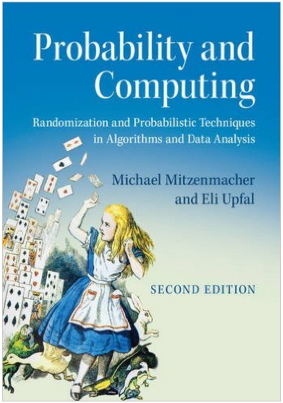
CS155/254: Probabilistic Methods in Computer Science

Chapters 2 & 3



Randome Variables and Expectation
Example: QuickSort

Procedure Q_S(S);
Input: An array S.
Output: The array S in sorted order.

1. Choose a random element y uniformly from S.
2. Compare all elements of S to y. Let

$$S_1 = \{x \in S - \{y\} \mid x \leq y\}, \qquad S_2 = \{x \in S - \{y\} \mid x > y\}.$$

3. Return the list:

$$Q\_S(S_1), y, Q\_S(S_2).$$

Let $T(n) =$ number of comparisons in a run of QuickSort on an array of size n.

$T(n)$ is a *random variable*.

Theorem

The expected number of steps in sorting an array of n elements using QuickSort is

$$E[T(n)] \; = \; O(n \log n).$$

Random Variable

Definition

A random variable X on a sample space Ω is a real-valued function on Ω; that is, $X : \Omega \to \mathcal{R}$
A vector random variable is $X^d : \Omega \to \mathcal{R}^d$
A discrete random variable is a random variable that takes on only a finite or countably infinite number of values.

Discrete random variable X and real value a: the event "$X = a$" represents the set $\{s \in \Omega : X(s) = a\}$.

$$\Pr(X = a) = Pr(\{s \in \Omega : X(s) = a\}) = \sum_{s \in \Omega : X(s) = a} \Pr(s)$$

Independence

Definition

Two events A and B are independent if and only if

$$\Pr(A \cap B) \;=\; \Pr(A) \cdot \Pr(B)$$

Two random variables X and Y are independent if and only if

$$\Pr((X = x) \cap (Y = y)) \;=\; \Pr(X = x) \cdot \Pr(Y = y)$$

for all values x and y. Similarly, random variables $X_1, X_2, \ldots X_k$ are mutually independent if and only if for any subset $I \subseteq [1, k]$ and any values x_i,$i \in I$,

$$\Pr\left(\bigcap_{i \in I} X_i = x_i\right) \;=\; \prod_{i \in I} \Pr(X_i = x_i).$$

Expectation

Definition

The expectation of a discrete random variable X, denoted by $\mathbf{E}[X]$, is given by

$$\mathbf{E}[X] = \sum_i i \Pr(X = i),$$

where the summation is over all values in the range of X. The expectation exists iff $\sum_i |i| \Pr(X = i)$ converges; otherwise, the expectation is unbounded or doesn't exist.

The expectation (or mean or average) is a weighted sum over all possible values of the random variable.

A rational agent will pay no more than the expected outcome to participate in a game.

Median

Definition

The **median** of a random variable X is a value m such

$$Pr(X < m) \leq 1/2 \quad \text{and} \quad Pr(X > m) < 1/2.$$

Quicksort

Procedure Q_S(S);

Input: An array S.

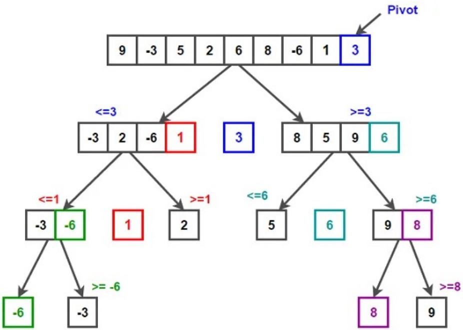
Output: The array S in sorted order.

1. Choose a random element y uniformly from S.
2. Compare all elements of S to y. Let
$$S_1 = \{x \in S - \{y\} \mid x \leq y\}, \qquad S_2 = \{x \in S - \{y\} \mid x > y\}.$$
3. Return the list:
$$Q\_S(S_1), y, Q\_S(S_2).$$

Theorem

The expected number of steps in sorting an array of n elements using QuickSort is

$$E[T(n)] = O(n \log n).$$



https://medium.com/@nathaldawson/unraveling-quicksort-the-fast-and-versatile-sorting-algorithm-2c1214755ce9

Proof:

Let $s_1, ..., s_n$ be the elements of S is sorted order.
For $i = 1, ..., n$, and $j > i$, define 0-1 random variable $X_{i,j}$, s.t.
$X_{i,j} = 1$ iff s_i is directly compared to s_j in the run of the algorithm, else $X_{i,j} = 0$.
The number of comparisons in running the algorithm is

$$T(n) = \sum_{i=1}^{n} \sum_{j>i} X_{i,j}.$$

We are interested in

$$E[T(n)] = E[\sum_{i=1}^{n} \sum_{j>i} X_{i,j}] = \sum_{i=1}^{n} \sum_{j>i} E[X_{i,j}].$$

Linearity of Expectation

Theorem

For any two random variables X *and* Y

$$E[X+Y] \;=\; E[X] \;+\; E[Y].$$

Lemma

For any constant c *and discrete random variable* X,

$$\mathbf{E}[cX] = c\mathbf{E}[X].$$

Linearity of Expectation

$D(x) =$ domain of random variable X.

$$\begin{aligned}
E[X+Y] &= \sum_{x \in D(X)} \sum_{y \in D(Y)} (x+y) Pr(X = x \cap Y = y) \\
&= \sum_{x \in D(X)} x \sum_{y \in D(Y)} Pr(X = x \cap Y = y) + \\
&\quad \sum_{y \in D(Y)} y \sum_{x \in D(X)} Pr(X = x \cap Y = y) \\
&= \sum_{x \in D(X)} x Pr(X = x) + \sum_{y \in D(Y)} y Pr(Y = y) \\
&= E[X] + E[Y]
\end{aligned}$$

We are interested in $E[T(n)] = \sum_{i=1}^{n} \sum_{j>i} E[X_{i,j}]$.

Since $X_{i,j}$ is a 0-1 random variable,

$$E[X_{i,j}] = 0 \cdot Pr(X_{i,j} = 0) + 1 \cdot Pr(X_{i,j} = 1) = Pr(X_{i,j} = 1).$$

What is the probability that $X_{i,j} = 1$?

s_i is compared to s_j iff either s_i or s_j is chosen as a "split item" before any of the $j - i - 1$ elements between s_i and s_j are chosen.

Elements are chosen uniformly at random $\rightarrow$ elements in the set $[s_i, s_{i+1}, ..., s_j]$ are chosen uniformly at random.

$X_{i,j} = 1$ iff the first split item chosen in the set $\{s_i, s_{i+1}, ..., s_j\}$ is either s_i or s_j.

$$E[X_{i,j}] = Pr(X_{i,j} = 1) = \frac{2}{j - i + 1}.$$

$$
\begin{aligned}
E[T] &= E[\sum_{i=1}^{n} \sum_{j>i} X_{i,j}] = \sum_{i=1}^{n} \sum_{j>i} E[X_{i,j}] = \sum_{i=1}^{n} \sum_{j>i} \frac{2}{j-i+1} \\
&\leq n \sum_{k=1}^{n} \frac{2}{k} \leq 2nH_n = 2n \log n + O(n)
\end{aligned}
$$

$$
H_n = \sum_{i=1}^{n} \frac{1}{i} \approx \int_1^n \frac{1}{x} dx = \log n
$$

$$
\log n \leq H_n \leq \log n + 1
$$

Theorem

The expected number of steps in sorting an array of n elements using QuickSort is $E[T(n)] = O(n \log n)$.

A Deterministic QuickSort

Procedure DQ_S(S);
Input: A set S.
Output: The set S in sorted order.

1. Let y be the first element in S.
2. Compare all elements of S to y. Let

   $$S_1 = \{x \in S - \{y\} \mid x \leq y\}, \qquad S_2 = \{x \in S - \{y\} \mid x > y\}.$$

   (Elements is S_1 and S_2 are in th same order as in S.)
3. Return the list:

   $$DQ\_S(S_1), y, DQ\_S(S_2).$$

Probabilistic Analysis of QuickSort

Theorem

The expected run time of DQ_S on a random input, uniformly chosen from all possible permutation of S is $O(n \log n)$.

Proof.

Set $X_{i,j}$ as before.
If all permutations have equal probability, all permutations of $S_i, \ldots, S_j$ have equal probability, thus

$$Pr(X_{i,j}) = \frac{2}{j-i+1}.$$

$$E[\sum_{i=1}^{n} \sum_{j>i} X_{i,j}] = O(n \log n).$$

□

Randomized Algorithms:

- Analysis is true for **any** input.
- The sample space is the space of random choices made by the algorithm.
- Repeated runs are independent.

Probabilistic Analysis:

- The sample space is the space of all possible inputs.
- If the algorithm is **deterministic** repeated runs give the same output.

Algorithm classification

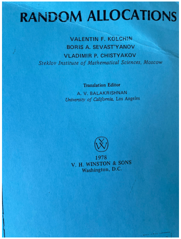
A **Monte Carlo Algorithm** is a randomized algorithm that may produce an incorrect solution.
For decision problems: A **one-side error** Monte Carlo algorithm errs only one one possible output, otherwise it is a **two-side error** algorithm.
A **Las Vegas** algorithm is a randomized algorithm that **always** produces the correct output.
In both types of algorithms the run-time is a random variable.

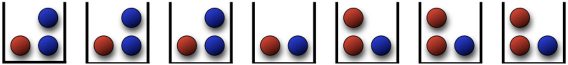
Balls and Bins and the Random Allocations Paradigm

Random Allocations

m marbles/balls/items
are placed independently and
uniformly at random into
n urns/bins/boxes

CS applications: hashing, load balancing, routing, distributed memory, lower bounds,…





https://dornsife.usc.edu/assets/sites/406/docs/505b/polya.urn.pdf

The Coupon Collector Problem

- We place balls independently and uniformly at random in n boxes.
- Let X be the number of balls placed until no box is empty..

We'll prove:

Theorem

1. $E[X] = nH_n = n \log n + \Theta(n)$
2. $Pr(X \geq 2H_n) = O(\frac{1}{n})$.

Application: How many trials are needed to verify encrypted membership?

- We place balls independently and uniformly at random in n boxes.
- Let X be the number of balls placed until no box is empty.
- Let $X_i =$ number of balls placed when there were exactly $i-1$ non-empty boxes.
- $X = \sum_{i=1}^{n} X_i$.
- X_i is the number of trials till we hit an empty box when $i-1$ of the n boxes are not empty.
- X_i is a geometric random variable with parameter $p_i = 1 - \frac{i-1}{n}$.

The Geometric Distribution

Definition

A geometric random variable X with parameter p is defined by the probability distribution on $n = 1, 2, \ldots$.

$$\Pr(X = n) = (1 - p)^{n-1} p.$$

Example: repeatedly draw independent Bernoulli random variables with parameter $p > 0$ until we get a 1. Let X be number of trials up to and including the first one. Then X is a geometric random variable with parameter p.

Memoryless Distribution

Lemma

For a geometric random variable with parameter p and $n > 0$,

$$\Pr(X = n + k \mid X > k) \quad = \quad \Pr(X = n).$$

Proof.

$$\begin{aligned}
\Pr(X = n + k \mid X > k) &= \frac{\Pr((X = n + k) \cap (X > k))}{\Pr(X > k)} \\
&= \frac{\Pr(X = n + k)}{\Pr(X > k)} = \frac{(1-p)^{n+k-1} p}{\sum_{i=k}^{\infty} (1-p)^i p} \\
&= \frac{(1-p)^{n+k-1} p}{(1-p)^k} = (1-p)^{n-1} \\
&= \Pr(X = n).
\end{aligned}$$

Conditional Expectation

Definition

$$\mathbf{E}[Y \mid Z = z] = \sum_y y \Pr(Y = y \mid Z = z),$$

where the summation is over all y in the range of Y.

Lemma

For any random variables X and Y,

$$\mathbf{E}[X] = E_y[E_X[X \mid Y]] = \sum_y \Pr(Y = y) E[X \mid Y = y],$$

where the sum is over all values in the range of Y.

Geometric Random Variable: Expectation

- Let X be a geometric random variable with parameter p.
- Let $Y = 1$ if the first trail is a success, $Y = 0$ otherwise.
-

$$\begin{aligned} \mathbf{E}[X] &= \Pr(Y = 0)\mathbf{E}[X \mid Y = 0] + \Pr(Y = 1)\mathbf{E}[X \mid Y = 1] \\ &= (1 - p)\mathbf{E}[X \mid Y = 0] + p\mathbf{E}[X \mid Y = 1]. \end{aligned}$$

- If $Y = 0$ let Z be the number of trials after the first one.
- $\mathbf{E}[X] = (1 - p)\mathbf{E}[Z + 1] + p \cdot 1 = (1 - p)\mathbf{E}[Z] + 1$
- But $\mathbf{E}[Z] = \mathbf{E}[X]$, giving

$$\mathbf{E}[X] = 1/p.$$

Lemma

Let X be a discrete random variable that takes on only non-negative integer values. Then

$$\mathbf{E}[X] = \sum_{i=1}^{\infty} \Pr(X \geq i).$$

Proof.

$$\begin{aligned}
\sum_{i=1}^{\infty} \Pr(X \geq i) &= \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} \Pr(X = j) \\
&= \sum_{j=1}^{\infty} \sum_{i=1}^{j} \Pr(X = j) \\
&= \sum_{j=1}^{\infty} j \Pr(X = j) = \mathbf{E}[X].
\end{aligned}$$

For a geometric random variable X with parameter p,

$$\Pr(X \geq i) = \sum_{n=i}^{\infty} (1-p)^{n-1} p = (1-p)^{i-1}.$$

$$\begin{aligned} \mathbf{E}[X] &= \sum_{i=1}^{\infty} \Pr(X \geq i) \\ &= \sum_{i=1}^{\infty} (1-p)^{i-1} \\ &= \frac{1}{1-(1-p)} \\ &= \frac{1}{p} \end{aligned}$$

Back to the Coupon Collector Problem

- Let $X_i =$ number of balls placed when there were exactly $i - 1$ non-empty boxes.
- $X = \sum_{i=1}^{n} X_i$.
- X_i is a geometric random variable with parameter $p_i = 1 - \frac{i-1}{n}$.

$$\mathbf{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}.$$

$$\begin{aligned}\mathbf{E}[X] &= E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{E}[X_i] \\ &= \sum_{i=1}^{n} \frac{n}{n-i+1} = n\sum_{i=1}^{n} \frac{1}{i} = nH_n = n \ln n + \Theta(n).\end{aligned}$$

Let $X(\alpha)$ be the number of balls placed till αn boxes are not empty:

$$\mathbf{E}[X(\alpha)] = \sum_{i=1}^{\alpha n} \frac{n}{n-i+1} \approx \sum_{i=(1-\alpha)n}^{n} \frac{n}{i} = n(\sum_{i=1}^{n} \frac{1}{i} - \sum_{i=1}^{(1-\alpha)n} \frac{1}{i}) = n \ln \frac{1}{1-\alpha}$$

Bounding Deviation from Expectation

Theorem

[Markov Inequality] *For any non-negative random variable*

$$Pr(X \geq a) \leq \frac{E[X]}{a}.$$

Proof.

$$E[X] = \sum iPr(X = i) \geq a \sum_{i \geq a} Pr(X = i) = aPr(X \geq a).$$

$\square$

Back to the Coupon Collector's Problem

- We place balls independently and uniformly at random in n boxes.
- Let X be the number of balls placed until all boxes are not empty.
- $E[X] = nH_n = n \ln n + \Theta(n)$
- What is $\Pr(X \geq 2\mathbf{E}[X])$?
- Applying Markov's inequality

$$\Pr(X \geq 2nH_n) \leq \frac{1}{2}.$$

- Can we do better?

Variance

Definition

The **variance** of a random variable X is

$$Var[X] \; = \; \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

Definition

The **standard deviation** of a random variable X is

$$\sigma(X) = \sqrt{Var[X]}.$$

Example: Let X be a 0-1 random variable with $Pr(X = 0) = Pr(X = 1) = 1/2$.

$$E[X] = 1/2.$$

$$Var[X] = \frac{1}{2}(1 - \frac{1}{2})^2 + \frac{1}{2}(0 - \frac{1}{2})^2 = \frac{1}{4}.$$

Chebyshev's Inequality

Theorem

For **any** *random variable*

$$Pr(|X - E[X]| \geq a) \leq \frac{Var[X]}{a^2}.$$

Proof.

$$Pr(|X - E[X]| \geq a) = Pr((X - E[X])^2 \geq a^2)$$

By Markov inequality

$$Pr((X - E[X])^2 \geq a^2) \leq \frac{E[(X - E[X])^2]}{a^2}$$

$$= \frac{Var[X]}{a^2}$$

Theorem

For **any** *random variable*

$$Pr(|X - E[X]| \geq a\sigma[X]) \leq \frac{1}{a^2}.$$

Theorem

For **any** *random variable*

$$Pr(|X - E[X]| \geq \epsilon E[X]) \leq \frac{Var[X]}{\epsilon^2 (E[X])^2}.$$

Theorem

If X and Y are independent random variable

$$E[XY] \;=\; E[X] \cdot E[Y],$$

Proof.

$$E[XY] = \sum_i \sum_j i \cdot jPr((X = i) \cap (Y = j)) \;=$$

$$\sum_i \sum_j ijPr(X = i) \cdot Pr(Y = j) =$$

$$(\sum_i iPr(X = i))(\sum_j jPr(Y = j)).$$

$\square$

Theorem

If X *and* Y *are independent random variable*

$$Var[X+Y] = Var[X] \ + \ Var[Y].$$

Proof.

$$Var[X+Y] = E[(X+Y-E[X]-E[Y])^2] =$$

$$E[(X-E[X])^2 + (Y-E[Y])^2 + 2(X-E[X])(Y-E[Y])] =$$

$$Var[X] + Var[Y] + 2E[X-E[X]]E[Y-E[Y]]$$

Since the random variables $X - E[X]$ and $Y - E[Y]$ are independent.
But $E[X-E[X]] = E[X] - E[X] = 0.$ □

Variance of a Geometric Random Variable

- We use
$$Var[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

- To compute $\mathbf{E}[X^2]$, let $Y = 1$ if the first trail is a success, $Y = 0$ otherwise.
-
$$\begin{aligned}\mathbf{E}[X^2] &= \Pr(Y = 0)\mathbf{E}[X^2 \mid Y = 0] + \Pr(Y = 1)\mathbf{E}[X^2 \mid Y = 1] \\ &= (1 - p)\mathbf{E}[X^2 \mid Y = 0] + p\mathbf{E}[X^2 \mid Y = 1].\end{aligned}$$

- If $Y = 0$ let Z be the number of trials after the first one.
-
$$\begin{aligned}\mathbf{E}[X^2] &= (1 - p)\mathbf{E}[(Z + 1)^2] + p \cdot 1 \\ &= (1 - p)\mathbf{E}[Z^2] + 2(1 - p)\mathbf{E}[Z] + (1 - p) + p,\end{aligned}$$

- $\mathbf{E}[Z] = 1/p$ and $\mathbf{E}[Z^2] = \mathbf{E}[X^2]$.

$$\begin{aligned}\mathbf{E}[X^2] &= (1-p)\mathbf{E}[(Z+1)^2] + p \cdot 1 \\ &= (1-p)\mathbf{E}[Z^2] + 2(1-p)\mathbf{E}[Z] + 1,\end{aligned}$$

$$\mathbf{E}[X^2] = (1-p)\mathbf{E}[X^2] + 2(1-p)\frac{1}{p} + 1 = (1-p)\mathbf{E}[X^2] + (2-p)/p,$$

- $\mathbf{E}[X^2] = (2-p)/p^2$.

$$Var[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

Back to the Coupon Collector's Problem

- We place balls independently and uniformly at random in n boxes.
- Let X be the number of balls placed until all boxes are not empty.
- $E[X] = nH_n = n \ln n + \Theta(n)$
- What is $\Pr(X \geq 2\mathbf{E}[X])$?
- Applying Markov's inequality

$$\Pr(X \geq 2nH_n) \leq \frac{1}{2}.$$

- Can we do better?

- Let $X_i =$ number of balls placed when there were exactly $i-1$ non-empty boxes.
- $X = \sum_{i=1}^{n} X_i$.
- X_i is a geometric random variable with parameter $p_i = 1 - \frac{i-1}{n}$.
- $Var[X_i] \leq \frac{1}{p^2} \leq (\frac{n}{n-i+1})^2$.
- $$Var[X] = \sum_{i=1}^{n} Var[X_i] \leq \sum_{i=1}^{n} \left(\frac{n}{n-i+1}\right)^2 = n^2 \sum_{i=1}^{n} \left(\frac{1}{i}\right)^2 \leq \frac{\pi^2 n^2}{6}.$$
- We used $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$.

- Let X_i = number of balls placed when there were exactly $i-1$ non-empty boxes.
- $X = \sum_{i=1}^{n} X_i$.
- X_i is a geometric random variable with parameter $p_i = 1 - \frac{i-1}{n}$.
- $Var[X_i] \le \frac{1}{p^2} \le (\frac{n}{n-i+1})^2$.
- $$Var[X] = \sum_{i=1}^{n} Var[X_i] \le \sum_{i=1}^{n} \left(\frac{n}{n-i+1}\right)^2 = n^2 \sum_{i=1}^{n} \left(\frac{1}{i}\right)^2 \le \frac{\pi^2 n^2}{6}.$$
- By Chebyshev's inequality
$$\Pr(|X - nH_n| \ge nH_n) \le \frac{n^2\pi^2/6}{(nH_n)^2} = \frac{\pi^2}{6(H_n)^2} = O\left(\frac{1}{\ln^2 n}\right).$$
- Can we do better?

Direct Bound

- The probability of not obtaining the i-th coupon after $n \ln n + cn$ steps:

$$\left(1 - \frac{1}{n}\right)^{n(\ln n + c)} < \mathrm{e}^{-(\ln n + c)} = \frac{1}{\mathrm{e}^{c} n}.$$

- By a union bound, the probability that some coupon has not been collected after $n \ln n + cn$ step is e^{-c}.
- The probability that all coupons are not collected after $2n \ln n$ steps is at most $1/n$.
- The probability that all coupons are not collected after $n(\ln n + 2 \log \log n)$ steps is $O\left(\frac{1}{\ln^2 n}\right)$.

Union Bound

Let $E_1, E_2, \ldots E_n$ be **any** n events in a probability space $(\Omega, \mathcal{F}, Pr)$,

$$Pr(\cup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} Pr(E_i).$$