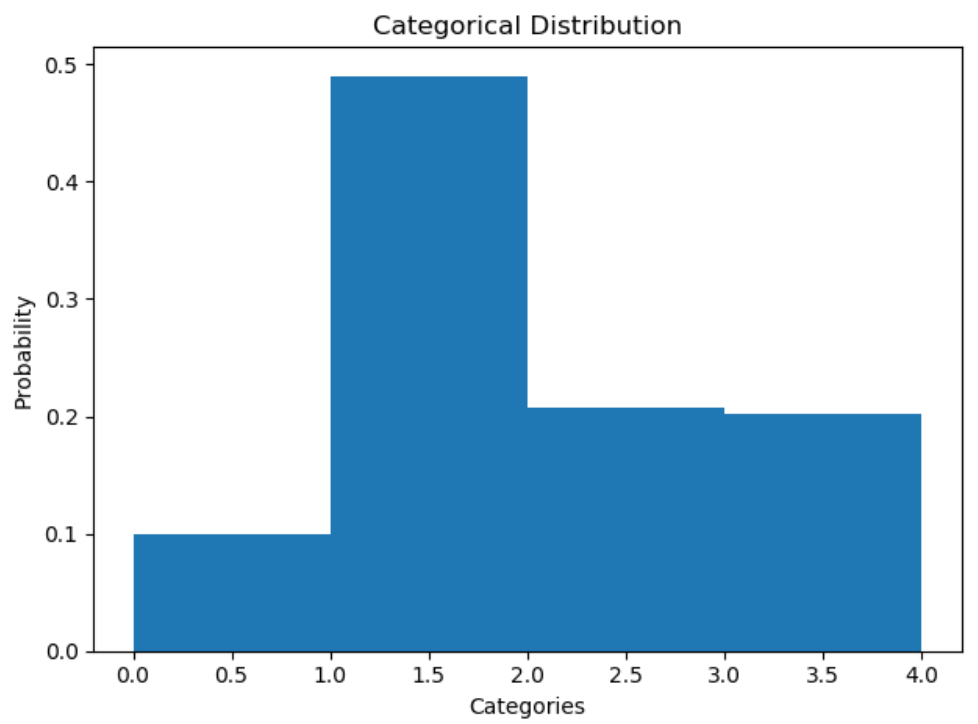
HW1

Zheyu Ding (zd12)

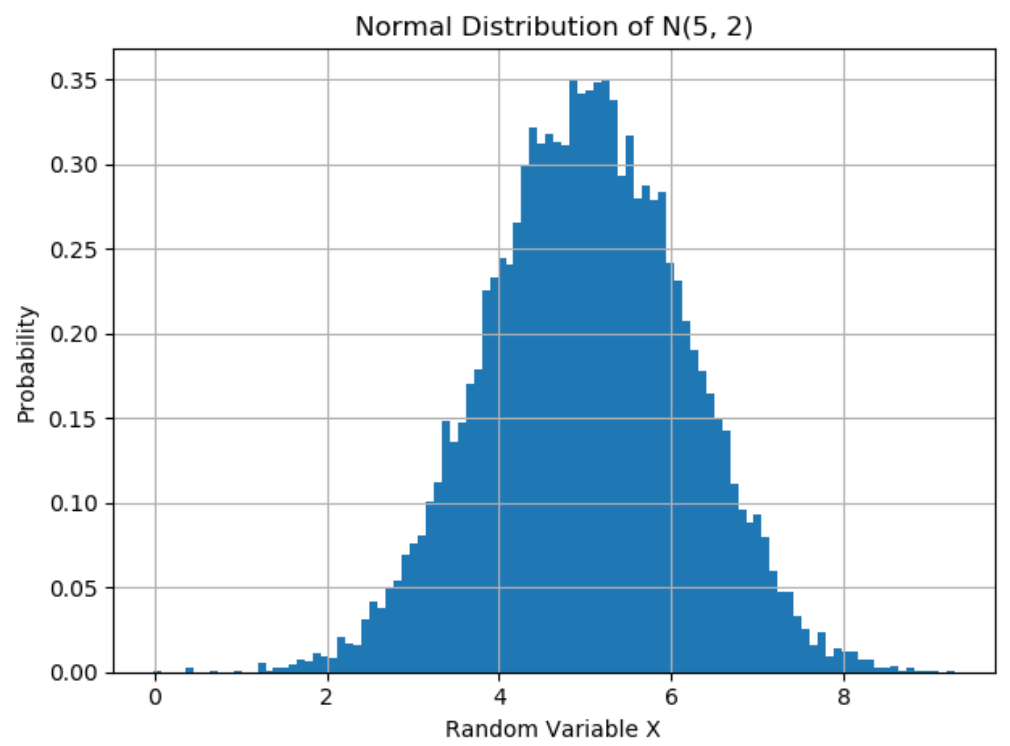
M.Viginesh Vaibhav (vm24)

**Problem 0: Background refresher**

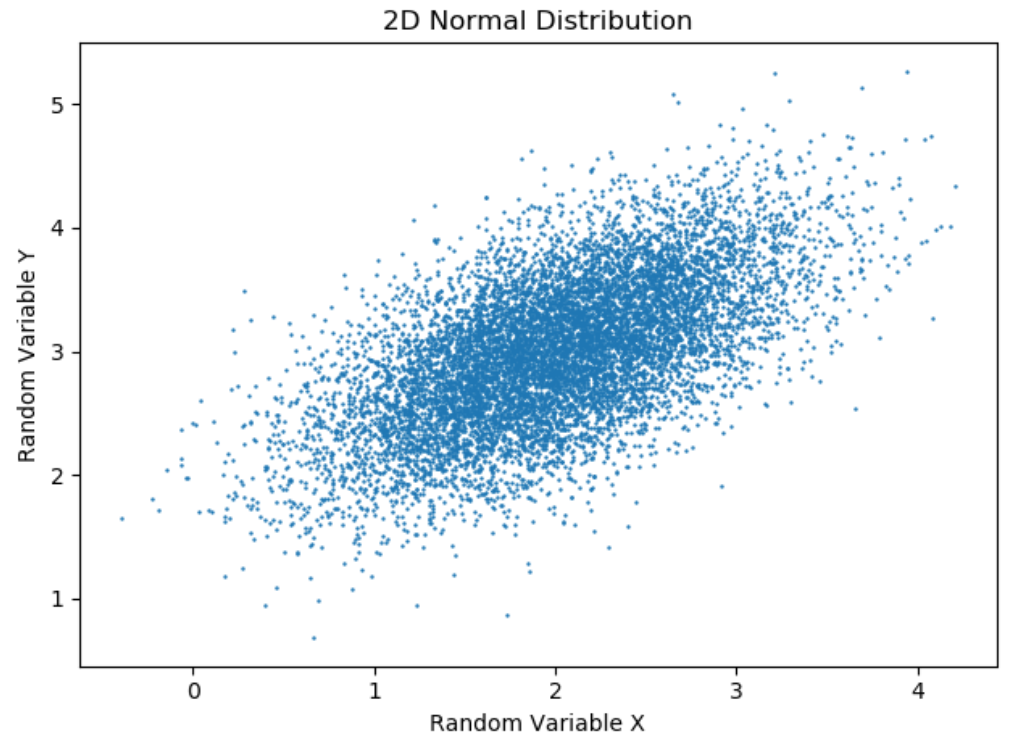
1. Plot 4 kinds of distribution
2. Plot the histogram of samples generated by a categorical distribution with probabilities [0.1,0.5,0.2,0.2]



1. Plot the univariate normal distribution with mean of 5 and standard deviation of 2.



1. Produce a scatter plot of the samples for a 2-D Gaussian with mean at [2,3] and a covariance matrix [[1,0.5], [0.2,1]].



1. Test your mixture sampling code by writing a function that implements an equal- weighted mixture of four Gaussians in 2 dimensions, centered at (±1, ±1) and having covariance I. Estimate the probability that a sample from this distribution lies within the unit circle centered at (0.1, 0.2) and include that number in your writeup.

The probability is 0.1689.

1. Prove that the sum of two independent Poisson random variables is also a Poisson random variable.

**Solution**

There are two random variables:

So, , , U=X+Y

That is to say .

1. Let and be continuous random variables. Show that if p ( = ) P ( = | = ) there exists α1, μ1 and σ1 such that. Write down expressions for these quantities in terms of α0, α, μ0, σ0 and σ.

there exists α1, μ1 and σ1 such that

Write down expressions for these quantities in terms of α0, α, μ0, σ0 and σ.

**Solution**

Because

So,

1. Find the eigenvalues and eigenvectors of the following 2 × 2 matrix A.

**Solution**

So, the eigenvalues of A are .

When eigenvector of A can be .

When eigenvector of A can be .

1. Provide one example for each of the following cases, where A and B are 2 × 2 matrices.

**Solution**

1. +2AB

, B =

1. AB=0, A0, B0

,

1. Let u denote a real vector normalized to unit length. That is, = 1. Show that

is orthogonal, i.e.,.

**Solution**

So, A is orthogonal.

1. A function f is convex on a given set S if and only if for λ ∈ [0,1] and for all

x, y ∈ S, the following holds.

Moreover, a univariate function f(x) is convex on a set S if and only if its second derivative ′′

f (x) is non-negative everywhere in the set. prove the following assertions.

**Solution**

1. is convex for. X ∈ R.

>0 X ∈ R

So, is convex.

1. is convex on .

is convex on

1. If univariate functions f and g are convex on S, then is convex on S.
2. If univariate functions and are convex and non-negative on S, and have their minimum within S at the same point, then is convex on S.

Then we can imply is convex on S.

1. The entropy of a categorical distribution on K values is defined as

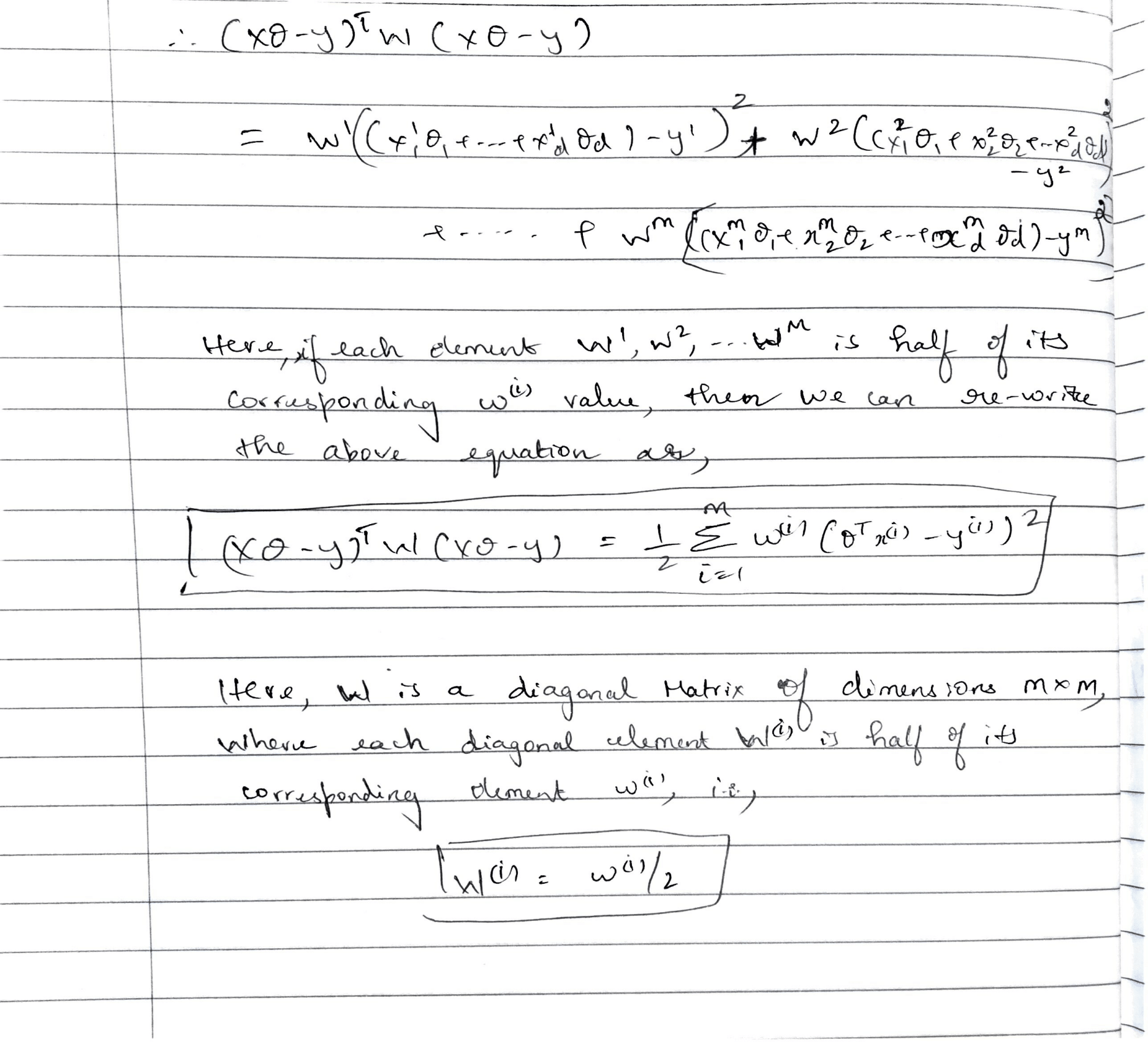
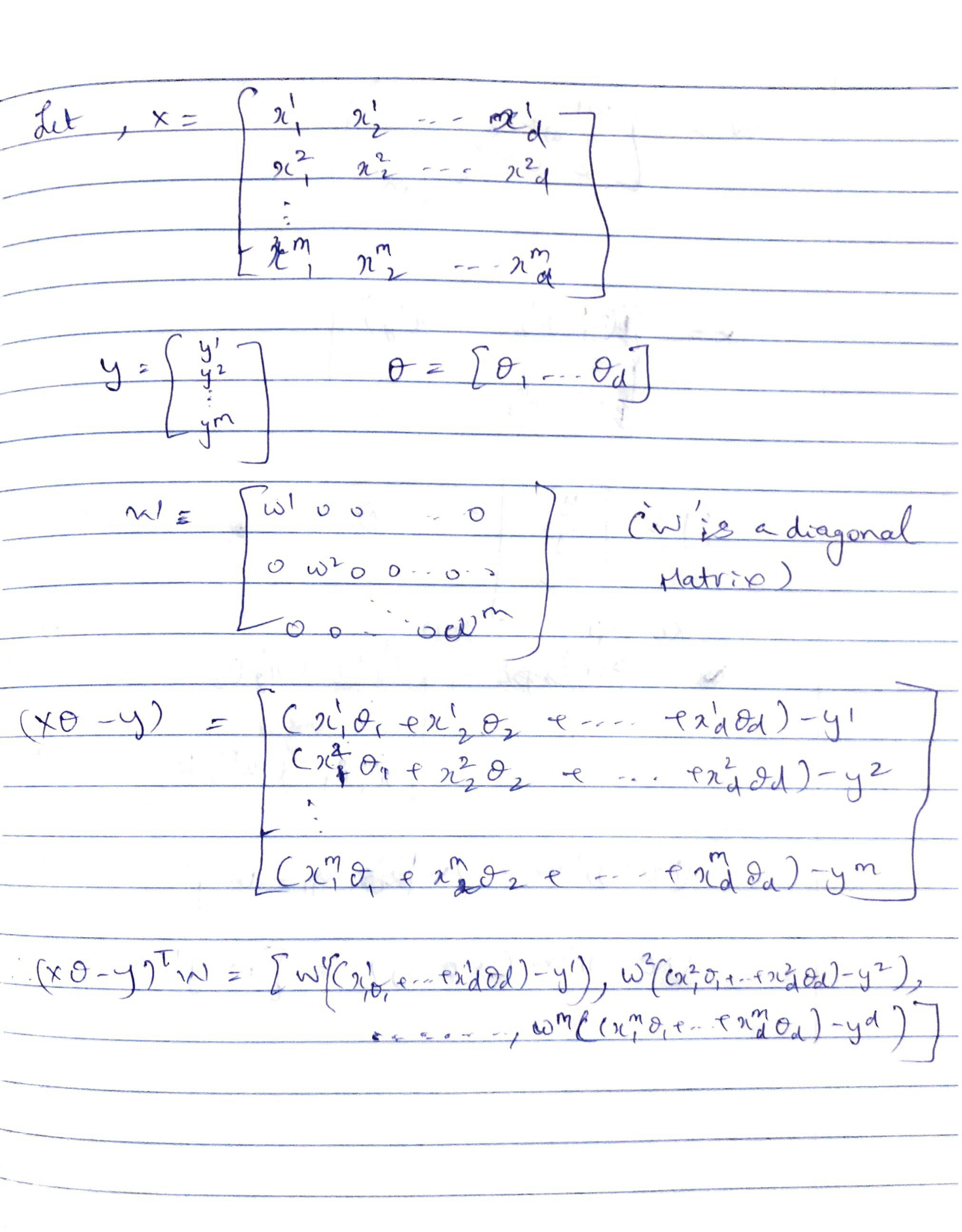
Using the method of Lagrange multipliers, find the categorical distribution that has the

highest entropy.

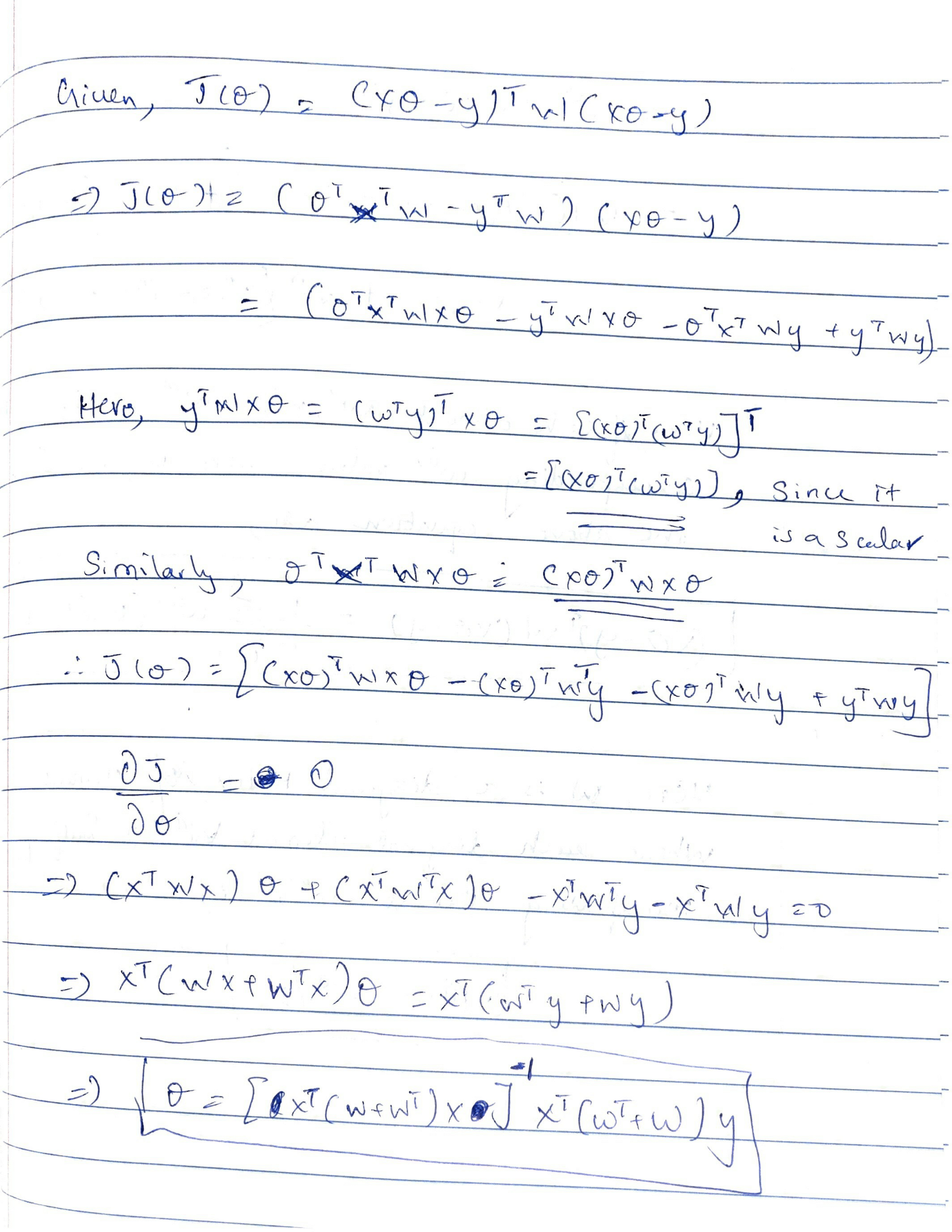
**Solution**

Therefore, the uniform distribution will get the highest entropy.

**Problem 1: Locally weighted linear regression**

1. The proof is attached in the pictures below. ‘W’ is a diagonal matrix of dimensions m x m, where each element of the diagonal is half of its corresponding w value, i.e., W(i) = w(i)/2.
2. The proof is attached in the picture below. The derived equation for is,





1. The algorithm for using batch gradient descent for locally weighted linear regression is as follows:
   1. Take the input vector x for the target to be predicted y
   2. Compute w(i) for every x(i) in the training data using x as the input vector, using the formula,

Here, is the bandwidth defining the sphere of influence around .

* 1. Initialize theta to a matrix of dimensions d x 1, with all the elements as zero.
  2. Compute the cost as,
  3. Compute the gradient as,
  4. Compute in the next step as,

Here, is the learning rate.

* 1. Repeat steps d to f over all of the training data, until J(θ) does not change, i.e., the value of converges.

Locally weighted linear regression is non-parametric, because the values of the model’s weights depends on the training dataset as well as the input of the test data.

**Problem 2: Properties of the linear regression estimator**

1. Show that for the least squares estimator.

**Solution**

So, for the least squares estimator.

1. Show that the variance of the least squares estimator is .

**Solution**

So, the variance of the least squares estimator is .

**Problem 3: Implementing linear regression and regularized linear regression**

**Problem 3.1: Implementing linear regression (45 points)**

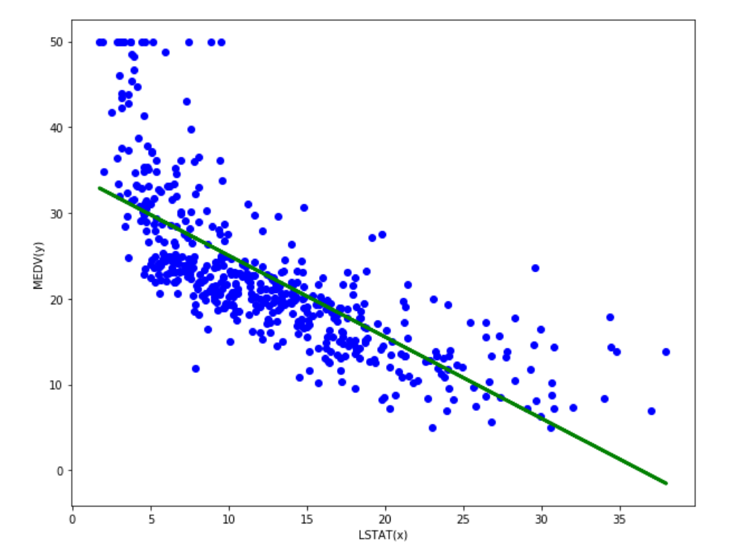
Problem 3.1.A: Linear regression with one variable (15 points)

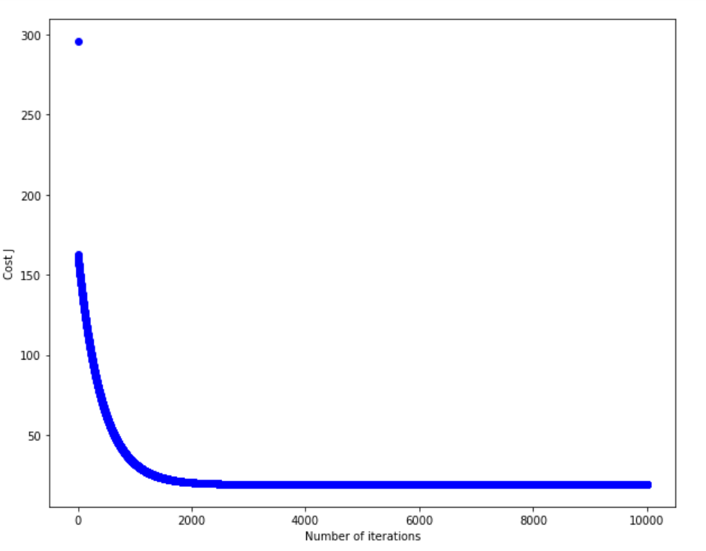
The predicted home values are as follows:

For **lower status percentage = 5**, we predict a median home value of **298034.49**

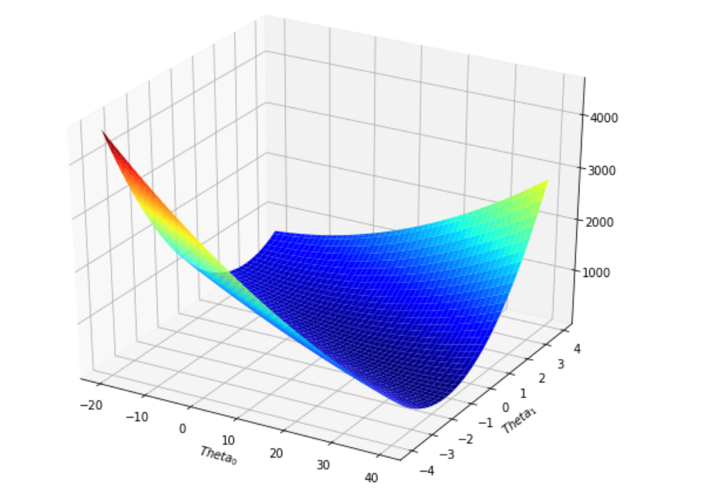
For **lower status percentage = 50**, we predict a median home value of **-129482.13**

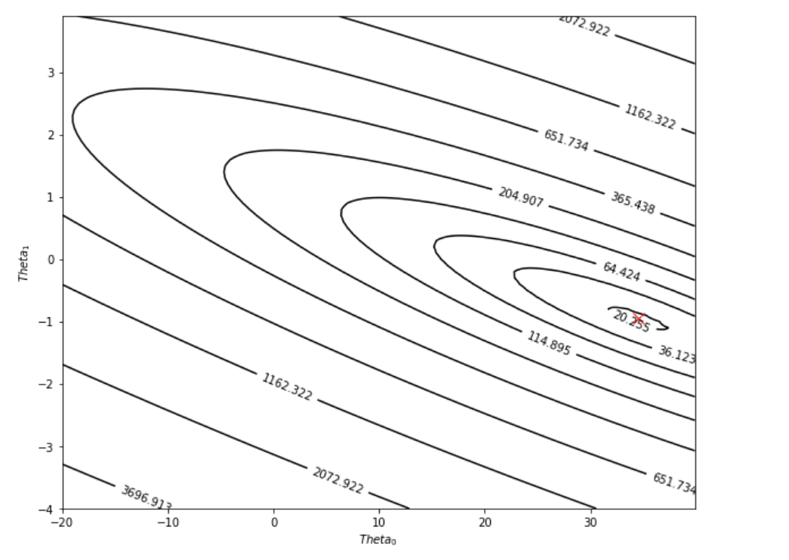
The plots produced for best fit curve and convergence of cost are as follows:





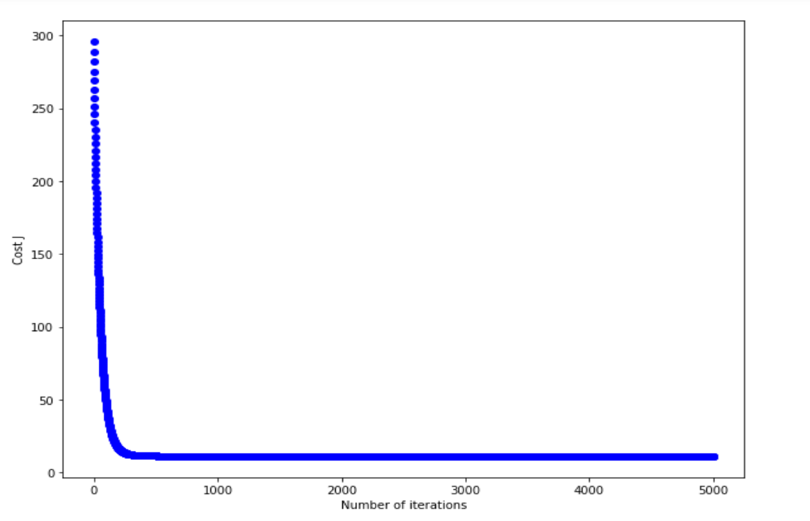
The visualization graphs of 0and θ1 are as follows:





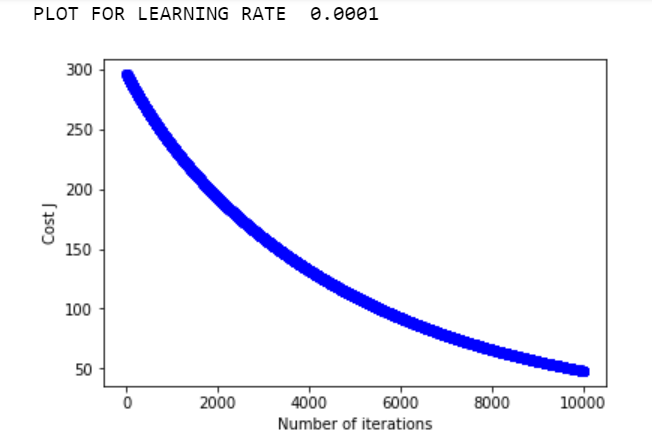
Problem 3.1.B: Linear regression with multiple variables

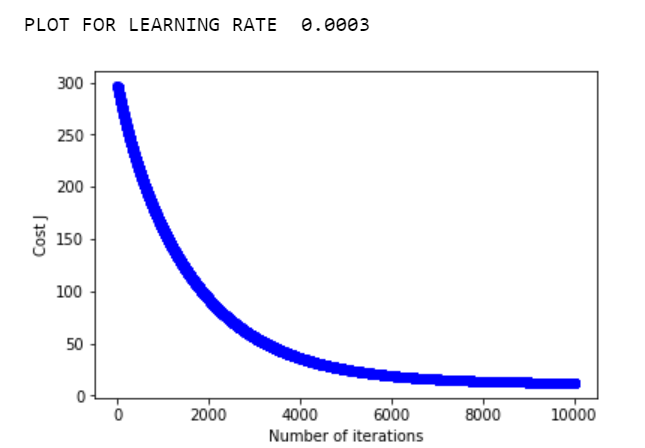
The plot showing the convergence of the cost with respect to the number of iterations is:



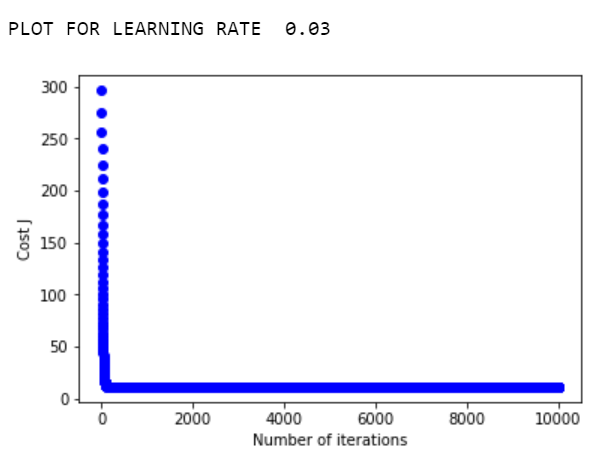
For the **average home in Boston** suburbs, we predict a median home value of **225328.06.** We obtain the same predicted result after computing theta using gradient descent and using the closed form solution.

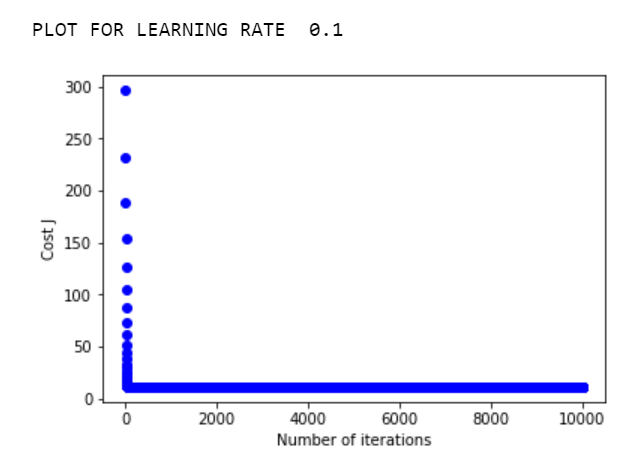
From what we have observed, the ideal learning rate for this problem is **0.01**, because it best resembles the curve we would like to obtain on the graph of convergence of cost versus the number of iterations. The initial cost isn’t too high for this learning rate, and it converges quickly. A similar convergence is obtained when we use a learning rate of 0.03, but the convergence occurs quickly and the curve does not resemble the shape we’d like to see. For higher learning rates such as 0.9, 1,5, and higher, the graph diverges and the cost increases exponentially with the number of iterations. For lower learning rates of 0.003 and 0.001, the convergence happens too slowly over the number of iterations. The following figures illustrate these points:

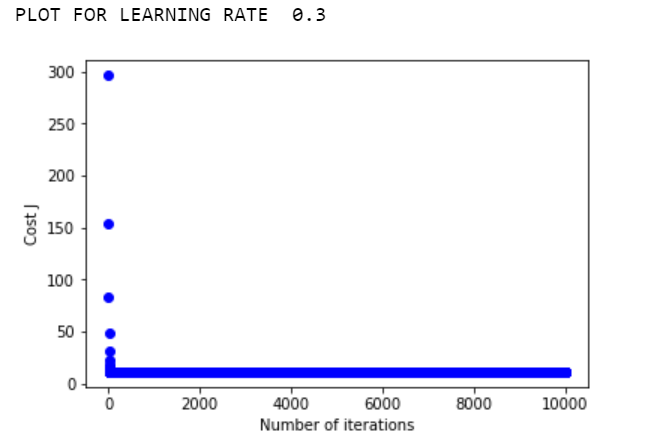


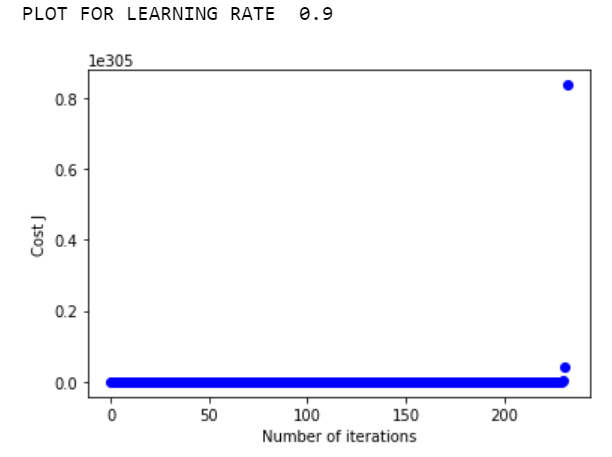


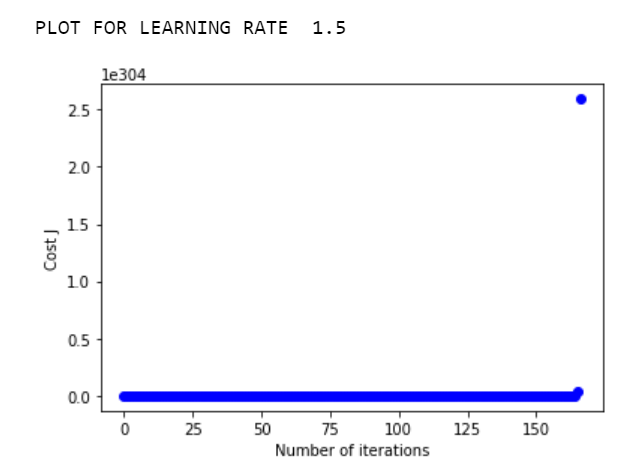








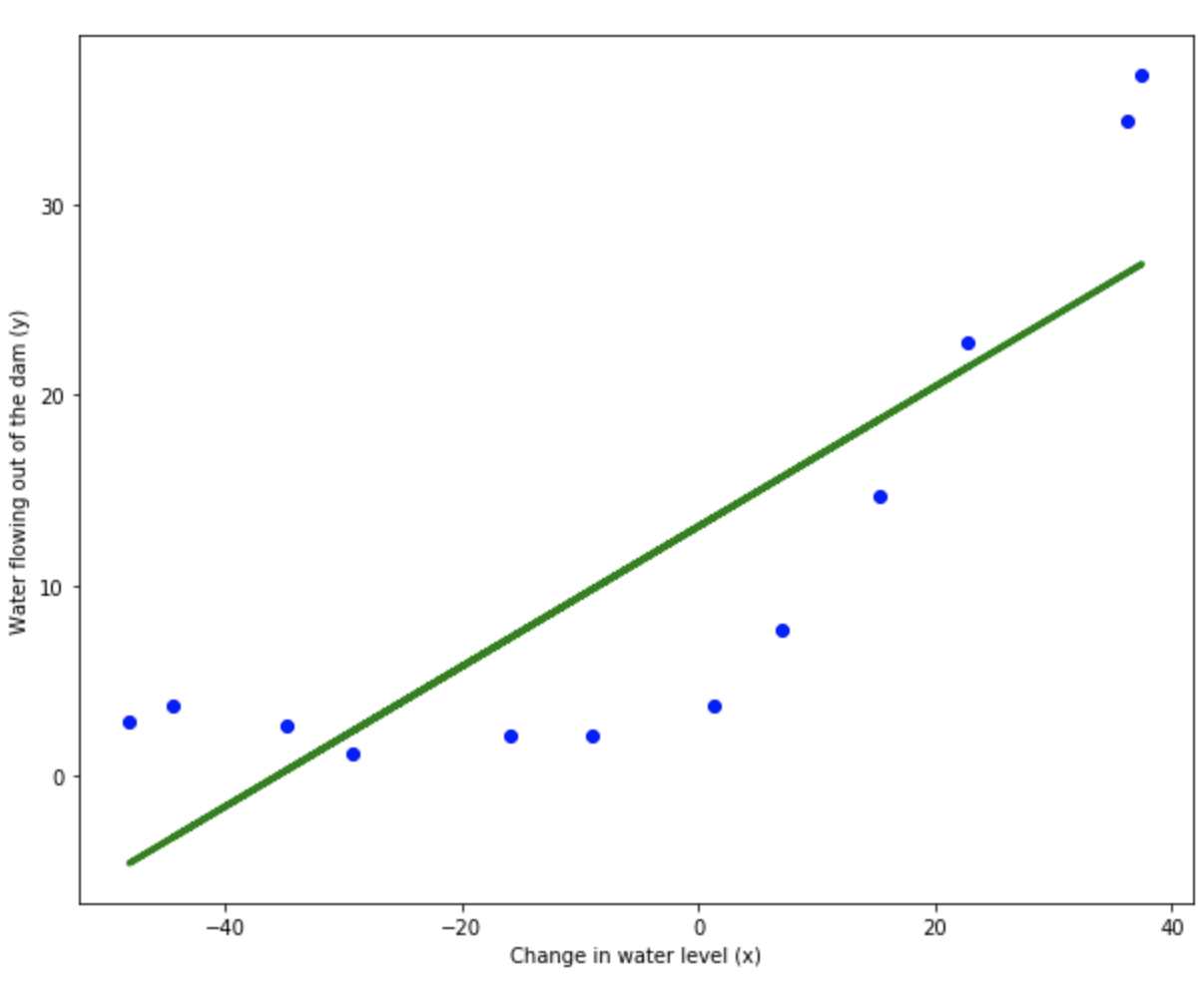




**Problem 3.2: Implementing regularized linear regression (35 points)**

Problem 3.2. A1: Regularized linear regression cost function (Figure 7)

Problem 3.2.A2: Gradient of the Regularized linear regression cost function

****Figure 7: The best fit line for the training data

Problem 3.2.A3: Learning curves

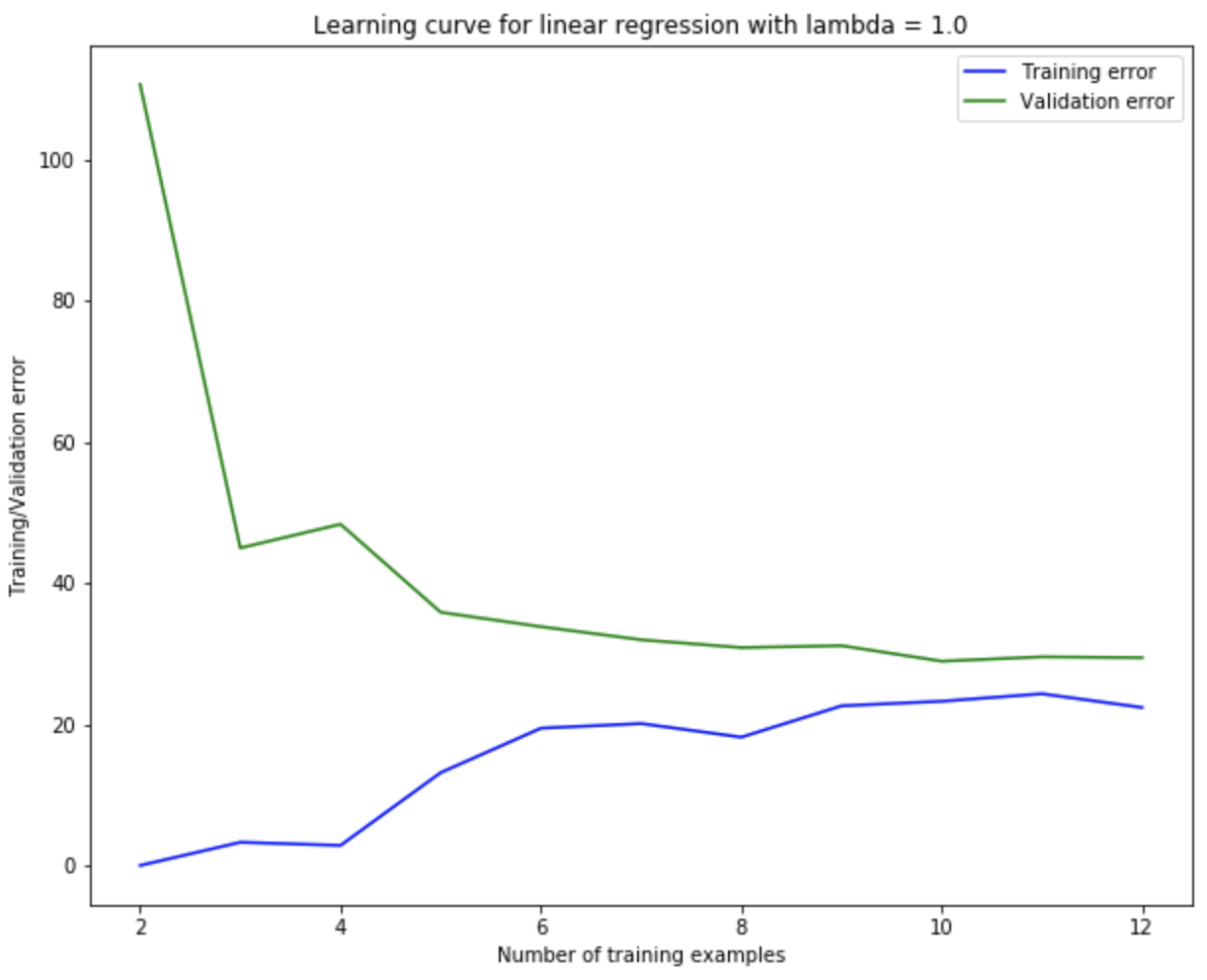


Figure 8: Learning curves

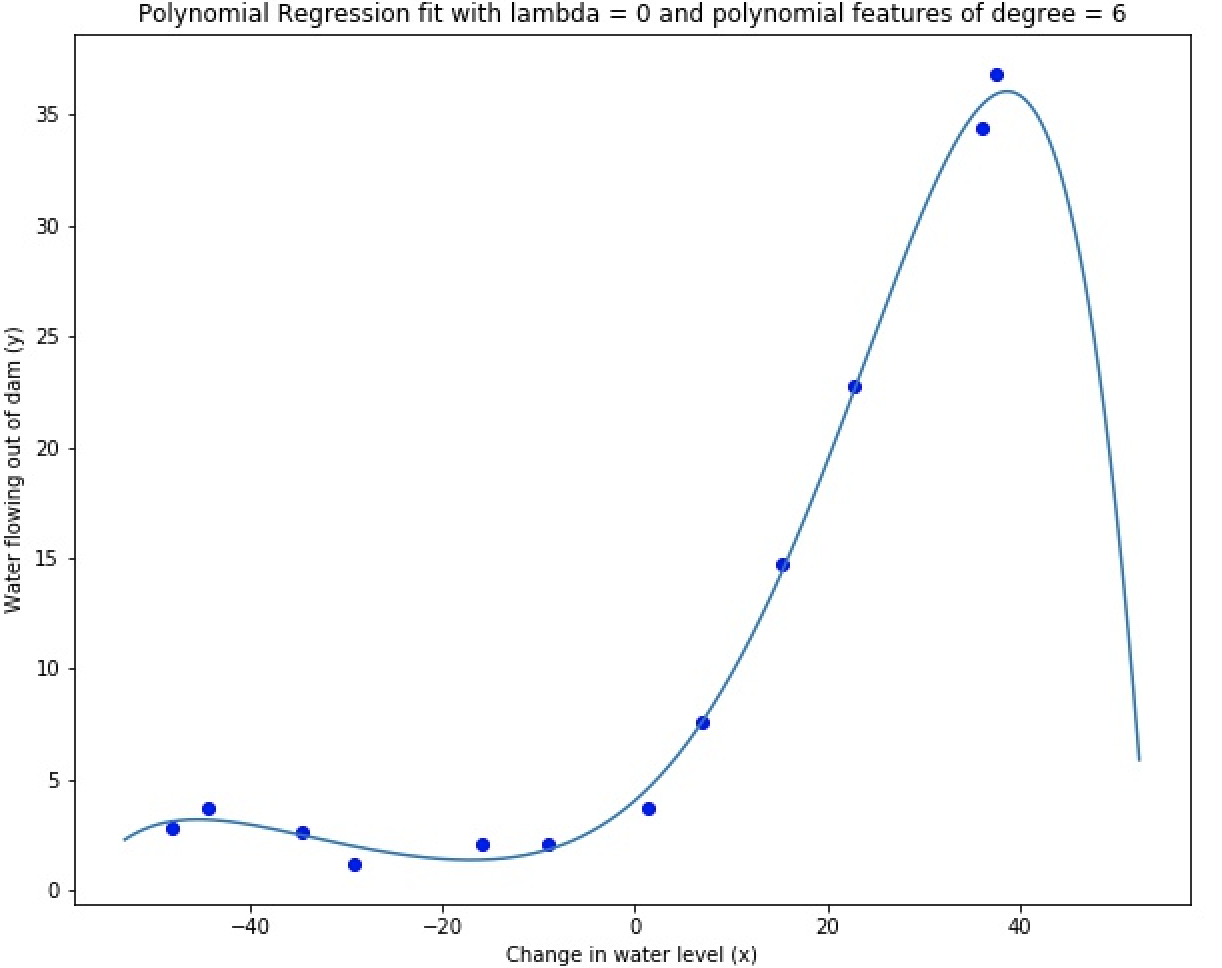
****

Figure 9: Polynomial fit for lambda = 0 with a p=6 order model.

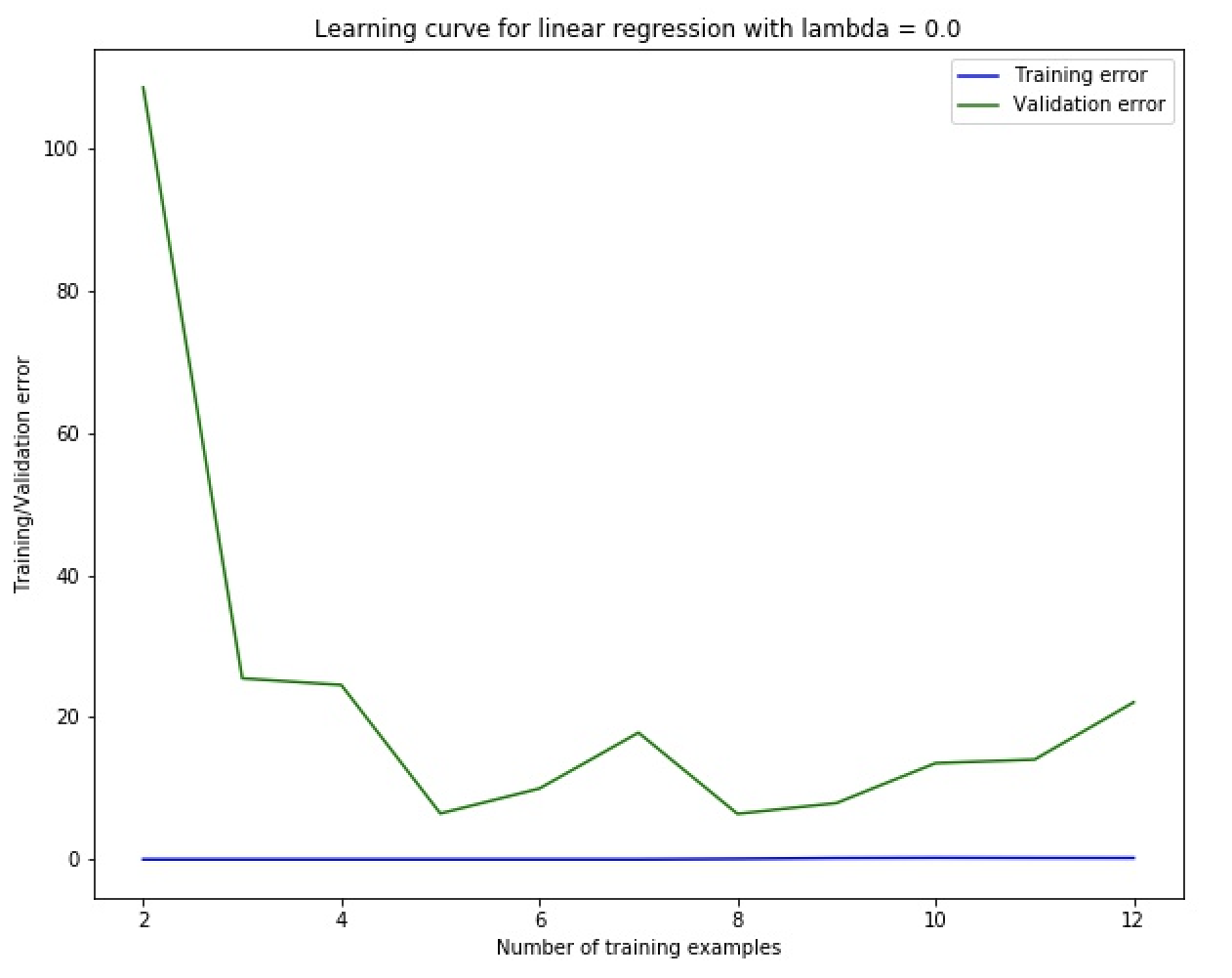
****

Figure 10: Learning curve for lambda = 0.

Problem 3.2.A4: Adjusting the regularization parameter

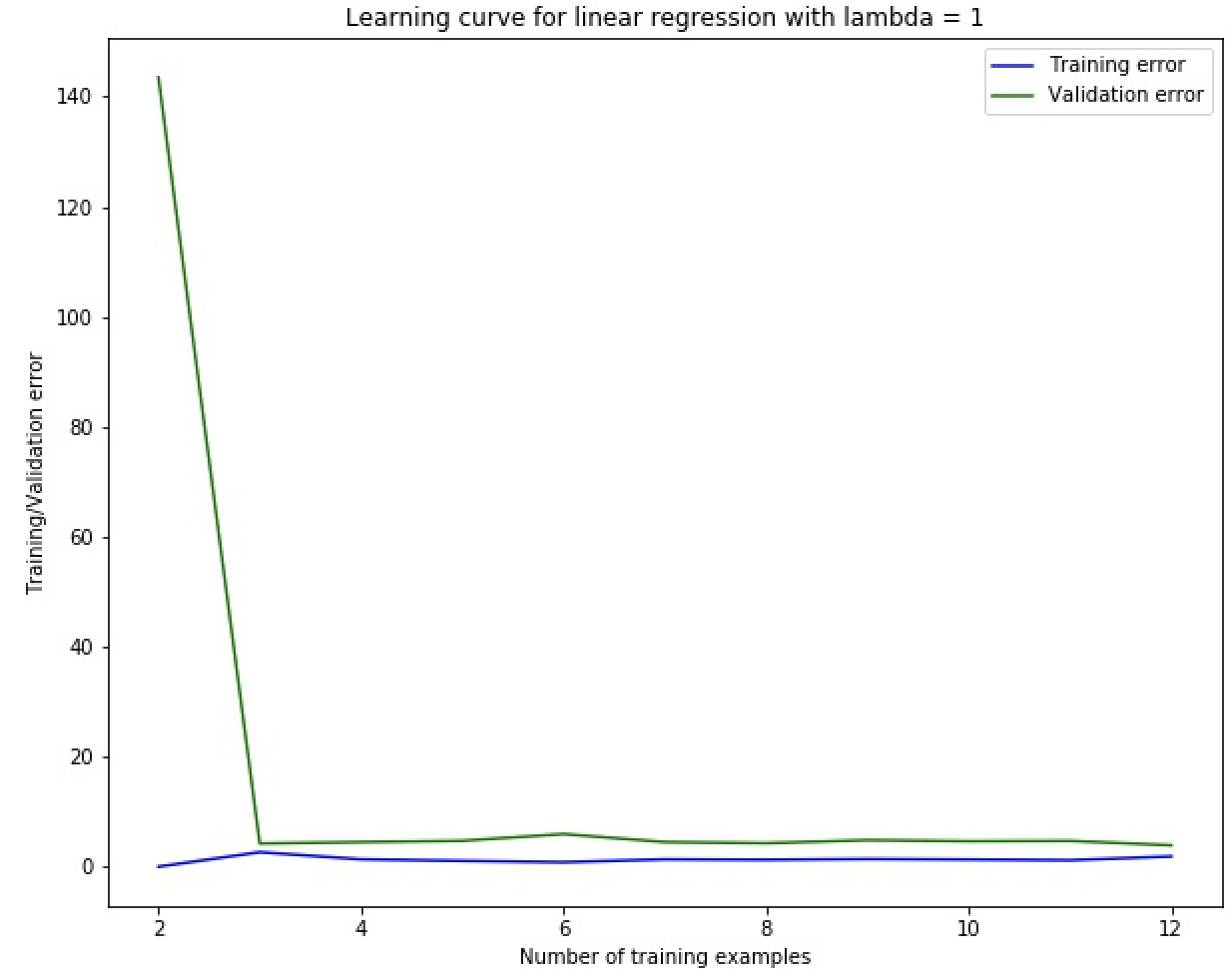
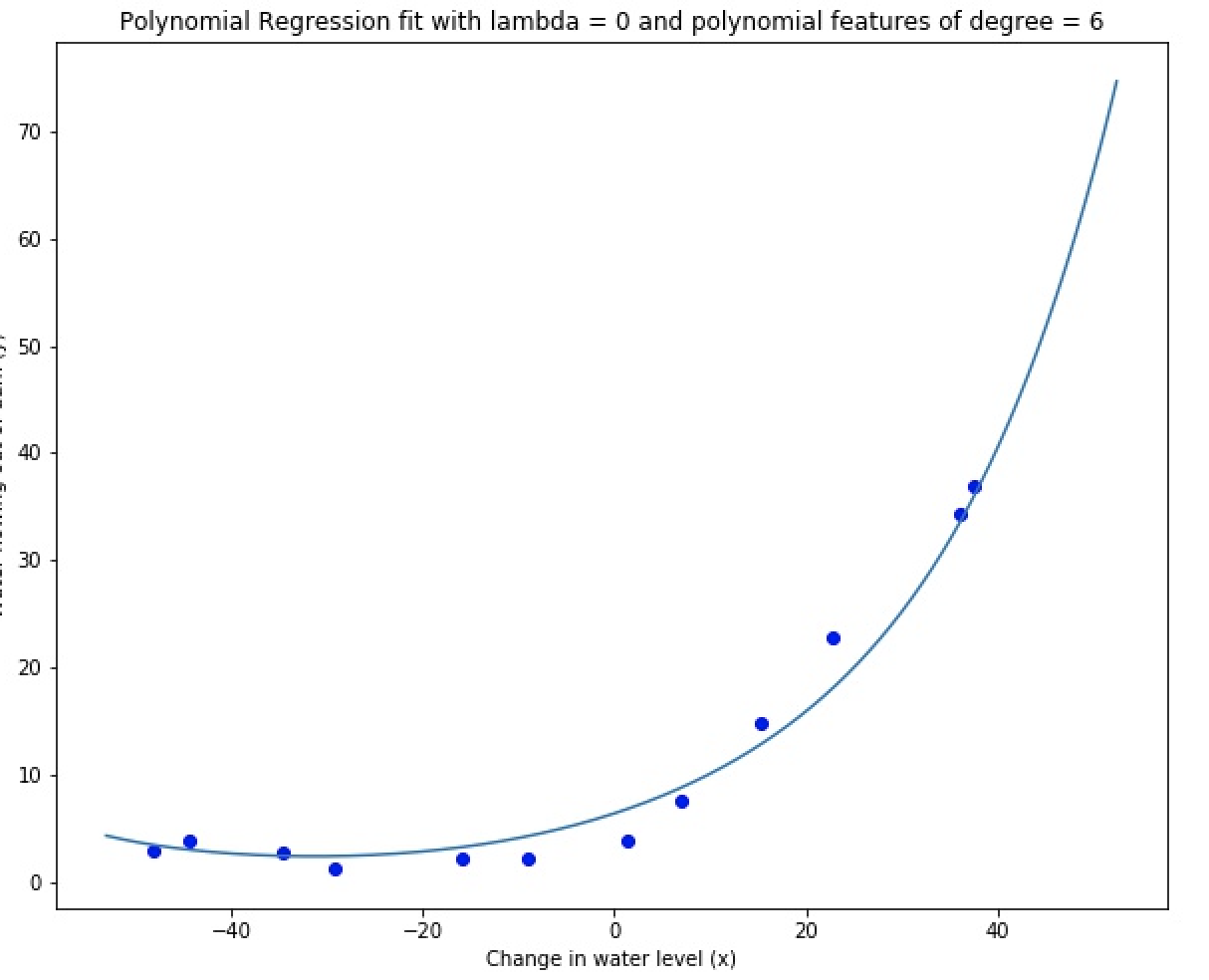


Figure 11: Polynomial fit and learning curve for lambda = 1 with a p=6 order model

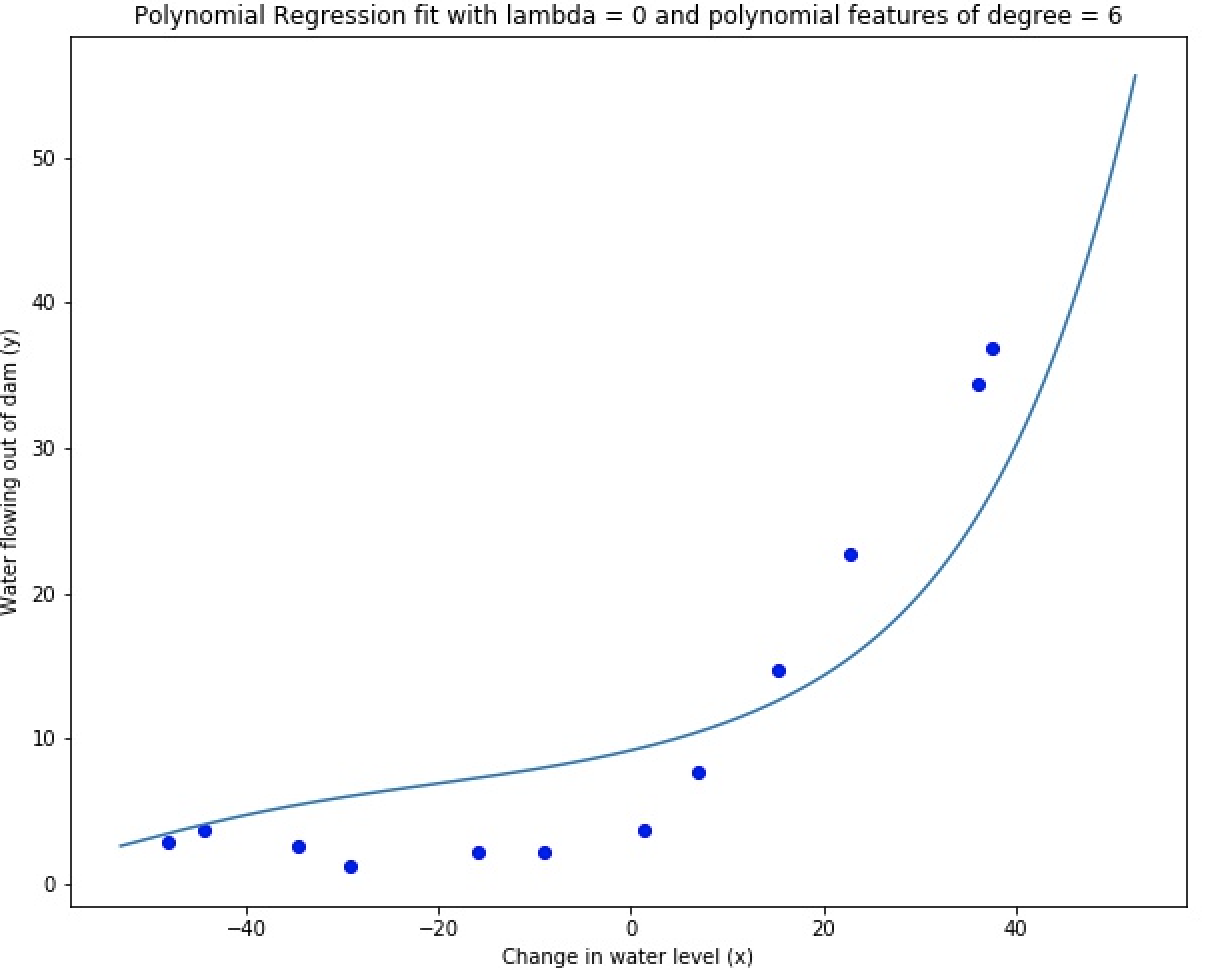
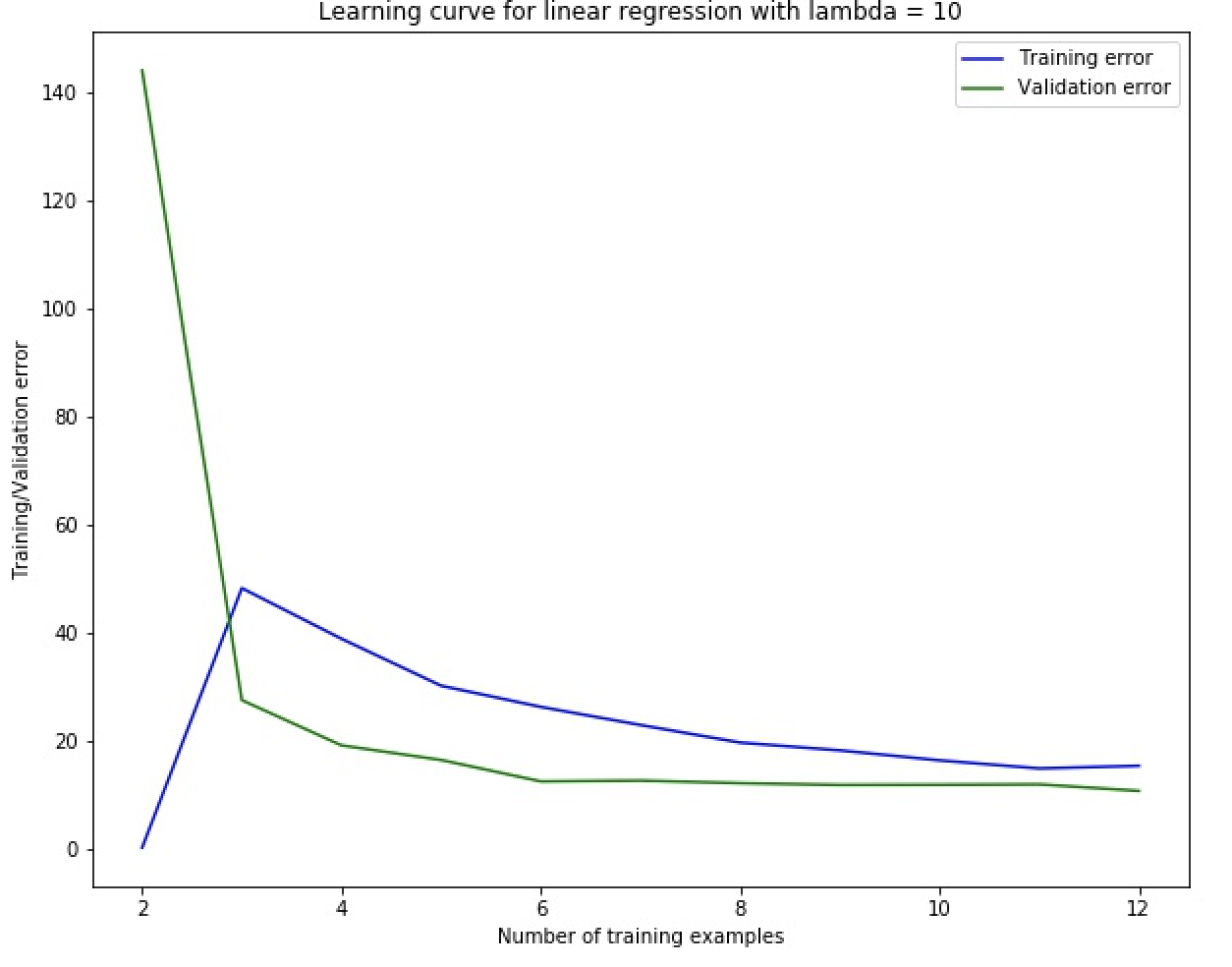
 

Figure 12: Polynomial fit and learning curve for lambda = 10 with a p=6 order model

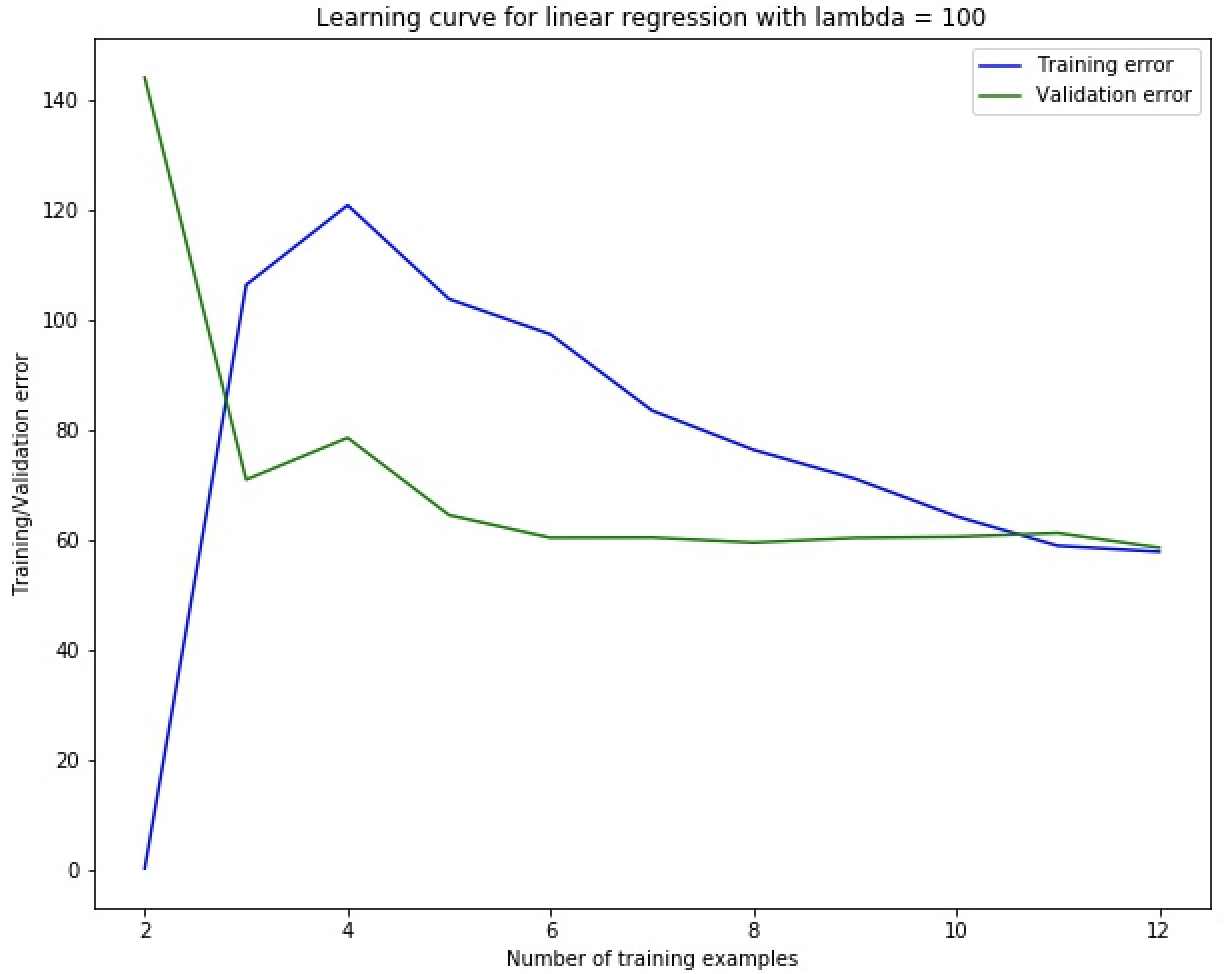
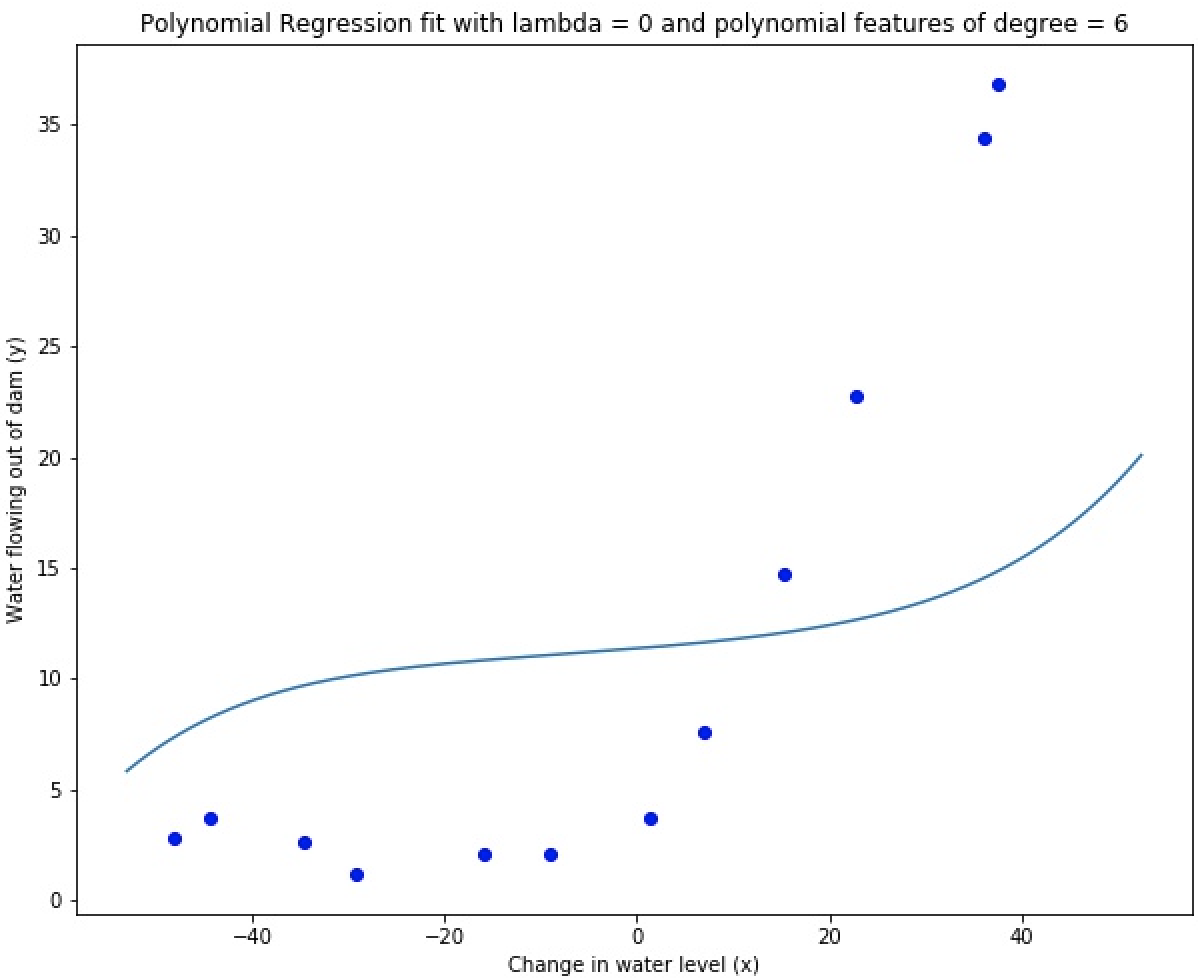


Figure 13: Polynomial fit and learning curve for lambda = 100 with a p=6 order model

According to these three plots, on the one hand, when the value of lambda increasing, it would be less likely to become overfitting. But on the other the hand, the test and train error of model would be larger, that means the bias and variance of model would be larger. Judging from these three lambdas, it would be better to choose lambda=1.

Problem 3.2.A5: Selecting λ using a validation set

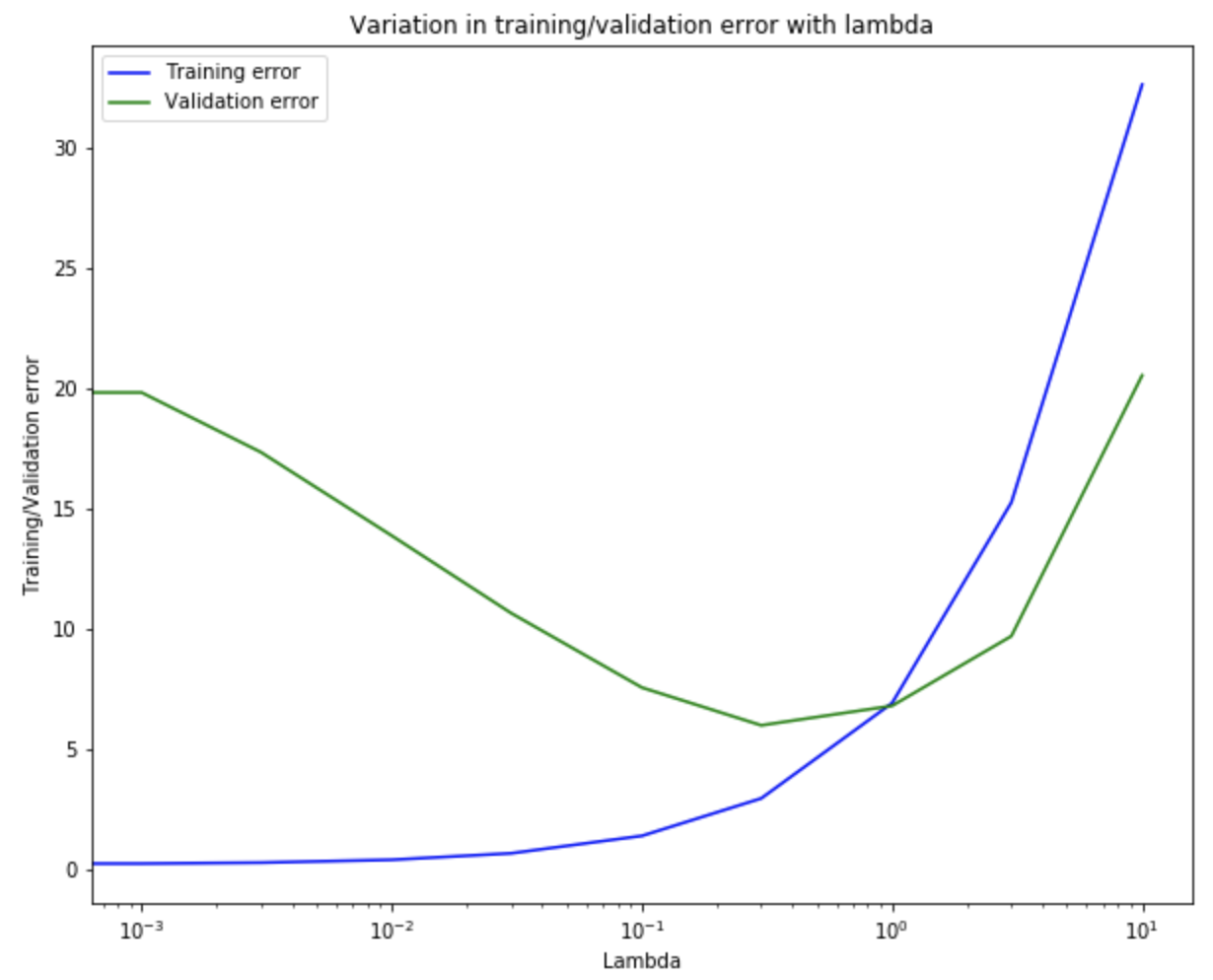


Figure 14: Variation of training/validation error with lambda

According to the plot, the plot of Validation Error is a convex function, so it exists a

lowest point. If lambda is too small, it cannot restrict the problem of overfitting enough.

If lambda is too large, the bias and variance of model would be larger. And the best lambda is 0.3. That’s the lowest point of Validation Error.

Problem 3.2.A6: Computing test set error

When lambda=0.3, the value of Validation Error will be lowest.

Error of training: 2.9313342465235688

Error of validation: 5.969807577475818

Problem 3.2.A7: Plotting learning curves with randomly selected examples

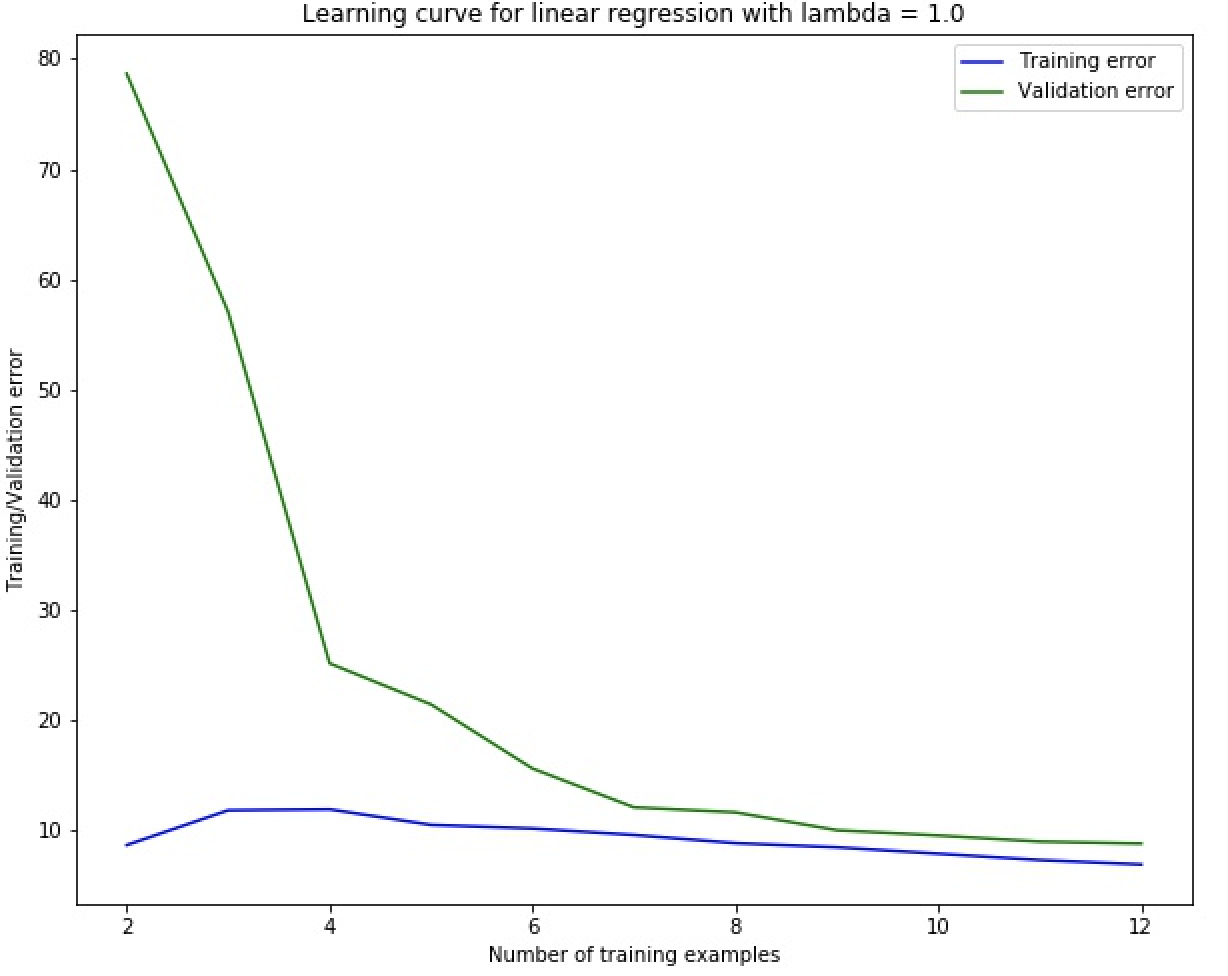


Figure 11: Averaged Learning curve for lambda = 1.