CSCE-658 Randomized Algorithms

Lecture #8, February 19, 2016

Lecturer: Professor Jianer Chen

8 Random variables and expectation

We have seen applications of randomized algorithms. Most of our analysis were based on "elementary" probability theory that you probably have seen in your high school math class. Now we would like to move forward a little bit and use techniques in probability theory that is a little bit less elementary.

8.1 Random variables

A *random variable* on a sample space Ω is just a function from Ω to the set $\mathbb{R}$ of real numbers. Indeed, the range of a random variable can be other things, such as a pair of integers, rather than just real numbers. However, this kind of random variables either is less interesting in our discussion or can be composed using functions on random variables as defined here. Thus, we will use our definition in the following discussion. Rather than writing a random variable as $f(\omega)$, the convention is to write a random variable as a capital letter such as X and Y and make the argument implicit: X is really $X(\omega)$.

Random variables can be any functions defined on a sample space. Consider the sample space we have seen in our Homework #1, Question 2: Henry and Tom play a game by repeatedly tossing a fair coin. Henry gains a point if the coin turns head and Tom gains a point if the coin turns tail. The game is over if either Henry gets two points or Tom gets three points. The outcomes of the sample space are (the parentheses indicate the probabilities of the outcomes):

$$HH\ (^1\!/_4),\ \ HTH\ (^1\!/_8),\ \ HTTH\ (^1\!/_{16}),\ \ HTTT\ (^1\!/_{16}),\ \ THH\ (^1\!/_8),$$
$$THTH\ (^1\!/_{16}),\ \ THTT\ (^1\!/_{16}),\ \ TTHH\ (^1\!/_{16}),\ \ TTHT\ (^1\!/_{16}),\ \ TTT\ (^1\!/_8).$$

Thus, if we define X_1 to be the random variable for the points gained by Henry, then

$$X_1(HH) = 2,\ \ X_1(HTH) = 2,\ \ X_1(HTTH) = 2,\ \ X_1(HTTT) = 1,\ \ X_1(THH) = 2,$$
$$X_1(THTH) = 2,\ \ X_1(THTT) = 1,\ \ X_1(TTHH) = 2,\ \ X_1(TTHT) = 1,\ \ X_1(TTT) = 0. \tag{18}$$

On the other hand, you may think that the points gained by Tom are paid off by Henry. Thus, if we define X_2 to be the "net" gain for Henry, then

$$X_2(HH) = 2,\ \ X_2(HTH) = 1,\ \ X_2(HTTH) = 0,\ \ X_2(HTTT) = -2,\ \ X_2(THH) = 1,$$
$$X_2(THTH) = 0,\ \ X_2(THTT) = -2,\ \ X_2(TTHH) = 0,\ \ X_2(TTHT) = -2,\ \ X_2(TTT) = -3. \tag{19}$$

Sometimes we introduce random variables without explicitly giving its values on individual outcomes. For instance, we may say "take an integer uniformly from $\{1, 2, \ldots, 100\}$", which really means that our sample space is $\{1, 2, \ldots, 100\}$ (where each outcome has a probability 1/100), while we are defining a random variable $X(\omega) = \omega$ for all ω in $\{1, 2, \ldots, 100\}$.

We will assume in the rest of this section that our probability space $(\Omega, \mathcal{F}, \Pr)$ is *discrete*, i.e., the sample space Ω is finite or countably infinite, and all subsets of Ω are events in $\mathcal{F}$.

A useful random variable defined based on events is given as follows.

Definition 8.1 Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space, and let $E \in \mathcal{F}$ be an event. Then the (event) *indicator* of E is the random variable that is defined as follows:

$$I_E(\omega) = \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases}$$

Conversely, events can be defined based on random variables. For example, let X be a random variable, then an event can be defined as $\{\omega \mid X(\omega) > 10\}$, which will simply be written as "$X > 10$".

New random variables can be constructed using functions on existing random variables. For example, suppose that both X and Y are random variables, then $X^2 + Y$, $\sin(X) + \sqrt{Y}$ are also random variables.

8.2 Mathematical expectation

The most useful concept on random variables is its expectation, which, intuitively, is its "average value" over all outcomes in the sample space (weighted by their probabilities). Formally,

Definition 8.2 Let X be a random variable over a probability space $\{\Omega, \mathcal{F}, \Pr\}$. The (mathematical) *expectation* of X is defined as

$$\mathbf{E}[X] = \sum_{\omega \in \Omega} \Pr[\omega] \cdot X(\omega),$$

provided that the series "converges absolutely," i.e., $\sum_{\omega\in\Omega} \Pr[\omega] \cdot |X(\omega)| < \infty$. Otherwise, we simply say that the expectation *does not exist*.

Since the sample space Ω is finite or countable, the sum in the above definition is well-defined.

As examples, consider the expectation of the number X_1 of points gained by Henry if he does not pay off Tom's gain (see (18)), we have

$$\mathbf{E}[X_1] = 2 \cdot \frac{1}{4} + 2 \cdot \frac{1}{8} + 2 \cdot \frac{1}{16} + 1 \cdot \frac{1}{16} + 2 \cdot \frac{1}{8} + 2 \cdot \frac{1}{16} + 1 \cdot \frac{1}{16} + 2 \cdot \frac{1}{16} + 1 \cdot \frac{1}{16} + 0 \cdot \frac{1}{8} = \frac{25}{16} = 1.5625.$$

On the other hand, the expectation of the number X_2 of "net points" gained by Henry is (see (19)):

$$\mathbf{E}[X_2] = 2\cdot\frac{1}{4}+1\cdot\frac{1}{8}+0\cdot\frac{1}{16}+(-2)\cdot\frac{1}{16}+1\cdot\frac{1}{8}+0\cdot\frac{1}{16}+(-2)\cdot\frac{1}{16}+0\cdot\frac{1}{16}+(-2)\cdot\frac{1}{16}+(-3)\cdot\frac{1}{8} = \frac{3}{16} = 0.1875.$$

We give another example. Let E be an event, and let I_E be the random variable that is the indicator of E. Since $I_E(\omega)$ is equal to 1 for $\omega \in E$ and is equal to 0 for $\omega \notin E$, we have

$$\mathbf{E}[I_E] = \sum_{\omega\in\Omega} \Pr[\omega] \cdot I_E(\omega) = \sum_{\omega\in E} \Pr[\omega] \cdot 1 = \sum_{\omega \in E} \Pr[\omega] = \Pr[E].$$

The expectation of a random variable X can also be given by the following equivalent definition, which sometimes simplifies the mathematical analysis. Let $r(X) = \{t \mid X(\omega) = t \text{ for some } \omega\}$ be the *range* of the random variable X:

$$\mathbf{E}[X] = \sum_{t \in r(X)} t \cdot \Pr[X = t], \tag{20}$$

where in $\Pr[X = t]$ we have interpreted $X = t$ as the event that consists of all outcomes ω that satisfy $X(\omega) = t$. Note that since the sample space Ω is either finite or countable, the range $r(X)$ of the random variable X is also finite or countable. Thus, the summation in the above definition is well-defined. The formula can be proved directly based on the original definition for expectation:

$$\begin{aligned}
\mathbf{E}[X] &= \sum_{\omega\in\Omega} \Pr[\omega] \cdot X(\omega) = \sum_{t\in r(X)} \left(\sum_{\omega : X(\omega) = t} \Pr[\omega] \cdot X(\omega) \right) \\
&= \sum_{t \in r(X)} \left(\sum_{\omega : X(\omega)=t} \Pr[\omega] \cdot t \right) = \sum_{t\in r(X)} t \left(\sum_{\omega : X(\omega) = t} \Pr[\omega] \right) = \sum_{t \in r(X)} t \cdot \Pr[X = t].
\end{aligned}$$

Again here the validity of regrouping the summation terms in the second equality in the above derivation is ensured by the assumption of the absolute convergence of the series.

8.3 Linearity of expectation

The linearity of expectations is probably the most useful trick when we play with probability and analyze randomized algorithms.

Theorem 8.1 (Linearity of Expectation) *Let $X_1, X_2, \ldots, X_n$ be random variables on a probability space $(\Omega, \mathcal{F}, \Pr)$, and let $c_1, c_2, \ldots, c_n$ be any constants. Then $\mathbf{E}[\sum_{i=1}^n c_i X_i] = \sum_{i=1}^n c_i \mathbf{E}[X_i]$.*

PROOF. First note that $Y = \sum_{i=1}^n c_i X_i$ is a random variable. The theorem can be easily proved based on the definition of expectations:

$$\mathbf{E}[Y] = \mathbf{E}\left[\sum_{i=1}^n c_i X_i\right] = \sum_{\omega\in\Omega} \Pr[\omega] \cdot \left(\sum_{i=1}^n c_i X_i(\omega)\right) = \sum_{i=1}^n \left(c_i \sum_{\omega\in\Omega} (\Pr[\omega] \cdot X_i(\omega))\right) = \sum_{i=1}^n c_i \mathbf{E}[X_i].$$

Note that in the third equality we were able to exchange the summations because of the assumption of the absolute convergence of the series. □

The most interesting (and most useful) property of Linearity of Expectation is that it enforces *no* conditions on the relationship among the random variables $X_1, X_2, \ldots, X_n$. In particular, they do not have to be "independent" (this concept for random variables will be defined later). In fact, it does not even exclude the possibility that some of them are identical such as $X_1 = X_2$.

We give a simple application for Linearity of Expectation. Recall the Max-Cut problem, which is for a given undirected graph $G = (V, E)$, to construct a partition of the vertices of G so that the number of crossing edges is maximized. Recall that the Max-Cut problem is NP-hard.

Let n and m be the number of vertices and the number of edges in the graph G, respectively. Our sample space consists of the 2^n partitions (L, R) of the vertex set V of G, $V = L \cup R$ and $L \cap R = \emptyset$, each with a probability $1/2^n$. For each edge e in G, we define a random variable $X_e(L, R)$ on each partition (L, R) of V such that if the edge e is crossing (i.e., if one end of e is in L and the other end of e is in R) then $X_e = 1$, otherwise $X_e = 0$. Therefore, for a partition (L, R), if we let $Y(L, R) = \sum_{e\in E} X_e(L, R)$, then $Y(L, R)$ is the total number of crossing edges for the partition (L, R). Now let us compute the expectation for Y (using Linearity of Expectation). First of all, it is easy to verify that for each edge e in G, we have $\mathbf{E}[X_e] = 1/2$. This gives:

$$\mathbf{E}[Y] = \mathbf{E}\left[\sum_{e\in E} X_e\right] = \sum_{e\in E} \mathbf{E}[X_e]. = \sum_{e\in E} \frac{1}{2} = \frac{m}{2}.$$

So a random partition of V, which can be implemented by randomly placing each vertex either in L or in R with an equal probability 1/2, will result in a cut (L, R) whose expected size is one half of the total number of edges in the graph G. Such a cut may not be a maximum cut, but is not very bad: its size is at least one half of the size of a maximum cut!

The above randomized algorithm is an *approximation algorithm* for the NP-hard problem Max-Cut, which produces a cut with an expected *approximation ratio* bounded by 2 (the approximation ratio is defined to be the ratio of the size of a maximum cut over the size of the approximated cut). Given the fact that optimally solving an NP-hard problem is difficult, this very simple and efficient (linear-time) randomized algorithm produces a result whose expected quality can be measured specifically. Surprisingly, this simple algorithm stood as the *best* approximation algorithms for the Max-Cut problem for more than 20 years! A very well-known open problem then was whether the approximation ratio of this trivial algorithm could be improved. This was eventually answered positively, again with a randomized algorithm for the problem (we will talk about this if time permits).

8.4 More on random variables and expectations

Based on the independence of events, we can define the independence of random variables.

Definition 8.3 Two random variables X and Y are *independent* if for all pairs (u, v) of real numbers, $\Pr[(X = u) \cap (Y = v)] = \Pr[X = u] \cdot \Pr[Y = v]$.

In the above definition, all "$(X = u) \cap (Y = v)$", "$X = u$", and "$Y = v$" are events defined based on random variables. The definition can be extended to more than two random variables.

Definition 8.4 Random variables $X_1, X_2, \ldots, X_k$ are *independent* if for any real numbers $u_1, u_2, \ldots, u_k$, $\Pr[(X_1 = u_1) \cap (X_2 = u_2) \cap \cdots \cap (X_k = u_k)] = \Pr[X_1 = u_1] \cdot \Pr[X_2 = u_2] \cdot \cdots \cdot \Pr[X_k = u_k]$.

Recall that Linearity of Expectation holds true for sums of any set of random variables, no matter whether the random variables are independent or not. On the other hand, it is easy to see that a similar result is not true for products of random variables. For example, let X and Y be random variables on the probability space $\{0, 1\}$ with $\Pr[0] = \Pr[1] = 1/2$ such that $X(0) = 0$, $X(1) = 1$, $Y(0) = 1$, and $Y(1) = 0$. Then $\mathbf{E}[XY] = 0$ while $\mathbf{E}[X] \cdot \mathbf{E}[Y] = 1/4$. However, for independent random variables, such a result holds true:

Lemma 8.2 *Let $X_1, X_2, \ldots, X_k$ be independent random variables. Then*

$$\mathbf{E}[X_1 X_2 \cdots X_k] = \mathbf{E}[X_1] \cdot \mathbf{E}[X_2] \cdot \cdots \cdot \mathbf{E}[X_k].$$

PROOF. We first prove the lemma for the case $k = 2$, based on the definition of expectations as given in (20). Let $r(X_1)$ and $r(X_2)$ be the ranges of the random variables X_1 and X_2, respectively.

$$\begin{aligned}
\mathbf{E}[X_1 X_2] &= \sum_{u_1 \in r(X_1), u_2 \in r(X_2)} \Pr[(X_1 = u_1) \cap (X_2 = u_2)] \cdot (u_1 u_2) \\
&= \sum_{u_1 \in r(X_1), u_2 \in r(X_2)} \Pr[X_1 = u_1] \cdot \Pr[X_2 = u_2] \cdot (u_1 u_2) \\
&= \sum_{u_2 \in r(X_2)} \sum_{u_1 \in r(X_1)} \Pr[X_1 = u_1] \cdot \Pr[X_2 = u_2] \cdot (u_1 u_2) \\
&= \sum_{u_2 \in r(X_2)} (\Pr[X_2 = u_2] \cdot u_2) \left(\sum_{u_1 \in r(X_1)} \Pr[X_1 = u_1] \cdot u_1 \right) \\
&= \sum_{u_2 \in r(X_2)} (\Pr[X_2 = u_2] \cdot u_2) \cdot \mathbf{E}[X_1] \\
&= \mathbf{E}[X_1] \sum_{u_2 \in r(X_2)} (\Pr[X_2 = u_2] \cdot u_2) \\
&= \mathbf{E}[X_1] \cdot \mathbf{E}[X_2].
\end{aligned}$$

The second equality has used the independence of the random variables X_1 and X_2. Now the case for general $k > 2$ can be proved using induction on k. □

We can also define the *conditional expectations* for random variables based on conditional probability.

Definition 8.5 Let F be an event with $\Pr[F] \neq 0$, and let X be a random variable. Then the *conditional expectation of X given F* is defined as $\mathbf{E}[X \mid F] = \sum_{\omega \in \Omega} \Pr[\omega \mid F] \cdot X(\omega)$.

A number of results for unconditional expectations also hold true for conditional expectations. We list some of them here but leave the verifications to the students.

1. The conditional expectation can also be defined based on the range $r(X)$ of the random variable:

$$\mathbf{E}[X \mid F] = \sum_{t \in r(X)} \Pr[X = t \mid F] \cdot t.$$

2. Linearity of Expectation: for any set of random variables $X_1, X_2, \ldots, X_k$:

$$\mathbf{E}[(X_1 + X_2 + \cdots + X_k) \mid F] = \mathbf{E}[X_1 \mid F] + \mathbf{E}[X_2 \mid F] + \cdots + \mathbf{E}[X_k \mid F].$$

We also have the following results relating the conditional and unconditional expectations. We say that a collection of events $F_1, F_2, \ldots, F_k$ is a *partition* of the sample space Ω if $\Omega = \cup_{i=1}^{k} F_i$ and $F_i \cap F_j = \emptyset$ for all $1 \leq i < j \leq k$.

Lemma 8.3 *Let $F_1, F_2, \ldots, F_k$ be a partition of the sample space. Then for any random variable X:*

$$\mathbf{E}[X] = \sum_{i=1}^{k} \Pr[F_i] \cdot \mathbf{E}[X \mid F_i].$$

PROOF. By definition, $\mathbf{E}[X \mid F_i] = \sum_{\omega \in \Omega} \Pr[\omega \mid F_i] \cdot X(\omega)$. Thus, we have

$$\sum_{i=1}^{k} \Pr[F_i] \cdot \mathbf{E}[X \mid F_i] = \sum_{i=1}^{k} \Pr[F_i] \cdot \left(\sum_{\omega \in \Omega} \Pr[\omega \mid F_i] \cdot X(\omega) \right) = \sum_{\omega \in \Omega} X(\omega) \sum_{i=1}^{k} (\Pr[F_i] \cdot \Pr[\omega \mid F_i]).$$

Since $F_1, F_2, \ldots, F_k$ make a partition of the sample space, $\sum_{i=1}^{k} (\Pr[F_i] \cdot \Pr[\omega \mid F_i]) = \Pr[\omega]$. Now the last expression above becomes $\sum_{\omega \in \Omega} X(\omega) \cdot \Pr[\omega] = \mathbf{E}[X]$. This completes the proof. □

For a random variable Y, let $r(Y)$ be the range of Y. For each $t \in r(Y)$, define $Y_t = \{\omega \mid Y(\omega) = t\}$. Then $\{Y_t \mid t \in r(Y)\}$ surely makes a partition of the sample space. Recall that the event Y_t can also be written as "$Y = t$". Thus, we have the following corollary for Lemma 8.3.

Corollary 8.4 *Let X and Y be random variables, and let $r(Y)$ be the range of Y. Then*

$$\mathbf{E}[X] = \sum_{t \in r(Y)} \Pr[Y = t] \cdot \mathbf{E}[X \mid Y = t].$$

8.5 Caution on expectation

The expectation of a random variable tells the "average" value of the random variable. If you interpret this as "this is the value I can expect, *very often*, to get at least (or to get at most, depending on your objective)," then you can be very wrong.

Image the following game offered to you by a gambler Mr. X. Suppose you toss six fair coins (at the same time), if all show head, then Mr. X pays you \$100, otherwise, you pay Mr. X \$1. Do you want to play?

Before you make a decision, you probably want to use the expectation to estimate your winning possibilities. There are totally $2^6 = 64$ outcomes, you win \$100 with a probability 1/64 and lose \$1 with a probability 63/64. Let Y be the random variable for the amount you win, then

$$\mathbf{E}[Y] = 100 \cdot \frac{1}{64} + (-1) \cdot \frac{63}{64} = \frac{37}{64} \approx 0.58.$$

Thus, you seems in favor since you would win about \$0.58 "in average", so you should play. On the other hand, if you ask the question "what is my chance to win at least \$0.58?" then you answer is

$$\Pr[Y \geq 0.58] = \frac{1}{64} < 2\%,$$

since the only way for you to win at least \$0.58 is to win \$100. Therefore, in this case, the expectation does not seem to tell you much. A statement based on the expectation such as "you would win about \$0.58 in average" seems quite encouraging, but the statement based on your winning possibility such as "the probability that you win \$0.58 is less than 2%" is kind of discouraging (you can make a statement that is equivalent to this one but sounds even more discouraging, such as "the probability that you lose less than \$1 is smaller than 2%.") . Now if you look at Mr. X's side, the statement based on the expectation as "Mr. X will lose about \$0.58 in average" is discouraging while the statement based on his winning possibility as "Mr. X will win \$1 with a probability larger than 98%" sounds quite attractive.

This seems terrible: the expectation may tell very title information for what we are interested in. Consider the example on the MAX-CUT problem given in subsection 8.3, where we showed that the expected size $\mu = \mathbf{E}[X]$ of a random cut of the given graph $G = (V, E)$ is $m/2$, where m is the number of edges in G, which is at least one half of the size of a maximum cut of the graph. Now the above example on the gambling game shows that it might be possible that even though the "average size" of a random cut is large (at least one half of a maximum cut), most random cuts of the graph G could have very small size. Therefore, by randomly picking a cut for G, even repeatedly, we might have a very small probability to find one that has a large size. In particular, how many times should we run the randomized algorithm in order to find a large cut, with a good probability?

Recall that the sample space for our randomized algorithm for MAX-CUT is the collection of all partitions of the vertex set V of the graph G. Define $E_{\geq m/2}$ to be the event that the number of crossing edges in a partition is at least $m/2$ (i.e., $E_{\geq m/2}$ is the set of those partitions in which the number of crossing edges is at least $m/2$), and $E_{\leq m/2-1}$ to be the event that the number of crossing edges in a partition is at most $m/2-1$. The events $E_{\geq m/2}$ and $E_{\leq m/2-1}$ obviously make a partition of the sample space. Recall that X is the random variable that is equal to the total number of crossing edges in a partition of V. By Lemma 8.3, we have

$$\begin{aligned}
\frac{m}{2} &= \mathbf{E}[X] = \Pr[E_{\geq m/2}] \cdot \mathbf{E}[X \mid E_{\geq m/2}] + \Pr[E_{\leq m/2-1}] \cdot \mathbf{E}[X \mid E_{\leq m/2-1}] \\
&= \Pr[E_{\geq m/2}] \cdot \mathbf{E}[X \mid E_{\geq m/2}] + (1 - \Pr[E_{\geq m/2}]) \cdot \mathbf{E}[X \mid E_{\leq m/2-1}] \\
&\leq \Pr[E_{\geq m/2}] \cdot m + (1 - \Pr[E_{\geq m/2}]) \cdot (\frac{m}{2} - 1). \qquad (21)
\end{aligned}$$

In deriving the last inequality, we have used the fact $X \leq m$ (unconditionally) so $\mathbf{E}[X \mid E_{\geq m/2}] \leq m$, and the fact that under the event $E_{\leq m/2-1}$, $X \leq m/2 - 1$ so $\mathbf{E}[X \mid E_{\leq m/2-1}] \leq m/2 - 1$. Solving the above inequality for $\Pr[E_{\geq m/2}]$, we get

$$\Pr[E_{\geq m/2}] \geq \frac{1}{1 + m/2}.$$

Therefore, by running the randomized algorithm for the MAX-CUT problem $t \cdot m$ times, where t is a reasonably large constant (e.g., $t = 5$), we should have a very hight probability (e.g., $\geq 99.99\%$) that one of these executions produces a cut whose size is at least $m/2$.

The above analysis suggested a randomized algorithm for the MAX-CUT problem, which runs in time $O(m^2)$, and produces, with a high probability, a cut of size at least one half of a maximum cut.

A critical observation on the above analysis is that in the case of the randomized algorithm for the MAX-CUT problem, the random variable X has an upper bound $2 \cdot \mathbf{E}[X]$ on its value (look at the above analysis and convince yourself why this is crucial in deriving the last inequality in (21)), which is not much larger than the expected value $\mathbf{E}[X]$ thus prevents the random variable X from having values far below the expected value $\mathbf{E}[X]$ for a large portion of the outcomes: there would not be sufficient outcomes with large X-values (the X-values are bounded by $2 \cdot \mathbf{E}[X]$) to "balance" this large number of outcomes with small X-values to make the "average" value equal to $\mathbf{E}[X]$.

On the other hand, for the example of the gambling game, the expected value of the random variable Y for your winning strategy is 0.58, while the highest value of the random variable Y is 100, which is much larger than the expected value 0.58. Thus, a single outcome with a large Y-value would allow many outcomes to have their Y-values much smaller than the expected value $\mathbf{E}[Y]$. This explains in that example why the random variable Y has a small probability to reach its expected value.

This observation has been used in the development of the famous *Markov Inequality*, which we will discuss in the next lecture.