HMC CS 158
Quiz 3: Linear Regression

1. Consider the problem of predicting how well a student does in her second year of college/university, given how well they did in their first year. Specifically, let x be equal to the number of "A" grades (including A-, A, and A+ grades) that a student receives in their first year of college (freshmen year). We would like to predict the value of y, which we define as the number of "A" grades they get in their second year (sophomore year).

   Consider the following training set of a small sample of different students' performances. Here each row is one training example.

x	y
1	0.5
2	1
4	2
0	0

   Recall that in linear regression, our hypothesis is $h_\theta(x) = \theta_0 + \theta_1 x$. What are the values of θ_0 and θ_1 that you would expect to obtain upon running gradient descent on this model? (Linear regression will be able to fit this data perfectly.)

   (a) $\theta_0 = 0.5$, $\theta_1 = 0$

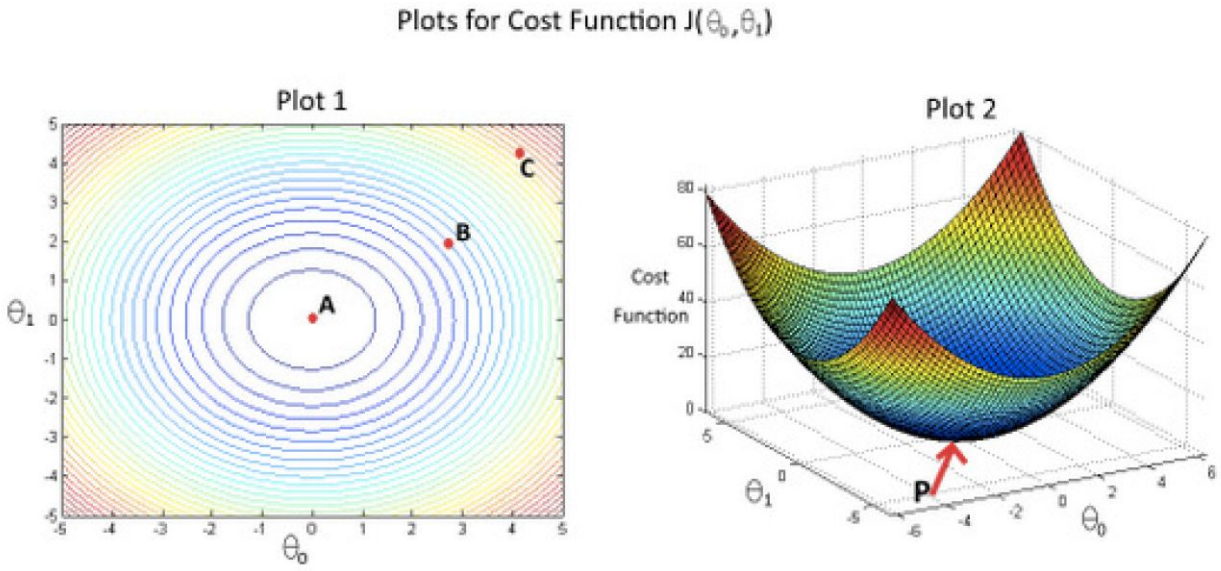
   (b) $\theta_0 = 1$, $\theta_1 = 0.5$

   (c) $\theta_0 = 0.5$, $\theta_1 = 0.5$

   (d) $\theta_0 = 0$, $\theta_1 = 0.5$

This quiz is adapted from course material by Andrew Ng (Stanford).

2. In the given figure, the cost function $J(\theta_0, \theta_1)$ has been plotted against θ_0 and θ_1, as shown in Plot 2. The contour plot for the same cost function is given in Plot 1. Based on the figure, choose the correct options (check all that apply).



(a) If we start from point B, gradient descent with a well-chosen learning rate will eventually help us reach at or near point A, as the value of cost function $J(\theta_0, \theta_1)$ is minimum at point A.

(b) If we start from point B, gradient descent with a well-chosen learning rate will eventually help us reach at or near point A, as the value of cost function $J(\theta_0, \theta_1)$ is maximum at point A.

(c) If we start from point B, gradient descent with a well-chosen learning rate will eventually help us reach at or near point C, as the value of cost function $J(\theta_0, \theta_1)$ is minimum at point C.

(d) Point P (the global minimum of Plot 2) corresponds to point A of Plot 1.

(e) Point P (the global minimum of Plot 2) corresponds to point C of Plot 1.

3. You run gradient descent for 15 iterations with $\alpha = 0.3$ and compute $J(\boldsymbol{\theta})$ after each iteration. You find that the value of $J(\boldsymbol{\theta})$ **increases** over time. Based on this, which of the following conclusions seems most plausible?

   (a) Rather than use the current value of α, it would be more promising to try a smaller value of α (say $\alpha = 0.1$).

   (b) Rather than use the current value of α, it would be more promising to try a larger value of α (say $\alpha = 1.0$).

   (c) $\alpha = 0.3$ is an effective choice of learning rate.

4. Suppose you have a dataset with $n = 1,000,000$ examples and $d = 15$ features for each example. You want to use multivariate linear regression to fit the parameters $\boldsymbol{\theta}$ to our data. Should you prefer gradient descent or the normal equation?

   (a) The normal equation, since it provides an efficient way to directly find the solution.

   (b) The normal equation, since gradient descent might be unable to find the optimal $\boldsymbol{\theta}$.

   (c) Gradient descent, since it will always converge to the optimal $\boldsymbol{\theta}$.

   (d) Gradient descent, since $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ will be very slow to compute in the normal equation.

5. Which of the following are reasons for using feature scaling?

   (a) It speeds up gradient descent by making it require fewer iterations to get to a good solution.

   (b) It speeds up solving for $\boldsymbol{\theta}$ using the normal equations.

   (c) It is necessary to prevent the normal equation from getting stuck in local optima.

   (d) It speeds up gradient descent by making each iteration of gradient descent less expensive to compute.