W05: Analysing Randomised Algorithms

CS3230 AY21/22 Sem 2

Table of Contents

Click on the link to jump to the relevant sections!

Question 1: Worst-cases

Question 1

The TA would like to encourage everyone in class to prepare before coming to class.

To do that, the TA wants to select students who have not prepared to answer questions in class.

The TA selects k students in each tutorial to answer the questions and wants to come up with an algorithm for selecting the students. The worst case for the TA would be the case where only the k selected students have prepared, but the other n-k students in the class did not prepare. The TA thinks that an uncooperative class can make the worst case happen every time if the algorithm for selecting the students is made known. Hence there is no choice but to hide the selection algorithm **(source code or pseudocode)** from the class. Is it true that if the TA makes the selection algorithm known, the students will be able to force the worst case to happen?

- ○ Yes
- ○ No

Question 1 (Answer)

No! I can come up with a randomised algorithm:

```
def choose_student(n):
  students = choose k numbers randomly from 1 to n
  return students
```

Question 1 (Answer)

No! I can come up with a randomised algorithm:

```
def choose_student(n):
  students = choose k numbers randomly from 1 to n
  return students
```

Idea: If your algorithm makes use of randomness to make decisions, even people knowing the algorithm **cannot find out a way to force worst-case**

Random Variable

A random variable *X,* is a **function** (not exactly a variable!) that maps a sample space *S* to a real number *R*. That is: $X: S \rightarrow R$

Random Variable

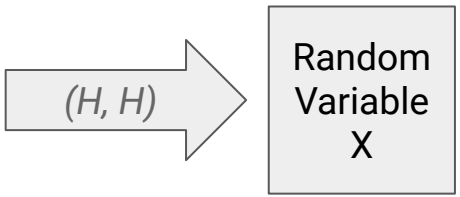
A random variable *X,* is a **function** (not exactly a variable!) that maps a sample space *S* to a real number *R*. That is: $X: S \rightarrow R$

e.g. Let *X* be a random variable of the **number of heads in 2 flips**.
S = { (H, H), (H, T), (T, H), (T, T) }.

Random Variable

A random variable *X,* is a **function** (not exactly a variable!) that maps a sample space *S* to a real number *R*. That is: $X: S \rightarrow R$

e.g. Let *X* be a random variable of the **number of heads in 2 flips**.
S = { (H, H), (H, T), (T, H), (T, T) }.



Random Variable

A random variable *X,* is a **function** (not exactly a variable!) that maps a sample space *S* to a real number *R*. That is: $X: S \rightarrow R$

e.g. Let *X* be a random variable of the **number of heads in 2 flips**.
S = { (H, H), (H, T), (T, H), (T, T) }.

Random Variable

A random variable *X,* is a **function** (not exactly a variable!) that maps a sample space *S* to a real number *R*. That is: $X: S \rightarrow R$

e.g. Let *X* be a random variable of the **number of heads in 2 flips**.
S = { (H, H), (H, T), (T, H), (T, T) }.

$$X = \begin{cases} 2 & \text{w.p. } \frac{1}{4} : \text{(H, H) case} \\ 1 & \text{w.p. } \frac{1}{2} : \text{(H, T) or (T, H) case} \\ 0 & \text{w.p. } \frac{1}{4} : \text{(T, T)} \end{cases}$$

"w.p." = with probability

Random Variable

We also write the probability this way: $\Pr(X = 2) = \frac{1}{4}$

$$X = \begin{cases} 2 & \text{w.p. } \frac{1}{4} : \text{(H, H) case} \\ 1 & \text{w.p. } \frac{1}{2} : \text{(H, T) or (T, H) case} \\ 0 & \text{w.p. } \frac{1}{4} : \text{(T, T)} \end{cases}$$

"w.p." = with probability

Bernoulli Trial

An instance of a ***Bernoulli trial*** has:

- Probability p of success and
- Probability $q = 1 - p$ of failure.

Bernoulli Trial

An instance of a ***Bernoulli trial*** has:

- Probability p of success and
- Probability $q = 1 - p$ of failure.

Think of it like a generalised coin flip!
e.g. Define “success” as getting 6 when rolling a die:

Bernoulli Trial

An instance of a ***Bernoulli trial*** has:

- Probability p of success and
- Probability $q = 1 - p$ of failure.

Think of it like a generalised coin flip!
e.g. Define “success” as getting 6 when rolling a die:

Bernoulli trial with

$p = \frac{1}{6}$

$q = \frac{5}{6}$

Geometric Distribution

They must all be the same probability!

Suppose we have a sequence of independent Bernoulli trials, each with prob p of success.

Geometric Distribution

They must all be the same probability!

Suppose we have a sequence of independent Bernoulli trials, each with prob *p* of success.

Let *X* be the random variable for **number of trials** needed to obtain **success for the first time**. "Keep trying until you get it"

Geometric Distribution

They must all be the same probability!

Suppose we have a sequence of independent Bernoulli trials, each with prob *p* of success.
Let *X* be the random variable for **number of trials** needed to obtain **success for the first time**. "Keep trying until you get it"

Then, *X* follows the *geometric distribution*:

$$\Pr(X = k) = q^{k-1}p$$

$$E[X] = \frac{1}{p}$$

Geometric Distribution

They must all be the same probability!

Suppose we have a sequence of independent Bernoulli trials, each with prob *p* of success.

Let *X* be the random variable for **number of trials** needed to obtain **success for the first time**. "Keep trying until you get it"

If you succeed on the k^{th} try for the first time, then you must have **failed (k-1)** times!

Then, *X* follows the *geometric distribution*:

$$\Pr(X = k) = q^{k-1}p$$

$$E[X] = \frac{1}{p}$$

Geometric Distribution

They must all be the same probability!

Suppose we have a sequence of independent Bernoulli trials, each with prob *p* of success.

Let *X* be the random variable for **number of trials** needed to obtain **success for the first time**. "Keep trying until you get it"

If you succeed on the k^{th} try for the first time, then you must have **failed (k-1)** times!

Then, *X* follows the *geometric distribution*:

$$\Pr(X = k) = q^{k-1}p$$

$$E[X] = \frac{1}{p}$$

Just remember this

Geometric Distribution: Coin Flips

Let X be the random variable of number of trials to reach head for the first time. What is the expected value of X?

Geometric Distribution: Coin Flips

Let X be the random variable of number of trials to reach head for the first time. What is the expected value of X?

Bernoulli Trial with $p = ½$

Geometric Distribution: Coin Flips

$$E[X] = \frac{1}{p}$$

Let *X* be the random variable of number of trials to reach head for the first time. What is the expected value of *X*?

Bernoulli Trial with *p* = ½

Thus: $E[X] = \frac{1}{\frac{1}{2}} = 2$

Geometric Distribution: Dice Roll

Let *Y* be the random variable of number of trials to roll a ‘2’ for the first time. What is the expected value of *Y*?

Geometric Distribution: Dice Roll

$$E[X] = \frac{1}{p}$$

Let *Y* be the random variable of number of trials to roll a ‘2’ for the first time. What is the expected value of *Y*?

Bernoulli Trial with $p = \frac{1}{6}$

Thus: $E[Y] = \frac{1}{\frac{1}{6}} = 6$

Question 2: Finding unprepared students

Question 2

The TA would like to encourage everyone in class to prepare before coming to class. To do that, the TA wants to select students who have not prepared to answer questions in class.

The TA decides to select students at random (with replacement) to answer the questions. Assume that there are n students and k of them have not prepared. What is the expected number of questions required for finding a student who has not prepared?

---

1. n
2. n/k
3. $n/2k$
4. k/n

I choose students at random (with replacement), and ask a question. I keep repeating until I find someone who **has not prepared**

Let's say we have *n* students and *k* students have not prepared.

Question 2 (Answer)

This is exactly a **geometric distribution**! Let *X* be the random variable on number of trials until first unprepared student is selected

I choose students at random (with replacement), and ask a question. I keep repeating until I find someone who **has not prepared**

Let's say we have n students and k students have not prepared.

Question 2 (Answer)

This is exactly a **geometric distribution**! Let X be the random variable on number of trials until first unprepared student is selected

Bernoulli Trial with $p = k / n$

I choose students at random (with replacement), and ask a question. I keep repeating until I find someone who **has not prepared**

Let's say we have *n* students and *k* students have not prepared.

Question 2 (Answer)

This is exactly a **geometric distribution**! Let *X* be the random variable on number of trials until first unprepared student is selected

Bernoulli Trial with $p = k / n$

Thus: $E[X] = \frac{1}{\frac{k}{n}} = \frac{n}{k}$

Linearity of Expectation

For any two events *X and Y, (can be independent or dependent),* and a constant *c*

$$E[X+Y] = E[X] + E[Y]$$

$$E[cX] = cE[X]$$

Linearity of Expectation

It's a little hammer that is frequently used! Remember this well!

$$E[X+Y] = E[X] + E[Y]$$

$$E[cX] = cE[X]$$

Harmonic Series

$$
\begin{aligned}
H_n &= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n} \\
&= \sum_{k=1}^{n} \frac{1}{k} \\
&= \ln n + O(1)
\end{aligned}
$$

Harmonic Series

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n}$$

$$= \sum_{k=1}^{n} \frac{1}{k}$$

$$= \ln n + O(1) \quad = \theta(\lg n)$$

The analysis “pattern” (1 & 2 are “interchangeable”)

Most (but not all) analysis in Randomised Algorithms follow this “pattern”.

1. Identify a Random Variable to “count” what you want (e.g. X. Goal: $E[X]$)

2. Express this RV as a **sum** of random variables (e.g. $X = X_1 + X_2 + \ldots + X_n$)
   a. Calculate the relevant probability for X_1, X_2, ...
   b. Calculate the individual expectation of the “sub”-random variables. ($E[X_1]$, $E[X_2]$, …)

3. Use linearity of expectations on $E[X]$. Then you add up the expectation of the “sub”-random variables (from step 2b)

Coupon Collector

- There are *n* types of coupon you want to collect. Randomly drawn with replacement

- What is the expected number of times you have to draw a coupon before having **at least one of each type of coupon?**

Coupon Collector

1. Identify a Random Variable to "count" what you want (e.g. X. Goal: $E[X]$)
2. Express this RV as a **sum** of random variables (e.g. $X = X_1 + X_2 + \ldots + X_n$)
   a. Calculate the relevant probability for X_1, X_2, ...
   b. Calculate the individual expectation of the "sub"-random variables. ($E[X_1]$, $E[X_2]$, …)
3. Use linearity of expectations on $E[X]$. Then you add up the expectation of the "sub"-random variables (from step 2b)

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Coupon Collector

1. Identify a Random Variable to "count" what you want (e.g. X. Goal: $E[X]$)
2. Express this RV as a **sum** of random variables (e.g. $X = X_1 + X_2 + \ldots + X_n$)
   a. Calculate the relevant probability for X_1, X_2, ...
   b. Calculate the individual expectation of the "sub"-random variables. ($E[X_1]$, $E[X_2]$, …)
3. Use linearity of expectations on $E[X]$. Then you add up the expectation of the "sub"-random variables (from step 2b)

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Important Clarification:
This is to collect the i^{th} ***new*** coupon. This is **NOT** assigning an ID to each coupon, and choosing the coupon that has ID = i.

For example, for the purpose of clarity. Let's say we have coupon type A, B, C, D, E.
If we have collected 2 coupons, for example, A and E.

T_3 is collecting ANY of B or C or D (the *third* new coupon)
it is NOT collecting a specific coupon (e.g. C)

1. Identify a Random Variable to "count" what you want (e.g. X. Goal: $E[X]$)
2. Express this RV as a **sum** of random variables (e.g. $X = X_1 + X_2 + \ldots + X_n$)
   a. Calculate the relevant probability for X_1, X_2, ...
   b. Calculate the individual expectation of the "sub"-random variables. ($E[X_1]$, $E[X_2]$, …)
3. Use linearity of expectations on $E[X]$. Then you add up the expectation of the "sub"-random variables (from step 2b)

Coupon Collector

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Example (5 coupon types A, B, C, D, E):

B B D B D C B C E C B D A

Coupon Collector

1. Identify a Random Variable to "count" what you want (e.g. X. Goal: $E[X]$)
2. Express this RV as a **sum** of random variables (e.g. $X = X_1 + X_2 + \ldots + X_n$)
   a. Calculate the relevant probability for X_1, X_2, ...
   b. Calculate the individual expectation of the "sub"-random variables. ($E[X_1]$, $E[X_2]$, …)
3. Use linearity of expectations on $E[X]$. Then you add up the expectation of the "sub"-random variables (from step 2b)

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Example (5 coupon types - A, B, C, D, E):

B B D B D C B C E C B D A

New coupon in 1 trial ($T_1 = 1$)

Coupon Collector

1. Identify a Random Variable to "count" what you want (e.g. X. Goal: $E[X]$)
2. Express this RV as a **sum** of random variables (e.g. $X = X_1 + X_2 + \ldots + X_n$)
   a. Calculate the relevant probability for X_1, X_2, ...
   b. Calculate the individual expectation of the "sub"-random variables. ($E[X_1]$, $E[X_2]$, …)
3. Use linearity of expectations on $E[X]$. Then you add up the expectation of the "sub"-random variables (from step 2b)

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Example (5 coupon types - A, B, C, D, E):

B B D B D C B C E C B D A

New coupon in 2 trials ($T_2 = 2$)

Coupon Collector

1. Identify a Random Variable to "count" what you want (e.g. X. Goal: $E[X]$)
2. Express this RV as a **sum** of random variables (e.g. $X = X_1 + X_2 + \ldots + X_n$)
   a. Calculate the relevant probability for X_1, X_2, ...
   b. Calculate the individual expectation of the "sub"-random variables. ($E[X_1]$, $E[X_2]$, …)
3. Use linearity of expectations on $E[X]$. Then you add up the expectation of the "sub"-random variables (from step 2b)

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Example (5 coupon types - A, B, C, D, E):

B B D B D C B C E C B D A

New coupon in 3 trials ($T_3 = 3$)

Coupon Collector

1. Identify a Random Variable to "count" what you want (e.g. X. Goal: $E[X]$)
2. Express this RV as a **sum** of random variables (e.g. $X = X_1 + X_2 + \ldots + X_n$)
   a. Calculate the relevant probability for X_1, X_2, ...
   b. Calculate the individual expectation of the "sub"-random variables. ($E[X_1]$, $E[X_2]$, …)
3. Use linearity of expectations on $E[X]$. Then you add up the expectation of the "sub"-random variables (from step 2b)

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) ***after i - 1 distinct coupons have been collected***

Example (5 coupon types - A, B, C, D, E):

B B D B D C B C E C B D A

New coupon in 3 trials ($T_4 = 3$)

1. Identify a Random Variable to "count" what you want (e.g. X. Goal: $E[X]$)
2. Express this RV as a **sum** of random variables (e.g. $X = X_1 + X_2 + \ldots + X_n$)
   a. Calculate the relevant probability for X_1, X_2, ...
   b. Calculate the individual expectation of the "sub"-random variables. ($E[X_1]$, $E[X_2]$, ...)
3. Use linearity of expectations on $E[X]$. Then you add up the expectation of the "sub"-random variables (from step 2b)

Coupon Collector

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Example (5 coupon types - A, B, C, D, E):

B B D B D C B C E C B D A

Last coupon in 4 trials ($T_5 = 4$)

Total number of trials is equal to the sum of all T_i!

1. Identify a Random Variable to “count” what you want (e.g. X. Goal: $E[X]$)
2. Express this RV as a **sum** of random variables (e.g. $X = X_1 + X_2 + \ldots + X_n$)
   a. Calculate the relevant probability for X_1, X_2, ...
   b. Calculate the individual expectation of the “sub”-random variables. ($E[X_1]$, $E[X_2]$, …)
3. Use linearity of expectations on $E[X]$. Then you add up the expectation of the “sub”-random variables (from step 2b)

Coupon Collector

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Success = collecting *i-th* coupon

Coupon Collector

1. Identify a Random Variable to "count" what you want (e.g. X. Goal: $E[X]$)
2. Express this RV as a **sum** of random variables (e.g. $X = X_1 + X_2 + \ldots + X_n$)
   a. Calculate the relevant probability for X_1, X_2, ...
   b. Calculate the individual expectation of the "sub"-random variables. ($E[X_1]$, $E[X_2]$, …)
3. Use linearity of expectations on $E[X]$. Then you add up the expectation of the "sub"-random variables (from step 2b)

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Success = collecting *i-th* coupon. Denote probability of success with p_i

$$p_i = \frac{n-(i-1)}{n}$$

Coupon Collector

1. Identify a Random Variable to "count" what you want (e.g. X. Goal: $E[X]$)
2. Express this RV as a **sum** of random variables (e.g. $X = X_1 + X_2 + \ldots + X_n$)
   a. Calculate the relevant probability for X_1, X_2, ...
   b. Calculate the individual expectation of the "sub"-random variables. ($E[X_1]$, $E[X_2]$, …)
3. Use linearity of expectations on $E[X]$. Then you add up the expectation of the "sub"-random variables (from step 2b)

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Success = collecting *i-th* coupon. Denote probability of success with p_i

$$p_i = \frac{n-(i-1)}{n}$$

Out of the *n* things, you have collected *(i - 1)*.
There are only *n - (i - 1)* "new" coupons left

Coupon Collector

1. Identify a Random Variable to "count" what you want (e.g. X. Goal: $E[X]$)
2. Express this RV as a **sum** of random variables (e.g. $X = X_1 + X_2 + \ldots + X_n$)
   a. Calculate the relevant probability for X_1, X_2, ...
   b. Calculate the individual expectation of the "sub"-random variables. ($E[X_1]$, $E[X_2]$, ...)
3. Use linearity of expectations on $E[X]$. Then you add up the expectation of the "sub"-random variables (from step 2b)

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Success = collecting *i-th* coupon. Denote probability of success with p_i

$$p_i = \frac{n-(i-1)}{n}$$

Out of the *n* things, you have collected *(i - 1)*.
There are only *n - (i - 1)* "new" coupons left

$$E[X] = \frac{1}{p}$$

T_i is a geometric distribution with: $E[T_i] = \frac{1}{p_i} = \frac{n}{n-(i-1)}$

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Let T be the **total number of trials** to collect all n coupons

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Let T be the **total number of trials** to collect all n coupons

1. Identify a Random Variable to "count" what you want (e.g. X. Goal: $E[X]$)
2. Express this RV as a **sum** of random variables (e.g. $X = X_1 + X_2 + \ldots + X_n$)
   a. Calculate the relevant probability for X_1, X_2, ...
   b. Calculate the individual expectation of the "sub"-random variables. ($E[X_1]$, $E[X_2]$, …)
3. Use linearity of expectations on $E[X]$. Then you add up the expectation of the "sub"-random variables (from step 2b)

$$T = T_1 + T_2 + \cdots + T_n$$

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Let *T* be the **total number of trials** to collect all n coupons

$$T = T_1 + T_2 + \cdots + T_n$$

$$E[T] = E[T_1 + T_2 + \cdots + T_n]$$

Applying expectations to both sides!

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Let *T* be the **total number of trials** to collect all n coupons

$$T = T_1 + T_2 + \cdots + T_n$$

$$E[T] = E[T_1 + T_2 + \cdots + T_n]$$

Applying expectations to both sides!

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Let T be the **total number of trials** to collect all n coupons

$$T = T_1 + T_2 + \cdots + T_n$$

$$E[T] = E[T_1 + T_2 + \cdots + T_n]$$

Applying expectations to both sides!

$$= E[T_1] + E[T_2] + \cdots + E[T_n]$$

LoE!

Let T_i be the number of trials to collect a new coupon (call it the *i-th* coupon) **after *i - 1 distinct coupons have been collected***

Let *T* be the **total number of trials** to collect all n coupons

A more concise version using the summation notation!

$$T = \sum_{i=1}^{n} T_i$$

Total number of draws

$$E[T] = E\left[\sum_{i=1}^{n} T_i\right]$$

Expected value

$$= \sum_{i=1}^{n} E[T_i]$$

Linearity of expectation

$$E[T] = \sum_{i=1}^{n} E[T_i]$$

Coupon Collector

Use the fact from the previous slide -- linearity of expectations

$$E[T] = \sum_{i=1}^{n} E[T_i]$$

$$= \sum_{i=1}^{n} \frac{n}{n-(i-1)}$$

Coupon Collector

We also previously analysed this!

T_i is a geometric distribution with:

$$E[T_i] = \frac{1}{p_i} = \frac{n}{n-(i-1)}$$

$$E[T] = \sum_{i=1}^{n} E[T_i]$$

$$= \sum_{i=1}^{n} \frac{n}{n-(i-1)}$$

$$= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1}$$

Coupon Collector

Just expanding the sums...

$$E[T] = \sum_{i=1}^{n} E[T_i]$$

$$= \sum_{i=1}^{n} \frac{n}{n-(i-1)}$$

$$= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1}$$

$$= n \cdot \left(\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{n} \right)$$

Coupon Collector

Factoring *n* out!

$$E[T] = \sum_{i=1}^{n} E[T_i]$$

$$= \sum_{i=1}^{n} \frac{n}{n-(i-1)}$$

$$= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1}$$

$$= n \cdot \left(\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{n} \right)$$

$$= n \cdot H_n$$

Coupon Collector

Harmonic Series (in reverse)!

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n}$$

$$= \sum_{k=1}^{n} \frac{1}{k}$$

$$= \ln n + O(1)$$

$$
\begin{aligned}
E[T] &= \sum_{i=1}^{n} E[T_i] \\
&= \sum_{i=1}^{n} \frac{n}{n-(i-1)} \\
&= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} \\
&= n \cdot \left(\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{n} \right) \\
&= n \cdot H_n \\
&= \Theta(n \lg n)
\end{aligned}
$$

Coupon Collector

The bound on harmonic series

$$
\begin{aligned}
H_n &= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n} \\
&= \sum_{k=1}^{n} \frac{1}{k} \\
&= \ln n + O(1)
\end{aligned}
$$

Question 3: Everyone shall participate!

Question 3

The TA would like to encourage everyone in class to prepare before coming to class. To do that, the TA wants to select students who have not prepared to answer questions in class.

The TA decides to select students at random to answer questions. However, the TA is worried that not everyone in class would get to participate even though many questions will be asked throughout the semester. Assume that there are n students in the class. What is the expected number of questions that need to be asked before every student has been asked a question?

- $\Theta(\lg n)$
- $\Theta(n)$
- $\Theta(n \lg n)$
- $\Theta(n^2)$

Question 3 (Answer)

This is exactly the **coupon collector** problem!

Question 3 (Answer)

This is exactly the **coupon collector** problem!

Coupons = Students

Collector choosing a coupon with replacement = Me selecting a student with replacement

Thus, the expected questions asked is $\boldsymbol{\theta(nlgn)}$

Question 4: Finding Males

Question 4

- Let A[1..n] be an array of n distinct names. Suppose m of them are male names. We hope to select q male names from A[1..n]. We propose the following algorithm to obtain q male names.

- Since personal data is sensitive, we hope to estimate the expected number of accesses to the array *A*.

- Please compute the expected number of access of Query(A, q).

```
Query(A, q)
Let S=Φ;
for j = 1 to q
    Repeat
        Randomly select k from
            {1, 2, …, n};
        Set B=A[k];
    Until B is a male and k∉S;
    S = {k}∪S;
Report S;
```

Question 4 (Answer)

This question should feel *very* familiar! It is yet **another coupon collector** (but modified to stop early if you have obtained q coupons):

Question 4 (Answer)

This question should feel *very* familiar! It is yet **another coupon collector** (but modified to stop early if you have obtained *q* coupons):

Coupon = Male students
Collector choosing a coupon with replacement = Algorithm accessing Array A

```
def Query(A, q):
1. Let S = {}
2. for j = 1 to q:
3.   repeat
4.     k = randomly selected from {1, 2, ..., n}
5.     B = A[k]
6.   until (B is male) and (k not in S)
7. return S
```

Question 4 (Answer)

This question should feel *very* familiar! It is yet **another coupon collector** (but modified to stop early if you have obtained *q* coupons):

Coupon = Male students
Collector choosing a coupon with replacement = Algorithm accessing Array A

Let X be the total number of accesses
Let X_j be the number of accesses to obtain *j-th* male

$X = X_1 + X_2 + \ldots + X_q$ (stop until q has been found!)

```
def Query(A, q):
1. Let S = {}
2. for j = 1 to q:
3.   repeat
4.     k = randomly selected from {1, 2, ..., n}
5.     B = A[k]
6.   until (B is male) and (k not in S)
7. return S
```

Finding 1st male: there are *m* males left out of *n* people

Finding 2nd male: there are *(m-1)* males left out of *n* people

…

Finding j-th male: there are (m - (j - 1)) males left out of *n* people

```
def Query(A, q):
1. Let S = {}
2. for j = 1 to q:
3.   repeat
4.     k = randomly selected from {1, 2, ..., n}
5.     B = A[k]
6.   until (B is male) and (k not in S)
7. return S
```

Finding 1st male: there are *m* males left out of *n* people

Finding 2nd male: there are *(m-1)* males left out of *n* people

…

Finding j-th male: there are (m - (j - 1)) males left out of *n* people

Hence, X_j is a geometric distribution with success probability $p_j = \frac{m-j+1}{n}$

Thus:

$$E[X_j] = \frac{1}{p_j} = \frac{n}{m-j+1}$$

```
def Query(A, q):
1. Let S = {}
2. for j = 1 to q:
3.    repeat
4.       k = randomly selected from {1, 2, ..., n}
5.       B = A[k]
6.    until (B is male) and (k not in S)
7. return S
```

$$E[X] = E[\sum_{j=1}^{q} X_j]$$

Let X be the total number of accesses

Let X_j be the number of accesses to obtain *j-th* male

$X = X_1 + X_2 + \ldots + X_q$ (stop until q has been found!)

Question 4

This is how we defined X

$$E[X] = E[\sum_{j=1}^{q} X_j]$$

$$= \sum_{j=1}^{q} E[X_j]$$

Let X be the total number of accesses

Let X_j be the number of accesses to obtain *j-th* male

$X = X_1 + X_2 + \ldots + X_q$ (stop until q has been found!)

Question 4

Linearity of Expectations!

Let X be the total number of accesses

Let X_j be the number of accesses to obtain *j-th* male

$X = X_1 + X_2 + \ldots + X_q$ (stop until q has been found!)

$$E[X] = E[\sum_{j=1}^{q} X_j]$$

$$= \sum_{j=1}^{q} E[X_j]$$

$$= \sum_{j=1}^{q} \frac{n}{m-j+1}$$

Question 4

We have calculated $E[X_j]$

$$E[X_j] = \frac{1}{p_j} = \frac{n}{m-j+1}$$

Let X be the total number of accesses

Let X_j be the number of accesses to obtain *j-th* male

$X = X_1 + X_2 + \ldots + X_q$ (stop until q has been found!)

Question 4

Just expanding the sums!

$$E[X] = E[\sum_{j=1}^{q} X_j]$$

$$= \sum_{j=1}^{q} E[X_j]$$

$$= \sum_{j=1}^{q} \frac{n}{m-j+1}$$

$$= \frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1}$$

$$E[X] = \frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1}$$

For this analysis, assume q < m.
The analysis for q = m is simpler

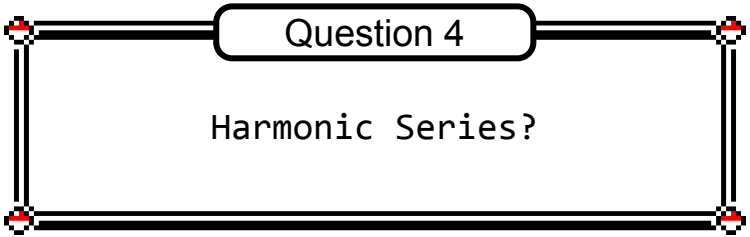
Question 4

Continue from previous slide

$$E[X] = \frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1}$$

Intuition: These seem to look like Harmonic Series… Right?
But we did not start from (1 / 1), so not quite!

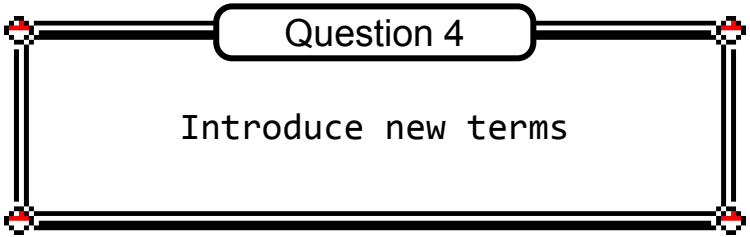
$$\begin{aligned} H_n &= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n} \\ &= \sum_{k=1}^{n} \frac{1}{k} \\ &= \ln n + O(1) \end{aligned}$$



Special case: If m=q, we get $\theta(n \lg(m))$

But for q < m we must add something extra…

$$E[X] = \frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1}$$
$$= (\frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1} + \frac{n}{m-q} + \cdots + \frac{n}{1}) - (\frac{n}{m-q} + \cdots \frac{n}{1})$$



$$E[X] = \frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1}$$

$$= (\frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1} + \frac{n}{m-q} + \cdots + \frac{n}{1}) - (\frac{n}{m-q} + \cdots \frac{n}{1})$$

Question 4

This is just adding terms that cancel!

Analogy:
(5 + 4 + 3)
= (5 + 4 + 3 + 2 + 1) - (2 + 1)

$$E[X] = \frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1}$$

$$= (\frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1} + \frac{n}{m-q} + \cdots + \frac{n}{1}) - (\frac{n}{m-q} + \cdots \frac{n}{1})$$

$$= n(\frac{1}{m} + \frac{1}{m-1} \cdots + \frac{1}{m-q+1} + \frac{1}{m-q} + \cdots + \frac{1}{1}) - n(\frac{1}{m-q} + \ldots \frac{1}{1})$$

Question 4

Just factoring n out

$$E[X] = \frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1}$$

$$= (\frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1} + \frac{n}{m-q} + \cdots + \frac{n}{1}) - (\frac{n}{m-q} + \ldots \frac{n}{1})$$

$$= n(\frac{1}{m} + \frac{1}{m-1} \cdots + \frac{1}{m-q+1} + \frac{1}{m-q} + \cdots + \frac{1}{1}) - n(\frac{1}{m-q} + \ldots \frac{1}{1})$$

$$= n(H_m) - n(H_{m-q})$$

Question 4

Harmonic Series (in reverse)!
But this time instead of ending at n, it ends differently

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n}$$

$$= \sum_{k=1}^{n} \frac{1}{k}$$

$$= \ln n + O(1)$$

$$
\begin{aligned}
E[X] &= \frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1} \\
&= (\frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1} + \frac{n}{m-q} + \cdots + \frac{n}{1}) - (\frac{n}{m-q} + \ldots \frac{n}{1}) \\
&= n(\frac{1}{m} + \frac{1}{m-1} \cdots + \frac{1}{m-q+1} + \frac{1}{m-q} + \cdots + \frac{1}{1}) - n(\frac{1}{m-q} + \ldots \frac{1}{1}) \\
&= n(H_m) - n(H_{m-q}) \\
&= (n\ln(m) - n\ln(m-q)) + O(1)
\end{aligned}
$$

Question 4

Bound on Harmonic Series

$$
\begin{aligned}
H_n &= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n} \\
&= \sum_{k=1}^{n} \frac{1}{k} \\
&= \ln n + O(1)
\end{aligned}
$$

$$E[X] = \frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1}$$

$$= (\frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1} + \frac{n}{m-q} + \cdots + \frac{n}{1}) - (\frac{n}{m-q} + \ldots \frac{n}{1})$$

$$= n(\frac{1}{m} + \frac{1}{m-1} \cdots + \frac{1}{m-q+1} + \frac{1}{m-q} + \cdots + \frac{1}{1}) - n(\frac{1}{m-q} + \ldots \frac{1}{1})$$

$$= n(H_m) - n(H_{m-q})$$

$$= (n \ln(m) - n \ln(m-q)) + O(1)$$

$$= \theta(n \lg(\frac{m}{m-q}))$$

Question 4

Logarithm Laws!

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n}$$

$$= \sum_{k=1}^{n} \frac{1}{k}$$

$$= \ln n + O(1)$$

$$
\begin{aligned}
E[X] &= \frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1} \\
&= (\frac{n}{m} + \frac{n}{m-1} \cdots + \frac{n}{m-q+1} + \frac{n}{m-q} + \cdots + \frac{n}{1}) - (\frac{n}{m-q} + \ldots \frac{n}{1}) \\
&= n(\frac{1}{m} + \frac{1}{m-1} \cdots + \frac{1}{m-q+1} + \frac{1}{m-q} + \cdots + \frac{1}{1}) - n(\frac{1}{m-q} + \ldots \frac{1}{1}) \\
&= n(H_m) - n(H_{m-q}) \\
&= (n\ln(m) - n\ln(m-q)) + O(1) \\
&= \theta(n \lg(\frac{m}{m-q}))
\end{aligned}
$$

Question 4

Done! Therefore this is the expected total number of accesses