Belief propagation is an algorithm for estimating conditional marginals in multivariate probability distributions specified by graphical models.

Estimating conditional marginals is closely related to sampling. An easy way to shuffle a deck of cards on a computer is to randomly pick the first card, then pick the second given the first, and so on. The probability of any given outcome, e.g. `2143` is a product of conditional probabilities

$$\Pr(\mathbf{X} = \texttt{2143}) = \Pr(X_1 = \texttt{2}) \Pr(X_2 = \texttt{1} | X_1 = \texttt{2}) \Pr(X_3 = \texttt{4} | X_1 X_2 = \texttt{21}) \Pr(X_4 = \texttt{3} | X_1 X_2 X_3 = \texttt{214}). \quad (1)$$

In the case of a card shuffle, we know exactly what the marginal distribution of the i-th card X_i is given the outcomes X_1 up to X_{i-1} of the previous cards is: It is uniform over the remaining values. For example, given that $X_1 X_2 = \texttt{21}$, X_3 is uniform over the remaining values `3` and `4`. This is the conditional marginal distribution of X_3 given X_1 and X_2. We can sample from it by choosing one at random. In general, sampling $\mathbf{X} = X_1 \cdots X_n$ efficiently reduces to calculating the probability mass function of random variable X_i given $X_1, \ldots, X_{i-1}$ for all possible values of X_1 up to X_{i-1} and every i.

The problem of conditional marginal estimation is to compute the conditional probability mass function of some random variable X given "observations" $Y_1, \ldots, Y_m$ for random variables $X, Y_1, \ldots, Y_m$ with a known joint probability distribution. A related problem is conditional sampling: Given outcomes Y_1 up to Y_m, sample X conditioned on them.

In many examples of interest, the random variable X takes one of a small number of values like a single bit (0 or 1). Then the complexities of conditional marginal estimation and conditional sampling are similar. If we can estimate the marginal probability mass function (p.m.f.) sufficiently accurately then we can sample from it via cumulative distribution sampling as described in the last lecture. Specifically, if X is a single bit, the marginal p.m.f. is specified by the single number $p = \Pr(X = 1 | Y_1, \ldots, Y_m)$. To sample from it we can pick a uniformly random number U in $[0, 1]$, output zero is $U \leq 1 - p$, and output 1 if $U > p$.

Conversely, if we can sample from this distribution then we can estimate the probability that X takes value x as the fraction of outcomes in which this is the case. The law of large numbers guarantees that the estimate is likely to be accurate given sufficiently many samples.

In general, the conditional distribution of X given Y_1 up to Y_m could be a complex function. Even if the values of $X, Y_1, \ldots, Y_m$ are single bits, it takes as many as 2^m numbers to specify all the conditional probabilities $\Pr(X = 1 | Y_1, \ldots, Y_m)$. It is therefore sensible to focus on special distributions that can be described more succinctly. One important class are directed graphical models. These are used to describe large-scale problems of statistical inference.

1 Statistical inference and posterior marginals

In statistical inference we start with some probabilistic model X of the world called the prior. For example, if X indicates the presence of COVID within a community, our prior could be $\Pr(X = 1) = 0.1$ and $\Pr(X = 0) = 0.9$, indicating a 10% chance of infection. We then carry out observation(s) Y described by the conditional probabilities $\Pr(Y = y | X)$. In our example, Y might be the probability that you test positive. This probability is different depending on your COVID status. For an ideal test, $\Pr(Y = 1 | X = 0)$ would be zero and $\Pr(Y = 1 | X = 1)$ would be one. However, owing to various errors a more realistic model might be $\Pr(Y = 1 | X = 0) = 0.05$ and $\Pr(Y = 1 | X = 1) = 0.9$.

Now suppose that you observe a given outcome, say $Y = 1$. You tested positive. What are the chances that you have COVID? Given that you observed an effect, what is the *posterior* probability of the cause?

The answer is provided by Bayes' formula

$$\Pr(X = x|Y = y) = \frac{\Pr(Y = y|X = x)\Pr(X = x)}{\Pr(Y = y)}. \tag{2}$$

In our model the probability that you have COVID given that you tested positive is then

$$\Pr(X = 1|Y = 1) = \frac{\Pr(Y = 1|X = 1)\Pr(X = 1)}{\Pr(Y = 1)} = \frac{0.9 \cdot 0.1}{\Pr(Y = 1)}.$$

To complete the calculation we would need to find the total probability $\Pr(Y = y)$ that a random person tests positive. It is not specified explicitly in the model. We need to compute it. In principle we can derive it using the total probability formula

$$\Pr(Y = 1) = \Pr(Y = 1|X = 0)\Pr(X = 0) + \Pr(Y = 1|X = 1)\Pr(X = 1) = 0.05 \cdot 0.9 + 0.9 \cdot 0.1 = 0.135$$

and $\Pr(X = 1|Y = 1) = 0.09/0.135 \approx 0.666$. There is a two thirds chance that you are infected in this scenario.

More generally, $\mathbf{X}$ can be a collection $X_1, \ldots, X_n$ of many random variables. For example, suppose you want to sample a 32×32 bitmap image with a cow in it. Then $X_1, \ldots, X_{32^2}$ would represent the image bits. The prior is that these are independent random bits. Any image is as likely as any other. Y would indicate the presence of a cow in the image. The conditional probabilities $\Pr(Y|X_1, \ldots, X_n)$ are too many to write down, but you may imagine having a neural network that calculates the probability a given image has a cow in it.

For simplicity will assume that the prior is uniform, namely all outcomes for $X_1, \ldots, X_n$ are equally likely. In particular this means $X_1, \ldots, X_n$ are independent random variables. These are easy to sample and their p.m.f. is easy to calculate. We will also assume that the probabilities $\Pr(Y|\mathbf{X})$ are easy to calculate. How about the posterior marginals $Pr(\mathbf{X} = \mathbf{x}|Y)$? In principle, these can again be obtained using Bayes' rule and the total probability formula. The latter, however, is now a sum over all 2^n possible assignments to $\mathbf{X}$:

$$\Pr(Y) = \sum_{x_1,\ldots,x_n} \Pr(Y|X_1 = x_1, \ldots, X_n = x_n)\Pr(X_1 = x_1, \ldots, X_n = x_n).$$

A naive evaluation takes time 2^n and is infeasible in practice. Is there an alternative?

For some inference problems, difficult calculations can be avoided. For instance, suppose we want to know which of the two outcomes $\mathbf{x}, \mathbf{x}'$ is more likely given Y. This is determined by the likelihood ratio $\Pr(\mathbf{X} = \mathbf{x}|Y = y)/\Pr(\mathbf{X} = \mathbf{x}'|Y = y)$. In Bayes' formula the total probabilities cancel out to give

$$\frac{\Pr(\mathbf{X} = \mathbf{x}|Y = y)}{\Pr(\mathbf{X} = \mathbf{x}'|Y = y)} = \frac{\Pr(Y = y|\mathbf{X} = \mathbf{x})}{\Pr(Y = y|\mathbf{X} = \mathbf{x}')} \cdot \frac{\Pr(\mathbf{X} = \mathbf{x})}{\Pr(\mathbf{X} = \mathbf{x}')}. \tag{3}$$

As all prior outcomes are equally likely, $\Pr(\mathbf{X} = \mathbf{x})$ and $\Pr(\mathbf{X} = \mathbf{x}')$ are the same and

$$\frac{\Pr(\mathbf{X} = \mathbf{x}|Y = y)}{\Pr(\mathbf{X} = \mathbf{x}'|Y = y)} = \frac{\Pr(Y = y|\mathbf{X} = \mathbf{x})}{\Pr(Y = y|\mathbf{X} = \mathbf{x}')}. \tag{4}$$

Thus the image that is most likely to have a cow in it is the image X for which the "cow predictor" $\Pr(Y|\mathbf{X})$ is the most confident.

To get started on sampling $\mathbf{X}$ given Y bit-by-bit using the chain rule (1), however, we need to know the posterior marginal probabilities $\Pr(X_1 = x_1|Y)$. Their *likelihood ratio* is also given by

$$LR[X_1|Y = y] = \frac{\Pr(X_1 = 1|Y = y)}{\Pr(X_1 = 0|Y = y)} = \frac{\Pr(Y = y|X_1 = 1)}{\Pr(Y = y|X_1 = 0)}.$$

It is again unclear how to calculate or estimate the terms on the right without resorting to the total probability formula, which entails summing over the 2^{n-1} possible values of $X_2 \cdots X_n$.

In this level of generality, there are instances of statistical inference on which we would not expect to solve efficiently. For example, if $\mathbf{X}$ is the binary representation of a quadratic residue modulo a safe prime and Y is the event that $g^{\mathbf{X}}$ equals h for some prespecified group elements g and h, then sampling $\mathbf{X}$ given Y amounts to finding the discrete logarithm of h in base g. We do not expect any efficient algorithm to succeed.

It is therefore sensible to further restrict the class of models on which we aim to carry out statistical inference efficiently.

2 Directed graphical models

In a directed graphical model, Y is a conjunction (AND) of multiple observations $\mathbf{Y} = (Y_1, \ldots, Y_m)$. Each Y_j in turn only depends on a limited number of base variables $\mathbf{X} = (X_1, \ldots, X_n)$. The dependence of Y_j on X_i is indicated by an edge in a bipartite graph with vertices $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ called the *causal graph*.

We can describe shuffled cards, i.e, random permutations, by a directed graphical model. Let's do the 4-card deck as an example. The base variables X_1, X_2, X_3, X_4 take values `1`, `2`, `3`, `4`. The prior distribution is uniform. Thus outcome `2322` is as likely as `2341`. To enforce the permutation constraint, we "observe" whether X_i and X_j are different for every pair $i \neq j$. In random variable notation, Y_{ij} takes value 1 (with probability 1) if X_i and X_j are different and value 0 if they are equal (see Figure 1a). The conditional distribution over $X_1X_2X_3X_4$ given $\mathbf{Y} = Y_{12}Y_{13}Y_{14}Y_{23}Y_{24}Y_{34}$ is then uniform over all permutations of `1234`. This is the distribution we want to sample from.

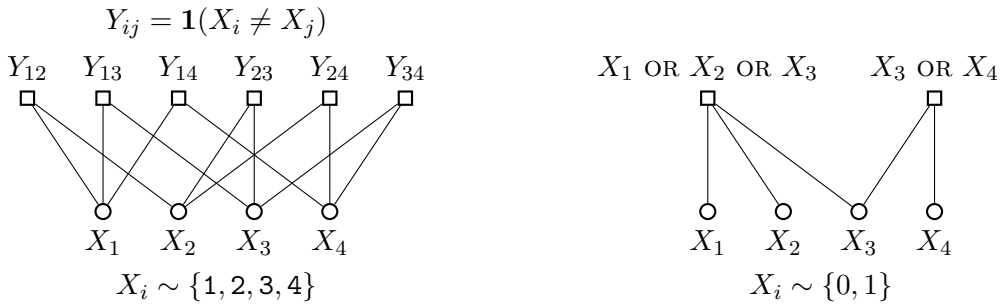

FIGURE 1: Two directed graphical models. Conditioned on $\mathbf{Y}$, X is (a) a random permutation of `1234`; (b) a random assignment consistent with both constraints.

Another class of examples are Boolean constraint satisfaction problems. Here both the base variables and the observations take $\{0, 1\}$ values. For example, Y_j can be the logical OR of its neighbors as in Figure 1b. For a generic instance of this type, finding any assignment $\mathbf{X}$ that is consistent with the constraints is believed to be computationally intractable. This is the famous P does not equal NP hypothesis. Sampling a random assignment and computing conditional marginals are even harder tasks. Belief propagation is an algorithm that attempts to do just that.

One class of graphical models that allow efficient computation of conditional marginals are trees. The model in Figure 1b is a tree.

3 Composition of graphical models

Conditional marginals are tricky to reason about. Adding an observation in a graphical model can dramatically affect base variables that do not participate in it. For example, conditioned on Y_1, X_1 in Figure 1b

is positive with probability 4/7. Once Y_2 is added in, this probability drops to 6/11 even though Y_2 does not directly depend on X_1.

To understand how changing the models affects conditional marginals we'll look at three operations that build more complicated models out of simpler ones.

The most basic operation is unconstrained composition. Given two models $(\mathbf{X}, \mathbf{Y})$, and $(\mathbf{X}', \mathbf{Y}')$ over disjoint sets of base variables $\mathbf{X}$ and $\mathbf{X}'$, their composition is the model whose base variables are the union of $\mathbf{X}$ and $\mathbf{X}'$ and whose observations are the union of $\mathbf{Y}$ and $\mathbf{Y}'$. In this model, $(\mathbf{X}, \mathbf{Y})$ is conditionally independent of $(\mathbf{X}', \mathbf{Y}')$. In particular, the conditional marginals of $\mathbf{X}$ and $\mathbf{X}'$ remain the same as in the original models:

$$\Pr(X_i = x_i|\mathbf{Y}, \mathbf{Y}') = \Pr(X_i = x_i|\mathbf{Y}) \qquad \text{and} \qquad \Pr(X'_i = x'_i|\mathbf{Y}, \mathbf{Y}') = \Pr(X'_i = x'_i|\mathbf{Y}').$$

Another operation is contracted composition. Suppose $(\mathbf{X}, \mathbf{Y})$ and $(\mathbf{X}', \mathbf{Y}')$ share a single base variable X that appears both in $\mathbf{X}$ and $\mathbf{X}'$. Apart from it, $\mathbf{X} \setminus \{X\}$ and $\mathbf{X}' \setminus \{X\}$ are disjoint (and so are $\mathbf{Y}$ and $\mathbf{Y}'$). The composed model has $\mathbf{X} \cup \mathbf{X}'$ as the base variables and $\mathbf{Y} \cup \mathbf{Y}'$ as the observations. For example, the model in Figure 1b is the contracted composition of $(X_1X_2X_3, Y_1)$ and (X_3X_4, Y_2).

Contracted composition could have a complicated effect on the conditional marginals of base variables other than X_i. As for X_i itself, the new conditional marginals are

$$\begin{aligned}
\frac{\Pr(X = x|\mathbf{Y}, \mathbf{Y}')}{\Pr(X = x'|\mathbf{Y}, \mathbf{Y}')} &= \frac{\Pr(\mathbf{Y}, \mathbf{Y}'|X = x)}{\Pr(\mathbf{Y}, \mathbf{Y}'|X = x')} && \text{by (4)} \\
&= \frac{\Pr(\mathbf{Y}|X = x) \cdot \Pr(\mathbf{Y}'|X = x')}{\Pr(\mathbf{Y}'|X = x) \cdot \Pr(\mathbf{Y}'|X = x')} \\
&= \frac{\Pr(\mathbf{Y}|X = x)}{\Pr(\mathbf{Y}|X = x')} \cdot \frac{\Pr(\mathbf{Y}'|X = x)}{\Pr(\mathbf{Y}'|X = x')} \\
&= \frac{\Pr(X = x|\mathbf{Y})}{\Pr(X = x'|\mathbf{Y})} \cdot \frac{\Pr(X = x|\mathbf{Y})}{\Pr(X = x'|\mathbf{Y}')} && \text{by (4).}
\end{aligned}$$

The key equality is the second one. It holds because when the value of X_i is fixed to the "constant" x_i, the remaining graphical model is a parallel composition of $(\mathbf{X} \setminus \{X\}, \mathbf{Y})$ and $(\mathbf{X}' \setminus \{X\}, \mathbf{Y}')$ and the probabilities factor.

The probability mass function of a Boolean-valued random variable X is completely specified by the likelihood ratio $LR[X] = \Pr(X = 1)/\Pr(X = 0)$. Variable merging multiplies the conditional marginals:

$$LR[X|\mathbf{Y}, \mathbf{Y}'] = LR[X|\mathbf{Y}] \cdot LR[X|\mathbf{Y}'].$$

Let's apply this formula to the example in Figure 1b. There are seven assignments to $X_1X_2X_3$ that satisfy $Y_1 = X_1$ OR X_2 OR X_3. Out of those four satisfy X_3. So $X_3|Y_1$ equals one with probability 4/7. (To minimize notation here we write $X_3|Y_1$ as shorthand for $X_3|Y_1 = 1$.) Alternatively, $LR[X_3|Y_1] = (4/7)/(3/7) = 4/3$. Similarly, there are three assignments to X_3X_4 that satisfy $Y_2 = X_3$ OR X_4. Out of them two satisfy X_3, so $LR[X_3|Y_2] = (2/3)/(1/3) = 2$. Therefore $LR[X_3|Y_1, Y_2] = 2 \cdot 4/3 = 8/3$. Boolean likelihood ratios can be converted back to probabilities using

$$\Pr(Z = 0) = \frac{1}{1 + LR[Z]} \qquad \Pr(Z = 1) = \frac{LR[Z]}{1 + LR[Z]},$$

so $\Pr(X_3 = 1|Y_1, Y_2) = (8/3)/(1 + 8/3) = 8/11$.

Constrained composition makes sense for multiple models $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_d, \mathbf{Y}_d)$ with shared base variable X and all other base variables disjoint. The new conditional marginal of X is

$$LR[X|\mathbf{Y}_1, \ldots, \mathbf{Y}_d] = LR[X|\mathbf{Y}_1] \cdots LR[X|\mathbf{Y}_d]. \tag{5}$$

The last operation of interest is constrained composition. Given two models $(\mathbf{X}, \mathbf{Y})$ and $(\mathbf{X}', \mathbf{Y}')$ over disjoint sets of base variables, the constrained composition is obtained by applying unconstrained

composition and introducing an additional observation Y that depends on $X \in \mathbf{X}$ and $X' \in \mathbf{X}'$. For example, the model in Figure 1b is the constrained composition of $(X_1X_2X_3, Y_1)$ and (X_4, ϵ) (where ϵ stands for "no observations") with additional observation Y_4.

The joint marginal of X and X' after constrained composition can be derived using (3). To simplify notation we look at the numerator only and write $\propto$ to mean "up to a normalization factor that does not affect the ratio":

$$\begin{aligned}\Pr(X, X'|\mathbf{Y}, \mathbf{Y}', Y) &\propto \Pr(Y|X, \mathbf{Y}, X', \mathbf{Y}') \cdot \Pr(X, X'|\mathbf{Y}, \mathbf{Y}') \\ &= \Pr(Y|X, X') \cdot \Pr(X|\mathbf{Y}) \cdot \Pr(X'|\mathbf{Y}').\end{aligned}$$

The reason for the first simplification is that X and X' completely determine Y so the conditioning on $\mathbf{Y}$ and $\mathbf{Y}'$ can be dropped. (More accurately, $\mathbf{Y}, \mathbf{Y}'$ is conditionally independent of Y given X, X'.) The reason for the second one is that $(X \cup X', \mathbf{Y} \cup \mathbf{Y}')$ is the unconstrained composition of $(X, \mathbf{Y})$ and $(X', \mathbf{Y}')$.

Let's apply this formula to the example in Figure 1b. We already saw that $X_3|Y_1$ equals one with probability 4/7. As X_4 is initially unconstrained, $\Pr(X_4 = 1|\epsilon) = 1/2$. Applying constrained composition gives

$$\Pr(X_3, X_4|Y_1, \epsilon, Y_2) \propto \Pr(Y_2|X_3, X_4) \cdot \Pr(X_3|Y_1) \cdot \Pr(X_4|\epsilon).$$

As $Y_2 = X_3$ OR X_4, we get (dropping ϵ from the notation)

$$\begin{aligned}
\Pr(X_3 = 1, X_4 = 1|Y_1, Y_2) &\propto \Pr(Y_2|X_3 = 1, X_4 = 1) \cdot \Pr(X_3 = 1|Y_1) \cdot \Pr(X_4 = 1) = 1 \cdot \tfrac{4}{7} \cdot \tfrac{1}{2} \\
\Pr(X_3 = 1, X_4 = 0|Y_1, Y_2) &\propto \Pr(Y_2|X_3 = 1, X_4 = 0) \cdot \Pr(X_3 = 1|Y_1) \cdot \Pr(X_4 = 0) = 1 \cdot \tfrac{4}{7} \cdot \tfrac{1}{2} \\
\Pr(X_3 = 0, X_4 = 1|Y_1, Y_2) &\propto \Pr(Y_2|X_3 = 0, X_4 = 1) \cdot \Pr(X_3 = 0|Y_1) \cdot \Pr(X_4 = 1) = 1 \cdot \tfrac{3}{7} \cdot \tfrac{1}{2} \\
\Pr(X_3 = 0, X_4 = 0|Y_1, Y_2) &\propto \Pr(Y_2|X_3 = 0, X_4 = 0) \cdot \Pr(X_3 = 0|Y_1) \cdot \Pr(X_4 = 0) = 0 \cdot \tfrac{3}{7} \cdot \tfrac{1}{2}
\end{aligned}$$

As these probabilities should add up to one, they must equal 4/11, 4/11, 3/11, and 0, respectively. We can now derive the conditional marginals by summing up the joint probabilities:

$$\Pr(X_3 = 1|Y_1, Y_2) = \tfrac{4}{11} + \tfrac{4}{11} = \tfrac{8}{11} \qquad \text{and} \qquad \Pr(X_4 = 1|Y_1, Y_2) = \tfrac{4}{11} + \tfrac{3}{11} = \tfrac{7}{11}.$$

The conditional marginals for X_3 we derived in contracted and constrained composition are the same.

There were quite a few cancellations in this calculation. This is no accident. We can directly relate the likelihood ratios of the constituent and composed models. Sparing the details, when X is Boolean-valued, its likelihood ratio after composition is

$$LR[X|\mathbf{Y}, \mathbf{Y}', Y] = \frac{\sum_{x'} \Pr(Y|X = 1, X' = x') \cdot LR[X|\mathbf{Y}] \cdot LR[X'|\mathbf{Y}']^{x'}}{\sum_{x'} \Pr(Y|X = 0, X' = x') \cdot 1 \cdot LR[X'|\mathbf{Y}']^{x'}}.$$

In the example we just did,

$$\begin{aligned}LR[X_3|Y_1, Y_2] &= \frac{\Pr(Y_2 = 1|X_3X_4 = 10) \cdot LR[X_3|Y_1] \cdot 1 + \Pr(Y_2 = 1|X_3X_4 = 11) \cdot LR[X_3|Y_1] \cdot LR[X_4]}{\Pr(Y_2 = 1|X_3X_4 = 00) \cdot 1 \cdot 1 + \Pr(Y_2 = 1|X_3X_4 = 01) \cdot 1 \cdot LR[X_4]} \\ &= \frac{4/3 \cdot 1 + 4/3 \cdot 1}{1 \cdot 1} = \frac{8}{3}\end{aligned}$$

as expected.

Constraint composition can be applied to more than two models. The models $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_d, \mathbf{Y}_d)$ should be disjoint, and Y is a new observation that depends on $X_1 \in \mathbf{X}_1, \ldots, X_d \in \mathbf{X}_d$ only. For Boolean-valued variables, the new conditional likelihood ratios are

$$\begin{aligned}LR[X_i|\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_d, Y] &= \frac{\sum_{\mathbf{x}_{-i}} F(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_d)}{\sum_{\mathbf{x}_{-i}} F(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_d)}, \\ \text{where} \quad F(\mathbf{x}) &= \Pr(Y|\mathbf{X} = \mathbf{x}) \cdot LR[X_1|\mathbf{Y}_1]^{x_1} \cdots LR[X_1|\mathbf{Y}_d]^{x_d},\end{aligned} \tag{6}$$

and $\sum_{\mathbf{x}_{-i}}$ is a sum over all assignments to $\mathbf{x}$ with its i-th bit removed.

4 Belief propagation on trees

When the directed graphical model G is a tree, formulas (5) and (6) can be applied to relate the conditional marginals of certain derived models obtained by removing constraints from G. Each edge (X, Y) in G specifies two such derived models:

- The y-cavity model $G_{Y \to X}$ obtained by removing Y, and
- The x-cavity model $G_{X \to Y}$ obtained by removing all observations that depend on X *except for* Y.

In both of these models variable X takes on the role of a "root variable". See Figure 2 for an example.

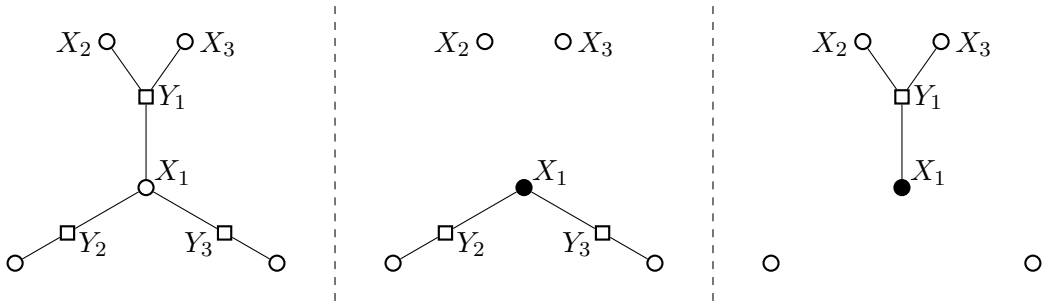

FIGURE 2: (a) A directed graphical model G. The cavity models (b) $G_{Y_1 \to X_1}$ and (c) $G_{X_1 \to Y_1}$.

In this example, the model $G_{Y_1 \to X_1}$ is the contracted composition of $G_{X_1 \to Y_2}$ and $G_{X_1 \to Y_3}$, plus some disconnected components (the isolated vertices X_2 and X_3). These components do not affect the marginals of the root variable X_1. We can therefore apply (5) to derive the equation

$$LR[X_1|G_{Y_1 \to X_1}] = LR[X_1|G_{X_1 \to Y_2}] \cdot LR[X_1|G_{X_1 \to Y_3}].$$

Here, X given G stands for the conditional distribution of X_1 given some fixing of all the observations in the model G. In general, for each edge $Y \to X$ in G, the marginal of X in $G_{Y \to X}$ is determined by the marginals of X in $G_{X \to Y'}$ among all neighbors Y' of X except Y itself:

$$LR[X|G_{Y \to X}] = \prod_{Y' \sim X \text{ except } Y} LR[X|G_{X \to Y'}]. \tag{7}$$

The model $G_{X_1 \to Y_1}$ can be viewed as the constrained composition of the isolated vertex X_1 and the models $G_{Y_1 \to X_2}$ and $G_{Y_1 \to X_3}$ (plus two isolated vertices). The marginal of X_1 in it can be calculated using (6):

$$LR[X_1|G_{X_1 \to Y_1}] = \frac{\sum_{x_2,x_3} \Pr(Y_1|X_1X_2X_3 = 1x_2x_3) LR[X_2|G_{Y_1 \to X_2}]^{x_2} LR[X_2|G_{Y_1 \to X_3}]^{x_3}}{\sum_{x_2,x_3} \Pr(Y_1|X_1X_2X_3 = 0x_2x_3) LR[X_2|G_{Y_1 \to X_2}]^{x_2} LR[X_2|G_{Y_1 \to X_3}]^{x_3}}.$$

(As X_1 was isolated before composition, its likelihood ratio is 1.) The general formula is

$$LR[X|G_{X \to Y}] = \frac{\sum_{\mathbf{x}'} \Pr(Y|X\mathbf{X}' = 1\mathbf{x}') \prod_{X' \sim Y \text{ except } X} LR[X|G_{Y \to X'}]^{x'}}{\sum_{\mathbf{x}'} \Pr(Y|X\mathbf{X}' = 0\mathbf{x}') \prod_{X' \sim Y \text{ except } X} LR[X|G_{Y \to X'}]^{x'}}. \tag{8}$$

Here $\mathbf{X}'$ stands for all neighbors of Y except X, X' is an entry in $\mathbf{X}'$, and $x', \mathbf{x}'$ are values that $X', \mathbf{X}'$ take.

Taken together, (5) and (6) is a system of equations in the likelihood ratios $LR[X|G_{X \to Y}], LR[X|G_{Y \to X}]$. There is one equation (5) for every edge $Y \to X$ and one equation (6) for every edge $X \to Y$. Thus there are as many equations as variables. The cavity marginals are a solution to this system.

How can we find this solution? The idea of Belief Propagation is to start with a guess $g(X \to Y)$ for $LR(X \to Y)$ and keep iterating.

Algorithm *BeliefPropagation*
Input: A directed graphical model G.

1 Choose guesses $g(X \to Y)$ for $LR[X|G_{X\to Y}]$ for every edge (X, Y) of G.

2 Until you are happy,

3 For every edge (X, Y), update $h(Y \to X)$ to $\prod_{Y' \sim_G X \text{ except } Y} g(X \to Y')$.

4 For every edge (Y, X), update $g(X \to Y)$ to

$$\frac{\sum_{\mathbf{x}'} \Pr_G(Y|X\mathbf{X}' = 1\mathbf{x}') \prod_{X' \sim_G Y \text{ except } X} h(Y \to X')^{x'}}{\sum_{\mathbf{x}'} \Pr_G(Y|X\mathbf{X}' = 0\mathbf{x}') \prod_{X' \sim_G Y \text{ except } X} h(Y \to X')^{x'}}. \tag{9}$$

5 Output $g(X \to Y)$ for all edges (X, Y).

The output is meant to estimate $LR[X|G_{X\to Y}]$. To obtain estimates of $LR[X|G]$ we can apply (5) one more time and replace the last line by

5' Output $\prod_{Y \sim_G X} g(X \to Y)$ for every base variable X.

A sensible initialization in line 1 is $g(X \to Y) = 1$ for all edges. That is, all assignments are equally likely. Here is how h and g evolve in the example on Figure 1b.

	h					g				
round	$Y_1 \to X_1$	$Y_1 \to X_2$	$Y_1 \to X_3$	$Y_2 \to X_3$	$Y_2 \to X_4$	$X_1 \to Y_1$	$X_2 \to Y_1$	$X_3 \to Y_1$	$X_3 \to Y_2$	$X_4 \to Y_2$
0						1	1	1	1	1
1	1	1	1	1	1	4/3	4/3	4/3	2	2
2	1	1	2	4/3	1	6/5	6/5	4/3	2	7/4
3	1	1	2	4/3	1	6/5	6/5	4/3	2	7/4

Belief Propagation converges after two rounds. Its estimates at this step are

$$LR[X_1|G] = g(X_1 \to Y_1) = \frac{6}{5} \qquad LR[X_2|G] = g(X_2 \to Y_1) = \frac{6}{5}$$
$$LR[X_3|G] = g(X_3 \to Y_1) \cdot g(X_3 \to Y_2) = \frac{8}{3} \qquad LR[X_4|G] = g(X_3 \to Y_2) = \frac{7}{4}.$$

This is correct. Out of the 16 possible assignments to $\mathbf{X}$, 11 satisfy both X_1 OR X_2 OR X_3 and X_3 OR X_4. Among those, X_1, X_2, X_3 and X_4 are positive 6, 6, 8, and 7 times, respectively. In fact, Belief Propagation is guaranteed to converge and output the correct answer provided the graphical model is a tree.

Theorem 1. *If G is a tree of depth d, Belief Propagation on G converges after d iterations and outputs the values $LR[X|G]$ for all base variables X in G.*

The variables $h(Y \to X)$ can be viewed as messages passed from constraint Y to variable X. They can be interpreted as Y's belief about the value of X (hence the name). For example $h(Y \to X) = 1$ indicates that Y thinks $X = 1$ and $X = 0$ should be equally likely. $h(Y \to X) = 9$ indicates the opinion that X should take value 1 with probability 9/10. Equation (9) is a rule for aggregating the opinions of all the constraints that X is involved in.

The variables $g(X \to Y)$ represent messages passed from variables to constraints. They are related to the influence that variable X has over enforcing constraint Y. I don't really understand what they mean.

5 Loopy belief propagation

Statistical inference on trees is not particularly interesting. Most models of interest have cycles. In our explanations about Belief Propagation we heavily relied on the assumption that the input model G is a

tree. In particular, the fixed point equations (5) and (6) might fail to hold if G has loops. What can we do then?

One great feature of Belief Propagation is that even though its *analysis* is very difficult in general graphs, the algorithm itself is perfectly valid! It can be tried on any directed graphical model, trees and graphs with cycles alike. If the model has loops however, undesirable outcomes are possible. The solution g that the algorithm converges to might not coincide with the conditional marginals. Or Belief Propagation might not converge at all.

The potential lack of convergence sounds serious. But it is also an opportunity. Suppose the algorithm does happen to converge. This may indicate that it found reasonable marginals. The marginals can be used to get the sampling going. We can sample one of the base variables according to its marginal, fix its value to derive a graphical model with one fewer variable, and iterate. At the end we obtain a candidate assignment $\mathbf{x}$ to all the base variables. While it is impossible to verify whether it was sampled correctly, we can at least check if it is consistent with all the observations $\mathbf{Y}$. Unlike Markov Chain Monte Carlo, Belief Propagation can be tested to some extent.

In many statistical inference problems of interest finding *any* assignment that is consistent with the observations, let alone a random one, is already a challenging problem. One famous such example is random 3SAT. There the "observations" are *random* ORs of three literals, namely constraints like X_3 OR (NOT X_5) OR (NOT X_6). In each constraint, the variables are chosen independently at random, and some of them are negated at random. The more such constraints are imposed, the fewer satisfying assignments survive. After sufficiently many constraints are sampled, they cease to exist altogether. Before this threshold is crossed how can we go about finding one?

No computer scientist knows the answer to this fascinating question. What we do know, however, is that among all the algorithms that have been tried, variants of Belief Propagation have had the most success. One explanation is that even though graphical model representing random 3SAT instances have cycles, very few of those cycles are *short*. In the presence of cycles, the cavity models $G_{X\to Y}$ and $G_{Y\to X}$ no longer split into disjoint components.

In the example in Figure 2, if vertex X_2 cycled back to X_1, those cycles would have survived the cavities at Y_1 and Y_2, Y_3. However, if the cycles are large, it would take many iterations for the messages out of X_2 in a run of Belief Propagation to make their way back to X_1. By the time they get there, they have to survive many multiplications, and their effect could be greatly diminished. On a good day we may hope that fixed-point equations (5) and (6) should still hold approximately, and that Belief Propagation converges to this approximate solution.

Another domain where directed graphical models are useful is error correction. In Lecture 7 we saw how "parity check" violations in Hamming's $[7, 4, 3]$ code informed error correction. Codeword corruptions can be described by a graphical model. The base variables $\mathbf{X}$ represent errors in transmission. In Shannon's theory of coding errors these are random and independent. The observed variables $\mathbf{Y}$ are the parity checks. The likelihood of a given error pattern is then precisely the conditional probability of $\mathbf{X}$ given $\mathbf{Y}$. When the error rate is reasonable, the probability of the true error pattern dwarfs all others. It is extremely close to one. The conditional marginals completely reveal the error pattern and enable decoding.

There is a special class of error-correcting codes called low-density parity-check codes in which the causal graph has few loops. Belief propagation is a natural error-correction algorithm in this setting. It can handle higher error rates than any other.

For me, understanding all of this has been very slow work in progress. I have struggled mightily with this Great Algorithm. Yet every time I see it, it makes a little more sense than before. As all Great Algorithms, it tells us a little story, and leaves the rest to the imagination.