Human Centered Multimedia
Institute of Computer Science

Universität Augsburg
Augsburg University

Multimodale Analyse

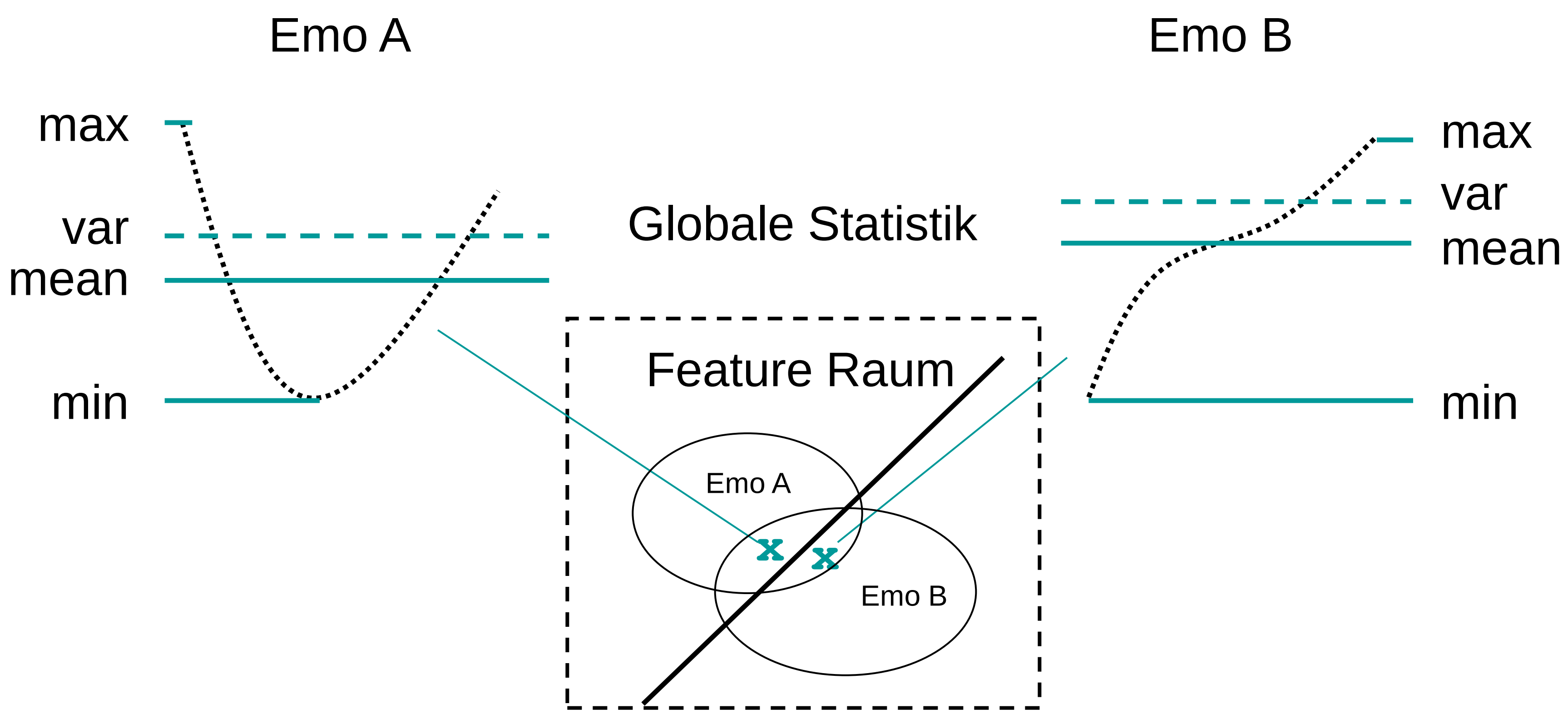
Hidden Markov Modelle

Elisabeth André
Chi Tai Dang
Stephan Hammer

HCM

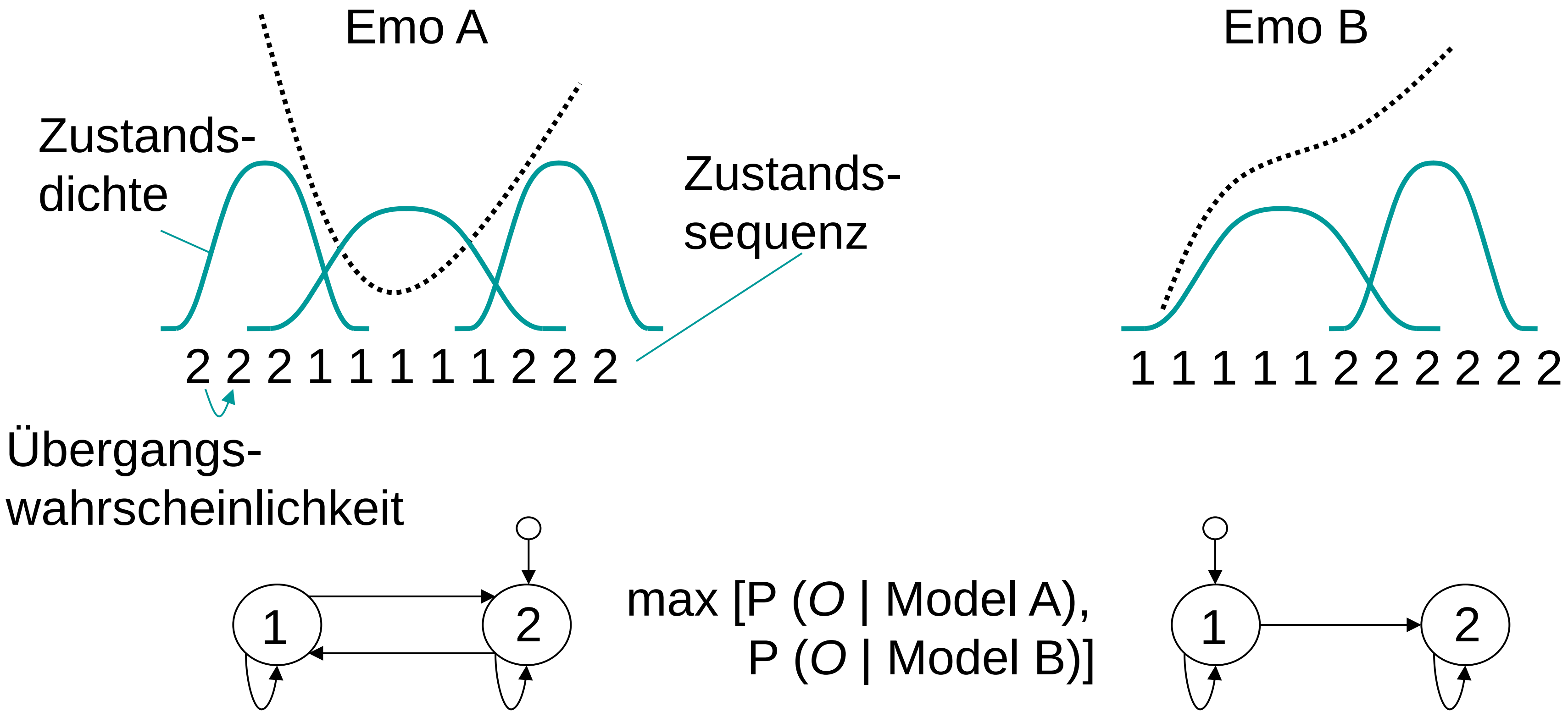
Human Centered Multimedia
Institute of Computer Science
Augsburg University
Universitätsstraße. 6a
86159 Augsburg, Germany

Statisches Modelieren



- kNN, GMM, Neural Networks, SVM, …

Dynamisches Modellieren



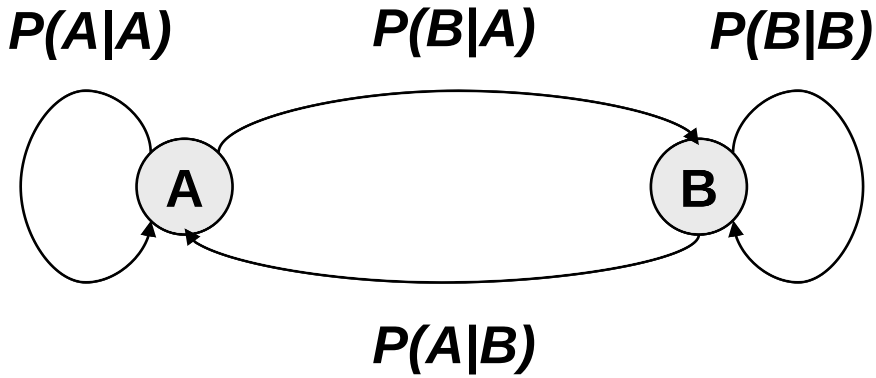
- Hidden Markov Models (HMM)

Charakterisierung von Signalen durch Modelle

- **Deterministische Modelle**
  - Bestimmung der Werte für Parameter des Signalmodells, z.B. Amplitude, Frequenz
- **Statistische Modelle**
  - Charakterisierung des Signals durch einen parametrisierten Zufallsprozess und Bestimmung der Parameter dieses Zufallprozesses

Einfache Markov-Modelle Definition

- Die *Zustände* des Markov Modells sind $S = \{S_1, S_2, \ldots, S_N\}$
- Die *Übergangswahrscheinlichkeiten* sind $P(q_t = S_j \mid q_{t-1} = S_i)$
- Ein Markov Modell kann dann als ein endlicher Automat gezeichnet werden, dessen Knoten die Zustände und an dessen Kanten die Übergangswahrscheinlichkeiten sind.



- Man nennt so etwas einen *stochastischen* Automaten.

Universität Augsburg University

- Die *Zustände* des Markov Modells sind $S = \{S_1, S_2, \ldots, S_N\}$
- Die *Übergangswahrscheinlichkeiten* sind $P(q_t = S_j \mid q_{t-1} = S_i)$

- Für Markov Modelle gilt immer die **Markov Annahme**, d.h. die Übergangswahrscheinlichkeiten hängen nur von dem aktuellen Zustand und nicht von den Vorherigen ab,

$$P(q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, \ldots) = P(q_t = S_j \mid q_{t-1} = S_i) = a_{ij}$$

wobei $a_{ij} \geq 0 \ \ \forall i, j$ und $\sum_{i=1}^{N} aji = 1 \ \ \forall j$

- Betrachte ein einfaches Markov-Modell für das Wetter, das sich aus folgenden drei Zuständen zusammensetzt:
  - S_1 : Niederschlag
  - S_2 : Bewölkung
  - S_3 : Sonnenschein



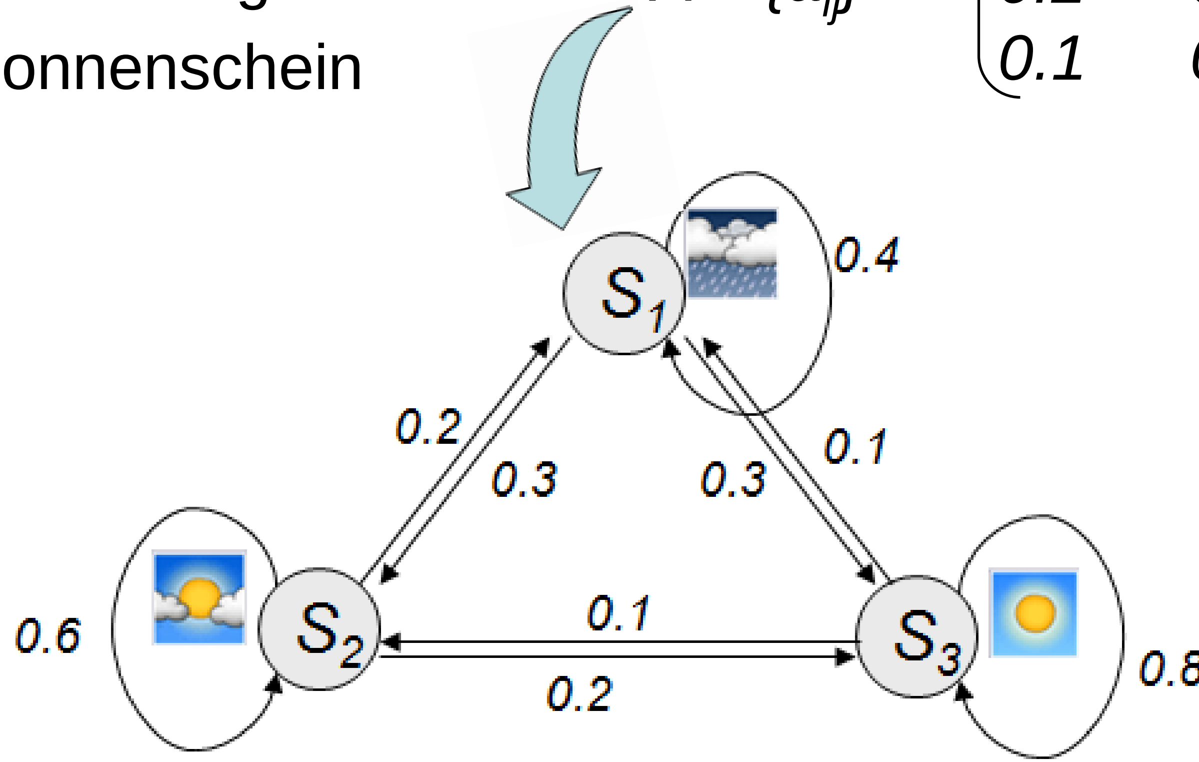
- Die Übergangswahrscheinlichkeiten sind in der Matrix *A:*

$$A = \{a_{ij}\} = \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$$

Einfache Markov-Modelle
Beispiel „Wettervorhersage“

- S_1 : Niederschlag
- S_2 : Bewölkung
- S_3 : Sonnenschein

$$A = \{a_{ij}\} = \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$$



- Wir nehmen an, dass an Tag $t = 1$ das Wetter sonnig ist.
- Frage: Wie hoch ist die Wahrscheinlichkeit, dass das Wetter für die nächsten *4* Tage aus der Zustandsfolge *„Sonne-Sonne-Regen-Bewölkung“* bestehen wird?
- Berechnung der Wahrscheinlichkeit für die Sequenz von Beobachtungen O unter Berücksichtigung des Modells λ

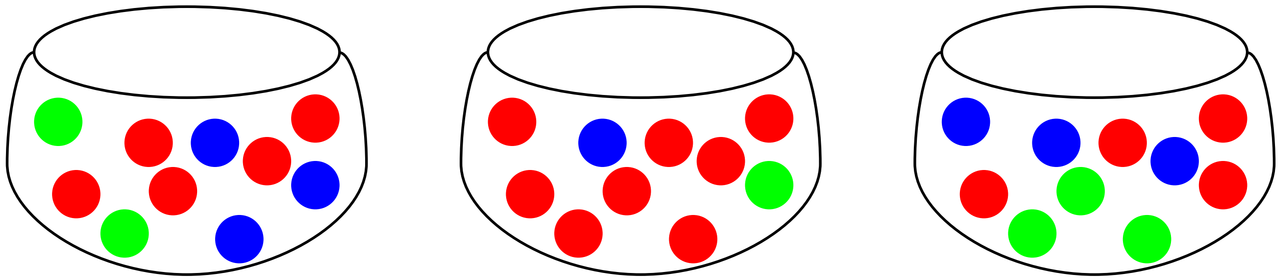
$$
\begin{aligned}
P(O|\lambda) &= P(S_3,S_3,S_3,S_1,S_2|\lambda) \\
&= P(S_3) \cdot P(S_3|S_3) \cdot P(S_3|S_3) \cdot P(S_1|S_3) \cdot P(S_2|S_1) \\
&= 1 \cdot 0.8 \cdot 0.8 \cdot 0.1 \cdot 0.3 \\
&= 0.0192
\end{aligned}
$$

Erweiterung zu Hidden Markov Modellen

- **Einfache Markov Modelle:**
  - Jeder Zustand des Modells korrespondiert zu einem beobachtbaren Ereignis, d.h. aus einer Beobachtung ergibt sich ein eindeutig bestimmter Zustand.

- **Hidden Markov Modelle:**
  - Die Beobachtung ist eine probabilistische Funktion des Zustands, so dass sich aus einer Reihe von Beobachtungen nicht eindeutig eine Reihe von Zuständen ableiten lässt.

- Eine Sequenz von Beobachtungen, z.B. *RRBBBGGRR* … korrespondiert nicht eindeutig zu einer Sequenz von Zuständen, da nicht bekannt ist, aus welche Urne die Kugel entnommen wurde.

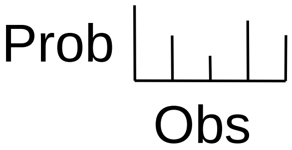


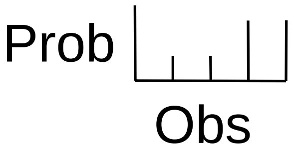
$P(R) = 0.5$	$P(R) = 0.8$	$P(R) = 0.4$
$P(G) = 0.2$	$P(G) = 0.1$	$P(G) = 0.3$
$P(B) = 0.3$	$P(B) = 0.1$	$P(B) = 0.3$

- Ein *Hidden Markov Modell* ist ein Tupel $\lambda = (S, O, A, B, \pi)$
  Dabei bezeichnet
  - $S = \{S_1, S_2, ..., S_N\}$ die *unsichtbaren* HMM-*Zustände*.
  - $O = \{o_1, o_2, ..., o_M\}$ die *sichtbaren Beobachtungen*.
  - $A = \{a_{ij}\}, \quad a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i)$ die *Übergangs-wahrscheinlichkeiten* zwischen den Zuständen.
  - $\pi = \{\pi_i\}, \quad \pi_i = P(q_1 = S_i)$ die initialen *Anfangs-wahrscheinlichkeiten* für die Zustände.
  - y_t die Beobachtung und q_t der Zustand zur Zeit t.
  - $B = \{b_j(k)\}, \quad \mathrm{b}_j(k) = P(y_t = o_k \mid q_t = S_j)$ die *Beobachtungswahrscheinlichkeiten* für die Beobachtungen y_t in den Zuständen q_t

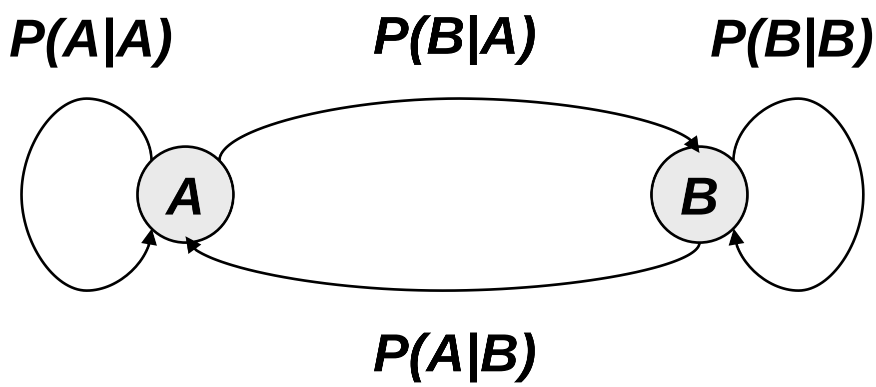
Hidden Markov Modelle
Formale Definition

- Man zeichnet das HMM dann als Automat wie folgt:





$P(o_1|A)$
$P(o_2|A)$
...
$P(o_M|A)$



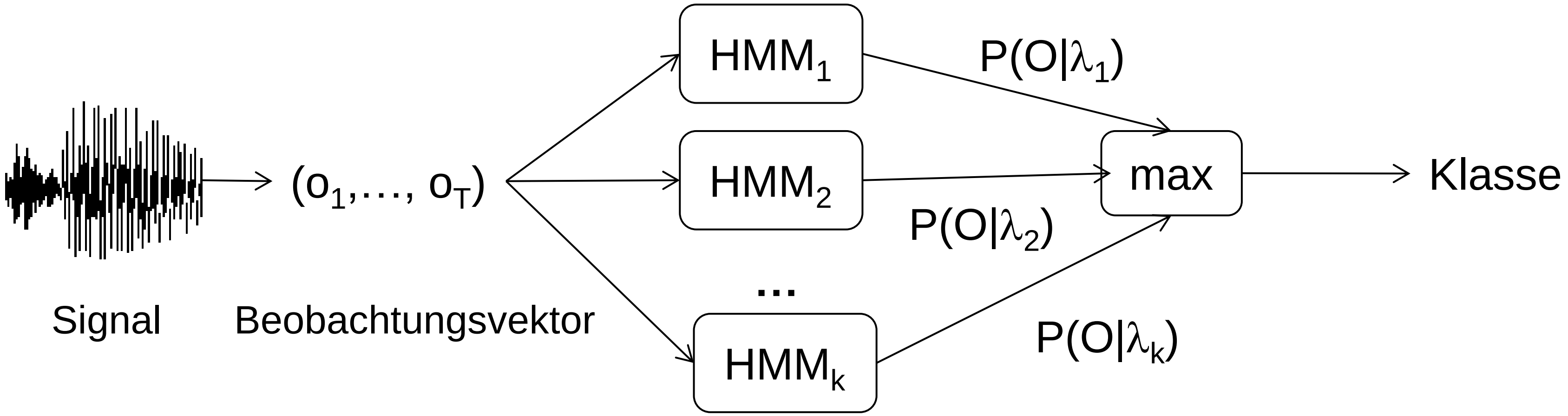
$P(o_1|B)$
$P(o_2|B)$
...
$P(o_M|B)$

Beobachtungswahrschein-lichkeiten im Zustand A

Beobachtungswahrschein-lichkeiten im Zustand B

- **Evaluation**
  - Berechne die Wahrscheinlichkeit für eine bestimmte Beobachtungssequenz bei einem gegebenen Modell.
- **Interpretation**
  - Finde Zustandssequenz, die die Wahrscheinlichkeit der Beobachtungssequenz maximiert, also diese Beobachtungssequenz am besten erklärt.
- **Training**
  - Passe die Modellparameter so an, dass die Wahrscheinlichkeit der Beobachtungssequenz maximiert wird.

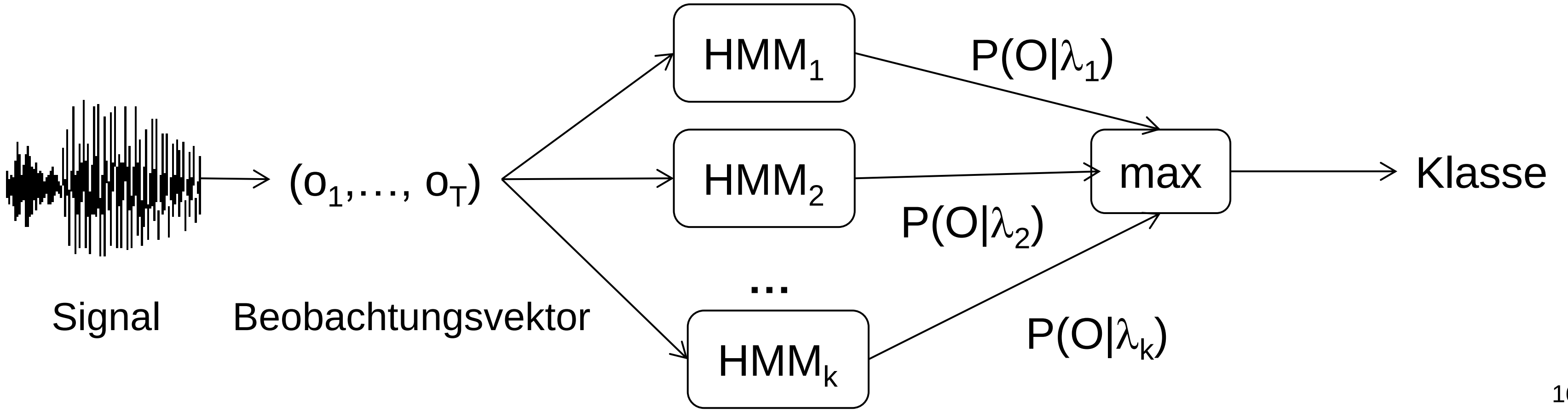
- Für Klassifikationsaufgaben wird für jede Klasse *k* ein eigenes HMM benötigt. Dabei müssen die Parameter für das Modell λ_k so angepasst werden, dass es die zur jeweiligen Klasse gehörenden Beobachtungen möglichst gut abbilden kann.



Hidden Markov Modelle
Klassifikation mit HMM

Universität Augsburg University

- Auskunft darüber, wie gut das HMM mit den Parametern λ_k eine Beobachtung O abbildet, wird über die Berechnung der Produktionswahrscheinlichkeiten $P(O|\lambda_k)$ erhalten.
- Zur Klassifikation wird die Beobachtung O allen HMM als Eingabe präsentiert und die jeweilige Produktionswahrscheinlichkeit $P(O|\lambda_k)$ berechnet.
- Das HMM, das die größte Produktwahrscheinlichkeit liefert, repräsentiert die gesuchte Klasse.

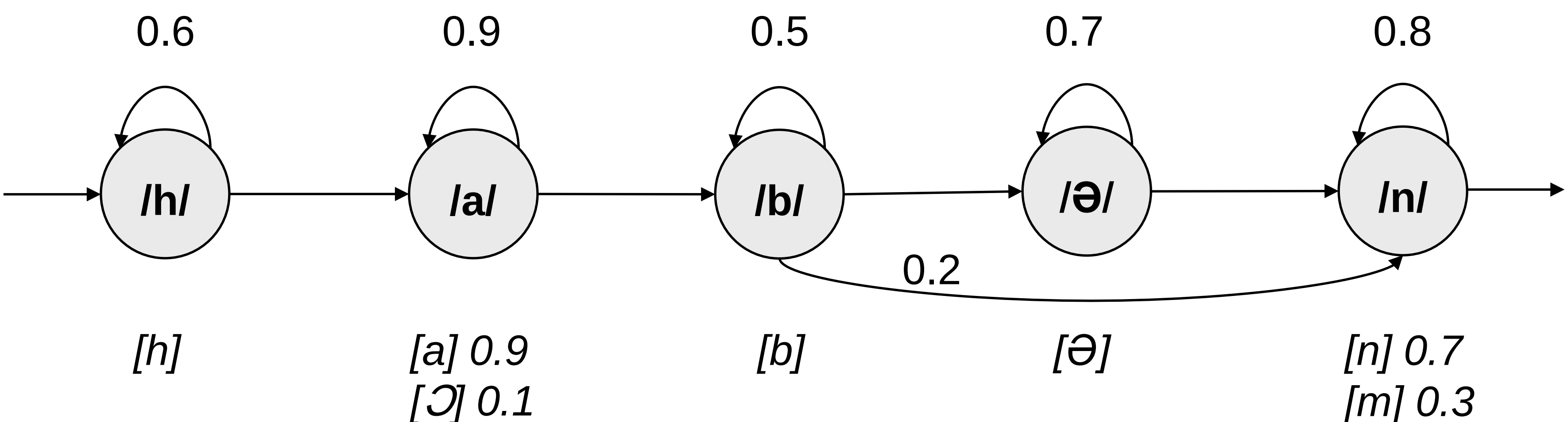


- **Evaluation:**
  Gegeben sei eine Sequenz von Merkmalsvektoren $X_1,\ldots,X_N$. Wie groß ist die Wahrscheinlichkeit dafür, dass die Sequenz von einem bestimmten Wort-Modell produziert wurde?

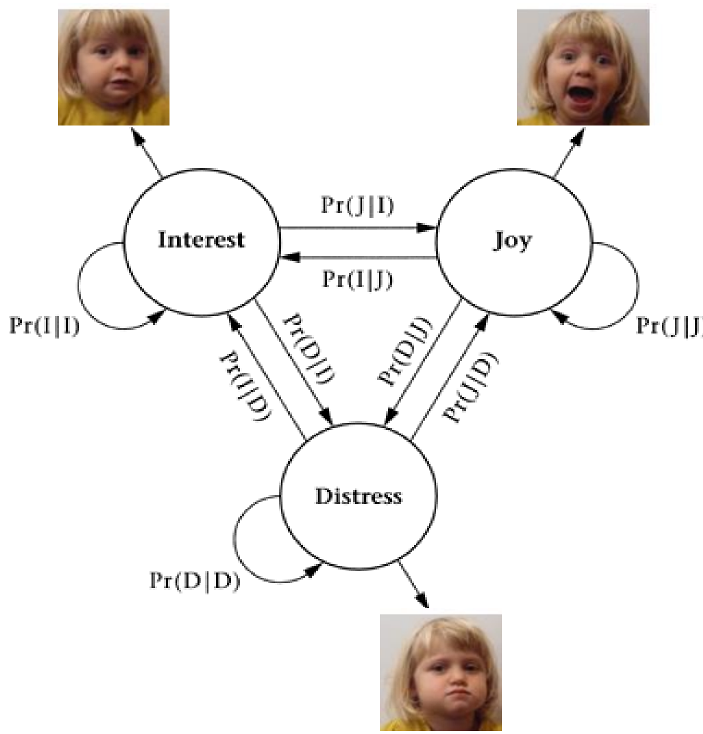
- **Interpretation:**
  Unter der Annahme, die Sequenz wurde von dem Wort-Modell produziert – welches war die wahrscheinlichste Sequenz eingenommener Zustände (die z. B. Phoneme beschreiben können)?

- **Training:**
  Wie müssen die Wahrscheinlichkeitsparameter justiert werden, damit das Modell die typischerweise auftretenden Realisierungen eines Wortes möglichst genau vorhersagt?

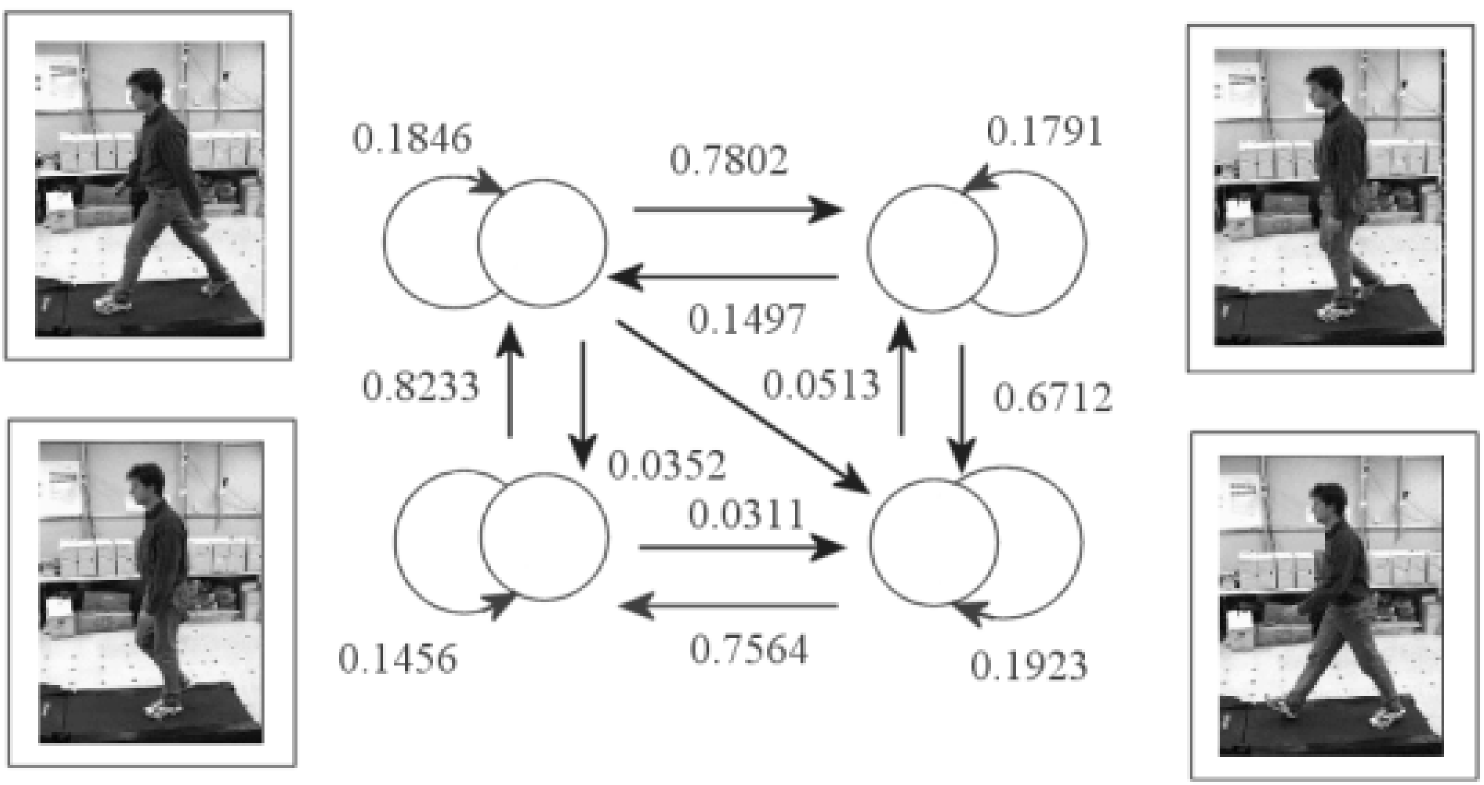
- **Beispiel:** Ein Vereinfachtes Modell für das Wort *„haben“*
- Das Model erfasst zeitliche Verzerrungen des Wortes.
- Das Model erfasst spektrale Variation (z.B. /a/ vs. /Ɔ/).
- Das Model berücksichtigt optionale Lautelimination (z.B. die Überbrückung des Zentrallauts /Ə/ in der Wortmitte)



- Sage Zustände anhand von Beobachtungsfolgen voraus



- Erkennung von *Aktivität* (z.B. Laufen vs. Stillstehen)



Universität Augsburg University

- Unterscheidung zwischen unterschiedlichen Formen von nonverbalem Feedback (z.B. Kopfnicken vs. -schütteln).

Zustimmung (Kopfnicken)

Ablehnung (Kopfschütteln)

Hidden Markov Modelle
Bonusaufgabe

- **Gefangener im Verlies**

Ein Gefangener im Kerkerverlies möchte das aktuelle Wetter herausfinden. Er schätzt, dass die Schuhe der Wärter **bei Regen zu 90 % dreckig**, bei **sonnigem Wetter aber nur zu 60 % dreckig** sind, so kann er durch Beobachtung der Wärterschuhe Rückschlüsse über das Wetter ziehen. Zu Beginn geht er davon aus, dass alle **Wetteränderungen gleichwahrscheinlich** sind.

- Aufgabe: Modellierung des Problems mittels einem **Hidden Markov Modell**

Universität Augsburg University

- Berechne die Wahrscheinlichkeit $P(O|\lambda)$ für die Sequenz von Beobachtungen $O = \{o_1, o_2, \ldots o_T\}$ bei einem Modell λ

- Betrachtet man eine Zustandsequenz $Q = q_1 \ldots q_T$, dann ist die Wahrscheinlichkeit für die Beobachtungssequenz

  $$P(O|Q,\lambda) = \prod_{t=1}^{T} P(o_t|q_t,\lambda) = b_{q_1}(o_1) \cdot b_{q_2}(o_2) \ \ldots \ b_{q_T}(o_T)$$

  wenn wir annehmen, dass die einzelnen Beobachtungen $o_1, o_2, \ldots o_T$ alle bedingt unabhängig voneinander sind.

- Die Wahrscheinlichkeit einer Zustandssequenz Q ist:

  $$P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \ldots a_{q_{T-1} q_T}$$

Universität Augsburg University

- Die **gemeinsame Wahrscheinlichkeit** von *O* und *Q*, d.h. die Wahrscheinlichkeit, dass *O* und *Q* simultan auftreten, ist einfach das Produkt der Terme:

$$P(O,Q|\lambda) = P(O|Q,\lambda) \cdot P(Q|\lambda)$$

- Die Wahrscheinlichkeit von *O* erhält man, indem man diese gemeinsamen Wahrscheinlichkeiten für alle möglichen Zustandssequenzen *Q* aufaddiert.

$$P(O|\lambda) = \sum_{\text{alle } Q} P(O|Q,\lambda)\, P(Q|\lambda) =$$

$$\sum_{\text{alle } Q} \pi_{q_1} b_{q_1}(o_1) \cdot a_{q_1 q_2} b_{q_2}(o_2) \cdot \ldots \cdot a_{q_{T-1} q_T} b_{q_T}(o_T)$$

- **Laufzeitabschätzung**: Zu jedem Zeitpunkt T gibt es N Zustände, die erreicht werden können, d.h. es gibt N^T mögliche Zustandssequenzen.

- Zur Berechnung von O braucht man also $(2T \cdot N^T) - 1$ Berechnungen, genauer $(2T-1) \cdot N^T$ Multiplikationen und zusätzlich $N^T - 1$ Additionen.

- Selbst für kleine Werte für N und T ist diese Berechnung **nicht praktikabel**. Für $N=5$ und $T=100$ braucht man z.B. bereits ca. 10^{72} Berechnungen.

Hidden Markov Modelle
Forward-Algorithmus

- Die Wahrscheinlichkeit, dass man sich zum Zeitpunkt *t* in Zustand S_i befindet und bis zu diesem Zeitpunkt die Beobachtungssequenz $o_1 \ldots o_t$ vorliegt:

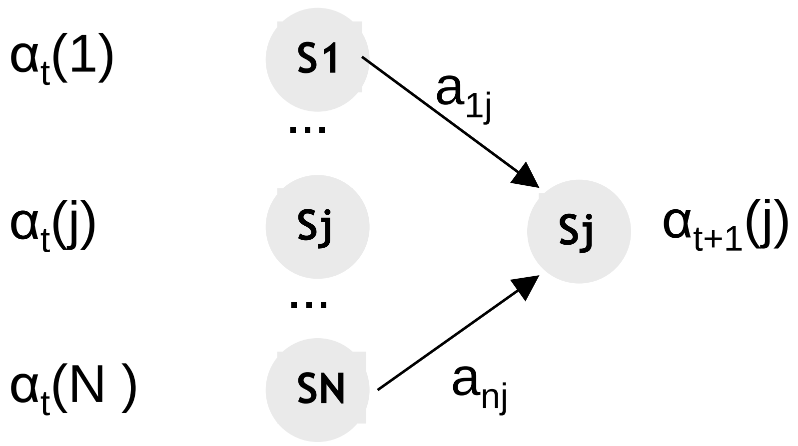
  $\alpha_t(i) = P(o_1 o_2 \ldots o_t, q_t = S_i | \lambda)$
- **Rekursive Berechnung von α:**

$$\alpha_1(i) = \pi_i * b_i(o_1) \quad 1 \le i \le N$$

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

$$1 \le t \le T-1,\ 1 \le j \le N$$

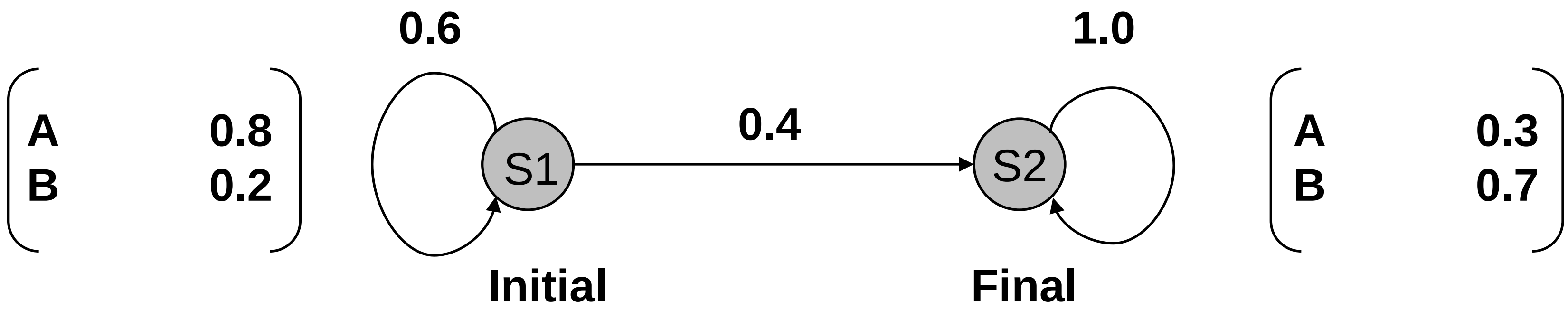
$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$



- **Laufzeit:** $O(N^2 T)$

Forward-Algorithmus

- Gegeben folgendes HMM-Modell:



- Wahrscheinlichkeit, dass man sich zum Zeitpunkt t2 in S1 befindet und bis dahin AA vorliegt.

$\alpha 1(S1) = \pi_{S1} * b_{S1}(o_{t1}) = 1.0*0.8=0.8$

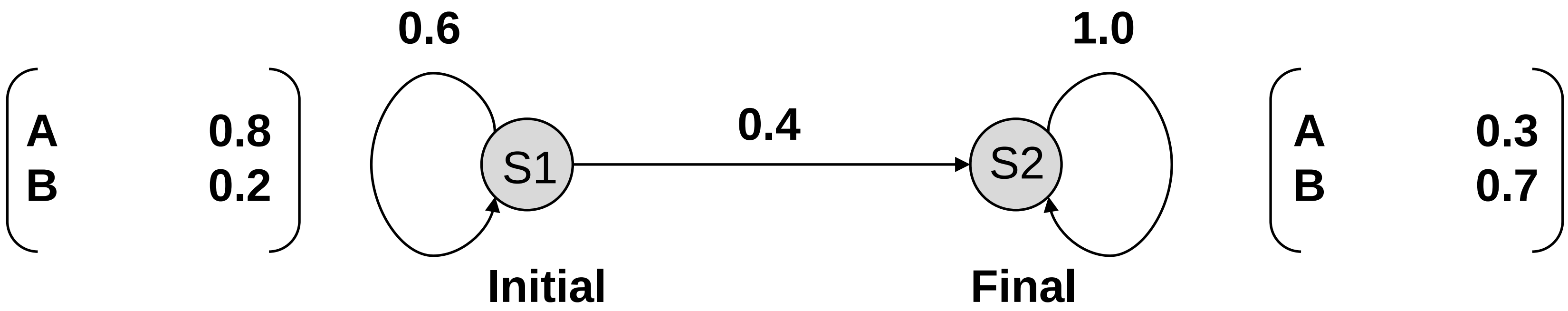
$\alpha 1(S2) = \pi_{S2} * b_{S2}(o_{t1}) = 0.0*0.3=0.0$

$\alpha 2(S1) = [\alpha 1(S1)*a_{S1,S1} + \alpha 1(S2)*a_{S2,S1}]*b_{S1}(o_{t2})$

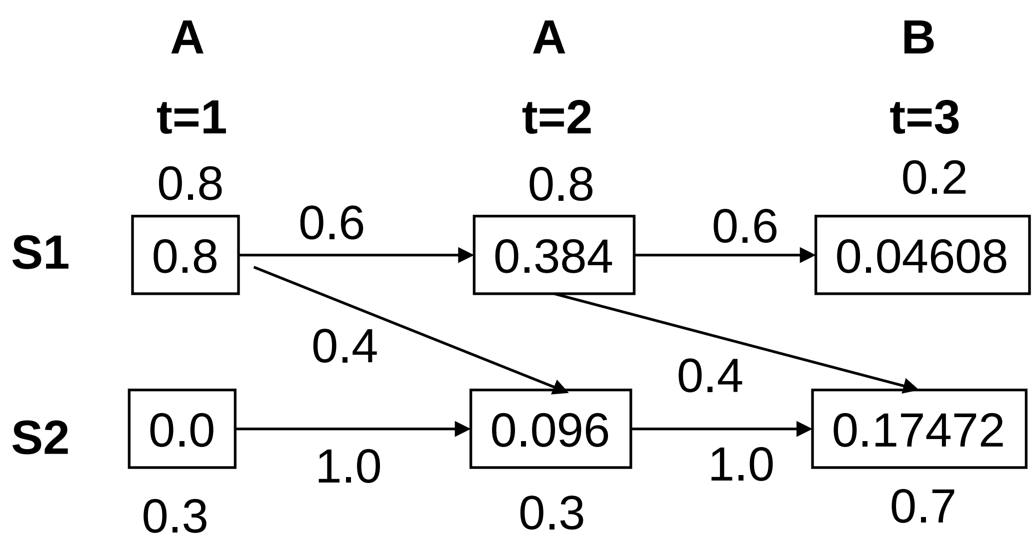
$= 0.8*0.6*0.8 = 0.384$

Forward-Algorithmus

- Gegeben folgendes HMM-Modell:



- Berechne Wahrscheinlichkeit der Beobachtungssequenz: AAB



Lösung:
0.04608 + 0.17472 = **0.2208**

Hidden Markov Modelle Backward-Algorithmus

- Wahrscheinlichkeit, dass man sich zum Zeitpunkt t im Zustand S_i befindet und ab diesem Zeitpunkt die Sequenz $o_{t+1} \ldots o_T$, beobachtet:

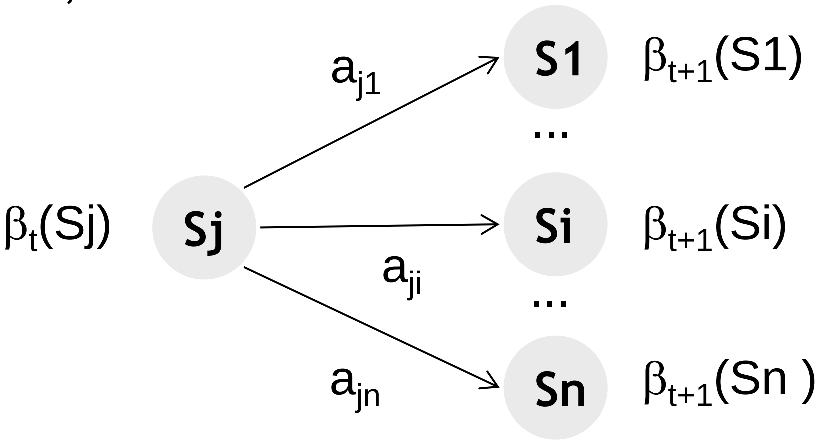
  $\beta_t(i) = P(o_{t+1} o_{t+2} \ldots o_T, q_t = S_i | \lambda)$

  $\beta_T(i) = 1$ für alle $1 \leq i \leq N$

- **Rekursive Berechnung von β:**

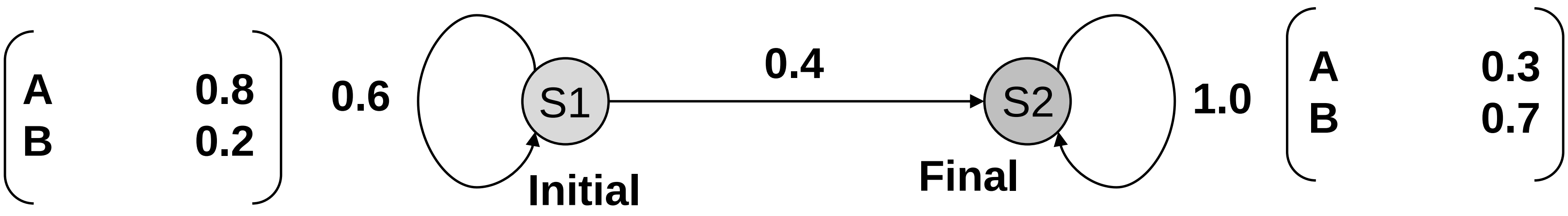
$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \quad 1 \leq t \leq T-1,\ 1 \leq i \leq N$$

$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i)$$



Backward-Algorithmus

- Gegeben folgendes HMM-Modell:



- Wahrscheinlichkeit, dass man sich zum Zeitpunkt t1 in S1 befindet und ab dann AB beobachtet.

$\beta3(S1) = 1,\ \beta3(S2) = 1$

$\beta2(S1) = a_{S1,S1} * b_{S1}(o_{t2}) * \beta3(S1) + a_{S1,S2} * b_{S2}(o_{t2}) * \beta3(S2)$

$ = 0.6 * 0.2 * 1.0 + 0.4 * 0.7 * 1.0 = 0.4$

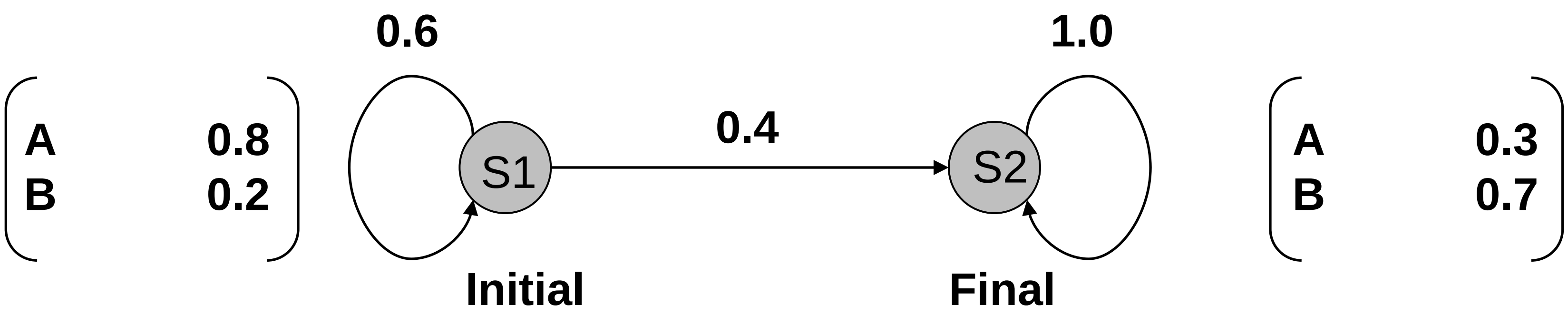
$\beta2(S2) = a_{S2,S1} * b_{S1}(o_{t2}) * \beta3(S1) + a_{S2,S2} * b_{S2}(o_{t2}) * \beta3(S2) = 0.7$

$\beta1(S1) = a_{S1,S1} * b_{S1}(o_{t1}) * \beta2(S1) + a_{S1,S2} * b_{S2}(o_{t1}) * \beta2(S2)$

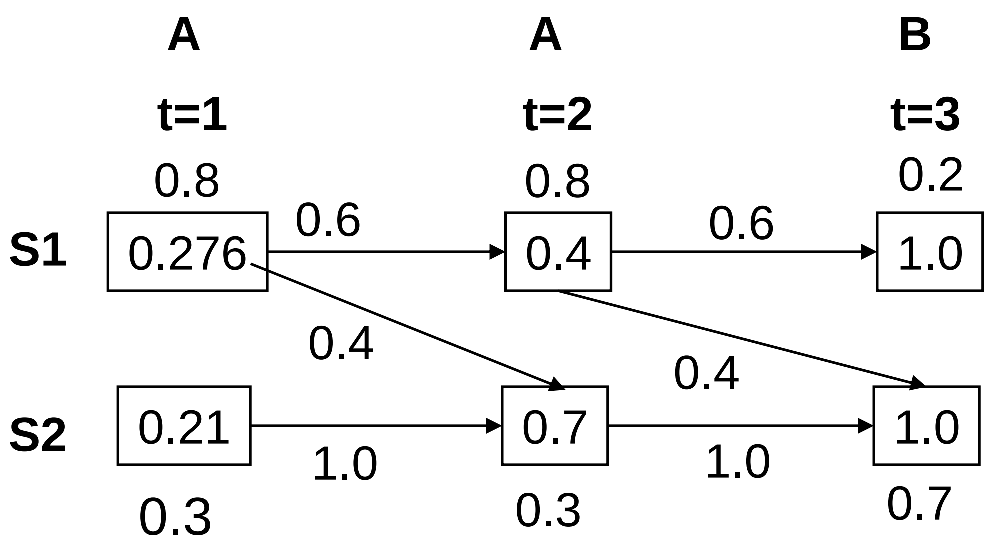
$ = 0.6*0.8*0.4+0.4*0.3*0.7=0.276$

Beispiel: Backward Algorithmus

- Gegeben folgendes HMM-Modell:



- Berechne Wahrscheinlichkeit der Beobachtungssequenz: AAB



Lösung:
1.0*0.8*0.276=
0.2208

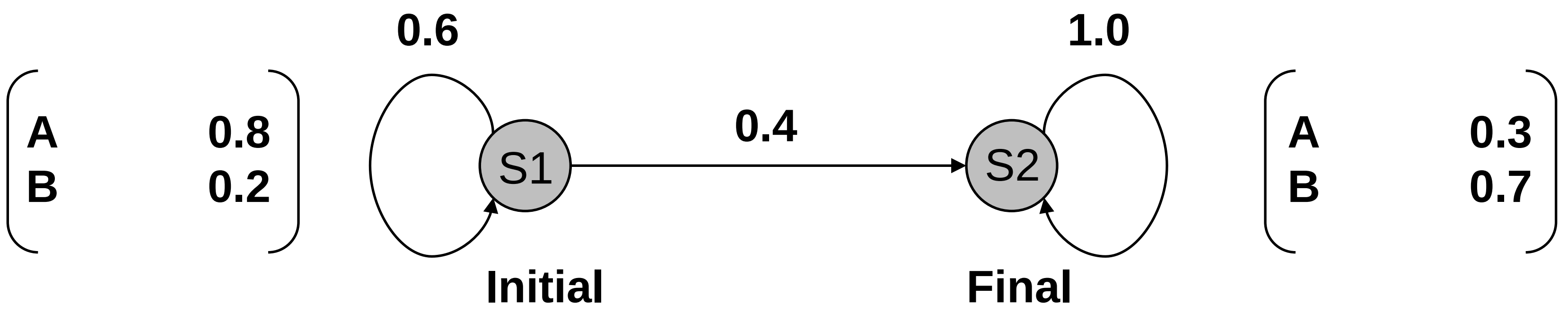
SCIENTIA ET CONSCIENTIA

UNIA Universität Augsburg University

- **Viterbi Algorithmus:**
  - Finde die Zustandssequenz Q, die $P(O,Q|\lambda)$ maximiert
  - Ähnlich zu Forward-Algorithmus, jedoch Maximumbestimmung anstelle von Summenbildung
    $\delta_t(i) = MAXq_1,\ldots,q_{t-1}P(o_1o_2\ldots o_t,q_t=S_i|\lambda)$
  - **Rekursive Berechnung:**
  - $\delta_1(j)= \pi_j{*}b_j(o_1)$, $1 \leq j \leq N$
    $\delta_{t+1}(j) = MAX_{i=1,\ldots,N}[\delta_t(i)\ a_{ij}]b_j(o_{t+1})\ \ 1 \leq j \leq N$
    $MAX\ P(O,Q|\lambda)=MAX_{1\leq i\leq N}\delta_T(i)$
    Speichere das Maximum für Backtracking am Ende

Viterbi Algorithmus

- Gegeben folgendes HMM-Modell



- Berechne wahrscheinlichste Zustandsfolge für die Beobachtungsfolge AA

$\delta_1(S1) = \pi_{S1} * b_{S1}(o_1) = 0.8$; $\delta_1(S2) = \pi_{S2} * b_{S2}(o_1) = 0.0$

$\delta_2(S1) = \mathrm{Max}(\delta_1(S1) * a_{S1,S1} * b_{S1}(o_{t2}) \text{ , } \delta_1(S2) * a_{S2,S1} * b_{S1}(o_{t2}))$

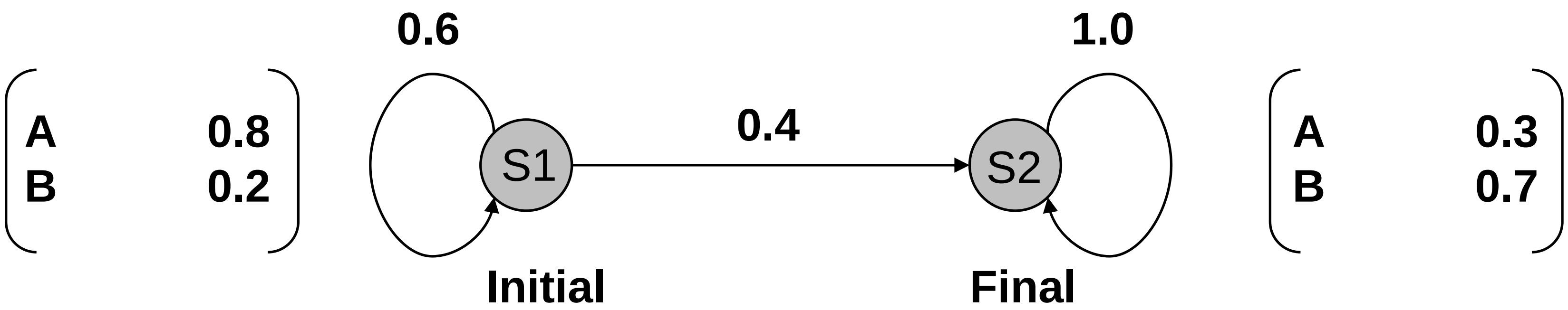
$\delta_2(S1) = \delta_1(S1) * a_{S1,S1} * b_{S1}(o_{t2}) = 0.8*0.6*0.8 = 0.384$

$\delta_2(S2) = \mathrm{Max}(\delta_1(S1) * a_{S1,S2} * b_{S2}(o_{t2}) \text{ , } \delta_1(S2) * a_{S2,S2} * b_{S2}(o_{t2}))$

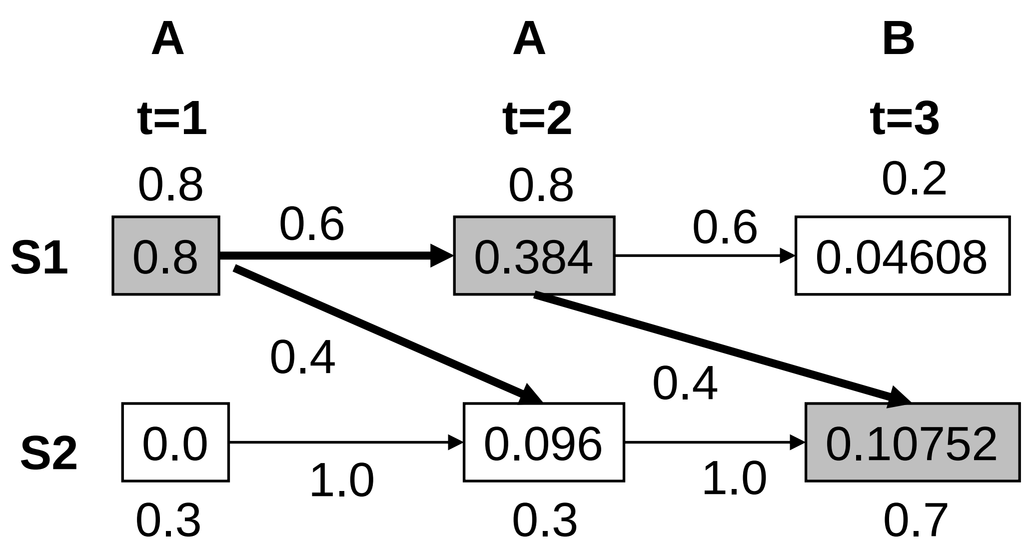
$\delta_2(S2) = \delta_1(S1) * a_{S1,S2} * b_{S2}(o_{t2}) = 0.8*0.4*0.3 = 0.096$

Viterbi-Algorithmus

- Gegeben folgendes HMM-Modell:



- Berechne wahrscheinlichste Zustandssequenz für AAB



Lösung: S1 S1 S2

- **Trainiere Parameter des HMM, um P(O|λ) zu maximieren**
  - Es gibt keinen effizienten Algorithmus, um ein globales Optimum zu bestimmen.
  - Es gibt einen effizienten iterativen Algorithmus, um ein lokales Optimum zu finden.
- **Baum-Welch (Forward-Backward)-Abschätzung**
  - Bestimme die Wahrscheinlichkeiten unter Verwendung des aktuellen Modells λ
  - Verfeinere $\lambda \rightarrow \lambda$‘ basierend auf den berechneten Werten
  - Verwende α und β von Forward- und Backward-Algorithmus

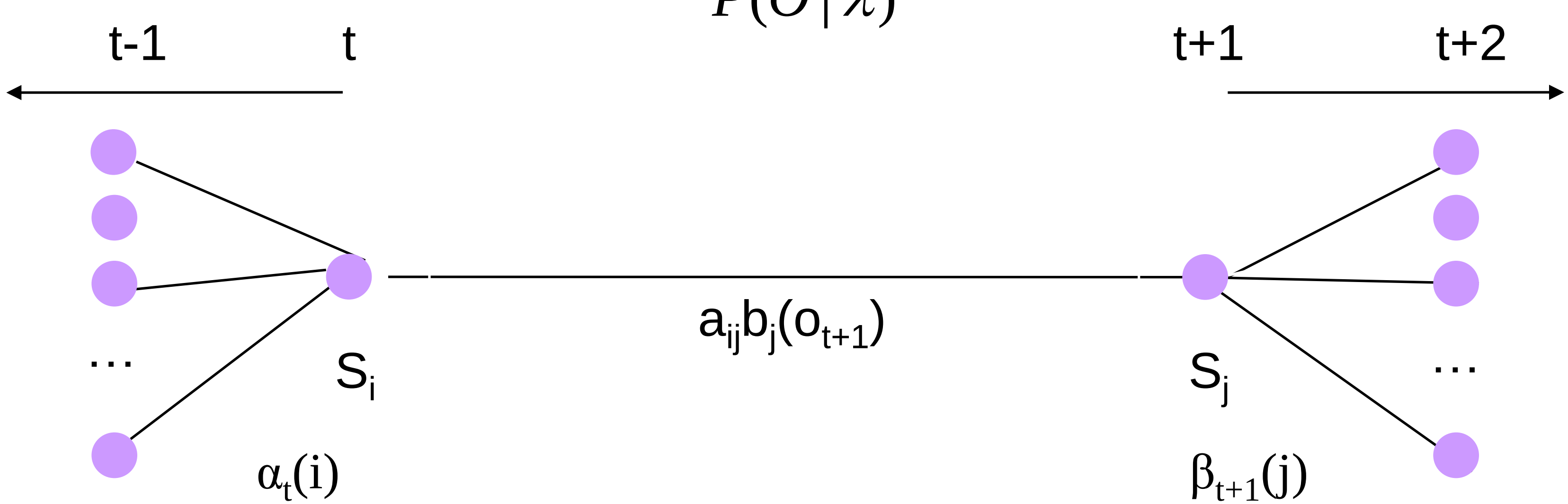
Universität Augsburg University

- Wahrscheinlichkeit des Übergangs von S_i nach S_j zur einer Zeit t bei einer gegebenen Beobachtungssequenz O

$$\xi_t(i,i) = P(q_t = S_i, q_{t+1} = S_j \mid O, \lambda)$$

$$= \frac{P(q_t = S_i, q_{t+1} = S_j, O \mid \lambda)}{P(O \mid \lambda)}$$

$$= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)}$$



- Sei $\gamma_t(i)$ die Wahrscheinlichkeit, dass man sich zur Zeit t im Zustand S_i befindet.
- Es gilt dann: $\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$
- Anzahl der erwarteten Übergänge von S_i $\sum_{t=1}^{T-1} \gamma_t(i)$
- Anzahl der erwarteten Übergänge von S_i nach S_j $\sum_{t=1}^{T-1} \xi_t(i,j)$

Hidden Markov Modelle
Baum-Welch Abschätzung

Universität Augsburg University

$$a'_{ij} = \frac{\text{Erwartete Zahl von Übergängen von } S_i \text{ nach } S_j}{\text{Erwartete Zahl von Übergängen von } S_i} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$b'_j(k) = \frac{\text{Erwartete Anzahl von Zeitpunkten im Zustand j mit Symbol k}}{\text{Erwartete Zahl von Zeitpunkten im Zustand j}} = \frac{\sum_{t: o_t = k} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

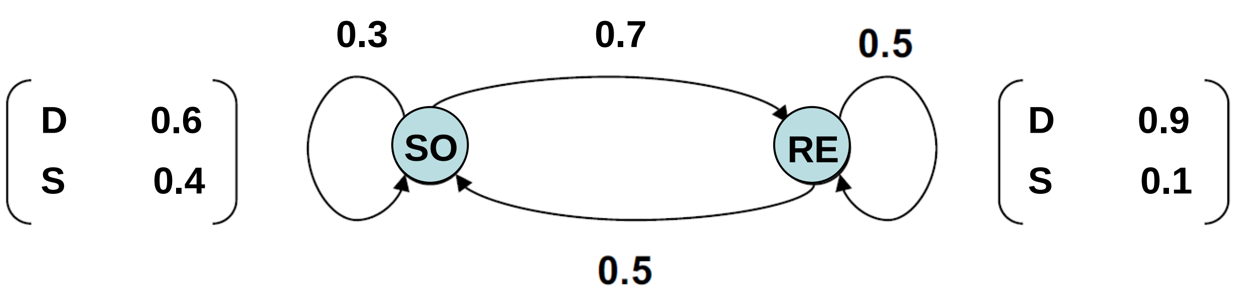
π'_i = Wahrscheinlichkeit, zur Zeit t=1 im Zustand S_i zu sein = $\gamma_1(i)$

- **Algorithmus:**
  1. Initialisiere $\lambda = (A,B)$
  2. Berechne α, β und ξ
  3. Schätze $\lambda' = (A',B')$
  4. Ersetze λ durch λ'
  5. Falls $\lambda \neq \lambda'$ gehe zu 2
- **Konvergenz des Verfahrens**
  - Man kann zeigen, dass entweder:
    1. $\lambda = \lambda'$ oder
    2. $P(O|\lambda) < P(O|\lambda')$ gilt.

- Gefangener im Verlies

Ein Gefangener im Kerkerverlies möchte das aktuelle Wetter herausfinden. Er schätzt, dass die Schuhe der Wärter bei Regen zu 90 % dreckig, bei sonnigem Wetter aber nur zu 60 % dreckig sind, so kann er durch Beobachtung der Wärterschuhe Rückschlüsse über das Wetter ziehen. Zu Beginn geht er davon aus, dass alle Wetteränderungen gleichwahrscheinlich sind.

Ergebnis nach Training (für O = Sauber, Dreckig, Dreckig):



Hidden Markov Modelle
Wahl der Wahrscheinlichkeiten

Universität Augsburg University

- Auch wenn man wenig Informationen hat, müssen die Wahrscheinlich-keiten vernünftig gewählt werden.
- Eine unbedachte Wahl kann ein Training des Modells unmöglich machen.

- Beispiel: