A Short Course on Graphical Models

2. Structured Representations

Mark Paskin
mark@paskin.org

Review: probability spaces and conditional probability

- A *probability space* (Ω, P) describes our uncertainty regarding an *experiment*; it consists of a *sample space* Ω of possible outcomes, and a *probability measure* P that quantifies how likely each outcome is.
- An *event* $A \subseteq \Omega$ is a set of *outcomes* of the experiment.
- The probability measure P must obey three axioms:
  1. $P(A) \geq 0$ for all events A
  2. $P(\Omega) = 1$
  3. $P(A \cup B) = P(A) + P(B)$ for disjoint events A and B
- When $P(B) > 0$, the *conditional probability of A given B* is defined as

$$P(A \,|\, B) \triangleq \frac{P(A \cap B)}{P(B)}$$

Review: random variables and densities

- A *random variable* $X : \Omega \rightarrow \Xi$ picks out some aspect of the outcome.
- The *density* $p_X : \Xi \rightarrow \Re$ describes how likely X is to take on certain values.
- We usually work with a set of random variables and their *joint density*; the probability space is implicit.
- The two types of densities suitable for computation are *table densities* (for finite-valued variables) and the *(multivariate) Gaussian* (for real-valued variables).
- Using a joint density, we can compute *marginal* and *conditional densities* over subsets of variables.
- *Inference* is the problem of computing one or more conditional densities given observations.

Independence

- Two events A and B are *independent* iff

$$P(A) = P(A \mid B)$$

  or, equivalently, iff

$$P(A \cap B) = P(A) \times P(B)$$

  Note that this definition is symmetric.

- If A and B are independent, learning that B happened does not make A more or less likely to occur.

- Two random variables X and Y are independent if for all x and y

$$p_X(x) = p_{X|Y}(x, y)$$

  In this case we write $X \perp\!\!\!\perp Y$. Note that this corresponds to (perhaps infinitely) many event independencies.

Example of independence

- Let's say our experiment is to flip two fair coins. Our sample space is

$$\Omega = \{(heads, heads), (heads, tails), (tails, heads), (tails, tails)\}$$

  and $P(\omega) = \frac{1}{4}$ for each $\omega \in \Omega$.

- The two events

$$\begin{aligned} A &= \{(heads, heads), (heads, tails)\} \\ B &= \{(heads, heads), (tails, heads)\} \end{aligned}$$

  are independent, since $P(A \cap B) = P(A) \times P(B) = \frac{1}{4}$.

Example of non-independence

- If our experiment is to draw a card from a deck, our sample space is

  $$\Omega = \{\text{A}\heartsuit, 2\heartsuit, \ldots, \text{K}\heartsuit, \text{A}\diamondsuit, 2\diamondsuit, \ldots, \text{K}\diamondsuit, \text{A}\clubsuit, 2\clubsuit, \ldots, \text{K}\clubsuit, \text{A}\spadesuit, 2\spadesuit, \ldots, \text{K}\spadesuit\}$$

  Let $P(\omega) = \frac{1}{52}$ for each $\omega \in \Omega$.

- The random variables

  $$N(\omega) = \begin{cases} n & \text{if } \omega \text{ is the number } n \\ 0 & \text{otherwise} \end{cases}$$

  $$F(\omega) = \begin{cases} \mathit{true} & \text{if } \omega \text{ is a face card} \\ \mathit{false} & \text{otherwise} \end{cases}$$

  are *not* independent, since

  $$\frac{4}{52} = p_N(3) \neq p_{N|F}(3, \mathit{true}) = 0$$

Independence is rare in complex systems

- Independence is very powerful because it allows us to reason about aspects of a system in isolation.
- However, it does not often occur in complex systems. For example, try and think of two medical symptoms that are independent.
- A generalization of independence is *conditional independence*, where two aspects of a system become independent once we observe a third aspect.
- Conditional independence does often arise and can lead to significant representational and computational savings.

Conditional independence

- Let A, B, and C be events. A and B are *conditionally independent given* C iff

$$P(A \mid C) = P(A \mid B \cap C)$$

  or, equivalently, iff

$$P(A \cap B \mid C) = P(A \mid C) \times P(B \mid C)$$

- If A and B are conditionally independent, then once we learn C, learning B gives us no *additional* information about A.

- Two random variables X and Y are conditionally independent given Z if for all x, y, and z

$$p_{X|Z}(x, z) = p_{X|YZ}(x, y, z)$$

  In this case we write $X \perp\!\!\!\perp Y \mid Z$. This also corresponds to (perhaps infinitely) many event conditional independencies.

Common sense examples of conditional independence

Some examples of conditional independence *assumptions*:

- The operation of a car's starter motor and its radio are conditionally independent given the status of the battery.
- The GPA of a student and her SAT score are conditionally independent given her intelligence.
- Symptoms are conditionally independent given the disease.
- The future and the past are conditionally independent given the present. *This is called the Markov assumption.*

An intuitive test of $X \perp\!\!\!\perp Y \mid Z$:

> *Imagine that you know the value of Z and you are trying to guess the value of X. In your pocket is an envelope containing the value of Y. Would opening the envelope help you guess X? If not, then $X \perp\!\!\!\perp Y \mid Z$.*

Example: the burglary problem

- Let's say we have a joint density over five random variables:
  1. $E \in \{true, false\}$: Has an earthquake happened? *Earthquakes are unlikely.*
  2. $B \in \{true, false\}$: Has a burglary happened? *Burglaries are also unlikely, but more likely than earthquakes.*
  3. $A \in \{true, false\}$: Did the alarm go off? *The alarm is usually tripped by a burglary, but an earthquake can also make it go off.*
  4. $J \in \{true, false\}$: Did my neighbor John call me at work? *John will call if he hears the alarm, but he often listens to music on his headphones.*
  5. $M \in \{true, false\}$: Did my neighbor Mary call me at work? *Mary will call if she hears the alarm, but she also calls just to chat.*
- I've just found out that John called. Has my house been burglarized?
- Problem: compute $p_{B|J}(\cdot, true)$ from the joint density p_{EBAJM}.

Inference by enumeration

- By the definition of conditional densities, we have

$$p_{EBAM|J}(e, b, a, true, m) = \frac{p_{EBAJM}(e, b, a, true, m)}{p_J(true)}$$

- By the definition of marginal densities, we have

$$\begin{aligned} p_{B|J}(b, true) &= \sum_e \sum_a \sum_m p_{EBAM|J}(e, b, a, true, m) \\ &= \frac{1}{p_J(true)} \sum_e \sum_a \sum_m p_{EBAJM}(e, b, a, true, m) \end{aligned}$$

- We don't need to compute $p_J(true)$. Since $p_{B|J}(\cdot, true)$ must be normalized, we can instead compute $p_{B|J}(\cdot, true) \times p_J(true)$ and renormalize.
- We do, however, need to perform the sums above; this gets expensive quickly, since they are essentially nested `for` loops.

A representational problem

- Note that the joint density p_{EBAJM} is a table density containing 32 probabilities. (We can store it using 31, since they must sum to one.)
- Where did those 32 probabilities come from? Specifying large densities by hand is difficult and error-prone.
- It is possible to learn the probabilities automatically from data; we will discuss this later. But this does not solve the problem, since the amount of data required scales quickly with the number of probabilities.
- The key to reducing the representational complexity is to make *conditional independence assumptions.*

Using conditional independence assumptions

- We start by applying the chain rule for random variables:

$$\underbrace{p_{EBAJM}}_{31} = \underbrace{p_E}_{1} \times \underbrace{p_{B|E}}_{2} \times \underbrace{p_{A|EB}}_{4} \times \underbrace{p_{J|EBA}}_{8} \times \underbrace{p_{M|EBAJ}}_{16}$$

- Now make the following independence assumptions:
  1. $B \perp\!\!\!\perp E$ *(E and B are independent.)*
  2. $J \perp\!\!\!\perp \{E, B, M\} \mid A$ *(Given A, J is independent of all other variables.)*
  3. $M \perp\!\!\!\perp \{E, B, J\} \mid A$ *(Given A, M is independent of all other variables.)*
- Then by the definition of conditional independence,

$$\underbrace{p_{EBAJM}}_{31} = \underbrace{p_E}_{1} \times \underbrace{p_B}_{1} \times \underbrace{p_{A|EB}}_{4} \times \underbrace{p_{J|A}}_{2} \times \underbrace{p_{M|A}}_{2}$$

This product of *factors* can be represented with only 10 probabilities!

Conditional independencies make densities modular

- Not only do conditional independencies reduce the space required to represent densities, they make it easier to specify them.
- For example, the factors of our BURGLARY density are easy to assess:
  - p_E represents the background frequency of earthquakes;
  - p_B represents the background frequency of burglaries;
  - $p_{A|BE}$ represents the conditional probability the alarm goes off given an earthquake and/or burglary does/doesn't occur;
  - $p_{J|A}$ is the conditional probability John calls given the alarm does or doesn't sound; and
  - $p_{M|A}$ is the conditional probability Mary calls given the alarm does or doesn't sound.

Conditional independencies speed inference

- Conditional independencies can also lead to efficient inference.
- Substitute in our factorized density:

$$p_{B|J}(b, true) \propto \sum_e \sum_a \sum_m p_{EBAJM}(e, b, a, true, m)$$
$$= \sum_e \sum_a \sum_m p_E(e) \cdot p_B(b) \cdot p_{A|EB}(a, e, b) \cdot p_{J|A}(true, a) \cdot p_{M|A}(m, a)$$

- Now, use the fact that $\times$ distributes over $+$:

$$\underbrace{xy + xz}_{\text{two multiplies and one addition}} = \underbrace{x(y + z)}_{\text{one multiply and one addition}}$$

The Variable Elimination Algorithm

Repeat: (1) choose a variable to eliminate; (2) push in its sum as far as possible; and (3) compute the sum, resulting in a new factor:

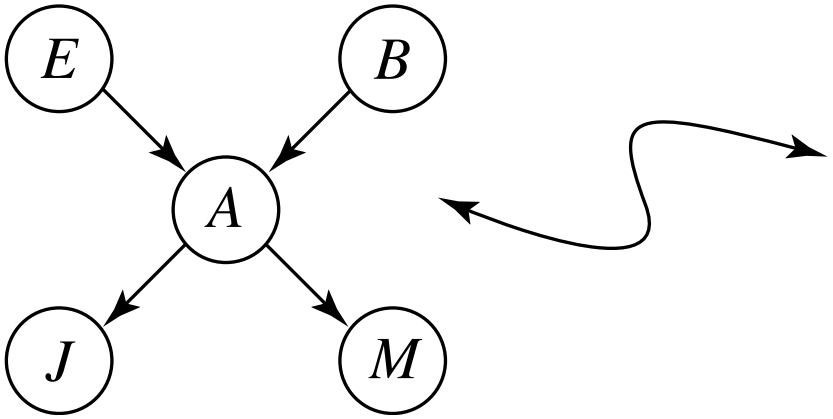
$$
\begin{aligned}
p_{B|J}(b, \mathit{true}) &\propto \sum_e \sum_a \sum_m p_E(e) \cdot p_B(b) \cdot p_{A|EB}(a, e, b) \cdot p_{J|A}(\mathit{true}, a) \cdot p_{M|A}(m, a) \\
&= \sum_e \sum_a p_E(e) \cdot p_B(b) \cdot p_{A|EB}(a, e, b) \cdot p_{J|A}(\mathit{true}, a) \cdot \sum_m p_{M|A}(m, a) \\
&= \sum_e \sum_a p_E(e) \cdot p_B(b) \cdot p_{A|EB}(a, e, b) \cdot p_{J|A}(\mathit{true}, a) \cdot \psi_A(a) \\
&= \sum_e p_E(e) \cdot p_B(b) \cdot \sum_a p_{A|EB}(a, e, b) \cdot p_{J|A}(\mathit{true}, a) \cdot \psi_A(a) \\
&= \sum_e p_E(e) \cdot p_B(b) \cdot \psi_{EB}(e, b) \\
&= p_B(b) \cdot \sum_e p_E(e) \cdot \psi_{EB}(e, b) \\
&= p_B(b) \cdot \psi_B(b)
\end{aligned}
$$

Complexity of Variable Elimination

- The time and space complexities of Variable Elimination are dominated by the size of the largest intermediate factor.
- Choosing the elimination ordering that minimizes this complexity is an NP-hard problem.
- Often, the sizes of the intermediate factors grow so large that inference is intractable, but there are many interesting models where it does not. The difference between these two cases lies in the independence properties.
- *Graphical models* are powerful tools for visualizing the independence properties of complex probability models. There are two kinds:
  - Directed graphical models (Bayesian networks)
  - Undirected graphical models (Markov random fields)

Bayesian networks

- A *Bayesian network* (a.k.a. *Bayes net*, *directed graphical model*, or *Belief network*) is a directed acyclic graph that encodes the independence properties of a joint density.
- It has a node for each random variable, and it has an edge $X \to Y$ if the factor for Y conditions on X (i.e., there is a factor of the form $p_{Y|\ldots X\ldots}$).
- The structure of a Bayes net determines the density's factorization, and vice versa.



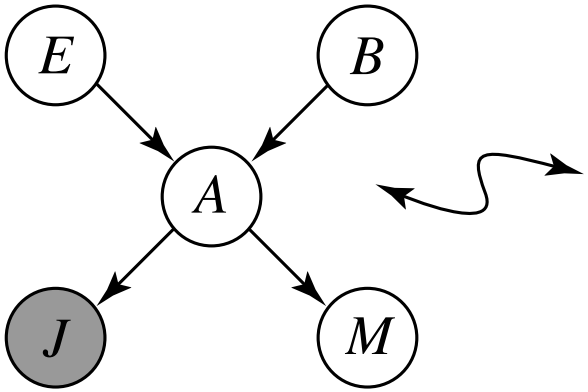
$$p_{EBAJM} = p_E \times p_B \times p_{A|EB} \times p_{J|A} \times p_{M|A}$$

Constructing Bayesian networks

- The formal technique has three steps:
  1. choose an ordering of the variables;
  2. apply the chain rule; and
  3. use conditional independence assumptions to prune parents.
- Be careful: the variable ordering matters. Choosing the wrong order can lead to a completely connected graph.
- We can think of the edges of a Belief network as representing *direct influence*. Another way to construct the network is to choose the parents of each node, and then ensure that the resulting graph is acyclic.
- Bayes net edges often emanate from causes and terminate at effects, but it is important to remember that *Bayesian networks are not causal models.*

Representing evidence in Bayesian networks

- When we condition on some of the variables, the result is a conditional density that can have different independence properties.
- When we condition on a variable, we shade the corresponding node in the Bayes net.



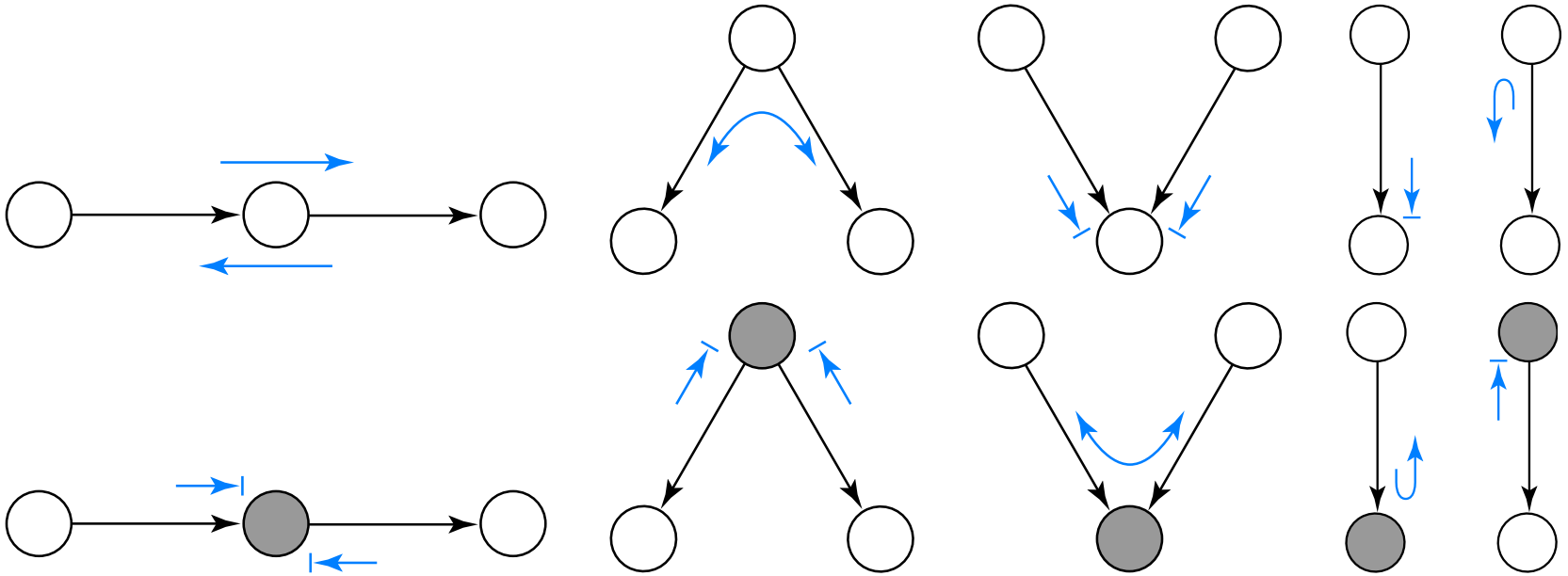
$$p_{EBAM|J} = \frac{p_E \times p_B \times p_{A|EB} \times p_{J|A} \times p_{M|A}}{\sum_e \sum_b \sum_a \sum_m p_E \times p_B \times p_{A|EB} \times p_{J|A} \times p_{M|A}}$$

Encoding conditional independence via d-separation

- Bayesian networks encode the independence properties of a density.
- We can determine if a conditional independence $X \perp\!\!\!\perp Y \,|\, \{Z_1, \ldots, Z_k\}$ holds by appealing to a graph separation criterion called *d-separation* (which stands for *direction-dependent separation*).
- X and Y are d-separated if there is no *active path* between them.
- The formal definition of active paths is somewhat involved. The *Bayes Ball Algorithm* gives a nice graphical definition.

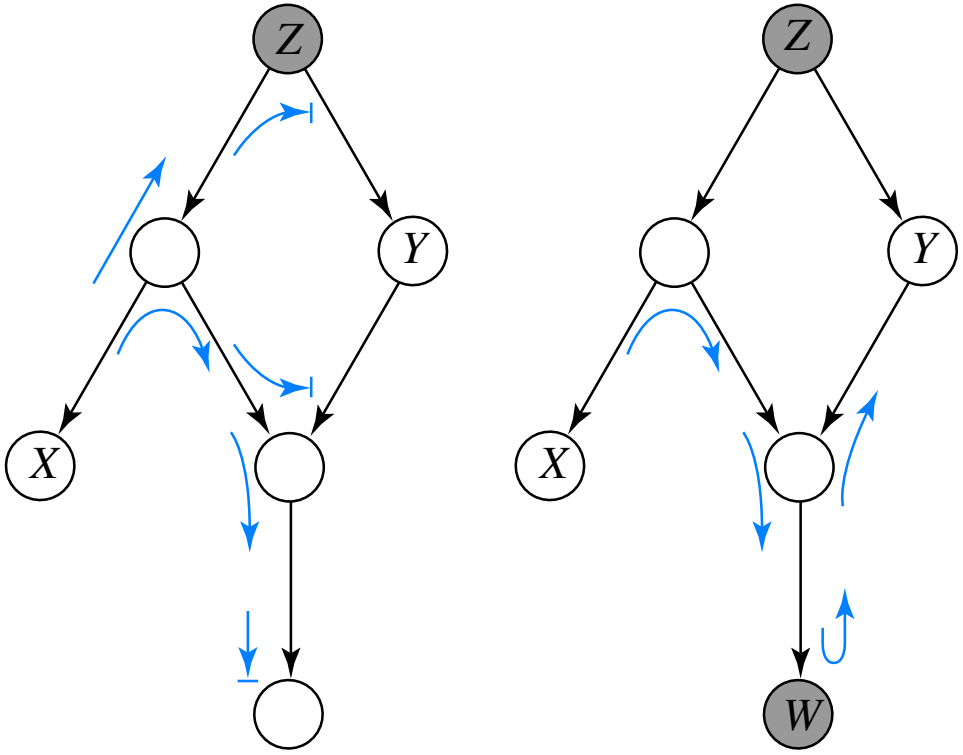
The ten rules of Bayes Ball

An undirected path is active if a Bayes ball travelling along it never encounters the "stop" symbol: $\rightarrow\!|$



If there are no active paths from X to Y when $\{Z_1, \ldots, Z_k\}$ are shaded, then $X \perp\!\!\!\perp Y \,|\, \{Z_1, \ldots, Z_k\}$.

A double-header: two games of Bayes Ball



no active paths

$X \perp\!\!\!\perp Y \mid Z$

one active path

$X \not\perp\!\!\!\perp Y \mid \{W, Z\}$

Undirected graphical models

- A *potential function* is a non-negative function.
- We can define a joint density by a *normalized product of potential functions.* For example, we could define the BURGLARY density as follows:

  $$p_{EBAJM}(e, b, a, j, m) = \frac{1}{Z}\psi_E(e) \cdot \psi_B(b) \cdot \psi_{AEB}(a, e, b) \cdot \psi_{JA}(j, a) \cdot \psi_{MA}(m, a)$$

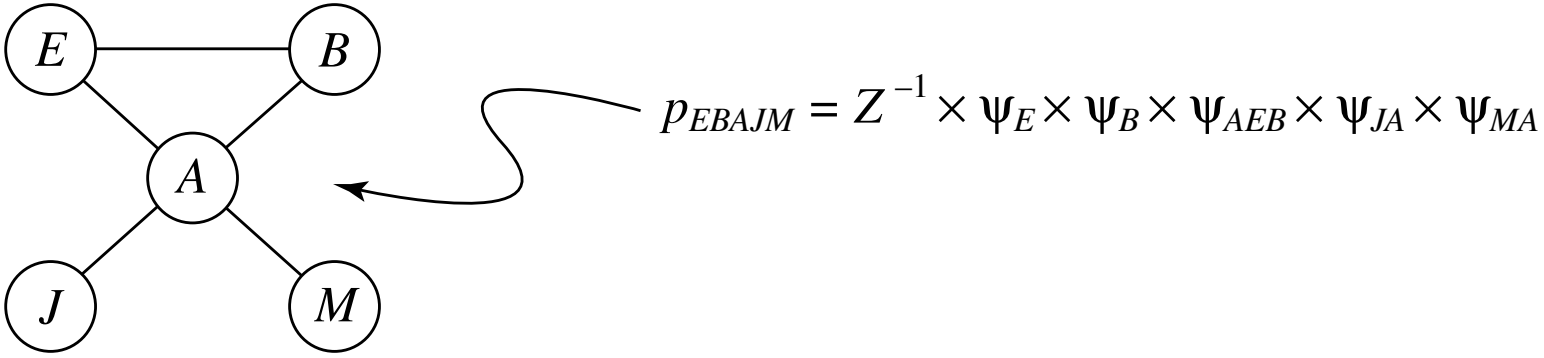
  where each ψ function is a potential and

  $$Z = \sum_e \sum_b \sum_a \sum_j \sum_m \psi_E(e) \cdot \psi_B(b) \cdot \psi_{AEB}(a, e, b) \cdot \psi_{JA}(j, a) \cdot \psi_{MA}(m, a)$$

  is the normalization constant (a.k.a. *partition function*).
- In general the potentials do not have a probabilistic interpretation, but they are interpretable: values with higher potential are more probable. The potentials trade-off with each other via the partition function.
- Multivariate Gaussians can be represented in this way.

Conditional independence in undirected graphical models

- For a density $p = \frac{1}{Z} \prod_i \psi_i$ we define an undirected graph G as follows:
  - Each variable of p becomes a node of G.
  - For each potential ψ_i we place a *clique* over its arguments in G.



- This is called an *undirected graphical model* (a.k.a. *Markov random field*).
- Then $X \perp\!\!\!\perp Y \mid \{Z_1, \ldots, Z_k\}$ if X is *separated* from Y by $Z_1, \ldots, Z_k$, i.e., if when $Z_1, \ldots, Z_k$ are removed there is no path between X and Y.

The Hammersley-Clifford Theorem

- When p is *strictly positive*, the connection between conditional independence and factorization is much stronger.
- Let G be an undirected graph over a set of random variables $\{X_1, \ldots, X_k\}$.
- Let $\mathcal{P}_1$ be the set of positive densities over $\{X_1, \ldots, X_k\}$ that are of the form

$$p = \frac{1}{Z} \prod_C \psi_C$$

  where each ψ_C is a potential over a clique of G.
- Let $\mathcal{P}_2$ be the set of positive densities with the conditional independencies encoded by graph separation in G.
- Then $\mathcal{P}_1 = \mathcal{P}_2$.

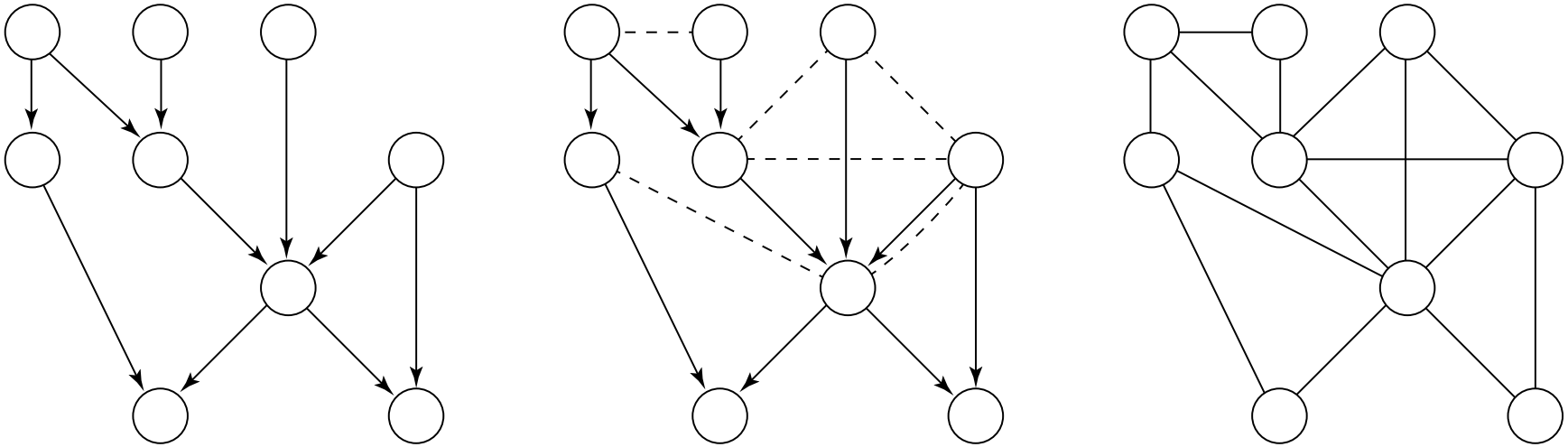
Comparing directed and undirected graphical models

- Specifying an undirected graphical model is easy (normalized product of potentials), but the factors don't have probabilistic interpretations. Specifying a directed graphical model is harder (we must choose an ordering of the variables), but the factors are marginal and conditional densities.
- Determining independence in undirected models is easy (graph separation), and in directed models it is hard (d-separation).
- Directed and undirected models are different languages: there are densities with independence properties that can be described only by directed models; the same is true for undirected models.
- In spite of this, inference in a directed model usually starts by converting it into an undirected graphical model with *fewer* conditional independencies.

Moralization

- Because the factors of a Bayesian network are marginal and conditional densities, they are also potential functions.
- Thus, a directed factorization is also an undirected factorization (with $Z = 1$). Each clique consists of a variable and its parents in the Bayes net.
- We can transform a Bayesian network into a Markov random field by placing a clique over each family of the Bayesian network and dropping the directed edges.
- This process is called *moralization* because we *marry* (or connect) the variable's parents and then drop the edge directions.

Example of moralization



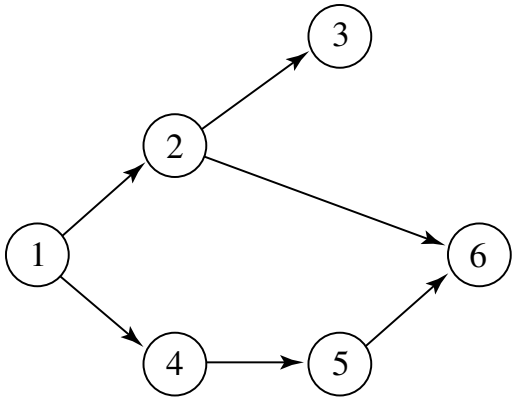
All of the conditional independencies represented by the undirected model are also present in the directed model, but the reverse is not true: *the directed model has conditional independencies that are not represented by the undirected model.*

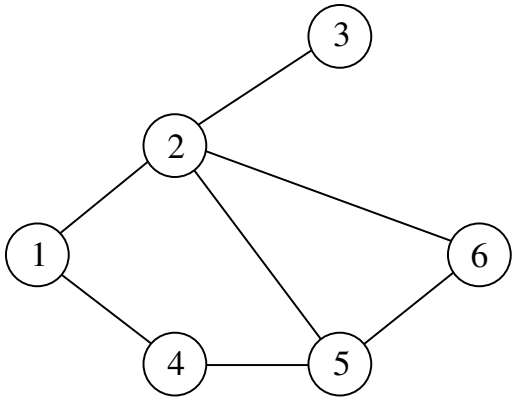
Node Elimination: a graphical view of Variable Elimination

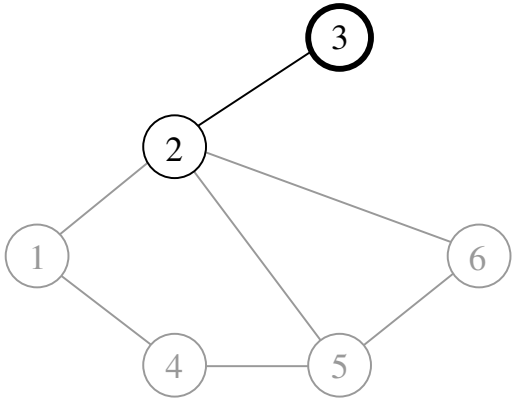
- The intermediate factors we created in the Variable Elimination algorithm are also potential functions.
- Thus, after each elimination step we are left with a density that can be visualized as an undirected graphical model.
- The Node Elimination algorithm is to repeatedly:
  1. choose a node (variable) to eliminate;
  2. create an *elimination clique* (intermediate factor) from the node and its neighbors; and
  3. remove the node and its incident edges.

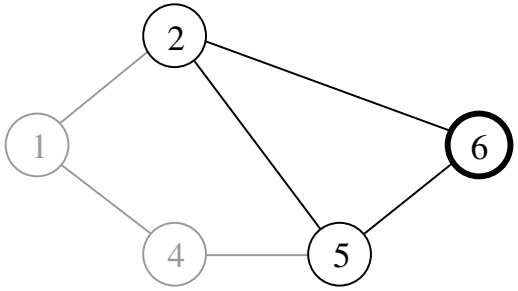
  The result is a set of elimination cliques that represent the arguments of the intermediate factors.

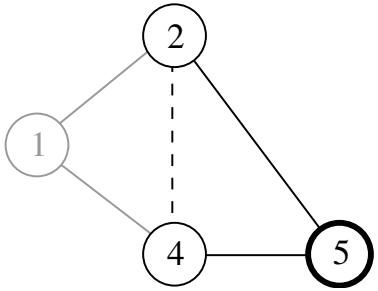
Example of node elimination

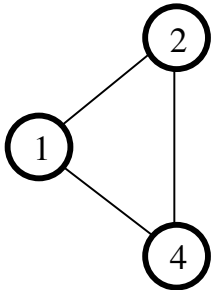






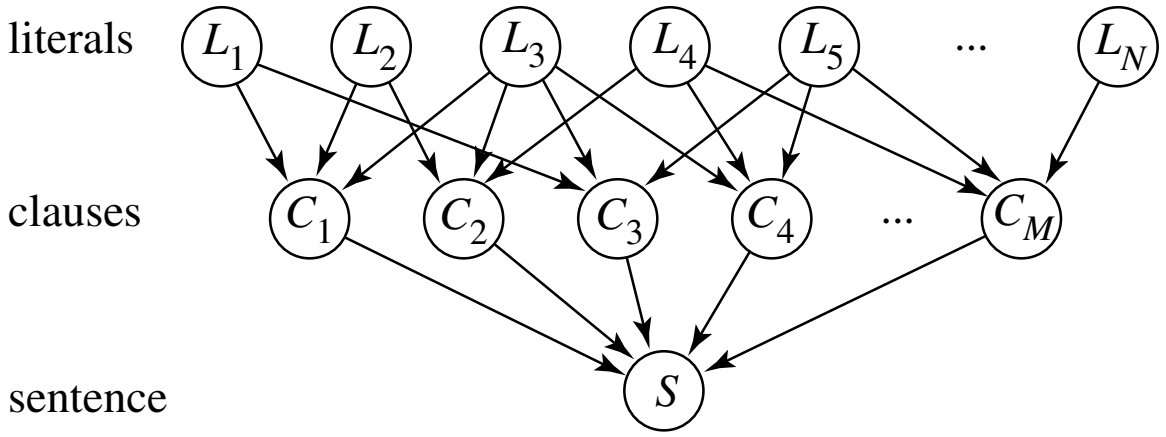






Complexity of the Inference problem

- Probabilistic inference can easily be shown to be NP and even #P hard.
- Reduction to 3-SAT: given a 3-CNF sentence ϕ, build the Bayes net:



where $p_{L_i}(true) = 0.5$, $p_{C_m|L_iL_jL_k}(true, \cdot, \cdot, \cdot) = 1$ iff C_m is satisfied by the assignment, and $p_{S|C_1 \cdots C_M}(true, \ldots) = 1$ iff the clauses are all true. Then

$$p_S(true) = \frac{\#\ of\ satisfying\ assignments\ of\ \phi}{2^N}$$

Looking ahead. . .

- Graphical models provide a language for characterizing independence properties. They have two important uses:
  - *Representational.* We can obtain compact joint densities by choosing a sparse graphical model and then choosing its local factors.
  - *Computational.* For example, Variable Elimination can be viewed in terms of operations on a graphical model. This model can be used to guide computation, e.g., it helps in choosing good elimination orders.
- Next time we will continue to explore the connection between Graph Theory and Probability Theory to obtain
  - graphical characterizations of tractable inference problems, and
  - the junction tree inference algorithms.

Summary

- Two random variables X and Y are *(marginally) independent* (written $X \perp\!\!\!\perp Y$) iff $p_X(\cdot) = p_{X|Y}(\cdot, y)$ for all y.
- If $X \perp\!\!\!\perp Y$ then Y gives us no information about X.
- X and Y are *conditionally independent given* Z (written $X \perp\!\!\!\perp Y \mid Z$) iff $p_{X|Z}(\cdot, z) = p_{X|YZ}(\cdot, y, z)$ for all y and z.
- If $X \perp\!\!\!\perp Y \mid Z$ then Y gives us no additional information about X once we know Z.
- We can obtain compact, factorized representations of densities by using the chain rule in combination with conditional independence assumptions.
- The Variable Elimination algorithm uses the distributivity of $\times$ over $+$ to perform inference efficiently.

Summary (II)

- A *Bayesian network* encodes the independence properties of a density using a directed acyclic graph.
- We can answer conditional independence queries by using the *Bayes Ball algorithm*, which is an operational definition of *d-separation*.
- A *potential* is a non-negative function. One simple way to define a joint density is as a *normalized product of potential functions.*
- An *undirected graphical model* for such a density has a clique of edges over the argument set of every potential.
- In an undirected graphical model, graph separation corresponds to conditional independence.
- *Moralization* is the process of converting a directed graphical model into an undirected graphical model. This process does not preserve all of the conditional independence properties.