# SCRIBE NOTES FOR OCTOBER 6, 2010 LECTURE

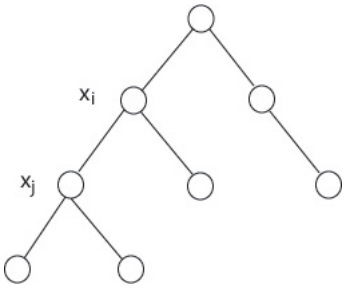
JOHN MYLES WHITE

## 1. ELIMINATE AND THE PROPAGATION ALGORITHM

During the first part of this lecture, we returned to the message-passing formulation of ELIMINATE. We noted that a message from a node $j$ to its neighbor $i$ has the form,

$$m_{ji} = \sum_{x_j} \psi(x_j)\psi(x_i, x_j) \prod_{k \in N(j) \setminus i} m_{kj}(x_j). \tag{1}$$

To illustrate ELIMINATE, we employed variants on the following graph:



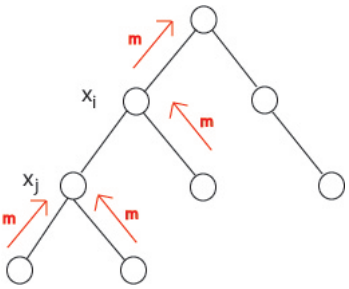

We then noted that the marginal probabilities at each node were defined using sums and products of potential functions as follows:

$$p(x) = \frac{1}{Z} \prod_i \psi(x_i) \prod_{(i,j) \in E} \psi(x_i, x_j). \tag{2}$$

In the preceding equation, the constant $Z$ was defined using the sum over all other marginal probabilities:

$$Z = \sum_{x'} \prod_i \psi(x_i^{'}) \prod_{(i,j) \in E} \psi(x_i^{'}, x_j^{'}). \tag{3}$$

In light of these equations, it was clear that ELIMINATE depends on using potential functions, which we can interpret as messages from a node that is being eliminated to its parents. To visualize this message-passing interpretation, we employed a variant of the graph we had drawn earlier:

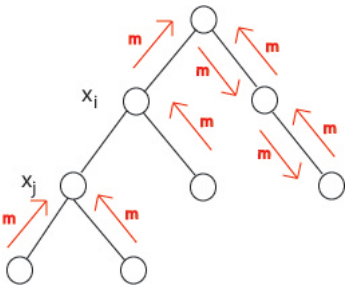



We then noted that ELIMINATE has an obvious flaw: when you consider different query nodes, you recompute some messages from scratch. By designing an algorithm that could exploit this redundancy, we arrived at the belief propagation algorithm. It is a message passing protocol, i.e. a rule for when we can pass messages.

Specifically, we have only one rule in our protocol: a node, $j$, can send a message, $m_{ji}$, to its neighbor, $i$, only after it has received messages from all of its other neighbors. In light of this, we rewrite our equation for marginal probabilities as follows:

$$p(x_f) \propto \psi(x_f) \prod_{e \in N(x_f)} m_{ef}(x_f). \tag{4}$$

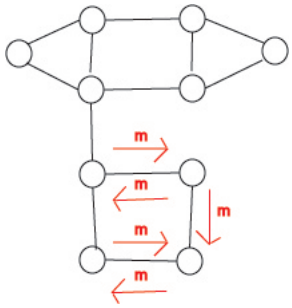
In contrast to ELIMINATE, when performing the propagation algorithm, we create messages in all directions at once. Thus, our graph is modified again to look like this:




Because the marginal equation shown earlier holds for any node in the graph, we can compute the marginal for any node after a single round of message-passing.

Having noted this, we asked, "what's being passed exactly when we pass these messages?" The messages we pass are marginal probabilities for each node, computed as if none of the rest of the graph existed.

Then we asked, "what happens if we attempt to perform the propagation algorithm on a graph with cycles?" For example, what happens if we run propagation with the following graph?




To answer this, we said that, being computer scientists, we can just try running propagation despite the obvious problems. The resulting algorithm is called loopy belief propagation. Sometimes it converges; sometimes it diverges. There is a research literature about when loopy belief propagation does or does not converge.

We then returned to a specific formulation of the algorithm using potential functions for the following graph:


Here we set the potential functions based on the conditional probabilities:

$$\psi(x_j, x_i) = p(x_i \mid x_j). \tag{5}$$

The marginal probabilities were then computed using the Law of Total Probability:

$$p(x_i) = \sum p(x_i)p(x_i \mid x_j). \tag{6}$$

## 2. Statistical Concepts

In statistics, we consider a general question: how can we use joint probability distributions to make meaningful statements about observed data?

One way is to employ graphical models. A graphical model (G.M.) is a family of distributions, in which the observed variables are shaded and the hidden variables are unshaded.

Setting parameters for a graphical model defines a particular member of the family, e.g. specific conditional probability tables (for directed models) or potential functions (for undirected models). The parameters are generally labeled as $\Theta$.

Inference occurs when we suppose that we know the model structure, but not the parameters. In this case, we observe data $X$. Because we have the joint distribution, we can compute $p(X \mid \Theta)$ for every $\Theta$. The general goal

of statistical inference is then to "invert" the natural relationship, i.e. to learn something about $\Theta$ given $X$.

## 3. Bayesian Statistics

A lot of statistics uses probability models. In Bayesian statistics, we only work with probability models. All statistical inference is formulated as a probabilistic computation.

Thus, the inversion that represents inference results from Bayes' Rule, where we consider the probability of $\Theta$ given $X$, i.e.

$$p(\Theta \mid X) = \frac{p(X \mid \Theta)p(\Theta)}{p(X)}. \tag{7}$$

When we treat $\Theta$ as a random variable, we need to posit a *prior*, $p(\Theta)$. To an "orthodox Bayesian", $p(\Theta)$ encodes your prior belief about $\Theta$ before seeing any data. Taste determines the palatability of using priors. For example, Freedman calls a prior an "opinion", while Jaynes calls a prior "common sense".

For any Bayesian, inference results in a *distribution* over $\Theta$ given the data $X$. This distribution is called the *posterior*. Computing the posterior is therefore the central problem for Bayesian statistics.

## 4. Frequentist Statistics

Frequentists don't like putting priors on parameters. They also don't want to be restricted to probabilistic computations. Thus, Frequentists consider *estimators* for $\Theta$, which are functions of $X$. They then try to understand various criteria like the *bias*, *variance* and *consistency* of the estimators. They do this by treating the data as random based on a true parameter, $\Theta$.

The core problem for Frequentists is understanding the relationship between the true parameter $\Theta$ and the estimator $\Theta_0$ as the number of data points increases. For example, consistency involves the convergence of $\Theta_0$ to $\Theta$ as the number of data points increases. Bias corresponds to systematic errors in the value of $\Theta_0$ relative to $\Theta$, i.e. an unbiased estimator is one such that $\mathbb{E}[\Theta_0 - \Theta] = 0$. And the variance of an estimator reflects the spread of the estimator, $\Theta_0$, around $\Theta$.

One particularly important estimator is the maximum likelihood estimator (MLE). To compute the MLE, we treat $p(X \mid \Theta)$, which we call the *likelihood*, as a function of $\Theta$. We then choose $\hat{\Theta} = \arg\max_{\Theta} p(X \mid \Theta)$ as our estimator. Notice that no prior is required for performing this computation.

In practice, we often use the *log likelihood* when defining the MLE: $\hat{\Theta} = \arg\max_{\Theta} \log p(X \mid \Theta)$. We use the log likelihood because the logarithm

is a monotonic function of the likelihood (and hence leaves the maximum unchanged), but it is easier to work with algebraically and computationally.

## 5. BLURRED LINES

The boundary between Bayesian and Frequentist approaches is not so clear. For example, Bayesian ideas can be used in Frequentist calculations. Consider the "Bayes Estimate":

$$\hat{\Theta}_{Bayes} = \mathbb{E}[\Theta \mid X], \tag{8}$$

alternatively defined as,

$$\hat{\Theta}_{Bayes} = \int \Theta p(\Theta \mid X)\, d\Theta. \tag{9}$$

Likewise, we can create the Bayesian analogue to the MLE, which is the maximum a posteriori (MAP) estimator:

$$\hat{\Theta}_{MAP} = \arg\max_{\Theta} p(\Theta \mid X). \tag{10}$$

Because $p(\Theta)$ is constant with respect to $\Theta$, we can use Bayes' Rule to simplify this problem, giving,

$$\hat{\Theta}_{MAP} = \arg\max_{\Theta} p(\Theta)p(X \mid \Theta). \tag{11}$$

After taking logs we find this equivalent equation:

$$\hat{\Theta}_{MAP} = \arg\max_{\Theta} [\log(p(\Theta)) + \log(p(X \mid \Theta))]. \tag{12}$$

Here the $\log(p(\Theta))$ term is called a *regularizer* or a *penalty*, and the equation as a whole becomes a *penalized likelihood*. We can also examine Frequentist properties of these estimators.

Returning to Bayesian statistics, it is worth noting that our priors on $\Theta$ themselves are distributions, and so they have parameters as well. These are called *hyperparameters* and are usually labelled $\alpha$. We then consider $p(\Theta \mid \alpha)$.

We can address this computation using Frequentist ideas. For example, the hyperparameters can be estimated with MLE, as shown below:

$$\alpha_{ML} = \arg\max_{\alpha} p(X \mid \alpha). \tag{13}$$

In this equation,

$$p(X \mid \alpha) = \int p(X \mid \Theta)p(\Theta \mid \alpha)\, d\Theta. \tag{14}$$

This approach is called “Empirical Bayes”. It is worth noting that Lindley, an orthodox Bayesian, said that “there is nothing less Bayesian than empirical Bayes.”