We thank the reviewers for extremely helpful suggestions. We will incorporate all of them in the final version of the paper. Below we discuss the most significant points.

Reviewer 1:

Eigenvalues for sine functions. The kernel is an even function, and so its decomposition includes only cosine functions. However, when it is applied as convolution, phase shifts are further included, and therefore the eigenvectors of H^{∞} include the sine functions with eigenvalues equal to those of the cosine functions. We will clarify this in the final version.

Mismatch at the low frequencies. For the very low frequencies the actual run time is affected strongly by the initialization, yielding slightly slower convergence times than predicted.

Stopping criterion and fitting. The target accuracy depends on both δ and ϵ. We stopped each experiment when the accuracy of the network reached within 5% of the desired output. (We tested other values as well and obtained similar results.) In each graph the predictions (in orange) were scaled by a single multiplicative constant to fit the measurements. This constant reflects the length of each gradient step (e.g., due to the learning rate and size of training set).

Clarity. We thank the reviewer for these suggestions. We will add example runs in the appendix and include a more intuitive explanation of the necessity of bias.

Reviewer 2:

Empirical evidence is not enough to justify its real practical value. We have added new experiments with deeper networks (5 FC layers), different architectures (Resnet-10) and different loss functions (cross-entropy). All results are consistent with our theory (see Figure 1).We will provide code for all experiments. We will include further results with different hyperparameters in supplementary material; these all show similar results.

Discussion of over-parametrization. We thank the reviewers for this point; we agree that it merits more discussion in our paper. We will explain that our results rely on lazy training induced by overparameterization. Chizat et al. provide experiments that question the ability of such networks to generalize well, which seems in contrast to theoretical results of Arora et al. and others. We will discuss this interesting question.

Intrinsic dimension. We agree that this is an exciting direction. As a first step, when the data lies in a lower dimensional linear space we can show analytically that the predicted times depend on the intrinsic dimension. This is supported in practice by experiments shown in Figure 1 (d) and (e).

Reviewer 3:

Insightful but not original. Our main contribution is to derive concrete predictions in all dimensions for the convergence time of gradient descent, as a function of frequency. We do this by providing explicit expressions for the eigenvalues of the gram matrix. This also allows us to identify instabilities in the bias-free model. Our predictions are validated empirically on both shallow and deep networks. In this rebuttal we provide more general experiments for different architectures and loss functions. Rev. 2 states: "As a major plus, the proposed rates seem to match well with the empirical observations". All of these contributions go beyond [Xie 2017 and Arora 2019], which are cited and discussed in the paper.

We also feel that our discussion of the role of bias goes beyond a "minor technical fix". It is clear from prior work that H^{∞} might contain eigenvectors with small eigenvalues, corresponding to functions that are hard to learn. However, that work did not show that these functions are intuitively complex; this was not possible because this is not true for networks without bias. By introducing bias, we are able to show for the first time that hard-to-learn functions in fact do correspond to high frequency functions.

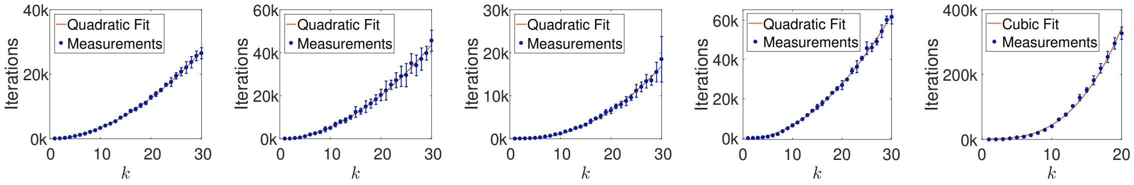

Figure 1: Number of iterations to convergence as a function of target frequency. From left to right: (a) input in S^1, MSE loss, 5 layers fully connected; (b) same, 10 layer Resnet; (c) Resnet-10, cross entropy loss; (d) input in S^1 embedded with a random rotation in $\mathbb{R}^{30}$, MSE loss, Resnet-10. (e) input in S^2 otherwise same as (d). For (c) the task is binary classification; so that for every k, class is 1 if $\cos(k\theta) > 2/3$, -1 if $\cos(k\theta) < -2/3$, training points with other θ values are not used. Consistent with our theory's predictions for two layer networks, we see quadratic growth in cases (a)-(d) and cubic growth for (e). To estimate the leading exponent in these graphs we fit a line to the corresponding log-log plots, obtaining, from left to right, $O(k^{1.93}), O(k^{1.95}), O(k^{2.37}), O(k^{2.08}), O(k^{3.09})$.