Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Sprachverarbeitung II / 7 FS 2017

DDHMM: Trellis-Diagramm, Forward-Algorithmus

Buch: Kapitel 5.1 bis 5.4.2

Beat Pfister

Programm heute:

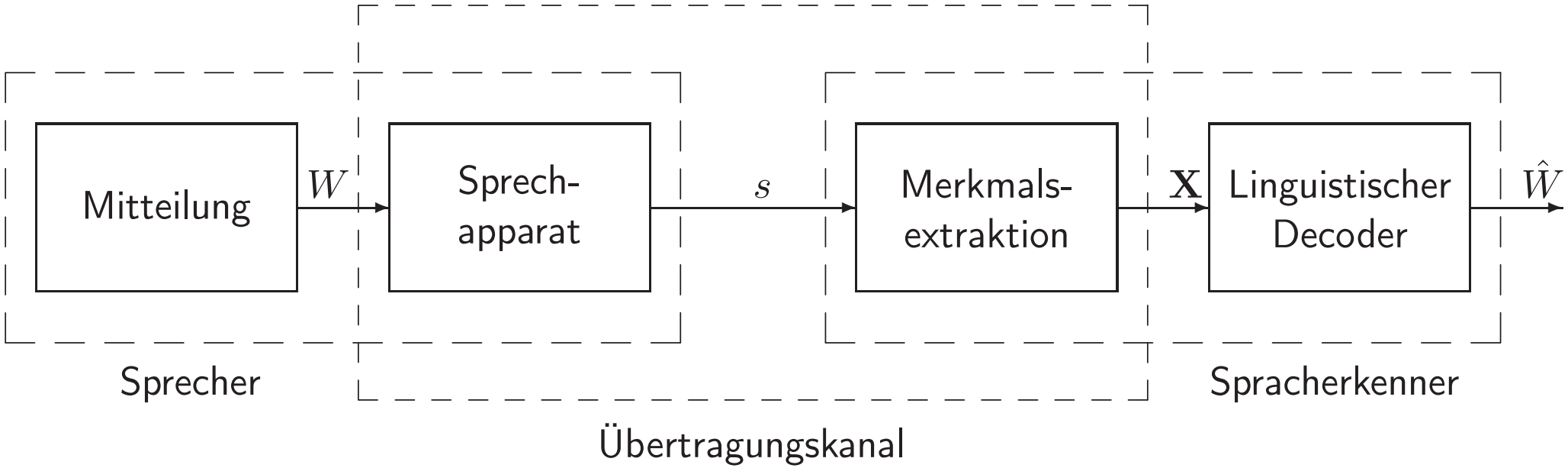
Vorlesung:

- Statistischer Ansatz (Repetition)
- diskrete vs. kontinuierliche Merkmale
- Hidden-Markov-Modelle (HMM)
- grundlegende HMM-Probleme
- Trellis-Diagramm
- Evaluationsproblem (Forward-Algorithmus)

Übung:

- ⋆ Experimente mit diskreten HMM

Informationstheoretische Sicht der Spracherkennung



Decodierungsproblem

Aufgabe des linguistischen Decoders

Gegeben: Merkmalssequenz $\mathbf{X} = \mathbf{x}_1 \mathbf{x}_2 \ldots \mathbf{x}_T$

Gesucht: Möglichst gute Schätzung $\hat{W}$ der geäusserten Wortfolge

Statistische Sicht

Bestimmen der optimalen Wortfolge $\hat{W} = w_1 w_2 \ldots w_K$ so,
dass Wahrscheinlichkeit eines Fehlentscheides minimal

Maximum-a-posteriori-Regel (MAP-Regel)

Wahrscheinlichkeit für Fehlentscheid ist minimal,
wenn die Wortfolge mit der höchsten
A-posteriori-Wahrscheinlichkeit $P(W|\mathbf{X})$ gewählt wird

$$\longrightarrow \quad \hat{W} = \underset{W}{\operatorname{argmax}} \; P(W|\mathbf{X})$$

Maximum-a-posteriori-Regel (MAP-Regel)

Wahrscheinlichkeit für Fehlentscheid ist minimal,
wenn die Wortfolge mit der höchsten
A-posteriori-Wahrscheinlichkeit $P(W|\mathbf{X})$ gewählt wird

$$\longrightarrow \quad \hat{W} = \underset{W}{\operatorname{argmax}} \; P(W|\mathbf{X})$$

Problem: $P(W|\mathbf{X})$ ist praktisch nicht ermittelbar!

MAP-Regel

Original: $\hat{W} = \underset{W}{\text{argmax}}\ P(W|\mathbf{X})$

mit: $P(A,B) = P(A|B)\,P(B) = P(B|A)\,P(A)$

$$\Rightarrow P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad \text{(Satz von Bayes)}$$

äquivalent: $$\hat{W} = \underset{W}{\text{argmax}}\ \frac{P(\mathbf{X}|W)\,P(W)}{P(\mathbf{X})} = \underset{W}{\text{argmax}}\ P(\mathbf{X}|W)\,P(W)$$

$P(\mathbf{X}|W)$: akustisches Modell

$P(W)$: A-priori-Wissen (Sprachmodell)

Sprachmodell $P(W)$

- Gibt Auskunft über die Wahrscheinlichkeit (relative Häufigkeit) von W
- Hilft insbes. bei akustisch nicht oder schlecht unterscheidbaren Wortfolgen:

  $$P(\text{“Gestern fiel viel Schnee.”}) \gg P(\text{“Gestern viel fiel Schnee.”})$$

  (Behandlung in Lektion 11; Buch Kapitel 14)

Akustisches Modell

Tatsache: Jede Äusserung von W ergibt ein etwas anderes Sprachsignal

Folge: Extrahierte Merkmalssequenz $\mathbf{X} = \mathbf{x}_1 \mathbf{x}_2 \ldots \mathbf{x}_T$ variiert auch

Merke: Sowohl die $\mathbf{x}_t$ als auch T variieren!

Nötig: Akustisches Modell $P(\mathbf{X}|W)$ muss

- sowohl die Variation des Merkmals $\mathbf{x}_t$ >>>
- als auch die zeitliche Variation von $\mathbf{X}$ beschreiben >>>

Akustisches Modell

Tatsache: Jede Äusserung von W ergibt ein etwas anderes Sprachsignal

Folge: Extrahierte Merkmalssequenz $\mathbf{X} = \mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_T$ variiert auch

Merke: Sowohl die $\mathbf{x}_t$ als auch T variieren!

Nötig: Akustisches Modell $P(\mathbf{X}|W)$ muss

- sowohl die Variation des Merkmals $\mathbf{x}_t$ >>>
- als auch die zeitliche Variation von $\mathbf{X}$ beschreiben >>>

Ansatz: HMM λ_W als Schätzung von $P(\mathbf{X}|W)$ verwenden

MAP-Erkenner mit HMM

Gegeben: ein HMM λ_W für jede Wortfolge W

MAP-Erkenner: $\hat{W} = \underset{W}{\operatorname{argmax}}\ P(\mathbf{X}|\lambda_W)\ P(W)$ >>>

MAP-Erkenner mit HMM

Gegeben: ein HMM λ_W für jede Wortfolge W

MAP-Erkenner: $\hat{W} = \underset{W}{\operatorname{argmax}}\ P(\mathbf{X}|\lambda_W)\ P(W)$ >>>

Frage: Wie kann $P(\mathbf{X}|\lambda_W)$ ermittelt werden ?

MAP-Erkenner mit HMM

Gegeben: ein HMM λ_W für jede Wortfolge W

MAP-Erkenner: $\hat{W} = \underset{W}{\operatorname{argmax}}\ P(\mathbf{X}|\lambda_W)\ P(W)$ >>>

Frage: Wie kann $P(\mathbf{X}|\lambda_W)$ ermittelt werden ?

$\longrightarrow$ Evaluationsproblem

Die grundlegenden HMM-Probleme

Schätzproblem:

Gegeben: Beobachtungssequenz $\mathbf{X} = \mathbf{x}_1\mathbf{x}_2 \ldots \mathbf{x}_T$
Gesucht: HMM $\lambda = (A, B)$, mit $\rightarrow P(\mathbf{X}|\lambda) =$ maximal
Lösung: Baum-Welch-Algorithmus

Evaluationsproblem:

Gegeben: HMM λ, Beobachtungssequenz $\mathbf{X} = \mathbf{x}_1\mathbf{x}_2 \ldots \mathbf{x}_T$
Gesucht: Produktionswahrscheinlichkeit $P(\mathbf{X}|\lambda)$
Lösung: Forward-Algorithmus

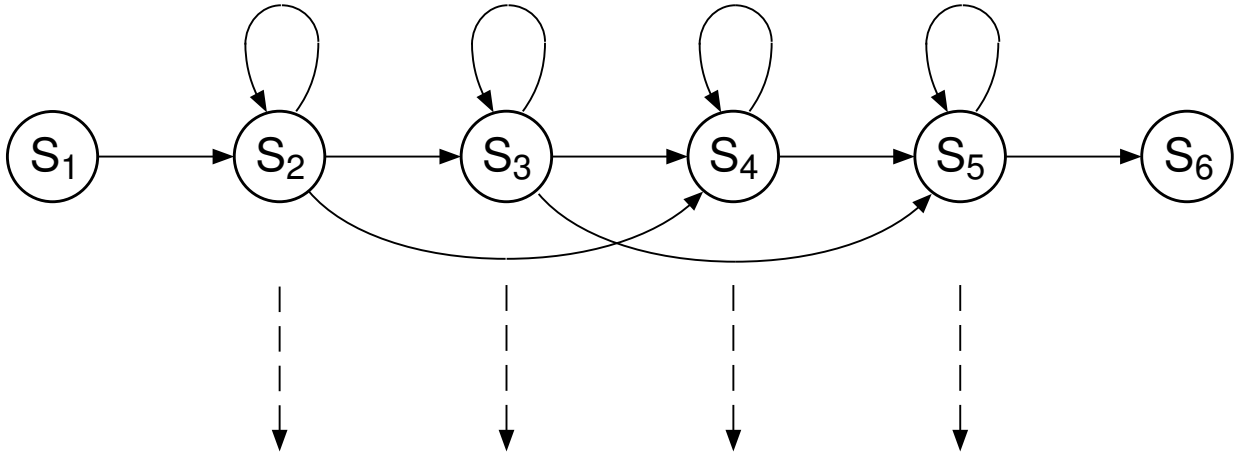
Decodierungsproblem:

Gegeben: HMM λ, Beobachtungssequenz $\mathbf{X} = \mathbf{x}_1\mathbf{x}_2 \ldots \mathbf{x}_T$
Gesucht: optimale Zustandssequenz $Q = S_1 q_1 \ldots q_T S_N$
Lösung: Viterbi-Algorithmus

Zusammenhang zwischen Zustand und Beobachtung

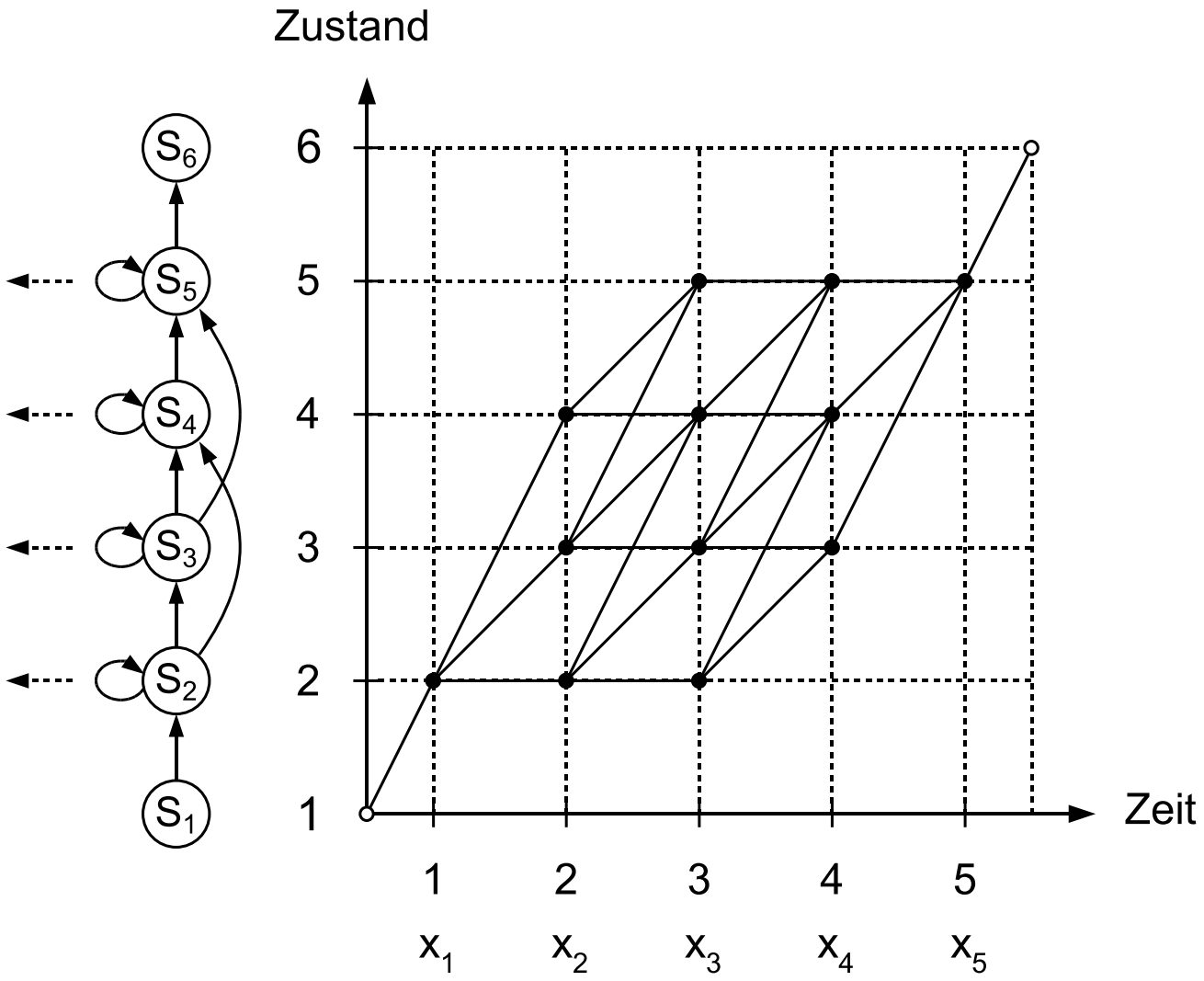
Grundsatz: Zustand des HMM aus Beobachtungssequenz nicht ermittelbar !

Gegeben:
- Beobachtungssequenz der Länge 5: $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_5$
- HMM mit 6 Zuständen:

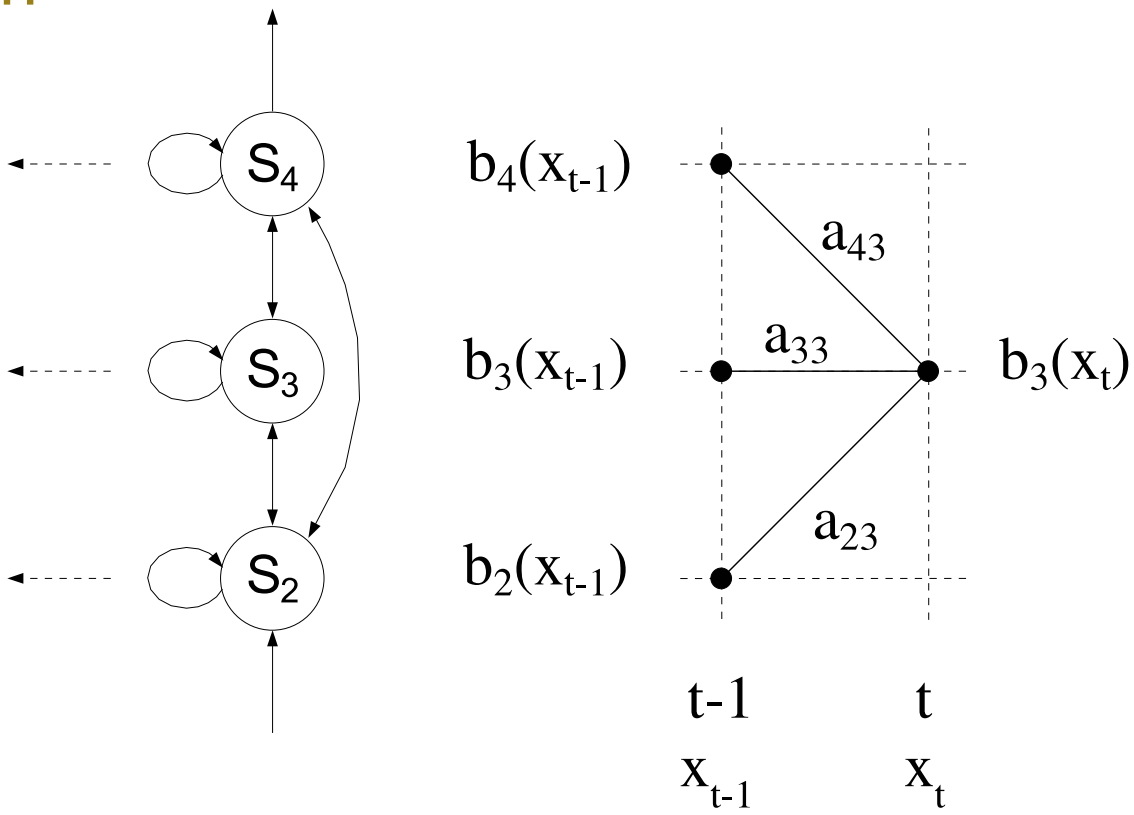


Frage: In welchen Zuständen kann sich das Modell zu einem bestimmten Zeitpunkt befinden ? $\longrightarrow$ Trellis-Diagramm

Trellis-Diagramm



Trellis-Diagramm



Kanten: Zustandsübergangswahrscheinlichkeiten

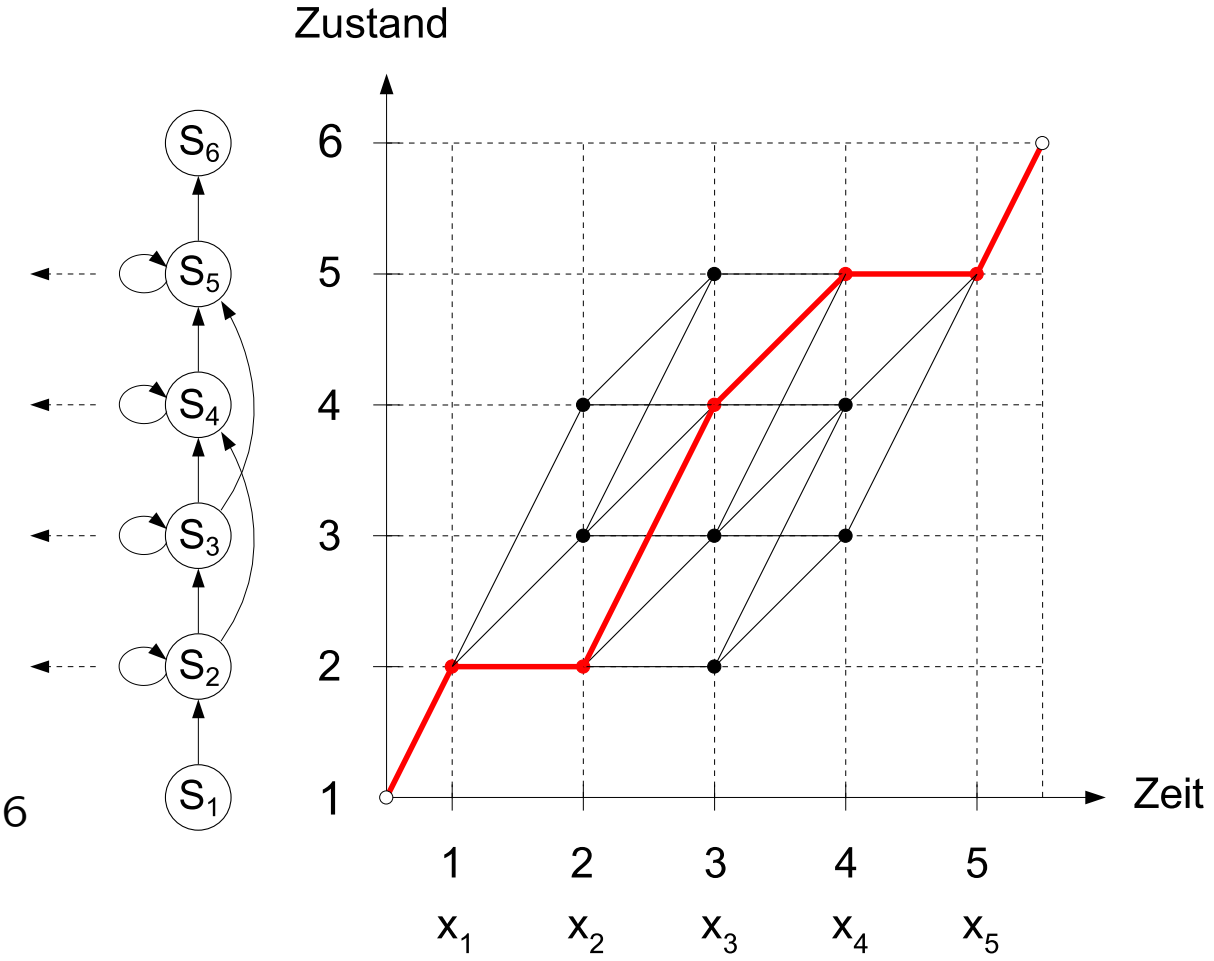
Knoten: Beobachtungswahrscheinlichkeiten

Beispiel 1

Gegeben: λ, Q

Gesucht: $P(Q|\lambda)$

(Wahrscheinlichkeit, dass das HMM λ die angegebene Zustandssequenz Q durchläuft)



$$P(Q|\lambda) = a_{12}\, a_{22}\, a_{24}\, a_{45}\, a_{55}\, a_{56}$$

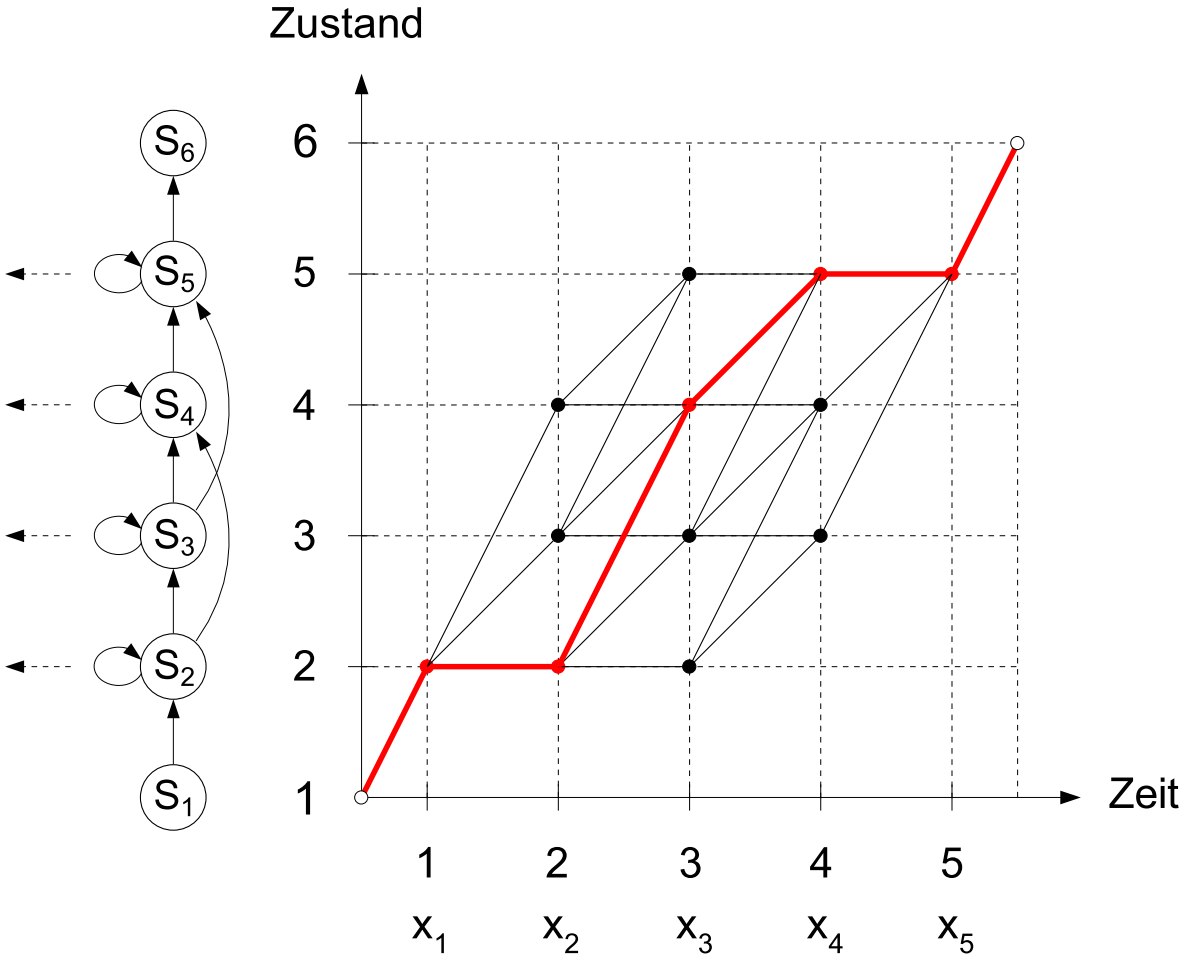
Merke: $\sum_Q P(Q|\lambda) = 1$

Beispiel 2

Gegeben: $\mathbf{X}$, λ, Q

Gesucht: $P(\mathbf{X}, Q|\lambda)$

(Wahrscheinlichkeit, dass das HMM λ die angegebene Zustandssequenz Q durchläuft und dabei die Beobachtungs-sequenz $\mathbf{X}$ erzeugt)

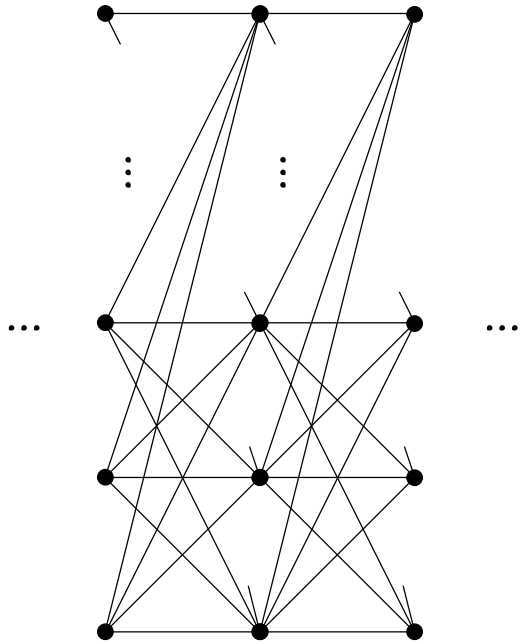


$$P(\mathbf{X}, Q|\lambda) = P(\mathbf{X}|Q, \lambda)\, P(Q|\lambda) = \ldots$$

Evaluationsproblem

Gegeben: $\lambda = (A, B), \quad \mathbf{X} = \mathbf{x}_1 \mathbf{x}_2 \ldots \mathbf{x}_T$

Gesucht: $P(\mathbf{X}|\lambda)$



Einfacher Ansatz:

$$
\begin{aligned}
P(\mathbf{X}|\lambda) &= \sum_{\text{alle } Q} P(\mathbf{X}, Q|\lambda) \\
&= \sum_{\text{alle } Q} a_{1q_1}\, b_{q_1}(\mathbf{x}_1)\, a_{q_1 q_2}\, b_{q_2}(\mathbf{x}_2) \ldots a_{q_{T-1} q_T}\, b_{q_T}(\mathbf{x}_T)\, a_{q_T N} \; .
\end{aligned}
$$

$\rightarrow$ Grosser Aufwand: $(N-2)^T$ Zustandssequenzen !!

Forward-Algorithmus

Wir definieren die Vorwärtswahrscheinlichkeit:

$$\alpha_t(j) = P(\mathbf{X}_1^t, q_t{=}S_j|\lambda)$$

$\alpha_t(j)$ ist die Verbundwahrscheinlichkeit, dass sich das HMM λ zum Zeitpunkt t im Zustand S_j befindet und die partielle Beobachtungssequenz $\mathbf{X}_1^t$ erzeugt hat.

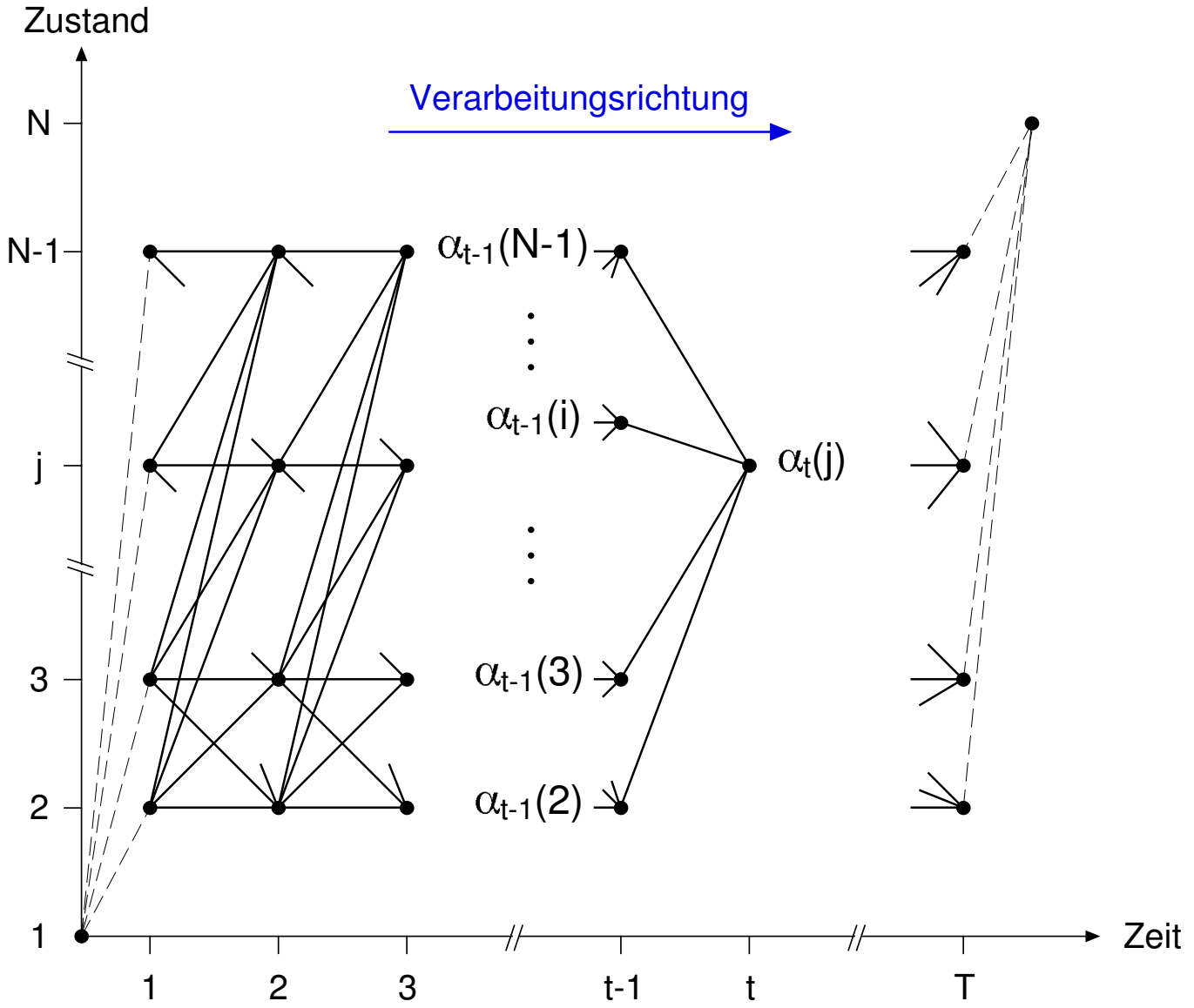
D.h. $\alpha_t(j)$ ergibt sich durch Summieren über alle Zustandssequenzen, die zum Zeitpunkt t im Zustand S_j enden:

$$\alpha_t(j) = \sum_{\text{alle } Q_1^t \text{ mit } q_t{=}S_j} P(\mathbf{X}_1^t, Q_1^t|\lambda) \; .$$

Forward-Algorithmus:

Prinzip:

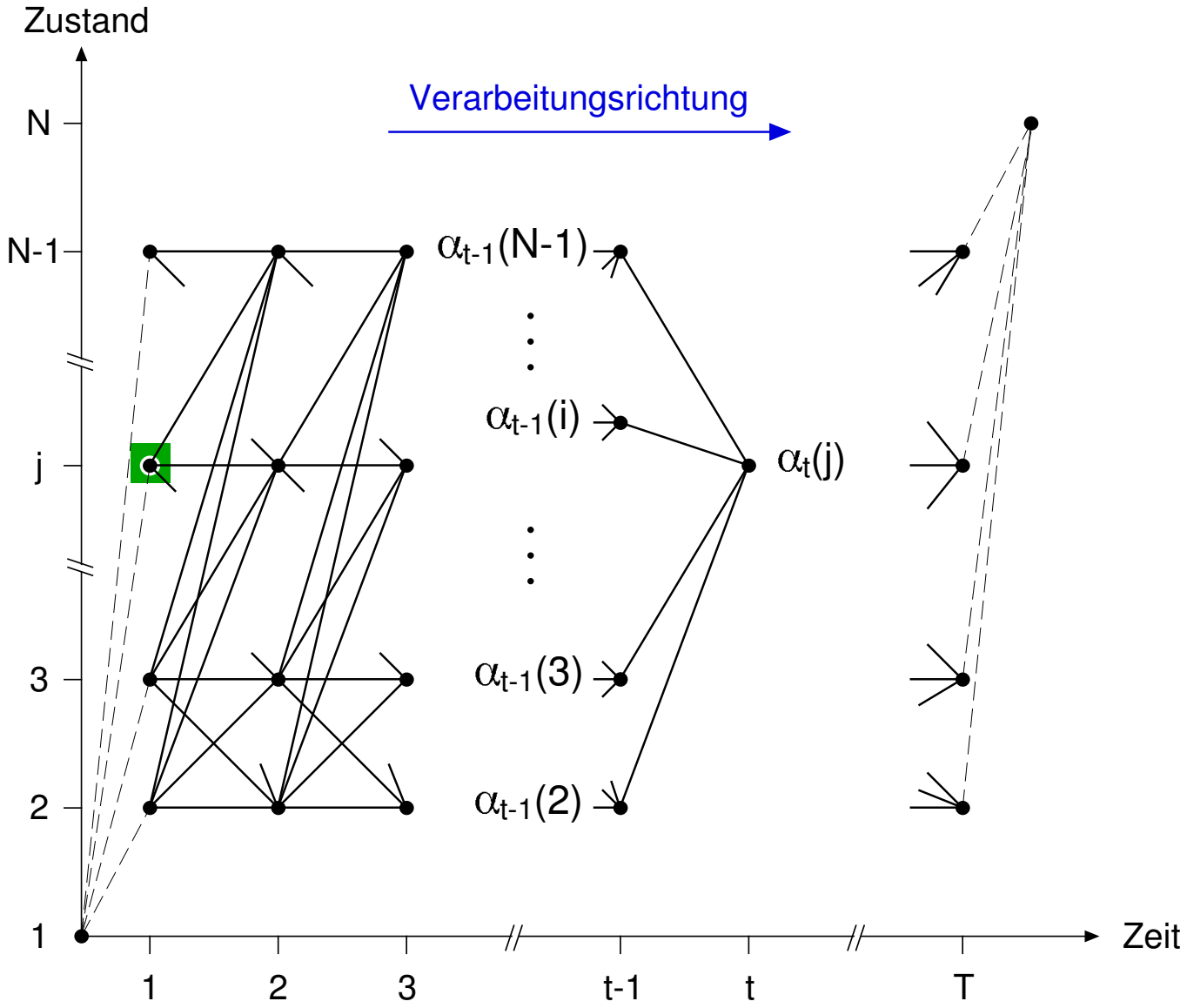
$\alpha_t(j)$ rekursiv ermitteln



$$\alpha_t(j) \;=\; P(\mathbf{X}_1^t, q_t{=}S_j|\lambda) \;=\; \left[\sum_{i=2}^{N-1} \alpha_{t-1}(i)\, a_{ij}\right] b_j(\mathbf{x}_t) \qquad \text{für} \quad \begin{cases} 1 < t \leq T \\ 1 < j < N \end{cases}$$

Forward-Algorithmus:

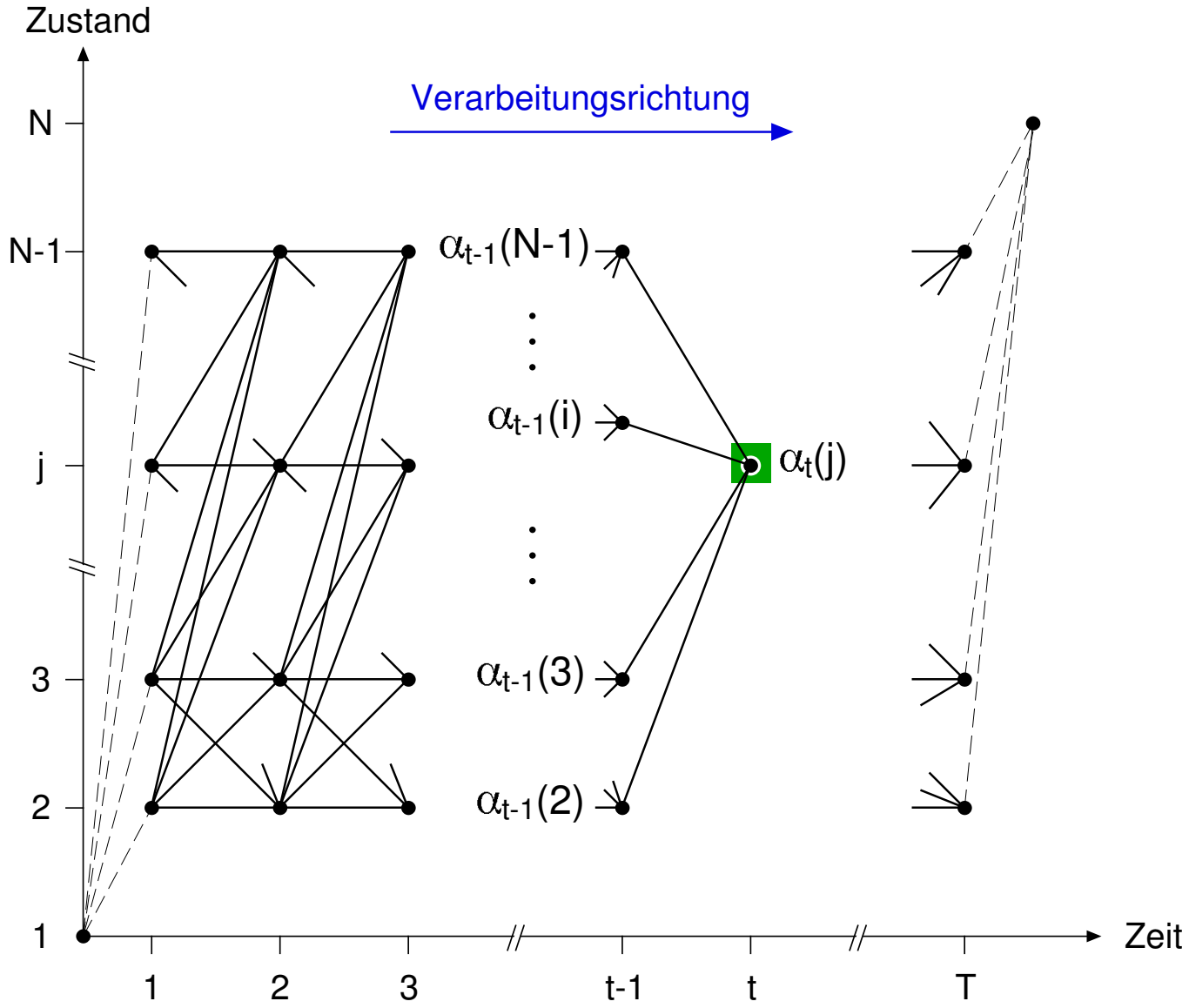
Initialisierung



$$\alpha_1(j) \;=\; a_{1j}\, b_j(\mathbf{x}_1)\,, \qquad 1 < j < N$$

Forward-Algorithmus:

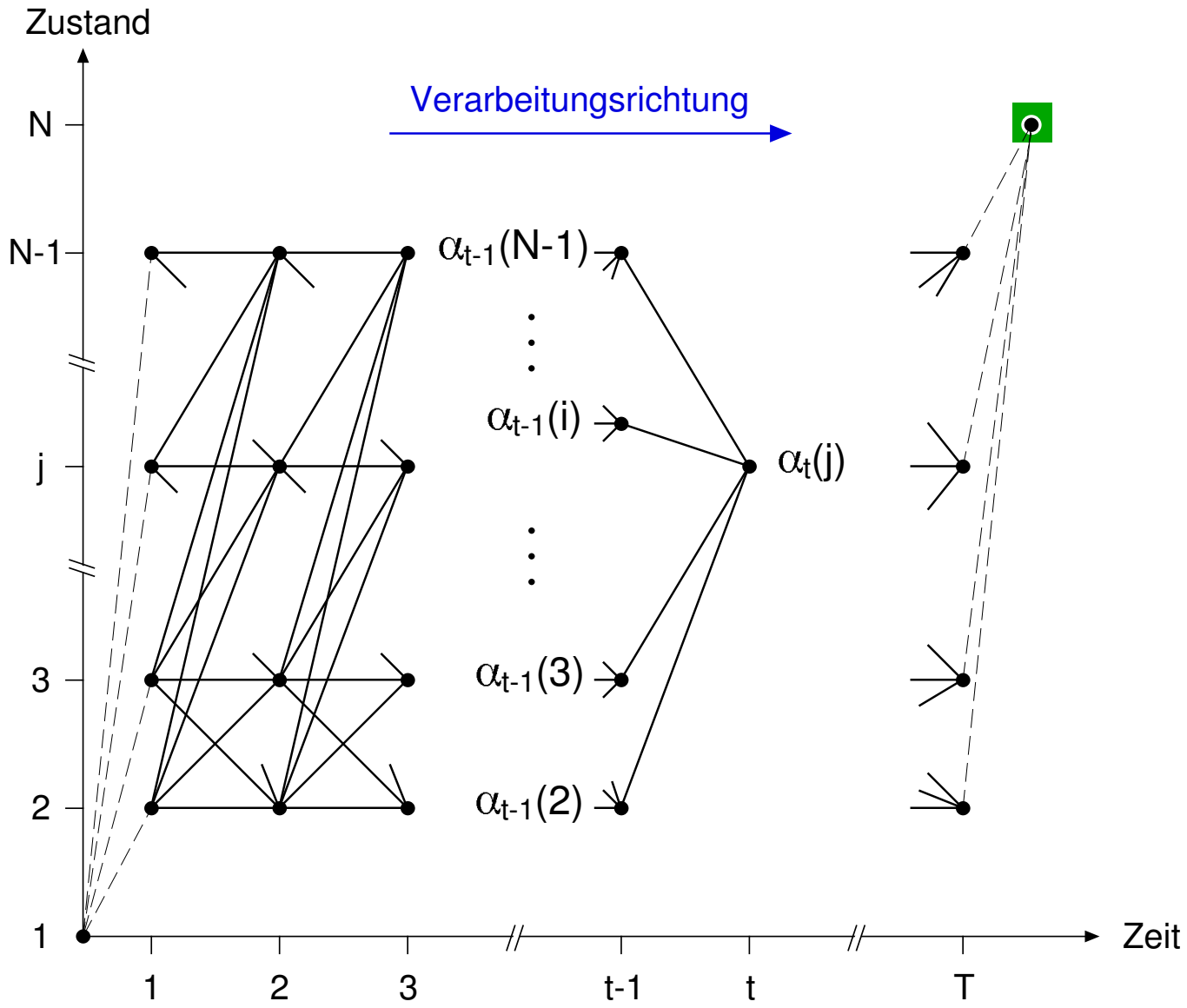
Rekursion



$$\alpha_t(j) = \left[\sum_{i=2}^{N-1} \alpha_{t-1}(i)\, a_{ij} \right] b_j(\mathbf{x}_t) \qquad \text{für} \quad \begin{cases} 1 < t \leq T \\ 1 < j < N \end{cases}$$

Forward-Algorithmus:

Terminierung



$$\alpha_{T+1}(N) = P(\mathbf{X}|\lambda) = \sum_{i=2}^{N-1} \alpha_T(i) a_{iN}$$ (Produktionswahrscheinlichkeit)

Forward-Algorithmus

Berechnung von $P(\mathbf{X}|\lambda)$

Initialisierung: $\alpha_1(j) = a_{1j}\, b_j(\mathbf{x}_1)\,, \qquad 1 < j < N$

Rekursion: $\alpha_t(j) = \left[\sum_{i=2}^{N-1} \alpha_{t-1}(i)\, a_{ij}\right] b_j(\mathbf{x}_t)\,, \qquad \begin{array}{l} 1 < j < N \\ 1 < t \leq T \end{array}$

Terminierung: $P(\mathbf{X}|\lambda) = \sum_{i=2}^{N-1} \alpha_T(i)\, a_{iN}$

Rechenaufwand

ohne Forward-Algorithmus: $\approx \mathcal{O}(T(N-2)^T) = \mathcal{O}(10^{102})$

mit Forward-Algorithmus: $\approx \mathcal{O}(T(N-2)^2) = \mathcal{O}(10^4)$

Beispiel für $N = 12$ und $T = 100$

Thema der nächsten Lektion:

Decodierungsproblem und Schätzproblem

(Viterbi-Algorithmus und Baum-Welch-Algorithmus)

Zur Übersicht der Vorlesung *Sprachverarbeitung II* >>>

Beschreibung der Variation der Sprachmerkmale

Sprachmerkmale für die Spracherkennung sind i.a. vieldimensional

A) Diskrete Sprachmerkmale

Abbildung der D-dimensionalen Vektoren auf M Werte >>>

(Vektorquantisierung mit einem Codebuch der Grösse M)

Statistische Beschreibung ? >>>

B) Kontinuierliche Sprachmerkmale

Statistische Beschreibung ? >>>

Beschreibung der Variation der Sprachmerkmale

Sprachmerkmale für die Spracherkennung sind i.a. vieldimensional

A) Diskrete Sprachmerkmale

Abbildung der D-dimensionalen Vektoren auf M Werte

(Vektorquantisierung mit einem Codebuch der Grösse M)

Statistische Beschreibung: diskrete Wahrscheinlichkeitsverteilung

B) Kontinuierliche Sprachmerkmale

Statistische Beschreibung ?

>>>

Beschreibung der Variation der Sprachmerkmale

Sprachmerkmale für die Spracherkennung sind i.a. vieldimensional

A) Diskrete Sprachmerkmale

Abbildung der D-dimensionalen Vektoren auf M Werte

(Vektorquantisierung mit einem Codebuch der Grösse M)

Statistische Beschreibung: diskrete Wahrscheinlichkeitsverteilung

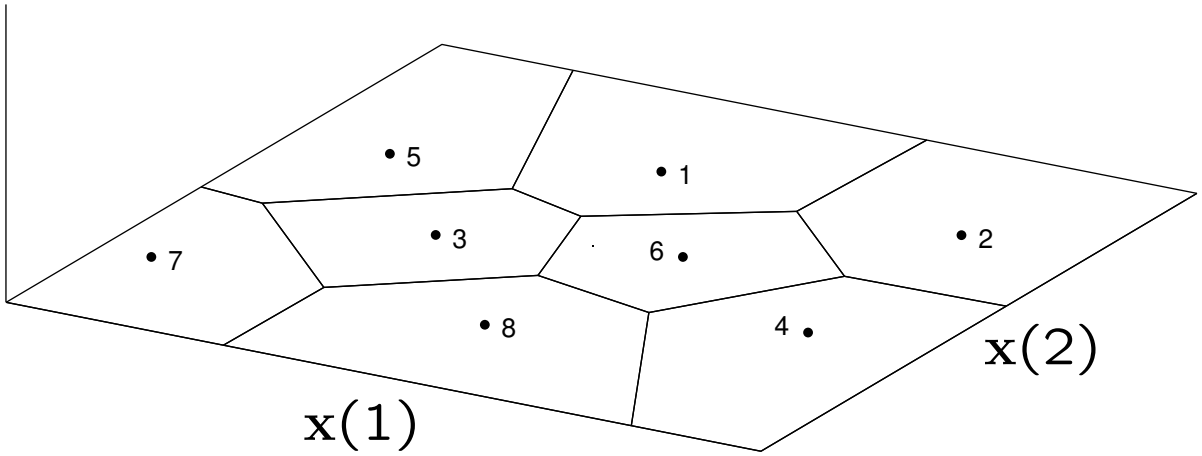
B) Kontinuierliche Sprachmerkmale

Statistische Beschreibung: multivariate Gauss-Mischverteilung

<<<

Diskrete Sprachmerkmale

Resultieren aus Vektorquantisierung



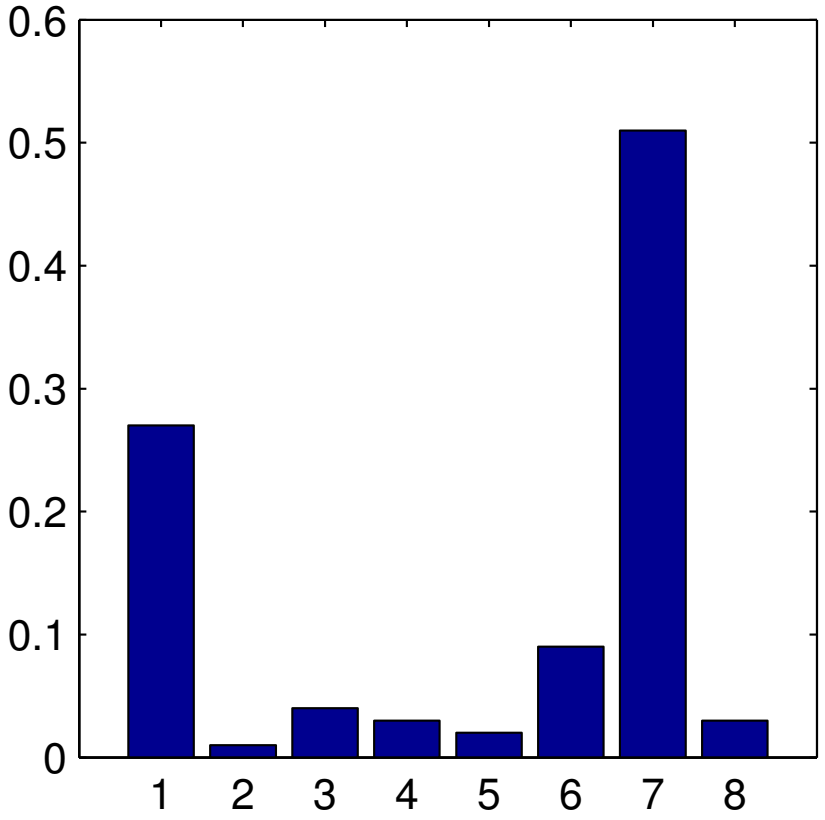
<<<

Diskrete Wahrscheinlichkeitsverteilung

Wahrscheinlichkeitsverteilung des diskreten Merkmals $\mathbf{x}_t$ mit $M = 8$ Werten:

$\mathbf{x}$	$P(\mathbf{x})$
1	0.27
2	0.01
3	0.04
4	0.03
5	0.02
6	0.09
7	0.51
8	0.03

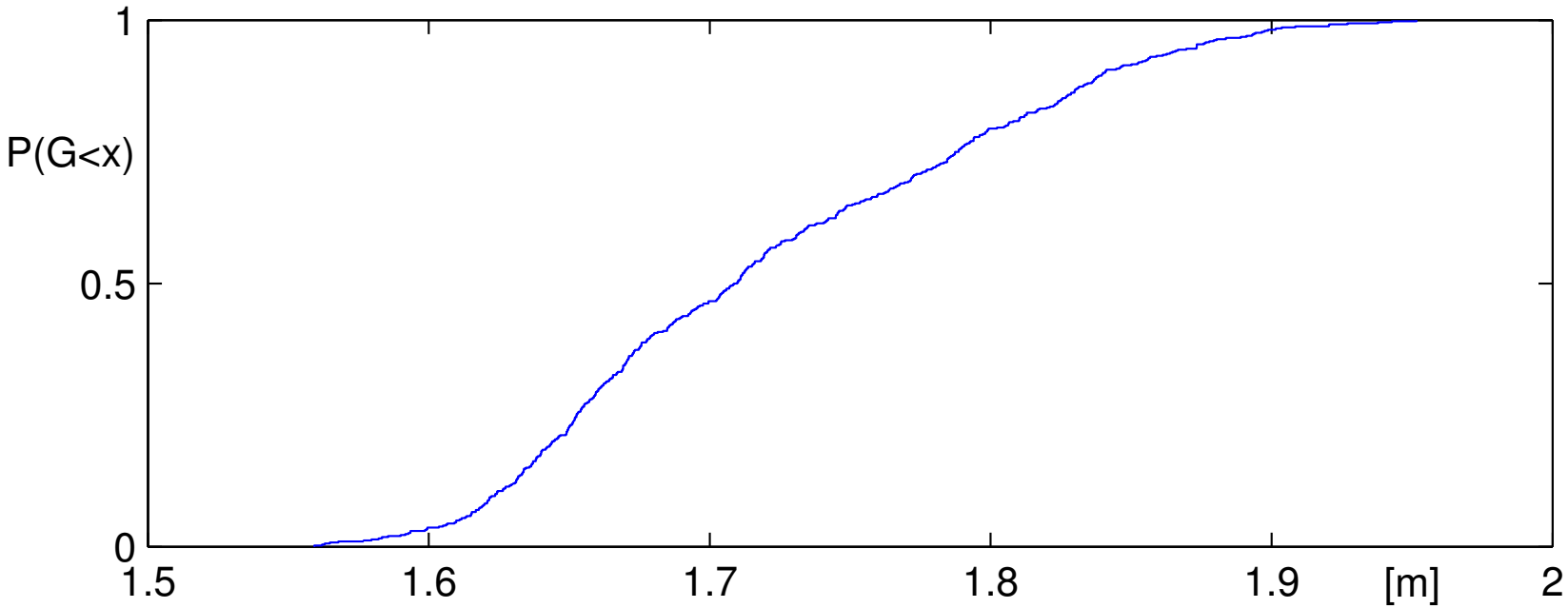
$$\sum_{\mathbf{x}} P(\mathbf{x}) \stackrel{!}{=} 1$$



<<<

Darstellung kontinuierlicher Grössen

Körpergrösse von 500 Personen dargestellt als kumulative Verteilung

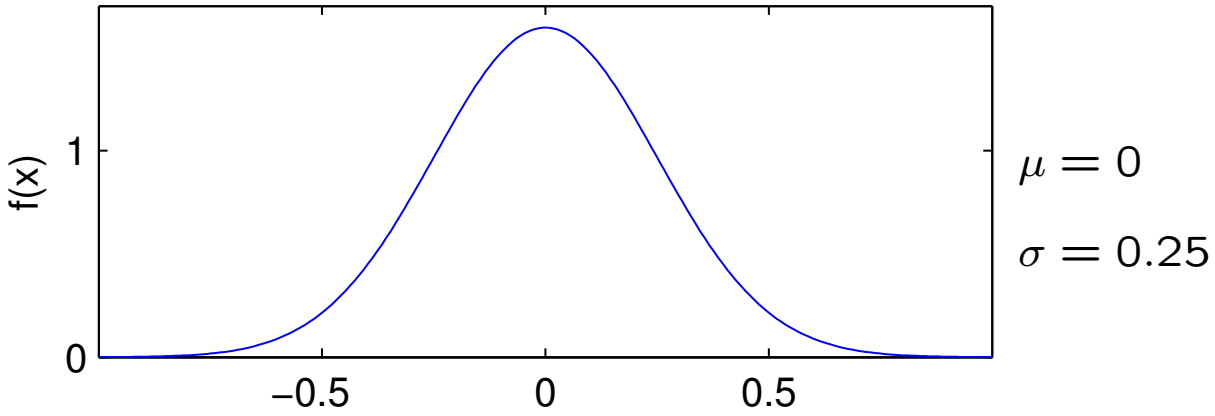


Nicht praktisch zum Rechnen! $\longrightarrow$ Approximation durch analytische Funktion

Normalverteilung (auch Gauss-Verteilung)

Dichtefunktion: $p(x) = \mathcal{N}(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Parameter:
- μ: Mittelwert
- σ: Standardabweichung (die Quadratwurzel der Varianz).
- σ^2: Varianz



Eigenschaften: $p(x) \geq 0, \quad \int_{-\infty}^{\infty} p(x)\, dx = 1$

Likelihood: Wert der Dichtefunktion $p(x)$ an der Stelle x

Gauss-Mischverteilung

Approximiert beliebige Verteilung durch gewichtete Summe von Normalverteilungen

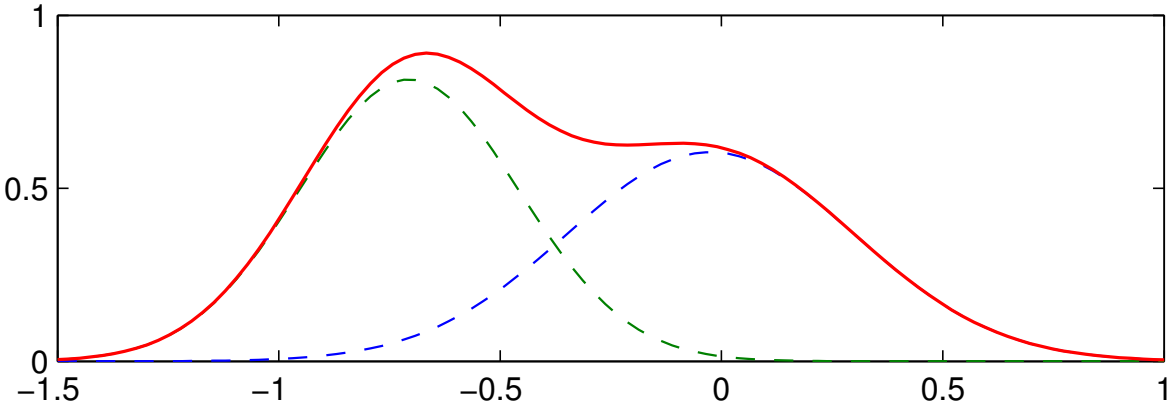
Dichtefunktion: $p(x) = \sum_{k=1}^{M} c_k \mathcal{N}(x, \mu_k, \sigma_k)$

c_k: Mischkoeffizient

M: Anzahl Mischkomponenten

Bedingungen: $c_k \geq 0\,, \quad \sum_{k=1}^{M} c_k = 1$

Beispiel:



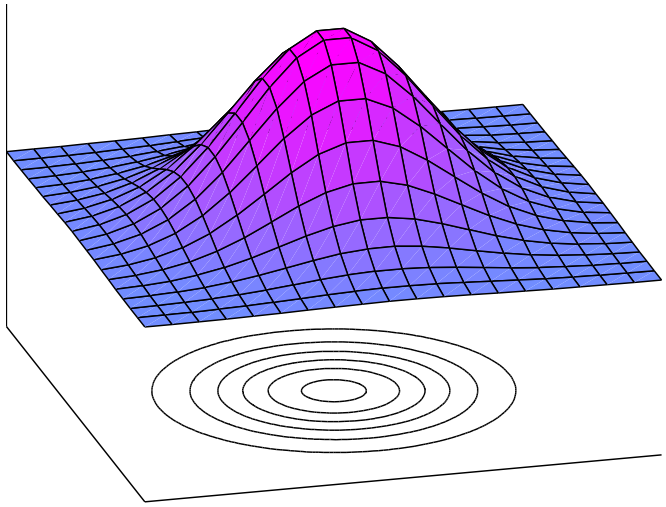
Multivariate Normalverteilung

Multivariat: Verteilung über mehrere Zufallsvariablen oder Zufallsvektor
$\boldsymbol{x} = [x(1)\, x(2)\, \ldots\, x(D)]^{\mathrm{t}}$

Dichtefunktion: $$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{t}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$

Parameter:

- $\boldsymbol{\mu}$: Mittelwertvektor
  (Länge: D)
- $\boldsymbol{\Sigma}$: Kovarianzmatrix
  ($D \times D$-Matrix; symmetrisch)



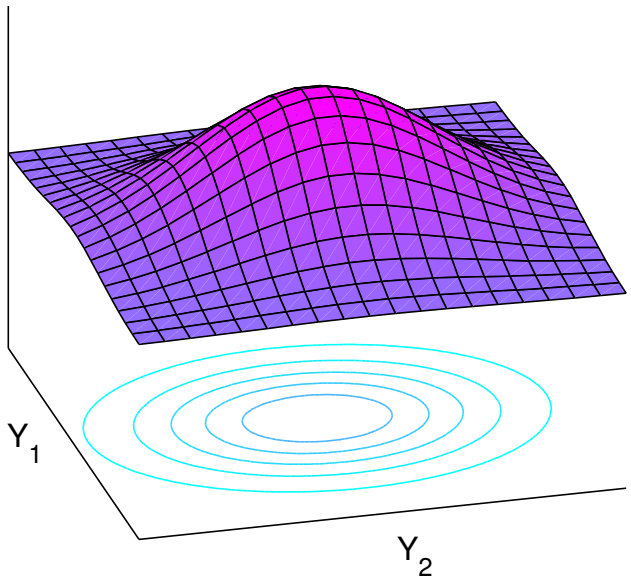
Kovarianzmatrix

Voll besetzt: $D(D+1)/2$ unabhängige Elemente

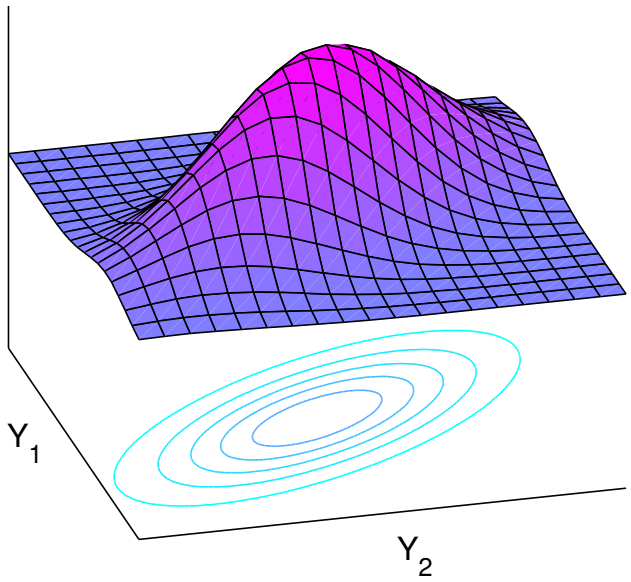
Diagonal: Wenn Elemente von Vektor $\boldsymbol{x}$ paarweise unkorreliert:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_D^2 \end{bmatrix}$$ (nur D Elemente)

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 0.6 \end{bmatrix}$$



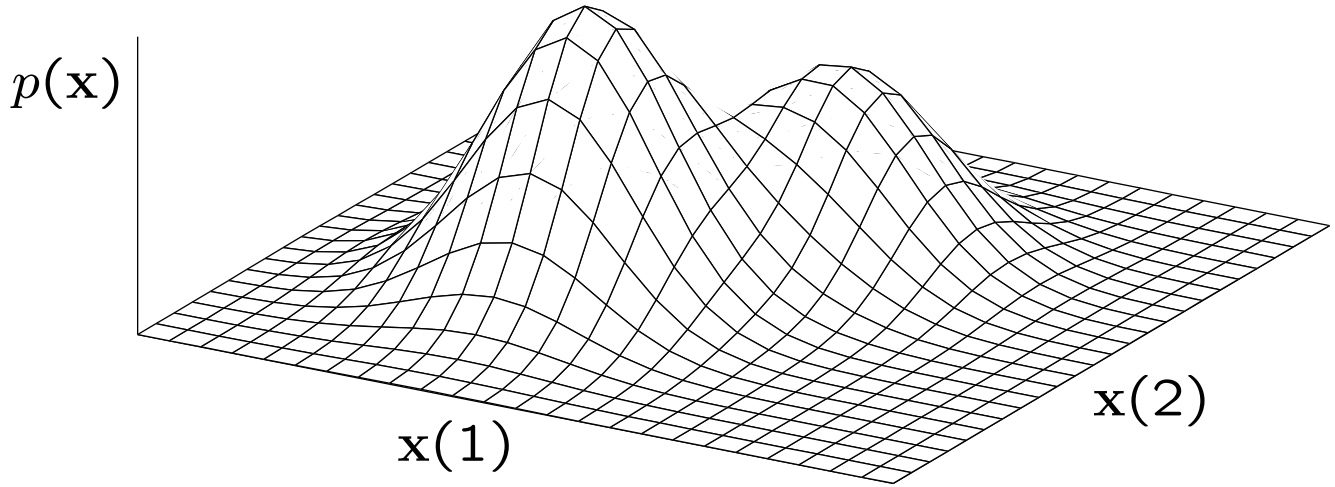
$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 0.6 \end{bmatrix}$$



Multivariate Gauss-Mischverteilung

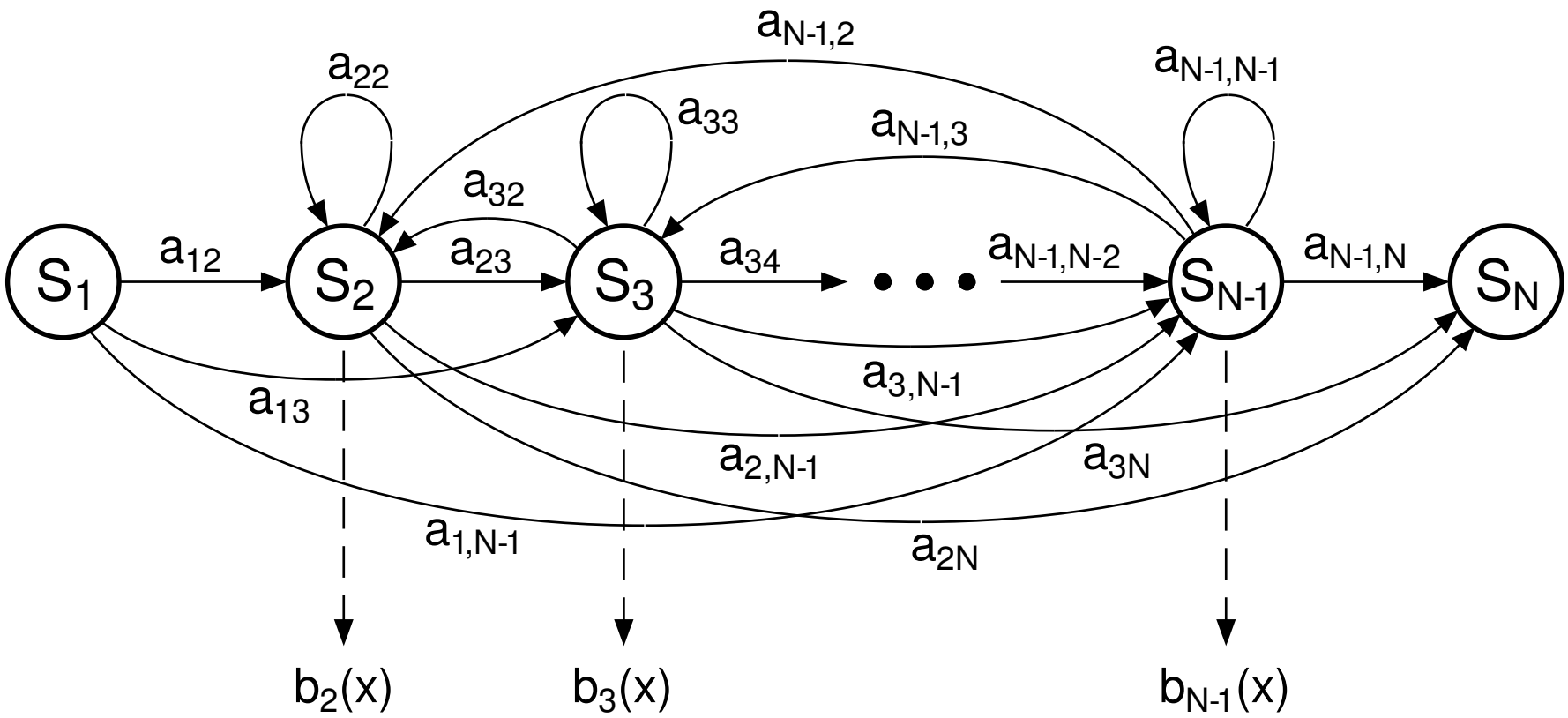
Approximiert beliebige multivatiate Verteilungen durch gewichtete Summe von multivariaten Gauss-Verteilungen

Beispiel: 2-dimensionale Gauss-Mischverteilung mit 2 Mischkomponenten



<<<

Hidden-Markov-Modell (mit $N-2$ emittierenden Zuständen)



Hidden-Markov-Modell

HMM λ mit N Zuständen vollständig beschrieben durch: $\lambda = (A, B)$

A: Zustandsübergangswahrscheinlichkeiten a_{ij} ($N \times N$-Matrix) >>>

B: Beobachtungswahrscheinlichkeitsverteilungen $b_j(\mathbf{x})$

(für jeden der $N-2$ emittierenden Zustände)

Diskrete Beobachtungen: $b_j(\mathbf{x})$ ist diskrete Wahrscheinlichkeitsverteilung

$\longrightarrow$ DDHMM (discrete density HMM) >>>

Kontinuierliche Beobachtungen: $b_j(\mathbf{x})$ ist multivatiate Gauss-Mischverteilung

$\longrightarrow$ CDHMM (continuous density HMM) >>>

HMM mit N Zuständen als Generator

Anfangszustand ist immer: S_1

Endzustand ist immer: S_N

Diskrete Zeit:	0	1	2	$\cdots$	t	$\cdots$	T	$T+1$
Zustandssequenz Q:	S_1	q_1	q_2	$\cdots$	q_t	$\cdots$	q_T	S_N
Beobachtungssequenz $\mathbf{X}$:	–	$\mathbf{x}_1$	$\mathbf{x}_2$	$\cdots$	$\mathbf{x}_t$	$\cdots$	$\mathbf{x}_T$	–

<<<

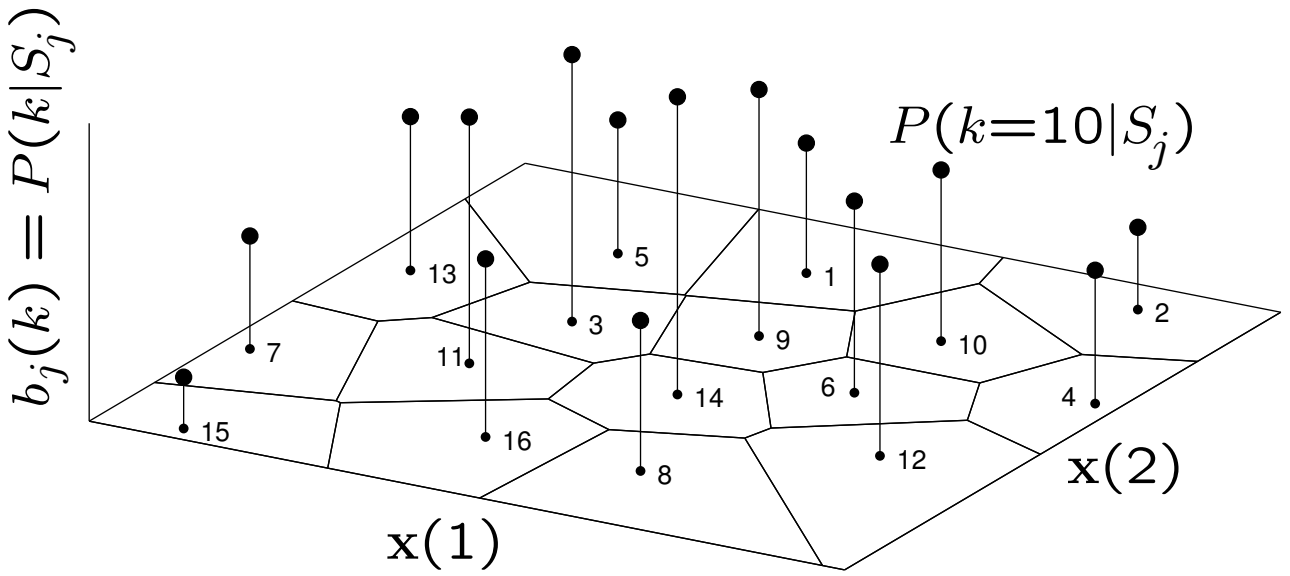
Zustandsübergangswahrscheinlichkeits-Matrix

$$A = \begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1,N-1} & 0 \\ 0 & a_{22} & a_{23} & \cdots & a_{2,N-1} & a_{2N} \\ 0 & a_{32} & a_{33} & \cdots & a_{3,N-1} & a_{3N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{N-1,2} & a_{N-1,3} & \cdots & a_{N-1,N-1} & a_{N-1,N} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

<<<

Diskrete Sprachmerkmale

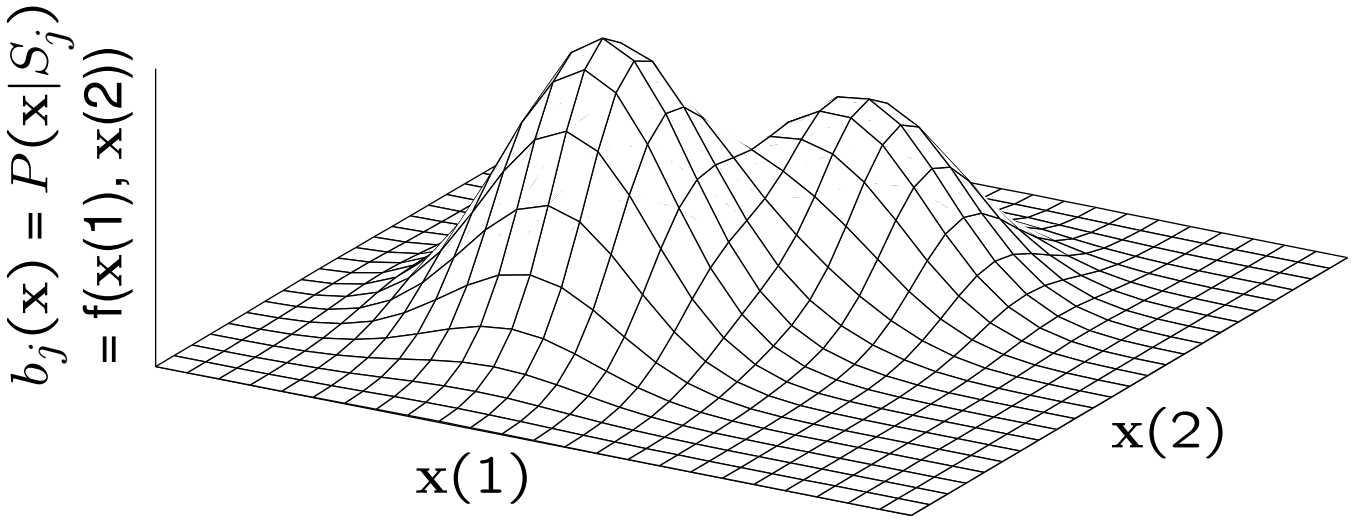
$b_j(\mathbf{x})$ ist eine diskrete Wahrscheinlichkeitsverteilung



<<<

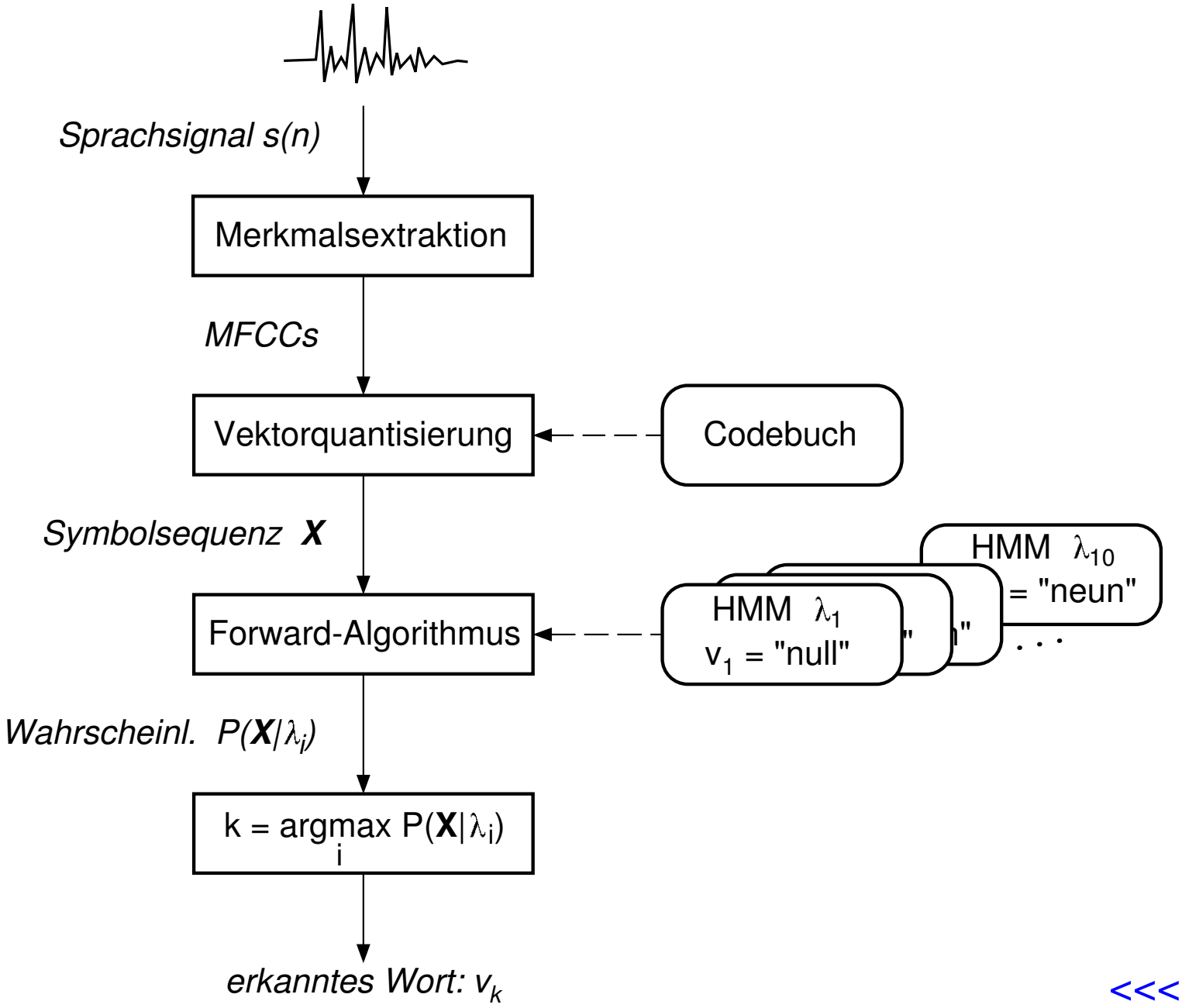
Kontinuierliche Sprachmerkmale

$b_j(\mathbf{x})$ ist eine D-dimensionale Wahrscheinlichkeitsdichte



<<<

Worterkennung



<<<