



MONASH
University

FIT 5215 Deep Learning

Quiz for:
Stochastic Gradient Descent and Optimization

Tutor Team

Department of Data Science and AI
Faculty of Information Technology, Monash University
Email: nglm@monash.edu/tuananh.bui@monash.edu



Question 1

- ❑ Consider the optimization problem to train a feed-forward NN

$$\min_{\theta} J(\theta) = \Omega(\theta) + \frac{1}{N} \sum_{i=1}^N CE(y_i, f(x_i; \theta))$$

where $\theta = [(W^k, b^k)]_{k=1}^L$ and $\Omega(\theta) = \lambda \sum_k \sum_{i,j} (W_{i,j}^k)^2 = \lambda \sum_k \|\mathbf{W}^k\|_F^2$. Choose the correct answers. (MC)

- ❑ A. Minimizing $\Omega(\theta)$ encourages more weights $W_{i,j}^k$ to approach to 0.
- ❑ B. Minimizing $\Omega(\theta)$ encourages complex models.
- ❑ C. Minimizing $\Omega(\theta)$ encourages simple models.
- ❑ D. Minimizing $\Omega(\theta)$ combats underfitting.
- ❑ E. Minimizing $\Omega(\theta)$ combats overfitting.

Question 2

- Consider the optimization problem to train a feed-forward NN

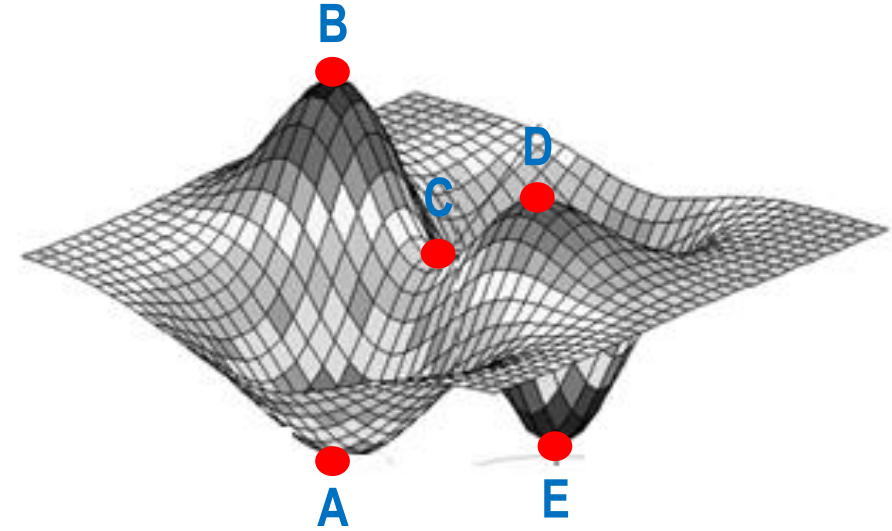
$$\min_{\theta} J(\theta) = \Omega(\theta) + \frac{1}{N} \sum_{i=1}^N CE(y_i, f(x_i; \theta))$$

where $\theta = [(W^k, b^k)]_{k=1}^L$ and $f(x_i; \theta)$ returns the prediction probabilities for x_i . Choose the correct answers. (MC)

- A. $\frac{1}{N} \sum_{i=1}^N CE(y_i, f(x_i; \theta))$ is known as a regularization term.
- B. $\frac{1}{N} \sum_{i=1}^N CE(y_i, f(x_i; \theta))$ is known as an empirical loss.
- C. Minimizing $\frac{1}{N} \sum_{i=1}^N CE(y_i, f(x_i; \theta))$ makes the model more fit to the training set.
- D. Minimizing $\frac{1}{N} \sum_{i=1}^N CE(y_i, f(x_i; \theta))$ makes the model more fit to the testing set.
- E. Minimizing $\frac{1}{N} \sum_{i=1}^N CE(y_i, f(x_i; \theta))$ can lead to overfitting.

Question 3

☐ Given a loss surface as showing, which statements are correct? (MC)



- ☐ A. **A**, **B**, **C**, **D**, and **E** are critical points.
- ☐ B. **A**, **E**, and **C** are local minima, while **B**, **D** are local maxima.
- ☐ C. **B** and **D** are local minima, **A** and **C** are local maxima, and **C** is a saddle point.
- ☐ D. **C** is a saddle point, **B** and **D** are local maxima, while **A** and **E** are local minima.
- ☐ E. **B** is global maxima, **D** is local maxima, and **C** is a saddle point.

Question 4

- ☐ Let $f(w) = w^2 - 3w + 1$. Assume that we use gradient descent with the learning rate $\eta = 0.1$ to solve $\min_w f(w)$. At the iteration t , we are at $w_t = 2$. What is the value of w_{t+1} at the next iteration? (SC)
- ☐ A. $w_{t+1} = 1$.
- ☐ B. $w_{t+1} = 1.8$
- ☐ C. $w_{t+1} = 1.9$.
- ☐ D. $w_{t+1} = 2.1$.

Question 5

- Given the function $f(w) = \frac{1}{1000} \sum_{i=1}^{1000} (w - x_i)^2$ where $x_i = i, \forall i = 1, \dots, 1000$. We need to solve $\min_w f(w)$ using stochastic gradient descent with the learning rate $\eta = 0.1$. Assume we sample a batch $b_1 = 1, b_2 = 3, b_3 = 5, b_4 = 7$ of indices and at the iteration t , we have $w_t = 10$. What is the value of w_{t+1} at the next iteration? (SC)
- A. $w_{t+1} = 8.9$.
- B. $w_{t+1} = 8.7$.
- C. $w_{t+1} = 8.5$.
- D. $w_{t+1} = 8.8$.

Question 6

□ Consider the optimization problem: $\min_{\theta} L(D; \theta) := \frac{1}{N} \sum_{i=1}^N l(x_i, y_i; \theta)$ with θ is the model parameter and $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$ is a training set. Let us sample a batch of indices i_1, \dots, i_b uniformly from $\{1, \dots, N\}$. Which statements are correct about the update rule of stochastic gradient descent? (SC)

- A. $\theta_{t+1} = \theta_t - \frac{\eta}{N} \sum_{i=1}^N \nabla_{\theta} l(x_i, y_i; \theta_t)$.
- B. $\theta_{t+1} = \theta_t + \frac{\eta}{N} \sum_{i=1}^N \nabla_{\theta} l(x_i, y_i; \theta_t)$.
- C. $\theta_{t+1} = \theta_t - \frac{\eta}{b} \sum_{k=1}^b \nabla_{\theta} l(x_{i_k}, y_{i_k}; \theta_t)$.
- D. $\theta_{t+1} = \theta_t + \frac{\eta}{b} \sum_{k=1}^b \nabla_{\theta} l(x_{i_k}, y_{i_k}; \theta_t)$.

Question 7

Given 4 implementations as below. What are f1/f2/f3/f4? (SC).

- ☐ A. sigmoid/tanh/softmax/relu
- ☐ B. softmax/tanh/relu/sigmoid
- ☐ C. sigmoid/tanh/relu/softmax
- ☐ D. softmax/tanh/sigmoid/relu

```
def f1(x):  
    return 1. / (1 + np.exp(-x))  
  
def f2(x):  
    return (np.exp(2*x)-1.) / (np.exp(2*x)+1.)  
  
def f3(x):  
    return x*(x>0)  
  
def f4(x):  
    return np.exp(x) / np.sum(np.exp(x))
```


Question 8

Which is $\frac{\partial h}{\partial \bar{h}}$ if $h = \sigma(\bar{h})$, \bar{h} is a vector and σ is an activation function? (SC)

- ☐ A. $\sigma'(\bar{h})$
- ☐ B. $\text{diag}(\bar{h})$
- ☐ C. $\text{diag}(\sigma(\bar{h}))$
- ☐ D. $\text{diag}(\sigma'(\bar{h}))$

Question 9

Which is $\frac{\partial h}{\partial \bar{h}}$ if $h = \sigma(\bar{h})$, $\bar{h} = [-1, 1, 2]$ and σ is ReLU activation function? (SC)

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

$$C = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad D = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Question 10

Which is $\frac{\partial l}{\partial \bar{h}}$ if $h = \sigma(\bar{h})$, $\bar{h} = [-1, 1, 2]$, σ is ReLU activation function and $\frac{\partial l}{\partial h} = [1, 1, 1]$?
(SC)

- A. $[-1, 1, 2]$
- B. $[0, 1, 2]$
- C. $[0, 1, 1]$
- D. $[-1, 1, 1]$

Question 11

Which is $\frac{\partial l}{\partial W}$ if $\bar{h} = Wx + b$, $\frac{\partial l}{\partial \bar{h}} = [0, 1, 1]$, $x = [1, 2, 3]^T$, $b = [0, 1, 2]^T$? (SC)

$$A = 6$$

$$B = [1, 2, 3]$$

$$C = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix}$$

$$D = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix}$$