Speech and Language Processing

Lecture 3

Maximum likelihood estimation and EM algorithm

Information and Communications Engineering Course

Takahiro Shinozaki

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. 10/4 (remote)
   Introduction and Preparation
2. 10/4 (remote)
   Probability Distributions, Markov Models, Samplings
3. 10/6 (remote)
   Maximum Likelihood Estimation and EM Algorithm
4. 10/6 (remote)
   Bayesian Networks and Bayesian Inference
5. 10/7 (remote)
   Neural networks
6. 10/7 (remote)
   Reinforcement Learning

Maximum Likelihood (ML) Method

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

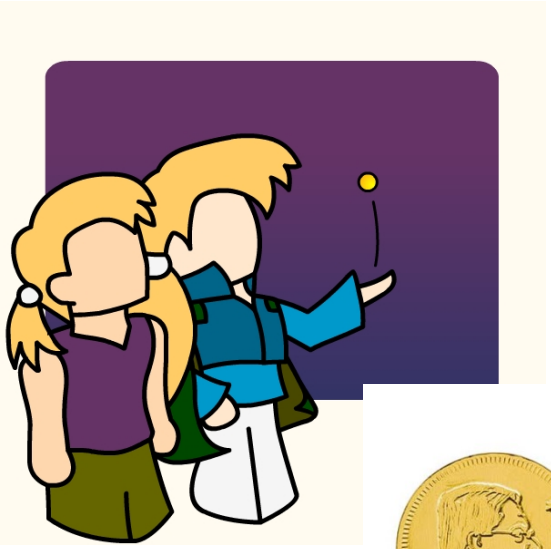
- Maximum likelihood method estimates the parameters by maximizing likelihood $p(D|\theta)$

$$\hat{\theta} = \arg\max_{\theta} p(D \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} p(x_i \mid \theta)$$

Probability of the data set D is decomposed to a product of samples when they are drawn independently

Bernoulli Distribution

- Probability distribution of a binary random variable which takes value 1 with probability μ and value 0 with probability 1-μ



x	0	1
Bern(x)	1-μ	μ

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

Is the result Head or Tail?

ML Estimation for Bernoulli Distribution

- Parameter θ in this case is : μ
- Training sample $x_i = 0 \; or \; 1$

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

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

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

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

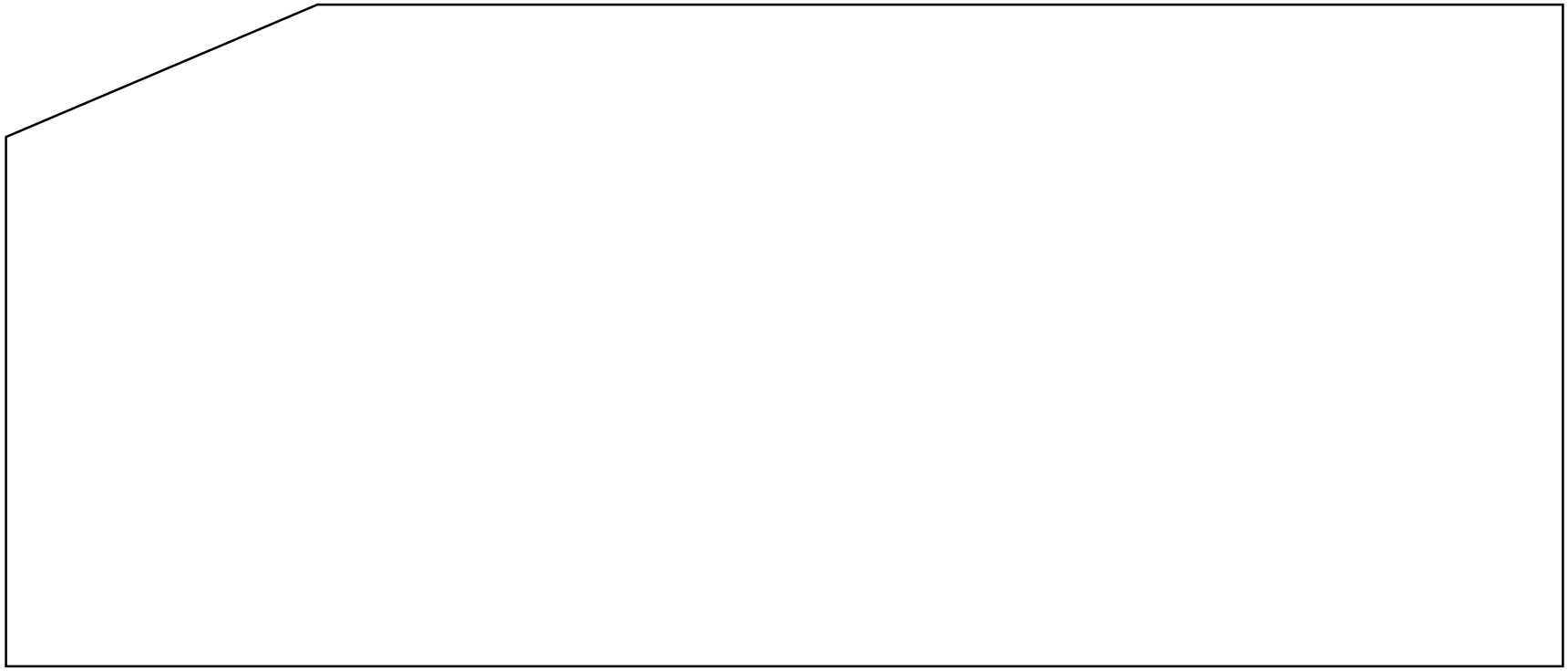
$$\Leftrightarrow \quad \mu = \frac{1}{n} \sum_{i} x_i$$

Taking log does not change the problem and makes the equation a bit easier

n: the number of samples

Exercise 3.1

- You tossed a winded coin 100 times, and got 62 heads and 38 tails. Estimate the probability μ of getting head with the coin by the ML method



Categorical Distribution

- As a generalization of the Bernoulli Distribution, lets consider a discrete random variable X that takes K values

X	1	2	…	K
1-of-K $<x_1,x_2,\ldots,x_K>$	$<1,0,..,0>$	$<0,1,..,0>$	…	$<0,0,..,1>$
$p(X)$	μ_1	μ_2	…	μ_K

$$p(\mathbf{x}\mid\boldsymbol{\mu})=\prod_{k=1}^{K}\mu_k^{\,x_k}$$

ML for Categorical Distribution

- Parameter θ in this case is : $\boldsymbol{\mu} = \{\mu_1, \mu_2, \ldots, \mu_K\}$
- 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 $j \neq k$

$$\hat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} p(D \mid \boldsymbol{\mu}) = \arg\max_{\boldsymbol{\mu}} \prod_{i=1}^{n} \prod_{k=1}^{K} \mu_k^{x_{i,k}} = \arg\max_{\boldsymbol{\mu}} \prod_{k=1}^{K} \mu_k^{m_k}$$

$$\sum_{k=1}^{K} \mu_k = 1$$

Constraint

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

$$m_k = \sum_{i=1}^{n} x_{i,k}$$

This is a maximization problem with a constraint

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

Solution

$$\hat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} \prod_{k=1}^{K} \mu_k^{\,m_k} = \arg\max_{\boldsymbol{\mu}} \sum_{k=1}^{K} m_k \log(\mu_k)$$

$$\sum_{k=1}^{K} \mu_k = 1$$

$$\Rightarrow \quad \hat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} \left\{ \sum_{k=1}^{K} m_k \log \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right) \right\}$$

$$\Rightarrow \quad \mu_k = \frac{m_k}{n}$$

Exercise 3.2

- Show the derivation process of obtaining $\mu_k = \frac{m_k}{n}$

  for the categorical distribution by maximizing

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

  where λ 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 θ in this case is : $\{\mu, \sigma\}$
Training sample x_i is a real value

ML estimation of Gaussian distribution

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} N(x_i \mid \theta) = \arg\max_{\theta} \sum_{i=1}^{n} \log\left(N(x_i \mid \theta)\right)$$

$$\int_{-\infty}^{\infty} N(x \mid \mu, \sigma^2)\, dx = 1$$

Exercise 3.3

- 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

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

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

$$\arg\max_{\mu_1, \sigma_1, w_1, \mu_2, \sigma_2, w_2} L(\mu_1, \sigma_1, w_1, \mu_2, \sigma_2, w_2) \quad \Leftrightarrow$$

$$\arg\max_{w_1, w_2} \sum_{i=1}^{n} \log(w_{m_i})$$

$$\arg\max_{\mu_1, \sigma_1} \sum_{i|m_i=1} \log(N(o_i \mid \mu_1, \sigma_1))$$

$$\arg\max_{\mu_2, \sigma_2} \sum_{i|m_i=2} \log(N(o_i \mid \mu_2, \sigma_2))$$

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

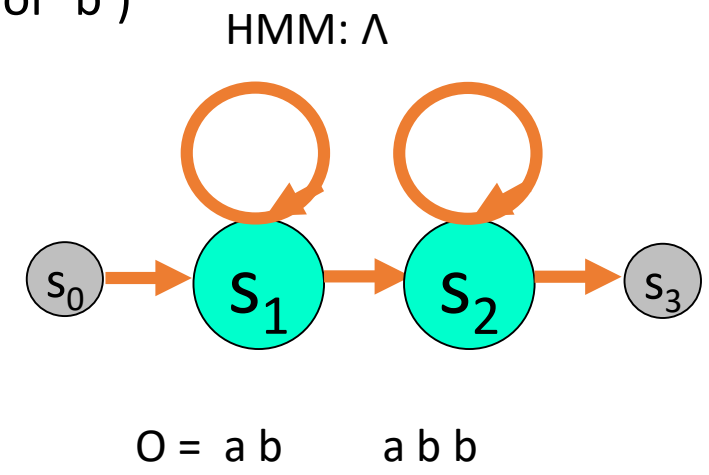
$p(s_1 \rightarrow s_1) = 1/2$, $p(s_1 \rightarrow s_2) = 1/2$

$p(s_2 \rightarrow s_2) = 2/3$, $p(s_2 \rightarrow s_3) = 1/3$

Emission probability

$p(a|s_1) = 1/2$, $p(b|s_1) = 1/2$

$p(a|s_2) = 1/3$, $p(b|s_2) = 2/3$



ML Estimation for GMM

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

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

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

GMM

ML Estimation for GMM (Cont.)

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

$$\frac{\partial L}{\partial \mu_m} =$$

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 $P(m, x) = P(m)N(x|\mu_m, \sigma_m)$

$$GMM(x) = \sum_{m=1}^{M} w_m N(x \mid \mu_m, \sigma_m) = \sum_{m=1}^{M} P(m) N(x \mid \mu_m, \sigma_m)$$

- 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 Σ for the marginalization is often problematic for optimization

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

$$= \arg\max_{\Theta}\left[\sum_i \log P(x_i \mid \Theta)\right]$$

$$= \arg\max_{\Theta}\left[\sum_i \log \sum_h P(x_i, h \mid \Theta)\right]$$

nuisance

Jensen Lower Bound of Likelihood

Let X be an observed variable, H be a hidden variable, and Θ be a parameter.

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

For arbitrary $q(H)$

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

By Jensen's inequality

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

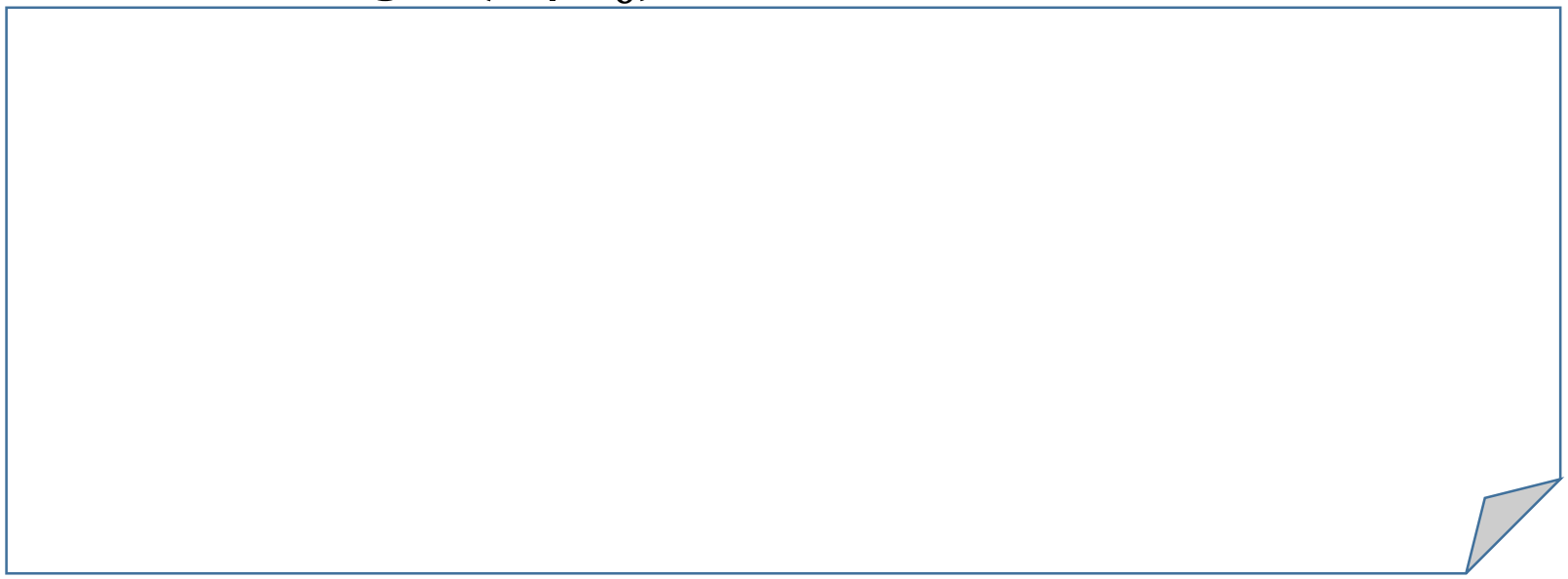
$$\equiv J(q, \Theta)$$

(Lower bound of likelihood)

This inequality holds for arbitrary q and arbitrary Θ

Exercise 3.4

- Assume you have an initial model parameter Θ_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 $logP(X|\Theta_0)$



$$\Rightarrow \quad \log P(X \mid \Theta_0) = J(P(H \mid X, \Theta_0), \Theta_0)$$

Maximization of the Lower Bound

- Assume we have an initial model parameter Θ_0.

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

$$\Rightarrow \log P\left(X \mid \Theta_0\right) \le \log P\left(X \mid \Theta_1\right)$$

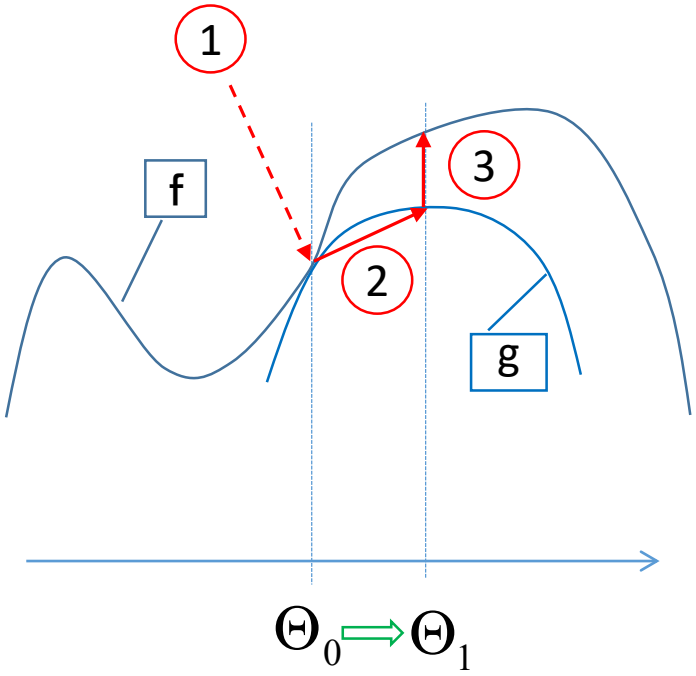
Because:

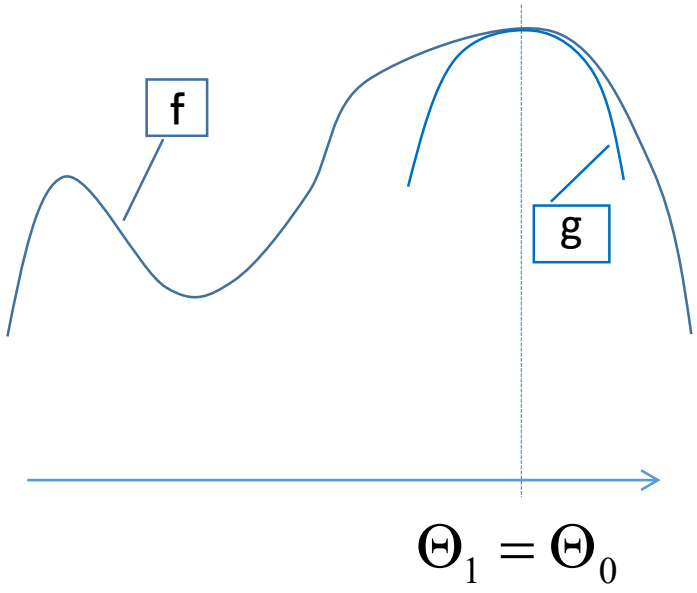
$$\begin{cases} \log P\left(X \mid \Theta_0\right) = J\left(P\left(H \mid X, \Theta_0\right), \Theta_0\right) & (1) \\ J\left(P\left(H \mid X, \Theta_0\right), \Theta_0\right) \le J\left(P\left(H \mid X, \Theta_0\right), \Theta_1\right) & (2) \\ J\left(P\left(H \mid X, \Theta_0\right), \Theta\right) \le \log P\left(X \mid \Theta\right) \quad for \quad \forall\Theta & (3) \end{cases}$$

By maximizing the lower bound J with respective to Θ, we can find Θ_1 that increases the log likelihood log $P(X|\Theta)$ from the initial value Θ_0

Relation Between the Likelihood and the Lower Bound

$$f(\Theta) \equiv \log P(X \mid \Theta)$$
$$g(\Theta) \equiv J(P(H \mid X, \Theta_0), \Theta)$$





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) Θ_0
2. Given a parameter Θ_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 Θ_{t+1}
   [M-step]
4. Go to step 2 until converge

The Process of the EM Algorithm

Θ_0 Initial model parameters.
May be just a random number.

$$\Theta_1 = \arg\max_{\Theta} \left[Q(\Theta, \Theta_0) \right]$$ Update the parameters

$$\Theta_2 = \arg\max_{\Theta} \left[Q(\Theta, \Theta_1) \right]$$ Update the parameters

$$\Theta_3 = \arg\max_{\Theta} \left[Q(\Theta, \Theta_2) \right]$$ Update the parameters

$$\Theta_\infty = local \arg\max_{\Theta} \left[\sum_H \log P(X, H \mid \Theta) \right]$$

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 $D=\{o_1, o_2, \ldots, o_n\}$, and an initial parameter set $\Theta_0 = \left\{\mu_1^{(0)}, \sigma_1^{(0)}, w_1^{(0)}, \mu_2^{(0)}, \sigma_2^{(0)}, w_2^{(0)}\right\}$ are given
- The posterior probability $P(m_i|D,\Theta_0)$ of the component index m for the i-th training sample is :

$$P(m_i \mid D,\Theta_0) = P(m_i \mid o_i,\Theta_0) = \frac{P(m_i, o_i \mid \Theta_0)}{\sum_{m=1}^{2} P(m, o_i \mid \Theta_0)} = \frac{w_{m_i} N(o_i \mid \mu_{m_i}^{(0)}, \sigma_{m_i}^{0})}{\sum_{m=1}^{2} w_m N(o_i \mid \mu_m^{(0)}, \sigma_m^{0})}$$

Exercise 3.5

- 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

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

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

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

i : sample index

m : 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:

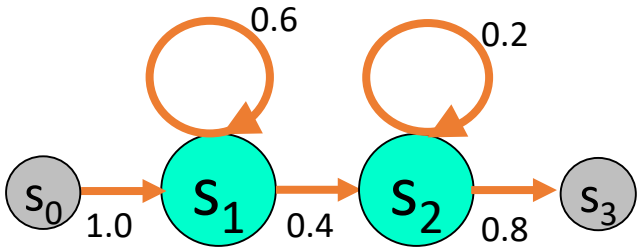
K1(s_0,s_1,s_1,s_2,s_3) and K2(s_0,s_1,s_2,s_2,s_3)

a b b a b b

$$P(\mathbf{O}, K_1 \mid \Lambda) = 0.016128$$

$$P(\mathbf{O}, K_2 \mid \Lambda) = 0.007168$$

Initial HMM: Λ



[p(a),p(b)] = S_1: [0.7, 0.3] S_2: [0.6, 0.4]

E-step (expectation step)

Posterior probability

$$P(K_1 \mid O, \Lambda) = \frac{P(K_1, O \mid \Lambda)}{P(O \mid \Lambda)} = \frac{P(K_1, O \mid \Lambda)}{\sum_K P(K, O \mid \Lambda)} = \frac{0.016128}{0.016128 + 0.007168} \approx 0.7, \quad P(K_2 \mid O, \Lambda) \approx 0.3$$

Expectations of transitions

$n(s_1 \to s_1)=1*0.7+0*0.3=0.7$ $n(s_1 \to s_2)=1*0.7+1*0.3=1$

$n(s_2 \to s_2)=0*0.7+1*0.3=0.3$ $n(s_2 \to s_3)=1*0.7+1*0.3=1$

Expectations of emissions

$n(a \mid s_1)=1*0.7+1*0.3=1$ $n(b \mid s_1)=1*0.7+0*0.3=0.7$

$n(a \mid s_2)=0*0.7+0*0.3=0$ $n(b \mid s_2)=1*0.7+2*0.3=1.3$

M-step (maximization step)

New transition probabilities

$p(s_1 \to s_1)=0.7/(0.7+1)=0.41$ $p(s_1 \to s_2)=1/(0.7+1)=0.59$

$p(s_2 \to s_2)=0.3/(0.3+1)=0.23$ $p(s_2 \to s_3)=1/(0.3+1)=0.77$

New emission probabilites

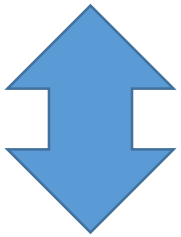
$p(a \mid s_1)=1/(1+0.7)=0.59$ $p(b \mid s_1)=0.7/(1+0.7)=0.41$

$p(a \mid s_2)=0/(0+1.3)=0.00$ $p(b \mid s_2)=1.3/(0+1.3)=1.00$

Appendix

Method of Lagrange Multiplier

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



Equivalent

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

Jensen's Inequality

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

$$\sum_i p(i) f(x_i) \le f\left(\sum_i p(i) x_i\right)$$

Weighted average of function value $f(x)$

Function value of weighted average of x

Example :

$$0.4 f(x_1) + 0.6 f(x_2) \le f(0.4x_1 + 0.6x_2)$$