Homework 2
CSCI4100

Han Hai
Rin:661534083
haih2@rpi.edu

September 17 2018

1. **[50 points] Exercise 1.8** If mu=0.9, what is the probability that a sample $\leq$0.1?

   $P = P[0] + P[1] = 0.1^{10} + C_{10}^{1} * 0.1^{9} * 0.9 = 9.1 * 10^{-9}$

2. **[50 points] Exercise 1.9** If mu=0.9, use the Hoeffding Inequality to bound the probability that a sample of 10 marbles will have v$\leq$0.1 and compare the answer to the previous exercise

   $P[|v - u| > \epsilon] \leq 2e^{-2\epsilon^2 N}$

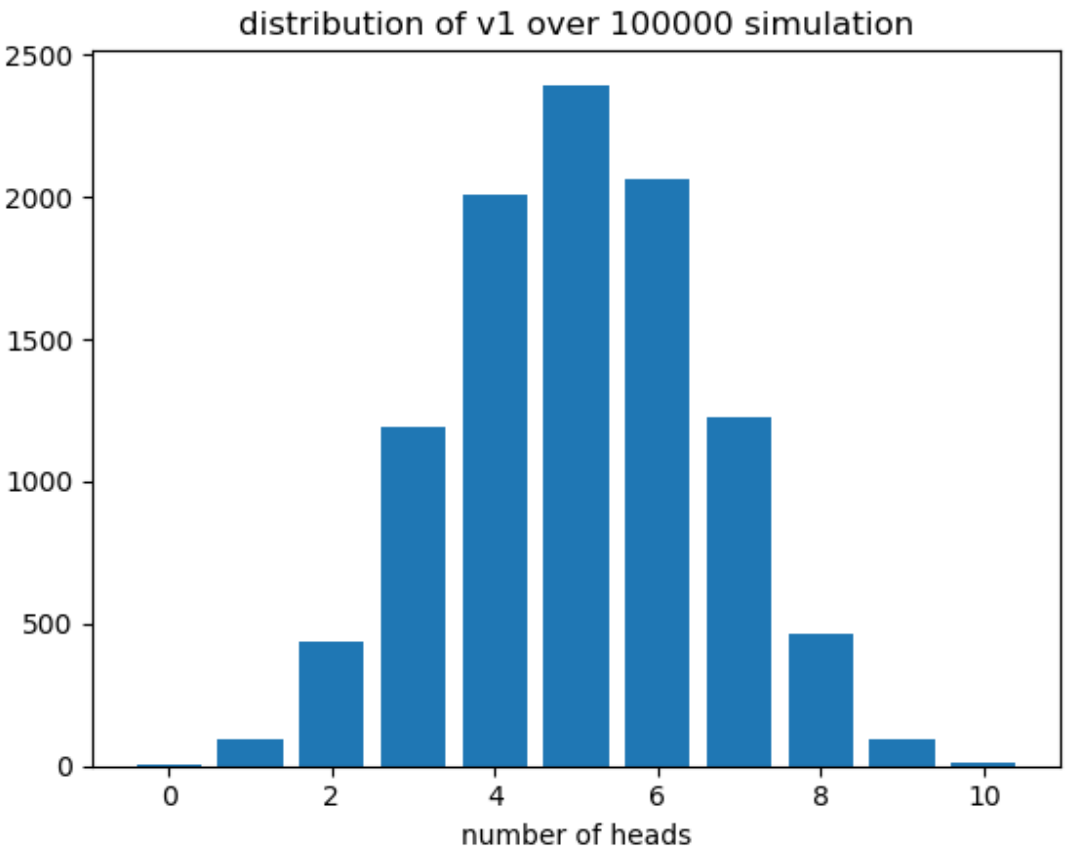
   $P[|0.1 - 0.9| > \epsilon]$ is bounded by ϵ 0.8

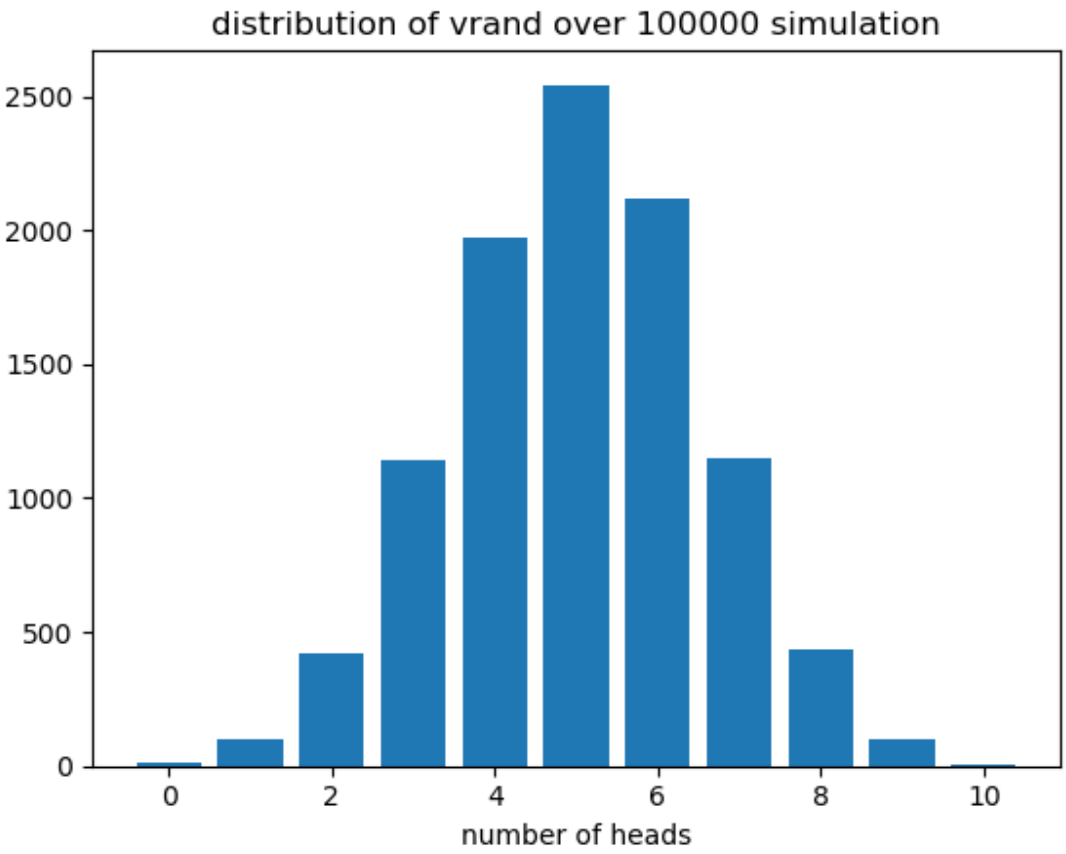
   $2e^{-2*0.8^2*10} = 5.522 * 10^{-6}$
   Comparing to Exercise 1.8, the probability is well within this boundary

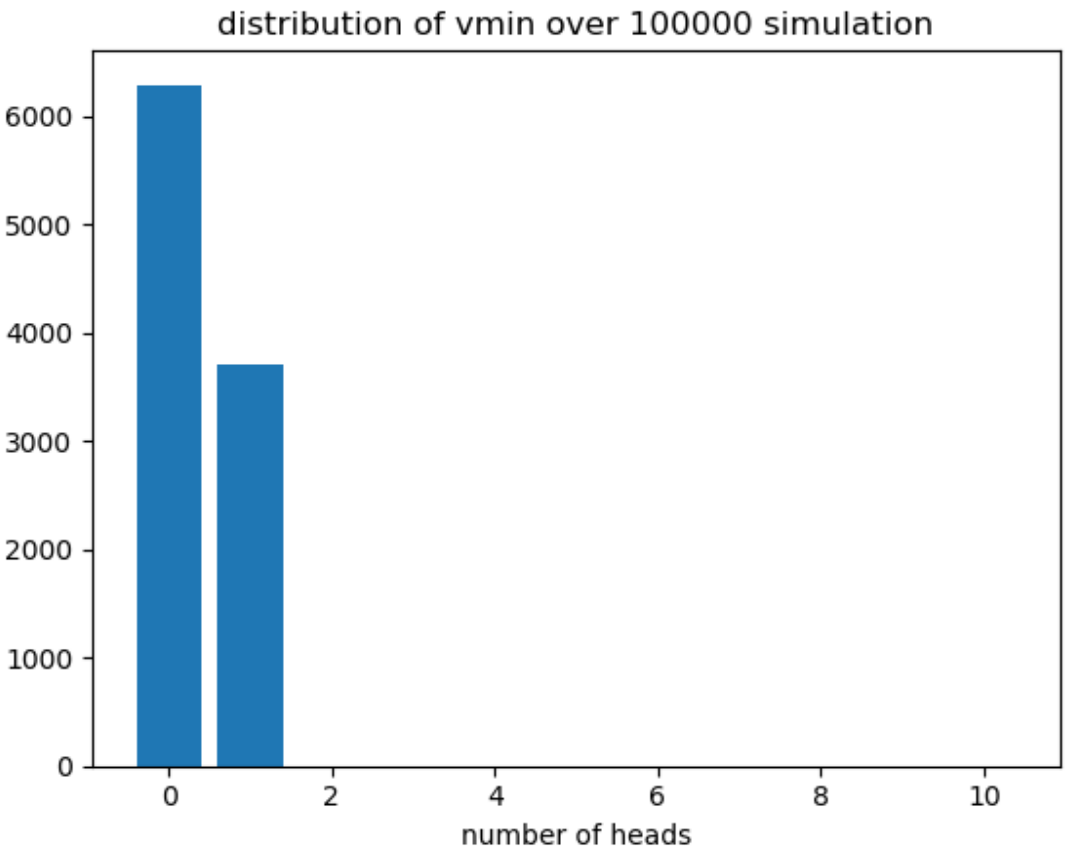
3. **[100 points] Exercise 1.10** Here is an experiment that illustrates the difference between a single bin and multiple bins. Run a computer simulation for flipping 1,000 fair coins. Flip each coin independently 10 times. Let's focus on 3 coins as follows: c_1 is the first coin flipped; c_{rand} is a coin you choose at random; c_{min} is the coin that had the minimum frequency of heads(pick the earlier one in case of a tie). Let v_1, v_{rand}, v_{min} be the fraction of heads you obtain for the respective three coins. For a coin, let μ be its probability of heads.

   (a) What is μ for the three coins selected?
   Since it's fair coin μ is 0.5

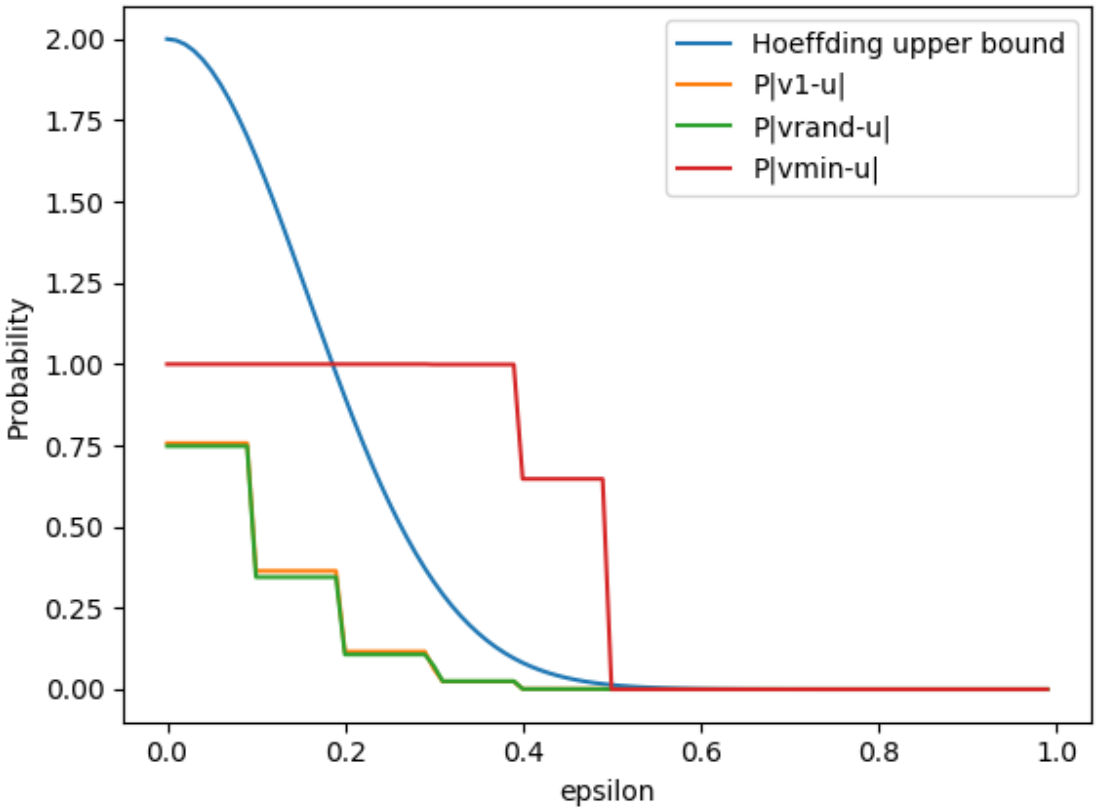
   (b) Repeat this entire experiment a large number of times, to get several instances of the three v and plot the histograms of the distribution of vs. Notice that which coins end up being c may differ from another


distribution of v1 over 100000 simulation
2500
2000
1500
1000
500
0
0
2
4
6
8
10
number of heads


distribution of vrand over 100000 simulation
2500
2000
1500
1000
500
0
0
2
4
6
8
10
number of heads


distribution of vmin over 100000 simulation
6000
5000
4000
3000
2000
1000
0
0
2
4
6
8
10
number of heads

(c) Using(b), plot estimates for $P[v - \mu] > \epsilon$ as a function ofϵ, together with the Hoeffding bound on the same graph.



(d) Which coins obey the Hoeffding bound, and which ones do not? Explain why.

v_1 and v_{rand} obey the bound, while v_{min} does not. Because v_{min} is a special case of sample, the minimum average of a 1000 coins is guaranteed to deviate from normal distribution, it therefore violates the rule that sample must be chosen at random in order for Hoeffding bound to work.

(e) Relate part (d) to the multiple bins in Figure 1.10.

Each of the three ways to pick a coinv_1, v_{rand}, v_{min} can be seen as a hypotheses. v_{min} is a biased bin that is a bad representation of the population.

4. **[100 points] Exercise 1.11** We are given a data set D of 25 training examples from an unknown target function $f : X \to Y$ where X is the set of real number and $Y = \{-1, +1\}$ To learn f, we use a simple hypothesis set$H = \{h1, h2\}$ where h1 is the constant +1 function and h2 is the constant -1. We consider two learning algorithms, S(smart) and C(crazy). S chooses the hypothesis that agrees the most with D and C chooses the other hypothesis deliberately. Let us see how these algorithms perform out of sample from the deterministic and probabilistic points of view. Assume in the probabilistic view that there is a probability distribution on X, and let $P[f(x) = +1] = p$

   (a) Can S produce a hypothesis that is guaranteed to perfrom better than random on any point outside D?

   No, according to Hoeffding, S can never guaranteed that. However, S can ensure it perform better than random some percent of the time

   (b) Suppose that all the examples in D have +1, Is it possible that the hypothesis that C produces turns out to be better than the hypothesis that S produces?

   It is possible, because the data might not be a good representation of the real population

   (c) If p=0.9, what is the probability that S will produce a better hypothesis than C?

   We want to find $P(P(S = f) > P(C = f))$ that is $P(0.9 > 0.1) = 1$

   (d) Is there any value of p for which it is more likely than not that C will produce a better hypothesis than S?

   Yes, any value of $p < 0.5$ that C will produce a better result

5. **[100 points] Exercise 1.12** A friend comes to you with a learning problem. She says the target function f is completely unknown, but she has 4000 data points. She is willing to pay you to solve her problem and produce for her a g which approximates f. What is the best that you can promise her among the following.

   (a) After learning you will provide her with a g that you will guarantee approximates f well out of sample.
   We cannot promise that

   (b) After learning you will provide her with a g, and with high probability the g which you produce will approximate f well out of sample.

   (c) One of the things will happen.

      i. You will produce a hyptohesis g;
      ii. You will declare that you failed.

   If you return a hypothesis g, then with high probability the g which you produce will approximate f well out of sample.

   Solution: I think the best we can promise is c. We cannot promise (a) according to Hoeffding inequality.b matches the principle of Hoeffding inequality. However, we cannot promise the learning will always result in a correct g, thus we can only promise c

6. **[200 points] Problem 1.3** Prove that the PLA eventually converges to a linear separator for separable data. The following steps will guide you through the proof. Let w^* be an optimal set of weights (one which separates the data). The essential idea in this proof is to show that the PLA weights w(t) get "more aligned" with w^* with every iteration. For simplicity, assume that w(0)=0.

   (a) Let $p = min_{1 \leq n \leq N} y_n(w^{*T} x_n)$ Show that $p > 0$.
   Since the data is linearly separable, and w^* correctly separate all the points, for each point n in data set, $y_n(w^{*T} x^n) > 0$. Therefore,$min_{1 \leq n \leq N} y_n(w^{*T} x_n) > 0$, so $p > 0$

   (b) Show that $w^T(t)w^* \geq w^T(t-1)w^* + p$

   From $w(t) = w(t-1) + y(t-1)x(t-1)$ we have LHS
   $(w^T(t-1) + y(t-1)x(t-1)^T)w^* = w^T(t-1)w^* + y(t-1)w^*Tx(t-1)$
   from a we know this is greater than $w^T(t-1)w^* + p$
   So the remain is to prove$w^T(t)w^* \geq tp$
   Proof by induction:

   Let f(t) be $w^T(t)w^* \geq tp$

   Base case:
   f(0) is $0w* \geq 0$ which certainly holds

   Induction step:
   f(t+1) is $w^T(t+1)w^* \geq (t+1)p$
   LHS is $w^T(t)w^* + y(t)w^{*T}x(t)$ RHS is $pt + p$
   Form IH, we know that $w^T(t)w^* \geq tp$
   $y(t)w^{*T}x(t) \geq p$ from a
   thus f(t+1) holds if f(t) is true
   By induction, we prove that f(t) is true

   (c) Show that $||w(t)||^2 \leq ||w(t-1)||^2 + ||x(t-1)||^2$

   $||w(t)||^2$
   $= ||w(t-1) + y(t-1)x(t-1)||^2$
   $= ||w(t-1)||^2 + ||y(t-1)x(t-1)||^2 + 2y(t-1)w(t-1)^T x(t-1)$ since it's misclassified $2y(t-1)w(t-1)^T x(t-1) \leq 0$ so the expression is $\leq$ than
   $||w(t-1)||^2 + ||y(t-1)x(t-1)||^2$
   and it's less than $||w(t-1)||^2 + ||x(t-1)||^2$

   (d) Show by induction that $||w(t)||^2 \leq tR^2$ where R=$max_{1 \leq n \leq N}|X_n|$

   Proof by induction:
   let f(t) be $||w(t)||^2 \leq tR^2$
   Base case: f(0) $0 \leq 0$ is true
   Induction step:
   f(t+1) is $||w(t+1)||^2 \leq t + 1R^2$
   LHS $||w(t+1)||^2 \leq ||w(t)||^2 + ||x(t)||^2$ by previous proof
   and it's $\leq tR^2 + ||x(t)||^2$ by IH
   RHS is $tR^2 + R^2$
   Since R is the maximum $||x(t)||^2 \leq ||R^2||$ there fore LHS $\leq$ RHS. Therefore $f(t) \rightarrow f(t-1)$
   By induction the original statement f(t) is true

(e) Using (b) and (d) Show that

$\frac{w^T(t)}{||w(t)||}w^* \geq \sqrt{t}\frac{p}{R}$

and hence prove that

$t \leq \frac{R^2||w^*||^2}{p^2}$

Solution:
By using b and d , we get $\frac{w^T(t)w^*}{||w(t)||} \geq \frac{tp}{\sqrt{t}R}$
$t \leq \frac{(w^T(t)w^*)^2R^2}{||w(t)||^2p^2}$

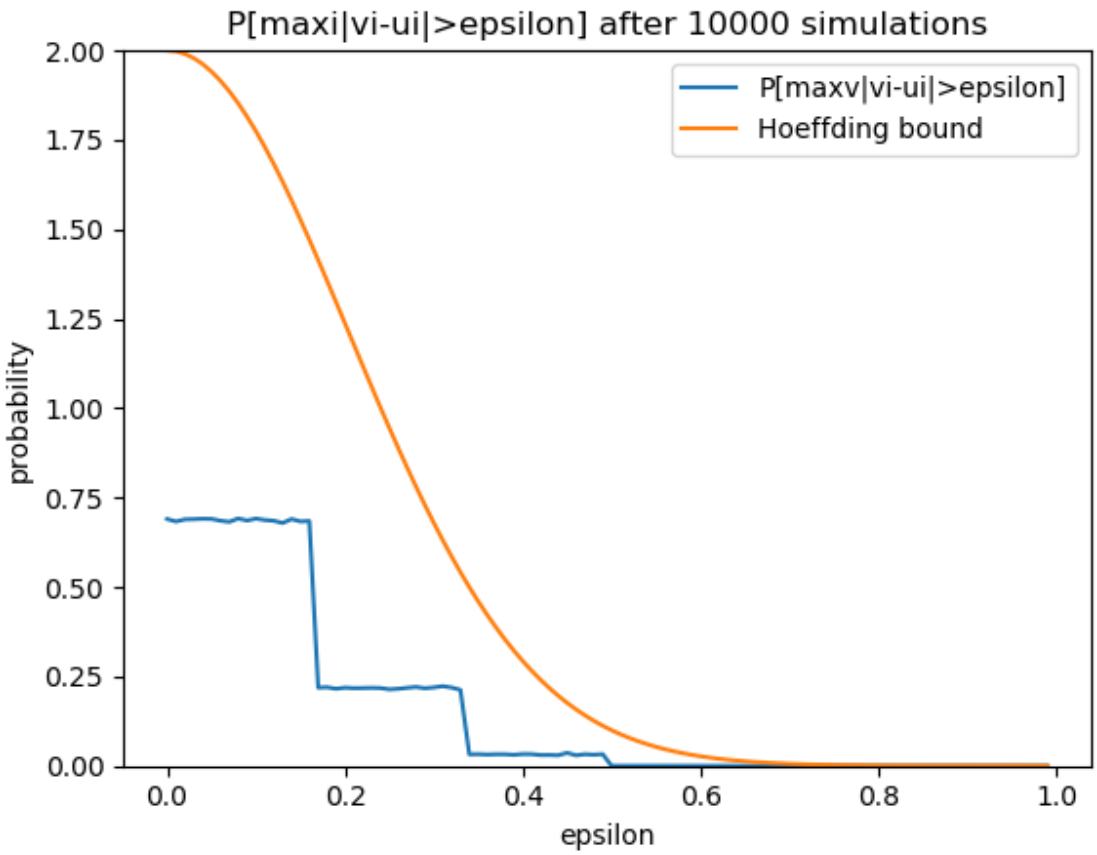
RHS= $\frac{(w^T(t)w^*)^2}{||w(t)||^2||w*||^2}\frac{R^2}{p^2}||w*||^2$
from hint the left part is smaller than one
therefore,
$t \leq \frac{R^2}{p^2}||w*||^2$

7. **[200 points] Problem 1.7** A sample of heads and tails is created by tossing a coin a number of times independently. Assume we have a number of coins that generate different samples independently. For a given coin, let the probability of heads (probability of error) be μ. The probability of obtaining k heads in N tosses of this coin is given by the binomial distribution: $P[k|N,\mu] = NCk\mu^k(1-\mu)^{n-K}$ Remember that training error v is $\frac{k}{N}$

(a) Assume the sample size N is 10. If all the coins have μ=0.05 compute the probability that at least one coin will have $v = 0$ for the case of1 coin, 1000 coins, 1000 coins and 1,000,000 coins. Repeat for $\mu = 0.8$

for μ=0.05:
P[at least one head in 10]=$(1-0.05)^{10} = 0.599$
P[at least one head in 10 for 1000 coins]=$1-(1-0.599)^{1000}$= approximately 1
P[at least one head in 10 throw for 1000000 coins]=$1-(1-0.599)^{1000000}$= approximately 1
for μ=0.8:
P[at least one head in 10 trows]=$(1-0.8)^{10} = 1.024 * 10^{-7}$
P[at least one head in 10 throw for 1000 coins]=$1-(1-1.024*10^{-7})^{1000}$= approximately 1
P[at least one head in for 1000000 coins]= approximately 1

(b) For the case N=6 and 2 coins with μ= 0.5 for both coins, plot the probability $P[max_i|vi-\mu i| \geq \epsilon]$ for ϵ in the range [0.1]. On the same plot show the bound of Hoeffding Inequality


P[maxi|vi-ui|>epsilon] after 10000 simulations
P[maxv|vi-ui|>epsilon]
Hoeffding bound
probability
epsilon
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0