

07-380 Notes

Pre-reading: Bayes Nets

Carnegie Mellon University
School of Computer Science

Contents

1	Probability Background (Assumed)	1
2	Notation and Conventions	2
2.1	Random Variables, Outcomes, and Values	2
2.2	Capital Letters Mean Tables	3
2.3	Discrete Distributions: Bernoulli and Categorical	3
3	Omega Pizzeria!	4
3.1	Partitions	4
3.2	Marginal Distribution	5
3.3	Joint Distribution	5
3.4	Conditional Distribution	6
3.5	When Does a Probability Table Sum to One?	7
4	Answering Any Query from a Joint Distribution	7
4.1	Joint Distributions Are the Best!	7
4.2	Example: Season, Temperature, and Weather	8
4.3	The General Recipe	9
5	Building a Joint Distribution with the Chain Rule	9
5.1	Two Problems with Joint Distributions	9
5.2	The Chain Rule	10
5.3	Caution: Multiplying Probability Tables!	10
5.4	From Conditional Probability Tables to Any Query	11
6	Bayes Nets	12
6.1	Nodes, Edges, and Tables	12
6.2	Building a Bayes Net from the Chain Rule	13
6.3	Missing Edges Encode Independence Assumptions	13
6.4	The Alarm Network	14
7	Independence Timeout	15
7.1	Independence	16
7.2	Conditional Independence	17
8	Bayes Net Triples	17
8.1	Causal Chain	18
8.2	Common Cause	18
8.3	Common Effect	19
8.4	The Three Triples, Side by Side	19
9	Summary	20

1 Probability Background (Assumed)

These notes build on the probability that you have already seen in 07-280 and in the probability course that is a prerequisite for this class. We will use the following definitions and rules without re-deriving them. If any of them look unfamiliar, review the [07-280 Probability Background](#) notes and the [probability reference sheet](#) before continuing.

Tool	Rule
Definition of conditional probability	$P(A B) = \frac{P(A, B)}{P(B)}$
Product rule	$P(A, B) = P(A B) P(B) = P(B A) P(A)$
Bayes' theorem	$P(B A) = \frac{P(A B) P(B)}{P(A)}$
Chain rule	$P(X_1, \dots, X_N) = \prod_{i=1}^N P(X_i X_1, \dots, X_{i-1})$
Marginalization (law of total probability)	$P(A) = \sum_b \sum_c P(A, b, c)$
Normalization	$P(B a) = \frac{P(a, B)}{P(a)} = \frac{1}{z} P(a, B), \quad z = P(a) = \sum_b P(a, b)$

All of these rules are related to each other. The product rule is the definition of conditional probability with both sides multiplied by $P(B)$, Bayes' theorem is the product rule written both ways and rearranged, and the chain rule is the product rule applied repeatedly. We will return to marginalization and normalization in Section 4 and to the chain rule in Section 5.

In this pre-reading, we use these rules on whole tables of probabilities rather than on one probability at a time, and we organize the tables with a Bayes net.

2 Notation and Conventions

2.1 Random Variables, Outcomes, and Values

There is a lot of notation in probability, and we will try to be consistent about ours. Consider the expression

$$P(B = +b, C) = \sum_{a \in \{a_1, a_2, a_3\}} P(A = a, B = +b, C),$$

which in this course we will almost always write in the shorthand

$$P(+b, C) = \sum_a P(a, +b, C).$$

Three kinds of symbols appear in these expressions:

- **Random variables** are capital letters, such as A , B , and C . A random variable stands for *all* of its potential outcomes at once.
- **Outcomes**, also called **values**, are lowercase, such as a_1 , a_2 , a_3 for the three outcomes of A . For a **binary** random variable such as B , we write its two outcomes as $+b$ and $-b$. This is not a common convention outside of this course, but it is a convenient shorthand for yes/no or true/false, and we use it constantly.
- **Variables for values** are lowercase letters with no subscript or sign, such as a . The letter a stands for one unspecified outcome of A , and it is what we loop over in a summation: $\sum_a$ plugs in a_1 , then a_2 , then a_3 .

The shorthand above uses two more conventions. First, we drop the " $A =$ " and write just the value: $P(a, +b, C)$ means $P(A = a, B = +b, C)$. Second, the commas mean **and**: $P(a, +b, C)$ is the probability that A takes value a *and* B takes value $+b$ *and* whatever we say about C .

The conditioning bar: In $P(A, B | C, D)$, everything to the right of the bar is **given**, and everything to the left is what we are asking about. It can help to picture the bar as a wall: this

expression is the probability of A and B , given both C and D . Other groupings, such as “ A , and then B given C ”, do not make sense.

2.2 Capital Letters Mean Tables

The most important convention in these notes is what a capital letter inside $P(\cdot)$ means. The probability $P(a_1)$ is one number. The expression $P(A)$, with a capital A , is *not* one number. It is a **probability table** with one entry for every outcome of A :

$$P(A) \quad \text{means} \quad \begin{array}{c|c} A & P(A) \\ \hline a_1 & P(a_1) \\ a_2 & P(a_2) \\ a_3 & P(a_3) \end{array}$$

The same goes for every capital letter on either side of the bar. With $A \in \{a_1, a_2, a_3\}$, $B \in \{+b, -b\}$, and $C \in \{+c, -c\}$:

- $P(A, B \mid C)$ is a table with $3 \times 2 \times 2 = 12$ entries, one for every combination of an outcome of A , an outcome of B , and an outcome of C .
- $P(A, +b \mid C)$ is a table with $3 \times 2 = 6$ entries. The value $+b$ is fixed, so only A and C vary.
- $P(a_2, +b \mid +c)$ is a single number.

The size of a table is the product of the number of outcomes of each capital letter in it. Whenever you see capital letters, ask yourself how many entries the table has. Not every table sums to one; Section 3.5 says which ones do.

2.3 Discrete Distributions: Bernoulli and Categorical

Every random variable in this pre-reading, and in the Bayes nets we study in this course, is **discrete**: it takes on one of a finite list of outcomes. A binary random variable, such as $B \in \{+b, -b\}$, has a **Bernoulli distribution**, which is described by a single probability $P(+b)$ (since $P(-b) = 1 - P(+b)$). A random variable with K outcomes, such as $A \in \{a_1, a_2, a_3\}$, has a **categorical distribution**, described by K probabilities that sum to one. We will rarely use these names. We will just say “ B is binary” or “ A takes three values” and write the probabilities in a table.

Sticking with discrete random variables has several advantages:

- We work with **probabilities** rather than densities, so every entry in every table is a number between 0 and 1.
- We **sum** rather than integrate, so marginalizing a variable out means adding up a few table entries.
- It is easy to see when a distribution sums to one: add up the entries of the table.
- It is easy to see how big a problem is. A random variable with d outcomes needs a table of d entries. A joint distribution over N random variables that each have d outcomes needs a table of d^N entries, and every extra variable multiplies the size of the table by d . Keeping track of these sizes is what motivates Bayes nets.

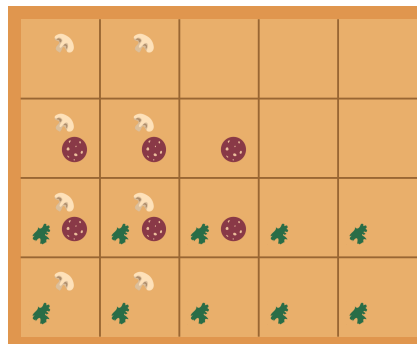
Bayes nets can also be built from continuous random variables, but we will not do that in this course.

3 Omega Pizzeria!

We will use the pizza below as a running example. It has 20 slices, and each slice may have mushrooms, spinach, and/or pepperoni on it. We are going to pick one slice uniformly at random, so every slice has probability $1/20$.

We call it the *Omega* Pizzeria because the whole pizza is our probability space, the set Ω of everything that can happen. What is the probability of getting a slice with

1. no mushrooms?
2. spinach and no mushrooms?
3. spinach, when we asked for a slice with no mushrooms?



Icons: CC, <https://openclipart.org/detail/296791/pizza-slice>

The first two are answered by counting slices out of 20. The third is a conditional probability, and the world we are counting in has shrunk to the 12 slices with no mushrooms. This section is about organizing questions like these into probability tables.

We define three binary random variables, one per topping:

$$M \in \{m_1, m_2\}, \quad S \in \{s_1, s_2\}, \quad R \in \{r_1, r_2\},$$

where m_2 means the slice has mushrooms and m_1 means it does not, s_2 and s_1 are the same for spinach, and r_2 and r_1 are the same for pepperoni. (The letter P is already taken by probability, so pepperoni gets R .) We could equally well have written $+m$ and $-m$; both conventions show up in this course.

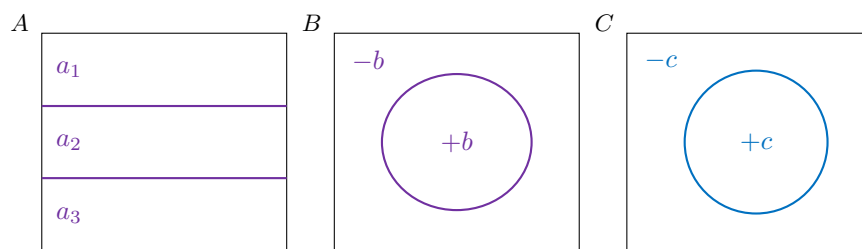
3.1 Partitions

Each random variable carves up the probability space. The outcomes of a random variable are

- discrete: there is a finite list of them,
- disjoint: exactly one of them happens, and
- exhaustive: together they account for the entire event space.

A collection of sets with these properties is called a **partition**. The variable M partitions the pizza into the 12 slices without mushrooms and the 8 slices with mushrooms, and every slice is in exactly one of the two parts. Partitions are not always binary. A random variable A for a person's age group could take three values, a_1 for kids, a_2 for adults, and a_3 for senior citizens, and everyone falls in exactly one of them.

The pictures below show three random variables partitioning the same event space Ω (the rectangle). A splits it into three pieces, and the binary variables B and C each split it into a piece inside a circle and a piece outside.



Because the outcomes of a random variable form a partition, their probabilities sum to one:

$$P(a_1) + P(a_2) + P(a_3) = 1, \quad P(+b) + P(-b) = 1, \quad P(+c) + P(-c) = 1.$$

3.2 Marginal Distribution

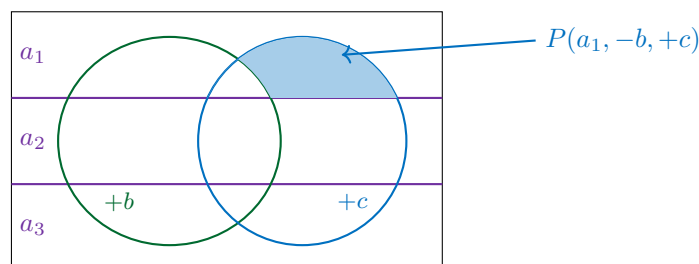
A **marginal distribution** is the distribution of one random variable on its own. It is the simplest probability table. $P(A)$ has three entries, $P(a_1)$, $P(a_2)$, and $P(a_3)$, one per piece of the partition; each is between 0 and 1, and they sum to one. For the pizza, the marginal distribution of mushrooms is

M	$P(M)$
m_1	$12/20$
m_2	$8/20$

since 12 of the 20 slices have no mushrooms and 8 do.

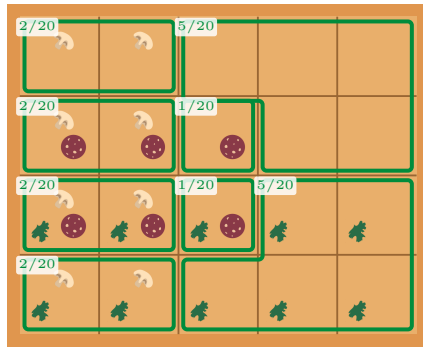
3.3 Joint Distribution

A **joint distribution** is the distribution of a collection of random variables considered all at once. If we draw the partitions for A , B , and C on top of each other, the event space is cut into $3 \times 2 \times 2 = 12$ pieces, and the joint distribution $P(A, B, C)$ is the table of the 12 probabilities of those pieces.



The shaded piece is the part of the top strip (a_1) that is inside the C circle ($+c$) but outside the B blob ($-b$). Its probability is the single entry $P(a_1, -b, +c)$ of the table. In a probability class, you spend most of your time computing one such entry: the probability of this particular hand of cards, or that particular sequence of coin flips. In this course, we almost always want the *whole* table, all 12 numbers, because we are going to hand the table to a program and ask it questions.

For the pizza, the joint distribution $P(M, S, R)$ has $2 \times 2 \times 2 = 8$ entries, one for each combination of toppings, and we can fill in every entry by counting slices:

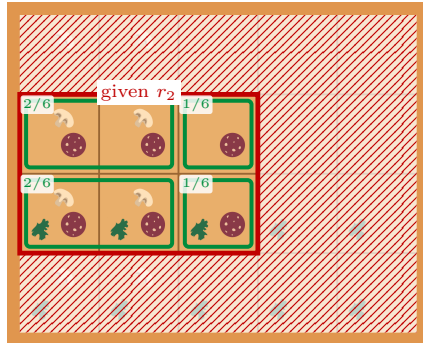


M	S	R	$P(M, S, R)$
m_1	s_1	r_1	$5/20$
m_1	s_1	r_2	$1/20$
m_1	s_2	r_1	$5/20$
m_1	s_2	r_2	$1/20$
m_2	s_1	r_1	$2/20$
m_2	s_1	r_2	$2/20$
m_2	s_2	r_1	$2/20$
m_2	s_2	r_2	$2/20$

For example, $P(m_1, s_1, r_1) = 5/20$ because 5 slices have no toppings at all, and $P(m_2, s_2, r_2) = 2/20$ because 2 slices have all three. The eight entries sum to $20/20 = 1$, as they must: the eight combinations partition the pizza.

3.4 Conditional Distribution

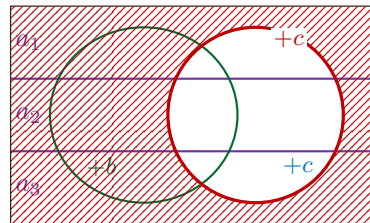
A **conditional distribution** is a distribution inside a world that has been given to us. Take $P(M, S \mid r_2)$, the distribution of mushrooms and spinach given that the slice has pepperoni. The given world is the 6 slices with pepperoni. We ignore the other 14 slices entirely and ask about the combinations of M and S inside those 6:



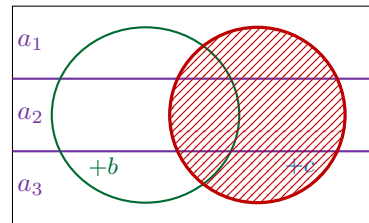
M	S	$P(M, S \mid r_2)$
m_1	s_1	$1/6$
m_1	s_2	$1/6$
m_2	s_1	$2/6$
m_2	s_2	$2/6$

For example, $P(m_1, s_1 \mid r_2) = 1/6$: of the 6 pepperoni slices, exactly one has neither mushrooms nor spinach. Compare this with the joint entry $P(m_1, s_1, r_2) = 1/20$ from Section 3.3. It is the same slice with a different denominator, because the world we are counting in has shrunk from the whole pizza to the pepperoni slices. The four entries of $P(M, S \mid r_2)$ sum to one, because inside the given world the four combinations of M and S form a partition.

The Venn diagram shows the same idea. To draw $P(A, B \mid +c)$ we zoom in on the world where $C = +c$, the inside of the C circle, and hatch out everything else. Inside that world, the partitions for A and B cut the circle into $3 \times 2 = 6$ pieces, and their six probabilities sum to one.



$P(A, B \mid +c)$: six entries, sum to one



$P(A, B \mid -c)$: six entries, sum to one

A capital letter on the right of the bar: What about $P(A, B \mid C)$? By the convention of Section 2.2, this is a table with all 12 combinations: the six entries of $P(A, B \mid +c)$ and the six entries of $P(A, B \mid -c)$. It is really two separate distributions stacked in one table, one for each given world, and the whole table sums to 2, not to 1. A table like this is called a **conditional probability table** (CPT), and Bayes nets are built out of them.

Why would we want both halves? Think about a hospital rather than a single patient. Suppose C is whether a patient has cardiac disease, and A and B are an age group and a blood pressure reading. A new patient walks in, and we run the test for C . If it comes back $+c$ we use the top half of the table, and if it comes back $-c$ we use the bottom half. We want the whole table ready to go before we know which patient walks in.

3.5 When Does a Probability Table Sum to One?

We have now seen tables that sum to one, $P(A)$, $P(M, S, R)$, and $P(A, B \mid +c)$, and one that does not, $P(A, B \mid C)$. The rule, which is also stated on the [probability reference sheet](#), is that a probability table is guaranteed to sum to one exactly when

1. there is exactly one specific combination of outcomes to the right of the bar (or nothing at all), so that we are inside exactly one given world, and
2. there are only capital letters to the left of the bar, so that we are considering *all* the pieces of the partition inside that world.

Another way to say it: no capital letters on the right, and no lowercase letters on the left.

Example: On the pizza, $P(M \mid r_2)$ sums to one: inside the pepperoni world, $P(m_1 \mid r_2) = 2/6$ and $P(m_2 \mid r_2) = 4/6$. The table $P(M, r_2)$ does not: its entries $P(m_1, r_2) = 2/20$ and $P(m_2, r_2) = 4/20$ sum to $P(r_2) = 6/20$, because r_2 is a fixed value on the left of the bar. The table $P(m_2 \mid R)$ is not a distribution at all: it is one number from each of two different worlds, $P(m_2 \mid r_1) = 4/14$ and $P(m_2 \mid r_2) = 4/6$, and there is no reason for those two numbers to sum to anything in particular.

A related warning: an expression such as $\sum_b P(a_1 \mid b) = P(a_1 \mid +b) + P(a_1 \mid -b)$ adds numbers from two different worlds, and it does not mean anything. We can sum over a variable on the left of the bar, as in $\sum_a P(a, +b \mid +c) = P(+b \mid +c)$, but never over a variable on the right.

4 Answering Any Query from a Joint Distribution

4.1 Joint Distributions Are the Best!

If someone hands us the joint distribution over all of the random variables in a problem, we can answer *any* question about those variables: any marginal, any conditional, any combination of the two. We only need two of the tools from Section 1, the definition of conditional probability and marginalization, plus the ability to look numbers up in the table and add them.

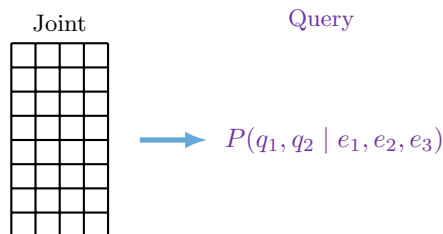


Illustration: Ketrina Yim / UC Berkeley AI

We write a **query** as $P(Q \mid e)$: the distribution of the **query variables** Q given specific values e of the **evidence variables** E . Any remaining variables in the joint are **hidden variables** H ; they appear in the joint table but not in the question, and we will have to sum them out. Each of Q , E , and H may stand for several random variables.

4.2 Example: Season, Temperature, and Weather

Here is a joint distribution over three binary random variables, the season $S \in \{\text{summer, winter}\}$, the temperature $T \in \{\text{hot, cold}\}$, and the weather $W \in \{\text{sun, rain}\}$. It has $2 \times 2 \times 2 = 8$ entries, and they sum to one.

Season	Temp	Weather	$P(S, T, W)$
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

We will compute three tables from it: $P(W)$, $P(W \mid \text{winter})$, and $P(W \mid \text{winter, hot})$. Notice that “what is the probability of weather?” is not a question with a single answer. $P(W)$ is a table with two entries, $P(\text{sun})$ and $P(\text{rain})$, so we compute the entries one at a time.

$P(W)$: To get $P(W = \text{sun})$ we marginalize out the season and the temperature, which means adding up every row of the joint table with weather equal to sun:

$$P(\text{sun}) = \sum_s \sum_t P(s, t, \text{sun}) = 0.30 + 0.10 + 0.10 + 0.15 = 0.65.$$

The other entry is $P(\text{rain}) = 0.05 + 0.05 + 0.05 + 0.20 = 0.35$, and the two sum to one.

$P(W \mid \text{winter})$: Now only the winter rows matter. By the definition of conditional probability,

$$P(\text{sun} \mid \text{winter}) = \frac{P(\text{winter, sun})}{P(\text{winter})}.$$

Neither the numerator nor the denominator is an entry of the joint table, so we *un-marginalize* each of them: we write each as a sum over the variables it is missing, until every term is an entry of the table.

$$\begin{aligned} P(\text{sun} \mid \text{winter}) &= \frac{\sum_t P(\text{winter}, t, \text{sun})}{\sum_t \sum_w P(\text{winter}, t, w)} \\ &= \frac{0.10 + 0.15}{0.10 + 0.05 + 0.15 + 0.20} = \frac{0.25}{0.50} = 0.5. \end{aligned}$$

The same steps give $P(\text{rain} \mid \text{winter}) = 0.25/0.50 = 0.5$.

$P(W \mid \text{winter, hot})$, **using the normalization trick**: Only two rows of the table have both winter and hot, and the two entries of $P(W \mid \text{winter, hot})$ are proportional to those two joint entries:

$$\begin{aligned} P(\text{sun} \mid \text{winter, hot}) &\propto P(\text{winter, hot, sun}) = 0.10, \\ P(\text{rain} \mid \text{winter, hot}) &\propto P(\text{winter, hot, rain}) = 0.05. \end{aligned}$$

Since the two entries must sum to one, we divide each by the sum of the two:

$$P(\text{sun} \mid \text{winter, hot}) = \frac{0.10}{0.10 + 0.05} = \frac{2}{3}, \quad P(\text{rain} \mid \text{winter, hot}) = \frac{0.05}{0.10 + 0.05} = \frac{1}{3}.$$

The normalization trick is not a new rule. The denominator $0.10+0.05$ is exactly $P(\text{winter, hot})$ written as a sum over w , so this is the definition of conditional probability followed by un-marginalizing the denominator, and $1/z = 1/0.15$ is the normalization constant from Section 1. The trick is that we never have to compute the denominator separately: it is the sum of the entries we already computed for the numerator.

You may be able to get these numbers by intuition. That is fine, but make sure you can also get them by following the rules step by step. Probability is notorious for answers that seem to make sense and are wrong, and, more importantly, we want a computer to do these calculations for us, which means we have to be able to write down the steps.

4.3 The General Recipe

Putting the example together, here is how to answer any query $P(Q \mid e)$ from a joint distribution $P(H, Q, E)$:

1. **Definition of conditional probability:** $P(Q \mid e) = \frac{P(Q, e)}{P(e)}$.
2. **Un-marginalize the numerator and the denominator** so that every term is an entry of the joint: sum over the hidden variables in the numerator, and over the hidden *and* query variables in the denominator,

$$P(Q, e) = \sum_h P(h, Q, e),$$

$$P(e) = \sum_q \sum_h P(h, q, e).$$

3. **Look up and add** the joint entries. If Q is a table, do this once per entry, or compute the numerator table $P(Q, e)$ and normalize it: $P(Q \mid e) = \frac{1}{z} P(Q, e)$ with $z = \sum_q P(q, e)$.

These three steps answer any query, and a program can follow them mechanically. The catch is the joint table itself, which is the subject of the rest of these notes.

5 Building a Joint Distribution with the Chain Rule

5.1 Two Problems with Joint Distributions

Joint distributions are the best, but there are two problems with them.

- **Nobody hands us the joint.** No one collects the probability that a patient is a kid *and* has high blood pressure *and* has cardiac disease, for every combination of every variable. What people do collect, and what we are usually handed, are **conditional probability tables**: the probability of high blood pressure given cardiac disease, and so on.
- **The joint is huge.** As we counted in Section 2.3, N random variables with d values each give a joint table with d^N entries. With 30 binary variables that is over a billion numbers.

The plan for the rest of these notes is to take the conditional probability tables we *are* given, multiply them together with the chain rule to get the joint, and then answer queries from the joint

as in Section 4. A Bayes net is a way of writing down which conditional probability tables we have and how they multiply together.

5.2 The Chain Rule

The product rule writes a joint distribution over two variables as a product of a conditional and a marginal, and it can be written in either order:

$$P(A, B) = P(A | B) P(B) = P(B | A) P(A).$$

Applying the product rule over and over gives the **chain rule**. Order the variables in some way, $X_1, X_2, \dots, X_N$; then

$$\begin{aligned} P(X_1, X_2, \dots, X_N) &= P(X_1) P(X_2 | X_1) P(X_3 | X_1, X_2) \cdots P(X_N | X_1, \dots, X_{N-1}) \\ &= \prod_{i=1}^N P(X_i | X_1, \dots, X_{i-1}). \end{aligned}$$

Each variable is conditioned on all the variables that came before it in the ordering. The chain rule holds for *any* ordering, with no assumptions, so a joint distribution over three variables can be written in $3! = 6$ ways:

$$\begin{aligned} P(A, B, C) &= P(A) P(B | A) P(C | A, B) && \text{for the ordering } A, B, C \\ P(A, B, C) &= P(A) P(C | A) P(B | A, C) && \text{for the ordering } A, C, B \\ P(A, B, C) &= P(C) P(B | C) P(A | B, C) && \text{for the ordering } C, B, A \\ &\vdots \end{aligned}$$

Which ordering should we use? The one whose factors are the tables we actually have.

Example: Three binary random variables: whether there is a fire, F ; whether there is smoke, S ; and whether the smoke alarm goes off, A . With the ordering F, S, A , the chain rule says

$$P(F, S, A) = P(F) P(S | F) P(A | F, S).$$

If we are handed those three tables, we can multiply them together to get the joint.

5.3 Caution: Multiplying Probability Tables!

We are going to write things like

$$P(A, B, C) = P(A) P(B | A) P(C | A, B)$$

quite a bit, and it is worth pausing on what the multiplication means. Each factor is a table, not a number, so this is not regular multiplication. The tables look a little like matrices, but it is not matrix multiplication either; the sizes do not even line up. What the equation really says is that the joint table is filled in *one entry at a time*: for each combination of values, go to each factor, find the entry that matches those values, and multiply the entries together.

Example: Take the following three tables. $P(B | A)$ has one column per given world ($+a$ or $-a$), and each column sums to one; $P(C | A, B)$ has one row per given world.

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

To fill in the entry of the joint for $+a, +b, +c$, we find the $+a$ entry of $P(A)$, the entry of $P(B \mid A)$ for $+b$ given $+a$, and the entry of $P(C \mid A, B)$ for $+c$ given $+a, +b$:

$$P(+a, +b, +c) = P(+a) P(+b \mid +a) P(+c \mid +a, +b) = (0.1)(0.5)(0.3) = 0.015.$$

Doing this for all $2 \times 2 \times 2 = 8$ combinations fills in the joint table:

		$+a$	$-a$
$P(A, B, C)$:	$+b, +c$	$(0.1)(0.5)(0.3) = 0.015$	$(0.9)(0.2)(0.0) = 0.000$
	$-b, +c$	$(0.1)(0.5)(0.7) = 0.035$	$(0.9)(0.8)(1.0) = 0.720$
	$+b, -c$	$(0.1)(0.5)(0.7) = 0.035$	$(0.9)(0.2)(1.0) = 0.180$
	$-b, -c$	$(0.1)(0.5)(0.3) = 0.015$	$(0.9)(0.8)(0.0) = 0.000$

The eight entries sum to one, as a joint distribution must. Whether we draw the tables on paper or store them in a dictionary or an array in a program, a product of probability tables is really a recipe for this entry-by-entry lookup and multiplication, repeated once for every entry of the result.

5.4 From Conditional Probability Tables to Any Query

We now have a complete process for going from the conditional probability tables we are given to the answer to a query:

1. **Construct the joint distribution** from the tables, using the product rule or the chain rule (Section 5.3).
2. **Answer the query from the joint**, using the definition of conditional probability and marginalization (Section 4.3).

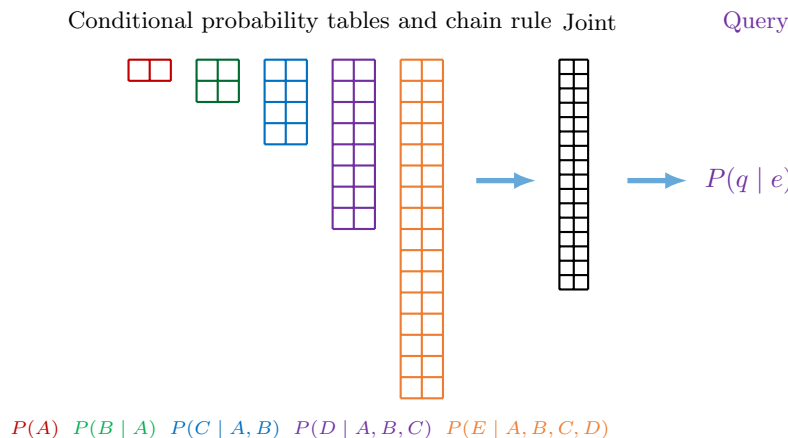
Example: Suppose we are given the tables $P(E \mid H, Q)$, $P(H \mid Q)$, and $P(Q)$ and we want $P(Q \mid e)$. The chain rule with the ordering Q, H, E builds the joint from exactly these tables,

$$P(E, H, Q) = P(E \mid H, Q) P(H \mid Q) P(Q),$$

and then the recipe from Section 4.3 answers the query:

$$P(Q \mid e) = \frac{P(e, Q)}{P(e)} = \frac{\sum_h P(e, h, Q)}{\sum_q \sum_h P(e, h, q)}.$$

The sizes of the tables: Look at what the chain rule asks for as the number of variables grows. With five binary variables and the ordering A, B, C, D, E , the factors are $P(A)$, $P(B \mid A)$, $P(C \mid A, B)$, $P(D \mid A, B, C)$, and $P(E \mid A, B, C, D)$, with 2, 4, 8, 16, and 32 entries.



The last table is as big as the joint itself, so the full chain rule has not saved us anything. Worse, $P(E \mid A, B, C, D)$ is not a table anyone is likely to hand us. If E is whether there is an earthquake, we are not going to find the probability of an earthquake given an alarm, a burglary, chest pain, and dehydration. We need a way to write the joint as a product of *small* conditional probability tables, the ones we can actually get. This is what Bayes nets give us.

6 Bayes Nets

Bayes nets are our answer to the two problems with joint distributions. A Bayes net describes a complex joint distribution, our model of the world, using simple, local distributions: one conditional probability table per random variable. It is one kind of **probabilistic graphical model**. Each table says how a variable relates to a few others, and these local relationships chain together to give global, indirect relationships between variables that are far apart in the net.

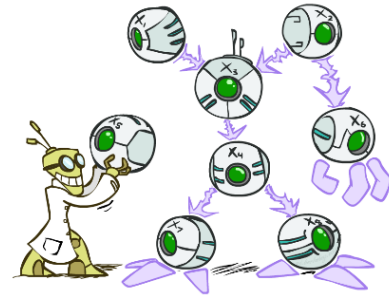


Illustration: Ketrina Yim / UC Berkeley AI

6.1 Nodes, Edges, and Tables

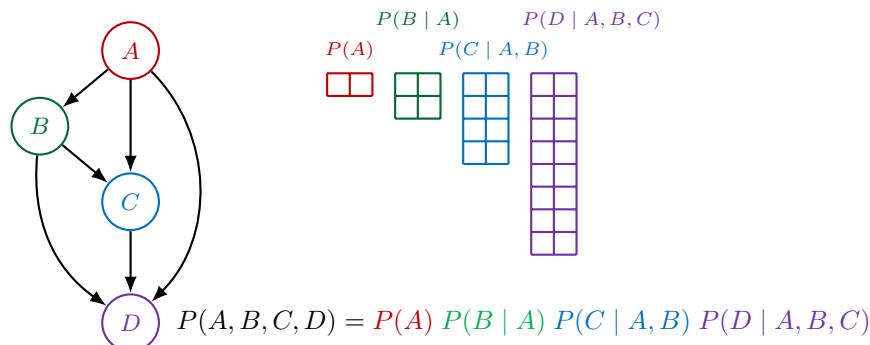
A **Bayes net** (also called a Bayesian network) is a **graphical model**: a picture of which conditional probability tables we have. It consists of

- one **node** per random variable,
- **directed edges** between nodes that form a **directed acyclic graph** (DAG): following the arrows, you can never get back to where you started, and
- one **conditional probability table** per node, $P(\text{node} \mid \text{Parents}(\text{node}))$, where the **parents** of a node are the nodes with edges pointing into it.

The joint distribution that a Bayes net represents is the product of all of its tables, “nodes given parents”:

$$P(X_1, \dots, X_N) = \prod_{i=1}^N P(X_i \mid \text{Parents}(X_i)).$$

Example: In the Bayes net below, A has no parents, B has parent A , C has parents A and B , and D has parents A , B , and C . The tables are $P(A)$, $P(B \mid A)$, $P(C \mid A, B)$, and $P(D \mid A, B, C)$, and the joint is their product:

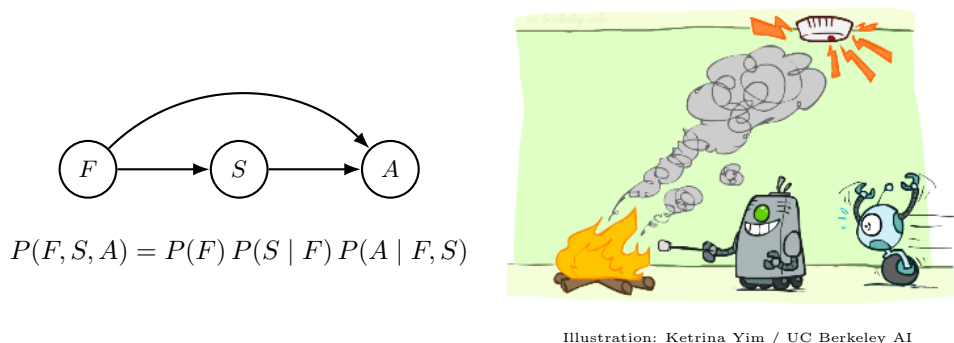


This particular product is exactly the chain rule for the ordering A, B, C, D . This is because the chain rule is itself a Bayes net, one in which every variable has an edge from every variable before it in the ordering.

6.2 Building a Bayes Net from the Chain Rule

We can also go the other way, from a chain-rule factorization to a picture. Draw one node per variable, and for each factor $P(X | Y, Z)$ draw an edge from each of the given variables, Y and Z , into X .

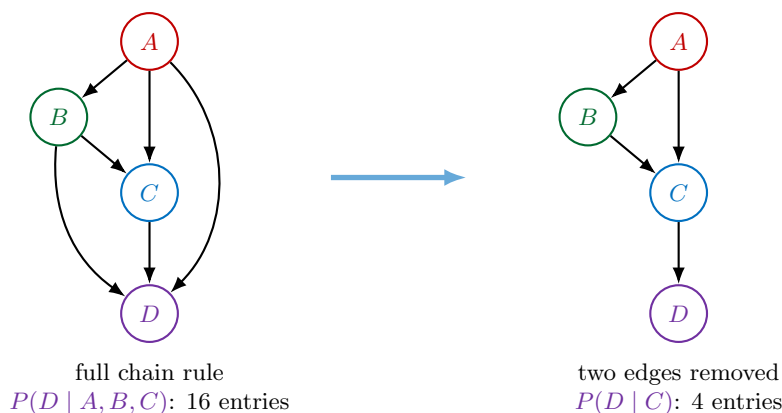
Example: For fire, smoke, and alarm with the ordering F, S, A , the chain rule gave $P(F, S, A) = P(F) P(S | F) P(A | F, S)$. The factor $P(S | F)$ gives an edge $F \rightarrow S$, and $P(A | F, S)$ gives edges $F \rightarrow A$ and $S \rightarrow A$:



Every ordering of the chain rule gives a **fully connected** DAG like this one, with an edge between every pair of nodes. Reading the joint off a Bayes net, fully connected or not, is always the same exercise: nodes given parents.

6.3 Missing Edges Encode Independence Assumptions

A fully connected Bayes net makes no assumptions, and it saves nothing: its last table is as big as the joint. What we care about in a Bayes net is the edges we *leave out*. Compare the four-node net from Section 6.1 with the same net after removing the edges $A \rightarrow D$ and $B \rightarrow D$:



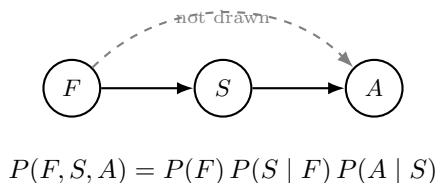
In the second net, D has only one parent, so its table is $P(D | C)$ instead of $P(D | A, B, C)$, and the joint is

$$P(A, B, C, D) = P(A) P(B | A) P(C | A, B) P(D | C).$$

This is cheaper, with 4 entries in the last table instead of 16, and it is also far more likely to be a table that someone can hand us: one variable given one other variable, rather than one variable given everything.

In exchange, we have made an assumption. **The edges that are missing, compared to the full chain rule, encode our independence assumptions.** By dropping $A \rightarrow D$ and $B \rightarrow D$ we assumed that once we know C , knowing A and B tells us nothing more about D . Section 8 makes this precise. The assumption may or may not be strictly true in the world; we make it because it lets us build the model from tables we can actually get.

Example: For fire, smoke, and alarm, a smoke alarm senses smoke, not fire. So we tell the causal story “fire causes smoke, and smoke causes the alarm” and draw only the edges $F \rightarrow S$ and $S \rightarrow A$, leaving out the chain-rule edge $F \rightarrow A$:



The tables we need are now $P(F)$, $P(S | F)$, and $P(A | S)$, with $2 + 4 + 4 = 10$ entries instead of $2 + 4 + 8 = 14$. The savings are small with three variables; they become enormous as the number of variables grows.

There is a subtlety in the other direction. An edge that *is* present does not assume that the two variables are dependent. It only means that we are not assuming they are independent. An absent edge makes an assumption, and a present edge does not.

If you would like a reminder of independence and conditional independence before going on, jump ahead to Section 7, which is short, and come back to the alarm network afterwards.

6.4 The Alarm Network

The alarm network is a classic example that we will keep coming back to. There is an alarm on your house, which is triggered by a burglary, and can also be shaken into going off by an earthquake. Two neighbors, John and Mary, may call you when they hear the alarm; each may also call you for no reason, and each may fail to call.

We build the Bayes net from the causal story. A burglary can cause the alarm, and so can an earthquake, so $B \rightarrow A$ and $E \rightarrow A$. The alarm is what makes John and Mary call, so $A \rightarrow J$ and $A \rightarrow M$. Every other edge is left out, and each one that is left out is an assumption: burglaries and earthquakes happen independently of each other; John and Mary do not see the burglar or feel the earthquake themselves, only the alarm; and John and Mary do not talk to each other before calling. Telling a causal story like this is a reliable way to build a valid Bayes net, although, as Section 8 explains, the arrows in a Bayes net do not have to be causal.

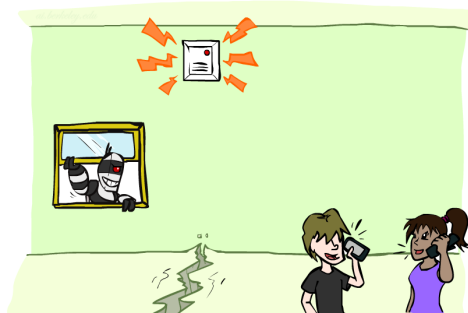


Illustration: Ketrina Yim / UC Berkeley AI

15

7.1 Independence

Two random variables X and Y are **independent**, written $X \perp\!\!\!\perp Y$, if

$$\forall x, y \quad P(x, y) = P(x) P(y).$$

That is, the joint distribution **factors** into a product of the two marginal distributions. The “for all” matters: independence is a statement about the random variables, the capital letters, and it has to hold for every combination of values.

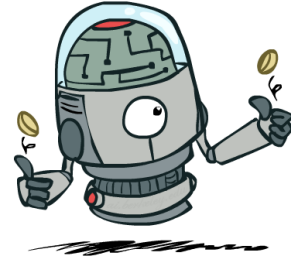


Illustration: Ketrina Yim / UC Berkeley AI

Compare this with the product rule, which always holds: $P(x, y) = P(x) P(y | x)$. Independence is the shortcut of replacing $P(y | x)$ with $P(y)$. Combining the definition with the product rule gives two equivalent forms,

$$\forall x, y \quad P(x | y) = P(x) \quad \text{or} \quad \forall x, y \quad P(y | x) = P(y),$$

which say the same thing in words: knowing the value of one variable does not change what we believe about the other.

Example: Two dice rolls R_1 and R_2 are modeled as independent, so $P(R_1 = 5, R_2 = 5) = P(R_1 = 5) P(R_2 = 5) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$, and $P(R_2 = 5 | R_1 = 5) = P(R_2 = 5)$. For physical dice this is an assumption, and we make it to keep the model simple.

Example: Are mushrooms and spinach independent on the pizza of Section 3? To check, we compute the joint $P(M, S)$ by counting slices, multiply the marginals entry by entry, and compare:

M	S	$P(M, S)$	M	$P(M)$	S	$P(S)$	M	S	$P(M) P(S)$
m_1	s_1	6/20	m_1	12/20	s_1	10/20	m_1	s_1	6/20
m_1	s_2	6/20	m_2	8/20	s_2	10/20	m_1	s_2	6/20
m_2	s_1	4/20					m_2	s_1	4/20
m_2	s_2	4/20					m_2	s_2	4/20

Every entry matches, for example $P(m_2) P(s_2) = \frac{8}{20} \cdot \frac{10}{20} = \frac{4}{20} = P(m_2, s_2)$, so $M \perp\!\!\!\perp S$: half of the mushroom slices have spinach, and so do half of the slices without mushrooms. Mushrooms and pepperoni are not independent. The very first check fails, $P(m_2) P(r_2) = \frac{8}{20} \cdot \frac{6}{20} = \frac{2.4}{20} \neq \frac{4}{20} = P(m_2, r_2)$, so M and R are not independent. One mismatched entry is enough to rule independence out; to rule it in, every entry has to match.

Example: With N fair, independent coin flips $X_1, \dots, X_N$, the joint distribution factors all the way down,

$$P(X_1, X_2, \dots, X_N) = P(X_1) P(X_2) \cdots P(X_N),$$

so instead of a joint table with 2^N entries we store N tables with 2 entries each, and we never need the conditional distribution of the thousandth flip given the previous 999. As a Bayes net, this is N nodes with no edges at all.

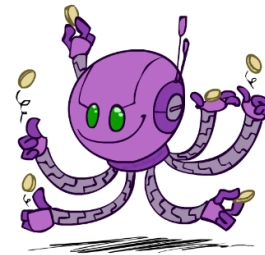


Illustration: Ketrina Yim / UC Berkeley AI

Example: Should an agent model whether it is raining, R , and whether there is traffic, T , as independent? Model 1 below assumes they are: two nodes, no edge, and the joint is $P(R) P(T)$. Model 2 draws the edge $R \rightarrow T$, and the joint is $P(R) P(T | R)$.

Model 1: independence



$$P(R, T) = P(R) P(T)$$

Model 2: rain affects traffic



$$P(R, T) = P(R) P(T | R)$$

Model 2 is the better model of the world, since rain really does affect traffic, and $P(T | R)$ is an easy table to collect. Notice again what the edge does and does not say. Model 1 makes an assumption, and it is wrong. Model 2 makes no assumption: if it happened that $P(R) P(T) = P(R, T)$, model 2 would still be correct, just with a table $P(T | R)$ whose columns are identical.

7.2 Conditional Independence

Absolute independence like the coin flips is rare. **Conditional independence** is much more common, and it is our most basic and robust form of knowledge about uncertain environments. X is conditionally independent of Y given Z , written $X \perp\!\!\!\perp Y | Z$, if

$$\forall x, y, z \quad P(x | y, z) = P(x | z),$$

or, equivalently,

$$\forall x, y, z \quad P(x, y | z) = P(x | z) P(y | z).$$

In words: once we know Z , learning Y tells us nothing more about X . If you understand independence, conditional independence is the same set of rules with “given Z ” inserted on the right of every bar:

If $X \perp\!\!\!\perp Y$, then	If $X \perp\!\!\!\perp Y Z$, then
$P(X, Y) = P(X) P(Y)$	$P(X, Y Z) = P(X Z) P(Y Z)$
$P(X Y) = P(X)$	$P(X Y, Z) = P(X Z)$
$P(Y X) = P(Y)$	$P(Y X, Z) = P(Y Z)$

Notation: The symbol $\perp\!\!\!\perp$ is read “is independent of.” As with the conditioning bar, precedence matters: $X \perp\!\!\!\perp Y | Z$ means “given Z , X and Y are independent,” not $X \perp\!\!\!\perp (Y | Z)$, which does not mean anything. Also as before, the statement is about the capital letters: $X \perp\!\!\!\perp Y | Z$ requires the equality to hold for every value z of Z , not just for one of them.

8 Bayes Net Triples

The three small Bayes nets in this section each have three nodes and one missing edge. For each one, we write the joint two ways, with the chain rule and with the Bayes net, and compare the two to find the independence assumption that the missing edge encodes. Each factor is colored by the variable on the left of its bar, matching the node colors.

8.1 Causal Chain

Fire F causes smoke S , and smoke causes the alarm A . Using the chain rule with the same ordering F, S, A as the Bayes net, so that the factors line up:

$$\text{Chain rule: } P(F, S, A) = P(F) P(S | F) P(A | F, S)$$

$$\text{Bayes net: } P(F, S, A) = P(F) P(S | F) P(A | S)$$

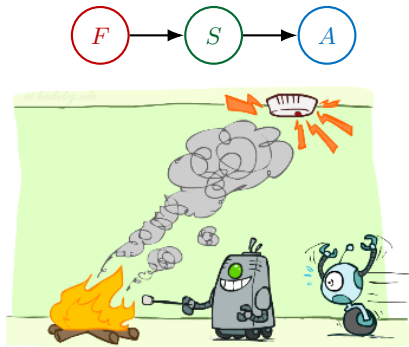


Illustration: Ketrina Yim / UC Berkeley AI

The first two factors are identical, so the only assumption is in the last one:

$$P(A | F, S) = P(A | S), \quad \text{that is, } A \perp\!\!\!\perp F | S.$$

Given smoke, the alarm is independent of fire. If we know whether there is smoke, we have everything we need to know about the alarm, and learning whether there is a fire changes nothing. Popcorn burning in a dorm room makes smoke without a fire, and the alarm responds to the smoke either way.

Note what is *not* assumed: nothing in the net says that A and F are independent unconditionally, and in this story they are not. If we know nothing about smoke, learning that there is a fire raises our belief that the alarm will go off. The connection runs through S , and it is observing S that blocks it.

Arrows and causation: The causal story is a good way to *build* a Bayes net, but arrows in a given Bayes net do not have to be causal. The two-node nets $R \rightarrow T$ and $T \rightarrow R$ are both fully connected, so neither makes any assumption, and both represent the same joint distribution, $P(R)P(T | R) = P(T)P(R | T)$. The second net is just as valid as the first, but its arrow is not a causal one. You can build a valid Bayes net from causal assumptions, but you cannot read causal conclusions off a Bayes net.

8.2 Common Cause

Rain R causes traffic T , and rain also causes umbrellas U , with no edge between T and U . With the ordering R, T, U :

$$\text{Chain rule: } P(R, T, U) = P(R) P(T | R) P(U | R, T)$$

$$\text{Bayes net: } P(R, T, U) = P(R) P(T | R) P(U | R)$$

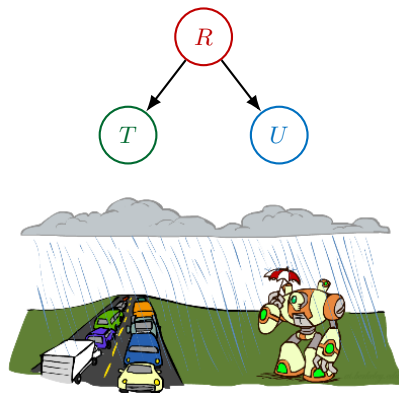


Illustration: Ketrina Yim / UC Berkeley AI

The assumption is

$$P(U | R, T) = P(U | R), \quad \text{that is, } U \perp\!\!\!\perp T | R.$$

Given the rain, umbrellas and traffic are independent. If we know it is raining, then seeing an umbrella tells us nothing more about the traffic, and the same holds if we know it is not raining.

Note what is *not* assumed: nothing in the net says that T and U are independent unconditionally, and in this story they are not. If we have been inside all day and know nothing about the weather, then seeing someone walk in with a wet umbrella raises our belief that it is raining, and with it our belief that there is traffic. It is observing the common cause that separates its two effects.

8.3 Common Effect

Rain R causes traffic T , and so does a hockey game H letting out downtown, with no edge between R and H . This case is different from the other two. With the ordering R, H, T :

Chain rule: $P(R, H, T) = P(R) P(H \mid R) P(T \mid R, H)$

Bayes net: $P(R, H, T) = P(R) P(H) P(T \mid R, H)$

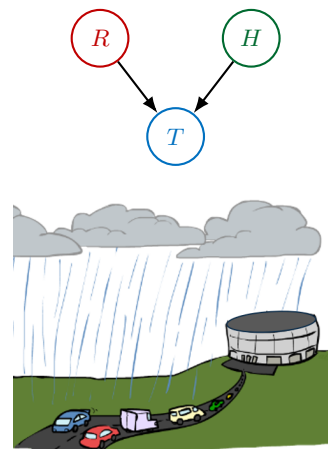


Illustration: Ketrina Yim / UC Berkeley AI

The assumption is in the middle factor this time,

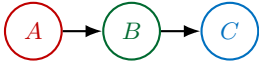
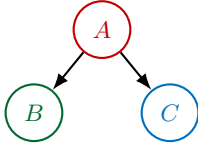
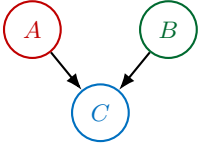
$$P(H \mid R) = P(H), \quad \text{that is,} \quad H \perp\!\!\!\perp R.$$

Rain and (indoor) hockey games are independent, given nothing. Knowing whether there is a game tells us nothing about whether it is raining.

What is *not* assumed is $H \perp\!\!\!\perp R \mid T$, and with a story like this one it does not hold. Suppose we know there is traffic. If we then learn that it is not raining, the rain can no longer explain the traffic, and our belief that there is a hockey game goes up. If instead we learn that it is raining, the rain explains the traffic, and our belief in a hockey game goes down a little. The two causes compete to explain the observed effect. This is backwards from the other two triples: in a causal chain and a common cause, the two ends are assumed independent given the middle node but not unconditionally; in a common effect, the two ends are assumed independent unconditionally but not given the middle node.

8.4 The Three Triples, Side by Side

Here are the same three comparisons with generic variable names. In each case the chain rule uses the top-down ordering A, B, C , one factor changes, and the change is the assumption.

Causal chain	Common cause	Common effect
		
$P(A) \ P(B A) \ P(C A, B)$	$P(A) \ P(B A) \ P(C A, B)$	$P(A) \ P(B A) \ P(C A, B)$
$P(A) \ P(B A) \ P(C B)$	$P(A) \ P(B A) \ P(C A)$	$P(A) \ P(B) \ P(C A, B)$
$P(C A, B) = P(C B)$	$P(C A, B) = P(C A)$	$P(B A) = P(B)$
$C \perp\!\!\!\perp A B$	$C \perp\!\!\!\perp B A$	$B \perp\!\!\!\perp A$

9 Summary

Expression	What it is
$P(A)$	A table with one entry per outcome of A . Capital letters mean tables; lowercase letters are single outcomes.
$P(A, B +c), P(A, B C)$	The first is a conditional probability table inside the world $C = +c$ and sums to one. The second stacks the tables for $+c$ and $-c$ and sums to two. A table sums to one only when the right of the bar is one specific combination of values and the left is all capital letters.
Joint $\rightarrow$ query	$P(Q e) = P(Q, e)/P(e)$, then un-marginalize the numerator and denominator into sums of entries of the joint, or normalize the numerator table.
Chain rule	$P(X_1, \dots, X_N) = \prod_i P(X_i X_1, \dots, X_{i-1})$, in any ordering. A product of tables is filled in entry by entry.
Bayes net	A DAG with one node per random variable and one conditional probability table per node: $P(X_1, \dots, X_N) = \prod_i P(X_i \text{Parents}(X_i))$, nodes given parents. The edges left out, compared with the full chain rule, are the independence assumptions.
$X \perp\!\!\!\perp Y, X \perp\!\!\!\perp Y Z$	$\forall x, y : P(x, y) = P(x)P(y)$, equivalently $P(x y) = P(x)$. Conditional: $\forall x, y, z : P(x, y z) = P(x z)P(y z)$, equivalently $P(x y, z) = P(x z)$.
Triples	Causal chain $A \rightarrow B \rightarrow C$ assumes $C \perp\!\!\!\perp A B$. Common cause $B \leftarrow A \rightarrow C$ assumes $C \perp\!\!\!\perp B A$. Common effect $A \rightarrow C \leftarrow B$ assumes $B \perp\!\!\!\perp A$, and <i>not</i> $B \perp\!\!\!\perp A C$.