# Statistical Computing

## Hidden Markov Models

## Expectation-Maximization (EM) Algorithm

## Maximum-A-Posteriori (MAP) Estimation

Uwe Menzel, 2008

uwe.menzel@matstat.org

www.matstat.org

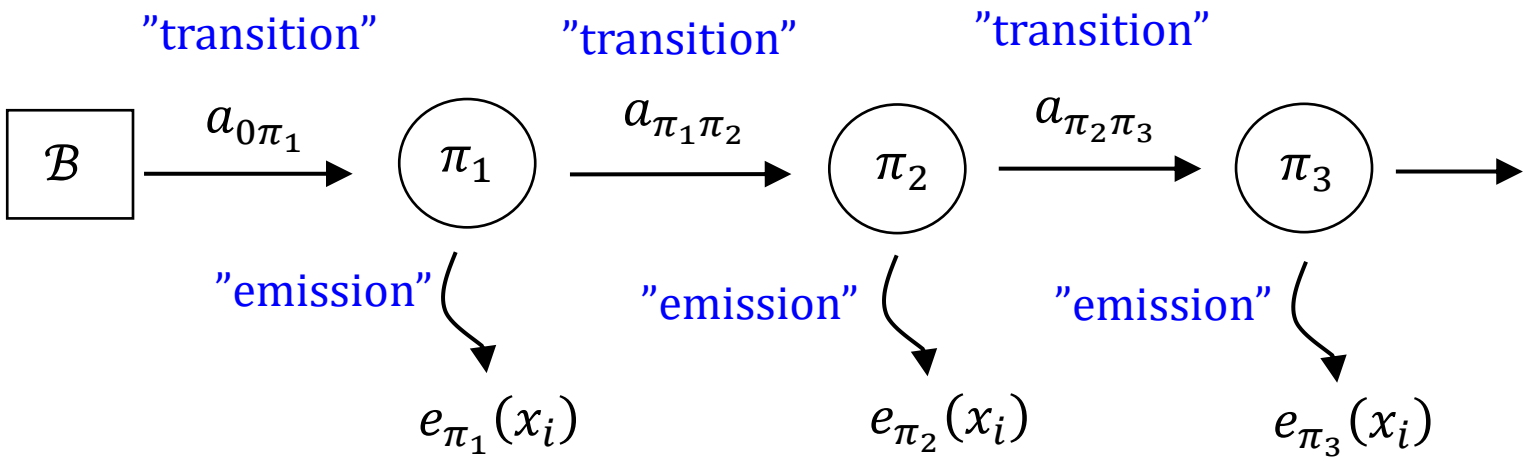
# Outline

1. Hidden Markov Models (HMM)
    - short repetition
2. Most probable state path
    - Viterbi algorithm
3. Forward-Backward algorithm
    - forward and backward probabilities
4. Expectation Maximization (EM)
    - notation adapted for HMM's
5. EM for HMM
    - with an alternative derivation of the general EM scheme
6. Maximum A Posteriori (MAP) estimation for HMM
    - very short introduction

# 1. Hidden Markov Models

see HMM_Lecture_2.pdf

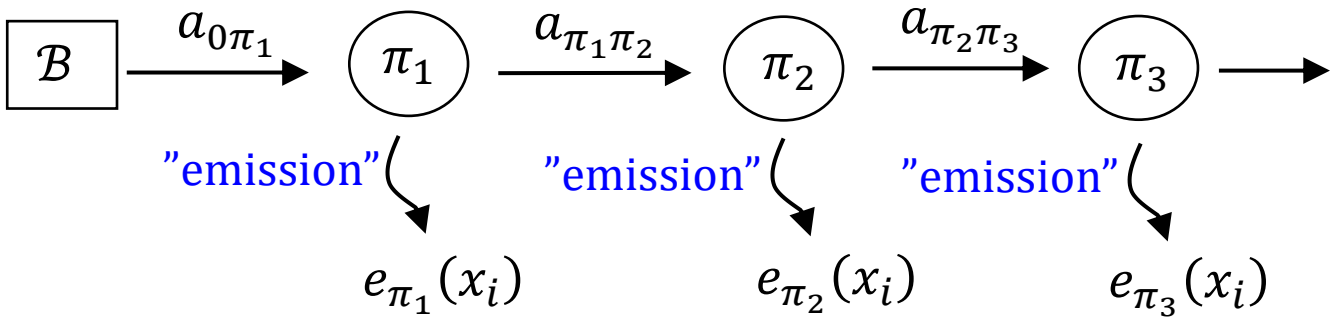
- we have a Markov chain, not visible for the observer (hidden)
- every state of the Markov chain can "emit" an element of a set of observable characters (symbols) with some probability
- the Markov chain itself forms the state path $\pi_i$
- the emission probabilities are labeled $e_{\pi_k}(x_i)$, defining the probability that the state $\pi_k$ emits the symbol $x_i$.




**Note**: transitions probabilities from the (virtual) begin state will be denoted $a_{\mathcal{B}\pi_i}$ or $a_{0\pi_i}$. They are also labelled start probabilities.

# Hidden Markov Model

see HMM_Lecture_2.pdf




Joint probability of the chain of states **and** symbols:

$$P(x, \pi \mid \theta) = a_{0\pi_1} \cdot e_{\pi_1}(x_1) \cdot a_{\pi_1\pi_2} \cdot e_{\pi_2}(x_2) \cdot a_{\pi_2\pi_3} \cdot \ldots$$

$$P(x, \pi \mid \theta) = a_{0\pi_1} \cdot \prod_{i=1}^{L} e_{\pi_i}(x_i) \cdot a_{\pi_i\pi_{i+1}}$$

$\theta = (a_{kl}, e_k\ a_{0k})$ stands for whole parameter set, $a_{0k}$ = start probabilities

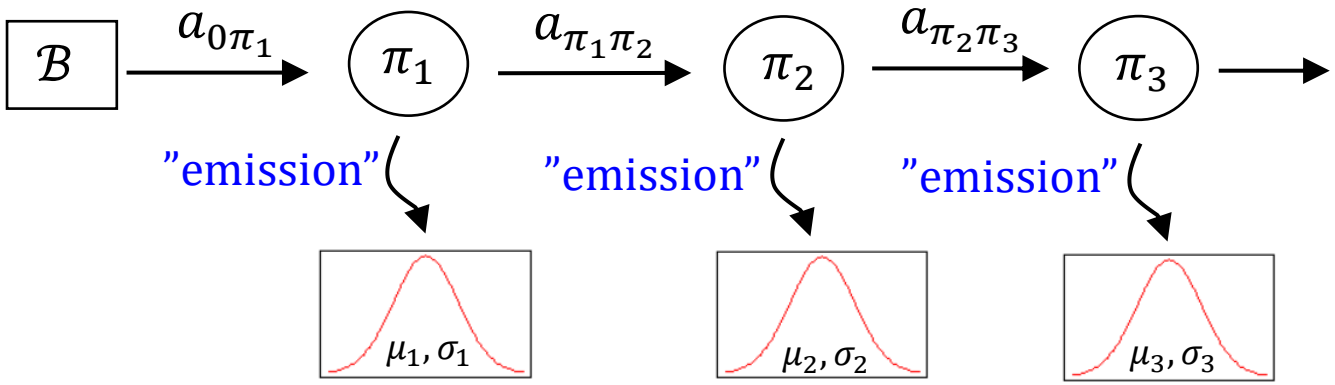
$a_{kl} = P(\pi_i = l \mid \pi_{i-1} = k)$ transition probability $k \rightarrow l$ (within state path)

$e_k(b) = P(x_i = b \mid \pi_i = k)$ emission probabilities, from state $k$ to symbol $b$

# Continuous Density Hidden Markov Model

The emission probabilities outgoing from state $k$ can also be modelled by a continuous random variable. A prominent example is Gaussian emission, where state $k$ emits the observation $x_i$ according to a normal distribution with mean $\mu_k$ and standard deviation $\sigma_k$:

$$P(x_i \mid \pi_i = k) = \frac{1}{\sqrt{2\pi} \cdot \sigma_k} \exp\left[-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right]$$




Having $k$ different states, we have $k$ Gaussians with different $\mu_k$ and $\sigma_k$. In many cases, it is favourable to model the emission probabilities with a mixture of Gaussians.

# Parameters of the Markov chain and their properties

$\theta = (a_{kl}, e_k, a_{0k})$

| | | |
|---|---|---|
| $a_{0k} = P(\pi_1 = k)$ | initial states | $1 \le k \le K$ |
| $a_{kl} = P(\pi_{i+1} = l \mid \pi_i = k)$ | transitions | $1 \le i \le L-1, 1 \le k, l \le K$ |
| $e_k(x_i) = P(x_i \mid \pi_i = k)$ | emissions | $1 \le i \le L, 1 \le k \le K$ |

$K$ : number of states ; $L$ : length of the chain

The $a_{kl}$ do not depend on $i$, because we assume a homogenous chain:

$$a_{kl} = P(\pi_{i+1} = l \mid \pi_i = k) = P(\pi_i = l \mid \pi_{i-1} = k)$$

$\sum_{x_i} e_k(x_i) = 1$ one of the $x_i$ **must** be emitted from state $k$ (for discrete $x_i$)

$$\int_x e_k(x)\,dx = 1 \quad \text{when } x \in \mathbb{R}$$

$\sum_l a_{kl} = 1$ the state $k$ **must** switch over to some other state $l$

# 2. Find the state path given the observations and the model

- observations: $\boldsymbol{x} = \{x_i\}$ ; the $x_i$ are known variables
- model: $\theta = (a_{kl}, e_k\ a_{0k})$ ; the symbol $\theta$ stands for all parameters

$$P(x, \pi \mid \theta) = P(\pi \mid x, \theta) \cdot P(x \mid \theta)$$ conditional probability

$$\pi^\star = \underset{\pi_1 \ldots \pi_L}{\operatorname{argmax}} P\left(\pi_1 \ldots \pi_L \mid x_1 \ldots x_L, \theta\right)$$

$$\pi^\star = \underset{\pi_1 \ldots \pi_L}{\operatorname{argmax}} P\left(\pi_1 \ldots \pi_L, x_1 \ldots x_L \mid \theta\right)$$

} most probable path

- $\pi^*$: state path that lies behind the observed data
- $K^L$ possible paths
  - $\rightarrow 2 \cdot L \cdot K^L$ arithmetic operations
  - $\rightarrow$ brute force approach fails

# Find the single best path for given $x$ and $\theta$
## - Viterbi algorithm -

$$\begin{aligned}\pi^\star &= \underset{\pi_1 \ldots \pi_L}{\operatorname{argmax}} P\left(\pi_1 \ldots \pi_L \mid x_1 \ldots x_L, \theta\right) \\ &= \underset{\pi_1 \ldots \pi_L}{\operatorname{argmax}} P\left(\pi_1 \ldots \pi_L, x_1 \ldots x_L \mid \theta\right)\end{aligned}$$

Define auxiliary variable $\delta_i(k)$:

$$\delta_i(k) = \max_{\pi_1, \ldots, \pi_{i-1}} P\left(\pi_1, \ldots, \pi_{i-1}, \pi_i = k, x_1, \ldots, x_i \mid \theta\right)$$

This is the probability of the chain based on the most probable path up to position $i-1$, ending in state $k$ at position $i$. It can be recursively calculated:

$$\delta_{i+1}(l) = \max_k \left\{\delta_i(k) \cdot a_{kl}\right\} \cdot b_l(x_{i+1})$$

- indices $l$ and $k$ run through all states 1 ... $K$
- index $i$ runs through positions in the Markov chain 1 ... $L$

# - Viterbi algorithm -

$$\delta_i(k) = \max_{\pi_1,\ldots,\pi_{i-1}} P\left(\pi_1,\ldots,\pi_{i-1},\pi_i = k, x_1,\ldots,x_i \mid \theta\right)$$ auxiliary variable

$$\delta_{i+1}(l) = \max_k \left\{\delta_i(k)\cdot a_{kl}\right\}\cdot b_l(x_{i+1})$$ recursion

**Proof of recursion:** $\pi_1,\ldots,\pi_i \equiv \pi_{1:i}$

$$\delta_{i+1}(l) = \max_{\pi_1,\ldots,\pi_i} P\left(\pi_{1:i},\pi_{i+1} = l, x_{1:i+1} \mid \theta\right)$$ extract $x_{i+1}$, using cond. probability

$$\delta_{i+1}(l) = \max_{\pi_1,\ldots,\pi_i} P\left(\pi_{1:i},\pi_{i+1} = l, x_{1:i} \mid \theta\right)\cdot P\left(x_{i+1} \mid \pi_{1:i},\pi_{i+1} = l, x_{1:i},\theta\right)$$ Markov property

$$\delta_{i+1}(l) = \max_{\pi_1,\ldots,\pi_i} P\left(\pi_{1:i},\pi_{i+1} = l, x_{1:i} \mid \theta\right)\cdot P\left(x_{i+1} \mid \pi_{i+1} = l,\theta\right)$$ extract $\pi_{i+1}$, cond. probability

$$\delta_{i+1}(l) = \max_{\pi_1,\ldots,\pi_i} P\left(\pi_{1:i}, x_{1:i} \mid \theta\right)\cdot P\left(\pi_{i+1} = l \mid \pi_{1:i}, x_{1:i},\theta\right)\cdot e_l\left(x_{i+1}\right)$$ Markov property

$$\delta_{i+1}(l) = \max_{\pi_1,\ldots,\pi_i} P\left(\pi_{1:i}, x_{1:i} \mid \theta\right)\cdot P\left(\pi_{i+1} = l \mid \pi_i,\theta\right)\cdot e_l\left(x_{i+1}\right)$$ choose maximal $\pi_i = k$

$$\delta_{i+1}(l) = \max_k \max_{\pi_1,\ldots,\pi_{i-1}} P\left(\pi_{1:i-1},\pi_i = k, x_{1:i} \mid \theta\right)\cdot P\left(\pi_{i+1} = l \mid \pi_i = k,\theta\right)\cdot e_l\left(x_{i+1}\right)$$

$\delta_i(k)$ $a_{kl}$

# The Viterbi algorithm

Initialise all $\delta$ at position 1 of the chain, for all states 1 ... $K$:

$$\delta_1(l) = a_{0l} \cdot e_l(x_1) \qquad 1 \leq l \leq K$$

$a_{0l} = P(\pi_1 = l)$ : probability that the chain starts with state $l$

$$\psi_1(l) = 0$$

pointer which will be used when tracing back from the end of the chain

$$\delta_{i+1}(l) = \max_k \{\delta_i(k) \cdot a_{kl}\} \cdot e_l(x_{i+1})$$

index $l$ and $k$ run through states
index $i$ runs through positions in chain
$1 \leq i \leq L-1; 1 \leq l, k \leq K$

$$\psi_i(l) = \operatorname*{argmax}_k \{\delta_{i-1}(k) \cdot a_{kl}\}$$

pointer $1 \leq i \leq L-1; 1 \leq l, k \leq K$

$$P^\star(x, \pi \mid \theta) = \max_l \delta_L(l)$$

the max. probability is the maximum $\delta$ (over all states) at the end of the chain

$$\pi_L^\star = \operatorname*{argmax}_l \delta_L(l)$$

the last element (at position L) of the most probable path

$$\pi_i^\star = \psi_{i+1}(\pi_{i+1}^\star)$$

the other elements of the most probable path can be backtracked using the pointer

# 3. The Forward and Backward probabilities

We'll need the forward and backward probabilities later when deriving the Expectation Maximization algorithm for Hidden Markov Models.

$$\alpha_i(k) = P(x_1, x_2, \ldots, x_i, \pi_i = k \mid \theta)$$ Defintion, forward probability

This is the probability of the observations up to position $i$ when having state $k$ at position $i$. It can be calculated recursively:

$$\alpha_{i+1}(l) = \left[ \sum_{k=1}^{K} \alpha_i(k) \cdot a_{kl} \right] \cdot e_l(x_{i+1})$$

- indices $l$ and $k$ run through all states $1 \ldots K$
- index $i$ runs through positions in the Markov chain $1 \ldots L - 1$

The recursion starts with

$$\alpha_1(l) = a_{0l} \cdot e_l(x_1) \qquad l = 1 \ldots K$$

$a_{0l}$ is the probability that the chain starts with state $l$ : $a_{0l} = P(\pi_1 = l)$.

# Forward probabilities

$$\alpha_i(k) = P(x_1, x_2, \ldots, x_i, \pi_i = k \mid \theta)$$ Definition, forward probability

Recursion:

$$\alpha_{i+1}(l) = \left[ \sum_{k=1}^{K} \alpha_i(k) \cdot a_{kl} \right] \cdot e_l(x_{i+1}) \qquad \alpha_1(l) = a_{0l} \cdot e_l(x_1)$$

**Proof of recursion:** $x_1, \ldots, x_i \equiv x_{1:i}$ observations 1 ... *i*

Initiation:

$$\begin{aligned} \alpha_1(k) &= P(x_1, \pi_1 = k \mid \theta) && \text{cond. probability} \quad k = 1 \ldots K \\ &= P(\pi_1 = k \mid \theta) \cdot P(x_1 \mid \pi_1 = k, \theta) && \text{use definitions} \\ &= a_{0k} \cdot e_k(x_1) \end{aligned}$$

# Forward probabilities

$x_1, \ldots, x_i \equiv x_{1:i}$

**Proof of recursion:**

$$\alpha_{i+1}(k) = P(x_{1:i+1}, \pi_{i+1} = k \mid \theta) = \sum_{j=1}^{K} P(x_{1:i+1}, \pi_i = j, \pi_{i+1} = k \mid \theta)$$ marginal rule

cond. probability

$$\alpha_{i+1}(k) = \sum_{j=1}^{K} P(x_{1:i}, \pi_i = j, \pi_{i+1} = k \mid \theta) \cdot P(x_{i+1} \mid x_{1:i}, \pi_i = j, \pi_{i+1} = k, \theta)$$

$$\alpha_{i+1}(k) = \sum_{j=1}^{K} P(x_{1:i}, \pi_i = j, \pi_{i+1} = k \mid \theta) \cdot P(x_{i+1} \mid \pi_{i+1} = k, \theta)$$ Markov property

cond. probability

$$\alpha_{i+1}(k) = \sum_{j=1}^{K} P(x_{1:i}, \pi_i = j \mid \theta) \cdot P(\pi_{i+1} = k \mid x_{1:i}, \pi_i = j, \theta) \cdot P(x_{i+1} \mid \pi_{i+1} = k, \theta)$$

Markov property

$$\alpha_{i+1}(k) = \sum_{j=1}^{K} P(x_{1:i}, \pi_i = j \mid \theta) \cdot P(\pi_{i+1} = k \mid \pi_i = j, \theta) \cdot P(x_{i+1} \mid \pi_{i+1} = k, \theta)$$

use definitions

$$\alpha_{i+1}(k) = \left[ \sum_{j=1}^{K} \alpha_i(j) \cdot a_{jk} \right] \cdot e_k(x_{i+1})$$ q.e.d.

# Forward probabilities

Probability of the complete chain: $$P\left(x_{1:L} \mid \theta\right) = \sum_{k=1}^{K} \alpha_L(k)$$

marginal rule

**Proof:** $$P\left(x_{1:L} \mid \theta\right) = \sum_{k=1}^{K} P\left(x_{1:L}, \pi_L = k \mid \theta\right) = \sum_{k=1}^{K} \alpha_L(k)$$

# The Forward algorithm

**Initiation:** $$\alpha_1(k) = a_{0k} \cdot e_k\left(x_1\right) \qquad k = 1 \ldots K$$

**Recursion:** $$\alpha_{i+1}(l) = \left[\sum_{k=1}^{K} \alpha_i(k) \cdot a_{kl}\right] \cdot e_l(x_{i+1}) \qquad \begin{array}{l} l = 1 \ldots K \\ i = 1 \ldots L-1 \end{array}$$

**Termination:** $$P\left(x \mid \theta\right) = \sum_{k=1}^{K} \alpha_L(k)$$ probability of the observation

# Backward probability

We'll need the forward and backward probabilities later when deriving the Expectation Maximization algorithm for Hidden Markov Models.

$$\beta_i(k) = P(x_{i+1}, x_{i+2}, \ldots, x_L \mid \pi_i = k, \theta)$$ Definition, backward probability

This is the probability of the observations from positions $i + 1$ up to the end of the chain, when having state $k$ at position $i$. It can be calculated recursively:

$$\beta_i(l) = \sum_{k=1}^{K} a_{lk} \cdot e_k(x_{i+1}) \cdot \beta_{i+1}(k)$$

- indices $l$ and $k$ run through all states 1 ... $K$
- index $i$ runs through positions in the Markov chain $L - 1$ ... 1

The recursion starts with

$$\beta_L(l) = 1 \qquad\qquad l = 1 \ldots K$$

# Backward probability

$$\beta_i(k) = P(x_{i+1}, x_{i+2}, \ldots, x_L \mid \pi_i = k, \theta)$$ backward probability

$$\beta_i(l) = \sum_{k=1}^{K} a_{lk} \cdot e_k(x_{i+1}) \cdot \beta_{i+1}(k) \qquad \beta_L(l) = 1$$

**Proof of recursion:**

marginal rule

$$\beta_i(l) = P(x_{i+1:L} \mid \pi_i = l, \theta) = \sum_{k=1}^{K} P\left(\underbrace{x_{i+1:L}}_{a}, \underbrace{\pi_{i+1} = k}_{b} \mid \underbrace{\pi_i = l, \theta}_{c}\right)$$

$$P(a,b) = P(a \mid b) \cdot P(b)$$
$$P(a,b \mid c) = P(a \mid b,c) \cdot P(b \mid c)$$

$$\beta_i(l) = \sum_{k=1}^{K} P\left(\overbrace{x_{i+1:L}}^{a} \mid \overbrace{\pi_{i+1} = k, \pi_i = l, \theta}^{b\ c}\right) \cdot P\left(\overbrace{\pi_{i+1} = k}^{b} \mid \overbrace{\pi_i = l, \theta}^{c}\right)$$

cond. probability, Markov property

$$\beta_i(l) = \sum_{k=1}^{K} P\left(x_{i+2:L} \mid \pi_{i+1} = k, \pi_i = l, \theta\right) \cdot P\left(x_{i+1} \mid \pi_{i+1} = k, \theta\right) \cdot P\left(\pi_{i+1} = k \mid \pi_i = l, \theta\right)$$

Markov property

$$\beta_i(l) = \sum_{k=1}^{K} P\left(x_{i+2:L} \mid \pi_{i+1} = k, \theta\right) \cdot P\left(x_{i+1} \mid \pi_{i+1} = k, \theta\right) \cdot P\left(\pi_{i+1} = k \mid \pi_i = l, \theta\right)$$

use definitions

$$\beta_i(l) = \beta_{i+1}(k) \cdot e_k\left(x_{i+1}\right) \cdot a_{lk}$$

# Backward probability

Show that the recursion must start with $\beta_L(l) = 1 \qquad l = 1 \dots K$

$$\beta_{L-1}(l) = P(x_L \mid \pi_{L-1} = l, \theta)$$ according to the definition of $\beta$

$$\beta_{L-1}(l) = \sum_{k=1}^{K} P(\underbrace{x_L}_{\text{a}}, \underbrace{\pi_L = k}_{\text{b}} \mid \underbrace{\pi_{L-1} = l}_{\text{c}}, \theta)$$ marginal rule

$$P(a,b) = P(a \mid b) \cdot P(b)$$
$$P(a,b \mid c) = P(a \mid b,c) \cdot P(b \mid c)$$

$$\beta_{L-1}(l) = \sum_{k=1}^{K} \overbrace{P(x_L}^{\text{a}} \mid \overbrace{\pi_L = k}^{\text{b}}, \overbrace{\pi_{L-1} = l}^{\text{c}}, \theta) \cdot \overbrace{P(\pi_L = k}^{\text{b}} \mid \overbrace{\pi_{L-1} = l}^{\text{c}}, \theta)$$

Markov property

$$\beta_{L-1}(l) = \sum_{k=1}^{K} P(x_L \mid \pi_L = k, \theta) \cdot P(\pi_L = k \mid \pi_{L-1} = l, \theta)$$

$$\beta_{L-1}(l) = \sum_{k=1}^{K} e_k(x_L) \cdot a_{lk}$$ with $\beta_L(k) = 1 \; \forall \; k$ we can write:

$$\beta_{L-1}(l) = \sum_{k=1}^{K} e_k(x_L) \cdot a_{lk} \cdot \beta_L(k)$$ which is the general recursion for $i = L - 1$

# The posterior probabilities

Another auxiliary variable is $\gamma_i(k)$:

$$\gamma_i(k) = P(\pi_i = k \mid x, \theta)$$ posterior probability for state $\pi_i$

This is the probability of having state $k$ at position $i$, given the observations $x$ and the model $\theta$. It can be expressed in terms of the forward- and backward probabilities:

$$\gamma_i(k) = \frac{\alpha_i(k) \cdot \beta_i(k)}{P(x \mid \theta)} = \frac{\alpha_i(k) \cdot \beta_i(k)}{\sum_{k=1}^{K} \alpha_i(k) \cdot \beta_i(k)} \qquad \sum_{k=1}^{K} \gamma_i(k) = 1$$

The individually most likely state at position $i$ is:

$$\pi_i^\star = \underset{k}{\operatorname{argmax}}\, \gamma_i(k) = \underset{k}{\operatorname{argmax}}\, P(\pi_i = k \mid x, \theta)$$

These expressions are interesting if the most probable state at some particular position of the chain is in the main focus, rather than the most probable path spanning the whole chain as calculated by the Viterbi algorithm.

# Show that $P(x \mid \theta) = \sum_k \alpha_i(k) \cdot \beta_i(k)$

forward probability:

$$\alpha_i(k) = P(x_{1:i}, \pi_i = k \mid \theta)$$

backward probability:

$$\beta_i(k) = P(x_{i+1:L} \mid \pi_i = k, \theta)$$

$$P(x \mid \theta) = \sum_{k=1}^{K} P(x, \pi_i = k \mid \theta)$$

marginal rule

conditional probability

$$P(x \mid \theta) = \sum_{k=1}^{K} P(x_{1:i}, \pi_i = k \mid \theta) \cdot P(x_{i+1:L} \mid x_{1:i}, \pi_i = k, \theta)$$

Markov property

$$P(x \mid \theta) = \sum_{k=1}^{K} P(x_{1:i}, \pi_i = k \mid \theta) \cdot P(x_{i+1:L} \mid \pi_i = k, \theta)$$

use definitions

$$P(x \mid \theta) = \sum_{k=1}^{K} \alpha_i(k) \cdot \beta_i(k)$$

# The posterior probabilities

$$\gamma_i(k) = P(\pi_i = k \mid x, \theta)$$

$$\gamma_i(k) = \frac{\alpha_i(k) \cdot \beta_i(k)}{P(x \mid \theta)} = \frac{\alpha_i(k) \cdot \beta_i(k)}{\sum_{k=1}^{K} \alpha_i(k) \cdot \beta_i(k)}$$

$$\alpha_i(k) = P(x_{1:i}, \pi_i = k \mid \theta)$$
$$\beta_i(k) = P(x_{i+1:L} \mid \pi_i = k, \theta)$$

**Proof:**

$$\gamma_i(k) = P(\pi_i = k \mid x, \theta) = \frac{P(\pi_i = k, x \mid \theta)}{P(x \mid \theta)}$$ conditional probability

$$\gamma_i(k) = \frac{P(\overbrace{\pi_i = k, x_{1:i}}^{a}, \overbrace{x_{i+1:L}}^{b} \mid \theta)}{P(x \mid \theta)}$$ $P(a, b) = P(a) \cdot P(b \mid a)$

$$\gamma_i(k) = \frac{P(\overbrace{\pi_i = k, x_{1:i}}^{a} \mid \theta) \cdot P(\overbrace{x_{i+1:L}}^{b} \mid \overbrace{\pi_i = k, x_{1:i}}^{a}, \theta)}{P(x \mid \theta)}$$ use Markov property

$$\gamma_i(k) = \frac{P(\pi_i = k, x_{1:i} \mid \theta) \cdot P(x_{i+1:L} \mid \pi_i = k, \theta)}{P(x \mid \theta)} = \frac{\alpha_i(k) \cdot \beta_i(k)}{P(x \mid \theta)}$$ definitions

# Probabilities $\xi_i(k, l)$

Another auxiliary variable is called $\xi_i(k, l)$:

$$\xi_i(k, l) = P(\pi_i = k, \pi_{i+1} = l \mid x, \theta) \qquad \text{Definition}$$

This is the probability of having state $k$ at position $i$, and having state $l$ at position $i + 1$, given the observations $x$ and the model $\theta$. It can be expressed in terms of the probabilities introduced earlier:

$$\xi_i(k, l) = \frac{\alpha_i(k) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot \beta_{i+1}(l)}{P(x \mid \theta)}$$

The denominator can also be written as a double sum over all states:

$$\xi_i(k, l) = \frac{\alpha_i(k) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot \beta_{i+1}(l)}{\sum_{k=1}^{K} \sum_{l=1}^{K} \alpha_i(k) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot \beta_{i+1}(l)}$$

$$\xi_i(k,l) = P(\pi_i = k, \pi_{i+1} = l \mid x, \theta)$$ Definition of $\xi$

$$\xi_i(k,l) = \frac{\alpha_i(k) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot \beta_{i+1}(l)}{P(x \mid \theta)}$$ Calculation of $\xi$

**Proof:**

$$\xi_i(k,l) = P(\pi_i = k, \pi_{i+1} = l \mid x, \theta) = \frac{P(x, \pi_i = k, \pi_{i+1} = l \mid \theta)}{P(x \mid \theta)}$$ conditional probability

$$\text{numerator} = P(x, \pi_i = k, \pi_{i+1} = l \mid \theta) = P(x_{1:i}, \pi_i = k, x_{i+1:L}, \pi_{i+1} = l \mid \theta)$$ reordered

$$= P(x_{i+1:L}, \pi_{i+1} = l \mid x_{1:i}, \pi_i = k, \theta) \cdot P(x_{1:i}, \pi_i = k, \theta)$$

$P(a,b) = P(b \mid a) \cdot P(a)$

$$= P(x_{i+1:L}, \pi_{i+1} = l \mid \pi_i = k, \theta) \cdot \underbrace{P(x_{1:i}, \pi_i = k, \theta)}_{\alpha_i(k)}$$ Markov property

$$= P(x_{i+1:L} \mid \pi_{i+1} = l, \pi_i = k, \theta) \cdot \underbrace{P(\pi_{i+1} = l \mid \pi_i = k, \theta)}_{a_{kl}} \cdot \alpha_i(k)$$

$$= P(x_{i+1}, x_{i+2:L} \mid \pi_{i+1} = l, \pi_i = k, \theta) \cdot a_{kl} \cdot \alpha_i(k)$$ Markov property

$$= P(x_{i+2:L} \mid \pi_{i+1} = l, \pi_i = k, \theta) \cdot P(x_{i+1} \mid x_{i+2:L}, \pi_{i+1} = l, \pi_i = k, \theta) \cdot a_{kl} \cdot \alpha_i(k)$$

$$= P(x_{i+2:L} \mid \pi_{i+1} = l, \theta) \cdot P(x_{i+1} \mid \pi_{i+1} = l, \theta) \cdot a_{kl} \cdot \alpha_i(k)$$

$$= \beta_{i+1}(l) \cdot e_l(x_{i+1}) \cdot a_{kl} \cdot \alpha_i(k)$$ q.e.d.

# Probabilities $\xi_i(k, l)$

$$\xi_i(k,l) = P(\pi_i = k, \pi_{i+1} = l \mid x, \theta)$$ Definition

From this definition, we see that suming up over $l$ yields the above defined variable $\gamma_i$ :

$$\gamma_i(k) = P(\pi_i = k \mid x, \theta) = \sum_{l=1}^{K} P(\pi_i = k, \pi_{i+1} = l \mid x, \theta) = \sum_{l=1}^{K} \xi(k,l)$$

by the marginal rule

$$\gamma_i(k) = \sum_{l=1}^{K} \xi(k,l)$$

# 4. Expectation-Maximization (EM) Algorithm

- calculates Maximum likelihood estimators for parameters in models that contain unobserved (latent, hidden) variables.
  - incomplete-data problems (HMM, missing data problems)
  - models with ''artificially'' introduced latent variables (Gaussian Mixture Models: GMM)
- EM emerged from a number of previous, ''intuitive'' algorithms
- generalized as EM by Dempster, Laird, and Rubin in 1977 [(1)]
- Recursive algorithm (E-step; M-step)
- EM is often computationally easier than other methods

[(1)] Dempster, A.P., Laird. N.M., Rubin, D.B.: Maximum-Likelihood from incomplete data via the EM algorithm. J. Royal Statist. Society, 1977

# General approach used in EM

- Observed data: $x$ (visible part of the chain)
- latent (hidden) data: $\pi$ (state path)
- parameters of the model: $\theta = \{a_{kl}, e_k(b), a_{0l}\}$ with $a_{0l} = P(\pi_1 = l)$

$$L(\theta) = P(x \mid \theta) = \frac{P(\pi, x \mid \theta)}{P(\pi \mid x, \theta)}$$ likelihood

$$l(\theta) = \log P(x \mid \theta) = \log P(\pi, x \mid \theta) - \log P(\pi \mid x, \theta)$$ log-likelihood

Expectation-Maximization algorithm:

- calculate the conditional expectation $E_{\pi \mid x,\theta_t}$ of the log-likelihood with respect to $\pi$ given the observation $x$ and some known $\theta_t$
- $\theta_t$ is either the 1st guess of the parameter set, or the parameter set obtained on the precedent iteration step
- this is eqivalent to generating $\pi$ according to the distribution $P(\pi \mid x, \theta_t)$ and average

# Conditional Expectation of the log-likelihood

$$\log P(x \mid \theta) = \log P(\pi, x \mid \theta) - \log P(\pi \mid x, \theta)$$

identity, cond. probability

This equation is valid for any value of $\pi \rightarrow$ whatever the value of $\pi$ on the right-hand side is, the left-hand side is still $\log P(x \mid \theta)$. It follows that the expectation of the right-hand side is also $\log P(x \mid \theta)$. Formally, this can be shown by applying the operator $\sum_\pi P(\pi \mid x, \theta_t) *$ to both sides of the equation. The left side is unchanged because it does not depend on $\pi$ and $\sum_\pi P(\pi \mid x, \theta_t) = 1$ :

Find expectation by applying $\sum_\pi P(\pi \mid x, \theta_t) *$ (with some known model $\theta_t$):

$$\log P(x \mid \theta) = \underbrace{\sum_\pi \log P(\pi, x \mid \theta) \cdot P(\pi \mid x, \theta_t)}_{Q(\theta, \theta_t)} - \underbrace{\sum_\pi \log P(\pi \mid x, \theta) \cdot P(\pi \mid x, \theta_t)}_{H(\theta, \theta_t)}$$

$$\log P(x \mid \theta) = Q(\theta, \theta_t) - H(\theta, \theta_t)$$

# Increasing the log-likelihood

**Task**: Find a new value of $\theta$ so that: $\Delta L = L(\theta) - L(\theta_t) \geq 0$ (likelihood increases)

$$\Delta L = \log P(x \mid \theta) - \log P(x \mid \theta_t)$$

$$\log P(x \mid \theta) = Q(\theta, \theta_t) - H(\theta, \theta_t)$$

$$\log P(x \mid \theta_t) = Q(\theta_t, \theta_t) - H(\theta_t, \theta_t)$$

$$\Delta L = Q(\theta, \theta_t) - H(\theta, \theta_t) - Q(\theta_t, \theta_t) + H(\theta_t, \theta_t)$$

$$\Delta L = Q(\theta, \theta_t) - Q(\theta_t, \theta_t) + \underbrace{H(\theta_t, \theta_t) - H(\theta, \theta_t)}$$

always $\geq 0$ (see below)

$$\Rightarrow \quad \theta_{t+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta_t)$$ increases the log likelihood !

with $$Q(\theta, \theta_t) = \sum_{\pi} \log P(\pi, x \mid \theta) \cdot P(\pi \mid x, \theta_t)$$

# EM - Iteration

1. **Initialise the parameters** (1st guess) : $\theta_0 \rightarrow \theta_t$
2. **Calculate the expectation (Q-function)**:

$$Q(\theta, \theta_t) = \sum_{\pi} \log P(\pi, x \mid \theta) \cdot P(\pi \mid x, \theta_t) \quad \textbf{E-step}$$

$Q(\theta, \theta_t)$ is the conditional expectation $E_{\pi \mid x, \theta_t}$ of the log-likelihood, $P(\pi, x \mid \theta)$, with respect to $\pi$ given the observation $x$ and the known parameters (model) $\theta_t$

3. **Find the parameters which maximize the expectation**:

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta_t) \quad \textbf{M-step}$$

$$\theta_{t+1} \rightarrow \theta_t \quad \text{return to E-step or terminate if } \theta_t \text{ is stationary}$$

Iterate through E-step and M-step until the relative change of $\theta$ gets small. EM improves the likelihood on each iteration (at least, it cannot decrease). EM guarantees that a stationary point of the likelihood is found. This point is not nessecarily a global maximum of $L(\theta)$, but can also be a local maximum or a saddle point.

# The upper bound for $H(\theta, \theta_t)$

Jensen's inequality (for –log, which is a convex function):

$$-\sum_{\pi} a_\pi \log B_\pi \geq -\log\left\{\sum_{\pi} a_\pi B_\pi\right\} \quad \text{if} \quad \sum_{\pi} a_\pi = 1 \quad a_\pi \geq 0$$ (see Appendix)

---

$$H(\theta_t, \theta_t) - H(\theta, \theta_t)$$

$$= \sum_{\pi} \log P(\pi \mid x, \theta_t) \cdot P(\pi \mid x, \theta_t) - \sum_{\pi} \log P(\pi \mid x, \theta) \cdot P(\pi \mid x, \theta_t)$$

$$= \sum_{\pi} -\log \underbrace{\left[\frac{P(\pi \mid x, \theta)}{P(\pi \mid x, \theta_t)}\right]}_{B_\pi} \cdot \underbrace{P(\pi \mid x, \theta_t)}_{a_\pi}$$ use Jensen's inequality

$$\geq -\log\left\{\sum_{\pi} \frac{P(\pi \mid x, \theta)}{P(\pi \mid x, \theta_t)} \cdot P(\pi \mid x, \theta_t)\right\} = -\log 1 = 0$$

$$\Longrightarrow \quad H(\theta, \theta_t) \leq H(\theta_t, \theta_t)$$ $H(\theta_t, \theta_t)$ is an upper bound for $H(\theta, \theta_t)$

# EM - What have we won?

$$L(\theta) = P(x \mid \theta)$$ likelihood, maximize wr.t. $\theta$

$$Q(\theta, \theta_t) = \sum_{\pi} \log P(\pi, x \mid \theta) \cdot P(\pi \mid x, \theta_t)$$ $Q$, maximize wr.t. $\theta$

At a first glance, the task of maximizing the expectation $Q(\theta, \theta_t)$ doesn't seem to be much of an easier task. However, in many situations, solving the second equation can be more convenient than maximizing $L(\theta)$ directly, if the latent variables $\pi$ are chosen in a beneficial manner. In Hidden Markov Models, the latent variables are given by the state path variables.

The EM scheme improves the likelihood on each iteration step (at least, it cannot decrease in any iteration). That makes EM appealing for practical applications.

# 5. EM for Hidden Markov Models

**Aim**: Adjust the model $\theta$ to maximize the likelihood. In order to do that, we have to calculate the conditional expectation (Q-function) first (**E-step**)

$$Q(\theta, \theta_t) = \sum_{\pi} \log P(\pi, x \mid \theta) \cdot P(\pi \mid x, \theta_t)$$

Probability of the "complete data" (= observations $x$ plus state path variables $\pi$) given the model (the parameters) $\theta$:

$$P(x, \pi \mid \theta) = a_{0\pi_1} \cdot e_{\pi_1}(x_1) \cdot a_{\pi_1\pi_2} \cdot e_{\pi_2}(x_2) \cdot a_{\pi_2\pi_3} \cdot \ldots$$

$$P(x, \pi \mid \theta) = a_{0\pi_1} \cdot \prod_{i=1}^{L} e_{\pi_i}(x_i) \cdot a_{\pi_i\pi_{i+1}}$$

Using the definitions of the variables, $P(x, \pi \mid \theta)$ can be written:

$$P(x, \pi \mid \theta) = P(\pi_1) \cdot \prod_{i=1}^{L} P(x_i \mid \pi_i) \cdot P(\pi_{i+1} \mid \pi_i)$$

# EM for HMM, E-step

$$P(x, \pi \mid \theta) = P(\pi_1) \cdot \prod_{i=1}^{L} P(x_i \mid \pi_i) \cdot P(\pi_{i+1} \mid \pi_i)$$ reorder product, log

$$\log P(x, \pi \mid \theta) = \log P(\pi_1) + \sum_{i=1}^{L} \log P(x_i \mid \pi_i) + \sum_{i=1}^{L} \log P(\pi_{i+1} \mid \pi_i)$$

recall that $$Q(\theta, \theta_t) = \sum_{\pi} \log P(\pi, x \mid \theta) \cdot P(\pi \mid x, \theta_t)$$

$Q$ separates into 3 parts:

$$Q(\theta, \theta_t) = \sum_{\pi} \left\{ \log P(\pi_1) + \sum_{i=1}^{L} \log P(x_i \mid \pi_i) + \sum_{i=1}^{L} \log P(\pi_{i+1} \mid \pi_i) \right\} \cdot P(\pi \mid x, \theta_t)$$

$$Q(\theta, \theta_t) = \sum_{\pi} \log P(\pi_1) \cdot P(\pi \mid x, \theta_t) + \sum_{\pi} \sum_{i=1}^{L} \log P(x_i \mid \pi_i) \cdot P(\pi \mid x, \theta_t)$$

$$+ \sum_{\pi} \sum_{i=1}^{L} \log P(\pi_{i+1} \mid \pi_i) \cdot P(\pi \mid x, \theta_t) = Q_A + Q_B + Q_C$$

# EM for HMM, E-step

$$Q(\theta, \theta_t) = Q_A + Q_B + Q_C$$

$$Q_A(\theta, \theta_t) = \sum_{\pi} \log P(\pi_1) \cdot P(\pi \mid x, \theta_t)$$

$$Q_B(\theta, \theta_t) = \sum_{\pi} \sum_{i=1}^{L} \log P(x_i \mid \pi_i) \cdot P(\pi \mid x, \theta_t)$$

$$Q_C(\theta, \theta_t) = \sum_{\pi} \sum_{i=1}^{L} \log P(\pi_{i+1} \mid \pi_i) \cdot P(\pi \mid x, \theta_t)$$

The model includes the parameter set: $\theta = \{a_{kl}, e_k(x_i), a_{0l}\}$

- $Q_A$ depends only on $a_{0l} = P(\pi_1 = l)$ - initial state
- $Q_B$ depends only on $e_k(x_i) = P(x_i \mid \pi_i = k)$ - emission probabilities
- $Q_C$ depends only on $a_{kl} = P(\pi_{i+1} = l \mid \pi_i = k)$ - transition probabilities
- That means that the 3 parts can be maximized separately →
- (Note that $P(\pi \mid x, \theta_t)$ only depends on the known model $\theta_t$).

# EM for HMM, M-step

**1. Maximization of $Q_A$**

$$Q_A(\theta, \theta_t) = \sum_{\pi} \log P(\pi_1) \cdot P(\pi \mid x, \theta_t)$$

We have to keep in mind that $\pi$ is a vector: $\pi = \{\pi_1, \pi_2, \ldots, \pi_L\}$

$$Q_A(\theta, \theta_t) = \sum_{\pi_1} \sum_{\pi_2} \ldots \sum_{\pi_L} \log P(\pi_1) \cdot P(\pi_{1:L} \mid x, \theta_t)$$

$$Q_A(\theta, \theta_t) = \sum_{\pi_1} \log P(\pi_1) \cdot P(\pi_1 \mid x, \theta_t)$$

all other sums add up to 1, marginal rule

$$Q_A(\theta, \theta_t) = \sum_{k} \log P(\pi_1 = k) \cdot P(\pi_1 = k \mid x, \theta_t)$$

$k$ runs over all possible states

$$Q_A(\theta, \theta_t) = \sum_{k} \log a_{0k} \cdot \gamma_1(k)$$

posterior probability, with known $\theta$:

$$\gamma_i(k) = P(\pi_i = k \mid x, \theta_t)$$

# EM for HMM, M-step

**Maximization of $Q_A$, continued**

$Q_A$ has to be maximized w.r.t. $a_{0k}$, with the constraint $\sum_k a_{0k} = 1$

Method of Lagrange multiplier (see Appendix):

$$L(a_{0k}, \lambda) = \sum_k \log a_{0k} \cdot \gamma_1(k) + \lambda \cdot \left[ \sum_k a_{0k} - 1 \right]$$

$$\frac{\delta L}{\delta a_{0l}} = \frac{1}{a_{0l}} \cdot \gamma_1(l) + \lambda = 0 \qquad a_{0l} = -\frac{\gamma_1(l)}{\lambda}$$

$$\frac{\delta L}{\delta \lambda} = \sum_k a_{0k} - 1 = 0 \qquad 1 = \sum_k a_{0k} = -\sum_k \frac{\gamma_1(k)}{\lambda}$$

$$\Rightarrow \quad \lambda = -\underbrace{\sum_k \gamma_1(k)}_{1} = -1 \quad \Rightarrow \quad \boxed{a_{0l} = \gamma_1(l)}$$

# EM for HMM, M-step

**2. Maximization of $Q_B$**

$$Q_B(\theta, \theta_t) = \sum_{\pi} \sum_{i=1}^{L} \log P(x_i \mid \pi_i) \cdot P(\pi \mid x, \theta_t)$$

We have to keep in mind that $\pi$ is a vector: $\pi = \{\pi_1, \pi_2, \ldots, \pi_L\}$

$$Q_B(\theta, \theta_t) = \sum_{\pi_1} \sum_{\pi_2} \ldots \sum_{\pi_L} \sum_{i=1}^{L} \log P(x_i \mid \pi_i) \cdot P(\pi_{1:L} \mid x, \theta_t)$$

$$Q_B(\theta, \theta_t) = \sum_{\pi_i} \sum_{i=1}^{L} \log P(x_i \mid \pi_i) \cdot P(\pi_i \mid x, \theta_t)$$

all other sums add up to 1, marginal rule

$$Q_B(\theta, \theta_t) = \sum_{k=1}^{K} \sum_{i=1}^{L} \log P(x_i \mid \pi_i = k) \cdot P(\pi_i = k \mid x, \theta_t)$$

$$Q_B(\theta, \theta_t) = \sum_{k=1}^{K} \sum_{i=1}^{L} \log P(x_i \mid \pi_i = k) \cdot \gamma_i(k)$$

posterior probability:

$$\gamma_i(k) = P(\pi_i = k \mid x, \theta_t)$$

# EM for HMM, M-step

**Maximization of $Q_B$ , continued**

$$Q_B(\theta, \theta_t) = \sum_{k=1}^{K} \sum_{i=1}^{L} \log P(x_i \mid \pi_i = k) \cdot \gamma_i(k)$$

For a detailed calculation of $Q_B$, it is more intructive to specify concrete arithmetic expressions for the emission probabilities $P(x_i \mid \pi_i)$. Here, two exemplary cases will be presented:

1) Gaussian emission probabilities, which means that state $k$ emits normally distributed observations $x_i$ with mean $\mu_k$ and standard deviation $\sigma_k$. A HMM with such emission probabilities is called Continuous Density Hidden Markov Model (CDHMM).
2) Multinomially distributed emission probabilities, which means that state $k$ emits one instance of a discrete random variable. The probability of emitting a particular symbol depends on the state $k$.

# EM for HMM, M-step

**Maximization of $Q_B$ , continued, Gaussian emissions**

$$Q_B(\theta, \theta_t) = \sum_{k=1}^{K} \sum_{i=1}^{L} \log P(x_i \mid \pi_i = k) \cdot \gamma_i(k)$$

A HMM with Gaussian emission probabilities means that state $k$ emits normally distributed observations $x_i$ with mean $\mu_k$ and standard deviation $\sigma_k$:

$$P(x_i \mid \pi_i = k) = \frac{1}{\sqrt{2\pi} \cdot \sigma_k} \exp\left[ -\frac{(x_i - \mu_k)^2}{2\sigma_k^2} \right]$$

so that we have

$$\log P(x_i \mid \pi_i = k) = -\frac{(x_i - \mu_k)^2}{2\sigma_k^2} - \log\left(\sqrt{2\pi}\sigma_k\right)$$

# EM for HMM, M-step

**Maximization of $Q_B$ , continued , Gaussian emissions**

$$Q_B(\theta, \theta_t) = -\sum_{k=1}^{K}\sum_{i=1}^{L}\left[\frac{(x_i - \mu_k)^2}{2\sigma_k^2} + \log\left(\sqrt{2\pi}\sigma_k\right)\right] \cdot \gamma_i(k)$$

Now, maximizing $Q_B$ with respect to the emission probabilities means that we have to maxmize with respect to the parameters $\mu_k$ and $\sigma_k$.

$$\frac{\delta Q_B}{\delta \mu_l} = \sum_{i=1}^{L} \frac{(x_i - \mu_l)}{\sigma_l^2} \cdot \gamma_i(l) = 0$$

$$\sum_{i=1}^{L} (x_i - \mu_l) \cdot \gamma_i(l) = 0$$

$$\sum_{i=1}^{L} x_i \cdot \gamma_i(l) = \mu_l \cdot \sum_{i=1}^{L} \gamma_i(l) \quad \Rightarrow \quad \mu_l = \frac{\sum_i x_i \cdot \gamma_i(l)}{\sum_i \gamma_i(l)}$$

This is similar to the common estimation of a Gaussian mean, with the distinction that the observations $x_i$ are weighted by the posterior probabilities that the chain is in state $l$ at position $i$.

# EM for HMM, M-step

**Maximization of $Q_B$ , continued , Gaussian emissions**

$$Q_B(\theta, \theta_t) = -\sum_{k=1}^{K}\sum_{i=1}^{L}\left[\frac{(x_i-\mu_k)^2}{2\sigma_k^2} + \log\left(\sqrt{2\pi}\sigma_k\right)\right]\cdot\gamma_i(k)$$

$$\frac{\delta Q_B}{\delta \sigma_l} = \sum_{i=1}^{L}\left[\frac{(x_i-\mu_l)^2}{\sigma_l^3} - \frac{1}{\sigma_l}\right]\cdot\gamma_i(l) = 0$$

$$\sum_{i=1}^{L}\left[(x_i-\mu_l)^2 - \sigma_l^2\right]\cdot\gamma_i(l) = 0$$

$$\sum_{i=1}^{L}(x_i-\mu_l)^2\cdot\gamma_i(l) = \sigma_l^2\cdot\sum_{i=1}^{L}\gamma_i(l) \quad \Rightarrow \quad \sigma_l^2 = \frac{\sum_i (x_i-\mu_l)^2\cdot\gamma_i(l)}{\sum_i \gamma_i(l)}$$

This is similar to the common estimation of a Gaussian variance, with the distinction that the squared residuals are weighted by the posterior probabilities that the chain is in state $l$ at position $i$.

# EM for HMM, M-step

**Maximization of $Q_B$ , continued, Multinomial emissions**

$$Q_B(\theta, \theta_t) = \sum_{k=1}^{K} \sum_{i=1}^{L} \log P(x_i \mid \pi_i = k) \cdot \gamma_i(k)$$

A HMM with multinomially distributed emission probabilities means that state $k$ emits an instance of a discrete random variable. The probability of emitting a particular symbol depends on the state $k$.

In general, the multinomial probability mass function reads:

$$P(X) = \frac{N!}{x_1! \cdot x_2! \cdot \ldots \cdot x_M!} \cdot p_1^{x_1} \cdot p_2^{x_2} \cdot \ldots \cdot p_M^{x_M}$$

Here, it is assumed that $N$ trials have been made, and each trial can lead to one of $M$ different outcomes, having probabilities $p_1, p_2, \ldots, p_M$. The following constraints apply:

$$\sum_n x_n = N \quad \text{and} \quad \sum_j p_j = 1$$

# EM for HMM, M-step

**Maximization of $Q_B$ , continued, Multinomial emissions**

When it comes to emissions from state $k$ of a Markov chain, only one symbol is emitted from that state, i.e. only one trial is made ($N = 1$), so that we have:

$$P(X) = p_1^{x_1} \cdot p_2^{x_2} \cdot \ldots \cdot p_M^{x_M} \quad \text{with} \quad \sum_n x_n = 1 \quad \text{and} \quad \sum_j p_j = 1$$

For our calculations, we have to introduce two additional indices:

- we need to summarize over the positions in the chain → additional index $i$ required
- the parameters $\{p_j\}$ are different for each state → additional index $k$ required

This leads to the conditional probability:

$$P(x_i \mid \pi_i = k) = p_{k1}^{x_{i1}} \cdot p_{k2}^{x_{i2}} \cdot \ldots \cdot p_{kM}^{x_{iM}} = \prod_{j=1}^{M} p_{kj}^{x_{ij}} \qquad \sum_j x_{ij} = 1 \;\forall i \qquad \sum_j p_{kj} = 1 \;\forall k$$

# EM for HMM, M-step

**Maximization of $Q_B$ , continued, Multinomial emissions**

$$P(x_i \mid \pi_i = k) = p_{k1}^{x_{i1}} \cdot p_{k2}^{x_{i2}} \cdot \ldots \cdot p_{kM}^{x_{iM}} = \prod_{j=1}^{M} p_{kj}^{x_{ij}}$$

$p_{kj}$ is the probability that the symbol $j$ is released by state $k$, and $x_{ij}$ indicates if symbol $j$ was observed at position $i$. Note that in this notation, $x_{ij}$ is one for exactly one symbol $j$ and zero for all others, which ensures that $\sum_j x_{ij} = 1$. Using these expression, we can specify $Q_B$:

$$Q_B(\theta, \theta_t) = \sum_{k=1}^{K} \sum_{i=1}^{L} \log P(x_i \mid \pi_i = k) \cdot \gamma_i(k)$$

$$Q_B(\theta, \theta_t) = \sum_{k=1}^{K} \sum_{i=1}^{L} \log \left[ \prod_{j=1}^{M} p_{kj}^{x_{ij}} \right] \cdot \gamma_i(k) = \sum_{k=1}^{K} \sum_{i=1}^{L} \sum_{j=1}^{M} x_{ij} \cdot \log p_{kj} \cdot \gamma_i(k)$$

with the remaining constraint $\sum_j p_{kj} = 1 \; \forall k$

# EM for HMM, M-step

**Maximization of $Q_B$ , continued, Multinomial emissions**

$$Q_B(\theta, \theta_t) = \sum_{k=1}^{K} \sum_{i=1}^{L} \sum_{j=1}^{M} x_{ij} \cdot \log p_{kj} \cdot \gamma_i(k) \qquad \sum_j p_{kj} = 1 \; \forall k$$

In order to optimize for $p_{kj}$, we establish the Lagrange function:

$$L = \sum_{k=1}^{K} \sum_{i=1}^{L} \sum_{j=1}^{M} x_{ij} \cdot \log p_{kj} \cdot \gamma_i(k) + \sum_k \lambda_k \left[ \sum_j p_{kj} - 1 \right]$$

$$\frac{\delta L}{\delta p_{kj}} = \sum_{i=1}^{L} x_{ij} \cdot \frac{1}{p_{kj}} \cdot \gamma_i(k) + \lambda_k = 0$$

$$\frac{\delta L}{\delta \lambda_k} = \sum_{j=1}^{M} p_{kj} - 1 = 0$$

# EM for HMM, M-step

**Maximization of $Q_B$ , continued, Multinomial emissions**

$$\frac{\delta L}{\delta p_{kj}} = \sum_{i=1}^{L} x_{ij} \cdot \frac{1}{p_{kj}} \cdot \gamma_i(k) + \lambda_k = 0 \quad (i) \qquad \frac{\delta L}{\delta \lambda_k} = \sum_{j=1}^{M} p_{kj} - 1 = 0 \quad (ii)$$

$$p_{kj} = -\frac{\sum_{i=1}^{L} x_{ij} \cdot \gamma_i(k)}{\lambda_k} \quad (i)'$$

$$1 = \sum_{j=1}^{M} p_{kj} = -\sum_{j} \sum_{i} \frac{x_{ij} \cdot \gamma_i(k)}{\lambda_k} \quad (ii)'$$

$$\lambda_k = -\sum_{i} \underbrace{\sum_{j} x_{ij}}_{=1} \cdot \gamma_i(k) = -\sum_{i} \gamma_i^k \quad (ii)''$$

$$\Rightarrow \boxed{p_{kj} = \frac{\sum_i x_{ij} \cdot \gamma_i(k)}{\sum_i \gamma_i(k)}}$$

# EM for HMM, M-step

**3. Maximization of $Q_C$**

$$Q_C(\theta, \theta_t) = \sum_{\pi} \sum_{i=1}^{L} \log P(\pi_{i+1} \mid \pi_i) \cdot P(\pi \mid x, \theta_t)$$

We have to keep in mind that $\pi$ is a vector: $\pi = \{\pi_1, \pi_2, \ldots, \pi_L\}$

$$Q_C(\theta, \theta_t) = \sum_{\pi_1} \sum_{\pi_2} \ldots \sum_{\pi_L} \sum_{i=1}^{L} \log P(\pi_{i+1} \mid \pi_i) \cdot P(\pi_{1:L} \mid x, \theta_t)$$

$$Q_C(\theta, \theta_t) = \sum_{\pi_i} \sum_{\pi_{i+1}} \sum_{i=1}^{L} \log P(\pi_{i+1} \mid \pi_i) \cdot P(\pi_i, \pi_{i+1} \mid x, \theta_t)$$ marginal rule

$$Q_C(\theta, \theta_t) = \sum_{k} \sum_{l} \sum_{i=1}^{L} \log P(\pi_{i+1} = l \mid \pi_i = k) \cdot P(\pi_i = k, \pi_{i+1} = l \mid x, \theta_t)$$

$$Q_C(\theta, \theta_t) = \sum_{k} \sum_{l} \sum_{i=1}^{L} \log a_{kl} \cdot \xi_i(k, l)$$

Definition $\xi(k, l)$

$$\xi_i(k, l) = P(\pi_i = k, \pi_{i+1} = l \mid x, \theta_t)$$

the $\xi_i$ can be trated as constants because they are taken for $\theta_t$

# EM for HMM, M-step

**Maximization of $Q_C$**

$$Q_C(\theta, \theta_t) = \sum_{k=1}^{K} \sum_{l=1}^{K} \sum_{i=1}^{L} \log a_{kl} \cdot \xi_i(k,l)$$

$K$ - number of states
$L$ – length of chain

$Q_C$ has to be maximized with respect to $a_{kl}$. The constraints are $\sum_l a_{kl} = 1$, which is valid for every $k$. That means that we actually have $K$ constraints. The Lagrange function is therefore:

$$L = \sum_k \sum_l \sum_{i=1}^{L} \log a_{kl} \cdot \xi_i(k,l) + \underbrace{\lambda_1 \left[ \sum_l a_{1l} - 1 \right] + \lambda_2 \left[ \sum_l a_{2l} - 1 \right] + \ldots}_{K \text{ summands}}$$

$$L = \sum_k \sum_l \sum_{i=1}^{L} \log a_{kl} \cdot \xi_i(k,l) + \sum_{k=1}^{K} \lambda_k \left[ \sum_l a_{kl} - 1 \right]$$

Lagrange function

# EM for HMM, M-step

**Maximization of $Q_C$, continued**

$$L = \sum_k \sum_l \sum_{i=1}^{L} \log a_{kl} \cdot \xi_i(k,l) + \sum_{k=1}^{K} \lambda_k \left[ \sum_l a_{kl} - 1 \right]$$

Lagrange function

$$\frac{\delta L}{\delta a_{mn}} = \sum_i \frac{1}{a_{mn}} \cdot \xi_i(m,n) + \lambda_m = 0 \quad \Rightarrow \quad a_{mn} = -\frac{\sum_i \xi_i(m,n)}{\lambda_m}$$

$$\frac{\delta L}{\delta \lambda_m} = \sum_l a_{ml} - 1 = 0 \quad m = 1, 2, \ldots, K$$

$$1 = \sum_l a_{ml} = -\frac{1}{\lambda_m} \sum_l \sum_i \xi_i(m,l) = -\frac{1}{\lambda_m} \sum_i \overbrace{\sum_l \xi_i(m,l)}^{\gamma_i(m)}$$

$$1 = -\frac{1}{\lambda_m} \sum_i \gamma_i(m) \quad \Rightarrow \quad \lambda_m = -\sum_i \gamma_i(m)$$

$$\Rightarrow \quad \boxed{a_{mn} = \frac{\sum_i \xi_i(m,n)}{\sum_i \gamma_i(m)}}$$

# EM for HMM, M-step

**Maximization of $Q_C$, continued**

$$a_{mn} = \frac{\sum_i \xi_i(m,n)}{\sum_i \gamma_i(m)}$$

Using the definitions of $\xi_i$ and $\gamma_i$,

$$\xi_i(m,n) = P(\pi_i = m, \pi_{i+1} = n \mid x, \theta)$$

$$\gamma_i(m) = P(\pi_i = m \mid x, \theta)$$

we can write:

$$a_{mn} = \frac{\sum_i P(\pi_i = m, \pi_{i+1} = n \mid x, \theta)}{\sum_i P(\pi_i = m \mid x, \theta)}$$

The estimated transition probability $a_{mn}$ is the ratio of the expected number of $m \to n$ transitions and the expected number of $m$- states in the chain.

# EM for HMM
## - Baum-Welch algorithm -

The Expectation-Maximization algorithm adapted to a Hidden Markov Model is called Baum-Welch algorithm.
Here comes a summary for the estimation of the model parameters $\theta$, when Gaussian emission probabilities are assumed:

$$a_{0l} = \gamma_1(l)$$ initial state probabilities

$$\mu_l = \frac{\sum_i x_i \cdot \gamma_i(l)}{\sum_i \gamma_i(l)}$$ mean of the Gaussian for state $l$

$$\sigma_l^2 = \frac{\sum_i (x_i - \mu_l)^2 \cdot \gamma_i(l)}{\sum_i \gamma_i(l)}$$ variance of the Gaussian for state $l$

$$a_{kl} = \frac{\sum_i \xi_i(k,l)}{\sum_i \gamma_i(k)}$$ transition probabilities

The $\xi_i$ and $\gamma_i$ are calculated based on the estimations for the preceding iteration step.

# EM for HMM
## - Baum-Welch algorithm -

Here comes a summary for the estimation of the model parameters $\theta$, when multinomial emission probabilities are assumed:

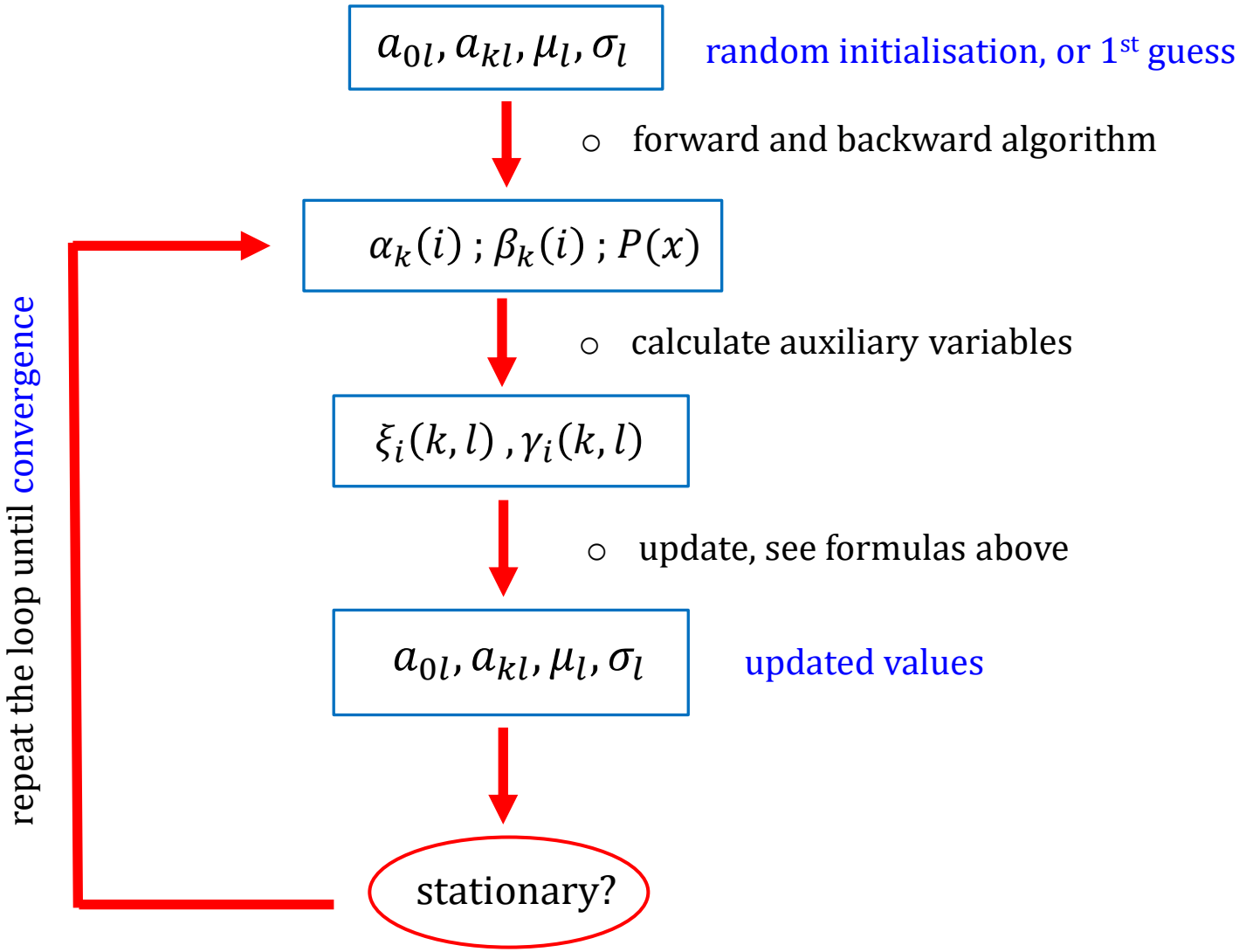
$$a_{0l} = \gamma_1(l)$$ initial state probabilities

$$p_{kj} = \frac{\sum_i x_{ij} \cdot \gamma_i(k)}{\sum_i \gamma_i(k)}$$ emission probabilities for state k

$$a_{kl} = \frac{\sum_i \xi_i(k,l)}{\sum_i \gamma_i(k)}$$ transition probabilities

The $\xi_i$ and $\gamma_i$ are calculated based on the estimations for the preceding iteration step.

# Baum-Welch iteration (CDHMM)



# 6. Maximum A Posteriori (MAP) estimation

Up to now, we have maximised the (log-) likelihood: $L(\theta) = P(x \mid \theta)$

MAP attempts to find the maximum of the posterior probability:

$$P(\theta|x) = \frac{P(x|\theta) \cdot P(\theta)}{P(x)}$$

The denominator $P(x)$ is not dependent on $\theta$ and can therefore be ignored:

$$\begin{aligned} L(\theta) &= \log P(x|\theta) + \log P(\theta) && \text{introduce latent variables. } P(x) = P(\pi, x)/P(\pi \mid x) \\ &= \log P(\pi, x|\theta) - \log P(\pi|x, \theta) + \log P(\theta) && \text{conditional expectation ...} \\ &= Q(\theta, \theta_t) - H(\theta, \theta_t) + \log P(\theta) && \text{additional term: } \log P(\theta) \end{aligned}$$

where the definitions of $Q$ and $H$ were used:

$$H(\theta, \theta_t) = \sum_{\pi} \log P(\pi \mid x, \theta) \cdot P(\pi \mid x, \theta_t)$$

$$Q(\theta, \theta_t) = \sum_{\pi} \log P(\pi, x \mid \theta) \cdot P(\pi \mid x, \theta_t)$$

# MAP - Iteration

1. **Initialise the parameters** (1st guess) : $\theta_0 \rightarrow \theta_t$
2. **Calculate the expectation (Q-function)**:

$$Q(\theta, \theta_t) = \sum_{\pi} \log P(\pi, x \mid \theta) \cdot P(\pi \mid x, \theta_t) \qquad \textbf{E-step}$$

$Q(\theta, \theta_t)$ is the conditional expectation $E_{\pi \mid x, \theta_t}$ of the log-likelihood, $P(\pi, x \mid \theta)$, with respect to $\pi$ given the observation $x$ and the known parameters (model) $\theta_t$

3. **Find the parameters which maximize the expectation**:

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}} \left[ Q(\theta, \theta_t) + \log P(\theta) \right] \qquad \textbf{M-step,} \text{ note the } \log P(\theta)$$

$$\theta_{t+1} \rightarrow \theta_t \qquad \text{return to E-step or terminate if } \theta_t \text{ is stationary}$$

Iterate through E-step and M-step until the relative change of $\theta$ gets small. EM improves the likelihood on each iteration (at least, it cannot decrease). Suitable priors must be chosen to make MAP not harder than MLE.

# Appendix

## Hidden Markov Models

## Expectation-Maximisation (EM) Estimation

## Maximum-A-Posteriori (MAP) Estimation

Uwe Menzel, 2008

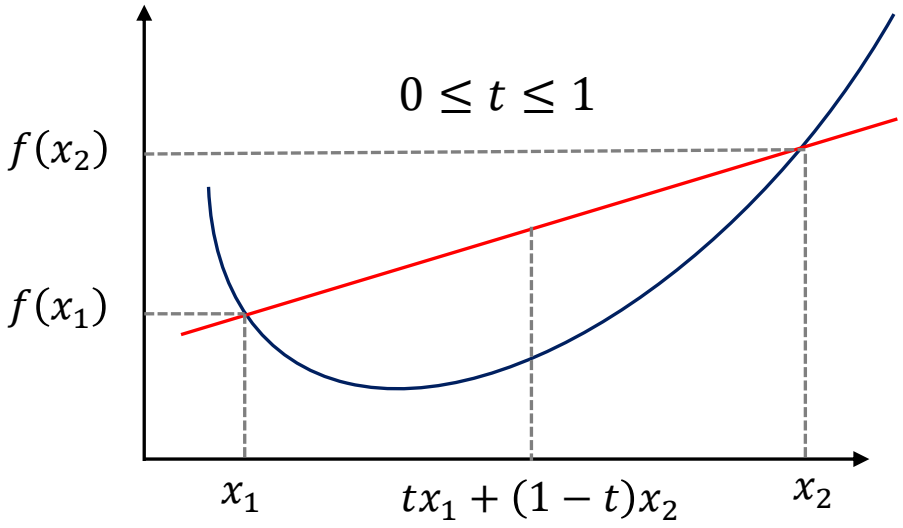
uwe.menzel@matstat.org

www.matstat.org

# The Jensen inequality

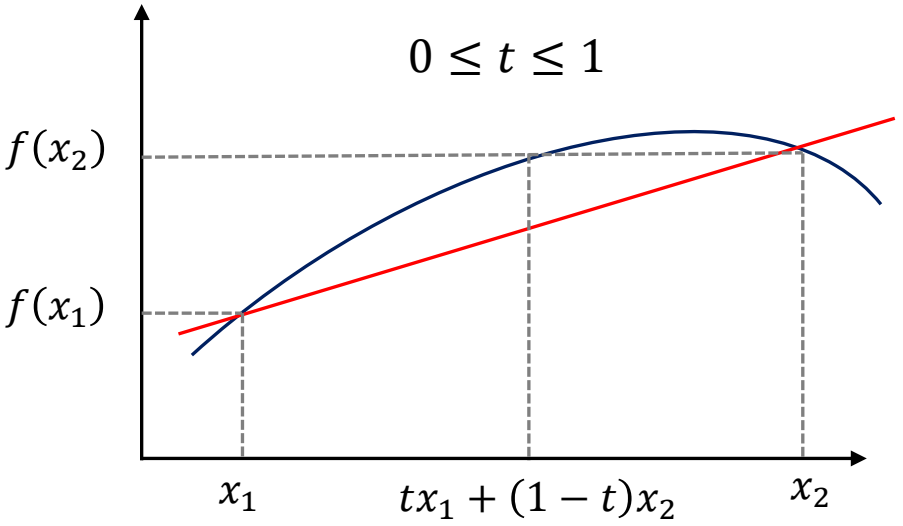
**Convex functions:**

$$\sum_i a_i f(x_i) \geq f\left(\sum_i a_i x_i\right)$$
$$\sum_i a_i = 1 \quad a_i > 0$$




**Concave functions:**

$$f\left(\sum_i a_i x_i\right) \geq \sum_i a_i f(x_i)$$
$$\sum_i a_i = 1 \quad a_i > 0$$



# The Jensen inequality

**Convex functions:**

$$\sum_i a_i f(x_i) \geq f\left(\sum_i a_i x_i\right)$$
$$\sum_i a_i = 1 \quad a_i > 0$$

**Concave functions:**

$$f\left(\sum_i a_i x_i\right) \geq \sum_i a_i f(x_i)$$
$$\sum_i a_i = 1 \quad a_i > 0$$

**Convex functions:**

$$\int \Phi\left[f(x)\right] p(x)dx \geq \Phi\left[\int f(x)p(x)dx\right]$$
$$\int p(x) = 1 \quad p(x) > 0$$

**Concave functions:**

$$\Phi\left[\int f(x)p(x)dx\right] \geq \int \Phi\left[f(x)\right] p(x)dx$$
$$\int p(x) = 1 \quad p(x) > 0$$

This is not more than a hint. An excellent presentation of Jensen's (and other) inequalities is by Dragos Hrimiuc (University of Alberta) can be found here: https://www.math.ualberta.ca/pi/issue4/ ("Pi in the sky" December 2001 issue).

# Extrema with constraints
## - Method of Lagrange multipliers -

**Aim**: Find the $\boldsymbol{x}$ that maximizes $f(\boldsymbol{x})$ with the contraint that $g(\boldsymbol{x}) = 0$
($\boldsymbol{x}$ might be a vector).

Construct **Lagrange function**: $\mathcal{L}(x, \lambda) = f(x) - \lambda \cdot g(x)$
The parameter $\lambda$ is called Lagrange multiplier.

Solve $\frac{\partial \mathcal{L}}{\partial x} = 0$ and $\frac{\partial \mathcal{L}}{\partial \lambda} = 0$ (a system of equations). The solution is an extremum of the function $f$ under the constraint $g$.

**Example:** Find extrema of $f(x, y) = x + y$ ; constraint $x^2 + y^2 = 1$

$$g(x, y) = x^2 + y^2 - 1 \qquad L(x, y, \lambda) = x + y + \lambda \cdot \left(x^2 + y^2 - 1\right)$$

$$\left.\begin{aligned} \frac{\delta L}{\delta x} &= 1 + 2\lambda x = 0 \\ \frac{\delta L}{\delta y} &= 1 + 2\lambda y = 0 \\ \frac{\delta L}{\delta \lambda} &= x^2 + y^2 - 1 = 0 \end{aligned}\right\} \quad x = \pm\frac{1}{\sqrt{2}} \quad y = \pm\frac{1}{\sqrt{2}}$$