**CSE422 Problem Set**
**Linear Regression and Gradient Descent.**
**Instructor: Ipshita Bonhi Upoma**

## Part 1: Basic Computation

1. Suppose you have a dataset with three training examples given by $(x, y)$:

$$(1, 2), \quad (2, 3), \quad (3, 6).$$

You are using the hypothesis (model)

$$h_\theta(x) = \theta_0 + \theta_1 x$$

and your parameters are initialized as $\theta_0 = 0$ and $\theta_1 = 1$. The mean squared error (MSE) Loss function is:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \big(h_\theta(x^{(i)}) - y^{(i)}\big)^2,$$

where $m = 3$ is the number of training samples.

   (a) Compute the predictions $\hat{y}$ for each $x$ using $\theta_0 = 0$ and $\theta_1 = 1$.
   (b) Compute the squared error for each example.
   (c) Compute the final cost $J(\theta_0, \theta_1)$.

2. Using the same data as above $\{(1, 2), (2, 3), (3, 6)\}$ and the same hypothesis $h_\theta(x) = \theta_0 + \theta_1 x$, assume the Loss function is again the MSE form. Let the learning rate (alpha) be $\alpha = 0.1$. Suppose your initial parameters are $\theta_0 = 0$ and $\theta_1 = 1$.

   (a) Write down the partial derivative expressions for

$$\frac{\partial J}{\partial \theta_0} \quad \text{and} \quad \frac{\partial J}{\partial \theta_1}.$$

   (b) Compute the partial derivatives given the three data points.
   (c) Perform one gradient descent update step, i.e., compute the new

$$\theta_0^{(\text{new})} = \theta_0 - \alpha \frac{\partial J}{\partial \theta_0}, \quad \theta_1^{(\text{new})} = \theta_1 - \alpha \frac{\partial J}{\partial \theta_1}.$$

   (d) Report the updated values of $\theta_0$ and $\theta_1$.

## Part 2: Simulation on given Dataset

**Problem:** Let the hypothesis model be,

$$h_\theta(x) = \theta_0 + \theta_1 x.$$

where, $\theta_0$ and $\theta_1$ are the parameters. The Loss function be of the MSE form:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

where, $m$ is the number of examples. Starting with $\theta_0 = 0$ and $\theta_1 = 1$, simulate the gradient descent algorithm on the following Dataset 1, Dataset 2, Dataset 3 for

1. The learning parameter, $\alpha = 0.1$.
2. Decaying learning parameter, $\alpha = \frac{1000}{1000+t}$, where $t$ is the iteration number.

- **Dataset 1:** A single-feature dataset where the feature is the number of hours a student studies, and the target is the student's test score on a 100-point exam.

| **Hours Studied** ($x$) | **Test Score** ($y$) |
|---|---|
| 1.0 | 45 |
| 2.0 | 50 |
| 3.0 | 60 |
| 4.0 | 72 |
| 5.0 | 80 |

Table 1: Hours studied vs. test score dataset.

**Interpretation:**

$$x = \text{Hours studied}, \quad y = \text{Final test score}.$$

- **Dataset 2:** A single-feature dataset where the feature is the size of a house in square feet, and the target is its price (in thousands of dollars).

| **House Size** ($x$, sq ft) | **House Price** ($y$, \$1000) |
|---|---|
| 800 | 120 |
| 900 | 135 |
| 1000 | 150 |
| 1200 | 185 |
| 1500 | 230 |

Table 2: House size vs. house price dataset.

**Interpretation:**

$$x = \text{House size (sq ft)}, \quad y = \text{House price (\$1,000)}.$$

- **Dataset 3:** A single-feature dataset where the feature is the age of a car and the target is its price.

| **Car Age** ($x$, years) | **Selling Price** ($y$, \$1000) |
|---|---|
| 1 | 22 |
| 2 | 18 |
| 4 | 15 |
| 5 | 12 |
| 7 | 8 |

Table 3: Car age vs. selling price.

**Interpretation:**

$$x = \text{Age of car (years)}, \quad y = \text{Selling price (\$1,000)}.$$

### Part 3: Find the answers to these conceptual questions

1. Explain the principle of gradient descent. What are the roles of the learning rate and the number of iterations in this algorithm?
2. What happens if the learning rate is too large?
3. What happens if the learning rate is too small?
4. Explain the differences between gradient descent, stochastic gradient descent (SGD), and mini-batch gradient descent.
5. What does it mean for gradient descent to converge? How can you tell if gradient descent has failed to converge in practice?
6. What are some potential drawbacks or challenges of using stochastic gradient descent, especially in terms of the final stages of convergence?
7. Describe the cost function used in linear regression. Why is it important to minimize this function?
8. What is the hypothesis function in linear regression? How does it differ in its formulation between simple linear regression and multiple linear regression?