CS 761: Randomized Algorithms Fall 2019

Lecture 1 — September 6, 2019

Prof. Gautam Kamath *By: Frédéric Bouchard, Patrick Li and Pranav Subramani*
Edited by Vedat Levi Alev

Disclaimer: These notes have not been subject to the usual scrutiny reserved for formal publications.

1 Overview of Probability

Before discussing randomized algorithms, we introduce the probability concepts necessary for this course. Rather than observing a single deterministic *event*, probability distributions simultaneously model all possible *events* of a random occurrence.

A random variable is a function that allows us to map *events* to another set. In this course, random variables will map *events* to the real number line. Using random variables, we can perform mathematical operations to infer properties about a random occurrence from a given distribution.

Definition 1 (Random Variable)**.** *A function from the set of events, $\mathcal{E}$ to the real numbers is a random variable, $X : \mathcal{E} \to \mathbb{R}$.*

$X \sim D$ means that X is sampled from distribution D.

Example 1.1. *A fair dice roll where $S = \{1, 2, 3, 4, 5, 6\}$, is written as $X \sim$ Uniform$([6])$, where each outcome is equiprobable.*

Example 1.2. *A random variable X that follows a Bernoulli distribution is written as $X \sim$ Bernoulli(p) which means that $\Pr[X = 1] = p$ (i.e., success) and $\Pr[X = 0] = 1 - p$ (i.e., failure).*

1.0.1 Probability Mass Function (PMF)

Probability mass functions (PMFs) are defined for discrete random variables. PMFs indicate the probability that the outcome of a random variable will equal a given value. A formal definition is provided below.

Definition 2. *The probability mass function of a discrete random variable ($\mathcal{E}$ is discrete) defined as $f_X(x) = \Pr[X = x] = \Pr[\{e \in \mathcal{E} : X(e) = x\}]$*

Remark 3. *Note that PMFs are constrained such that $\sum_{x \in X(e)} f_X(x) = 1$.*

Example 1.3. *In the example where $X \sim$ Bernoulli(p), $f_X(1) = p$ and $f_X(0) = 1-p$. Here, p and $1 - p$ indicate the probability that the random variable X **realizes** the values 1 and 0 respectively.*

1.0.2 Cumulative Distribution Function (CDF)

The cumulative distribution function (CDF) is defined for any distribution over a totally ordered set, and can often be computed using the PMF. The CDF of a random variable X looks to compute the probability that a realization of X will be below a given value.

Definition 4. *The cumulative distribution function of a* $F_X(x) = \Pr[X \leq x_j] = \sum_{i \leq j} \Pr[X = x_i] = \sum_{i \leq j} f_X(i)$.

1.0.3 Probability Density Function (PDF)

Definition 5. *The probability density function is defined as* $\frac{d}{dx} F_X(x) = f_X(x)$.

The PDF is often viewed as the continuous analogue of the PMF, since the notion of mass functions over continuous sample spaces does not make sense. Explicitly, if X is a continuous random variable, $\Pr(X = x) = 0$ for all $x \in \mathbb{R}$.

Remark 6. *This equation has the following discrete equivalence:* $F_X(x) - F_X(x-1) = f_X(x)$.

1.0.4 Binomial Distribution

The binomial distribution is one of the most common discrete probability distributions. With parameters n, and p, the binomial distribution models the number of successes from a set of n independent binary events, each occurring with probability p.

This is analogous to the the sum of n i.i.d. Bernoulli random variables with parameter p.

Definition 7. $X \sim \mathit{Binomial}(n, p) \iff X = \sum_{i=1}^{n} X_i$ *subject to* $X_i \sim \mathit{Bernoulli}(p)$

The following are properties of the binomial distribution:

1. $X = 0$, with probability $(1-p)^n$.
2. $X = 1$, with probability $n \cdot p \cdot (1-p)^{n-1}$.

The PMF for the binomial distribution is given as $f_X(k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$.

1.0.5 Events

Definition 8. *An event is a set of realizations taken from the sample space of a random variable.*

Example 1.4. *Subsetting the set of possible outcomes from a fair dice roll, we can denote* ξ_{even} *as the event for rolling an even number.*

$$Pr[\xi_{even}] = \frac{1}{2}.$$

1.0.6 Conditional Probability

Definition 9. *Given two events* ξ_A, ξ_B, *the probability of* ξ_A *given* ξ_B *is denoted as* $Pr[\xi_A | \xi_B] = \frac{Pr[\xi_A \cap \xi_B]}{Pr[\xi_B]}$.

Conditional probability is defined as the probability of an event happening under the assumption that another event has happened. We can think of conditional probabilities as the probability distribution of ξ_A on the restricted sample space where ξ_B has occurred.

Example 1.5. *Going back to the fair dice example, let ξ_6 be the event that the outcome of a dice roll is 6. Under an unconditional probability, we get that $Pr[\xi_6] = \frac{1}{6}$. If we wish to compute the conditional probability of ξ_6 given that all dice rolls in a sequence are even (ξ_{even}), we get that*

$$Pr[\xi_6|\xi_{even}] = \frac{Pr[\xi_6 \cap \xi_{even}]}{Pr[\xi_{even}]} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}.$$

1.0.7 Independence

In statistics, independence is used to ascertain whether or not two random variables A, B depend on one another. A, B are said to be independent if the occurrence of the event ξ_A is not influenced by the event ξ_B.

Definition 10. *Independence Two events, ξ_A, ξ_B are called independent $\iff$ $Pr[\xi_A \cap \xi_B] = Pr[\xi_A] \cdot Pr[\xi_B]$. To denote that random variables A, B are independent, we use the notation $A \perp B$.*

An equivalent statement of the above definition is that $Pr[\xi_A|\xi_B] = Pr[\xi_A]$.

1.0.8 Expected Value

The expected value of a random variable X, also known as the mean or average, is the probability weighted sum of the realization of all possible outcomes of X.

Definition 11. *If S is the sample space of all possible events for X, we the expected value of a function $g(\cdot)$ on X is equal to*

$E[g(X)] = \sum_{x \in S} g(x) \cdot f_X(x) = \int_{x \in S} g(x) \cdot f_X(x) dx.$

Example 1.6. *Suppose that $X \sim Uniform([6])$.*

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

Example 1.7. *Suppose that $X \sim Bernoulli(p)$.*

$$E[X] = 0 \cdot (1 - p) + 1 \cdot (p) = p$$

Example 1.8. *Suppose that $X \sim Binomial(n, p)$.*

$$
\begin{aligned}
E[X] &= \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} && (1) \\
&= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} && (2) \\
&= np \sum_{x=0}^{n-1} \frac{(n-1)!}{(x-1)!((n-1)-(x-1))!} p^{x-1} (1-p)^{n-x} && (3) \\
&= np && (4)
\end{aligned}
$$

The last line holds true, as $\sum_{x=0}^{n-1} \frac{(n-1)!}{(x-1)!((n-1)-(x-1))!} p^{x-1}(1-p)^{n-x}$ sums over the PMF of a binomial random variable with parameters $(n-1, p)$, and thus equals 1. To demonstrate a less tedious way of computing the expectation of binomial random variables, the concept of linearity of expectations is introduced.

Theorem 12. *(Linearity of Expectations) Let X be the weighted sum of a set of random variables. If $X = \sum_{i=1}^{n} a_i X_i$, where $a_i \in \mathbb{R}$, and X_i is a random variable, then*

$$E[X] = E[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i E[X_i]$$

Due to the linearity of expected values, we can solve the expected value of a binomial random variable $X \sim Binomial(n, p)$ using the fact that a binomial random variable X is equivalent to the sum of n Bernoulli random variables $X_i \sim Bernoulli(p)$

By linearity of expectations,

$$E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p = np$$

1.0.9 Product of Expectations

Due to the convenience of linearity of expectations, one would hope that a similar result exists for products of random variables X and Y. Unlike linearity of expectations, $E[XY] = E[X] \cdot E[Y] \iff X$ and Y are uncorrelated.

Example 1.9. *The following is an example of the identity breaking down for correlated random variables.*

Let $X = Y \sim Bernoulli(\frac{1}{2})$

$$E[X] = \frac{1}{2}, E[Y] = \frac{1}{2}, E[XY] = \frac{1}{2} \neq \frac{1}{4} = E[X] \cdot E[Y]$$

1.0.10 Independence and Correlation

Correlation $\rho_{X,Y}$ between two random variables X, Y measures the strength of the linear relationship between X and Y. $|\rho_{X,Y}| \in [0,1]$, with larger values indicating a stronger linear relationship, and a $\rho_{X,Y} = 0$ representing no linear relationship. A pair of random variables X, Y is said to be uncorrelated when $\rho_{X,Y} = 0$.

Theorem 13. *If X and Y are independent random variables, then X and Y are uncorrelated.*

Proof.

$$\begin{aligned}
E[XY] &= \sum_{x\in S_x}\sum_{y\in S_y} xy\cdot \Pr[X = x \cap Y = y] \\
&= \sum_{x\in S_x}\sum_{y\in S_y} xy\cdot \Pr[X = x]\Pr[Y = y] \\
&= \sum_{x\in S_x} x\cdot \Pr[X = x] \sum_{y\in S_y} y\cdot \Pr[Y = y] \\
&= \sum_{x\in S_x} x\cdot \Pr[X = x]\cdot E[Y] \\
&= E[Y]\cdot \sum_{x\in S_x} x\cdot \Pr[X = x] \\
&= E[X]\cdot E[Y]
\end{aligned}$$

□

This theorem is not an if and only if. The following example shows a pair of uncorrelated random variables that are not independent.

Example 1.10. *Let $X \sim$ Uniform[-1,1], and $Y \sim X^2$*

$$E[X] = \int_{-1}^{1} \frac{1}{2} x f(x)dx = \frac{1}{2}\int_{-1}^{1} x f(x)dx = \frac{1}{2}\left[\frac{x^2}{2}\right]_{-1}^{1} = \frac{1}{4}(1^2 - (-1)^2) = 0$$

$$E[Y] = \int_{-1}^{1} \frac{1}{2} x^2 f(x)dx = \frac{1}{2}\int_{-1}^{1} x^2 f(x)dx = \frac{1}{2}\left[\frac{x^3}{3}\right]_{-1}^{1} = \frac{1}{6}(1^3 - (-1)^3) = \frac{1}{3}$$

$$E[XY] = \int_{-1}^{1} \frac{1}{2} x\cdot x^2 f(x)dx = \frac{1}{2}\int_{-1}^{1} x^3 f(x)dx = \frac{1}{2}\left[\frac{x^4}{4}\right]_{-1}^{1} = \frac{1}{4}(1^4 - (-1)^4) = 0$$

Thus, $E[XY] = E[X]E[Y]$, which implies that X and Y are uncorrelated.

However, $f_Y(1)f_X(0) > 0$, but $Pr[(X = 0) \cap (Y = 1)] = 0$. Thus, we have shown that a pair of random variables can be uncorrelated, but not independent.

1.0.11 Geometric Distribution

$X \sim \text{Geo}(p)$ denotes that X follows a geometric distribution. This is analogous to treating X as the number of tosses it takes before a weighted coin with probability p of landing heads lands on heads. The PMF of this distribution is as follows:

$$f_X(k) = (1-p)^{k-1}p, \ \forall k \geq 1$$

We can compute the expected value of a Geometric random variable:

$$\begin{aligned}
E[X] &= \sum_{i=0}^{\infty}(1 - \Pr[X \leq i]) \\
&= \sum_{i=0}^{\infty}\Pr[X > i] \\
&= \sum_{i=1}^{\infty}\Pr[X \geq i] \\
&= \sum_{i=1}^{\infty}\sum_{k=i}^{\infty}(1-p)^{k-1}p \\
&= \sum_{i=1}^{\infty}(1-p)^{i-1}\sum_{j=1}(1-p)^{j-i}p \\
&= \sum_{i=1}^{\infty}(1-p)^{i-1} \\
&= \sum_{i=1}^{\infty}(1-p)^{i-1}\frac{p}{p} \\
&= \frac{1}{p}\sum_{i=1}^{\infty}(1-p)^{i-1}p \\
&= \frac{1}{p}
\end{aligned}$$

Alternatively, $\sum_{i=1}^{\infty}(1-p)^{i-1} = \frac{1}{1-(1-p)} = \frac{1}{p}$.

1.0.12 Dice Problem

The following dice problem from Elchanan Mossel's blog illustrates the counter-intuitive nature of conditional probabilities.

A dice is repeatedly thrown until it lands on a 6. Let T be the number of rolls it takes for a dice to roll a 6, and let $\xi_{\text{all even}}$ be the event that that all dice rolls in a sequence are even. What is $E[T|\xi_{\text{all even}}]$?

Answering using intution from the geometric distribution, one might claim that $E[T|\xi_{\text{all even}}] = 3$. However, upon further inspection, we see that $E[T|\xi_{\text{all even}}]$ is equivalent to finding the expected number of throws until the result is not a 2 or 4. Using the geometric distribution, we realize that $p = \Pr[not\{2,4\}] = \frac{2}{3}$. Hence, $E[T|\xi_{\text{all even}}] = \frac{1}{p} = \frac{3}{2}$.

1.1 Quicksort

Quicksort is a famous sorting algorithm that emphasizes the use of randomness and probabilistic analysis. The algorithm is carried out in a divide and conquer manner by selecting a pivot and partitioning elements on either side of the pivot. In algorithms like quicksort, and in future randomized algorithms that we will see in the course, the emphasis will be on minimizing the expected runtime and not the worst-case runtime.

Algorithm 1 Deterministic Quicksort

```
procedure QUICKSORT($[x_1, x_2, ..., x_n]$)
    $pivot \leftarrow x_1$
    $\mathbf{S}_{smaller} \leftarrow [\ ]$, $\mathbf{S}_{larger} \leftarrow [\ ]$
    for i in 1:n do
        if $x_i \leq pivot$ then
            $\mathbf{S}_{smaller}.append(x_i)$
        else
            $\mathbf{S}_{larger}.append(x_i)$
    return[Quicksort($\mathbf{S}_{smaller}$), $pivot$, Quicksort($\mathbf{S}_{larger}$)]
```

In the Deterministic Quicksort presented in Algorithm 1, the runtime is dictated by arbitrary user input. For example, if a user were to feed the array $\{n, n-1, ..., 2, 1\}$, then there will be a total of $\frac{n\cdot(n-1)}{2} = \Theta(n^2)$ comparisons.

Alternatively, we can select the pivot uniformly at random. We will demonstrate that, in expectation, this will lead to a significantly faster running time.

Theorem 14. *Let $y_1, ..., y_n$ be the correctly sorted list, and $X = \sum_{i<j} X_{ij}$ denote the number of comparisons for randomized quicksort where*

$$X_{ij} = \begin{cases} 1 & y_i, y_j \textit{ are compared} \\ 0 & \textit{otherwise} \end{cases}$$

then

$$E[X] = 2n \log n + O(n)$$

Intuition: we usually will separate the sequence fairly evenly $(\frac{n}{2}, \frac{n}{2})$. That said, even when the split is far from balanced (say, $(0.9n, 0.1n)$), we will still obtain the desired runtime. Specifically, these two types of splits lead to recurrences describing the number of comparisons which are, respectively,

$$C(n) = n - 1 + 2C(n/2) \qquad \text{and} \qquad C(n) = n - 1 + C(0.1n) + C(0.9n)$$

Solving either recurrences leads to a run time of $O(n \log n)$. We proceed with the more formal argument.

Proof. To determine the probability that y_i and y_j are compared, we need either y_i or y_j to be chosen as a pivot before any of $\{y_{i+1}, ..., y_{j-1}\}$. If one of the other pivots are chosen, then y_i and y_j are split into two different sets, and will never be compared. This probability equals $\frac{2}{j-i+1}$.

$$\begin{aligned}
E[X] &= \sum_{i<j} E[X_{ij}] \\
&= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{2}{j-i+1} \\
&= \sum_{i=1}^{n-1} \sum_{k=2}^{n-i+1} \frac{2}{k} \\
&= \sum_{k=2}^{n-i+1} \sum_{i=1}^{n-1} \frac{2}{k} \\
&= \sum_{k=2}^{n} \sum_{i=1}^{n+1-k} \frac{2}{k} \\
&= \sum_{k=2}^{n} (n+1-k)\frac{2}{k} \\
&= (2n+2)\sum_{k=1}^{n} \frac{1}{k} - 4n
\end{aligned}$$

Using the definition of the harmonic numbers,

$$H_n = \sum_{k=1}^{n} \frac{1}{k} = \Theta(\log n) = \log n + \gamma + \frac{1}{2n} + O\left(\frac{1}{n^2}\right)$$

Thus,

$$E[X] = (2n+2)H_n - 4n = 2n \log n + O(n).$$

□

1.1.1 Coupon Collector

In the coupon collector problem, we repeatedly sample from a set of n objects until at least one copy of each distinct object is obtained. More explicitly, if $X_i \sim Uniform([n])$ and T is the number of draws before the every item $\{1, ..., n\}$ is seen, the coupon collector attempts to solve for $E[T]$.

Let t_i = time to collect the i^{th} unique coupon after collecting $i-1$ unique coupons. $t_i \sim$ Geometric(p).

$$p = 1 - \frac{i-1}{n} = \frac{n-i+1}{n}$$

Thus,

$$E[T] = \sum_{i=1}^{n} E[t_i] = \frac{n}{n} + \frac{n}{n-1} + ... + \frac{n}{1} = n(1 + \frac{1}{2} + ... + \frac{1}{n}) = nH_n$$

$$E[T] = n(\log n + \gamma)$$

The expectation is not the only quantity of interest in the coupon collector problem. We also wish to study its tail probabilities; we wish to create an upper bound R such that T exceeds R with low probability (we define low probability as $p \leq \frac{1}{n}$.)

Given a sequence of R draws for the coupon collector problem, T exceeds R if at least one of the n distinct objects has not been selected. Let $\mathcal{E}_i^R$ denote when item i is not observed in the first R draws ($Pr(\mathcal{E}_i^R) = (1 - \frac{1}{n})^R$).

$$\text{Pr(not done in first R draws)} = \Pr[\cup_{i=1}^{n} \mathcal{E}_i^R] \leq \sum_{i=1}^{n} \Pr[\mathcal{E}_i^R] = \sum_{i=1}^{n} (1 - \frac{1}{n})^R = n(1 - \frac{1}{n})^R$$

Set $R = \beta n \log n$

$$n(1 - \frac{1}{n})^R \leq n(e^{-\frac{1}{n}})^{\beta n \log n} = ne^{-\beta \log n} = n(e^{\log n})^{-\beta} = n^{-\beta+1}$$

since

$$1 + x \leq e^x, \forall x \in \mathbb{R}$$

Thus,

$$\text{Pr(not done in first R draws)} \leq n^{-\beta+1}$$

1.2 Concentration Inequalities

Concentration inequalities define the magnitude of deviation that a random variable deviates from its expected value. Though many random variables have high probability mass surrounding its expectation, many random variables can strongly deviate from their expected values. For example, consider

$$X = \begin{cases} n & \text{subject to } \Pr[X = n] = \frac{1}{2} \\ -n & \text{subject to } \Pr[X = -n] = \frac{1}{2} \end{cases}$$

$E[X] = \frac{1}{2}(n - n) = 0$, but $|X - E[X]| = n$, where n can be taken to be arbitrarily large.

Most random variables are not as ill-behaved as the example given above. Specifically, if a random variable is sufficiently “nice,” we can say that it will be “close” to its expected value. Our first definition of “nice” will be fairly weak: simply that the random variable is non-negative. Later, we will enforce stronger notions of “niceness,” leading to sharper concentration.

Theorem 15 (Markov's Inequality)**.** *If X is a non-negative random variable, then*

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

In other words,

$$\Pr[X \geq \beta E[X]] \leq \frac{1}{\beta}.$$

Proof.

$$\begin{aligned} E[X] &= \int_{x \in S} x f(x) dx \\ &= \int_{x<a} x f(x) dx + \int_{x \geq a} x f(x) dx \\ &\leq 0 + \int_{x \geq a} f(x) dx \\ &= a \int_{x \geq a} f(x) dx \\ &= a \Pr[X \geq a] \end{aligned}$$

Thus,

$$\frac{E[X]}{a} \geq \Pr[X \geq a]$$

□

Note that Markov's inequality is tight, realized by the following example:

$$X = \begin{cases} a & \text{subject to } \Pr[X = a] = \frac{t}{a} \\ 0 & \text{subject to } \Pr[X = 0] = 1 - \frac{t}{a} \end{cases}$$

Then,

$$E[X] = t, \Pr[X \geq a] = \frac{t}{a} = \frac{E[X]}{a}.$$

Markov's inequality is the first of many concentration inequalities that we will be covering in this course. Intuitively, it indicates how well behaved a random variable is with respect to the first moment, the mean. Understandably, we will be using Markov's inequality as a tool in the coupon collector problem as follows:

$$\Pr[T > \beta n \log n] \leq \frac{1}{\beta}$$

Notice that this bound is different than the one we obtained above.

1.2.1 Variance

The variance of a random variable is the ‘centered second moment.’ It serves as a measure of how much the random variable fluctuates about its expected value.

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

Proof.

$$\begin{aligned} Var(X) = E[(X - E[X])^2] &= E[X^2 - 2XE[X] + E[X]^2] \\ &= E[X^2] - 2E[XE[X]] + E[E[X]^2] \\ &= E[X^2] - 2E[X]E[X] + E[X^2] \\ &= E[X^2] - E[X]^2 \end{aligned}$$

□