomially dependent on the size of the training dataset and the minimum eigenvalue of the neural tangent kernel. And the polynomial bounds we obtained match the state-of-art results [3][4] etc., if we set the regularization parameters to 0. We believe the constraint on the regularization parameter can be indeed relaxed by more advanced analysis, which is left as a future work.

- *“some minor suggestions..."*

RE: Thank you for the suggestions. We will address these notation issues and typos in the final version.

Reference

[1] Priors for infinite networks, Radford M. Neal

[2] https://blog.smola.org/post/10572672684/the-neal-kernel-and-random-kitchen-sinks

[3]A convergence theory for deep learning via over-parameterization, Zeyuan Allen-Zhu et. al.

[4] Gradient descent provably optimizes over-parameterized neural networks, Simon S Du et. al.

[5] On exact computation with an infinitely wide neural net, Sanjeev Arora et. al.