BKI PGMs – Cheat Sheet

Fall 2025

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 $(x, y) \in \mathrm{Val}(X, Y)$ in the joint outcome space. The joint outcome space is the cross-product (or Cartesian product) of the rvs' outcome spaces: $\mathrm{Val}(X, Y) = \mathrm{Val}(X) \times \mathrm{Val}(Y) = \{(x^0, y^0), (x^0, y^1), (x^1, y^0), (x^1, y^1)\}$. 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 $x \in \mathrm{Val}(X)$, $P(Y|X = x)$ 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 $y \in \mathrm{Val}(Y)$ given that X takes on $x \in \mathrm{Val}(X)$.

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 $P(X, Y) = \sum_{z \in \mathrm{Val}(Z)} P(X, Y, Z = z)$. 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: $P(X) = \sum_{y \in \mathrm{Val}(Y)} \sum_{z \in \mathrm{Val}(Z)} P(X, Y = y, Z = z)$. The quantity $P(X = x)$ then is the marginal probability that X takes on $x \in \mathrm{Val}(X)$ 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.

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.

Bayes rule. For a joint distribution $P(X, Y)$ factorised as $P(X)P(Y|X)$ we can express beliefs about X given an assignment of $Y = y$ via: $P(X|Y = y) = \frac{P(X,Y=y)}{P(Y=y)} = \frac{P(X)P(Y=y|X)}{\sum_{x \in \mathrm{Val}(x)} P(X=x)P(Y=y|X=x)}$.

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 $\phi_1 : \mathrm{Val}(X) \to \mathbb{R}_{\geq 0}$ and $\phi_2 : \mathrm{Val}(Y) \to \mathbb{R}_{\geq 0}$. 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 $\phi_1 : \mathrm{Val}(X, Z) \to \mathbb{R}_{\geq 0}$ and $\phi_2 : \mathrm{Val}(Y, Z) \to \mathbb{R}_{\geq 0}$. 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 $z \in \mathrm{Val}(Z)$ 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 $z \in \mathrm{Val}(Z)$.

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.

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.

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. The parameters of the model are the conditional probabilities in the CPDs.

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 $P_{X|Y}$.

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, a BN associates a specific CPD with each node and hence with each factor $P(X_i|\text{Pa}_{\mathcal{G}}(X_i))$.

Direct influence. If X and Y are connected by an edge they exert influence on one another regardless of the direction of the edge.

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. A trail should pass by a node only once.

3-node trails. There are 3 essential types of trails in BNs: the *chain* $A \to C \to B$, the *fork* $A \leftarrow C \to B$, and the *v-structure* $A \to C \leftarrow B$. In the v-structure the C node is called a *collider*, in the chain and fork C is called a *non-collider*. The end nodes of a trail (A and C) are also called non-colliders. If we are considering influence between A and B, it matters how the 3 nodes are arranged, since the influence is indirect (it's mediated by C). In the chain and fork, where C is non-collider, influence flows from A to B and back, but influence is interrupted when we condition on the non-collider C. There's a technical jargon for this: A and B are *marginally dependent* but *conditionally independent*. In the v-structure, where C is a collider, the situation is reversed: influence cannot flow from A to B until we observe the collider C and/or any one of the descendants of C in the graph. The jargon for this is: A and B are *marginally independent* but *conditionally dependent*.

Active trail. A trail made of 3 nodes or more has a sequence of 3-nodes chains, forks and v-structures in it, we need to check each one of them. If they are all active, then the trail is active. If even one 3-node trail is inactive then the whole trail is inactive.

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.

2 Markov Networks

Undirected Graphs. A collection of nodes and edges. Unlike DAGs, the edges are *undirected* indicating a symmetric interaction between nodes. We denoted one such graph by a calligraphic letter such as $\mathcal{H}$.

Complete sub-graphs or cliques. A collection of nodes in an undirected graph $\mathcal{H}$ such that any two nodes in the collection are connected by an edge in $\mathcal{H}$. For any one graph $\mathcal{H}$, there can be multiple sets of complete sub-graphs, each set being able to cover all nodes and edges in the graph.

Factor. A real-valued function $\phi(\mathbf{X})$ of a set $\mathbf{X}$ of rvs, called the scope of the factor and denoted Scope$[\phi]$. A general factor maps outcomes in Val$(\mathbf{X})$ to arbitrary real values, but, in this course, we are only concerned with positive factors, that is $\phi : \text{Val}(\mathbf{X}) \to \mathbb{R}_{\geq 0}$.

Tabular Factor. A table-like representation of a factor over discrete variables with finite outcome space. We list row by row the joint outcomes of the variables in the scope and their factor value.

Markov networks. A representation of joint probability distributions parameterised by factors (symmetrical interaction) rather than CPDs. An MN combines an undirected graph $\mathcal{H}$ and a collection Φ of factors. The nodes of $\mathcal{H}$ correspond to a set of rvs $\mathbf{X}$. Two variables that interact directly are connected by an edge, but this edge is undirected. The MN structure $\mathcal{H}$ is a representation of a set $\mathcal{I}(\mathcal{H})$ of conditional independencies, which we will cover later in this document. Each factor ϕ_i in Φ has scope $\mathbf{D}_i$, a set of nodes of $\mathcal{H}$ which must correspond to a complete sub-graph of $\mathcal{H}$. The collection as a whole must account for each node of $\mathcal{H}$ at least once and each edge of $\mathcal{H}$ at least once. With its graph and factors, an MN identifies a probability distribution P_Φ such that $P_\Phi(\mathbf{X}) = \frac{1}{Z}\tilde{P}_\Phi(\mathbf{X})$, where the unnormalised measure $\tilde{P}_\Phi$ is obtained by multiplication over all factors: $\tilde{P}_\Phi(\mathbf{X}) = \prod_{\phi_i \in \Phi} \phi_i(\mathbf{D}_i)$. And the normaliser, also known as partition function, is defined as $Z = \sum_{\mathbf{X}} \tilde{P}_\Phi(\mathbf{X})$.

From graphs to factorisation. The MN structure does not identify a unique factorisation. But valid factorisations are constrained by certain rules (see MN definition above). Instead the graph imposes a unique set of conditional independencies, we will discuss those later in the document.

From factors to graph. A collection Φ of factors (or even just their scopes) can, on its own, be used to identify a graph $\mathcal{H}$. That's because of the rules we agreed upon when describing the set Φ (see MN definition above).

Marginalisation. If $\tilde{P}(A, B)$ is an unnormalised joint distribution, then $\tilde{P}(A) = \sum_B \tilde{P}(A, B)$ is its unnormalised marginal over A. The normaliser $Z = \tilde{P}(A, B)$ of the joint also normalises the marginal.

Conditioning. If $\tilde{P}(A, B)$ is an unnormalised joint distribution, then $P(B|A = a) = \frac{\tilde{P}(A=a,B)}{\tilde{P}(A=a)}$, and we can also say that $P(B|A = a) \propto \tilde{P}(A = a, B)$.

Factor product. If $\phi_1(A, B)$ and $\phi_2(B, C)$ are two factors, we can multiply them to obtain $\psi(A, B, C) = \phi_1(A, B)\phi_2(B, C)$. The scope of the new factor is the union of the scopes of the input factors. Factor product always produces useful factors (e.g., useful in building joint distributions).

Factor reduction. If $\phi_1(A, B)$ is a factor and we want to assign one of its variables $B = b$, we can consider a factor whose scope is 'reduced' to A alone, and whose values are that of $\phi_1(A, B = b)$. We denote this reduced factor by $\phi_1[b](A)$. If we attempt to reduce an irrelevant variable, we just obtain the same factor without any reduction. Factor reduction is useful when conditioning on observations. If Φ contains 3 factors $\phi_1(A, B)$, $\phi_2(B, C)$ and $\phi_3(C, A)$, then $\tilde{P}(A, B, C) = \phi_1(A, B)\phi_2(B, C)\phi_3(C, A)$, and $P(A, C|B = b) \propto \tilde{P}(A, B = b, C) = \phi_1[b](A)\phi_2[b](C)\phi_3(C, A)$. Moreover, we can compute products and then reduce, or reduce and then take products: $(\phi_1(A, B)\phi_2(B, C)\phi_3(C, A))[b] = \phi_1[b](A)\phi_2[b](C)\phi_3(C, A)$, reducing first usually spares effort and computation.

Factor normalisation. Divides the factor values by the total sum of all of its values, normalising it. Again, if Φ contains three factors and we take their product $\pi(A, B, C) = \phi_1(A, B)\phi_2(B, C)\phi_3(C, A)$, this is $\tilde{P}(A, B, C)$, and its normalised version $\eta(A, B, C) = \frac{1}{Z}\pi(A, B, C)$ where $Z = \sum_{A,B,C} \pi(A, B, C)$ is in fact $P(A, B, C)$. You should only apply factor normalisation if you have accounted for all factors covering the the variables in the scope of the unnormalised factor and those variables are separated from the rest of the variables in the MN. Separation is covered later in the document. So for example, we can normalise this $\phi_1[b](A)\phi_2[b](C)\phi_3(C, A)$ to obtain $P(A, C|B = b)$.

Factor marginalisation. Sum along one of the variables in the scope of the factor, returning a factor over fewer rvs. If Φ contains three factors and we take their product $\pi(A, B, C) = \phi_1(A, B)\phi_2(B, C)\phi_3(C, A)$, this is $\tilde{P}(A, B, C)$, and its marginal version $\sigma(A, B) = \sum_C \pi(A, B, C)$ is in fact $\tilde{P}(A, B)$. If we normalise it we get $P(A, B)$. You should only apply factor marginalisation if you have accounted for all factors covering the the variables in the scope of the factor to be marginalised and the variables in the scope are separated from the rest of the variables in the MN. Separation is covered later in the document.

Direct influence. In an MN, any two variables that are directly connected by an undirected edge can influence one another.

Indirect influence. In an MN, two rvs can influence one another indirectly (that is, via their cascade of influence on other rvs), so long as the variables are connected by at least one path in the MN structure that's free of observations. MNs have no equivalent of the v-structure's 'collider' node.

Active path. When a path between two nodes has no observed nodes, this path is called *active*. Otherwise it's called inactive or blocked.

Separation. If three disjoint sets of nodes $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ in the MN structure $\mathcal{H}$ are such that every path connecting $X \in \mathbf{X}$ to $Y \in \mathbf{Y}$ is blocked by observations in Z, then $\mathbf{Z}$ *separates* $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{H}$, denoted $\text{sep}_{\mathcal{H}}(\mathbf{X}; \mathbf{Y}|\mathbf{Z})$. Separation implies that for any P that factorises over $\mathcal{H}$, $P \models \mathbf{X}, \mathbf{Y}|\mathbf{Z}$.

Markov blanket. With respect to a node X_i in $\mathcal{H}$, the Markov blanket (MB) is the set of nodes in $\mathcal{H}$ that are direct neighbours of X_i. We denote this $\text{MB}_{\mathcal{H}}(X_i)$.

Local independencies. The MN structure $\mathcal{H}$ implies a set of independencies that can be state from each node's perspective: a node is independent of all other nodes given its Markov blanket. This set is called $\mathcal{I}_l(\mathcal{H}) = \{X_i \perp \text{Rest}_{\mathcal{H}}(X_i)|\text{MB}_{\mathcal{H}}(X_i)\}$ and it is a subset of a larger set $\mathcal{I}(\mathcal{H})$ discussed below.

Global independencies. The MN structure $\mathcal{H}$ also implies a set of independencies that can be state for any three sets of nodes $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$. This global set is $\mathcal{I}(\mathcal{H}) = \{\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z} : \text{sep}_{\mathcal{H}}(\mathbf{X}; \mathbf{Y}|\mathbf{Z})\}$.