We thank all reviewers for their positive reviews and valuable comments. We begin by addressing questions that are of general concern among multiple reviewers, and later respond to questions individually to each reviewer.

Role of the data distribution. Reviewers 1 and 2 raise the interesting question of the data distribution and its effect on the spectral properties of the kernel. However, note that our study of the Mercer decomposition is aimed at providing a clear description of the functions in the RKHS, their norm, as well as their approximation properties, *irrespectively* of the specific data distribution. In particular, one can obtain a Mercer decomposition for any measure, and the approximation properties given in Corollary 6 and 7 rely specifically on the spherical harmonic decomposition with a uniform distribution on the sphere.

While our analysis does not provide a precise study of estimation error for specific data distributions, it does yield insight for the (somewhat crude) uniform bounds based on Rademacher complexity, which only depend on the data through the norm of the learned function $\hat{f}$ (and the radius R), taking the form $O(\|\hat{f}\|R/\sqrt{n})$. In the context of NTK, the papers [1, 2] derive such bounds where $\hat{f}$ is the minimum-norm interpolating solution. For more refined bounds based on eigenvalues of the integral operator, one would then require a spectral decomposition of the kernel w.r.t. the data distribution, which is more difficult to obtain (though the eigenvalue decay may be preserved, e.g., if the data distribution is absolutely continuous w.r.t. the uniform distribution on the sphere). We will be happy to clarify this further in the paper.

Role of depth. We agree that a limitation of the paper is that the approximation results of Section 3.2 are limited to two-layer fully-connected networks. The extension to a two-layer CNN with global average pooling is straightforward, with an eigenvalue decay similar to the fully-connected case but which only depends on the dimension of a patch rather than the full signal. The study of approximation for deeper networks is more complicated and is left for future work. We will state this more explicitly in the paper. The smoothness and stability results do apply to deep CNNs, and in particular depth is important for deformation stability, since the bound in Proposition 12 improves with smaller patches (i.e., small β): indeed, with appropriate pooling and downsampling, a deeper architecture is needed in order to reach a fixed target level of translation invariance with small patches at each layer [8]. We will clarify this further in the paper.

Empirical validation. We conducted numerical experiments in order to assess stability and approximation properties, comparing the NTK to the simpler kernel with all layers fixed but the last, for a three-layer convolutional architecture on MNIST digits. Considering deformations from the "infinite MNIST" dataset, we indeed observe that the stability of the NTK kernel mapping is weaker, with a faster growth as a function of the deformation size for small deformations. Regarding approximation, we computed interpolating solutions $\hat{f}$ on binary classification problems with a dozen digits in each class for the two kernels, and found that the quantity $\|\hat{f}\|_{\mathcal{H}}R$ (where R is the average of norms $\|\Phi(x_i)\|_{\mathcal{H}}$, for normalization purposes) is always smaller for the NTK, suggesting it indeed has better approximation properties. We will be happy to include these results in the paper, if it is accepted.

R1. We thank the reviewer for his positive comments. The questions related to data distribution and depth are addressed above.

- "... is it surprising ...": given an appropriate CNN architecture, stability is indeed not surprising, but we find it interesting to characterize stability for the NTK, contrast it with approximation results, and compare it with known deep convolutional kernels.
- "... lower bound on stability ...": this is an interesting point, which is partly discussed in earlier papers on deformation stability (e.g. Section 3.2 in [8]), though without a precise lower bound. One way to see this instability is by constructing a function f in the RKHS based on a large filter with very high frequencies, but with norm less than one. This yields a lower bound on $\|\Phi(x) - \Phi(x')\| \geq f(x) - f(x')$ which can be made arbitrarily unstable due to high frequencies, even when x' is a small deformation of x.

R2. We thank the reviewer for pointing out the Matthews reference, which we will include in the paper. The concerns on the data distribution, depth and empirics are addressed above.

R3. We thank the reviewer for his remarks. We will state the requirement on sequential limit more explicitly in the paper. While our stability bounds partially illustrate the benefits of depth and convolution, we agree that extending the approximation results to deep CNNs would be interesting (see above).

References

[1] S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *ICML*, 2019.

[2] Y. Cao and Q. Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *preprint arXiv:1905.13210*, 2019.