MARTHEY Antoine
am4665
COMS 4903
Homework 2
Certificate Program

Problem 1:

a) Derive $\hat{\pi}$

$$\sum_{i=1}^{n} \ln p(y_i|\pi) = \sum_{i=1}^{n} \ln\left(\pi^{y_i}(1-\pi)^{(1-y_i)}\right)$$

$$= \sum_{i=1}^{n} y_i \ln \pi + \sum_{i=1}^{n}(1-y_i)\ln(1-\pi)$$

$$= \ln \pi \cdot \left(\sum_{i=1}^{n} y_i\right) + \ln(1-\pi) \cdot \left(\sum_{i=1}^{n}(1-y_i)\right)$$

So to find a maximum, we need to derive this expression and equal it to 0:

$\rightarrow (n - \sum_{i=1}^{n} y_i)$

so $$\frac{\sum_{i=1}^{n} y_i}{\hat{\pi}} + \frac{\sum_{i=1}^{n}(1-y_i)}{1-\hat{\pi}} = 0$$

so $$\boxed{\hat{\pi} = \frac{\sum_{i=1}^{n} y_i}{n}}$$

b) Since $\theta_y^{(1)}$ follow a Bernoulli distribution as well, we have:

$$\boxed{\hat{\theta}_y^{(1)} = \frac{\sum_{i=1}^{n} x_{i,1}}{n}}$$

c) With the same idea, we have:

$$\sum_{i=1}^{n} \ln\left(\theta_y^{(2)} (x_{i,2})^{-(\theta_y^{(2)}+1)}\right) = \sum_{i=1}^{n} \ln \theta_y^{(2)} + \sum_{i=1}^{n} \left(-(\theta_y^{(2)}+1) \ln(x_{i,2})\right)$$

so $$\frac{\sum_{i=1}^{n} 1}{\hat{\theta}_y^{(2)}} - \sum_{i=1}^{n} \ln(x_{i,2}) = 0.$$

so $$\boxed{\hat{\theta}_y^{(2)} = \frac{n}{\sum_{i=1}^{n} \ln(x_{i,2})}}$$

Problem 2 :

a) See code.

Solution :

	0	1
0	54	5
1	2	32

The accuracy is equal to $(32+54)/93 = \underline{92,47\%}$.

b) See Figure 1.

See code

Observation :

word

$\theta_1[\boxed{\text{Free}}] = 0,54$

$\theta_0[\boxed{\text{Free}}] = 0,09$

character

$\theta_1[\,!\,] = 0,83$

$\theta_0[\,!\,] = 0,26$.

We can conclude that those two features are really likely to be found in a spam $(y=1)$. Since their $\theta_1 \gg \theta_0$ it means that the probability (linked to a Bernoulli distribution) of apperance in a spam is high.

c) See figure 2.
See code.

d) See figure 3
See code..

e) See figure 4
See code.

Accuracy for the newton's method is:

$$\frac{85}{93} \cdot 100 = 91,39\,\%$$

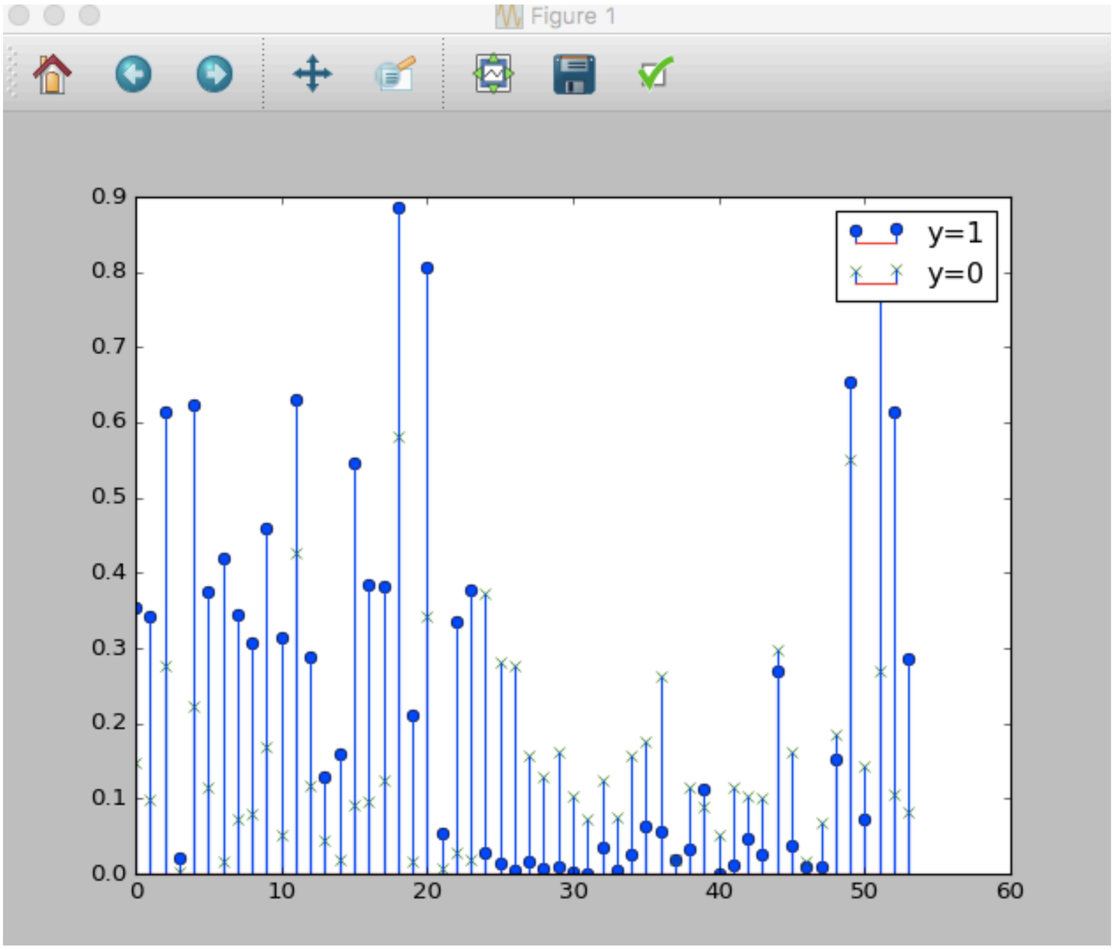

Figure 1

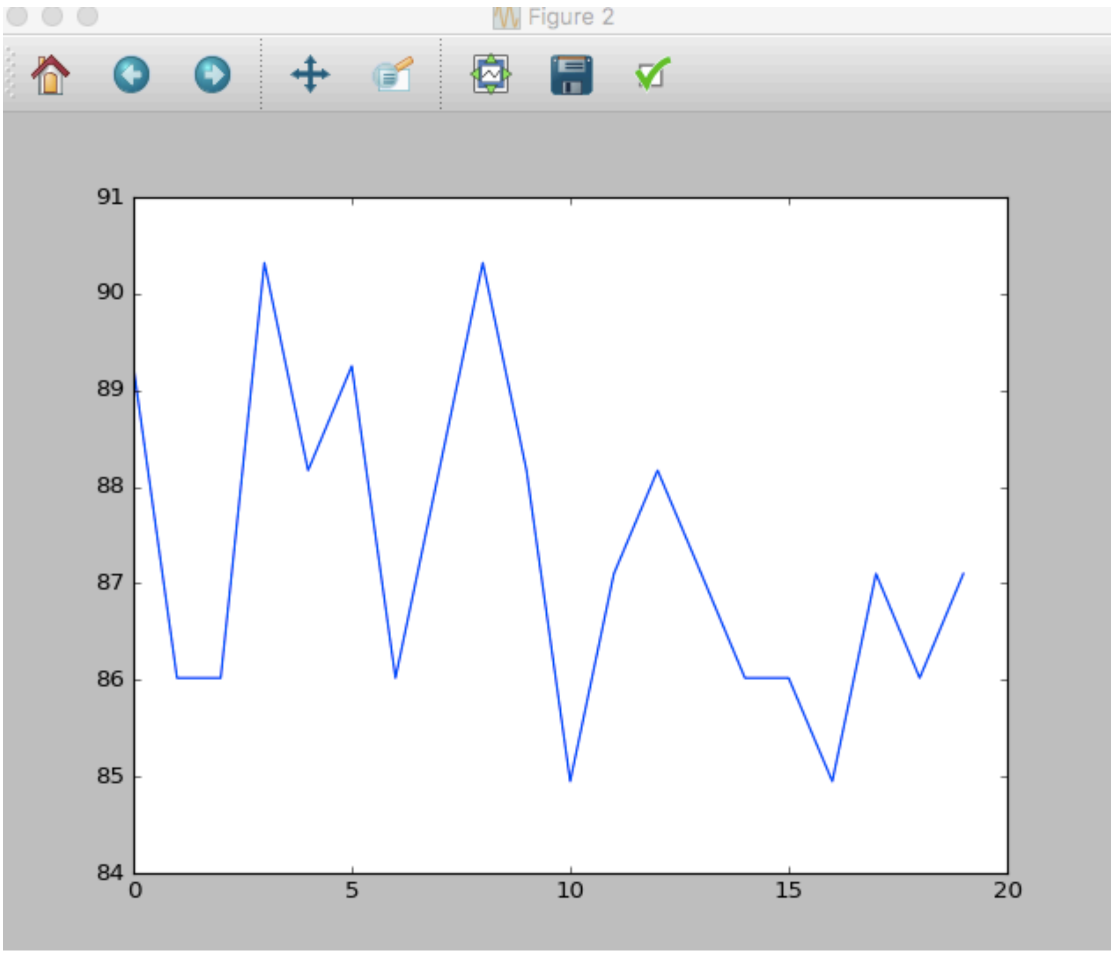

Figure 2

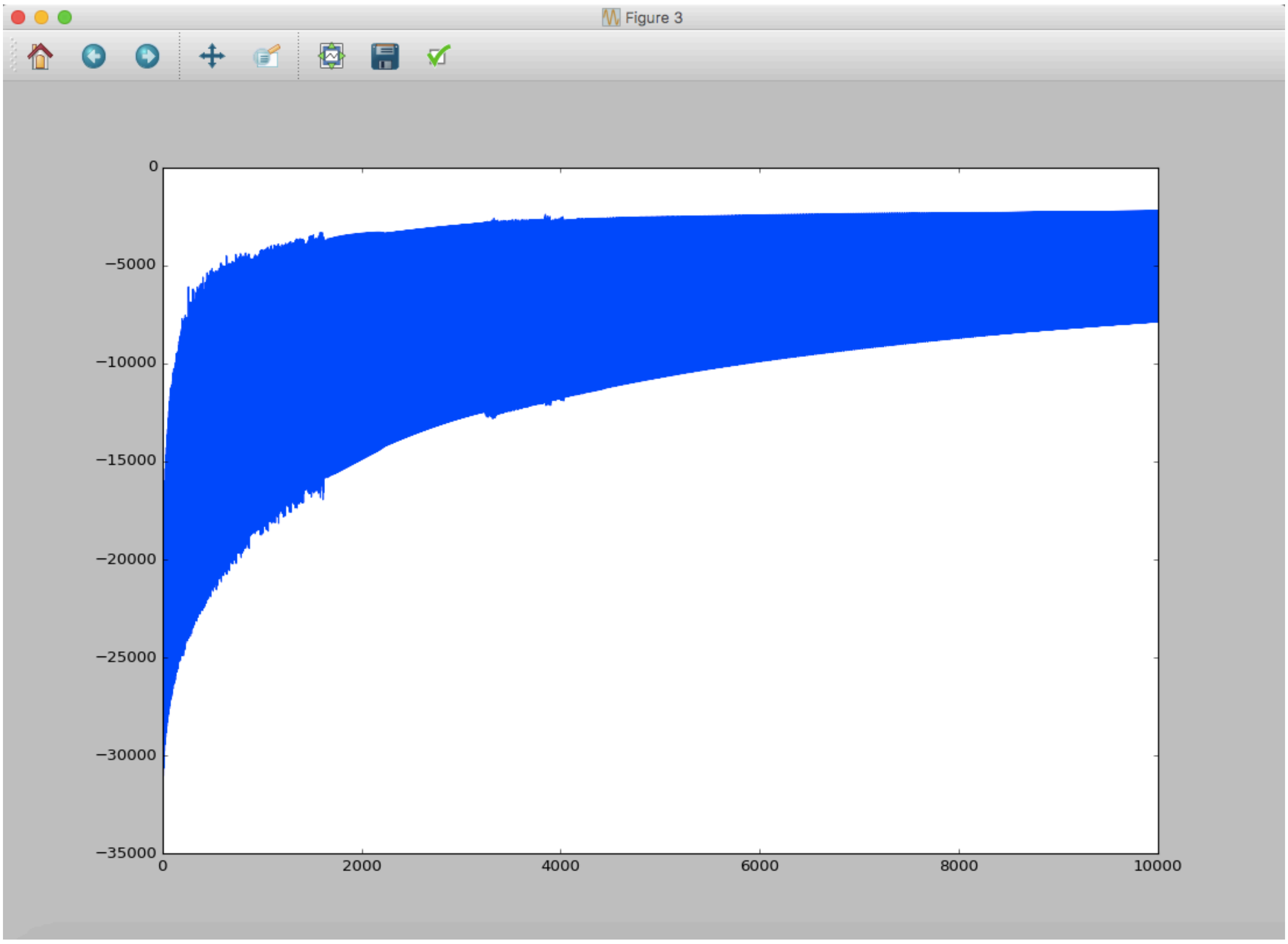

Figure 3

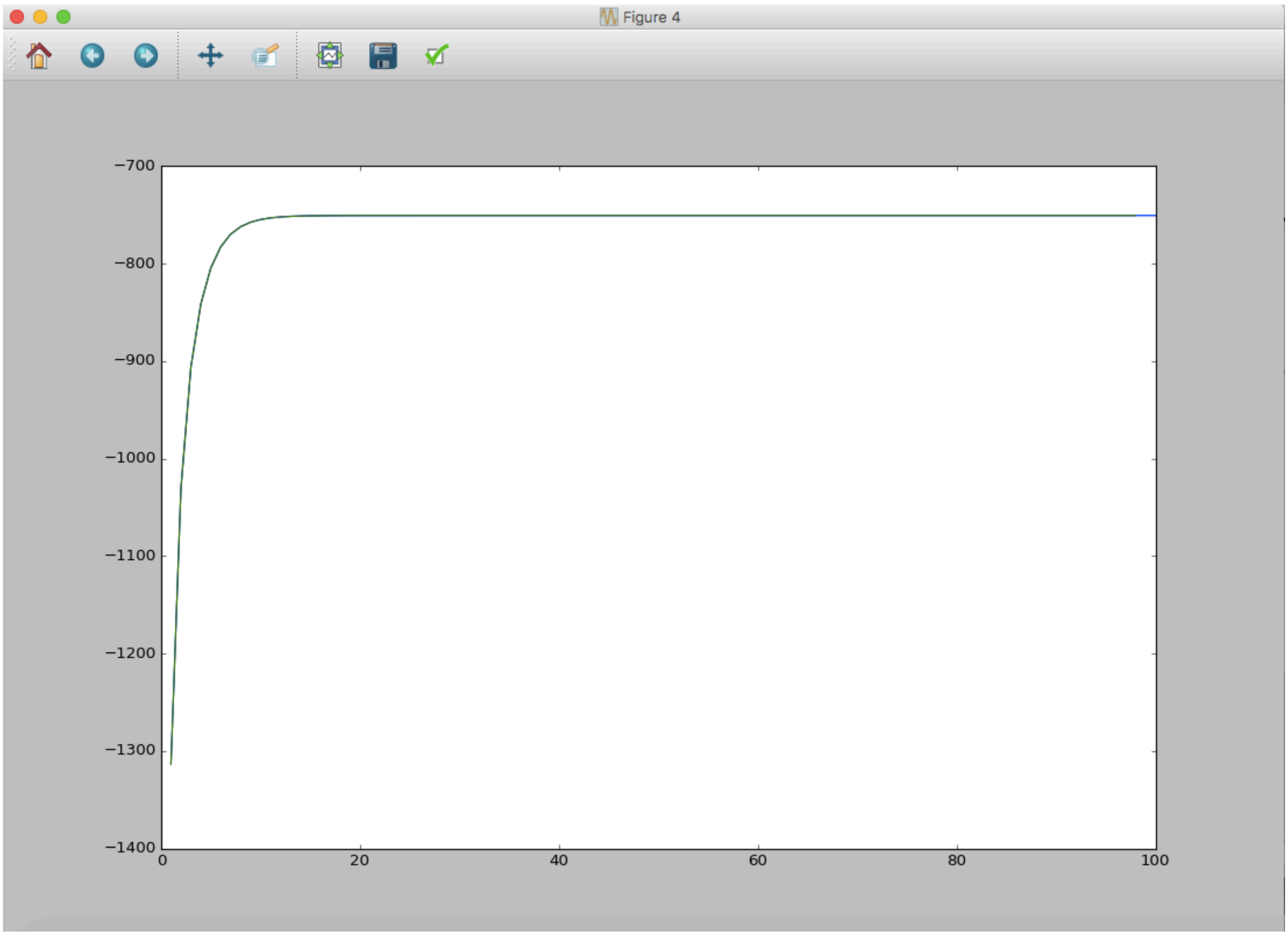

Figure 4