Response to Reviewer #1

The goal of this paper is to determine whether neural networks (NN) are equivalent to kernel methods or not. While at first sight one would guess that neural networks are more powerful than 'simple' kernel methods, the recently developed neural tangent (NT) kernel theory (with a dozen papers published by several groups) argue that SGD-trained networks are well approximated by the NT kernel, for very large networks.

In this paper, we argue in the opposite direction, namely NN is superior to NT. Our argument consists in rigorously proving that, in a concrete example, a significant gap exists between NT and NN. In view of this, *a simpler example results in a stronger conclusion.* This being said, a series of extensions are available:

1. RF model. The analysis of the random features model can be generalized to arbitrary target functions f_*, and activation function σ (under mild technical conditions). Also, the analysis can be generalized to the case of $\boldsymbol{x}$ and $\boldsymbol{w}$ distributed uniformly over the d-dimensional sphere.

All of these generalizations are accessible to the same proof techniques developed in our paper. We focused on the quadratic case uniquely to make the comparison more transparent.

2. NT model. This analysis can be generalized to arbitrary f_*, if $\boldsymbol{x}$ and $\boldsymbol{w}$ are distributed uniformly on the sphere.

3. NN model. Generalizing the analysis of the NN model would require proving global convergence for gradient descent beyond quadratic activations. This has been an open problem for a long time, and it probably requires additional assumptions. Before our paper, this problem was open for quadratic activations as well.

4. We have extensive experimental results comparing NN and NT for RELU and Tanh activations (see figure). Our experiments indicate that these nonlinearities behave in general as predicted by the theory developed for the square nonlinearity.

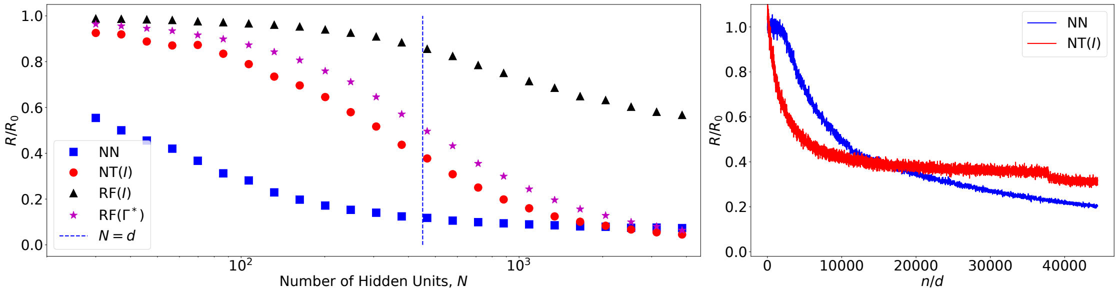

Figure 1: Left: Test error of NN, NT and RF in fitting a quadratic function in d = 450 dimensions. Here the experiments are performed using **ReLU** nonlinearity. Similar to the case of square nonlinearity studied in the paper, there is a significant gap between NT and NN. Right: Evolution of the risk for NT and NN with the number of samples.

Response to Reviewer #2

We agree that a direct comparison of RF and NT is not meaningful. We find the difference in behavior quite interesting, but not indicating the superiority of one method over the other.

Response to Reviewer #3

1. We believe that the case of Gaussians with unequal means can also be analyzed within the RF and NT model (under certain technical assumptions) although at the cost of non-negligible technical complications.

2. We believe that the analysis of RF and NT can be generalized to certain other models beyond Gaussian covariates $\boldsymbol{x}$. For instance, the case of $\boldsymbol{x}$ and $\boldsymbol{w}$ uniform over the d-dimensional sphere can be treated using similar techniques. Also certain derivations can be extended to $\boldsymbol{x} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}$ with $\boldsymbol{z}$ having i.i.d. (non-Gaussian) components.

3. Our proof uses the gradient flow dynamics for studying the convergence of the NN. In the small learning rate regime, both SGD and mini-batch SGD dynamics converge to the gradient flow dynamics. We use the mini-batch SGD setting because (1) the convergence happens with larger learning rates and (2) experiments with minibatch SGD can be done in a computationally efficient, GPU-friendly manner.

4. We have performed an extensive set of experiments with different types of $\boldsymbol{\Delta}$, $\boldsymbol{\Sigma}$ and activation functions. Due to space constraints, we did not include these in the initial submission. We will definitely add these to the camera-ready version / appendix.