1: Gradient Descent

By using a constant value for alpha, we run the risk of either choosing too large of a value and actually having theta overshoot the solution, or to choose too small of a value and take a very long time. By using a value of alpha that varies with k, we can cause it to decrease with iterations such that when theta is far from the solution, it will converge quickly, but as theta converges, alpha decreases in order to slow the rate of change and ensure an accurate result.

2: Fitting an SVM by Hand

(a)

$$\begin{aligned}
\vec{V_1} &= \langle 1, 0, 0 \rangle \\
\vec{V_2} &= \langle 1, 2, 2 \rangle \\
\vec{V_2} - \vec{V_1} &= \langle 1-1, 2-2, 2-0 \rangle = \langle 0, 2, 2 \rangle
\end{aligned}$$

(b)

$$||\langle 0, 2, 2 \rangle|| = \sqrt{0^2 + 2^2 + 2^2} = 2\sqrt{2} = 2.828$$

(c)

$$2\sqrt{2} = \frac{2}{\sqrt{0^2 + W_1^2 + W_2^2}}$$

$$W_1^2 + W_2^2 = \frac{4}{8}$$

$$\frac{2}{2} = \frac{W_2}{W_1}$$

$$2W_1^2 = \frac{1}{2}$$

$$W_1 = \frac{1}{2}$$

$$W_2 = W_1$$

$$\vec{W} = \langle 0, \frac{1}{2}, \frac{1}{2} \rangle$$

(d)

$$-1\left(\begin{bmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} + w_0\right) \geq 1$$

$$w_0 \geq -1$$

$$- + 1\left(\begin{bmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}\begin{bmatrix} 1 & 2 & 2 \end{bmatrix} + w_0\right) \geq 1$$

$$w_0 \geq -1$$

$$w_0 = -1$$

(e)

$$h(x) = \frac{1}{2}x^2 + \frac{\sqrt{2}}{2}x - 1$$

3: VC Dimension (519 Only)

(a)

The VC dimension of this classifier must be one because if we have two points such that point one is radially closer to the origin and negative while point two is positive and radially further away from the origin, there is no way to classify point two as positive without also classifying point one as positive.

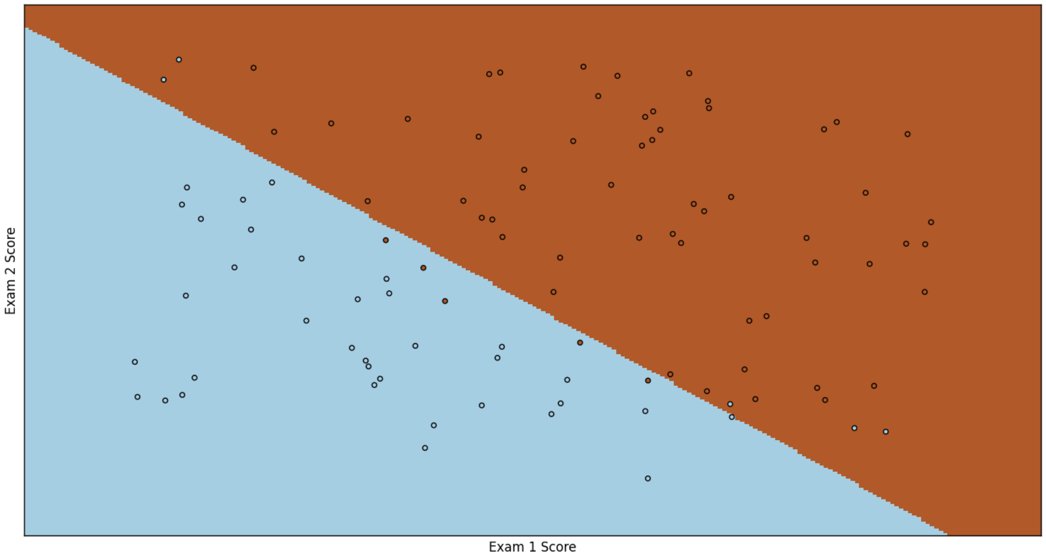
(b)

The VC dimension of the classifier must now be two because if we have three points (1, 2, and 3) in increasing radial distance from the origin, and points 2 has a different classification from both point 1 and point 3, then it will be impossible to properly classify point 2 without also miss-classifying at least one of point 1 or point 3.

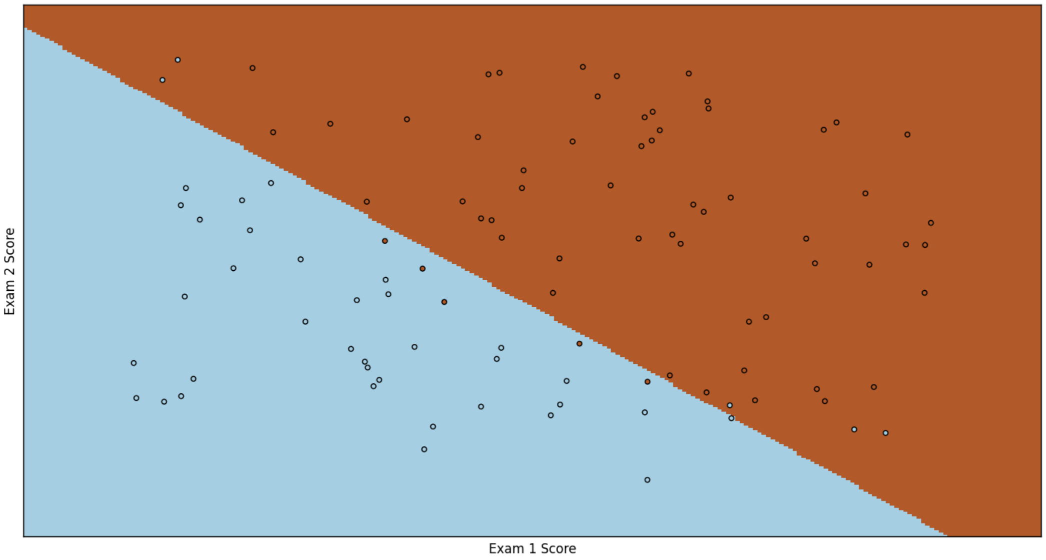
2.3 Analysis: Plots of Varying Lambda

Because there is a relatively linear split between the data, small changes in lambda do not seem to have huge impacts on the decision boundary. However, it is plan that increasing the value of lambda does increase the bias - in effect smoothing the result. Or in other words, a larger lambda makes the decision boundary less sensitive to individual pieces of data being miss-classified. This is because lambda acts as a penalty on thetas and as a result, larger values of lambda word to drive thetas to zero.

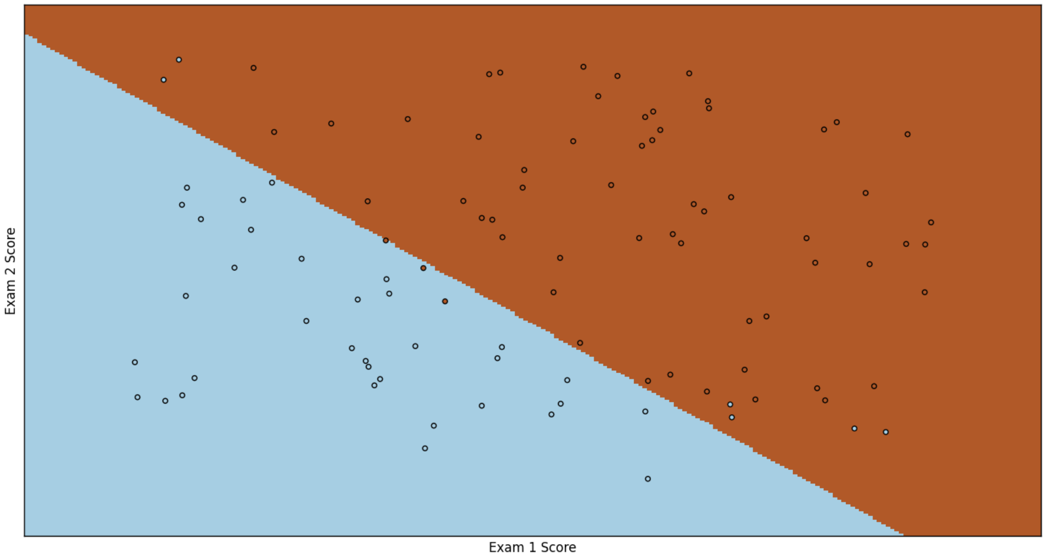
$\lambda = 0.00000001$



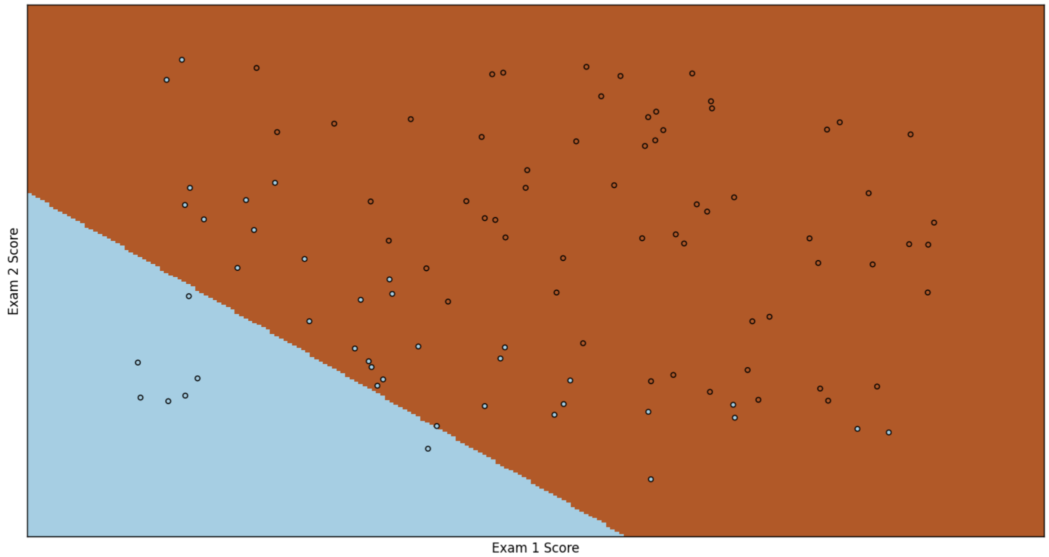
$\lambda = 0.00001$



$\lambda = 0.1$



$\lambda = 1.0$



3.4 Analysis: Implementing the Gaussian Radial Basis Function Kernel

Low values of C cause the SVM to be highly generalized, or have high bias. This is because, in the opposite manner of lambda, C causes us to ignore theta and weight the miss-classification of points, so when C is zero, we see the inverse and put high weight on theta. As a result, with small values of C, we see a smooth form, while with large values, we see lower testing error. Similarly, increasing both the dimensionality, d and the gaussian gives us more flexibility in classifying the points in the space and results in a line that is less smooth (low bias) but has lower testing error.