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

$$
\begin{aligned}
x_r^{(i)} &= \frac{x^{(i)} - \mu(X)}{\mathrm{SD}(X)} \\
y_r^{(i)} &= \frac{y^{(i)} - \mu(Y)}{\mathrm{SD}(Y)}
\end{aligned}
$$

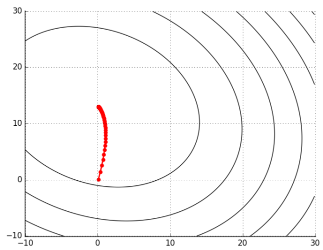
where $\mu(X)$ is the mean, or average, of the data values $x^{(i)}$ and $\mathrm{SD}(X)$ is the standard deviation. Then, they perform ordinary least squares regression using the $(x_r^{(i)}, y_r^{(i)})$ data points, and get the parameters θ and θ_0.

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, $\mu(X)$, $\mathrm{SD}(X)$, Y, $\mu(Y)$, $\mathrm{SD}(Y)$, θ and θ_0.

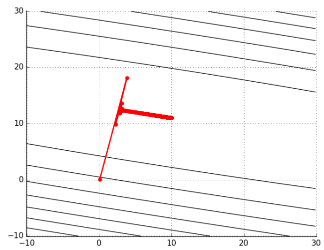
(b) Reggie ran ridge regression using several different parameter settings, but scrambled the graphs! The dimension of the data is $d = 1$, so there are two parameters, θ and θ_0, which are the axes of the graphs. The contour lines indicate the value of the overall objective J, and the connected points 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?

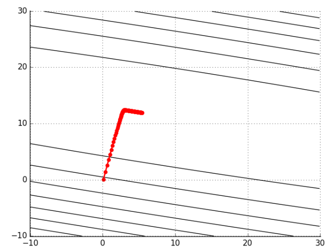
- Step size: 0.05, 0.3, 0.7
- lambda : 0.0, 1.0



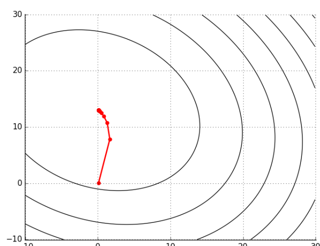
step size: ______________
lambda: ______________



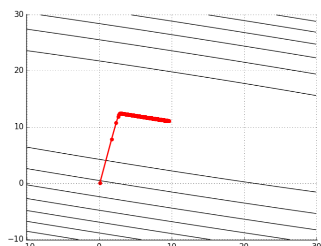
step size: ______________
lambda: ______________



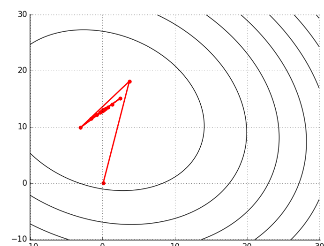
step size:______________
lambda: ______________



step size: ______________
lambda: ______________



step size: ______________
lambda: ______________



step size: ______________
lambda: ______________

(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