



SHARIF UNIVERSITY OF TECHNOLOGY

---

## Deep Learning Assignment 2

---

Saeedreza Zouashkiani

November 21, 2022

1

### 1.1

Consider the Taylor expansion of the loss function at  $\omega$ :

$$J(\omega + \delta\omega) - J(\omega) \approx \nabla_{\omega} L \delta\omega + \frac{1}{2} \delta\omega^T H \delta\omega \quad (1)$$

The first term goes to zero when the algorithm converges. So, our problem is to minimize:

$$C = \frac{1}{2} \delta\omega^T H \delta\omega \quad (2)$$

Our goal is then to set one of the weights to zero which is denoted by  $\omega_i$  to minimize the increase in error given by equation (2). Then to eliminate  $\omega_i$ :

$$\omega_i + \mathbf{e}_i^T \cdot \delta\omega = 0 \quad (3)$$

Where  $\mathbf{e}_i$  is a vector with a 1 corresponding to  $\omega_i$  and zero elsewhere. To minimize equation (2) constrained on equation (3), we use Lagrangian multipliers:

$$L = \frac{1}{2} \delta\omega^T H \delta\omega + \lambda (\omega_i + \mathbf{e}_i^T \cdot \delta\omega) \quad (4)$$

Differentiating with respect to  $\delta\omega$  and setting to zero:

$$H \delta\omega + \lambda \mathbf{e}_i = 0 \quad (5)$$

Then if we solve for  $\lambda$ :

$$\lambda = \frac{\omega_i}{(\mathbf{H}^{-1})_{ii}} \quad (6)$$

Where  $\mathbf{H}^{-1}$ , corresponds to the  $i^{th}$  element in the inverse Hessian matrix. Then by plugging (6) into (5)

$$\delta\boldsymbol{\omega} = -\frac{\omega_i}{(\mathbf{H}^{-1})_{ii}} \mathbf{H}^{-1} \cdot \mathbf{e}_i \quad (7)$$

Then plugging (7) into (2):

$$\mathbf{C} = \frac{1}{2} \frac{\omega_i^2}{(\mathbf{H}^{-1})_{ii}} \quad (8)$$

## 1.2

If  $\mathbf{H} = \mathbf{I}$ , then  $\mathbf{H}^{-1} = \mathbf{I}$ . Therefore:

$$\delta\boldsymbol{\omega} = -\omega_i \quad (9)$$

$$\mathbf{C} = \frac{1}{2} \omega_i^2 \quad (10)$$

In the previous part, we needed to compute the inverse Hessian first, and compute  $\frac{1}{2} \frac{\omega_i^2}{(\mathbf{H}^{-1})_{ii}}$  for each element and remove the one with the least increase in the loss, but in here we compute  $\frac{1}{2} \omega_i^2$  and set  $\delta\boldsymbol{\omega} = -\omega_i$  for that  $i$  that minimizes the loss.

## 2

### 2.1

This is solved in part 3.1

### 2.2

$$\begin{aligned} & \mathbb{E}[\frac{1}{N} \|(X(X^T X)^{-1} X^T - \mathbf{I})\boldsymbol{\epsilon}\|_2^2] \\ &= \frac{1}{N} \mathbb{E}[\boldsymbol{\epsilon}^T (X(X^T X)^{-1} X^T - \mathbf{I})^T (X(X^T X)^{-1} X^T - \mathbf{I}) \boldsymbol{\epsilon}] \\ &= \frac{1}{N} \mathbb{E}[\boldsymbol{\epsilon}^T (\mathbf{B})^T (\mathbf{B}) \boldsymbol{\epsilon}] \end{aligned} \quad (11)$$

Since  $\mathbf{A}$  is symmetric and  $\mathbf{A}^T \mathbf{A} = \mathbf{A}^2$ , then it is idempotent. That is its eigenvalues consist of  $d$  ones and  $n - d$  zeros. Therefore using 11:

$$\mathbf{B}^T \mathbf{B} = (\mathbf{A} - \mathbf{I})^T (\mathbf{A} - \mathbf{I}) = \mathbf{A}^2 - 2\mathbf{A} + \mathbf{I} = \mathbf{I} - \mathbf{A} \quad (12)$$

Since  $\boldsymbol{\epsilon}^T \mathbf{B}^T \mathbf{B} \boldsymbol{\epsilon}$  is a scalar, then its trace equals itself:

$$\begin{aligned} & \frac{1}{N} \mathbb{E}[\boldsymbol{\epsilon}^T \mathbf{B}^T \mathbf{B} \boldsymbol{\epsilon}] \\ &= \frac{1}{N} \mathbb{E}[\text{Trace}(\boldsymbol{\epsilon}^T \mathbf{B}^T \mathbf{B} \boldsymbol{\epsilon})] \\ &= \frac{1}{N} \mathbb{E}[\text{Trace}(\mathbf{B}^T \mathbf{B} \boldsymbol{\epsilon}^T \boldsymbol{\epsilon})] \\ &= \frac{1}{N} \text{Trace}(\mathbf{B}^T \mathbf{B} \boldsymbol{\Sigma}) \\ &= \frac{\sigma^2}{N} \text{Trace}(\mathbf{I} - \mathbf{A}) \\ &= \frac{N - d}{N} \sigma^2 \end{aligned} \quad (13)$$

Where the last part comes from the property of idempotent matrices. That is  $\text{Trace}(\mathbf{A}) = \text{rank}(\mathbf{X})$ .

### 2.3

As  $d$  increases and gets closer to  $N$ , the trace of  $\mathbf{I} - \mathbf{A}$  approaches zero. That is the number of independent columns of  $\mathbf{X}$  increases, so we can have a better projection matrix than before. Naturally the training error will decrease as well.

## 3

### 3.1

We want to find:

$$\begin{aligned}
 & \underset{\boldsymbol{\beta}}{\text{argmin}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\
 &= \underset{\boldsymbol{\beta}}{\text{argmin}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
 &= \underset{\boldsymbol{\beta}}{\text{argmin}} \mathbf{Y}^T \mathbf{Y} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{Y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} \\
 &= \underset{\boldsymbol{\beta}}{\text{argmin}} -2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{Y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}
 \end{aligned} \tag{14}$$

Define  $L(\boldsymbol{\beta}) = 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{Y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}$ , and differentiate with respect to  $\boldsymbol{\beta}$  and set to zero:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T \mathbf{Y} + 2\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} \tag{15}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \tag{16}$$

### 3.2

Add the L2 norm of the weights to the minimization problem of the previous part:

$$\begin{aligned}
 & \underset{\boldsymbol{\beta}}{\text{argmin}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_2^2 \\
 &= \underset{\boldsymbol{\beta}}{\text{argmin}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \alpha \boldsymbol{\beta}^T \boldsymbol{\beta} \\
 &= \underset{\boldsymbol{\beta}}{\text{argmin}} \mathbf{Y}^T \mathbf{Y} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{Y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} + \alpha \boldsymbol{\beta}^T \boldsymbol{\beta} \\
 &= \underset{\boldsymbol{\beta}}{\text{argmin}} -2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{Y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} + \alpha \boldsymbol{\beta}^T \boldsymbol{\beta}
 \end{aligned} \tag{17}$$

Define  $L(\boldsymbol{\beta}) = 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{Y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} + \alpha \boldsymbol{\beta}^T \boldsymbol{\beta}$ , and differentiate with respect to  $\boldsymbol{\beta}$  and set to zero:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T \mathbf{Y} + 2\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} + 2\alpha \boldsymbol{\beta} \tag{18}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X} + \alpha \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y} \tag{19}$$

### 3.3

We know that both  $\boldsymbol{\beta}^*$  and  $\hat{\boldsymbol{\beta}}$  have an expectation value of  $\boldsymbol{\beta}$ . Let  $\boldsymbol{\beta}^* = \mathbf{A}\mathbf{y}$  and  $\hat{\boldsymbol{\beta}} = \mathbf{B}\mathbf{y}$ . Where  $\mathbf{A} = (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \boldsymbol{\Sigma}^{-1} \mathbf{X}^T$  and  $\mathbf{B} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ .

$$\begin{aligned} \text{Var}(\boldsymbol{\beta}^*) &= \mathbf{B} \text{Var}(\mathbf{y}) \mathbf{B}^T = \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}^{-1} \\ \text{Var}(\hat{\boldsymbol{\beta}}) &= \mathbf{A} \text{Var}(\mathbf{y}) \mathbf{A}^T = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \end{aligned} \quad (20)$$

If  $\mathbf{X}\mathbf{F} = \boldsymbol{\Sigma}\mathbf{X}$ , then:

$$(\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} = \mathbf{F}(\mathbf{X}^T \mathbf{X})^{-1} \quad (21)$$

It is obvious from 21 that  $\mathbf{F}$  is non-singular. Using (21), we can write:

$$\begin{aligned} \text{Var}(\boldsymbol{\beta}^*) &= \mathbf{F}(\mathbf{X}^T \mathbf{X})^{-1} \\ \text{Var}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \mathbf{F} (\mathbf{X}^T \mathbf{X})^{-1} = \mathbf{F}(\mathbf{X}^T \mathbf{X})^{-1} \end{aligned} \quad (22)$$

Therefore  $\boldsymbol{\beta}^*$  and  $\hat{\boldsymbol{\beta}}$  estimate the same quantity. On the other hand, to prove when the estimators are equal, we need to show that:

$$\begin{aligned} \boldsymbol{\beta}^* &= \hat{\boldsymbol{\beta}} \\ (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} &= \mathbf{X}^T \boldsymbol{\Sigma} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ \mathbf{X}^T (\mathbf{X} (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} - \boldsymbol{\Sigma} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}) &= \mathbf{0} \\ \mathbf{X} (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} &= \boldsymbol{\Sigma} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ \mathbf{X} (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} &= \boldsymbol{\Sigma} \mathbf{X} \\ \mathbf{X} \mathbf{F} &= \boldsymbol{\Sigma} \mathbf{X} \end{aligned} \quad (23)$$

### 3.4

Extend the error function without the L-1 norm:

$$\begin{aligned} L(\boldsymbol{\beta}, \lambda_1) &= \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_2^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_1 \boldsymbol{\beta}^T \boldsymbol{\beta} \\ &= \mathbf{Y}^T \mathbf{Y} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{Y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} + \lambda_1 \boldsymbol{\beta}^T \boldsymbol{\beta} \end{aligned} \quad (24)$$

Define the augmented  $\hat{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_1} \mathbf{I} \end{bmatrix}$ , and  $\hat{\mathbf{Y}} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{0} \end{bmatrix}$ . Then equation (24) can be written as:

$$L(\boldsymbol{\beta}, \lambda_1) = \|\hat{\mathbf{Y}} - \hat{\mathbf{X}}\boldsymbol{\beta}\|_2^2 \quad (25)$$

Therefore the loss function  $L(\boldsymbol{\beta}, \lambda_1, \lambda_2)$  will become:

$$L(\boldsymbol{\beta}, \lambda_1, \lambda_2) = \|\hat{\mathbf{Y}} - \hat{\mathbf{X}}\boldsymbol{\beta}\|_2^2 \quad (26)$$

## 4

### 4.1

Let  $J_D = (y_d - O_D)^2$  and  $J_N = (y_d - O_N)^2$  denote the loss function of the Dropout and without Dropout network, respectively, and  $O_D = \sum_{k=1}^n \delta_k \omega_k x_k$

$$\begin{aligned} \frac{\partial J_N}{\partial \omega_i} &= -2(y_d - O_N) \frac{\partial O_N}{\partial \omega_i} = -2(y_d - O_N) x_i = -2y_d x_i + 2 \sum_{k=1}^n \omega_k x_k x_i \\ &= -2y_d x_i + 2\omega_i x_i^2 + 2 \sum_{k=1, k \neq i}^n \omega_k x_k x_i \end{aligned} \quad (27)$$

$$\begin{aligned}
\frac{\partial J_D}{\partial \omega_i} &= -2(y_d - O_D) \frac{\partial O_D}{\partial \omega_i} = -2(y_d - O_D) \delta_i x_i \\
&= -2y_d \delta_i x_i + 2\omega_i^2 \delta_i^2 x_i^2 + 2 \sum_{k=1, k \neq i}^n \omega_k \delta_k \delta_i x_k x_i
\end{aligned} \tag{28}$$

Since equation (12) is probabilistic, we take the expectation:

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial J_D}{\partial \omega_i}\right] &= -2y_d x_i + 2\omega_i x_i^2 + 2\sigma^2 \omega_i^2 x_i^2 + 2 \sum_{k=1, k \neq i}^n \omega_k x_k x_i \\
&= \frac{\partial J_N}{\partial \omega_i} + 2\sigma^2 \omega_i^2 x_i^2
\end{aligned} \tag{29}$$

## 4.2

The overall regularized loss function can be written as:

$$J_R = (y_d - \sum_{k=1}^n \delta_k \omega_k x_k)^2 + \sigma^2 \sum_{k=1}^n \omega_k^2 x_k^2 \tag{30}$$

It means that the network is regularized by the square of multiplication of weights by inputs with a regularization factor of  $\sigma^2$ .

## 5

Let the cost function be:

$$L = \frac{1}{2} \omega^T H \omega \tag{31}$$

Computing the first and second order derivatives:

$$\nabla_{\omega} L = \frac{\partial L}{\partial \omega} = H \omega \tag{32}$$

$$\nabla_{\omega}^2 L = \frac{\partial^2 L}{\partial \omega^2} = H \tag{33}$$

### 5.1

$$\omega^{(t+1)} = \omega^{(t)} - \epsilon H \omega^{(t)} = Q(I - \epsilon \Lambda) Q^T \omega^{(t)} \tag{34}$$

### 5.2

$$\omega^{(t+1)} = Q(I - \epsilon \Lambda) Q^T \omega^{(0)} \tag{35}$$

### 5.3

$|(I - \epsilon \Lambda)| < 1$  must be satisfied. Therefore:

$$\epsilon < \frac{2}{\lambda_{max}} \tag{36}$$

### 5.4

Because of 33, the newtons method will be:

$$\omega^{(t+1)} = (1 - \epsilon) \omega^{(t)} \tag{37}$$

Then by choosing  $\epsilon = 1$ , the newton method will converge in one step.

## 5.5

Because in Newton's method, calculating the inverse of the Hessian matrix is expensive, gradient descent is used instead.

## 6

### 6.1

It will learn a similarity measure between 2 inputs, since the inputs are both mapped to a latent space and the distance between the 2 latent vectors is minimized. This can be used for images and text, for example.

### 6.2

$$\begin{aligned}
 \mathbf{z}_1 &= \mathbf{W}\mathbf{x}_1 + \mathbf{b} \\
 \mathbf{z}_2 &= \mathbf{W}\mathbf{x}_2 + \mathbf{b} \\
 \Delta_1 &= \frac{\partial J}{\partial \mathbf{h}_1} = 2\mathbf{h}_1 - \mathbf{h}_2 \\
 \Delta_2 &= \frac{\partial J}{\partial \mathbf{h}_2} = 2\mathbf{h}_2 - \mathbf{h}_1 \\
 \Delta_3 &= \frac{\partial J}{\partial \mathbf{z}_1} = \Delta_1 \odot (1 - \mathbf{h}_1^2) \\
 \Delta_4 &= \frac{\partial J}{\partial \mathbf{z}_2} = \Delta_2 \odot (1 - \mathbf{h}_2^2) \\
 \frac{\partial J}{\partial \mathbf{W}} &= \Delta_3 \mathbf{x}_1^T + \Delta_4 \mathbf{x}_2^T + 2\mathbf{W} \\
 \frac{\partial J}{\partial \mathbf{b}} &= \Delta_3 + \Delta_4
 \end{aligned} \tag{38}$$

Then the gradient descent update rule for  $m$  batch size is:

$$\begin{aligned}
 \mathbf{W}^{(t+1)} &= \mathbf{W}^{(t)} - \epsilon \frac{1}{m} \sum_{i=1}^m \frac{\partial J_i}{\partial \mathbf{W}} \\
 \mathbf{b}^{(t+1)} &= \mathbf{b}^{(t)} - \epsilon \frac{1}{m} \sum_{i=1}^m \frac{\partial J_i}{\partial \mathbf{b}}
 \end{aligned} \tag{39}$$