Exercise 3: The Expressive Power of Depth: XOR with a 1-Hidden-Layer MLP

The XOR (exclusive-or) Boolean function takes two binary inputs $x = (x_1, x_2)$ with $x_i \in \{0, 1\}$ and outputs 1 if exactly one of the inputs is 1, and 0 otherwise:

$$f(0,0) = 0, \qquad f(1,1) = 0, \qquad f(0,1) = 1, \qquad f(1,0) = 1.$$

XOR is the canonical example of a classification problem that cannot be solved by a linear model but can be solved by an MLP with at least one hidden layer.

(a) Prove that the XOR points are not linearly separable in $\mathbb{R}^2$.

(b) Consider a 1-hidden-layer ReLU MLP: $f_0(x) = x$, $A^{(0)} = W_0 f_0(x) + b_0$, $f_1 = \sigma(A^{(0)})$ with $\sigma(z) = \max(0, z)$, and output $f(x) = \beta^\top f_1 + b_1$, where b_1 is a scalar output bias. Show that with hidden width 2 there exist parameters (W_0, b_0, β, b_1) that implement XOR under the decision rule "predict Class 1 if $f(x) > 0$." Give one explicit working choice and verify it on the four inputs.

(c) Can you still implement XOR with the same architecture and hidden width 2 when $b_1 = 0$? Give an explicit (W_0, b_0, β) if that is the case or argue why not for your mapping.

Solution

(a) Assume toward a contradiction that there is a linear classifier $g(x) = w^\top x + b$ such that $g(x) > 0$ on positives $\{(1,0),(0,1)\}$ and $g(x) < 0$ on negatives $\{(0,0),(1,1)\}$. Then

$$\begin{aligned} (1,0): &\quad w_1 + b > 0, \\ (0,1): &\quad w_2 + b > 0, \\ (0,0): &\quad b < 0, \\ (1,1): &\quad w_1 + w_2 + b < 0. \end{aligned}$$

Adding the two positive inequalities yields $w_1 + w_2 + 2b > 0$, while adding the two negative ones gives $w_1 + w_2 + 2b < 0$, a contradiction. Hence XOR is not linearly separable in $\mathbb{R}^2$. (Equivalently, the convex hulls of positives $\{(1,0),(0,1)\}$ and negatives $\{(0,0),(1,1)\}$ both contain $(\frac{1}{2}, \frac{1}{2})$, so they intersect.)

(b) Choose hidden width 2 and set

$$W_0 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \qquad b_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \beta = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad b_1 = -\tfrac{1}{2}.$$

For input $x = (x_1, x_2)$ the hidden pre-activations are

$$A^{(0)} = \begin{pmatrix} x_1 - x_2 \\ x_2 - x_1 \end{pmatrix}, \quad f_1 = \sigma(A^{(0)}) = \begin{pmatrix} \max(0, x_1 - x_2) \\ \max(0, x_2 - x_1) \end{pmatrix}.$$

Evaluating on the four inputs:

x	$f_1(x)$	$\beta^\top f_1(x)$	$f(x) = \beta^\top f_1(x) + b_1$
$(0,0)$	$(0,0)$	0	$-\frac{1}{2} < 0$
$(1,1)$	$(0,0)$	0	$-\frac{1}{2} < 0$
$(1,0)$	$(1,0)$	1	$\frac{1}{2} > 0$
$(0,1)$	$(0,1)$	1	$\frac{1}{2} > 0$

Thus, under the decision rule "predict Class 1 if $f(x) > 0$," the network implements XOR.

(c) We use a 1-hidden-layer ReLU MLP with two hidden units and no output bias:

$$h(x) = \mathrm{ReLU}(W_0 x + b_0) \in \mathbb{R}^2, \qquad f(x) = \beta^\top h(x), \qquad \mathrm{ReLU}(z) = \max\{0, z\} \text{ (entrywise)}.$$

Explicit realization of XOR under the rule "predict Class 1 if $f(x) > 0$". Choose

$$W_0 = \begin{pmatrix} 1.5 & -0.6 \\ -0.6 & 1.5 \end{pmatrix}, \qquad b_0 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \qquad \beta = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad b_1 = 0.$$

For $x \in \{(0,0), (1,0), (0,1), (1,1)\}$ we get

$$W_0 x + b_0 \in \{(-1,-1), (0.5,-1.6), (-1.6,0.5), (-0.1,-0.1)\},$$

hence

$$h(x) \in \{(0,0), (0.5,0), (0,0.5), (0,0)\}, \qquad f(x) \in \{0,\ 0.5,\ 0.5,\ 0\}.$$

Therefore $f(1,0) > 0$, $f(0,1) > 0$, and $f(0,0) = f(1,1) = 0$, which implements XOR with the decision rule "Class 1 if $f(x) > 0$" (and Class 0 otherwise).