

DAA II Lecture 0 Review of Elementary Probability Theory

Or, all that you probably need to know about elementary probability to analyze the algorithms in this course. The average behavior $A(n)$ of an algorithm is defined to be the expectation of the random variable t , where $t : I_n \rightarrow \{0, 1, 2, \dots, W(n)\}$ maps an input $I \in I_n$ of size n into how many basic operations are performed by the algorithm with the given input I . In this lecture we review the notions from elementary probability theory which make the previous sentence make sense.

Recall that a *finite sample space* is simply a finite set S together with a mapping $P : S \rightarrow [0, 1]$ such that $\sum_{s \in S} P(s) = 1$. If $A \subseteq S$, then $P(A) = \sum_{s \in A} P(s)$. P is called a *probability distribution* on S . P is called the *uniform distribution* on S if it is a **constant** mapping, that is, $P(s) = 1/|S|$ for all elements s in S . Thus, probabilities when we have the uniform distribution are calculated by simply counting.

Interest in probability theory goes back to ancient times, but was only put on a solid mathematical foundation by Pascal and others in the 1600s. People were interested in games of chance, and wanted to compute the odds for various outcomes of the games. For example, suppose we consider the game of poker. Recall that a poker hand is 5 cards drawn from a deck of 52. We assume a fair game, so that each of the $1/C(52, 5)$ different hands is equally likely (it is no accident that Pascal's name is so integrally tied to binomial coefficients $C(n, m)$, since he was an early developer of the theory of probability, and we have Pascal's Triangle based on the recurrence $C(n, k) = C(n - 1, k) + C(n - 1, k - 1)$, as well as his recurrence for $S(n, k) = 1^k + 2^k + \dots + n^k$, which also involves binomial coefficients). Now consider the question of whether or not a flush (5 cards of the same suit) beats a straight (5 cards in numerical order, but NOT ALL in the same suit, that is, not a straight flush). This becomes the question of which event is less likely. Note that

$$P(\text{flush}) = 4 * C(13, 5),$$

since there are 4 suits, and once the suit is chosen, there are $C(13, 5)$ ways to get a flush in that suit. On the other hand,

$$P(\text{straight}) = (10 * 4^5 - 10 * 4) / C(52, 5),$$

since there are 10 choices for the lowest numerical value in a straight, then 4^5 straights having a given numerical lowest value. The reason we must subtract $10 * 4$ is to account for straight flushes, which indeed beat regular (non-straight) flushes. Since we have the same denominator in both probabilities, it becomes a question of which numerator is smaller. A quick calculation shows that $4 * C(13, 5) = 3744$, whereas $10 * 4^5 - 10 * 4 = 450 - 40 = 4096$.

Thus, $P(\text{flush}) < P(\text{straight})$, which means that a flush beats a straight.

Remarks. 1. In computing these numbers, we used the so-called *Rule of Product*, which says if there are m ways for an event A to occur, and *independently* there are n ways for a second event B to occur, then there are $m * n$ ways for both events to occur (this has the obvious generalization to $k > 2$ events). This is actually a special case of conditional probability, which that we now discuss.

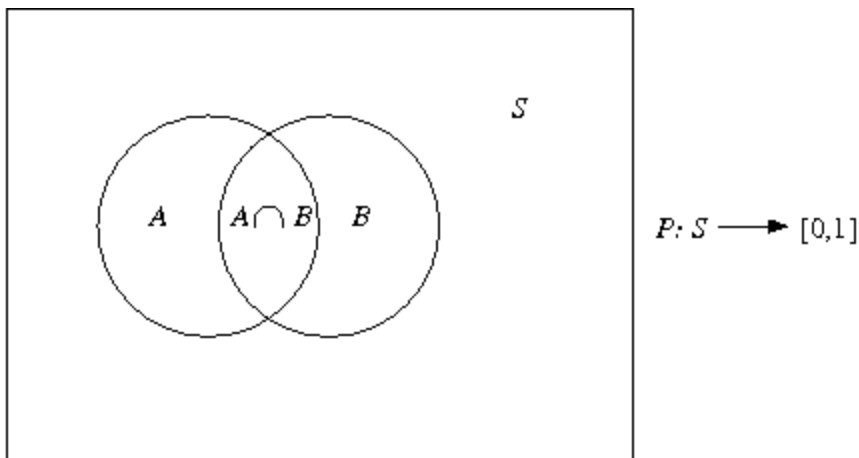
Conditional Probability

Give two subsets (events) A and B of S , the *conditional probability* $P(A|B)$ is the probability that A occurs **given that B has occurred**. **Keep this explanation of $P(A|B)$ in mind, even though I will now define it more formally.** The question is, what is a formula for $P(A|B)$? Well, if P is the uniform distribution, then $P(A|B)$ is nothing more

than $P(A|B) = \frac{|A \cap B|}{|B|}$ (see the figure below). However, writing this equivalently as

$$P(A|B) = \frac{|A \cap B|/|S|}{|B|/|S|} = \frac{P(A \cap B)}{P(B)},$$
 we see how to define $P(A|B)$ in the general case,

namely, $P(A|B) = \frac{P(A \cap B)}{P(B)}$.



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

This result is often used in the form $P(A \cap B) = P(A|B)P(B)$.

Two events A and B are called *independent* if $P(A|B) = P(A)$. Then the above equation becomes

$$P(A \cap B) = P(A)P(B),$$

which is reminiscent of the Rule of Product we stated earlier when we have the uniform distribution so that probabilities are computed by simple counting the elements in the given event, and then dividing by $|S|$.

Another useful formula, this time involving union as well as intersection, is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

You were asked to prove this as part of Homework Assignment 1.

Now consider the example of rolling a pair of fair dice. Then each die can come up 1,2,3,4,5 or 6, with each of these events being equally likely. Let B be the event that die 1 has come up 5, and A the event that the sum of the two dice is > 8 . Then $P(A|B)$ is intuitively equal to $3/6$, since there are 3 ways for die 2 to come up (namely, 4,5,6) so that the sum is > 8 given that die 1 came up 5. Note that this calculation can also be done in a purely formal way, using the formula for condition expectation:

$$p(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{3/36}{1/6} = 3/6,$$

but our intuitive calculation was correct and easier.

Expectation of a Random Variable

Given a finite sample space S , a random variable defined on S is nothing more than a fancy name for a function X mapping S to the real numbers, $X: S \rightarrow \text{Reals}$. Hence, if $S = \{s_1, s_2, \dots, s_m\}$, the X maps S to the m real numbers $X(s_1), X(s_2), \dots, X(s_m)$. Now the “man on the street” average of these $X(s_1), X(s_2), \dots, X(s_m)$, namely,

$$\text{AVE}(X(s_1), X(s_2), \dots, X(s_m)) = (X(s_1) + X(s_2) + \dots + X(s_m))/m$$

is exactly what we call the *expectation of the random variable* X , given that we have the **uniform distribution** on S . We denote this average by $E[X]$. But how should we define $E[X]$ if we do NOT have the uniform distribution? Well, analogous to the ordinary average, we define

$$E[X] = X(s_1)P(s_1) + X(s_2)P(s_2) + \dots + X(s_m)P(s_m),$$

which agrees with how we defined $E[X]$ when we have the uniform distribution, since then $P(s_i) = 1/m$ for each $i = 1, 2, \dots, m$, and the $1/m$ can be factored out of the above expression using the distributive law for addition and multiplication.

Probably the most useful elementary result about $E[X]$ is that it is “additive.” More precisely, suppose $X = X_1 + X_2 + \dots + X_k$. Then

$$E[X] = E[X_1 + X_2 + \dots + X_k] = E[X_1] + E[X_2] + \dots + E[X_k].$$

To prove this result, we have

$$\begin{aligned}
& E[X_1 + X_2 + \dots + X_k] \\
&= (X_1 + X_2 + \dots + X_k)(s_1)P(s_1) + \dots + (X_1 + X_2 + \dots + X_k)(s_m)P(s_m) \\
&= ((X_1(s_1) + X_2(s_2) + \dots + X_k(s_m))P(s_1) + \dots + ((X_1(s_1) + X_2(s_2) + \dots + X_k(s_m))P(s_m)) \\
&\quad \text{(by definition of } X_1 + X_2 + \dots + X_k) \\
&= ((X_1(s_1)P(s_1) + X_2(s_2)P(s_1) + \dots + X_k(s_m)P(s_1)) \\
&\quad + \dots + ((X_1(s_1)P(s_m) + X_2(s_2)P(s_m) + \dots + X_k(s_m)P(s_m))) \\
&\quad \text{(by the distributive law for addition and multiplication)} \\
&= ((X_1(s_1)P(s_1) + X_1(s_2)P(s_1) + \dots + X_1(s_m)P(s_1)) \\
&\quad + \dots + ((X_k(s_1)P(s_1) + X_k(s_2)P(s_1) + \dots + X_k(s_m)P(s_1))) \\
&\quad \text{(by rearranging the sum using the commutative law for addition)} \\
&= E[X_1] + E[X_2] + \dots + E[X_k] \quad \text{(by definition of expectation).}
\end{aligned}$$

When applied to the problem of computing $A(n) = E[t]$ for an algorithm, the above formula with $t = t_1 + t_2 + \dots + t_k$ is applied in the situation where the code for the algorithm is divided up into k disjoint code segments, and t_i is the number of basic operation performed by the algorithm during the execution of the i^{th} code segment, $i = 1, 2, \dots, k$ (which we have called Formulation IV).

Another useful fact about expectation is $E[cX] = cE[X]$, and is easily proved (verify this!), again using the distributive law of addition and multiplication.

To convince you of the utility of the additive property of expectation, consider again the sample space S consisting of the 36 possible outcomes of rolling a pair of fair dice. Denote these outcomes by $(D1, D2)$, where $D1$ is the number showing on die 1, and $D2$ is the number showing on $D2$. Let X be the random variable defined by $X(D1, D2) = D1 + D2$, so that X maps S to the set $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. Then by simply counting, the number of ways to get these various sums is 1, 2, 3, 4, 5, 6, 7, 6, 5, 4, 3, 2, 1, respectively. If each of these numbers is divided by 36, we have the probability that X maps $(D1, D2)$ to the given sum. Thus, we have

$$E[X] = (1/36)[1*2 + 2*3 + 3*4 + 4*5 + 5*6 + 6*7 + 5*8 + 4*9 + 3*10 + 2*11 + 1*12],$$

which is a sum not easily calculated in your head, but it turns out to be 7.

Remarks. This calculation is actually using Formulation II for computing expectation. Indeed, given any random variable, if we let p_i denote the probability that $X = i$, (where we use the notation $P(X=i)$ to denote this probability), then Formulation II says that

$$(II) \quad E[X] = \sum_i i p_i \cdot$$

In our example, $i = 2, 3, \dots, 12$, and $P(X=i) = 1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 5/36, 4/36, 3/36, 2/36, 1/36$, respectively. Hence, (II) agrees with our previous calculation, in which we simply factored out the common multiplier $1/36$.

We now use the additivity of expectation to get the previous result much easier. Note that $X = X_1 + X_2$, where $X_1(D1, D2) = D1$, and $X_2(D1, D2) = D2$. Moreover, the $E[X_i]$, $i = 1, 2$ are simply the expected value of the outcome of rolling a single fair die, since this event does not depend on the outcome of the other die. Hence

$$E[X_i] = (1 + 2 + 3 + 4 + 5 + 6)/6 = (6 \cdot 7/2)/6 \quad (\text{which we can do in our head using our old friend for summing } 1 + 2 + \dots + n = \underline{n(n + 1)/2})$$

$$= 7/2,$$

so that $E[X_1 + X_2] = E[X_1] + E[X_2] = 7/2 + 7/2 = 7$.

Conditional Expectation

Analogous to conditional probability, given a subset $F \subseteq S$, together with a random variable X defined on S , we now define the conditional expectation $E[X|F]$. The description in words of this expectation is “the expectation of X given that F has occurred.” This intuitive notion is very helpful when computing, but we need a formal definition. The idea is simple. First we define a probability distribution P_F on F by

$$P_F(s) = P(s)/P(F) \text{ for all } s \in F.$$

Note that this is indeed a probability distribution on F

$$\text{since } \sum_{s \in F} (P(s)/P(F)) = \frac{1}{P(F)} \sum_{s \in F} P(s) = \frac{1}{P(F)} (P(F)) = 1$$

It is very important to note that if P is the uniform distribution on S , then P_F is the uniform distribution on F , that is, $P_F(s) = 1/|F|$ for all s in F .

Remember from calculus (or somewhere) that the *restriction* $(X|F)$ of X to F is the function defined on F which agrees everywhere on F with X . We then simply define $E[X|F]$ to be $E[(X|F)]$ with respect to P_F . In symbols, we have

$$E[X | F] = \sum_{s \in F} ((X | F)(s) P_F(s)) = \sum_{s \in F} (X(s) P(s) / P(F)) = \frac{1}{P(F)} \sum_{s \in F} (X(s) P(s))$$

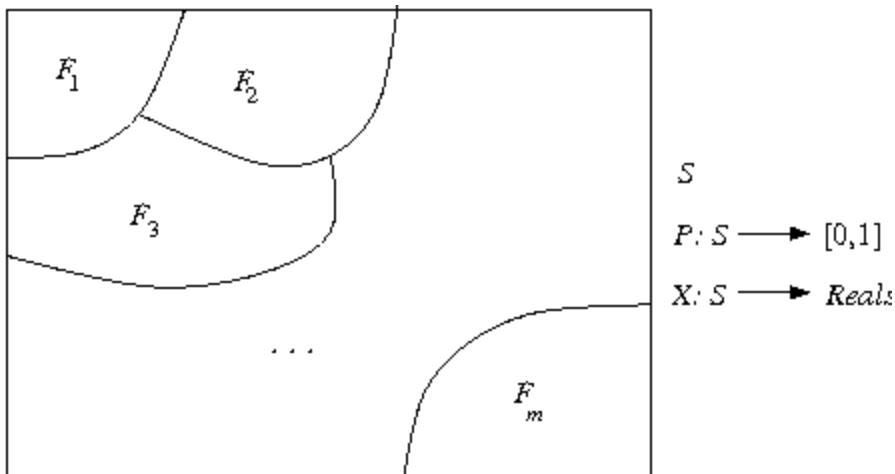
Note that when X is the uniform distribution, then the above formula becomes

$$E[X | F] = \sum_{s \in F} ((X | F)(s) P_F(s)) = \sum_{s \in F} (X(s) (\frac{1}{|F|})) = \frac{1}{|F|} \sum_{s \in F} X(s)$$

Conditional Expectation and Partitioning the Sample Space

The formulas for conditional expectation are generally used when the sample space S is broken up into a union of m disjoint subsets (called a *partition of S*)

$$S = \bigcup_{s \in F_i} S.$$



We then have

$$\begin{aligned} E[X] &= \sum_{s \in S} (X(s) P(s)) = \sum_{i=1}^m (\sum_{s \in F_i} (X(s) P(s))) = \sum_{i=1}^m (\sum_{s \in F_i} (X(s) (P(s) / P(F_i)) P(F_i))) \\ &= \sum_{i=1}^m ((\sum_{s \in F_i} (X(s) P_{F_i}(s))) P(F_i)) = \sum_{i=1}^m ((E[X | F_i]) P(F_i)). \end{aligned}$$

This is Formulation V for expectation. Of course, this latter formula is only useful when the quantities $E[X|F_i]$ and $P(F_i)$ are readily computable, $i = 1, 2, \dots, m$. Even when these quantities are readily computable, it may not be the easiest way to compute $E[X]$. To illustrate this point, as well as to interpret conditional expectation in a easily understandable setting, let's reexamine the problem of computing $E[X]$ for the sample space of rolling two fair dice and X is the sum showing on the two dice. Divide up the 36 possible outcomes into 6 disjoint sets, depending on the value showing on die 1. In other words, $F_i = \{(i, D2) : D2 = 1, 2, 3, 4, 5, 6\}$

Note that $P(F_i) = 1/6$, $i = 1, 2, 3, 4, 5, 6$. $E[X | F_i]$ is also readily computable, since we have

$$E[X | F_i] = (1/6)[i+1 + i+2 + i+3 + i+4 + i+5 + i+6] = (1/6)[(i+6)(i+7)/2 - i(i+1)/2] \\ = (1/12)[12i + 42] = (2i + 7)/2. \text{ We then have}$$

$$E[X] = \sum_{i=1}^6 ((E[X | F_i])P(F_i)) = \frac{1}{6} \sum_{i=1}^6 E[X | F_i] = \frac{1}{6} \sum_{i=1}^6 (2i + 7)/2 = \frac{1}{12} [9 + 11 + 13 + 15 + 17 + 19] \\ = 84/12 = 7.$$

Well, it certainly was easier to compute $E[X]$ using the formula $E[X] = E[X_1] + E[X_2]$, where $X_i(D1, D2) = Di, i = 1, 2$, as was done earlier in the lecture, but the example is a nice concrete illustration of the use of conditional expectation. Moreover, while computing $E[X]$ using conditional expectation was not helpful in this case, there are situations in algorithm analysis where breaking up the input space I_n into disjoint subsets is the only way to solve the problem reasonably.

We finish with the comment that partitions of a set S into m disjoint subsets is really the same thing as defining a map $Y: S \rightarrow \{1, 2, \dots, m\}$. Indeed, given a partition $S = \bigcup_{i=1}^m F_i$, merely define $Y(s) = i$, where $s \in F_i$. Since we have a partition, Y is a well-defined mapping, and $F_i = \{s \in S \mid Y(s) = i\}$. On the other hand, given a mapping $Y: S \rightarrow \{1, 2, \dots, m\}$, we get a partition of S into m subsets by defining $F_i = \{s \in S \mid Y(s) = i\}$. Hence, we often compute $E[X]$ using a second random variable $Y: S \rightarrow \{1, 2, \dots, m\}$, in which case our formula using conditional expectations becomes

$$E[X] = \sum_{i=1}^m ((E[X | Y = i])P(Y = i))$$