I assume the reviewers are familiar with the notations in the paper.

Summary of How to Choose an Activation Function The paper mainly focuses on the universal approximation of the classes W_r^s and W_r^{s*}. There are two main contributions of this paper: 1. The estimation of number of neurons required for a neural network with a single hidden layer to ensure a given accuracy of approximation to every function in a given function class, under the assumption that the functions are continuously differentiable to some order. 2. The way to construct networks with the indicated number of neurons evaluating standard activation functions.
The Section 1 gives a brief overview of the previous work on the approximation capabilities of feedforward neural networks, and then it introduces the focus of the paper, which is mentioned above in the discussion about the contributions of the paper.
In section 2, the paper introduces some necessary notations and reviews some certain known facts. It also introduces the assumptions on the targeting function class. The remark here is that the periodic activation function $sin(x)$ provides optimal convergence rates for the class W_r^{s*}. Another highlight is the problem of approximating an r times continuously differentiable function can be reduced to that of approximating another function from the corresponding periodic class, and the order of approximation is the same for both. The paper also mentions that an optimal network can be constructed using a higher order sigmoidal function.
In section 3, the paper first investigates the degree of approximation by neural networks in terms of a general periodic activation function. The main theorem, Theorem 3.1 states that it is possible to construct a neural network by using a periodic activation function ϕ to approximate a targeting periodic function $f \in C^{s*}$, with an error bound. The error bound relates the degree of approximation of f by neural networks explicitly in terms of the degree of approximation of f and ϕ by trigonometric polynomials, and indicates that the smoother the function ϕ, the better will be the degree of approximation. It also mentions how to construct an approximator given the Fourier coefficients of the target function in another paper.
In section 4, the paper introduces a process of how to construct a periodic function for which Theorem 3.1 can be applied, given the activation function σ is not periodic and satisfies certain decay conditions. Under an assumption, a 2π-periodic function integrable on $[-\pi, \pi]^s$ is derived by a transformation of a function in the linear span of $A_{\sigma,J}$. The main result is Table 1, a table showing the order of approximation for different activation functions.
In section 5, the paper derives dimension independent bounds of L^2 degree of approximation in a specific function class SF_s for both periodic and squashing activation functions.

Detailed comments
Comments: We should be aware of the limitation of the paper; it does not cover other popular activation functions such as Rectified Linear Units and Hyperbolic Tangent Class. Since the line of research is focusing on squashing function such as Gaussian and sigmoidal functions, it is acceptable. Mhaskar and Poggio (2016) further discussed new results for ReLU, a non-smooth activation function.
In section 4, the paper "suppose that there exists a function ψ", which should be discussed in detail.
It's unclear to me how to go from Theorem 3.1 and Section 4 to Table 1; Theorem 3 assumes a target f is periodic ($f \in W^{s*}_r$), while the functions in the table does not have this assumption ($f \in W_r^s$). The paper should elaborate the derivation.

Advantages:
One significant result of the paper is that it theoretically proves the result that the smoother the function ϕ, the better will be the degree of approximation, given the assumptions of the paper. This result is of great importance both for theoretical analysis and applications.
Using table to show the order of approximation for different activation functions is good. The results in the table is also of great importance.

Weakness:
Although the paper is pure theoretical, it would be better if it can provide any numerical experiments. Considering the paper was published in 1994, where the computational power is far less than today, it is acceptable.
The paper is not self-contained for two reasons: 1. Some notations used in the paper are not properly defined. For example, L^2 error of Fourier transform was mentioned in section 1, but it is not defined in the paper. 2. There is no proof in the paper, but it seems okay because the proofs might be too complicate.
It would be great if the paper can provide a detailed example of how to construct an approximation with an activation function.
The paper should emphasis more about the fact, mentioned in Section 2, that a function can achieve the same order of approximation for a general r times continuously differentiable function and the corresponding periodic function. I suppose this is the reason why we can derive from Theorem 3.1 to Table 1.

1 Citations

Mhaskar, H. N., Poggio, T. (2016). Deep vs. shallow networks: An approximation theory perspective. Analysis and Applications, 14(06), 829-848.