Name: _____

3 Regression

3. (15 points) (a) Reggie heard about standardizing features for classification and thought they'd try it for regression, too. Reggie has a one-dimensional linear regression data set (so $d = 1$) and so they decide to compute the transform

$$x_r^{(i)} = \frac{x^{(i)} - \mu(X)}{\text{SD}(X)}$$
$$y_r^{(i)} = \frac{y^{(i)} - \mu(Y)}{\text{SD}(Y)}$$

where $\mu(X)$ is the mean, or average, of the data values $x^{(i)}$ and $\text{SD}(X)$ is the standard deviation. Then, they perform ordinary least squares regression using the $(x_r^{(i)}, y_r^{(i)})$ data points, and get the parameters θ and θ_0.

Now they have to perform a transformation on θ and θ_0 to obtain the θ^*, θ_0^* that solve the original problem (that is, so that it will work correctly on the original $(x^{(i)}, y^{(i)})$ data).

Write an expression for θ^* in terms of X, $\mu(X)$, $\text{SD}(X)$, Y, $\mu(Y)$, $\text{SD}(Y)$, θ and θ_0.

Solution:

$$\theta^* = \frac{\text{SD}(Y)}{\text{SD}(X)}\theta$$

See next page for more detailed explanations.

(b) Reggie ran ridge regression using several different parameter settings, but scrambled the graphs! The dimension of the data is $d = 1$, so there are two parameters, θ and θ_0, which are the axes of the graphs. The contour lines indicate the value of the overall objective J, and the connected points indicate the trajectory of the (θ, θ_0) values during the process of gradient update. It always starts near $(0, 0)$, with θ plotted on the x axis and θ_0 on the y axis.

Which graph corresponds to which parameter settings?

- Step size: 0.05, 0.3, 0.7
- lambda : 0.0, 1.0

3a Explanation

Originally we had

$$y = \theta^* x + \theta_0^*$$

and now we have

$$y_r = \theta^T x_r + \theta_0$$

Expanding the second equation, we have

$$\frac{y-\mu(Y)}{SD(Y)} = \theta^T \frac{x-\mu(X)}{SD(X)} + \theta_0$$

which simplifies to

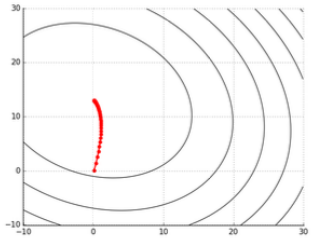
$$y = \frac{SD(Y)}{SD(X)}\theta^T(x - \mu(X)) + SD(Y)\theta_0 + \mu(Y)$$

The part that corresponds to θ^* is the coefficient of x above, so $\theta^* = \frac{SD(Y)}{SD(X)}\theta$.

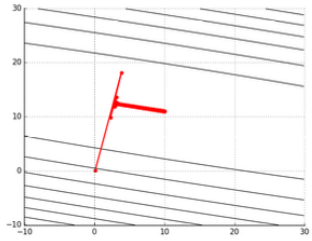
Intuitive explanation:
In this 1d linear regression, theta is just the slope (since theta.T x + theta_0 = y). When the y values get scaled by SD(Y), the slope of the data (x_r, y_r) gets scaled by SD(Y). When the x values get scaled by SD(X), the slope gets scaled by 1/SD(X). The mean (subtraction of a constant) has no effect on the slope since it is just shifts the data somewhere else on the graph (but it does affect theta_0). Since y_r is y_i divided by SD(Y), theta* gets multiplied by 1/SD(Y). Since x_r is x_i divided by SD(X), theta* gets multiplied by SD(X). So theta*(SD(X)/SD(Y)) = the new theta (the new slope). Solving for theta* gives theta*SD(Y)/SD(X).

Name: _____

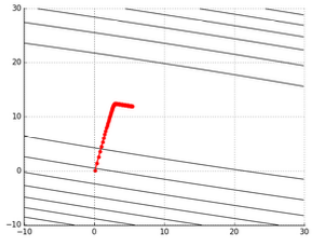
A lambda of zero (unregularized regression) will cause the optimal value of theta to approach infinity and the objective to approach 0. This can be seen in the plots with straight contour lines and increasing theta. The rest of the plots can be differentiated from each other according to the step size. Smaller step sizes have less oscillation and the red dots are closer together.



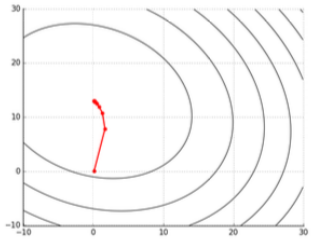
step size: **0.05** lambda: **1.0**



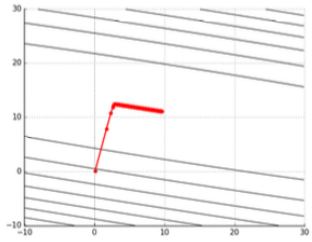
step size: **0.7** lambda: **0.0**



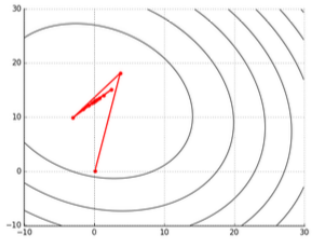
step size: **0.05** lambda: **0.0**



step size: **0.3** lambda: **1.0**



step size: **0.3** lambda: **0.0**



step size: **0.7** lambda: **1.0**

(c) We are considering formulating our machine-learning problem as an optimization problem with the following objective function

$$J(\theta) = \sum_{i=1}^{n} (\theta^T x^{(i)} - y^{(i)})^2 + \lambda R(\theta) \ ,$$

but we are not sure what regularizer R to use. For each of the possible choices listed below, answer the questions.

i. $R(\theta) = \sum_{j=1}^{d} \theta_j$
Is this equivalent to ridge regression? ◯ Yes √ **No**
Is this a reasonable choice for a regularizer? ◯ Yes √ **No**

ii. $R(\theta) = \sum_{j=1}^{d} |\theta_j|$
Is this equivalent to ridge regression? ◯ Yes √ **No**
Is this a reasonable choice for a regularizer? √ **Yes** ◯ No

iii. $R(\theta) = \sum_{j=1}^{d} \theta_j^2$
Is this equivalent to ridge regression? √ **Yes** ◯ No
Is this a reasonable choice for a regularizer? √ **Yes** ◯ No

iv. $R(\theta) = \sum_{j=1}^{d} \theta_j^3$
Is this equivalent to ridge regression? ◯ Yes √ **No**
Is this a reasonable choice for a regularizer? ◯ Yes √ **No**

v. $R(\theta) = \theta^T\theta$
Is this equivalent to ridge regression? √ **Yes** ◯ No
Is this a reasonable choice for a regularizer? √ **Yes** ◯ No