We thank the reviewers for their helpful comments and we will fix accordingly. Below we address some comments and questions that have been raised.

Reviewers 1 and 3

The reviewers wrote that requiring only polynomial versus not polynomial in the separation definition is a bit limiting. The benefits of depth can be studies from different angles. We focus on the notion of depth-separation that has been extensively studied in many prior works, and requires polynomial vs. exponential width. We agree that other notions of depth-separation should also be explored, and we mention in the "related work" section an example for such a different notion that has been studied in [23,16,31].

Reviewer 2

The reviewer asked: "bounded functions can oscillate exponentially and has an $\exp(d)$ Lipschitz norm. However, the first theorem suggests they can be approximated with only $\mathrm{poly}(d)$ weights. This looks very counterintuitive. How should we understand it?". This observation is indeed surprising, however, it has a simple intuitive explanation. Consider a univariate function $f : \mathbb{R} \to \mathbb{R}$ that is expressible by a neural network N of constant depth and $\mathrm{poly}(d)$ width (for some integer d). Such a function is piecewise-linear with a $\mathrm{poly}(d)$ number of pieces. Since the weights in N might be exponential in d, then some of the linear pieces might have exponential derivatives. Thus, f might oscillate quickly in some intervals. However, since f is bounded, then an interval where f has an exponential derivative must be very small (exponentially small). Hence, f consists of $\mathrm{poly}(d)$ linear pieces, but may oscillate quickly only in exponentially-small intervals. A network of constant depth and $\mathrm{poly}(d)$ weights cannot approximate such f in the L_∞ sense. However, since we assume that the input distribution μ is not too concentrated in very small intervals, then it is possible to approximate f in the $L_2(\mu)$ sense. The case where $f : \mathbb{R}^d \to \mathbb{R}$ is more complicated, but follows the same intuition.

Reviewer 4

Thanks for your feedback and comments, we will incorporate them into the final version. Below we address your specific questions:

(2) Not for $k > 2$.

(5) Since we focus on constant-depth networks, then it does not matter. We will comment on that in the final version.

(7) There is no special reason for this choice of notations. We will change.

(9) The results hold also for approximation w.r.t. $L_1(\mu)$ (it follows easily from our proof), but do not hold for approximation w.r.t. L_∞.