

Section 5 Solution

With questions by Will Monroe and Julia Daniel

Congratulations on making it through the midterm! :)

1. **Binary Tree:** Consider the following function for constructing binary trees:

```
struct Node {
    Node *left;
    Node *right;
};

Node *randomTree(float p) {
    if (randomBool(p)) { // returns true with probability p
        Node *newNode = new Node;
        newNode->left = randomTree(p);
        newNode->right = randomTree(p);
        return newNode;
    } else {
        return nullptr;
    }
}
```

The `if` branch is taken with probability p (and the `else` branch with probability $1 - p$). A tree with no nodes is represented by `nullptr`; so a tree node with no left child has `nullptr` for the `left` field (and the same for the right child).

Let X be the number of nodes in a tree returned by `randomTree`. You can assume $0 < p < 0.5$. What is $E[X]$, in terms of p ?

Let X_1 and X_2 be number of nodes the left and right calls to `randomTree`.
 $E[X_1] = E[X_2] = E[X]$.

$$\begin{aligned}
 E[X] &= p \cdot E[X \mid \text{if}] + (1 - p)E[X \mid \text{else}] \\
 &= p \cdot E[1 + X_1 + X_2] + (1 - p) \cdot 0 \\
 &= p \cdot (1 + E[X] + E[X]) \\
 &= p + 2pE[X] \\
 (1 - 2p)E[X] &= p \\
 E[X] &= \frac{p}{1 - 2p}
 \end{aligned}$$

2. **Girl Scout Cookies** Three girl scouts are out selling cookies. The number of boxes of cookies each girl sells in a day in the neighborhood is randomly distributed with mean μ and standard deviation σ .

- a. Noa and Chaya are sisters, but each out selling cookies on their own today. What are the parameters for the distribution of the total number of boxes sold by families that have two daughters both out selling cookies? Assume they are in different neighborhoods and their sales results are independent.

$$X = \text{Noa} + \text{Chaya} \sim N(\mu + \mu = 2\mu, \sigma^2 + \sigma^2 = 2\sigma^2)$$

- b. Maria's grandmother has told her that for every box Maria sells today, her grandmother will also buy a box to donate to the food bank. What are the parameters for the distribution of the total number of boxes Maria will sell?

$$Y = 2 \cdot \text{Maria} \sim N(\mu \cdot 2 = 2\mu, (2\sigma)^2 = 4\sigma^2)$$

3. Timing Attack:

In this problem we are going to show you how to crack a password in linear time, by measuring how long the password check takes to execute (see code below). Assume that our server takes T ms to execute any line in the code where $T \sim N(\mu = 5, \sigma^2 = 0.5)$ seconds. The amount of time taken to execute a line is always independent of other values of T .

```
# An insecure string comparison
def stringEquals(guess, password):
    nGuess = len(guess)
    nPassword = len(password)
    if nGuess != nPassword:
        return False
    for i in range(nGuess):
        if guess[i] != password[i]:
            return False
    return True
```

4 lines executed to get here

6 + 2i lines executed to get here

5 + 2n lines executed to get here

On our site all passwords are length 5 through 10 (inclusive) and are composed of lower case letters only. A hacker is trying to crack the root password which is "gobayes" by carefully measuring how long we take to tell them that her guesses are incorrect.

- a. What is the distribution of time that it takes our server to execute k lines of code? Recall that each line independently takes $T \sim N(\mu = 5, \sigma^2 = 0.5)$ ms.

Let Y be the amount of time to execute k lines. $Y = \sum_{i=1}^k X_i$ where X_i is the amount of time to execute line i . $X_i \sim N(\mu = 5, \sigma^2 = 0.5)$.

Since Y is the sum of independent normals:

$$\begin{aligned} Y &\sim N(\mu = \sum_{i=1}^k 5, \sigma^2 = \sum_{i=1}^k 0.5) \\ &\sim N(\mu = 5k, \sigma^2 = 0.5k) \end{aligned}$$

- b. First the hacker needs to find out the length of the password. What is the probability that the time taken to test a guess of correct length (server executes 6 lines) is longer than the time taken to test a guess of an incorrect length (server executes 4 lines)? Assume that the first letter of the guess does not match the first letter of the password. Hint: $P(A > B)$ is the same as $P(A - B > 0)$.

From last problem:

Time to run 6 lines of code $A \sim N(\mu = 30, \sigma^2 = 3)$

Time to run 4 lines of code $B \sim N(\mu = 20, \sigma^2 = 2)$

$$\begin{aligned} -B &\sim N(\mu = -20, \sigma^2 = 2) \\ A - B &\sim N(\mu = 10, \sigma^2 = 5) \\ P(A > B) &= P(A - B > 0) \\ &= 1 - F_{A-B}(0) \\ &= 1 - \Phi\left(\frac{0 - 10}{\sqrt{5}}\right) \\ &\approx 1.0 \end{aligned}$$

- c. Now that our hacker knows the length of the password, to get the actual string she is going to try and figure out each letter one at a time, starting with the first letter. The hacker tries the string “aaaaaaa” and it takes 27s. Based on this timing, how much more probable is it that first character did not match (server executes 6 lines) than the first character did match (server executes 8 lines)? Assume that all letters in the alphabet are equally likely to be the first letter.

Let M be the event that the first letter matched.

$$\begin{aligned}
 \frac{P(M^C|T = 27)}{P(M|T = 27)} &= \frac{f(T = 27|M^C)P(M^C)}{f(T = 27|M)P(M)} \\
 &= \frac{f(T = 27|M^C)\frac{25}{26}}{f(T = 27|M)\frac{1}{26}} \\
 &= 25 \cdot \frac{f(T = 27|M^C)}{f(T = 27|M)} \\
 &= 25 \cdot \frac{\frac{1}{\sqrt{6\pi}}e^{-\frac{(27-30)^2}{6}}}{\frac{1}{\sqrt{8\pi}}e^{-\frac{(27-40)^2}{8}}} \\
 &= 25 \cdot \frac{\sqrt{8}}{\sqrt{6}} \cdot \frac{e^{-\frac{9}{6}}}{e^{-\frac{169}{8}}} \\
 &\approx 9.6 \text{ million}
 \end{aligned}$$

- d. If it takes the hacker 6 guesses to find the length of the password, and 26 guesses per letter to crack the password string, how many attempts does she need to crack our password, “gobayes”? Yikes!

$$7 \cdot 26 + 6 = 188$$