HW2

Juan Arroyo Miranda

Problem 1: Softmax

Using the definition of softmax, we know that $\hat{p}(y = c_1|\mathbf{x};\mathbf{W}) = \dfrac{exp(\mathbf{w_{c_1}} \cdot \mathbf{x})}{\sum_{y=1}^{2} exp(\mathbf{w_y} \cdot \mathbf{x})}$ and $\hat{p}(y = c_2|\mathbf{x};\mathbf{W}) = \dfrac{exp(\mathbf{w_{c_2}} \cdot \mathbf{x})}{\sum_{y=1}^{2} exp(\mathbf{w_y} \cdot \mathbf{x})}$. This allows to compute the log-odds in the as following:

$$
\begin{aligned}
log\left(\frac{\hat{p}(y = c_1|\mathbf{x};\mathbf{W})}{\hat{p}(y = c_2|\mathbf{x};\mathbf{W})}\right) &= log(\hat{p}(y = c_1|\mathbf{x};\mathbf{W}) - log(\hat{p}(y = c_2|\mathbf{x};\mathbf{W}) \\
&= log\left(\frac{exp(\mathbf{w_{c_1}} \cdot \mathbf{x})}{\sum_{y=1}^{2} exp(\mathbf{w_y} \cdot \mathbf{x})}\right) - log\left(\frac{exp(\mathbf{w_{c_2}} \cdot \mathbf{x})}{\sum_{y=1}^{2} exp(\mathbf{w_y} \cdot \mathbf{x})}\right) \\
&= \mathbf{w_{c_1}} \cdot \mathbf{x} - log\left(\sum_{y=1}^{2} exp(\mathbf{w_y} \cdot \mathbf{x})\right) - \mathbf{w_{c_2}} \cdot \mathbf{x} + log\left(\sum_{y=1}^{2} exp(\mathbf{w_y} \cdot \mathbf{x})\right) \\
&= \mathbf{w_{c_1}} \cdot \mathbf{x} - \mathbf{w_{c_2}} \cdot \mathbf{x} \\
&= \mathbf{w_{c_1}^T}\mathbf{x} - \mathbf{w_{c_2}^T}\mathbf{x} \\
&= (\mathbf{w_{c_1}^T} - \mathbf{w_{c_2}^T})\mathbf{x} \\
&= \mathbf{vx} \quad \text{where} \quad \mathbf{v} = \mathbf{w_{c_1}^T} - \mathbf{w_{c_2}^T}
\end{aligned}
$$

From this, we know that we can model the log-odds with the following linear function:

$$log\left(\frac{\hat{p}(y=c_1|\mathbf{x};\mathbf{W})}{\hat{p}(y=c_2|\mathbf{x};\mathbf{W})}\right) = \mathbf{w_{c_1}} \cdot \mathbf{x} - \mathbf{w_{c_2}} \cdot \mathbf{x}$$

$$\frac{\hat{p}(y=c_1|\mathbf{x};\mathbf{W})}{\hat{p}(y=c_2|\mathbf{x};\mathbf{W})} = exp(\mathbf{w_{c_1}} \cdot \mathbf{x} - \mathbf{w_{c_2}} \cdot \mathbf{x})$$

$$\frac{\hat{p}(y=c_1|\mathbf{x};\mathbf{W})}{1-\hat{p}(y=c_1|\mathbf{x};\mathbf{W})} = exp(\mathbf{w_{c_1}} \cdot \mathbf{x} - \mathbf{w_{c_2}} \cdot \mathbf{x})$$

$$\frac{1-\hat{p}(y=c_1|\mathbf{x};\mathbf{W})}{\hat{p}(y=c_1|\mathbf{x};\mathbf{W})} = \frac{1}{exp(\mathbf{w_{c_1}} \cdot \mathbf{x} - \mathbf{w_{c_2}} \cdot \mathbf{x})}$$

$$\frac{1}{\hat{p}(y=c_1|\mathbf{x};\mathbf{W})} = 1 + \frac{1}{exp(\mathbf{w_{c_1}} \cdot \mathbf{x} - \mathbf{w_{c_2}} \cdot \mathbf{x})}$$

$$\frac{1}{\hat{p}(y=c_1|\mathbf{x};\mathbf{W})} = \frac{1 + exp(\mathbf{w_{c_1}} \cdot \mathbf{x} - \mathbf{w_{c_2}} \cdot \mathbf{x})}{exp(\mathbf{w_{c_1}} \cdot \mathbf{x} - \mathbf{w_{c_2}} \cdot \mathbf{x})}$$

$$\hat{p}(y=c_1|\mathbf{x};\mathbf{W}) = \frac{exp(\mathbf{w_{c_1}} \cdot \mathbf{x} - \mathbf{w_{c_2}} \cdot \mathbf{x})}{1 + exp(\mathbf{w_{c_1}} \cdot \mathbf{x} - \mathbf{w_{c_2}} \cdot \mathbf{x})}$$

$$\hat{p}(y=c_1|\mathbf{x};\mathbf{W}) = \frac{exp(\mathbf{w_{c_1}} \cdot \mathbf{x})exp(-\mathbf{w_{c_2}} \cdot \mathbf{x})}{1 + exp(\mathbf{w_{c_1}} \cdot \mathbf{x})exp(-\mathbf{w_{c_2}} \cdot \mathbf{x})}$$

$$\hat{p}(y=c_1|\mathbf{x};\mathbf{W}) = \frac{exp(\mathbf{w_{c_1}} \cdot \mathbf{x})}{exp(\mathbf{w_{c_1}} \cdot \mathbf{x}) + exp(\mathbf{w_{c_2}} \cdot \mathbf{x})}$$

Problem 2

To show that the softmax model as stated in (1) is *overparametrized* we can write the probabilities as follows:

$$p(y=1|\mathbf{x};\mathbf{W}) = \frac{exp(\mathbf{w_1'})}{1+\sum_{y=1}^{C-1} exp(\mathbf{w_y'}\cdot\mathbf{x})}$$

$$\text{where} \quad exp(\mathbf{w_i'}) = exp(\mathbf{w_i})exp(-\mathbf{w_C}) \quad \forall i \neq C \quad \text{and}$$

$$exp(\mathbf{w_C'}) = exp(\mathbf{w_C})exp(-\mathbf{w_C}) = 1$$

$$p(y=2|\mathbf{x};\mathbf{W}) = \frac{exp(\mathbf{w_2'})}{1+\sum_{y=1}^{C-1} exp(\mathbf{w_y'}\cdot\mathbf{x})}$$

$$\vdots$$

$$p(y=C-1|\mathbf{x};\mathbf{W}) = \frac{exp(\mathbf{w_{C-1}'})}{1+\sum_{y=1}^{C-1} exp(\mathbf{w_y'}\cdot\mathbf{x})}$$

Using the fact that the probabilities from 1 to C must sum to one, we get:

$$p(y=C|\mathbf{x};\mathbf{W}) = 1-\sum_{c=1}^{C-1} p(y=c|\mathbf{x};\mathbf{W})$$

$$= 1-\frac{\sum_{c=1}^{C-1} exp(\mathbf{w_c}'\cdot\mathbf{x})}{1+\sum_{y=1}^{C-1} exp(\mathbf{w_y}'\cdot\mathbf{x})}$$

$$= \frac{1}{1+\sum_{c=1}^{C-1} exp(\mathbf{w_c'}\cdot\mathbf{x})}$$

This expression shows that we can write $p(y=c|\mathbf{x};\mathbf{W})$ for any c by using $C-1$ parameter vectors, where we interpret each vector parameter as a difference with respect to our base category.

Problem 3

In order to show that **H** is a positive definite matrix, we have to start with the solution of the maximum likelihood problem for the logistic regression

$$\underset{\mathbf{w}}{\operatorname{argmin}}\ log\, p(\mathbf{y}|\mathbf{X};\mathbf{w}) = -\sum_{i=1}^{d} y_i\, log\, \sigma(w_0 + \mathbf{w}\cdot\mathbf{x}_i) + (1-y_i)\, log(1-\sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i))$$

$$\text{where} \quad \sigma(w_0 + \mathbf{w}\cdot\mathbf{x}_i) = \frac{1}{1+exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)}$$

It will be helpful to compute first the following derivatives:

$$\begin{aligned}
\frac{\delta}{\delta w_0} log\, \sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i) &= \frac{1}{\sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i)}\frac{\delta}{\delta w_0}\sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i) \\
&= \frac{(1+exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i))^{-2} exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)}{(1+exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i))^{-1}} \\
&= \frac{exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)}{1+exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)}
\end{aligned}$$

$$\begin{aligned}
\frac{\delta}{\delta w_j} log\, \sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i) &= \frac{1}{\sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i)}\frac{\delta}{\delta w_j}\sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i) \\
&= \frac{exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)x_{ij}}{1+exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)}
\end{aligned}$$

and define $\gamma_i \equiv \dfrac{1}{1+exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)} = \sigma(w_0+\mathbf{w}\cdot\mathbf{x}_\mathrm{i})$.

Similarly,

$$\begin{aligned}
\frac{\delta}{\delta w_0} log\, (1-\sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i)) &= \frac{\delta}{\delta w_0} log \left(\frac{exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)}{1+exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)}\right) \\
&= \frac{\delta}{\delta w_0}\Big[-w_0-\mathbf{w}\cdot\mathbf{x}_i - log\,(1+exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i))\Big] \\
&= -1 - \frac{exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)}{1+exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)} \\
&= -\frac{1}{1+exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)} \\
&= -\gamma_i
\end{aligned}$$

$$\frac{\delta}{\delta w_j} log\left(1-\sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i)\right) = \frac{\delta}{\delta w_j} log\left(\frac{exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)}{1+exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)}\right)$$

$$= \frac{\delta}{\delta w_j}\left[-w_0-\mathbf{w}\cdot\mathbf{x}_i - log\left(1+exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)\right)\right]$$

$$= -x_{ij} - \frac{exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)}{1+exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)} x_{ij}$$

$$= -\frac{x_{ij}}{1+exp(-w_0-\mathbf{w}\cdot\mathbf{x}_i)}$$

$$= -\gamma_i x_{ij}$$

Now, taking the derivatives with respect to w_0 and w_j in our original function, we get:

$$\frac{\delta}{\delta w_0} log\, p(\mathbf{y}|\mathbf{X};\mathbf{w}) = -\sum_{i=1}^{d} y_i \frac{\delta}{\delta w_0} log\,\sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i) + (1-y_i)\frac{\delta}{\delta w_0} log(1-\sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i)) = 0$$

$$= -\sum_{i=1}^{d} y_i(1-\gamma_i) + (1-y_i)(-\gamma_i) = 0$$

$$= -\sum_{i=1}^{d} y_i - \gamma_i = 0$$

$$= \quad \sum_{i=1}^{d} \gamma_i - y_i = 0$$

$$\frac{\delta}{\delta w_j} log\, p(\mathbf{y}|\mathbf{X};\mathbf{w}) = -\sum_{i=1}^{d} y_i \frac{\delta}{\delta w_j} log\,\sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i) + (1-y_i)\frac{\delta}{\delta w_j} log(1-\sigma(w_0+\mathbf{w}\cdot\mathbf{x}_i)) = 0$$

$$= -\sum_{i=1}^{d} (y_i - \gamma_i) x_{ij} = 0$$

$$= \quad \sum_{i=1}^{d} (\gamma_i - y_i) x_{ij} = 0$$

We can rewrite all the derivates different from w_0 in matrix notation considering the following:

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \vdots \\ \gamma_d \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_d \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ x_{31} & x_{32} & \cdots & x_{3d} \\ \vdots & \ddots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dd} \end{pmatrix}$$

Therefore, the gradient for our problem becomes $\nabla L = \frac{\delta}{\delta \mathbf{w}} \, log \, p(\mathbf{y}|\mathbf{X};\mathbf{w}) = (\boldsymbol{\gamma} - \mathbf{y})^T \mathbf{X}$, which is equivalent to $\nabla L = \mathbf{X}^T(\boldsymbol{\gamma} - \mathbf{y})$.

Now, we want to compute the second derivatives with respect to some $w_k \neq w_j$. Before continuing, it will be useful to compute $\dfrac{\delta}{\delta w_k}\gamma_i$. We know from our previous calculations that $\dfrac{\delta}{\delta w_k} log \, \gamma_i = (1-\gamma_i)x_{ik}$.

We also know that, in general, $\delta log \, x = \dfrac{\delta x}{x}$, which implies that $\delta x = x \, \delta log \, x$.

Therefore, $\dfrac{\delta}{\delta w_k}\gamma_i = \gamma_i(1-\gamma_i)x_{ik}$.

With this information, the second derivative with respect to w_k becomes:

$$
\begin{aligned}
\frac{\delta^2}{\delta w_j w_k} log \, p(\mathbf{y}|\mathbf{X};\mathbf{w}) &= \sum_{i=1}^{d} \frac{\delta}{\delta w_k}\gamma_i x_{ij} \\
&= \sum_{i=1}^{d} \gamma_i(1-\gamma_i)x_{ik}x_{ij} \quad \text{this is a quadratic form} \\
&= \mathbf{a}_{\mathrm{k}}^{\mathrm{T}} \mathbf{R} \, \mathbf{a}_{\mathrm{j}} > 0 \\
\text{where} \quad \mathbf{a}_{\mathrm{k}} &= [x_{1k}, x_{2k}, \ldots, x_{dk}]^T, \\
\mathbf{R} &= \begin{pmatrix} \gamma_1(1-\gamma_1) & 0 & 0 & \cdots & 0 \\ 0 & \gamma_2(1-\gamma_2) & 0 & \cdots & 0 \\ 0 & 0 & \gamma_3(1-\gamma_3) & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \gamma_d(1-\gamma_d) \end{pmatrix}
\end{aligned}
$$

Here, $\mathbf{a_k}$ and $\mathbf{a}_{\mathrm{j}}$ represent columns of our matrix **X**. In particular, the expression $\mathbf{a}_{\mathrm{k}}^{\mathrm{T}} \mathbf{R} \, \mathbf{a}_{\mathrm{j}}$ gives us the derivative for the kth-jth entry. Thus, the Hessian matrix can be expressed as $\mathbf{H} = \mathbf{X}^{\mathrm{T}}\mathbf{R}\mathbf{X}$.

To incorporate the constant, we only have to take the dimensions into account as follows

$$\underset{(d+1)\times(d+1)}{\mathrm{H}} = \underset{(d+1)\times d}{X^T} \times \underset{d\times d}{\mathrm{R}} \times \underset{d\times(d+1)}{\mathrm{X}}$$

One way of showing that **H** is positive definite is to rewrite it as $\mathbf{H} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$ for some matrix **A** with independent columns.

The first step to show this is to make sure that all the the entries in **R**, $\gamma_i(1-\gamma_i)$, are greater than zero.

Recall that $\gamma_i = \dfrac{1}{1 + exp(-w_0 - \mathbf{w} \cdot \mathbf{x}_i)}$, which implies that $0 \leq \gamma_i \leq 1$ and $0 \leq (1 - \gamma_i) \leq 1$, so $0 \leq \gamma_i(1 - \gamma_i) \leq 1$. This means that we can take the square root of all the elements in **R** and rewrite the Hessian as follows

$$\mathbf{H} = \mathbf{X}^{\mathrm{T}}\mathbf{R}^{1/2}\mathbf{R}^{1/2}\mathbf{X} \quad \text{since} \quad \mathbf{R}^{1/2} \quad \text{is a diagonal matrix it is true that} \quad \mathbf{R}^{1/2} = (\mathbf{R}^{1/2})^T$$
$$= (\mathbf{R}^{1/2}\mathbf{X})^T(\mathbf{R}^{1/2}\mathbf{X})$$

Given that the columns of $\mathbf{R}^{1/2}$ are linearly independent, then, by definition, any linear combination of the columns is independent, in particular $\mathbf{R}^{1/2}\mathbf{X}$. Therefore, the matrix **H** is a positive definite matrix.

Finally, we can use the Newton-Raphson method to approximate the log-loss function around a minimum

$$\mathbf{w}_{\mathrm{t+1}} = \mathbf{w}_{\mathrm{t}} + \mathbf{H}^{-1}\frac{\delta}{\delta\mathbf{w}}\, log\, p(X; \mathbf{w})$$

$$= \mathbf{w}_{\mathrm{t}} + (\mathbf{X}^{\mathrm{T}}\mathbf{R}\mathbf{X})^{-1}\mathbf{X}^T(\boldsymbol{\gamma} - \mathbf{y})$$

$$= (\mathbf{X}^{\mathrm{T}}\mathbf{R}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{R}(\mathbf{X}\mathbf{w}_{\mathrm{t}} + \mathbf{R}^{-1}(\boldsymbol{\gamma} - \mathbf{y}))$$

$$= (\mathbf{X}^{\mathrm{T}}\mathbf{R}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{R}\mathbf{z}_{\mathrm{t}}$$

$$\text{where} \quad \mathbf{z}_{\mathrm{t}} = (\mathbf{X}\mathbf{w}_{\mathrm{t}} + \mathbf{R}^{-1}(\boldsymbol{\gamma} - \mathbf{y}))$$

The solution for $\mathbf{w}_{\mathrm{t+1}}$ looks very similar to the optimal solution to least squares problem. In fact, it can be shown that it is the optimal solution for the following problem:

$$\underset{\mathbf{w}}{\operatorname{argmin}} = \sum_{i=1}^{d} \gamma_i(1 - \gamma_i)(z_i - \mathbf{w} \cdot \mathbf{x}_{\mathrm{i}})^2$$

Therefore, in each iteration we are computing the global minimum given that the loss function for this problem is convex.

Problem 4

In this problem, we will represent the label for y_i by an indicator vector t_i. For example, if t_1 refers to $c = 1$, then $t_1^T = [1, 0, \ldots, 0]$.

Therefore, t_i represents a basis vector for a class c, containing a 1 at the jth position and 0 elsewhere. This vectors will be useful later to compute the the derivatives.

Befor we continue, lets define $\hat{p}_i$ as a vector of probabilities for the ith row as $\hat{p}_i = \hat{p}(y_i|\mathbf{x}_{\mathrm{i}}; \mathbf{W})$ and $a_i = \mathbf{w}_{\mathrm{i}} \cdot \mathbf{x}_{\mathrm{i}}$. This last definition allows us to rewrite the softmax model as follows:

$$\hat{p}(y_i = c|\mathbf{x}_\mathrm{i}; \mathbf{W}) \frac{exp(a_c)}{\sum_{y=1}^{C} exp(a_y)}$$

We have to rewrite now our cost function in terms of the log-loss as follows

$$\mathbf{W}^* = \underset{\mathrm{W}}{\mathrm{argmin}}\ J(\mathbf{x}_\mathrm{i}, y_i, \mathbf{W}) = -\frac{1}{N}\sum_{i=1}^{N} t_i\, log\, \hat{p}(y_i|\mathbf{x}_\mathrm{i}; \mathbf{W}) + \lambda \sum_i \sum_j \mathrm{w_{ij}}^2$$

$$= -\frac{1}{N}\sum_{i=1}^{N} t_i log\, \hat{p}_i + \lambda \sum_i \sum_j \mathrm{w_{ij}}^2$$

First, lets take the derivative of the log-loss function **J** with respect to $\hat{p}_i$.

$$(1) \quad \frac{\delta J}{\delta \hat{p}_i} = -\frac{1}{N}\frac{t_i}{\hat{p}_i}$$

$$(2) \quad \frac{\delta \hat{p}_i}{\delta a_k} = \begin{cases} \dfrac{exp(a_i)}{\sum_{y=1}^{C} exp(a_y)} - \left(\dfrac{exp(a_i)}{\sum_{y=1}^{C} exp(a_y)}\right)^2 & \text{if} \quad i = k \\ -\dfrac{exp(a_i)exp(a_k)}{\left(\sum_{y=1}^{C} exp(a_y)\right)^2} & \text{if} \quad i \neq k \end{cases}$$

$$= \begin{cases} \hat{p}_i(1-\hat{p}_i) & \text{if} \quad i = k \\ \hat{p}_i\hat{p}_k & \text{if} \quad i \neq k \end{cases}$$

$$(3) \quad \frac{\delta J}{\delta a_i} = \sum_{k=1}^{C} \frac{\delta J}{\delta \hat{p}_k}\frac{\delta \hat{p}_k}{\delta a_i}$$

$$= \frac{\delta J}{\delta \hat{p}_i}\frac{\delta \hat{p}_i}{\delta a_i} - \sum_{k\neq i} \frac{\delta J}{\delta \hat{p}_k}\frac{\delta \hat{p}_k}{\delta a_i}$$

$$= -\frac{1}{N}\frac{t_i}{\hat{p}_i}\hat{p}_i(1-\hat{p}_i) + \frac{1}{N}\sum_{k\neq i}\frac{t_k}{\hat{p}_k}\hat{p}_k\hat{p}_i$$

$$= -\frac{1}{N}t_i(1-\hat{p}_i) + \frac{1}{N}\sum_{k\neq i} t_k\hat{p}_i$$

$$= -\frac{1}{N}t_i(1-\hat{p}_i) + \frac{1}{N}\hat{p}_i\sum_{k\neq i} t_k$$

$$= \frac{1}{N}\left[\hat{p}_i\sum_{k\neq i} t_k - t_i(1-\hat{p}_i)\right]$$

$$= \frac{1}{N}\left[\hat{p}_i\left(\sum_{k\neq i} t_k + t_i\right) - t_i\right]$$

$$= \frac{1}{N}\left[\hat{p}_i\left(\sum_{k} t_k\right) - t_i\right] \quad \text{where} \quad \sum_{k} t_k = 1$$

$$= \frac{1}{N}\left[\hat{p}_i - t_i\right]$$

$$
\begin{aligned}
(4) \qquad \frac{\delta J}{\delta w_{ij}} &= \sum_{i=1}^{N} \frac{\delta J}{\delta a_i} \frac{\delta a_i}{\delta w_{ij}} + \lambda \sum_i \sum_j \frac{\delta}{\delta w_{ij}} \mathrm{w_{ij}}^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} \left[\hat{p}_i - t_i \right] \frac{\delta a_i}{\delta w_{ij}} + 2\lambda w_{ij} \\
&= \frac{1}{N} \sum_{i=1}^{N} \left[\hat{p}_i - t_i \right] x_{ij} + 2\lambda w_{ij} \\
&= \frac{1}{N} \sum_{i=1}^{N} \left[\hat{p}_i - t_i \right] x_{ij} + 2\lambda \sum_{i=1}^{N} w_{ij} \\
&= \frac{1}{N} (\mathbf{X}^{\mathrm{T}} (\widehat{\mathbf{p}} - \mathbf{t})) + 2\lambda \mathbf{W} \mathrm{t}_i \\
&= \frac{1}{N} (\mathbf{X}^{\mathrm{T}} (\widehat{\mathbf{p}} - \mathbf{t})) + 2\lambda \mathbf{w}_{\mathrm{i}}
\end{aligned}
$$

$$
\hat{\boldsymbol{p}} = \begin{pmatrix} \hat{p}_{11} & \hat{p}_{12} & \cdots & \hat{p}_{1C} \\ \hat{p}_{21} & \hat{p}_{22} & \cdots & \hat{p}_{2C} \\ \hat{p}_{31} & \hat{p}_{32} & \cdots & \hat{p}_{3C} \\ \vdots & \ddots & \ddots & \vdots \\ \hat{p}_{N1} & \hat{p}_{N2} & \cdots & \hat{p}_{NC} \end{pmatrix}, \quad
\mathbf{t} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1C} \\ t_{21} & t_{22} & \cdots & t_{2C} \\ t_{31} & t_{32} & \cdots & t_{3C} \\ \vdots & \ddots & \ddots & \vdots \\ t_{N1} & t_{N2} & \cdots & t_{NC} \end{pmatrix}, \quad
\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ x_{31} & x_{32} & \cdots & x_{3N} \\ \vdots & \ddots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NN} \end{pmatrix}
$$

$$
\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & w_{2N} \\ w_{31} & w_{32} & \cdots & w_{3N} \\ \vdots & \ddots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{NN} \end{pmatrix}
$$

Therefore, we can write the equation for the stochastic gradient descent as follows

$$
\mathbf{w}_{\mathrm{t+1}} = \eta_t \frac{1}{N} (\mathbf{X}^{\mathrm{T}} (\widehat{\mathbf{p}} - \mathbf{t})) + 2\lambda \mathbf{w}_{\mathrm{t}}
$$