EE2211 Tutorial 8

Dr Feng LIN

Q1

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?

Q1

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?

- The gradient of $f(x)$ is $4x^3$
- At x = 2, the gradient is $4 \times 2^3 = 32$
- After first iteration of gradient descent, value of x will be $x = 2 - 0.1 \times 32 = -1.2$

Q2

(Python)

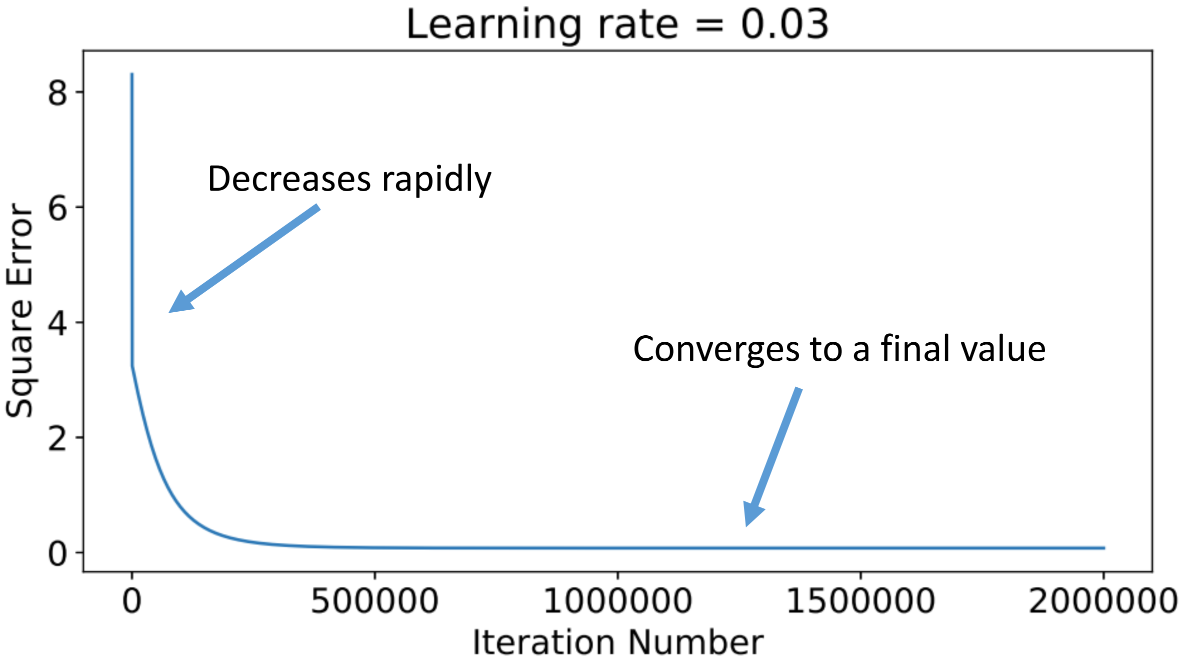
- Please consider the csv file (government-expenditure-on-education.csv), which depicts the government's educational expenditure over the years. We would like to predict expenditure as a function of year. To do this, fit an exponential model $f(x, w) = \exp(-x^T w)$ with squared error loss to estimate w based on the csv file and gradient descent.
  a) Plot the cost function $C(w)$ as a function of the number of iterations
  b) Use the fitted parameters to plot the predicted educational expenditure from year 1981 to year 2023
  c) Repeat (a) using a learning rate of 0.1 and learning rate of 0.001. What do you observe relative to (a)?

Q2

$$
\begin{aligned}
\nabla_{\mathbf{w}} C(\mathbf{w}) &= \nabla_{\mathbf{w}} \sum_{i=1}^{m} (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2 \\
&= \sum_{i=1}^{m} \nabla_{\mathbf{w}} (f(\mathbf{x}_i, \mathbf{w}) - y_i)^2 \\
&= \sum_{i=1}^{m} 2(f(\mathbf{x}_i, \mathbf{w}) - y_i) \nabla_{\mathbf{w}} f(\mathbf{x}_i, \mathbf{w}) \qquad \text{chain rule} \\
&= \sum_{i=1}^{m} 2(f(\mathbf{x}_i, \mathbf{w}) - y_i) \nabla_{\mathbf{w}} \exp(-\mathbf{x}_i^T \mathbf{w}) \\
&= -\sum_{i=1}^{m} 2(f(\mathbf{x}_i, \mathbf{w}) - y_i) \exp(-\mathbf{x}_i^T \mathbf{w}) \nabla_{\mathbf{w}} (\mathbf{x}_i^T \mathbf{w}) \qquad \text{chain rule} \\
&= -\sum_{i=1}^{m} 2(f(\mathbf{x}_i, \mathbf{w}) - y_i) \exp(-\mathbf{x}_i^T \mathbf{w}) \mathbf{x}_i \\
&= -\sum_{i=1}^{m} 2(f(\mathbf{x}_i, \mathbf{w}) - y_i) f(\mathbf{x}_i, \mathbf{w}) \mathbf{x}_i
\end{aligned}
$$

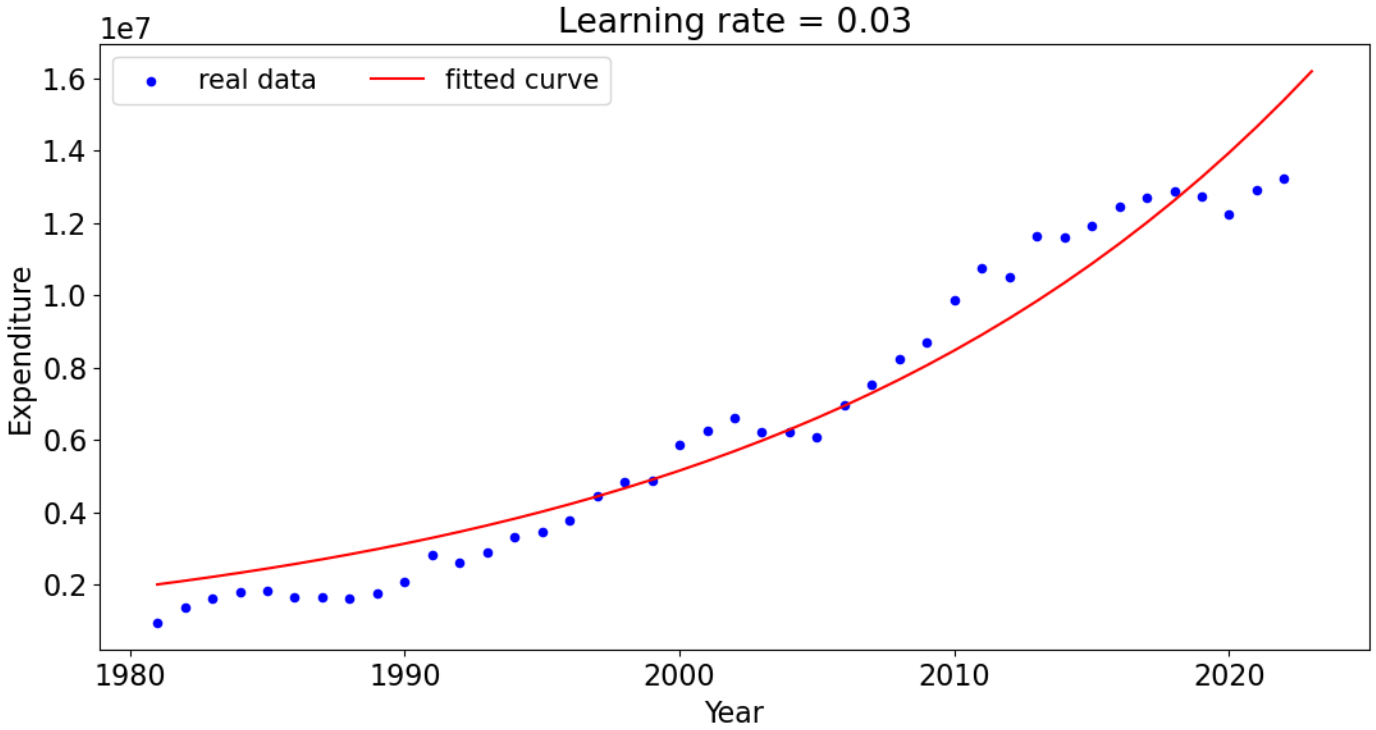
Q2

a) Plot the cost function $C(w)$ as a function of the number of iterations



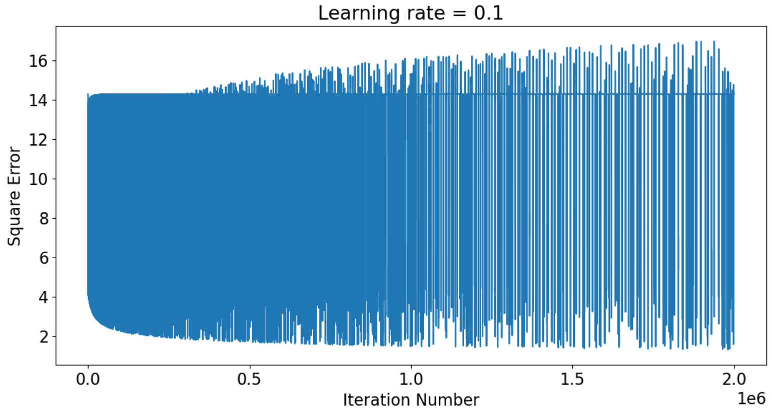
Q2

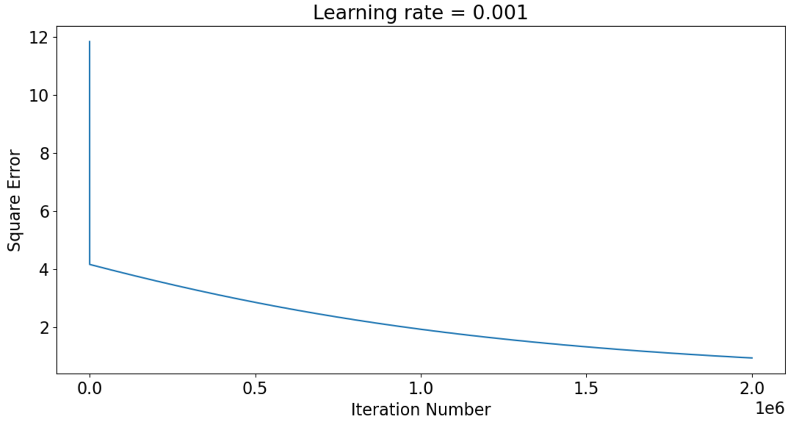
b) Use the fitted parameters to plot the predicted educational expenditure from year 1981 to year 2023



Q2

c) Repeat (a) using a learning rate of 0.1 and learning rate of 0.001. What do you observe relative to (a)?





A learning rate of 0.1 is too big, so the cost function does not decrease monotonically with increasing iterations, but instead fluctuate a lot without convergence. The final cost function value is much worse.

A learning rate of 0.001 is too small. So even though the cost function decreases monotonically with increasing iterations, gradient descent has not converged even after 2000000 iterations. The final cost function value is much worse.

Q3

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

Q3

$$\nabla_{\mathbf{w}} C(\mathbf{w}) = \nabla_{\mathbf{w}} \sum_{i=1}^{m} (f(\mathbf{x}_i, \mathbf{w}) - y_i)^4$$

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

$$= \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \nabla_{\mathbf{w}} f(\mathbf{x}_i, \mathbf{w})$$

$$= \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \nabla_{\mathbf{w}} (\mathbf{x}_i^T \mathbf{w})$$

$$= \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \mathbf{x}_i$$

Chain Rule:

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$$

Let $f(x_i, w) - y_i$ be a

Let $f(x_i, w)$ be u

$$\frac{da}{dw} = \frac{da}{du} \times \frac{du}{dw}$$

Lecture 5 slide 8

b^T

$$\frac{d(\boxed{\mathbf{y}^T\mathbf{A}}\mathbf{x})}{d\mathbf{x}} = \underline{\mathbf{A}^T\mathbf{y}}$$

Q4

Repeat Question 3 using $f(x, w) = \sigma(x^T w)$,

where $\sigma\,(a) = \frac{1}{1+\exp(-\beta a)}$

Q4

$$
\begin{aligned}
\nabla_{\mathbf{w}} C(\mathbf{w}) &= \nabla_{\mathbf{w}} \sum_{i=1}^{m} (f(\mathbf{x}_i, \mathbf{w}) - y_i)^4 \\
&= \sum_{i=1}^{m} \nabla_{\mathbf{w}} (f(\mathbf{x}_i, \mathbf{w}) - y_i)^4 \\
&= \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \nabla_{\mathbf{w}} f(\mathbf{x}_i, \mathbf{w}) \qquad \text{chain rule} \\
&= \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \nabla_{\mathbf{w}} \sigma(\mathbf{x}_i^T \mathbf{w}) \\
&= \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \frac{\partial \sigma(a)}{\partial a} \nabla_{\mathbf{w}} (\mathbf{x}_i^T \mathbf{w}) \qquad \text{chain rule} \\
&= \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \frac{\partial \sigma(a)}{\partial a} \mathbf{x}_i
\end{aligned}
$$

Q4

So we just have to evaluate $\frac{\partial \sigma(a)}{\partial a}$ and plug it into the above equation. Note that $\frac{\partial \sigma(a)}{\partial a}$ is evaluated at $a = \mathbf{x}_i^T \mathbf{w}$, so

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

From previous slide

$$\nabla_{\mathbf{w}} C(\mathbf{w}) = \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \frac{\partial \sigma(a)}{\partial a}\mathbf{x}_i$$

Therefore,

$$\nabla_{\mathbf{w}} C(\mathbf{w}) = \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \beta\sigma(\mathbf{x}_i^T\mathbf{w})(1-\sigma(\mathbf{x}_i^T\mathbf{w}))\mathbf{x}_i$$

Q5

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

Q5

$$\nabla_{\mathbf{w}} C(\mathbf{w}) = \nabla_{\mathbf{w}} \sum_{i=1}^{m} (f(\mathbf{x}_i, \mathbf{w}) - y_i)^4$$

First part is the same as Q4

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

$$= \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \nabla_{\mathbf{w}} f(\mathbf{x}_i, \mathbf{w}) \qquad \text{chain rule}$$

$$= \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \nabla_{\mathbf{w}} \sigma(\mathbf{x}_i^T \mathbf{w})$$

$$= \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \frac{\partial \sigma(a)}{\partial a} \nabla_{\mathbf{w}} (\mathbf{x}_i^T \mathbf{w}) \qquad \text{chain rule}$$

$$= \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \frac{\partial \sigma(a)}{\partial a} \mathbf{x}_i$$

Q5

$$\sigma\ (a) = \max(0, a)$$

So we just have to evaluate $\frac{\partial \sigma(a)}{\partial a}$ and plug it into the above equation. Note that $\frac{\partial \sigma(a)}{\partial a}$ is evaluated at $a = \mathbf{x}_i^T \mathbf{w}$. When $a < 0$, $\sigma(a) = 0$, so $\frac{\partial \sigma(a)}{\partial a} = 0$. When $a > 0$, $\sigma(a) = a$, so $\frac{\partial \sigma(a)}{\partial a} = 1$. Let us define $\delta(\mathbf{x}_i^T \mathbf{w} > 0) = \begin{cases} 1 & \text{if } \mathbf{x}_i^T \mathbf{w} > 0 \\ 0 & \text{if } \mathbf{x}_i^T \mathbf{w} < 0 \end{cases}$, so we get

$$\frac{\partial \sigma(a)}{\partial a} = \delta(\mathbf{x}_i^T \mathbf{w} > 0)$$

From previous slide

$$\nabla_{\mathbf{w}} C(\mathbf{w}) = \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \frac{\partial \sigma(a)}{\partial a} \mathbf{x}_i$$

Therefore, $$\nabla_{\mathbf{w}} C(\mathbf{w}) = \sum_{i=1}^{m} 4(f(\mathbf{x}_i, \mathbf{w}) - y_i)^3 \mathbf{x}_i \delta(\mathbf{x}_i^T \mathbf{w} > 0),$$

THANK YOU