1 Implementing Symbol Manipulation in MLPs

When specifying the structure of a neural network, we restrict our hypothesis class, i.e. we restrict the possible mappings from the input to the output space. Furthermore, we also encode a *bias* into the network, i.e. a certain tendency to prefer some mappings over others.

TODO: example interpolation etc.

Our goal is to encode the premised human bias for UQOTOMs (at least for certain tasks) to MLPs in order to achieve good generalization of these models. We hope that by examining successful models we might be able to draw conclusions to the inner mechanisms of the human brain.

Understanding the bias and learning behavior of multilayer perceptrons is rather difficult, which is why empirical testing is used to assess different architectures.

1.1 MLPs with Linear Activation Functions

When restricting our models to only using linear activation function, though, we can infer some information:

Proposition 1.1. *The hypothesis class of a MLP with linear activation functions is restricted to linear functions $f : \mathbb{R}^m \mapsto \mathbb{R}^n$, i.e. there exists a matrix $\boldsymbol{M} \in \mathbb{R}^{n \times m}$ s.t. $f(\boldsymbol{v}) = \boldsymbol{M}\boldsymbol{v}$.*

Proof. Consider a multi-layer perceptron with L layers and linear activation functions $\phi(x) = \alpha x$. Let the input be $\boldsymbol{v} \in \mathbb{R}^m$, and let each layer i have a weight matrix $\boldsymbol{W}^{(i)}$. The output of the network is:

$$f(\boldsymbol{v}) = \phi\left(\boldsymbol{W}^{(L)}\phi\left(\boldsymbol{W}^{(L-1)}\phi\left(\cdots\phi\left(\boldsymbol{W}^{(1)}\boldsymbol{v}\right)\cdots\right)\right)\right) \quad ,$$

where ϕ is applied element-wise. Since $\phi(x) = \alpha x$, this simplifies to:

$$f(\boldsymbol{v}) = \alpha^L \boldsymbol{W}^{(L)}\boldsymbol{W}^{(L-1)}\cdots\boldsymbol{W}^{(1)}\boldsymbol{v} \quad .$$

Let $\boldsymbol{M} = \alpha^L \boldsymbol{W}^{(L)}\boldsymbol{W}^{(L-1)}\cdots\boldsymbol{W}^{(1)} \in \mathbb{R}^{n \times m}$. Then $f(\boldsymbol{v}) = \boldsymbol{M}\boldsymbol{v}$, which is a linear function. Thus, the hypothesis class is restricted to linear mappings from $\mathbb{R}^m$ to $\mathbb{R}^n$. □

Remark 1.1. This proposition also holds when allowing α to vary by layer, or even with every node.

Lemma 1.1. *Let $\boldsymbol{M} \in \mathbb{R}^{n \times m}$ be a matrix inducing the mapping $f(\boldsymbol{v}) \coloneqq \boldsymbol{M}\boldsymbol{v}$. Then we have*

$$f \textit{ injective} \iff \operatorname{rank} \boldsymbol{M} = m \quad .$$

Proof. The function f is injective iff $\ker(\boldsymbol{M}) = \{\mathbf{0}\}$:

' $\implies$ ' is trivial. For ' $\impliedby$ ', consider the contraposition 'f is not injective $\implies \ker(\boldsymbol{M}) \neq \{\mathbf{0}\}$'. Since f is not injective, there must be $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$ with $\boldsymbol{u} \neq \boldsymbol{v}$ s.t. $\boldsymbol{M}\boldsymbol{u} = \boldsymbol{M}\boldsymbol{v}$. Hence, $\boldsymbol{M}(\boldsymbol{u} - \boldsymbol{v}) = \mathbf{0}$, and thus $(\boldsymbol{u} - \boldsymbol{v}) \in \ker \boldsymbol{M}$. Note that $(\boldsymbol{u} - \boldsymbol{v}) \neq \mathbf{0}$.

Now, by the rank-nullity theorem, we have:

$$\dim(\ker(\boldsymbol{M})) + \operatorname{rank}(\boldsymbol{M}) = m \quad .$$

Finally, we see that

$$\operatorname{rank}(\boldsymbol{M}) = m \iff \dim(\ker(\boldsymbol{M})) = 0 \iff \ker(\boldsymbol{M}) = \{\mathbf{0}\} \iff f \text{ injective} \quad .$$

□

Corollary 1.1. *A MLP with only one input node and linear activations is forced to learn either $f : \mathbb{R} \mapsto \mathbb{R}^n, f(x) \equiv \mathbf{0}$ or an UQOTOM.*

Proof. Based on proposition 1.1 we know that f can be written as $f(x) = \boldsymbol{v}x$ for some $v \in \mathbb{R}^{n \times 1}$. Furthermore, based on lemma 1.1 we know that f injective $\iff \operatorname{rank} \boldsymbol{v} = 1$. Since $\boldsymbol{v}$ is a vector, we have $\operatorname{rank} \boldsymbol{v} = 1 \iff \boldsymbol{v} \neq \mathbf{0}$.

Hence, for $\boldsymbol{v} \neq \mathbf{0}$ we have that f is injective. When restricting the image domain accordingly we also have that f is surjective, and hence UQOTOM.

On the other hand, if $\boldsymbol{v} = \mathbf{0}$, then $f(x) \equiv \mathbf{0}$. □

Remark 1.2. If a MLP with linear activations has multiple input nodes, it will depend on the properties of matrix $\boldsymbol{M}$ whether or not the MLP implements an UQOTOM based on lemma 1.1.

For example, consider the mapping

$$f : \mathbb{R}^2 \mapsto \mathbb{R}^2, \boldsymbol{v} \mapsto \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \boldsymbol{v} \quad .$$

It is not injective, since $f \begin{pmatrix} 1 \\ 0 \end{pmatrix} = f \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.