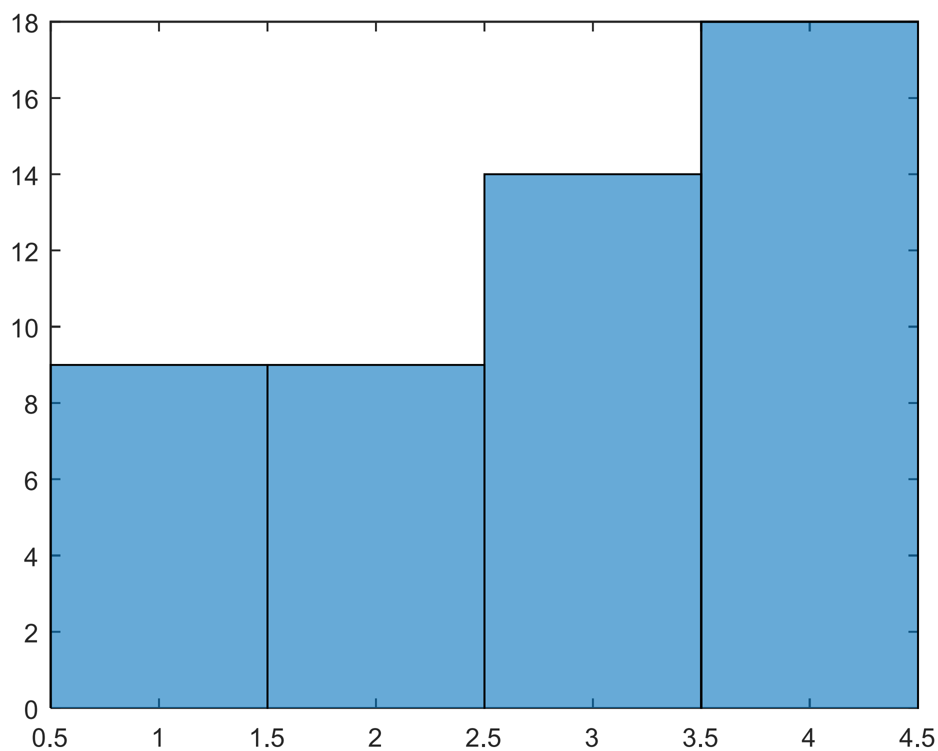
Mark Gluzman
Mg3627
ELENE 4903
Homework 3
Due 3/25

- Part 1 (20 points)

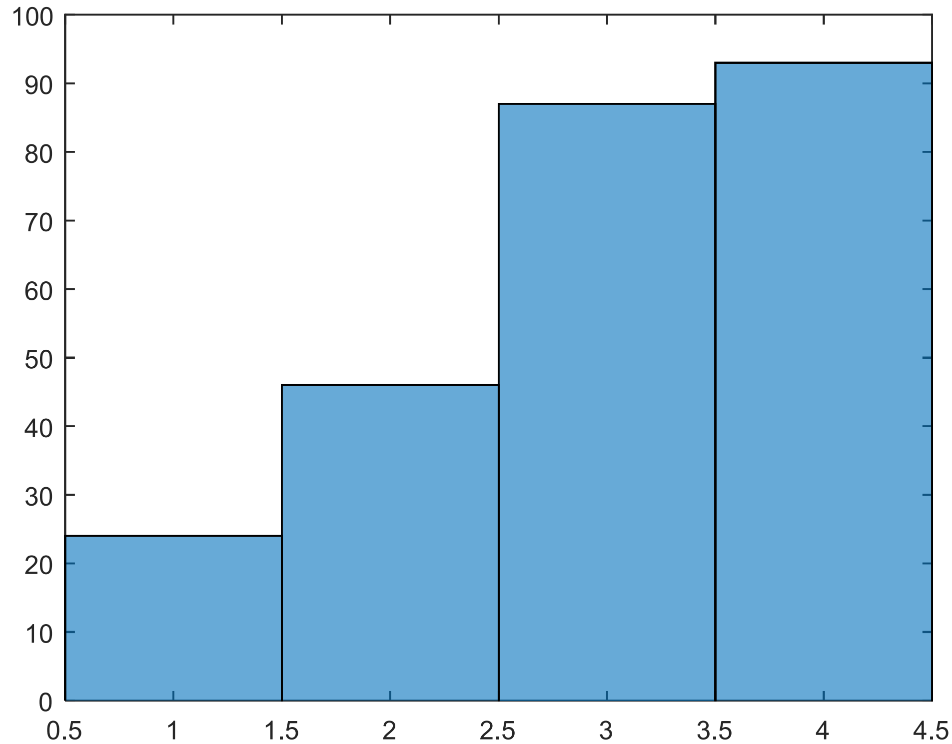
  Write a function that samples a discrete random variable. You will use this to implement Step 1 of the boosting algorithm given above. The function should take in a positive integer n and a k-dimensional probability distribution w. It should return a $1 \times n$ vector c, where each $c_i \in \{1, \ldots, k\}$ and $\text{Prob}(c_i = j|w) = w(j)$. The entries of c should be independently generated. For the distribution $w = [0.1,\ 0.2,\ 0.3,\ 0.4]$, show the histogram of one sampled vector c when $n = 50, 250, 500$.

Solution

N=50

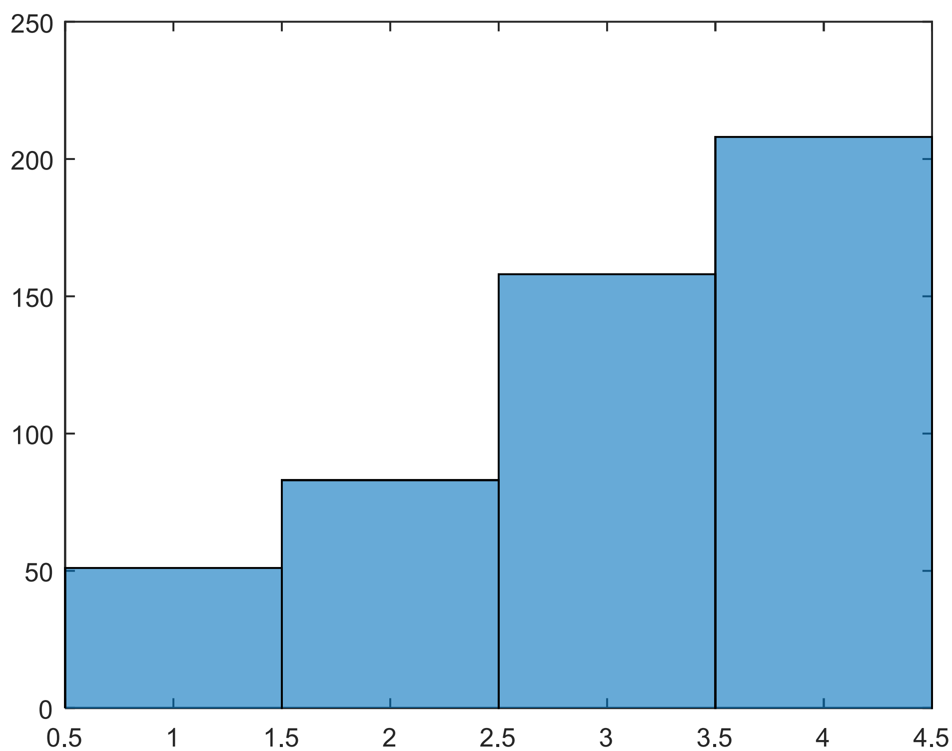


N=250



100
90
80
70
60
50
40
30
20
10
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5

N=500



250
200
150
100
50
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5

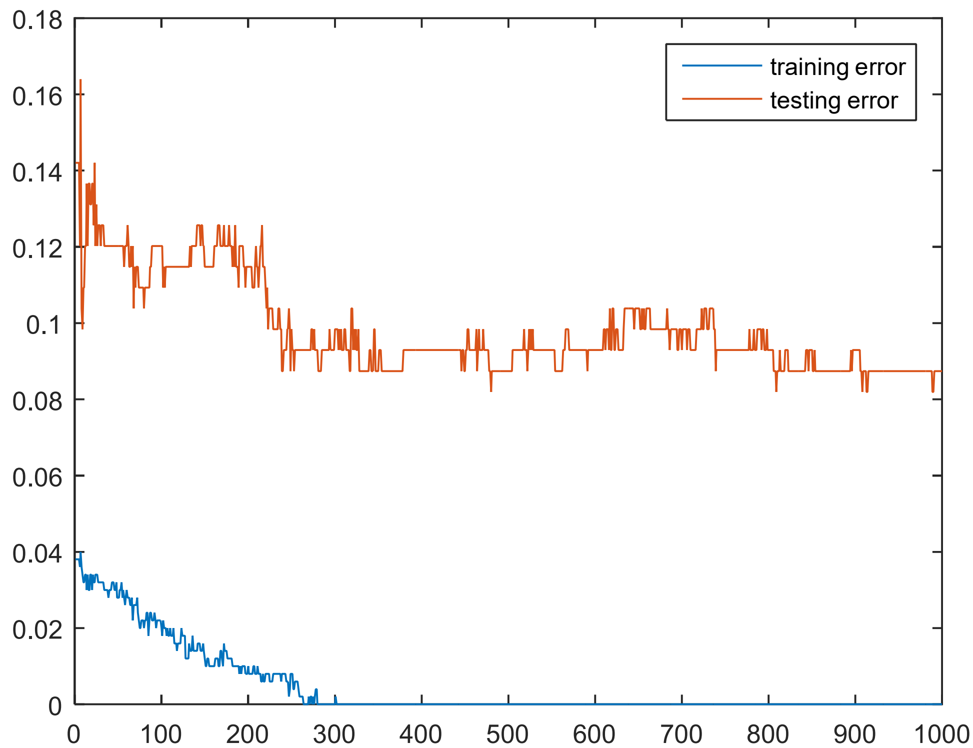
- Part 2 (50 points)

For $T = 1000$ iterations of boosting, do the following:

1. Implement a boosted version of this Bayes classifier, where class-specific π and μ, and shared Σ are learned on the bootstrap set $\mathcal{B}_t$. Notice that you only need to store w_0 and w for this problem, as written in the equation above. Since the data already contains a bias dimension equal to 1, you can store a single "augmented" vector where w_0 and w are combined. (When calculating μ_1, μ_0 and Σ, make sure you don't use this extra dimension!)

2. On a single plot, show the training and testing error as a function of iteration t for $t = 1, \ldots, T$.

Soluiton



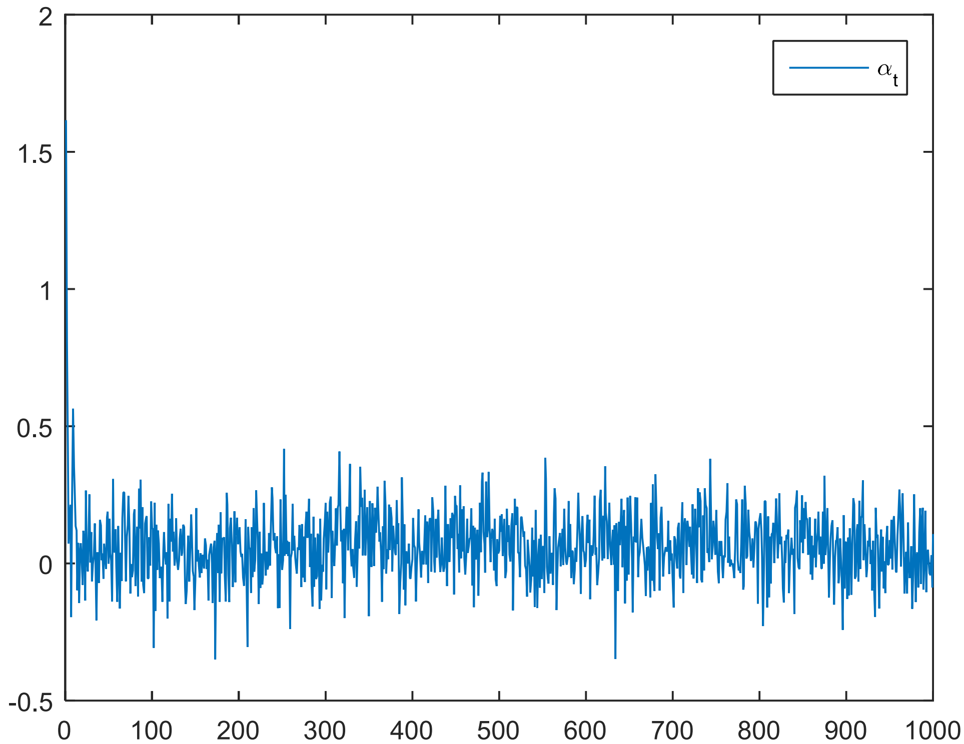
3. What is the testing accuracy for this Bayes classifier *without* boosting?

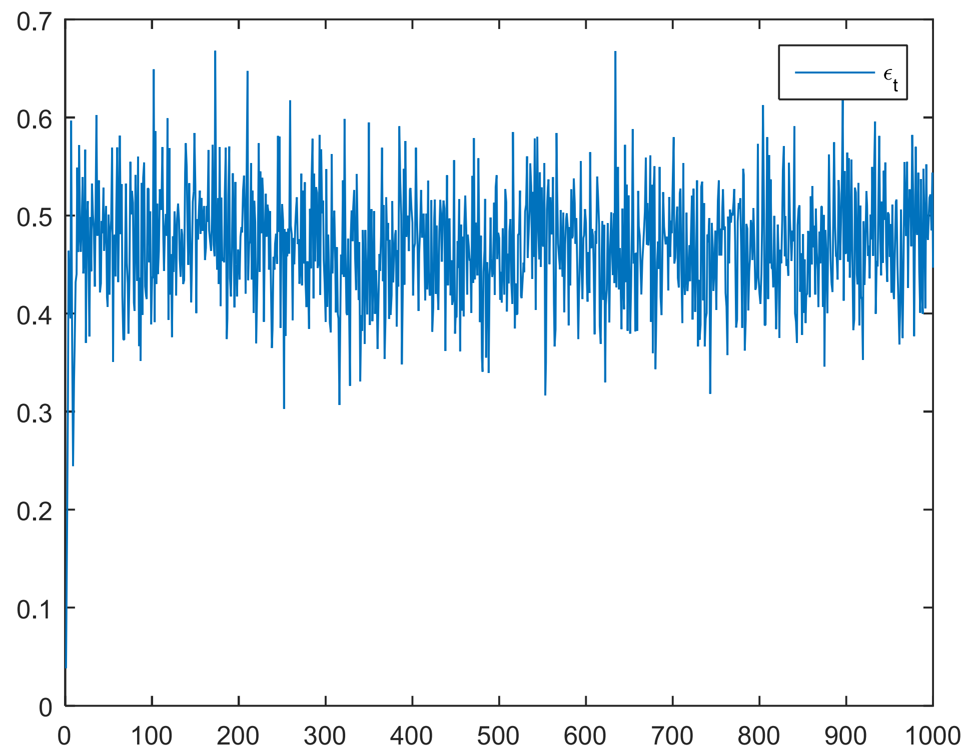
Solution

error_bayes =

0.1585

4. Plot α_t and ϵ_t as a function of t on different plots.

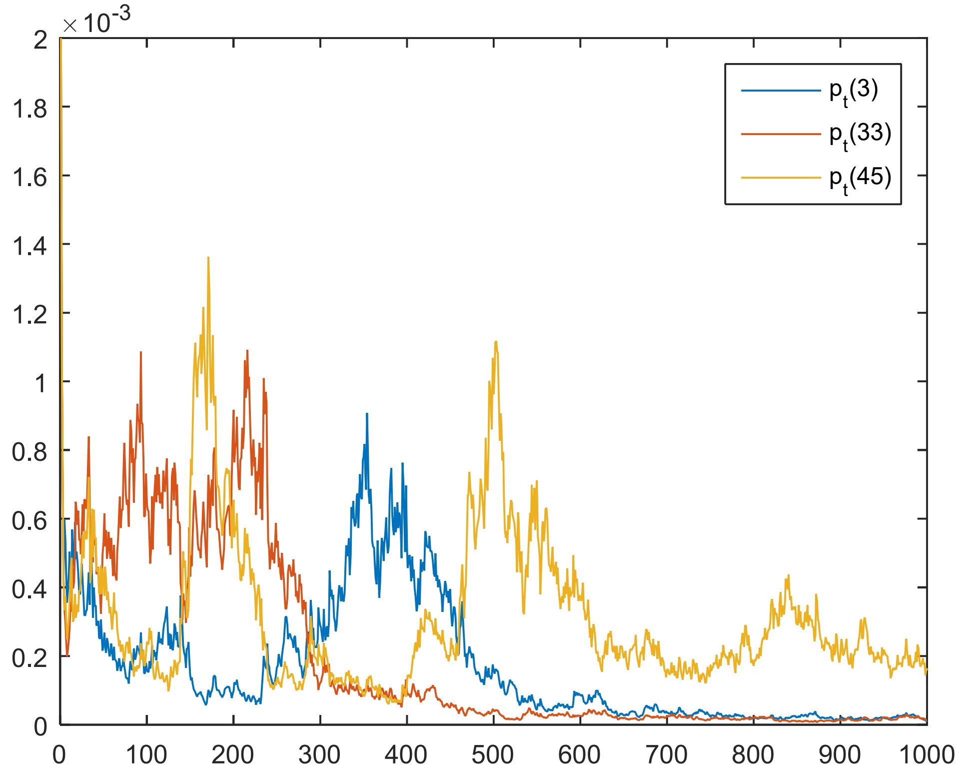




5. Pick 3 data points and plot their corresponding $p_t(i)$ as a function of t on the same plot. Select the points such that there is some variation in these values.

Solution

i=3,33,45

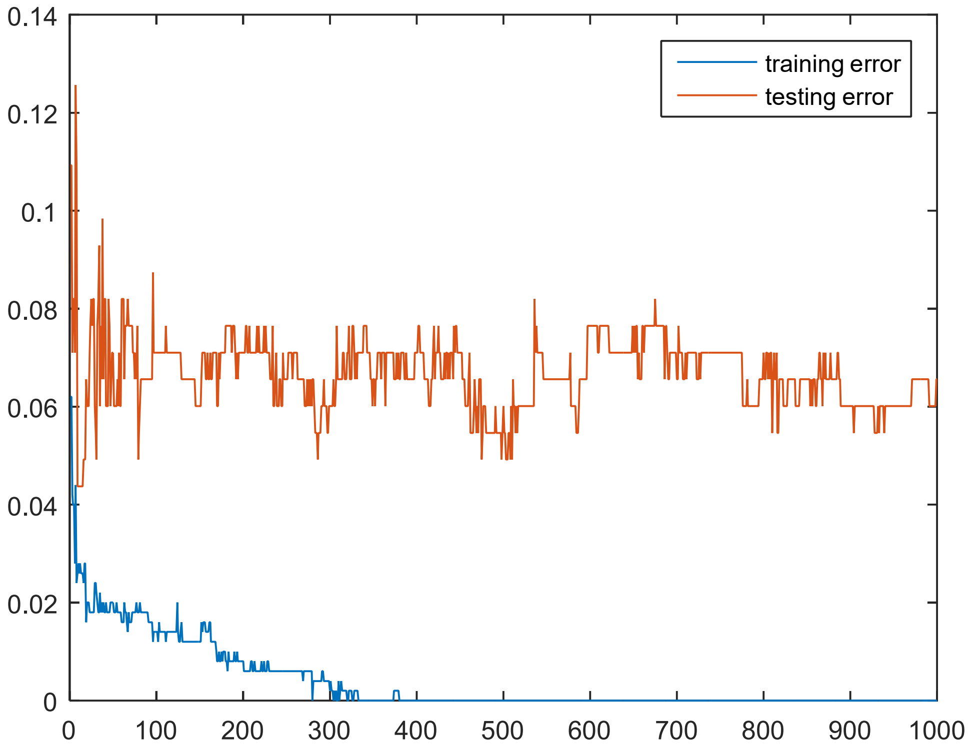


- Part 3 (50 points)

For $t = 1, \ldots, 1000$ iterations of boosting, do the following:

1. Implement the online logistic regression classifier described above.
2. On a single plot, show the training and testing error as a function of iteration t for $t = 1, \ldots, T$.

Solution



3. What is the testing accuracy of the logistic regression model *without* boosting? (You can use the two-class version of your logistic regression code from Homework 2, or your own implementation of binary logistic regression to do this.)

Solution

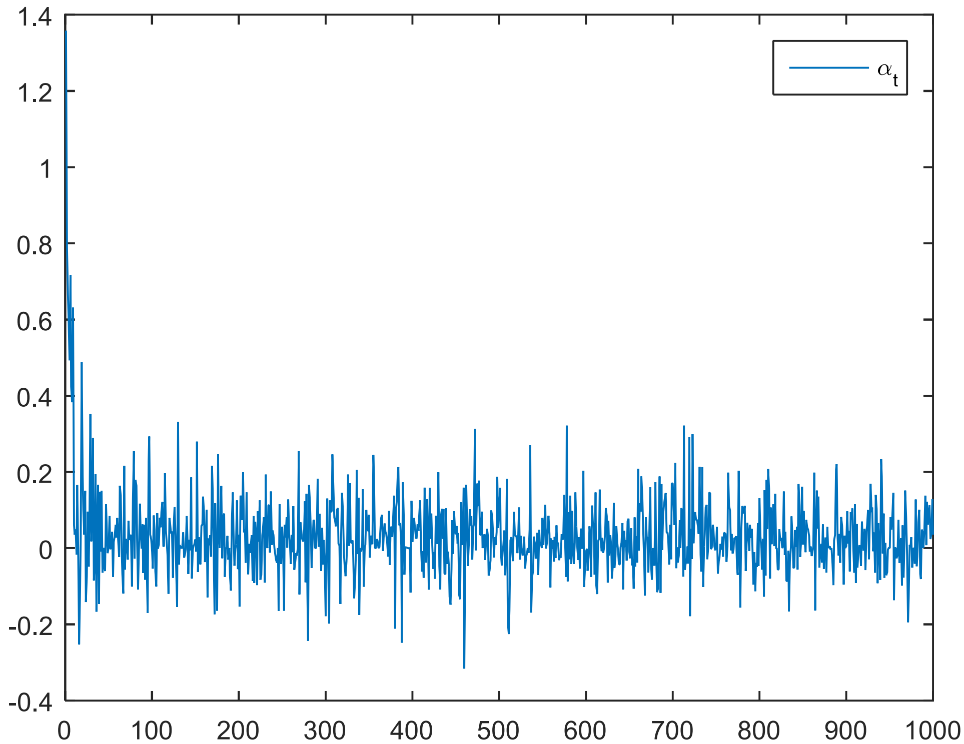
I used Logistic Regression Algorithm (Steepest Ascent) with 1000 iterations.

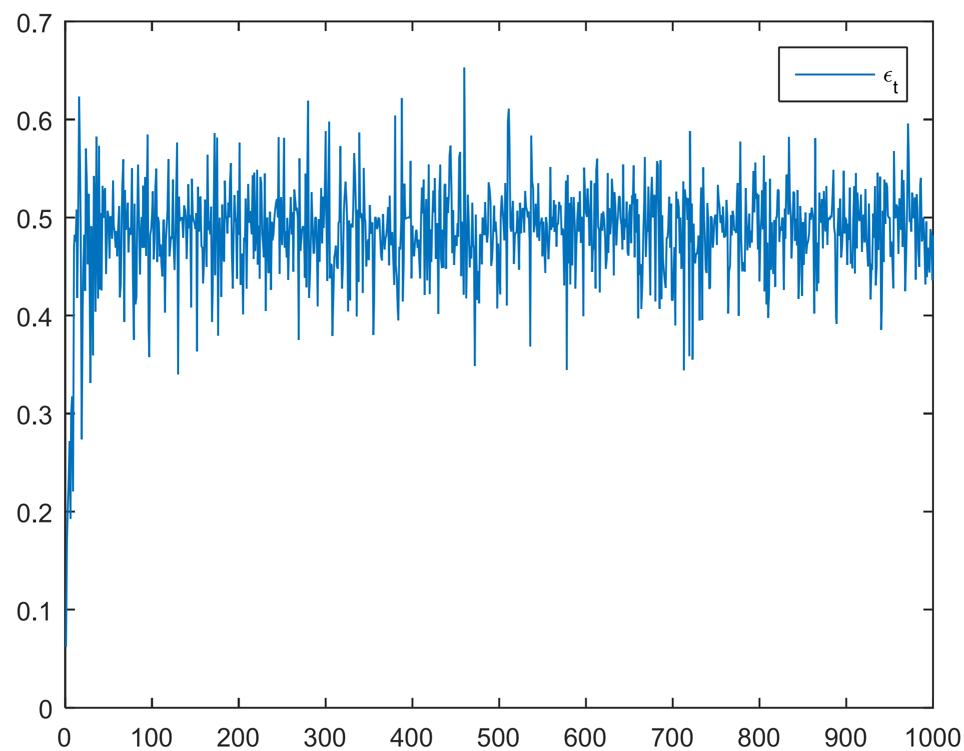
ans =

0.0273

4. Plot α_t and ϵ_t as a function of t on different plots.

Solution





5. Pick 3 data points and plot their corresponding $p_t(i)$ as a function of t on the same plot. Select the points such that there is some variation in these values.

Solution
I=27,38,42