**Response to reviewer 1.** *Weakness 1: The result for neural network is an upper bound which may not be tight.* Indeed: Theorem 4 is just an upper bound on the approximation error, and a lower bound for NN is currently missing. The upper bound is sufficient to prove a separation result for $\kappa$ large enough. While the upper bound is not tight pointwise (per function $f_*$), we believe it is almost tight in a *minimax* sense. Namely, if $N \ll d_0^\ell$, then there is a degree-$(\ell+1)$ polynomial $f_*$ such that the corresponding approximation error $R_{\mathsf{NN},N}(f_*)$ is bounded below by a positive constant. This seems quite intuitive by parameter counting. We agree that this is an important question to be addressed. *Weakness 2: Our theory requires a smooth activation $\sigma$ while our experiments uses ReLU activation.* We believe that the smoothness $\sigma$ is a technical issue and can be relaxed using soft approximation arguments. We did not relax this requirement since the paper is already technically involved and this point is tangential to the main questions we address. Experimentally, it is also easy to check that replacing ReLU by a smoothed ReLU does not change the numerical results.

**Response to reviewer 2.** We will improve the readability, clarify the notations, and implications in the final version of the paper. *Additional comment 1: How does the result of Theorem 1 depend on $\lambda$?* Theorem 1 implies that —to the leading order– the test error is independent of $\lambda$ as long as $\lambda = \lambda(d) = O_d(1)$. In particular, $\lambda(d) = 0$ is optimal up to subleading terms. On the contrary, our proofs also imply that $\lambda(d) > d^\iota$ leads to suboptimal test error for any $\iota > 0$. $\lambda(d) = 0$ still leads to good generalization because, in high dimension, the non-vanishing high order derivatives of the kernel function induce regularization effects (hence the assumption that $h^{(k)}(0) > 0$ for some $k \geq \ell$). *Additional comment 2: What is the 'high dimensional regime'?* We considered two regimes: 1) $n = \infty$, $N$ and $d$ are polynomially related and goes to $\infty$ together. 2) $N = \infty$, $n$ and $d$ are polynomially related and go to $\infty$ together. *Additional comment 3: Figure 3 shows that at finite samples the test error corresponding to the neural network appears to be more sensitive to $\kappa$ which is puzzling.* We agree that the test error of the SGD trained NN deserves further investigation. However notice that the test error of NTK is substantially larger, for each given $\kappa$. Since the error is normalized to be between 0 and 1, it cannot grow much, whence the apparent lack of variability.

**Response to reviewer 3.** The reviewer's summary of our work is incomplete. We considered the following model: the feature vector $\boldsymbol{x} = \boldsymbol{U}(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^d$ is composed of an isotropic signal feature $\boldsymbol{x}_1$ in $d_0$ dimensions and an isotropic junk feature $\boldsymbol{x}_2$ in $d - d_0$ dimensions; the target function only depends on $\boldsymbol{x}_1$. We characterize the risk of KRR, RF, NT in terms of the effective dimension $d_{\text{eff}}$ which depends on $d$, $d_0$, and feature SNR (the variance of the signal feature over the variance of the junk feature). Crucially, the effective dimension $d_{\text{eff}}$ does not depend uniquely on '$d_1$' (perhaps $d_0$ is meant here), the dimension of subspace of the feature vector 'concentrated in', as written by the reviewer.

Further, we do not agree with the reviewer that our results are incremental. We agree: it is natural to expect that the behaviors of RF, NT, KRR, only depend on some sort of 'intrinsic dimension'. The result would be trivial if the variance in the junk subspace was zero, and feature vectors lied exactly on a $d_0$-dimensional subspace. Our contribution is non-trivial in extending the result to the case when there is non-vanishing variance of junk features: we quantified how the effective dimension $d_{\text{eff}}$ depends on the signal dimension $d_0$ and feature SNR. (Often $d_{\text{eff}} \neq d_0$.) Let us also emphasize that a model in which the data lie exactly on a $d_0$-dimensional subspace would be extremely crude: in particular it would not explain the experiments in which we gradually add noise to the junk subspace.

*Weakness: Why do not just study a single neuron $g(\boldsymbol{w}^\mathsf{T}\boldsymbol{x})$?* First, limiting the analysis to the target function $g(\boldsymbol{w}^\mathsf{T}\boldsymbol{x})$ would not simplify the proofs: we would still need to develop the whole machinery and then apply them to this specific case. The real technical challenge arises in dealing with non-vanishing variance in the junk subspace. Second, the target function $g(\boldsymbol{w}^\mathsf{T}\boldsymbol{x})$ is too restrictive: we cannot consider this a remotely realistic model. On the other hand, functions of a low-dimensional subspace are a classical model in non-parametric theory.

*Clarity: why choosing $\lambda = O_d(1)$ instead of $\lambda \to 0$?* Notice that $\lambda \to 0$ is a special case of $\lambda = O_d(1)$. (See comments to 2nd reviewer.) Finally, we agree that there is some room to improve the clarity and readability of this paper. We will make these changes in the final version.

**Response to reviewer 4.** *Weakness 1: What do we expect for multi-layers fully-connected networks?* For KRR with multi-layers NT kernels with Gaussian weights (infinitely wide networks), the resulting kernel will still be an inner product kernel. Hence our Theorem 1 can be applied directly. We have also conducted experiments comparing 3-layers NN and 3-layers NT KRR in Figure 4. On the other hand, analyzing finite width NT KRR for 3-layers networks is beyond current techniques. *Weakness 2: Does the model $\varphi(\boldsymbol{U}^\mathsf{T}\boldsymbol{x})$ tend to trivially favor NN models?* As the reviewer writes, this is a natural way to model 'low-dimensionality' of the target function. In fact it is a standard model in the non-parametric regression literature and is directly related to other canonical functional classes. [Bac17] provides several pointer to the vast literature on this topic, and connections with other functional classes. *Weakness 4: How would the performance of RKHS methods be in these experiments?* We compared NN and KRR (RKHS methods) in Figure 1, 3, and 4 (in these figures, we use the shorthand NT KRR for the RKHS induced by the infinite-width NT kernel). We thank the reviewer for the additional remarks. We will make these changes in the final version.

[Bac17] Breaking the Curse of Dimensionality with Convex Neural Networks. F. Bach.