IE531: Algorithms for Data Analytics
Spring, 2023
Homework 1: Recursion
Due Date: February 3, 2023

Instructions

1. You will submit a PDF-version of your answers on Compass on-or-before midnight of the due date.

Instructions

1. (*10 points*) (**Recursive GCD Computation**) The *greatest common divisor* (GCD) of two positive numbers m and n is the largest number k that divides m and n. That is, $\frac{m}{k}$ and $\frac{n}{k}$ leaves no remainder. Consider the Python code shown in figure 1. Using the fact that the GCD of two numbers does not change if the smaller number is subtracted from the larger number, show that the recursive function gcd(m, n) in the code shown in figure 1 implements the GCD computation.

   (**Note**: Please do not give me an illustrative example as a "proof" of that the code does what it is supposed to do. *I am looking for at least a semi-rigorous attempt at showing the correctness of the code for any pair of numbers*.)

```
def gcd(x , y):
    if y == 0:
        return x
    else:
        return gcd(y, x % y)
```

Figure 1: Function gcd(m, n) for Problem 1.

If two numbers m and n are divisible by p then it follows that $m + n$ and $m - n$ (assuming $m > n$, without loss of generality) is also divisible by p, It is also easy[1] to see that $gcd(n, n) = n$. So, it follows that,

$$gcd(m, n) = \begin{cases} gcd(m - n, n) & \text{if } m > n \\ gcd(m, n - m) & \text{if } n > m \\ n & \text{if } n = m. \end{cases}$$

So, at the i-th stage of any trace, we are replacing $gcd(m.n)$ with either $gcd(m - n, n), gcd(m, n - m)$ or $gcd(n, n)(= n)$ depending on whether $m > n, n > m$ or $n = m$. Since each stage is correct, and one of the two arguments of m or n decreases in size in each step, it follows that the recursion will stop eventually (with the correct answer).

[1] If $m > n$, and $p = gcd(m, n) \Rightarrow gcd(m - n, n) \geq p$. Also, $q = gcd(m - n, n) \Rightarrow gcd(m, n) \geq q$ (why? because $(m - n) = q \times \alpha, n = q \times \beta \Rightarrow m = q \times (\alpha + \beta)$). Therefore $gcd(m, n) = gcd(m - n, n)$.

2. (*40 points*) (**Modular Exponentiation**) Suppose $x, n \in \mathcal{N}$, where n is very very large, we can use the recursive procedure of *Repeated Squaring* to compute x^n in $O(logn)$ time. Oftentimes, we need to compute

$$(x^n)_{mod\ m},$$

where both the exponent $n \in \mathcal{N}$ and the modulus $m \in \mathcal{N}$ are very large[2].

(a) (*10 points*) Show that

$$(x^n)_{mod\ m} = \left\{ \underbrace{(x_{mod\ m}) \times (x_{mod\ m}) \times \cdots \times (x_{mod\ m})}_{n \text{ times}} \right\}_{mod\ m} = \{(x_{mod\ m})^n\}_{mod\ m}$$

If $x = (\alpha \times m) + \beta$ then

$$x^n = \binom{n}{0}(\alpha \times m)^n \beta^0 + \binom{n}{1}(\alpha \times m)^{n-1}\beta + \binom{n}{2}(\alpha \times m)^{n-2}\beta^2 + \cdots + \binom{n}{n}(\alpha \times m)^0 \beta^n.$$

Therefore,

$$\Rightarrow (x^n)_{mod\ m} = (\beta^n)_{mod\ m} = \{(x_{mod\ m})^n\}_{mod\ m}.$$

(b) (*10 points*) Present a pseudo-code sample that uses *Repeated Squaring* to implement the observation in problem 2a. It might help to write some Python code that correctly implements this algorithm. This way you will make sure you have covered all the subtleties (and trust me, there are a few!) in your answer.
See figure 2.

```
def modular_exponent(x, n, m) :
  if (n == 1) :
    return x%m
  else :
    if (n%2 == 1) :
      return (((x%modulus) * modular_exponent(((x%m)*(x%m)), int((n-1)/2), m))%m)
    else :
      return ((modular_exponent(((x%m)*(x%m)), int(n/2), m))%m)
```

Figure 2: Problem 2b.

(c) (*10 points*) (**Chinese Remainder Theorem**): As a follow-on to problems 2a and 2b, let us suppose that the modulus m can be factored as $m = p_1 \times p_2 \times \cdots \times p_k$, where $\forall i, j, gcd(p_i, p_j) = 1$. That is, we have a pairwise

[2] In Public-Key Cryptography, for example.

coprime factorization of the modulus m. Show that the system of k-many equations

$$\begin{aligned}(x^n)_{mod\ p_1} &= a_1\\ (x^n)_{mod\ p_2} &= a_2\\ &\cdots\\ (x^n)_{mod\ p_k} &= a_k\end{aligned}$$

has a unique solution for $(x^n)_{mod\ m}$.

There are many ways to showing this result. Consider the case where $k = 2$. That is, $(x^n)_{mod\ p_1} = a_1$ and $(x^n)_{mod\ p_2} = a_2$. Suppose $q_1 = (p_1^{-1})_{mod\ p_2}$ (i.e. $(p_1q_1)_{mod\ p_2} = 1$) and $q_2 = (p_2^{-1})_{mod\ p_1}$ (i.e. $(p_2q_2)_{mod\ p_1} = 1$) , then without loss of generality we can say $x^n = a_1p_2q_2 + a_2p_1q_1$. This way,

$$(x^n)_{mod\ p_1} = (a_1p_2q_2 + a_2p_1q_1)_{mod\ p_1} = (a_1p_2q_2)_{mod\ p_1} = (a_1)_{mod\ p_1} = a_1,$$

likewise,

$$(x^n)_{mod\ p_2} = (a_1p_2q_2 + a_2p_1q_1)_{mod\ p_2} = (a_2p_1q_1)_{mod\ p_2} = (a_2)_{mod\ p_2} = a_2.$$

All that is left is to show that there is no other solution other than $a_1p_2q_2 + a_2p_1q_1$. If $z_{mod\ p_1} = a_1$, then it follows that $z - a_1p_2q_2 + a_2p_1q_1$ (or, equivalently, $a_1p_2q_2 + a_2p_1q_1 - z$) is a multiple of p_1. This follows from the observation

$$(z-a_1p_2q_2+a_2p_1q_1)_{mod\ p_1} = (z-a_1p_2q_2)_{mod\ p_1} = (z)_{mod\ p_1}-(a_1p_2q_2)_{mod\ p_1} = a_1-a_1 = 0.$$

Similarly, if $z_{mod\ p_1} = a_2$ as well, then it follows that $z-a_1p_2q_2+a_2p_1q_1$ (or, equivalently, $a_1p_2q_2 + a_2p_1q_1 - z$) is a multiple of p_2. Since p_1 and p_2 are co-prime, it follows that $z - a_1p_2q_2 + a_2p_1q_1$ is a multiple of p_1p_2, which would mean that $(z)_{mod\ p_1p_2} = (a_1p_2q_2 + a_2p_1q_1)_{mod\ p_1p_2}$, which establishes uniqueness for the case where $k = 2$.

For the general case, let $q_i = m/p_i = p_1 \times p_2 \times \cdots \times p_{i-1} \times p_{i+1} \times \cdots p_k$. That is, q_i is the product of all co-prime factors, except for p_i. Also, let $\widehat{q_i} = (q_i^{-1})_{mod\ p_i}$ (i.e. $(\widehat{q_i}q_i)_{mod\ p_i} = 1$)). Then the unique solution is

$$\left(\sum_{i=1}^{k} a_iq_i\widehat{q_i}\right)_{mod\ m}.$$

The next problem is about computing the *Prime Factorization* of any number, which together with the *Chinese Remainder Theorem* and the *Repeated Squaring Algorithm*, presents a very efficient algorithm to compute $(x^n)_{mod\ m}$ for large n and m.

(d) (*10 points*) (**Prime Factorization**) A Prime number is a positive integer, greater than 1, that cannot be divided by any other integer other than itself

and 1. The Fundamental Theorem of Arithmetic says every number $n \in \mathcal{N}$ has a unique *Prime Factorization*. For example,

$$\begin{aligned} 123456789 &= 3^2 \times 3607 \times 3803 \\ 987654321 &= 3^2 \times 17^2 \times 379721, \end{aligned}$$

where the numbers $\{3, 17, 3607, 3803, 379721\}$ are all prime numbers. Consider the Python code shown in figure 3. Present a cogent argument that this code will generate the prime factorization of any number n.

```
In [1]: def primeFactors(n):
            c = 2
            while(n > 1):
                if(n % c == 0):
                    print(c, end=" ")
                    n = n / c
                else:
                    c = c + 1

In [2]: primeFactors(987654321)
        3 3 17 17 379721

In [3]: primeFactors(123456789)
        3 3 3607 3803
```

Figure 3: Python Code to compute the Prime Factors of a number.

The Wikipedia page contains a proof that any (composite) number can be expressed a product of primes. The code shown in figure 3 starts with dividing the number by the smallest possible prime number (i.e. $c = 2$). The while-loop "removes" all multiples, and prints the prime number as many times at it divides n. When the while-loop is exited, c is incremented, and

(a) if the new value of c is a prime number, n will be repeatedly divided by c as often as needed

(b) if the new value of c is a composite number, the if statement will never be executed (Why?),

and c will be incremented, and the process repeats. The while-loop is guaranteed to halt as n is reduced in value at each execution.

In principle, we can compute the *Prime Factors* of any modulus m, and then compute the values of $a_1, a_2, \ldots, a_k$ using the Modular Exponentiation with Repeated Squaring. Following this, we can use the efficient algorithms (that you can discover yourself on the web) to solve the equations of the Chinese Remainder Theorem. All this will result in an extremely efficient procedure to compute $(x^n)_{mod\ m}$ for very large n and m.

3. (50 points) For the *Randomized Selection Algorithm* we picked a random member of the array as the *pivot* p (as opposed to p being the *Median-of-Medians*, when we did the *Deterministic Selection Algorithm*). Assume there is only one such p in the array – we have to compare with $(n - 1)$ many others to split the

original array into three arrays. This takes $c * (n-1)$ time. In the worst case, we will recurse on the longer of two arrays (i.e. the array that contains numbers smaller than p, and the array that contains numbers larger than p[3]. The average value of the number of elements in the longer array cannot be larger than $\frac{3n}{4}$. This means the average running time of the algorithm must satisfy the recursion

$$A(n) \leq c(n-1) + A\left(\frac{3n}{4}\right) \leq cn + A\left(\frac{3n}{4}\right) \Rightarrow A(n) \leq 4cn.$$

That is, the average running time of the *Randomized Selection Algorithm* must be linear in n.

To see the fact that the average value of the longer array is bounded by $\frac{3n}{4}$, we used an analogy of splitting a candy-bar of length L into two pieces by picking a uniformly distributed random variable in the interval $[0, L]$. It is not hard to see that the length of the longer-piece is uniformly distributed in the interval $[L/2, L]$, which means the average-value of the length of the longer-piece is $\frac{3L}{4}$.

In class we discussed the possibility of using multiple pivots, and improving the average running time of the Randomized Selection Algorithm. For example, if we had two (random) pivots p_1 and p_2. Using p_1, after $(n-1)$-many comparisons, we split the original array into three parts; this is followed by identifying the appropriate sub-array that contains the k-th smallest elements (that we wish to find). A similar process can be done with the other pivot p_2 – after another $(n-1)$-many comparisons – we identify yet another sub-array that contains the k-th smallest element (that we wish to find). The logical thing to do, would be to recurse on the smaller of these two sub-arrays that contains the k-th smallest element.

We would like to get an estimate of the average running time of the Randomized Selection Algorithm in the presence of m-many randomly-chosen pivots. We will do this in multiple steps in this homework.

(a) (*25 points*) You have m-many candy-bars, each of them has a length L. Each candy bar is broken into two pieces using a uniformly distributed random variable in the interval $x \in [0, L]$ where the two pieces have a length of x and $L - x$, respectively. We throw away the shorter of these two pieces and keep the longer one. In the end, we will have m-many longer-pieces of candy-bars. Show that the average value of the smallest of these m-many longer-pieces of candy-bars is

$$\frac{L(m+2)}{2(m+1)}$$

The larger of the two pieces of a randomly cut length of L will be uniformly distributed in $[L/2, L]$. The probability that any one of these larger pieces

[3]For the worst-case running-time we can glibly ignore the array that contains of elements in the original array that are equal to p , why?

is less than $t \in [L/2, L]$ is given by

$$Prob(x_i \leq t) = \frac{t - L/2}{L/2} = \frac{2t}{L} - 1 \Rightarrow Prob(x_i \geq t) = 2 - \frac{2t}{L}$$

If there are m-many such randomly cut pieces and we want the probability of the minimum of the larger m-many pieces to be less than some $t \in [L/2, L]$, we are looking to compute

$$\begin{aligned} Prob(\min_i x_i \leq t) &= 1 - Prob(\min_i x_i \geq t) \\ &= 1 - \prod_{i=1}^{n} Prob(\min_i x_i \geq t) \\ &= 1 - \left(2 - \frac{2t}{L}\right)^m \end{aligned}$$

The PDF of this statistic is the derivative of the above expression with respect to t, which is

$$\frac{d}{dt}\left(1 - \left(2 - \frac{2t}{L}\right)^m\right) = -m \times \left(2 - \frac{2t}{L}\right)^{m-1} \times -\frac{2}{L} = \frac{2m}{L}\left(2 - \frac{2t}{L}\right)^{m-1}$$

The average value of this statistic is therefore

$$\int_{t=L/2}^{t=L} t \times \frac{2m}{L}\left(2 - \frac{2t}{L}\right)^{m-1} \times dt = \frac{L}{2} \times \frac{m+2}{m+1},$$

The value of this average for different values of m is listed below

$$\underbrace{\frac{2L}{3}}_{m=2}, \underbrace{\frac{5L}{8}}_{m=3}, \underbrace{\frac{3L}{5}}_{m=4}, \text{etc}$$

To see this,

$$\begin{aligned} &\int_{t=L/2}^{t=L} t \times \frac{2m}{L}\left(2 - \frac{2t}{L}\right)^{m-1} \times dt \\ &= \frac{2^m m}{L} \int_{t=L/2}^{t=L} \underbrace{t}_{=u} \times \underbrace{\left(1 - \frac{t}{L}\right)^{m-1} dt}_{=dv \Rightarrow v = -\frac{L}{m}(1-\frac{t}{L})^m} \\ &= \frac{\cancel{2^m}\cancel{m}}{\cancel{L}} \times \frac{L^{\cancel{2}}(m+2)}{2 \times \cancel{2^m} \times \cancel{m}(m+1)} \end{aligned}$$

(b) (*25 Points*) Show that the version of the Randomized Selection Algorithm that uses m-many pivots to split the original array into m-many sub-arrays,

which is then followed by a recursion on the array of smallest-size that contains the k-th smallest element will have an average running time

$$A_m(n) \leq \frac{2(m+1)}{m} \times c \times n.$$

The recursion for the average running-time is

$$A_m(n) \leq c_m(n-1)+A_m\left(\frac{n(m+2)}{2(m+1)}\right) \leq cn+A_m\left(\frac{n(m+2)}{2(m+1)}\right) \Rightarrow A_m(n) \leq \frac{2(m+1)}{m}cn.$$