# LASSO & Ridge Regularization

## Regularizing Linear Models (Shrinkage methods)

When we have too many parameters and exposed to curse of dimensionality, we resort to dimensionality reduction techniques such as transforming to PCA and eliminating the PCA with least magnitude of eigen values. This can be a laborious process before we find the right number principal components. Instead, we can employ the shrinkage methods.

Shrinkage methods attempt to shrink the coefficients of the attributes and lead us towards simpler yet effective models. The two shrinkage methods are :

Ridge regression is similar to the linear regression where the objective is to find the best fit surface. The difference is in the way the best coefficients are found. Unlike linear regression where the optimization function is SSE, here it is slightly different

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

Linear Regression cost function

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

Ridge Regression with additional term in the cost function

The term $\lambda$ is like a penalty term used to penalize large magnitude coefficients $\beta_j$ when it is set to a high number, coefficients are suppressed significantly. When it is set to 0, the cost function becomes same as linear regression cost function

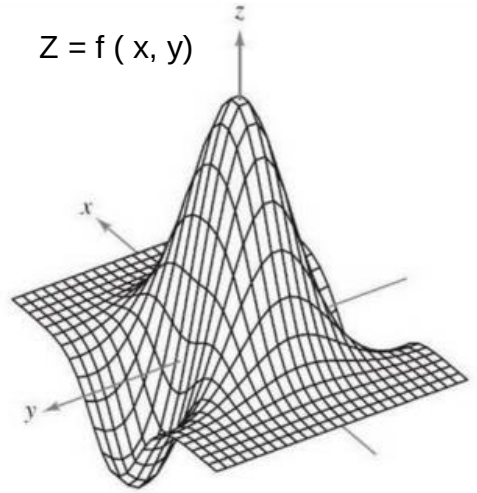
## Regularizing Linear Models (Shrinkage methods)

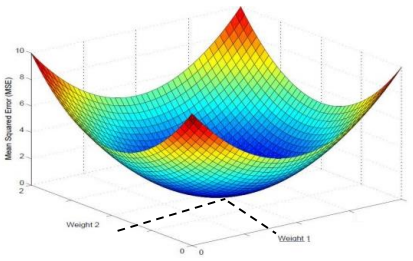
Why should we be interested in shrinking the coefficients? How does it help?

When we have large number of dimensions and few data points, the models are likely to become complex, over fit and prone to variance errors. When you print out the coefficients of the attributes of such complex model, you will notice that the magnitude of the different coefficients become large

Large coefficients indicate a case where for a unit change in the input variable, the magnitude of change in the target column is very large.

## Regularizing Linear Models (Shrinkage methods)

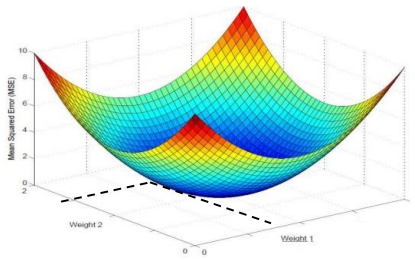



$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2=0$$

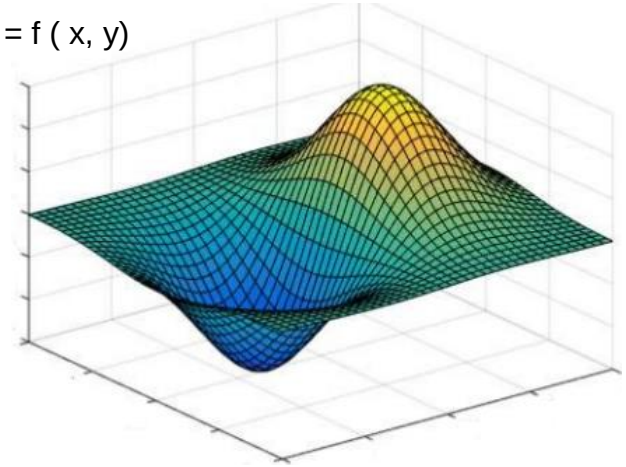
1. Curse of dimensionality results in large magnitude coefficients which results in a complex undulated surface / model.
2. This complex surface has the data points occupying the peaks and the valleys
3. The model gives near 100% accuracy in training but poor result in testing and the testing scores also vary a lot from one sample to another.
4. The model is supposed to have absorbed the noise in the data distribution!
5. Large magnitudes of the coefficient give the least SSE and at times SSE = 0! A model that fits the training set 100%!
6. Such models do not generalize

## Regularizing Linear Models (Shrinkage methods)




$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

Z = f ( x, y)



1. In Ridge Regression, the algorithm while trying to find the best combination of coefficients which minimize the SSE on the training data, is constrained by the penalty term

2. The penalty term is akin to cost of magnitude of the coefficients. Higher the magnitude, more the cost. Thus to minimize the cost, the coefficient are suppressed

3. Thus the resulting surface tends to be relatively much more smoother than the unconstrained surface. This means we have settled for a model which will make errors in the training data

4. This is fine as long as the errors can be attributed to the random fluctuations i.e. because the model does not absorb the random fluctuations in the data

5. Such model will perform equally well on unseen data i.e. test data. The model will generalize better than the complex model

## Regularizing Linear Models (Shrinkage methods)

1. Lasso Regression is similar to the Ridge regression with a difference in the penalty term. Unlike Ridge, the penalty term here is raised to power 1. Also known as L1 norm.

$$\sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} |w_j|$$

2. The term $\lambda$ continues to be the input parameter which will decide how high penalties would be for the coefficients. Larger the value more diminished the coefficients will be.

3. Unlike Ridge regression, where the coefficients are driven towards zero but may not become zero, Lasso Regression penalty process will make many of the coefficients 0. In other words, literally drop the dimensions