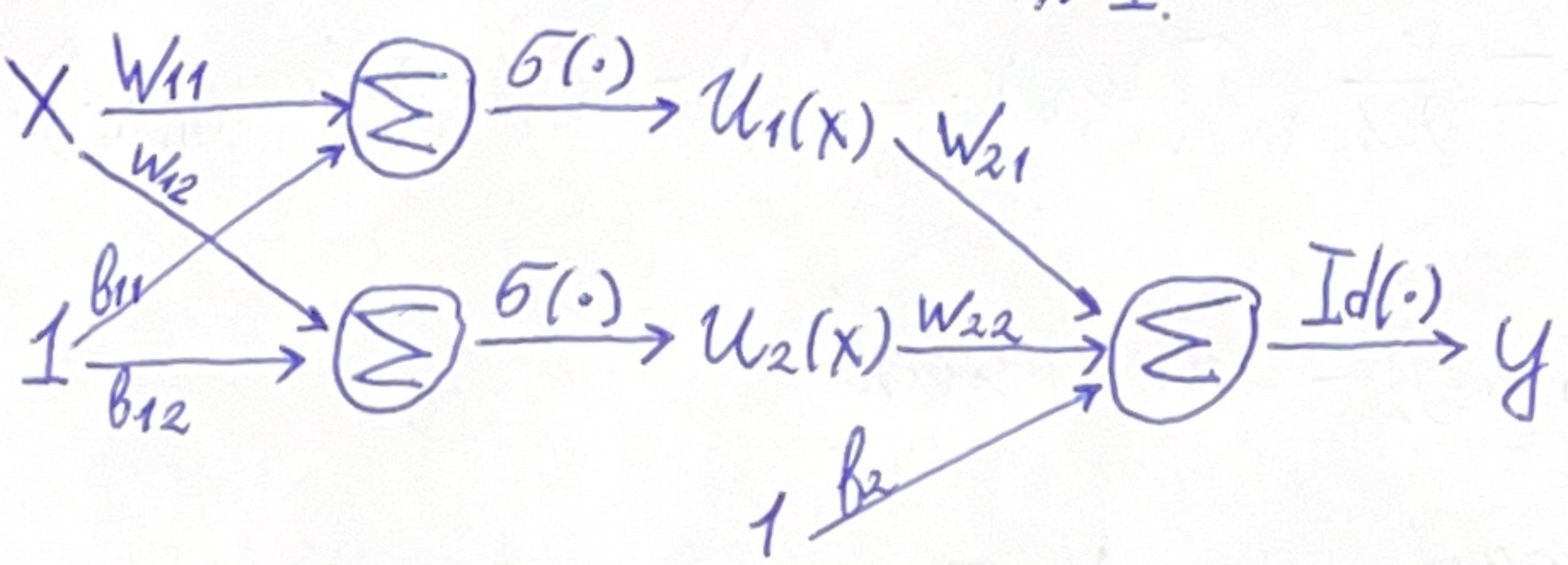
Задача 1

№1.



где $U = \begin{pmatrix} U_1(x) \\ U_2(x) \end{pmatrix}$

$Id(\cdot)$ - тождественное преобразование (ф-ия)

Как видно из схемы, у ~~нас~~ модели 7 параметров: смещения (b_{12}, b_{11}, b_2) и веса ($W_{11}, W_{12}, W_{21}, W_{22}$).

№2.

① Начнём с производной MSE по выходам сети:

$$\frac{\partial MSE}{\partial \hat{y}(X_i)} = \frac{2}{n}\sum_{i=1}^{n}\left(\hat{y}(X_i) - Y_i\right)$$

② Возьмём производные выходов сети по входам посл. слоя:

$$\frac{\partial \hat{y}(X_i)}{\partial w_{2h}} = \frac{\partial}{\partial w_{2h}}\left(\sum_{h=1}^{2} w_{2h}\cdot u_h(X_i) + b_2\right) = u_h(X_i); \quad \frac{\partial \hat{y}(X_i)}{\partial b_2} = 1$$

Также, выпишем производные выходов сети по входам последнего слоя: $\frac{\partial \hat{y}(X_i)}{\partial u_h(X_i)} = w_{2h}$

Теперь выпишем производные MSE по параметрам и входам последнего слоя:

$$\frac{\partial MSE}{\partial w_{2h}} = \sum_{i=1}^{n} \frac{\partial MSE}{\partial \hat{y}(X_i)} \cdot \frac{\partial \hat{y}(X_i)}{\partial w_{2h}}$$

$$\frac{\partial MSE}{\partial b_2} = \sum_{i=1}^{n} \frac{\partial MSE}{\partial \hat{y}(X_i)} \cdot \frac{\partial \hat{y}(X_i)}{\partial b_2}$$

$$\frac{\partial MSE}{\partial u_h} = \sum_{i=1}^{n} \frac{\partial MSE}{\partial \hat{y}(X_i)} \cdot \frac{\partial \hat{y}(X_i)}{\partial u_h}$$

③ Теперь будем разбираться с производными по параметрам первого слоя. Для начала нам понадобится производная функции активации: $\sigma'(x) = -(1+e^{-x})^{-2}\cdot(-1)\cdot e^{-x} = (1+e^{-x})^{-2}\cdot(1+e^{-x}-1) =$
$= \sigma^2(x)\cdot(\sigma^{-1}(x)-1) = \sigma(x)\cdot(1-\sigma(x)) \Rightarrow \boxed{\sigma'(x) = \sigma(x)\cdot(1-\sigma(x))}$

Теперь возьмём производные выходов первого слоя по его параметрам:

$$\cancel{\frac{\partial MSE}{\partial w_{1h}} = \sum_{i=1}^{n}}$$

$$\frac{\partial u_h(x_i)}{\partial w_{1h}} = \frac{\partial}{\partial w_{1h}}\left(\sigma(w_{1h}x_i + b_{1h})\right) = u_h(x_i)\cdot(1-u_h(x_i))\cdot x_i$$

$$\frac{\partial u_h(x_i)}{\partial b_{1h}} = u_h(x_i)\cdot(1-u_h(x_i))$$ – аналогично

Наконец, вычислим производные MSE по параметрам первого слоя:

$$\frac{\partial MSE}{\partial w_{1h}} = \sum_{i=1}^{n}\frac{\partial MSE}{\partial u_h(x_i)}\cdot\frac{\partial u_h(x_i)}{\partial w_{1h}}$$

$$\frac{\partial MSE}{\partial b_{1h}} = \sum_{i=1}^{n}\frac{\partial MSE}{\partial u_h(x_i)}\cdot\frac{\partial u_h(x_i)}{\partial b_{1h}}$$

№3.

Пусть обучающая ~~выборка~~ выборка очень большая. Тогда можно, например, использовать стохастический градиентный спуск. А именно,

1. Разбиваем данные на блоки (батчи)
2. Для каждого блока считаем градиент и обновляем параметры.

$$\cancel{\theta_{t+1} = \theta_t - \eta\cdot\nabla L_t,\ \text{где } L = \frac{1}{B}\cdot\sum_{b=1}^{B}\mathcal{L}(\hat{y}_{\theta_t}(x_{i_b}}$$

$$\theta_{t+1} = \theta_t - \eta\cdot\nabla L,\ \text{где } L = \frac{1}{B}\cdot\sum_{b=1}^{B}\mathcal{L}(\hat{y}_{\theta_t}(x_{i_b}), y_{i_b}),$$ где B – размер блока

$(x_1,\dots,x_b)$ – текущий блок.