# 1 Loss

$$L(W,b)=\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\lambda\|W\|_F^2\right) \tag{1}$$

where $W$ and $b$ are the weights and biases in the network, $\lambda$ the regularization strength.

# 2 Activation

## 2.1 Hyperbolic tangent (tanh)

$$\tanh(x)=\frac{e^x-e^{-x}}{e^x+e^{-x}} \tag{2}$$

$$\tanh'(x)=\frac{4}{(e^x+e^{-x})^2} \tag{3}$$

# 3 Example

We use the following example to test this implementation.

$$y(k+1)=\underbrace{\frac{1.5y(k)y(k-1)}{1+y^2(k)+y^2(k-1)}+\sin\big(y(k)+y(k-1)\big)+0.8u(k-1)}_{N_1}+\underbrace{\cos\big(y(k)-u(k-1)\big)}_{N_2}u(k) \tag{4}$$

The regressor vector is thus

$$x(k)=\begin{bmatrix}y(k) & y(k-1) & u(k-1)\end{bmatrix} \tag{5}$$

# 4 Gradient descent

## 4.1 Naive stochastic gradient descent

$$w \leftarrow w-\epsilon\,\Delta_w$$

where $\epsilon$ is the learning rate.

## 4.2 Momentum

$$v \leftarrow \gamma v-\epsilon\Delta_w$$

$$w \leftarrow w+v$$

where $\gamma$ is the forgetting factor leading to exponential averaging.