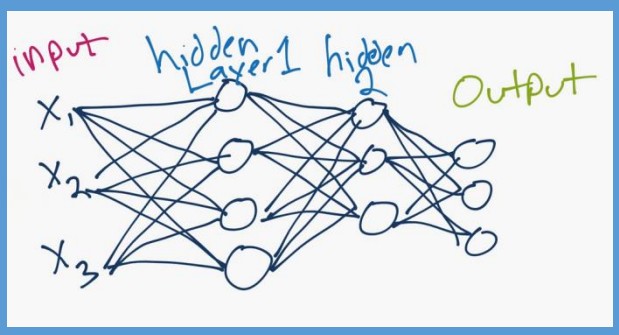
EMS702P Statistical Thinking and Applied Machine Learning

Week 9.2 – Neural Networks Training

Yunpeng Zhu



Neural Networks

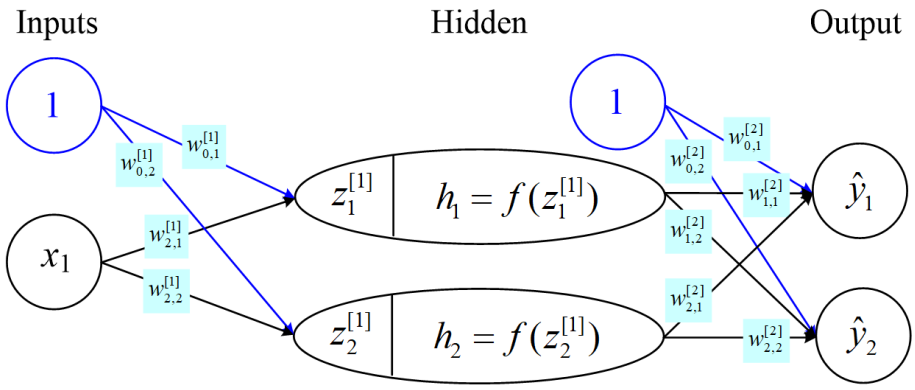
Edition: v1.1

Table of Contents

1 Neural Networks training

1.1 Neural Networks – Regression

BPNN Illustrate_122_GD.py



Labels:

$$(x_1, y_1, y_2) = \{(1,3,4),(2,6,5)\}$$

Activation function (sigmoid):

$$f(z) = \frac{1}{1+\exp(-z)}$$

Initialization:

$$w_{0,1}^{[1]} = 0.1,\ w_{0,2}^{[1]} = 0.1,\ w_{1,1}^{[1]} = 0.1,\ w_{1,2}^{[1]} = 0.1$$

$$w_{0,1}^{[2]} = 0.1,\ w_{0,2}^{[2]} = 0.1,\ w_{1,1}^{[2]} = 0.1,\ w_{1,2}^{[2]} = 0.1,\ w_{2,1}^{[2]} = 0.1,\ w_{2,2}^{[2]} = 0.1$$

and $\lambda = 0.5$

Cost function:

$$\bar{J} = \frac{1}{2}\sum_{t=1}^{2} J_{(t)} = \frac{1}{2}\sum_{t=1}^{2}\sum_{k=1}^{2} J_{k,(t)} = \frac{1}{2}\sum_{t=1}^{2}\sum_{k=1}^{2}\frac{1}{2}\left(\hat{y}_{k,(t)} - y_{k,(t)}\right)^2$$

1) Feed-forward calculation

$$\begin{cases} z_{1,(1)}^{[1]} = w_{0,1}^{[1]} + w_{1,1}^{[1]} x_{1,(1)} = 0.2 \\ z_{2,(1)}^{[1]} = w_{0,2}^{[1]} + w_{1,2}^{[1]} x_{1,(1)} = 0.2 \end{cases}, \begin{cases} z_{1,(2)}^{[1]} = w_{0,1}^{[1]} + w_{1,1}^{[1]} x_{1,(2)} = 0.3 \\ z_{2,(2)}^{[1]} = w_{0,2}^{[1]} + w_{1,2}^{[1]} x_{1,(2)} = 0.3 \end{cases}$$

$$\begin{cases} h_{1,(1)} = f(z^{[1]}_{1,(1)}) = 0.55 \\ h_{2,(1)} = f(z^{[1]}_{2,(1)}) = 0.55 \end{cases}, \begin{cases} h_{1,(2)} = f(z^{[1]}_{1,(2)}) = 0.574 \\ h_{2,(2)} = f(z^{[1]}_{2,(2)}) = 0.574 \end{cases}$$

$$\begin{cases} h_{1,(1)} = f(z^{[1]}_{1,(1)}) = 0.55 \\ h_{2,(1)} = f(z^{[1]}_{2,(1)}) = 0.55 \end{cases}, \begin{cases} h_{1,(2)} = f(z^{[1]}_{1,(2)}) = 0.574 \\ h_{2,(2)} = f(z^{[1]}_{2,(2)}) = 0.574 \end{cases}$$

$$\begin{cases} \begin{cases} \hat{y}_{1,(1)} = w^{[2]}_{0,1} + w^{[2]}_{1,1} h_{1,(1)} + w^{[2]}_{2,1} h_{2,(1)} = 0.21 \\ \hat{y}_{2,(1)} = w^{[2]}_{0,2} + w^{[2]}_{1,2} h_{1,(1)} + w^{[2]}_{2,2} h_{2,(1)} = 0.21 \end{cases} \\ \begin{cases} \hat{y}_{1,(2)} = w^{[2]}_{0,1} + w^{[2]}_{1,1} h_{1,(2)} + w^{[2]}_{2,1} h_{2,(2)} = 0.215 \\ \hat{y}_{2,(2)} = w^{[2]}_{0,2} + w^{[2]}_{1,2} h_{1,(2)} + w^{[2]}_{2,2} h_{2,(2)} = 0.215 \end{cases} \end{cases}$$

$$\overline{J} = \frac{1}{2}\sum_{t=1}^{2}\sum_{k=1}^{2}\frac{1}{2}\left(\hat{y}_{k,(t)} - y_{k,(t)}\right)^2 = 19.628$$

2) Backward propagation

Definition:

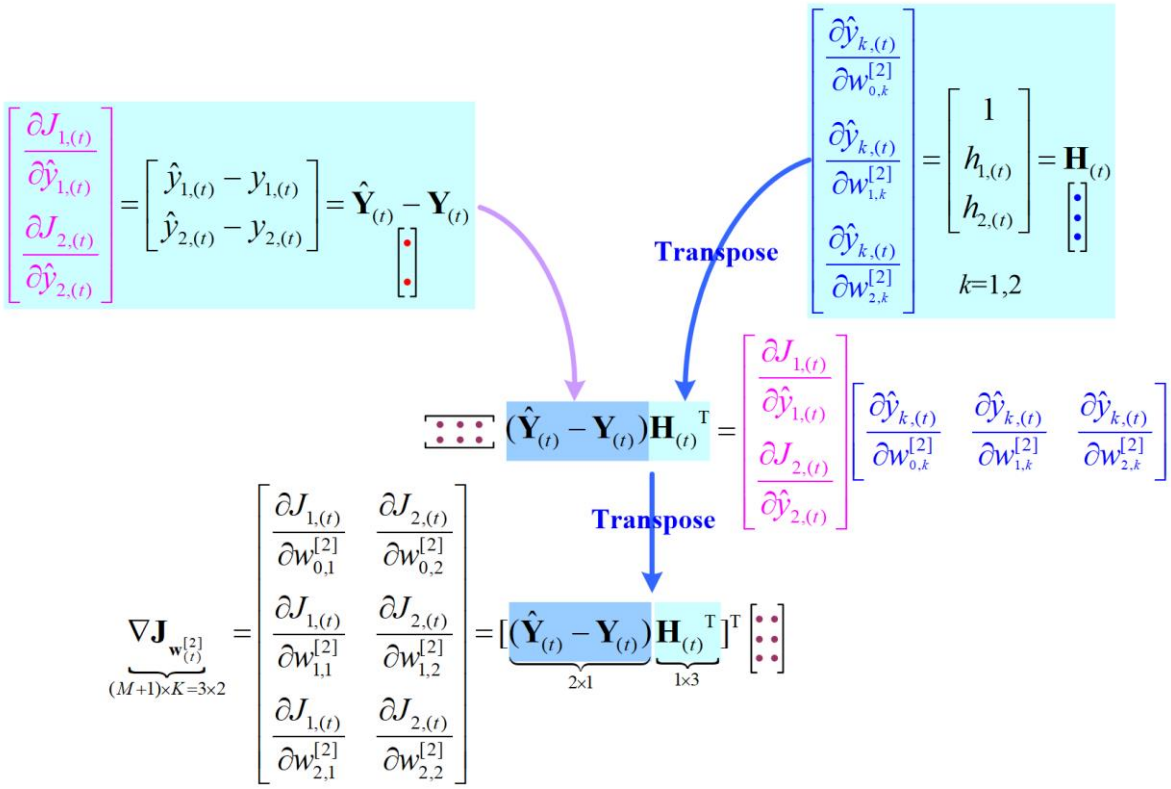
$$\underbrace{\mathbf{W}^{[n]}}_{(M+1)\times K} = \begin{bmatrix} w^{[n]}_{0,1} & \cdots & w^{[n]}_{0,K} \\ \vdots & \ddots & \vdots \\ w^{[n]}_{M+1,1} & \cdots & w^{[n]}_{M+1,K} \end{bmatrix}, n = 1,2$$

a) Output layer gradient

$$\underbrace{\nabla \mathbf{J}_{\mathbf{w}^{[2]}_{(t)}}}_{(M+1)\times K=3\times 2} = \begin{bmatrix} \frac{\partial J_{1,(t)}}{\partial w^{[2]}_{0,1}} & \frac{\partial J_{2,(t)}}{\partial w^{[2]}_{0,2}} \\ \frac{\partial J_{1,(t)}}{\partial w^{[2]}_{1,1}} & \frac{\partial J_{2,(t)}}{\partial w^{[2]}_{1,2}} \\ \frac{\partial J_{1,(t)}}{\partial w^{[2]}_{2,1}} & \frac{\partial J_{2,(t)}}{\partial w^{[2]}_{2,2}} \end{bmatrix} = [\underbrace{(\hat{\mathbf{Y}}_{(t)} - \mathbf{Y}_{(t)})}_{2\times 1} \underbrace{\mathbf{H}_{(t)}}_{1\times 3}{}^{\mathrm{T}}]^{\mathrm{T}}$$

where

$$\hat{\mathbf{Y}}_{(t)} - \mathbf{Y}_{(t)} = \begin{bmatrix} \hat{y}_{1,(t)} - y_{1,(t)} \\ \hat{y}_{2,(t)} - y_{2,(t)} \end{bmatrix},\ \mathbf{H}_{(t)} = \begin{bmatrix} 1 \\ h_{1,(t)} \\ h_{2,(t)} \end{bmatrix}$$



$$\nabla \mathbf{J}_{\mathbf{w}_{(t)}^{[2]}} = \underbrace{\begin{bmatrix} \frac{\partial J_{1,(t)}}{\partial w_{0,1}^{[2]}} & \frac{\partial J_{2,(t)}}{\partial w_{0,2}^{[2]}} \\ \frac{\partial J_{1,(t)}}{\partial w_{1,1}^{[2]}} & \frac{\partial J_{2,(t)}}{\partial w_{1,2}^{[2]}} \\ \frac{\partial J_{1,(t)}}{\partial w_{2,1}^{[2]}} & \frac{\partial J_{2,(t)}}{\partial w_{2,2}^{[2]}} \end{bmatrix}}_{(M+1)\times K=3\times 2} = [\underbrace{(\hat{\mathbf{Y}}_{(t)} - \mathbf{Y}_{(t)})}_{2\times 1}\underbrace{\mathbf{H}_{(t)}^{\mathrm{T}}}_{1\times 3}]^{\mathrm{T}} \begin{bmatrix} \\ \\ \end{bmatrix}$$

i.e.

$$\nabla \mathbf{J}_{\mathbf{w}_{(1)}^{[2]}} = \begin{bmatrix} \frac{\partial J_{1,(1)}}{\partial w_{0,1}^{[2]}} & \frac{\partial J_{2,(1)}}{\partial w_{0,2}^{[2]}} \\ \frac{\partial J_{1,(1)}}{\partial w_{1,1}^{[2]}} & \frac{\partial J_{2,(1)}}{\partial w_{1,2}^{[2]}} \\ \frac{\partial J_{1,(1)}}{\partial w_{2,1}^{[2]}} & \frac{\partial J_{2,(1)}}{\partial w_{2,2}^{[2]}} \end{bmatrix},\ \begin{cases} \frac{\partial J_{1,(1)}}{\partial w_{0,1}^{[2]}} = \hat{y}_{1,(1)} - y_{1,(1)} = -2.79 \\ \frac{\partial J_{1,(1)}}{\partial w_{1,1}^{[2]}} = \left(\hat{y}_{1,(1)} - y_{1,(1)}\right) h_{1,(1)} = -1.534 \\ \frac{\partial J_{1,(1)}}{\partial w_{2,1}^{[2]}} = \left(\hat{y}_{1,(1)} - y_{1,(1)}\right) h_{2,(1)} = -1.534 \\ \frac{\partial J_{2,(1)}}{\partial w_{0,2}^{[2]}} = \hat{y}_{2,(1)} - y_{2,(1)} = -3.79 \\ \frac{\partial J_{2,(1)}}{\partial w_{1,2}^{[2]}} = \left(\hat{y}_{2,(1)} - y_{2,(1)}\right) h_{1,(1)} = -2.08 \\ \frac{\partial J_{2,(1)}}{\partial w_{2,2}^{[2]}} = \left(\hat{y}_{2,(1)} - y_{2,(1)}\right) h_{2,(1)} = -2.08 \end{cases}$$

$$\nabla \mathbf{J}_{\mathbf{w}_{(2)}^{[2]}} = \begin{bmatrix} \frac{\partial J_{1,(2)}}{\partial w_{0,1}^{[2]}} & \frac{\partial J_{2,(2)}}{\partial w_{0,2}^{[2]}} \\ \frac{\partial J_{1,(2)}}{\partial w_{1,1}^{[2]}} & \frac{\partial J_{2,(2)}}{\partial w_{1,2}^{[2]}} \\ \frac{\partial J_{1,(2)}}{\partial w_{2,1}^{[2]}} & \frac{\partial J_{2,(2)}}{\partial w_{2,2}^{[2]}} \end{bmatrix} = \begin{bmatrix} -2.79 & -1.534 \\ -1.534 & -3.79 \\ -2.08 & -2.08 \end{bmatrix}$$

b) Input layer gradient

$$\underbrace{\nabla \mathbf{J}_{\mathbf{w}_{(t)}^{[1]}}}_{(N+1)\times M=2\times 2} = \begin{bmatrix} \frac{\partial J_{1,(t)}}{\partial w_{0,1}^{[1]}} + \frac{\partial J_{2,(t)}}{\partial w_{0,1}^{[1]}} & \frac{\partial J_{1,(t)}}{\partial w_{0,2}^{[1]}} + \frac{\partial J_{2,(t)}}{\partial w_{0,2}^{[1]}} \\ \frac{\partial J_{1,(t)}}{\partial w_{1,1}^{[1]}} + \frac{\partial J_{2,(t)}}{\partial w_{1,1}^{[1]}} & \frac{\partial J_{1,(t)}}{\partial w_{1,2}^{[1]}} + \frac{\partial J_{2,(t)}}{\partial w_{1,2}^{[1]}} \end{bmatrix}$$

$$= \left(\{ [\underbrace{(\hat{\mathbf{Y}}_{(t)} - \mathbf{Y}_{(t)})^{\mathrm{T}}}_{1\times 2} \underbrace{(\bar{\mathbf{W}}^{[2]})^{\mathrm{T}}}_{2\times 2}]^{\mathrm{T}} \odot \underbrace{\mathbf{H}_{\mathrm{d},(t)}}_{2\times 1} \} \underbrace{\mathbf{X}_{(t)}^{\mathrm{T}}}_{1\times 2} \right)^{\mathrm{T}}$$

$$\frac{\partial J}{\partial \hat{y}} \quad \frac{\partial \hat{y}}{\partial h} \quad \frac{\partial h}{\partial z^{[1]}} \quad \frac{\partial z^{[1]}}{\partial w^{[1]}}$$

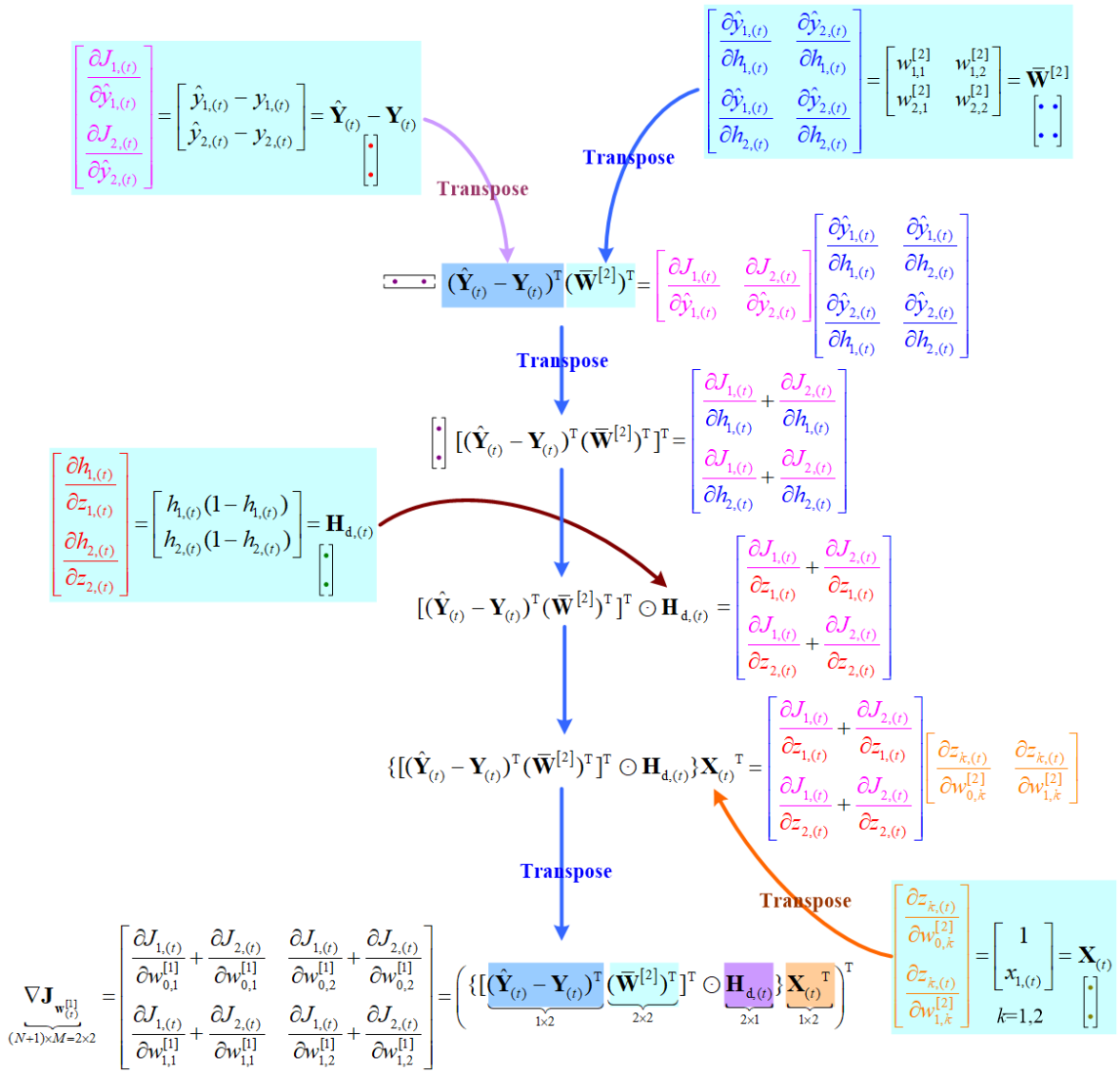
where

$$\hat{\mathbf{Y}}_{(t)} - \mathbf{Y}_{(t)} = \begin{bmatrix} \hat{y}_{1,(t)} - y_{1,(t)} \\ \hat{y}_{2,(t)} - y_{2,(t)} \end{bmatrix}, \quad \underset{M\times K=2\times 2}{\bar{\mathbf{W}}^{[2]}} = \begin{bmatrix} w_{1,1}^{[2]} & w_{1,2}^{[2]} \\ w_{2,1}^{[2]} & w_{2,2}^{[2]} \end{bmatrix}, \quad \mathbf{H}_{\mathrm{d},(t)} = \begin{bmatrix} h_{1,(t)}(1-h_{1,(t)}) \\ h_{2,(t)}(1-h_{2,(t)}) \end{bmatrix},$$

$$\mathbf{X}_{(t)} = \begin{bmatrix} 1 \\ x_{1,(t)} \end{bmatrix}$$

and $\odot$ is the hadamard product,

$$\begin{bmatrix} a_{11} & \cdots & a_{1,m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{n,n} \end{bmatrix} \odot \begin{bmatrix} b_{11} & \cdots & b_{1,m} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{n,n} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & \cdots & a_{1,m}b_{1,m} \\ \vdots & \ddots & \vdots \\ a_{n1}b_{n1} & \cdots & b_{n,n}b_{n,n} \end{bmatrix}$$



i.e.

$$\nabla \mathbf{J}_{\mathbf{w}^{[1]}_{(1)}} = \begin{bmatrix} \frac{\partial J_{1,(1)}}{\partial w^{[1]}_{0,1}} + \frac{\partial J_{2,(1)}}{\partial w^{[1]}_{0,1}} & \frac{\partial J_{1,(1)}}{\partial w^{[1]}_{0,2}} + \frac{\partial J_{2,(1)}}{\partial w^{[1]}_{0,2}} \\ \frac{\partial J_{1,(1)}}{\partial w^{[1]}_{1,1}} + \frac{\partial J_{2,(1)}}{\partial w^{[1]}_{1,1}} & \frac{\partial J_{1,(1)}}{\partial w^{[1]}_{1,2}} + \frac{\partial J_{2,(1)}}{\partial w^{[1]}_{1,2}} \end{bmatrix} = \begin{bmatrix} -0.163 & -0.163 \\ -0.163 & -0.163 \end{bmatrix}$$
$$(N+1)\times M = 2\times 2$$

$$\nabla \mathbf{J}_{\mathbf{w}^{[1]}_{(2)}} = \begin{bmatrix} \frac{\partial J_{1,(2)}}{\partial w^{[1]}_{0,1}} + \frac{\partial J_{2,(2)}}{\partial w^{[1]}_{0,1}} & \frac{\partial J_{1,(2)}}{\partial w^{[1]}_{0,2}} + \frac{\partial J_{2,(2)}}{\partial w^{[1]}_{0,2}} \\ \frac{\partial J_{1,(2)}}{\partial w^{[1]}_{1,1}} + \frac{\partial J_{2,(2)}}{\partial w^{[1]}_{1,1}} & \frac{\partial J_{1,(2)}}{\partial w^{[1]}_{1,2}} + \frac{\partial J_{2,(2)}}{\partial w^{[1]}_{1,2}} \end{bmatrix} = \begin{bmatrix} -0.163 & -0.163 \\ -0.163 & -0.163 \end{bmatrix}$$
$$(N+1)\times M = 2\times 2$$

3) Gradient descent method (or Batch Gradient Descent method [1])

$$\begin{cases} \mathbf{W}_{\text{new}}^{[2]} = \mathbf{W}_{\text{old}}^{[2]} - \lambda \dfrac{1}{T} \displaystyle\sum_{t=1}^{T} \nabla \mathbf{J}_{\mathbf{w}_{(t)}^{[2]}} \\ \mathbf{W}_{\text{new}}^{[1]} = \mathbf{W}_{\text{old}}^{[1]} - \lambda \dfrac{1}{T} \displaystyle\sum_{t=1}^{T} \nabla \mathbf{J}_{\mathbf{w}_{(t)}^{[1]}} \end{cases}$$

where

$$\underset{(M+1)\times K=3\times 2}{\mathbf{W}^{[2]}} = \begin{bmatrix} w_{0,1}^{[2]} & w_{0,2}^{[2]} \\ w_{1,1}^{[2]} & w_{1,2}^{[2]} \\ w_{2,1}^{[2]} & w_{2,2}^{[2]} \end{bmatrix}, \quad \underset{(N+1)\times M=2\times 2}{\mathbf{W}^{[1]}} = \begin{bmatrix} w_{0,1}^{[1]} & w_{0,2}^{[1]} \\ w_{1,1}^{[1]} & w_{1,2}^{[1]} \end{bmatrix}$$

As a result, there is

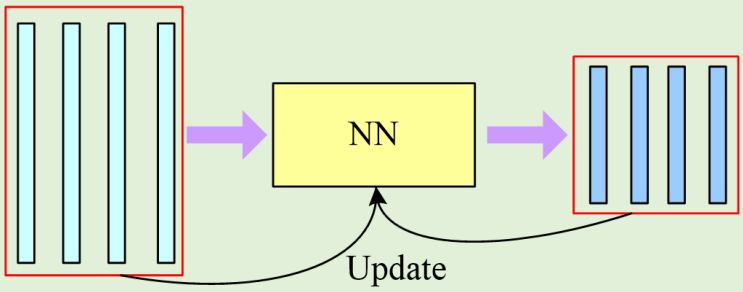
$$\mathbf{W}_{\text{new}}^{[2]} = \begin{bmatrix} 2.244 & 2.244 \\ 1.314 & 1.308 \\ 1.314 & 1.308 \end{bmatrix}, \mathbf{W}_{\text{new}}^{[1]} = \begin{bmatrix} 0.205 & 0.205 \\ 0.270 & 0.270 \end{bmatrix}$$

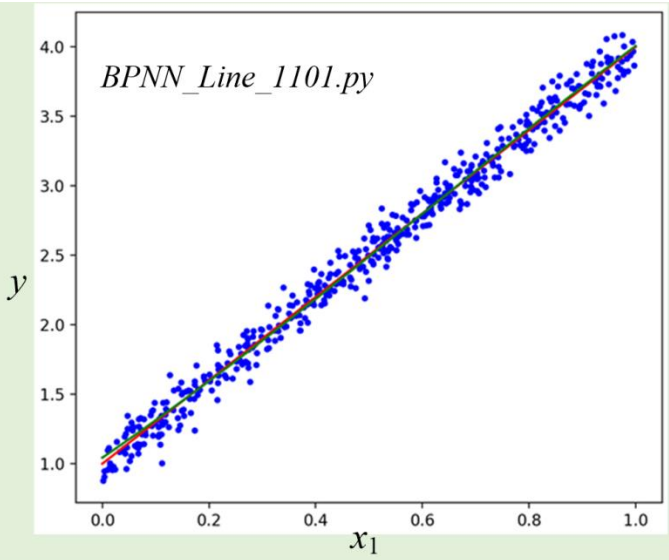
IT class (Python code):

BPNN_Line_1101_GD.py and BPNN_Sine_181_SGD.py on QM+

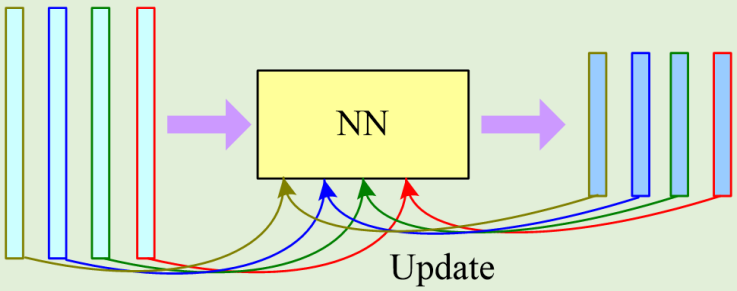
Example: Application of Back-Propagation Neural Networks

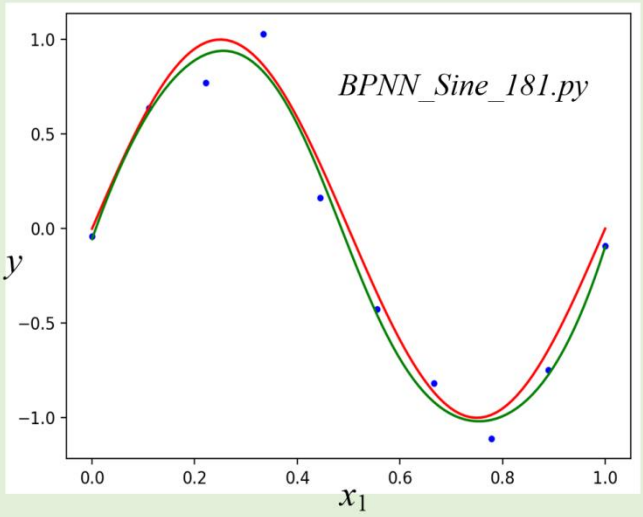
- NN regression of a straight line (Gradient descent):
  Net: 1 input, 10 hidden, 1 output, with 2 bias nodes





- NN regression of a sine wave (Stochastic gradient descent):
  Net: 1 input, 8 hidden, 1 output, with 2 bias nodes

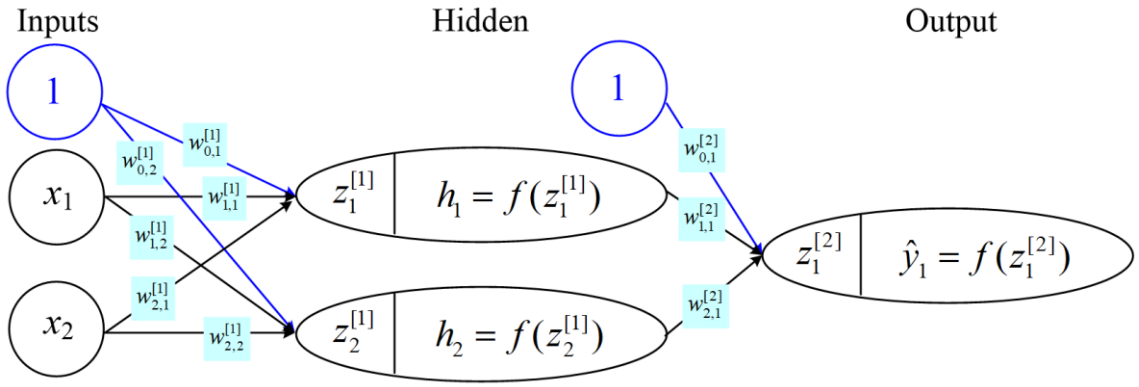




1.2 Neural Networks – Classification

Consider a binary classification problem,

$$y_1 = \begin{cases} 1 & \text{True} \\ 0 & \text{False} \end{cases}$$



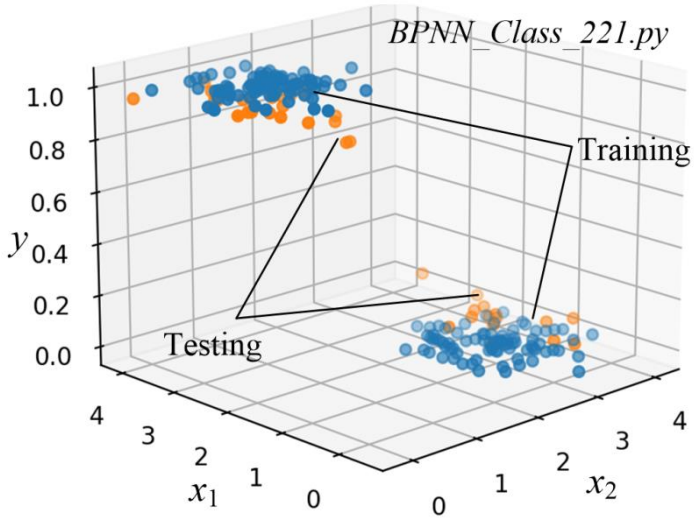
where $f(z) = \dfrac{1}{1+\exp(-z)}$.

For classification, denote the cost function as:

$$J(\mathbf{W}) = -\sum_{t=1}^{T} \left\{ y_{1,(t)} \ln \hat{y}_{1,(t)} + \left[1 - y_{1,(t)}\right] \ln\left[1 - \hat{y}_{1,(t)}\right] \right\}$$

$$\frac{\partial J_{(t)}}{\partial w_{m,1}^{[2]}} = \frac{\partial J_{(t)}}{\partial \hat{y}_1} \frac{\partial \hat{y}_1}{\partial z_1^{[2]}} \frac{\partial z_1^{[2]}}{\partial w_{m,1}^{[2]}} = \begin{cases} \dfrac{\hat{y}_1 - y_1}{\hat{y}_1(1-\hat{y}_1)} \hat{y}_1(1-\hat{y}_1) = \hat{y}_1 - y_1 & m = 0 \\ \dfrac{\hat{y}_1 - y_1}{\hat{y}_1(1-\hat{y}_1)} \hat{y}_1(1-\hat{y}_1) h_m = (\hat{y}_1 - y_1) h_m & m = 1,2 \end{cases}$$

$$\frac{\partial J_{(t)}}{\partial w_{n,m}^{[1]}} = \frac{\partial J_{(t)}}{\partial \hat{y}_1} \frac{\partial \hat{y}_1}{\partial z_1^{[2]}} \frac{\partial z_1^{[2]}}{\partial h_m} \frac{\partial h_m}{\partial z_m^{[1]}} \frac{\partial z_m^{[1]}}{\partial w_{n,m}^{[1]}} = \begin{cases} (\hat{y}_1 - y_1) w_{m,1}^{[2]} h_m (1-h_m) & n = 0 \\ (\hat{y}_1 - y_1) w_{m,1}^{[2]} h_m (1-h_m) x_n & n = 1,2 \end{cases}, \; m = 1,2$$



Summarize:

a) Output layer gradient

$$\underset{(M+1)\times K}{\nabla \mathbf{J}_{\mathbf{w}_{(t)}^{[2]}}} = [\underbrace{(\hat{\mathbf{Y}}_{(t)} - \mathbf{Y}_{(t)})}_{K\times 1} \underbrace{\mathbf{H}_{(t)}{}^{\mathrm{T}}}_{1\times(M+1)}]^{\mathrm{T}}, \quad \frac{\partial J}{\partial y}\ \frac{\partial y}{\partial w^{[2]}}$$

b) Input layer gradient

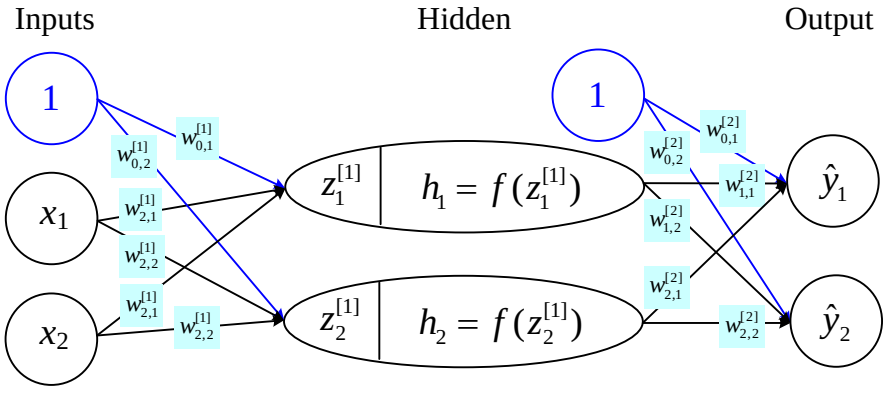
$$\underset{(N+1)\times M}{\nabla \mathbf{J}_{\mathbf{w}_{(t)}^{[1]}}} = \left(\{ [\underbrace{(\hat{\mathbf{Y}}_{(t)} - \mathbf{Y}_{(t)})^{\mathrm{T}}}_{1\times K} \underbrace{(\bar{\mathbf{W}}^{[2]})^{\mathrm{T}}}_{K\times M}]^{\mathrm{T}} \odot \underbrace{\mathbf{H}_{\mathrm{d},(t)}}_{M\times 1} \} \underbrace{\mathbf{X}_{(t)}{}^{\mathrm{T}}}_{1\times(N+1)} \right)^{\mathrm{T}}, \quad \frac{\partial J}{\partial y}\ \frac{\partial y}{\partial h}\ \frac{\partial h}{\partial z^{[1]}}\ \frac{\partial z^{[1]}}{\partial w^{[1]}}$$

c) Gradient descent method

$$\begin{cases} \mathbf{W}_{\text{new}}^{[2]} = \mathbf{W}_{\text{old}}^{[2]} - \lambda \dfrac{1}{T} \displaystyle\sum_{t=1}^{T} \nabla \mathbf{J}_{\mathbf{w}_{(t)}^{[2]}} \\ \mathbf{W}_{\text{new}}^{[1]} = \mathbf{W}_{\text{old}}^{[1]} - \lambda \dfrac{1}{T} \displaystyle\sum_{t=1}^{T} \nabla \mathbf{J}_{\mathbf{w}_{(t)}^{[1]}} \end{cases}$$

Quiz 1.1: (group work)

Derivate the BP training of the neural network



where $f(z) = \dfrac{1}{1+\exp(-z)}$, and the cost function is:

$$J = \frac{1}{3}\sum_{t=1}^{3}\sum_{k=1}^{2}\frac{1}{2}\left(\hat{y}_{k,(t)} - y_{k,(t)}\right)^2$$

Denote the weight matrices are:

$$\underset{(M+1)\times K=3\times 2}{\mathbf{W}^{[2]}} = \begin{bmatrix} w_{0,1}^{[2]} & w_{0,2}^{[2]} \\ w_{1,1}^{[2]} & w_{1,2}^{[2]} \\ w_{2,1}^{[2]} & w_{2,2}^{[2]} \end{bmatrix}, \quad \underset{(N+1)\times M=3\times 2}{\mathbf{W}^{[1]}} = \begin{bmatrix} w_{0,1}^{[1]} & w_{0,2}^{[1]} \\ w_{1,1}^{[1]} & w_{1,2}^{[1]} \\ w_{2,1}^{[1]} & w_{2,2}^{[1]} \end{bmatrix}$$

a) Output layer gradient

For the t th training data set,

$$\begin{bmatrix} \frac{\partial J_{1,(t)}}{\partial y_{1,(t)}} \\ \frac{\partial J_{2,(t)}}{\partial y_{2,(t)}} \end{bmatrix} = \hat{\mathbf{Y}}_{(t)} - \mathbf{Y}_{(t)} = \begin{bmatrix} \hat{y}_{1,(t)} - y_{1,(t)} \\ \hat{y}_{2,(t)} - y_{2,(t)} \end{bmatrix}, \ \mathbf{H}_{(t)} = \begin{bmatrix} \frac{\partial y_{(t)}}{\partial w_{0,1}^{[2]}} \\ \frac{\partial y_{(t)}}{\partial w_{1,1}^{[2]}} \\ \frac{\partial y_{(t)}}{\partial w_{2,1}^{[2]}} \end{bmatrix} = \begin{bmatrix} 1 \\ h_{1,(t)} \\ h_{2,(t)} \end{bmatrix}$$

$$\nabla \mathbf{J}_{\mathbf{w}_{(t)}^{[2]}} = \begin{bmatrix} \frac{\partial J_{1,(t)}}{\partial w_{0,1}^{[2]}} & \frac{\partial J_{2,(t)}}{\partial w_{0,2}^{[2]}} \\ \frac{\partial J_{1,(t)}}{\partial w_{1,1}^{[2]}} & \frac{\partial J_{2,(t)}}{\partial w_{1,2}^{[2]}} \\ \frac{\partial J_{1,(t)}}{\partial w_{2,1}^{[2]}} & \frac{\partial J_{2,(t)}}{\partial w_{2,2}^{[2]}} \end{bmatrix} = [(\hat{\mathbf{Y}}_{(t)} - \mathbf{Y}_{(t)})\mathbf{H}_{(t)}{}^{\mathrm{T}}]^{\mathrm{T}}$$

$$= \begin{bmatrix} \hat{y}_{1,(t)} - y_{1,(t)} & \hat{y}_{2,(t)} - y_{2,(t)} \\ (\hat{y}_{1,(t)} - y_{1,(t)})h_{1,(t)} & (\hat{y}_{2,(t)} - y_{2,(t)})h_{1,(t)} \\ (\hat{y}_{1,(t)} - y_{1,(t)})h_{2,(t)} & (\hat{y}_{2,(t)} - y_{2,(t)})h_{2,(t)} \end{bmatrix}$$

b) Input layer gradient

$$\begin{bmatrix} \frac{\partial J_{1,(t)}}{\partial y_{1,(t)}} \\ \frac{\partial J_{2,(t)}}{\partial y_{2,(t)}} \end{bmatrix} = \hat{\mathbf{Y}}_{(t)} - \mathbf{Y}_{(t)} = \begin{bmatrix} \hat{y}_{1,(t)} - y_{1,(t)} \\ \hat{y}_{2,(t)} - y_{2,(t)} \end{bmatrix}, \begin{bmatrix} \frac{\partial y_{1,(t)}}{\partial h_{1,(t)}} & \frac{\partial y_{2,(t)}}{\partial h_{1,(t)}} \\ \frac{\partial y_{1,(t)}}{\partial h_{2,(t)}} & \frac{\partial y_{2,(t)}}{\partial h_{2,(t)}} \end{bmatrix} = \bar{\mathbf{W}}^{[2]} = \begin{bmatrix} w_{1,1}^{[2]} & w_{1,2}^{[2]} \\ w_{2,1}^{[2]} & w_{2,2}^{[2]} \end{bmatrix}$$

Then

$$[(\hat{\mathbf{Y}}_{(t)} - \mathbf{Y}_{(t)})^{\mathrm{T}}(\bar{\mathbf{W}}^{[2]})^{\mathrm{T}}]^{\mathrm{T}} = \bar{\mathbf{W}}^{[2]}(\hat{\mathbf{Y}}_{(t)} - \mathbf{Y}_{(t)}) = \begin{bmatrix} (\hat{y}_{1,(t)} - y_{1,(t)})w_{1,1}^{[2]} + (\hat{y}_{2,(t)} - y_{2,(t)})w_{1,2}^{[2]} \\ (\hat{y}_{1,(t)} - y_{1,(t)})w_{2,1}^{[2]} + (\hat{y}_{2,(t)} - y_{2,(t)})w_{2,2}^{[2]} \end{bmatrix}$$

Consider

$$\begin{bmatrix} \frac{\partial h_{1,(t)}}{\partial z_{1,(t)}} \\ \frac{\partial h_{2,(t)}}{\partial z_{2,(t)}} \end{bmatrix} = \mathbf{H}_{\mathrm{d},(t)} = \begin{bmatrix} h_{1,(t)}(1-h_{1,(t)}) \\ h_{2,(t)}(1-h_{2,(t)}) \end{bmatrix}$$

$$[(\hat{\mathbf{Y}}_{(t)} - \mathbf{Y}_{(t)})^{\mathrm{T}}(\bar{\mathbf{W}}^{[2]})^{\mathrm{T}}]^{\mathrm{T}} \odot \mathbf{H}_{\mathrm{d},(t)}$$

$$= \begin{bmatrix} [(\hat{y}_{1,(t)} - y_{1,(t)})w_{1,1}^{[2]} + (\hat{y}_{2,(t)} - y_{2,(t)})w_{1,2}^{[2]}]h_{1,(t)}(1-h_{1,(t)}) \\ [(\hat{y}_{1,(t)} - y_{1,(t)})w_{2,1}^{[2]} + (\hat{y}_{2,(t)} - y_{2,(t)})w_{2,2}^{[2]}]h_{2,(t)}(1-h_{2,(t)}) \end{bmatrix}$$

$$\begin{bmatrix} \frac{\partial z_{k,(t)}}{\partial w_{0,k}^{[1]}} \\ \frac{\partial z_{k,(t)}}{\partial w_{1,k}^{[1]}} \\ \frac{\partial z_{k,(t)}}{\partial w_{2,k}^{[1]}} \end{bmatrix} = \mathbf{X}_{(t)} = \begin{bmatrix} 1 \\ x_{1,(t)} \\ x_{2,(t)} \end{bmatrix}$$

$$\nabla \mathbf{J}_{\mathbf{w}_{(t)}^{[1]}} = \begin{bmatrix} \frac{\partial J_{1,(t)}}{\partial w_{0,1}^{[1]}} + \frac{\partial J_{2,(t)}}{\partial w_{0,1}^{[1]}} & \frac{\partial J_{1,(t)}}{\partial w_{0,2}^{[1]}} + \frac{\partial J_{2,(t)}}{\partial w_{0,2}^{[1]}} \\ \frac{\partial J_{1,(t)}}{\partial w_{1,1}^{[1]}} + \frac{\partial J_{2,(t)}}{\partial w_{1,1}^{[1]}} & \frac{\partial J_{1,(t)}}{\partial w_{1,2}^{[1]}} + \frac{\partial J_{2,(t)}}{\partial w_{1,2}^{[1]}} \\ \frac{\partial J_{1,(t)}}{\partial w_{2,1}^{[1]}} + \frac{\partial J_{2,(t)}}{\partial w_{2,1}^{[1]}} & \frac{\partial J_{1,(t)}}{\partial w_{2,2}^{[1]}} + \frac{\partial J_{2,(t)}}{\partial w_{2,2}^{[1]}} \end{bmatrix} = \{[(\hat{\mathbf{Y}}_{(t)} - \mathbf{Y}_{(t)})^{\mathrm{T}}(\bar{\mathbf{W}}^{[2]})^{\mathrm{T}}]^{\mathrm{T}} \odot \mathbf{H}_{\mathrm{d},(t)}\mathbf{X}_{(t)}{}^{\mathrm{T}}\}^{\mathrm{T}}$$

$$= \begin{bmatrix} [(\hat{y}_{1,(t)} - y_{1,(t)})w_{1,1}^{[2]} + (\hat{y}_{2,(t)} - y_{2,(t)})w_{1,2}^{[2]}]h_{1,(t)}(1-h_{1,(t)}) \\ [(\hat{y}_{1,(t)} - y_{1,(t)})w_{2,1}^{[2]} + (\hat{y}_{2,(t)} - y_{2,(t)})w_{2,2}^{[2]}]h_{2,(t)}(1-h_{2,(t)}) \end{bmatrix} \begin{bmatrix} 1 & x_{1,(t)} & x_{2,(t)} \end{bmatrix}$$

$$= \begin{bmatrix} [(\hat{y}_{1,(t)} - y_{1,(t)})w_{1,1}^{[2]} + (\hat{y}_{2,(t)} - y_{2,(t)})w_{1,2}^{[2]}] & [(\hat{y}_{1,(t)} - y_{1,(t)})w_{2,1}^{[2]} + (\hat{y}_{2,(t)} - y_{2,(t)})w_{2,2}^{[2]}] \\ \times h_{1,(t)}(1-h_{1,(t)}) & \times h_{2,(t)}(1-h_{2,(t)}) \\ [(\hat{y}_{1,(t)} - y_{1,(t)})w_{1,1}^{[2]} + (\hat{y}_{2,(t)} - y_{2,(t)})w_{1,2}^{[2]}] & [(\hat{y}_{1,(t)} - y_{1,(t)})w_{2,1}^{[2]} + (\hat{y}_{2,(t)} - y_{2,(t)})w_{2,2}^{[2]}] \\ \times h_{1,(t)}(1-h_{1,(t)})x_{1,(t)} & \times h_{2,(t)}(1-h_{2,(t)})x_{1,(t)} \\ [(\hat{y}_{1,(t)} - y_{1,(t)})w_{1,1}^{[2]} + (\hat{y}_{2,(t)} - y_{2,(t)})w_{1,2}^{[2]}] & [(\hat{y}_{1,(t)} - y_{1,(t)})w_{2,1}^{[2]} + (\hat{y}_{2,(t)} - y_{2,(t)})w_{2,2}^{[2]}] \\ \times h_{1,(t)}(1-h_{1,(t)})x_{2,(t)} & \times h_{2,(t)}(1-h_{2,(t)})x_{2,(t)} \end{bmatrix}$$

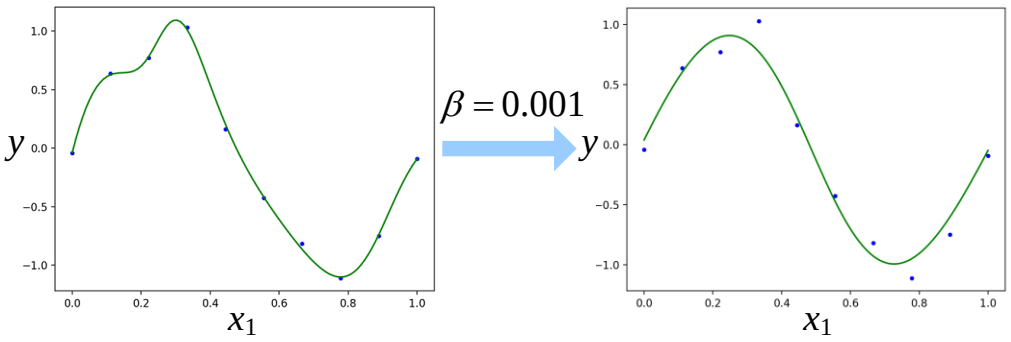
c) Gradient descent

$$\begin{cases} \mathbf{W}_{\text{new}}^{[2]} = \mathbf{W}_{\text{old}}^{[2]} - \lambda \dfrac{1}{3}\displaystyle\sum_{t=1}^{3} \nabla \mathbf{J}_{\mathbf{w}_{(t)}^{[2]}} \\ \mathbf{W}_{\text{new}}^{[1]} = \mathbf{W}_{\text{old}}^{[1]} - \lambda \dfrac{1}{3}\displaystyle\sum_{t=1}^{3} \nabla \mathbf{J}_{\mathbf{w}_{(t)}^{[1]}} \end{cases}$$

2 Regularization of neural networks

Consider the NN regression of a sine wave in the above example. If we apply 1-10-1 (1 input node, 10 hidden nodes, 1 output node) neural network, there is an over fitting problem.

BPNN_RegSine_1101_SDG.py



Similar to the regularization based LS method, the over fitting problem can be addressed by using the regularization. In the regularized Neural Network, the cost function for regression is

$$\tilde{J} = \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{k=1}^{K}\frac{1}{2}\left(\hat{y}_{k,(t)} - y_{k,(t)}\right)^2 + \frac{1}{2}\beta\left\|\mathbf{W}_{\text{All}}\right\|_2\right]$$

where $\mathbf{W}_{\text{All}} = \left[w_{0,1}^{[1]}, \ldots, w_{M,K}^{[2]}\right]^{\text{T}}$.

The gradients of the cost function become

$$\frac{\partial \tilde{J}_{(t)}}{\partial w_{i,j}^{[s]}} = \frac{\partial J_{(t)}}{\partial w_{i,j}^{[s]}} + \beta w_{i,j}^{[s]}, s = 1,2; i, j = 1,2,\ldots$$

For example, with the regularization parameter $\beta = 0.001$, the over fitting problem in the sine wave regression is addressed.

3 Further Readings

[1] Batch and stochastic gradient descent.

https://sebastianraschka.com/faq/docs/gradient-optimization.html