BKI PGMs – Glossary

Wilker Aziz

October 27, 2025

Abstract

Each section covers key concepts introduced in the corresponding module in a not-so-formal way. These are intended as concise reminders, not as formal definitions. In some cases, we also have a hint to help you avoid common mistakes.

Random variable (rv). A mathematical tool to represent claims (or situations or propositions) we are uncertain about. An rv may take on any one of a set of possible values. We usually use capital letters to denote rvs and corresponding lowercase letters do denote outcomes, for example an rv T used to model the outside temperature may take on values such as t^1 (for 'low'), t^2 (for 'moderate') or t^3 (for 'high').

No matter what situation in the world we choose to give random treatment to, this situation need not be stochastic in nature. When we model something as an rv, we attempt to represent in a mathematical language our own incomplete knowledge about the part of the world that's relevant to our decisions.

Outcome space. The set of all possible values that an rv may take on is called the rv's outcome space. We follow the textbook and use $\text{Val}(X)$ to denote the set of outcomes of the rv X. For example, if we say that the possible outcomes of X are 'high' (x^1) and 'low' (x^0), then $\text{Val}(X) = \{x^0, x^1\}$. We can use $x \in \text{Val}(X)$ to denote an outcome in the set, without specifying which. In this course we will concentrate on outcome spaces that are countably finite. The size of the outcome space is regarded as the *cardinality* of the rv.

Random variable assignment. An rv X may take on any one value in its outcome space $\text{Val}(X)$, which happens with a certain probability, when this happens we denote $X = x$ for some $x \in \text{Val}(X)$.

It looks superfluous and maybe it feels tedious to write $X = x$, as opposed to only x, but it's in fact necessary to avoid ambiguities: X alone is one thing (a random variable), x alone is another thing (for example, an outcome such as a number), and $X = x$ is yet something else. Specifically, $X = x$ implies that, of all possible situations involving of a specific rv and of all possible values it may take on, we are concentrating on those situations where X turns out to take on the value x (for example, because we observed it to be the case, or because we are wondering about what happens when that is the case).

Distribution of an rv. A function that tells us with what probability an rv takes on one of its possible values. We typically denote the distribution of an rv X by $P(X)$ or P_X for brevity, then $P(X = x)$ is the probability that X takes on an outcome $x \in \text{Val}(X)$.

Joint distribution. A function that tells us with what probability a collection of rvs takes on one of its possible values. For example, if we have two binary rvs X and Y, then a joint outcome is a pair of values $(x, y) \in \text{Val}(X, Y)$ in the joint outcome space. The joint outcome space is the cross-product (or Cartesian product) of the rvs' outcome spaces: $\text{Val}(X, Y) = \text{Val}(X) \times \text{Val}(Y) = \{(x^0, y^0), (x^0, y^1), (x^1, y^0), (x^1, y^1)\}$. The joint distribution is commonly denoted $P(X, Y)$ or P_{XY} for brevity, and $P(X = x, Y = y)$ is the probability of the (joint) assignment.

Conditional probability distribution (CPD). The distribution of one or more rvs when considered in the context of other rvs (so-called the conditioning context). For example, we have two rvs X and Y, then $P(Y|X)$, or $P_{Y|X}$ for brevity, is the CPD for Y given X, and $P(X|Y)$ is the CPD for X given Y. Then, for some $x \in \text{Val}(X)$, $P(Y|X = x)$ is the distribution of Y in the specific situations where $X = x$, and $P(Y = y|X = x)$ is the conditional probability of Y taking on $y \in \text{Val}(Y)$ given that X takes on $x \in \text{Val}(X)$.

Marginalisation. From a joint distribution such as $P(X, Y, Z)$ we can obtain the distribution of any subset of its rvs via what is known as marginalisation. For example, the so-called 'marginal distribution' of (X, Y) is given by $P(X, Y) = \sum_{z \in \text{Val}(Z)} P(X, Y, Z = z)$. Sometimes, for brevity and if no confusion is possible, we follow the textbook and write the marginalisation as $P(X, Y) = \sum_Z P(X, Y, Z)$. We can also marginalise multiple rvs out at once, for example: $P(X) = \sum_{y \in \text{Val}(Y)} \sum_{z \in \text{Val}(Z)} P(X, Y = y, Z = z)$. The quantity $P(X = x)$ then is the marginal probability that X takes on $x \in \text{Val}(X)$ with respect to the joint distribution of (X, Y, Z).

Conditioning. From a joint distribution such as $P(X, Y, Z)$ we can obtain the distribution of any subset of its rvs given the remaining rvs via what is known as conditioning. For example, the distribution of Z given X and Y can be expressed as follows: $P(Z|X, Y) = \frac{P(X,Y,Z)}{P(X,Y)}$, where the denominator is a marginal of the joint distribution. In another situation, we may condition on some rvs and the result is another joint distribution: $P(X, Y|Z) = \frac{P(X,Y,Z)}{P(Z)}$, again the denominator is a marginal of P_{XYZ}. It is possible to combine marginalisation and conditioning in more ways, for example: $P(Z|X) = \frac{P(X,Z)}{P(X)}$, this time both the numerator and denominator are marginals of the joint distribution.

Continuing with the previous example, when attempting to express the conditional distribution of Z given X, it is a common mistake to sum over the values of Y in $P(Z|X, Y)$, something like ~~$\sum_{y \in \text{Val}(Y)} P(Z|X, Y = y)$~~. Marginalisation can only be performed over jointly distributing variables (in a conditional distribution over multiple rvs, those are the rvs to the left of the conditioning bar) as in, for example, $P(X|Z) = \sum_Y P(X, Y|Z)$.

Chain rule. As a direct consequence of the definition of conditional probability, we can obtain the result known as chain rule. Consider a joint distribution P_{XY}, we can express it in any of the following two ways: $P(X, Y) = P(X)P(Y|X) = P(Y)P(X|Y)$. And the result generalises to any number of rvs, no matter the order in which we enumerate them: $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_{<i})$, where $X_{<i}$ is a shorthand for the rvs that precede X_i in the order we chose to iterate over them.

Let's analyse this result and try to clear up some confusion.

Starting from the joint distribution. Assume we have an explicit representation for P_{XY}, in the form of a table that contains the probabilities for all possible joint assignments of X and Y, each value of x indexes a row, each value of y indexes a column:
then we can obtain P_X via marginalisation (in the example, we sum along the columns of the table P_{XY}, obtaining

	P_{XY}		$P_X = \sum_Y P_{XY}$	$P_{Y\|X} = P_{XY}/P_X$	
	y^0	y^1		y^0	y^1
x^0	$P(X = x^0, Y = y^0)$	$P(X = x^0, Y = y^1)$	$\overset{\text{sum}}{\to} P(X = x^0)$	$P(Y = y^0\|X = x^0)$	$P(Y = y^1\|X = x^0)$
x^1	$P(X = x^1, Y = y^0)$	$P(X = x^1, Y = y^1)$	$\overset{\text{sum}}{\to} P(X = x^1)$	$P(Y = y^0\|X = x^1)$	$P(Y = y^1\|X = x^1)$
$P_Y = \sum_X P_{XY}$	$\downarrow^{\text{sum}}$ $P(Y = y^0)$	$\downarrow^{\text{sum}}$ $P(Y = y^1)$			
	$P_{X\|Y} = P_{XY}/P_Y$				
x^0	$P(X = x^0\|Y = y^0)$	$P(X = x^0\|Y = y^1)$			
x^1	$P(X = x^1\|Y = y^0)$	$P(X = x^1\|Y = y^1)$			

the shaded column P_X), and, with P_{XY} and P_X, we can obtain $P_{Y|X}$ (see top-right corner of the figure), or, conversely, we can obtain P_Y via marginalisation (in the example, we sum along the columns of the P_{XY} table, obtaining the shaded row P_Y), and, with P_{XY} and P_Y, we can obtain $P_{X|Y}$ (see bottom-left corner of the figure).

Prescribing the joint distribution. Sometimes, we have access to P_X and $P_{Y|X}$, for example because for the variables we are interested in it's more natural to think of them as if X had a causal effect on Y (for example, rain X affects road traffic Y). Now, we can use these two distributions to prescribe a joint distribution over X and Y via chain rule: $P(X, Y) = P(X)P(Y|X)$. Bayesian networks will take this idea to amazing ends!

Bayes rule. There is very little to say about Bayes rule, what makes it a celebrated result is its application in statistical inference, more so than its derivation from first principles. Suppose we are working in a setting where it is more natural or intuitive to prescribe P_X and $P_{Y|X}$ and then combine them into a joint distribution $P(X, Y) = P(X)P(Y|X)$, as opposed to starting from P_{XY} directly or to build it by first prescribing P_Y and $P_{X|Y}$.

Then, we can use evidence about Y to update our beliefs about X via the so-called Bayes rule:

$$P(X|Y=y) = \frac{P(X,Y=y)}{P(Y=y)} = \frac{P(X)P(Y=y|X)}{\sum_{x\in \text{Val}(x)} P(X=x)P(Y=y|X=x)} \ .$$

Bayes rule allows us to express the distribution of X conditioned on having obtained evidence about Y, namely that $Y=y$, it follows by direct application of conditional probability, chain rule and marginalisation with a fixed outcome for Y.

Independence. An rv X is said to be independent of an rv Y when X has no effect on the distribution of Y, no matter what value X takes on. When this is true, the other way around is also true (namely, no matter what outcome Y takes on, the distribution of X remains the same). Their independence is denoted $X \perp Y$. Independence has the following implications: $P(X,Y) = P(X)P(Y)$, $P(X|Y) = P(X)$, $P(Y|X) = P(Y)$, and the more subtle but equally true $P(X,Y) \propto \phi_1(X)\phi_2(Y)$ for any $\phi_1 : \text{Val}(X) \to \mathbb{R}_{\geq 0}$ and $\phi_2 : \text{Val}(Y) \to \mathbb{R}_{\geq 0}$. If one of these is true, all of them are, hence any of these can be used to ascertain the independence between two rvs.

Conditional independence. A form of independence that's licensed in a certain conditioning context. For example, given Z, no outcome of X can affect the distribution of Y. This is denoted $X \perp Y \mid Z$. The implications of conditional independence are: $P(X,Y|Z) = P(X|Z)P(Y|Z)$, $P(X|Z,Y) = P(X|Z)$, $P(Y|Z,X) = P(Y|Z)$, and the more subtle but equally true $P(X,Y|Z) \propto \phi_1(X,Z)\phi_2(Y,Z)$ for any $\phi_1 : \text{Val}(X,Z) \to \mathbb{R}_{\geq 0}$ and $\phi_2 : \text{Val}(Y,Z) \to \mathbb{R}_{\geq 0}$. As with independence, if one of these is true, they all are. Hence, any one of them can be used to test whether the conditional independence holds.

There's an important asymmetry in testing conditional independence: to reject the conditional independence of X and Y given Z, it is enough to find one outcome of $z \in \text{Val}(Z)$ for which $P(Y|X,Z=z)$ differs from $P(Y|Z=z)$. But, to accept the statement, the equality of $P(Y|X,Z=z)$ and $P(Y|Z=z)$ must hold for all possible $z \in \text{Val}(Z)$.

0.1 Bayesian Networks

Directed acyclic graph (DAG). A graph whose edges have directionality, and where no directed cycles are permitted. We usually use calligraphic capital letters to denote graphs, like $\mathcal{G}$ or $\mathcal{H}$.

Topological ordering. Any ordering of the nodes in a DAG, such that the parents of the node in position i of the order are guaranteed to precede it. There can be multiple orderings meeting this requirement, and there is always at least one such ordering.

Topological orderings are relevant to analysing BNs because it guarantees that the descendants of a node come *after* that node in a topologically-sorted sequence.

Tabular CPD. The representation of a CPD using a table of probability values where each column is associated with an outcome of the rv being modelled and each row is associated with an outcome of the conditioning context. Hence, rows sum to 1.0.

Tabular CPDs are only possible when the joint outcome space of the random variable and its context is countably finite (otherwise it would require infinitely many rows and/or columns).

Bayesian network (BN). A representation of probability distributions achieved by combining a DAG and a collection of CPDs. The nodes in the DAG correspond to rvs. The edges represent direct dependence of a child node on its parents (an edge always points from parent to child). We must have one CPD per node, where the CPD's is modelling the node's rv in the context of the node's parents.

A BN allows us to represent our uncertainty about variables we are interested in reasoning about, in light of how we assume they depend directly on one another. These dependencies are *assumed* to make sense and they help us design distributions that are more compact than the most complex distribution possible over the rvs of interest. If we have N rvs, each with cardinality K, the most complex distribution possible has a probability value per joint outcome, hence it takes $\mathcal{O}(K^N)$ parameters to represent it. In a BN with tabular CPDs, this number is dominated

by the size of the largest table: if the node with most parents has L parents, then the number of parameters of the BN scales with $\mathcal{O}(K^{L+1})$.

Local independence. The BN structure (*i.e.*, its DAG) represents a collection of conditional independence statements. Within this collection, there's an important set called "local independencies of the graph", denoted $\mathcal{I}_l(\mathcal{G})$ for a DAG $\mathcal{G}$. This set contains one independence statement for each node X_i conveying that: X_i is independent of its non-descendants in the graph, given its parents in the graph. Mathematically, this is denoted: $X_i \perp \text{NonDesc}_{\mathcal{G}}(X_i) \mid \text{Pa}_{\mathcal{G}}(X_i)$. The implications are those of any conditional independence statement: $P(X_i, \text{NonDesc}_{\mathcal{G}}(X_i) \mid \text{Pa}_{\mathcal{G}}(X_i)) = P(X_i|\text{Pa}_{\mathcal{G}}(X_i))P(\text{NonDesc}_{\mathcal{G}}(X_i)|\text{Pa}_{\mathcal{G}}(X_i))$ or, equivalently, $P(X_i|\text{NonDesc}_{\mathcal{G}}(X_i), \text{Pa}_{\mathcal{G}}(X_i)) = P(X_i|\text{Pa}_{\mathcal{G}}(X_i))$.

Factorisation. A decomposition of a probability value into a product of other values. For example, chain rule *factorises* $P(X = x, Y = y)$ as $P(X = x)P(Y = y|X = x)$ or $P(Y = y)P(X = x|Y = y)$, a product of two other probabilities. We can also talk about factorisation of probability distributions, in this case chain rule tell us that it is possible to find two CPDs, P_X and $P_{Y|X}$, such that P_{XY} is expressible via $P(X)P(Y|X)$, hence it is possible to factorise P_{XY} like that. By chain rule, it is also possible to factorise P_{XY} in terms of P_Y and $P_{X|Y}$.

Factorisation plays an essential role in PGMs because it is a tool to represent a joint distribution in terms of other objects.

Chain rule for BNs. The local independencies $I_l(\mathcal{G})$ implied by the BN structure $\mathcal{G}$ are sufficient to uniquely specify a factorisation of a probability distribution P over the rvs $X_1, \ldots, X_n$. A joint distribution P represented by the BN then factorises as follows: $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Pa}_{\mathcal{G}}(X_i))$. This result identifies a class of distributions, namely, all of those for which the statements in $I_l(\mathcal{G})$ are known to (or assumed to) hold. To identify a specific distribution, rather than an entire family, a BN associates a specific CPD with each node and hence with each factor $P(X_i|\text{Pa}_{\mathcal{G}}(X_i))$.

Causal reasoning. When we reason about Z, upon having observed one of its ancestors Y in the BN, we are reasoning 'causally'. The term alludes to Y being one of the 'causes' of Z.

Evidential reasoning. When we reason about Y, upon having observed one of its descendants Z in the BN, we are reasoning 'evidentially'. The term alludes to Z being one of the 'effects' of Y.

Intercausal reasoning. Consider X and Y which are not ancestors or descendants of one another. When we reason about Y, upon having observed X and Z, and Z is a common descendant of X and Y, we are reasoning intercausally. The term alludes to X and Y being common 'causes' of Z, or equivalently Z being a common 'effect' of X and Y. Without observing Z, X and Y are independent, but observing Z makes them dependent, hence the outcome of one helps us explain the outcome of the other.

Direct influence. The effect of observing X on the distribution of Y when X and Y are connected via a directed edge. Consider the BN: $X \to Z \leftarrow Y$. If we observe $Y = y$, we can reason causally and conclude that Y exerts influence on the distribution $P_{Z|Y=y}$; and, if we observe $Z = z$, we can reason evidentially and conclude that Z exerts influence on the distribution $P_{Y|Z=z}$. These are examples of Y's direct influence on Z and vice-versa. Variables that are directly connected influence one another no matter the direction we reason about (namely, that of the edge, or its reverse).

Indirect influence. The effect of observing X on the distribution of Y when their dependence is only 'activated' given evidence about a common effect. Consider the BN: $X \to Z \leftarrow Y$. With Z unobserved, X and Y are independent of one another, but with Z observed we can reason intercausally and conclude that X exerts influence on $P(Y|Z = z, X = x)$. This is an example of X's indirect influence on Y by the observation of Z. In fact, X exerts influence on Y so long as Z and/or any of its descendants is observed. This follows by application of direct influence: a descendant of Z exerts influence on Z, which then 'activates' the dependency between X and Y, and then X influences the distribution of Y given the evidence we have.

Trail. A trail between two nodes in a DAG is a sequence of edges that forms a path connecting the two nodes, where we traverse edges regardless of their direction. For example, in the BN $X \to Z \leftarrow Y$, the sequence (X, Z, Y) is a trail connecting X and Y, it has 2 edges (one from X to Z and one from Y to Z) whose directions we ignore to form the trail.

v-structure. Let's consider all possible kinds of 2-edge BNs involving 3 nodes as these form all possible 2-edge trails connecting two nodes (such trails are important for they help us analyse indirect influence). They look as follows: $A \to C \to B$, $A \leftarrow C \leftarrow B$, $A \leftarrow C \to B$ and $A \to C \leftarrow B$. The last of these is what we call a v-structure, and we often refer to it as 'the v-structure at C' (singling out the node that plays the role of a 'common effect' of the other two). By convention, in the v-structure, we call C a collider node, and, in the remaining 3 structures, C is a non-collider. We say the v-structure is 'blocked' if its collider C and all of its descendants are unobserved, in this case influence cannot flow from A to B and back. We say the v-structure is 'activated' if its collider C or any of its descendants is observed, in this case influence flow from A to B and back.

Active trail. It is useful to think of influence as something 'flowing' through the BN structure. There are many trails between any two nodes, an active trail is one in which influence can flow from one of its end to the other and back. Influence flows up and down (direct influence) through unobserved non-colliders and it bounces up (indirect influence) through colliders in activated v-structures; direct influence is blocked by observed non-colliders and by blocked v-structures.

Directed separation. A set of nodes $\mathbf{X}$ is d-separated from a set of nodes $\mathbf{Y}$ given outcomes for a set of nodes $\mathbf{Z}$ (*i.e.*, the 'evidence') if there is no active trail between any node in $\mathbf{X}$ and any node in $\mathbf{Y}$ given the evidence $\mathbf{Z}$. In other words, direct separation (or d-separation for short), denoted d-sep$(\mathbf{X}; \mathbf{Y} | \mathbf{Z})$, holds when, given $\mathbf{Z}$, no rv in $\mathbf{X}$ can influence any rv in $\mathbf{Y}$. This is a powerful tool because i) d-separation implies conditional independence and vice-versa, and ii) testing d-separation only requires knowledge of the BN structure (not of its CPDs), offering a powerful alternative to testing from first principles.

Earlier in this section, when discussing local independencies, we mentioned that a BN structure represents a collection of conditional independence statements and that within this collection an important set is the set $I_l(\mathcal{G})$. That sentence suggested that a BN codes many more independence statements, and indeed it does, and d-separation is a powerful tool for us to figure out what those statements are.