We thank the reviewers for their encouraging reviews and detailed comments, and are pleased that they all recommend the paper be accepted.

R2 *Theoretical contribution and proof technique.* We would like to emphasize our contributions: (1) Deriving the eigenvalue decay rate of NTK with one hidden layer and bias and a lower bound for deeper networks for data distributed uniformly on the hypersphere (Thm. 1). (2) Deriving eigenvalue decay under the same conditions for the Laplace kernel (Thm. 2). (3) Theoretical comparison of the class of functions (RKHS) that correspond to NTK (deep and shallow), Laplace and Gaussian kernels (Thm. 3). (4) Results on the eigenfunctions of NTK outside the sphere (Thm. 5). (5) Empirical comparison of NTK to exponential kernels on various standard datasets. (6) Empirical comparison of CNTK and convolutional versions of exponential kernels.

Regarding our proof techniques, the proof in Thm. 1 for NTK with two layers and bias borrows techniques from [6]. Our proof technique for deep networks uses the algebra of RKHSs and is therefore novel in this context. Our proof of Thm. 2 derives bounds that result from the relation between the Fourier expansion of the Laplace kernel in $\mathbb{R}^d$ to its harmonic expansion in $\mathbb{S}^{d-1}$. Finally, Thm. 5 derives the eigenfunctions of NTK in $\mathbb{R}^d$ by using invariant properties of NTK (established in Thm. 4) and identifying the spaces fixed under the appropriate integral transform.

R2 *"why they need additional parameters a, b, c."* The hyperparameter c is used commonly to determine the sharpness of the Laplace kernel. We note that analogously NTK becomes sharper for deeper networks. Controlling c is therefore analogous to depth selection. The remaining parameters apply an affine modulation to the Laplace kernel. In regression scaling a kernel has no effect, while an affine shift only affects the DC component. Therefore a and b have only little effect on the results of regression. They are however important in repeated application of the kernel, such as in the convolutional C-Exp algorithm. The bias term β in the C-Exp experiments is chosen through cross validation in $\{1, ..., 10\}$.

R3 *"Some treatment of non-uniform data."* Analyzing the similarity between the kernels under nonuniform distributions is a worthwhile future objective. Figure 1 below indicates that both the eigenvalues and eigenfunctions of NTK and the Laplace kernel closely match also for nonuniform distributions. This is shown for a piecewise uniform density with three bins of ratios 5:12:1 in $\mathbb{S}^1$ (left and middle panels). The right panel shows the eigenvalues obtained for the UCI Abalone dataset.

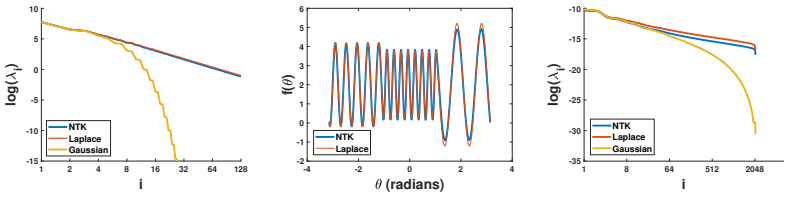

Figure 1: An overlay of the eigenvalues (left, log-log plot) and example eigenfunctions (middle) of NTK and the Laplace kernels for data in $\mathbb{S}^1$. Right: eigenvalues on the UCI Abalone dataset.

R1 and **R3** *"Which activation functions are being treated."* Our paper indeed focuses on the ReLU function only – we apologize for the unintended omission. With different activation functions the RKHS of NTK would differ, depending on the degree of smoothness of the function. For example, the eigenvalues for NTK with tanh activation appears to decay exponentially fast.

R1 *"the obtained accuracies on Cifar10 seem quite low."* Our experiments are run under the same conditions as in [2]. Incorporating an average pooling layer is likely to lead to better accuracies, as is demonstrated for CNTK in [28].

R1 and **R4** *"The dependence on the number of examples is also unclear."* This generalization result depends on the mesh norm, which captures the minimal distance between training points. Alternative bounds, such as in [10] page 9, provide such dependence.

R1: *Eigenvalues in* $\mathbb{R}^d$. Compared to $\mathbb{S}^{d-1}$, the eigenvalues are scaled uniformly, depending on the radial distribution.

R4: *"Convexity versus non-convexity is overlooked here."* Solving kernel ridge regression using the Representer theorem involves least squares minimization, which is convex. Since the kernel matrix is positive definite we can use a faster, primal-dual formulation [16], where both are convex.

We will modify the paper to improve the discussion of previous work, change Eq. (2) to include ridge regression, and address the rest of the reviewers' comments.