1 Introduction

1.1 Notation

Capital letters denote random variables:

- $X_1, X_2, \ldots, X_n$ random variables
- $x_1, x_2, \ldots, x_n$ observations of the random variables

Their joint probability mass function is written

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

or as shorthand, we can write this as $p(x_1, x_2, \ldots, x_n)$. We will use these interchangeably.

1.1.1 Example

Let $A = \{1, 4\}$, so $X_A = \{X_1, X_4\}$; $A$ is the set of indices for the r.v. $X_A$.

Also let $B = \{2\}$, $X_B = \{X_2\}$ so we can write

$$P(X_A|X_B) = P(X_1 = x_1, X_4 = x_4|X_2 = x_2).$$

1.2 Graphical Models

Graphs can be represented as a pair: $G = (V, E)$. 
• \( V \) is the set of nodes (vertices)

• \( E \) is the set of edges.

\[
\begin{align*}
&\text{a directed graph} & \text{an undirected graph}
\end{align*}
\]

1.2.1 Directed graphical models

We assume our directed graphs are acyclic, so that edge direction represents “causation”.

• \( A \rightarrow B: A \) “causes” \( B \)

• cycles are meaningless

\[
\begin{align*}
&\text{cyclic: contains a cycle} & \text{an acyclic directed graph}
\end{align*}
\]

We consider a 1-1 map between our graph’s vertices and random values.

1.2.2 Example

Wet grass could be caused by rain, or a sprinkler.
\[ V = \{C, R, S, W\} \]

This directed graph shows the relation between 4 random variables. If we have the joint probability \( P(C, R, S, W) \), then we can answer any queries about this system.

However, the joint probability function grows exponentially with the number of variables.

\[
P(C, R, S, W): \begin{array}{c|c|c|c|c}
 & C & R & S & W \\
\hline
C & & & & \\
R & & & & \\
S & & & & \\
W & & & & \\
\end{array}
\]

Our graph helps us avoid this intractability.

Let’s define local relationships in our graph. For each \( i \in V \),

- \( \pi_i \): parents of \( i \)
  - ex. \( \pi_R = C \) (the parent of \( R = C \))

- \( f_i(x_i, x_{\pi_i}) \): joint p.d.f. of \( i \) and \( \pi_i \)
  - nonnegative
  - \( \sum_{x_i} f_i(x_i, x_{\pi_i}) = 1 \)
Claim: there is a family of probability functions \( P(X_V) = \prod_{i=1}^{n} f_i(x_i, x_{\pi_i}). \)

Note that this function is nonnegative, and

\[
\sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} P(X_V) = 1
\]  \quad (1)

Proof of (1):

\[
P(X_V) = P(C, R, S, W)
\]

\[
= f(C) f(R, C) f(S, C) f(W, S, R)
\]

We want to show that

\[
\sum_{C} \sum_{R} \sum_{S} \sum_{W} P(C, R, S, W) = \sum_{C} \sum_{R} \sum_{S} \sum_{W} f(C) f(R, C) f(S, C) f(W, S, R)
\]

\[
= 1.
\]

Consider factors \( f(C), f(R, C), f(S, C) \): they do not depend on \( W \), so we can write this all as

\[
\sum_{C} \sum_{R} \sum_{S} f(C) f(R, C) f(S, C) \sum_{W} f(W, S, R)^{-1}
\]

\[
= \sum_{C} \sum_{R} \sum_{S} f(C) f(R, C) \sum_{S} f(S, C)^{1}
\]

\[
= \sum_{C} \sum_{R} f(C) \sum_{R} f(R, C)^{1}
\]

\[
= 1
\]

since we had set \( \sum_{x_i} f_i(x_i, x_{\pi_i}) = 1. \)
1.2.3 Example

Consider the simple directed graph:

\[ X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \]

\[ p(x_3|x_2) = ? \]

\[ p(x_1, x_2, x_3, x_4) = f(x_1)f(x_2, x_1)f(x_3, x_2)f(x_4, x_3) \]

Bayes’ Rule: \[ p(x_3|x_2) = \frac{p(x_2, x_3)}{p(x_2)} \]

\[ p(x_2, x_3) = \sum_{x_1} \sum_{x_4} p(x_1, x_2, x_3, x_4) \quad \text{(marginalization)} \]

\[ = \sum_{x_1} \sum_{x_4} f(x_1)f(x_2, x_1)f(x_3, x_2)f(x_4, x_3) \]

\[ = \sum_{x_1} f(x_1)f(x_2, x_1)f(x_3, x_2) \sum_{x_4} f(x_4, x_3)^{-1} \]

\[ = f(x_3, x_2) \sum_{x_1} f(x_1)f(x_2, x_1). \]

We also need

\[ p(x_2) = \sum_{x_1} \sum_{x_3} \sum_{x_4} f(x_1)f(x_2, x_1)f(x_3, x_2)f(x_4, x_3) \]

\[ = \sum_{x_1} \sum_{x_3} f(x_1)f(x_2, x_1)f(x_3, x_2) \]

\[ = \sum_{x_1} f(x_1)f(x_2, x_1). \]
Thus,

\[
p(x_3|x_2) = \frac{\sum_{x_1} f(x_1) f(x_2, x_1)}{\sum_{x_1} f(x_1) f(x_2, x_1)} = f(x_3, x_2).
\]

1.2.4 Theorem

\[f_i(x_i, x_{\pi_i}) = p(x_i|x_{\pi_i}).\]

\[\therefore P(X_V) = \prod_{i=1}^{n} p(x_i|x_{\pi_i}).\]

In our simple graph, the joint probability is

\[p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3).\]

Instead, using the chain rule:

\[p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1)p(x_4|x_3, x_2, x_1).\]

The number of parameters required here is far more than what we were able to derive.

We’re using the Markov Property: the history of \(x_4\) is completely determined by \(x_3\). Simply applying the Markov Property to the chain-rule formula would have obtained our result.