Notes on the paper by Park and Sandberg:

Page 305, bottom: The convolution operation * for two functions f(x) and g(x) on $\mathbb{R}^r$ is an operation on the two functions which yields a third function h(x), defined by

$$h(x) = (f*g)(x) = \int_{\mathbb{R}^r} f(x - y)\, g(y)\, dy.$$

The Fourier transform of a function f(x) of r variables $x = (x_1, x_2, \ldots, x_r)$ is denoted by

$$\mathcal{F}f(\xi) \equiv \hat{f}(\xi) = (2\pi)^{-r/2} \int_{\mathbb{R}^r} f(x)\, e^{i\xi \cdot x}\, dx,$$

where $\xi \in \mathbb{R}^r$.

By *hidden layer* is meant the middle layer of neurons. The reference to the centroid of a unit of the hidden layer is made clear later. Specifically, suppose that x_i is the activation level of the i^{th} neuron in the bottom (input) layer. Let

$$x = (x_1, x_2, \ldots, x_r).$$

Then the i^{th} neuron in the middle (hidden, or kernel node) layer computes the function $K(x - z_i)$, where z_i is a fixed vector which depends on i, and K(x) is a single predefined function (usually a function in the shape of a "bump" near the origin). The value of this function becomes the "activation level" of this neuron. The vector z_i is called the *centroid* of the i^{th} neuron. Generally, the function K depends on $x - z_i$ only through the value of $|x - z_i|$, which is the distance between the two vectors (i.e., K is really just a function of this distance).

Page 306, middle:

"... some nonlinear function of that distance..."

The distance referred to is $x - z_i$, and the nonlinear function of it is $K(x - z_i)$, which is a radially symmetric function about the point z_i if it only depends on $|x - z_i|$, as is assumed here.

"... the strongest output is obtained when the input is at the centroid of the node."

That is, the output of the i^{th} neuron in the middle (kernel node) layer is $K(x - z_i)$, which is largest when x is at z_i (the centroid), since we assume generally that K has its maximum near the origin, i.e., when its argument $x - z_i$ is near 0.

"Each output node gives a weighted summation of the outputs of kernel nodes."

This paper assumes (without any real loss of generality) that there is only one output node (contrary to the diagram). The output of this node depends on the activations of the nodes in the hidden layer. Recall from that the activation of the i^{th} node of the hidden layer is $K(x - z_i)$. The output node is assumed to have an activation level given by

$$q(x) = \sum_{i=1}^{M} w_i \cdot K(x - z_i), \tag{1}$$

(here we will have $\sigma = 1$, so we omit it), which is the weighted summation of outputs of kernel nodes referred to. Note that each kernel node has an output $K(x - z_i)$.

Page 306, bottom:

The reference to σ as a smoothing factor just means that for large values of the constant σ, the function $K\left(\frac{x - z_i}{\sigma}\right)$ is "wider", and hence "smoother." Again, we may fix σ to be 1 for our purposes.

"we call this family $S_0(K)$. . . "

The family referred to is the set of all functions in the form (1), with arbitrary choices of w_i and z_i *as well as* σ.

Page 307, top:

Note that the general function q(x) of this form differs from the previous one in that the scalings σ_i of the function K are allowed to vary inside the sum of the translates $K(x - z_i)$.

The one-dimensional output space which is referred to means that there is only one output neuron, and that its activation q(x) is given by (1).

Page 307, middle:

Recall that by a function K(x) being *radially symmetric* we mean that the function can be written $K(x) = H(|x|)$, where H is a function defined on the real numbers.

Note also that this paper uses $\| x \|$ to denote the Euclidean norm of the vector x; we will denote this by $|x|$.

Page 307, bottom:

Note that our goal is to show that the output function q(x) given in (1) for the neural network can approximate any other desired input-output function f(x). The first theorem tries to prove that if the desired function f is in L^1 (i.e., is integrable), then it can be approximated arbitrarily well by functions of the form q(x) in (1), in the L^1 norm. Equivalently, the goal of the theorem is to show that functions of the form q(x) in (1) are dense in the space of all L^1 functions, in using the L^1 norm.

"Since $C_c(\mathbb{R}^r)$ is dense in $L^1(\mathbb{R}^r)$. . . "

The proof of this fact (referred to in the analysis book of Rudin) is very similar to the proof of exercise 18 in Chapter 1 of Reed and Simon.

Note that the notation supp f_c denotes the support of the function f_c, i.e., the set of points where f_c is not equal to 0.

"Since f_c has compact support. . . "

Recall that compact support for f means that the set of points where f does not vanish is a bounded set in $\mathbb{R}^r$.

"supp $f_c = [-T,T]^r$ "

The set [-T, T] = $\{(x_1,x_2,\ldots,x_r) \mid -T \le x_i \le T \quad \forall\, i\}$. Since the support of f is bounded, we can make T large enough so that the support is contained in $[-T,T]^r$.

Page 308, top:

Equation (2): Note that since $K \in L^1$, it can be approximated by a continuous compactly supported function K_c arbitrarily well.

Page 308, bottom:

Note that the notation $\phi_\sigma(\alpha - \cdot)\, f_c(\cdot)$ simply refers to the function $\phi_\sigma(\alpha - x)\, f_c(x)$. Here α is another variable to be varied later on.

Recall the Riemann integral: Given a function $g(x_1,x_2,\ldots, x_n) = g(x)$ defined for x in

$$[-T,T]^r = \{x \mid -T \le x_i \le T \text{ for all } i\},$$

the Riemann integral consists of dividing the domain (consisting of the set $[-T,T]^r$) into a finite number of pieces $\{V_i\}_{i=1}^m$ of small measure, finding a point α_i in each piece V_i and forming the sum

$$\sum_{i=1}^{m} g(\alpha_i)\, \Delta V_i, \tag{2}$$

where ΔV_i denotes the volume of the piece V_i. We then take the limit as the number of pieces ΔV_i becomes infinite and the sizes of the volumes goes to 0, and define that as the Riemann integral:

$$\lim_{\Delta V \to 0} \sum_{i=1}^{m} g(\alpha_i)\, \Delta V_i \;\equiv\; \int_V g(x)\, dx.$$

Recall this is a slightly different way of defining an integral than the Lebesgue integral, but it works as well if the function g being integrated is continuous.

Here, the r-dimensional cube [-T,T] is divided into n smaller cubes along each of its r directions (imagine this in r = 3 dimensions if you like), and so the total cube is sliced evenly into n^r cubes, whose sides are 1/n the length of the original cube. Here the smaller cubes overlap on their boundaries, but it is not important to correct this since the boundary of a cube has measure 0. One can check that if we consider the set of points given in the paper:

$$x = [x_1,x_2,x_3,\ldots, x_r] = \left[-T+\tfrac{2i_1T}{n}, -T+\tfrac{2i_2T}{n}, -T+\tfrac{2i_3T}{n}, \ldots, -T+\tfrac{2i_rT}{n}\right], \tag{3}$$

where the values $i_1,i_2, i_3,\ldots, i_r$ are allowed to take all combinations of integer values from 1 to n, then this set of n^r points contains one point in each of the smaller n^r cubes mentioned above. Therefore if in the definition of the Riemann integral given above we choose the volumes to be the above cubes (so $m = n^r$), and the points to be as in (3), the function integrated is $g(x) = \phi_\sigma(\alpha - x) f_c(x)$ then the sum (2) becomes (plugging in the points α_i for x, which is the variable we are integrating over):

(4) $$\sum_{i=1}^{n^r} \phi_\sigma(\alpha - \alpha_i) f_c(\alpha_i)\, \Delta V_i = \sum_{i=1}^{n^r} \phi_\sigma(\alpha - \alpha_i) f_c(\alpha_i) \left(\tfrac{2T}{n}\right)^r,$$

since the sides of each of the small cubes are 2T/n, and thus their volumes are $\Delta V_i = \left(\tfrac{2T}{n}\right)^r$. Call this sum (which depends on the parameter α) $v_n(\alpha)$. In the limit this sum takes the value of the integral, so we know

(5) $$\lim_{n\to\infty} v_n(\alpha) = \lim_{n\to\infty} \sum_{i=1}^{n^r} \phi_\sigma(\alpha - \alpha_i) f_c(\alpha_i) \left(\tfrac{2T}{n}\right)^r = \int_{[-T,T]^r} \phi_\sigma(\alpha - x)\, f_c(x)\, dx,$$

$$= (\phi_\sigma * f)(\alpha)$$

where the last quantity is just the integral over the original cube $[-T,T]^r$.

Note finally that the convolution operation on the bottom of the page is defined earlier in these notes. Note that in the last equality on page 308, since f_c is equal to 0 outside the cube $[-T,T]^r$, we can reduce the integration over all of $\mathbb{R}^r$ to an integration over only this cube.

Page 309, top:

The statement $v_n(\alpha) \underset{n\to\infty}{\to} (\phi_\sigma * f_c)(\alpha)$ is just a restatement of the fact that the Riemann sums (4) converge to the Riemann integral (5), renaming the Riemann sums as v_n and the integral as the convolution, according to definitions.

"Since $(\phi_\sigma * f_c)$ and the v_n are dominated by an integrable bounded function with compact support. . . "

Note that the function $(\phi_\sigma * f_c)(\alpha)$, viewed *as a function of* α, has bounded support. Indeed, notice that the functions $\phi_\sigma(x)$ and $f_c(x)$ by their definitions have bounded support (i.e., are 0 for x outside of some fixed bounded set). To be precise, assume that there is a constant A such that $\phi_\sigma(x)$ and $f_c(x)$ are both equal to 0 if $|x|>A$ (i.e., outside of a ball of radius A about the origin in r dimensional space). Then notice that for fixed α $\phi_\sigma(\alpha - x)$ vanishes if $|\alpha - x|>A$, and $f_c(x)$ vanishes if $|x|>A$. Therefore, it is easy to check that when the (for now fixed) point α satisfies $|\alpha|>2A$, then every point x in $\mathbb{R}^r$ satisfies either $|\alpha - x|>A$ or $|x|>A$ (i.e., every point is a distance greater than A from either the point α or from the origin. Thus if $|\alpha|>2A$, then for each point x, either the function $\phi_\sigma(\alpha - x)$ or the function $f_c(x)$ is 0, so that the product $\phi_\sigma(\alpha - x)\, f_c(x)$ is zero for *all x*. Thus the integral defining $(\phi_\sigma * f_c)(\alpha)$ is 0 for $|\alpha|>2A$. Therefore the function $(\phi_\sigma * f_c)(\alpha)$ is zero for α outside of a bounded set (now viewing the function as depending on α).

In addition, it is easy to check from the definition that since the functions $\phi_\sigma(\alpha - x)$ and $f_c(x)$ are both bounded as functions of x, and have bounded (compact) support as functions of x, the integral

$$(\phi_\sigma * f_c)(\alpha) = \int_{[-T,T]^r} \phi_\sigma(\alpha - x)\, f_c(x)\, dx$$

is also bounded for all α (simply replace the integrand by its maximum possible value to show that the full integral cannot be larger than some fixed constant, so that $(\phi_\sigma * f)(\alpha)$ is bounded for all α). Since we have established that as a function of α, $(\phi_\sigma * f)(\alpha)$ is bounded and has bounded support (i.e. is zero outside of a bounded set), it is clear that $(\phi_\sigma * f)(\alpha)$, as a function of α, is bounded by an integrable function (such a function could just be chosen to be the maximum value of $(\phi_\sigma * f)(\alpha)$ on the set where it is non-zero, and zero elsewhere).

Similarly, we can show that $v_n(\alpha)$ (*as functions of* α) are all bounded by the same integrable function of α (when we say bounded by an integrable function, we always mean that the absolute value $|v_n(\alpha)|$ is what's bounded; the term "dominated by" is synonymous with "bounded by" here). To show this, notice that if B is larger than the maximum values of both the functions ϕ_σ and f_c, then

$$|v_n(\alpha)| = \left| \sum_{i=1}^{n^r} \phi_\sigma(\alpha - \alpha_i) f_c(\alpha_i) \left(\frac{2T}{n}\right)^r \right|$$

$$\leq \left| \sum_{i=1}^{n^r} B^2 \left(\frac{2T}{n}\right)^r \right|$$

$$= B^2 n^r (2T/n)^r$$

$$= B^2/(2T)^r$$

(since the last sum is just a sum of constants), which is certainly bounded. Furthermore, all of the functions $v_n(\alpha)$ are 0 for $|\alpha|>2A$, for the same reason as the integral above was. Therefore, for the same reasons as above, all of the functions $v_n(\alpha)$ are bounded by a single integrable function (the function equal to $B^2/(2T)^r$ on the ball $|\alpha| \leq 2A$). Thus by the dominated convergence theorem and (5) above, we have that

$$\int_{\mathbb{R}^r} |(\phi_\sigma * f_c)(\alpha) - v_n(\alpha)| d\alpha \xrightarrow[n\to\infty]{} 0,$$

as desired.

Page 309, middle:

Note that

$$\left\| \frac{1}{\sigma^r} K_c\left(\frac{\cdot - \alpha_i}{\sigma}\right) - \frac{1}{\sigma^r} K_c\left(\frac{\cdot - \alpha_i}{\sigma}\right) \right\| \equiv \int \left| \frac{1}{\sigma^r} K_c\left(\frac{x - \alpha_i}{\sigma}\right) - \frac{1}{\sigma^r} K_c\left(\frac{x - \alpha_i}{\sigma}\right) \right| dx$$

$$\equiv \int \left| K_c(x') - K_c(x') \right| dx',$$

as follows from the change of variables $x' = (x - \alpha_i)/\sigma$ (note that $dx' = \sigma^r\, dx$, since there are r variables of integration). This gives the second equation after (4).

To obtain equation (5), just plug in the definition of v_N and $\tilde{v}_N$, and replace f_c by its maximum value $\| f_c \|_\infty$ (recall the second equation after (2), which gives lower bound on $\int K_c(x)\, dx$).

Now since we have shown that an arbitrary function $f \in L^1$ can be approximated by aa function $\tilde{v}_n \in S_0$ arbitrarily well, we conclude that S_0 is dense in L^1.