Lecture 23: The Expectation-Maximization Algorithm and Hidden Markov Models

ELE 525: Random Processes in Information Systems

Hisashi Kobayashi

Department of Electrical Engineering
Princeton University
December 11, 2013

19.2  Expectation-Maximization Algorithm for
Maximum-Likelihood Estimation

19.2.1  EM Algorithm for Transformed Data

We consider the case where the observable variable $Y$ is a transformed variable of $X$, which may or may not be completely unobservable

$$Y = T(X), \quad X \in \mathcal{X}, \quad Y \in \mathcal{Y}, \quad (19.17)$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the sample spaces of the variables $X$ and $Y$.

The variable $Y$ is called incomplete, while $X$ is a complete variable.

Our final goal is to find an MLE of $\theta$, given an observation instance $y$.

Let us start with the joint variable $(X, Y)$, and denote its probability by $p_{X,Y}(x, y; \theta)$. We assume that the parameter $\theta$ is fixed.
If we substitute $y = T(x)$,

$$p_X(x; \theta) = p_{X,Y}(x, T(x); \theta). \quad (19.18)$$

Then

$$p_{X|Y}(x|y; \theta) = \frac{p_{X,Y}(x, y; \theta)}{p_Y(y; \theta)} = \frac{p_X(x; \theta)}{p_Y(y; \theta)}, \quad (19.19)$$

from which we obtain the log-likelihood function $\log L_y(\theta) = \log p_Y(y; \theta)$ as

$$\log L_y(\theta) = \log p_X(x; \theta) - \log p_{X|Y}(x|y; \theta). \quad (19.20)$$

Since the complete data $x$ is not available to us, $\log L_y(\theta)$ is a function of the complete variable $X$ given by

$$\log L_y(X; \theta) = \log p_X(X; \theta) - \log p_{X|Y}(X|y; \theta). \quad (19.21)$$

Now we take expectations on both sides of (19.21) with respect to $p_{X|Y}(x|y; \theta^{(p)})$, and note

$$E \left[ \log L_y(X; \theta) | y; \theta^{(p)} \right] = \log L_y(\theta).$$

$$\log L_y(\theta) = E \left[ \log p_X(X; \theta) | y; \theta^{(p)} \right] - E \left[ \log p_{X|Y}(X|y; \theta) | y; \theta^{(p)} \right], \quad (19.22)$$
\[ \log L_y(\theta) = E \left[ \log p_X(X; \theta) | y; \theta^{(p)} \right] - E \left[ \log p_{X|Y}(X|y; \theta) | y; \theta^{(p)} \right], \quad (19.22) \]

The last equation can be written as

\[ \log L_y(\theta) = Q(\theta | \theta^{(p)}) + H(\theta | \theta^{(p)}), \quad (19.23) \]

where \( Q(\theta | \theta^{(p)}) \) is an auxiliary function, called the Q-function

\[ Q(\theta | \theta^{(p)}) \triangleq E \left[ \log p_X(X; \theta) | y; \theta^{(p)} \right], \quad (19.24) \]

and

\[ H(\theta | \theta^{(p)}) \triangleq -E \left[ \log p_{X|Y}(X|y; \theta) | y; \theta^{(p)} \right]. \quad (19.25) \]

The function \( H \) of (19.25) satisfies the following inequality (see Problem 19.1 (a))

\[ H(\theta | \theta^{(p)}) \geq H(\theta^{(p)} | \theta^{(p)}) = \mathcal{H}(X | y; \theta^{(p)}), \quad (19.26) \]

where \( \mathcal{H}(X | y; \theta^{(p)}) \) is the conditional entropy.
Relationship between the log-likelihood function and $Q$-function

The auxiliary function defined in (19.24)

$$Q(\theta|\theta^{(p)}) \triangleq E \left[ \log p_X(X; \theta) | y; \theta^{(p)} \right], \quad (19.24)$$

should read as the expectation of $\log p_X(X; \theta)$ under the probability measure $p_{X|Y}(x|y; \theta^{(p)})$ for the complete variable $X$, with the observation instance $y$ fixed.

The auxiliary function $Q$ will help us find an iterative algorithm to arrive at the MLE because the following important relation holds between $\log L_y(\theta)$ and $Q(\theta|\theta^{(p)})$

$$Q(\theta|\theta^{(p)}) - Q(\theta^{(p)}|\theta^{(p)}) \leq \log L_y(\theta) - \log L_y(\theta^{(p)}), \quad (19.27)$$

where the equality holds when $\theta = \theta^{(p)}$. 

12/11/2013

Copyright Hisashi Kobayashi 2013

5
Equation (19.26)

\[
H(\theta|\theta^{(p)}) \geq H(\theta^{(p)}|\theta^{(p)}) = \mathcal{H}(X|y; \theta^{(p)}),
\]

shows that \(H(\theta|\theta^{(p)})\) attains its minimum at \(\theta = \theta^{(p)}\), and thus \(\nabla_{\theta} H(\theta|\theta^{(p)})|_{\theta=\theta^{(p)}} = 0\).

Then differentiating both sides of (19.23).

\[
\log L_y(\theta) = Q(\theta|\theta^{(p)}) + H(\theta|\theta^{(p)}),
\]

we obtain

\[
\nabla_{\theta} \log L_y(\theta)|_{\theta=\theta^{(p)}} = \nabla_{\theta} Q(\theta|\theta^{(p)})|_{\theta=\theta^{(p)}}.
\]
Derivation of the E-step and the M-step

Suppose that we have obtained an estimate \( \theta^{(p)} \) after \( p \) iterations.

We want to improve upon this estimate using \( Q(\theta|\theta^{(p)}) \) as a guide.

\[
Q(\theta|\theta^{(p)}) = E \left[ \log p_X(X; \theta) | y, \theta^{(p)} \right]. \quad \text{(E-step)}
\]  

(19.29)

This \( Q \) function is a function of \( \theta \) and the observation \( y \), given the current model parameter \( \theta^{(p)} \). The variable \( X \) is averaged out by taking the expectation.

The inequality (19.27)

\[
Q(\theta|\theta^{(p)}) - Q(\theta^{(p)}|\theta^{(p)}) \leq \log L_y(\theta) - \log L_y(\theta^{(p)}),
\]  

(19.27)

implies that if \( Q(\theta|\theta^{(p)}) > Q(\theta^{(p)}, \theta^{(p)}) \), then \( L_y(\theta) > L_y(\theta^{(p)}) \).

Therefore, the best choice for the next estimate \( \theta^{(p+1)} \) will be found by maximizing \( Q(\theta|\theta^{(p)}) \) with respect to \( \theta \):

\[
\theta^{(p+1)} = \arg \max_{\theta} Q(\theta|\theta^{(p)}). \quad \text{(M-step)}
\]  

(19.30)
If $Q(\theta|\theta^{(p)})$ is differentiable, the next estimate $\theta^{(p+1)}$ satisfies
\[ \nabla_{\theta} Q(\theta|\theta^{(p)})|_{\theta=\theta^{(p+1)}} = 0. \]
Thus, if the algorithm converges to $\theta^*$, $\nabla_{\theta} Q(\theta|\theta^*)|_{\theta=\theta^*} = 0$, which implies
\[ \nabla_{\theta} L(\theta)|_{\theta=\theta^*} = 0. \] (19.31)
according to (19.28)
\[ \nabla_{\theta} \log L_y(\theta)|_{\theta=\theta^{(p)}} = \nabla_{\theta} Q(\theta|\theta^{(p)})|_{\theta=\theta^{(p)}}. \] (19.28)
Thus, the algorithm converges to the *stationary point* of the likelihood function.

Therefore, an MLE $\hat{\theta}$ can be found by the above iterative procedure of alternating the *expectation step* (E-step) (19.29) and the *maximization step* (M-step) (19.30).
Algorithm 19.2  EM Algorithm for an MLE

1: Denote the initial estimate of the model parameters as \( \theta^{(0)} \). Set the iteration number \( p = 0 \).
2: Assume the \( p \)-th estimate \( \theta^{(p)} \).
3: Evaluate

\[
E \left[ \log p_X(X; \theta) | y, \theta^{(p)} \right] \triangleq Q(\theta | \theta^{(p)}). \quad (E\text{-step})
\]

4: Find

\( \theta^{(p+1)} = \arg \max_{\theta} Q(\theta | \theta^{(p)}) \). \quad (M\text{-step})

5: If any of the stopping conditions is met, go to step 6. Otherwise, replace \( \theta^{(p)} \) by \( \theta^{(p+1)} \), set \( p \leftarrow p + 1 \), and repeat the Steps 2 through 5.
6: Output \( \theta^{(p+1)} \) as an MLE.
Geometrical interpretation of the EM algorithm

If we apply inequality (19.26)

\[ H(\theta|\theta^{(p)}) \geq H(\theta^{(p)}|\theta^{(p)}) = \mathcal{H}(X|y; \theta^{(p)}), \]  

(19.26)

to

\[ \log L_y(\theta) = Q(\theta|\theta^{(p)}) + H(\theta|\theta^{(p)}), \]  

(19.23)

we obtain

\[ \log L_y(\theta) \geq Q(\theta|\theta^{(p)}) + H(\theta^{(p)}|\theta^{(p)}) \triangleq B(\theta|\theta^{(p)}). \]  

(19.32)

\( B(\theta|\theta^{(p)}) \) differs from the \( Q(\theta|\theta^{(p)}) \) by the term \( H(\theta^{(p)}|\theta^{(p)}) \) which does not depend on \( \theta \), so they both can be used in the M-step of the algorithm.

\( B(\theta|\theta^{(p)}) \) coincides with the log-likelihood at the point \( \theta = \theta^{(p)} \), since

\[ B(\theta^{(p)}|\theta^{(p)}) = Q(\theta^{(p)}|\theta^{(p)}) + H(\theta^{(p)}|\theta^{(p)}) = \log L_y(\theta^{(p)}), \]  

(19.33)
Figure 19.1 The M-step of the EM algorithm
Recall the definition of the B-function:

$$\log L_y(\theta) \geq Q(\theta|\theta^{(p)}) + H(\theta^{(p)}|\theta^{(p)}) \triangleq B(\theta|\theta^{(p)}).$$  \hspace{1cm} (19.32)

According to equation (19.28),

$$\nabla_{\theta} \log L_y(\theta)|_{\theta=\theta^{(p)}} = \nabla_{\theta} Q(\theta)|_{\theta^{(p)}}|_{\theta=\theta^{(p)}}.$$  \hspace{1cm} (19.28)

we have

$$\nabla_{\theta} \log L_y(\theta)|_{\theta=\theta^{(p)}} = \nabla_{\theta} B(\theta|\theta^{(p)})|_{\theta=\theta^{(p)}},$$  \hspace{1cm} (19.34)

thus, \(\log L_y(\theta)\) and its lower bound \(B(\theta|\theta^{(p)})\) have a common tangential plane.

At \(\theta = \theta^{(p)}\), the log-likelihood is \(\log L_y(\theta^{(p)})\), and the M-step finds the next point \(\theta^{(p+1)}\) which maximizes the auxiliary function \(B(\theta|\theta^{(p)})\).

The M-step increases the log-likelihood as well:

$$\log L_y(\theta^{(p)}) = B(\theta^{(p)}|\theta^{(p)}) \leq B(\theta^{(p+1)}|\theta^{(p)}) \leq \log L_y(\theta^{(p+1)}).$$  \hspace{1cm} (19.35)
Like any iterative algorithm there is no guarantee that the point of convergence is the true MLE value: it can be a local maximum or a saddle point.

But the EM algorithm guarantees that the likelihood-function $L_y(\theta^{(p)})$ increases in every step because of (19.35)

$$\log L_y(\theta^{(p)}) = B(\theta^{(p)}|\theta_1^{(p)}) \leq B(\theta^{(p+1)}|\theta^{(p)}) \leq \log L_y(\theta^{(p+1)}).$$

or (19.27),

$$Q(\theta|\theta^{(p)}) - Q(\theta^{(p)}|\theta^{(p)}) \leq \log L_y(\theta) - \log L_y(\theta^{(p)}),$$

In order to reach a global maximum, we must carefully choose the initial estimate $\theta^{(u)}$ close to the MLE, and if this cannot be arranged, we must try a number of choices
Dempster et al. [79] show that the rate of convergence of the EM algorithm depends on the proportion of information in the observed data.
Several methods have been proposed to speed up the EM algorithm.

The logarithm of $p_x(x; \theta)$ may take a simpler form when the probability distribution belongs to the exponential family.

When it is difficult to find $\theta$ that maximizes $Q(\theta|\theta^{(p)})$, we can adopt the Newton-Raphson algorithm or hill-climbing method and choose an appropriate $\theta$

$$Q(\theta|\theta^{(p)}) \geq Q(\theta^{(p)}, \theta^{(p)}).$$

It will give a greater likelihood function, i.e., $L_y(\theta) \geq L_y(\theta^{(p)})$.

Such a method is often referred to as the generalized EM (GEM) method.
19.2.2 EM Algorithm for Missing Data

Let us consider a special case namely, the complete variable $X$ can be written as

$$X = (Y, Z),$$  \hspace{1cm} (19.36)

where $Y$ is the observable variable and $Z$ is a latent variable.

$$T(x) = T(y, z) = y.$$ \hspace{1cm} (19.37)

An instance $x$ is called complete data;
- $y$, observed data or incomplete data,
- $z$, missing data.

Many practical applications of the EM algorithm, such as parameter estimation in a hidden Markov model (HMM) and parameter estimation for pattern classification, can be formulated under this special case defined by (19.36) and (19.37).
The $Q$-function is defined, similar to (19.24)

\[ Q(\theta | \theta^{(p)}) \triangleq E \left[ \log p_X(X; \theta) | y; \theta^{(p)} \right] , \tag{19.24} \]

by

\[ Q(\theta | \theta^{(p)}) \triangleq E \left[ \log p_{YZ}(y, Z; \theta) | y; \theta^{(p)} \right] = \sum_z p_{Z|Y}(z | y; \theta^{(p)}) \log p_{YZ}(y, z; \theta) , \tag{19.38} \]

which reads as the expectation of $\log p_{YZ}(y, Z; \theta)$ under $p_{Z|Y}(z | y; \theta^{(p)})$ for the unobserved variable $Z$, with the observation instance $y$ fixed.

This $Q$-function also satisfies the important property (19.27) (Problem 19.5).

We consider a tight lower bound function $B^*(\theta | \theta^{(p)})$ defined by (19.53)

\[ B^*(\theta | \theta^{(p)}) = Q(\theta | \theta^{(p)}) + H(Z | y; \theta^{(p)}) , \tag{19.53} \]

Then maximization of this $B^*$-function is equivalent to maximization of $Q$-function

(see Problem 19.8).
Bayesian EM algorithm for MAP estimation

Thus far, we assumed that the parameter $\theta$ is unknown but fixed.
If we have a prior probability distribution $p_\Theta(\theta)$ of the parameter variable $\Theta$, then we should take the Bayesian approach, i.e., find the MAP estimate:

$$\hat{\theta}_{\text{MAP}} = \arg \max_\theta p_\Theta|Y(\theta|y).$$ (19.39)

According to equation (18.102)

$$\hat{\theta}^*(x) = \arg \max_\theta [\log L_x(\theta) + \log \pi(\theta)].$$ (18.102)

the MAP estimate can be obtained by

$$\hat{\theta}_{\text{MAP}} = \arg \max_\theta [\log L_y(\theta) + \log \pi(\theta)].$$ (19.40)

Therefore, the M-step of (19.30) should be modified (see Problem 19.9 (a)) to

$$\theta^{(p+1)} = \arg \max_\theta \left[ Q(\theta|\theta^{(p)}) + \log p_\Theta(\theta) \right].$$ (19.41)
Monte Carlo EM Algorithm

The EM algorithm is most effective when both the E-step and the M-step can be performed analytically. However, it is not always possible to do so in practical applications. In such a case we need to resort to some kind of simulation method, such as Markov Chain Monte Carlo (MCMC), which is discussed in Section 21.7.

The E-step to compute the $Q$-function of (19.38)

$$Q(\theta|\theta^{(p)}) = E \left[ \log p(y, Z; \theta) | y, \theta^{(p)} \right]$$  \hspace{1cm} (19.42)

can be approximated, by generating $N^{(p)}$ samples, $z_1, z_2, \ldots, z_t, \ldots, z_{N^{(p)}}$ from the conditional distribution $p(z|y, \theta^{(p)})$ and then by substituting the empirical average for the conditional expectation:

$$Q(\theta|\theta^{(p)}) \approx \frac{1}{N^{(p)}} \sum_{t=1}^{N^{(p)}} \log p(y, z_t; \theta).$$  \hspace{1cm} (19.43)

Maximization of this function with respect to $\theta$ can be performed using the simulated annealing method discussed in Section 21.7.5.
20 Hidden Markov Models and Applications

20.1 Introduction

In some applications of a Markov chain model, we may not be able to directly observe a sequence of states, or may not even know the structure of the Markov model and the model parameters.

The observable variable may be a probabilistic function of the underlying Markov state. Such a model is called a hidden Markov model (HMM).

In this chapter we study how to estimate states (or a sequence of states) and HMM parameters such as state transition probabilities based on observable data.
20.2 Formulation of a Hidden Markov Model

20.2.1 Discrete-Time Markov Chain and Hidden Markov Model

Consider a homogeneous (i.e., time-invariant) DTMC \( \{S_t\} \), where the number of states \( M \) We label the states by integers \( \{0, 1, \ldots, M - 1\} \)

\[
S \triangleq \{0, 1, \ldots, M - 1\}. \quad (20.1)
\]

The subscript \( t \) of \( S_t \) is the discrete-time index

\[
t \in \{0, 1, 2, \ldots, T\} \triangleq T. \quad (20.2)
\]

Let the initial state distribution be denoted by

\[
\pi_0 = (\pi_0(0), \pi_0(1), \ldots, \pi_0(M - 1)) \triangleq (\pi_0(i); i \in S), \quad (20.3)
\]

where \( \pi_0(i) = P[S_0 = i], \ i \in S. \quad (20.4) \]
In a Markov chain

\[ P[S_t = s_t | S_{t-1} = s_{t-1}, S_{t-2} = s_{t-2}, \ldots] = P[S_t = s_t | S_{t-1} = s_{t-1}], \]

or more compactly,

\[ p(s_t | s_{t-1}, s_{t-2}, \ldots) = p(s_t | s_{t-1}). \]

Thus, the (one-step) state transition probabilities \( \{a(i, j)\} \)

\[ A \triangleq [a(i, j); \ i, j \in S], \]

where \( a(i, j) \triangleq P[S_t = j | S_{t-1} = i], \ i, j \in S. \)

Clearly \( \sum_{j \in S} a(i, j) = 1 \), hence \( A \cdot 1 = 1. \)

In the figure of next slide, we show the state transition diagram of time-homogenous Markov chain, and its trellis diagram.
We denote the state sequence in the period $\mathcal{T}$ defined in (20.2)

$$S \triangleq (S_t : t \in \mathcal{T}) \in S^{[T]}, \quad \text{and} \quad s \triangleq (s_t : t \in \mathcal{T}),$$

where $|\mathcal{T}| = T + 1$, the size of set $\mathcal{T}$.

For every instance of the state sequence $s$, there is a unique path in the trellis diagram.
Hidden Markov Model

Let us assume that the Markov chain $S_t$ is not observable,

There is a random process $Y_t$, which is a probabilistic function of $S_t$ and is observable. $Y_t \in \mathcal{Y}$, where

$$\mathcal{Y} \triangleq \{0, 1, 2, \ldots, K - 1\}. \quad (20.9)$$

We denote the observed sequence variable over the period $\mathcal{T}$ and its instance by

$$Y \triangleq (Y_t; \ t \in \mathcal{T}), \quad \text{and} \quad y \triangleq (y_t; \ t \in \mathcal{T}). \quad (20.10)$$

If $Y_t$ is a probabilistic function of only $S_{t-1}$ and $S_t$, then a pair process

$$X_t = (S_t, Y_t), \ t \in \mathcal{T} \quad (20.11)$$

is also a Markov process, since it depends on the past only through $X_{t-1} = (S_{t-1}, Y_{t-1})$. $Y_t$ by itself is generally not a Markov process.
Definition 20.1 (Hidden Markov Model). A Markov process \( X_t = (S_t, Y_t) \) is called a partially observable Markov process or hidden Markov model (HMM), if its state transition probability does not depend on \( Y_{t-1} \), i.e.,

\[
p(x_t | x_{t-1}) = p(x_t | s_{t-1}), \quad (20.12)
\]

i.e.,

\[
p(s_t, y_t | s_{t-1}, y_{t-1}) = p(s_t, y_t | s_{t-1}), \quad (20.13)
\]

Any state \( S_t \in S \) is called a hidden state and the process \( S = (S_t; \ t \in T) \) is a hidden process; and \( y = (y_t; \ t \in T) \) is called an observation.
Definition 20.2 (Model parameters of a HMM). For a homogeneous HMM, we denote

\[ P[S_0 = i, Y_0 = k] \triangleq \alpha_0(i, k), \quad i \in \mathcal{S}, \quad k \in \mathcal{Y}, \]  
(20.14)

\[ P[S_t = j, Y_t = k | S_{t-1} = i] \triangleq c(i; j, k), \quad i, j \in \mathcal{S}, \quad k \in \mathcal{Y}, \quad \text{for } t = 1, 2, \ldots, T. \]  
(20.15)

Denote the sets of initial distribution vectors \( \alpha_0(k) \)'s and the state transition probability matrices \( C(k) \)'s by \( \alpha_0 \) and \( C \), respectively:

\[ \alpha_0 \triangleq (\alpha_0(k); \quad k \in \mathcal{Y}), \]  
(20.16)

where \( \alpha_0(k) \triangleq (\alpha_0(i, k); \quad i \in \mathcal{S}), \quad k \in \mathcal{Y}, \)  
(20.17)

\[ C \triangleq (C(k); \quad k \in \mathcal{Y}), \]  
(20.18)

where \( C(k) \triangleq [c(i; j, k); \quad i, j \in \mathcal{S}], \quad k \in \mathcal{Y}. \)  
(20.19)

We then define the model parameters \( \theta \) associated with the HMM by the set

\[ \theta = (\alpha_0, C). \]  
(20.20)

Note that

\[ \sum_{k \in \mathcal{Y}} \alpha_0(k) = \pi_0. \]  
(20.21)

\[ \sum_{k \in \mathcal{Y}} C(k) = A. \]  
(20.22)
Now we can rewrite $p(s_t, y_t|s_{t-1})$ of (20.13) as

$$p(s_t, y_t|s_{t-1}) = p(s_t|s_{t-1})p(y_t|s_{t-1}, s_t).$$  \hspace{1cm} (20.23)

Then we define a special type of HMM

**Definition 20.3 (State-based HMM versus transition-based HMM).** A homogeneous HMM is said to be **state-based**, if a special condition

$$p(y_t|s_{t-1}, s_t) = p(y_t|s_t)$$ \hspace{1cm} (20.24)

holds. A state-based HMM is defined by

$$\theta = (\pi_0, A, B),$$ \hspace{1cm} (20.25)

where $\pi_0$ and $A$ are defined in (20.3) and (20.7), respectively, and $B$ is an $M \times K$ matrix given by

$$B \triangleq [b(j,k)]; \ j \in S, \ k \in \mathcal{Y},$$ \hspace{1cm} (20.26)

where

$$b(j,k) = P[Y_t = k|S_t = j].$$ \hspace{1cm} (20.27)

If the special condition (20.24) does not hold, the HMM is said to be **transition-based**.
Note that state-based HMMs form a subclass of transition-based HMMs.

Most existing literature on HMMs, however, deal with only state-based HMMs.

But, there is an advantage in working with a transition-based HMM.

Thus, we assume a general transition-based HMM, unless stated otherwise.

**The case where the observable \( Y_t \) is a continuous random variable.**

We should replace \( c(i; j, k) \) of the transition-based HMM by

\[
p_{S_t, Y_t|S_{t-1}}(j, y|i) \, dy = P[S_t = j, y < Y_t \leq y + dy|S_{t-1} = i]. \tag{20.31}
\]

Similarly, \( b(j; k) \) should be replaced by \( f_{Y_t|S_t}(y|j) \, dy \):

\[
f_{Y_t|S_t}(y|j) \, dy = P[y < Y_t \leq y + dy|S_t = j]. \tag{20.32}
\]

The restrictive condition (20.24) for the state-based model is

\[
p_{S_t, Y_t|S_{t-1}}(j, y|i) = a(i, j) \, f_{Y_t|S_t}(y|j). \tag{20.33}
\]
20.2.2 Examples of Hidden Markov Model

Example 20.1: Convolutional encoder and binary symmetric channel

\[ O_t = O_t^{(1)} O_t^{(2)}, \quad t \geq 0, \quad (20.34) \]

where \[ O_t^{(1)} = I_t \oplus I_{t-1} \oplus I_{t-2}, \quad \text{and} \quad O_t^{(2)} = I_t \oplus I_{t-2}, \quad (20.35) \]

\[ I_{-2} = I_{-1} = I_0 = 0. \quad (20.36) \]

\[ 0 \oplus 0 = 0, 0 \oplus 1 = 1, 1 \oplus 0 = 1 \quad \text{and} \quad 1 \oplus 1 = 0. \]
A natural choice of the state for this encoder would be

\[ S_t = I_t I_{t-1} I_{t-2}, \quad t = 0, 1, 2, \ldots. \]

We define, instead, state \( S_t \) as

\[ S_t = I_t I_{t-1}, \quad t = 0, 1, 2, \ldots. \tag{20.37} \]

Then, the state space is

\[ S = \{00, 01, 10, 11\} = \{0, 1, 2, 3\}. \]

If we assume that \( \{I_t\} \) is an i.i.d. sequence, then the state sequence \( \{S_t\} \) is a simple Markov chain
Figure 20.3  (a) The state transition diagram of the rate 1/2 convolutional encoder.  
(b) The trellis diagram (all transitions have probability 1/2).
If 0 and 1 appear in $I_t$ with equal probability, i.e., $P[I_t = 0] = P[I_t = 1] = 1/2$ for all $t$ the state transition matrix associated with the convolutional encoder is

$$A \equiv \begin{bmatrix}
1/2 & 0 & 1/2 & 0 \\
1/2 & 0 & 1/2 & 0 \\
0 & 1/2 & 0 & 1/2 \\
0 & 1/2 & 0 & 1/2
\end{bmatrix}. \quad (20.38)$$

**Discrete Memoryless Channel:**

$$Y_t = O_t \oplus E_t, \quad (20.39)$$

We assume that the channel output $Y_t$ has the same alphabet as the channel input (the encoder output) $O_t$, i.e.,

$$\mathcal{Y} = \mathcal{O} = \{00, 01, 10, 11\}.$$ 

where $E_t$ is called an error pattern, meaning a “1” in $E_t$ will result in an error
We assume that errors at different times occur independently of each other.

\[ p(\mathbf{y}_0^T | \mathbf{o}_0^T) = \prod_{t=0}^{T} p(y_t | o_t) \quad (20.40) \]

Such a channel is referred to as a discrete memoryless channel. Since the channel input \( O_t \) is a function of the encoder states \( S_{t-1} \) and \( S_t \) only, the memoryless property implies that \( Y_t \) depends probabilistically only on \( S_{t-1} \) and \( S_t \). \( X_t = (S_t, Y_t) \) depends only on \( S_{t-1} \), hence it is a hidden Markov process.

A discrete memory channel is called a binary symmetric channel (BSC) if

\[ p(1|0) = p(0|1) = \epsilon \]

Thus, the model parameter of this HMM is given by \( \theta = C = (A, \epsilon) \).