Suggestions - Problem set 4

1. In this case

$$\widehat{f}(\boldsymbol{x}) = \sum_j b_j \cos\,(V^{(j)}\boldsymbol{x} - \theta_j);$$

Notice that if $\{b_j\}$ are square summable (i.e. $\sum_{j=1}^{\infty} b_j^2 < \infty$), and $V^{(j)} = j$, and $\theta_j = 0$, we get a *cosine series*, which can clearly approximate any L^2 function on $[0, \pi]$.

You may use the following version of a basic theorem of Fourier series:

Theorem: Given any function $f(x) \in L^2(0, \pi)$, the cosine series of f given by

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} a_n \cos\, nx$$

converges to $f(x)$ in L^2, if $a_n = \frac{2}{\pi}\int_0^{\pi} f(x) \cos\, nx\, dx$. That is, given the partial sum

$$f_N(x) = \frac{a_0}{2} + \sum_{k=1}^{N} a_n \cos\, nx,$$

we have

$$||f_N(x) - f(x)||_2 \underset{N \to \infty}{\longrightarrow} 0.$$

2. What relationship can you prove regarding the two norms $||f||_\infty$ and $||f||_p$? Note that we are assuming that f is *compactly supported*, i.e. that its support K (i.e. where it is non-zero) is a compact (i.e. closed and bounded) set in $\mathbb{R}^k$.

3. Notice that in such a network, for given *input* activations $x_1, \dots, x_k$, there is a desired *output* set of activations $q_1, \dots, q_m$. Note for example that the given desired output q_1 for the first neuron in the last layer will depend on the inputs $(x_1, \dots, x_k) \equiv \boldsymbol{x}$ through a given function $f_1(\boldsymbol{x})$. Similarly we will want q_2 (the desired output of neuron 2) to be a function $f_2(x_1, x_2, \dots, x_k)$, so we write $q_2 = f_2(\boldsymbol{x})$. In general then show we need $q_l = f_l(\boldsymbol{x})$ for all l, with $f_l(\boldsymbol{x})$ just the l^{th} component of the desired full vector input-output function $\boldsymbol{f}(x) = (f_1(\boldsymbol{x}), \dots, f_m(\boldsymbol{x}))$.

Now show (you can use the in-class version of Funahashi's theorem, for single output neuron) that each of these single output functions f_l can be approximated as above, i.e.,

$$f_l(x) \;\approx\; \sum_{j=1}^{n} w_j^l \; H(V_l^j \cdot x - \theta_j^l) = q_l \qquad (1)$$

with the proper choice of w's and V's. Show we need the extra index l on the right side, since all parameters w, V, and θ may need to be distinct for different i-o functions f_l.

Thus show a *family* of functions $f_l(x)$ for $l = 1, 2, \ldots, m$ representing desired activations of each output q_l can all be approximated as in (1), with a different choice of w, V, and θ's for each function f_l.

We then still need to know if there is a neural network which can be constructed so that the outputs q_l in the final layer are given by formulas as in (1), i.e., an actual neural network that implements the formulas in (1). Show this can be done as follows. Assume that the first layer of our new network still contains the k neurons (whose activations are) $x_1 \ldots x_k$. Assume the middle layer is made up of m collections of neurons y_{11}, $y_{12}, \ldots y_{1n}$; $y_{21}, y_{22}, \ldots, y_{2n}$; $y_{31}, y_{32}, \ldots, y_{3n}$; $\ldots$ $y_{m1}, y_{m2}, \ldots, y_{mn}$. Assume that each of these collections is connected to all of the neurons $x_1 \ldots x_k$ in the first layer, but assume that the l^{th} collection of hidden neurons $y_{l1}, \ldots, y_{lr}$ is only connected to the l^{th} neuron (with activation) q_l in the third layer. Verify then that this l^{th} collection of neurons can be chosen to have activations $H(V_l^j - \theta_j^l)$ for any choice of V_l^j and θ_j^l we might want (since the connections from the input x to the l^{th} collection for different l are completely different from each other, so that the V_1^j can be chosen independently, as can the θ_j^l). Finally show the same for the w_j^l, i.e., that the connections for each group $y_{l1}, \ldots, y_{lr}$ to the third layer are completely independent for each l, so that for different l we may choose the w_j^l completely independently.

Thus show that even for different f_l as l varies, we can find approximations $\widehat{f}_l(\boldsymbol{x}) \equiv \sum_{j=1}^{n} w_j^l \, H(V_l^j \cdot \boldsymbol{x} - \theta_j^l)$ to these desired outputs which can be implemented in the multi-component network described above.

Finally, this allows you to re-state Funahashi's theorem for an underlying (target) vector function $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^k \to \mathbb{R}^m$. Now carefully extend the original theorem in class for approximating a function $f(\boldsymbol{x})$: $\mathbb{R}^k \to \mathbb{R}$ using a single output neuron to the current case of multiple output neurons. This can follow by considering each component function $f_l(\boldsymbol{x})$ of $\boldsymbol{f}(\boldsymbol{x})$ and showing that, as above, our single neural network can approximate each f_l with an $\widehat{f}_l$ within error ϵ. Show the full vector function $\boldsymbol{f}(\boldsymbol{x})$ can be approximated by an $\widehat{\boldsymbol{f}}(\boldsymbol{x})$ computed by the above network within error ϵ. Be sure to state this error estimate precisely.

4. Consider using the above relation between p and ∞ norms.