Exercise sheet 2

Federico Santona, Radek Vasicek Ruiz, Isreal Cabrerizo Garcia

Exercise 2.1

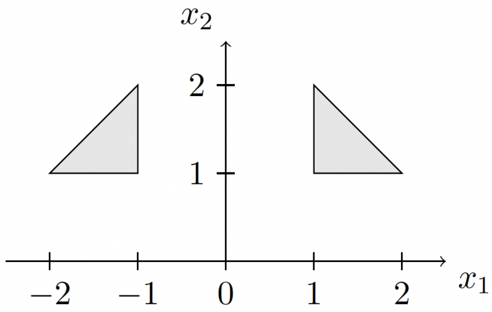

Figure 1: Visualization of the problem

We design a 3-layer perceptron using hard-threshold (Heaviside) activations to output $y = 1$ if and only if the input point $x = (x_1, x_2)$ lies in one of two shaded triangles, and $y = 0$ otherwise.
Since the input is bidimensional, the input layer will feature 2 nodes. In the following subsections we deal with the 2 hidden layers and the final output layer.

Layer 1: Half-space Detectors (8 Units)

Each unit computes

$$h_i^{(1)} = \theta\big(w_i^{(1)T} x + b_i^{(1)}\big), \quad \theta(u) = \begin{cases} 1, & u \geq 0, \\ 0, & u < 0. \end{cases}$$

Where $w_i^{(1)T}$ is a column 2-dimensional vector and x is the row 2-dimensional vector described above. We chose to use 8 nodes for the first hidden layer, the weights and biases implement the following linear inequalities:

Unit	$w_i^{(1)}$	$b_i^{(1)}$	Inequality
1	$[1, 0]$	$+2$	$x_1 + 2 \geq 0 \iff x_1 \geq -2$
2	$[-1, 0]$	-1	$-x_1 - 1 \geq 0 \iff x_1 \leq -1$
3	$[0, 1]$	-1	$x_2 - 1 \geq 0 \iff x_2 \geq 1$
4	$[1, -1]$	$+3$	$x_1 - x_2 + 3 \geq 0 \iff x_2 \leq x_1 + 3$
5	$[1, 0]$	-1	$x_1 - 1 \geq 0 \iff x_1 \geq 1$
6	$[-1, 0]$	$+2$	$-x_1 + 2 \geq 0 \iff x_1 \leq 2$
7	$[0, 1]$	-1	$x_2 - 1 \geq 0 \iff x_2 \geq 1$
8	$[-1, -1]$	$+3$	$-x_1 - x_2 + 3 \geq 0 \iff x_2 \leq -x_1 + 3$

Table 1: Layer 1 implements four boundary tests for each of the two triangles.

Units 1–4 carve the left triangle; units 5–8 carve the right triangle.

Layer 2: Triangle Indicators (2 Units)

In the second hidden layer we combine ("AND") the four half-space tests for each triangle by a single linear threshold unit. Denote by $h^{(1)} = (h_1^{(1)}, \ldots, h_8^{(1)})^T$ the outputs of Layer 1. We set:

$$\begin{aligned} h_1^{(2)} &= \theta\big(h_1^{(1)} + h_2^{(1)} + h_3^{(1)} + h_4^{(1)} - 3.5\big), \\ h_2^{(2)} &= \theta\big(h_5^{(1)} + h_6^{(1)} + h_7^{(1)} + h_8^{(1)} - 3.5\big). \end{aligned}$$

where θ was defined in the previous layer.

Interpretation of the threshold:

Each sum $S_j = \sum_{i \in I_j} h_i^{(1)}$ for $j = 1, 2$ can only take integer values $0, 1, 2, 3, 4$. By choosing the bias $b_j^{(2)} = -3.5$, the argument of ϕ is

$$S_j - 3.5 \;\geq\; 0 \quad \iff \quad S_j \;\geq\; 3.5 \quad \iff \quad S_j = 4.$$

Thus $h_j^{(2)} = 1$ if and only if *all four* relevant half-space tests are satisfied simultaneously—i.e. the point lies inside (or on the boundary of) the corresponding triangle.
This can also be visualized in matrix form. Let:

$$W^{(2)} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}, \quad b^{(2)} = \begin{pmatrix} -3.5 \\ -3.5 \end{pmatrix}.$$

Then

$$h^{(2)} = \phi\big(W^{(2)}\, h^{(1)} + b^{(2)}\big) \quad \Longrightarrow \quad h^{(2)}_j = 1 \iff \sum_{i \in I_j} h^{(1)}_i = 4, \quad j = 1, 2.$$

Layer 3: OR Operation (1 Unit)

We "OR" the two triangle indicators:

$$y = \phi\big(h^{(2)}_1 + h^{(2)}_2 - 0.5\big),$$

which yields $y = 1$ if either $h^{(2)}_1 = 1$ or $h^{(2)}_2 = 1$. In weights:

$$W^{(3)} = [1 \;\; 1], \quad b^{(3)} = -0.5.$$

This network classifies exactly the two shaded triangles.

Exercise 2.2: Equivalent Networks

Proof that any multi-layer perceptron with identity activations collapses to a single layer

Setup. Label the layers by $k = 1, 2, \ldots, L$. Let the input be $x \in \mathbb{R}^n$, and the final output $y \in \mathbb{R}^m$. Since the activation function is the identity,

$$\phi(u) = u,$$

each layer computes an affine map

$$h^{(k)} \;=\; W^{(k)}\, h^{(k-1)} \;+\; b^{(k)}, \qquad h^{(0)} = x,$$

and the network output is

$$y = h^{(L)} = W^{(L)}\, h^{(L-1)} + b^{(L)}.$$

Folding in layers. We now show by repeated substitution that every bias term $b^{(k)}$ is carried forward through all higher-numbered weight matrices. Starting from

$$y = W^{(L)}\, h^{(L-1)} + b^{(L)},$$

and using $h^{(L-1)} = W^{(L-1)}\, h^{(L-2)} + b^{(L-1)}$, we get

$$\begin{aligned} y &= W^{(L)}\big(W^{(L-1)}\, h^{(L-2)} + b^{(L-1)}\big) + b^{(L)} \\ &= \big(W^{(L)} W^{(L-1)}\big)\, h^{(L-2)} + \underbrace{W^{(L)}\, b^{(L-1)}}_{\text{bias from layer } L-1} + b^{(L)}. \end{aligned}$$

Next, substitute $h^{(L-2)} = W^{(L-2)}\, h^{(L-3)} + b^{(L-2)}$ to carry forward $b^{(L-2)}$:

$$y = W^{(L)}W^{(L-1)}W^{(L-2)}\, h^{(L-3)} + W^{(L)}W^{(L-1)}\, b^{(L-2)} + W^{(L)}\, b^{(L-1)} + b^{(L)},$$

and so on down to layer 1. After folding in all layers $k = L-1, L-2, \ldots, 1$, we arrive at

$$y = \underbrace{W^{(L)}W^{(L-1)} \cdots W^{(1)}}_{W_{\text{total}}}\, x + \sum_{k=1}^{L} \Bigl(\underbrace{W^{(L)}W^{(L-1)} \cdots W^{(k+1)}}_{\text{propagate } b^{(k)} \text{ forward}}\Bigr) b^{(k)}.$$

Here by convention the empty product for $k = L$ is the identity:

$$\prod_{j=L}^{L+1} W^{(j)} = I.$$

Thus defining

$$W_{\text{total}} = W^{(L)}W^{(L-1)} \cdots W^{(1)}, \quad b_{\text{total}} = \sum_{k=1}^{L} \Bigl(\prod_{j=L}^{k+1} W^{(j)}\Bigr)\, b^{(k)},$$

we obtain the single-layer form

$$y = W_{\text{total}}\, x + b_{\text{total}}.$$

Conclusion. Since each layer with identity activation is affine, and the composition of affine maps is affine, the entire L-layer network is equivalent to a single affine transformation $x \mapsto W_{\text{total}}x + b_{\text{total}}$. Hence no hidden layer increases expressivity beyond that of a single layer.

(b) Construction of an equivalent single-layer perceptron

The given network has one hidden layer of size 3 and an output layer of size 2, with identity activations:

$$h^{(1)} = W^{(1)}x + b^{(1)}, \qquad y = W^{(2)}\, h^{(1)} + b^{(2)}.$$

By the result of part (a), the two layers collapse to one affine map

$$y = W_{\text{total}}\, x + b_{\text{total}},$$

where

$$W_{\text{total}} \;=\; W^{(2)}\, W^{(1)}, \qquad b_{\text{total}} \;=\; W^{(2)}\, b^{(1)} \;+\; b^{(2)}.$$

Substitute the given matrices

$$W^{(1)} = \begin{pmatrix} 1 & 2 & -2 \\ -1 & -1 & 2 \\ 3 & 2 & 3 \end{pmatrix}, \quad b^{(1)} = \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}, \quad W^{(2)} = \begin{pmatrix} 2 & -1 & -1 \\ 1 & 2 & 3 \end{pmatrix}, \quad b^{(2)} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}.$$

Then

$$W_{\text{total}} = W^{(2)} W^{(1)} = \begin{pmatrix} 2 & -1 & -1 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 & -2 \\ -1 & -1 & 2 \\ 3 & 2 & 3 \end{pmatrix} = \begin{pmatrix} 0 & 3 & -9 \\ 8 & 6 & 11 \end{pmatrix},$$

$$b_{\text{total}} = W^{(2)}\, b^{(1)} + b^{(2)} = \begin{pmatrix} 2 & -1 & -1 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} -5 \\ 7 \end{pmatrix} + \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} -4 \\ 5 \end{pmatrix}.$$

Hence the equivalent single-layer perceptron is

$$y \;=\; \begin{pmatrix} 0 & 3 & -9 \\ 8 & 6 & 11 \end{pmatrix} x \;+\; \begin{pmatrix} -4 \\ 5 \end{pmatrix}.$$

As required.

Exercise 2.3: Matrix-valued functions and gradients

Let $X = R^{n\times n}$ be the vector space of real $n \times n$ matrices. We equip X with the *Frobenius inner product*, defined entry-wise by

$$\langle A, B \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}\, B_{ij} = tr\big(A\, B^T\big).$$

Here $tr(M)$ denotes the trace of M, i.e. the sum of its diagonal entries. Indeed,

$$tr(A\, B^T) = \sum_{i=1}^{n} (A\, B^T)_{ii} = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}\, B_{ij},$$

which shows that $tr(A\, B^T)$ coincides with the standard dot-product of the n^2 entries of A and B.
For a (possibly matrix-valued) function $f\colon X \to Y$, its *Fréchet derivative* $\underset{\cdot}{f}(A)\colon X \to Y$ at A is the unique linear map satisfying

$$f(A + H) \;=\; f(A) + df(A)[H] + o(\|H\|).$$

When $Y = R$, the *gradient* $\nabla f(A) \in X$ is defined implicitly by

$$df(A)[H] = \langle \nabla f(A),\, H \rangle \quad \text{for all } H \in X.$$

(a) **Function:**

$$f(A) \;=\; A : B \;=\; \langle A, B\rangle, \quad B \in X \text{ fixed.}$$

Derivation: For a small perturbation H,

$f(A+H) = tr\big((A+H)B^T\big) = tr(AB^T) + tr(HB^T) = f(A) + tr(HB^T).$

Hence the linear part is

$$df(A)[H] \;=\; tr(HB^T) = \langle H, B\rangle,$$

and by the Riesz representation the gradient is $\nabla f(A) = B$.

(b) **Function:**

$$f(A) = B\,A, \quad B \in X \text{ fixed.}$$

Derivation: Since f is itself matrix-valued, we perturb by H and get

$$f(A+H) - f(A) = B\,(A+H) - B\,A = B\,H.$$

No higher-order terms appear, so the Fréchet derivative is the linear map

$$df(A)[H] = B\,H.$$

(c) **Function:**

$$f(A) = A^2 = A \cdot A.$$

Derivation: Perturbing A by H gives

$$(A+H)^2 - A^2 = AH + HA + H^2.$$

The term H^2 is $o(\|H\|)$, so the Fréchet derivative is

$$df(A)[H] = A\,H + H\,A.$$

(d) **Function:**

$$f(A) = A : A = \langle A, A\rangle = tr(A\,A^T).$$

Derivation: Under $A \mapsto A + H$,

$f(A+H) - f(A) = tr\big((A+H)(A+H)^T\big) - tr(A\,A^T) = tr(AH^T + HA^T) + tr(HH^T).$

The last term is $o(\|H\|)$, so

$$df(A)[H] = tr(AH^T) + tr(HA^T) = \langle H, A\rangle + \langle A, H\rangle = 2\,\langle H, A\rangle.$$

Thus by the Riesz representation the gradient is $\nabla f(A) = 2A$.