IEEE Information Theory Society Newsletter

Vol. 53, No.4, December 2003 Editor: Lance C. Pérez ISSN 1059-2362

The Shannon Lecture

Hidden Markov Models and the Baum-Welch Algorithm

Lloyd R. Welch

Content of This Talk

The lectures of previous Shannon Lecturers fall into several categories such as introducing new areas of research, resuscitating areas of research, surveying areas identified with the lecturer, or reminiscing on the career of the lecturer. In this talk I decided to restrict the subject to the Baum-Welch "algorithm" and some of the ideas that led to its development.

I am sure that most of you are familiar with Markov chains and Markov processes. They are natural models for various communication channels in which channel conditions change with time. In many cases it is not the state sequence of the model which is observed but the effects of the process on a signal. That is, the states are not observable but some functions, possibly random, of the states are observed. In some cases it is easy to assign the values of the parameters to model a channel. All that remains is to determine what probabilities are desired and derive the necessary algorithms to compute them.

In other cases, the choice of parameter values is only an estimate and it is desired to find the "best" values. The usual criterion is maximum likelihood. That is: find the values of parameters which maximizes the probability of the observed data. This is the problem that the Baum-Welch computation addresses.

Preliminaries

Let $\mathcal{N}$ be the set of non-negative integers. Let's introduce some useful notation to replace the usual n-tuple notations:

$$[a_k]_{k=i}^{j} \equiv (a_i, a_{i+1}, \ldots, a_j)$$
$$[a(k)]_{k=i}^{j} \equiv (a(i), a(i+1), \ldots, a(j))$$

The '$k =$' will be dropped from the subscript when it is clear what the 'running variable' is.

Of particular use will be the concept of conditional probability and recursive factorization. The recursive factorization idea says that the joint probability of a collection of events can be expressed as a product of conditional probabilities, where each is the probability of an event conditioned on all previous events. For example, let A, B, and C be three events. Then

$$\Pr(A \cap B \cap C) = \Pr(A)\Pr(B \mid A)\Pr(C \mid A \cap B)$$

Using the bracket notation, we can display the recursive factorization of the joint probability distribution of a sequence of discrete random variables:

$$\Pr\big([\mathbb{X}(k)]_0^N = [x_k]_0^N\big) = \Pr(\mathbb{X}(0) = x_0) \cdot \prod_{n=0}^{N} \Pr\big(\mathbb{X}(n) = x_n \mid [\mathbb{X}(k)]_0^{n-1} = [x_k]_0^{n-1}\big)$$

Markov Chains and Hidden Markov Chains

We will treat only Markov Chains which have finite state spaces. The theory is more general, but to cover the more general case will only obscure the basic ideas.

Let **S** be a finite set, the set of states. Let the number of elements in **S** be M. It will be convenient to identify the elements of **S** with the integers from 1 to M.

Let $\{\mathbb{S}(t) : t \in \mathcal{N}\}$ be a sequence of random variables with $\Pr(\mathbb{S}(t) \in \mathbf{S}) = 1$ for all $t \in \mathcal{N}$. That is, the values of $\mathbb{S}(t)$ are confined to **S**.

Applying the above factorization to the joint distribution of the first $N+1$ random variables gives:

continued on page 10

Hidden Markov Models and the Baum-Welch Algorithm (continued from page 1)

$$\Pr\big([\mathbb{S}(k)]_0^N{=}[s_k]_0^N\big) = \Pr(\mathbb{S}(0) = s_0)\cdot \prod_{n=1}^{N} \Pr\big(\mathbb{S}(n){=}s_n \,|\, [\mathbb{S}(k)]_0^{n-1} = [s_k]_0^{n-1}\big) \tag{1}$$

For the sequence of random variables to be a Markov Chain the conditional probabilities must only be a function of the last random variable in the condition so that equation (1) reduces to

$$\Pr\big([\mathbb{S}(k)]_0^N{=}[s_k]_0^N\big) = \Pr(\mathbb{S}(0) = s_0)\cdot \prod_{n=1}^{N} \Pr(\mathbb{S}(n){=}s_n \,|\, \mathbb{S}(n-1) = s_{n-1}) \tag{2}$$

In my work, the transition probabilities were stationary, that is they are constant functions of time:

$$\Pr(\mathbb{S}(n) = j \,|\, \mathbb{S}(n-1) = i) = \Pr(\mathbb{S}(1) = j \,|\, \mathbb{S}(0) = i) \stackrel{\text{def}}{=} p_{ij}$$

In addition to the Markov Chain, let $\{\mathbb{Y}(t) : t \in \mathcal{N}\}$ be a sequence of random variables (called random observations). It will be convenient to assume that the values are confined to a discrete set and an experiment consists of observing values of T consecutive random variables. Again, treating a more general case will only obscure the basic ideas.

Applying recursive factorization to the joint distribution of the first $T+1$ random states and first T random observations:

$$\begin{aligned}\Pr\big([\mathbb{S}(t)]_0^T &= [s_t]_0^T \text{ and } [\mathbb{Y}(t)]_1^T = [y_t]_1^T\big) = \\ \Pr(\mathbb{S}(0) &= s_0)\cdot \prod_{t=1}^{T} \Pr\big(\mathbb{S}(t) = s_t \,|\, [\mathbb{S}(k)]_0^{t-1} = [s_k]_0^{t-1}\big)\cdot \\ \prod_{t=1}^{T} \Pr\big(\mathbb{Y}(t) &= y_t \,|\, [\mathbb{S}(k)]_0^T = [s_k]_0^T \text{ and } [\mathbb{Y}(k)]_1^{t-1} = [y_k]_1^{t-1}\big)\end{aligned} \tag{3}$$

The next simplifying assumption is that the conditional probability distribution of $\mathbb{Y}(t)$ given all states and all previous random observations is only a function of $\mathbb{S}(t)$ (and not of time). I considered also the case when the distribution of $\mathbb{Y}(t)$ depends on $\mathbb{S}(t)$ and $\mathbb{S}(t-1)$. However, though it added little to the computational complexity, it added significantly to the number of parameters to be estimated. Making use of the above conditions,

$$\begin{aligned}\Pr\big([\mathbb{S}(t)]_0^T = [s_t]_0^T &\text{ and } [\mathbb{Y}(t)]_1^T = [y_t]_1^T\big) = \\ \Pr(\mathbb{S}(0) = s_0)\cdot &\prod_{t=1}^{T} \Pr(\mathbb{S}(t) = s_t \,|\, \mathbb{S}(t-1) = s_{t-1})\cdot \\ &\prod_{t=1}^{T} \Pr(\mathbb{Y}(t) = y_t \,|\, \mathbb{S}(t) = s_t)\end{aligned} \tag{4}$$

To simplify notation, define

$$\begin{aligned} f(y \,|\, s) &\stackrel{\text{def}}{=} \Pr(\mathbb{Y}(t) = y \,|\, \mathbb{S}(t) = s), \\ \underline{s} &\stackrel{\text{def}}{=} [s_t]_0^T, \\ \underline{y} &\stackrel{\text{def}}{=} [y_t]_1^T, \text{ and} \\ p(\underline{s}, \underline{y}) &\stackrel{\text{def}}{=} \Pr([\mathbb{S}(t)]_0^T = \underline{s} \text{ and } [\mathbb{Y}(t)]_1^T = \underline{y}). \end{aligned}$$

With this notation we have

$$\Pr\big([\mathbb{S}(t)]_0^T = [s_t]_0^T \text{ and } [\mathbb{Y}(t)]_1^T = [y_t]_1^T\big) = p(\underline{s}, \underline{y}) = p_{s_0} \prod_{t=1}^{T} p_{s_{t-1}s_t} f(y_t \,|\, s_t) \tag{5}$$

This formula gives the probability of the atoms of the model, that is, those events that can not be subdivided into smaller events. The probability of any event describable in the model is the sum of the probability of the atoms.

Of course, the probabilities are functions of the parameters of the model, which we will denote by:

$$\lambda \stackrel{\text{def}}{=} \left\{ \begin{array}{l} p_s : 1 \le s \le M, \\ p_{s\sigma} : 1 \le s, \sigma \le M, \\ f(y \,|\, s) : 1 \le y \le Y, 1 \le s \le M \end{array} \right\}$$

and use λ as a function argument where appropriate.

Questions

What questions are of interest? One question is what is the probability that an T-tuple of $\mathbb{Y}(t)$ will be observed. This, of course, is a function of the parameters. The probability of an T-tuple of observations is just the sum over all state sequences of the probabilities of the corresponding atoms:

$$p(\underline{y}; \lambda) \stackrel{\text{def}}{=} \Pr\big([\mathbb{Y}(t)]_1^T = [y_t]_1^T; \lambda\big) = \sum_{\underline{s}} p(\underline{s}, \underline{y}; \lambda) \tag{6}$$

Then $p(\underline{y}; \lambda)$ is the probability of the observations, $\underline{y}$. It is also the likelihood function for λ given the observations, $\underline{y}$. A standard problem is to choose λ to maximize the likelihood function.

It is frequently the case that the random observables, $\{\mathbb{Y}(t) : t \in \mathcal{N}\}$, are observed for some period of time and it is desired to find some information about the state sequence from those observations. For example, given the event, $\{[\mathbb{Y}(t)]_1^T = [y_t]_1^T\}$, we may wish to find the probability distribution of the state at a specified time, τ. That is we wish to find, for a given sequence of observations, the probability distribution of s_τ given those observations:

$$Pr\big(\mathbb{S}(\tau) = s_\tau \,|\, [\mathbb{Y}(t)]_1^T = [y_t]_1^T\big)$$

Since,

$$Pr\big(\mathbb{S}(\tau) = s_\tau \,|\, [\mathbb{Y}(t)]_1^T = [y_t]_1^T\big) = \frac{Pr\big(\mathbb{S}(\tau) = s_\tau \text{ and } [\mathbb{Y}(t)]_1^T = [y_t]_1^T\big)}{Pr\big([\mathbb{Y}(t)]_1^T = [y_t]_1^T\big)}$$

the computation of the a posteriori probability is equivalent to computing the joint probability. Referring to equation (5), we see that this reduces to computing

$$\begin{aligned}\Gamma_\tau(s_\tau) &\stackrel{\text{def}}{=} Pr\big(\mathbb{S}(\tau) = s_\tau \text{ and } [\mathbb{Y}(t)]_1^T = [y_t]_1^T\big) \\ &= \sum_{[s_k]_0^{\tau-1}} \sum_{[s_k]_{\tau+1}^T} p_{s_0} \prod_{t=1}^{N} p_{s_{t-1}s_t} f(y_t \,|\, s_t)\end{aligned} \tag{7}$$

for each choice of s_τ. With the exception of s_τ, each indexed state is a summation variable and, with the exception of s_0, occurs in exactly two factors. It is easily deduced that the equation can be expressed in terms of the product of matrices.

However, there is a better (at least to me) approach to computing this probability. We begin with the joint probability of the T-tuple of observations and the state at time τ and apply recursive factorization where the first event is the set of observations up through time, τ, and the state at time τ.

$$\begin{aligned}&\Pr\big([\mathbb{Y}(t)]_1^T = [y_t]_1^T \text{ and } \mathbb{S}(\tau) = s_\tau\big)\\ &\quad = \Pr\big([\mathbb{Y}(t)]_1^\tau = [y_t]_1^\tau \text{ and } \mathbb{S}(\tau) = s_\tau\big)\cdot\\ &\qquad \Pr\big([\mathbb{Y}(t)]_{\tau+1}^T = [y_t]_{\tau+1}^T \,|\, [\mathbb{Y}(t)]_1^\tau = [y_t]_1^\tau \text{ and } \mathbb{S}(\tau) = s_\tau\big)\end{aligned}$$

Now Markovity of the state sequence implies that the probablility of $[\mathbb{S}_t]_{\tau+1}^T$ and therefore the probability of $[\mathbb{Y}_t]_{\tau+1}^T$ are independent of history prior to time τ. So the condition on the $\mathbb{Y}$ in the second term drop out and the factorization reduces to

$$\begin{aligned}&\Pr\big([\mathbb{Y}(t)]_1^T = [y_t]_1^T \text{ and } \mathbb{S}(\tau) = s_\tau\big)\\ &\quad = \Pr\big([\mathbb{Y}(t)]_1^\tau = [y_t]_1^\tau \text{ and } \mathbb{S}(\tau) = s_\tau\big)\cdot\\ &\qquad \Pr\big([\mathbb{Y}(t)]_{\tau+1}^T = [y_t]_{\tau+1}^T \,|\, \mathbb{S}(\tau) = s_\tau\big)\end{aligned}$$

We next address the problem of computing these factors,

$$\begin{aligned}\alpha_\tau(s) &\stackrel{\text{def}}{=} \Pr\big([\mathbb{Y}(t)]_1^\tau = [y_t]_1^\tau \text{ and } \mathbb{S}(\tau) = s\big),\\ \beta_\tau(s) &\stackrel{\text{def}}{=} \Pr\big([\mathbb{Y}(t)]_{\tau+1}^T = [y_t]_{\tau+1}^T \,|\, \mathbb{S}(\tau) = s\big)\end{aligned}$$

and

$$\Gamma_\tau(s) = \alpha_\tau(s)\cdot\beta_\tau(s)$$

I remark that

$$\sum_s \alpha_T(s) = \Pr\big([\mathbb{Y}(t)]_1^T = [y_t]_1^T\big) = p(\underline{y};\lambda).$$

Now recursive factoring of $\alpha_\tau(s)$ where the first factor is the observations up through time $\tau-1$ and the state at time $\tau-1$ gives

$$\begin{aligned}\alpha_\tau(s) &\equiv \Pr\big([\mathbb{Y}(t)]_1^\tau = [y_t]_1^\tau \text{ and } \mathbb{S}(\tau) = s\big)\\ &= \sum_\sigma \Pr\big([\mathbb{Y}(t)]_1^{\tau-1} = [y_t]_1^{\tau-1} \text{ and } \mathbb{S}(\tau-1) = \sigma\big)\cdot\\ &\Pr\big(\mathbb{Y}(\tau) = y_\tau \text{ and } \mathbb{S}(\tau) = s \,|\, [\mathbb{Y}(t)]_1^{\tau-1} = [y_t]_1^{\tau-1} \text{ and } \mathbb{S}(\tau-1) = \sigma\big)\end{aligned}$$

Again, Markovity implies that the condition, $[\mathbb{Y}(t)]_1^{\tau-1} = [y_t]_1^{\tau-1}$ can be dropped from the second factor:

$$\begin{aligned}\alpha_\tau(s) &= \sum_\sigma \Pr\big([\mathbb{Y}(t)]_1^{\tau-1} = [y_t]_1^{\tau-1} \text{ and } \mathbb{S}(\tau-1 = \sigma\big)\cdot\\ \Pr\big(\mathbb{Y}(\tau) &= y_\tau \text{ and } \mathbb{S}(\tau) = s \,|\, \mathbb{S}(\tau-1) = \sigma\big)\end{aligned}$$

The first factor is just $\alpha_{\tau-1}(\sigma)$ and the second factor is $p_{\sigma s}\cdot f(y_\tau|s)$ and above equation becomes the recursion:

$$\alpha_\tau(s) = \sum_\sigma \alpha_{\tau-1}(\sigma)p_{\sigma s}f(y_\tau \,|\, s) \tag{8}$$

Similarly, a reverse time recursion exists for $\beta_\tau(s)$:

$$\beta_\tau(s) = \sum_\sigma p_{s\sigma} f(y_{\tau+1} \,|\, \sigma)\beta_{\tau+1}(\sigma) \tag{9}$$

Finally we have

$$\begin{aligned}\Pr\big(\mathbb{S}(\tau) = s \text{ and } [\mathbb{Y}(t)]_1^T = [y_t]_1^T\big) &= \Gamma_\tau(s)\\ &= \alpha_\tau(s)\beta_\tau(s)\end{aligned}$$

and

$$\Pr(\mathbb{S}(\tau) = s \,|\, \underline{y}) = \frac{\alpha_\tau(s)\cdot\beta_\tau(s)}{\sum_\sigma \alpha_T(\sigma)}.$$

Once we have routines for computing α and β we can compute not only $\Pr(\mathbb{S}(t) = s \,|\, \underline{y})$ but also the a posteriori probability of other 'local' events, such as the event, $\{\mathbb{S}(t) = s \text{ and } \mathbb{S}(t+1) = \sigma\}$. In this case the expression is

$$\begin{aligned}&\Pr(\mathbb{S}(t-1) = s, \mathbb{S}(t) = \sigma \text{ and } [\mathbb{Y}(\tau)]_1^T = \underline{y})\\ &\quad = \alpha_t(s)p_{s\sigma}f(y_{t+1} \,|\, \sigma)\beta_{t+1}(\sigma)\\ &\quad \equiv \Gamma_t(s,\sigma)\end{aligned} \tag{10}$$

Improving on Estimates of Parameters

At this point Leonard Baum and I found that we had both been working independently on Hidden Markov Chains and had both come up with essentially the same calculation for a posteriori probabilities of 'local' events. At that point we joined forces.

Now, the above calculations were based upon specified parameter values. What if those parameter values did not adequately represent the phenomena under investigation? My thoughts proceeded as follows. While the parameters may not be accurate, the a posteriori probabilities may translate to better parameters.

For example, from the frequency interpretation of probability, if we could observe the state sequence over a long period of time and count the number of times the state, s, occurs, the frequency of occurrence should be approximately $\Pr(\mathbb{S}(t) = s)$ If the assumed parameters are correct, we will get p_s, where for a large enough period of time p will be the stationary distribution (the eigenvector with eigenvalue 1) of the transition matrix.

Furthermore, if the observation sequence, $[y_t]_1^T$, is a typical sequence in the Information Theoretic sense, that is, it has high probability using the parameters of the model, then the expected frequencies of states given the observations should also be approximately p_s, where p_s is the stationary distribution, not the initial distribution. Expressed in equation form:

$$\frac{\sum_{t=1}^T \Pr\big(\mathbb{S}(t) = s \,|\, [\mathbb{Y}(\tau)]_1^T = [y_\tau]_1^T\big)}{T} \approx p_s\,.$$

Similarly, from the frequency interpretation of probability, if we could observe the state sequence over a long period of time and count the number of times the state, s, occurs followed by σ, the frequency of occurrence should be approximately $\Pr(\mathbb{S}(t-1) = s$ and $\mathbb{S}(t) = \sigma)$. If the assumed parameters are correct, we will get $p_s\cdot p_{s\sigma}$.

Again if the observation sequence, $[y_t]_1^T$, is a typical sequence in

the Information Theoretic sense, then the expected frequencies of state transitions given the observations should also be approximately $p_s \cdot p_{s\sigma}$. Expressed in equation form:

$$\frac{\sum_{t=1}^{T} \Pr(\mathbb{S}(t-1)=s, \mathbb{S}(t)=\sigma | [\mathbb{Y}(\tau)]_1^T = [y_\tau]_1^T)}{T} \approx p_s \cdot p_{s\sigma}\ .$$

Finally, if we could observe the state sequence and the observation sequence and count the number of times $\mathbb{S}(t)=s$, $\mathbb{Y}(t)=y$, and the frequency should approximate $p_s f(y|s)$. Again, if the observed sequence is a typical sequence for the given parameters, the a posteriori expected frequencies should approximate $p_s f(y|s)$, i.e.,

$$\begin{aligned} p_s f(y\,|\,s) &\approx \frac{\sum_{t=1}^{T} \Pr(\mathbb{S}(t)=s, \mathbb{Y}(t)=y\,|\,[\mathbb{Y}(\tau)]_1^T=[y_\tau]_1^T)}{T} \\ &\approx \frac{\sum_{t\in\{t:y=y_t\}} \Pr(\mathbb{S}(t)=s\,|\,[\mathbb{Y}(\tau)]_1^T=[y_\tau]_1^T)}{T} \end{aligned}$$

My next thought was that if $\underline{y}$ was generated by a model with different parameter values and therefore not a typical sequence for the assumed values, the a posteriori frequencies, influenced by behavior of $[y_\tau]_1^T$, may be a better indication of the true parameters that the initial guess. So I replaced the parameter values by the expected frequencies and recomputed $p(\underline{y};\lambda')$ where

$$\lambda' \stackrel{\text{def}}{=} \left\{ \begin{array}{l} p_s(\lambda') = \frac{\sum_{t=1}^{T} \Pr(\mathbb{S}(t)=s\,|\,[\mathbb{Y}(\tau)]_1^T=[y_\tau]_1^T)}{T} \\ p_{s\sigma}(\lambda') = \frac{\sum_{t=1}^{T} \Pr(\mathbb{S}(t-1)=s, \mathbb{S}(t)=\sigma\,|\,[\mathbb{Y}(\tau)]_1^T=[y_\tau]_1^T)}{T} \\ f(y\,|\,s;\lambda') \leftarrow \frac{\sum_{t=1}^{T} \Pr(\mathbb{S}(t)=s, \mathbb{Y}(t)=y\,|\,[\mathbb{Y}(\tau)]_1^T=[y_\tau]_1^T)}{T p_s(\lambda')} \end{array} \right\} \quad (11)$$

I was please to find that $p(\underline{y};\lambda') > p(\underline{y};\lambda)$. In other words, this substitution increased the likelihood function! I tried this transformation on several Hidden Markov Models and the likelihood function always increased. Leonard Baum tried it on a number of examples and again the likelihood function always increased.

That is my contribution to the Baum-Welch `algorithm', the easy part. I tried to provide a mathematical proof that the likelihood always increases but I failed.

Baum, in cooperation with J. Eagon did the hard part by proving that this transformation either increases the likelihood function or leaves it constant. In the latter case, λ is a fixed point of the transformation. Their proof involved rather complex computations and applications of Hölder's inequality and the fact that the geometric mean is less than or equal to the arithmetic mean.

Later Baum, together with T. Petrie, G. Soules and N. Weiss, all at CRD/IDA at the time found a more elegant proof with the flavor of Information Theory which I will now discuss.

The Q Function

They began with the Kullback-Leibler divergence of two distributions:

$$D(p_1, p_2) \equiv \sum_{\omega} p_1(\omega) \log\left(\frac{p_1(\omega)}{p_2(\omega)}\right)$$

where p_1 and p_2 are two probability distributions on a discrete space and ω is summed over that space. The interpretation of this number is that for an experiment consisting of multiple selections from the distribution p_1, $D(p_1, p_2)$ is the expected log factor of the probability in favor of p_1 against p_2. It is an information theoretic measure and is known to be non-negative, equaling zero only when the $p_2(\omega) = p_1(\omega)$ for all ω for which $p_1(\omega) > 0$.

How does this apply to Hidden Markov Models? We let

$$p_1(s) = \frac{p(\underline{s}, \underline{y}; \lambda)}{p(\underline{y};\lambda)} \text{ and } p_2(s) = \frac{p(\underline{s}, \underline{y}; \lambda')}{p(\underline{y};\lambda')}.$$

Then p_1 and p_2 are distributions and

$$\begin{aligned} 0 \le D(\lambda, \lambda') &= \sum_s \frac{p(\underline{s}, \underline{y}; \lambda)}{p(\underline{y};\lambda)} \log\left(\frac{p(\underline{s}, \underline{y}; \lambda) p(\underline{y};\lambda')}{p(\underline{s}, \underline{y}; \lambda') p(\underline{y};\lambda)}\right) \\ &= \log\left(\frac{p(\underline{y};\lambda')}{p(\underline{y};\lambda)}\right) + \sum_s \frac{p(\underline{s}, \underline{y}; \lambda)}{p(\underline{y};\lambda)} \log\left(\frac{p(\underline{s}, \underline{y}; \lambda)}{p(\underline{s}, \underline{y}; \lambda')}\right). \end{aligned}$$

We simplify this by defining

$$Q(\lambda, \lambda') \equiv \sum_{\underline{s}} p(\underline{s}, \underline{y}; \lambda) \log(p(\underline{s}, \underline{y}; \lambda')).$$

Then

$$0 \le D(\lambda, \lambda') = \log\left(\frac{p(\underline{y};\lambda')}{p(\underline{y};\lambda)}\right) + \frac{Q(\lambda, \lambda) - Q(\lambda, \lambda')}{p(\underline{y};\lambda)} \quad (12)$$

and rearranging the inequality we have

$$\frac{Q(\lambda, \lambda') - Q(\lambda, \lambda)}{p(\underline{y};\lambda)} \le \log\left(\frac{p(\underline{y};\lambda')}{p(\underline{y};\lambda)}\right)$$

and this implies that if $Q(\lambda, \lambda') > Q(\lambda, \lambda)$ then $p(\lambda') > p(\lambda)$

Hill Climbing

We obtain a "hill climbing" algorithm by finding that λ' which maximizes $Q(\lambda, \lambda')$ as a function of its second argument. If $Q(\lambda, \lambda') > Q(\lambda, \lambda)$ then $p(\lambda') > p(\lambda)$ and we have succeeded in increasing $p(\lambda)$ which is the probability of the observations.

To maximize $Q(\lambda, \lambda')$ we begin by finding the critical points of Q as a function of λ' and subject to the stochastic constraints on the components of λ'. (A sample constraint is $\sum_j p_{ij} = 1$).

Before proceeding, let's manipulate the expression for Q. In equation (5) the expression for $p(\underline{s}, \underline{y}; \lambda)$ is a product, so its logarithm is a sum of log factors. Replacing λ by λ' in equation(5) and taking logarithms we have:

$$\begin{aligned} \log(p(\underline{s}, \underline{y}; \lambda')) &= \log(p_{s(0)}(\lambda')) \\ &\quad + \sum_{t=1}^{T} \log(p_{s(t-1)s(t)}(\lambda') f(y(t); \lambda' \,|\, s(t-1)s(t))) \end{aligned}$$

Substituting into the definition for Q gives:

$$Q(\lambda, \lambda') = \sum_{\underline{s}} p(\underline{s}, \underline{y}, \lambda) \log(p_{s(0)}(\lambda')) \quad (13)$$
$$+ \sum_{\underline{s}} p(\underline{s}, \underline{y}, \lambda) \sum_{t=1}^{T} \log(p_{s(t-1)s(t)}(\lambda') f(y(t); \lambda' \mid s(t-1)s(t))).$$

From this equation it can be seen that it is easy to differentiate with respect to the components of λ', add the appropriate Lagrange factors and solve. The result has already been displayed in equation (11).

Applications

There are too many papers published on applications to list here. In the area of speech recognition, here is a small sample:

F. Jelinek, L. Bahl, and R. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Trans. Inform. Theory*, vol. 21, May 1975.

L.R. Rabiner, "A tutorial on Hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE,* vol. 77, no. 2, Feb. 1989.

A. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proceedings of ICASSP '82,* May 1982.

EM Theory

In 1977, Dempster, Laird and Rubin collected a variety of maximum likelihood problems and methods of solving these problems that occurred in the literature. They found that all of these methods had some ideas in common and they named it the EM Algorithm, (standing for "Expectation, Maximization".)

The common problem is to maximize Prob(y; Φ), as a function of Φ where y is observed. (The probability of y is used in the case of discrete random variables and a density is maximized in the case of continuous random variables.) The observation, y, is viewed as "incomplete data" in the sense that there is a larger model containing "complete data" and y inherits its distribution by way of a mapping from the larger model to the observation model.

Mathematically: There is a probability space, X with a family of probability measures, $p(x; \Phi)$, and a mapping function, F with $F(x) = y$. The distribution, q, of y is

$$q(y; \Phi) = \sum_{\{x:F(x)=y\}} p(x; \Phi).$$

The conditional distribution of x given y is

$$p(x \mid y; \Phi) = \frac{p(x; \Phi)}{q(y; \Phi)},$$

provided $F(x) = y$.

Given a second value of Φ, say Φ' they define

$$Q(\Phi, \Phi') = \mathcal{E}\{\log(p(x; \Phi')) \mid y; \Phi\},$$

where $\mathcal{E}$ is the notation for the expected value function. This is the Expectation step. It gives a formula in Φ'. The Maximization step is to vary Φ' to maximize Q. In many problems the maximization step is easy and in many others it is as difficult as the original maximum likelihood problem. This leads to a "Generalized Estimation Maximization" which simply finds any Φ' which increases Q.

The Baum-Welch algorithm fits right into the EM scheme. x is $(\underline{s}, \underline{y})$ and the observation is y. The Q function is exactly the Q function that Baum *et al*. introduced to prove that the transformation increases the likelihood.

Recommended Reading

There are many papers published on these subjects. A few are:

L.E. Baum and J. Eagon "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bulletin of the American Mathematical Soc.,* vol. 70, pp. 360–363.

L.E. Baum, T. Petrie, G. Soules and N. Weiss, "A maximization technique occurring in the statictical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.,* vol. 41, 1970.

A. Dempster, N. Laird and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, B, vol. 39, 1977.

A paper which extends theory to observations with continuous distributions.

L. Liporace, "Maximum likelihood estimates for multivariate observations Markov sources," *IEEE Trans. Inform. Theory*, vol. 28, Sept. 1982.