

CSC120 Algorithms and Data Structures

Prof. Kamberova

Overview of Probability for Average Case Analysis of Algorithms

Terms we will use

- Sample space, probability measure
- Random variables, cumulative distribution functions (CDF)
- Discrete random variables, discrete distributions and point mass functions (PMF)
- Expected value (*mean*) of a random variable
- Variance

Average case analysis

- Typically, assume uniformly distribution of the input (equally likely)
- The run time is considered as a random variable
- Compute the expected (average) run time under those assumptions, and its complexity

Sample space set of all possible outcomes of an experiment

Ex:

1. Rolling a die: $S = \{1, 2, 3, 4, 5, 6\}$
2. Flipping a coin: $S = \{H, T\}$
3. Flipping a coin twice in order: $S = \{HH, HT, TH, TT\}$

Event: a collection of outcomes (a subset of S).

For example, “rolling even number” is an event in rolling a die.

Probabilities: specify the likelihoods of different events

Probability distribution (function) is an assignment of weights to possible outcomes in S .

For example, we may assume that in a coin flip example, H and T are equally likely.

Definition For given S , probability measure P is a function $P : S \rightarrow [0, 1]$ which satisfies the following three axioms:

- $P(s) \geq 0$, for all $s \in S$; nonnegative.
- $\sum_{s \in S} P(s) = 1$.

- $P(\cup_{i=1}^{\infty} A_i : A_i \cup A_j = \delta_{ij}) = \sum_{i=1}^{\infty} P(A_i)$; countably additive, i.e. the probability of the union of at most countably many mutually disjoint events is sum the individual probabilities of the events.

For a sample space S , $|S|$ denotes the number of elements in S .

Ex: Given a finite sample space S . If all outcomes are equally likely, then the so called *Uniform probability* assigns to each element of S the same weight, $1/|S|$.

Random variable Given S and P , we may talk about random variables. These are real-valued functions defined on the sample space. There are some special restrictions, but since all functions we consider satisfy them, we'll not focus on that.

X is a random variable, i.e., X is a function $X : S \rightarrow R$.

We will be mostly interested in discrete random variables which take on positive integer values only

$$X : S \rightarrow N \cup \{0\}$$

Random variable are characterized by their *cumulative distribution functions*(CDF). The CDF, F_X described the likelihood that the random variable takes certain values, $F_X : S \rightarrow [0, 1]$,

$$F_X(x) \stackrel{\text{def}}{=} P[X = x] \stackrel{\text{def}}{=} P(X^{-1}(x))$$

Random variables X and Y are independent if F

$$P[X = x \text{ and } Y = y] = P[X = x]P[Y = y]$$

That is, knowing the value of one of the two RVs does not tell us anything about the value of the other.

Mean (Expectation) of a RV.

Given (S, P, F_X) , the mean or expected value μ_X of X is

$$\mu_X = E[X] = \sum_{x \in X(S)} x F_X(x) = \sum_{x \in X(S)} x P[X = x]$$

Variance of a RV. Given (S, P, F_X) , the variance σ_X^2 is

$$\sigma_X^2 = Var[X] = E[X - E[X]]^2 = E[X^2] - (E[X])^2$$

Standard deviation: $\sigma_X = \sqrt{Var[X]}$

The standard divination characterizes the dispersion around the mean of the values X takes.

Facts:

$$E[aX_1 + X_2] = aE[X_1] + E[X_2]$$

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

If X and Y are independent RVs, then

$$E[XY] = E[X]E[Y]$$

$$\text{Var}[XY] = \text{Var}[X]\text{Var}[Y]$$

Rolling a 6 sided die

$$S = \{1, 2, 3, 4, 5, 6\}$$

P the uniform probability measure, $P(i) = 1/6$

X a random variable $X : S \rightarrow \{0, 1\}$,

X is 0 if the number rolled is even, and 1 otherwise

$$P[X = 0] = P(\text{outcome is even}) = P(\{2, 4, 6\}) = P(2) + P(4) + P(6) = 3/6 = 1/2$$

$$P[X = 1] = P(\text{outcome is odd}) = P(\{1, 3, 5\}) = P(1) + P(3) + P(5) = 3/6 = 1/2$$

$$E[X] = 0 \cdot P[X = 0] + 1 \cdot P[X = 1] = 1/2$$

Lets do another random variable for the same (S, P) . Y denotes the outcome of one roll, i.e. $Y(k) = k$. In this case

$$F_Y(y) = P[Y = y] = P(Y^{-1}(y)) = 1/6, \text{ for } y = 1, \dots, 6$$

$$E[Y] = \sum_{y=1}^6 yP[Y = y] = (1/6) * \sum_{y=1}^6 y = (1/6) * 6 * 7/2 = 3.5$$

Average case analysis of algorithms, expected run time

- Example: linear search in an array A of size n .
- Assumptions: we assume that it is equally likely for the key to appear at any position, or not at all.
- Sample space, $S = \{1, 2, 3, \dots, n\} \cup \{0\}$
- Probability measure on S , $P(i) = 1/(n + 1)$, $i = 0, 1, 2, \dots, n$
- Random variable, X taking as values the positions at which the key could be found, or 0 if the key is not there
- The probability distribution of X , uniform, i.e.,
 $P[X = i] = 1/(n + 1)$, $i = 0, 1, 2, \dots, n$
- Worst case asymptotic time complexity:
 $T_{worst}(n) = \Theta(n)$ - time to look when the key is not there

- The average case run time analysis:

Often we will use simply $T_{avg}(n)$ or $T(n)$ to denote $E[T(n)]$, i.e the expected run time on an input of size n

$$T_{avg}(n) = \sum_{i=1}^n iP[key\ found\ at\ pos.\ i] + 0P[key\ not\ in] \quad T_{avg}(n) = \sum_{i=0}^n iP[X = i] = (1/(n+1)) * n * (n+1)/2 = n/2,$$

i.e., if we make many repeated searches for different key values (all equally likely as stated above), we search, on average, half of the array to get the answer.

$$T_{avg}(n) = \Theta(n).$$

We say that linear search has expected (average) asymptotic time complexity, $\Theta(n)$.