Machine Learning- COL774
Assignment 1

Ankesh Gupta
2015CS10435

Question 1

What was done:

1. Batch Gradient Descent was implemented to train the given dataset.
2. Different *learning rates* and *convergence values* were engineered.
3. Loss functions were plotted and visualized.

Observations:

1. The algorithm diverged for learning rate(η) above 0.2 .
2. While observing the contours for different learning rates, it was observed that for lower learning rate, the jump between successive *epochs* was comparatively lower that jump for rates.
3. Direct consequence of above phenomenon was increased *convergence time* as well as number of iterations for lower rates. However, increasing η again raised the epochs because of oscillations.
4. Another interesting observation was that for low learning rates, algorithm *stably* converged to the minima, whereas for $\eta = 0.017$, it first *oscillated* about the minima a bit and then converged.
5. It failed convergence for next $\eta = 0.021$ as oscillations only pushed it further away from the minima.
6. Epochs also increased with decrease in error condition(ϵ) kept for convergence.
7. Optimal value of learning rate was around $\eta = 0.09$ for below mentioned ϵ.

Here is a tabulated form for η vs *epochs* on the dataset, with $\epsilon = 10^{-7}$, that is difference of $J(\theta)$ became less than ϵ***(Stopping Criteria)***. Obtained θ were:

$$\theta_0 = 0.9965$$

$$\theta_1 = 0.0013$$

η	Epochs
0.001	89
0.005	16
0.009	6
0.013	10
0.017	29

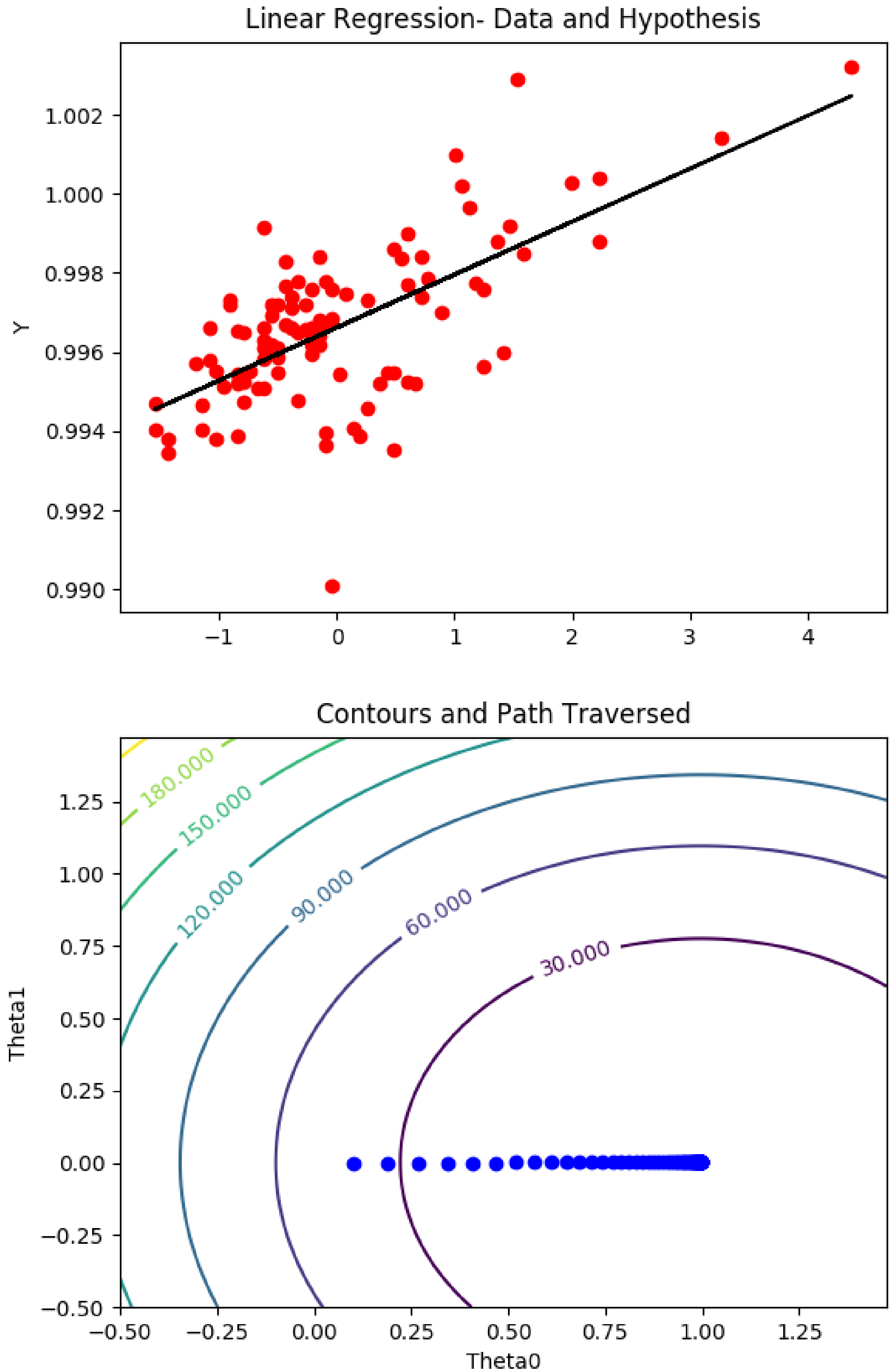

Figure 2: Example Plot of Hypothesis and data for $\eta = 0.001$

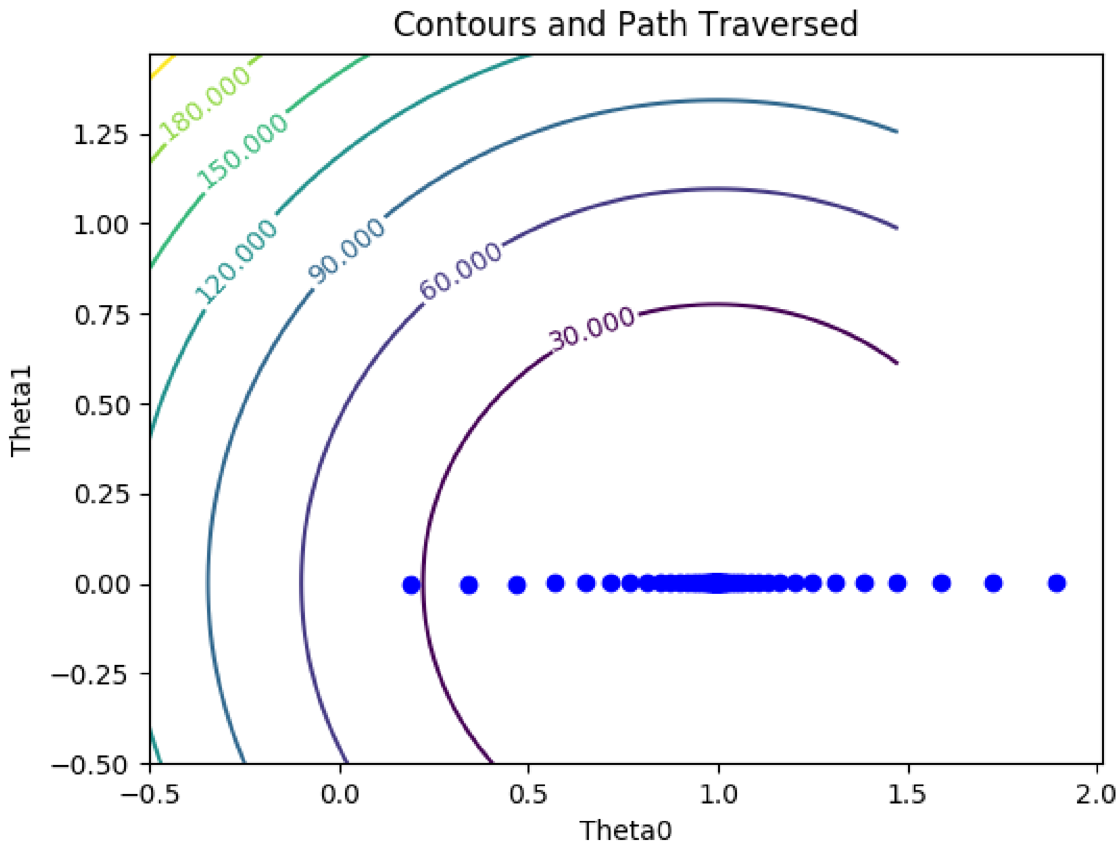

Figure 3: Example Plot of Hypothesis and data for $\eta = 0.019$

Question 2

What was done:

1. Closed form equations were implemented for linear regression and locally weighted linear regression($LWLR$).***Bandwidth parameter*** for LWLR(τ) was engineered.
2. Inferences were drawn on extreme values of τ.

Observations:

1. The best value of τ was 0.3.
2. Too low $\tau(< 0.2)$ gave rise to $overfitting$. Too high $\tau(> 5)$ resulted in $underfitting$. Example figure is shown.
3. High τ makes the value in power of exponent ≈ 0 which resulted in equal weightage of all sample points, a reduction of its power to ***Linear Regression***.
4. Low τ assigns too much weight to the point in consideration and hence forces the hypothesis to fit as many points as it could, resulting in underfitting.
5. Although powerful, this technique is ***computationally expensive***, as evaluation at each point is quite expensive. In our case, it required complete data lookup as well as *inverse* computations.
6. Closed form θ for Linear Case: $\theta_0 = 0.9966$**,** $\theta_1 = 0.0013$
7. Closed form expression for LWLR is

$$\theta = (W^T W X)^{-1}(X^T W^T Y)$$

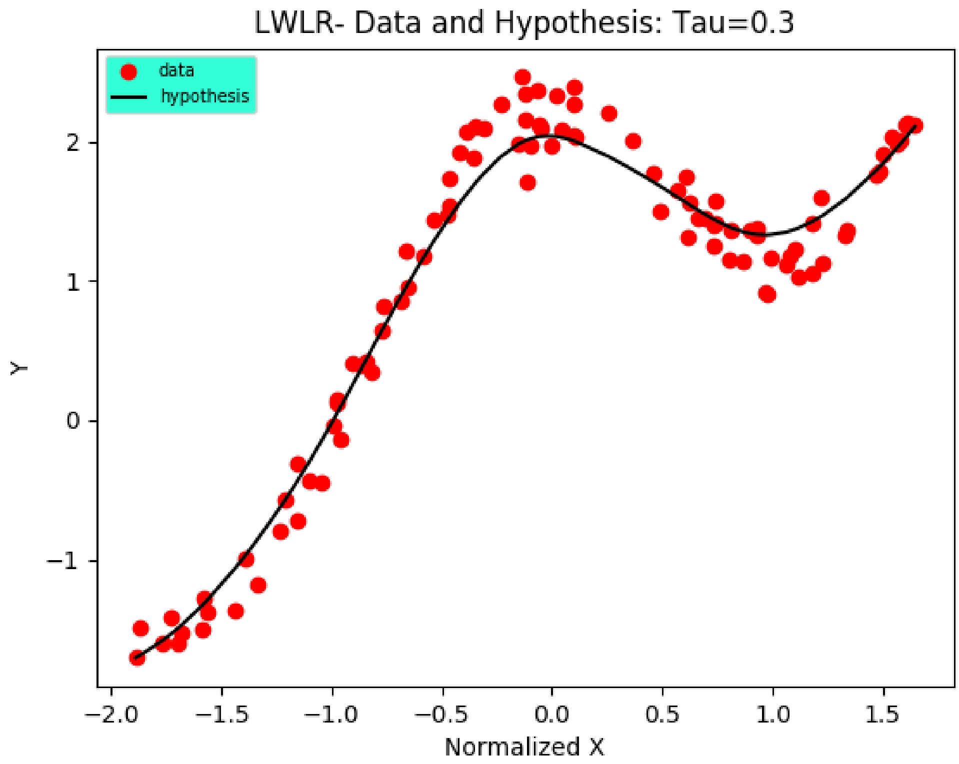

Figure 4: Best Fit

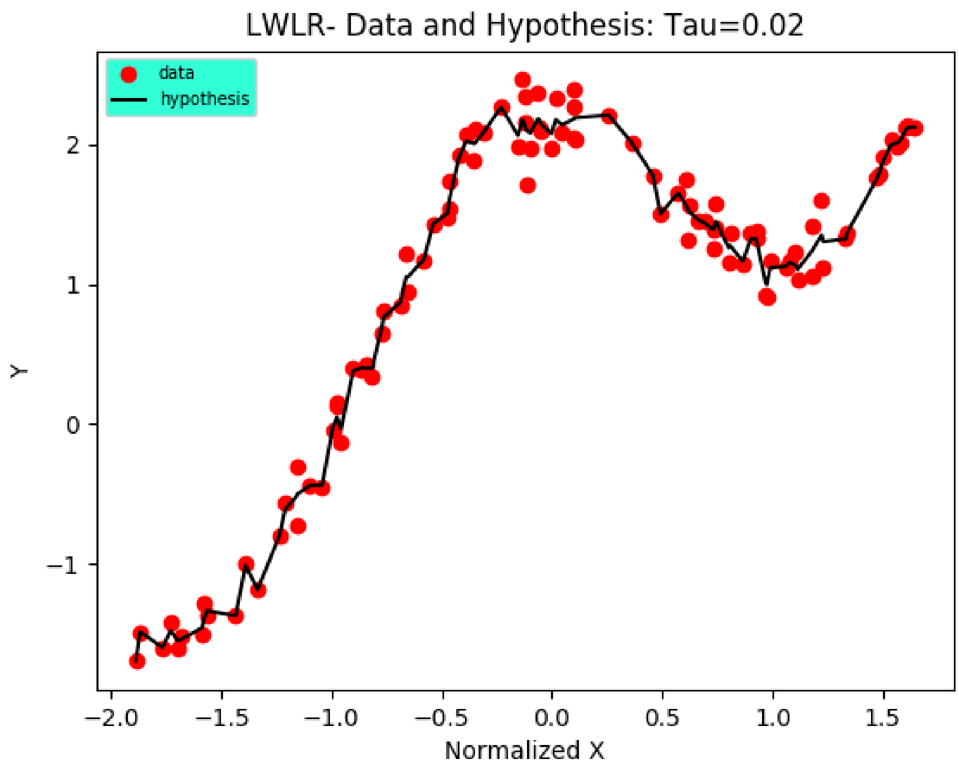

Figure 5: An example of small τ

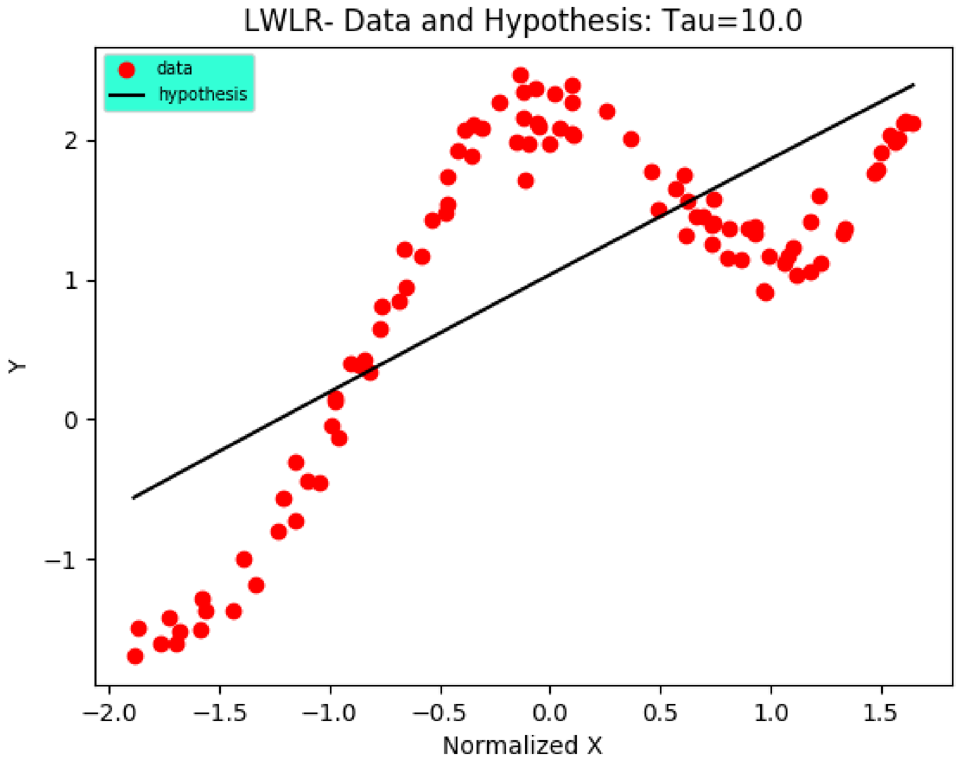

Figure 6: An example of large τ

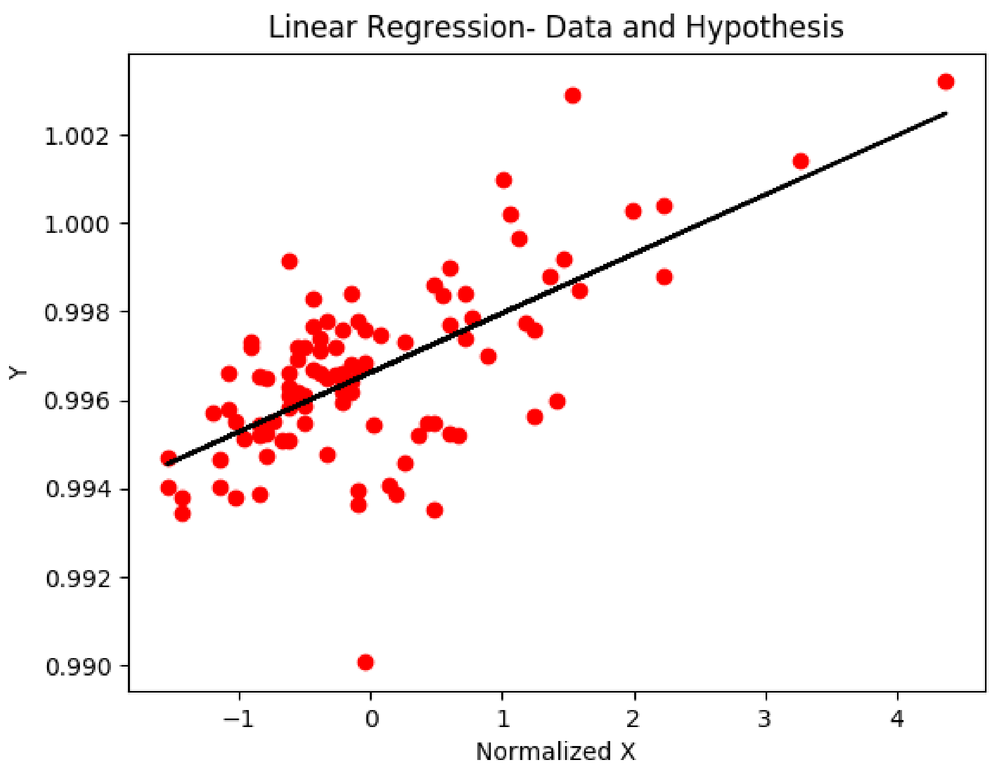

Figure 7: Hypothesis plot for Closed form of Linear Regression

Question 3

What was done:

1. Newton's method was applied to optimize/maximise *likelihood* for the problem.
2. Decision boundary was plotted and analyzed.

Observations:

1. Obtained θ leading to covergence:

$$\theta_0 = 0.401$$
$$\theta_1 = 2.588$$
$$\theta_2 = -2.725$$

2. The stopping criteria was when $\|\theta^t - \theta^{t-1}\|_2 < \epsilon$
3. Generally, Newton takes lower epochs than *GradientDescent* but computationally suffers because of computation of inverse of *Hessian*.
4. A benefit of using it is we don't have to engineer any *step_size* or learning rate, as in *Gradient Descent*.

General tabulation of experiments done with different converging criteria(ϵ)

ϵ	Epochs
0.1	8
0.001	258
10^{-6}	968
10^{-9}	1687

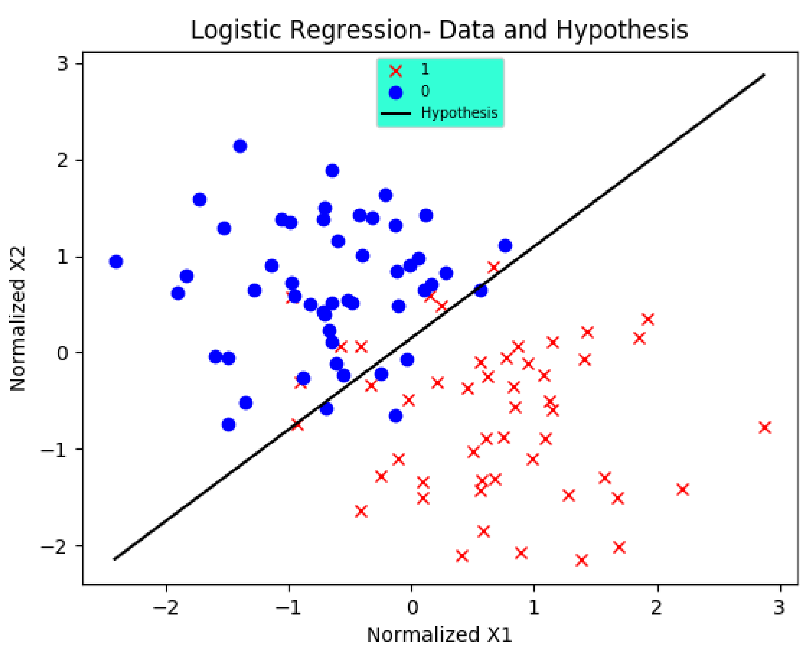

Figure 8: Decision Boundary and Points $\epsilon = 10^{-6}$

Question 4

What was done:

1. Closed form equations were derived for $\mu_0, \mu_1, \Sigma's, \phi$.
2. Decision boundary equations were derived and plotted once assuming $\Sigma_0 = \Sigma_1$, and other time, removing the restriction.
3. The obtained boundaries were analyzed.
4. Alaska is numbered 1 and Canada, 0.

Obtained values of $\mu's, \Sigma's, \phi$ on *normalized* data are:

$$\begin{aligned}
\phi &= 0.5 \\
\mu_0 &= \begin{bmatrix} -0.7553 & 0.685 \end{bmatrix} \\
\mu_1 &= \begin{bmatrix} 0.7553 & -0.685 \end{bmatrix} \\
\Sigma &= \begin{bmatrix} 2.333 & 0.0988 \\ 0.0988 & 1.8886 \end{bmatrix} \\
\Sigma_0 &= \begin{bmatrix} 0.4774 & 0.1099 \\ 0.1099 & 0.4135 \end{bmatrix} \\
\Sigma_1 &= \begin{bmatrix} 0.3816 & -0.1548 \\ -0.1548 & 0.6477 \end{bmatrix}
\end{aligned}$$

Obtained Decision Boundary when $\Sigma_0 = \Sigma_1$ is linear given by:

$$(\mu_1^T \Sigma^{-1} - \mu_0^T \Sigma^{-1})x = 0.5 * (\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0) + \log \frac{1-\phi}{\phi}$$

Obtained Decision Boundary without the above assumption is quadratic in nature given by:

$$x^T(\Sigma_1^{-1} - \Sigma_0^{-1})x + 2(\mu_0^T \Sigma_0^{-1} - \mu_1^T \Sigma_1^{-1})x = (\mu_0^T \Sigma_0^{-1} \mu_0 - \mu_1^T \Sigma_1^{-1} \mu_1) + 2 * \log \frac{\phi}{1-\phi} - \log \frac{|\Sigma_1|}{|\Sigma_0|}$$

Analyzing the decision boundary to comment which of them better fits is non-trivial given so few instances. But given so few a datapoints, if indeed the underlying $x/y \sim N(\mu, \sigma^2)$, then GDA is one of the better discriminator's because of its strong modelling assumptions. Both the ***boundaries*** do a decent role in separating the given set of instances. Since the points are concentrated in a local neighbourhood for both the classes, their is a stray possibility that the underlying distribution is indeed *Gaussian*.

Since the quadratic curve allows ***larger class of functions*** to be estimated, we can say that quadratic boundary is a better estimator as it misclassifies very few data points. Also its quadratic curve is in a sense, better/strongly separating the 2 classes. But again, we definitely need more instances to comment something stronger.

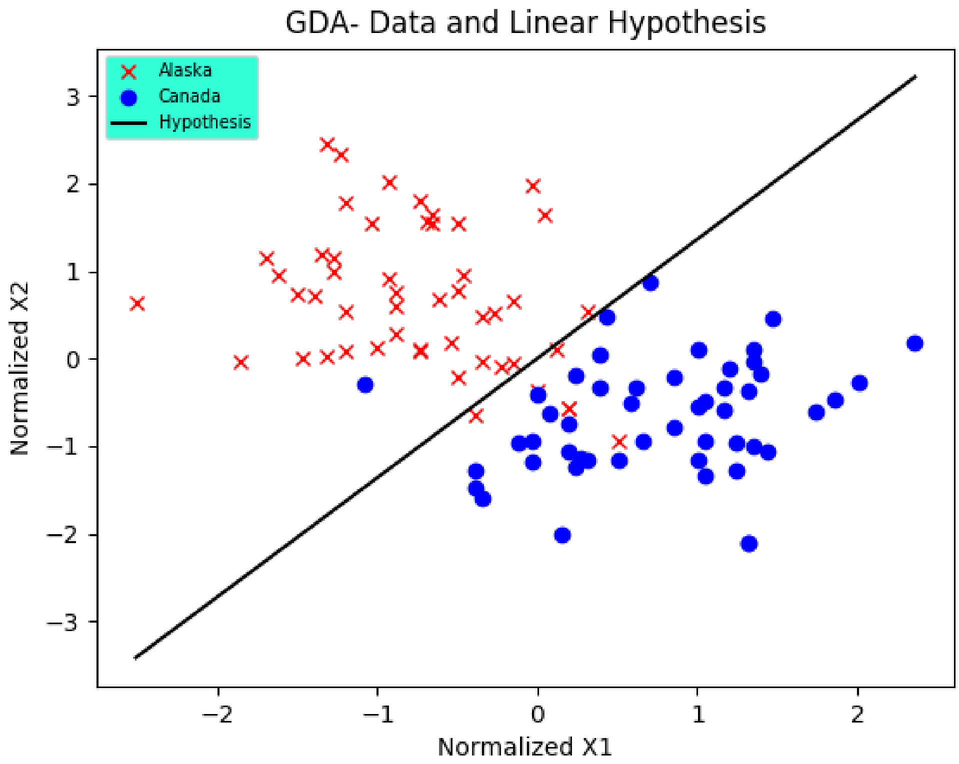

Figure 9: Linear Fit when $\Sigma_0 = \Sigma_1$

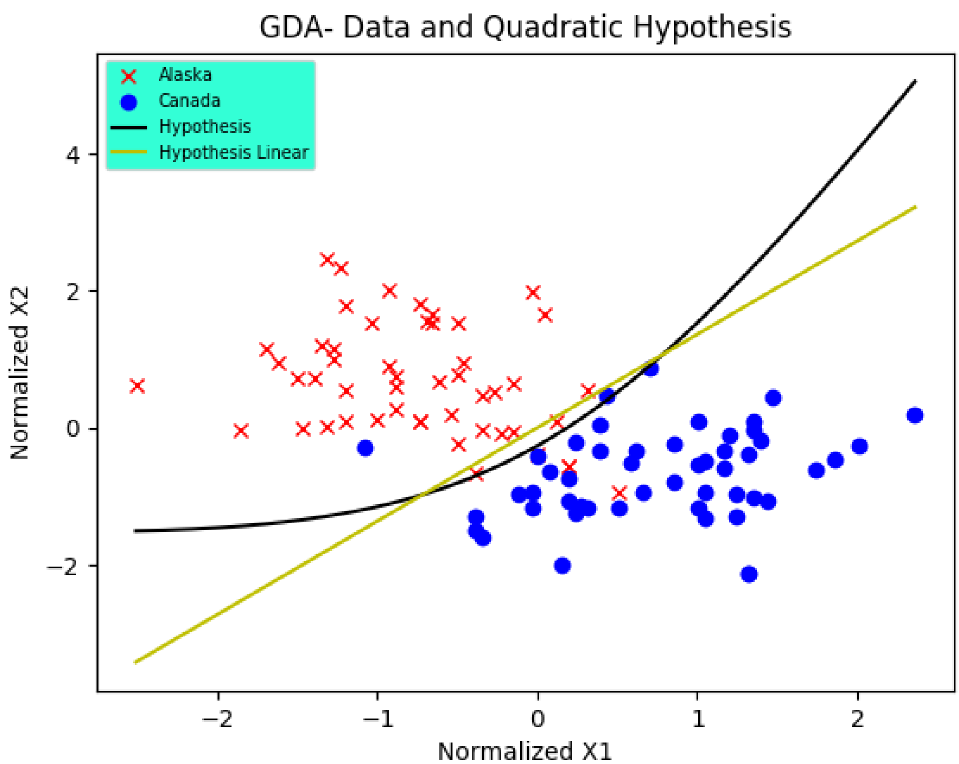

Figure 10: General GDA Quadratic Fit