Welcome Back Our Friend: Expectation

- Recall expectation for discrete random variable:
  \[ E[X] = \sum_x xP(x) \]

- Analogously for a continuous random variable:
  \[ E[X] = \int_{-\infty}^{\infty} xf(x) \, dx \]

- Note: If \( X \) always between \( a \) and \( b \) then so is \( E[X] \)
  - More formally:
    \[ \text{if } P(a \leq X \leq b) = 1 \text{ then } a \leq E[X] \leq b \]

Generalizing Expectation

- Let \( g(X, Y) \) be real-valued function of two variables

\[ E[g(X, Y)] = \sum_{x,y} g(x,y)P_{X,Y}(x,y) \]

- Let \( X \) and \( Y \) be discrete jointly distributed RVs:
  \[ E[g(X,Y)] = \sum_{x,y} g(x,y)p_{X,Y}(x,y) \]

- Analogously for continuous random variables:
  \[ E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y) \, dx \, dy \]

Expected Values of Sums

- Let \( g(X, Y) = X + Y \). Compute \( E[g(X, Y)] = E[X + Y] \)

\[ E[X + Y] = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} f_{X,Y}(x,y) \, dx \, dy \]

\[ = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{\infty} f_{X}(x) \, dx \, dy + \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} f_{Y}(y) \, dx \, dy \]

\[ = E[X] + E[Y] \]

- Generalized: \( E\left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} E[X_i] \)
  - Holds regardless of dependency between \( X_i \)'s

Tie Me Up! : Bounding Expectation

- If random variable \( X \geq a \) then \( E[X] \geq a \)
  \[ \text{if } P(a \leq X \leq \infty) = 1 \text{ then } a \leq E[X] \leq \infty \]
  - Often useful in cases where \( a = 0 \)
  - But, \( E[X] \geq a \) does not imply \( X \geq a \) for all \( X = x \)
    - E.g., \( x \) is equally likely to take on values -1 or 3. \( E[X] = 1 \).
  - If random variables \( X \geq Y \) then \( E[X] \geq E[Y] \)
    - \( x \geq y \implies X - Y \geq 0 \implies E[X - Y] \geq 0 \)
    - Substituting: \( E[X] - E[Y] \geq 0 \implies E[X] \geq E[Y] \)
    - But, \( E[X] \geq E[Y] \) does not imply \( X \geq Y \) for all \( X = x, Y = y \)

Sample Mean

- Consider \( n \) random variables \( X_1, X_2, ..., X_n \)
  - \( X_i \) are all independently and identically distributed (I.I.D.)
  - Have same distribution function \( F \) and \( E[X] = \mu \)
  - We call sequence of \( X_i \) a sample from distribution \( F \)
  - Sample mean:
    \[ \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \]
    - Compute \( E[\bar{X}] \)
    \[ E[\bar{X}] = E\left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \frac{1}{n} n \mu = \mu \]
    - \( \bar{X} \) is “unbiased” estimate of \( \mu \) (\( E[\bar{X}] = \mu \))

Boole was so Cool!

- Let \( E_1, E_2, ..., E_n \) be events with indicator RVs \( X_i \)
  - if event \( E_i \) occurs, then \( X_i = 1 \), else \( X_i = 0 \)
  - Recall \( E[X] = P(E) \)
  - Now, let \( X = \sum_{i=1}^{n} X_i \) and let \( Y = 1 \) if \( X \geq 1, 0 \) otherwise
  - Note: \( X \geq Y \implies E[X] \geq E[Y] \)
  - \[ E[X] = E\left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} P(E_i) \]
  - \[ E[Y] = P(\text{at least one event } E_i \text{ occurs}) = P\left( \bigcup_{i=1}^{n} E_i \right) \]
  - Boole’s inequality: \( \sum_{i=1}^{n} P(E_i) \geq P\left( \bigcup_{i=1}^{n} E_i \right) \)
    - Boole died from being too cool (literally)!
Expected of (Negative) Binomial

- Let \( Y \sim \text{Bin}(n, p) \)
- \( n \) independent trials
- Let \( X_i = 1 \) if \( i \)-th trial is “success”, 0 otherwise
- \( X_i \sim \text{Ber}(p) \)
- \( E[X_i] = p \) (\( = 1p + 0(1 - p) \))
- \( E[Y] = E[X_1] + E[X_2] + \ldots + E[X_n] = np \)

- Let \( Y \sim \text{NegBin}(r, p) \)
- Recall \( Y \) is number of trials until \( r \) “successes”
- Let \( X_i \) = # of trials to get success after \((i - 1)\)th success
- \( X_i \sim \text{Geo}(p) \) (i.e., Geometric RV) \( E[X_i] = 1/p \)
- \( E[Y] = E[X_1] + E[X_2] + \ldots + E[X_r] = r/p \)

Hash Tables (a.k.a. Coupon Collecting)

- Consider a hash table with \( n \) buckets
- Each string equally likely to get hashed into any bucket
- Let \( X = \# \) strings to hash until each bucket \( \geq 1 \) string
- What is \( E[X] \)?
- \( E[X] = E[X_0] + E[X_1] + \ldots + E[X_{n-1}] \)

Course Mean

\[ E[CS109] \]

This is actual midpoint of course
(Just wanted you to know)

QuickSort

5 3 7 4 8 6 2 1

select “pivot”

Recursive Insight

5 3 7 4 8 6 2 1

Partition array so:
- everything smaller than pivot is on left
- everything greater than or equal to pivot is on right
- pivot is in-between
Partition array so:
• everything smaller than pivot is on left
• everything greater than or equal to pivot is on right
• pivot is in-between

Now recursive sort “red” sub-array

Now recursive sort “red” sub-array
Then, recursive sort “blue” sub-array

Everything is sorted!
Complexity QuickSort

- QuickSort is $O(n \log n)$, where $n$ is # elements to sort.
- But in “worst case” it can be $O(n^2)$.
- Worst case occurs when every time pivot is selected, it is maximal or minimal remaining element.
- What is $P($QuickSort worst case$)$?
  - On each recursive call, pivot = max/min element, so we are left with $n-1$ elements for next recursive call.
  - 2 possible "bad" pivots (max/min) on each recursive call.
  - $P($Worst case$) = \frac{2}{n} \sum_{i=1}^{n-1} \frac{1}{2} \frac{2^{n-2}}{n}!$.
  - Saw similar behavior for BSTs on problem set #1.

- Complexity of algorithm determined by number of comparisons made to pivot.

Expected Running Time of QuickSort

- Let $X = \#$ comparisons made when sorting $n$ elements.
- $E[X]$ gives us expected running time of algorithm.
- Given $X_1, X_2, ..., X_n$ in random order to sort.
- Let $Y_1, Y_2, ..., Y_n$ be $X_1, X_2, ..., X_n$ in sorted order.
- Let $L_0 = 1$ if $Y_0$ and $Y_1$ are compared, 0 otherwise.
- Order where $Y_{a+1} > Y_a$ so we have: $X = \sum_{a=1}^{n-1} L_a.$

$$E[X] = E\left[\sum_{a=1}^{n-1} L_a\right] = \sum_{a=1}^{n-1} E[L_a] = \sum_{a=1}^{n-1} P(Y_a \text{ and } Y_{a+1} \text{ ever compared})$$

Determining $P(Y_a \text{ and } Y_{a+1} \text{ ever compared})$

- Consider when $Y_a$ and $Y_{a+1}$ are directly compared.
  - If pivot chosen is $< Y_a$ or $> Y_a$, then $Y_a$ and $Y_{a+1}$ are not directly compared (since values only compared to pivot).
  - So, we only care about case where pivot chosen from set: {$Y_a, Y_{a+1}, Y_{a+2}, ..., Y_{n}$}.
  - From that set either $Y_a$ and $Y_{a+1}$ must be selected as pivot (with equal probability) in order to be compared.
  - So, $P(Y_a \text{ and } Y_{a+1} \text{ ever compared}) = \frac{2}{n-a+1}.

$$E[X] = \sum_{a=1}^{n-1} \sum_{b=a+1}^{n} P(Y_a \text{ and } Y_{b-a+1} \text{ ever compared}) = \sum_{a=1}^{n-1} \sum_{b=a+1}^{n} \frac{2}{n-a+1}$$

Bring it on Home... (i.e., Solve the Sum)

$$E[X] = \sum_{a=1}^{n-1} \sum_{b=a+1}^{n} \frac{2}{n-a+1} = \int_{2}^{n} \frac{2}{b-a+1} db$$

Recall: $\int_{a}^{b} \frac{1}{x} dx = \ln(x)$

$$= \ln(a) - \ln(b) \bigg|_{a=1}^{b=n} = 2\ln(n-a+1) - 2\ln(2)$$

For large $n$:

$$E[X] = \sum_{a=1}^{n-1} \ln(n-a+1) da \quad \text{Let } y = n-a+1$$

$$= \int_{1}^{n} \frac{1}{y} \ln(y) dy \quad \text{Integration by parts:}$$

$$= -2\ln(n-a+1) da$$

$$= -2\ln(n-a+1) \bigg|_{a=1}^{b=n} = 2a\ln(n-a+1) - 2n = O(n \log n)$$