CS 70 Discrete Mathematics and Probability Theory
Spring 2015 Vazirani HW 11

Reminder: Deadline for HW10 self-grade is Thursday, April 9 at noon.

Reading: Note 15, 16.

Due Monday April 13

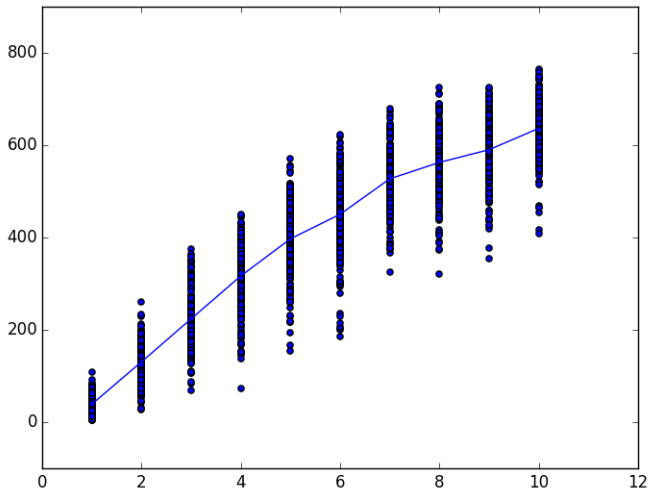
1. **Virtual Lab**

   We have provided a tool for you to explore hashing and load balancing (balls and bins). Follow the instructions on the tool, and answer the questions in your homework solutions. Include screenshots of the virtual lab.

   Answer:
   Note: You were not required to have plots for this question. But for better depiction of the concepts we have provided various plots throughout the solution.

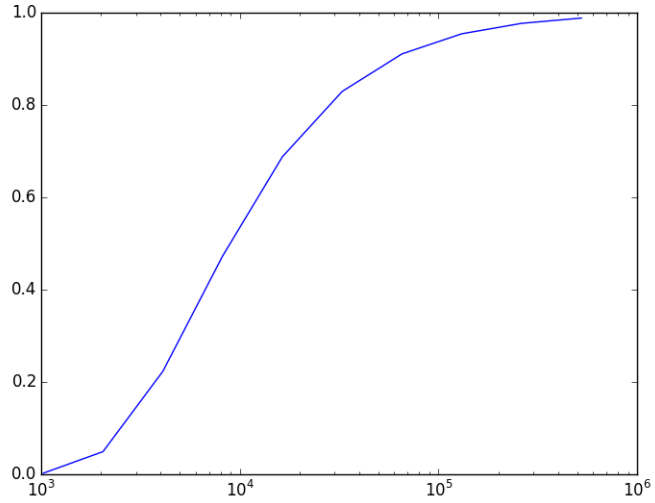
   Hashing

   1. Here you can see a scatter plot of this experiment (100 trials for each choice). On the x axis, you see the number of random bin choices, and on the y axis, you see the result (i.e. the number of balls thrown before we stop). For clarity a trend is also depicted that goes through the mean value for each x value. It is clear that going from 1 to 2 more than doubles the number of balls, while going from 2 to 3, 3 to 4 and so on does not have that effect (they all less than double the number of balls). So the biggest change happens when we go from 1 to 2.



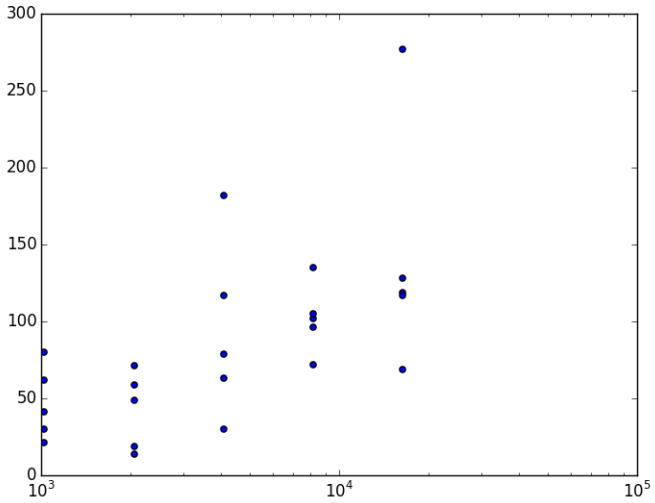
   2. One can actually compute the probability that 5 trials succeed with the help of a compute program (you were not required to do this). Below you can find these probabilities. The x axis is scaled

logarithmically and represents the number of bins and the y axis is the probability that 5 trials succeed.

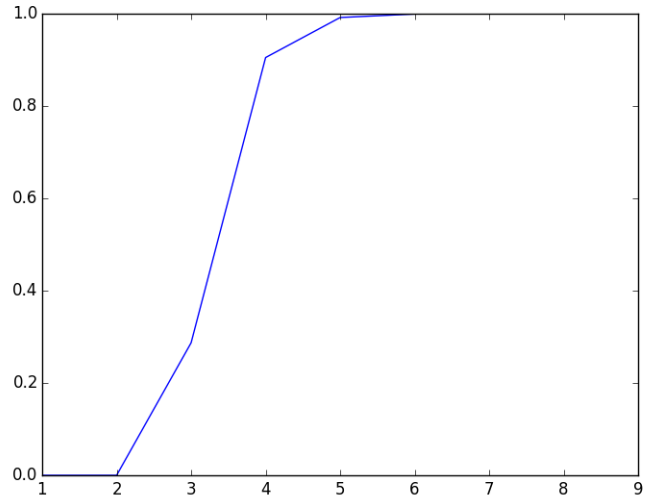


As it can be seen a wide range of values could be an acceptable answer. At about 2000 bins, the probability of success is less than 5% and at about 150000 bins, the probability of success is more than 95%. Although a fair value would probably be closer to 10000 (depending on the strategy you chose to pinpoint the threshold).

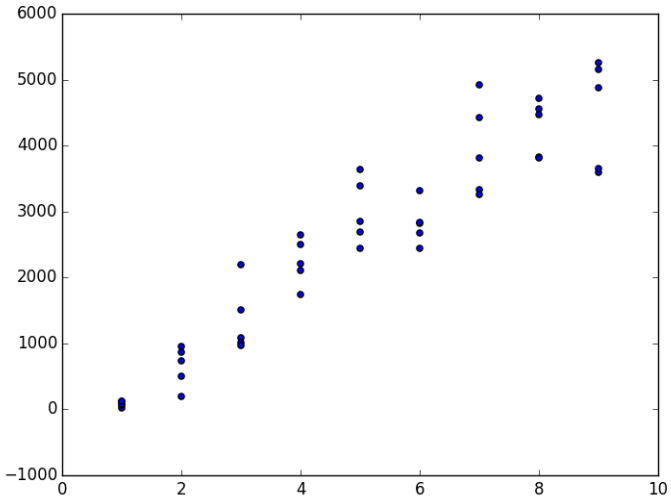
Below you can find scatter plots of 5 trials for some choices for the number of bins. As you can see, in this instance we stopped seeing values below 50 at about 10000 bins.



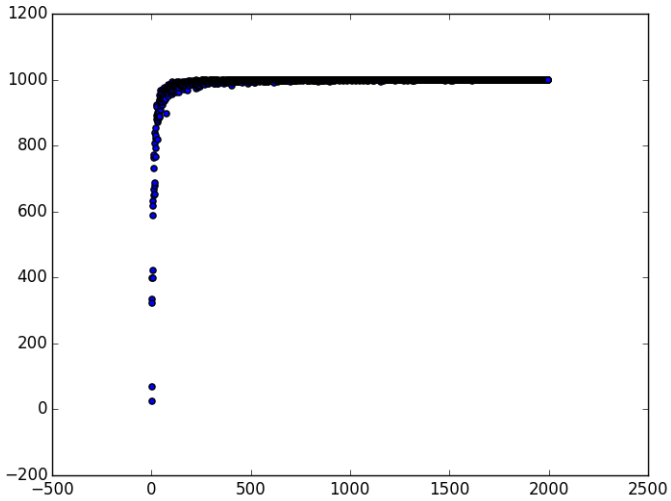
3. Similar to the previous question, we can actually compute the probability of the success of 5 trials with the help of a computer program (you were not required to do this). The x axis shows the number of random bin choices and the y axis shows the probability of the success of 5 trials. Based on this plot, the answers 3, 4, or 5 all seem reasonable.



Here are the scatter plots of the number of balls for these choice (with 5 trials per choice). As you can see in this instance, after 4 we stop seeing numbers below 1000.

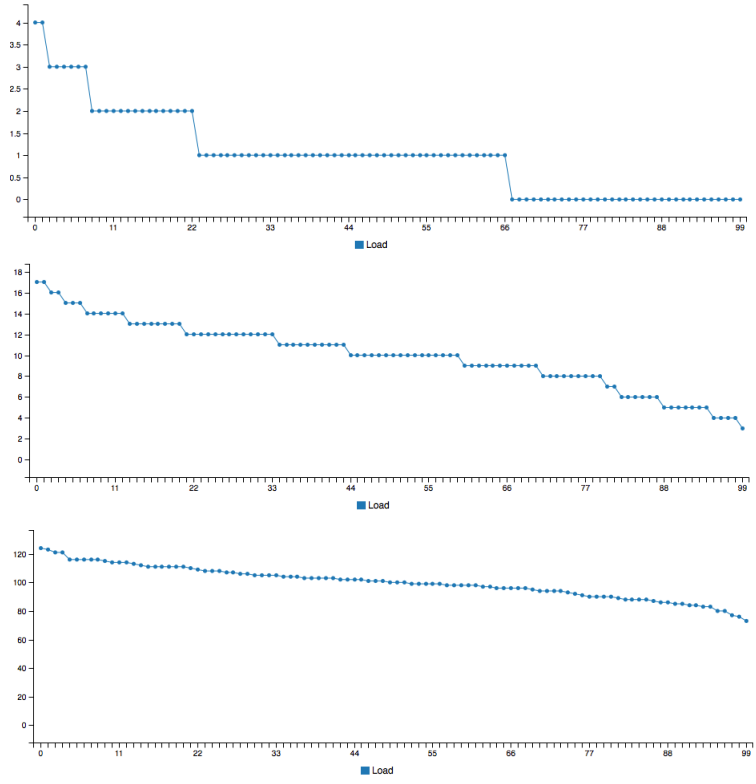


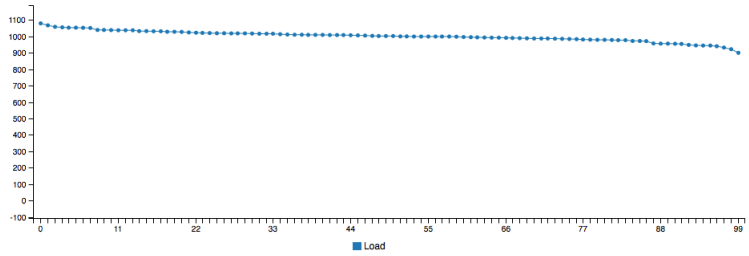
4. When the number of random choices grows larger and larger the probability of choosing all of the bins becomes closer and closer to 1. In the limit, for all balls we look at all bins. Therefore we won't stop until all bins are filled, which means that we throw as many balls as there are bins. So in the limit the result of the experiment becomes equal to the number of bins. Below you can see this happening when the number of bins is fixed to 1000. The x axis is the number of random choices. And the y axis is the result of the experiment.



Load Balancing

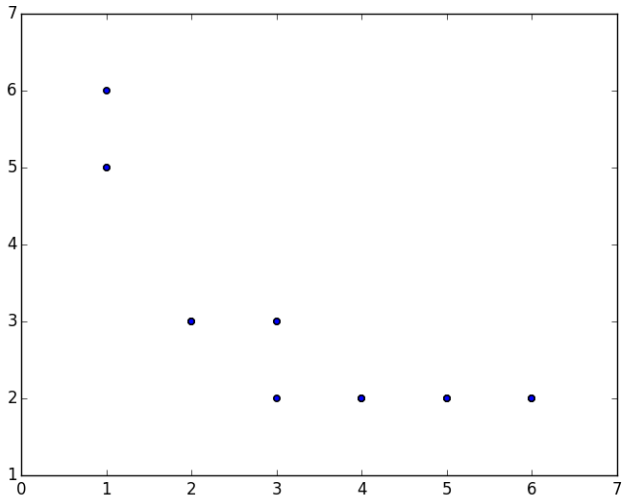
1. Below you can find sample charts for the 100, 1000, 10000, and 100000 balls.



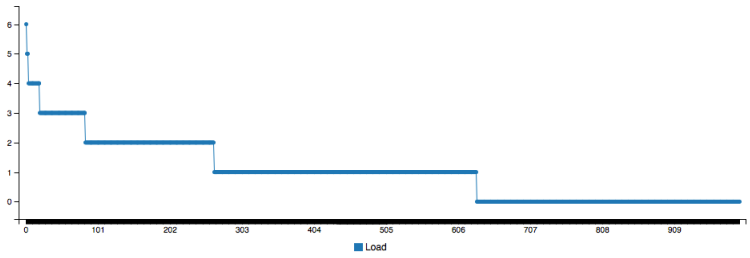


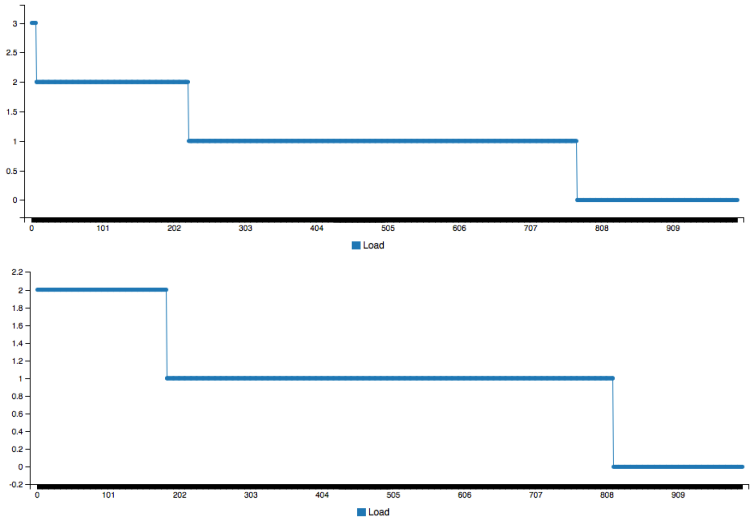
As you can see the graph becomes flatter and flatter. Here is an intuitive explanation: the load in each bin is a random variable X that can be written as a sum of n different random variables $X_1, \ldots, X_n$ where n is the number of balls. Here each X_i is 1 when the i-th ball has been thrown into our particular bin and 0 otherwise. The random variables $X_1, \ldots, X_n$ are independent of each other. Note that $E[X] = nE[X_1] = \frac{n}{m}$ where m is the number of bins. So we expect the value of X to lie around $E[X]$, its expectation. But of course there is a lot of noise which come from $X_1, \ldots, X_n$. But as n grows larger and larger, these noises tend to cancel each other out. So as n grows larger and larger the average of $X_1, \ldots, X_n$ which is simply X/n becomes more and more stable (or in other words becomes closer and closer to $1/m$ which is a constant). This is called the law of large numbers.

2. Here the largest change again should happen when going from 1 to 2. Below you can find scatter plots of the maximum load as a function of the number of random choices (for each choice the experiment was done 5 times).



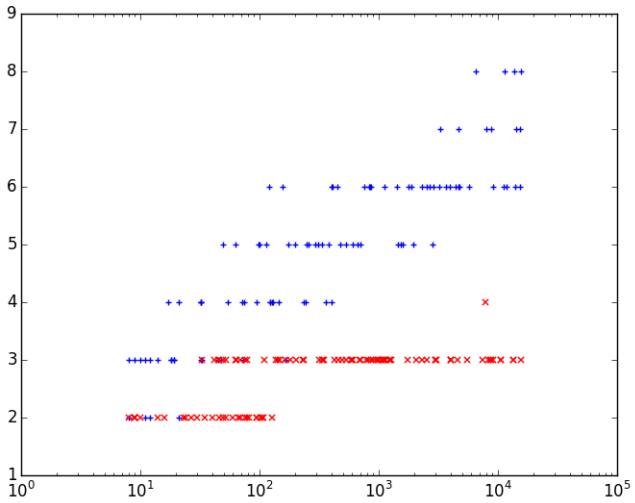
When the number of choices is 1, the typical value for maximum load is 5 or 6. This drops to 3 when the number of choices is 2 and then to 2-3 when the number of choices is 3 and from then on it stays at 2 for a while. Below are pictures of the graph for 1, 2, 3 choices.





The graph becomes flatter as the number of choices grows. The intuitive explanation is that when the number of choices grows, the procedure tries to equate the load in different bins more. It is less likely for the maximum to grow (since each of the random choices has to be one of the bins with the maximum load) and therefore it is more likely for bins with less load to grow, until a lot of bins become the maximum.

3. As long as the maximum load is 1, load balancing is similar to hashing. In other words, the maximum load is 1 if and only if no collision occurs. Therefore this is very similar to question 3 of the hashing part of this exercise. And indeed the answers should look similar. In particular 3, 4, and 5 all seem reasonable answers. This can be verified experimentally as well.
4. Again similar to the previous part, the maximum load being 1 is similar to having no collision in hashing. Therefore the answers here should be similar to the answers in question 2 of the hashing part of this exercise. In particular the answer should most likely be in the range 2000-150000 and somewhat close to 10000.
5. Below you can see a scatter plot containing the results of this experiment. On the x axis you can find the number of balls/bins and on the y axis you can find the maximum load of any bin in one trial. The blue dots represent the trials when the number of random bin choices were 1 and the red crosses represent the trials with that number set to 2. As it is obvious from the graph, the maximum load grows fastest when the number of random choices is 1.



2. **The Power of Choice**

Suppose that I have *two* hash functions, H_1 and H_2. Each of these hash functions sends each key to a uniformly random location in a table of size n.

To try to make my hash table smaller, I consider the following optimization. When I want to put a key x in the table:

- First I try to put it in the location $H_1(x)$.
- If there is already an item in $H_1(x)$, I try to put x in $H_2(x)$.
- If there is an item in both $H_1(x)$ and $H_2(x)$, then I give up and throw away my hash table.

I'd like to know how many keys I can add before I have to throw away my hash table.

(a) Show that if the hash table already contains k elements, the probability that I'm unable to put x in the hash table is at most $\frac{k^2}{n^2}$.

(b) Show that if I add k elements to an initially empty hash table, one at a time, the probability that I have to throw away the hash table at any point is at most $\frac{k^3}{3n^2}$. This means I can put about $n^{2/3}$ items in my hash table before I have to throw it away.
Recall that $\sum_{i=1}^{k} i^2 = \frac{k(k+1)(2k+1)}{6}$.

Answer:

(a) I am unable to put x in the hash table if and only if both $H_1(x)$ and $H_2(x)$ are already occupied. Since $H_1(x)$ and $H_2(x)$ are uniformly random and k of the n locations are occupied, $\Pr[H_1(x) \text{ is occupied}] = \Pr[H_2(x) \text{ is occupied}] = \frac{k}{n}$. Hence, the probability that both locations are occupied is $\frac{k^2}{n^2}$.

(b) Let A_i denote the event that I throw away the hash table when I am placing the i-th element. Then the probability that I throw away the hash table at any point is $\Pr[A_1 \cup A_2 \cup \cdots \cup A_k]$. Using the union bound,

$$\Pr[A_1 \cup A_2 \cup \cdots \cup A_k] \leq \sum_{i=1}^{k} \Pr[A_i].$$

Also,

$$\begin{aligned}\Pr[A_i] &= \Pr[\text{I do not fail in the first } i-1 \text{ rounds} \\ &\qquad \cap \text{ both } H_1(x) \text{ and } H_2(x) \text{ are occupied in the } i\text{-th round}] \\ &\leq \Pr[\text{both } H_1(x) \text{ and } H_2(x) \text{ are occupied in the } i\text{-th round}] \\ &\leq \frac{(i-1)^2}{n^2}.\end{aligned}$$

Hence,

$$\Pr[A_1 \cup A_2 \cup \cdots \cup A_k] \leq \sum_{i=1}^{k} \Pr[A_i] = \sum_{i=1}^{k} \frac{(i-1)^2}{n^2} = \sum_{i=0}^{k-1} \frac{i^2}{n^2} = \frac{(k-1)k(2k-1)}{6n^2} \leq \frac{2k^3}{6n^2} = \frac{k^3}{3n^2}$$

3. **Balls and bins, again**

Suppose that I throw n balls into n bins.

(a) What is the probability that I throw the i^{th} ball into the same bin as the j^{th} ball?

(b) Suppose that every time I throw a ball into a bin, I lose \$1 for every ball that was already in the bin. How much money should I expect to lose?
Hint: Note that when you throw ball i in the bin, you can only pay because of the presence of balls numbered $j < i$. Define an indicator random variable $X_{i,j}$ for all $j < i$ and express the total money you lose in terms of the $X_{i,j}$.

Answer:

(a) Assume without loss of generality that $j < i$ (the case where $i < j$ is similar). Once the j^{th} ball has been thrown into some bin, the i^{th} ball lands in that bin with probability $\frac{1}{n}$.

(b) Let $X_{i,j} = 1$ if balls i, j land in the same bin, and 0 otherwise. Observe that the amount of money that I lose by throwing ball i is $\sum_{j<i} X_{i,j}$. It follows that the total money I lose is $X = \sum_{i=1}^{n} \sum_{j=1}^{i-1} X_{i,j}$. By linearity of expectation,

$$\mathbf{E}[X] = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathbf{E}[X_{i,j}] = \binom{n}{2} \frac{1}{n} = \frac{n-1}{2}$$

4. **Runs**

Suppose I have a biased coin which comes up heads with probability p, and I flip it n times. A "run" is a sequence of coin flips all of the same type, which is not contained in any longer sequence of coin flips all of the same type. For example, the sequence "HHHTHH" has three runs: "HHH," "T," and "HH."

Compute the expected number of runs in a sequence of n flips.

Answer: Let X_i be the indicator variable for the event that position i is the beginning of a run. Then the total number of runs is simply $X = \sum_{i=1}^{n} X_i$. Now, obviously X_1 is always 1, whereas for $i \geq 2$, $X_i = 1$ if and only if the outcome of the i-th flip is different from that of the $(i-1)$-th flip. This happens with probability $p(1-p) + (1-p)p = 2p(1-p)$. By linearity of expectation,

$$\mathbf{E}[X] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \mathbf{E}[X_3] + \cdots + \mathbf{E}[X_n] = 1 + (n-1) \cdot 2p(1-p) = 1 + 2(n-1)p(1-p).$$

5. **Second Chance**

What is the average number of dots shown by a die tossed once at random? You wish to maximize the value shown by the die. If you are allowed to throw the die a second time, when should you do so? What is the expected value shown by a die for which one is allowed one rethrowing?

Answer: The expected number of dots shown by a die tossed once at random is

$$\frac{1}{6}(1+2+3+4+5+6) = 3.5$$

We should use the rethrowing opportunity when we got less dots than the expected value, i.e., we should use it when we got one, two, or three dots. In fact, using the rethrowing opportunity corresponds to replacing some of the terms in the above formula with 3.5. So if we want to maximize the expected value, we should replace exactly those terms that are smaller than 3.5. To see this, we can also explicitly write down the formula as follows:

$$\frac{1}{6}\left(\overbrace{\frac{1}{6}(1+2+3+4+5+6)}^{\text{rethrow if 1 dot on first throw}} + \overbrace{\frac{1}{6}(1+2+3+4+5+6)}^{\text{rethrow if 2 dots on first throw}} + \overbrace{\frac{1}{6}(1+2+3+4+5+6)}^{\text{rethrow if 3 dots on first throw}} + \overbrace{4+5+6}^{\text{otherwise settle}}\right)$$
$$= \frac{1}{6}(3.5+3.5+3.5+4+5+6) = 4.25$$

6. **James Bond**

James Bond is imprisoned in a cell from which there are three possible ways to escape: an air-conditioning duct, a sewer pipe and the door (which is unlocked). The air-conditioning duct leads him on a one-hour trip whereupon he falls through a trap door onto his head, much to the amusement of his captors. The sewer pipe is similar but takes two hours to traverse. Each fall produces temporary amnesia and he is returned to the cell immediately after each fall. Assume that he always immediately chooses one of the three exits from the cell with probability $\frac{1}{3}$. On average, how long does it take before he realizes that the door is unlocked and escapes?

Hint: If you are doing complicated calculations you're taking the wrong approach.

Answer: Let X be the number of hours that James Bond takes to escape. If he chooses the duct or the pipe, he wastes one or two hours respectively and then he is back to the original situation. Hence,

$$\mathbf{E}[X] = \overbrace{\frac{1}{3}\cdot 0}^{\text{door}} + \overbrace{\frac{1}{3}\cdot(1+\mathbf{E}[X])}^{\text{duct}} + \overbrace{\frac{1}{3}\cdot(2+\mathbf{E}[X])}^{\text{pipe}}.$$

Solving this, we get $\mathbf{E}[X] = 3$ hours.

The solution above is complete, but more rigorously one can let Y be the number of hours spent after the first escape attempt and Z as the number of hours spent during that first escape attempt. Clearly $X = Y + Z$. Note that

$$\mathbf{E}[Z] = \frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 2 = 1.$$

One can also see that

$$\mathbf{E}[Y] = \Pr[\text{door is chosen}]\mathbf{E}[Y|\text{door is chosen}] + \Pr[\text{door is not chosen}]\mathbf{E}[Y|\text{door is not chosen}].$$

But note that $\mathbf{E}[Y|\text{door is chosen}] = 0$ and $\mathbf{E}[Y|\text{door is not chosen}] = \mathbf{E}[X]$. The last equality follows from the fact that if the door is not chosen then Y behaves the same way as X and certainly has the same expectation. Therefore

$$\mathbf{E}[Y] = \Pr[\text{door is not chosen}]\mathbf{E}[X] = \frac{2}{3}\mathbf{E}[X].$$

Now using the linearity of expectation we have

$$\mathbf{E}[X] = \mathbf{E}[Y] + \mathbf{E}[Z] = \frac{2}{3}\mathbf{E}[X] + 1,$$

which again implies that $\mathbf{E}[X] = 3$.

7. **Extra credit**

There are two pouches, one containing exactly twice as much gold as the other. You select one of these pouches at random and find x amount of gold in it. You do a quick calculation and figure that the expected amount of gold in the other pouch is $\frac{1}{2} \cdot \frac{1}{2}x + \frac{1}{2} \cdot 2x = 1.25x$, and wish you had picked the other pouch. But then your friend reminds you that you chose the pouch at random and you would have had the same thought if you had picked the other pouch. You are not convinced, so your friend points out that if you were allowed to repeatedly exchange the pouch, by your reasoning your expected reward would increase by a factor of 1.25 with each exchange, which is absurd. Can you explain the flaw in your reasoning?

Answer: The flaw in the reasoning is with the assumption that $\frac{1}{2} \cdot \frac{1}{2}x + \frac{1}{2} \cdot 2x$ gives the expected gold in the other pouch.

The value x is a random variable, since in every outcome of the world it has a value. The problem does not describe the rules of the probability space enough to uniquely identify the distribution of x, but that does not matter. Assume that y is the amount of gold in the second pouch (again a random variable). The flawed reasoning is trying to compute $\mathbf{E}[y]$ by conditioning on the events $x > y$ and $x < y$. To do that one generally writes the following formula

$$\mathbf{E}[y] = \mathbf{E}[y|x < y]\Pr[x < y] + \mathbf{E}[y|x > y]\Pr[x > y] = \frac{1}{2}\mathbf{E}[y|x < y] + \frac{1}{2}\mathbf{E}[y|x > y].$$

The probabilities $\Pr[x < y]$ and $\Pr[x > y]$ are both $\frac{1}{2}$ because we chose a pouch at random. But then, the flawed reasoning goes on to assume that $\mathbf{E}[y|x < y] = 2E[x]$ and $\mathbf{E}[y|x > y] = \frac{1}{2}\mathbf{E}[x]$. This is the part that's not correct. We know that once $x < y$ it is true that $y = 2x$, so at best that implies $\mathbf{E}[y|x < y] = 2\mathbf{E}[x|x < y]$. But it is not necessarily true that $\mathbf{E}[x|x < y] = \mathbf{E}[x]$. Similarly once we condition on $x > y$, it is true that $y = \frac{1}{2}x$, but at best that implies $\mathbf{E}[y|y < x] = \frac{1}{2}\mathbf{E}[x|y < x]$. But $\mathbf{E}[x|y < x]$ is not the same as $\mathbf{E}[x]$. To see concretely why these values are not the same, let us construct an example: assume that we further restrict the pouches to have 1 or 2 ounces of gold. Then $\Pr[x = 1, y = 2] = \Pr[x = 2, y = 1] = \frac{1}{2}$. In this case $\mathbf{E}[x] = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 2 = 1.5$. But note that $\mathbf{E}[x|x < y] = 1$ and $\mathbf{E}[x|x > y] = 2$, and these are not the same as $\mathbf{E}[x]$.