



## EE2211 Tutorial 8

$$\begin{aligned}
 1) \quad w &= w - \eta \nabla_w C \\
 x &= x - \eta \frac{\partial f(x)}{\partial x} \rightarrow x^4 \\
 &= 2 - 0.1(4x^3) \\
 &= 2 - 0.1(4(2)^3) = -1.2
 \end{aligned}$$

### Question 1

Suppose we are minimizing  $f(x) = x^4$  with respect to  $x$ . We initialize  $x$  to be 2. We perform gradient descent with learning rate 0.1. What is the value of  $x$  after the first iteration?

### Question 2

Please consider the csv file (government-expenditure-on-education.csv), which depicts the government's educational expenditure over the years. We would like to predict expenditure as a function of year. To do this, fit an exponential model  $f(\mathbf{x}, \mathbf{w}) = \exp(-\mathbf{x}^T \mathbf{w})$  with squared error loss to estimate  $\mathbf{w}$  based on the csv file and gradient descent. In other words,  $C(\mathbf{w}) = \sum_{i=1}^m (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2$ .

Note that even though year is one dimensional, we should add the bias term, so  $\mathbf{x} = [1 \text{ year}]^T$ . Furthermore, optimizing the exponential function is tricky (because a small change in  $\mathbf{w}$  can lead to large change in  $f$ ). Therefore for the purpose of optimization, divide the "year" variable by the largest year (2018) and divide the "expenditure" by the largest expenditure, so that the resulting normalized year and normalized expenditure variables have maximum values of 1. Use a learning rate of 0.03 and run gradient descent for 2000000 iterations.

- Plot the cost function  $C(\mathbf{w})$  as a function of the number of iterations.
- Use the fitted parameters to plot the predicted educational expenditure from year 1981 to year 2023.
- Repeat (a) using a learning rate of 0.1 and learning rate of 0.001. What do you observe relative to (a)?

The goal of this question is for you to code up gradient descent, so I will provide you with the gradient derivation. First, please note that in general,  $\nabla_{\mathbf{w}}(\mathbf{x}^T \mathbf{w}) = \mathbf{x}$ . To see this:

$$\nabla_{\mathbf{w}}(\mathbf{x}^T \mathbf{w}) = \begin{bmatrix} \frac{\partial(\mathbf{x}^T \mathbf{w})}{\partial w_1} \\ \frac{\partial(\mathbf{x}^T \mathbf{w})}{\partial w_2} \\ \vdots \\ \frac{\partial(\mathbf{x}^T \mathbf{w})}{\partial w_d} \end{bmatrix} = \begin{bmatrix} \frac{\partial(w_1 x_1 + w_2 x_2 + \dots + w_d x_d)}{\partial w_1} \\ \frac{\partial(w_1 x_1 + w_2 x_2 + \dots + w_d x_d)}{\partial w_2} \\ \vdots \\ \frac{\partial(w_1 x_1 + w_2 x_2 + \dots + w_d x_d)}{\partial w_d} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \mathbf{x} \quad (1)$$

The above equality will be very useful for the other questions as well. Now, going back to our question,

$$\nabla_{\mathbf{w}} C(\mathbf{w}) = \nabla_{\mathbf{w}} \sum_{i=1}^m (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2 \quad (2)$$

$$= \sum_{i=1}^m \nabla_{\mathbf{w}} (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2 \quad (3)$$

$$= \sum_{i=1}^m 2(f(\mathbf{x}_i, \mathbf{w}) - y_i) \nabla_{\mathbf{w}} f(\mathbf{x}_i, \mathbf{w}) \quad \text{chain rule} \quad (4)$$

$$= \sum_{i=1}^m 2(f(\mathbf{x}_i, \mathbf{w}) - y_i) \nabla_{\mathbf{w}} \exp(-\mathbf{x}_i^T \mathbf{w}) \quad (5)$$

$$= - \sum_{i=1}^m 2(f(\mathbf{x}_i, \mathbf{w}) - y_i) \exp(-\mathbf{x}_i^T \mathbf{w}) \nabla_{\mathbf{w}} (\mathbf{x}_i^T \mathbf{w}) \quad \text{chain rule} \quad (6)$$

$$= - \sum_{i=1}^m 2(f(\mathbf{x}_i, \mathbf{w}) - y_i) \exp(-\mathbf{x}_i^T \mathbf{w}) \mathbf{x}_i \quad (7)$$

$$= - \sum_{i=1}^m 2(f(\mathbf{x}_i, \mathbf{w}) - y_i) f(\mathbf{x}_i, \mathbf{w}) \mathbf{x}_i \quad (8)$$

### Question 3

Given the linear learning model  $f(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T \mathbf{w}$ , where  $\mathbf{x} \in \mathbb{R}^d$ . Consider the loss function  $L(f(\mathbf{x}_i, \mathbf{w}), y_i) = (f(\mathbf{x}_i, \mathbf{w}) - y_i)^4$ , where  $i$  indexes the  $i$ -th training sample. The final cost function is  $C(\mathbf{w}) = \sum_{i=1}^m L(f(\mathbf{x}_i, \mathbf{w}), y_i)$ , where  $m$  is the total number of training samples. Derive the gradient of the cost function with respect to  $\mathbf{w}$ .

### Question 4

Repeat Question 3 using  $f(\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{x}^T \mathbf{w})$ , where  $\sigma(a) = \frac{1}{1 + \exp(-\beta a)}$

### Question 5

Repeat Question 3 using  $f(\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{x}^T \mathbf{w})$ , where  $\sigma(a) = \max(0, a)$

$$3) f(x, w) = x_i^T w \quad L(f(x_i, w), y_i) = (f(x_i, w) - y_i)^4$$

(learning  
model)

(loss function)

$$\text{Final cost function: } C(w) = \sum_{i=1}^m L(f(x_i, w), y_i)$$

$$\begin{aligned} \text{Gradient of cost function: } \nabla_w C(w) &= \nabla_w \sum_{i=1}^m L(f(x_i, w), y_i) \\ &= \sum_{i=1}^m \nabla_w L(f(x_i, w), y_i) \\ &= \sum_{i=1}^m \nabla_w (f(x_i, w) - y_i)^4 \\ &= \sum_{i=1}^m 4(f(x_i, w) - y_i)^3 \nabla_w f(x_i, w) \\ &= \sum_{i=1}^m 4(f(x_i, w) - y_i)^3 \nabla_w x_i^T w \\ &= \sum_{i=1}^m 4(f(x_i, w) - y_i)^3 x_i \end{aligned}$$

$$4) f(x, w) = \sigma(x^T w) \quad \sigma(a) = \frac{1}{1 + \exp(-\beta a)} \quad L(f(x_i, w), y_i) = (f(x_i, w) - y_i)^4$$

(learning model)

(loss function)

$$\text{Final cost function: } C(w) = \sum_{i=1}^m L(f(x_i, w), y_i)$$

$$\text{Gradient of cost function: } \nabla_w C(w) = \nabla_w \sum_{i=1}^m L(f(x_i, w), y_i)$$

$$= \nabla_w \sum_{i=1}^m (f(x_i, w) - y_i)^4$$

$$= \sum_{i=1}^m 4(f(x_i, w) - y_i)^3 \nabla_w f(x_i, w)$$

$$= \sum_{i=1}^m 4(f(x_i, w) - y_i)^3 \nabla_w \sigma(x_i^T w)$$

$$= \sum_{i=1}^m 4(f(x_i, w) - y_i)^3 \frac{\partial \sigma(a)}{\partial a} \nabla_w (x_i^T w)$$

$$= \sum_{i=1}^m 4(f(x_i, w) - y_i)^3 \frac{\beta}{(1 + e^{-\beta a})^2} e^{-\beta a} x_i$$

$$\begin{aligned} \frac{\partial \sigma(a)}{\partial a} &= \frac{\partial}{\partial a} \left( \frac{1}{1 + e^{-\beta a}} \right) \\ &= -1(1 + e^{-\beta a})^{-2} \frac{\partial}{\partial a} (e^{-\beta a}) \\ &= -1(1 + e^{-\beta a})^{-2} e^{-\beta a} \frac{\partial}{\partial a} (-\beta a) \\ &= -1(1 + e^{-\beta a})^{-2} e^{-\beta a} \frac{\partial}{\partial a} (-\beta a) \\ &= \frac{\beta}{(1 + e^{-\beta a})^2} e^{-\beta a} \end{aligned}$$

$$5) \quad f(x, w) = \sigma(x^T w) \quad \sigma(a) = \max(0, a) \quad L(f(x_i, w), y_i) = (f(x_i, w) - y_i)^2$$

(learning model) (loss function)

$$\text{Final cost function: } C(w) = \sum_{i=1}^M L(f(x_i, w), y_i)$$

$$\begin{aligned} \text{Gradient of cost function: } \nabla_w C(w) &= \nabla_w \sum_{i=1}^M L(f(x_i, w), y_i) \\ &= \nabla_w \sum_{i=1}^M (f(x_i, w) - y_i)^2 \\ &= \sum_{i=1}^M 4(f(x_i, w) - y_i)^3 \nabla_w f(x_i, w) \\ &= \sum_{i=1}^M 4(f(x_i, w) - y_i)^3 \nabla_w \sigma(x_i^T w) \\ &= \sum_{i=1}^M 4(f(x_i, w) - y_i)^3 \frac{\partial \sigma(a)}{\partial a} \nabla_w (x_i^T w) \\ &= \sum_{i=1}^M 4(f(x_i, w) - y_i)^3 \frac{\partial \sigma(a)}{\partial a} x_i \end{aligned}$$

$$\therefore \text{Since } \sigma(a) = \max(0, a) \rightarrow \begin{array}{l} a > 0 : a \\ a \leq 0 : 0 \end{array}$$

$$\frac{\partial \sigma(a)}{\partial a} = \begin{cases} 1 & \text{if } a > 0 \\ 0 & \text{if } a \leq 0 \end{cases} \quad [\text{Lect 8 pg 33}]$$

$$\text{Since } a = x_i^T w,$$

$$\frac{\partial \sigma(x_i^T w)}{\partial w} = \begin{cases} x_i & \text{if } x_i^T w > 0 \\ 0 & \text{if } x_i^T w \leq 0 \end{cases} \rightarrow \text{Indicator function}$$

$$\begin{aligned} \therefore \nabla_w C(w) &= \sum_{i=1}^M 4(f(x_i, w) - y_i)^3 \frac{\partial \sigma(a)}{\partial a} x_i \\ &= \sum_{i=1}^M 4(f(x_i, w) - y_i)^3 \delta(x_i^T w > 0) x_i \end{aligned}$$