LFD Problem Set 2

John Cohen

September 16, 2024

Exercise 1.8

(If $\mu = 0.9$, what is the probability that a sample of 10 marbles will have $\nu \leq 0.1$? [*Hints: 1. Use binomial distribution. 2. The answer is a very small number.*])

$$P(X = k) = \binom{n}{k} \mu^k (1-\mu)^{n-k}$$

$$X \leq 1$$

$$P(X \leq 1) = P(X = 0) + P(X = 1)$$

$$P(0) = \binom{10}{0} (0.9)^0 (0.1)^1 0 = 0.1 * 10^{-9}$$

$$P(1) = \binom{10}{1} (0.9)^1 (0.1)^9 = 9 * 10^{-9}$$

$$P(X \leq 1) = 0.1 * 10^{-9} + 9 * 10^{-9} = 9.1 * 10^{-9}$$

Exercise 1.9

(If $\mu = 0.9$, use the Hoeffding Inequality to bound the probability that a same of 10 marbles will have $\nu \leq 0.1$ and compare the answer to the previous exercise.)

if $\nu \leq 0.1$ and $\mu = 0.9$, than it deviates from μ by $\epsilon = 0.8$

$$P(v \leq 0.1) \leq P(|\nu - 0.9| \geq 0.8) \leq 2e^{-2(0.8)^2 10}$$

$$P(v \leq 0.1) \leq P(|\nu - 0.9| \geq 0.8) \leq 2e^{-12.8}$$
$$P(v \leq 0.1) \leq P(|\nu - 0.9| \geq 0.8) \leq 5.5 * 10^{-6}$$

Hoeffding bound is an upper bound of $5.5 * 10^{-6}$. While small, this is still much larger than the exact probability given by the binomial distribution. This makes sense as $P(v \leq 0.1) \leq P(|\nu - 0.9| \geq 0.8)$ still holds as $9.1 * 10^{-9} \leq 5.5 * 10^{-6}$. The bound ensures that the probability is below this value, but the exact probability can be much smaller, as in this case.

Exercise 1.10

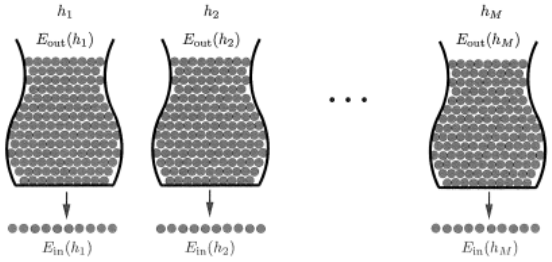


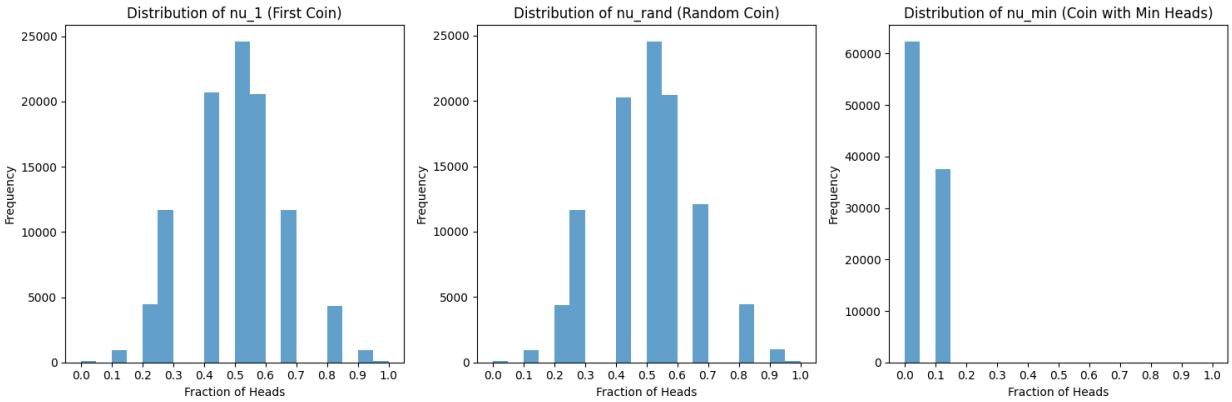
Figure 1.10: Multiple bins depict the learning problem with M hypotheses

(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 (pic the earlier one in case of a tie). Let ν_1, ν_{rand} and ν_{min} be the fraction of heads you obtain for the respective three coins.

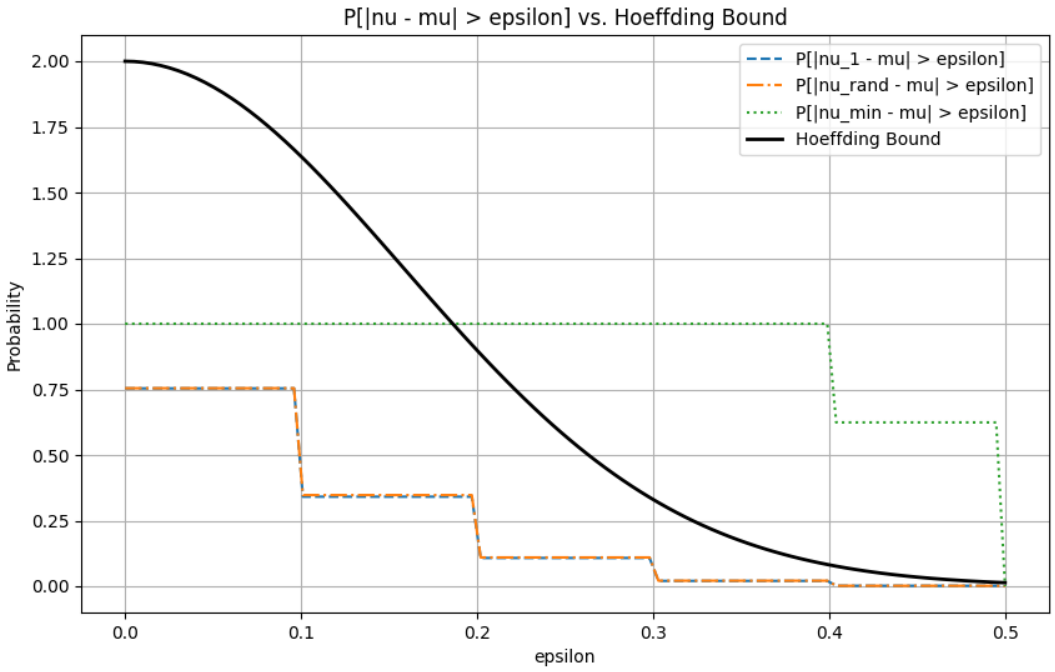
(a) What is μ for the three coins selected?
A fair coin has a probability of flipping heads half the time ($\mu = 0.5$). The coins are fair and the flips are independent so the expected fraction of heads for all three coins (c_1,c_{rand},c_{min}) is:

$$\mu = 0.5$$

(b) Repeat this entire experiment a large number of times (e.g. 100,000 runs of the entire experiment) to get several instances of ν_1, ν_{rand} and ν_{min} and plot the histograms of the distributions of ν_1, ν_{rand} and ν_{min}. Notice that which coins end up being c_{rand} and c_{min} may differ from one run to another.



(c) Using (b), plot estimates for $P[|\nu - \mu| > \epsilon]$ as a function of ϵ, together with the Hoeffding bound $2e^{-2\epsilon^2 N}$ (on the same graph).



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

(i) c_1 and c_{rand} should obey the Hoeffding bound. They are fair coins chosen independently during the flipping process. Therefore, the distribution of their fraction of heads (ν_1 and ν_{rand}) follows typical behavior of a fair coin. The Hoeffding bound gives a good estimate of the probability that their ν values deviate from the expected value $\mu = 0.5$.

(ii) c_{min}, however, does NOT obey the Hoeffding bound. This coin is selected for having the minimum number of heads across 1000 coins. This is a bias. This coin is selected for having the most extreme outcome. Thus, the probability of it deviating from $\mu = 0.5$ is consequentially much larger than the Hoeffding bound would predict for a regular, independent coin. This shows why independent and identically distributed (i.i.d.) random variables is so important as we learned in class. Once we change the setting so that one coin is chosen based on its outcome, the assumption of i.i.d. is violated and thus the problems, approach, and what we are able to conclude changes as well, including the validity of the Hoeffding bound. Hoeffding's bound is designed for single hypothesis testing (or a single bin). What we see here is the "multiple bins" effect we discussed in class which we know the standard Hoeffding bound ($\leq 2e^{-2\epsilon^2 N}$) does not account for. When we test many hypotheses or look across many bins (like selecting c_{min} from 1,000 coins), the chances of finding an extreme deviation by chance increase. The Hoeffding bound doesn't account for this selection, so it underestimates the probability of large deviations for c_{min}. This further explains why the deviation in the experiment shown in Part (c) for c_{min} exceeds the "expected" bound. This bound should not be the "expected" bound for c_{min} as c_{min} by definition does not meet the requirements for the Hoeffding bound to be applicable to it. In conclusion: Since c_{min} does not satisfy the i.i.d. assumption, the Hoeffding bound is NOT applicable to it. This is an important point because applying the bound to non-i.i.d. data gives misleading results. The importance of i.i.d. and the failure of the Hoeffding bound in cases involving selection bias (like c_{min}) are key concepts in understanding why c_{min} deviates more than expected.

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

c_{min} represents the outcome of selecting from multiple bins (specifically, the coin with the fewest heads among many coins). Since the Hoeffding bound applies only to single bins, it does not hold in this case. By selecting for an extreme outcome, the probability of deviation from the expected value $\mu = 0.5$ becomes much larger. This situation relates to the "multiple bins" effect depicted in Figure 1.10, where multiple hypotheses or bins are considered. In such cases, Hoeffding's Inequality breaks down because with many bins, the likelihood of finding an extreme outcome by chance increases significantly.
c_1 and c_{rand} represent single bins so the inequality holds for them.

Exercise 1.11

(We are given a data set D of 25 training examples form an unknown target function $f : X \implies Y$, where $X = R$ and $Y = \{-1, +1\}$. To learn f, we use a simple hypothesis set $H = \{h_1, h_2\}$ where h_1 is the constant +1 function and h_2 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 perform better than random on any point outside D?
No, Algorithm S cannot produce a hypothesis that is guaranteed to perform better than random on any point outside the training data D. Without additional information about the target function f, we cannot ensure that the hypothesis chosen by S will outperform random guessing on unseen data. Since S selects the constant hypothesis that best fits D, it might still misclassify new points if the distribution of f changes outside D.

(b) Assume for the rest of the exercise that all the examples in D have $y_n = +1$. Is is *possible* that the hypothesis that C produces turns out to be better than the hypothesis that S produces?
Yes, it is possible that the hypothesis produced by Algorithm C turns out to be better than the one produced by Algorithm S. Given that all examples in D have $y_n = +1$, S will choose h_1 the constant +1 function), and C will choose h_2 (the constant -1 function). If the true target function f outputs -1 for most points outside D, then h_2 will perform better than h_1 on new data, despite h_2 disagreeing with all training examples.

(c) If p = 0.9, what is the probability that S will produce a better hypothesis than C?
The probability that S will produce a better hypothesis than C is **0.9** (or **90%**) for any new point. Given p=0, the probability that f(x)=+1 is 0.9. Since all training examples have $y_n = +1$, S chooses h_1, which predicts +1 everywhere. The out-of-sample error for h_1 is $E_{out}(h_1) = P[f(x) \neq +1] = 0.1$, while for h_2 its $E_{out}(h_2) = P[f(x) \neq -1] = 0.9$. Therefore, h_1 will almost always outperform h_2, making the probability that S's hypothesis is better equal 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, there are values of p for which it is more likely than not that C will produce a better hypothesis than S. Specifically, when $p < 0.5$, the probability that $f(x) = +1$ is less than the probability that $f(x) = -1$. In this case h_2 (chosen by C) has a lower out-of-sample error than h_1 (chosen by S). Therefore when $p < 0.5$, it is more likely that C's hypothesis will perform better than S's on new data.

Exercise 1.12

(A friend comes to you with a learning problem. She says the target function f is completely unknown, but she has 4,000 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.

(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 two things will happen.
(i) You will produce a hypothesis g;
(ii) You will declare that you failed.
If you do return a hypothesis g, then with high probability the g which you produce will approximate f well out of sample.

Given that the target function f is completely unknown and we have finite dataset of 4,000 points, the best promise we can make is **option (c)**. (a) is too strong and (b) is problematic because it assumes the data is sufficiently representative and that f falls within the capacity of the learning model. (c) is best because it allows for the possibility of failure if the data isn't sufficient or if f is too complex to be approximated with the given resources. (c) also correctly states that if we do produce a hypothesis g, we can ensure with high probability under the learning model's assumptions that g will approximate g well out of sample.

Problem 1.3

(Prove that 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 w^* is by definition an optimal set of weights, it thus must separate the data perfectly. For all n:

$$y_n(w^{*T} x_n) > 0$$

Defining p as $p = min_{1 \leq n \leq N} y_n(w^{*T} x_n)$ and since $y_n(w^{*T} x_n)$ is strictly positive for all n, the minimum value p must be at least greater than 0, $p > 0$.

(b) Show that $w^T(t)w^* \geq w^T(t-1)w^* + p$, and conclude that $w^T(t)w^* \geq tp$. [*Hint: Use induction.*]
At iteration t, suppose the algorithm makes a mistake on the example $(x_n, y_n$. The update rule is:

$$w(t) = w(t-1) + y_n x_n.$$

Expanding the dot product with w^*:

$$w^T(t)w^* = (w(t-1) + y_n x_n)^T w^* = w^T(t-1)w^* + y_n x_n^T w^*$$

Since we know $y_n(w^{*T} x_n) \geq p$,

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

With induction:

Base case (t=0):

w(0)=0, so:

$$w^T(0)w^* = 0$$

Inductive step:

Assumption: Suppose $w^T(t-1)w^* \geq (t-1)p$.

From part (a):

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

Substitute the inductive assumption:

$$w^T(t)w^* \geq (t-1)p + p = tp$$

Conclusion:

$$w^T(t)w^* \geq tp$$

(c) Show that $||w(t)||^2 \leq ||w(t-1)||^2 + ||x(t-1)||^2$.
The weight update rule is $w(t) = w(t-1) + y_n x_n$. Taking the squared norm of both sides:

$$||w(t)||^2 = ||w(t-1) + y_n x_n||^2 = ||w(t-1)||^2 + 2y_n w^T(t-1)x_n + ||y_n x_n||^2.$$

Now since $y_n^2 = 1$:

$$||w(t)||^2 = ||w(t-1)||^2 + ||x_n||^2 + 2y_n w^T(t-1)x_n.$$

The algorithm makes a mistake, $y_n w^T(t-1)x_n < 0$, so the term $2y_n w^T(t-1)x_n$ is negative or zero. Thus:

$$||w(t)||^2 \leq ||w(t-1)||^2 + ||x_n||^2$$

(d) Show by induction that $||w(t)||^2 \leq tR^2$, where $R = max_{1 \leq n \leq N} ||x_n||$.
From part (c):

$$||w(t)||^2 \leq ||w(t-1)||^2 + ||x_n||^2$$

Taking $R = max_{1 \leq n \leq N} ||x_n||$, bound $||x_n||^2 \leq R^2$. So:

$$||w(t)||^2 \leq ||w(t-1)||^2 + R^2$$

By induction, after t updates:

$$||w(t)||^2 \leq tR^2.$$

(e) Using (b) and (d), show that

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

and hence prove that

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

[*Hint:* $\frac{w^T(t)w^*}{||w(t)||||w^*||} \leq 1$. *Why?*]
Part (b) shows us:

$$w^T(t)w^* \geq tp$$

and part (d) proved:

$$||w(t)|| \leq \sqrt{t}R.$$

Thus:

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

We know $\frac{w^T(t)w^*}{||w(t)||||w^*|| \leq 1}$ since it's the cosine of the angle between the vectors, which is at most 1). This gives us:

$$\frac{\sqrt{t}\frac{p}{R}}{||w^*||} \leq 1,$$

implying:

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

Squaring it gives:

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

Thus the PLA converges in t iterations bounded by $t \leq \frac{R^2||w^*||^2}{p^2}$ iterations. This is just an upper bound however.
In practice, PLA converges more quickly than the bound $\frac{R^2||w^*||^2}{p^2}$ suggests. Nevertheless, because we do not know p in advance, we can't determine the number of iterations to

convergence, which does pose a problem if the data is non-separable.

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 is N tosses of this coin is given by the binomial distribution:

$$P[k|N,\mu] = \binom{N}{k}\mu^k(1-\mu)^{N-k}.$$

Remember that the training error ν is $\frac{k}{N}$.

(a) Assume the sample size (N) is 10. If all the coins have $\mu = 0.05$ compute the probability that at least one coin will have $\nu = 0$ for the case of 1 coin, 1,000 coins, 1,000,000 coins. Repeat for $\mu = 0.8$.
$\nu = 0 = \frac{k}{N} \implies k = 0$. With this

$$P[k = 0|N = 10,\mu] = \binom{10}{0}\mu^0(1-\mu)^{10-0} = (1-\mu)^{10}.$$

This is the probability that a single coin will have $\nu = 0$ (no heads) after 10 tosses. Now, we want to find the probability that at least one coin out of a given number of coins (1, 1,000, or 1,000,000 coins) will have $\nu = 0$.
If the probability that a single coin has $\nu = 0$ is $P[\nu = 0]$, then the probability that a single coin has $\nu = 0$ (i.e., at least one head appears) is:

$$P[\nu > 0] = 1 - P[\nu = 0]$$

With M independent coins, the probability that all M coins have $\nu > 0$ is:

$$P[all\nu > 0] = (P[\nu > 0])^M = (1 - P[\nu = 0])^M = (1 - (1-\mu)^{10})^M$$

The probability that at least one coin out of M coins will have $\nu = 0$ is the complement of the above probability (the chance that all coins $\nu > 0$ does not happen):

$$P[atleastone\nu = 0] = 1 - P[all\nu > 0] = 1 - (1 - (1-\mu)^{10})^M$$

Calculating for $\mu = 0.05$ at $M = 1, 1000, 1000000$:

$$P[atleastone\nu = 0|M = 1] = 1 - (1 - (1 - 0.05)^{10})^1 = 0.5987$$

$$P[atleastone\nu = 0|M = 1] = 1 - (1 - (1 - 0.05)^{10})^{1000} = 0.99999999999$$

$$P[atleastone\nu = 0|M = 1] = 1 - (1 - (1 - 0.05)^{10})^{1000000} = 0.99999999999$$

Calculating for $\mu = 0.8$ at $M = 1, 1000, 1000000$:

$$P[atleastone\nu = 0|M = 1] = 1 - (1 - (1 - 0.8)^{10})^{1} = 1.024 * 10^{-7}$$

$$P[atleastone\nu = 0|M = 1] = 1 - (1 - (1 - 0.8)^{10})^{1000} = 1.02 * 10^{-4}$$

$$P[atleastone\nu = 0|M = 1] = 1 - (1 - (1 - 0.8)^{10})^{1000000} = 0.09733$$

These results show that for $\mu = 0.05$, the probability that at least one coin will have no heads becomes very high with a large number of coins. For $\mu = 0.8$, the probability remains small even for many coins. However, as the number of coins increases, the likelihood of observing an improbable event (such as no heads or all heads) increases. This is because even if each individual coin has a low probability of such an extreme outcome, with enough trials, the chance of at least one coin exhibiting this improbable event grows significantly.

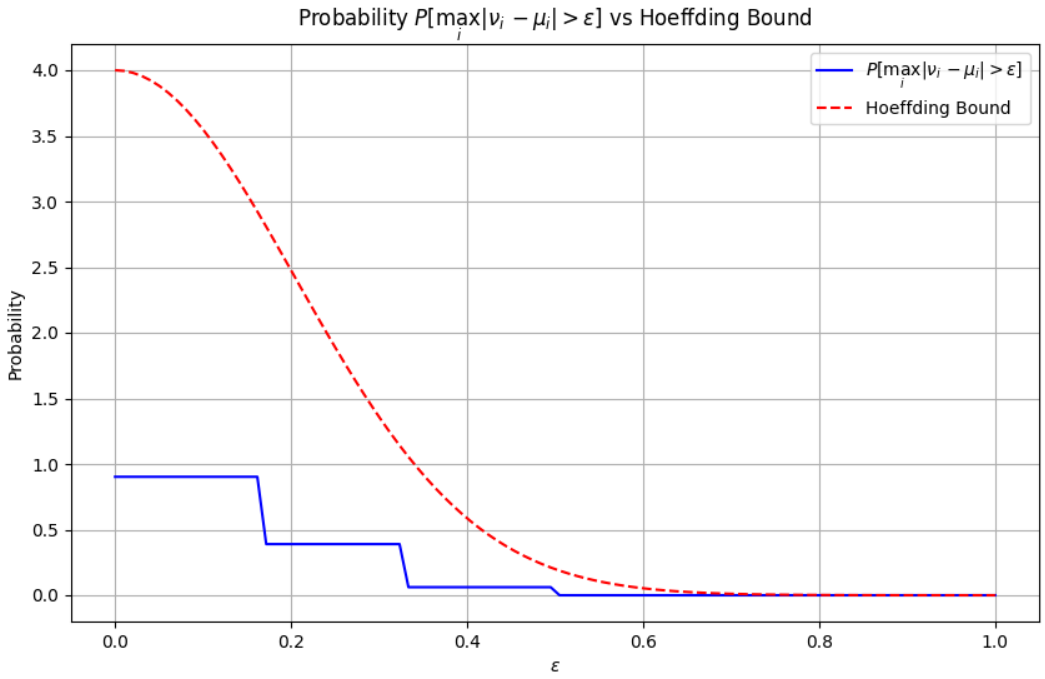
(b) For the case N = 6 and 2 coins with $\mu = 0.5$ for both coins, plot the probability

$$P[max_i|\nu_i - \mu_i| > \epsilon]$$

for ϵ in range [0,1] (the max is over coins). On the same plot show the bound that would be obtained using the Hoeffding Inequality. Remember that for a single coin, the Hoeffding bound is

$$P[|\nu - \mu| > \epsilon] <= 2e^{-2Ne^2}.$$

[*Hint: Use* P[A or B] = P[A] + P[B] - P[A and B] = P[A] + P[B] - P[A]P[B], *where the last equality follows by independence, to evaluate* P[max...]]



The formula $2P[|\nu - \mu| > \epsilon] - (P[|\nu - \mu| > \epsilon])^2$ is used, taking into account that the coins are independent.