COS513 LECTURE 2
CONDITIONAL INDEPENDENCE AND FACTORIZATION

HAIPENG ZHENG, JIEQI YU

Let $\{X_1, X_2, \cdots, X_n\}$ be a set of random variables. Now we have a series of questions about them:

- What are the conditional probabilities (answered by normalization / marginalization)?
- What are the independencies (answered by *factorization* of the joint distribution)?

For now, we assume that all random variables are discrete. Then the joint distribution is a table:

$$p(x_1, x_2, \cdots, x_n). \tag{0.1}$$

Therefore, Graphic Model (GM) is a *economic representation of joint distribution taking advantage of the local relationships between the random variables.*

1. Directed Graphical Models (DGM)

A directed GM is a Directed Acyclic Graph (DAG) $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where

- $\mathcal{V}$, the set of vertices, represents the random variables;
- $\mathcal{E}$, the edges, denotes the "parent of" relationships.

We also define

$$\mathit{\Pi}_i = \text{set of parents of } X_i. \tag{1.1}$$

For example, in the graph shown in Figure 1.1, The random variables are

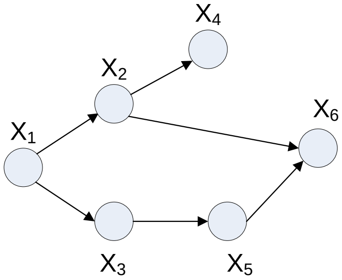

FIGURE 1.1. Example of Graphical Model

$\{X_1, X_2 \cdots, X_6\}$, and

$$\mathit{\Pi}_6 = \{X_2, X_3\}. \tag{1.2}$$

This DAG represents the following joint distribution:

$$p(x_{1:6}) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5). \tag{1.3}$$

In general,

$$p(x_{1:n}) = \prod_{i=1}^{n} p(x_i|x_{\pi_i}) \tag{1.4}$$

specifies a particular joint distribution. Note here π_i stands for the set of indices of the parents of i.

For those who are familiar with the term Bayesian Network, it is worth pointing out that DGM is basically a Bayesian Network. Why cycles is not allowed for DGM? We will address the reason for this "acyclic" requirement in the later lectures.

The GM is essentially a *family of distributions*, it is a family of *those who respect the factorization implied by the graph.*

If we assume that $X_{1:6}$ are all binary random variables, then the naive representation (the complete table of the joint distribution) has 2^6 entries in the table. Yet if we notice that the representation for $p(x_3|x_1)$ has only 4 entries, then the GM representation for the joint distribution has only $\sum_{i=1}^{6} 2^{|\pi_i|+1} = 24$ entries. Thus, we replace an exponential growth in n, the total number of nodes, to an exponential growth in $|\pi_i|$,the number of parents. Therefore, GM representation provides eminent saving in space.

1.1. **Conditional Independence.**

First we define independence and conditional independence:

- Independence:

$$X_A \perp\!\!\!\perp X_B \iff p(x_A, x_B) = p(x_A)p(x_B)$$

- Conditional Independence:

$$\begin{aligned} X_A \perp\!\!\!\perp X_B|X_C &\iff p(x_A, x_B|x_C) = p(x_A|x_c)p(x_B|x_C) \\ &\iff p(x_A|x_B, x_C) = p(x_A|x_c) \end{aligned}$$

Independence is akin to factorization, hence akin to examination of the structure of the graph.

1.2. **Basic Independencies.** The *Chain Rule* of the probability is:

$$p(x_{1:n}) = \prod_{i=1}^{n} p(x_i|x_{1:i-1}). \tag{1.5}$$

For example,

$$p(x_{1:6}) = p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)\cdots p(x_6|x_1,\cdots,x_5). \tag{1.6}$$

This is suggestive to independencies. *Conditional independencies are imbedded in the graph.* By comparing equations (1.3) and (1.6), it suggests

$$p(x_6|x_1,\cdots,x_5) = p(x_6|x_2,x_5), \tag{1.7}$$

which is equivalent to

$$X_6 \perp\!\!\!\perp X_1, X_3, X_4 \,|\, X_2, X_5. \tag{1.8}$$

By strict computation, we can show that this is indeed true. (See Appendix A.1)

Let I be a *topological ordering*, this is equivalent to π_i appears before i, $\forall i$. Let ν_i be the set of indices appearing before i in I. Then we have a series of conditional independencies, given a topological ordering:

$$\{X_i \perp\!\!\!\perp X_{\nu_i}|X_{\pi_i}\}. \tag{1.9}$$

And these conditional independencies are called *basic independencies*. For the GM in Figure 1.1, a possible topological ordering is

$$I = \{X_1, X_2, X_3, X_4, X_5, X_6\}, \tag{1.10}$$

then,

$$\begin{aligned}
X_1 &\perp\!\!\!\perp \emptyset|\emptyset, \\
X_2 &\perp\!\!\!\perp \emptyset|X_1, \\
X_3 &\perp\!\!\!\perp X_2|X_1, \\
X_4 &\perp\!\!\!\perp \{X_1, X_3\}|X_2, \\
X_5 &\perp\!\!\!\perp \{X_4, X_2, X_1\}|X_3.
\end{aligned}$$

Basic independencies are *attached* to topological ordering. However, they are *not* all the conditional independencies implied by the graph.

By simple computation (See Appendix A.2), we can confirm that

$$p(x_4|x_1,x_2,x_3) = p(x_4|x_2). \tag{1.11}$$

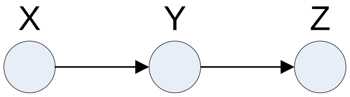

FIGURE 2.1

2. BAYES BALL ALGORITHM

2.1. **Three Simple Graphs.**

(1) *A little sequence*

As shown in Figure 2.1, X can be deemed as the "past", Y the "present" and Z the "future". Based on the graph, we have the joint distribution

$$p(x, y, z) = p(x)p(y|x)p(z|y). \tag{2.1}$$

By simple derivation, we know that

$$X \perp\!\!\!\perp Z|Y. \tag{2.2}$$

This is illustrated in Figure 2.2. Note that shading denotes conditioning.

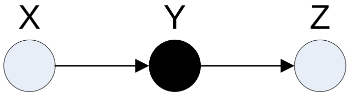

FIGURE 2.2

Here we have the following observations:

(a) This is the only conditional independency implied by this graph.
(b) It suggests that graph separation can be related to the conditional probabilities.
(c) The interpretation of the graph is "the past is independent of the future given the present", which is the famous *Markov Assumption*. The above GM is a simplified version of Markov Chain.

Remark: The "only" conditional independency does not mean that other independencies cannot hold. For some settings, other independencies may hold, but they do not hold for all joint distributions represented by the GM. So, the arrow in the graph between x and y does not mean that y has to depend on x. For some settings of $p(y|x)$, we may have $X \perp\!\!\!\perp Z$, but this independency does not hold for all of the settings.

(2) *A little tree*

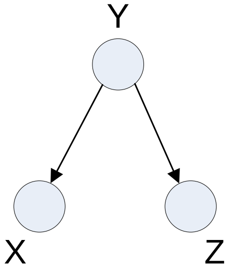

FIGURE 2.3

The joint distribution implied by the GM shown in Figure 2.3 is

$$p(x, y, z) = p(y)p(x|y)p(z|y). \tag{2.3}$$

Notice that

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(y)p(x|y)p(z|y)}{p(y)} = p(x|y)p(z|y), \tag{2.4}$$

so we have $X \perp\!\!\!\perp Z|Y$, as illustrated in Figure 2.4.

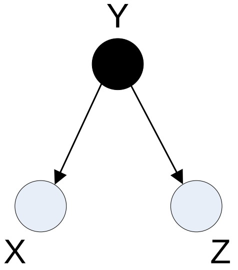

FIGURE 2.4

Again, we have the following observations:

(a) This is the only conditional independency implied by this graph.
(b) It suggests that graph separation can be related to the conditional probabilities.
(c) An interesting interpretation of this conditional independence is like this: Obviously, the "shoe size" (represented by X) and the "amount of gray hair" (represented by Z) of a person is highly dependent, because a boy, with no gray hair, wears small shoes, while an old man, with many gray hairs, wears large

shoes. However, when the “age” (represented by Y) is given, the correlation between “shoe size” and “amount of gray hair” suddenly disappears, since “age” provides all the information that “shoe size” can provide to infer “amount of gray hair”. The same is true for “amount of gray hair”. Thus, given “age”, “shoe size” and “amount of gray hair” are independent. This GM provides us with the intuition for *hidden cause*.

(3) *Inverse Tree*

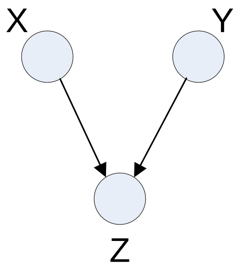

FIGURE 2.5

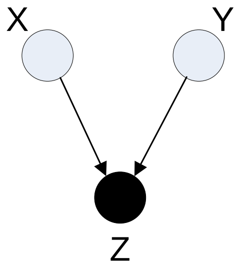

FIGURE 2.6

For the graph shown in Figure 2.5, we observe:

(a) $X \perp\!\!\!\perp Y$ is the only implied independency.
(b) Graph separation is opposed in this case.
(c) It is not necessarily true that $X \perp\!\!\!\perp Y|Z$, as illustrated in Figure 2.6 . This looks a little less intuitive than the previous cases, yet we have a wonderful example: let’s define
   - Z: late for lunch
   - X: lost my watch
   - Y: aliens abduction

> Here, given Z, X and Y "explain away" each other. They decreased each other's possibility if Z is given, hence they are not conditionally independent. To put it more specifically, say, if we know that David is late for lunch, then the two seemingly independent events, "late for lunch" and "lost the watch" suddenly explains each other away. If David comes to lunch and tells us that he lost his watch, then the probability of aliens abduction is very slim; but if we cannot find David and his watch, then the probability of aliens abduction increases.

2.2. **Bayes Ball Algorithm.** Description: Bayes Ball Algorithm is a notion of separability that let us determine the validity of an independency statement in a GM.

Bayes Ball Algorithm is not for actual implementation, but we use it quite often in our minds.

Is there any way to get all the conditional independencies? The only way is to try all configurations on the GM, then run Bayes Ball Algorithm to verify the independencies.

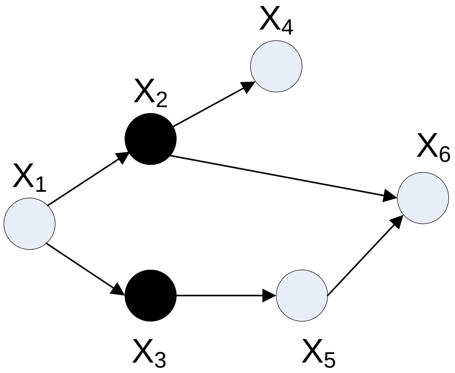

FIGURE 2.7

The basic idea of Bayes Ball Algorithm is like this, say, we want to verify the independency of $X_1 \perp\!\!\!\perp X_6 | X_2, X_3$, as illustrated in Figure 2.7. Then we start a ball at X_1 and try to bounce it to X_6. According to the three rules as shown in Figure 2.8, 2.9 and 2.10. If the ball can bounce from X_1 to X_6, then the conditional independency is not true. Otherwise, the conditional independency is verified.

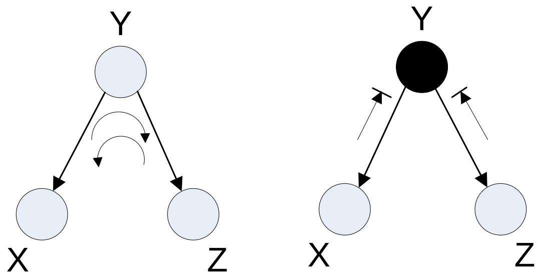

FIGURE 2.8. Rule 1 for Bayes Ball Algorithm. A ball can bounce between X and Z if Y is not given. However, when Y is given, it blocks the way between X and Z and the ball can no longer bounce in between.

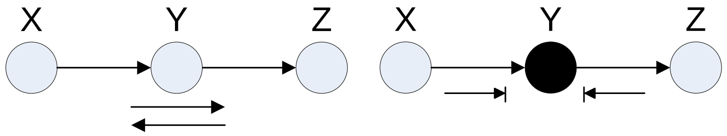

FIGURE 2.9. Rule 2 for Bayes Ball Algorithm. A ball can bounce between X and Z if Y is not given. However, when Y is given, it blocks the way between X and Z and the ball can no longer bounce in between.

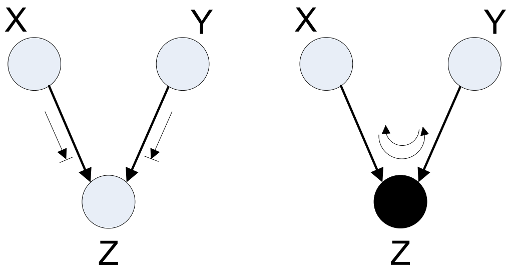

FIGURE 2.10. Rule 3 for Bayes Ball Algorithm. A ball cannot bounce between X and Y if Z is not given. However, when Z is given, it connects X and Y and the ball can bounce in between.

3. CONCLUSION

The punch line of this lecture (H-C Theorem):
Consider two families:

(1) Family of joint distributions found by ranging over all conditional probability tables associated with $\mathcal{G}$ (via factorization);

(2) All joint distributions that respect *all* conditional independence statements implied by $\mathcal{G}$ and d-separation (via Bayes Ball Algorithm).

Theorem. *(H-C Theorem)*
Family 1 and 2 are the same.

We will try to prove this theorem in the next lecture.

APPENDIX A. PROOF IN THIS LECTURE

A.1. **Proof of 1.8.** We only need to show that $p(x_6|x_{1:5}) = p(x_6|x_2, x_5)$.

$$\begin{aligned}
p(x_6|x_{1:5}) &= \frac{p(x_{1:6})}{p(x_{1:5})} \\
&= \frac{p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2,x_5)}{\sum_{x_6} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2,x_5)} \\
&= \frac{p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2,x_5)}{p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)\sum_{x_6} p(x_6|x_2,x_5)} \\
&= \frac{p(x_6|x_2,x_5)}{\sum_{x_6} p(x_6|x_2,x_5)} \\
&= p(x_6|x_2,x_5).
\end{aligned}$$

A.2. **Proof of 1.11.** Here is the derivation:

$$\begin{aligned}
p(x_4|x_1,x_2,x_3) &= \frac{p(x_4,x_1,x_2,x_3)}{p(x_1,x_2,x_3)} \\
&= \frac{\sum_{x_5}\sum_{x_6} p(x_{1:6})}{\sum_{x_4}\sum_{x_5}\sum_{x_6} p(x_{1:6})} \\
&= \frac{\sum_{x_5}\sum_{x_6} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2,x_5)}{\sum_{x_4}\sum_{x_5}\sum_{x_6} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2,x_5)} \\
&= \frac{p(x_4|x_2)\sum_{x_5} p(x_5|x_3)\sum_{x_6} p(x_6|x_2,x_5)}{\sum_{x_4} p(x_4|x_2)\sum_{x_5} p(x_5|x_3)\sum_{x_6} p(x_6|x_2,x_5)} \\
&= \frac{p(x_4|x_2)}{\sum_{x_4} p(x_4|x_2)} \\
&= p(x_4|x_2).
\end{aligned}$$