Speech and Language Processing

Lecture 2
Maximum likelihood estimation and EM algorithm

Information and Communications Engineering Course
Takahiro Shinozaki

2018/10/26
Lecture Plan (Shinozaki’s part)

I gives the first 6 lectures about speech recognition. Through these lectures, the backbone of the latest speech recognition techniques is explained.

1. 10/19 (remote)
   Speech recognition based on GMM, HMM, and N-gram
2. 10/26 (remote)
   Maximum likelihood estimation and EM algorithm
3. 11/5 (remote)
   Bayesian network and Bayesian inference
4. 11/5 (@TAIST)
   Variational inference and sampling
5. 11/6 (@TAIST)
   Neural network based acoustic and language models
6. 11/6 (@TAIST)
   Weighted finite state transducer (WFST) and speech decoding
Today’s Topic

• Answers for the previous exercises
• Brief review of probability theory
• Maximum likelihood estimation
• Expectation maximization (EM) algorithm
Answers for the Previous Exercises
Suppose $W$ is a vowel and $O$ is an MFCC feature vector. Suppose that $P_{AM}(O \mid W)$ is an acoustic model and $P_{LM}(W)$ is a language model. Obtain a vowel $\hat{W}$ that maximizes $P(W \mid O)$ when the acoustic and language model log likelihoods are given as in the following table.

$$\hat{W} = \arg \max_{W \in \{a, i, u, e, o\}} \{P(W \mid O)\}$$

<table>
<thead>
<tr>
<th>Vowel $W$</th>
<th>$a$</th>
<th>$i$</th>
<th>$u$</th>
<th>$e$</th>
<th>$o$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log(P(O \mid W))$</td>
<td>-13.4</td>
<td>-10.5</td>
<td>-30.1</td>
<td>-15.2</td>
<td>-17.0</td>
</tr>
<tr>
<td>$\log(P(W))$</td>
<td>-1.61</td>
<td>-2.30</td>
<td>-1.61</td>
<td>-1.39</td>
<td>-1.39</td>
</tr>
</tbody>
</table>
Exercise 1.2 (Answer)

The following table defines a Bi-gram \( P(\text{Word} | \text{Context}) \)

<table>
<thead>
<tr>
<th>C \ W</th>
<th>today</th>
<th>is</th>
<th>sunny</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>0.6</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>today</td>
<td>0.1</td>
<td>0.5</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>is</td>
<td>0.1</td>
<td>0.1</td>
<td>0.7</td>
<td>0.1</td>
</tr>
<tr>
<td>sunny</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>0.6</td>
</tr>
</tbody>
</table>

*\( P(\text{Start}) = 1.0 \)*

Example:

\[
\text{Start} \rightarrow \text{today} \rightarrow \text{is} \rightarrow \text{sunny} \rightarrow \text{End}
\]

\[
1.0 \times 0.6 \times 0.5 \times 0.7 \times 0.6 = 0.126
\]
Exercise 1.2 (Cont.) (Answer)

• Based on the bigram definition of the previous slide, compute the probability of the following sentences

1) \[ P(\text{"Start today sunny today sunny End"}) \]
   \[
   = \frac{0.6 \times 0.3 \times 0.1 \times 0.3 \times 0.6}{0.6 \times 0.3 \times 0.1 \times 0.3 \times 0.6} = 1
   \]

2) \[ P(\text{"Start today today sunny sunny End"}) \]
   \[
   = \frac{0.6 \times 0.1 \times 0.3 \times 0.2 \times 0.6}{0.6 \times 0.1 \times 0.3 \times 0.2 \times 0.6} = 1
   \]
Brief Review of Probability Theory
Probability Space

• Sample space ($\Omega$)
  • Set of all possible outcomes of an experiment

• Probability function ($f(x)$)
  • A function that maps each outcome to a probability
    \[ f(x) \in [0, 1] \quad \text{for all } x \in \Omega \]
    \[ \sum_{x \in \Omega} f(x) = 1 \]

• Event (E)
  • Subset of the sample space
  Probability of an event $E$ is:
    \[ P(E) = \sum_{x \in E} f(x) \]
Random Variable

• A function that maps an outcome of an experiment to a value
  • Notation:
    \[ P[X=x] = p \] means “the probability of a random variable \( X \) takes a value \( x \) is \( p \)”

Example

\[ \begin{array}{c|cccccc}
\text{Random Variable} & 1 & 2 & 3 & 4 & 5 & 6 \\
\hline
X & 1 & 2 & 3 & 4 & 5 & 6 \\
Y & 1 & 0 & 1 & 0 & 1 & 0 \\
Z & 0 & 0 & 1 & 1 & 1 & 1 \\
\end{array} \]

X: The value of a die
Y: Whether the value of a die is odd number or not
Z: Whether the value of a die is larger than 2 or not
Joint Probability

• Probability that more than one events jointly occur

Example

\[ X : \text{Dice 1} \]
\[ Y : \text{Dice 2} \]

\[ P(X = i, Y = j) : \text{Probability that the value of } X \text{ is } i \text{ and the value of } Y \text{ is } j \]

Note: \[ P(X=i, Y=j) = P(Y=j, X=i) \]
Conditional Probability

• Probability of an event given that another event has occurred

Example

Randomly picks up two balls sequentially from a box containing 4 blue and 6 green balls

\(X\) : Color of the first ball

\(Y\) : Color of the second ball

\[P(Y = \text{blue} \mid X = \text{green}) = \frac{4}{9}\]

\[P(Y = \text{blue} \mid X = \text{blue}) = \frac{3}{9}\]
Two Principal Rules

• Sum rule
  • Summing joint probability $P(X,Y)$ for all possible values of $Y$ gives probability of $P(X)$
  • $P(X)$ is called the marginal probability

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$$

• Product rule
  • Multiplying probability $P(X)$ and joint probability $P(Y|X)$ is equal to joint probability $P(X,Y)$

$$P(X = x_i, Y = y_j) = P(Y = y_j | X = x_i)P(X = x_i)$$
Bayes’ Theorem

- From the product rule, we obtain:

\[ P(Y = y_j \mid X = x_i) = \frac{P(X = x_i \mid Y = y_j)P(Y = y_j)}{P(X = x_i)} \quad \text{for } \forall x_i, y_i \]

If we denote the distributions as \( P(X) \) etc., we have:

\[ P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \]

Using the sum rule, it can be expressed as:

\[ P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} = \frac{P(X \mid Y)P(Y)}{\sum_Y P(X,Y)} = \frac{P(X \mid Y)P(Y)}{\sum_Y P(X \mid Y)P(Y)} \]
Independence

• If the joint distribution of two variables $X$ and $Y$ factorizes into the product of the marginals, then $X$ and $Y$ are said to be “independent”

\[ P(X, Y) = P(X | Y)P(Y) = P(X)P(Y) \quad \Rightarrow \quad X \text{ and } Y \text{ are independent} \]

\[ P(X, Y) = P(X | Y)P(Y) \neq P(X)P(Y) \quad \Rightarrow \quad X \text{ and } Y \text{ are not independent} \]
Probability Densities

- If the probability of a real-valued variable $x$ falling in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ when $\delta x \to 0$, $p(x)$ is called the probability density of $x$.

$p(x)\delta x$ is probability $\Rightarrow$ $p(x) \geq 0$ and $\int_{-\infty}^{\infty} p(x)dx = 1$
The Sum and The Product Rules For Continuous Variable

\[ P(X = x_i) = \sum_j P(X = x_i, Y = y_j) \]

\[ p(x) = \int p(x, y)dy \]

\[ P(X = x_i, Y = y_j) = P(Y = y_j \mid X = x_i)P(X = x_i) \]

\[ p(x, y) = p(y \mid x)p(x) \]
Expectation

- Expectation of a function $f(x)$ under a probability distribution $p(x)$ is denoted by $E[f]$

$$E[f] = \sum_x p(x)f(x) \quad (x \text{ is discrete})$$

$$E[f] = \int p(x)f(x)dx \quad (x \text{ is continuous})$$
Mean and Variance

• Mean
  • Synonym of the expectation $E[f(x)]$

• Variance
  • A measure of how much variability there is in $f(x)$ around its mean value $E[f(x)]$

$$\text{var}[f] \equiv E\left[\left(f(x) - E[f(x)]\right)^2\right] = E[f(x)^2] - E[f(x)]^2$$

• In particular, the variance of the variable $x$ itself is:

$$\text{var}[x] = E\left[(x - E[x])^2\right]$$
Covariance

- Covariance
  - The extent to which $x$ and $y$ vary together

$$\text{cov}[x, y] \equiv E_{x,y} [(x - E[x])(y - E[y])]$$

$$= E_{x,y} [xy] - E[x]E[y]$$

Expectation with respect to joint probability of $x$ and $y$
Entropy

• Amount of randomness in the random variable

\[ H[x] = E[-\log(p(x))] = -\sum_x p(x)\log p(x) \]

Example

<table>
<thead>
<tr>
<th>x</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(x)</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

\[ H[x] = -0.5\log(0.5) - 0.5\log(0.5) \]
\[ = 0.693 \]

<table>
<thead>
<tr>
<th>x</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(x)</td>
<td>0.1</td>
<td>0.9</td>
</tr>
</tbody>
</table>

\[ H[x] = -0.1\log(0.1) - 0.9\log(0.9) \]
\[ = 0.325 \]
Relative Entropy

• A measure of dissimilarity of two distributions $p$ and $q$
  • Also called as kullback-Leibler (KL) divergence

\[
KL(p \parallel q) = E_p \left[ \log \left( \frac{p(x)}{q(x)} \right) \right] = -\int p(x) \log \left( \frac{q(x)}{p(x)} \right) dx
\]

• $KL(p \parallel q)$ is nonnegative.
  $KL(p \parallel q) = 0$ if and only if $p(x) = q(x)$

Note: $KL(p \parallel q) \neq KL(q \parallel p)$
Maximum Likelihood Estimation
Maximum Likelihood (ML) Method

• Assume that we have a set of samples \( D = \{x_1, x_2, \ldots, x_n\} \) drawn from a distribution \( p(x|\theta) \) with unknown parameters \( \theta \), and we want to estimate \( \theta \)

• Maximum likelihood method estimates the parameters by maximizing likelihood \( p(D|\theta) \)

\[
\hat{\theta} = \arg \max_{\theta} p(D | \theta) = \arg \max_{\theta} \prod_{i=1}^{n} p(x_i | \theta)
\]

Decomposed to a product when the samples are drawn independently
Bernoulli Distribution

• Probability distribution of a binary random variable which takes value 1 with probability $\mu$ and value 0 with probability $1-\mu$

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bern($x$)</td>
<td>$1-\mu$</td>
<td>$\mu$</td>
</tr>
</tbody>
</table>

$Bern(x) = \mu^x (1 - \mu)^{1-x}$

Is the result Head or Tail?
ML Estimation for Bernoulli Distribution

- Parameter $\theta$ in this case is: $\mu$
- Training sample $x_i = 0 \text{ or } 1$

\[
\hat{\mu} = \arg \max_{\mu} p(D \mid \mu) = \arg \max_{\mu} \prod_{i=1}^{n} \mu^{x_i} (1 - \mu)^{1-x_i}
\]

\[
= \arg \max_{\mu} \log \left( \prod_{i} \mu^{x_i} (1 - \mu)^{1-x_i} \right)
\]

\[
= \arg \max_{\mu} \left\{ \sum_{i} x_i \log(\mu) + \sum_{i} (1-x_i) \log(1-\mu) \right\}
\]

\[
\frac{\partial}{\partial \mu} \left( \sum_{i} x_i \log(\mu) + \sum_{i} (1-x_i) \log(1-\mu) \right) = 0
\]

\[
\mu = \frac{1}{n} \sum_{i} x_i \quad \text{n: the number of samples}
\]
Example

• You tossed a winded coin 100 times, and got 62 heads and 38 tails. Estimate the probability $\mu$ of getting head with the coin by the ML method.
Categorical Distribution

As a generalization of the Bernoulli Distribution, let's consider a discrete random variable $X$ that takes $K$ values.

<table>
<thead>
<tr>
<th>$X$</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-of-K</td>
<td>$&lt;1,0,...,0&gt;$</td>
<td>$&lt;0,1,...,0&gt;$</td>
<td>...</td>
<td>$&lt;0,0,...,1&gt;$</td>
</tr>
<tr>
<td>$p(X)$</td>
<td>$\mu_1$</td>
<td>$\mu_2$</td>
<td>...</td>
<td>$\mu_K$</td>
</tr>
</tbody>
</table>

$p(x | \mu) = \prod_{k=1}^{K} \mu_k^{x_k}$
ML for Categorical Distribution

- Parameter $\theta$ in this case is: $\mu = \{\mu_1, \mu_2, \ldots, \mu_K\}$
- Training sample $x_i$ is a vector of 1-of-K representation. When $x_i$ represents $k$-th value, $x_{i,k} = 1$, and $x_{i,j} = 0$ for $j \neq k$

\[
\hat{\mu} = \arg \max_\mu p(D | \mu) = \arg \max_\mu \prod_{i=1}^{n} \prod_{k=1}^{K} \mu_k^{x_{i,k}} = \arg \max_\mu \prod_{k=1}^{K} \mu_k^{m_k}
\]

\[
\sum_{k=1}^{K} \mu_k = 1 \quad \text{Constraint}
\]

$m_k$ is the number of the occurrence of $k$-th value, where $n$ is the number of samples

This is a maximization problem with a constraint

Use the method of Lagrange multiplier (c.f. Appendix)
Solution

\[
\hat{\mu} = \arg \max_{\mu} \prod_{k=1}^{K} \mu_k^{m_k} = \arg \max_{\mu} \sum_{k=1}^{K} m_k \log(\mu_k)
\]

\[
\sum_{k=1}^{K} \mu_k = 1
\]

\[
\hat{\mu} = \arg \max_{\mu} \left\{ \sum_{k=1}^{K} m_k \log \mu_k + \lambda \left( \sum_{k=1}^{K} \mu_k - 1 \right) \right\}
\]

\[
\mu_k = \frac{m_k}{n}
\]
Exercise 2.1

• Show the derivation process of obtaining \( \mu_k = \frac{m_k}{n} \)

for the categorical distribution by maximizing

\[
L(\mu, \lambda) = \sum_{k=1}^{K} m_k \log \mu_k + \lambda \left( \sum_{k=1}^{K} \mu_k - 1 \right)
\]

where \( \lambda \) is the Lagrange multiplier.
ML Estimation for Gaussian Distribution

Gaussian distribution:

\[
N(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}
\]

Parameter \( \theta \) in this case is: \( \{\mu, \sigma\} \)

Training sample \( x_i \) is a real value

ML estimation of Gaussian distribution

\[
\hat{\theta} = \text{arg max}_\theta \prod_{i=1}^{n} N(x_i \mid \theta) = \text{arg max}_\theta \sum_{i=1}^{n} \log(N(x_i \mid \theta))
\]

\[
\int_{-\infty}^{\infty} N(x \mid \mu, \sigma^2)dx = 1
\]
Exercise 2.2

• Derive the ML solution \( \{\hat{\mu}, \hat{\sigma}\} \) of the Gaussian distribution. The derivation process must be described.
ML Estimation for GMM with Known Index

- Let’s consider 2-mix GMM (component index $m$ is 1 or 2)
- A training sample $x_i = <o_i, m_i>$ is a pair of an observation $o_i$ and an index of Gaussian component $m_i$, where $i$ is a sample index

$$L(\mu_1, \sigma_1, w_1, \mu_2, \sigma_2, w_2) = \log \prod_{i=1}^{n} w_{m_i} N(o_i | \mu_{m_i}, \sigma_{m_i}) = \sum_{i=1}^{n} \log(w_{m_i} N(o_i | \mu_{m_i}, \sigma_{m_i}))$$

$$= \sum_{i=1}^{n} \log(w_{m_i}) + \sum_{i|m_i=1} \log(N(o_i | \mu_1, \sigma_1)) + \sum_{i|m_i=2} \log(N(o_i | \mu_2, \sigma_2)), \quad w_1 + w_2 = 1.0$$

$$\text{arg max } L(\mu_1, \sigma_1, w_1, \mu_2, \sigma_2, w_2) \quad \leftrightarrow \quad \text{arg max } \sum_{i=1}^{n} \log(w_{m_i})$$

The components can be optimized independently
ML Estimation for HMM with Known Path

• Both observation and state sequences are given
  • Transition probability: Transition probability from state \( i \) to \( j \) is obtained by dividing the number of transitions from state \( i \) to \( j \) by the number of transition from state \( i \)
  • Emission probability: ML estimate of the emission distribution based on the observations assigned to the state

Example: (Observation is a binary value taking ‘a’ or ‘b’)
When \( O=(a,b,a,b,b) \) and \( K=(s_0,s_1,s_1,s_2,s_2,s_2,s_3) \)

Transition probability
\[
\begin{align*}
p(s_1 \rightarrow s_1) &= 1/2, \\
p(s_1 \rightarrow s_2) &= 1/2, \\
p(s_2 \rightarrow s_2) &= 2/3, \\
p(s_2 \rightarrow s_3) &= 1/3
\end{align*}
\]

Emission probability
\[
\begin{align*}
p(a \mid s_1) &= 1/2, \\
p(b \mid s_1) &= 1/2, \\
p(a \mid s_2) &= 1/3, \\
p(b \mid s_2) &= 2/3
\end{align*}
\]
Exercise 2.3

• Given a training data $D$ with $n$ training samples $D=\{x_1, x_2, ..., x_n\}$, obtain ML estimation for GMM with $M$ mixtures. You can assume the variance is 1 for simplicity.

\[
\hat{M} = \arg\max_{\langle\mu_1, \mu_2, \ldots, \mu_M\rangle} \left[ \prod_{i=1}^{N} \left\{ \sum_{m=1}^{M} w_m \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{(X_i - \mu_m)^2}{2} \right) \right\} \right]
\]

\[
= \arg\max_{\langle\mu_1, \mu_2, \ldots, \mu_M\rangle} \left[ \sum_{i=1}^{N} \log \left\{ \sum_{m=1}^{M} w_m \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{(X_i - \mu_m)^2}{2} \right) \right\} \right]
\]
Exercise 2.3 (Cont.)

\[ L = \sum_{i=1}^{N} \log \left\{ \sum_{m=1}^{M} w_m \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(X_i - \mu_m)^2}{2} \right) \right\} \]

\[ \frac{\partial L}{\partial \mu_m} = \text{EM algorithm} \]
Expectation Maximization (EM) Algorithm
Hidden Variable

• The mixture weight $w_m$ of GMM can be regarded as a probability $P(m)$ since it is non-negative and sum to one.

• Then, the GMM can be seen as a marginal probability of $P(m, x) = P(m)N(x | \mu_m, \sigma_m)$

\[
GMM(x) = \sum_{m=1}^{M} w_m N(x | \mu_m, \sigma_m) = \sum_{m=1}^{M} P(m)N(x | \mu_m, \sigma_m)
\]

• In general, when a probability model is defined as a marginal probability, the summed-out variable is not seen from the outside, and it is called a hidden variable.

• The mixture weight of GMM is a hidden variable.
ML Estimation for Models with Hidden Variables

- The summation $\Sigma$ for the marginalization is often problematic for optimization

\[
\hat{\Theta} = \arg \max_\Theta \left[ P(D | \Theta) \right]
\]

\[
= \arg \max_\Theta \left[ \sum_i \log P(x_i | \Theta) \right]
\]

\[
= \arg \max_\Theta \left[ \sum_i \log \sum_h P(x_i, h | \Theta) \right]
\]

nuisance
Jensen Lower Bound of Likelihood

Let $X$ be an observed variable, $H$ be a hidden variable, and $\Theta$ be a parameter.

\[
\log P(X \mid \Theta) = \log \sum_H P(X, H \mid \Theta)
\]

\[
= \log \sum_H q(H) \frac{P(X, H \mid \Theta)}{q(H)}
\]

\[
\geq \sum_H q(H) \log \frac{P(X, H \mid \Theta)}{q(H)}
\]

This inequality holds for arbitrary $q$ and arbitrary $\Theta$.

Let

\[
J(q, \Theta) = \sum_H q(H) \log \frac{P(X, H \mid \Theta)}{q(H)}
\]

(Lower bound of likelihood)
Exercise 2.4

• Assume you have an initial model parameter $\Theta_0$. Prove that if you take $q(H) = q_0(H) = P(H|X, \Theta_0)$, then the lower bound $J(q_0, \Theta_0)$ is equal to the log likelihood $logP(X|\Theta_0)$

\[ \log P(X | \Theta_0) = J(P(H | X, \Theta_0), \Theta_0) \]
Maximization of the Lower Bound

• Assume we have an initial model parameter $\Theta_0$.

$$\Theta_1 = \arg \max_{\Theta} J(P(H \mid X, \Theta_0), \Theta)$$

Because:

$$\log P(X \mid \Theta_0) = J(P(H \mid X, \Theta_0), \Theta_0)$$
$$J(P(H \mid X, \Theta_0), \Theta) \leq \log P(X \mid \Theta) \quad \text{for} \quad \forall \Theta$$
$$J(P(H \mid X, \Theta_0), \Theta_0) \leq J(P(H \mid X, \Theta_0), \Theta_1)$$

By maximizing the lower bound $J$ with respective to $\Theta$, we can find $\Theta_1$ that increases the log likelihood $\log P(X \mid \Theta)$ from the initial value $\Theta_0$.
Relation Between the Likelihood and the Lower Bound

\[ f(\Theta) \equiv \log P(X | \Theta) \]
\[ g(\Theta) \equiv J(P(H | X, \Theta_0), \Theta) \]
Q-function

Let \[ Q(\Theta, \Theta_0) \equiv \sum_{H} P(H \mid X, \Theta_0) \log P(X, H \mid \Theta) \]

\[ \arg \max_{\Theta} J(P(H \mid X, \Theta_0), \Theta) \]

\[ = \arg \max_{\Theta} \sum_{H} P(H \mid X, \Theta_0) \log \frac{P(X, H \mid \Theta)}{P(H \mid X, \Theta_0)} \]

\[ = \arg \max_{\Theta} \left\{ \sum_{H} P(H \mid X, \Theta_0) \log P(X, H \mid \Theta) - \sum_{H} P(H \mid X, \Theta_0) \log P(X, H \mid \Theta_0) \right\} \]

\[ = \arg \max_{\Theta} Q(\Theta, \Theta_0) \]

Finding the argmax of \( J \) is equal to finding the argmax of \( Q \)-function \( Q(\Theta, \Theta_0) \)
Expectation Maximization (EM) Algorithm

1. Prepare an initial parameter (or parameter set) $\Theta_0$

2. Given a parameter $\Theta_t$, obtain a Q-function $Q(\Theta, \Theta_t)$, which is an expectation of the log joint probability $\log P(X, H | \Theta)$ with $P(H|X, \Theta_t)$
   [E-step]

3. Maximizing the Q-function $Q(\Theta, \Theta_t)$ and obtain an updated parameter $\Theta_{t+1}$
   [M-step]

4. Go to step 2 until converge
The Process of the EM Algorithm

\( \Theta_0 \)  
Initial model parameters.  
May be just a random number.

\( \Theta_1 = \arg\max_\Theta [Q(\Theta, \Theta_0)] \)  
Update the parameters

\( \Theta_2 = \arg\max_\Theta [Q(\Theta, \Theta_1)] \)  
Update the parameters

\( \Theta_3 = \arg\max_\Theta [Q(\Theta, \Theta_2)] \)  
Update the parameters

\( \Theta_\infty = \text{local } \arg\max_\Theta \left[ \sum_H \log P(X, H | \Theta) \right] \)  
Gives a local maximum  
(not necessarily the global maximum)
EM for GMM

• Let’s consider 2-mix GMM

• Assume a training data $D$ with $n$ training samples $D=\{o_1, o_2, ..., o_n\}$, and an initial parameter set $\Theta_0 = \{\mu_1^{(0)}, \sigma_1^{(0)}, w_1^{(0)}, \mu_2^{(0)}, \sigma_2^{(0)}, w_2^{(0)}\}$ are given

• The posterior probability $P(m_i | D, \Theta_0)$ of the component index $m$ for the $i$-th training sample is:

$$P(m_i | D, \Theta_0) = P(m_i | o_i, \Theta_0) = \frac{P(m_i, o_i | \Theta_0)}{\sum_{m=1}^{2} P(m, o_i | \Theta_0)} = \frac{w_m N(o_i | \mu_m^{(0)}, \sigma_m^{(0)})}{\sum_{m=1}^{2} w_m N(o_i | \mu_m^{(0)}, \sigma_m^{(0)})}$$
Exercise 2.5

• Consider the 2-mix GMM of the previous page. Let $\gamma_m(i) = P(m | o_i, \Theta_0)$. Obtains the followings.

$\mu_1^{(1)} = \arg \max_{\mu_1} Q(\Theta, \Theta_0)$

$\sigma_1^{(1)} = \arg \max_{\sigma_1} Q(\Theta, \Theta_0)$

$w_1^{(1)} = \arg \max_{w_1} Q(\Theta, \Theta_0)$

Where

$Q(\Theta, \Theta_0) = \sum_{M = <m_1, m_2, \cdots m_n>} P(M | D, \Theta_0) \log P(D, M | \Theta)$

$= \sum_{i=1}^{n} \sum_{m=1}^{2} P(m | o_i, \Theta_0) \log P(o_i, m | \Theta) = \sum_{i=1}^{n} \sum_{m=1}^{2} \gamma_m(i) \log P(o_i, m | \Theta)$

$\Theta = \{\mu_1, \sigma_1, w_1, \mu_2, \sigma_2, w_2\}$

$i : \text{sample index}$

$m : \text{mixture component index}$
EM Estimation for HMM with Unknown Path

For HMM, path is a hidden variable

When O=\((a, b, b)\), possible paths are:
\(K_1(s_0, s_1, s_1, s_2, s_3)\) and \(K_2(s_0, s_1, s_2, s_2, s_3)\)

\[
P(O, K_1 | \Lambda) = 0.016128
\]
\[
P(O, K_2 | \Lambda) = 0.007168
\]

\begin{align*}
\text{E-step (expectation step)} \\
\text{Posterior probability} & \quad P(K_1 | O, \Lambda) = \frac{P(K_1, O | \Lambda)}{P(O | \Lambda)} = \frac{P(K_1, O | \Lambda)}{\sum_k P(K, O | \Lambda)} = \frac{0.016128}{0.016128 + 0.007168} \approx 0.7, \\
\text{Expectations of transitions} & \quad n(s_1 \rightarrow s_1) = 1 \times 0.7 + 0 \times 0.3 = 0.7 \quad n(s_1 \rightarrow s_2) = 1 \times 0.7 + 1 \times 0.3 = 1 \\
& \quad n(s_2 \rightarrow s_2) = 0 \times 0.7 + 1 \times 0.3 = 0.3 \quad n(s_2 \rightarrow s_3) = 1 \times 0.7 + 1 \times 0.3 = 1 \\
\text{Expectations of emissions} & \quad n(a | s_1) = 1 \times 0.7 + 1 \times 0.3 = 1 \quad n(b | s_1) = 1 \times 0.7 + 0 \times 0.3 = 0.7 \\
& \quad n(a | s_2) = 0 \times 0.7 + 0 \times 0.3 = 0 \quad n(b | s_2) = 1 \times 0.7 + 2 \times 0.3 = 1.3
\end{align*}

\begin{align*}
\text{M-step (maximization step)} \\
\text{New transition probabilities} & \quad p(s_1 \rightarrow s_1) = 0.7 / (0.7 + 1) = 0.41 \quad p(s_1 \rightarrow s_2) = 1 / (0.7 + 1) = 0.59 \\
& \quad p(s_2 \rightarrow s_2) = 0.3 / (0.3 + 1) = 0.23 \quad p(s_2 \rightarrow s_3) = 1 / (0.3 + 1) = 0.77 \\
\text{New emission probabilites} & \quad p(a | s_1) = 1 / (1 + 0.7) = 0.59 \quad p(b | s_1) = 0.7 / (1 + 0.7) = 0.41 \\
& \quad p(a | s_2) = 0 / (0 + 1.3) = 0.00 \quad p(b | s_2) = 1.3 / (0 + 1.3) = 1.00
\end{align*}
Appendix
Method of Lagrange Multiplier

Maximize $f(X)$ subject to $g(X) = 0$

Maximize $f(X) - \lambda g(X)$
with respect to $X$ and $\lambda$,
where $\lambda$ is a new parameter
Jensen’s Inequality

• If \( f(x) \) is a concave function, the following equation holds for arbitrary probability distribution of \( i \)

\[
\sum_{i} p(i)f(x_i) \leq f\left(\sum_{i} p(i)x_i\right)
\]

| Weighted average of function value \( f(x) \) | Function value of weighted average of \( x \) |

Example:

\[0.4f(x_1) + 0.6f(x_2) \leq f(0.4x_1 + 0.6x_2)\]