Problem 1:

1.

We can begin by simplifying the given expressing using the definition of the 2 norm.

$$\left|\left|Xw-y\right|\right|^2 = (Xw-y)^T(Xw-y) = (w^TX^T-y^T)(Xw-y) = w^TX^TXw - y^TXw - w^TX^Ty + y^Ty$$

We can simplify further, since all terms are scalars, and a scalar transposed is itself:

$$w^TX^TXw - 2w^TX^Ty + y^Ty$$

Taking the derivative with respect to w yields

$$2X^TXw - 2X^Ty$$

Setting that expression to 0 allows us to solve for w

$$2X^TXw - 2X^Ty = 0$$

$$X^TXw - X^Ty = 0$$

$$X^TXw = X^Ty$$

$$w = (X^TX)^{-1}X^Ty$$

2.

The derivative of this new expression will be the same but with an added term:

$$2X^TXw - 2X^Ty + \frac{d}{dw}\lambda\left|\left|w\right|\right|^2 = 2X^TXw - 2X^Ty + \lambda\frac{d}{dw}w^Tw$$

Solving that and setting to 0 yields:

$$2X^TXw - 2X^Ty + 2\lambda w = 0$$

$$X^TXw - X^Ty + \lambda w = 0$$

$$(X^TX + \lambda I)w = X^Ty$$

$$w = (X^TX + \lambda I)^{-1}X^Ty$$

Problem 2:

1.

The probability of observing HHTTH is:

$$pp(1-p)(1-p)p = p^3(1-p)^2$$

Taking the natural log of this gives

$$\ln p^3(1-p)^2 = 3\ln p + 2\ln(1-p)$$

2.

a)

When p = ½, 1 – p = 1/2. Then the probability of the given sequence is (1/2)^5 = 1/32. However, since there is a ½ probability of choosing the fair coin, the joint probability is ½ * 1/32 = 1/64.

b)

When p = 2/3, 1 – p = 1/3. Therefore, the probability of the sequence is:

$$\frac{2}{3}^3 \frac{1}{3}^2 = \frac{8}{27} * \frac{1}{9} = \frac{8}{243}$$

However, the joint probability is that multiplied by the probability of getting the biased coin, which is ½. Therefore, the joint probability is 4/234.

3.

Maximizing the log of the sequence's probability with respect to p will also maximize the sequence's probability with respect to p because log grows monotonically. This is very convenient, because to maximize this function we must take the derivate and set it to 0, which becomes trivial with the log:

$$\frac{d}{dp}(3\ln p + 2\ln(1-p)) = \frac{3}{p} - \frac{2}{1-p} = \frac{3(1-p) - 2p}{p(1-p)}$$

Setting that to 0, and assuming p is not 0 or 1, we get:

$$3(1-p) - 2p = 0$$

$$3 - 5p = 0$$

$$3 = 5p$$

$$p = 3/5$$

It remains only to verify that this is a global maximum and not a global minimum, which can be done by sampling a point on both sides of p = 3/5. However, using p = 0 and p = 1 yields a probability of 0, so we know this is a global maximum. Therefore, the optimal probability for the coin to see HHTTH is 3/5.

Problem 3:

As instructed, there are three images provided: one for the perceptron initialization, one for the perceptron convergence, and one for the linear programming problem (data is linearly unseparable).

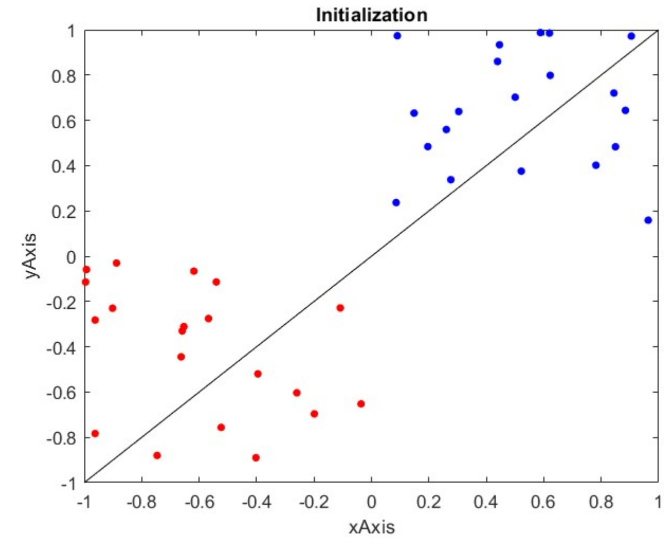

Figure 1: initialization of perceptron on linearly separable data.

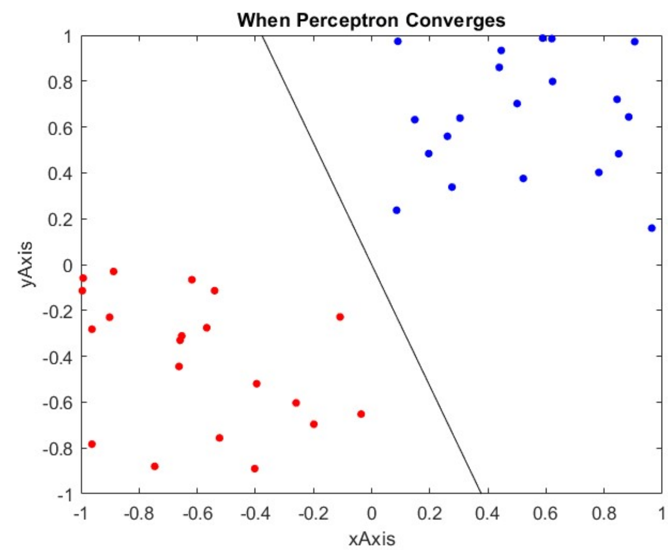

Figure 2: Perceptron on linearly separarable data once it has converged. We can see a clear gap between the data.

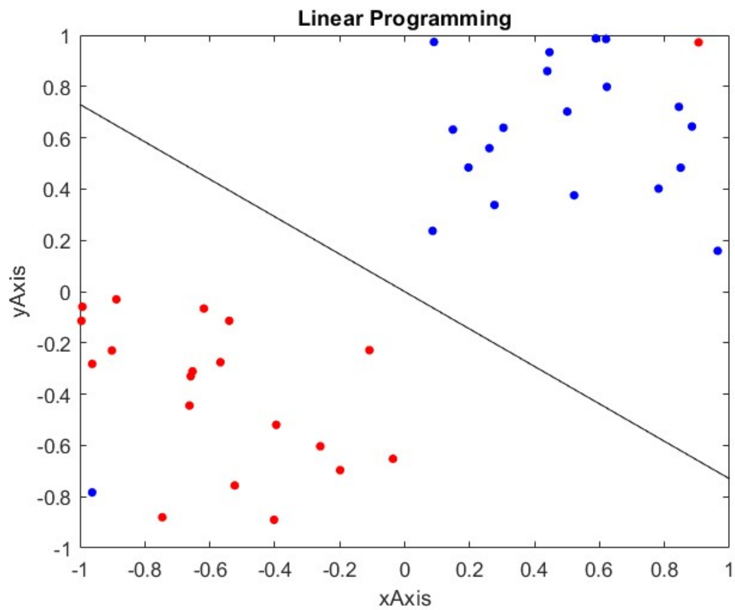

Figure 3: Linear programming applied to a linearly inseparable data set. We can see that though the data are mostly separated, there are two outliers; this is a good application of linear programming.