Roll No:.....

Name:.....

Indian Institute of Information Technology Kottayam

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY KOTTAYAM

Department of Computational Science & Humanities

MID TERM EXAMINATION- SEPTEMBER, 2024

COURSE TITLE: IMA 314 (Optimization Techniques for Data Science)

Time: 2:30 PM-4:00 PM (Date 24/09/2024) **Max. Marks: 50**

Course instructor: Dr. Susheel Kumar Joshi **Batch (2022 AI&DS)**

Answer all questions.

Part A: Each question carries 2 marks. $5 \times 2 = 10$

1: Subgradient of the function $f(x_1, x_2) = |x_1| + 2|x_2|$ at point (0,0) is _____

2: Find the directional derivative of $f(x, y) = x^2 + 3xy^2$ at $(2, -2)$ in the direction of the unit vector $u = \begin{bmatrix} \frac{3}{5} & \frac{4}{5} \end{bmatrix}^T$. Is this direction a descent direction? Justify your answer.

3: Given the following Hessian matrix of a function $f(x_1, x_2)$:
$H = \begin{bmatrix} 5 & 0 \\ 0 & -6 \end{bmatrix}$. Is $f(x_1, x_2)$ convex? Justify your answer.

4: Is it true or false that Gradient Descent always guarantees to find the global minimum for non-convex loss functions? Justify your answer.

5: Given the weight vector $W = [3 \ -4 \ \ 0 \ \ 2\ ]^T$. Calculate $\| W \|_1$ and $\| W \|_2$.

Part B: Each question carries 5 marks. $4 \times 5 = 20$

1: Given the following gradient values of the loss function $L(W)$ with respect to the model weight vector $W \in \mathbb{R}^2$ over two iterations:

$$\nabla L(W)^{(1)} = [0.4 \ -0.3]^T, \nabla L(W)^{(2)} = [-0.2 \ \ 0.1]^T$$

(a): Apply the Momentum method to obtain the weight updates over two iterations using the following settings: Initial weight vector $W^{(0)} = [1 \ \ 1]^T$, learning rate $\alpha = 0.1$, momentum cofficient $\beta = 0.9$, initial velocity $v^{(0)} = [0 \ \ 0]^T$. (Provide all intermediate steps with concepts). [3 marks]

(b): What are the limitations of the Momentum method, and how does the Nesterov Accelerated Gradient method overcome these limitations? [2 marks]

2: You are implementing the Exponentially Weighted Moving Average (EWMA) to accumulate gradients for a weight vector $W \in \mathbb{R}^2$ during model training. Given the gradient values of the loss function $L(W)$ with respect to W over four iterations as follows:

$\nabla L(W)^{(1)} = [0.1 \ \ -0.2]^T, \nabla L(W)^{(2)} = [0.2 \ \ -0.1]^T$,
$\nabla L(W)^{(3)} = [-0.1 \ \ 0.1]^T, \nabla L(W)^{(4)} = [0.3 \ \ -0.2]^T$

(a): With initial moving average $G^{(0)} = [0 \ \ 0]^T$ and decay facor $\beta = 0.9$, Calculate the moving average of the gradients after each iteration using EWMA. Provide the final moving average after

PTO

the 4^{th} iteration. [3 marks]
(b): Discuss the impact of changing β on the sensitivity of the moving average to recent gradients? (Provide all intermediate steps with concepts). [2 marks]

3: (a): Explain the Armijo rule for selecting the step size α in Gradient Descent to ensure sufficient decrease. [2 marks]
(b): Consider the function $f(x,y) = x^2 + y^2$ with given initial point $(x_0, y_0) = (3,4)$. Apply Gradient Descent with Armijo rule as a line search method for one iteration with the setting: Initial step size $\alpha^{(0)} = 1$, Sufficient decrease parameter $c_1 = 0.1$, Reduction factor $\beta = 0.4$ (Provide all intermediate steps with concepts). [3 marks]

4: Consider the two dimensional function $f(x,y) = 3x^2 + 2xy + y^2$
(a): Compute the gradient $\nabla f(x,y)$ at point $(1,2)$. What does the obtained numerical value represent? [2.5 marks]
(b): Compute the Hassian $\nabla^2 f(x,y)$ at point $(1,2)$. What does the Hessian represent? [2.5 marks]

Part C: Each question carries 10 marks. **$2 \times 10 = 20$**

1: You are working on a real estate price prediction model (linear regression model) using the following dataset:
Feature matrix $X = \begin{bmatrix} 1 & 0 \\ 0 & 5 \\ 2 & 0 \end{bmatrix}$ and Target price $Y = \begin{bmatrix} 3 \\ 7 \\ 5 \end{bmatrix}$.
(a): Why are sparse datasets problematic for traditional gradient optimization methods, and how does Adam algorithm address these issues more effectively? [3 marks]
(b): Apply the Adam algorithm to train the model with the following setting:
$W^{(0)} = [0 \quad 0]^T$, Bias term $b = 0$, $\alpha = 0.01, \beta_1 = 0.9, \beta_2 = 0.99$ and $\varepsilon = 10^{-8}$. Stopping criteria: maximum iterations = 2 (Provide all intermediate steps with concepts). [7 marks]

2: You are training a logistic regression model to predict whether a device will pass or fail a quality check based on two features: temperature and pressure. The dataset is as follows:
Feature matrix $X = \begin{bmatrix} 2.0 & 0.3 \\ 2.5 & 0.4 \end{bmatrix}$. Target labels $Y = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ where 0 means the device failed the test and 1 means the device passed.
With $W^{(0)} = [0.1 \quad -0.05]^T$, Bias term $b = 0.5$ and $\alpha = 0.01$:
(a): Calculate the weight updates using Gradient Descent (GD) after two complete iterations over the entire dataset (Provide all intermediate steps with concepts). [4 marks]
(b): Calculate the weight updates using Stochastic Gradient Descent (SGD) after processing a single randomly chosen data point for two iterations (Provide all intermediate steps with concepts). How does the update differ from the GD update? [6 marks]

****

PTO