

## Lecture 12: October 14, 2009

CS 330 Discrete Structures  
Fall Semester, 2009

### 1 Analysis of algorithms

We now return to the problem that originally led us into the discussion of probability:

For example, consider the algorithm to find the maximum of  $n$  numbers  $x_1, x_2, \dots, x_n$  listed below. To analyze this algorithm fully we want to know exactly how many times each of the instructions is executed; but for the moment, let us ask only how many times the statement “ $m \leftarrow k$ ” is executed. The answer, of course, depends not on the particular values of the  $x_i$ , but rather on their relative order: If  $x_1$  is the maximum of the  $n$  elements, the statement will *never* be executed. If  $x_1 < x_2 < \dots < x_n$  then the statement will be executed  $n - 1$  times. For how many of the relative arrangements of the inputs  $x_1, x_2, \dots, x_n$  will the statement be executed exactly  $i$  times? The answer to this question is at the heart of the analysis of the algorithm.

**ALGORITHM 1:**

```

/* Selecting the largest of  $x[1], x[2], \dots, x[n]$ . */
 $m \leftarrow 1$ 
FOR  $k \leftarrow 2$  TO  $n$  DO
    IF  $x_k > x_m$  THEN  $m \leftarrow k$ 
                        “assignment” of interest
    
```

We have shown that the best case of algorithm 1 is zero assignments and in the worst case there are  $n - 1$  assignments. But what about the average case? Let’s examine the case where  $n = 3$ . The following is list of all the possible outcomes of relative  $x_i$ , for  $i = 1, 2, 3$

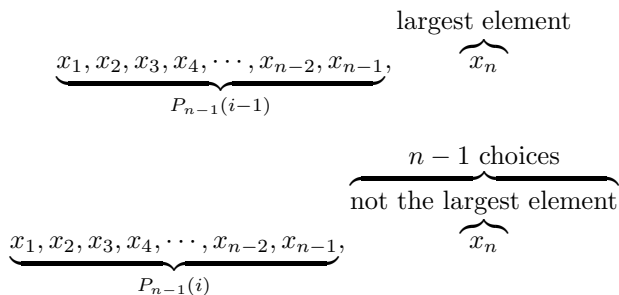
$$\text{average number of assignments} = \left\{ \begin{array}{ll} x_1 < x_2 < x_3 & \text{has 2 assignments} \\ x_1 < x_3 < x_2 & \text{has 1 assignments} \\ x_2 < x_1 < x_3 & \text{has 1 assignments} \\ x_2 < x_3 < x_1 & \text{has 0 assignments} \\ x_3 < x_1 < x_2 & \text{has 1 assignments} \\ x_3 < x_2 < x_1 & \text{has 0 assignments} \end{array} \right\} = \frac{2 + 1 + 1 + 0 + 1 + 0}{6} = \frac{5}{6}$$

It is possible to do this for three, maybe even four, but when we get a list of ten the number of permutations is quite large:  $10! = 3628800$  and a list of ten names is considered very short. We need to generalize this argument for a list of length  $n$ .

We can make some easy observations first: How many items in a list of  $n$  have 0 swaps? There are  $(n - 1)!$ , since there are  $(n - 1)!$  permutations with  $x_1$  as the maximum. How many items in a list of  $n$  have  $n - 1$  swaps? There is 1, since only when the elements are in increasing order does this occur.

We now need to know the general case: How many ways are there to make  $i$  assignments in a list of length  $n$ ? Let's call this number  $P_n(i)$ .

We have the following two possibilities (either with an execution at the last step or without):



But applying the rules of sum and product, we have:

$$P_n(i) = P_{n-1}(i - 1) + (n - 1)P_{n-1}(i)$$

## 2 Expected value

Now we know the number of permutations with  $i$  swaps, but what about the average case? We've see that it's just the sum of the  $i$  divided by the number of permutations, or:

$$\text{average case} = \frac{\sum_{i=0}^{n-1} iP_n(i)}{n!} = \sum_{i=0}^{n-1} i \frac{P_n(i)}{n!} = \sum_{i=0}^{n-1} i \Pr\{i\}$$

This is known as the **expected value** of  $i$ .

By the definition of probability,  $\Pr\{i\} = \frac{P_n(i)}{n!} = p_n(i)$ , where  $p_n(i)$  is the probability of executing  $i$  assignments in a list of length  $n$ .

Since  $P_n(i) = P_{n-1}(i - 1) + (n - 1)P_{n-1}(i)$ :

$$p_n(i) = \frac{P_n(i)}{n!} = \frac{P_{n-1}(i - 1)}{n!} + \frac{(n - 1)P_{n-1}(i)}{n!} = \frac{1}{n}p_{n-1}(i - 1) + \frac{n - 1}{n}p_{n-1}(i), i \geq 1$$

This gives us the recurrence relation between the probabilities. You can see this from the below table showing some values of  $p_n(i)$ :

	$i = 0$	$i = 1$	$i = 2$	$i = 3$	...
$n = 1$	1	0	0	0	...
$n = 2$	1/2	1/2	0	0	...
$n = 3$	1/3	1/2	1/6	0	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Let us consider the **probability generating function** defined as:

$$\mathcal{P}_n(x) = p_n(0)x^0 + p_n(1)x^1 + p_n(2)x^2 + p_n(3)x^3 + \dots = \sum_{i=0}^{\infty} p_n(i)x^i$$

Notice that

$$\mathcal{P}'_n(1) = e_n = 0 \cdot p_n(0) + 1 \cdot p_n(1) + 2 \cdot p_n(2) + \cdots = \sum_{i=0}^{\infty} i \cdot p_n(i)$$

We would like to determine  $\mathcal{P}_n(x)$ . First, note the special case of  $p_n(0) = \frac{1}{n}$ . We have:

$$\begin{aligned} \mathcal{P}_n(x) &= \sum_{i=0}^{\infty} p_n(i)x^i \\ &= p_n(0)x^0 + \sum_{i=1}^{\infty} p_n(i)x^i \\ &= \frac{1}{n}x^0 + \sum_{i=1}^{\infty} p_n(i)x^i \\ &= \frac{1}{n} + \sum_{i=1}^{\infty} \frac{1}{n}p_{n-1}(i-1)x^i + \sum_{i=1}^{\infty} \frac{n-1}{n}p_{n-1}(i)x^i \\ &= \frac{1}{n} + \frac{1}{n} \sum_{i=1}^{\infty} p_{n-1}(i-1)x^i + \frac{n-1}{n} \sum_{i=1}^{\infty} p_{n-1}(i)x^i \\ &= \frac{1}{n} + \frac{x}{n} \sum_{i=1}^{\infty} p_{n-1}(i-1)x^{i-1} + \frac{n-1}{n} \sum_{i=1}^{\infty} p_{n-1}(i)x^i \\ &= \frac{1}{n} + \frac{x}{n} \sum_{i=0}^{\infty} p_{n-1}(i)x^i + \frac{n-1}{n} \sum_{i=1}^{\infty} p_{n-1}(i)x^i \\ &= \frac{1}{n} + \frac{x}{n} \mathcal{P}_{n-1}(x) + \frac{n-1}{n} \sum_{i=1}^{\infty} p_{n-1}(i)x^i \\ &= \frac{1}{n} + \frac{x}{n} \mathcal{P}_{n-1}(x) + \frac{n-1}{n} \left( \sum_{i=0}^{\infty} p_{n-1}(i)x^i - p_{n-1}(0) \right) \\ &= \frac{1}{n} + \frac{x}{n} \mathcal{P}_{n-1}(x) + \frac{n-1}{n} \left( \sum_{i=0}^{\infty} p_{n-1}(i)x^i - \frac{1}{n-1} \right) \\ &= \frac{x}{n} \mathcal{P}_{n-1}(x) + \frac{n-1}{n} \mathcal{P}_{n-1}(x) + \frac{1}{n} - \frac{1}{n} \\ &= \frac{x+n-1}{n} \mathcal{P}_{n-1}(x) \end{aligned}$$

We can now find the derivative of  $\mathcal{P}_n(x)$  with respect to  $x$ :

$$\begin{aligned} \mathcal{P}'_n(x) &= \frac{d}{dx} \left( \frac{x+n-1}{n} \right) \mathcal{P}_{n-1}(x) + \left( \frac{x+n-1}{n} \right) \mathcal{P}'_{n-1}(x) \\ &= \frac{1}{n} \mathcal{P}_{n-1}(x) + \left( \frac{x+n-1}{n} \right) \mathcal{P}'_{n-1}(x) \\ \mathcal{P}'_n(1) &= \frac{1}{n} \mathcal{P}_{n-1}(1) + \mathcal{P}'_{n-1}(1) \end{aligned}$$

By the definition of  $\mathcal{P}(n)$ , we know that  $\mathcal{P}_{n-1}(1) = 1$ . This gives us:

$$\mathcal{P}'_n(1) = \frac{1}{n} + \mathcal{P}'_{n-1}(1)$$

Recall that  $\mathcal{P}'_n(1) = e_n$ . We can substitute as follows:

$$\begin{aligned} e_n &= \frac{1}{n} + e_{n-1} \\ &= \frac{1}{n} + \frac{1}{n-1} + e_{n-1} \\ &\quad \vdots \\ &= \frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \cdots + e_0 \\ &= H_n - 1 \\ &\approx \ln n \end{aligned}$$

When we get such a simple answer after a complicated derivation, we should ask ourselves whether a simpler derivation is possible. The answer is yes (though our more complicated method above will be needed later, when we ask for more sophisticated information). Notice that an important basic principle holds:

The expectation of the sum = the sum of the expectations.

That means that if  $q_k$  is the probability that (in the algorithm) a new maximum element is found at iteration  $k > 1$ , then the its contribution to the expected value is  $1 \times q_k = q_k$ . Of the  $k$  possible relative sizes of  $x_k$  among the first  $k - 1$  values, all equally likely, the assignment statement is made only for one relative size:  $x_k$  must be larger than the first  $k - 1$  values. Thus  $q_k = 1/k$  and the expected number of executions of the assignments statement is  $\sum_{1 < k \leq n} 1/n = H_n - 1$ .

### 3 Two Similar Problems

With that analysis under our belts, we can look at two similar problems.

First, how often are records set? That is, suppose we look at expected rainfall and ask how often does the rainfall in a given year exceed the rainfall in all previous years (thus setting a new record)? To address this problem we make the (not entirely convincing) assumption that amount rainfall in a year is independent of rainfall in prior years. The problem is then identical to our algorithm analysis above because we can view each time we see a new largest element as setting a record; the only difference is that the first year's rainfall is (by definition) a new record. Thus the expected number of record-breaking years in a sequence of  $n$  years is  $H_n \approx \ln n$ . For example, in New York City's Central Park there were 6 record rainfall years in the 160-year period 1835–1994;  $H_{160} \approx 5.7$ .

Our second problem is a bit more complex. Cracker Jack, the carmel-covered popcorn treat, has a small toy as a prize in each box; if there  $n$  different prizes, how many boxes would one expect to buy to get a complete set of prizes? Because the expectation of the sum is the sum of the expectations, the answer is the expected number of boxes to get the first prize, plus the expected number of boxes to get the second prize (that is, a prize that differs from the first prize we got), plus the expected number of boxes to get the third prize (that is, a prize that differs from the first two prizes we got), etc.

Let  $E_i$  be the expected number of boxes we must buy to get  $i$ th new prize. Clearly,  $E_1 = 1$  because the first prize we get is not a duplicate of a previously obtained prize. Using the rules of sum and product, let's examine what happens as we buy more boxes hoping to get a second toy. Getting a new toy on the first box we buy has probability  $(n-1)/n$  because any of  $n-1$  prizes would be okay; if that fails (with probability  $1/n$  we get the toy we already had), the probability of getting a new toy on the next box we buy is  $(n-1)/n$  for a total probability of their product  $(n-1)/n^2$ . In general, to get a new toy on the  $k$ th box means we failed  $k-1$  times (that is, we got a duplicate of the first toy  $k-1$  times) and succeeded on the  $k$ th try; that probability is  $(n-1)/n^k$ . The expected value is thus

$$E_2 = \sum_{k=1}^{\infty} k \times \frac{n-1}{n^k} = \frac{n-1}{n} \sum_{k=1}^{\infty} \frac{k}{n^{k-1}} = \frac{n}{n-1}.$$

The last summation comes from substituting  $x = 1/n$  in the derivative of  $(1-x)^{-1}$ ,

$$1 + 2x + 3x^2 + 4x^3 + \dots = \frac{1}{(1-x)^2}.$$

What about getting a third toy differing from our first two toys? The analysis is similar: To get such a toy in the  $k$ th box we buy we must get a duplicate of one of our two toys  $k-1$  times [probability  $(2/n)^{k-1}$ ], then succeed in getting a new toy in the  $k$ th box [probability  $(n-2)/n$ ]. Thus

$$E_3 = \sum_{k=1}^{\infty} k \times \left(\frac{2}{n}\right)^{k-1} \frac{n-2}{n} = \frac{n-2}{n} \sum_{k=1}^{\infty} k \times \left(\frac{2}{n}\right)^{k-1} = \frac{n}{n-2}$$

[substituting  $x = 2/n$  in the derivative of  $(1-x)^{-1}$ ].

Doing this calculation for the  $t$ th new toy, we find

$$E_t = \frac{n}{n-t+1}$$

and hence the expected number of boxes we must purchase to get  $n$  different toys is

$$\sum_{t=1}^n E_t = \sum_{t=1}^n \frac{n}{n-t+1} = n \sum_{t=1}^n \frac{1}{n-t+1} = n \left( \frac{1}{n} + \frac{1}{n-1} + \dots + \frac{1}{1} \right) = nH_n.$$

The above analysis tells, for example, that we have to roll a normal 6-sided die an average of  $6H_6 = 6 \times 49/20 = 14.7$  times to see all 6 sides come up.