1 Review for the Bayes-Ball Algorithm

Recall the primary rules for Bayes-Ball algorithm:

- In the case of a chain, if we know (“given”) a middle node, then the ball is blocked. Otherwise, the ball passes through.
- Consider the case of a tree with one parent node and two children. If the parent is known, then the ball is blocked. Otherwise, the ball passes through.
- Consider two parent nodes pointing to one child node. If the child node is given, the ball passes through. Otherwise, it is blocked.

If the ball passes through, then the nodes are conditionally dependent. Otherwise, they are conditionally independent.

2 Undirected Graphical Models

The main focus of the class was on undirected graphical models. These are also known as Markov Random Fields (MRFs). Undirected graphical models have a variety of applications. These include image processing, natural language processing, and networking. For instance, in image processing, Figure 1 could represent the image, with the nodes representing the pixels (or potentially blocks). A potential use of this would be to distinguish the foreground from the background, such as the foreground denoted by the circled nodes in Figure 1.

It is important to distinguish that undirected graphical models encode correlation, not causation. Directed graphical models, on the other hand, are better suited for encoding causation.

2.1 Conditional Independence in Undirected Graphical Models

The fundamental concept covered in this lecture is conditional independence in undirected graphical models. Consider an undirected graph denoted as the pair $G = (V, E)$, where $V$ is the set of nodes
and $E$ is the set of undirected edges. Each node $i$ is associated with a random variable $X_i$. Let $A$, $B$, and $C$ be three disjoint sets of nodes and let $X_A$, $X_B$, and $X_C$ be the corresponding sets of random variables. Then, $X_A$ is conditionally independent of $X_C$ given $X_B$ if the nodes in $A$ are separated from the nodes in $C$ by the nodes in $B$. In other words, in order to be conditionally independent, then all paths between nodes in $A$ and nodes in $C$ must go through one or more nodes in $B$. This also means that if we remove the nodes in $X_B$ and the edges connected to them, then the graph is split into two disconnected sets of nodes and the graph is no longer a connected graph. For instance, consider Figure 2 where $X_A$ is conditionally independent from $X_C$ given $X_B$.

2.2 Undirected and Directed Graphs Equivalency

Now that we have established the idea of conditional independence in undirected graphs and directed graphs (previous lectures), a natural question is whether we can reduce a given directed graph to an equivalent undirected graph (or vice versa). For instance, consider the directed graph shown in
Figure 3: Example of a directed graph and an attempt for an equivalent undirected graph.

This graph encodes these conditional independence assertions: \( x \perp z, x \not\perp z \mid y \). Figure 3b shows an undirected graph that might be able to encode the same set of independence assertions. However, this fails to encode \( x \perp z \). It turns out that no undirected graph can exactly represent the set of conditional independence assertions given by the directed graph in Figure 3a.

In general, neither undirected or directed graphical models are reducible to the other. However, there is a set of graphs that can be represented by both. These are known as chordal graphs.

An interesting difference between a directed an undirected graphs is that the only way to have marginal independence in an undirected graph is for the graph not to be connected. With a connected, undirected graph, no node is marginally independent of the others.

3 Undirected Graphical Model Families

An undirected graph provides a set of conditional independence assertions using graph separation. The undirected graphical model family is the set of distributions that satisfy all of the assertions given in the graph. We would like to be able to represent the joint distributions of random variables that satisfy all of the conditional independence assertions implied by an undirected graphical model.

3.1 Local Parametrization

To compactly represent a joint distribution, we want to obtain a local parametrization of a given undirected graphical model. We had such a parametrization in the case of directed graphs. In directed graphs, the local conditional probabilities of nodes given their parents provided the local parametrization. However, this interpretation of “local”, meaning a node and its parents, is not available in undirected graphical models, in which a node does not have parents.

One obvious choice for the meaning of “local” in an undirected graphical model is a node and its neighbors where we would give the conditional probability of a node given its neighbors. This type of local parametrization, though, does not ensure consistency of conditional probabilities across
nodes, which means that the product of these conditional probabilities will not necessarily be a valid 
joint distribution. Defining conditional probabilities for each node dependent on its neighbors is 
thus not a good way to locally parametrize the distribution represented by an undirected graphical 
model.

To find a more convenient way to locally parametrize, we first establish a different meaning of 
“local”. To do so, we must first understand “cliques” in graphs. A clique is a set of nodes wherein 
each node has an edge connecting it to each other node. For example, in Figure 4, the nodes A, B, 
and C form a clique, but nodes B, C, and D do not because there is no edge between B and D. A 
maximal clique is a clique to which no node can be added to make a larger clique. In Figure 4, nodes 
C and D are a maximal clique, as are nodes A, B, and C, whereas nodes A and B by themselves are 
not a maximal clique because node C can be added and the nodes still form a clique. From here on, 
“local” will refer to a maximal clique. It should be noted that while maximum cliques will be 
convenient to use as we develop our theory, the problem of finding maximum cliques is known to be 
nontrivial.

3.2 Potential Functions

Now that we have decided on a meaning of local, we will determine how to find functions that are 
local but which can be chosen independently and still lead to a valid joint distribution. Note that if 
random variables $X_A$ and $X_C$ are conditionally independent given variable $X_B$, then 

$$P(X_A|X_B,X_C) = P(X_A|X_C),$$ 

and 

$$P(X_A,X_B,X_C) = P(X_A|X_C)P(X_B,X_C).$$
We see that the joint distribution $P(X_A, X_B, X_C)$ can be factored such that

$$P(X_A, X_B, X_C) = f(X_A, X_B)g(X_B, X_C)$$

for some functions $f$ and $g$. Using this idea, we will factor the joint distribution as a product of functions, and each function in the factorization will depend only on variables within a maximal clique. So, if there is no edge between variables $X_i$ and $X_j$, for example, then the joint distribution cannot have a factor such as $\psi(X_i, X_j, X_k)$ because the three variables do not form a clique.

We will refer to the functions in the factorization as “potential functions”. If $C$ is a set containing the indices of a maximal clique in an undirected graph $G$, then the potential function $\psi_C(x_C)$ is a function that depends only on values $x_c$ of maximal clique variables $X_C$. It turns out that we need only loose restrictions on our potential functions, namely that the functions are non-negative and real-valued. Aside from these restrictions, our potential functions can be arbitrary. Let us denote the set of all maximal cliques $M$. We then let the probability of some outcome $x$, $p(x)$, be

$$p(x) \triangleq \frac{1}{Z} \prod_{C \in M} \psi_{X_C}(x_C),$$

where $Z$ is the normalization constant and is given by

$$Z \triangleq \sum_{x} \prod_{C \in M} \psi_{X_C}(x_C).$$

These potential functions, being arbitrary, do not have a probabilistic interpretation. However, they do have a quasi-probabilistic interpretation, in that a configuration of random variables with a large potential-function value will be favored over a configuration with a smaller value. Because of this relationship, one can think of potential functions as “pre-probabilistic”.

### 3.3 Exponentiated Potential Functions

To ensure that the potential function is non-negative, it is often convenient to make it the exponential of an arbitrary real-valued function. That is, have a potential function $\psi$ as

$$\psi_{X_C}(x_C) = \exp\{-H_C(x_C)\},$$

where we either define $H_C$ and exponentiate or directly define $\psi$. In this case, the actual probability is

$$p(x) = \frac{1}{Z} \prod_{C \in M} \exp\{-H_C(x_C)\} \triangleq \frac{1}{Z} \exp\left\{-\sum_{C \in M} H_C(x_C)\right\}.$$
Thus, we see that using the function $H_C$ and the exponentiation instead of $\psi$ also allows us to transform the product into a more convenient summation. The two representations, defining $\psi$ directly or defining $H_C$, are equivalent, so using the $H$ functions is done for convenience and does not gain any ability to represent different potential functions. As an aside, the representation using $H_C$ instead of $\psi$ comes up often in physics, where it is known as the Boltzmann distribution.

### 3.4 Hammersley-Clifford Theorem

The above use of potential functions with maximal cliques is justified by the Hammersley-Clifford theorem. This theorem states that our convenient representation, namely, expressing distributions as the product of potential functions over maximal cliques, is equivalent to representing distributions that satisfy conditional independence statements specified by graph separation. In other words, the factorized representation we use allows us to represent all distributions specified by using graph separation on our undirected graphs.

### 4 Reduced Parametrizations

It is often useful to add additional structure that is not specified in a graphical model in order to further simplify complicated joint distributions. Instead of imposing constraints only based on conditional independence, we can impose an algebraic structure on our conditional probabilities and potential functions. For example, look at the graphical model in Figure 5. Since node $Y$ has 50 parents, the conditional probability table at $Y$ would have $2^{50}$ entries. Since this is unacceptably large, we instead make an additional assumption about $P(Y|X)$ that is not specified in the graph. For example, we could specify that it is a function of a linear combination of the variables $X_i$, i.e.

$$p(Y = 1|x) = f(\theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_m x_m).$$

This reduces our formula down to just the 50 parameters $\theta_i$. This loses some representational power, but is now feasible to use.

![Figure 5: An example demonstrating why we might need reduced factorization](image-url)