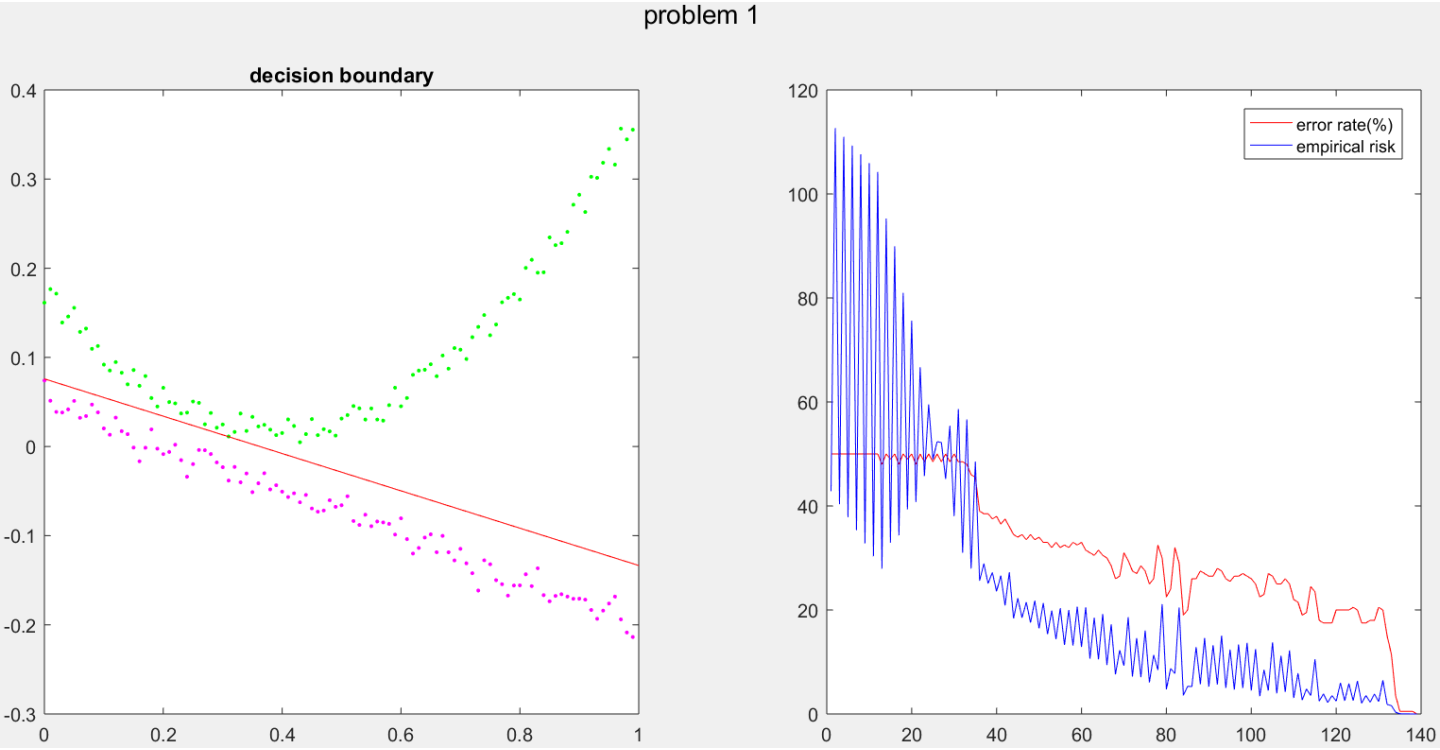
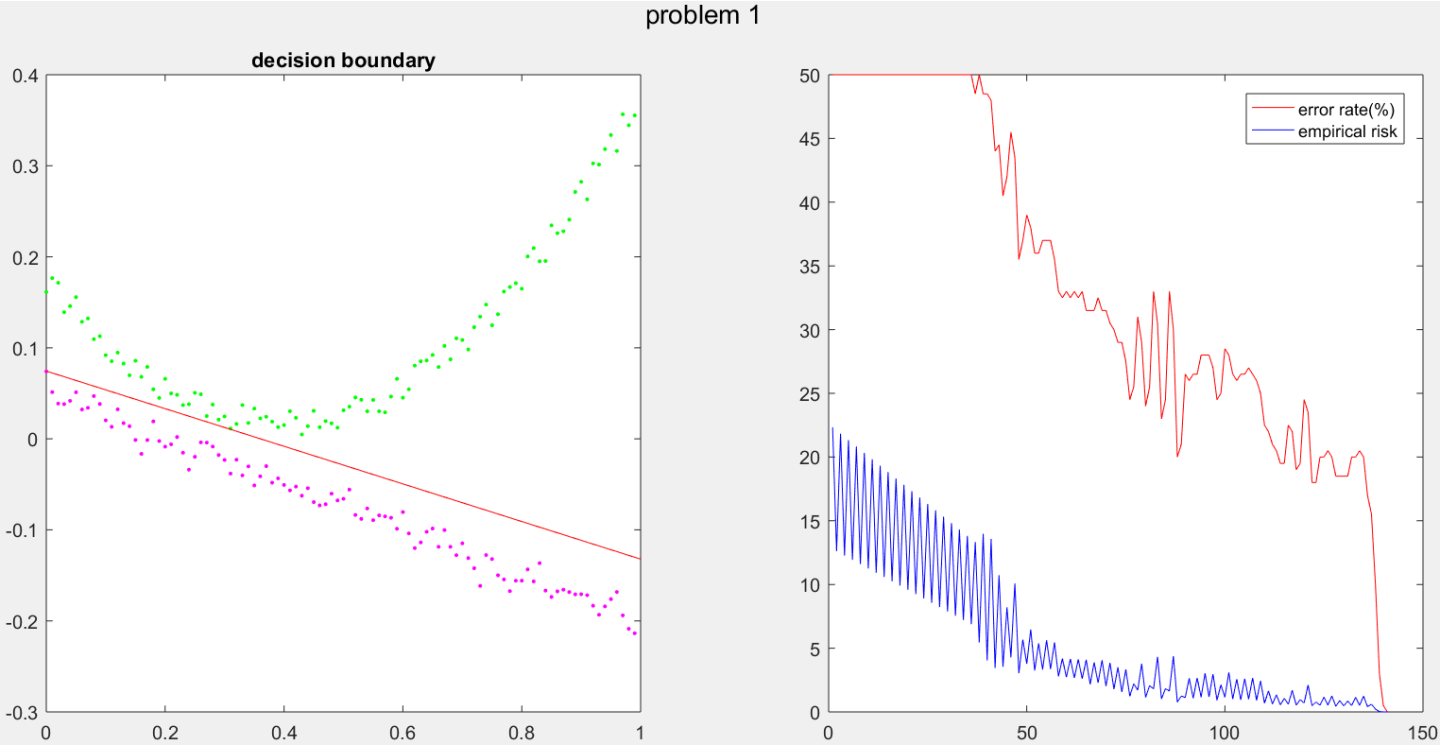
Problem 1

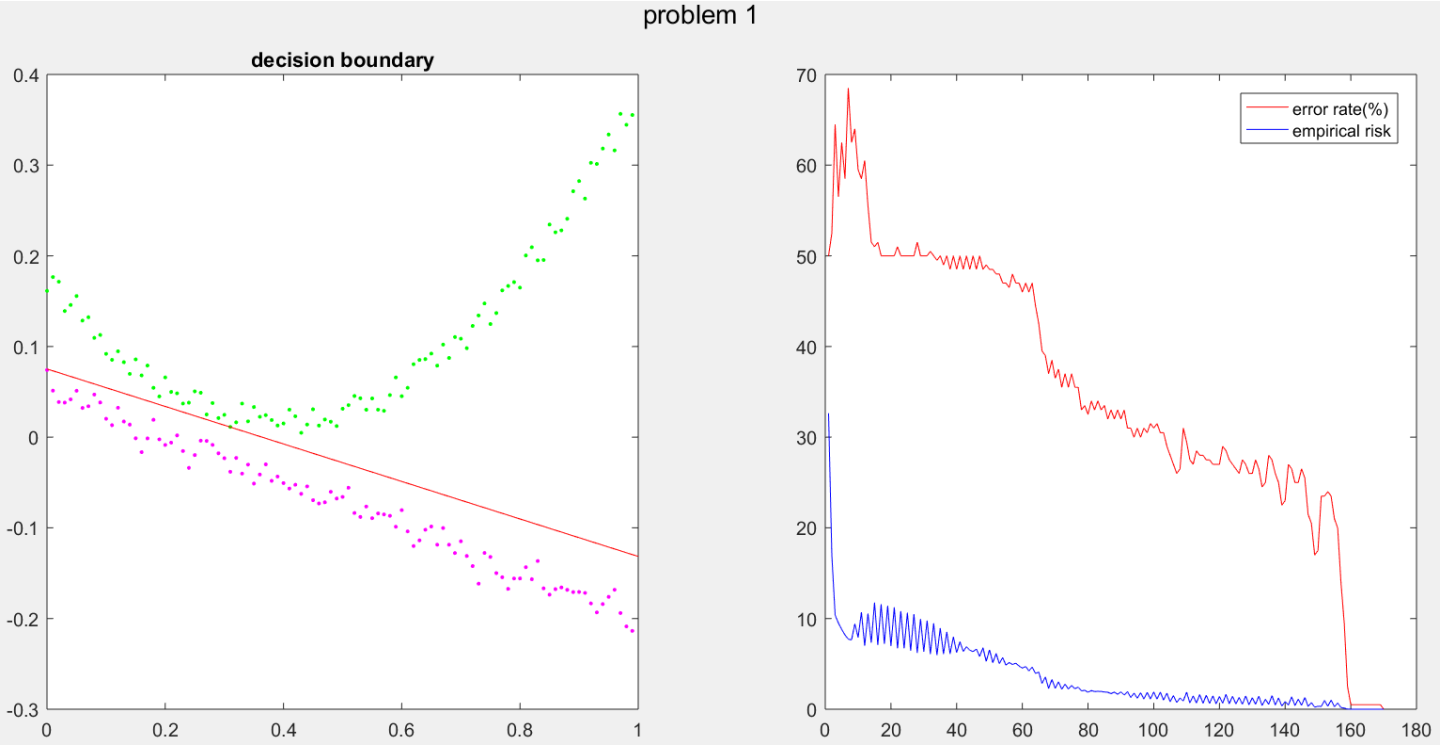
```
n = 5; % step size n
e = 0.0001; % tolerance e
```



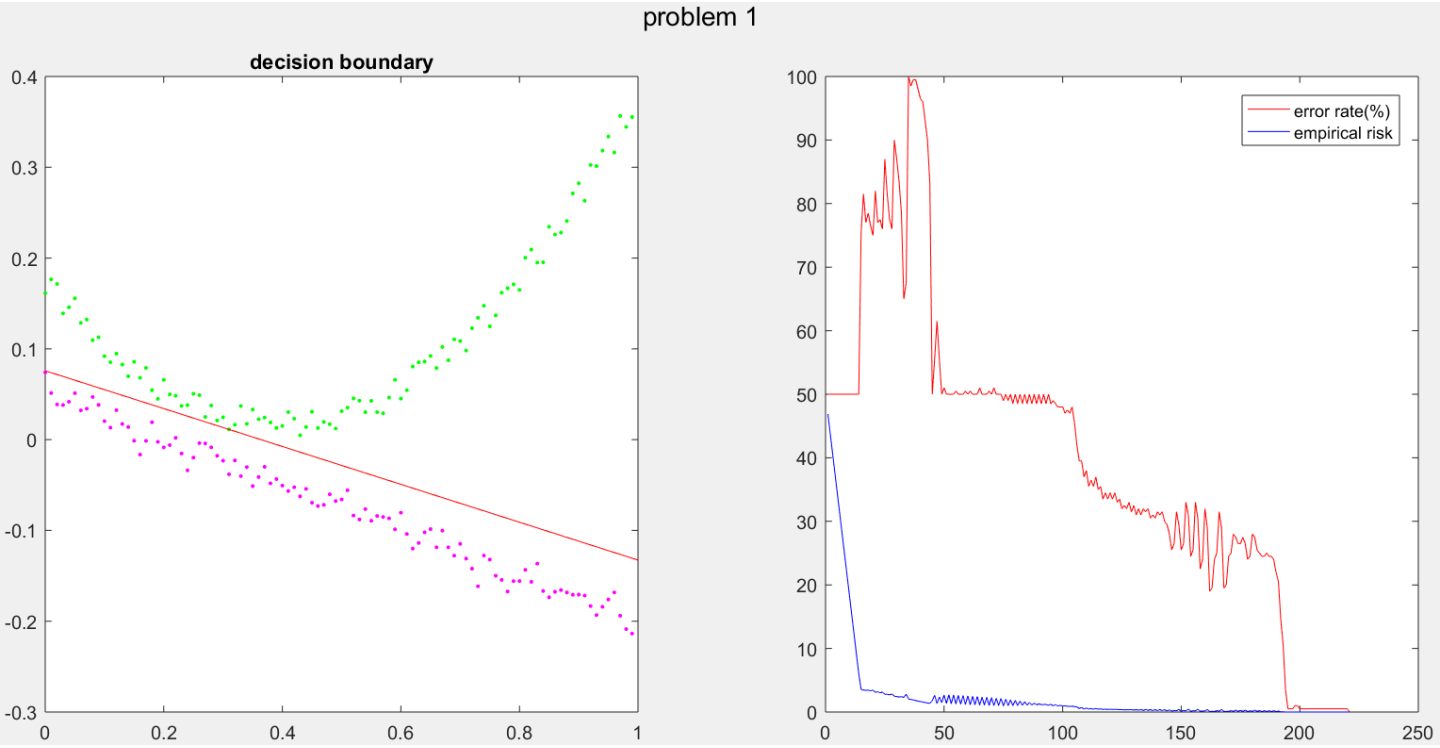

```
n = 1; % step size n
e = 0.0001; % tolerance e
```



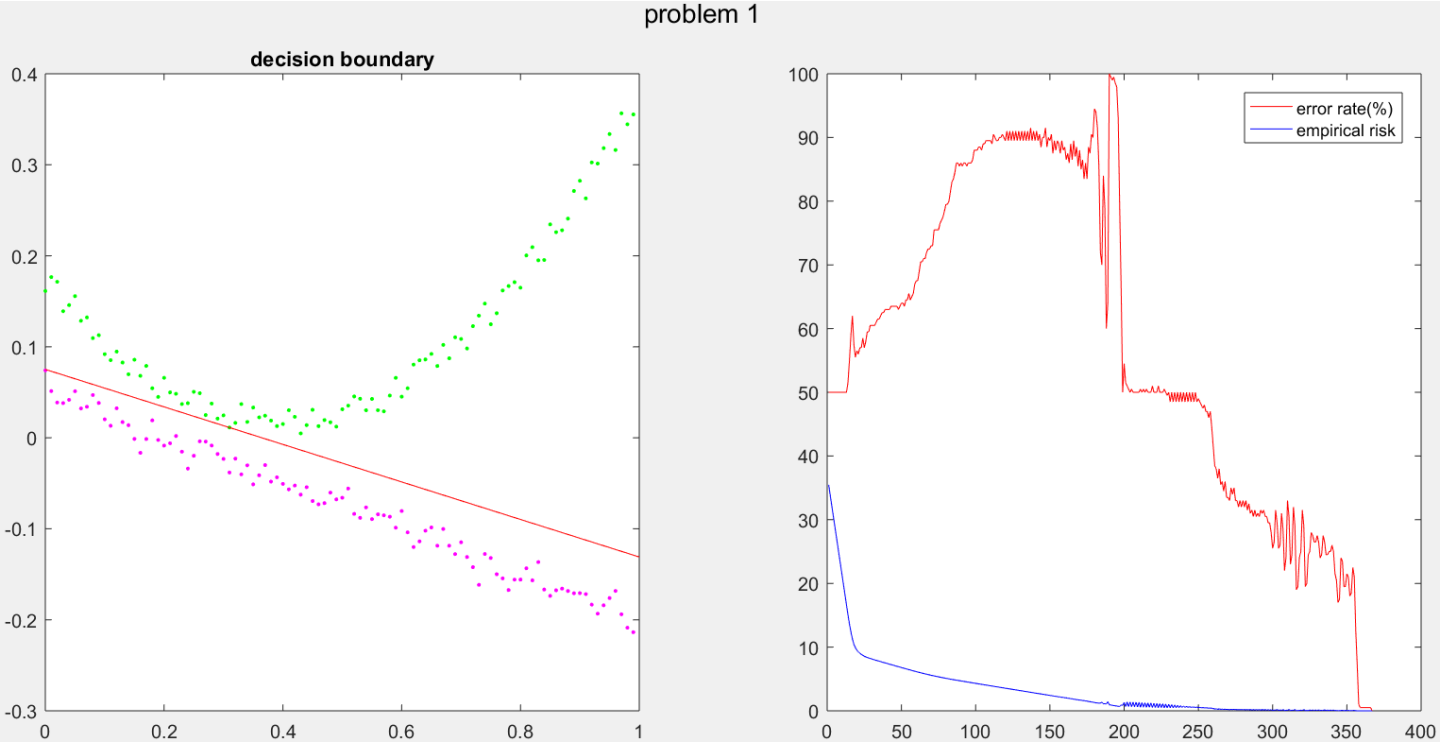
```
n = 0.5; % step size n
e = 0.0001; % tolerance e
```



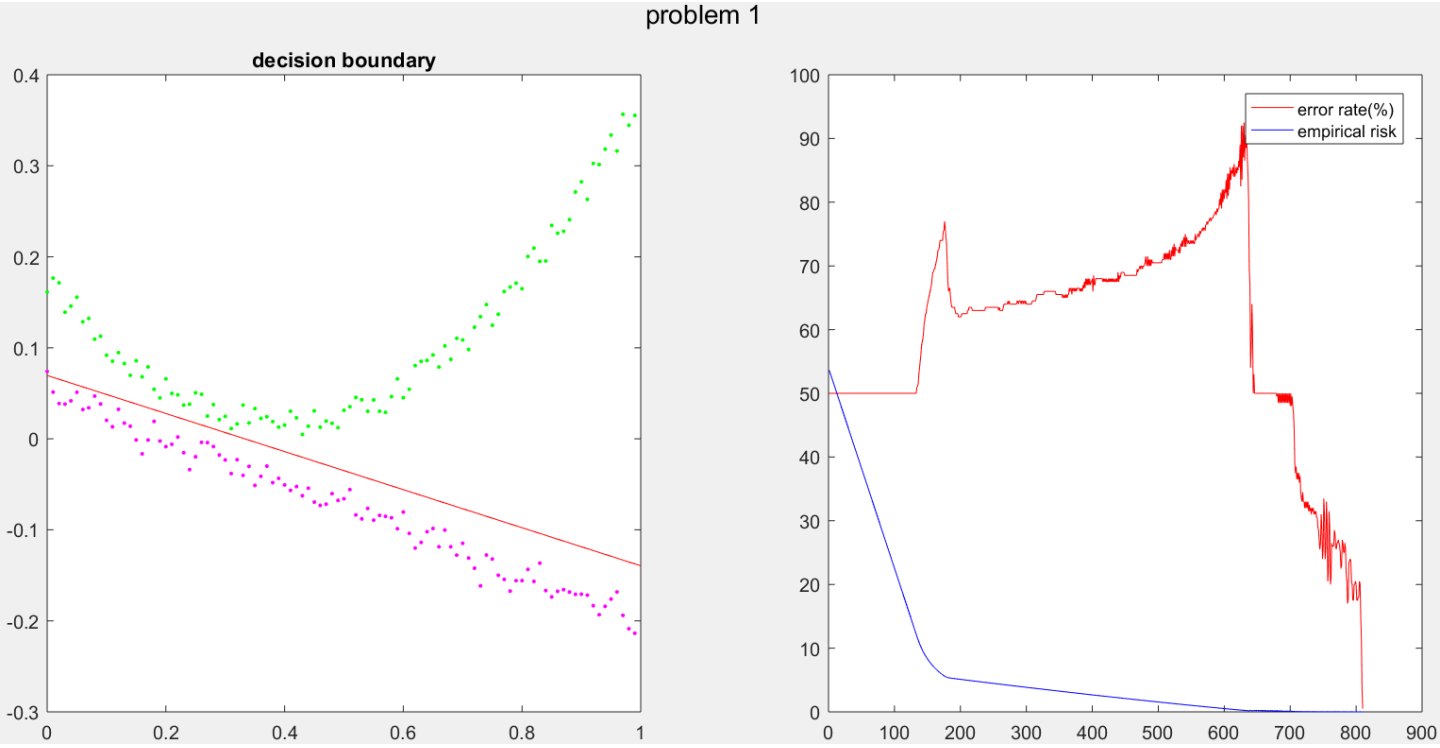

```
n = 0.1; % step size n
e = 0.0001; % tolerance e
```



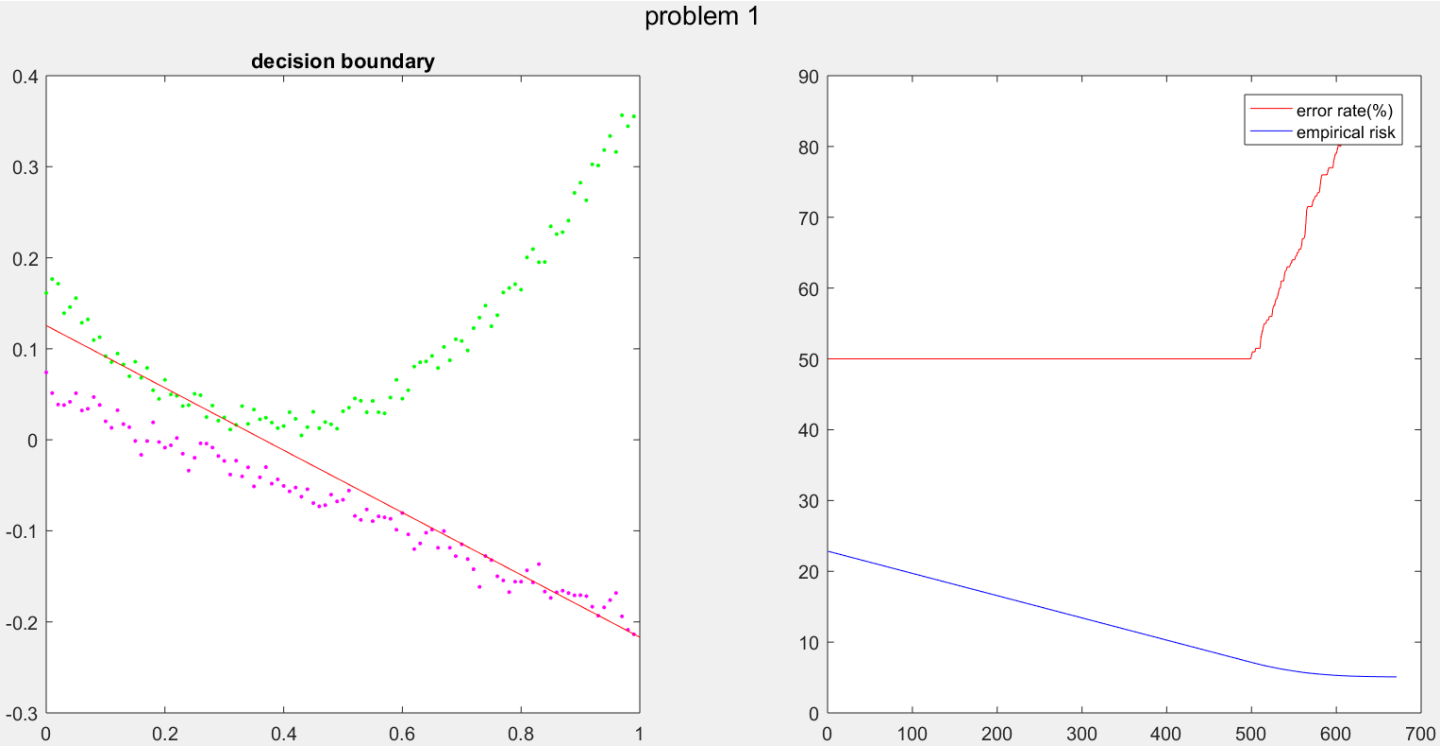
```
n = 0.05; % step size n
e = 0.0001; % tolerance e
```




```
n = 0.01; % step size n
e = 0.0001; % tolerance e
```



```
n = 0.001; % step size n
e = 0.0001; % tolerance e
```

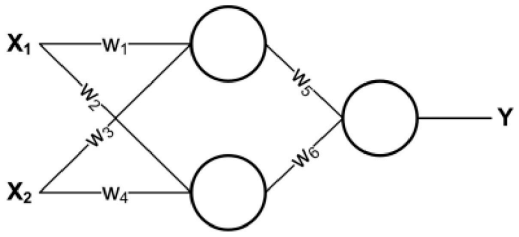


As the results shown above, the empirical risk becomes smoother as step size n decreases. (When step size is large, the empirical risk will oscillate strongly.)

As for error rate, it will be more unstable as step size n decreases. In the last case where step size is 0.001, the final error rate is even high(up to about 80%) when perceptron error become convergent. And error rate will reach a high point(from 60% up to 100%) during the iterations.

1.

2.



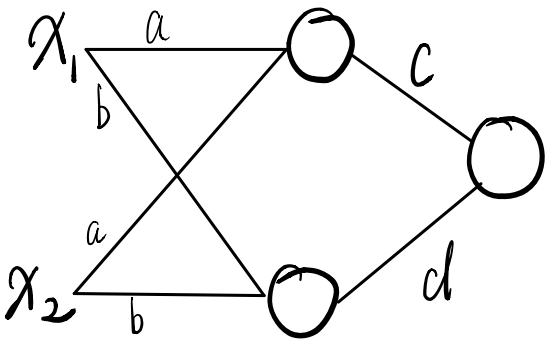
(a) In the old neural network:

$$\begin{cases} U_{11} = (w_1^T X_1 + w_3^T X_2) C \\ U_{12} = (w_2^T X_1 + w_4^T X_2) C \end{cases}$$

$$\Rightarrow Y = (w_5^T U_{11} + w_6^T U_{12}) C = (w_1^T w_5^T + w_2^T w_6^T) X_1 C^2 + (w_3^T w_5^T + w_4^T w_6^T) X_2 C^2$$

(b). Yes, it's possible. If a neural network is made up of only linear units, like the one above, the weight vectors are all constant vectors. Then the whole neural network can be seen as $CW^T X$, which W is the expression of w_i after multiplication and addition.

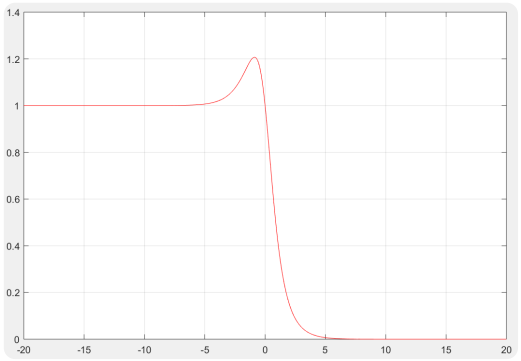
(c)

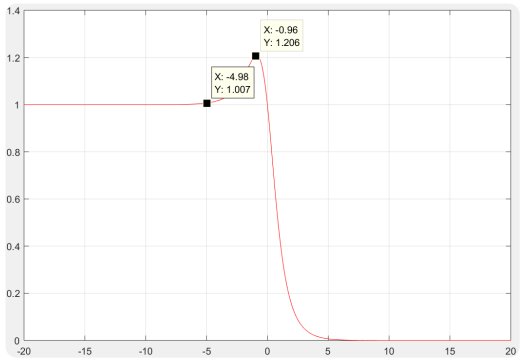


Set the weights as above, then:

x_1	x_2	output t	W^TA	
0	0	0	$(c+d)/2 \le 0$	①
1	1	0	$\frac{c}{1+e^{2a}} + \frac{d}{1+e^{2b}} \le 0$	②
1	0	1	$\frac{c}{1+e^{a}} + \frac{d}{1+e^{b}} > 0$	③
0	1	1		

now find the trend of $(1+e^{2x})$ and $(1+e^{x})$, assume $f(x) = \frac{1+e^{x}}{1+e^{2x}}$
plot the $f(x)$ in MATLAB we get following figure



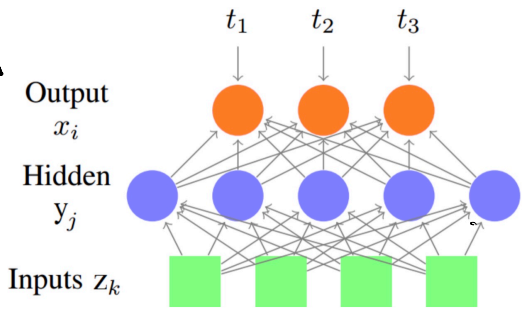


Obviously, $c \times d < 0$, then assume $c > 0$, $d < 0$.
If a is a negatively large number (less than -5), then
$f(a) = \frac{1+e^{a}}{1+e^{2a}} \approx 1 \Rightarrow 1+e^{a} \approx 1+e^{2a} \Rightarrow \frac{c}{1+e^{a}} \approx \frac{c}{1+e^{2a}} \approx c$

Next let $b = -1$, where is roughly the max point of $f(x)$.
As long as $\frac{-d}{1+e^{b}} < c \le \frac{-d}{1+e^{2b}}$, inequality ② & ③ hold.

say $a = -50$, $b = -1$, $d = -10$, $c = 8$ ($w_1 = w_3 = -50$, $w_2 = w_4 = -1$, $w_5 = 8$, $w_6 = -10$)
such network can compute the XOR problem

3.



(a)

$$E = -\sum_i [t_i \log x_i + (1-t_i)\log(1-x_i)]$$

$$x_i = \frac{1}{1+e^{-\sum_j y_j w_{ji}}}, \quad y_j = \frac{1}{1+e^{-\sum_k z_k w_{kj}}}$$

$$E = -\sum_i \left\{ t_i \left[\sum_j y_j w_{ji} - \log(1+e^{\sum_j y_j w_{ji}}) \right] + (1-t_i)\cdot\left[-\log(1+e^{\sum_j y_j w_{ji}}) \right] \right\}$$

$$= -\sum_i \left[t_i \sum_j y_j w_{ji} - \log(1+e^{\sum_j y_j w_{ji}}) \right]$$

$$\frac{\partial E}{\partial w_{ji}} = -\sum_i \left(t_i y_j - \frac{1}{1+e^{\sum_j y_j w_{ji}}} \times e^{\sum_j y_j w_{ji}} \times y_j \right)$$

note: $\frac{\partial (\sum_j y_j w_{ji})}{\partial w_{ji}} = y_j$

$$= -\sum_i \left[\left(t_i - \frac{1}{1+e^{-\sum_j y_j w_{ji}}} \right) y_j \right] = -\sum_i \left[(t_i - x_i) y_j \right]$$

$$\frac{\partial E}{\partial w_{kj}} = \frac{\partial E}{\partial z_k w_{kj}} \frac{\partial z_k w_{kj}}{\partial w_{kj}} = \frac{\partial E}{\partial z_k w_{kj}} \cdot z_k = \frac{\partial E}{\partial y_j w_{ji}} \cdot \frac{\partial y_j w_{ji}}{\partial z_k w_{kj}} \cdot z_k$$

$$= -\sum_i (t_i - x_i)\cdot(1-y_j) y_j w_{ji} z_k$$

(note: $\frac{\partial E}{\partial y_j w_{ji}} = -\sum_i (t_i - x_i)$, $\frac{\partial y_j}{\partial z_k w_{kj}} = \frac{e^{-\sum_k z_k w_{kj}}}{(1+e^{-\sum_k z_k w_{kj}})^2} = y_j(1-y_j)$)

(b) $E = -\sum_i t_i \log(x_i)$

$$x_i = \frac{e^{-\sum_j y_j w_{ji}}}{\sum_{c=1}^{m} e^{-\sum_j y_j w_{jc}}}, \quad y_j = \frac{1}{1+e^{-\sum_k z_k w_{kj}}}$$

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial y_j w_{ji}} \cdot \frac{\partial y_j w_{ji}}{\partial w_{ji}} = \frac{\partial E}{\partial y_j w_{ji}} \cdot y_j$$

$$= \sum_i t_i \left(1 - \frac{e^{-\sum_j y_j w_{ji}}}{\sum_{c=1}^{m} e^{-\sum_j y_j w_{jc}}}\right) \cdot y_j = \sum_i t_i (1 - x_i) \cdot y_j$$

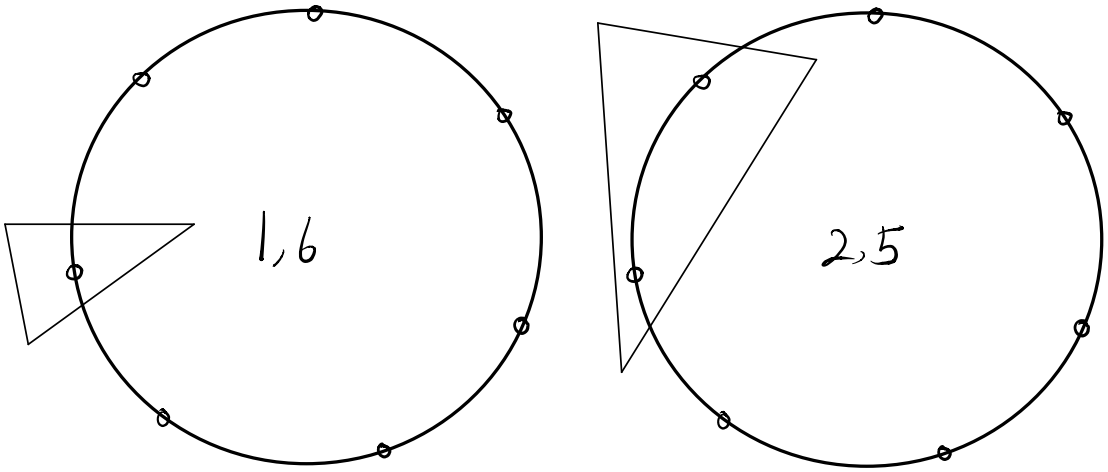
$$\frac{\partial E}{\partial w_{kj}} = \frac{\partial E}{\partial z_k w_{kj}} \frac{\partial z_k w_{kj}}{\partial w_{kj}} = \frac{\partial E}{\partial z_k w_{kj}} \cdot z_k = \frac{\partial E}{\partial y_j w_{ji}} \cdot \frac{\partial y_j w_{ji}}{\partial z_k w_{kj}} \cdot z_k$$

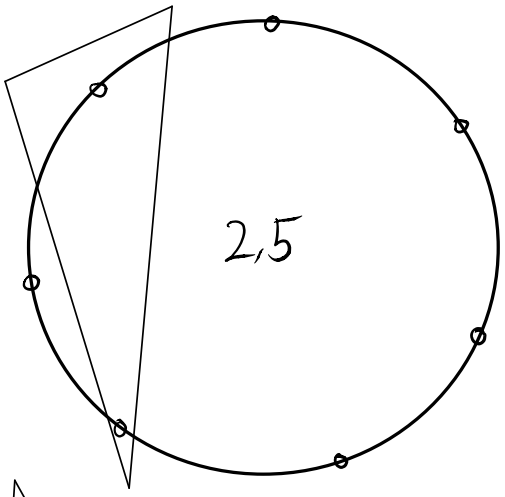
$$= \sum_i t_i (1 - x_i) \cdot (1 - y_i)\, y_i\, w_{ji}\, z_k$$

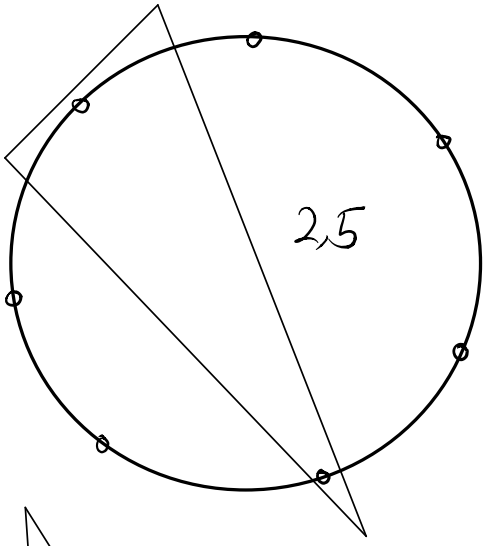
(note: $\frac{\partial y_i}{\partial z_k w_{kj}} = \frac{e^{-\sum_k z_k w_{kj}}}{(1 + e^{-\sum_k z_k w_{kj}})^2} = y_i (1 - y_i)$)

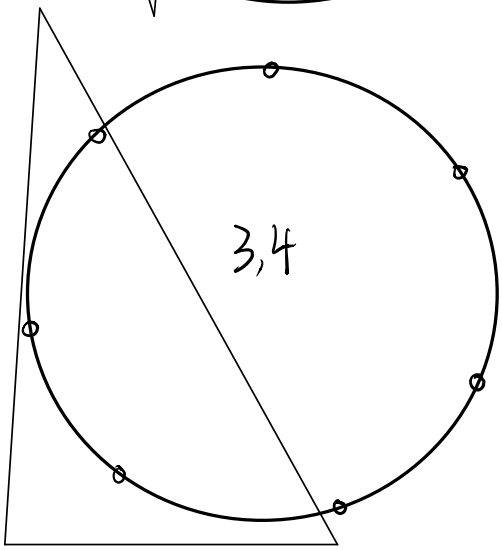
4. the VC-dimension of a triangle is at least 7.
All possible labelings of the seven points aligned on a circle can be seperated using the triangles.

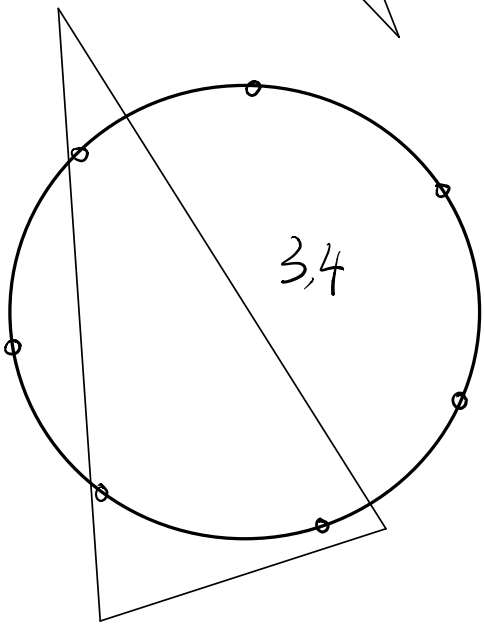
I can't find a scenary of eight points whose all possible labelings are seperable using a triangle. Also, VC is not neccessarily proportional to # of parameter.

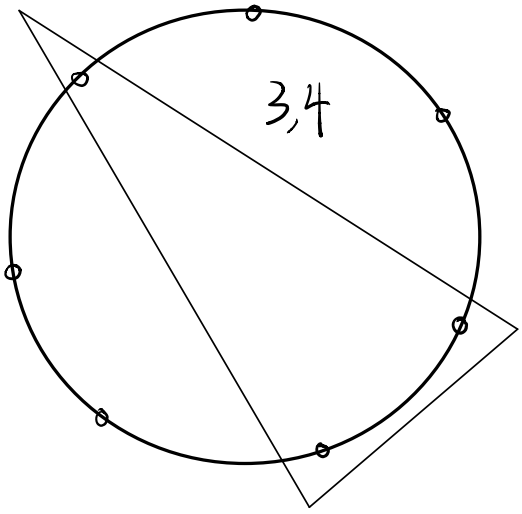


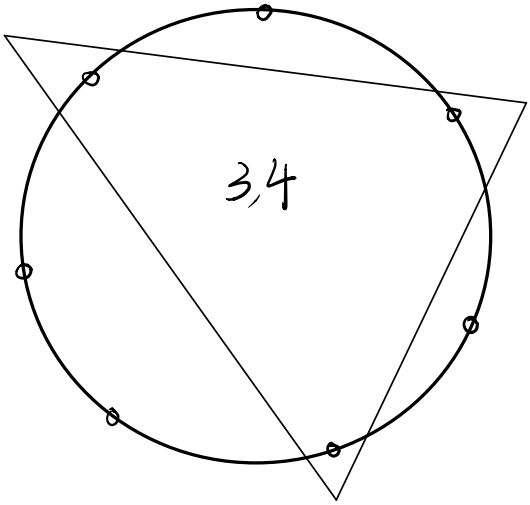

2,5


2,5


3,4


3,4


3,4


3,4