We thank all reviewers for their helpful and detailed comments.

Review 1

Regarding the improvement suggestions:
- We will present the main results and limitations more explicit in the introduction section of the paper.

Review 2

Regarding the remarks under the "quality" bullet:
(1) We will add a discussion on the requirement that the u_i coefficients are exponentially large. In a nutshell, existing analyses of stochastic gradient descent, even for convex functions, imply that the required number of iterations scales polynomially with the norm of the target solution, which would mean exponentially many iterations in our case. Moreover, practically speaking, such huge coefficients can cause overflow when running SGD on a computer with standard floating point formats.
(2) c_1 is a small numerical constant that does not depend on any parameter of the theorem (it comes from Lemma 17 in [1], quantifies concentration of measure on a sphere, and can be explicitly upper bounded by 40). We will try to make this clearer.
(3) This is a proof by contradiction, and this assumption is what we want to show to be invalid. We will write this in a clearer way.

Regarding the remarks under the "clarity" bullet:
(1) We agree that “explicitly or implicitly” is not sufficiently clear, and we will rephrase. What we meant is simply that all the papers discussed in section 2 use the random features idea in various ways.
(2) The goal of section 3 is to give a simple self-contained proof on how neural networks can be explained using neural networks, and to give motivation to the forthcoming section. As we explicitly point out, the proof methods are not that novel, which is why this section is only about half a page (although we do improve on previous results regarding approximations of polynomials).
(3) We will rephrase this notation to make it clearer. It should be written that (3a) $\psi : \mathbb{R} \to \mathbb{R}$ is a real periodic function.
(3b,c) It is the norm defined at the beginning of the section (lines 199-200).
(4) We will try to add a conclusions section in the final version (appropriate to the page limit).

Review 3

Regarding the improvement suggestions:
- Regarding the "uniformly spherically distributed" assumption on W: The theorem can be readily extended so that W has a standard Gaussian distribution, though it would require more complicated calculations as we would need to bound the norm of the function f_W w.h.p, instead of an absolute bound which we used in the theorem. We prefer to keep the theorem that way to make the proofs easier to understand. However, we can add a comment if the reviewer feels it is needed.
- In the relevant theorems, we will make it clearer when $\mathbf{x}$ is assumed to have a Gaussian distribution.
- Regarding extension to the "linearized" neural tangent kernel model: In fact, Theorem 4.6 applies to this model (it does not make any specific assumptions on the feature class $\mathcal{F}$). We will add an explicit comment on that.
- We will fix the boldface notation where relevant.

References

[1] Shamir, Ohad. "Distribution-specific hardness of learning neural networks." The Journal of Machine Learning Research 19.1 (2018): 1135-1163.