FRIEDRICH-SCHILLER-UNIVERSITÄT JENA
Fakultät für Mathematik und Informatik
INSTITUT für INFORMATIK

Vorlesung im Sommersemester

MUSTERERKENNUNG

E.G. Schukat-Talamazzini

5. Klassifikation zusammengesetzter Muster

Geometrische Verfahren, Stochastische Grammatiken, Hidden Markov Modelle, Kontextuelle Klassifikation,

NUMERISCHE KLASSIFIKATION

$$\delta : \mathbb{R}^D \rightarrow \{1, \ldots, K\}$$

Und was tun wir mit Mustern unbekannter **Länge**, **Dauer**, **Meßstelligkeit** *?*

- Konkurrierende Messungen *eines* Objektes
- ZEIT — Sprach- und Biosignale, Börsenkurse
- ORT — Schriftzüge, Fahrtrouten, „Konturen“
- KONTEXT — Zeichenfolgen, Nukleinsäuren

Definition 5.1
Eine Folge $\boldsymbol{X} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ von Vektoren $\boldsymbol{x}_t$ des $\mathbb{R}^D$ heißt (numerisches) **Longitudinalmuster** *der Länge (Dauer) T.*

Eine Folge $w_1, \ldots, w_T$ von Zeichen eines Alphabets $\mathcal{A}$ heißt **diskretes** *(symbolwertiges)* **Longitudinalmuster** *der Länge T.*

Die Länge $T = T_{\boldsymbol{X}}$ eines longitudinalen Musters ist i.a. von Objekt zu Objekt verschieden. Andernfalls verfahre gemäß $\mathbb{R}^{D \times T} = \mathbb{R}^{D \cdot T}$.

Konsortialklassifikation (Rucksackklassifikation)

- Wir setzen ein K-Klassenproblem des $\mathbb{R}^D$ voraus.
- Ein Objekt der (unbekannten) Klasse Ω_κ wird ausgewählt.
- Das Objekt wird T-mal unabhängig vermessen:

$$\boldsymbol{x}_1 \sim \mathrm{P}(\boldsymbol{y}|\Omega_\kappa), \quad \boldsymbol{x}_2 \sim \mathrm{P}(\boldsymbol{y}|\Omega_\kappa), \quad \ldots, \quad \boldsymbol{x}_T \sim \mathrm{P}(\boldsymbol{y}|\Omega_\kappa)$$

Satz 5.1
Ein Objekt mit den konkurrierenden Meßvektoren $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ wird mit minimaler Fehlerrate klassifiziert unter Verwendung der **konsortialen** *Bayesregel:*

$$\delta([\boldsymbol{x}_t]) \;=\; \operatorname*{argmax}_{\kappa} \left(\mathrm{P}(\Omega_\kappa) \cdot \prod_{t=1}^{T} \mathrm{P}(\boldsymbol{x}_t|\Omega_\kappa) \right)$$

Der obenstehende, zu maximierende Ausdruck entspricht dem Zähler der a posteriori Klassenwahrscheinlichkeit nach Beobachtung der T unabhängig gezogenen Proben.

Satz 5.2 *Für klassenweise normalverteilte Mustervektoren lautet die Bayesregel*

$$\delta([\boldsymbol{x}_t]) \;=\; \operatorname*{argmax}_{\kappa} \left(\mathrm{P}(\Omega_\kappa) \cdot \exp(-\tfrac{1}{2} Q_\kappa) \right)$$

mit den logarithmierten Verbundwahrscheinlichkeiten

$$Q_\kappa \;=\; T \cdot \left\{ \log |2\pi \boldsymbol{S}_\kappa| + \operatorname{spur}\left(\boldsymbol{S}_\kappa^{-1} \cdot \left(\hat{\boldsymbol{S}} + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_\kappa)(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_\kappa)^\top \right) \right) \right\}$$

Dabei sind $\hat{\boldsymbol{\mu}}$ und $\hat{\boldsymbol{S}}$ Mittelwertvektor und Kovarianzmatrix der Meßvektorenprobe und $\boldsymbol{\mu}_\kappa$ und $\boldsymbol{S}_\kappa$ die klassenbedingten Normalverteilungsstatistiken.

Der Levenshtein-Abstand

ELEMENTAROPERATIONEN auf Zeichenketten:

- **Substitution** — Ersetzung eines Zeichens durch ein anderes
- **Deletion** — Löschung eines Zeichens
- **Insertion** — Einfügung eines Zeichens

Definition 5.2
Ist $\mathcal{A}$ ein endliches Alphabet und sind $\boldsymbol{v}$, $\boldsymbol{w}$ zwei Zeichenfolgen aus $\mathcal{A}^\star$, so bezeichnet der **Levenshtein-Abstand** *$d_{\mathrm{LD}}(\boldsymbol{v}, \boldsymbol{w})$ die minimale Anzahl von Elementaroperationen, mit denen $\boldsymbol{v}$ in $\boldsymbol{w}$ überführt werden kann.*

Definition 5.3
Quantifizieren die nichtnegativen reellen Zahlen $r(a, b)$, $r(a, \$)$ und $r(\$, b)$ die **Editierkosten** *für die Ersetzung des Zeichens $a \in \mathcal{A}$ durch $b \in \mathcal{A}$ bzw. für die Löschung von a oder die Einfügung von b, so bezeichnet der* **gewichtete Levenshtein-Abstand** *$d_{\mathrm{WLD}}(\boldsymbol{v}, \boldsymbol{w})$ die minimale Kostensumme von Elementaroperationen zur Überführung von $\boldsymbol{v}$ in $\boldsymbol{w}$.*

OPTIMALITÄTSPRINZIP:

> *„Jeder Teilpfad eines kostenoptimalen Zuordnungspfades ist selbst wieder eine kostenoptimale Zuordnung.“*

Diese Aussage gilt insbesondere für die kostenoptimale Zuordnung der Anfangsstücke $v_1, \ldots, v_n$ und $w_1, \ldots, w_m$ zweier Zeichenketten $\boldsymbol{v}$ und $\boldsymbol{w}$.

Satz 5.3 (Dynamische Programmierung)
Für zwei Zeichenfolgen $\boldsymbol{v}$ und $\boldsymbol{w}$ der Länge N und M seien die **partiellen kumulativen Distanzen**

$$D_{n,m} \stackrel{\text{def}}{=} d_{\text{WLD}}(v_1, \ldots, v_n, w_1, \ldots, w_m)\ , \quad (n,m) \in [0,N] \times [0,M]$$

definiert. Dann gelten die Rekursionsgleichungen:

$$D_{n,m} = \begin{cases} 0 & n,m = 0 \\ D_{n-1,0} + r(v_n, \$) & m = 0 \\ D_{0,m-1} + r(\$, w_m) & n = 0 \\ \min \begin{pmatrix} D_{n-1,m-1} & + & r(v_n, w_m) \\ D_{n-1,m} & + & r(v_n, \$) \\ D_{n,m-1} & + & r(\$, w_m) \end{pmatrix} & \textit{sonst} \end{cases}$$

Für den ungewichteten Levenshtein-Abstand vereinfacht sich der Rekursionsschritt zu

$$D_{n,m} = \min \begin{pmatrix} D_{n-1,m-1} + \begin{cases} 0 & v_n = w_m \\ 1 & v_n \neq w_m \end{cases} \\ D_{n-1,m} + 1 \\ D_{n,m-1} + 1 \end{pmatrix}$$

Kumulative Abstandsmaße

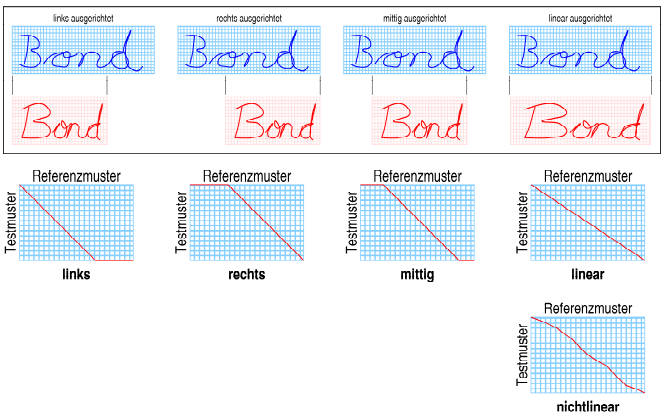
zwischen longitudinalen Mustern

Minimum-Abstand-Klassifikator

$$\delta(\boldsymbol{X}) = \operatorname*{argmin}_{\kappa} d(\boldsymbol{X}, \boldsymbol{Z}_\kappa)$$

Nächster-Nachbar-Regel

$$\delta(\boldsymbol{X}) = \operatorname*{argmin}_{\kappa} \left(\min_{\boldsymbol{Z} \in \omega_\kappa} d(\boldsymbol{X}, \boldsymbol{Z}) \right)$$



Lokale und **kumulative** Abstände

$$d(\boldsymbol{X}, \boldsymbol{Z}) = F\left([d(\boldsymbol{x}_t, \boldsymbol{z}_s)]_{t=1,\ldots,|\boldsymbol{X}|}^{s=1,\ldots,|\boldsymbol{Z}|} \right) = \sum_{r=1}^{T_\phi} d(\boldsymbol{x}_{\phi_1(r)}, \boldsymbol{z}_{\phi_2(r)})$$

mit einer geeigneten Skalenausrichtung (ϕ_1, ϕ_2)

Die optimale Skalenausrichtungsfunktion

Definition 5.4 *Die Abbildung*

$$\phi : \begin{cases} \{1, \ldots, R\} & \to & \{1, \ldots, T\} \times \{1, \ldots, S\} \\ r & \mapsto & \phi(r) = \langle \phi_1(r), \phi_2(r) \rangle \end{cases}$$

heißt **Skalenausrichtungsfunktion** *mit der Länge $R = T_\phi$ und der Dimension $T \times S$, wenn die folgenden Bedingungen erfüllt sind:*

$$\begin{aligned} \phi(1) &= \langle 1, 1 \rangle \\ \phi(R) &= \langle T, S \rangle \end{aligned}$$

$$\begin{aligned} r \leq r' &\Rightarrow \phi_1(r) \leq \phi_1(r') \text{ und } \phi_2(r) \leq \phi_2(r') \\ r \neq r' &\Rightarrow \phi(r) \neq \phi(r') \end{aligned}$$

Die Funktion ϕ richtet die beiden Skalenbereiche *vollständig* und *monoton* gegeneinander aus. ϕ heißt auch Zeitverzerrungsfunktion oder *„time-warping function“*.

Definition 5.5 *Für zwei Vektorfolgen $\boldsymbol{X}, \boldsymbol{Y}$ und eine Ausrichtungsfunktion ϕ der Dimension $T_X \times T_Y$ heißt*

$$d_\phi(\boldsymbol{X}, \boldsymbol{Y}) \stackrel{\text{def}}{=} \sum_{r=1}^{T_\phi} \left\| \boldsymbol{x}_{\phi_1(r)} - \boldsymbol{y}_{\phi_2(r)} \right\|$$

der **kumulative Abstand** *zwischen $\boldsymbol{X}$ und $\boldsymbol{Y}$ bezüglich der Ausrichtung ϕ.*
Ist Φ eine Menge von $(T_X \times T_Y)$-Ausrichtungsfunktionen, so heißt der Wert

$$d_\Phi(\boldsymbol{X}, \boldsymbol{Y}) = \min_{\phi \in \Phi} d_\phi(\boldsymbol{X}, \boldsymbol{Y})$$

der kumulative Abstand zwischen $\boldsymbol{X}$ und $\boldsymbol{Y}$ bezüglich Φ. Ein $\phi^* \in \Phi$ mit diesem Abstandswert heißt **optimale Ausrichtungsfunktion** *für $\boldsymbol{X}$ und $\boldsymbol{Y}$.*

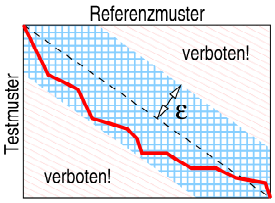
Erlaubte Skalenausrichtungen

Endpunktkonsistenz

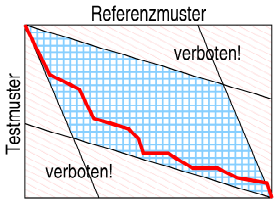
Monotonie

Zyklenfreiheit

Globale Einschränkungen



ε-Streifen um die Diagonale

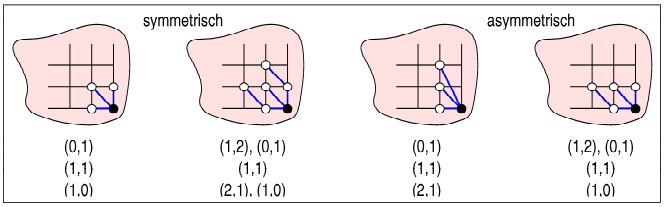


Steigungsschranken

Lokale Einschränkungen („Stetigkeit“)

Vorgabe der lokalen Differenzen für benachbarte Gitterpunkte — Φ enthält nur Ausrichtungen ϕ mit

$$\Delta_\phi(r) \stackrel{\text{def}}{=} \begin{pmatrix} \phi_1(r) - \phi_1(r-1) \\ \phi_2(r) - \phi_2(r-1) \end{pmatrix} \in \Delta$$



Der DTW-Algorithmus

„Dynamic Time Warping“ (Itakura '75, Sakoe '78)

Satz 5.4

Für zwei Vektorfolgen $\boldsymbol{X}$ und $\boldsymbol{Y}$ der Länge N und M seien die **partiellen kumulativen Distanzen**

$$D_{n,m} \stackrel{\text{def}}{=} d_{\Phi}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m)\ , \quad (n,m)\in[0,N]\times[0,M]$$

für eine geeignete Menge Φ_{Δ} von Zeitverzerrungsfunktionen mit den lokalen Differenzen $\boldsymbol{\Delta}$ definiert. Dann gilt mit den Initialisierungen

$$D_{n,m} = \begin{cases} 0 & n=m=0 \\ -\infty & n=0 \text{ oder } m=0 \end{cases}$$

die folgende Rekursionsgleichung:

$$D_{n,m} = \|\boldsymbol{x}_n-\boldsymbol{y}_m\| + \min_{(\nu,\mu)\in\Delta} D_{n-\nu,m-\mu}$$

Die Rekursion lautet also für $\boldsymbol{\Delta}_1=\{(1,1),(0,1),(1,0)$ speziell

$$D_{n,m} = \|\boldsymbol{x}_n-\boldsymbol{y}_m\| + \min\begin{pmatrix} D_{n-1,m-1} \\ D_{n-1,m} \\ D_{n,m-1}\end{pmatrix}$$

und für $\boldsymbol{\Delta}_3=\{(0,1),(1,1),(2,1)$

$$D_{n,m} = \|\boldsymbol{x}_n-\boldsymbol{y}_m\| + \min\begin{pmatrix} D_{n,m-1} \\ D_{n-1,m-1} \\ D_{n-2,m-1}\end{pmatrix}$$

Für die mehrstufige Differenzenmenge $\boldsymbol{\Delta}_4$ der Abbildung ergibt sich der Ausdruck

$$D_{n,m} = \min\begin{pmatrix} D_{n-1,m} + \|\boldsymbol{x}_n-\boldsymbol{y}_m\| \\ D_{n-1,m-1} + \|\boldsymbol{x}_n-\boldsymbol{y}_m\| \\ D_{n-1,m-2} + \|\boldsymbol{x}_n-\boldsymbol{y}_m\| + \|\boldsymbol{x}_n-\boldsymbol{y}_{m-1}\| \end{pmatrix}$$

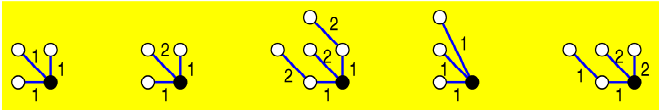
Gewichtete DTW-Abstandsmaße

Diagonalferne Ausrichtungen ϕ besitzen eine größere Länge T_ϕ; ihre Abstandssumme nimmt also tendenziell höhere Werte an.

Gewichteter kumulativer Abstand

$$d_\phi(\boldsymbol{X},\boldsymbol{Y}) \stackrel{\text{def}}{=} \sum_{r=1}^{T_\phi} \phi_w(r)\cdot\left\|\boldsymbol{x}_{\phi_1(r)}-\boldsymbol{y}_{\phi_2(r)}\right\|$$

$\phi_w(r)$ gewichtet den aktuellen Übergang der Ausrichtung und hängt nur von den lokalen Differenzen $\phi(r)-\phi(r-1)$ ab.



REKURSIONSGLEICHUNGEN:

$$D_{n,m} = \min\begin{pmatrix} D_{n,m-1} & + & d_{n,m} \\ D_{n-1,m-1} & + & 2\cdot d_{n,m} \\ D_{n-1,m} & + & d_{n,m}\end{pmatrix}$$

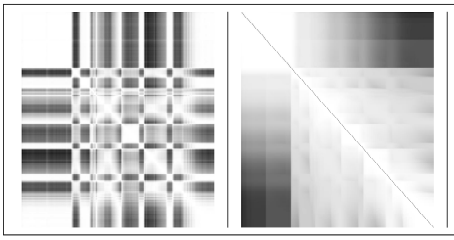
$$D_{n,m} = \min\begin{pmatrix} D_{n-1,m-2} & + & 2\cdot d_{n,m-1}+d_{n,m} \\ D_{n-1,m-1} & + & 2\cdot d_{n,m} \\ D_{n-2,m-1} & + & 2\cdot d_{n-1,m}+d_{n,m}\end{pmatrix}$$

$$D_{n,m} = \min\begin{pmatrix} D_{n,m-1} & + & d_{n,m} \\ D_{n-1,m-1} & + & d_{n,m} \\ D_{n-2,m-1} & + & d_{n,m}\end{pmatrix}$$

$$D_{n,m} = \min\begin{pmatrix} D_{n-1,m-2} & + & d_{n,m-1}+d_{n,m} \\ D_{n-1,m-1} & + & 2\cdot d_{n,m} \\ D_{n-1,m} & + & 2\cdot d_{n,m}\end{pmatrix}$$

Die Gewichtsummen der letzten vier abgebildeten Ausrichtungsmodelle betragen konstant $(N+M)$, $(N+M)$, M und $2N$.

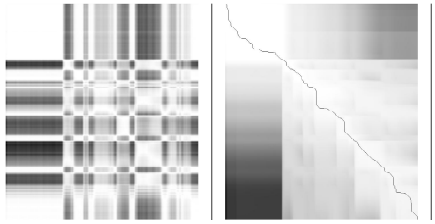
Spektraler Wortvergleich mit DTW



„Edmund Stoiber“
versus
„Edmund Stoiber“

$D(\boldsymbol{X},\boldsymbol{Y})=0$

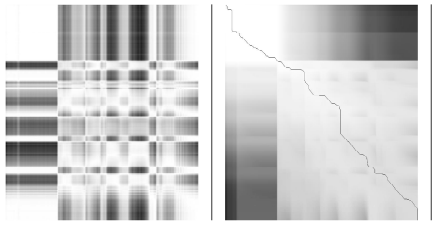
$(\mathbb{R}^{774}\times\mathbb{R}^{774})$



„Edmund Stoiber“
versus
„Edmund Stoiber“

$D(\boldsymbol{X},\boldsymbol{Y})=13081.4$

$(\mathbb{R}^{774}\times\mathbb{R}^{671})$



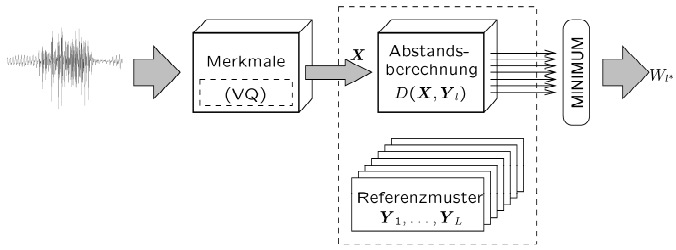
„Edmund Stoiber“
versus
„Angela Merkel“

$D(\boldsymbol{X},\boldsymbol{Y})=42011.9$

$(\mathbb{R}^{774}\times\mathbb{R}^{743})$

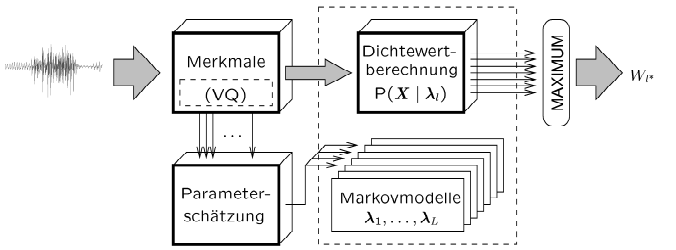
BEISPIEL: **Einzelworterkennung**

Geometrische Modellierung



Dynamic Time Warping — Levenshtein-Abstand

Wahrscheinlichkeitstheoretische Modellierung



Kontinuierliches HMM — Diskret(wertig)es HMM

Wahrscheinlichkeitsmodelle

„Mathematische Maschinen zur (algorithmischen) Berechnung der Erzeugungswahrscheinlichkeiten von Zeichen- oder Vektorfolgen variabler Länge“

Zufallsprozeß

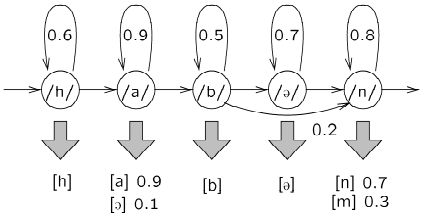
$$\mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_3, \ldots, \mathbb{X}_t, \mathbb{X}_{t+1}, \ldots$$

Produktionswahrscheinlichkeit

$$\mathsf{P}(\boldsymbol{X}) \;=\; \mathsf{P}(\mathbb{X}_1 = \boldsymbol{x}_1, \mathbb{X}_2 = \boldsymbol{x}_2, \ldots, \mathbb{X}_T = \boldsymbol{x}_T)$$

MARKOVKETTEN	(diskret)
STOCHASTISCHE AUTOMATEN	(diskret)
STOCHASTISCHE GRAMMATIKEN	(diskret)
HIDDEN MARKOV MODELLE	(kontinuierlich)

BEISPIEL: Wortaussprachemodelle



Markovketten

Definition 5.6 *Ein Zufallsprozeß*

$$\langle \mathbb{X}_t \mid t \in \mathbb{N} \rangle$$

mit diskreten Variablen $\mathbb{X}_t$ aus einem endlichen Zustandsalphabet

$$\mathcal{S} \;=\; \{s_1, \ldots, s_N\}\;, \qquad N \in \mathbb{N}$$

heißt **Markovkette** *(oder Markovquelle) mit N* **Zuständen**.

Für die gemeinsame Produktionswahrscheinlichkeit einer Zustandsfolge $\boldsymbol{q}$ der Dauer T gilt die bekannte *Kettenregel*:

$$\begin{aligned}\mathsf{P}(\boldsymbol{q}) &= \mathsf{P}(q_1, \ldots, q_T)\\ &= \mathsf{P}(\mathbb{X}_1 = q_1) \cdot \mathsf{P}(\mathbb{X}_2 = q_2 \mid q_1) \cdot \ldots \cdot \mathsf{P}(\mathbb{X}_T = q_T \mid q_1 \ldots q_{T-1})\\ &= \prod_{t=1}^{T} \mathsf{P}(\mathbb{X}_t = q_t \mid q_1 \ldots q_{t-1})\end{aligned}$$

Definition 5.7
Eine Markovkette $[\mathbb{X}_t]$ heißt **einfach** *oder* **erster Ordnung**, *wenn die Auftretenswahrscheinlichkeit eines jeden Zustandes ausschließlich vom Vorgängerzustand abhängt:*

$$\mathsf{P}(\mathbb{X}_t = q_t \mid q_1, \ldots, q_{t-1}) \;=\; \mathsf{P}(\mathbb{X}_t = q_t \mid \mathbb{X}_{t-1} = q_{t-1})$$

Das (statistische) Verhalten einer einfachen Markovkette ist vollständig durch den Vektor $\boldsymbol{\pi}$ der **Anfangszustandswahrscheinlichkeiten**

$$\pi_i \stackrel{\text{def}}{=} \mathsf{P}(\mathbb{X}_1 = s_i)\;, \qquad i = 1, \ldots, N$$

und die zeitabhängige Matrix $\boldsymbol{A}(t)$ der **Übergangswahrscheinlichkeiten**

$$a_{ij}(t) \stackrel{\text{def}}{=} \mathsf{P}(\mathbb{X}_t = s_j \mid \mathbb{X}_{t-1} = s_i)\;, \qquad i, j = 1, \ldots, N$$

charakterisiert.

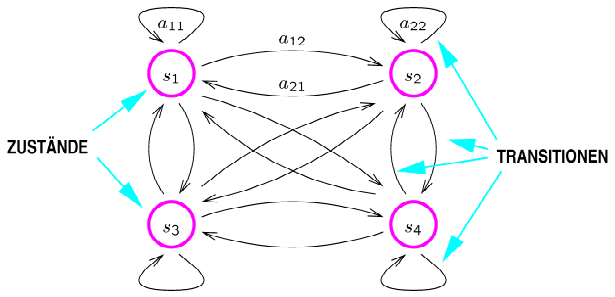
Definition 5.8 *Eine einfache Markovkette $[\mathbb{X}_t]$ heißt* **stationär** *oder* **homogen**, *wenn die Wahrscheinlichkeiten der Zustandsübergange unabhängig vom Zeitindex sind:*

$$\mathsf{P}(\mathbb{X}_t = s_j \mid \mathbb{X}_{t-1} = s_i) \;=\; a_{ij}(t) \;=\; a_{ij} \qquad \textit{für alle } t$$

Eine einfache stationäre Markovkette mit N Zuständen ist durch die Parameter

$$(\boldsymbol{\pi}, \boldsymbol{A})\;, \qquad \boldsymbol{\pi} \in \mathbb{R}^N,\; \boldsymbol{A} \in \mathbb{R}^{N \times N}$$

charakterisiert.



NORMIERUNGSBEDINGUNGEN

$$\sum_{i=1}^{N} \pi_i \;=\; 1\;, \qquad \sum_{j=1}^{N} a_{ij} \;=\; 1 \qquad (\forall i)$$

ZUSTANDSWAHRSCHEINLICHKEITEN IN t

$$\boldsymbol{\pi}^{(t)} \;=\; \left(\boldsymbol{A}^{\top}\right)^{t-1} \boldsymbol{\pi}\;, \qquad \pi_i^{(t)} \stackrel{\text{def}}{=} \mathsf{P}(\mathbb{X}_t = s_i)$$

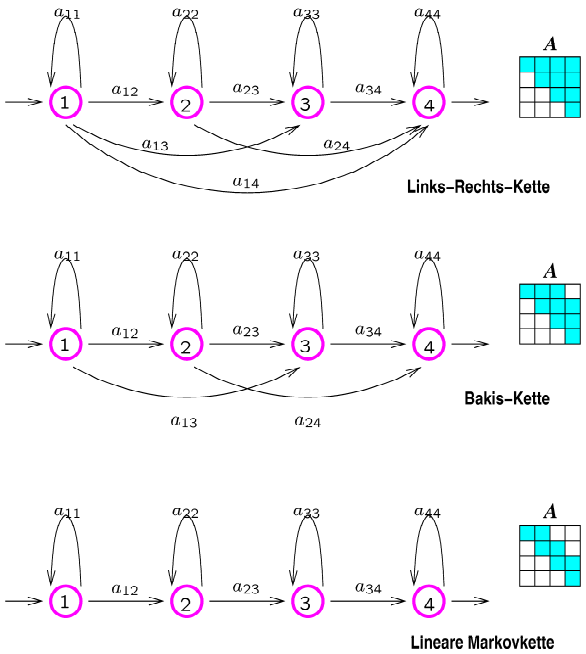
ERGODIZITÄT

$$\lim_{t \to \infty} \mathsf{P}(\mathbb{X}_t = s_j \mid \mathbb{X}_1 = s_i) \;=:\; p_j$$

existiert und ist unabhängig vom Anfangszustand s_i.

Modelltopologien

Markovketten unterschiedlicher Verbindungsstruktur



Die Zustände von LR-Ketten haben eine natürliche Reihenfolge. Ihr letzter Zustand hat die Verbleibenswahrscheinlichkeit $a_{N,N} = 1$. Endzustand s_N ist wesentlich, alle anderen Zustände sind transient.

Gegeben ist eine **Zustandsfolge**

$$\boldsymbol{q} \;=\; q_1,\ldots,q_T \;\in\; \{1,\ldots,N\}^T$$

Die **Verbundwahrscheinlichkeit** von $\boldsymbol{q}$ beträgt

$$\begin{aligned}\mathrm{P}(\boldsymbol{q}) &= \mathrm{P}(\mathbb{X}_1 = q_1, \mathbb{X}_2 = q_2, \ldots, \mathbb{X}_T = q_T)\\ &= \pi_{q_1}\cdot a_{q_1q_2}\cdot a_{q_2q_3}\cdot a_{q_3q_4}\cdot\ldots\cdot a_{q_{T-1}q_T}\\ &= \pi_{q_1}\cdot\prod_{t=2}^{T} a_{q_{t-1}q_t}\end{aligned}$$

Es gilt (für jedes $T\in\mathbb{N}$) die **Normierungsbedingung**

$$\sum_{\boldsymbol{q}\in\mathcal{S}^T}\mathrm{P}(\boldsymbol{q}) \;=\; 1$$

Gegeben ist eine **Lernstichprobe**

$$\omega \;\subseteq\; \mathcal{S}^\star \;=\; \bigcup_{T=1}^{\infty}\mathcal{S}^T$$

Die **ML-Zielgröße**

$$\begin{aligned}\ell_{\pi,A}(\omega) &= \log\prod_{\boldsymbol{q}\in\omega}\left\{\pi_{q_1}\cdot\prod_{t=2}^{T(\boldsymbol{q})} a_{q_{t-1}q_t}\right\} = \sum_{\boldsymbol{q}\in\omega}\log\pi_{q_1} + \sum_{\boldsymbol{q}\in\omega}\sum_{t=2}^{T(\boldsymbol{q})}\log a_{q_{t-1}q_t}\\ &= \sum_{i=1}^{N}\gamma_1(i)\cdot\log\pi_i + \sum_{i=1}^{N}\sum_{j=1}^{N}\xi(i,j)\cdot\log a_{ij}\end{aligned}$$

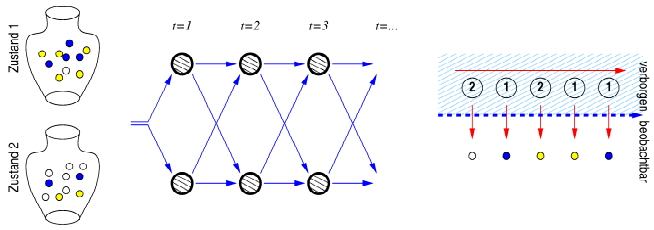
wird maximiert durch die **Parameterschätzwerte**

$$\hat{\pi}_i \;=\; \gamma_1(i)\Big/\sum_{k=1}^{N}\gamma_1(k)\;,\qquad \hat{a}_{ij} \;=\; \xi(i,j)\Big/\sum_{k=1}^{N}\xi(i,k)$$

mit den Stichprobenstatistiken $[\gamma_1(i)]$ und $[\xi(i,j)]$.

Zufallsprozesse in der Mustererkennung

- *Einfache Markovketten*
  Zeichenfolgen, „kurzes Gedächtnis“
- *n-Gramme*
  Markovketten n-ter Ordnung, Gedächtnisdauer $n-1$
- *Stochastische Grammatiken*
  Stochastische Automaten, stochast. kontextfreie Grammatiken
- *Hidden Markov Modelle*
  Latente Zustände — zeichen- oder vektorwertige Ausgabe



Zweistufiger Zufallsprozeß

Wahl eines (verborgenen) ZUSTANDES

$$\mathbb{U}_t \;\in\; \mathcal{S} \;=\; \{s_1,\ldots,s_N\}$$

Wahl eines (beobachtbaren) AUSGABEZEICHENS/-vektors

$$\mathbb{O}_t \;\in\; \mathcal{A} \;=\; \{z_1,\ldots,z_M\} \quad\text{oder}\quad \mathbb{X}_t \;\in\; \mathbb{R}^D$$

Diskrete und kontinuierliche HMM

Definition 5.9
Es sei $\mathcal{S}$ eine Menge von N inneren Zuständen und $\mathcal{A}$ eine Menge von M Ausgabezeichen. Der zweistufige Zufallsprozeß $[\mathbb{U}_t,\mathbb{O}_t]$ heißt **diskretes HMM** *(„hidden markov model“), wenn $[\mathbb{U}_t]$ eine (einfache, stationäre) Markovkette über $\mathcal{S}$ ist und die Ausgabewahrscheinlichkeit eines Zeichens o_t zum Zeitpunkt t nur vom aktuellen Systemzustand abhängt:*

$$\mathrm{P}(\mathbb{O}_t = o_t \mid q_1,\ldots,q_t,o_1,\ldots,o_{t-1}) \;=\; \mathrm{P}(\mathbb{O}_t = o_t \mid \mathbb{U}_t = q_t)$$

Das statistische Verhalten eines diskreten HMM ist durch das Parameterfeld

$$\boldsymbol{\lambda} \stackrel{\text{def}}{=} (\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B})$$

charakterisiert:

$$\begin{aligned}\boldsymbol{\pi}\in\mathbb{R}^N &\quad\text{mit}\quad \pi_i = \mathrm{P}(\mathbb{U}_1 = s_i)\\ \boldsymbol{A}\in\mathbb{R}^{N\times N} &\quad\text{mit}\quad a_{ij} = \mathrm{P}(\mathbb{U}_t = s_j \mid \mathbb{U}_{t-1} = s_i)\\ \boldsymbol{B}\in\mathbb{R}^{N\times M} &\quad\text{mit}\quad b_{jk} = \mathrm{P}(\mathbb{O} = z_k \mid \mathbb{U}_t = s_j)\end{aligned}$$

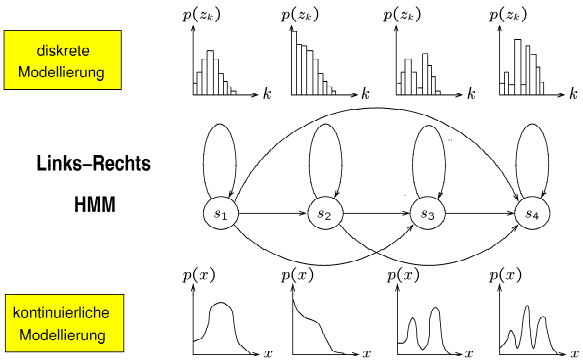
Definition 5.10
Es sei $\mathcal{S}$ eine Menge von N inneren Zuständen und $D\in\mathbb{N}$. Der zweistufige Zufallsprozeß $[\mathbb{U}_t,\mathbb{O}_t]$ über $\mathcal{S}$ und $\mathbb{R}^D$ heißt **kontinuierliches HMM**, *wenn $[\mathbb{U}_t]$ eine (einfache, stationäre) Markovkette über $\mathcal{S}$ ist und die Ausgabewahrscheinlichkeit eines Vektors $\boldsymbol{x}_t\in\mathbb{R}^D$ zum Zeitpunkt t nur vom aktuellen Systemzustand abhängt:*

$$\mathrm{f}_j(\boldsymbol{x}) \;=\; \mathrm{P}(\mathbb{X}_t = \boldsymbol{x} \mid \mathbb{U}_t = s_j)$$

Das statistische Verhalten eines diskreten HMM ist durch die Parameter und Dichtefunktionen

$$\boldsymbol{\lambda} \stackrel{\text{def}}{=} (\boldsymbol{\pi},\boldsymbol{A},[\mathrm{f}_j])$$

charakterisiert.



Drei offene Fragen zum Thema „HMM“:

- Berechnung der *Produktionswahrscheinlichkeit*

  $$\mathrm{P}(\boldsymbol{o}|\boldsymbol{\lambda}) \;=\; ? \qquad\text{bzw.}\qquad \mathrm{P}(\boldsymbol{X}|\boldsymbol{\lambda}) \;=\; ?$$

- Bestimmung der *wahrscheinlichsten Zustandsfolge*

  $$\mathrm{P}(\boldsymbol{q},\boldsymbol{X}\mid\boldsymbol{\lambda}) \quad \xrightarrow{\;\boldsymbol{q}\;} \quad \text{„}\textit{Maximum}\text{“}$$

- Schätzung der optimalen *Modellparameter*

  $$\tilde{\boldsymbol{\lambda}} \;=\; (\tilde{\boldsymbol{\pi}},\tilde{\boldsymbol{A}},\tilde{\boldsymbol{B}}) \quad\text{mit}\quad \mathrm{P}(\omega|\tilde{\boldsymbol{\lambda}}) \;=\; \max_{\boldsymbol{\lambda}}\mathrm{P}(\omega|\boldsymbol{\lambda})$$

Produktionswahrscheinlichkeiten

- GEGEBEN:
  die Parameter $\boldsymbol{\lambda}$ eines HMM
  eine Beobachtungsfolge $\boldsymbol{o}$ der Dauer T
- GESUCHT:
  die Wahrscheinlichkeit $\mathrm{P}(\boldsymbol{o}|\boldsymbol{\lambda})$, daß $\boldsymbol{o}$ von $\boldsymbol{\lambda}$ erzeugt wurde

Modellbedingte Verteilung der Zustandsfolgen

$$\mathrm{P}(\boldsymbol{q} \mid \boldsymbol{\lambda}) \;=\; \mathrm{P}(q_1 \ldots q_T \mid \boldsymbol{\lambda}) \;=\; \pi_{q_1} \cdot \prod_{t=2}^{T} a_{q_{t-1} q_t}$$

Zustandsbedingte Verteilung der Ausgabezeichenfolgen

$$\mathrm{P}(\boldsymbol{o} \mid \boldsymbol{q}, \boldsymbol{\lambda}) \;=\; \mathrm{P}(o_1 \ldots o_T \mid q_1 \ldots q_T, \boldsymbol{\lambda}) \;=\; \prod_{t=1}^{T} b_{q_t}(o_t)$$

Verbundwahrscheinlichkeit für $\boldsymbol{q}$ und $\boldsymbol{o}$

$$\mathrm{P}(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\lambda}) \;=\; \mathrm{P}(\boldsymbol{o} \mid \boldsymbol{q}, \boldsymbol{\lambda}) \cdot \mathrm{P}(\boldsymbol{q} \mid \boldsymbol{\lambda}) \;=\; \pi_{q_1} b_{q_1}(o_1) \cdot \prod_{t=2}^{T} a_{q_{t-1} q_t} b_{q_t}(o_t)$$

... und die gesuchte Produktionswahrscheinlichkeit

$$\mathrm{P}(\boldsymbol{o}|\boldsymbol{\lambda}) \;=\; \sum_{\boldsymbol{q} \in \mathcal{S}^T} \mathrm{P}(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\lambda}) \;=\; \sum_{\boldsymbol{q} \in \mathcal{S}^T} \pi_{q_1} b_{q_1}(o_1) \cdot \prod_{t=2}^{T} a_{q_{t-1} q_t} b_{q_t}(o_t)$$

⇨ ca. $2T \cdot N^T$ Multiplikationen (exponentielle Komplexität!)

VORWÄRTSALGORITHMUS

Definition 5.11
Sei $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ ein HMM mit N Zuständen und $\boldsymbol{o}$ eine Beobachtungsfolge der Dauer T. Die Zahlen

$$\alpha_t(j) \stackrel{\text{def}}{=} \mathrm{P}(o_1 \ldots o_t, q_t = j \mid \boldsymbol{\lambda})$$

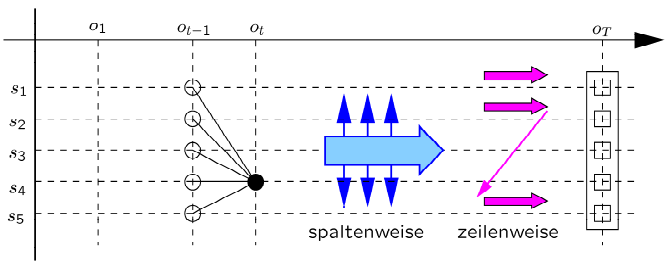
für $t = 1, \ldots, T$ und $j = 1, \ldots, N$ heißen **Vorwärtswahrscheinlichkeiten** *von $\boldsymbol{\lambda}$ für $\boldsymbol{o}$.*

Satz 5.5
Die Vorwärtswahrscheinlichkeiten eines HMM $\boldsymbol{\lambda}$ für eine Eingabefolge $\boldsymbol{o} = o_1 \ldots o_T$ berechnen sich nach dem **Vorwärtsalgorithmus**:

1. Initialisierung: *Für alle $j = 1, \ldots, N$ setze*
$$\alpha_1(j) \;=\; \pi_j \cdot b_j(o_1)$$
2. Rekursion: *Für $t > 1$ und alle $j = 1, \ldots, N$ setze*
$$\alpha_t(j) \;=\; \left(\sum_{i=1}^{N} \alpha_{t-1}(i) \cdot a_{ij} \right) \cdot b_j(o_t)$$
3. Terminierung: *Berechne die Endspaltensumme*
$$\mathrm{P}(\boldsymbol{o}|\boldsymbol{\lambda}) \;=\; \sum_{j=1}^{N} \alpha_T(j)$$

Rechenschema für den Vorwärtsalgorithmus

(Baum, Welch, Eagon 1967)



BEMERKUNGEN

- Rechenaufwand (Additionen/Multiplikationen)
$$\mathrm{O}(2T \cdot N^T) \quad \rightsquigarrow \quad \mathrm{O}(2N^2T)$$
- Für Links-Rechts-Modelle initialisieren wir
$$\alpha_1(j) \;=\; \begin{cases} 1 & j = 1 \\ 0 & j \neq 1 \end{cases}, \quad \beta_T(i) \;=\; \begin{cases} 1 & i = N \\ 0 & i \neq N \end{cases}$$
- Zeilenweise Berechnung bei Links-Rechts-HMMs;
- Aufwand bei Links-Rechts-HMMs reduziert:
$$\mathrm{O}(2n \cdot NT)$$

RÜCKWÄRTSALGORITHMUS

Definition 5.12
Sei $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ ein HMM mit N Zuständen und $\boldsymbol{o}$ eine Beobachtungsfolge der Dauer T. Die Zahlen

$$\beta_t(i) \stackrel{\text{def}}{=} \mathrm{P}(o_{t+1} \ldots o_T \mid q_t = i, \boldsymbol{\lambda})$$

für $t = 1, \ldots, T$ und $i = 1, \ldots, N$ heißen **Rückwärtswahrscheinlichkeiten** *von $\boldsymbol{\lambda}$ für $\boldsymbol{o}$.*

Satz 5.6
Die Rückwärtswahrscheinlichkeiten eines HMM $\boldsymbol{\lambda}$ für eine Eingabefolge $\boldsymbol{o} = o_1 \ldots o_T$ berechnen sich nach dem **Rückwärtsalgorithmus**:

1. Initialisierung: *Für alle $i = 1, \ldots, N$ setze*
$$\beta_T(i) \;=\; 1$$
2. Rekursion: *Für $t < T$ und alle $i = 1, \ldots, N$ setze*
$$\beta_t(i) \;=\; \sum_{j=1}^{N} a_{ij} \cdot b_j(o_{t+1}) \cdot \beta_{t+1}(j)$$
3. Terminierung: *Berechne die gewichtete Summe*
$$\mathrm{P}(\boldsymbol{o}|\boldsymbol{\lambda}) \;=\; \sum_{j=1}^{N} \pi_j \cdot b_j(o_1) \cdot \beta_1(j)$$

Aufdeckung der verborgenen Zustandsfolge

- GEGEBEN:
  die Parameter $\boldsymbol{\lambda}$ eines HMM
  eine Beobachtungsfolge $\boldsymbol{o}$ der Dauer T
- GESUCHT:
  1. die a posteriori wahrscheinlichsten HMM-Zustände $q_t^\star$
  2. die a posteriori wahrscheinlichste Zustandsfolge $\boldsymbol{q}^\star$

Satz 5.7
Sei $\boldsymbol{\lambda}$ ein HMM und $\boldsymbol{o} = o_1 \dots o_T$ eine Beobachtungsfolge. Die **a posteriori Zustandswahrscheinlichkeit**

$$\gamma_t(j) \stackrel{\text{def}}{=} \mathrm{P}(q_t = j \mid \boldsymbol{o}, \boldsymbol{\lambda})$$

für das Erreichen des Zustandes s_j zum Zeitpunkt t berechnet sich aus den Vor- und Rückwärtswahrscheinlichkeiten von $\boldsymbol{o}$ gemäß

$$\gamma_t(j) \;=\; \frac{\alpha_t(j) \cdot \beta_t(j)}{\sum_{i=1}^{N} \alpha_t(i) \cdot \beta_t(i)} \;.$$

Im Spezialfall $T = t$ erhalten wir mit

$$\mathrm{P}(q_t = j \mid o_1 \dots o_t, \boldsymbol{\lambda}) \;=\; \frac{\alpha_t(j)}{\sum_{i=1}^{N} \alpha_t(i)}$$

die **kausalen** *a posteriori Zustandswahrscheinlichkeiten.*

BEMERKUNGEN:

- Die Folge $\hat{\boldsymbol{q}}(\boldsymbol{o})$ von Zuständen

  $$\hat{q}_t(\boldsymbol{o}) \stackrel{\text{def}}{=} \operatorname*{argmax}_j \mathrm{P}(q_t = j \mid \boldsymbol{o}, \boldsymbol{\lambda})$$

  heißt **Maximum a posteriori** Zustandsfolge von $\boldsymbol{o}$.

- Die MAP-Zustandsfolge einer Beobachtung $\boldsymbol{o}$ kann unter Umständen die Wahrscheinlichkeit Null besitzen!

- Zur Berechnung der *kausalen* MAP-Zustände müssen wir nicht „in die Zukunft schauen".

BEISPIEL: **Kontextuelle Schriftzeichenerkennung**

- HMM-Zustände ⇨ Schriftzeichen
- HMM-Ausgabe ⇨ Merkmalvektoren $\boldsymbol{x}_t$ gescannter Zeichen
- Multivariate Ausgabeverteilungsdichte
  $b_j(\boldsymbol{x}) \;=\; \mathrm{P}(\text{„Zeichenmerkmale } \boldsymbol{x}\text{"} \mid \text{„Schriftzeichen } j\text{"})$
- Übergangswahrscheinlichkeiten („Zeichenpaarwahrscheinl.")
  $a_{ij} \;=\; \mathrm{P}(\text{„Zeichen } j\text{"} \mid \text{„letztes Zeichen war } i\text{"})$
  häufig: „au", „er", „sch", „ung"
  selten: „hh", „fv", „ea", ...
  niemals: „xx", „scp", „äö", ...

VITERBI-ALGORITHMUS

Definition 5.13
Sei $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ ein HMM mit N Zuständen und $\boldsymbol{o}$ eine Beobachtungsfolge der Dauer T. Die Zustandsfolge $\boldsymbol{q}^\star = q^\star{}_1 \dots q^\star{}_T$ heißt **Viterbi-optimal** *für $\boldsymbol{o}$, wenn sie unter allen Konkurrenten die maximale a posteriori Folgenwahrscheinlichkeit besitzt:*

$$\mathrm{P}(\boldsymbol{q}^\star \mid \boldsymbol{o}, \boldsymbol{\lambda}) \;=\; \max_{\boldsymbol{q} \in \mathcal{S}^T} \mathrm{P}(\boldsymbol{q} \mid \boldsymbol{o}, \boldsymbol{\lambda})$$

Wegen der Beziehung

$$\mathrm{P}(\boldsymbol{q} \mid \boldsymbol{o}, \boldsymbol{\lambda}) \;=\; \frac{\mathrm{P}(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\lambda})}{\mathrm{P}(\boldsymbol{o} \mid \boldsymbol{\lambda})}$$

besitzt eine Viterbi-optimale Zustandsfolge auch die maximale Verbundwahrscheinlichkeit. Man bezeichnet

$$\mathrm{P}^\star(\boldsymbol{o} \mid \boldsymbol{\lambda}) \stackrel{\text{def}}{=} \max_{\boldsymbol{q} \in \mathcal{S}^T} \mathrm{P}(\boldsymbol{o}, \boldsymbol{q} \mid \boldsymbol{\lambda})$$

als **Viterbi-Bewertung** von $\boldsymbol{o}$. Es ist $\mathrm{P}^\star(\boldsymbol{o} \mid \boldsymbol{\lambda})$ der größte Summand in $\mathrm{P}(\boldsymbol{o} \mid \boldsymbol{\lambda})$.

Definition 5.14
Sei $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ ein HMM mit N Zuständen und $\boldsymbol{o}$ eine Beobachtungsfolge der Dauer T. Die Zahlen

$$\vartheta_t(j) \stackrel{\text{def}}{=} \max_{\boldsymbol{q} \in \mathcal{S}^{t-1} \times \{j\}} \mathrm{P}(o_1 \dots o_t, q_1 \dots q_t \mid \boldsymbol{\lambda})$$

heißen **partielle Viterbi-Bewertungen** *von $\boldsymbol{o}$.*

Offensichtlich gilt die Aussage

$$\mathrm{P}^\star(\boldsymbol{o} \mid \boldsymbol{\lambda}) \;=\; \max_j \vartheta_T(j)$$

Satz 5.8 (Viterbi 1967)
Die partiellen Viterbi-Bewertungen eines HMM $\boldsymbol{\lambda}$ für eine Eingabefolge $\boldsymbol{o} = o_1 \dots o_T$ berechnen sich nach dem **Viterbi-Algorithmus***:*

1 Initialisierung: *Für alle $j = 1, \dots, N$ setze*

$$\vartheta_1(j) \;=\; \pi_j \cdot b_j(o_1) \qquad \psi_1(j) \;=\; 0$$

2 Rekursion: *Für $t > 1$ und alle $j = 1, \dots, N$ setze*

$$\vartheta_t(j) \;=\; \Big(\max_i \vartheta_{t-1}(i) \cdot a_{ij}\Big) \cdot b_j(o_t)$$
$$\psi_t(j) \;=\; \operatorname*{argmax}_i \vartheta_{t-1}(i) \cdot a_{ij}$$

3 Terminierung: *Berechne*

$$\mathrm{P}^\star(\boldsymbol{o} \mid \boldsymbol{\lambda}) \;=\; \max_j \vartheta_T(j) \qquad q_T^\star \;=\; \operatorname*{argmax}_j \vartheta_T(j)$$

4 Rückverfolgung:
Rekonstruiere eine optimale Zustandsfolge mittels

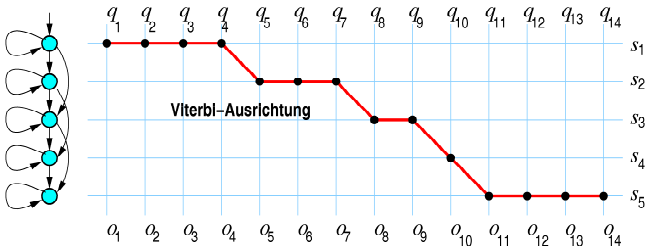
$$q_t^\star \;=\; \psi_{t+1}(q_{t+1}^\star) \;, \qquad t = T-1, \dots, 1$$

Insbesondere ist obiges $q^\star{}_1 \dots q^\star{}_T$ eine Viterbi-optimale Zustandsfolge für $\boldsymbol{o}$.

BEISPIEL:

$$\boldsymbol{\vartheta} = \begin{pmatrix} \vartheta_1(1) & \vartheta_2(1) & \vartheta_3(1) & \vartheta_4(1) \\ \vartheta_1(2) & \vartheta_2(2) & \vartheta_3(2) & \vartheta_4(2) \\ \vartheta_1(3) & \vartheta_2(3) & \vartheta_3(3) & \vartheta_4(3) \\ \vartheta_1(4) & \vartheta_2(4) & \vartheta_3(4) & \vartheta_4(4) \end{pmatrix} \qquad \boldsymbol{\psi} = \begin{pmatrix} ? & \leftarrow & \leftarrow & \leftarrow \\ ? & \nwarrow & \nwarrow & \leftarrow \\ ? & \nwarrow & \leftarrow & \nwarrow \\ ? & \nwarrow & \nwarrow & \leftarrow \end{pmatrix}$$

Entscheidungsüberwachtes Lernen
der HMM-Parameter



Optimiere $\boldsymbol{\lambda}$ bzgl. $\boldsymbol{o}$ durch schrittweise Verbesserung von

$$\mathrm{P}^\star(\boldsymbol{o}|\boldsymbol{\lambda}) \;=\; \max_{\boldsymbol{q}\in\mathcal{S}^T} \mathrm{P}(\boldsymbol{o},\boldsymbol{q}\mid\boldsymbol{\lambda})$$

Entscheidungsüberwachter Iterationsschritt:

- FINDE VITERBI-OPTIMALE ZUSTANDSFOLGE $\boldsymbol{q}^\star$
- BERECHNE NEUE ML-PARAMETER
  bezüglich der **etikettierten Stichprobe**

$$(\boldsymbol{o},\boldsymbol{q}^\star) \stackrel{\text{def}}{=} \{\langle o_t, q_t^\star\rangle \mid t=1,\ldots,T\}$$

Dann gilt die Abschätzung

$$\begin{aligned}\mathrm{P}^\star(\boldsymbol{o}|\boldsymbol{\lambda}^{(\nu)}) &= \mathrm{P}(\boldsymbol{o},\boldsymbol{q}^\star\mid\boldsymbol{\lambda}^{(\nu)})\\ &\le \mathrm{P}(\boldsymbol{o},\boldsymbol{q}^\star\mid\boldsymbol{\lambda}^{(\nu+1)})\\ &\le \mathrm{P}^\star(\boldsymbol{o}|\boldsymbol{\lambda}^{(\nu+1)})\end{aligned}$$

Viterbi-Training (für diskrete HMMs)

1 *Initialisierung:*
Wähle geeignetes Startmodell $\boldsymbol{\lambda}^{[0]}$ und setze $\nu=1$.

2 *Viterbi-Etikettierung:*
Bestimme zu allen $\boldsymbol{o}\in\omega$ eine optimale Zustandsfolge

$$\boldsymbol{q}^\star(\boldsymbol{o}) \stackrel{\text{def}}{=} \max_{\boldsymbol{q}\in\mathcal{S}^T} \mathrm{P}(\boldsymbol{o},\boldsymbol{q}\mid\boldsymbol{\lambda}^{(\nu)})$$

mit dem Viterbi-Algorithmus.

3 *Akkumulierung der Statistiken:*
Inkrementiere für jedes $\langle\boldsymbol{o},\boldsymbol{q}^\star\rangle$ gemäß

$$\begin{aligned}\bar\pi_i &\;+\!=\; \text{„Ist } q_1^\star = s_i \text{ ?“}\\ \bar a_{ij} &\;+\!=\; \text{„Anzahl der Übergänge } s_i \rightsquigarrow s_j\text{“}\\ \bar b_{jk} &\;+\!=\; \text{„Anzahl der Ausgaben } z_k \text{ in Zustand } s_j\text{“}\end{aligned}$$

4 *Normierung der Statistiken:*

$$\hat\pi_i = \frac{\bar\pi_i}{\sum_{i=1}^N \bar\pi_i}\,,\quad \hat a_{ij} = \frac{\bar a_{ij}}{\sum_{j=1}^N \bar a_{ij}}\,,\quad \hat b_{jk} = \frac{\bar b_{jk}}{\sum_{k=1}^M \bar b_{jk}}$$

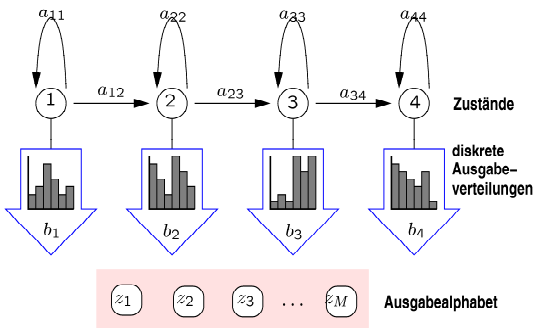
und setze $\boldsymbol{\lambda}^{(\nu+1)} = (\hat{\boldsymbol{\pi}},\hat{\boldsymbol{A}},\hat{\boldsymbol{B}})$.

5 *Abbruchbedingung:*
ENDE oder $\nu \Leftarrow \nu+1$ und weiter bei $\rightsquigarrow$ $\boxed{2}$

- Der Algorithmus findet ein lokales Maximum der Viterbi-Bewertung von ω.
- Als Startparameter lassen sich *nicht* die Gleichverteilungen verwenden!
- Für kontinuierliche HMMs gibt es einen entsprechenden Algorithmus.

Parameterschätzung für diskrete HMMs

- GEGEBEN:
  die Struktur (N Zustände, M Symbole) eines HMM
  eine Stichprobe ω von Beobachtungsfolgen $\boldsymbol{o}\in\mathcal{A}^\star$
- GESUCHT:
  das bestpassende Parameterfeld $\boldsymbol{\lambda}=(\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B})$



Maximiere die **Maximum-Likelihood-Zielgröße**

$$\ell_\lambda(\omega) \;=\; \log \mathrm{P}(\omega|\boldsymbol{\lambda}) \;=\; \sum_{\boldsymbol{o}\in\omega} \log \mathrm{P}(\boldsymbol{o}|\boldsymbol{\lambda})$$

auf der (konvexen) **Mannigfaltigkeit**

$$\mathcal{M}_\lambda \;=\; \left\{(\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B}) \mid \pi_i, a_{ij}, b_{jk} \ge 0 \text{ und } \sum_i \pi_i = \sum_j a_{ij} = \sum_k b_{jk} = 1\right\}$$

EM-Algorithmus

Die iterative Verbesserung der Parameter $\boldsymbol{\lambda}^{(\nu)}$ eines HMM — beginnend bei den Startwerten $\boldsymbol{\lambda}^{(0)}$ — unter Verwendung des Satzes 5.9 garantiert ein lokales Maximum der ML-Zielgröße.

Satz 5.9 (Baum, Welch 1966)
Gegeben ist ein HMM mit N Zuständen, M Ausgabesymbolen und den Modellparametern $\boldsymbol{\lambda}$ sowie eine beobachtete Zeichenfolge $\boldsymbol{o}=o_1\ldots o_T$. Es bezeichnen ferner $\alpha_t(j)$, $\beta_t(i)$ die Vorwärts- und Rückwärtswahrscheinlichkeiten,

$$\gamma_t(j) \stackrel{\text{def}}{=} \mathrm{P}(q_t=j\mid\boldsymbol{o},\boldsymbol{\lambda}) \;=\; \frac{\alpha_t(j)\cdot\beta_t(j)}{\sum_{i=1}^N \alpha_t(i)\cdot\beta_t(i)}$$

die **a posteriori Zustandswahrscheinlichkeiten** *und*

$$\xi_t(i,j) \stackrel{\text{def}}{=} \mathrm{P}(q_t=i, q_{t+1}=j\mid\boldsymbol{o},\boldsymbol{\lambda}) \;=\; \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^N \alpha_t(i)\cdot\beta_t(i)}$$

die **a posteriori Übergangswahrscheinlichkeiten**.

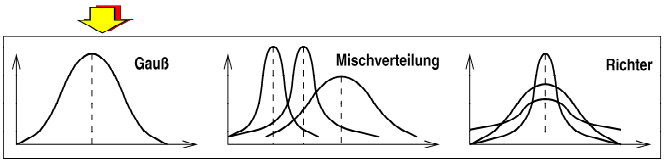
Dann gilt für die neugeschätzten Parameter $\hat{\boldsymbol{\lambda}}=(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{A}},\hat{\boldsymbol{B}})$ mit

$$\begin{aligned}\hat\pi_i &= \gamma_1(i)\\ \hat a_{ij} &= \sum_{t=1}^{T-1}\xi_t(i,j) \Big/ \sum_{t=1}^{T-1}\gamma_t(i)\\ \hat b_{jk} &= \sum_{t=1}^{T-1}\gamma_t(j)\cdot\mathbf{I}_{o_t,z_k} \Big/ \sum_{t=1}^{T-1}\gamma_t(j)\end{aligned}$$

die Ungleichung

$$\ell_{\hat\lambda}(\boldsymbol{o}) \;\ge\; \ell_{\boldsymbol{\lambda}}(\boldsymbol{o})$$

Parameterschätzung für kontinuierliche HMMs



Die HMM-Ausgabe ist **normalverteilt**

$$b_j(\boldsymbol{x}) \;=\; \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_j, \boldsymbol{S}_j)$$

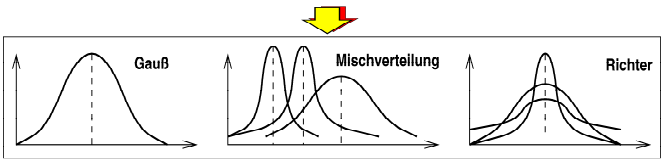
Die Zerlegung der Kullback-Leibler-Statistik ergibt

$$\mathrm{Q}(\boldsymbol{\lambda}, \{\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{S}}_j\}) \;=\; \sum_{t=1}^{T} \gamma_t(j) \cdot \log \mathcal{N}(\boldsymbol{x}_t \mid \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{S}}_j)$$

Schätzformeln f.d. **Normalverteilungsparameter** von s_j:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_j &= \sum_{t=1}^{T} \gamma_t(j) \cdot \boldsymbol{x}_t \Big/ \sum_{t=1}^{T} \gamma_t(j) \\
\hat{\boldsymbol{S}}_j &= \sum_{t=1}^{T} \gamma_t(j) \cdot (\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_j)(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_j)^\top \Big/ \sum_{t=1}^{T} \gamma_t(j) \\
&= \sum_{t=1}^{T} \gamma_t(j) \cdot \boldsymbol{x}_t \boldsymbol{x}_t^\top \Big/ \sum_{t=1}^{T} \gamma_t(j) \;-\; \hat{\boldsymbol{\mu}}_j \hat{\boldsymbol{\mu}}_j^\top
\end{aligned}$$

Mischverteilte HMMs



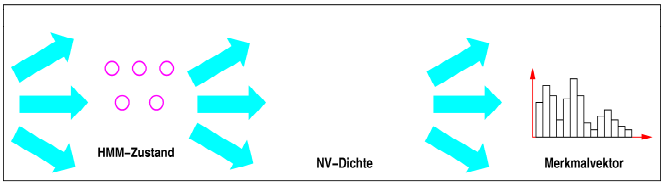
Die HMM-Ausgabe ist **gemischt normalverteilt**

$$b_j(\boldsymbol{x}) \;=\; \sum_{\kappa=1}^{K} c_{j\kappa} \cdot \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_{j\kappa}, \boldsymbol{S}_{j\kappa}) \;, \qquad \sum_{\kappa=1}^{K} c_{j\kappa} = 1$$

DREISTUFIGER Zufallsprozeß

$$\langle \mathbb{U}_t, \mathbb{K}_t, \mathbb{X}_t \rangle \;\rightsquigarrow\; \langle q_t, \kappa_t, \boldsymbol{x}_t \rangle \;\in\; \mathcal{S} \times \mathcal{K} \times \mathbb{R}^D$$

1. **Modellzustand** $q_t \in \mathcal{S}$ *(verborgen)*
2. **Mischungskomponente** $\kappa_t \in \mathcal{K}$ *(verborgen)*
3. **Ausgabevektor** $\boldsymbol{x}_t \in \mathbb{R}^D$ *(beobachtbar)*



Verbundverteilungsdichte des HMM

$$\mathrm{P}(\boldsymbol{X}, \boldsymbol{q}, \boldsymbol{\kappa} \mid \boldsymbol{\lambda}) \;=\; \pi_{q_1} \cdot \prod_{t=2}^{T} a_{q_{t-1} q_t} \cdot \prod_{t=1}^{T} c_{q_t \kappa_t} \mathsf{f}_{q_t \kappa_t}(\boldsymbol{x}_t)$$

Produktionswahrscheinlichkeiten in „doppelt marginaler“ Form

$$\mathrm{P}(\boldsymbol{X} | \boldsymbol{\lambda}) \;=\; \sum_{\boldsymbol{q} \in \mathcal{S}^T} \sum_{\boldsymbol{\kappa} \in \mathcal{K}^T} \mathrm{P}(\boldsymbol{X}, \boldsymbol{q}, \boldsymbol{\kappa} \mid \boldsymbol{\lambda})$$

Definition 5.15 *Ist $\boldsymbol{\lambda}$ ein kontinuierliches HMM über $\mathbb{R}^D$ mit N Zuständen und K Mischungskomponenten $\mathsf{f}_{j\cdot}$ je Zustand, und ist $\boldsymbol{X}$ eine Vektorfolge der Dauer T, so bezeichnen wir*

$$\zeta_t(j,k) \stackrel{\text{def}}{=} \mathrm{P}(q_t = j, \kappa_t = k \mid \boldsymbol{X}, \boldsymbol{\lambda})$$

als **Selektionswahrscheinlichkeit** *der Komponente $\mathsf{f}_{j\kappa}$ im Zustand s_j bei der Erzeugung der Ausgabe $\boldsymbol{X}$.*

Satz 5.10
Die oben genannten Selektionswahrscheinlichkeiten von $\boldsymbol{X}$ in $\boldsymbol{\lambda}$ berechnen sich gemäß

$$\begin{aligned}
\zeta_t(j,k) &= \frac{1}{\mathrm{P}(\boldsymbol{X}|\boldsymbol{\lambda})} \cdot \frac{\alpha_t(j) \cdot c_{jk} \cdot \mathsf{f}_{jk}(\boldsymbol{x}_t) \cdot \beta_t(j)}{b_j(\boldsymbol{x}_t)} \\
&= \gamma_t(j) \cdot \frac{c_{jk} \cdot \mathsf{f}_{jk}(\boldsymbol{x}_t)}{\sum_{k=1}^{K} c_{jk} \cdot \mathsf{f}_{jk}(\boldsymbol{x}_t)}
\end{aligned}$$

Satz 5.11 (Juang & Rabiner, 1985)
Es sei ein Mischverteilungs-HMM mit normalverteilten Dichtekomponenten und Parametern $\boldsymbol{\lambda}$ gegeben sowie eine Vektorfolge $\boldsymbol{X}$ der Dauer T. Die neugeschätzten Parameter

$$\hat{\boldsymbol{\lambda}} \;=\; (\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{A}}, \{\hat{c}_j, \hat{\boldsymbol{\mu}}_{jk}, \hat{\boldsymbol{S}}_{jk}\})$$

mit den Werten

$$\begin{aligned}
\hat{\pi}_i &= \gamma_1(i) \\
\hat{a}_{ij} &= \sum_{t=1}^{T-1} \xi_t(i,j) \Big/ \sum_{t=1}^{T-1} \gamma_t(i) \\
\hat{c}_{jk} &= \sum_{t=1}^{T} \zeta_t(j,k) \Big/ \sum_{t=1}^{T} \gamma_t(j) \\
\hat{\boldsymbol{\mu}}_{jk} &= \sum_{t=1}^{T} \zeta_t(j,k) \cdot \boldsymbol{x}_t \Big/ \sum_{t=1}^{T} \zeta_t(j,k) \\
\hat{\boldsymbol{S}}_{jk} &= \sum_{t=1}^{T} \zeta_t(j,k) \cdot \boldsymbol{x}_t \boldsymbol{x}_t^\top \Big/ \sum_{t=1}^{T} \zeta_t(j,k) \;-\; \hat{\boldsymbol{\mu}}_{jk} \hat{\boldsymbol{\mu}}_{jk}^\top
\end{aligned}$$

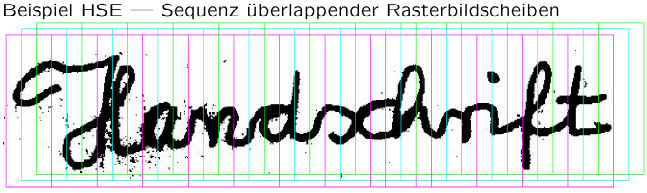
maximieren die Kullback-Leibler-Statistik $\mathrm{Q}(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$, d.h., es gilt insbesondere $\ell_{\hat{\lambda}}(\boldsymbol{o}) \;\geq\; \ell_{\lambda}(\boldsymbol{o})$.

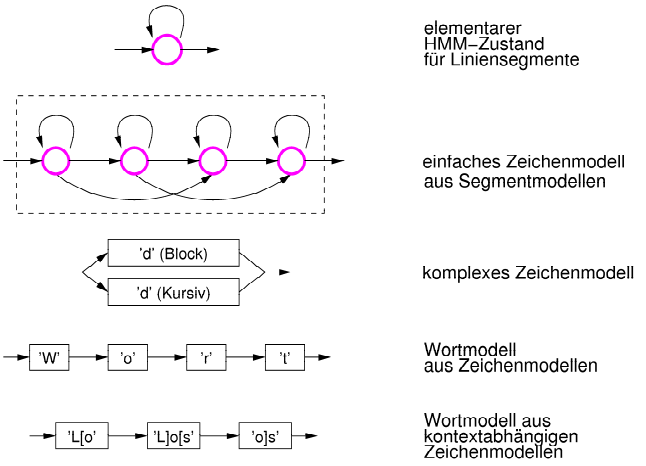
Zustandsweises Clustern — *„segmental K-means“*

- Lernen eines Vektorquantisierers
- Lernen eines diskreten HMMs auf den VQ-Daten
- Viterbi-Etikettierung der Daten nach Zuständen
- je Zustand alle zugehörigen Vektoren clustern
- Dichtekomponenten des Mischverteilungs-HMM initialisieren

Basiseinheiten

Beispiel HSE — Sequenz überlappender Rasterbildscheiben





Definition 5.16
Es sei λ ein Hidden Markov Modell mit N Zuständen.
Ein Zustand, der keine Ausgabe produziert heißt **konfluenter** *Zustand.*
Ein konfluenter Zustand s_i mit $\pi_i = 1$ heißt **Anfangszustand** *des Modells.*
Ein konfluenter Zustand s_j mit $a_{jj} = 1$ heißt **Endzustand** *des Modells.*
Ein Modell mit genau einem Anfangszustand und einem Endzustand heißt **E/A-Modell**.

O.B.d.A. seien immer s_1 und s_N Anfangs- und Endzustand.

Definition 5.17
Es seien λ', λ'' zwei E/A-Modelle mit N' bzw. N'' Zuständen.
Das Vereinigungsmodell $\lambda' \circ \lambda''$ mit dem zusätzlichen Übergang

$$s'_{N'} \mapsto s''_1 \quad \text{mit} \quad \mathrm{P}(q_t = s''_1 \mid q_{t-1} = s'_{N'}) \;=\; 1$$

ist ein E/A-Modell mit Anfangs/Endezustand s'_1 bzw. $s''_{N''}$ und heißt **sequentielle Verschaltung** *von λ' und λ''.*

Das Vereinigungsmodell $\lambda' \parallel \lambda''$ mit dem neuen Anfangs- und Endezustand s_α bzw. s_β mit

$$\mathrm{P}(q_t = s'_1 \mid q_{t-1} = s_\alpha) \;=\; \mathrm{P}(q_t = s''_1 \mid q_{t-1} = s_\alpha) \;=\; 1/2$$

und

$$\mathrm{P}(q_t = s_\beta \mid q_{t-1} = s'_{N'}) \;=\; \mathrm{P}(q_t = s_\beta \mid q_{t-1} = s''_{N''}) \;=\; 1$$

ist ein E/A-Modell und heißt **parallele Verschaltung** *von λ' und λ''.*

Überwachte Segmentierung mit HMMs

GEGEBEN:

- Longitudinalmuster $o = o_1 \ldots o_T$
- Sequentielle Beschreibung $\kappa = \kappa_1 \ldots \kappa_S \in \{1, \ldots, K\}^S$

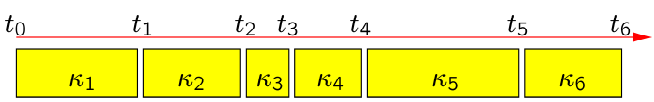
GESUCHT:

- Segmentierung von o in Abschnitte

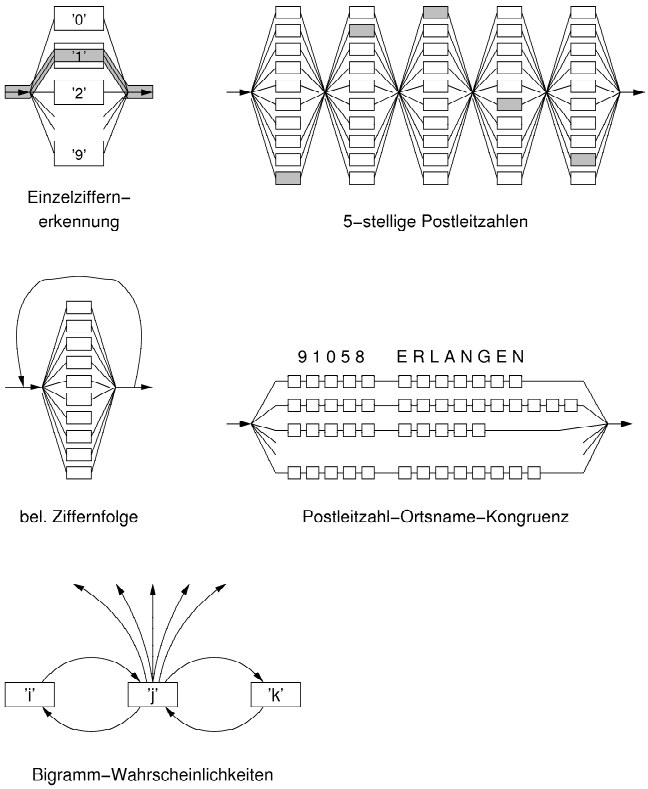
$$\kappa_s \;\mapsto\; [o_{t_{s-1}+1}, \ldots, o_{t_s}]$$

mit den Segmentgrenzen

$$0 = t_0 \le t_1 \le \ldots \le t_{S-1} \le t_S = T$$



Zeitliche Segmentierung des Eingabemusters

Lösungsansätze

- Segmente ⇔ Zustände
  Viterbi-Segmentierung; MAP-Zustand-Segmentierung
- Segmente ⇔ E/A-HMMs
  Viterbi-Annotation; konfluente MAP-Annotation