

## Graphical Models

*Lecturer: Drew Bagnell**Scribe: Ryan Waliany*

## 1 Graphical Models

### 1.1 What is graphical model?

Graphical model is a language that can be used to represent uncertain quantities and the relationships between them. Nodes represent random variables and edges represent the links, or causal relationships between these random variables.

### 1.2 Key ideas of graphical models

- **Communication tool** that helps to *compactly* express my beliefs of a system.
- **Reasoning tool** that can be used to *extract* relationships that maybe were not obvious when formulating the problem.
- **Computational framework** that helps organize how we perform *inference*.

## 2 Bayes' nets

### 2.1 Introduction

One of the most common graphical models is called a Bayes' net, also known as: a belief network, directed graphical model, directed independence diagram. In short, a Bayes' net is a directed acyclic graph with nodes representing uncertain quantities.

In Figure 1, we have uncertain quantities A, B, C, and we draw directed arrows between them to represent causal relationships. A bayesian network encodes a joint probability distribution over all the nodes in the graph. In this case, our Bayes' net encodes the joint probability distribution,  $P(A, B, C, D)$ .

The basic factorization of the probability distribution using the Chain Rule of Probability can be seen in Equation 1. If no graphical model is specified, this factorization is always true.

$$P(A, B, C, D) = P(A) * P(B|A) * P(C|A, B) * P(D|A, B, C) \quad (1)$$

In particular with Figure 1, we can use the causal information contained in the graph to eliminate unnecessary conditional dependencies, see equation 2.

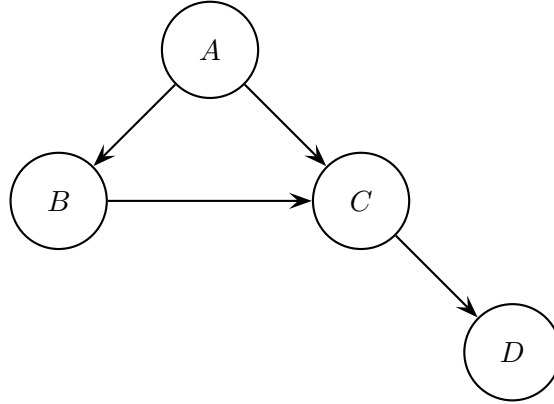


Figure 1: A Bayesian network.

$$P(A, B, C, D) = P(A)P(B|A)P(C|A, B)P(D|C) \quad (2)$$

For an arbitrary Bayes' net with nodes  $\in X$  where  $X$  is defined as  $x_1, x_2, \dots, x_n$ . We can derive the joint distribution of such a graph,  $P(X)$  as the product of each node ( $x_i$ ) given it's parents ( $\pi(x_i)$ ), see equation 3.

$$P(X) = \prod_{x_i} P(x_i|\pi(x_i)) \quad (3)$$

Looking back at equation 2, we can derive that quantity using equation 3 and 1.

The mental model of what a Bayes' net should be that it is a language for encoding a causal relationship. One should think that A is influencing B and C.

## 2.2 Communication tool

In the classical burglar problem seen in Figure 2, you have three states, earthquake, burglar, and burglar alarm. Earthquake and Burglar are events that happen independently. If one of the two things happen, then the burglar alarm goes off. One should think that if either happens, then the burglar alarm should go off, but it doesn't describe the actual relationship but what happens.

Whenever you write down a Bayes' net, you are thinking of a simplified causal model of the world. This will allow us to infer the probability of such quantities like when there is a burglar. There are many algorithms that will learn Bayes' nets from data, which we will see in this class. If you don't specific a Bayes' net down deliberately you may get a Bayes' net that is doesn't preserve causality.

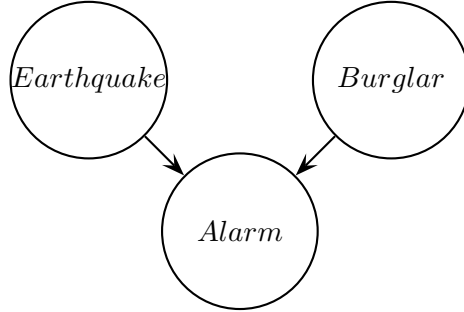


Figure 2: A classic example in graphical models.

For example, there are many Bayes' net factorization that can represent the same structure in the data. In general, if we have an independent set of variables  $x_1, x_2, x_3$ , we can always add causal relationships (edges) in a Bayes' net that are not necessarily without perturbing the data incorrectly. However, removing dependencies in a Bayes' net may destruct the validity of a joint distribution. The primary reason to write down a Bayes' net is to represent causality. But it doesn't necessarily have to be causal.

### 2.3 Reasoning tool

Bayes' nets have the ability that you can read from them conditional independence relationships. What is an example of a representation that can encode the Markov relationship between states?

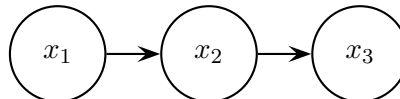


Figure 3: A Bayesian network that demonstrates the Markov assumption.

If you think of this as the robot location,  $x_1$  is causally influenced by  $x_0$ . You don't necessarily need to know  $x_1$  to know  $x_3$  given information about  $x_2$ . The key idea is that we are going to be able to determine conditional independencies by operations on the graph. Conditional independence can be defined as: Given C, A and B are independent of each other. This relationship is explored mathematically in equation 4

$$P(A|B, C) = P(A|C) \tag{4}$$

What about the case that A and B are independent? Does that necessarily mean that A and B are conditionally independent? No.

Let's look at the burglar graph in Figure 2 and when we condition on the burglar alarm. If we know that the burglar didn't come, how does that influence the probability of the an earthquake happening. It increases the probability of an earthquake, because we know that a burglar didn't arrive. This phenomena is called the explaining away effect.

Pretend you can get into CMU because your math GRE's were really good, you wrote a really good paper, or you like bag-pipes and that they were independent events. You may really like bag-pipes. Now, I told you that you got into CMU. If I told you that you really liked bag-pipe music, the probability of you having strong GRE scores may not go up as much. Your GRE math scores may be higher than the normal population because you got into CMU, but they might be lower because part of the reason you may have gotten into CMU was due to liking bag-pipes.

The idea behind reading conditional independencies is that you're trying to find a path between two random variables.

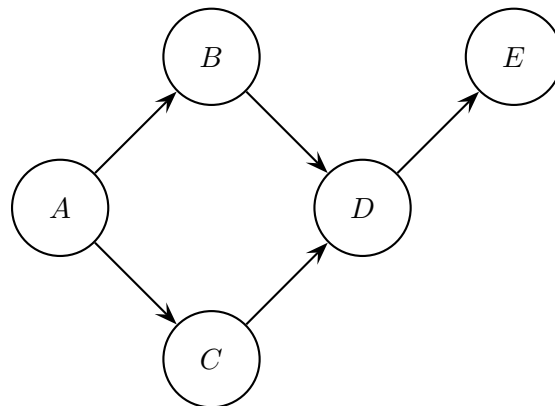


Figure 4:  $A$  is  $\perp$  to  $E$  given  $B$  and  $C$ .

Let's say that you want to find the relationship between  $A$  and  $E$  in Figure 5, so there are two paths (this is not just any path, but a directed path). For two things to possibly not be dependent, you need to block all paths to  $E$  from  $A$ . That is, if there is a path from  $A$  to  $E$ ,  $A$  and  $E$  are dependent.

For the possibility of dependencies, there must be an **unblocked** path between the two variables.  $A$  and  $B$  are said to be dependent on each other if there is an unblocked path.

There are two rules for blocked path:

### 2.3.1 Rule #1

Additionally, say you look at a node along a path and there are no converging arrows (see Figure 7), the path is then blocked. Figure 5 and 6 show the two cases in which conditional independence is achieved absent of converging arrows. In these graphs, there are no conditional dependence between  $A$  and  $B$ .

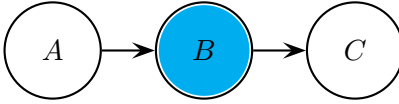


Figure 5: **BLOCKED**  $A$  is  $\perp$  to  $C$  given  $B$ .

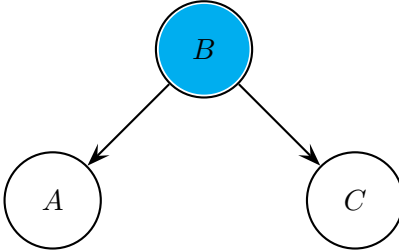


Figure 6: **BLOCKED**  $A$  is  $\perp$  to  $C$  given  $B$ .

### 2.3.2 Rule #2

Remember the burglar example, if I knew that the burglar alarm went off, then what happens to  $A$  (Earthquake) and  $C$  (Burglar)? The answer is that they become dependent! Knowing that a burglar came, explains away the probability that the alarm was caused by an earthquake. The general relationship for this case can be seen in Figure 7. This is the converging arrows case that causes conditional dependencies in the landmark localization formulation, seen later in Figure 11.

Additionally, with the converging arrows case, we must be careful since any **descendent** of a node with converging arrows point to it, such as  $B$  in Figure 7, will facilitate a path from  $A$  to  $C$ . This can be seen more formally in Figure 8. We can model our intuition using the burglar alarm and by adding a new variable  $D$ , which represents a police car. If, we observe that a police car is driving by, there is a higher likelihood that there was an alarm. If there is a higher likelihood that there was an alarm and there was actually a burglary—we could reason that there would be a lower chance of an earthquake and thus, they would not be conditionally independent.

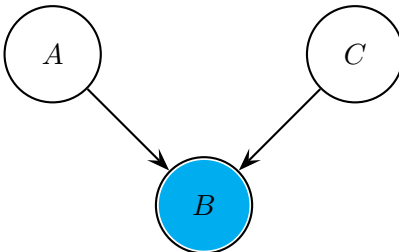


Figure 7: **UNBLOCKED**  $A$  is conditionally dependent on  $C$  given  $B$ .

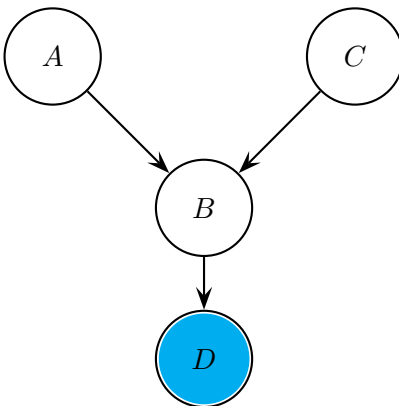


Figure 8: **UNBLOCKED**  $A$  is conditionally dependent on  $C$  given  $D$ .

## 2.4 Examples

### 2.4.1 Localization

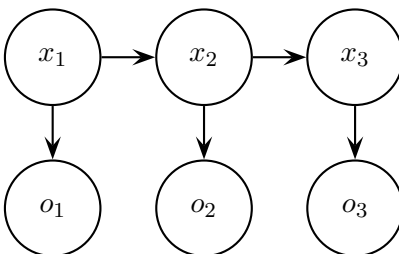


Figure 9: A sample localization bayesian network with internal states  $x_i$  and observations  $o_i$ .

The standard localization model can be seen as represented in a Bayesian network in Figure ?? . Pretend that you are given where a robot is located at time  $x_3$ . We may want to ask whether two quantities are conditionally independent. For example, we may ask whether  $o_2$  is conditionally independent of  $o_1$  given  $x_3$ . We can write this more formally as seen in Equation 5.

$$o_2 \perp o_1 | x_3 \tag{5}$$

We often use the symbol  $\perp$  to represent independence in graphical models. Equation 5 is actually false. There exists a path from  $o_1$  to  $x_1$ ,  $x_2$ , and back to  $o_2$ . Similarly, can we say that  $o_1 \perp o_3 | x_3$ ? Yes, because the path from  $x_2$  to  $o_3$  is blocked by  $x_3$ .

## 2.4.2 SLAM

Let's look at SLAM, how can we generalize this picture. See Figure 10.

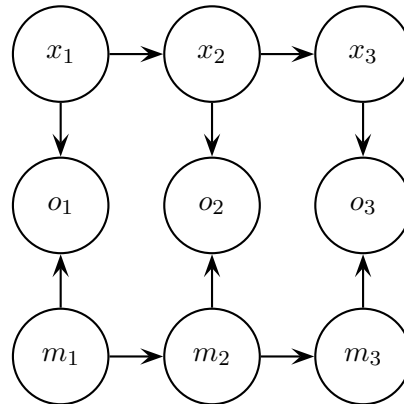


Figure 10: A sample localization and mapping bayesian network with internal states  $x_i$ , observations  $o_i$ , and mapping states  $m_i$ .

In this case, we're thinking of two uncertain quantities, the position ( $x_i$ ) and the map ( $m_i$ ). Tell me whether there exists conditional independence between  $o_1$  and  $o_3$  now, if I tell you  $x_2$ . There is, because the observation tells you something about the map, and the map influences the likelihood of our future observation. There now exists a path from  $o_1$  to  $m_1, m_2, m_3$ , and finally to  $o_3$ .

### 2.4.3 Landmark

Now let's think of a landmark map.

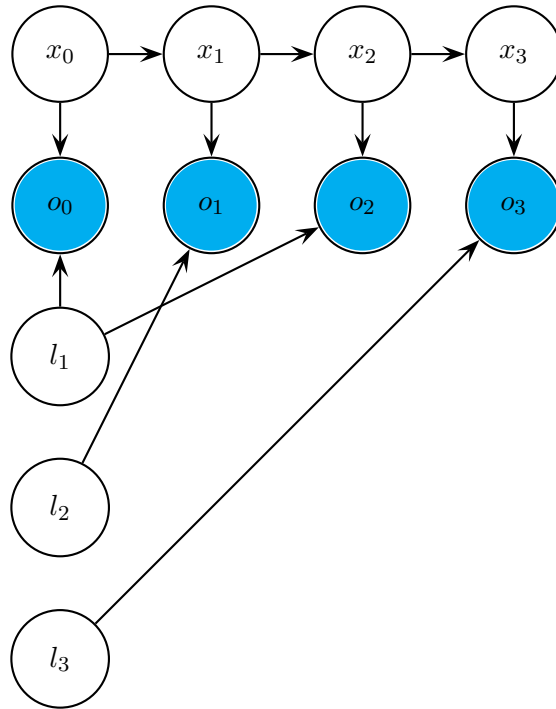


Figure 11: A sample landmark bayesian network with internal states  $x_i$  and observations  $o_i$  and landmarks  $l_i$ .

If we pose the case that we see all the observations,  $o_0, o_1, o_2, \dots$ , are the landmarks conditionally independent of each other? That is, is  $l_1 \perp l_2 | O$ ? This is a tricky case, where the converging arrows at  $o_i$  create a path between  $l_1$  and  $x_0$ . This path extends through  $x_1$  and back from  $o_1$  to  $l_2$ . Thus,  $l_1$  and  $l_2$  are not conditionally independent given  $o_0$  and  $o_1$ .

We will see later that the more conditional independencies that our graphical models exhibit, the simpler computationally complexity we will have. Likewise, if we can simply observe the values of  $X$ , we can prevent conditional dependencies between  $l_1$  and  $l_2$ . This formalization is called Fast SLAM and can be seen in Figure 12. We will discuss this in more depth later.



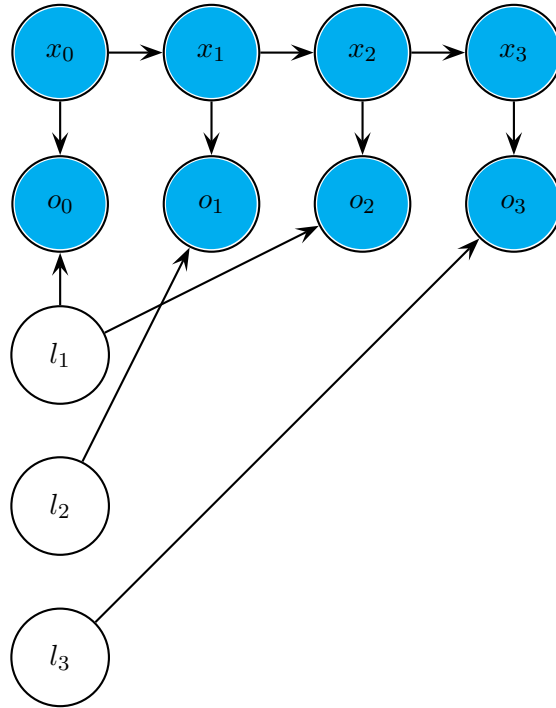


Figure 12: The Fast SLAM formulation.

### 3 Gibbs Field

A Gibbs Field is a collection of nodes that have undirected edges between them, this can be seen in Figure 13.

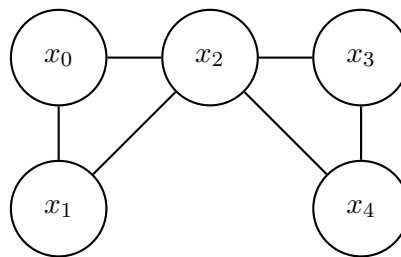


Figure 13: A Gibbs Field with nodes  $x_0, x_1, x_2, x_3, x_4$ .

A clique is a fully connected subset of the the graph.  $x_0, x_1$ , and  $x_2$  forms a clique and  $x_2, x_3$ , and  $x_4$  also forms a clique. These cliques are known as maximal cliques because they are not part of any larger clique. We only need to pay attention to maximal cliques, even though several individual cliques exist, e.g.  $x_0, x_1, \dots$

We can represent a joint probability distribution, much like a Bayesian networks by multiplying a set of clique potential functions  $\phi$  (see Equation 6).

$$P(\vec{x}) = \frac{\prod_{cliques} \phi_{cliques}(X_{cliques})}{Z} \quad (6)$$

The first clique has a function representing the probability distribution of the set of three variables  $x_0, x_1, x_2$ :  $\phi(x_0, x_1, x_2)$ . Unlike normal probability distributions, these potential functions need not necessarily normalize. There is also a potential function  $\phi(x_2, x_3, x_4)$ . Since the potential functions are not normalized, we must enforce a normalization constant,  $Z$  (see Equation 7) in our equation 6 to create a valid probability distribution.

$$Z = \sum_x \prod_{cliques} \phi_{cliques}(x_{cliques}) \quad (7)$$

Additionally, you might think of the potential functions as energy functions, such as  $\phi(x_1, x_2) = e^{-(x_1 - x_2)^2}$ . Lower energies are obtained when  $x_1$  and  $x_2$  have almost the same value. We could derive the probability distribution then, by simply multiplying these cliques together and normalizing. The minimum energy state in this case is going to be flat. It's often common to have bigger cliques that capture the notion that not only is the ground flat, but it has constant slope.

### 3.1 Conditional Independence

In a Gibbs field, conditional independency is simple. If there is any **unblocked** path between a set of nodes, they are conditionally dependent. Blocking is defined as simply observing a random variable. Conditionally independent is then defined iff there is no path to get from A to B.