SciencePG
Science Publishing Group

Tutorial on Hidden Markov Model

Loc Nguyen

Sunflower Soft Company, Ho Chi Minh city, Vietnam

Email address:
ng_phloc@yahoo.com

Abstract: Hidden Markov model (HMM) is a powerful mathematical tool for prediction and recognition. Many computer software products implement HMM and hide its complexity, which assist scientists to use HMM for applied researches. However comprehending HMM in order to take advantages of its strong points requires a lot of efforts. This report is a tutorial on HMM with full of mathematical proofs and example, which help researchers to understand it by the fastest way from theory to practice. The report focuses on three common problems of HMM such as evaluation problem, uncovering problem, and learning problem, in which learning problem with support of optimization theory is the main subject.

Keywords: Hidden Markov Model, Optimization, Evaluation Problem, Uncovering Problem, Learning Problem

1. Introduction

There are many real-world phenomena (so-called states) that we would like to model in order to explain our observations. Often, given sequence of observations symbols, there is demand of discovering real states. For example, there are some states of weather: *sunny*, *cloudy*, *rainy* [1, p. 1]. Suppose you are in the room and do not know the weather outside but you are notified observations such as wind speed, atmospheric pressure, humidity, and temperature from someone else. Basing on these observations, it is possible for you to forecast the weather by using hidden Markov model (HMM). Before discussing about HMM, we should glance over the definition of Markov model (MM). First, MM is the statistical model which is used to model the stochastic process. MM is defined as below [2]:

- Given a finite set of state $S=\{s_1, s_2,\ldots, s_n\}$ whose cardinality is n. Let $\prod$ be the *initial state distribution* where $\pi_i \in \prod$ represents the probability that the stochastic process begins in state s_i. In other words π_i is the initial probability of state s_i, where

$$\sum_{s_i \in S} \pi_i = 1$$

- The stochastic process which is modeled gets only one state from S at all time points. This stochastic process is defined as a finite vector $X=(x_1, x_2,\ldots, x_T)$ whose element x_t is a state at time point t. The process X is called *state stochastic process* and $x_t \in S$ equals some state $s_i \in S$. Note that X is also called *state sequence*. Time point can be in terms of second, minute, hour, day, month, year, etc. It is easy to infer that the initial probability $\pi_i = P(x_1=s_i)$ where x_1 is the first state of the stochastic process. The state stochastic process X must meet fully the *Markov property*, namely, given previous state x_{t-1} of process X, the conditional probability of current state x_t is only dependent on the previous state x_{t-1}, not relevant to any further past state $(x_{t-2}, x_{t-3},\ldots, x_1)$. In other words, $P(x_t \mid x_{t-1}, x_{t-2}, x_{t-3},\ldots, x_1) = P(x_t \mid x_{t-1})$ with note that $P(.)$ also denotes probability in this report. Such process is called first-order Markov process.

- At time point, the process changes to the next state based on the *transition probability distribution a_{ij}*, which depends only on the previous state. So a_{ij} is the probability that the stochastic process changes current state s_i to next state s_j. It means that $a_{ij} = P(x_t=s_j \mid x_{t-1}=s_i) = P(x_{t+1}=s_j \mid x_t=s_i)$. The probability of transitioning from any given state to some next state is 1, we have

$$\forall s_i \in S, \sum_{s_j \in S} a_{ij} = 1$$

All transition probabilities a_{ij} (s) constitute the *transition probability matrix A*. Note that A is n by n matrix because there are n distinct states. It is easy to infer that matrix A represents state stochastic process X. It is possible to

understand that the initial probability matrix $\prod$ is degradation case of matrix A.

Briefly, MM is the triple $\langle S, A, \prod \rangle$. In typical MM, states are observed directly by users and transition probabilities (A and $\prod$) are unique parameters. Otherwise, hidden Markov model (HMM) is similar to MM except that the underlying states become hidden from observer, they are hidden parameters. HMM adds more output parameters which are called observations. Each state (hidden parameter) has the conditional probability distribution upon such observations. HMM is responsible for discovering hidden parameters (states) from output parameters (observations), given the stochastic process. The HMM has further properties as below [2]:

- Suppose there is a finite set of possible observations $\Phi = \{\varphi_1, \varphi_2,\ldots, \varphi_m\}$ whose cardinality is m. There is the second stochastic process which produces *observations* correlating with hidden states. This process is called *observable stochastic process*, which is defined as a finite vector $O = (o_1, o_2,\ldots, o_T)$ whose element o_t is an observation at time point t. Note that $o_t \in \Phi$ equals some φ_k. The process O is often known as *observation sequence*.

- There is a probability distribution of producing a given observation in each state. Let $b_i(k)$ be the probability of observation φ_k when the state stochastic process is in state s_i. It means that $b_i(k) = b_i(o_t=\varphi_k) = P(o_t=\varphi_k \mid x_t=s_i)$. The sum of probabilities of all observations which observed in a certain state is 1, we have

$$\forall s_i \in S, \sum_{\theta_k \in \Phi} b_i(k) = 1$$

All probabilities of observations $b_i(k)$ constitute the *observation probability matrix B*. It is convenient for us to use notation b_{ik} instead of notation $b_i(k)$. Note that B is n by m matrix because there are n distinct states and m distinct observations. While matrix A represents state stochastic process X, matrix B represents observable stochastic process O.

Thus, HMM is the 5-tuple $\Delta = \langle S, \Phi, A, B, \prod \rangle$. Note that components S, Φ, A, B, and $\prod$ are often called parameters of HMM in which A, B, and $\prod$ are essential parameters. Going back weather example, suppose you need to predict how weather tomorrow is: *sunny*, *cloudy* or *rainy* since you know only observations about the humidity: *dry*, *dryish*, *damp*, *soggy*. The HMM is totally determined based on its parameters S, Φ, A, B, and $\prod$ according to weather example. We have $S = \{s_1=sunny, s_2=cloudy, s_3=rainy\}$, $\Phi = \{\varphi_1=dry, \varphi_2=dryish, \varphi_3=damp, \varphi_4=soggy\}$. Transition probability matrix A is shown in table 1.

Table 1. *Transition probability matrix A.*

		Weather current day (Time point t)		
		sunny	***cloudy***	***rainy***
Weather previous day (Time point $t-1$)	*sunny*	a_{11}=0.50	a_{12}=0.25	a_{13}=0.25
	cloudy	a_{21}=0.30	a_{22}=0.40	a_{23}=0.30
	rainy	a_{31}=0.25	a_{32}=0.25	a_{33}=0.50

From table 1, we have $a_{11}+a_{12}+a_{13}=1$, $a_{21}+a_{22}+a_{23}=1$, $a_{31}+a_{32}+a_{33}=1$.

Initial state distribution specified as uniform distribution is shown in table 2.

Table 2. *Uniform initial state distribution Π.*

sunny	cloudy	rainy
π_1=0.33	π_2=0.33	π_3=0.33

From table 2, we have $\pi_1+\pi_2+\pi_3=1$.

Observation probability matrix B is shown in table 3.

Table 3. *Observation probability matrix B.*

		Humidity			
		dry	***dryish***	***damp***	***soggy***
Weather	*sunny*	b_{11}=0.60	b_{12}=0.20	b_{13}=0.15	b_{14}=0.05
	cloudy	b_{21}=0.25	b_{22}=0.25	b_{23}=0.25	b_{24}=0.25
	rainy	b_{31}=0.05	b_{32}=0.10	b_{33}=0.35	b_{34}=0.50

From table 3, we have $b_{11}+b_{12}+b_{13}+b_{14}=1$, $b_{21}+b_{22}+b_{23}+b_{24}=1$, $b_{31}+b_{32}+b_{33}+b_{34}=1$.

The whole weather HMM is depicted in fig. 1.
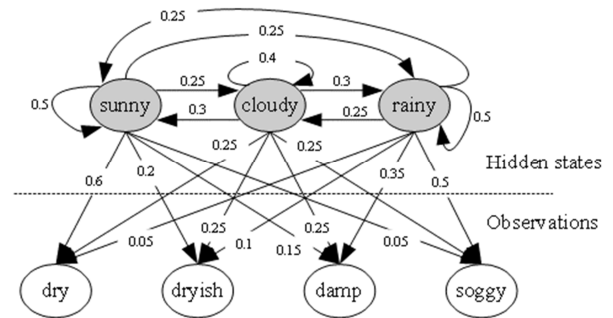
Figure 1. *HMM of weather forecast (hidden states are shaded).*

There are three problems of HMM [2] [3, pp. 262-266]:

1. Given HMM Δ and an observation sequence $O = \{o_1, o_2,\ldots, o_T\}$ where $o_t \in \Phi$, how to calculate the probability $P(O|\Delta)$ of this observation sequence. Such probability $P(O|\Delta)$ indicates how much the HMM Δ affects on sequence O. This is *evaluation problem* or *explanation problem*. Note that it is possible to denote $O = \{o_1 \rightarrow o_2 \rightarrow\ldots\rightarrow o_T\}$ and the sequence O is aforementioned observable stochastic process.

2. Given HMM Δ and an observation sequence $O = \{o_1, o_2,\ldots, o_T\}$ where $o_t \in \Phi$, how to find the sequence of states $X = \{x_1, x_2,\ldots, x_T\}$ where $x_t \in S$ so that X is most likely to have produced the observation sequence O. This is *uncovering problem*. Note that the sequence X is aforementioned state stochastic process.

3. Given HMM Δ and an observation sequence $O = \{o_1, o_2,\ldots, o_T\}$ where $o_t \in \Phi$, how to adjust parameters of Δ such as initial state distribution $\prod$, transition probability matrix A, and observation probability matrix B so that the quality of HMM Δ is enhanced. This is *learning problem*.

These problems will be mentioned in sections 2, 3, and 4, in turn.

2. HMM Evaluation Problem

The essence of evaluation problem is to find out the way to compute the probability $P(O|\Delta)$ most effectively given the observation sequence $O = \{o_1, o_2,..., o_T\}$. For example, given HMM Δ whose parameters A, B, and $\prod$ specified in tables 1, 2, and 3, which is designed for weather forecast. Suppose we need to calculate the probability of event that humidity is *soggy* and *dry* in days 1 and 2, respectively. This is evaluation problem with sequence of observations $O = \{o_1=\varphi_4=soggy,\ o_2=\varphi_1=dry,\ o_3=\varphi_2=dryish\}$. There is a complete set of $3^3=27$ mutually exclusive cases of weather states for three days: $\{x_1=s_1=sunny,\ x_2=s_1=sunny,\ x_3=s_1=sunny\}$, $\{x_1=s_1=sunny,\ x_2=s_1=sunny,\ x_3=s_2=cloudy\}$, $\{x_1=s_1=sunny,\ x_2=s_1=sunny,\ x_3=s_3=rainy\}$, $\{x_1=s_1=sunny,\ x_2=s_2=cloudy,\ x_3=s_1=sunny\}$, $\{x_1=s_1=sunny,\ x_2=s_2=cloudy,\ x_3=s_2=cloudy\}$, $\{x_1=s_1=sunny,\ x_2=s_2=cloudy,\ x_3=s_3=rainy\}$, $\{x_1=s_1=sunny,\ x_2=s_3=rainy,\ x_3=s_1=sunny\}$, $\{x_1=s_1=sunny,\ x_2=s_3=rainy,\ x_3=s_2=cloudy\}$, $\{x_1=s_1=sunny,\ x_2=s_3=rainy,\ x_3=s_3=rainy\}$, $\{x_1=s_2=cloudy,\ x_2=s_1=sunny,\ x_3=s_1=sunny\}$, $\{x_1=s_2=cloudy,\ x_2=s_1=sunny,\ x_3=s_2=cloudy\}$, $\{x_1=s_2=cloudy,\ x_2=s_1=sunny,\ x_3=s_3=rainy\}$, $\{x_1=s_2=cloudy,\ x_2=s_2=cloudy,\ x_3=s_1=sunny\}$, $\{x_1=s_2=cloudy,\ x_2=s_2=cloudy,\ x_3=s_2=cloudy\}$, $\{x_1=s_2=cloudy,\ x_2=s_2=cloudy,\ x_3=s_3=rainy\}$, $\{x_1=s_2=cloudy,\ x_2=s_3=rainy,\ x_3=s_1=sunny\}$, $\{x_1=s_2=cloudy,\ x_2=s_3=rainy,\ x_3=s_2=cloudy\}$, $\{x_1=s_2=cloudy,\ x_2=s_3=rainy,\ x_3=s_3=rainy\}$, $\{x_1=s_3=rainy,\ x_2=s_1=sunny,\ x_3=s_1=sunny\}$, $\{x_1=s_3=rainy,\ x_2=s_1=sunny,\ x_3=s_2=cloudy\}$, $\{x_1=s_3=rainy,\ x_2=s_1=sunny,\ x_3=s_3=rainy\}$, $\{x_1=s_3=rainy,\ x_2=s_2=cloudy,\ x_3=s_1=sunny\}$, $\{x_1=s_3=rainy,\ x_2=s_2=cloudy,\ x_3=s_2=cloudy\}$, $\{x_1=s_3=rainy,\ x_2=s_2=cloudy,\ x_3=s_3=rainy\}$, $\{x_1=s_3=rainy,\ x_2=s_3=rainy,\ x_3=s_1=sunny\}$, $\{x_1=s_3=rainy,\ x_2=s_3=rainy,\ x_3=s_2=cloudy\}$, $\{x_1=s_3=rainy,\ x_2=s_3=rainy,\ x_3=s_3=rainy\}$.

According to total probability rule [4, p. 101], the probability $P(O|\Delta)$ is:

$$P(O|\Delta) = P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_1, x_2 = s_1, x_3 = s_1)$$
$$* P(x_1 = s_1, x_2 = s_1, x_3 = s_1)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_1, x_2 = s_1, x_3 = s_2)$$
$$* P(x_1 = s_1, x_2 = s_1, x_3 = s_2)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_1, x_2 = s_1, x_3 = s_3)$$
$$* P(x_1 = s_1, x_2 = s_1, x_3 = s_3)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_1, x_2 = s_2, x_3 = s_1)$$
$$* P(x_1 = s_1, x_2 = s_2, x_3 = s_1)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_1, x_2 = s_2, x_3 = s_2)$$
$$* P(x_1 = s_1, x_2 = s_2, x_3 = s_2)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_1, x_2 = s_2, x_3 = s_3)$$
$$* P(x_1 = s_1, x_2 = s_2, x_3 = s_3)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_1, x_2 = s_3, x_3 = s_1)$$
$$* P(x_1 = s_1, x_2 = s_3, x_3 = s_1)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_1, x_2 = s_3, x_3 = s_2)$$
$$* P(x_1 = s_1, x_2 = s_3, x_3 = s_2)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_1, x_2 = s_3, x_3 = s_3)$$
$$* P(x_1 = s_1, x_2 = s_3, x_3 = s_3)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_2, x_2 = s_1, x_3 = s_1)$$
$$* P(x_1 = s_2, x_2 = s_1, x_3 = s_1)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_2, x_2 = s_1, x_3 = s_2)$$
$$* P(x_1 = s_2, x_2 = s_1, x_3 = s_2)$$

$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_2, x_2 = s_1, x_3 = s_3)$$
$$* P(x_1 = s_2, x_2 = s_1, x_3 = s_3)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_2, x_2 = s_2, x_3 = s_1)$$
$$* P(x_1 = s_2, x_2 = s_2, x_3 = s_1)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_2, x_2 = s_2, x_3 = s_2)$$
$$* P(x_1 = s_2, x_2 = s_2, x_3 = s_2)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_2, x_2 = s_2, x_3 = s_3)$$
$$* P(x_1 = s_2, x_2 = s_2, x_3 = s_3)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_2, x_2 = s_3, x_3 = s_1)$$
$$* P(x_1 = s_2, x_2 = s_3, x_3 = s_1)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_2, x_2 = s_3, x_3 = s_2)$$
$$* P(x_1 = s_2, x_2 = s_3, x_3 = s_2)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_2, x_2 = s_3, x_3 = s_3)$$
$$* P(x_1 = s_2, x_2 = s_3, x_3 = s_3)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_3, x_2 = s_1, x_3 = s_1)$$
$$* P(x_1 = s_3, x_2 = s_1, x_3 = s_1)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_3, x_2 = s_1, x_3 = s_2)$$
$$* P(x_1 = s_3, x_2 = s_1, x_3 = s_2)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_3, x_2 = s_1, x_3 = s_3)$$
$$* P(x_1 = s_3, x_2 = s_1, x_3 = s_3)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_3, x_2 = s_2, x_3 = s_1)$$
$$* P(x_1 = s_3, x_2 = s_2, x_3 = s_1)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_3, x_2 = s_2, x_3 = s_2)$$
$$* P(x_1 = s_3, x_2 = s_2, x_3 = s_2)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_3, x_2 = s_2, x_3 = s_3)$$
$$* P(x_1 = s_3, x_2 = s_2, x_3 = s_3)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_3, x_2 = s_3, x_3 = s_1)$$
$$* P(x_1 = s_3, x_2 = s_3, x_3 = s_1)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_3, x_2 = s_3, x_3 = s_2)$$
$$* P(x_1 = s_3, x_2 = s_3, x_3 = s_2)$$
$$+P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_3, x_2 = s_3, x_3 = s_3)$$
$$* P(x_1 = s_3, x_2 = s_3, x_3 = s_3)$$

We have:
$$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2|x_1 = s_1, x_2 = s_1, x_3 = s_1)$$
$$* P(x_1 = s_1, x_2 = s_1, x_3 = s_1)$$
$$= P(o_1 = \varphi_4|x_1 = s_1, x_2 = s_1, x_3 = s_1)$$
$$* P(o_2 = \varphi_1|x_1 = s_1, x_2 = s_1, x_3 = s_1)$$
$$* P(o_3 = \varphi_2|x_1 = s_1, x_2 = s_1, x_3 = s_1)$$
$$* P(x_1 = s_1, x_2 = s_1, x_3 = s_1)$$

(Because observations o_1, o_2, and o_3 are mutually independent)
$$= P(o_1 = \varphi_4|x_1 = s_1) * P(o_2 = \varphi_1|x_2 = s_1)$$
$$* P(o_3 = \varphi_2|x_3 = s_1)$$
$$* P(x_1 = s_1, x_2 = s_1, x_3 = s_1)$$

(Because an observation is only dependent on the day when it is observed)
$$= P(o_1 = \varphi_4|x_1 = s_1) * P(o_2 = \varphi_1|x_2 = s_1)$$
$$* P(o_3 = \varphi_2|x_3 = s_1)$$
$$* P(x_3 = s_1|x_1 = s_1, x_2 = s_1)$$
$$* P(x_1 = s_1, x_2 = s_1)$$

(Due to multiplication rule [4, p. 100])
$$= P(o_1 = \varphi_4|x_1 = s_1) * P(o_2 = \varphi_1|x_2 = s_1)$$
$$* P(o_3 = \varphi_2|x_3 = s_1)$$
$$* P(x_3 = s_1|x_2 = s_1)$$
$$* P(x_1 = s_1, x_2 = s_1)$$

(Due to Markov property, current state is only dependent on right previous state)

$= P(o_1 = \varphi_4 | x_1 = s_1) * P(o_2 = \varphi_1 | x_2 = s_1)$
$$* P(o_3 = \varphi_2 | x_3 = s_1)$$
$$* P(x_3 = s_1 | x_2 = s_1)$$
$$* P(x_2 = s_1 | x_1 = s_1) * P(x_1 = s_1)$$
(Due to multiplication rule [4, p. 100])
$= b_{14} b_{11} b_{12} a_{11} a_{11} \pi_1$
(According to parameters A, B, and $\prod$ specified in tables 1, 2, and 3)

Similarly, we have:
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_1, x_2 = s_1, x_3 = s_2)$
$$* P(x_1 = s_1, x_2 = s_1, x_3 = s_2)$$
$$= b_{14} b_{11} b_{22} a_{12} a_{11} \pi_1$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_1, x_2 = s_1, x_3 = s_3)$
$$* P(x_1 = s_1, x_2 = s_1, x_3 = s_3)$$
$$= b_{14} b_{11} b_{32} a_{13} a_{11} \pi_1$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_1, x_2 = s_2, x_3 = s_1)$
$$* P(x_1 = s_1, x_2 = s_2, x_3 = s_1)$$
$$= b_{14} b_{21} b_{12} a_{21} a_{12} \pi_1$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_1, x_2 = s_2, x_3 = s_2)$
$$* P(x_1 = s_1, x_2 = s_2, x_3 = s_2)$$
$$= b_{14} b_{21} b_{22} a_{22} a_{12} \pi_1$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_1, x_2 = s_2, x_3 = s_3)$
$$* P(x_1 = s_1, x_2 = s_2, x_3 = s_3)$$
$$= b_{14} b_{21} b_{32} a_{23} a_{12} \pi_1$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_1, x_2 = s_3, x_3 = s_1)$
$$* P(x_1 = s_1, x_2 = s_3, x_3 = s_1)$$
$$= b_{14} b_{31} b_{12} a_{31} a_{13} \pi_1$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_1, x_2 = s_3, x_3 = s_2)$
$$* P(x_1 = s_1, x_2 = s_3, x_3 = s_2)$$
$$= b_{14} b_{31} b_{22} a_{32} a_{13} \pi_1$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_1, x_2 = s_3, x_3 = s_3)$
$$* P(x_1 = s_1, x_2 = s_3, x_3 = s_3)$$
$$= b_{14} b_{31} b_{32} a_{33} a_{13} \pi_1$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_2, x_2 = s_1, x_3 = s_1)$
$$* P(x_1 = s_2, x_2 = s_1, x_3 = s_1)$$
$$= b_{24} b_{11} b_{12} a_{11} a_{21} \pi_2$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_2, x_2 = s_1, x_3 = s_2)$
$$* P(x_1 = s_2, x_2 = s_1, x_3 = s_2)$$
$$= b_{24} b_{11} b_{22} a_{12} a_{21} \pi_2$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_2, x_2 = s_1, x_3 = s_3)$
$$* P(x_1 = s_2, x_2 = s_1, x_3 = s_3)$$
$$= b_{24} b_{11} b_{32} a_{13} a_{21} \pi_2$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_2, x_2 = s_2, x_3 = s_1)$
$$* P(x_1 = s_2, x_2 = s_2, x_3 = s_1)$$
$$= b_{24} b_{21} b_{12} a_{21} a_{22} \pi_2$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_2, x_2 = s_2, x_3 = s_2)$
$$* P(x_1 = s_2, x_2 = s_2, x_3 = s_2)$$
$$= b_{24} b_{21} b_{22} a_{22} a_{22} \pi_2$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_2, x_2 = s_2, x_3 = s_3)$
$$* P(x_1 = s_2, x_2 = s_2, x_3 = s_3)$$
$$= b_{24} b_{21} b_{32} a_{23} a_{22} \pi_2$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_2, x_2 = s_3, x_3 = s_1)$
$$* P(x_1 = s_2, x_2 = s_3, x_3 = s_1)$$
$$= b_{24} b_{31} b_{12} a_{31} a_{23} \pi_2$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_2, x_2 = s_3, x_3 = s_2)$
$$* P(x_1 = s_2, x_2 = s_3, x_3 = s_2)$$
$$= b_{24} b_{31} b_{22} a_{32} a_{23} \pi_2$$

$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_2, x_2 = s_3, x_3 = s_3)$
$$* P(x_1 = s_2, x_2 = s_3, x_3 = s_3)$$
$$= b_{24} b_{31} b_{32} a_{33} a_{23} \pi_2$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_3, x_2 = s_1, x_3 = s_1)$
$$* P(x_1 = s_3, x_2 = s_1, x_3 = s_1)$$
$$= b_{34} b_{11} b_{12} a_{11} a_{31} \pi_3$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_3, x_2 = s_1, x_3 = s_2)$
$$* P(x_1 = s_3, x_2 = s_1, x_3 = s_2)$$
$$= b_{34} b_{11} b_{22} a_{12} a_{31} \pi_3$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_3, x_2 = s_1, x_3 = s_3)$
$$* P(x_1 = s_3, x_2 = s_1, x_3 = s_3)$$
$$= b_{34} b_{11} b_{32} a_{13} a_{31} \pi_3$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_3, x_2 = s_2, x_3 = s_1)$
$$* P(x_1 = s_3, x_2 = s_2, x_3 = s_1)$$
$$= b_{34} b_{21} b_{12} a_{21} a_{32} \pi_3$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_3, x_2 = s_2, x_3 = s_2)$
$$* P(x_1 = s_3, x_2 = s_2, x_3 = s_2)$$
$$= b_{34} b_{21} b_{22} a_{22} a_{32} \pi_3$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_3, x_2 = s_2, x_3 = s_3)$
$$* P(x_1 = s_3, x_2 = s_2, x_3 = s_3)$$
$$= b_{34} b_{21} b_{32} a_{23} a_{32} \pi_3$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_3, x_2 = s_3, x_3 = s_1)$
$$* P(x_1 = s_3, x_2 = s_3, x_3 = s_1)$$
$$= b_{34} b_{31} b_{12} a_{31} a_{33} \pi_3$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_3, x_2 = s_3, x_3 = s_2)$
$$* P(x_1 = s_3, x_2 = s_3, x_3 = s_2)$$
$$= b_{34} b_{31} b_{22} a_{32} a_{33} \pi_3$$
$P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2 | x_1 = s_3, x_2 = s_3, x_3 = s_3)$
$$* P(x_1 = s_3, x_2 = s_3, x_3 = s_3)$$
$$= b_{34} b_{31} b_{32} a_{33} a_{33} \pi_3$$

It implies
$P(O|\Delta) = P(o_1 = \varphi_4, o_2 = \varphi_1, o_3 = \varphi_2)$
$= b_{14} b_{11} b_{12} a_{11} a_{11} \pi_1 + b_{14} b_{11} b_{22} a_{12} a_{11} \pi_1$
$$+ b_{14} b_{11} b_{32} a_{13} a_{11} \pi_1$$
$+ b_{14} b_{21} b_{12} a_{21} a_{12} \pi_1 + b_{14} b_{21} b_{22} a_{22} a_{12} \pi_1$
$$+ b_{14} b_{21} b_{32} a_{23} a_{12} \pi_1$$
$+ b_{14} b_{31} b_{12} a_{31} a_{13} \pi_1 + b_{14} b_{31} b_{22} a_{32} a_{13} \pi_1$
$$+ b_{14} b_{31} b_{32} a_{33} a_{13} \pi_1$$
$+ b_{24} b_{11} b_{12} a_{11} a_{21} \pi_2 + b_{24} b_{11} b_{22} a_{12} a_{21} \pi_2$
$$+ b_{24} b_{11} b_{32} a_{13} a_{21} \pi_2$$
$+ b_{24} b_{21} b_{12} a_{21} a_{22} \pi_2 + b_{24} b_{21} b_{22} a_{22} a_{22} \pi_2$
$$+ b_{24} b_{21} b_{32} a_{23} a_{22} \pi_2$$
$+ b_{24} b_{31} b_{12} a_{31} a_{23} \pi_2 + b_{24} b_{31} b_{22} a_{32} a_{23} \pi_2$
$$+ b_{24} b_{31} b_{32} a_{33} a_{23} \pi_2$$
$+ b_{34} b_{11} b_{12} a_{11} a_{31} \pi_3 + b_{34} b_{11} b_{22} a_{12} a_{31} \pi_3$
$$+ b_{34} b_{11} b_{32} a_{13} a_{31} \pi_3$$
$+ b_{34} b_{21} b_{12} a_{21} a_{32} \pi_3 + b_{34} b_{21} b_{22} a_{22} a_{32} \pi_3$
$$+ b_{34} b_{21} b_{32} a_{23} a_{32} \pi_3$$
$+ b_{34} b_{31} b_{12} a_{31} a_{33} \pi_3 + b_{34} b_{31} b_{22} a_{32} a_{33} \pi_3$
$$+ b_{34} b_{31} b_{32} a_{33} a_{33} \pi_3$$
$= 0.012980859375$

It is easy to explain that given weather HMM modeled by parameters A, B, and $\prod$ specified in tables 1, 2, and 3, the event that it is *soggy*, *dry*, and *dryish* in three successive days is rare because the probability of such event $P(O|\Delta)$ is low ($\approx 1.3\%$). It is easy to recognize that it is impossible to browse all combinational cases of given observation sequence $O = \{o_1, o_2, \ldots, o_T\}$ as we knew that it is necessary to survey $3^3 = 27$

mutually exclusive cases of weather states with a tiny number of observations {*soggy*, *dry*, *dryish*}. Exactly, given n states and T observations, it takes extremely expensive cost to survey n^T cases. According to [3, pp. 262-263], there is a so-called *forward-backward procedure* to decrease computational cost for determining the probability $P(O|\Delta)$. Let $\alpha_t(i)$ be the joint probability of partial observation sequence {$o_1, o_2,…, o_t$} and state $x_t=s_i$ where $1 \leq t \leq T$, specified by (1).

$$\alpha_t(i) = P(o_1, o_2, …, o_t, x_t = s_i|\Delta) \qquad (1)$$

The joint probability $\alpha_t(i)$ is also called *forward variable* at time point t and state s_i. The product $\alpha_t(i)a_{ij}$ where a_{ij} is the transition probability from state i to state j counts for probability of join event that partial observation sequence {$o_1, o_2,…, o_t$} exists and the state s_i at time point t is changed to s_j at time point $t+1$.

$\alpha_t(i)a_{ij} = P(o_1, o_2, …, o_t, x_t = s_i|\Delta)P(x_{t+1} = s_j|x_t = s_i)$
$= P(o_1, o_2, …, o_t|x_t = s_i)P(x_t = s_i)P(x_{t+1} = s_j|x_t = s_i)$
(Due to multiplication rule [4, p. 100])
$= P(o_1, o_2, …, o_t|x_t = s_i)P(x_{t+1} = s_j|x_t = s_i)P(x_t = s_i)$
$= P(o_1, o_2, …, o_t, x_{t+1} = s_j|x_t = s_i)P(x_t = s_i)$
(Because the partial observation sequence {$o_1, o_2,…, o_t$} is independent from next state x_{t+1} given current state x_t)
$= P(o_1, o_2, …, o_t, x_t = s_i, x_{t+1} = s_j)$
(Due to multiplication rule [4, p. 100])

Summing product $\alpha_t(i)a_{ij}$ over all n possible states of x_t produces probability of join event that partial observation sequence {$o_1, o_2,…, o_t$} exists and the next state is $x_{t+1}=s_j$ regardless of the state x_t.

$$\sum_{i=1}^{n} \alpha_t(i)a_{ij} = \sum_{i=1}^{n} P(o_1, o_2, …, o_t, x_t = s_i, x_{t+1} = s_j)$$
$$= P(o_1, o_2, …, o_t, x_{t+1} = s_j)$$

The forward variable at time point $t+1$ and state s_j is calculated on $\alpha_t(i)$ as follows:

$\alpha_{t+1}(j) = P(o_1, o_2, …, o_t, o_{t+1}, x_{t+1} = s_j|\Delta)$
$= P(o_{t+1}|o_1, o_2, …, o_t, x_{t+1} = s_j)P(o_1, o_2, …, o_t, x_{t+1} = s_j)$
(Due to multiplication rule)
$= P(o_{t+1}|x_{t+1} = s_j)P(o_1, o_2, …, o_t, x_{t+1} = s_j)$
(Due to observations are mutually independent)
$$= b_j(o_{t+1}) \sum_{i=1}^{n} \alpha_t(i)a_{ij}$$

Where $b_j(o_{t+1})$ is the probability of observation o_{t+1} when the state stochastic process is in state s_j, please see an example of observation probability matrix shown in table 3. In brief, please pay attention to recurrence property of forward variable specified by (2).

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i)a_{ij}\right)b_j(o_{t+1}) \qquad (2)$$

The aforementioned construction of forward recurrence equation (2) is essentially to build up Markov chain, illustrated by fig. 2 [3, p. 262].
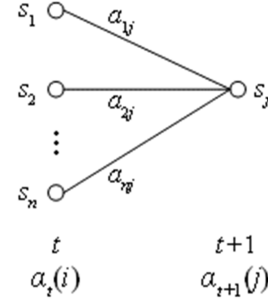

Figure 2. *Construction of recurrence formula for forward variable.*

According to the forward recurrence equation (2), given observation sequence $O = \{o_1, o_2,…, o_T\}$, we have:

$$\alpha_T(i) = P(o_1, o_2, …, o_T, x_T = s_i|\Delta)$$

The probability $P(O|\Delta)$ is sum of $\alpha_T(i)$ over all n possible states of x_T, specified by (3).

$$P(O|\Delta) = P(o_1, o_2, …, o_T) =$$
$$\sum_{i=1}^{n} P(o_1, o_2, …, o_T, x_T = s_i|\Delta) = \sum_{i=1}^{n} \alpha_T(i) \qquad (3)$$

The forward-backward procedure to calculate the probability $P(O|\Delta)$, based on forward equations (2) and (3), includes three steps as shown in table 4 [3, p. 262].

Table 4. *Forward-backward procedure based on forward variable to calculate the probability $P(O|\Delta)$.*

1. Initialization step: Initializing $\alpha_1(i) = b_i(o_1)\pi_i$ for all $1 \leq i \leq n$
2. Recurrence step: Calculating all $\alpha_{t+1}(j)$ for all $1 \leq j \leq n$ and $1 \leq t \leq T - 1$ according to (2). $$\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i)a_{ij}\right)b_j(o_{t+1})$$
3. Evaluation step: Calculating the probability $P(O|\Delta) = \sum_{i=1}^{n} \alpha_T(i)$

It is required to execute $n+2n^2(T-1)+n-1 = 2n^2(T-1)+2n-1$ operations for forward-backward procedure based on forward variable due to:

- There are n multiplications at initialization step.
- There are n multiplications, $n-1$ additions, and 1 multiplication over the expression $\left(\sum_{i=1}^{n} \alpha_t(i)a_{ij}\right)b_j(o_{t+1})$. There are n cases of values $\alpha_{t+1}(j)$ for all $1 \leq j \leq n$ at time point $t+1$. So, there are $(n+n-1+1)$ $n = 2n^2$ operations over values $\alpha_{t+1}(j)$ for all $1 \leq j \leq n$ at time point t. The recurrence step runs over $T-1$ times and so, there are $2n^2(T-1)$ operations at recurrence step.
- There are $n-1$ additions at evaluation step.

Inside $2n^2(T-1)+2n-1$ operations, there are $n+(n+1)n(T-1)$ $= n+(n^2+n)(T-1)$ multiplications and $(n-1)n(T-1)+n-1 = (n^2+n)(T-1)+n-1$ additions.

Going back example with weather HMM whose parameters A, B, and $\prod$ are specified in tables 1, 2, and 3. We need to re-calculate the probability of observation sequence $O = \{o_1=\varphi_4=soggy, \quad o_2=\varphi_1=dry, \quad o_3=\varphi_2=dryish\}$ by forward-backward procedure shown in table 4 (based on forward variable). According to initialization step of forward-backward procedure based on forward variable, we have:

$$\alpha_1(1) = b_1(o_1 = \varphi_4)\pi_1 = b_{14}\pi_1 = 0.0165$$

$\alpha_1(2) = b_2(o_1 = \varphi_4)\pi_2 = b_{24}\pi_2 = 0.0825$
$\alpha_1(3) = b_3(o_1 = \varphi_4)\pi_3 = b_{34}\pi_3 = 0.165$

According to recurrence step of forward-backward procedure based on forward variable, we have:

$$\alpha_2(1) = \left(\sum_{i=1}^{3}\alpha_1(i)a_{i1}\right)b_1(o_2 = \varphi_1) = \left(\sum_{i=1}^{3}\alpha_1(i)a_{i1}\right)b_{11}$$
$$= 0.04455$$

$$\alpha_2(2) = \left(\sum_{i=1}^{3}\alpha_1(i)a_{i2}\right)b_2(o_2 = \varphi_1) = 0.01959375$$

$$\alpha_2(3) = \left(\sum_{i=1}^{3}\alpha_1(i)a_{i3}\right)b_3(o_2 = \varphi_1) = 0.00556875$$

$$\alpha_3(1) = \left(\sum_{i=1}^{3}\alpha_2(i)a_{i1}\right)b_1(o_3 = \varphi_2) = 0.0059090625$$

$$\alpha_3(2) = \left(\sum_{i=1}^{3}\alpha_2(i)a_{i2}\right)b_2(o_3 = \varphi_2) = 0.005091796875$$

$$\alpha_3(3) = \left(\sum_{i=1}^{3}\alpha_2(i)a_{i3}\right)b_3(o_3 = \varphi_2) = 0.00198$$

According to evaluation step of forward-backward procedure based on forward variable, the probability of observation sequence $O = \{o_1=s_4=soggy, o_2=s_1=dry, o_3=s_2=dryish\}$ is:
$P(O|\Delta) = \alpha_3(1) + \alpha_3(2) + \alpha_3(3) = 0.012980859375$

The result from the forward-backward procedure based on forward variable is the same to the one from aforementioned brute-force method that browses all $3^3=27$ mutually exclusive cases of weather states.

There is interesting thing that the forward-backward procedure can be implemented based on so-called *backward variable*. Let $\beta_t(i)$ be the backward variable which is conditional probability of partial observation sequence $\{o_t, o_{t+1},\ldots, o_T\}$ given state $x_t=s_i$ where $1 \leq t \leq T$, specified by (4).

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, \ldots, o_T | x_t = s_i, \Delta) \qquad (4)$$

We have
$a_{ij}b_j(o_{t+1})\beta_{t+1}(j)$
$$= P(x_{t+1} = s_j | x_t = s_i)$$
$$* P(o_{t+1} | x_{t+1} = s_j)$$
$$* P(o_{t+2}, o_{t+3}, \ldots, o_T | x_{t+1} = s_j, \Delta)$$
$$= P(x_{t+1} = s_j | x_t = s_i)$$
$$* P(o_{t+1}, o_{t+2}, o_{t+3}, \ldots, o_T | x_{t+1} = s_j, \Delta)$$
(Because observations $o_{t+1}, o_{t+2},\ldots, o_T$ are mutually independent)
$$= P(x_{t+1} = s_j | x_t = s_i)$$
$$* P(o_{t+1}, o_{t+2}, o_{t+3}, \ldots, o_T | x_t = s_i, x_{t+1} = s_j, \Delta)$$
(Because partial observation sequence $o_{t+1}, o_{t+2},\ldots, o_T$ is independent from state x_t at time point t)
$$= P(o_{t+1}, o_{t+2}, o_{t+3}, \ldots, o_T, x_{t+1} = s_j | x_t = s_i, \Delta)$$
(Due to multiplication rule [4, p. 100])
Summing the product $a_{ij}b_j(o_{t+1})\beta_{t+1}(j)$ over all n possible states of $x_{t+1}=s_j$, we have:

$$\sum_{j=1}^{n} a_{ij}b_j(o_{t+1})\beta_{t+1}(j)$$
$$= \sum_{j=1}^{n} P(o_{t+1}, o_{t+2}, o_{t+3}, \ldots, o_T, x_{t+1} = s_j | x_t = s_i, \Delta)$$
$$= P(o_{t+1}, o_{t+2}, o_{t+3}, \ldots, o_T | x_t = s_i, \Delta)$$
(Due to the total probability rule [4, p. 101])
$$= \beta_t(i)$$
In brief, the recurrence property of backward variable specified by (5).

$$\beta_t(i) = \sum_{j=1}^{n} a_{ij}b_j(o_{t+1})\beta_{t+1}(j) \qquad (5)$$

Where $b_j(o_{t+1})$ is the probability of observation o_{t+1} when the state stochastic process is in state s_j, please see an example of observation probability matrix shown in table 3. The construction of backward recurrence equation (5) is essentially to build up Markov chain, illustrated by fig. 3 [3, p. 263].
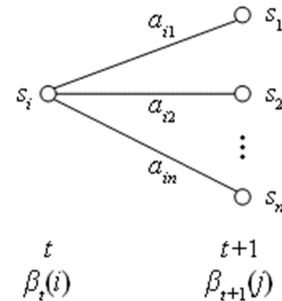

Figure 3. *Construction of recurrence equation for backward variable.*

According to the backward recurrence equation (5), given observation sequence $O = \{o_1, o_2,\ldots, o_T\}$, we have:

$$\beta_1(i) = P(o_2, o_3, \ldots, o_T | x_1 = s_i, \Delta)$$

The product $\pi_i b_i(o_1)\beta_1(i)$ is:
$\pi_i b_i(o_1)\beta_1(i)$
$$= P(x_1 = s_i)P(o_1 | x_1 = s_i)P(o_2, o_3, \ldots, o_T | x_1 = s_i, \Delta)$$
$$= P(x_1 = s_i)P(o_1, o_2, o_3, \ldots, o_T | x_1 = s_i, \Delta)$$
(Because observations $o_1, o_2,\ldots, o_T$ are mutually independent)
$$= P(o_1, o_2, o_3, \ldots, o_T, x_1 = s_i | \Delta)$$
It implies that the probability $P(O|\Delta)$ is:
$P(O|\Delta) = P(o_1, o_2, \ldots, o_T)$
$$= \sum_{i=1}^{n} P(o_1, o_2, \ldots, o_T, x_1 = s_i | \Delta)$$
(Due to the total probability rule [4, p. 101])
$$= \sum_{i=1}^{n} \pi_i b_i(o_1)\beta_1(i)$$
Shortly, the probability $P(O|\Delta)$ is sum of product $\pi_i b_i(o_1)\beta_1(i)$ over all n possible states of $x_1=s_i$, specified by (6).

$$P(O|\Delta) = \sum_{i=1}^{n} \pi_i b_i(o_1)\beta_1(i) \qquad (6)$$

The forward-backward procedure to calculate the probability $P(O|\Delta)$, based on backward equations (5) and (6), includes three steps as shown in table 5 [3, p. 263].

Table 5. *Forward-backward procedure based on backward variable to calculate the probability P(O|Δ).*

1. Initialization step: Initializing $\beta_T(i) = 1$ for all $1 \leq i \leq n$
2. Recurrence step: Calculating all $\beta_t(i)$ for all $1 \leq i \leq n$ and $t=T–1$, $t=T–2,\ldots, t=1$, according to (5).
3. Evaluation step: Calculating the probability $P(O|\Delta)$ according to (6), $P(O|\Delta) = \sum_{i=1}^{n} \pi_i b_i(o_1)\beta_1(i)$

It is required to execute $(3n–1)n(T–1)+2n+n–1 = 3n^2(T–1)–n(T–4)–1$ operations for forward-backward procedure based on forward variable due to:

- There are $2n$ multiplications and $n–1$ additions over the sum $\sum_{j=1}^{n} a_{ij}b_j(O_{t+1})\beta_{t+1}(j)$. So, there are $(2n+n–1)n = (3n–1)n$ operations over values $\beta_t(i)$ for all $1 \leq i \leq n$ at time point t. The recurrence step runs over $T–1$ times and so, there are $(3n–1)n(T–1)$ operations at recurrence step.
- There are $2n$ multiplications and $n–1$ additions over the sum $\sum_{i=1}^{n} \pi_i b_i(o_1)\beta_1(i)$ at evaluation step.

Inside $3n^2(T–1)–n(T–4)–1$ operations, there are $2n^2(T–1)+2n$ multiplications and $(n–1)n(T–1)+n–1 = n^2(T–1)–n(T–2)–1$ additions.

Going back example with weather HMM whose parameters A, B, and $\prod$ are specified in tables 1, 2, and 3. We need to re-calculate the probability of observation sequence $O = \{o_1=\varphi_4=soggy, \quad o_2=\varphi_1=dry, \quad o_3=\varphi_2=dryish\}$ by forward-backward procedure shown in table 5 (based on backward variable). According to initialization step of forward-backward procedure based on backward variable, we have:

$$\beta_3(1) = \beta_3(2) = \beta_3(3) = 1$$

According to recurrence step of forward-backward procedure based on backward variable, we have:

$$\beta_2(1) = \sum_{j=1}^{n} a_{1j}b_j(o_3 = \varphi_2)\beta_3(j) = \sum_{j=1}^{n} a_{1j}b_{j2}\beta_3(j) = 0.1875$$

$$\beta_2(2) = \sum_{j=1}^{n} a_{2j}b_j(o_3 = \varphi_2)\beta_3(j) = 0.19$$

$$\beta_2(3) = \sum_{j=1}^{n} a_{3j}b_j(o_3 = \varphi_2)\beta_3(j) = 0.1625$$

$$\beta_1(1) = \sum_{j=1}^{n} a_{1j}b_j(o_2 = \varphi_1)\beta_2(j) = 0.07015625$$

$$\beta_1(2) = \sum_{j=1}^{n} a_{2j}b_j(o_2 = \varphi_1)\beta_2(j) = 0.0551875$$

$$\beta_1(3) = \sum_{j=1}^{n} a_{3j}b_j(o_2 = \varphi_1)\beta_2(j) = 0.0440625$$

According to evaluation step of forward-backward procedure based on backward variable, the probability of observation sequence $O = \{o_1=\varphi_4=soggy, o_2=\varphi_1=dry, o_3=\varphi_2=dryish\}$ is:

$$P(O|\Delta) = \sum_{i=1}^{3} \pi_i b_i(o_1 = \varphi_4)\beta_1(i) = \sum_{i=1}^{3} \pi_i b_{i4}\beta_1(i)$$
$$= 0.012980859375$$

The result from the forward-backward procedure based on backward variable is the same to the one from aforementioned brute-force method that browses all $3^3=27$ mutually exclusive cases of weather states and the one from forward-backward procedure based on forward variable.

The evaluation problem is now described thoroughly in this section. The uncovering problem is mentioned particularly in successive section.

3. HMM Uncovering Problem

Recall that given HMM Δ and observation sequence $O = \{o_1, o_2,\ldots, o_T\}$ where $o_t \in \Phi$, how to find out a state sequence $X = \{x_1, x_2,\ldots, x_T\}$ where $x_t \in S$ so that X is most likely to have produced the observation sequence O. This is the uncovering problem: which sequence of state transitions is most likely to have led to given observation sequence. In other words, it is required to establish an *optimal criterion* so that the state sequence X leads to maximizing such criterion. The simple criterion is the conditional probability of sequence X with respect to sequence O and model Δ, denoted $P(X|O,\Delta)$. We can apply brute-force strategy: "go through all possible such X and pick the one leading to maximizing the criterion $P(X|O,\Delta)$".

$$X = \underset{X}{\text{argmax}}\big(P(X|O, \Delta)\big)$$

This strategy is impossible if the number of states and observations is huge. Another popular way is to establish a so-called *individually optimal criterion* [3, p. 263] which is described right later.

Let $\gamma_t(i)$ be joint probability that the stochastic process is in state s_i at time point t with observation sequence $O = \{o_1, o_2,\ldots, o_T\}$, equation (7) specifies this probability based on forward variable α_t and backward variable β_t.

$$\gamma_t(i) = P(o_1, o_2, \ldots, o_T, x_t = s_i|\Delta) = \alpha_t(i)\beta_t(i) \quad (7)$$

The variable $\gamma_t(i)$ is also called *individually optimal criterion* with note that forward variable α_t and backward variable β_t are calculated according to (2) and (5), respectively.

Following is proof of (7).
$\gamma_t(i) = P(o_1, o_2, \ldots, o_T, x_t = s_i|\Delta)$
$= P(x_t = s_i, o_1, o_2, \ldots, o_T|\Delta)$
         (Due to Bayes' rule [4, p. 99])
$= P(o_1, o_2, \ldots, o_t, x_t = s_i, o_{t+1}, o_{t+2}, \ldots, o_T|\Delta)$
$= P(o_1, o_2, \ldots, o_t, x_t = s_i|\Delta)$
$* P(o_{t+1}, o_{t+2}, \ldots, o_T|o_1, o_2, \ldots, o_t, x_t = s_i, \Delta)$
       (Due to multiplication rule [4, p. 100])
$= P(o_1, o_2, \ldots, o_t, x_t = s_i|\Delta)P(o_{t+1}, o_{t+2}, \ldots, o_T|x_t = s_i, \Delta)$
   (Because observations $o_1, o_2,\ldots, o_T$ are observed independently)
$= \alpha_t(i)\beta_t(i)$
 (According to (1) and (4) for determining forward variable

and backward variable)

The state sequence $X = \{x_1, x_2,\ldots, x_T\}$ is determined by selecting each state $x_t \in S$ so that it maximizes $\gamma_t(i)$.

$$x_t = \underset{i}{\operatorname{argmax}}\, P(x_t = s_i | o_1, o_2, \ldots, o_T, \Delta)$$

$$= \underset{i}{\operatorname{argmax}}\, \frac{\alpha_t(i)\beta_t(i)}{P(o_1, o_2, \ldots, o_T | \Delta)}$$

$$= \underset{i}{\operatorname{argmax}}\, \frac{P(o_1, o_2, \ldots, o_T, x_t = s_i | \Delta)}{P(o_1, o_2, \ldots, o_T | \Delta)}$$

(Due to Bayes' rule [4, p. 99])

$$= \underset{i}{\operatorname{argmax}}\, \frac{\gamma_t(i)}{P(o_1, o_2, \ldots, o_T | \Delta)}$$

(Due to (7))

Because the probability $P(o_1, o_2, \ldots, o_T | \Delta)$ is not relevant to state sequence X, it is possible to remove it from the optimization criterion. Thus, equation (8) specifies how to find out the optimal state x_t of X at time point t.

$$x_t = \underset{i}{\operatorname{argmax}}\, \gamma_t(i) = \underset{i}{\operatorname{argmax}}\, \alpha_t(i)\beta_t(i) \qquad (8)$$

Note that index i is identified with state $s_i \in S$ according to (8). The optimal state x_t of X at time point t is the one that maximizes product $\alpha_t(i)\,\beta_t(i)$ over all values s_i. The procedure to find out state sequence $X = \{x_1, x_2,\ldots, x_T\}$ based on individually optimal criterion is called *individually optimal procedure* that includes three steps, shown in table 6.

Table 6. *Individually optimal procedure to solve uncovering problem.*

1. Initialization step:
   - Initializing $\alpha_1(i) = b_i(o_1)\pi_i$ for all $1 \le i \le n$
   - Initializing $\beta_T(i) = 1$ for all $1 \le i \le n$
2. Recurrence step:
   - Calculating all $\alpha_{t+1}(i)$ for all $1 \le i \le n$ and $1 \le t \le T-1$ according to (2).
   - Calculating all $\beta_t(i)$ for all $1 \le i \le n$ and $t=T-1, t=T-2,\ldots, t=1$, according to (5).
   - Calculating all $\gamma_t(i)=\alpha_t(i)\beta_t(i)$ for all $1 \le i \le n$ and $1 \le t \le T$ according to (7).
   - Determining optimal state x_t of X at time point t is the one that maximizes $\gamma_t(i)$ over all values s_i.
     $$x_t = \underset{i}{\operatorname{argmax}}\, \gamma_t(i)$$
3. Final step: The state sequence $X = \{x_1, x_2,\ldots, x_T\}$ is totally determined when its partial states x_t (s) where $1 \le t \le T$ are found in recurrence step.

It is required to execute $n+(5n^2-n)(T-1)+2nT$ operations for individually optimal procedure due to:

- There are n multiplications for calculating $\alpha_1(i)$ (s).
- The recurrence step runs over $T-1$ times. There are $2n^2(T-1)$ operations for determining $\alpha_{t+1}(i)$ (s) over all $1 \le i \le n$ and $1 \le t \le T-1$. There are $(3n-1)n(T-1)$ operations for determining $\beta_t(i)$ (s) over all $1 \le i \le n$ and $t=T-1, t=T-2,\ldots, t=1$. There are nT multiplications for determining $\gamma_t(i)=\alpha_t(i)\beta_t(i)$ over all $1 \le i \le n$ and $1 \le t \le T$. There are nT comparisons for determining optimal state $x_t = \underset{i}{\operatorname{argmax}}\, \gamma_t(i)$ over all $1 \le i \le n$ and $1 \le t \le T$. In general, there are $2n^2(T-1)+ (3n-1)n(T-1) + nT + nT = (5n^2-n)(T-1) + 2nT$ operations at the recurrence step.

Inside $n+(5n^2-n)(T-1)+2nT$ operations, there are $n+(n+1)n(T-1)+2n^2(T-1)+nT = (3n^2+n)(T-1)+nT+n$ multiplications and $(n-1)n(T-1)+(n-1)n(T-1) = 2(n^2-n)(T-1)$ ad-

ditions and nT comparisons.

For example, given HMM Δ whose parameters A, B, and $\prod$ specified in tables 1, 2, and 3, which is designed for weather forecast. Suppose humidity is *soggy* and *dry* in days 1 and 2, respectively. We apply individual optimal procedure into solving the uncovering problem that finding out the optimal state sequence $X = \{x_1, x_2\}$ with regard to observation sequence $O = \{o_1=\varphi_4=soggy,\ o_2=\varphi_1=dry,\ o_3=\varphi_2=dryish\}$. According to (2) and (5), forward variable and backward variable are calculated as follows:

$$\alpha_1(1) = b_1(o_1 = \varphi_4)\pi_1 = b_{14}\pi_1 = 0.0165$$
$$\alpha_1(2) = b_2(o_1 = \varphi_4)\pi_2 = b_{24}\pi_2 = 0.0825$$
$$\alpha_1(3) = b_3(o_1 = \varphi_4)\pi_3 = b_{34}\pi_3 = 0.165$$

$$\alpha_2(1) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i1}\right)b_1(o_2 = \varphi_1) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i1}\right)b_{11}$$
$$= 0.04455$$

$$\alpha_2(2) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i2}\right)b_2(o_2 = \varphi_1) = 0.01959375$$

$$\alpha_2(3) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i3}\right)b_3(o_2 = \varphi_1) = 0.00556875$$

$$\alpha_3(1) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i1}\right)b_1(o_3 = \varphi_2) = 0.0059090625$$

$$\alpha_3(2) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i2}\right)b_2(o_3 = \varphi_2) = 0.005091796875$$

$$\alpha_3(3) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i3}\right)b_3(o_3 = \varphi_2) = 0.00198$$

$$\beta_3(1) = \beta_3(2) = \beta_3(3) = 1$$

$$\beta_2(1) = \sum_{j=1}^{n} a_{1j}b_j(o_3 = \varphi_2)\beta_3(j) = \sum_{j=1}^{n} a_{1j}b_{j2}\beta_3(j)$$
$$= 0.1875$$

$$\beta_2(2) = \sum_{j=1}^{n} a_{2j}b_j(o_3 = \varphi_2)\beta_3(j) = 0.19$$

$$\beta_2(3) = \sum_{j=1}^{n} a_{3j}b_j(o_3 = \varphi_2)\beta_3(j) = 0.1625$$

$$\beta_1(1) = \sum_{j=1}^{n} a_{1j}b_j(o_2 = \varphi_1)\beta_2(j) = 0.07015625$$

$$\beta_1(2) = \sum_{j=1}^{n} a_{2j}b_j(o_2 = \varphi_1)\beta_2(j) = 0.0551875$$

$$\beta_1(3) = \sum_{j=1}^{n} a_{3j}b_j(o_2 = \varphi_1)\beta_2(j) = 0.0440625$$

According to recurrence step of individually optimal procedure, individually optimal criterion $\gamma_t(i)$ and optimal state x_t are calculated as follows:

$$\gamma_1(1) = \alpha_1(1)\beta_1(1) = 0.001157578125$$
$$\gamma_1(2) = \alpha_1(2)\beta_1(2) = 0.00455296875$$
$$\gamma_1(3) = \alpha_1(3)\beta_1(3) = 0.0072703125$$

$$x_1 = \underset{i}{\mathrm{argmax}}\{\gamma_1(i)\} = \underset{i}{\mathrm{argmax}}\{\gamma_1(1), \gamma_1(2), \gamma_1(3)\} = s_3$$
$$= rainy$$
$$\gamma_2(1) = \alpha_2(1)\beta_2(1) = 0.008353125$$
$$\gamma_2(2) = \alpha_2(2)\beta_2(2) = 0.0037228125$$
$$\gamma_2(3) = \alpha_2(3)\beta_2(3) = 0.000904921875$$
$$x_2 = \underset{i}{\mathrm{argmax}}\{\gamma_2(i)\} = \underset{i}{\mathrm{argmax}}\{\gamma_2(1), \gamma_2(2), \gamma_2(3)\} = s_1$$
$$= sunny$$
$$\gamma_3(1) = \alpha_3(1)\beta_3(1) = 0.0059090625$$
$$\gamma_3(2) = \alpha_3(2)\beta_3(2) = 0.005091796875$$
$$\gamma_3(3) = \alpha_3(3)\beta_3(3) = 0.00198$$
$$x_3 = \underset{i}{\mathrm{argmax}}\{\gamma_3(i)\} = \underset{i}{\mathrm{argmax}}\{\gamma_3(1), \gamma_3(2), \gamma_3(3)\} = s_1$$
$$= sunny$$

As a result, the optimal state sequence is $X = \{x_1=rainy, x_2=sunny, x_3=sunny\}$.

The individually optimal criterion $\gamma_t(i)$ does not reflect the whole probability of state sequence X given observation sequence O because it focuses only on how to find out each partially optimal state x_t at each time point t. Thus, the indi-

vidually optimal procedure is heuristic method. Viterbi algorithm [3, p. 264] is alternative method that takes interest in the whole state sequence X by using joint probability $P(X,O|\Delta)$ of state sequence and observation sequence as optimal criterion for determining state sequence X. Let $\delta_t(i)$ be the maximum joint probability of observation sequence O and state $x_t=s_i$ over $t-1$ previous states. The quantity $\delta_t(i)$ is called *joint optimal criterion* at time point t, which is specified by (9).

$$\delta_t(i) = \max_{x_1,x_2,\dots,x_{t-1}}\big(P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_t = s_i|\Delta)\big) \quad (9)$$

The recurrence property of *joint optimal criterion* is specified by (10).

$$\delta_{t+1}(j) = \big(\max_i(\delta_t(i)a_{ij})\big)b_j(o_{t+1}) \quad (10)$$

The semantic content of joint optimal criterion δ_t is similar to the forward variable α_t. Following is the proof of (10).

$$\delta_{t+1}(j) = \max_{x_1,x_2,\dots,x_t}\big(P(o_1, o_2, \dots, o_t, o_{t+1}, x_1, x_2, \dots, x_t, x_{t+1} = s_j|\Delta)\big)$$

$$= \max_{x_1,x_2,\dots,x_t}\big(P(o_{t+1}|o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_t, x_{t+1} = s_j) * P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_t, x_{t+1} = s_j)\big)$$
(Due to multiplication rule [4, p. 100])

$$= \max_{x_1,x_2,\dots,x_t}\big(P(o_{t+1}|x_{t+1} = s_j) * P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_t, x_{t+1} = s_j)\big)$$
(Due to observations are mutually independent)

$$= \max_{x_1,x_2,\dots,x_t}\big(b_j(o_{t+1}) * P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_t, x_{t+1} = s_j)\big)$$

$$= \max_{x_1,x_2,\dots,x_t}\big(P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_t, x_{t+1} = s_j)\big) * b_j(o_{t+1})$$
(The probability $b_j(o_{t+1})$ is moved out of the maximum operation because it is independent from states $x_1, x_2,\dots, x_t$)

$$= \max_{x_1,x_2,\dots,x_t}\big(P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_{t-1}, x_{t+1} = s_j|x_t) * P(x_t)\big) * b_j(o_{t+1})$$
(Due to multiplication rule [4, p. 100])

$$= \max_{x_1,x_2,\dots,x_t}\big(P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_{t-1}|x_{t+1} = s_j, x_t) * P(x_{t+1} = s_j|x_t) * P(x_t)\big) * b_j(o_{t+1})$$
(Due to multiplication rule [4, p. 100])

$$= \max_{x_1,x_2,\dots,x_t}\big(P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_{t-1}|x_t) * P(x_{t+1} = s_j|x_t) * P(x_t)\big) * b_j(o_{t+1})$$
(Because observation x_{t+1} is dependent from $o_1, o_2,\dots, o_t, x_1, x_2,\dots, x_{t-1}$)

$$= \max_{x_1,x_2,\dots,x_t}\big(P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_{t-1}|x_t) * P(x_t) * P(x_{t+1} = s_j|x_t)\big) * b_j(o_{t+1})$$

$$= \max_{x_1,x_2,\dots,x_t}\big(P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_{t-1}, x_t) * P(x_{t+1} = s_j|x_t)\big) * b_j(o_{t+1})$$
(Due to multiplication rule [4, p. 100])

$$= \max_{x_t}\Big(\max_{x_1,x_2,\dots,x_{t-1}}\big(P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_{t-1}, x_t) * P(x_{t+1} = s_j|x_t)\big)\Big) * b_j(o_{t+1})$$

$$= \max_{x_t}\Big(\big(\max_{x_1,x_2,\dots,x_{t-1}} P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_{t-1}, x_t)\big) * P(x_{t+1} = s_j|x_t)\Big) * b_j(o_{t+1})$$

$$= \max_i\Big(\big(\max_{x_1,x_2,\dots,x_{t-1}} P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_{t-1}, x_t = s_i)\big) * P(x_{t+1} = s_j|x_t = s_i)\Big) * b_j(o_{t+1})$$

$$= \max_i\Big(\big(\max_{x_1,x_2,\dots,x_{t-1}} P(o_1, o_2, \dots, o_t, x_1, x_2, \dots, x_{t-1}, x_t = s_i)\big) * a_{ij}\Big) * b_j(o_{t+1})$$

$$= \max_i\big(\delta_t(i) * a_{ij}\big) * b_j(o_{t+1})$$

$$= \big(\max_i(\delta_t(i)a_{ij})\big)b_j(o_{t+1})$$

Given criterion $\delta_{t+1}(j)$, the state $x_{t+1}=s_j$ that maximizes $\delta_{t+1}(j)$ is stored in the backtracking state $q_{t+1}(j)$ that is specified by (11).

$$q_{t+1}(j) = \mathrm{argmax}_i\big(\delta_t(i)a_{ij}\big) \quad (11)$$

Note that index i is identified with state $s_i \in S$ according to (11). The Viterbi algorithm based on joint optimal criterion $\delta_t(i)$ includes three steps described in table 7.

Table 7. *Viterbi algorithm to solve uncovering problem.*

1. Initialization step:
   - Initializing $\delta_1(i) = b_i(o_1)\pi_i$ for all $1 \le i \le n$
   - Initializing $q_1(i) = 0$ for all $1 \le i \le n$
2. Recurrence step:
   - Calculating all
   $$\delta_{t+1}(j) = \left(\max_i(\delta_t(i)a_{ij})\right)b_j(o_{t+1})$$
   for all $1 \le i,j \le n$ and $1 \le t \le T-1$ according to (10).
   - Keeping tracking optimal states
   $$q_{t+1}(j) = \underset{i}{\text{argmax}}(\delta_t(i)a_{ij})$$
   for all $1 \le j \le n$ and $1 \le t \le T-1$ according to (11).
3. State sequence backtracking step: The resulted state sequence $X = \{x_1, x_2, \ldots, x_T\}$ is determined as follows:
   - The last state $x_T = \text{argmax}_j(\delta_T(j))$
   - Previous states are determined by backtracking: $x_t = q_{t+1}(x_{t+1})$ for $t=T-1, t=T-1, \ldots, t=1$.

The total number of operations inside the Viterbi algorithm is $2n+(2n^2+n)(T-1)$ as follows:
- There are n multiplications for initializing n values $\delta_1(i)$ when each $\delta_1(i)$ requires 1 multiplication.
- There are $(2n^2+n)(T-1)$ operations over the recurrence step because there are $n(T-1)$ values $\delta_{t+1}(j)$ and each $\delta_{t+1}(j)$ requires n multiplications and n comparisons for maximizing $\max_i(\delta_t(i)a_{ij})$ plus 1 multiplication.
- There are n comparisons for constructing the state sequence X, $x_T = \max_j(q_T(j))$.

Inside $2n+(2n^2+n)(T-1)$ operations, there are $n+(n^2+n)(T-1)$ multiplications and $n^2(T-1)+n$ comparisons. The number of operations with regard to Viterbi algorithm is smaller than the number of operations with regard to individually optimal procedure when individually optimal procedure requires $(5n^2-n)(T-1)+2nT+n$ operations. Therefore, Viterbi algorithm is more effective than individually optimal procedure. Besides, individually optimal procedure does not reflect the whole probability of state sequence X given observation sequence O.

Going backing the weather HMM Δ whose parameters A, B, and $\prod$ are specified in tables 1, 2, and 3. Suppose humidity is *soggy* and *dry* in days 1 and 2, respectively. We apply Viterbi algorithm into solving the uncovering problem that finding out the optimal state sequence $X = \{x_1, x_2, x_3\}$ with regard to observation sequence $O = \{o_1=\varphi_4=soggy, o_2=\varphi_1=dry, o_3=\varphi_2=dryish\}$. According to initialization step of Viterbi algorithm, we have:

$\delta_1(1) = b_1(o_1 = \varphi_4)\pi_1 = b_{14}\pi_1 = 0.0165$
$\delta_1(2) = b_2(o_1 = \varphi_4)\pi_2 = b_{24}\pi_2 = 0.0825$
$\delta_1(3) = b_3(o_1 = \varphi_4)\pi_3 = b_{34}\pi_3 = 0.165$
$q_1(1) = q_1(2) = q_1(3) = 0$

According to recurrence step of Viterbi algorithm, we have:
$\delta_1(1)a_{11} = 0.00825$
$\delta_1(2)a_{21} = 0.02475$
$\delta_1(3)a_{31} = 0.04125$

$$\delta_2(1) = \left(\max_i\{\delta_1(i)a_{i1}\}\right)b_1(o_2 = \varphi_1)$$
$$= \left(\max_i\{\delta_1(i)a_{i1}\}\right)b_{11} = 0.04125 * 0.6$$
$$= 0.02475$$
$$q_2(1) = \underset{i}{\text{argmax}}\{\delta_1(i)a_{i1}\}$$
$$= \underset{i}{\text{argmax}}\{\delta_1(1)a_{11}, \delta_1(2)a_{21}, \delta_1(3)a_{31}\}$$
$$= s_3 = rainy$$

$\delta_1(1)a_{12} = 0.004125$
$\delta_1(2)a_{22} = 0.033$
$\delta_1(3)a_{32} = 0.04125$

$$\delta_2(2) = \left(\max_i\{\delta_1(i)a_{i2}\}\right)b_2(o_2 = \varphi_1)$$
$$= \left(\max_i\{\delta_1(i)a_{i2}\}\right)b_{21} = 0.04125 * 0.25$$
$$= 0.0103125$$
$$q_2(2) = \underset{i}{\text{argmax}}\{\delta_1(i)a_{i2}\}$$
$$= \underset{i}{\text{argmax}}\{\delta_1(1)a_{12}, \delta_1(2)a_{22}, \delta_1(3)a_{32}\}$$
$$= s_3 = rainy$$

$\delta_1(1)a_{13} = 0.004125$
$\delta_1(2)a_{23} = 0.02475$
$\delta_1(3)a_{33} = 0.0825$

$$\delta_2(3) = \left(\max_i\{\delta_1(i)a_{i3}\}\right)b_3(o_2 = \varphi_1)$$
$$= \left(\max_i\{\delta_1(i)a_{i3}\}\right)b_{31} = 0.0825 * 0.05$$
$$= 0.004125$$
$$q_2(2) = \underset{i}{\text{argmax}}\{\delta_1(i)a_{i3}\}$$
$$= \underset{i}{\text{argmax}}\{\delta_1(1)a_{13}, \delta_1(2)a_{23}, \delta_1(3)a_{33}\}$$
$$= s_3 = rainy$$

$\delta_2(1)a_{11} = 0.012375$
$\delta_2(2)a_{21} = 0.00309375$
$\delta_2(3)a_{31} = 0.00103125$

$$\delta_3(1) = \left(\max_i\{\delta_2(i)a_{i1}\}\right)b_1(o_3 = \varphi_2)$$
$$= \left(\max_i\{\delta_2(i)a_{i1}\}\right)b_{12} = 0.012375 * 0.2$$
$$= 0.002475$$
$$q_3(1) = \underset{i}{\text{argmax}}\{\delta_2(i)a_{i1}\}$$
$$= \underset{i}{\text{argmax}}\{\delta_2(1)a_{11}, \delta_2(2)a_{21}, \delta_2(3)a_{31}\}$$
$$= s_1 = sunny$$

$\delta_2(1)a_{12} = 0.0061875$
$\delta_2(2)a_{22} = 0.004125$
$\delta_2(3)a_{32} = 0.00103125$

$$\delta_3(2) = \left(\max_i\{\delta_2(i)a_{i2}\}\right)b_2(o_3 = \varphi_2)$$
$$= \left(\max_i\{\delta_2(i)a_{i2}\}\right)b_{22}$$
$$= 0.0061875 * 0.25 = 0.001546875$$
$$q_3(2) = \underset{i}{\text{argmax}}\{\delta_2(i)a_{i2}\}$$
$$= \underset{i}{\text{argmax}}\{\delta_2(1)a_{12}, \delta_2(2)a_{22}, \delta_2(3)a_{32}\}$$
$$= s_1 = sunny$$

$\delta_2(1)a_{13} = 0.0061875$
$\delta_2(2)a_{23} = 0.00309375$
$\delta_2(3)a_{33} = 0.0020625$

$$\delta_3(3) = \left(\max_i\{\delta_2(i)a_{i3}\}\right)b_3(o_3 = \varphi_2)$$
$$= \left(\max_i\{\delta_2(i)a_{i3}\}\right)b_{32}$$
$$= 0.0061875 * 0.1 = 0.00061875$$
$$q_3(3) = \text{argmax}_i\{\delta_2(i)a_{i3}\}$$
$$= \text{argmax}_i\{\delta_2(1)a_{13}, \delta_2(2)a_{23}, \delta_2(3)a_{33}\}$$
$$= s_1 = sunny$$

According to state sequence backtracking of Viterbi algorithm, we have:

$$x_3 = \text{argmax}_j\{\delta_3(j)\} = \text{argmax}_j\{\delta_3(1), \delta_3(2), \delta_3(3)\} = s_1$$
$$= sunny$$
$$x_2 = q_3(x_3 = s_1) = q_3(1) = s_1 = sunny$$
$$x_1 = q_2(x_2 = s_1) = q_2(1) = s_3 = rainy$$

As a result, the optimal state sequence is $X = \{x_1=rainy, x_2=sunny, x_3=sunny\}$. The result from the Viterbi algorithm is the same to the one from aforementioned individually optimal procedure described in table 6.

The uncovering problem is now described thoroughly in this section. Successive section will mention the learning problem of HMM which is the main subject of this tutorial.

4. HMM Learning Problem

The learning problem is to adjust parameters such as initial state distribution $\prod$, transition probability matrix A, and observation probability matrix B so that given HMM Δ gets more appropriate to an observation sequence $O = \{o_1, o_2,\ldots, o_T\}$ with note that Δ is represented by these parameters. In other words, the learning problem is to adjust parameters by maximizing probability of observation sequence O, as follows:

$$(A, B, \Pi) = \text{argmax}_{A,B,\Pi} P(O|\Delta)$$

The Expectation Maximization (EM) algorithm is applied successfully into solving HMM learning problem, which is equivalently well-known Baum-Welch algorithm [3]. Successive sub-section 4.1 describes EM algorithm in detailed before going into Baum-Welch algorithm.

4.1. EM Algorithm

Expectation Maximization (EM) is effective parameter estimator in case that incomplete data is composed of two parts: observed part and missing (or hidden) part. EM is iterative algorithm that improves parameters after iterations until reaching optimal parameters. Each iteration includes two steps: E(xpectation) step and M(aximization) step. In E-step the missing data is estimated based on observed data and current estimate of parameters; so the lower-bound of likelihood function is computed by the expectation of complete data. In M-step new estimates of parameters are determined by maximizing the lower-bound. Please see document [5] for short tutorial of EM. This sub-section focuses on practice general EM algorithm; the theory of EM algorithm is described comprehensively in article "Maximum Likelihood from Incomplete Data via the EM algorithm" by authors [6].

Suppose O and X are observed data and missing (hidden) data, respectively. Note O and X can be represented in any form such as discrete values, scalar, integer number, real number, vector, list, sequence, sample, and matrix. Let Θ represent parameters of probability distribution. Concretely, Θ includes initial state distribution $\prod$, transition probability matrix A, and observation probability matrix B inside HMM. In other words, Θ represents HMM Δ itself. EM algorithm aims to estimate Θ by finding out which $\widehat{\Theta}$ maximizes the likelihood function $L(\Theta) = P(O|\Theta)$.

$$\widehat{\Theta} = \text{argmax}_\Theta L(\Theta) = \text{argmax}_\Theta P(O|\Theta)$$

Where $\widehat{\Theta}$ is the optimal estimate of parameters which is called usually *parameter estimate*. Because the likelihood function is product of factors, it is replaced by the log-likelihood function $LnL(\Theta)$ that is natural logarithm of the likelihood function $L(\Theta)$, for convenience. We have:

$$\widehat{\Theta} = \text{argmax}_\Theta LnL(\Theta) = \text{argmax}_\Theta ln(L(\Theta))$$
$$= \text{argmax}_\Theta ln(P(O|\Theta))$$

Where,

$$LnL(\Theta) = ln(L(\Theta)) = ln(P(O|\Theta))$$

The method finding out the parameter estimate $\widehat{\Theta}$ by maximizing the log-likelihood function is called maximum likelihood estimation (MLE). Of course, EM algorithm is based on MLE.

Suppose the current parameter is Θ_t after the tth iteration. Next we must find out the new estimate $\widehat{\Theta}$ that maximizes the next log-likelihood function $LnL(\Theta)$. In other words it maximizes the deviation between current log-likelihood $LnL(\Theta_t)$ and next log-likelihood $LnL(\Theta)$ with regard to Θ.

$$\widehat{\Theta} = \text{argmax}_\Theta (LnL(\Theta) - LnL(\Theta_t)) = \text{argmax}_\Theta (Q(\Theta, \Theta_t))$$

Where $Q(\Theta, \Theta_t) = LnL(\Theta) - LnL(\Theta_t)$ is the deviation between current log-likelihood $LnL(\Theta_t)$ and next log-likelihood $LnL(\Theta)$ with note that $Q(\Theta, \Theta_t)$ is function of Θ when Θ_t was determined.

Suppose the total probability of observed data can be determined by marginalizing over missing data:

$$P(O|\Theta) = \sum_X P(O|X, \Theta)P(X|\Theta)$$

The expansion of $P(O|\Theta)$ is total probability rule [4, p. 101]. The deviation $Q(\Theta, \Theta_t)$ is re-written:

$$Q(\Theta, \Theta_t) = LnL(\Theta) - LnL(\Theta_t)$$
$$= ln(P(O|\Theta)) - ln(P(O|\Theta_t))$$
$$= ln\left(\sum_X P(O|X, \Theta)P(X|\Theta)\right) - ln(P(O|\Theta_t))$$
$$= ln\left(\sum_X P(O, X|\Theta)\right) - ln(P(O|\Theta_t))$$

(Due to multiplication rule [4, p. 100])

$$= ln\left(\sum_X P(X|O,\Theta_t)\frac{P(O,X|\Theta)}{P(X|O,\Theta_t)}\right) - ln\big(P(O|\Theta_t)\big)$$

Because hidden X is the complete set of mutually exclusive variables, the sum of conditional probabilities of X is equal to 1 given O and Θ_t.

$$\sum_X P(X|O,\Theta_t) = 1$$

Applying Jensen's inequality [5, pp. 3-4]

$$ln\left(\sum_i \lambda_i x_i\right) \geq \sum_i \lambda_i ln(x_i) \text{ where } \sum_i \lambda_i = 1$$

into deviation $Q(\Theta,\Theta_t)$, we have:

$$Q(\Theta,\Theta_t) \geq \left(\sum_X P(X|O,\Theta_t)ln\left(\frac{P(O,X|\Theta)}{P(X|O,\Theta_t)}\right)\right)$$
$$- ln\big(P(O|\Theta_t)\big)$$
$$= \left(\sum_X P(X|O,\Theta_t)\big(ln(P(O,X|\Theta)) - ln(P(X|O,\Theta_t))\big)\right)$$
$$- ln\big(P(O|\Theta_t)\big)$$
$$= \left(\sum_X P(X|O,\Theta_t)ln\big(P(O,X|\Theta)\big)\right)$$
$$- \left(\sum_X P(X|O,\Theta_t)ln\big(P(X|O,\Theta_t)\big)\right)$$
$$- ln\big(P(O|\Theta_t)\big)$$
$$= \sum_X P(X|O,\Theta_t)ln\big(P(O,X|\Theta)\big) + C$$

Where,

$$C = -\left(\sum_X P(X|O,\Theta_t)ln\big(P(X|O,\Theta_t)\big)\right) - ln\big(P(O|\Theta_t)\big)$$

Because C is constant with regard to Θ, it is possible to eliminate C in order to simplify the optimization criterion as follows:

$$\widehat{\Theta} = \underset{\Theta}{\text{argmax}}\big(Q(\Theta,\Theta_t)\big)$$
$$\approx \underset{\Theta}{\text{argmax}}\left(\sum_X P(X|O,\Theta_t)ln\big(P(O,X|\Theta)\big)\right.$$
$$\left. - C\right)$$
$$= \underset{\Theta}{\text{argmax}} \sum_X P(X|O,\Theta_t)ln\big(P(O,X|\Theta)\big)$$

The expression $\sum_X P(X|O,\Theta_t)ln\big(P(O,X|\Theta)\big)$ is essentially expectation of $ln\big(P(O,X|\Theta)\big)$ given conditional probability distribution $P(X|O,\Theta_t)$ when $P(X|O,\Theta_t)$ is totally determined. Let $E_{X|O,\Theta_t}\{ln\big(P(O,X|\Theta)\big)\}$ denote this conditional expectation, equation (12) specifies EM optimization criterion for determining the parameter estimate, which is the most important aspect of EM algorithm.

$$\widehat{\Theta} = \text{argmax}_\Theta E_{X|O,\Theta_t}\{ln\big(P(O,X|\Theta)\big)\} \qquad (12)$$

Where,

$$E_{X|O,\Theta_t}\{ln\big(P(O,X|\Theta)\big)\} = \sum_X P(X|O,\Theta_t)ln\big(P(O,X|\Theta)\big)$$

If $P(X|O,\Theta_t)$ is continuous density function, the continuous version of this conditional expectation is:

$$E_{X|O,\Theta_t}\{ln\big(P(O,X|\Theta)\big)\} = \int_X P(X|O,\Theta_t)ln\big(P(O,X|\Theta)\big)$$

Finally, the EM algorithm is described in table 8.

Table 8. *General EM algorithm.*

Starting with initial parameter Θ_0, each iteration in EM algorithm has two steps:
    1. *E-step*: computing the conditional expectation $E_{X|O,\Theta_t}\{ln\big(P(O,X|\Theta)\big)\}$ based on the current parameter Θ_t according to (12).
    2. *M-step*: finding out the estimate $\widehat{\Theta}$ that maximizes such conditional expectation. The next parameter Θ_{t+1} is assigned by the estimate $\widehat{\Theta}$, we have:
 $$\Theta_{t+1} = \widehat{\Theta}$$ Of course Θ_{t+1} becomes current parameter for next iteration. How to maximize the conditional expectation is optimization problem which is dependent on applications. For example, the popular method to solve optimization problem is Lagrangian duality [7, p. 8].
 EM algorithm stops when it meets the terminating condition, for example, the difference of current parameter Θ_t and next parameter Θ_{t+1} is smaller than some pre-defined threshold ε.
 $$|\Theta_{t+1} - \Theta_t| < \varepsilon$$ In addition, it is possible to define a custom terminating condition.

General EM algorithm is simple but please pay attention to the concept of lower-bound and what the essence of EM is. Recall that the next log-likelihood function $LnL(\Theta)$ is current likelihood function $LnL(\Theta_t)$ plus the deviation $Q(\Theta,\Theta_t)$. We have:

$$LnL(\Theta) = LnL(\Theta_t) + Q(\Theta,\Theta_t)$$
$$\geq LnL(\Theta_t) + E_{X|O,\Theta_t}\{ln\big(P(O,X|\Theta)\big)\} + C$$

Where,

$$C = -\left(\sum_X P(X|O,\Theta_t)ln\big(P(X|O,\Theta_t)\big)\right) - ln\big(P(O|\Theta_t)\big)$$

Let $lb(\Theta,\Theta_t)$ denote the lower-bound of the log-likelihood function $LnL(\Theta)$ given current parameter Θ_t [5, pp. 7-8]. The lower-bound $lb(\Theta,\Theta_t)$ is the function of Θ as specified by (13):

$$lb(\Theta,\Theta_t) = LnL(\Theta_t) + E_{X|O,\Theta_t}\{ln\big(P(O,X|\Theta)\big)\} + C \quad (13)$$

Determining $lb(\Theta,\Theta_t)$ is to calculate the EM conditional expectation $E_{X|O,\Theta_t}\{ln\big(P(O,X|\Theta)\big)\}$ because terms $LnL(\Theta_t)$ and C were totally determined. The lower-bound $lb(\Theta,\Theta_t)$ has a feature where its evaluation at $\Theta = \Theta_t$ equals the log-likelihood function $LnL(\Theta)$.

$$lb(\Theta_t,\Theta_t) = LnL(\Theta_t)$$

In fact,
$$lb(\Theta, \Theta_t) = LnL(\Theta_t) + E_{X|O,\Theta_t}\{ln(P(O, X|\Theta))\} + C$$

$$= LnL(\Theta_t) + \left(\sum_X P(X|O, \Theta_t)ln(P(O, X|\Theta))\right)$$
$$- \left(\sum_X P(X|O, \Theta_t)ln(P(X|O, \Theta_t))\right)$$
$$- ln(P(O|\Theta_t))$$

$$= LnL(\Theta_t) + \left(\sum_X P(X|O, \Theta_t)ln\left(\frac{P(O, X|\Theta)}{P(X|O, \Theta_t)}\right)\right)$$
$$- ln(P(O|\Theta_t))$$

$$= LnL(\Theta_t) + \left(\sum_X P(X|O, \Theta_t)ln\left(\frac{P(X|O, \Theta)P(O|\Theta)}{P(X|O, \Theta_t)}\right)\right)$$
$$- ln(P(O|\Theta_t))$$

(Due to multiplication rule [4, p. 100])

It implies
$$lb(\Theta_t, \Theta_t)$$
$$= LnL(\Theta_t) + \left(\sum_X P(X|O, \Theta_t)ln\left(\frac{P(X|O, \Theta_t)P(O|\Theta_t)}{P(X|O, \Theta_t)}\right)\right)$$
$$- ln(P(O|\Theta_t))$$
$$= LnL(\Theta_t) + \left(\sum_X P(X|O, \Theta_t)ln(P(O|\Theta_t))\right)$$
$$- ln(P(O|\Theta_t))$$
$$= LnL(\Theta_t) + ln(P(O|\Theta_t))\sum_X P(X|O, \Theta_t) - ln(P(O|\Theta_t))$$
$$= LnL(\Theta_t) + ln(P(O|\Theta_t)) - ln(P(O|\Theta_t))$$
$$\left(\text{due to } \sum_X P(X|O, \Theta_t) = 1\right)$$
$$= LnL(\Theta_t)$$

Fig. 4. [8, p. 7] shows relationship between the log-likelihood function $LnL(\Theta)$ and its lower-bound $lb(\Theta, \Theta_t)$.
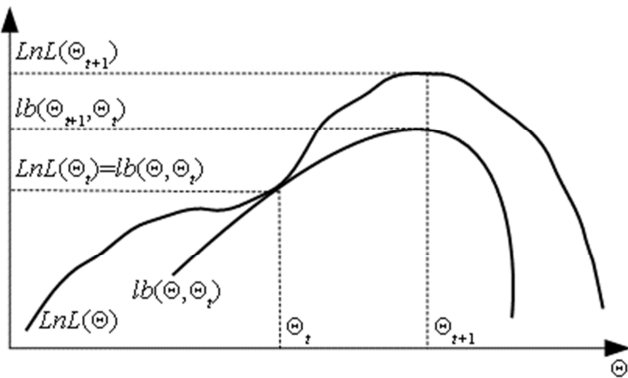

Figure 4. *Relationship between the log-likelihood function and its lower-bound.*

The essence of maximizing the deviation $Q(\Theta, \Theta_t)$ is to maximize the lower-bound $lb(\Theta, \Theta_t)$ with respect to Θ. For each iteration the new lower-bound and its maximum are computed based on previous lower-bound. A single iteration in EM algorithm can be understood as below:

1. E-step: the new lower-bound $lb(\Theta, \Theta_t)$ is determined based on current parameter Θ_t according to (13). Of course, determining $lb(\Theta, \Theta_t)$ is to calculate the EM conditional expectation $E_{X|O,\Theta_t}\{ln(P(O, X|\Theta))\}$.
2. M-step: finding out the estimate $\widehat{\Theta}$ so that $lb(\Theta, \Theta_t)$ reaches maximum at $\widehat{\Theta}$. The next parameter Θ_{t+1} is assigned by the estimate $\widehat{\Theta}$, we have:

$$\Theta_{t+1} = \widehat{\Theta}$$

Of course Θ_{t+1} becomes current parameter for next iteration. Note, maximizing $lb(\Theta, \Theta_t)$ is to maximize the EM conditional expectation $E_{X|O,\Theta_t}\{ln(P(O, X|\Theta))\}$.

In general, it is easy to calculate the EM expectation $E_{X|O,\Theta_t}\{ln(P(O, X|\Theta))\}$ but finding out the estimate $\widehat{\Theta}$ based on maximizing such expectation is complicated optimization problem. It is possible to state that the essence of EM algorithm is to determine the estimate $\widehat{\Theta}$. Now the EM algorithm is introduced with full of details. How to apply it into solving HMM learning problem is described in successive sub-section.

4.2. Applying EM Algorithm into Solving Learning Problem

Now going back the HMM learning problem, the EM algorithm is applied into solving this problem, which is equivalently well-known Baum-Welch algorithm [3]. The parameter Θ becomes the HMM model $\Delta = (A, B, \prod)$. Recall that the learning problem is to adjust parameters by maximizing probability of observation sequence O, as follows:

$$\widehat{\Delta} = (\widehat{A}, \widehat{B}, \widehat{\Pi}) = (\widehat{a}_{ij}, \widehat{b}_j(k), \widehat{\pi}_j) = \underset{\Delta}{\text{argmax}}\, P(O|\Delta)$$

Where $\widehat{a}_{ij}$, $\widehat{b}_j(k)$, $\widehat{\pi}_j$ are parameter estimates and so, the purpose of HMM learning problem is to determine them.

The observation sequence $O = \{o_1, o_2, …, o_T\}$ and state sequence $X = \{x_1, x_2, …, x_T\}$ are observed data and missing (hidden) data within context of EM algorithm, respectively. Note O and X is now represented in sequence. According to EM algorithm, the parameter estimate $\widehat{\Delta}$ is determined as follows:

$$\widehat{\Delta} = (\widehat{a}_{ij}, \widehat{b}_j(k), \widehat{\pi}_j) = \underset{\Delta}{\text{argmax}}\, E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\}$$

Where $\Delta_r = (A_r, B_r, \prod_r)$ is the known parameter at the current iteration. Note that we use notation Δ_r instead of popular notation Δ_t in order to distinguish iteration indices of EM algorithm from time points inside observation sequence O and state sequence X. The EM conditional expectation in accordance with HMM is:

$$E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\} = \sum_X P(X|O, \Delta_r)ln(P(O, X|\Delta))$$
$$= \sum_X P(X|O, \Delta_r)ln(P(O|X, \Delta)P(X|\Delta))$$
$$= \sum_X P(X|O, \Delta_r)ln(P(o_1, o_2, …, o_T|x_1, x_2, …, x_T, \Delta)$$
$$* P(x_1, x_2, …, x_T|\Delta))$$

$$= \sum_X P(X|O,\Delta_r)ln\big(P(o_1|x_1,x_2,\dots,x_T,\Delta)$$

$$* P(o_2|x_1,x_2,\dots,x_T,\Delta) * \cdots$$
$$* P(o_T|x_1,x_2,\dots,x_T,\Delta)$$
$$* P(x_1,x_2,\dots,x_T|\Delta)\big)$$

(Because observations $o_1, o_2,\dots, o_T$ are mutually independent)

$$= \sum_X P(X|O,\Delta_r)ln\big(P(o_1|x_1,\Delta) * P(o_2|x_2,\Delta) * \cdots$$

$$* P(o_T|x_T,\Delta) * P(x_1,x_2,\dots,x_T|\Delta)\big)$$

(Because each observations o_t is only dependent on state x_t)

$$= \sum_X P(X|O,\Delta_r)ln\left(\left(\prod_{t=1}^T P(o_t|x_t,\Delta)\right)\right.$$
$$\left.* P(x_1,x_2,\dots,x_T|\Delta)\right)$$

$$= \sum_X P(X|O,\Delta_r)ln\left(\left(\prod_{t=1}^T P(o_t|x_t,\Delta)\right)\right.$$
$$* P(x_T|x_1,x_2,\dots,x_{T-1},\Delta)$$
$$\left.* P(x_1,x_2,\dots,x_{T-1}|\Delta)\right)$$

$$= \sum_X P(X|O,\Delta_r)ln\left(\left(\prod_{t=1}^T P(o_t|x_t,\Delta)\right)* P(x_T|x_{T-1},\Delta)\right.$$
$$\left.* P(x_1,x_2,\dots,x_{T-1}|\Delta)\right)$$

(Because each state x_t is only dependent on previous state x_{t-1})

$$= \sum_X P(X|O,\Delta_r)ln\left(\left(\prod_{t=1}^T P(o_t|x_t,\Delta)\right)* P(x_T|x_{T-1},\Delta)\right.$$
$$* P(x_{T-1}|x_{T-2},\Delta) * \cdots * P(x_2|x_1,\Delta)$$
$$\left.* P(x_1|\Delta)\right)$$

(Due to recurrence on probability $P(x_1, x_2,\dots, x_t)$)

$$= \sum_X P(X|O,\Delta_r)ln\left(\left(\prod_{t=1}^T P(o_t|x_t,\Delta)\right)\right.$$
$$\left.*\left(\prod_{t=2}^T P(x_t|x_{t-1},\Delta)\right)* P(x_1|\Delta)\right)$$

It is conventional that $P(x_1|x_0,\Delta) = P(x_1|\Delta)$ where x_0 is pseudo-state, equation (14) specifies general EM conditional expectation for HMM:

$$E_{X|O,\Delta_r}\{ln\big(P(O,X|\Delta)\big)\} =$$
$$\sum_X P(X|O,\Delta_r)ln(\prod_{t=1}^T P(x_t|x_{t-1},\Delta)P(o_t|x_t,\Delta)) =$$
$$\sum_X P(X|O,\Delta_r)\sum_{t=1}^T \big(ln\big(P(x_t|x_{t-1},\Delta)\big) + ln\big(P(o_t|x_t,\Delta)\big)\big)$$
$$(14)$$

Let $I\big(x_{t-1} = s_i, x_t = s_j\big)$ and $I\big(x_t = s_j, o_t = \varphi_k\big)$ are

two index functions so that

$$I\big(s_i = x_{t-1}, s_j = x_t\big) = \begin{cases}1 \text{ if } s_i = x_{t-1} \text{ and } s_j = x_t \\ 0 \text{ otherwise}\end{cases}$$

$$I\big(x_t = s_j, o_t = \varphi_k\big) = \begin{cases}1 \text{ if } x_t = s_j \text{ and } o_t = \varphi_k \\ 0 \text{ otherwise}\end{cases}$$

We have:
$$E_{X|O,\Delta_r}\{ln\big(P(O,X|\Delta)\big)\}$$
$$= \sum_X P(X|O,\Delta_r)\left(\sum_{t=1}^T ln\big(P(x_t|x_{t-1},\Delta)\big)\right.$$
$$\left.+ \sum_{t=1}^T ln\big(P(o_t|x_t,\Delta)\big)\right)$$
$$= \sum_X P(X|O,\Delta_r)\left(\sum_{i=1}^n\sum_{j=1}^n\sum_{t=1}^T I\big(x_{t-1} = s_i, x_t\right.$$
$$= s_j\big)ln\big(P(s_j|s_i,\Delta)\big)$$
$$+ \sum_{j=1}^n\sum_{k=1}^m\sum_{t=1}^T I\big(x_t = s_j, o_t$$
$$\left.= \varphi_k\big)ln\big(P(\varphi_k|s_j,\Delta)\big)\right)$$
$$= \sum_X P(X|O,\Delta_r)\left(\sum_{i=1}^n\sum_{j=1}^n\sum_{t=1}^T I\big(x_{t-1} = s_i, x_t = s_j\big)ln\big(a_{ij}\big)\right.$$
$$+ \sum_{j=1}^n\sum_{k=1}^m\sum_{t=1}^T I\big(x_t = s_j, o_t$$
$$\left.= \varphi_k\big)ln\big(b_j(k)\big)\right)$$

Because of the convention $P(x_1|x_0,\Delta) = P(x_1|\Delta)$, matrix $\prod$ is degradation case of matrix A at time point $t=1$. In other words, the initial probability π_j is equal to the transition probability a_{ij} from pseudo-state x_0 to state $x_1 = s_j$.

$$P\big(x_1 = s_j|x_0,\Delta\big) = P\big(x_1 = s_j|\Delta\big) = \pi_j$$

Note that $n=|S|$ is the number of possible states and $m=|\Phi|$ is the number of possible observations.

Shortly, the EM conditional expectation for HMM is specified by (15).

$$E_{X|O,\Delta_r}\{ln\big(P(O,X|\Delta)\big)\} =$$
$$\sum_X P(X|O,\Delta_r)\Big(\sum_{i=1}^n\sum_{j=1}^n\sum_{t=1}^T I\big(x_{t-1} = s_i, x_t =$$
$$s_j\big)ln\big(a_{ij}\big) + \sum_{j=1}^n\sum_{k=1}^m\sum_{t=1}^T I\big(x_t = s_j, o_t = \varphi_k\big)ln\big(b_j(k)\big)\Big)$$
$$(15)$$

Where,

$$I\big(x_{t-1} = s_i, x_t = s_j\big) = \begin{cases}1 \text{ if } x_{t-1} = s_i \text{ and } x_t = s_j \\ 0 \text{ otherwise}\end{cases}$$

$$I\big(x_t = s_j, o_t = \varphi_k\big) = \begin{cases}1 \text{ if } x_t = s_j \text{ and } o_t = \varphi_k \\ 0 \text{ otherwise}\end{cases}$$

$$P(x_1 = s_j | x_0, \Delta) = P(x_1 = s_j | \Delta) = \pi_j$$

Note that the conditional expectation $E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\}$ is function of Δ. There are two constraints for HMM as follows:

$$\sum_{j=1}^{n} a_{ij} = 1, \forall i = \overline{1, n}$$

$$\sum_{k=1}^{m} b_j(k) = 1, \forall k = \overline{1, m}$$

Maximizing $E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\}$ with subject to these constraints is optimization problem that is solved by Lagrangian duality theorem [7, p. 8]. Original optimization problem mentions minimizing target function but it is easy to infer that maximizing target function shares the same methodology. Let $l(\Delta, \lambda, \mu)$ be Lagrangian function constructed from $E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\}$ together with these constraints [9, p. 9], we have (16) for specifying HMM Lagrangian function as follows:

$$l(\Delta, \lambda, \mu) = l(a_{ij}, b_j(k), \lambda_i, \mu_j) = E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\} + \sum_{i=1}^{n} \lambda_i\left(1 - \sum_{j=1}^{n} a_{ij}\right) + \sum_{j=1}^{n} \mu_j\left(1 - \sum_{k=1}^{m} b_j(k)\right) \quad (16)$$

Where λ is n-component vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ and μ is m-component vector $\mu = (\mu_1, \mu_2, \ldots, \mu_m)$. Factors $\lambda_i \geq 0$ and $\mu_j \geq 0$ are called Lagrange multipliers or Karush-Kuhn-Tucker multipliers [10] or dual variables. The expectation $E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\}$ is specified by (15).

The parameter estimate $\widehat{\Delta}$ is extreme point of the Lagrangian function. According to Lagrangian duality theorem [11, p. 216] [7, p. 8], we have:

$$\widehat{\Delta} = (\widehat{A}, \widehat{B}) = (\hat{a}_{ij}, \hat{b}_j(k)) = \underset{A,B}{\operatorname{argmax}}\, l(\Delta, \lambda, \mu)$$

$$(\hat{\lambda}, \hat{\mu}) = \underset{\lambda,\mu}{\operatorname{argmin}}\, l(\Delta, \lambda, \mu)$$

The parameter estimate $\widehat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k))$ is determined by setting partial derivatives of $l(\Delta, \lambda, \mu)$ with respect to a_{ij} and $b_j(k)$ to be zero. The partial derivative of $l(\Delta, \lambda, \mu)$ with respect to a_{ij} is:

$$\frac{\partial l(\Delta, \lambda, \mu)}{\partial a_{ij}} = \frac{\partial E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\}}{\partial a_{ij}}$$

$$+ \frac{\partial}{\partial a_{ij}}\left(\sum_{i=1}^{n} \lambda_i\left(1 - \sum_{j=1}^{n} a_{ij}\right)\right)$$

$$+ \frac{\partial}{\partial a_{ij}}\left(\sum_{j=1}^{n} \mu_j\left(1 - \sum_{k=1}^{m} b_j(k)\right)\right)$$

$$= \frac{\partial E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\}}{\partial a_{ij}} - \lambda_i$$

$$= \frac{\partial}{\partial a_{ij}}\left(\sum_{X} P(X|O, \Delta_r)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} I(x_{t-1} = s_i, x_t = s_j)ln(a_{ij}) + \sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{t=1}^{T} I(x_t = s_j, o_t = \varphi_k)ln(b_j(k))\right)\right) - \lambda_i$$

$$= \left(\sum_{X} P(X|O, \Delta_r)\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} I(x_{t-1} = s_i, x_t = s_j)\frac{\partial ln(a_{ij})}{\partial a_{ij}}\right) - \lambda_i$$

$$= \left(\sum_{X} P(X|O, \Delta_r)\sum_{t=1}^{T} I(x_{t-1} = s_i, x_t = s_j)\frac{1}{a_{ij}}\right) - \lambda_i$$

$$= \frac{1}{a_{ij}}\left(\sum_{X} P(X|O, \Delta_r)\sum_{t=1}^{T} I(x_{t-1} = s_i, x_t = s_j)\right) - \lambda_i$$

Setting the partial derivative $\frac{\partial l(\Delta, \lambda, \mu)}{\partial a_{ij}}$ to be zero:

$$\frac{\partial l(\Delta, \lambda, \mu)}{\partial a_{ij}} = 0 \Leftrightarrow \frac{1}{a_{ij}}\left(\sum_{X} P(X|O, \Delta_r)\sum_{t=1}^{T} I(x_{t-1} = s_i, x_t = s_j)\right) - \lambda_i = 0$$

The parameter estimate $\hat{a}_{ij}$ is solution of equation $\frac{\partial l(\Delta, \lambda, \mu)}{\partial a_{ij}} = 0$, we have:

$$\hat{a}_{ij} = \frac{\sum_{X} P(X|O, \Delta_r)\sum_{t=1}^{T} I(x_{t-1} = s_i, x_t = s_j)}{\lambda_i}$$

It is required to estimate the Lagrange multiplier λ_i. The multiplier estimate $\hat{\lambda}_i$ is determined by setting the partial derivative of $l(\Delta, \lambda, \mu)$ with respect to λ_i to be zero as follows:

$$\frac{\partial l(\Delta, \lambda, \mu)}{\partial \lambda_i} = 0$$

$$\Rightarrow \frac{\partial E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\}}{\partial \lambda_i} + \frac{\partial}{\partial \lambda_i}\left(\sum_{i=1}^{n} \lambda_i\left(1 - \sum_{j=1}^{n} a_{ij}\right)\right)$$

$$+ \frac{\partial}{\partial \lambda_i}\left(\sum_{j=1}^{n} \mu_j\left(1 - \sum_{k=1}^{m} b_j(k)\right)\right) = 0$$

$$\Rightarrow 1 - \sum_{j=1}^{n} a_{ij} = 0$$

Substituting $\hat{a}_{ij}$ for a_{ij}, we have:

$$1 - \sum_{j=1}^{n} \hat{a}_{ij}$$

$$= 1 - \sum_{j=1}^{n} \frac{\sum_{X} P(X|O, \Delta_r)\sum_{t=1}^{T} I(x_{t-1} = s_i, x_t = s_j)}{\lambda_i}$$

$$= 1 - \frac{1}{\lambda_i}\sum_{j=1}^{n}\sum_{X}P(X|O,\Delta_r)\sum_{t=1}^{T}I\big(x_{t-1}=s_i, x_t=s_j\big) = 0$$

It implies:

$$\hat{\lambda}_i = \sum_{j=1}^{n}\sum_{X}P(X|O,\Delta_r)\sum_{t=1}^{T}I\big(x_{t-1}=s_i, x_t=s_j\big)$$

$$= \sum_{X}P(X|O,\Delta_r)\sum_{t=1}^{T}\sum_{j=1}^{n}I\big(x_{t-1}=s_i, x_t=s_j\big)$$

$$= \sum_{X}P(X|O,\Delta_r)\sum_{t=1}^{T}I(x_{t-1}=s_i)$$

Where, $I(s_i = x_{t-1})$ is index function.

$$I(x_{t-1}=s_i) = \begin{cases}1 \text{ if } x_{t-1}=s_i\\0 \text{ otherwise}\end{cases}$$

Substituting $\hat{\lambda}_i$ for λ_i inside

$$\hat{a}_{ij} = \frac{\sum_{X}P(X|O,\Delta_r)\sum_{t=1}^{T}I\big(x_{t-1}=s_i, x_t=s_j\big)}{\lambda_i}$$

We have:

$$\hat{a}_{ij} = \frac{\sum_{X}P(X|O,\Delta_r)\sum_{t=1}^{T}I\big(x_{t-1}=s_i, x_t=s_j\big)}{\hat{\lambda}_i}$$

$$= \frac{\sum_{X}P(X|O,\Delta_r)\sum_{t=1}^{T}I\big(x_{t-1}=s_i, x_t=s_j\big)}{\sum_{X}P(X|O,\Delta_r)\sum_{t=1}^{T}I(x_{t-1}=s_i)}$$

Evaluating the numerator, we have:

$$\sum_{X}P(X|O,\Delta_r)\sum_{t=1}^{T}I\big(x_{t-1}=s_i, x_t=s_j\big)$$

$$= \sum_{t=1}^{T}\sum_{X}I\big(x_{t-1}=s_i, x_t=s_j\big)P(X|O,\Delta_r)$$

$$= \sum_{t=1}^{T}\sum_{X}I\big(x_{t-1}=s_i, x_t = s_j\big)P(x_1,\dots,x_{t-1},x_t,\dots,x_T|O,\Delta_r)$$

$$= \sum_{t=1}^{T}P\big(x_{t-1}=s_i, x_t=s_j|O,\Delta_r\big)$$

(Due to total probability rule [4, p. 101])

$$= \sum_{t=1}^{T}\frac{P\big(O, x_{t-1}=s_i, x_t=s_j|\Delta_r\big)}{P(O|\Delta_r)}$$

(Due to multiplication rule [4, p. 100])

$$= \frac{1}{P(O|\Delta_r)}\sum_{t=1}^{T}P\big(O, x_{t-1}=s_i, x_t=s_j|\Delta_r\big)$$

Evaluating the denominator, we have:

$$\sum_{X}P(X|O,\Delta_r)\sum_{t=1}^{T}I(x_{t-1}=s_i)$$

$$= \sum_{t=1}^{T}\sum_{X}I(x_{t-1}=s_i)P(X|O,\Delta_r)$$

$$= \sum_{t=1}^{T}\sum_{X}I(x_{t-1}=s_i)P(x_1,\dots,x_{t-1},x_t,\dots,x_T|O,\Delta_r)$$

$$= \sum_{t=1}^{T}P(x_{t-1}=s_i|O,\Delta_r)$$

(Due to total probability rule [4, p. 101])

$$= \sum_{t=1}^{T}\frac{P(O, x_{t-1}=s_i|\Delta_r)}{P(O|\Delta_r)}$$

(Due to multiplication rule [4, p. 100])

$$= \frac{1}{P(O|\Delta_r)}\sum_{t=1}^{T}P(O, x_{t-1}=s_i|\Delta_r)$$

It implies

$$\hat{a}_{ij} = \frac{\sum_{X}P(X|O,\Delta_r)\sum_{t=1}^{T}I\big(x_{t-1}=s_i, x_t=s_j\big)}{\sum_{X}P(X|O,\Delta_r)\sum_{t=1}^{T}I(x_{t-1}=s_i)}$$

$$= \frac{\sum_{t=1}^{T}P\big(O, x_{t-1}=s_i, x_t=s_j|\Delta_r\big)}{\sum_{t=1}^{T}P(O, x_{t-1}=s_i|\Delta_r)}$$

Because of the convention $P(x_1|x_0,\Delta) = P(x_1|\Delta)$, the estimate $\hat{a}_{ij}$ is fixed as follows:

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T}P\big(O, x_{t-1}=s_i, x_t=s_j|\Delta_r\big)}{\sum_{t=2}^{T}P(O, x_{t-1}=s_i|\Delta_r)}$$

The estimate of initial probability $\hat{\pi}_j$ is known as specific estimate $\hat{a}_{ij}$ from pseudo-state x_0 to state $x_1=s_j$. It means that

$$\hat{\pi}_j = \frac{P(O, x_1=s_j|\Delta_r)}{\sum_{i=1}^{n}P(O, x_1=s_i|\Delta_r)}$$

Recall that the parameter estimate $\hat{\Delta} = \big(\hat{a}_{ij}, \hat{b}_j(k)\big)$ is determined by setting partial derivatives of $l(\Delta, \lambda, \mu)$ with respect to a_{ij} and $b_j(k)$ to be zero. The parameter estimate $\hat{a}_{ij}$ was determined. Now it is required to calculate the parameter estimate $\hat{b}_j(k)$. The partial derivative of Lagrangian function $l(\Delta, \lambda, \mu)$ with respect to $b_j(k)$ is:

$$\frac{\partial l(\Delta, \lambda, \mu)}{\partial b_j(k)} = \frac{\partial E_{X|O,\Delta_r}\big\{ln\big(P(O,X|\Delta)\big)\big\}}{\partial b_j(k)}$$

$$+ \frac{\partial}{\partial b_j(k)}\left(\sum_{i=1}^{n}\lambda_i\left(1 - \sum_{j=1}^{n}a_{ij}\right)\right)$$

$$+ \frac{\partial}{\partial b_j(k)}\left(\sum_{j=1}^{n}\mu_j\left(1 - \sum_{k=1}^{m}b_j(k)\right)\right)$$

$$= \frac{\partial E_{X|O,\Delta_r}\big\{ln\big(P(O,X|\Delta)\big)\big\}}{\partial b_j(k)} - \mu_j$$

$$= \frac{\partial}{\partial b_j(k)}\left(\sum_{X}P(X|O,\Delta_r)\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T}I\big(x_{t-1}=s_i, x_t\right.\right.$$

$$= s_j\big)ln(a_{ij})$$

$$+ \sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{t=1}^{T}I\big(x_t=s_j, o_t$$

$$\left.\left.= \varphi_k\big)ln\big(b_j(k)\big)\right)\right) - \mu_j$$

$$= \left(\sum_X P(X|O, \Delta_r) \sum_{j=1}^{n} \sum_{k=1}^{m} \sum_{t=1}^{T} I(x_t = s_j, o_t \right.$$

$$\left. = \varphi_k) \frac{\partial ln\left(b_j(k)\right)}{\partial b_j(k)} \right) - \mu_j$$

$$= \left(\sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j, o_t = \varphi_k) \frac{1}{b_j(k)} \right) - \mu_j$$

$$= \frac{1}{b_j(k)} \left(\sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j, o_t = \varphi_k) \right) - \mu_j$$

Setting the partial derivative $\frac{\partial l(\Delta, \lambda, \mu)}{\partial b_j(k)}$ to be zero:

$$\frac{\partial l(\Delta, \lambda, \mu)}{\partial b_j(k)} = 0 \Leftrightarrow \frac{1}{b_j(k)} \left(\sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j, o_t \right.$$

$$\left. = \varphi_k) \right) - \mu_j = 0$$

The parameter estimate $\hat{a}_{ij}$ is solution of equation $\frac{\partial l(\Delta, \lambda, \mu)}{\partial b_j(k)} = 0$, we have:

$$\hat{b}_j(k) = \frac{\sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j, o_t = \varphi_k)}{\mu_j}$$

It is required to estimate the Lagrange multiplier μ_j. The multiplier estimate $\hat{\mu}_j$ is determined by setting the partial derivative of $l(\Delta, \lambda, \mu)$ with respect to μ_j to be zero as follows:

$$\frac{\partial l(\Delta, \lambda, \mu)}{\partial \mu_j} = 0$$

$$\Rightarrow \frac{\partial E_{X|O, \Delta_r}\{ln\left(P(O, X|\Delta)\right)\}}{\partial \mu_j} + \frac{\partial}{\partial \mu_j} \left(\sum_{i=1}^{n} \lambda_i \left(1 - \sum_{j=1}^{n} a_{ij} \right) \right)$$

$$+ \frac{\partial}{\partial \mu_j} \left(\sum_{j=1}^{n} \mu_j \left(1 - \sum_{k=1}^{m} b_j(k) \right) \right) = 0$$

$$\Rightarrow 1 - \sum_{k=1}^{m} b_j(k) = 0$$

Substituting $\hat{b}_j(k)$ for $b_j(k)$ we have:

$$1 - \sum_{k=1}^{m} \hat{b}_j(k) = 1 - \sum_{k=1}^{m} \frac{\sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j, o_t = \varphi_k)}{\mu_j}$$

$$= 1$$

$$- \frac{1}{\mu_j} \sum_{k=1}^{m} \sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j, o_t = \varphi_k)$$

$$= 0$$

It implies:

$$\hat{\mu}_j = \sum_{k=1}^{m} \sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j, o_t = \varphi_k)$$

$$= \sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} \sum_{k=1}^{m} I(x_t = s_j, o_t = \varphi_k)$$

$$= \sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j)$$

Where, $I(s_j = x_t)$ is index function.

$$I(x_t = s_j) = \begin{cases} 1 \text{ if } x_t = s_j \\ 0 \text{ otherwise} \end{cases}$$

Substituting $\hat{\mu}_j$ for μ_j inside

$$\hat{b}_j(k) = \frac{\sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j, o_t = \varphi_k)}{\mu_j}$$

We have:

$$\hat{b}_j(k) = \frac{\sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j, o_t = \varphi_k)}{\hat{\mu}_j}$$

$$= \frac{\sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j, o_t = \varphi_k)}{\sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j)}$$

Evaluating the numerator, we have:

$$\sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j, o_t = \varphi_k)$$

$$= \sum_{t=1}^{T} \sum_X I(x_t = s_j, o_t = \varphi_k) P(X|O, \Delta_r)$$

$$= \sum_{t=1}^{T} \sum_X I(x_t = s_j, o_t = \varphi_k) P(x_1, \dots, x_t, \dots, x_T|O, \Delta_r)$$

$$= \sum_{\substack{t=1 \\ o_t=\varphi_k}}^{T} P(x_t = s_j|O, \Delta_r)$$

(Due to total probability rule [4, p. 101])

$$= \sum_{\substack{t=1 \\ o_t=\varphi_k}}^{T} \frac{P(O, x_t = s_j|\Delta_r)}{P(O|\Delta_r)}$$

(Due to multiplication rule [4, p. 100])

$$= \frac{1}{P(O|\Delta_r)} \sum_{\substack{t=1 \\ o_t=\varphi_k}}^{T} P(O, x_t = s_j|\Delta_r)$$

Note, the expression $\sum_{\substack{t=1 \\ o_t=\varphi_k}}^{T} P(O, x_t = s_j|\Delta_r)$ expresses the sum of probabilities $P(O, x_t = s_j|\Delta_r)$ over T time points with condition $o_t = \varphi_k$.

Evaluating the denominator, we have:

$$\sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j)$$

$$= \sum_{t=1}^{T} \sum_X I(x_t = s_j) P(X|O, \Delta_r)$$

$$= \sum_{t=1}^{T} \sum_X I(x_t = s_j) P(x_1, \dots, x_t, \dots, x_T|O, \Delta_r)$$

$$= \sum_{t=1}^{T} P(x_t = s_j|O, \Delta_r)$$

(Due to total probability rule [4, p. 101])

$$= \sum_{t=1}^{T} \frac{P(O, x_t = s_j|\Delta_r)}{P(O|\Delta_r)}$$

(Due to multiplication rule [4, p. 100])

$$= \frac{1}{P(O|\Delta_r)} \sum_{t=1}^{T} P(O, x_t = s_j | \Delta_r)$$

It implies

$$\hat{b}_j(k) = \frac{\sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j, o_t = \varphi_k)}{\sum_X P(X|O, \Delta_r) \sum_{t=1}^{T} I(x_t = s_j)}$$

$$= \frac{\sum_{\substack{t=1 \\ o_t=\varphi_k}}^{T} P(O, x_t = s_j | \Delta_r)}{\sum_{t=1}^{T} P(O, x_t = s_j | \Delta_r)}$$

In general, the parameter estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ is totally determined as follows:

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} P(O, x_{t-1} = s_i, x_t = s_j | \Delta_r)}{\sum_{t=2}^{T} P(O, x_{t-1} = s_i | \Delta_r)}$$

$$\hat{b}_j(k) = \frac{\sum_{\substack{t=1 \\ o_t=\varphi_k}}^{T} P(O, x_t = s_j | \Delta_r)}{\sum_{t=1}^{T} P(O, x_t = s_j | \Delta_r)}$$

$$\hat{\pi}_j = \frac{P(O, x_1 = s_j | \Delta_r)}{\sum_{i=1}^{n} P(O, x_1 = s_i | \Delta_r)}$$

As a convention, we use notation Δ instead of Δ_r for denoting known HMM at current iteration of EM algorithm. We have (17) for specifying HMM parameter estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ given current parameter $\Delta = (a_{ij}, b_j(k), \pi_j)$ as follows:

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} P(O, x_{t-1} = s_i, x_t = s_j | \Delta)}{\sum_{t=2}^{T} P(O, x_{t-1} = s_i | \Delta)}$$

$$\hat{b}_j(k) = \frac{\sum_{\substack{t=1 \\ o_t=\varphi_k}}^{T} P(O, x_t=s_j | \Delta)}{\sum_{t=1}^{T} P(O, x_t=s_j | \Delta)} \qquad (17)$$

$$\hat{\pi}_j = \frac{P(O, x_1 = s_j | \Delta)}{\sum_{i=1}^{n} P(O, x_1 = s_i | \Delta)}$$

The parameter estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ is the ultimate solution of the learning problem. As seen in (17), it is necessary to calculate probabilities $P(O, x_{t-1}=s_i, x_t=s_j)$ and $P(O, x_{t-1}=s_i)$ when other probabilities $P(O, x_t=s_j)$, $P(O, x_1=s_i)$, and $P(O, x_1=s_j)$ are represented by the joint probability γ_t specified by (7).

$$P(O, x_t = s_j | \Delta) = \gamma_t(j) = \alpha_t(j)\beta_t(j)$$

$$P(O, x_1 = s_i | \Delta) = \gamma_1(i) = \alpha_1(i)\beta_1(i)$$

$$P(O, x_1 = s_j | \Delta) = \gamma_1(j) = \alpha_1(j)\beta_1(j)$$

Let $\xi_t(i, j)$ is the joint probability that the stochastic process receives state s_i at time point $t-1$ and state s_j at time point t given observation sequence O [3, p. 264].

$$\xi_t(i, j) = P(O, x_{t-1} = s_i, x_t = s_j | \Delta)$$

Given forward variable α_t and backward variable β_t, if

$t \geq 2$, we have:

$\alpha_{t-1}(i)a_{ij}b_j(o_t)\beta_t(j)$

$= P(o_1, o_2, ..., o_t, x_{t-1} = s_i | \Delta) * P(x_t = s_j | x_{t-1} = s_i)$
$\qquad * b_j(o_t) * \beta_t(j)$

$= P(o_1, o_2, ..., o_t | x_{t-1} = s_i, \Delta) * P(x_{t-1} = s_i | \Delta)$
$\qquad * P(x_t = s_j | x_{t-1} = s_i) * b_j(o_t) * \beta_t(j)$

(Due to multiplication rule [4, p. 100])

$= P(o_1, o_2, ..., o_t | x_{t-1} = s_i, \Delta) * P(x_t = s_j | x_{t-1} = s_i)$
$\qquad * P(x_{t-1} = s_i | \Delta) * b_j(o_t) * \beta_t(j)$

$= P(o_1, o_2, ..., o_t, x_t = s_j | x_{t-1} = s_i, \Delta) * P(x_{t-1} = s_i | \Delta)$
$\qquad * b_j(o_t) * \beta_t(j)$

(Because the partial observation sequence $\{o_1, o_2,..., o_t\}$ is independent from current state x_t given previous state x_{t-1})

$= P(o_1, o_2, ..., o_t, x_{t-1} = s_i, x_t = s_j | \Delta) * b_j(o_t) * \beta_t(j)$

$= P(o_1, o_2, ..., o_t, x_{t-1} = s_i | x_t = s_j, \Delta) * P(x_t = s_j | \Delta)$
$\qquad * b_j(o_t) * \beta_t(j)$

(Due to multiplication rule [4, p. 100])

$= P(o_1, o_2, ..., o_t, x_{t-1} = s_i | x_t = s_j, \Delta) * P(x_t = s_j | \Delta)$
$\qquad * P(o_t | x_t = s_j)$
$\qquad * P(o_{t+1}, o_{t+2}, ..., o_T | x_t = s_j, \Delta)$

$= P(o_1, o_2, ..., o_t, x_{t-1} = s_i | x_t = s_j, \Delta) * P(x_t = s_j | \Delta)$
$\qquad * P(o_t, o_{t+1}, o_{t+2}, ..., o_T | x_t = s_j, \Delta)$

(Because observations $o_t, o_{t+1}, o_{t+2},..., o_T$ are mutually independent)

$= P(o_1, o_2, ..., o_t, x_{t-1} = s_i | x_t = s_j, \Delta)$
$\qquad * P(o_t, o_{t+1}, o_{t+2}, ..., o_T | x_t = s_j, \Delta)$
$\qquad * P(x_t = s_j | \Delta)$

$= P(o_1, o_2, ..., o_t, o_{t+1}, o_{t+2}, ..., o_T, x_{t-1} = s_i | x_t = s_j, \Delta)$
$\qquad * P(x_t = s_j | \Delta)$

(Due to multiplication rule [4, p. 100])

$= P(o_1, o_2, ..., o_T, x_t = s_i, x_{t+1} = s_j | \Delta)$

(Due to multiplication rule [4, p. 100])

$= P(O, x_{t-1} = s_i, x_t = s_j | \Delta) = \xi_t(i, j)$

In general, equation (18) determines the joint probability $\xi_t(i, j)$ based on forward variable α_t and backward variable β_t.

$$\xi_t(i, j) = \alpha_{t-1}(i)a_{ij}b_j(o_t)\beta_t(j) \text{ where } t \geq 2 \quad (18)$$

Where forward variable α_t and backward variable β_t are calculated by previous recurrence equations (2) and (5).

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i)a_{ij} \right) b_j(o_{t+1})$$

$$\beta_t(i) = \sum_{j=1}^{n} a_{ij}b_j(o_{t+1})\beta_{t+1}(j)$$

Shortly, the joint probability $\xi_t(i, j)$ is constructed from forward variable and backward variable, as seen in fig. 5 [3, p. 264].
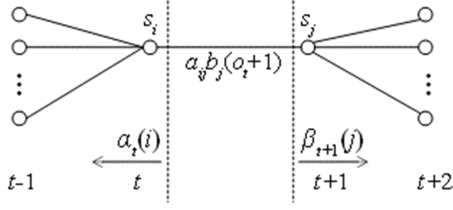
Figure 5. *Construction of the joint probability $\xi_t(i, j)$.*

Recall that $\gamma_t(j)$ is the joint probability that the stochastic process is in state s_j at time point t with observation sequence $O = \{o_1, o_2,\dots, o_T\}$, specified by (7).

$$\gamma_t(j) = P(O, x_t = s_j|\Delta) = \alpha_t(j)\beta_t(j)$$

According to total probability rule [4, p. 101], it is easy to infer that γ_t is sum of ξ_t over all states with $t \geq 2$, as seen in (19).

$$\forall t \geq 2, \gamma_t(j) = \sum_{i=1}^n \xi_t(i,j) \text{ and } \gamma_{t-1}(i) = \sum_{j=1}^n \xi_t(i,j) \tag{19}$$

Deriving from (18) and (19), we have:

$$P(O, x_{t-1} = s_i, x_t = s_j|\Delta) = \xi_t(i,j)$$

$$P(O, x_{t-1} = s_i|\Delta) = \sum_{j=1}^n \xi_t(i,j), \forall t \geq 2$$

$$P(O, x_t = s_j|\Delta) = \gamma_t(j)$$

$$P(O, x_1 = s_j|\Delta) = \gamma_1(j)$$

By extending (17), we receive (20) for specifying HMM parameter estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_i(k), \hat{\pi}_i)$ given current parameter $\Delta = (a_{ij}, b_i(k), \pi_i)$ in detailed.

$$\hat{a}_{ij} = \frac{\sum_{t=2}^T \xi_t(i,j)}{\sum_{t=2}^T \sum_{l=1}^n \xi_t(i,l)}$$

$$\hat{b}_j(k) = \frac{\sum_{\substack{t=1 \\ o_t=\varphi_k}}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)} \tag{20}$$

$$\hat{\pi}_j = \frac{\gamma_1(j)}{\sum_{i=1}^n \gamma_1(i)}$$

Followings are interpretations relevant to the joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ with observation sequence O.

- The sum $\sum_{t=2}^T \xi_t(i,j)$ expresses expected number of transitions from state s_i to state s_j [3, p. 265].
- The double sum $\sum_{t=2}^T \sum_{l=1}^n \xi_t(i,l)$ expresses expected number of transitions from state s_i [3, p. 265].
- The sum $\sum_{\substack{t=1 \\ o_t=\varphi_k}}^T \gamma_t(j)$ expresses expected number of times in state s_j and in observation φ_k [3, p. 265].
- The sum $\sum_{t=1}^T \gamma_t(j)$ expresses expected number of times in state s_j [3, p. 265].

Followings are interpretations of the parameter estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$:

- The transition estimate $\hat{a}_{ij}$ is expected frequency of transitions from state s_i to state s_j.
- The observation estimate $\hat{b}_j(k)$ is expected frequency of times in state s_j and in observation φ_k.
- The initial estimate $\hat{\pi}_j$ is (normalized) expected frequency of state s_j at the first time point ($t=1$).

It is easy to infer that the parameter estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ is based on joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ which, in turn, are based on current parameter $\Delta = (a_{ij}, b_j(k), \pi_j)$. The EM conditional expectation $E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\}$ is determined by joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$; so, the main task of E-step in EM algorithm is essentially to calculate the joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ according to (18) and (7). The EM conditional expectation $E_{X|O,\Delta_r}\{ln(P(O, X|\Delta))\}$ gets maximal at estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ and so, the main task of M-step in EM algorithm is essentially to calculate $\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j$ according to (20). The EM algorithm is interpreted in HMM learning problem, as shown in table 9.

Table 9. *EM algorithm for HMM learning problem.*

Starting with initial value for Δ, each iteration in EM algorithm has two steps:

1. *E-step*: Calculating the joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ according to (18) and (7) given current parameter $\Delta = (a_{ij}, b_j(k), \pi_j)$.

$$\xi_t(i,j) = \alpha_{t-1}(i)a_{ij}b_j(o_t)\beta_t(j) \text{ where } t \geq 2$$
$$\gamma_t(j) = P(O, x_t = s_j|\Delta) = \alpha_t(j)\beta_t(j)$$

Where forward variable α_t and backward variable β_t are calculated by (2) and (5).

2. *M-step*: Calculating the estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ based on the joint probabilities $\xi_t(i, j)$ and $\gamma_t(j)$ determined at E-step, according to (20).

$$\hat{a}_{ij} = \frac{\sum_{t=2}^T \xi_t(i,j)}{\sum_{t=2}^T \sum_{l=1}^n \xi_t(i,l)}$$
$$\hat{b}_j(k) = \frac{\sum_{\substack{t=1 \\ o_t=\varphi_k}}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)}$$
$$\hat{\pi}_j = \frac{\gamma_1(j)}{\sum_{i=1}^n \gamma_1(i)}$$

The estimate $\hat{\Delta}$ becomes the current parameter for next iteration.

EM algorithm stops when it meets the terminating condition, for example, the difference of current parameter Δ and next parameter $\hat{\Delta}$ is insignificant. It is possible to define a custom terminating condition.

The algorithm to solve HMM learning problem shown in table 9 is known as Baum-Welch algorithm [3]. Please see document "Hidden Markov Models Fundamentals" by [9, pp. 8-13] for more details about HMM learning problem. As aforementioned in sub-section 4.1, the essence of EM algorithm applied into HMM learning problem is to determine the estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$.

As seen in table 9, it is not difficult to run E-step and M-step of EM algorithm but how to determine the terminating condition is considerable problem. It is better to establish a computational terminating criterion instead of applying the general statement "EM algorithm stops when it meets the terminating condition, for example, the difference of current parameter Δ and next parameter $\hat{\Delta}$ is insignificant". Going back the learning problem that EM algorithm solves, the EM algorithm aims to maximize probability $P(O|\Delta)$ of given observation sequence $O=(o_1, o_2,\dots, o_T)$ so as to find out the estimate $\hat{\Delta}$. Maximizing the probability $P(O|\Delta)$ is equivalent to max-

imizing the conditional expectation. So it is easy to infer that EM algorithm stops when probability $P(O|\Delta)$ approaches to maximal value and EM algorithm cannot maximize $P(O|\Delta)$ any more. In other words, the probability $P(O|\Delta)$ is terminating criterion. Calculating criterion $P(O|\Delta)$ is evaluation problem described in section 2. Criterion $P(O|\Delta)$ is determined according to forward-backward procedure; please see tables 4 and 5 for more details about forward-backward procedure.

At the end of M-step, the next criterion $P(O|\hat{\Delta})$ that is calculated based on the next parameter (also estimate) $\hat{\Delta}$ is compared with the current criterion $P(O|\Delta)$ that is calculated in the previous time. If these two criterions are the same or there is no significantly difference between them then, EM algorithm stops. This implies EM algorithm cannot maximize $P(O|\Delta)$ any more. However, calculating the next criterion $P(O|\hat{\Delta})$ according to forward-backward procedure causes EM algorithm to run slowly. This drawback is overcome by following comment and improvement. The essence of forward-backward procedure is to determine forward variables α_t while EM algorithm must calculate all forward variables and backward variables in its learning process (E-step). Thus, the evaluation of terminating condition is accelerated by executing forward-backward procedure inside the E-step of EM algorithm. In other words, when EM algorithm results out forward variables in E-step, the forward-backward procedure takes advantages of such forward variables so as to determine criterion $P(O|\Delta)$ the at the same time. As a result, the speed of EM algorithm does not decrease. However, there is always a redundant iteration; suppose that the terminating criterion approaches to maximal value at the end of the rth iteration but the EM algorithm only stops at the E-step of the $(r+1)^{th}$ iteration when it really evaluates the terminating criterion. In general, the terminating criterion $P(O|\Delta)$ is calculated based on the current parameter Δ at E-step instead of the estimate $\hat{\Delta}$ at M-step. Table 10 shows the proposed implementation of EM algorithm with terminating criterion $P(O|\Delta)$. Pseudo-code like programming language C is used to describe the implementation of EM algorithm. Variables are marked as *italic words* and comments are followed by the signs // and /*. Please pay attention to programming language keywords: *while*, *for*, *if*, [], ==, !=, &&, //, /*, */, etc. For example, notation [] denotes array index operation; concretely, $\alpha[t][i]$ denotes forward variable $\alpha_t(i)$ at time point t with regard to state s_i.

Table 10. *Proposed implementation of EM algorithm for learning HMM with terminating criterion $P(O|\Delta)$.*

```
/*
Input:
    HMM with current parameter Δ = {a_ij, π_j, b_jk}
    Observation sequence O = {o_1, o_2,…, o_T}
Output:
    HMM with optimized parameter Δ = {a_ij, π_j, b_jk}
*/

Allocating memory for two matrices α and β representing forward variables
and backward variables.
previous_criterion = –1
current_criterion = –1
iteration = 0
/*Pre-defined number MAX_ITERATION is used to prevent from infinite
```

```
loop.*/
MAX_ITERATION = 10000
While (iteration < MAX_ITERATION)

    //Calculating forward variables and backward variables
    For t = 1 to T
        For i = 1 to n
            Calculating forward variables α[t][i] and backward variables
            β[T–t+1][i] based on observation sequence O according to (2) and
            (5).
        End for i
    End for t

    //Calculating terminating criterion current_criterion = P(O|Δ)
    current_criterion = 0
    For i = 1 to n
        current_criterion = current_criterion + α[T][i]
    End for i

    //Terminating condition
    If previous_criterion >= 0 && previous_criterion == current_criterion
    then
        break //breaking out the loop, the algorithm stops
    Else
        previous_criterion = current_criterion
    End if

    //Updating transition probability matrix
    For i = 1 to n
        denominator = 0
        Allocating numerators as a 1-dimension array including n zero ele-
        ments.
        For t = 2 to T
            For k = 1 to n
                ξ = α[t–1][i] * a_ik * b_k(o_t) * β[t][k]
                numerators[k] = numerators[k] + ξ
                denominator = denominator + ξ
            End for k
        End for t

        If denominator != 0 then
            For j = 1 to n
                a_ij = numerators[j] / denominator
            End for j
        End if
    End for i

    //Updating initial probability matrix
    Allocating g as a 1-dimension array including n elements.
    sum = 0
    For j = 1 to n
        g[j] = α[1][j] * β[1][j]
        sum = sum + g[j]
    End for j

    If sum != 0 then
        For j = 1 to n
            π_j = g[j] / sum
        End for j
    End if

    //Updating observation probability distribution
    For j = 1 to n
        Allocating γ as a 1-dimension array including T elements.
        denominator = 0
        For t = 1 to T
            γ[t] = α[t][j] * β[t][j]
            denominator = denominator + γ[t]
        End for t

        Let m be the columns of observation distribution matrix B.
        For k = 1 to m
            numerator = 0
            For t = 1 to T
```

```
          If o_t == k then
                numerator = numerator + γ[t]
          End if
        End for t

        b_jk = numerator / denominator
     End for k
  End for j

   iteration = iteration + 1
End while
```

According to table 10, the number of iterations is limited by a pre-defined maximum number, which aims to solve a so-called infinite loop optimization. Although it is proved that EM algorithm always converges, maybe there are two different estimates $\hat{\Delta}_1$ and $\hat{\Delta}_2$ at the final convergence. This situation causes EM algorithm to alternate between $\hat{\Delta}_1$ and $\hat{\Delta}_2$ in infinite loop. Therefore, the final estimate $\hat{\Delta}_1$ or $\hat{\Delta}_2$ is totally determined but the EM algorithm does not stop. This is the reason that the number of iterations is limited by a pre-defined maximum number.

Going back given weather HMM Δ whose parameters A, B, and $\prod$ are specified in tables 1, 2, and 3, suppose observation sequence is $O = \{o_1=\varphi_4=soggy,\ o_2=\varphi_1=dry,\ o_3=\varphi_2=dryish\}$, the EM algorithm and its implementation described in tables 9 and 10 are applied into calculating the parameter estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ which is the ultimate solution of the learning problem, as below.

At the first iteration ($r=1$) we have:

$\alpha_1(1) = b_1(o_1 = \varphi_4)\pi_1 = b_{14}\pi_1 = 0.0165$
$\alpha_1(2) = b_2(o_1 = \varphi_4)\pi_2 = b_{24}\pi_2 = 0.0825$
$\alpha_1(3) = b_3(o_1 = \varphi_4)\pi_3 = b_{34}\pi_3 = 0.165$

$\alpha_2(1) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i1}\right)b_1(o_2 = \varphi_1) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i1}\right)b_{11} = 0.04455$

$\alpha_2(2) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i2}\right)b_2(o_2 = \varphi_1) = 0.01959375$

$\alpha_2(3) = \left(\sum_{i=1}^{3} \alpha_1(i)a_{i3}\right)b_3(o_2 = \varphi_1) = 0.00556875$

$\alpha_3(1) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i1}\right)b_1(o_3 = \varphi_2) = 0.0059090625$

$\alpha_3(2) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i2}\right)b_2(o_3 = \varphi_2) = 0.005091796875$

$\alpha_3(3) = \left(\sum_{i=1}^{3} \alpha_2(i)a_{i3}\right)b_3(o_3 = \varphi_2) = 0.00198$

$\beta_3(1) = \beta_3(2) = \beta_3(3) = 1$

$\beta_2(1) = \sum_{j=1}^{n} a_{1j}b_j(o_3 = \varphi_2)\beta_3(j) = \sum_{j=1}^{n} a_{1j}b_{j2}\beta_3(j) = 0.1875$

$\beta_2(2) = \sum_{j=1}^{n} a_{2j}b_j(o_3 = \varphi_2)\beta_3(j) = 0.19$

$\beta_2(3) = \sum_{j=1}^{n} a_{3j}b_j(o_3 = \varphi_2)\beta_3(j) = 0.1625$

$\beta_1(1) = \sum_{j=1}^{n} a_{1j}b_j(o_2 = \varphi_1)\beta_2(j) = 0.07015625$

$\beta_1(2) = \sum_{j=1}^{n} a_{2j}b_j(o_2 = \varphi_1)\beta_2(j) = 0.0551875$

$\beta_1(3) = \sum_{j=1}^{n} a_{3j}b_j(o_2 = \varphi_1)\beta_2(j) = 0.0440625$

Within the E-step of the first iteration ($r=1$), the terminating criterion $P(O|\Delta)$ is calculated according to forward-backward procedure (see table 4) as follows:

$$P(O|\Delta) = \alpha_3(1) + \alpha_3(2) + \alpha_3(3) \approx 0.013$$

Within the E-step of the first iteration ($r=1$), the joint probabilities $\xi_t(i,j)$ and $\gamma_t(j)$ are calculated based on (18) and (7) as follows:

$\xi_2(1,1) = \alpha_1(1)a_{11}b_1(o_2 = \varphi_1)\beta_2(1) = \alpha_1(1)a_{11}b_{11}\beta_2(1)$
$\qquad\qquad = 0.000928125$
$\xi_2(1,2) = \alpha_1(1)a_{12}b_2(o_2 = \varphi_1)\beta_2(2) = 0.0001959375$
$\xi_2(1,3) = \alpha_1(1)a_{13}b_3(o_2 = \varphi_1)\beta_2(3) = 0.000033515625$
$\xi_2(2,1) = \alpha_1(2)a_{21}b_1(o_2 = \varphi_1)\beta_2(1) = 0.002784375$
$\xi_2(2,2) = \alpha_1(2)a_{22}b_2(o_2 = \varphi_1)\beta_2(2) = 0.0015675$
$\xi_2(2,3) = \alpha_1(2)a_{23}b_3(o_2 = \varphi_1)\beta_2(3) = 0.00020109375$
$\xi_2(3,1) = \alpha_1(3)a_{31}b_1(o_2 = \varphi_1)\beta_2(1) = 0.004640625$
$\xi_2(3,2) = \alpha_1(3)a_{32}b_2(o_2 = \varphi_1)\beta_2(2) = 0.001959375$
$\xi_2(3,3) = \alpha_1(3)a_{33}b_3(o_2 = \varphi_1)\beta_2(3) = 0.0006703125$
$\xi_3(1,1) = \alpha_2(1)a_{11}b_1(o_3 = \varphi_2)\beta_3(1) = 0.004455$
$\xi_3(1,2) = \alpha_2(1)a_{12}b_2(o_3 = \varphi_2)\beta_3(2) = 0.002784375$
$\xi_3(1,3) = \alpha_2(1)a_{13}b_3(o_3 = \varphi_2)\beta_3(3) = 0.00111375$
$\xi_3(2,1) = \alpha_2(2)a_{21}b_1(o_3 = \varphi_2)\beta_3(1) = 0.001175625$
$\xi_3(2,2) = \alpha_2(2)a_{22}b_2(o_3 = \varphi_2)\beta_3(2) = 0.001959375$
$\xi_3(2,3) = \alpha_2(2)a_{23}b_3(o_3 = \varphi_2)\beta_3(3) = 0.0005878125$
$\xi_3(3,1) = \alpha_2(3)a_{31}b_1(o_3 = \varphi_2)\beta_3(1) = 0.0002784375$
$\xi_3(3,2) = \alpha_2(3)a_{32}b_2(o_3 = \varphi_2)\beta_3(2) = 0.000348046875$
$\xi_3(3,3) = \alpha_2(3)a_{33}b_3(o_3 = \varphi_2)\beta_3(3) = 0.0002784375$
$\gamma_1(1) = \alpha_1(1)\beta_1(1) = 0.001157578125$
$\gamma_1(2) = \alpha_1(2)\beta_1(2) = 0.00455296875$
$\gamma_1(3) = \alpha_1(3)\beta_1(3) = 0.0072703125$
$\gamma_2(1) = \alpha_2(1)\beta_2(1) = 0.008353125$
$\gamma_2(2) = \alpha_2(2)\beta_2(2) = 0.0037228125$
$\gamma_2(3) = \alpha_2(3)\beta_2(3) = 0.000904921875$
$\gamma_3(1) = \alpha_3(1)\beta_3(1) = 0.0059090625$
$\gamma_3(2) = \alpha_3(2)\beta_3(2) = 0.005091796875$
$\gamma_3(3) = \alpha_3(3)\beta_3(3) = 0.00198$

Within the M-step of the first iteration ($r=1$), the estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ is calculated based on joint probabilities $\xi_t(i,j)$ and $\gamma_t(j)$ determined at E-step.

$\hat{a}_{11} = \dfrac{\sum_{t=2}^{3} \xi_t(1,1)}{\sum_{t=2}^{3}\sum_{l=1}^{3} \xi_t(1,l)} \approx 0.5660$

$\hat{a}_{12} = \dfrac{\sum_{t=2}^{3} \xi_t(1,2)}{\sum_{t=2}^{3}\sum_{l=1}^{3} \xi_t(1,l)} \approx 0.3134$

$\hat{a}_{13} = \dfrac{\sum_{t=2}^{3} \xi_t(1,3)}{\sum_{t=2}^{3}\sum_{l=1}^{3} \xi_t(1,l)} \approx 0.1206$

$\hat{a}_{21} = \dfrac{\sum_{t=2}^{3} \xi_t(2,1)}{\sum_{t=2}^{3}\sum_{l=1}^{3} \xi_t(2,l)} \approx 0.4785$

$\hat{a}_{22} = \dfrac{\sum_{t=2}^{3} \xi_t(2,2)}{\sum_{t=2}^{3}\sum_{l=1}^{3} \xi_t(2,l)} \approx 0.4262$

$\hat{a}_{23} = \dfrac{\sum_{t=2}^{3} \xi_t(2,3)}{\sum_{t=2}^{3}\sum_{l=1}^{3} \xi_t(2,l)} \approx 0.0953$

$\hat{a}_{31} = \dfrac{\sum_{t=2}^{3} \xi_t(3,1)}{\sum_{t=2}^{3}\sum_{l=1}^{3} \xi_t(3,l)} \approx 0.6017$

$\hat{a}_{32} = \dfrac{\sum_{t=2}^{3} \xi_t(3,2)}{\sum_{t=2}^{3}\sum_{l=1}^{3} \xi_t(3,l)} \approx 0.2822$

$$\hat{a}_{33} = \frac{\sum_{t=2}^{3} \xi_t(3,3)}{\sum_{t=2}^{3} \sum_{l=1}^{3} \xi_t(3,l)} \approx 0.1161$$

$$\hat{b}_2(1) = \frac{\sum_{\substack{t=1 \\ o_t=\varphi_1}}^{3} \gamma_t(2)}{\sum_{t=1}^{3} \gamma_t(2)} = \frac{\gamma_2(2)}{\gamma_1(2) + \gamma_2(2) + \gamma_3(2)} \approx 0.2785$$

$$\hat{b}_2(2) = \frac{\sum_{\substack{t=1 \\ o_t=\varphi_2}}^{3} \gamma_t(2)}{\sum_{t=1}^{3} \gamma_t(2)} = \frac{\gamma_3(2)}{\gamma_1(2) + \gamma_2(2) + \gamma_3(2)} \approx 0.3809$$

$$\hat{b}_2(3) = \frac{\sum_{\substack{t=1 \\ o_t=\varphi_3}}^{3} \gamma_t(2)}{\sum_{t=1}^{3} \gamma_t(2)} = \frac{0}{\gamma_1(2) + \gamma_2(2) + \gamma_3(2)} = 0$$

$$\hat{b}_2(4) = \frac{\sum_{\substack{t=1 \\ o_t=\varphi_4}}^{3} \gamma_t(2)}{\sum_{t=1}^{3} \gamma_t(2)} = \frac{\gamma_1(2)}{\gamma_1(2) + \gamma_2(2) + \gamma_3(2)} \approx 0.3406$$

$$\hat{b}_3(1) = \frac{\sum_{\substack{t=1 \\ o_t=\varphi_1}}^{3} \gamma_t(3)}{\sum_{t=1}^{3} \gamma_t(3)} = \frac{\gamma_2(3)}{\gamma_1(3) + \gamma_2(3) + \gamma_3(3)} \approx 0.0891$$

$$\hat{b}_3(2) = \frac{\sum_{\substack{t=1 \\ o_t=\varphi_2}}^{3} \gamma_t(3)}{\sum_{t=1}^{3} \gamma_t(3)} = \frac{\gamma_3(3)}{\gamma_1(3) + \gamma_2(3) + \gamma_3(3)} \approx 0.1950$$

$$\hat{b}_3(3) = \frac{\sum_{\substack{t=1 \\ o_t=\varphi_3}}^{3} \gamma_t(3)}{\sum_{t=1}^{3} \gamma_t(3)} = \frac{0}{\gamma_1(3) + \gamma_2(3) + \gamma_3(3)} = 0$$

$$\hat{b}_3(4) = \frac{\sum_{\substack{t=1 \\ o_t=\varphi_4}}^{3} \gamma_t(3)}{\sum_{t=1}^{3} \gamma_t(3)} = \frac{\gamma_1(3)}{\gamma_1(3) + \gamma_2(3) + \gamma_3(3)} \approx 0.7159$$

$$\hat{\pi}_1 = \frac{\gamma_1(1)}{\gamma_1(1) + \gamma_1(2) + \gamma_1(3)} \approx 0.0892$$

$$\hat{\pi}_2 = \frac{\gamma_1(2)}{\gamma_1(1) + \gamma_1(2) + \gamma_1(3)} \approx 0.3507$$

$$\hat{\pi}_3 = \frac{\gamma_1(3)}{\gamma_1(1) + \gamma_1(2) + \gamma_1(3)} \approx 0.5601$$

At the second iteration ($r=2$), the current parameter $\Delta = (a_{ij}, b_j(k), \pi_j)$ is received values from the estimate $\hat{\Delta} = (\hat{a}_{ij}, \hat{b}_j(k), \hat{\pi}_j)$ above. By repeating the similar calculation, it is easy to determine HMM parameters at the second iteration. Table 11 summarizes HMM parameters resulted from the first iteration and the second iteration of EM algorithm.

Table 11. *HMM parameters resulted from the first iteration and the second iteration of EM algorithm.*

Iteration	HMM parameters			
1st	$\hat{a}_{11} = 0.5660$	$\hat{a}_{12} = 0.3134$	$\hat{a}_{13} = 0.1206$	
	$\hat{a}_{21} = 0.4785$	$\hat{a}_{22} = 0.4262$	$\hat{a}_{23} = 0.0953$	
	$\hat{a}_{31} = 0.6017$	$\hat{a}_{32} = 0.2822$	$\hat{a}_{33} = 0.1161$	
	$\hat{b}_1(1)$ = 0.5417	$\hat{b}_1(2)$ = 0.3832	$\hat{b}_1(3)$ = 0	$\hat{b}_1(4)$ = 0.0751
	$\hat{b}_2(1)$ = 0.2785	$\hat{b}_2(2)$ = 0.3809	$\hat{b}_2(3)$ = 0	$\hat{b}_2(4)$ = 0.3406
	$\hat{b}_3(1)$ = 0.0891	$\hat{b}_3(2)$ = 0.1950	$\hat{b}_3(3)$ = 0	$\hat{b}_3(4)$ = 0.7159
	$\hat{\pi}_1 = 0.0892$	$\hat{\pi}_2 = 0.3507$	$\hat{\pi}_3 = 0.5601$	
	Terminating criterion $P(O\|\Delta) = 0.013$			
2nd	$\hat{a}_{11} = 0.6053$	$\hat{a}_{12} = 0.3299$	$\hat{a}_{13} = 0.0648$	
	$\hat{a}_{21} = 0.5853$	$\hat{a}_{22} = 0.3781$	$\hat{a}_{23} = 0.0366$	
	$\hat{a}_{31} = 0.7793$	$\hat{a}_{32} = 0.1946$	$\hat{a}_{33} = 0.0261$	
	$\hat{b}_1(1)$ = 0.5605	$\hat{b}_1(2)$ = 0.4302	$\hat{b}_1(3)$ = 0	$\hat{b}_1(4)$ = 0.0093

$\hat{b}_2(1)$ = 0.2757	$\hat{b}_2(2)$ = 0.4517	$\hat{b}_2(3)$ = 0	$\hat{b}_2(4)$ = 0.2726
$\hat{b}_3(1)$ = 0.0283	$\hat{b}_3(2)$ = 0.0724	$\hat{b}_3(3)$ = 0	$\hat{b}_3(4)$ = 0.8993
$\hat{\pi}_1 = 0.0126$	$\hat{\pi}_2 = 0.2147$	$\hat{\pi}_3 = 0.7727$	
Terminating criterion $P(O\|\Delta) = 0.0776$			

As seen in table 11, the EM algorithm does not converge yet when it produces two different terminating criterions (0.013 and 0.0776) at the first iteration and the second iteration. It is necessary to run more iterations so as to gain the most optimal estimate. Within this example, the EM algorithm converges absolutely after 10 iterations when the criterion $P(O\|\Delta)$ approaches to the same value 1 at the 9th and 10th iterations. Table 12 shows HMM parameter estimates along with terminating criterion $P(O\|\Delta)$ at the 9th and 10th iterations of EM algorithm.

Table 12. *HMM parameters along with terminating criterions after 10 iterations of EM algorithm.*

Iteration	HMM parameters			
9th	$\hat{a}_{11} = 0$	$\hat{a}_{12} = 1$	$\hat{a}_{13} = 0$	
	$\hat{a}_{21} = 0$	$\hat{a}_{22} = 1$	$\hat{a}_{23} = 0$	
	$\hat{a}_{31} = 1$	$\hat{a}_{32} = 0$	$\hat{a}_{33} = 0$	
	$\hat{b}_1(1) = 1$	$\hat{b}_1(2) = 0$	$\hat{b}_1(3) = 0$	$\hat{b}_1(4) = 0$
	$\hat{b}_2(1) = 0$	$\hat{b}_2(2) = 1$	$\hat{b}_2(3) = 0$	$\hat{b}_2(4) = 0$
	$\hat{b}_3(1) = 0$	$\hat{b}_3(2) = 0$	$\hat{b}_3(3) = 0$	$\hat{b}_3(4) = 1$
	$\hat{\pi}_1 = 0$	$\hat{\pi}_2 = 0$	$\hat{\pi}_3 = 1$	
	Terminating criterion $P(O\|\Delta) = 1$			
10th	$\hat{a}_{11} = 0$	$\hat{a}_{12} = 1$	$\hat{a}_{13} = 0$	
	$\hat{a}_{21} = 0$	$\hat{a}_{22} = 1$	$\hat{a}_{23} = 0$	
	$\hat{a}_{31} = 1$	$\hat{a}_{32} = 0$	$\hat{a}_{33} = 0$	
	$\hat{b}_1(1) = 1$	$\hat{b}_1(2) = 0$	$\hat{b}_1(3) = 0$	$\hat{b}_1(4) = 0$
	$\hat{b}_2(1) = 0$	$\hat{b}_2(2) = 1$	$\hat{b}_2(3) = 0$	$\hat{b}_2(4) = 0$
	$\hat{b}_3(1) = 0$	$\hat{b}_3(2) = 0$	$\hat{b}_3(3) = 0$	$\hat{b}_3(4) = 1$
	$\hat{\pi}_1 = 0$	$\hat{\pi}_2 = 0$	$\hat{\pi}_3 = 1$	
	Terminating criterion $P(O\|\Delta) = 1$			

As a result, the learned parameters A, B, and $\prod$ are shown in table 13:

Table 13. *HMM parameters of weather example learned from EM algorithm.*

		Weather current day (Time point t)		
		sunny	***cloudy***	***rainy***
Weather previous day (Time point $t-1$)	*sunny*	$a_{11}=0$	$a_{12}=1$	$a_{13}=0$
	cloudy	$a_{21}=0$	$a_{22}=1$	$a_{23}=0$
	rainy	$a_{31}=1$	$a_{32}=0$	$a_{33}=0$

sunny	cloudy	rainy
$\pi_1=0$	$\pi_2=0$	$\pi_3=1$

	Humidity			
	dry	*dryish*	*damp*	*soggy*
sunny	$b_{11}=1$	$b_{12}=0$	$b_{13}=0$	$b_{14}=0$
cloudy	$b_{21}=0$	$b_{22}=1$	$b_{23}=0$	$b_{24}=0$
rainy	$b_{31}=0$	$b_{32}=0$	$b_{33}=0$	$b_{34}=1$

(Weather — left row label spanning the three weather rows of the Humidity table)

Such learned parameters are more appropriate to the training observation sequence $O = \{o_1=\varphi_4=soggy,\ o_2=\varphi_1=dry,\ o_3=\varphi_2=dryish\}$ than the original ones shown in tables 1, 2, and 3 when the terminating criterion $P(O|\Delta)$ corresponding to its optimal state sequence is 1.

Now three main problems of HMM are described; please see an excellent document "A tutorial on hidden Markov models and selected applications in speech recognition" written by the author Rabiner [3] for advanced details about HMM.

5. Conclusion

In general, there are three main problems of HMM such as evaluation problem, uncovering problem, and learning problem. For evaluation problem and uncovering problem, researchers should pay attention to forward variable and backward variable. Most computational operations are relevant to them. They reflect unique aspect of HMM. The Viterbi algorithm is very effective to solve the uncovering problem. The Baum-Welch algorithm is often used to solve the learning problem. It is easier to explain Baum-Welch algorithm by combination of EM algorithm and optimization theory, in which the Lagrangian function is maximized so as to find out optimal parameters of EM algorithms when such parameters are also learned parameters of HMM.

Observations of normal HMM described in this report are quantified by discrete probability distribution which is observation probability matrix B. In the most general case, observation is represented by continuous variable and matrix B is replaced by probability density function. At that time the normal HMM becomes continuously observational HMM. Readers are recommended to research continuously observational HMM, an enhanced variant of normal HMM.

References

[1] E. Fosler-Lussier, "Markov Models and Hidden Markov Models: A Brief Tutorial," 1998.

[2] J. G. Schmolze, "An Introduction to Hidden Markov Models," 2001.

[3] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[4] L. Nguyen, "Mathematical Approaches to User Modeling," Journals Consortium, 2015.

[5] B. Sean, "The Expectation Maximization Algorithm - A short tutorial," Sean Borman's Homepage, 2009.

[6] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, Series B (Methodological), vol. 39, no. 1, pp. 1-38, 1977.

[7] Y.-B. Jia, "Lagrange Multipliers," 2013.

[8] S. Borman, "The Expectation Maximization Algorithm - A short tutorial," Sean Borman's Home Page, South Bend, Indiana, 2004.

[9] D. Ramage, "Hidden Markov Models Fundamentals," 2007.

[10] Wikipedia, "Karush–Kuhn–Tucker conditions," Wikimedia Foundation, 4 August 2014. [Online]. Available: http://en.wikipedia.org/wiki/Karush–Kuhn–Tucker_conditions. [Accessed 16 November 2014].

[11] S. Boyd and L. Vandenberghe, Convex Optimization, New York, NY: Cambridge University Press, 2009, p. 716.G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references).