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

ELE 525: Random Processes in Information Systems

Hisashi Kobayashi

Department of Electrical Engineering
Princeton University
December 11, 2013

Textbook: Hisashi Kobayashi, Brian L. Mark and William Turin, ***Probability, Random Processes and Statistical Analysis*** (Cambridge University Press, 2012)

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**

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

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 θ, 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 θ is fixed.

If we substitute $\boldsymbol{y} = T(\boldsymbol{x})$,

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

Then

$$p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}) = \frac{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})}{p_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})} = \frac{p_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})}{p_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}, \tag{19.19}$$

from which we obtain the log-likelihood function $\log L_{\boldsymbol{y}}(\boldsymbol{\theta}) = \log p_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})$ as

$$\log L_{\boldsymbol{y}}(\boldsymbol{\theta}) = \log p_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta}) - \log p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}). \tag{19.20}$$

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

$$\log L_{\boldsymbol{y}}(\boldsymbol{X};\boldsymbol{\theta}) = \log p_{\boldsymbol{X}}(\boldsymbol{X};\boldsymbol{\theta}) - \log p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta}). \tag{19.21}$$

Now we take expectations on both sides of (19.21) with respect to $p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta}^{(p)})$, and note $E\left[\log L_{\boldsymbol{y}}(\boldsymbol{X};\boldsymbol{\theta})|\boldsymbol{y};\boldsymbol{\theta}^{(p)}\right] = \log L_{\boldsymbol{y}}(\boldsymbol{\theta})$.

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

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

The last equation can be written as

$$\boxed{\log L_{\boldsymbol{y}}(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) + H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}),} \tag{19.23}$$

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

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

and

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

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

$$\boxed{H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) \geq H(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}^{(p)}) = \mathcal{H}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta}^{(p)}),} \tag{19.26}$$

where $\mathcal{H}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta}^{(p)})$ is the **conditional entropy**

Relationship between the log-likelihood function and Q-function

The auxiliary function defined in (19.24)

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

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

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

$$\boxed{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) - Q(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}^{(p)}) \leq \log L_{\boldsymbol{y}}(\boldsymbol{\theta}) - \log L_{\boldsymbol{y}}(\boldsymbol{\theta}^{(p)}),} \tag{19.27}$$

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

Equation (19.26)

$$\boxed{H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) \geq H(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}^{(p)}) = \mathcal{H}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta}^{(p)}),} \tag{19.26}$$

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

Then differentiating both sides of (19.23)

$$\boxed{\log L_{\boldsymbol{y}}(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) + H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}),} \tag{19.23}$$

we obtain

$$\nabla_{\boldsymbol{\theta}} \log L_{\boldsymbol{y}}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(p)}} = \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})|\boldsymbol{\theta}^{(p)})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(p)}}. \tag{19.28}$$

Derivation of the E-step and the M-step

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

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

$$\boxed{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) = E\left[\log p_{\boldsymbol{X}}(\boldsymbol{X};\boldsymbol{\theta})|\,\boldsymbol{y},\boldsymbol{\theta}^{(p)}\right]. \quad \textbf{(E-step)}} \tag{19.29}$$

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

The inequality (19.27)

$$\boxed{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) - Q(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}^{(p)}) \leq \log L_{\boldsymbol{y}}(\boldsymbol{\theta}) - \log L_{\boldsymbol{y}}(\boldsymbol{\theta}^{(p)}),} \tag{19.27}$$

implies that if $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) > Q(\boldsymbol{\theta}^{(p)},\boldsymbol{\theta}^{(p)})$, then $L_{\boldsymbol{y}}(\boldsymbol{\theta}) > L_{\boldsymbol{y}}(\boldsymbol{\theta}^{(p)})$.

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

$$\boxed{\boldsymbol{\theta}^{(p+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}). \qquad \textbf{(M-step)}} \tag{19.30}$$

If $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)})$ is differentiable, the next estimate $\boldsymbol{\theta}^{(p+1)}$ satisfies

$$\nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(p+1)}} = \mathbf{0}.$$

Thus, if the algorithm converges to $\boldsymbol{\theta}^*$, $\nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \mathbf{0}$, which implies

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \mathbf{0}. \tag{19.31}$$

according to (19.28)

$$\nabla_{\boldsymbol{\theta}} \log L_{\boldsymbol{y}}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(p)}} = \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})|\boldsymbol{\theta}^{(p)})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(p)}}. \tag{19.28}$$

Thus, the algorithm converges to the *stationary point* of the likelihood function.

Therefore, an MLE $\hat{\boldsymbol{\theta}}$ 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 $\boldsymbol{\theta}^{(0)}$. Set the iteration number $p = 0$.

2: Assume the p-th estimate $\boldsymbol{\theta}^{(p)}$.

3: Evaluate

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

4: Find

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

5: If any of the stopping conditions is met, go to step 6. Otherwise, replace $\boldsymbol{\theta}^{(p)}$ by $\boldsymbol{\theta}^{(p+1)}$, set $p \leftarrow p + 1$, and repeat the Steps 2 through 5.

6: Output $\boldsymbol{\theta}^{(p+1)}$ as an MLE.

Geometrical interpretation of the EM algorithm

If we apply inequality (19.26)

$$\boxed{H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) \geq H(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}^{(p)}) = \mathcal{H}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta}^{(p)}),} \tag{19.26}$$

to

$$\boxed{\log L_{\boldsymbol{y}}(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) + H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}),} \tag{19.23}$$

we obtain

$$\log L_{\boldsymbol{y}}(\boldsymbol{\theta}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) + H(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}^{(p)}) \triangleq B(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}). \tag{19.32}$$

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

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

$$\boxed{B(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}^{(p)}) = Q(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}^{(p)}) + H(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}^{(p)}) = \log L_{\boldsymbol{y}}(\boldsymbol{\theta}^{(p)})}, \tag{19.33}$$

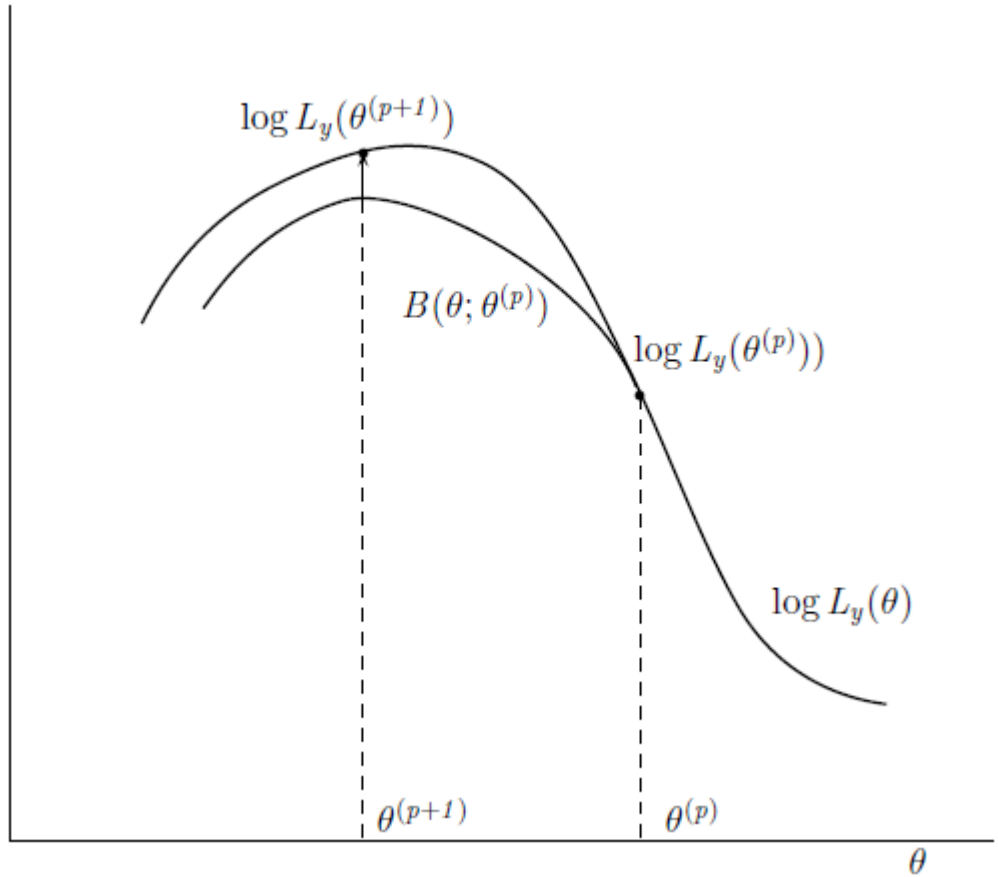

Figure 19.1 The M-step of the EM algorithm

Recall the definition of the B-function:

$$\log L_{\boldsymbol{y}}(\boldsymbol{\theta}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) + H(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}^{(p)}) \triangleq B(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}). \tag{19.32}$$

According to equation (19.28),

$$\nabla_{\boldsymbol{\theta}} \log L_{\boldsymbol{y}}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(p)}} = \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})|\boldsymbol{\theta}^{(p)})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(p)}}. \tag{19.28}$$

we have

$$\nabla_{\boldsymbol{\theta}} \log L_{\boldsymbol{y}}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(p)}} = \nabla_{\boldsymbol{\theta}} B(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(p)}}, \tag{19.34}$$

thus, $\log L_{\boldsymbol{y}}(\boldsymbol{\theta})$ and its *lower bound* $B(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)})$ have a common tangential plane

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

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

$$\log L_{\boldsymbol{y}}(\boldsymbol{\theta}^{(p)}) = B(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}^{(p)}) \leq B(\boldsymbol{\theta}^{(p+1)}|\boldsymbol{\theta}^{(p)}) \leq \log L_{\boldsymbol{y}}(\boldsymbol{\theta}^{(p+1)}). \tag{19.35}$$

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 $L_{\boldsymbol{y}}(\boldsymbol{\theta}^{(p)})$ increases in every step because of (19.35)

$$\log L_{\boldsymbol{y}}(\boldsymbol{\theta}^{(p)}) = B(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}_1^{(p)}) \leq B(\boldsymbol{\theta}^{(p+1)}|\boldsymbol{\theta}^{(p)}) \leq \log L_{\boldsymbol{y}}(\boldsymbol{\theta}^{(p+1)}). \tag{19.35}$$

or (19.27),

$$\boxed{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) - Q(\boldsymbol{\theta}^{(p)}|\boldsymbol{\theta}^{(p)}) \leq \log L_{\boldsymbol{y}}(\boldsymbol{\theta}) - \log L_{\boldsymbol{y}}(\boldsymbol{\theta}^{(p)}),} \tag{19.27}$$

In order to reach a global maximum, we must carefully choose the initial estimate $\boldsymbol{\theta}^{(0)}$ 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 $p_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta})$ may take a simpler form when the probability distribution belongs to the *exponential family*

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

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

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

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 $\boldsymbol{X}$ can be written as

$$\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{Z}), \tag{19.36}$$

where $\boldsymbol{Y}$ is the **observable variable** and $\boldsymbol{Z}$ is a **latent variable**.

$$T(x) = T(\boldsymbol{y}, \boldsymbol{z}) = \boldsymbol{y}. \tag{19.37}$$

An instance x is called **complete data**;

$\boldsymbol{y}$, *observed data* or *incomplete data*,

$\boldsymbol{z}$, **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 $\boldsymbol{Q}$-function is defined, similar to (19.24)

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

by

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

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

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

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

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

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 $\boldsymbol{\theta}$ is unknown but fixed.

If we have a *prior probability* distribution $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ of the parameter variable $\boldsymbol{\Theta}$, then we should take the Bayesian approach, i.e., find the MAP estimate:

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

According to equation (18.102)

$$\hat{\boldsymbol{\theta}}^{\star}(\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \; [\log L_{\boldsymbol{x}}(\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta})]. \tag{18.102}$$

the MAP estimate can be obtained by

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} \; [\log L_{\boldsymbol{y}}(\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta})]. \tag{19.40}$$

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

$$\boxed{\boldsymbol{\theta}^{(p+1)} = \arg\max_{\boldsymbol{\theta}} \left[Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) + \log p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \right].} \tag{19.41}$$

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)

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

can be approximated, by generating $N^{(p)}$ samples, $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_t, \ldots, \boldsymbol{z}_{N^{(p)}}$ from the conditional distribution $p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}^{(p)})$ and then by substituting the empirical average for the conditional expectation:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(p)}) \approx \frac{1}{N^{(p)}} \sum_{t=1}^{N^{(p)}} \log p(\boldsymbol{y}, \boldsymbol{z}_t; \boldsymbol{\theta}). \tag{19.43}$$

Maximization of this function with respect to $\boldsymbol{\theta}$ 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 $\{0, 1, \ldots, M-1\}$

$$\mathcal{S} \triangleq \{0, 1, \ldots, M-1\}. \tag{20.1}$$

The subscript t of S_t is the discrete-time index

$$t \in \{0, 1, 2, \ldots, T\} \triangleq \mathcal{T}. \tag{20.2}$$

Let the *initial state distribution* be denoted by

$$\boldsymbol{\pi}_0 = (\pi_0(0), \pi_0(1), \ldots, \pi_0(M-1)) \triangleq (\pi_0(i); i \in \mathcal{S}), \tag{20.3}$$

$$\text{where } \pi_0(i) = P[S_0 = i], \quad i \in \mathcal{S}. \tag{20.4}$$

In a Markov chain

$$P[S_t = s_t | S_{t-1} = s_{t-1}, S_{t-2} = s_{t-2}, \ldots] = P[S_t = s_t | S_{t-1} = s_{t-1}], \tag{20.5}$$

or more compactly,

$$p(s_t | s_{t-1}, s_{t-2}, \ldots) = p(s_t | s_{t-1}). \tag{20.6}$$

Thus, the (one-step) **state transition probabilities**[3] $\{a(i,j)\}$

$$\boldsymbol{A} \triangleq [a(i,j);\ \ i,j \in \mathcal{S}], \tag{20.7}$$

$$\text{where } \ a(i,j) \triangleq P[S_t = j \,|S_{t-1} = i],\ \ i,j \in \mathcal{S}. \tag{20.8}$$

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

In the figure of next slide, we show the **state transition diagram** of time-homogenous Markov chain, and its **trellis diagram**.

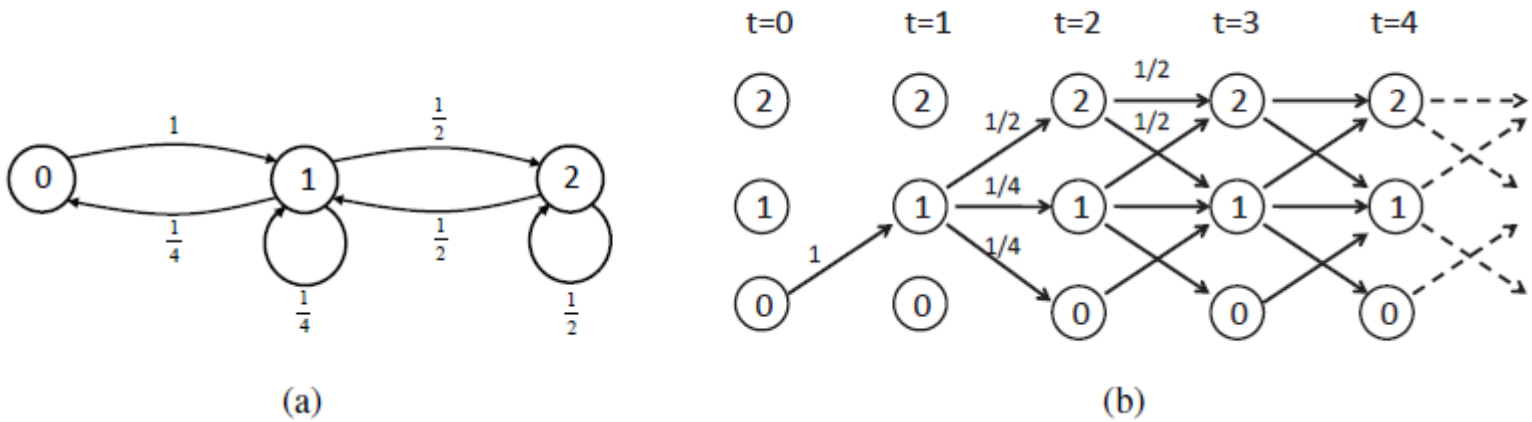

Figure 20.1 (a) The state transition diagram of a 3-state Markov chain $\mathcal{S} = \{0, 1, 2\}$, and (b) its trellis diagram.

We denote the state sequence in the period $\mathcal{T}$ defined in (20.2)

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

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

For every instance of the state sequence $\boldsymbol{s}$, 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

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

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

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

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

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

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.,*

$$p(x_t|x_{t-1}) = p(x_t|s_{t-1}), \tag{20.12}$$

i.e.,

$$p(s_t, y_t|s_{t-1}, y_{t-1}) = p(s_t, y_t|s_{t-1}), \tag{20.13}$$

Any state $S_t \in \mathcal{S}$ *is called a* hidden state *and the process* $\boldsymbol{S} = (S_t;\ t \in \mathcal{T})$ *is a hidden process; and* $\boldsymbol{y} = (y_t;\ t \in \mathcal{T})$ *is called an* observation. □

Definition 20.2 (Model parameters of a HMM). *For a homogeneous HMM, we denote*

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

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

Denote the sets *of initial distribution vectors* $\boldsymbol{\alpha}_0(k)$*'s and the* state transition probability matrices $\boldsymbol{C}(k)$*'s by* $\boldsymbol{\alpha}_0$ *and* $\mathbf{C}$*, respectively:*

$$\boldsymbol{\alpha}_0 \triangleq (\boldsymbol{\alpha}_0(k); \quad k \in \mathcal{Y}), \tag{20.16}$$

$$\textit{where} \quad \boldsymbol{\alpha}_0(k) \triangleq (\alpha_0(i,k); \; i \in \mathcal{S}), \quad k \in \mathcal{Y}, \tag{20.17}$$

$$\mathbf{C} \triangleq (\boldsymbol{C}(k); k \in \mathcal{Y}), \tag{20.18}$$

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

We then define the ***model parameters*** $\boldsymbol{\theta}$ *associated with the HMM by the set*

$$\boldsymbol{\theta} = (\boldsymbol{\alpha}_0, \mathbf{C}). \tag{20.20}$$

□

Note that

$$\sum_{k \in \mathcal{Y}} \boldsymbol{\alpha}_0(k) = \boldsymbol{\pi}_0. \tag{20.21}$$

$$\sum_{k \in \mathcal{Y}} \boldsymbol{C}(k) = \boldsymbol{A}. \tag{20.22}$$

Now we can rewrite $p(s_t, y_t|s_{t-1})$ of (20.13) as

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

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*

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

holds. A state-based HMM is defined by

$$\boldsymbol{\theta} = (\boldsymbol{\pi}_0, \boldsymbol{A}, \boldsymbol{B}), \tag{20.25}$$

where $\boldsymbol{\pi}_0$ *and* $\boldsymbol{A}$ *are defined in* (20.3) *and* (20.7)*, respectively, and* $\boldsymbol{B}$ *is an* $M \times K$ *matrix given by*

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

where

$$\text{where}\ \ b(j;k) = P[Y_t = k|S_t = j]. \tag{20.27}$$

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

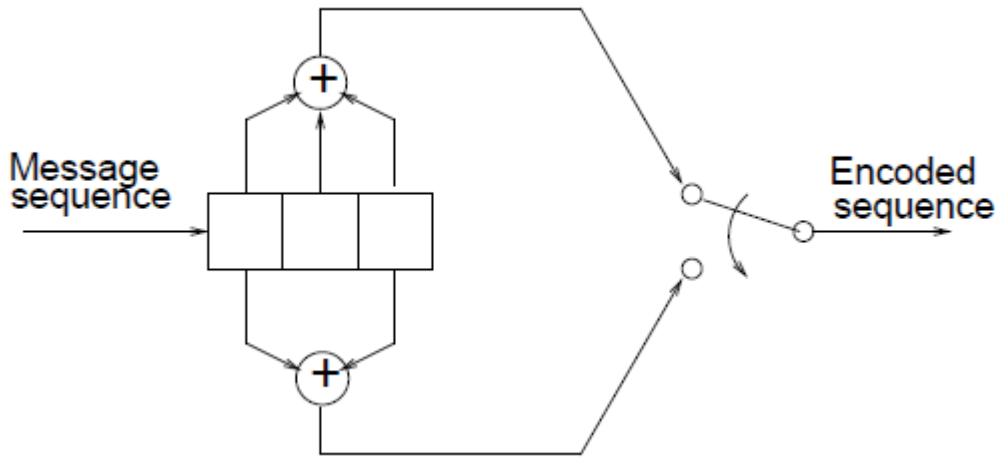

Figure 20.2 A rate 1/2 convolutional encoder with input I_t and output O_t

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

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

$$I_{-2} = I_{-1} = I_0 = 0. \tag{20.36}$$

$$0 \oplus 0 = 0, 0 \oplus 1 = 1, 1 \oplus 0 = 1 \text{ and } 1 \oplus 1 = 0.$$

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

$$\mathcal{S} = \{00, 01, 10, 11\} = \{0, 1, 2, 3\}.$$

If we assume that $\{I_t\}$ is an i.i.d. sequence, then the state sequence $\{S_t\}$ is a simple Markov chain

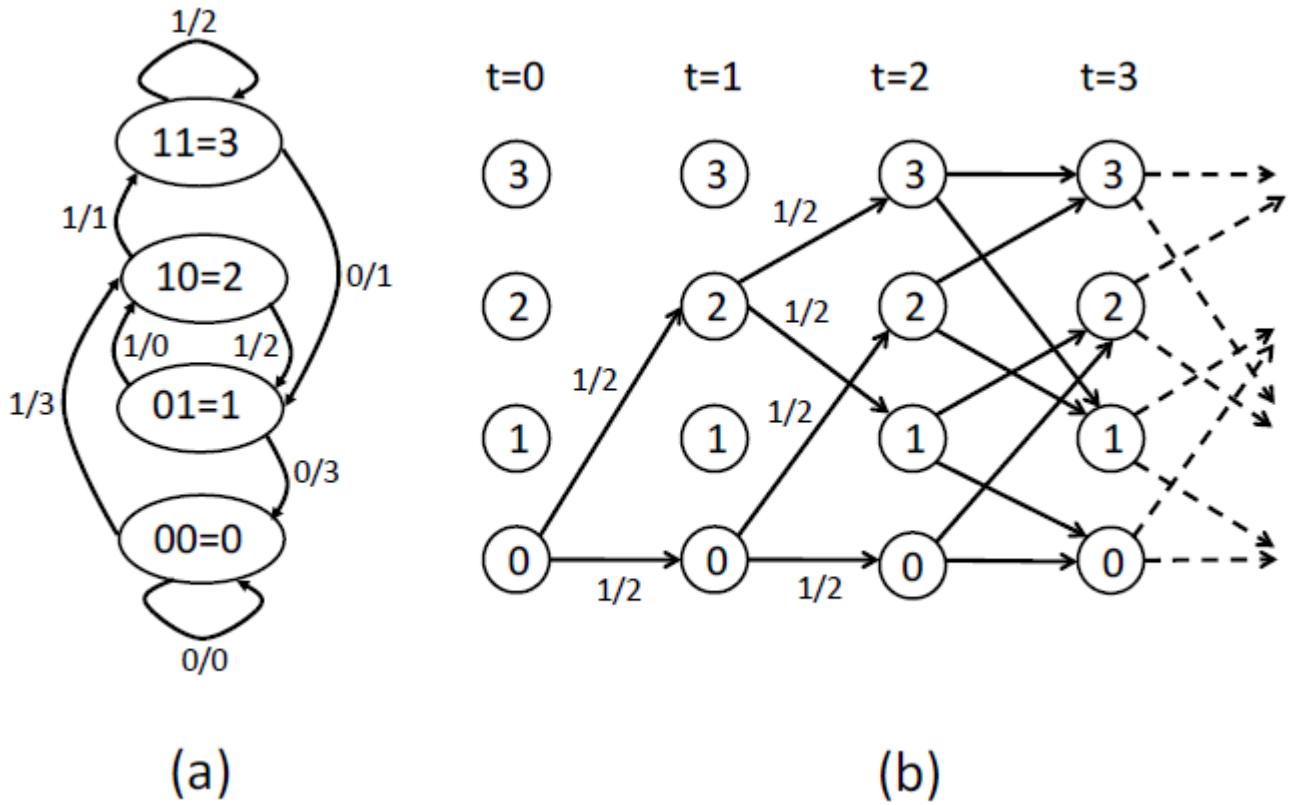

Figure 20.3 (a) The state transition diagram of the rate 1/2 convolutional encoder. (b) The trellis diagram (all transitions have probability $1/2$).

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

$$\boldsymbol{A} \triangleq \begin{bmatrix} 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 \end{bmatrix}. \tag{20.38}$$

Discrete Memoryless Channel:

$$Y_t = O_t \oplus E_t, \tag{20.39}$$

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.

$$p(\boldsymbol{y}_0^T|\boldsymbol{o}_0^T) = \prod_{t=0}^{T} p(y_t|o_t) \tag{20.40}$$

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

$$p(1|0)=p(0|1)=\varepsilon$$

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