EE 126 — Probability and Random Processes

January 18, 2018 — Discrete Random Variables

1 Reading

2.1. Basic Concepts

In a given experiment, perhaps outcomes are not numerical but are associated with some numerical values of interest. A **random variable** is a **real-valued function of the experimental outcome**, and allows us to model such a setup. Specifically, a random variable associates a particular number, known as the **numerical value** or **value**, with each outcome of the experiment. Each value might be any real number. For example, in an experiment involving the transmission of a message, the time needed to transmit the message might become a random variable.

- A **function of a random variable** defines another random variable.
- We can associate with each random variable certain "averages" of interest, e.g. mean and variance.
- A random variable can be **conditioned** on an event or on another random variable.
- A random variable can be **independent** from an event or from another random variable.

A random variable is **discrete** if its range (the set of values it can take) is either finite or countably infinite. Even if the domain is not discrete, a random variable can still be discrete.

- A discrete random variable has an associated **probability mass function (PMF)**, which gives the probability of each numerical value that the random variable can take.
- A **function of a discrete random variable** defines another discrete random variable, whose PMF can be obtained from the PMF of the original random variable.

In general, this chapter consists of taking concepts from Chapter 1 (probabilities, conditioning, independence, etc.) and applying them to random variables rather than events.

2.2. Probability Mass Functions

The PMF p_X of a discrete random variable X is a function mapping X's values to their probabilities. This is the most important way to characterize a random variable.

If x is any possible value of X, the **probability mass** of x (denoted $p_X(x)$) is the probability of the event $\{X = x\}$, which contains all outcomes giving rise to a value of X equal to x. Technically the probability of that event should be written $P(\{X = x\})$, but for brevity we'll usually refer to it as $P(X = x)$.

Note that $\sum_x p_X(x) = 1$, and $P(X \in S) = \sum_{x \in S} p_X(x)$.

Calculation of the PMF for a Random Variable X
For each possible value x of X:

1. Collect all outcomes that give rise to the event $\{X = x\}$.
2. Add their probabilities to obtain $p_X(x)$.

2.2.1. The Bernoulli Random Variable

The **Bernoulli random variable** describes the success or failure in a single trial. Consider a coin toss, where the coin comes up heads with probability p, and tails with probability $1-p$. The Bernoulli random variable takes the value 1 if the outcome is heads and 0 if the outcome is tails. Its PMF is

$$p_X(k) = \begin{cases} p, & \text{if } k = 1 \\ 1-p, & \text{if } k = 0 \end{cases}$$

The Bernoulli random variable is used to model generic probabilistic situations with just two outcomes, such as the state of a telephone at a given time which can be either busy or free. Its mean is $\mathbb{E}[X] = p$, and its variance is $var(X) = p(1-p)$.

2.2.2. The Binomial Random Variable

The **binomial random variable** describes the number of successes in n independent Bernoulli trials. In this setup, a coin is tossed n times and each toss is independent of prior tosses. Let X be the number of heads in the n-toss sequence. We would then refer to X as a binomial random variable with parameters n and p. The PMF consists of binomial probabilities:

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, ..., n$$

The normalization property is written as $\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$.
The binomial random variable's mean is $\mathbb{E}[X] = np$, and its variance is $var(X) = np(1-p)$.

2.2.3. The Geometric Random Variable

The **geometric random variable** describes the number of trials until the first success, in a sequence of independent Bernoulli trials. Here, we repeatedly and independently toss the coin (again with heads probability p). The geometric random variable is the number of tosses X needed for a head to come up for the first time. Its PMF is

$$p_X(k) = (1-p)^{k-1} p, \qquad k = 1, 2, ...$$

since $(1-p)^{k-1}p$ is the probability of the sequence consisting of $k-1$ successive tails followed by a head.
Note that $\sum_{k=1}^{\infty} (1-p)^{k-1} p = 1$.

The geometric random variable's mean is $\mathbb{E}[X] = \frac{1}{p}$, and its variance is $var(X) = \frac{1-p}{p^2}$.

2.2.4. The Poisson Random Variable

The **Poisson random variable** approximates the binomial PMF when n is large, p is small, and $\lambda = np$. A Poisson random variable has a PMF given by

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, ...$$

where λ is a positive parameter characterizing the PMF. Note that $\sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} = 1$.

The Poisson random variable can be a good approximation for a binomial random variable with very small p and very large n (for example, the number of typos in a book with n words). The advantage of using the Poisson PMF comes from potentially simpler models and calculations.

The Poisson random variable's mean is $\mathbb{E}[X] = \lambda$, and its variance is $var(X) = \lambda$.

2.3. Functions of Random Variables

A real-valued function of a random variable defines another random variable. For example, if X is a random variable and Y is of the form $g(X) = aX + b$ (for arbitrary scalars a and b), then Y is a random variable in its own right. In general, if $Y = g(X)$ then we can obtain the PMF $p_Y(y)$ as

$$p_Y(y) = \sum_{\{x \mid g(x)=y\}} p_X(x)$$

2.4. Expectation, Mean, and Variance

The PMF p_X of a random variable X provides us with several numbers: the probabilities of all the possible values of X. Sometimes we might want to summarize this information into a single representative number, that being the **expectation** of X. This is just a weighted average of the possible values of X, and can be viewed as the center of gravity of the PMF. Formally, the expected value is defined as $\mathbb{E}[X] = \sum_x x p_X(x)$.

Aside from the mean, there are other quantities that can be associated with a random variable and its PMF. The **second moment** of the random variable X is defined as the expected value of the random variable X^2, while the **nth moment** is defined as $\mathbb{E}[X^n]$. There is also the **variance** of X, denoted $var(X)$ and defined as the expected value of the random variable $(X - \mathbb{E}[X])^2$:

$$var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

The variance provides a measure of dispersion of X around its mean.
Note that we can also calculate the variance via a method based on the following rule:

Expected Value Rule for Functions of Random Variables
If $g(X)$ is a function of the random variable X, then the expected value of $g(X)$ is $\mathbb{E}[g(X)] = \sum_x g(x) p_X(x)$.

This rule allows us to write the variance of X as $var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_x (x - \mathbb{E}[X])^2 p_X(x)$.
It also allows us to write the nth moment as $\mathbb{E}[X^n] = \sum_x x^n p_X(x)$.

2.4.1. Properties of Mean and Variance

Let X be a random variable, let $Y = aX + b$, and let a and b be [given] scalars. Then

- $\mathbb{E}[Y] = a\mathbb{E}[X] + b$
- $var(Y) = a^2 var(X)$
- $var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

2.4.2. The Discrete Uniform Random Variable

The **discrete uniform random variable** takes one out of a range of contiguous integer values (from a to b) with equal probability. It has a PMF of the form

$$p_X(k) = \begin{cases} \frac{1}{b-a+1}, & \text{if } k = a, a+1, ..., b \\ 0, & \text{otherwise} \end{cases}$$

Its mean is $\mathbb{E}[X] = \frac{a+b}{2}$, and its variance is $var(X) = \frac{(b-a)(b-a+2)}{12}$.

2.5. Joint PMFs of Multiple Random Variables

There are often several random variables associated with the same experiment, sample space, and probability law. Since their values might relate in interesting ways, we would like to consider probabilities involving multiple random variables at the same time. Thus, if we have two discrete random variables X and Y associated with the same experiment, we can say that the probabilities of the values that X and Y can take are captured by the **joint PMF of X and Y**, denoted $p_{X,Y}$. In particular, if (x, y) is a pair of possible values of X and Y, the probability mass of (x, y) is the probability of the event $\{X = x, Y = y\}$.

The joint PMF determines the probability of any event that can be specified in terms of X and Y. For example, if A is the set of all pairs (x, y) that have a certain property, then $P((X, Y) \in A) = \sum_{(x,y)\in A} p_{X,Y}(x, y)$.

Also, the PMFs of X and Y can be calculated using the formulas $p_X(x) = \sum_y p_{X,Y}(x, y)$ and $p_Y(y) = \sum_x p_{X,Y}(x, y)$. p_X and p_Y are known as the **marginal** PMFs. In general, we can calculate the marginal PMFs from the joint PMF by using the **tabular method**. Here, the joint PMF of X and Y is arranged in a 2D table, and the marginal PMF of X or Y at a given value is obtained by adding the table entries along a corresponding column or row.

2.5.1. Functions of Multiple Random Variables

A function $Z = g(X, Y)$ of the random variables X and Y defines another random variable, and its PMF can be calculated from the joint PMF $p_{X,Y}$ according to $p_Z(z) = \sum_{\{(x,y)\ |\ g(x,y)=z\}} p_{X,Y}(x, y)$. Also, the expected value for functions takes the form

$$\mathbb{E}[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y)$$

In the event that g is linear and of the form $aX + bY + c$, we have $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$.

2.5.2. More than Two Random Variables

The joint PMF of three random variables X, Y, and Z is defined in analogy with the above as $p_{X,Y,Z}(x, y, z) = P(X = x, Y = y, Z = z)$ for all possible triplets of values (x, y, z). Corresponding marginal PMFs are obtained by equations such as $p_{X,Y}(x, y) = \sum_z p_{X,Y,Z}(x, y, z)$ and $p_X(x) = \sum_y \sum_z p_{X,Y,Z}(x, y, z)$.

The expectation of a function of three random variables is given by

$$\mathbb{E}[g(X, Y, Z)] = \sum_x \sum_y \sum_z g(x, y, z) p_{X,Y,Z}(x, y, z)$$

2.6. Conditioning

2.6.1. Conditioning a Random Variable on an Event

The **conditional PMF** of a random variable X, conditioned on a particular event A with $P(A) > 0$, is defined by

$$p_{X|A}(x) = P(X = x \mid A) = \frac{P(\{X = x\} \cap A)}{P(A)}$$

Since the events $\{X = x\} \cap A$ are disjoint for different values of x, and their union is A, we have

$$P(A) = \sum_x P(\{X = x\} \cap A)$$

Therefore, $\sum_x p_{X|A}(x) = 1$ and $p_{X|A}$ is a legitimate PMF.

2.6.2. Conditioning One Random Variable on Another

Let X and Y be two random variables associated with the same experiment. If we know that the value of Y is some particular y (with $p_Y(y) > 0$), this may provide partial knowledge about the value of X. Such knowledge is captured by the **conditional PMF** $p_{X|Y}$ of X given Y, which is defined as $p_{X|Y}(x \mid y) = P(X = x \mid Y = y)$. According to the definition of conditional probabilities, we have

$$p_{X|Y}(x \mid y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

The conditional PMF of X given $Y = y$ can be used to calculate the joint PMF via

$$p_{X,Y}(x, y) = p_Y(y) p_{X|Y}(x \mid y)$$

This can be extended to the case of three or more random variables as

$$p_{X,Y,Z}(x, y, z) = p_Z(z) p_{Y|Z}(y \mid z) p_{X|Y,Z}(x \mid y, z)$$

Also, the conditional PMF of X given Y can be used to calculate the marginal of X via the total probability formula

$$p_X(x) = \sum_y p_{X,Y}(x, y) = \sum_y p_Y(y) p_{X|Y}(x \mid y)$$

Finally, the above can be extended to compute expected values using the total expectation theorem:

$$\mathbb{E}[X] = \sum_y p_Y(y) \mathbb{E}[X \mid Y = y]$$

Summary of Facts About Conditional PMFs
Let X and Y be random variables associated with the same experiment.

- Conditional PMFs are similar to ordinary PMFs, but pertain to a universe where the conditioning event is known to have occurred.
- If $A_1, ..., A_n$ are disjoint events that form a partition of the sample space, with $P(A_i) > 0$ for all i, then

  $$p_X(x) = \sum_{i=1}^{n} P(A_i) p_{X|A_i}(x)$$

  Furthermore, for any event B, with $P(A_i \cap B) > 0$ for all i, we have

  $$p_{X|B}(x) = \sum_{i=1}^{n} P(A_i|B) p_{X|A_i \cap B}(x)$$

2.6.3. Conditional Expectation

Summary of Facts About Conditional Expectations
Let X and Y be random variables associated with the same experiment.

- The conditional expectation of X given an event A with $P(A) > 0$ is defined by $\mathbb{E}[X|A] = \sum_x x p_{X|A}(x)$. For a function $g(X)$, we have $\mathbb{E}[g(X)|A] = \sum_x g(x) p_{X|A}(x)$.
- The conditional expectation of X given a value y of Y is defined by $\mathbb{E}[X|Y=y] = \sum_x x p_{X|Y}(x|y)$.
- If $A_1, ..., A_n$ are disjoint events that form a partition of the sample space, with $P(A_i) > 0$ for all i, then

  $$\mathbb{E}[X] = \sum_{i=1}^{n} P(A_i)\mathbb{E}[X|A_i]$$

  Furthermore, for any event B, with $P(A_i \cap B) > 0$ for all i, we have

  $$\mathbb{E}[X|B] = \sum_{i=1}^{n} P(A_i|B)\mathbb{E}[X|A_i \cap B]$$

- We have $\mathbb{E}[X] = \sum_y p_Y(y)\mathbb{E}[X|Y=y]$.

2.7. Independence

2.7.1. Independence of a Random Variable from an Event

For a random variable X to be independent of the event A, knowing the occurrence of the conditioning event A should provide no new information on the value of the random variable X:

$$P(X = x \text{ and } A) = P(X = x)P(A) = p_X(x)P(A), \quad \text{for all } x$$

(In other words, the two events $\{X = x\}$ and A should be independent for any choice of x.)

From the definition of the conditional PMF, we have $P(X = x \text{ and } A) = p_{X|A}(x)P(A)$. Thus, as long as $P(A) > 0$, independence is the same as the condition $p_{X|A}(x) = p_X(x)$ for all x (i.e. the PMFs p_X and $p_{X|A}$ are the same).

2.7.2. Independence of Random Variables

Two random variables X and Y are independent if $p_{X,Y} = p_X(x)p_Y(y)$ for all x, y. This is the same as requiring that the two events $\{X = x\}$ and $\{Y = y\}$ be independent for all pairs (x, y). Also, the formula $p_{X,Y}(x, y) = p_{X|Y}(x \mid y)p_Y(y)$ shows that independence is equivalent to the condition "$p_{X|Y}(x \mid y) = p_X(x)$ for all x and all y with $p_Y(y) > 0$."

X and Y are **conditionally independent**, given a positive probability event A, if

$$P(X = x, Y = y \mid A) = P(X = x \mid A)P(Y = y \mid A) \text{ for all } x \text{ and } y$$

or, in other words, $p_{X,Y|A}(x, y) = p_{X|A}(x)p_{Y|A}(y)$ for all x and y. This is also equivalent to $p_{X|Y,A}(x \mid y) = p_{X|A}(x)$ for all x and y such that $p_{Y|A}(y) > 0$.

Conditional independence may not imply unconditional independence.

If X and Y are independent random variables, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Also, $var(X + Y) = var(X) + var(Y)$.

Functions of independent random variables are also independent. $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$.

2.7.3. Sum of Independent Random Variables

If $X_1, X_2, ..., X_n$ are independent random variables, then

$$var(X_1 + X_2 + ... + X_n) = var(X_1) + var(X_2) + ... + var(X_n)$$

(Though it does not require independence of the X_i, it is also true that $\mathbb{E}[X_1 + ... + X_n] = \mathbb{E}[X_1] + ... + \mathbb{E}[X_n]$.)

2 Lecture

Discrete Random Variables

A random variable on a sample space Ω is a function $X : \Omega \mapsto \mathbb{R}$. If $\text{Im}(X) = \{X(\omega) \mid \omega \in \Omega\}$ is finite or countable, then the random variable is discrete.

Example: toss a fair coin three times. Let $X =$ the number of heads. Then X is a random variable. For example, $X(HHT) = 2$, while $X(TTT) = 0$. Also, $P(X = 2) = P(\{\omega \in \Omega \mid X(\omega) = 2\}) = P(HHT, HTH, THH) = 3/8$.

Probability Mass Functions

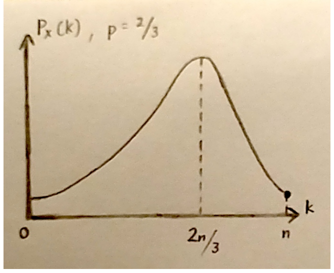
A probability mass function (PMF) must assign probabilities to each possible value of a random variable. Let X be a discrete random variable. Then $p_X(k) = P(X = k),\ \ k \in \text{Im}(X)$.

Example: flip a coin with bias p. Let $X = 1$ if heads, and 0 if tails. Then $p_X(k) = p$ if $k = 1$ and $(1 - p)$ if $k = 0$. Such is an example of a **Bernoulli random variable**. We might say that $X \sim$ Bernoulli(p).

Binomial Distribution

Binomial random variable: toss a coin with bias p, n times. We are then interested in $X =$ the number of heads $= X_1 + X_2 + ... + X_n$, where each $X_i \sim$ Bernoulli(p).

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$



A binomial PMF with $p = 2/3$. It is left-skewed, since $p > 1/2$ and we therefore expect more heads (1) than tails (0).

Poisson Distribution

The **Poisson distribution** arises in the following setting: We know in advance that in the time interval from 2-3pm, $\lambda = 5$ customers *on average* will go to the bank. We want to define a random variable X to be the number of customers that indeed went to the bank.

$$p_X(k) = \text{(empirically)}\ \ e^{-\lambda} \frac{\lambda^k}{k!},\ \ k = 0, 1, 2, ...$$

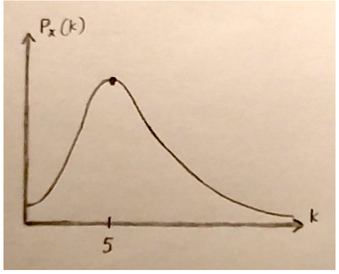
Say we split the total time interval into n subintervals:

$$[0,1) = \left[0, \frac{1}{n}\right) \cup \left[\frac{1}{n}, \frac{2}{n}\right) \cup ... \cup \left[\frac{n-1}{n}, 1\right) = I_1 \cup I_2 \cup ... \cup I_n$$

In each I_j, either 1 event or 0 events will occur. (The intervals are so extremely small that we can make the assumption that only one event is able to occur in each of them.) On average, $\frac{\lambda}{n} = 1 \cdot p + 0 \cdot (1-p)$ events will occur. Notice that since each result is either 1 or 0, our setup is analogous to flipping a coin. Thus we can let $X_i \sim \text{Bernoulli}(p)$ and define what we're interested in as $S_n = X_1 + X_2 + ... + X_n \sim \text{Binomial}(n, p)$. ($S_n$ represents the number of customers that came in total, over the entire interval.) We know that

$$p_{S_n}(k) = \binom{n}{k} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \xrightarrow{n\to\infty} e^{-\lambda}\frac{\lambda^k}{k!}$$

Here, in the case of the Poisson random variable, the sample space consists of binary sequences of infinite length, e.g. 0010101..., where $X(0010101...)$ might equal 4.. Each 1 signals that a customer showed up in that infinitesimal interval. Note that the probability of seeing a customer in any of these minuscule subintervals is almost 0.



A Poisson PMF, assuming $\lambda = 5$. The probability of infinite customers is 0 (because as k goes to ∞, $p_X(k)$ goes to 0).

Expectation

The PMF is a function which only needs to satisfy two requirements: $p_X(k) \geq 0$ for all k, and $\sum_k p_X(k) = 1$. However, notice that if k lives in the real numbers, we need infinite values to describe the full PMF! (Even if the random variable is discrete, we'll probably still need a lot of numbers.) Since the PMF is so complex, we'd like to create a more concise summary of the function. One possible summary is the **expectation**, a real number whose goal is to describe well the whole function. The expectation of a discrete random variable X is $\mathbb{E}[X] = \sum_k k \cdot p_X(k)$.

Obviously, on its own the expectation is not enough to describe the function very adequately. We can hardly distinguish an entire PMF from just an expectation, which might be shared by many different distributions. Therefore we make our summary richer by embedding it with one more quantity: the **variance**. The variance encodes the spread of the distribution, and is defined $var(X) = \sum_k (k - \mathbb{E}[x])^2 \cdot p_X(k)$.

Generally speaking, we introduce these real numbers in order to enrich our summary about the function. Using them, we can distinguish more random variables!

Example. $X \sim \text{Uniform}\{0, 1, ..., n\}$. In this case, we have

$$\mathbb{E}[X] = 0 \cdot \frac{1}{n+1} + 1 \cdot \frac{1}{n+1} + ... + n \cdot \frac{1}{n+1} = \frac{n}{2}$$

Note: we can interpret the expectation as the "center of gravity." Also, according to the Law of Large Numbers, the sample average should converge to the expectation:

$$\frac{X_1 + X_2 + ... + X_n}{n} \to \mathbb{E}[x]$$

Below are some properties of expectation. Assume that $a, b \in \mathbb{R}$ and that X, Y are random variables.

- $\mathbb{E}[a] = a$
- **Linearity of expectation:** $\mathbb{E}[aX + bY] = a\mathbb{E}[x] + b\mathbb{E}[y]$
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if X and Y are independent or the covariance is 0
- $\mathbb{E}[g(X) \cdot h(X)] = \mathbb{E}[g(X)]\mathbb{E}[h(X)]$
- $\mathbb{E}[g(X)] = \mathbb{E}[Y] = \sum_l l \cdot p_Y(l) = \sum_k g(k)p_X(k)$
  - Note that we let $g(X) =$ the random variable Y.
  - This property is useful if calculating $p_Y(l)$ is hard, or if calculating $p_X(k)$ is easy.

Example. $X \sim \text{Binomial}(n, p)$. Thus $X = X_1 + ... + X_n,\ \ X_i \sim \text{Bernoulli}(p)$, and $\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i] = np$.

Example. $X \sim \text{Poisson}(\lambda)$. For modeling to be accurate, $\mathbb{E}[X] = \lambda$. Since $p_X(k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$, we have

$$\sum_{k=0}^{+\infty} k \cdot e^{-\lambda} \frac{\lambda^k}{k!} = \lambda$$

Fundamentally (and due to a Taylor expansion), $e^x = \sum_{k=0}^{+\infty} \frac{x^k}{k!}$.

Joint PMFs

In this case we have two random variables over the same experiment. Say that X, Y are two discrete random variables over the same experiment. Then $p_{X,Y}(k, l) = P(X = k \text{ and } Y = l)$. Now the random variables are able to interact, to talk with each other... and from this we are able to derive useful information.

Example. There is a family, which made the following agreement: the family will stop giving birth to kids either when they have a girl or when they have three kids. Let $G =$ number of girls and $C =$ number of children. The joint PMF $p_{G,C}(k, l)$ can be expressed as a table:

$p_{G,C}(k, l)$	**1**	**2**	**3**
0	0	0	1/8
1	1/2	1/4	1/8

Joint PMF properties:

- $p_{X,Y}(k, l) \geq 0$
- $\sum_k \sum_l p_{X,Y}(k, l) = 1$

If we have access to the joint $p_{X,Y}(k, l)$, can we extract the individual PMFs? **Yes** – by marginalizing out any undesired variable. In the previous example, $P(G = 1) = 1/2 + 1/4 + 1/8$. The **marginals** are $p_X(k) = \sum_l p_{X,Y}(k, l)$ and $p_Y(l) = \sum_k p_{X,Y}(k, l)$. In short, we *can* make the conversion from the joint to the marginals.

Can we then also make the conversion from the marginals to the joint? **No.** If each random variable were able to take on n values, the marginals would involve about $2n$ entries (i.e. one probability for each value). However, the joint would be a table with n^2 entries! Since $n^2 > 2n$, we simply wouldn't have sufficient information to reconstruct the joint from the marginals (in a unique, non-underdetermined fashion). But as a special case, if we know that the random variables are independent then we can say that the joint is the *product* of the marginals.

Deriving Linearity of Expectation

We can derive [one aspect of] linearity of expectation using the joint. $\mathbb{E}[X+Y]$ can be defined as

$$\mathbb{E}[X+Y] = \sum_k \sum_l (k+l) \cdot p_{X,Y}(k,l)$$

Starting from this definition, we will derive $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_k \sum_l k \cdot p_{X,Y}(k,l) + \sum_l \sum_k l \cdot p_{X,Y}(k,l) \\
&= \sum_k k \cdot \boxed{\sum_l p_{X,Y}(k,l)} + \sum_l l \cdot \boxed{\sum_k p_{X,Y}(k,l)} \\
&= \sum_k k \cdot \boxed{p_X(k)} + \sum_l l \cdot \boxed{p_Y(l)} \\
&= \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}$$

References

[1] D.P. Bertsekas and J.N. Tsitsiklis. *Introduction to Probability.* Athena Scientific books. Athena Scientific, 2002.