Conditional probabilities and graphical models

Thomas Mailund

Bioinformatics Research Centre (BiRC), Aarhus University

Probability theory allows us to describe uncertainty in the processes we model within the exact and formal language of mathematics. While what we model might be full of uncertainty — either intrinsically or because we lack a full understanding of what we choose to model, or simply choose not to model at sufficient detail to eliminate this uncertainty — the language we choose to reason with is formal and enable exact reasoning.

This note is not intended as an introduction to probability theory, not even an informal one — if that was the intend it would be substantially longer — but is simply intended to remind you of the basic rules for manipulating multi-variable (joint and conditional) probabilities and to introduce the notation we will use in this class when manipulating probabilities.

The treatment of random variables and probabilities is intended to be informal. For instance, I will generally pretend that the only difference between discrete and continuous random variables is whether you sum or integrate over events. This might make some mathematicians uncomfortable, and those I will recommend getting a propper text book on probability and measure theory. For the purpose of this class, we can get away with pretending that everything looks like a discrete set, even when it does not, because we only work with "nice" sets. We leave it for the mathematicians to define what "nice" means in this context.

1 Stochastic ("random") variables

We will only be concerned with random variables and not define an underlying universe of events and measures; we simply will have random variables that can take on different values with different probabilities. So while formally a random/stochastic variable is a function from events to values, we will only care about the values they take and the probabilities of taking those values. So we can define a random variable X as a variable that can take values in some domain D_X if for each value in this domain we can assign a probability/density $p(X = x) \geq 0$ for all $x \in D_X$. If D_X is discrete we require $\sum_{x \in D_X} p(X = x) = 1$ and if D_X is continuous we require $\int p(X = x)\, \mathrm{d}x = 1$ (but as I mentioned above we generally won't care much about whether we work with probabilities or densities or whether we sum or integrate).

We will follow the tradition of probability theory and be lazy with our notation. We will just write $p(X)$ for the function that assigns probabilities to possible outcomes $x \mapsto p(X = x)$. Keep in mind, though, that this is just a bit of laziness. The variable X is random in the sense that it can take on different values in D_X. The probability we assign to each $x \in D_X$ tells us how likely it is that X take on any of those values, but $p(X = x)$ is a number and not random at all. Also, there is a huge difference between a random variable we haven't observed yet and one we have: Once we have seen that X takes the value x it is no longer random.

If you are so inclined — as I personally am — you can consider statistics the craft of updating your knowledge of the un-observed from that which you have observed. We initially build some model for how we think the probability of various outcomes could be, and how they are related. This model can be more or less based on prior knowledge of the domain that we model and often somewhat restricted by computational efficiency and mathematical convenience. When we then observe the outcome of some of the random variables in our model we learn more about the system and we can update the probabilities of the as-yet unobserved random variables. This is played straight in so-called Bayesian statistics, where unknown underlying parameters are considered random variables and where inference exactly *is* updating conditional probabilities as we observe more and more outcomes of random variables. It is not exactly the same in so-called frequentists statistics where we update parameters that are considered unknown but not random, but still here inference is also a question of updating our knowledge based on the observed outcomes of random variables.

We return to how to do inference from observed data in another note, though. The rest of this note concerns joint and conditional probabilities when we have more than one variable.

2 Working with probabilities

If we want to express the probability that a random variable X falls in some subset of its domain, $A \subseteq D_X$, we have the probability $p(X \in A) = \sum_{x \in A} p(X = x)$. That is, we sum over all the possible outcomes of X that are in A. If A is empty this is a special case with $p(X \in A) = 0$; if $A = D_X$ we have, by the definition of a random variable, $p(X \in A) = 1$. These properties are typically given as part of the defintion of what a probability is, and here we just take them as given as well.

If we have two sets, $A \subseteq D_X$ and $B \subseteq D_X$, we similarly have $p(X \in A \cup B) = \sum_{x \in A \cup B} p(X = x)$. In general, $p(X \in A \cup B) \neq p(X \in A) + p(X \in B)$ since that would count elements in $A \cap B$ more than once. In fact,

$$p(X \in A \cup B) = p(X \in A) + p(X \in B) - p(X \in A \cap B),$$

with $p(X \in A \cup B) = p(X \in A) + p(X \in B)$ only when $p(X \in A \cap B) = 0$. The latter will always be the case when $A \cap B = \emptyset$, so whenever we have a set, $A_1, \ldots, A_k$, of disjoint subsets of D_X we have

$$p\left(X \in \bigcup_k A_k\right) = \sum_k p(X \in A_k).$$

Whenever it is clear from context that we are working with the variable X we will again be lazy with notation and write $P(A)$ to mean $P(X \in A)$ in which case we get

$$p\left(\bigcup_k A_k\right) = \sum_k p(A_k).$$

2.1 Joint probabilities

If we have two random variables, X and Y, that can take values in domains D_X and D_Y, we can assign probabilities for their joint outcome: $p(X = x, Y = y)$ for $x \in D_X$ and $y \in D_Y$. With our lazy notation we write simply $p(X, Y)$, although if we want to make explicit that, say, X is actually observed to be x, we will write $p(X = x, Y)$. This is best thought of as a function of Y that gives the joint probability of $p(X = x, Y = y)$ for all y in the domain of Y. The variable Y is still random even though we have observed x.

To express the probability that X falls in some subset $A \subseteq D_X$ and Y falls in some subset $B \subseteq D_Y$ we write $p(X \in A, Y \in B)$ — or just $p(A, B)$ if the context makes it clear which variables goes with which sets — and we compute these probabilities by summing over all the events in these sets:

$$p(A, B) = \sum_{x \in A} \sum_{y \in B} p(X = x, Y = y).$$

Special cases of this general rule are when either A or B are singleton sets where we are expressing, for example, that X falls in the set A while Y takes the value y: $p(A, y) = \sum_{x \in A} p(x, y)$.

Another special case is when any of the sets is the full domain of the random variable. For

$$p(A, D_Y) = \sum_{x \in A} \sum_{y \in D_Y} p(X = x, Y = y) = p(X \in A)$$

we are in essence saying "X takes a value in A while Y takes any possible value" which of course is the same as saying that X takes a value in A.

Summing over all possible values of a random variable as we just did here is called marginalisation and is remarkably useful for something that seems so trivial.

This way of considering the joint outcome of two variables, and summing over the possibilities that they fall in different subsets of their respective domains, readily generalises to three or more variables: $p(A, B, C)$, $p(A, B, C, D)$, ...

2.2 Conditional probabilities

We define the "conditional probability" $p(Y \mid X)$ as the value that makes the following equation true:

$$p(X)p(Y \mid X) = p(X, Y).$$

When $p(X) \neq 0$ this means $p(Y \mid X) = p(X, Y)/p(X)$, which has a useful intuitive interpretation, but I prefer the definition above since it makes it less likely that we divide by zero. When $p(X)$ *is* zero, any value of $p(Y \mid X)$ satisfy the equation (the marginalisation rule above guarantees that $p(X) = 0$ implies $p(X, Y) = 0$) and the conditional probability is not well defined.

The conditional probability captures updated expectations about the values Y can take after we have observed X (or at least observed that it falls in sum subset of D_X). To see this, think of $p(X = x) = \sum_{y \in D_Y} p(X = x, Y = y)$ as the total "probability mass" on all pairs $(x, y) \in \{x\} \times D_Y$ and $p(X = x, Y = y)$ as the "probability mass" on the single point (x, y). It is thus the mass on (x, y) relative to all points with $X = x$ rather than the mass on (x, y) relative to all possible points (the unconditional probability). For sets $p(Y \in B \mid X \in A)$ the interpretation is the same: it is the mass on points $A \times B$ relative to the mass on $A \times D_Y$ rather than relative to all possible points, $D_X \times D_Y$.

Since we explicitly defined $p(Y \mid X)$ this way, it is not surprising that we can always write the joint probability as the marginal times the conditional: $p(X, Y) = p(X)p(Y \mid X)$. It is, after all, just a notation trick. It is a helpful trick, though, when we are building models. We don't need to specify the joint probability of two variables explicitly — sometimes that can be difficult if we want the model to capture certain aspects of the system we are interested in — instead we can just specify a marginal and a conditional probability and then get the joint probability from that.

Of course, X is not special here and we might as well have chosen Y as the one to have a marginal probability and X to be conditioned on Y:

$$p(X, Y) = p(Y)p(X \mid Y)$$

The best choice is usually what is mathematical convenient when modelling and not much more.

2.3 Bayes' Theorem

Since both forms of marginal times conditional probabilities are are just notation for the joint probability we have $p(X)p(Y \mid X) = p(Y)p(X \mid Y)$. If we have one of the conditionals and want the other, we can therefore rewrite (assuming that it doesn't trick us into dividing by zero, which we should really try to avoid):

$$p(Y \mid X) = \frac{p(Y)p(X \mid Y)}{p(X)}.$$

This rewrite of conditionals is know as *Bayes' Theorem.* Little more than a notation trick, but immensely useful because of the way we typically build models. For some reason, it is just generally easier to build models where you have marginal probabilities for the variables you are unlikely to directly observe and then specify the observed data as conditional on these hidden variables. When you do inference you want to go in the other direction: now you will have observed the variables you specified as conditional on the hidden variables and you want to know the likely values of the hidden variables conditional on the observations.

Even though I have called the conditional probability a notational trick a couple of times I should stress that it *is* a probability distribution still. At least whenever the outcome we condition on does not have zero probability. You can check yourself that it assings non-negative values to all outcomes and sums to one: $\sum_{y\in D_Y} p(Y = y \mid X) = 1$ whenever $p(X)$ is not zero.

Your intuition should be that $p(Y)$ is the probability distribution that reflects the likely outcomes of the random variable Y before you have observed the outcome of X while $P(Y \mid X)$ is the updated probability distribution that reflects the information you got from observing X. This intuition is reflected in terminology from Bayesian statistics where the marginal probability is called the *prior* probability and the conditional probability the *posterior* probability.

Conditional probabilities generalises to more than two variables as well. For X, Y, and Y we have

$$p(X, Y, Z) = p(X)p(Y \mid X)p(Z \mid X, Y)$$

or

$$p(X, Y, Z) = p(Z)p(Y \mid Z)p(X \mid Y, Z)$$

or

$$p(X, Y, Z) = p(Y)p(Z \mid Y)p(X \mid Y, Z)$$

or any of the other compositions. Notice, however, that you have to condition on *all* the variables you have specified earlier in the line of probabilities you multiply! In general you cannot leave any of them out and still get the right conditional probability.

In case you are wondering what $p(Z \mid X, Y)$ means, aside from the value that would the equation true — which is as good a definition as any — it is of course equal to $p(X, Y, Z)/p(X, Y)$ as well (whenever $p(X, Y) \neq 0$). This follows trivially from observing that $p(X)p(Y \mid X) = p(X, Y)$ and then dividing on both side by this.

Generalisations to four or more random variables are trivial.

3 You have a joint probability — Now what?

Regardless of how we come up with a joint probability, once we have one we have a complete specification of our mathematical/statistical model. Pretty much everything we want to do from here on boils down to manipulations of this structure. (Again, this is literally true for Bayesian statistics, I would claim, but not quite for frequentist statistics; it isn't all wrong there either, though).

So what are those manipulations? We can split our random variables into three kinds: Those we have observed (and that are thus no longer random), those we don't care about (called nuisance parameters; they are just there to build the model but we don't care about them as such), and those parameters we are interested in (that can be underlying parameters we are interested in by themselves or future events we want to predict).

Condition on observed parameters. If we have the joint probability $p(X, Y, Z)$ but we have already observed $Z = z$ we are not really interested in the outcome of Z any longer: we already know that it came out as z. We will still be interested in X and Y, so when observing Z we move our interest from $p(X, Y, Z)$ to $p(X, Y \mid Z)$. If there are clever tricks for doing this we will exploit them — and some times there are as we will see — but in general we can do it just from marginalisation. Remember

$$p(X, Y \mid Z) = \frac{p(X, Y, Z)}{p(Z)}$$

and

$$p(Z) = \sum_X \sum_Y p(X, Y, Z)$$

(where I took yet another chance to introduce lazy notation: when we want to sum over all values of random variable we often write $\sum_X p(X)$ to mean $\sum_{x \in D_X} p(X = x)$. There is a lot of such sloppy notation in the literature so you won't be able to avoid it even if I did it in this note. To be fully formal I should also have used $p(Z = z)$ or $p(Z \in C)$ for $z \in D_Z$ and $C \subseteq D_Z$, respectively, since that is what we mean when we say that Z was observed, but again this is the way it will typically be written and in general it is clear from context if Z is unknown or known whenever it actually matters).

Marginalise away nuisance parameters. Now, say I am really only interested in the outcome of X. The Y variable was needed to connect the variable of interest, X, to the variable I could observe, Z, but in itself it holds no interest to me. In that case I am not interested in $p(X, Y \mid Z)$ but in $p(X \mid Z)$ which I can again get from marginalisation

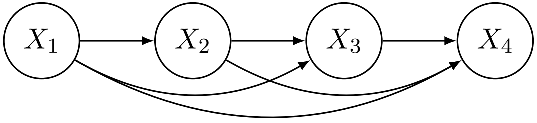
$$p(X \mid Z) = \sum_Y p(X, Y \mid Z).$$

Those are the general manipulations we do on a joint probability: when we observe a random variable we move it from the joint probability so we condition on it instead and when we don't care about a variable we marginalise it away. What we are left with is a joint probability of those variables we are still interested in, updated to best reflect the observations we have made.

The summing over all possible values used in the marginalisation (and the implicit marginalisation used in the conditioning) is potentially computational intractable, so we cannot always simply do this. Some times clever algorithms are necessary, but they will typically always just do one of those two things in a clever way.

4 Dependency graphs

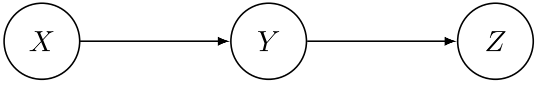
We have a graphical notation we use to describe the dependency relationships when specifying a joint probability as we did a section back. Each random variable is represented as a node, and whenever the the composition of the joint probability has a term $p(Y \mid X_1, \ldots, X_k)$ we have directed edges from all the X_i's to Y. (We don't really handle terms like $p(X, Y \,|Z)$ in this notation, so if you want to represent that simply replace the pair $X \times Y$ with a single random variable $W = X \times Y$ representing the pairs and you are fine).



If these graphs were only used to represent the kind of compositions we have seen above we probably wouldn't use them. For those compositions — that handle the full general case of writing up a joint probability from the product of a list of conditionals — the composition is given by the order at which we add variables. For each new variable to add we must condition on all the previous. Dependency graphs are really most useful for specifying joint probabilities when you do *not* have to condition on all earlier variables in a composition. The interesting aspect for a dependency graph is that it shows the *independence* relationships (which we discuss in more detail in the next section).

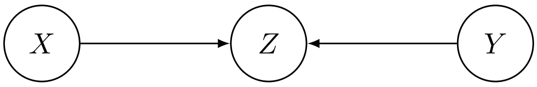
A general dependency graph is a directed acyclic graph where the nodes are our random variables. It tells us how to compute the joint probability of these: if $a(n)$ denotes the (direct) ancestors of node n (the nodes with a direct edge to n) then the joint probability of the graph is $\prod p(n \mid a(n))$ — where we will be a bit sloppy and equate the nodes with the random variables they represent.

Consider the joint probability $p(X, Y, Z)$ specified by the graph below:



The graph specifies that $p(X, Y, Z) = p(X)p(Y \mid X)p(Z \mid Y)$ where the probability for Z depends on Y but not on X as it would in the general case.

The graph below on the other hand specifies that $p(X, Y, Z) = p(X)p(Y)p(Z \mid X, Y)$ where Z depends on both X and Y but Y does not depend on X as it would in the general case.

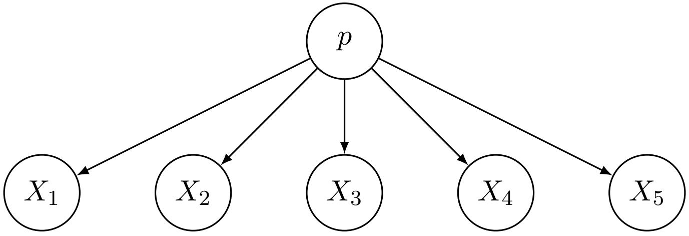


There are no general rules telling you when you can remove edges in dependency graphs. That is a model choice and up to the modeller. The graphs simply specify which marginal and which conditional probabilities are needed to compute the joint probability.

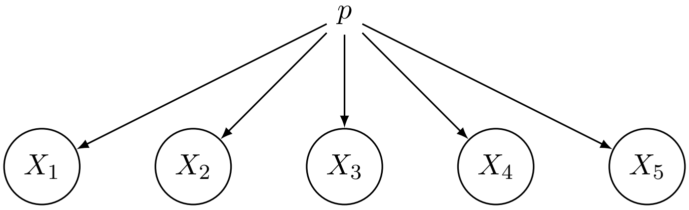
4.1 Example graphs

Independent and identically distributed random variables. Let's consider a typical inference setup: We have n random variables, $X_1, X_2, \ldots, X_n$ all independent and identically distributed (IID). The "independent" here, obviously, means that there are no dependence edges between them. Identically distributed means that they are drawn from some distribution we can easily specify with a few parameters. One example is so-called Bernoulli distributions that are random variables that turn out either true or false and where a single parameter determines the probability for true instead of false. If the X_i's are Bernoulli distributed with parameter p it means $p(X_i = 1) = p$ for all i. This dependence of p is, of course, a dependency we can put in the dependency graph, and a statistical question would then be: given observed values for the n random variables, what is p?

In a Bayesian setting we assume that p is itself a random variable, perhaps uniformly distributed between 0 and 1: $p \sim U[0,1]$. We can then add p to the dependency graph with edges to all the X_i's. The uniform distribution would be the prior distribution of p and after observing all the outcomes of $X_1, \ldots, X_i$ we can update this distribution to the conditional, or posterior, distribution to narrow the range of likely p values.

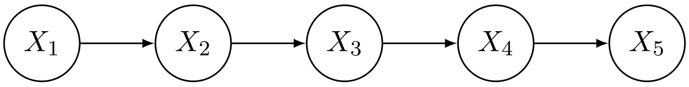


We do the same thing in frequentist statistics, there we just do not consider p a random variable, so it doesn't have a distribution. Instead it is initially considered completely unknown and after we observe the X_i's we have some knowledge of what it is likely to be (but we do not represent this as uncertainty in the sense of randomness; this is the big philosophical difference between Bayesian and frequentist statistics). We can still represent p as a node in the graph, but then we will typically use a different kind of nodes to make it clear that it is a parameter (known or unknown) and not a random variable.



Markov chains. As another example consider $X_1, \ldots, X_n$ but this time we do not consider then independent. Rather their dependency is

$$p(X_1, \ldots, X_n) = p(X_1)p(X_2 \mid X_1)p(X_3 \mid X_2) \cdots p(X_n \mid x_{n-1}).$$



This type of structure, where the next variable depends on only one previous variable, is called a *Markov chain* and is frequently used to model sequential data because of its simple structure.

5 Independence and conditional independence

Above I called the lack of an edge we would normally expect a lack of dependency. If the edge was there it would be an explicit dependency but the lack of it doesn't actually mean there isn't a dependency.

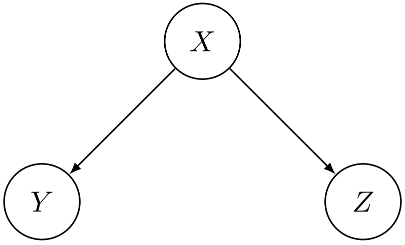
To understand this we need to define what independence is. Whenever variables are not independent there is a dependency, explicit or not.

The usual definition of independence is that X and Y are independent if $p(X, Y) = p(X)p(Y)$, that is that their joint probability equals the product of their marginal probabilities. To see that this is a definition that matches well with our intuition consider the compositions of $p(X, Y) = p(X)p(Y \mid X) = p(Y)p(X \mid Y) = p(X)p(Y)$. This equation states that the conditional probability for Y equals its marginal $p(Y \mid X) = p(Y)$ and similar for X: $p(X \mid Y) = p(X)$. So knowning the outcome of X doesn't change the distribution for Y and vice versa. That is what we mean by independence.

We also have *conditional* independence. This is when we have independence of two conditional probabilities: $p(X, Y \mid Z) = p(X \mid Z)p(Y \mid Z)$.

Independence of two variables does not guarantee conditional independence when a third variable is introduced, nor does conditional independence guarantee dependence when we don't condition.

Example Consider $p(X, Y, Z) = p(X)p(Y \mid X)p(Z \mid X)$:



Here, Y and Z are not (unconditional) independent since in general

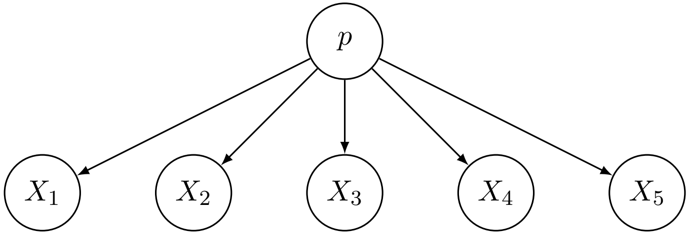
$$\sum_X p(X)p(Y \mid X)p(Z \mid X) \neq p(Y)p(Z).$$

If we condition on X, however, we have

$$p(Y, Z \mid X) = \frac{p(X)p(Y \mid X)p(Z \mid X)}{p(X)} = p(Y \mid X)p(Z \mid X)$$

so Y and Z are independent once X is observed. This is actually a general pattern: if two variables are dependent because they have a common ancestral variable in the dependency graph, then observing that common ancestor makes them independent. (If they are connected by more than this one path the relationship is less clear, of course).

The IID example we saw earlier has this form:

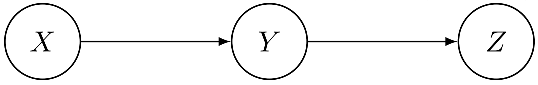


From the observations we just made it follows that these *independent* identically distributed random variables are not, in fact, independent unless we know the parameter that determines their distribution. In the frequentist setting we do not consider p a random variable — it is unknown but not random — so here they really are independent, but in a Bayesian setting, where p is random, not knowing the parameter introduces a dependency between the random variables.

This shouldn't come as a surprise, though: if the parameters were independent then observing the first $n-1$ outcomes should not give us information about the n'th outcome. This means we wouldn't be able to learn from observations. The dependency between the variables is exactly the reason why we can use previous observations to make better predictions about future observations.

Once the underlying parameter is know, however, the variables really *are* independent. Seeing more outcomes will not help us get better estimates of p — we already know what the value of p is — and in that case we can only predict the next observation within the uncertainty of its distribution.

Example Consider $p(X,Y,Z) = p(X)p(Y\,|\,X)p(Z\,|\,Y)$:



Here, X and Z are not independent, since marginalising with respect to Y

$$p(X,Z) = \sum_Y p(X)p(Y\,|\,X)p(Z\,|\,Y)$$

is not, in general, equal to $p(X)p(Z)$.

We can, however, rewrite the joint probability, using Bayes' theorem

$$p(X,Y,Z) = p(X)p(Y\,|\,X)p(Z\,|\,Y) = p(X)\frac{p(Y)p(X\,|\,Y)}{p(X)}p(Z\,|\,Y) = p(Y)p(X\,|\,Y)p(Z\,|\,Y)$$

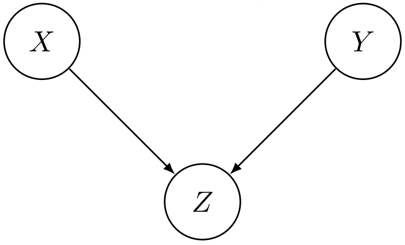
and so if we condition on Y we get

$$p(X,Z\,|\,Y) = \frac{p(X,Y,Z)}{p(Y)} = p(X\,|\,Y)p(Z\,|\,Y)$$

showing that X and Z are independent if Y is observed.

This case has important implications for Markov models. Markov models have the general structure of a sequence of pairwise dependent random variables ($p(X)p(Y \mid X)p(Z \mid Y)$ and so on), and in general they are not independent. If, however, we observe any of the variables in the chain then the sequence before that observation and the sequence following that observation are now independent. We can thus split a long chain into smaller, independent, chains by conditioning on values along the chain.

Example Consider $p(X, Y, Z) = p(X)p(Y)p(Z \mid X, Y)$.

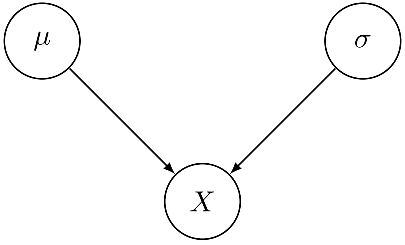


Clearly X and Y are independent because we can marginalise away Z to get $p(X, Y) = \sum_Z p(X)p(Y)p(Z \mid X, Y) = p(X)p(Y)$. If we condition on Z, however, we get $p(X, Y \mid Z) = p(X)p(Y)p(Z \mid X, Y)/p(Z)$ where in general $p(Z \mid X, Y)/p(Z)$ is not identical to one (it would be if Z was independent of the joint X and Y but in general it is not). So X and Y are not conditional independent given Z.

An example of such a system could be this: X is 1 with probability p and 0 with probability $1 - p$, while Y is 1 with probability q and 0 with probability $1 - q$, and $Z = X$ xor Y. If you build a table of all the possible outcomes and check, you will find that $p(X, Y) = p(X)p(Y)$ (because we did this by construction). If we know that Z is 1, however, X and Y are no longer independent. If X is 1 then Y has to be 0 and vice versa. Knowing Z introduces a dependency between X and Y that wasn't there before we observed Z.

Again, this is a general pattern: whenever otherwise independent random variables have shared descendants in the dependency graph, then observing these will introduce a conditional dependence.

Consider a normally distributed random variable $X \sim N(\mu, \sigma)$ where the parameters are themselves random variables. A priori we consider μ and σ independent, but when we observe X we update their posterior probability to reflect the observation.



These updates will not be independent: If we update μ to be closer to the observed X then we would expect a smaller variance, σ, then if we move μ less towards X.