A. Proof of Theorems

Theorem 1. *Suppose a shift neural network in the form* $\hat{f}_{\mathbf{w}}(\mathbf{x}) = \sum_{j=1}^{J}\sum_{k=1}^{d_j} w_k^j h_k^j(\mathbf{x})$, *where* $w_k^j \in \{0\} \cup \{\pm 2^p\}$ *approximates the true unknown function* $f(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d$. *There is another dense neural network with* $w_k'^j \in \{\pm 2^p\}$ *that has a similar form* $\hat{f}_{\mathbf{w}'}(\mathbf{x}) = \sum_{j=1}^{J}\sum_{k=1}^{d'_j} \mathbf{w}_k'^j h_k'^j(\mathbf{x})$, *such that* $\forall \mathbf{x} \in \mathbb{R}^d, \hat{f}_{\mathbf{w}}(\mathbf{x}) = \hat{f}_{\mathbf{w}'}(\mathbf{x})$.

Proof. The proof is straightforward by isolating zero shifts, and recreating these null weights in a larger dense shift network with opposite weight signs. Define the j^{th} layer as

$$h^j(\mathbf{x}) = a(\mathbf{W}^j h^{j-1}(\mathbf{x}) + \mathbf{b}^j),$$

where $h^0(\mathbf{x}) = \mathbf{x}$, J is the total number of layers each of output dimension d_j and $\mathbf{W}$ of size $d_j \times d_{j-1}$ can be a Toeplitz matrix for a convolutional layer, $a(\cdot)$ is the activation function, and $\mathbf{b}$ is the bias term. Define the shift network approximation of $f(\mathbf{x})$ as

$$\hat{f}_{\mathbf{w}}(\mathbf{x}) = \sum_{j=1}^{J}\sum_{k=1}^{d_j} w_k^j h_k^j(\mathbf{x})$$

in which $w_k^j \in \{0\} \cup \{\pm 2^p\}$ defines a shift network. We re-create an equivalent desne shift network by isolating null weights $w_k^j = 0$ and replacing them with a larger dense shift network of an arbitrary weights but with opposite signs. Now suppose $\mathbf{w}_0 = \{k \mid w_k^j = 0\}$ and $\mathbf{w}_1 = \{k \mid w_k^j \neq 0\}$ where $\mathbf{w} = [\mathbf{w}_0^\top \mathbf{w}_1^\top]$, $d_j = d_{0j} + d_{1j}$

$$\hat{f}_{\mathbf{w}}(\mathbf{x}) = \sum_{j=1}^{J}\left[\sum_{k=1}^{d_{0j}} w_{0k}^j h_{0k}^j(\mathbf{x}) + \sum_{k=1}^{d_{1j}} w_{1k}^j h_{1k}^j(\mathbf{x})\right].$$

Assume $\tilde{\mathbf{w}} \in \{\pm 2^p\}$ is a nonzero shift arbitrary vector of elements $\tilde{w}_{0k}$,

$$\begin{aligned}
&\sum_{j=1}^{J}\left[\sum_{k=1}^{d_{0j}} \mathbf{w}_{0k}^j h_{0k}^j(\mathbf{x}) + \sum_{k=1}^{d_{1j}} \mathbf{w}_{1k} h_{1k}^j(\mathbf{x})\right] \\
= &\sum_{j=1}^{J}\left[\sum_{k=1}^{d_{0j}} \tilde{\mathbf{w}}_k^j h_{0k}^j(\mathbf{x}) - \tilde{\mathbf{w}}_k^j h_{0k}^j(\mathbf{x}) + \sum_{k=1}^{d_{1j}} \mathbf{w}_{1k} h_{1k}^j(\mathbf{x})\right].
\end{aligned}$$

By defining $\mathbf{w}' = [\tilde{\mathbf{w}}, -\tilde{\mathbf{w}}, \mathbf{w}_0]$ of increased size $d'_j = d_{0j} + d_j \geq d_j$, one may rearrange terms and rewrite the neural approximate as

$$\hat{f}_{\mathbf{w}'}(\mathbf{x}) = \sum_{j=1}^{J}\sum_{k=1}^{d'_j} \mathbf{w}_k'^j h_k'^j(\mathbf{x}),$$

where $h_k'^j$ is either h_{0k}^j or h_{1k}^j depending on the dense shift weight $\mathbf{w}_k'^j \in \{\pm 2^p\}$. □

Theorem 2. *Dense shift network with a Lischitz activation function is a universal approximator on a compact set* K *for any measurable continuous function* $f \in C(K)$ *with respect to the measure* μ, *given that its weight and activations* $\mathbf{w}, h(\mathbf{x})$ *remain close to the regular network in the following sense*

$$\int_K \left(\sum_{j=1}^{J}\sum_{k=1}^{d_j}\left(\mathbf{w}_k^j h_k^j(\mathbf{x}) - \mathbf{w}_k'^j h_k'^j(\mathbf{x})\right)\right)^2 d\mu < \frac{\epsilon}{4},$$

where $\frac{\epsilon}{4}$ *is the approximation quality of the regular neural network,* $\left(\mathbf{w}, h^j(\mathbf{x})\right)$ *and* $\left(\mathbf{w}', h'^j(\mathbf{x})\right)$ *are the weight and activations of the regular and DenseShift networks in layer* j, *respectively.*

Proof. It is well-known that shallow networks are universal approximator [18] as well as deep networks [50]. These results holds in infinity norm, so is also valid in ℓ_p norm with $p < \infty$. For the simplicity of the mathematical mechanics here we only focus on the multilayer perceptron [18] on ℓ_2 norm

$$\left\|f - \hat{f}\right\| = \int_K \left|f(\mathbf{x}) - \hat{f}(\mathbf{x})\right|^2 d\mu \tag{9}$$

Suppose $\hat{f}_{\mathbf{w}}$ is a real weight neural network and $\hat{f}_{\mathbf{w}'}$ is a dense shift version of the same network $\hat{f}_{\mathbf{w}}$, of course with weights $\mathbf{w}' \in \{\pm 2^p\}$.

$$\begin{aligned}
&\left\|f - \hat{f}_{\mathbf{w}'}\right\| \\
= &\int_K \left|f(\mathbf{x}) - \hat{f}_{\mathbf{w}'}(\mathbf{x}) \pm \hat{f}_{\mathbf{w}}(\mathbf{x})\right|^2 d\mu && (10)\\
= &\int_K |f(\mathbf{x}) - \hat{f}_{\mathbf{w}}(\mathbf{x})|^2 d\mu && (11)\\
&+ \int_K |\hat{f}_{\mathbf{w}}(\mathbf{x}) - \hat{f}_{\mathbf{w}'}(\mathbf{x})|^2 d\mu && (12)\\
&+ 2\int_K |[f(\mathbf{x}) - \hat{f}_{\mathbf{w}}(\mathbf{x})][\hat{f}_{\mathbf{w}}(\mathbf{x}) - \hat{f}_{\mathbf{w}'}(\mathbf{x})]| d\mu && (13)
\end{aligned}$$

In order to show that shift networks are universal approximator it is enough to show that $\left\|f - \hat{f}_{\mathbf{w}'}\right\|$ is bounded by an arbitrarily small $\epsilon > 0$. The first term (11) is bounded by $\frac{\epsilon}{4}$ [18]. The second term (12) is bounded given the shift net closeness assumption

$$\begin{aligned}
&\int_K |\hat{f}_{\mathbf{w}}(\mathbf{x}) - \hat{f}_{\mathbf{w}'}(\mathbf{x})|^2 d\mu \\
= &\int_K \left(\sum_{j=1}^{J}\sum_{k=1}^{d_j} \mathbf{w}_k^j h_k^j(\mathbf{x}) - \sum_{j=1}^{J}\sum_{k=1}^{d_j} \mathbf{w}_k'^j h_k'^j(\mathbf{x})\right)^2 d\mu \\
< &\frac{\epsilon}{4}
\end{aligned}$$

The last term (13) is bounded by $\frac{\epsilon}{2}$ thanks to the Cauchy-Schwartz inequality. So (9) is bounded by ϵ by merging the pieces together. $\square$