

Universität Potsdam  
Institut für Informatik  
Lehrstuhl Maschinelles Lernen



# Graphische Modelle

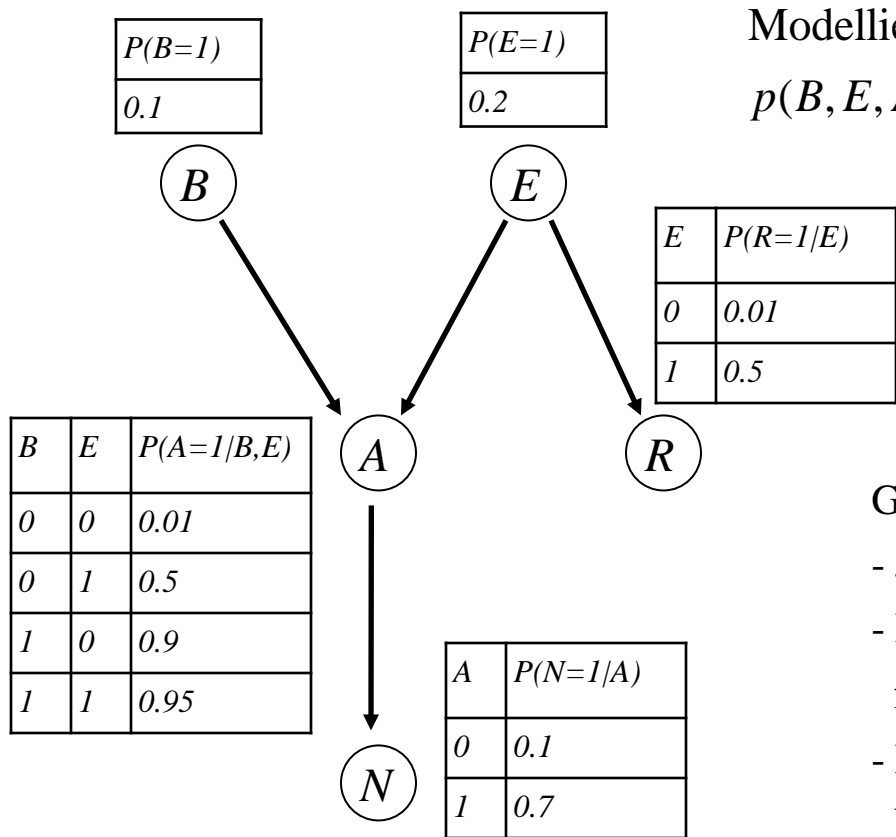
Christoph Sawade/Niels Landwehr/Tobias Scheffer

# Überblick

- Gerichtete Graphische Modelle: Bayessche Netze
- Graphische Modelle im Maschinellen Lernen
- Inferenz in Graphischen Modellen
- Ungerichtete Graphische Modelle: Markov Netze

# Graphische Modelle: Beispiel

## ■ Graphisches Modell für „Alarm“ Szenario



Modellierte Verteilung:

$$p(B, E, A, N, R) = p(B)p(E)p(A | E, B)p(N | A)p(R | E)$$

Graphisches Modell:

- Jede ZV ist ein Knoten
- Für jeden Faktor der Form  $p(X | X_1, \dots, X_k)$  fügen wir gerichtete Kanten von den  $X_i$  zu  $X$  ein
- Modell ist parametrisiert mit den bedingten Verteilungen  $p(X | X_1, \dots, X_k)$

# Überblick

- Gerichtete Graphische Modelle: Bayessche Netze
- **Graphische Modelle im Maschinellen Lernen**
- Inferenz in Graphischen Modellen
- Ungerichtete Graphische Modelle: Markov Netze

# Erinnerung: Lernproblem

- Erinnerung: Lernproblem

- ◆ Trainingsdaten

$$L = \langle (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N) \rangle$$

$\mathbf{x}_i \in \mathbb{R}^m$  Merkmalsvektoren

$y_i \in \{0, 1\}$  binäre Klassenlabel

$y_i \in \mathbb{R}$  reelles Label

- ◆ Matrixschreibweise

Merkmalsvektoren

$$X = (\mathbf{x}_1 \quad \dots \quad \mathbf{x}_N) = \begin{pmatrix} x_{11} & \dots & x_{N1} \\ \vdots & \ddots & \vdots \\ x_{1m} & \dots & x_{Nm} \end{pmatrix}$$

Zugehörige Klassenlabels

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \dots \\ y_N \end{pmatrix}$$

- ◆ Ziel: Vorhersage des Klassenlabels für Testinstanz  $\mathbf{x}$

$$\mathbf{x} \mapsto y$$

# Erinnerung: Bayes'sches Lernen

- Wir betrachten probabilistische Modelle

$$p(y | \mathbf{x}, \theta) \text{ [diskriminativ]}$$

$$p(\mathbf{x}, y | \theta) \text{ [generativ]}$$

- A-priori Verteilung über Modelle  $p(\theta)$  (bekannt)

- Vorhersageproblem: MAP Lösung

$$\theta_* = \arg \max_{\theta} p(\theta | L) \quad y_* = \arg \max_y p(y | \mathbf{x}, \theta_*)$$

- Vorhersageproblem: Bayes Lösung

$$y_* = \arg \max_y p(y | \mathbf{x}, L) = \arg \max_y \int p(y | \mathbf{x}, \theta) p(\theta | L) d\theta$$

# Erinnerung: Parameterschätzung

## Münzwurf

- Erinnerung: Münzwurf
  - ◆ Einzelner Münzwurf Bernouilli-verteilt mit Parameter  $\mu$

$$X \sim \text{Bern}(X | \mu) = \mu^X (1 - \mu)^{1-X}$$

$\mu = p(X = 1 | \mu)$  unbekannter Parameter

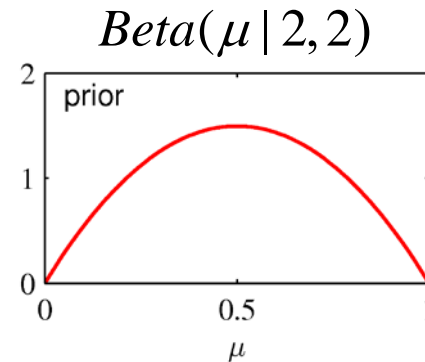
- Parameterschätzproblem:
  - ◆ Wir haben  $N$  unabhängige Münzwürfe gesehen, als Ausprägung  $L = \{x_1, \dots, x_N\}$  der ZV  $X_1, \dots, X_N$
  - ◆ Der echte Parameter  $\mu$  ist unbekannt, wir wollen eine Schätzung  $\hat{\mu}$  bzw. eine posterior-Verteilung  $p(\mu | L)$
  - ◆ Bayesscher Ansatz: Posterior  $\propto$  Prior  $\times$  Likelihood

$$\underbrace{p(\mu | L)}_{\text{posterior}} \propto \underbrace{p(L | \mu)}_{\text{likelihood}} \underbrace{p(\mu)}_{\text{prior}}$$

# Erinnerung: Parameterschätzung Münzwurf

- Prior: Beta-Verteilung über Münzparameter  $\mu$

$$\begin{aligned} p(\mu | \alpha_k, \alpha_z) &= \text{Beta}(\mu | \alpha_k, \alpha_z) \\ &= \frac{\Gamma(\alpha_k + \alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)} \mu^{\alpha_k-1} (1-\mu)^{\alpha_z-1} \end{aligned}$$



- Likelihood  $N$  unabhängige Münzwürfe:

$$\begin{aligned} P(X_1, \dots, X_N | \mu) &= \prod_{i=1}^N p(X_i | \mu) \quad i.i.d. \\ &= \prod_{i=1}^N \text{Bern}(X_i | \mu) \\ &= \prod_{i=1}^N \mu^{X_i} (1-\mu)^{1-X_i} \end{aligned}$$



# Erinnerung: Parameterschätzung Münzwurf

- Zufallsvariablen in Münzwurfszenario sind  $X_1, \dots, X_N, \mu$
- Gemeinsame Verteilung von Daten und Parameter:  
Prior x Likelihood

$$p(X_1, \dots, X_N, \mu) = p(\mu) \prod_{i=1}^N \underbrace{p(X_i | \mu)}_{\text{Bernoulli}}$$

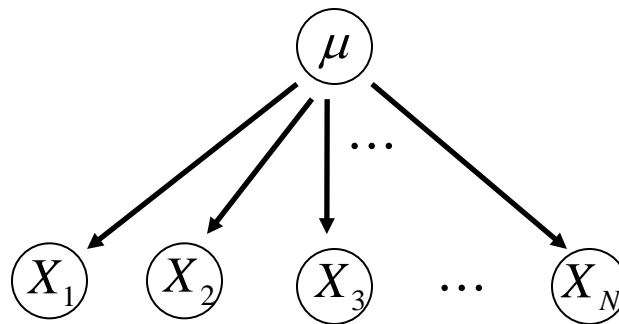
- Darstellung als graphisches Modell:

# Erinnerung: Parameterschätzung Münzwurf

- Zufallsvariablen in Münzwurfszenario sind  $X_1, \dots, X_N, \mu$
- Gemeinsame Verteilung von Daten und Parameter:  
Prior x Likelihood

$$p(X_1, \dots, X_N, \mu) = p(\mu) \prod_{i=1}^N \underbrace{p(X_i | \mu)}_{\text{Bernoulli}}$$

- Darstellung als graphisches Modell:

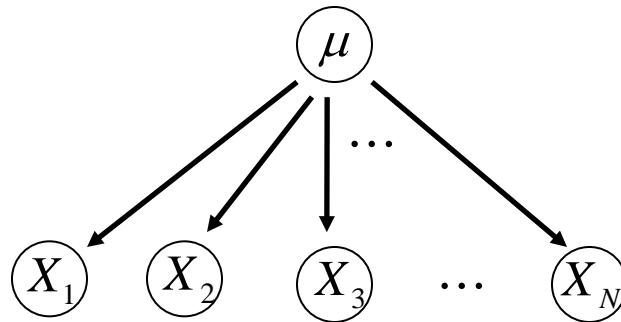


$$pa(\mu) = \emptyset$$

$$pa(X_i) = \{\mu\}$$

# Schätzung eines Münzparameters als Graphisches Modell

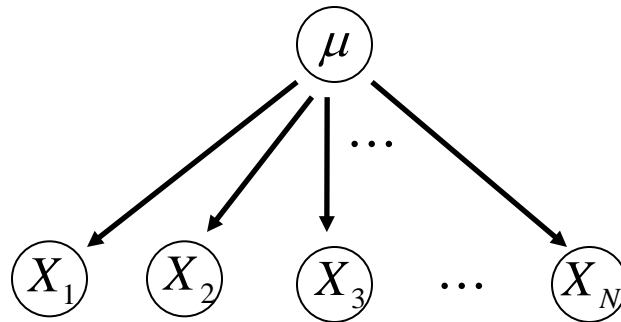
- Unabhängige Münzwürfe: Darstellung als graphisches Modell



- D-separation
  - ◆ Gilt  $X_N \perp X_1, \dots, X_{N-1} \mid \emptyset$  ?

# Schätzung eines Münzparameters als Graphisches Modell

- Unabhängige Münzwürfe: Darstellung als graphisches Modell



- D-separation

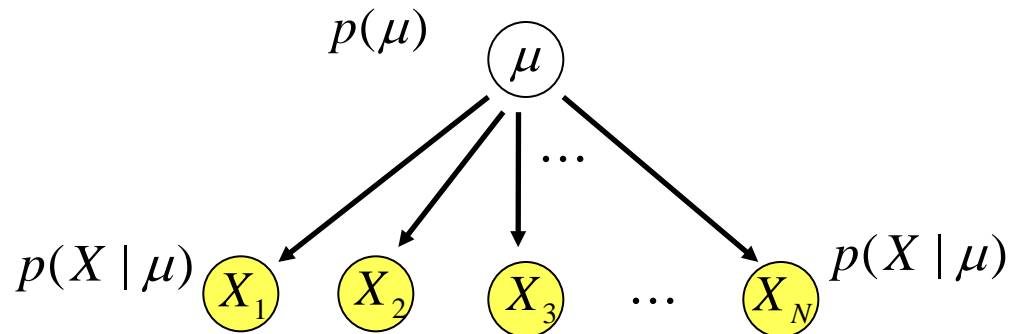
- ◆ Gilt  $X_N \perp X_1, \dots, X_{N-1} \mid \emptyset$  ?
- ◆ Nein, Pfad durch  $\mu$  ist nicht blockiert.
- ◆ Intuitiv:  
 $X_1 = X_2 = \dots = X_{N-1} = 1 \Rightarrow$  Wahrscheinlich  $\mu > 0.5 \Rightarrow$  Wahrscheinlich  $X_N = 1$
- ◆ Der versteckte Parameter  $\mu$  koppelt ZV  $X_1, \dots, X_N$ .
- ◆ Aber es gilt  $X_N \perp X_1, \dots, X_{N-1} \mid \mu$

# Parameterschätzung als Inferenzproblem

- MAP-Parameterschätzung Münzwurf

$$\hat{\mu} = \arg \max_{\mu} p(\mu | x_1, \dots, x_N)$$

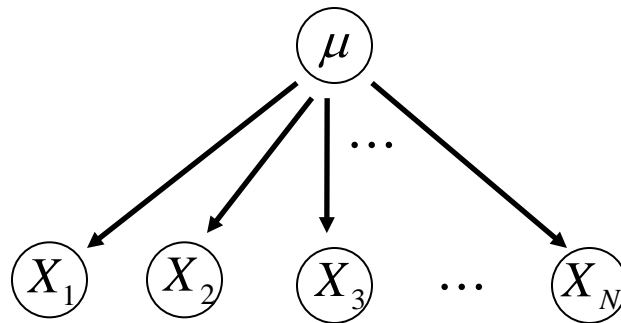
- Inferenzproblem:



- ◆ Evidenz auf den Knoten  $X_1, \dots, X_N$
- ◆ Wahrscheinlichster Zustand des Knotens  $\mu$  gegeben  $X_1, \dots, X_N$

# Plate-Modelle

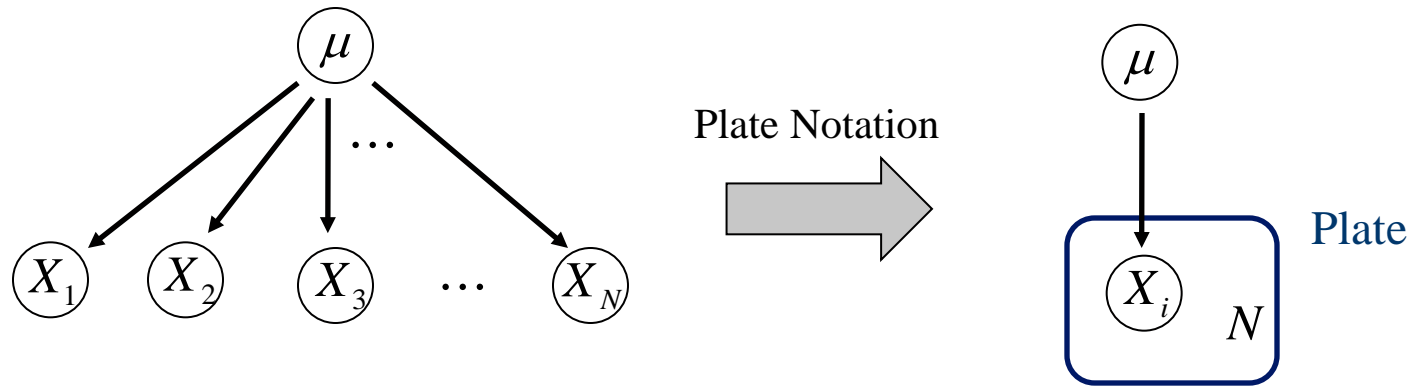
- Unabhängige Münzwürfe: Darstellung als graphisches Modell



- Knoten  $X_1, \dots, X_N$  sind von gleicher Form
  - ◆ Gleicher Wertebereich
  - ◆ Gleiche bedingte Verteilungen  $p(X_i | \mu) = p(X_j | \mu)$ .
- Kurznotation in der Form einer „Schablone“: Plate Notation

# Plate-Modelle

- Plate Notation



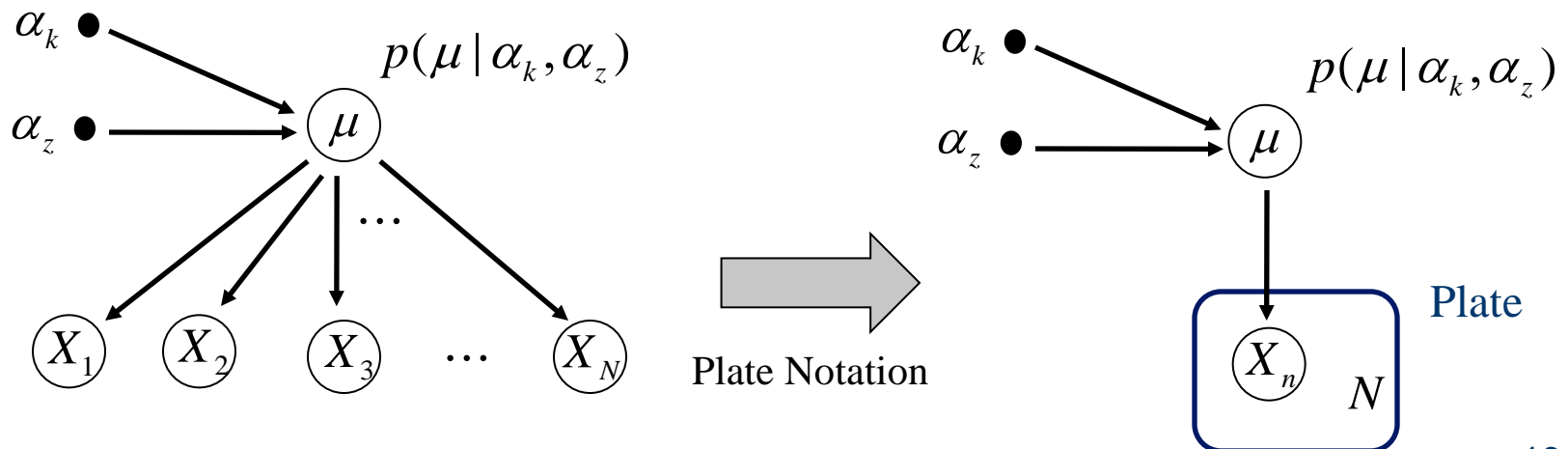
- Ein „Plate“ ist eine abkürzende Notation für  $N$  Variablen der gleichen Form
  - ◆ Bezeichnet mit Anzahl der Variablen,  $N$
  - ◆ Variablen haben Index (z.B.  $X_i$ ).
- Plate-Modelle werden im Maschinellen Lernen oft verwendet

# Plate-Modelle: Hyperparameter

- „Hyperparameter“  $\alpha_k, \alpha_z$  sind keine Zufallsvariablen
  - ◆ Wir modellieren nur die gemeinsame Verteilung über  $X_1, \dots, X_N, \mu$  gegeben Hyperparameter

$$p(X_1, \dots, X_N, \mu | \alpha_k, \alpha_z) = p(\mu | \alpha_k, \alpha_z) \prod_{i=1}^N p(X_i | \mu)$$

- ◆ Hyperparameter keine Knoten im GM, werden aber oft zusätzlich angegeben (Notation: Punkt statt Kreis)





# Erinnerung: Bayessche Lineare Regression

- Regressionslernen

$$L = \langle (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N) \rangle$$

$\mathbf{x}_i \in \mathbb{R}^m$  Merkmalsvektoren

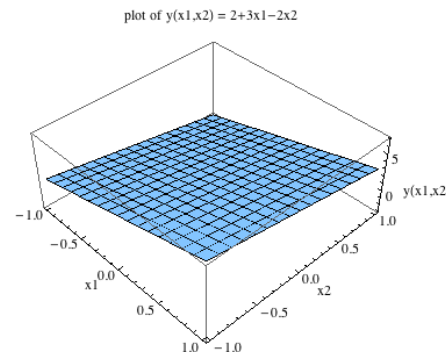
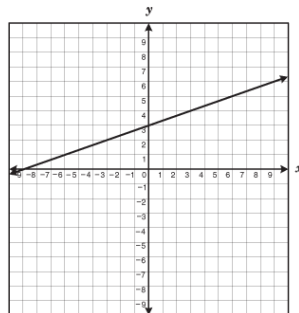
$y_i \in \mathbb{R}$  reelles Zielattribut

- Lineare Regression

$$f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$$

$$= \sum_{i=1}^m w_i x_i$$

$\mathbf{w}$  „Parametervektor“, „Gewichtsvektor“



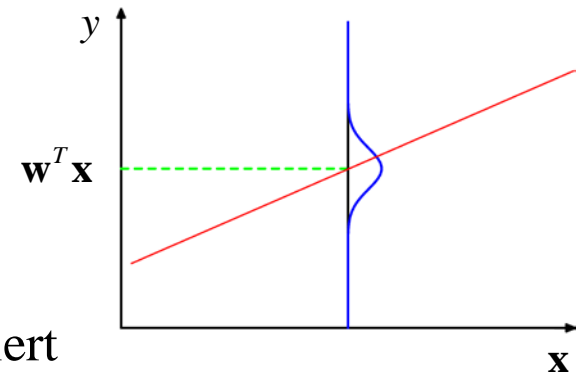
# Erinnerung: Bayessche Lineare Regression

- Diskriminatives Setting:  $\mathbf{x}_i$  fest,  $y_i$  generiert aus  $\mathbf{x}_i$  und  $\mathbf{w}$  plus Gaußschem Rauschen

$$\begin{aligned} p(y | \mathbf{x}, \mathbf{w}) &= \mathbf{w}^T \mathbf{x} + N(y | 0, \sigma^2) \\ &= N(y | \mathbf{w}^T \mathbf{x}, \sigma^2) \end{aligned}$$

$$y_i \sim p(y | \mathbf{x}_i, \mathbf{w})$$

diskriminatives Modell:  $p(\mathbf{x})$  nicht modelliert



- Bayessches Setting: Posterior  $\propto$  Prior x Likelihood

$$\underbrace{p(\mathbf{w} | L)}_{\text{posterior}} \propto \underbrace{p(L | \mathbf{w})}_{\text{likelihood}} \underbrace{p(\mathbf{w})}_{\text{prior}}$$

# Erinnerung: Bayessche Lineare Regression

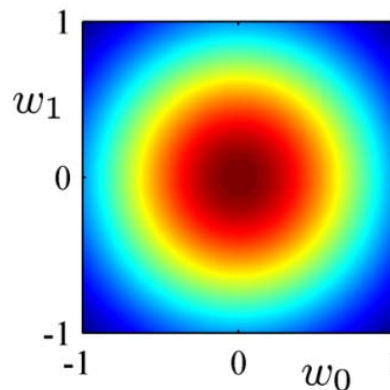
- Likelihood der Daten  $L$  unter einem Modell  $\mathbf{w}$ :

$$\begin{aligned} p(L | X, \mathbf{w}, \sigma^2) &= \prod_{i=1}^N p(y_i | \mathbf{x}_i, \mathbf{w}, \sigma^2) \quad i.i.d. \\ &= \prod_{i=1}^N N(y_i | \mathbf{w}^T \mathbf{x}_i, \sigma^2) \end{aligned}$$

- Normalverteilter Prior über Modelle

$$p(\mathbf{w} | \tau^2) = N(\mathbf{w} | \mathbf{0}, \tau^2 I)$$

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$



Isotrope multivariate  
Normalverteilung,  
Mittelwert  $\mathbf{0}$ , Varianz  $\tau^2$

# Bayessche Lineare Regression als Graphisches Modell

- Was sind Zufallsvariablen?
  - ◆ Datenpunkte  $y_1, \dots, y_N$ , Parameter  $\mathbf{w}$
  - ◆ Nicht:  $\mathbf{x}_1, \dots, \mathbf{x}_N$ , Hyperparameter
- Gemeinsame Verteilung über Daten und Parameter

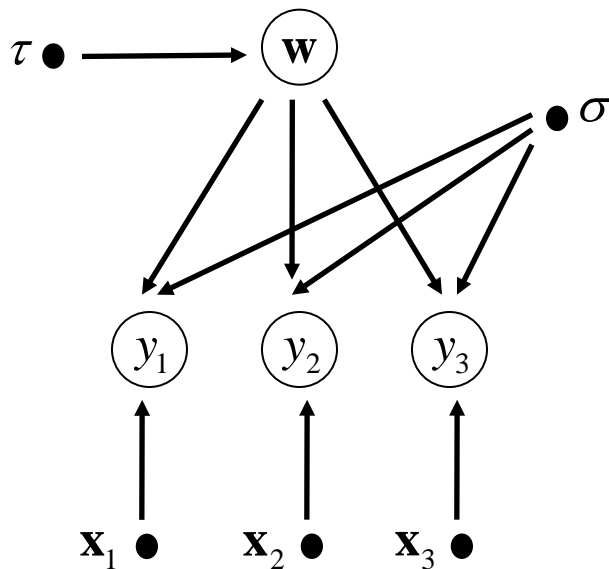
$$\begin{aligned} p(y_1, \dots, y_N, \mathbf{w} \mid \mathbf{x}_1, \dots, \mathbf{x}_N, \sigma^2, \tau^2) &= p(\mathbf{w} \mid \tau^2) p(y_1, \dots, y_N \mid \mathbf{w}, \mathbf{x}_1, \dots, \mathbf{x}_N, \sigma^2) \\ &= p(\mathbf{w} \mid \tau^2) \prod_{i=1}^N p(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2) \end{aligned}$$

- Darstellung von Bayesscher Linearer Regression als graphisches Modell: Ablesen der Struktur aus gemeinsamer Verteilung

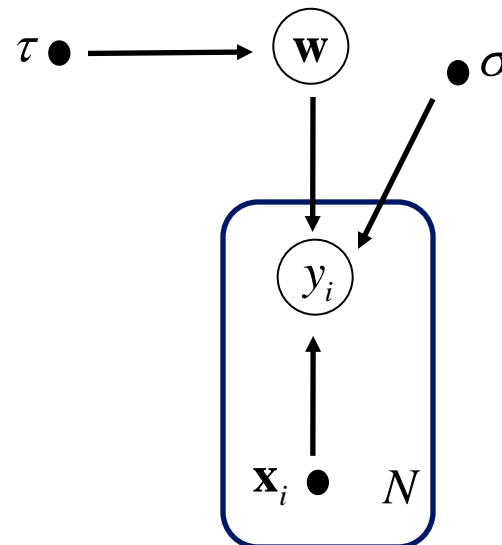
# Bayessche Lineare Regression als Graphisches Modell

$$p(y_1, \dots, y_N, \mathbf{w} \mid \mathbf{x}_1, \dots, \mathbf{x}_N, \sigma^2, \tau^2) = p(\mathbf{w} \mid \tau^2) \prod_{i=1}^N p(y_i \mid \mathbf{x}_i, \mathbf{w}, \sigma^2)$$

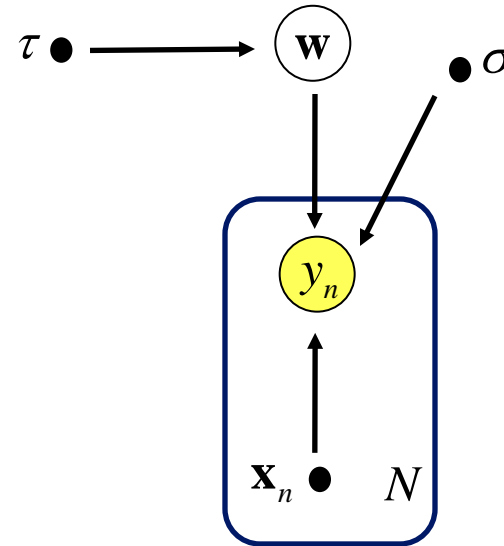
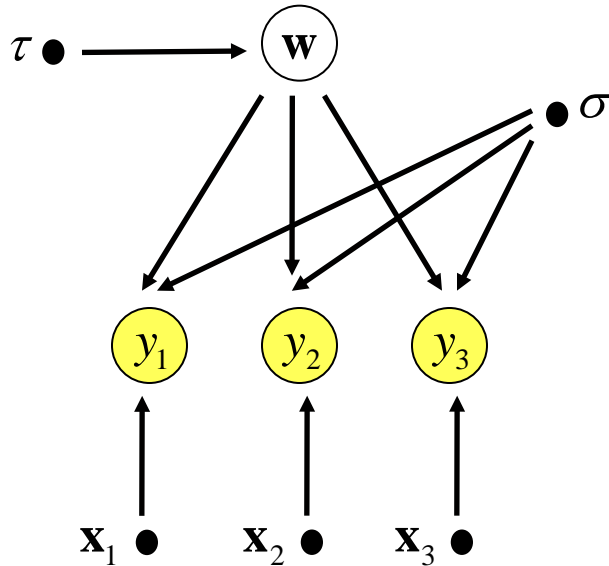
Graphisches Modell,  $N=3$



Graphisches Modell, Plate-Notation



# MAP Parameterschätzung als Inferenzproblem



- MAP Parameterschätzung: wahrscheinlichstes Modell gegeben Daten
  - ◆  $\mathbf{w}_* = \arg \max_{\mathbf{w}} p(\mathbf{w} \mid y_1, \dots, y_N, \mathbf{x}_1, \dots, \mathbf{x}_N, \sigma^2, \tau^2)$
  - ◆ Inferenzproblem: was ist der wahrscheinlichste Zustand für Knoten  $\mathbf{w}$ , gegeben beobachtete Knoten  $y_1, \dots, y_N$ ?

# Bayes-optimale Vorhersage

- Vorhersage mit MAP Modell:

$$\mathbf{w}_* = \arg \max_{\mathbf{w}} p(\mathbf{w} | L, X, \sigma^2, \tau^2)$$

$$y_* = \arg \max_y p(y | \mathbf{x}, \mathbf{w}_*, \sigma^2)$$

$$= \mathbf{w}_*^T \mathbf{x}$$

$X = (\mathbf{x}_1 \quad \dots \quad \mathbf{x}_N)$  Merkmalsvektoren

- Besser als Vorhersage mit MAP Modell ist Bayessche Vorhersage:

$$y_* = \arg \max_y p(y | \mathbf{x}, L, X, \sigma^2, \tau^2)$$

$$= \arg \max_y \int p(y | \mathbf{x}, \mathbf{w}, \sigma^2) p(\mathbf{w} | L, X, \sigma^2, \tau^2) d\mathbf{w}$$

Nicht nötig, sich auf ein Modell fest zu legen

# Bayessche Lineare Regression als Graphisches Modell

- Bayessche Vorhersage: Erweiterung des Modells durch neue Testinstanz (neue Zufallsvariable  $y$ )

$$p(y_1, \dots, y_N, y, \mathbf{w} \mid \mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{x}, \sigma^2, \tau^2) = p(\mathbf{w} \mid \tau^2) \left( \prod_{i=1}^N p(y_i \mid \mathbf{w}, \mathbf{x}_i, \sigma^2) \right) p(y \mid \mathbf{w}, \mathbf{x}, \sigma^2)$$

Graphisches Modell,  $N=3$

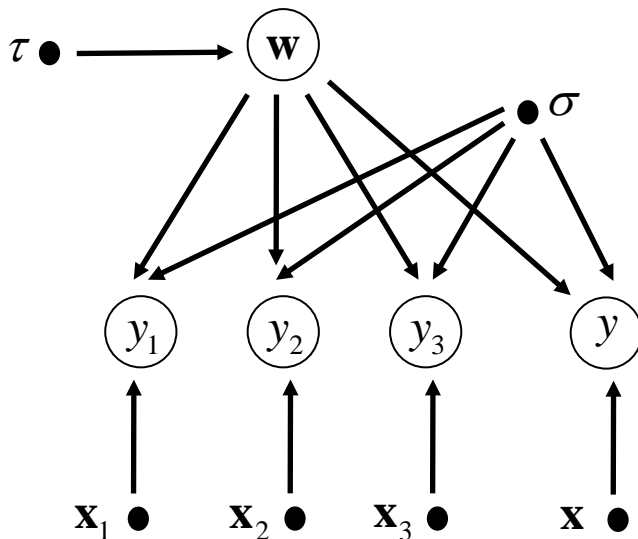
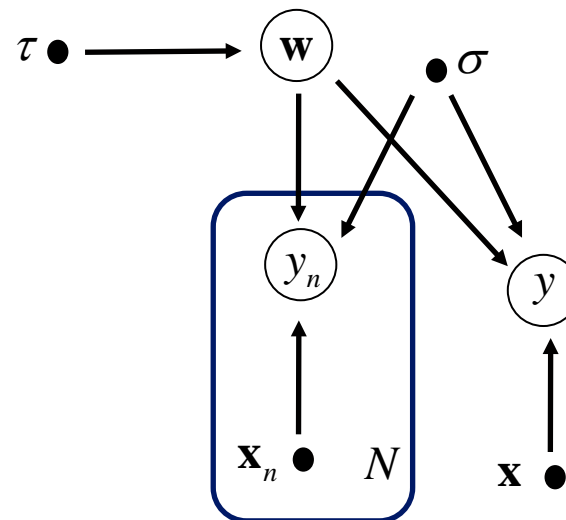
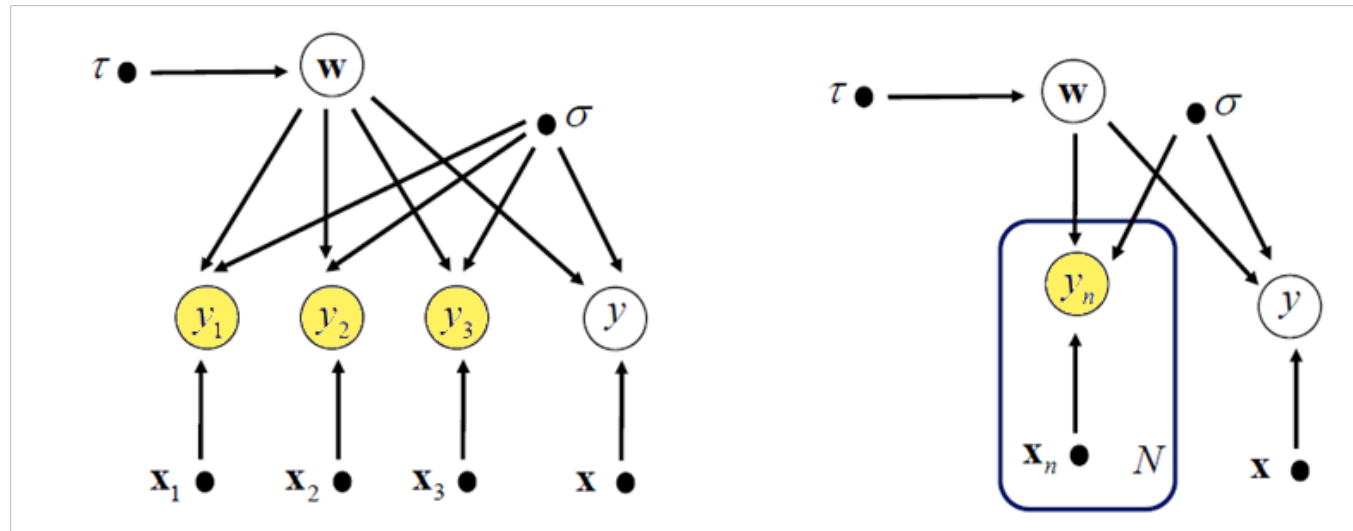


Plate Notation





# Bayessche Lineare Regression als Graphisches Modell



## ■ Bayessche Vorhersage

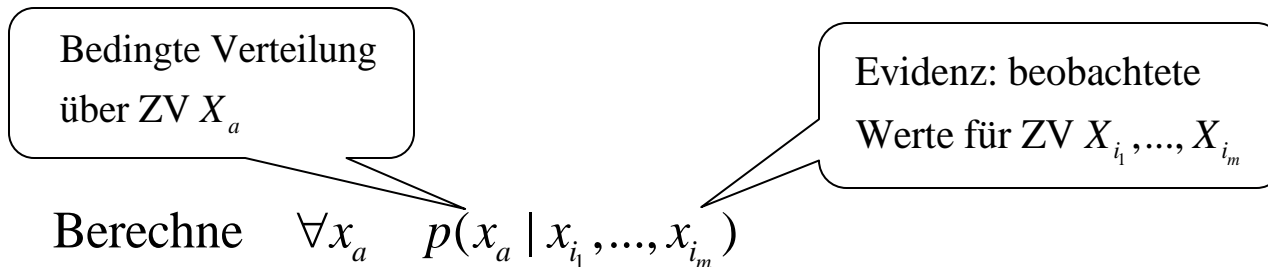
- ◆  $y_* = \arg \max_y p(y | \mathbf{x}, L, X, \sigma^2, \tau^2)$
- ◆ Inferenzproblem: was ist der wahrscheinlichste Zustand für Knoten  $y$ , gegeben beobachtete Knoten  $y_1, \dots, y_N$ ?

# Überblick

- Gerichtete Graphische Modelle: Bayessche Netze
- Graphische Modelle im Maschinellen Lernen
- Inferenz in Graphischen Modellen
- Ungerichtete Graphische Modelle: Markov Netze

# Problemstellung Inferenz

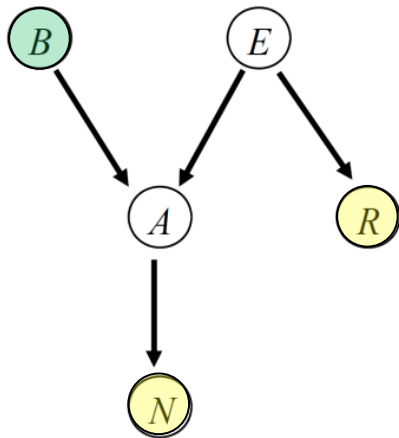
- Gegeben graphisches Modell über Menge von ZV  $\{X_1, \dots, X_N\}$ .
- Problemstellung Inferenz:
  - ◆ Variablen mit Evidenz  $X_{i_1}, \dots, X_{i_m}$   $\{i_1, \dots, i_m\} \subseteq \{1, \dots, N\}$
  - ◆ Anfrage-Variable  $X_a$   $a \in \{1, \dots, N\} \setminus \{i_1, \dots, i_m\}$
  - ◆ Berechne Randverteilung über Anfrage-Variable gegeben Evidenz



Allgemeiner auch  $\forall x_{a_1}, \dots, x_{a_k} \quad p(x_{a_1}, \dots, x_{a_k} \mid x_{i_1}, \dots, x_{i_m})$

# Graphische Modelle: Inferenz

- Beispiel „Alarm“ Domäne
  - ◆ Variablen mit Evidenz:  $N, R$
  - ◆ Anfrage-Variablen:  $B$



Wahrscheinlichkeit für Einbruch gegeben dass der Nachbar uns angerufen hat?

Zum Beispiel:

$$p(B = 1 \mid N = 1, R = 0) = 0.6$$

$$p(B = 0 \mid N = 1, R = 0) = 0.4$$

$$p(B = 1 \mid N = 1, R = 1) = 0.2$$

$$p(B = 0 \mid N = 1, R = 1) = 0.8$$

- Posterior über Parameter, Bayessche Vorhersage, ...

# Graphische Modelle: Inferenz

- Inferenz schwieriges Problem
  - ◆ Allgemeine graphische Modelle: exakte Inferenz NP-hart
  - ◆ Es gibt Algorithmen für exakte Inferenz in allgemeinen graphischen Modellen, deren Laufzeit von den Eigenschaften der Graphstruktur abhängt („Message-Passing“)
  - ◆ Es gibt verschiedene Techniken für approximative Inferenz (Sampling, Variational Inference, Expectation Propagation)
- Wir betrachten
  - ◆ Message-Passing Algorithmus: in Spezialfällen
  - ◆ Sampling-basierte approximative Inferenz

# Inferenz: Diskrete vs. Kontinuierliche Variablen

- Wir diskutieren Inferenz nur für diskrete Variablen
- Betrachtete Inferenzalgorithmen sind auch auf kontinuierliche Variablen anwendbar
  - ◆ Summen ersetzen durch Integrale
  - ◆ Verteilungen müssen so gewählt sein, dass sich die entsprechenden Integrale in geschlossener Form ausrechnen lassen

# Überblick

- Gerichtete Graphische Modelle: Bayessche Netze
- Graphische Modelle im Maschinellen Lernen
- Inferenz in Graphischen Modellen
  - ◆ Exakte Inferenz: Message-Passing
  - ◆ Approximative Inferenz: Sampling
- Ungerichtete Graphische Modelle: Markov Netze

# Überblick

- Gerichtete Graphische Modelle: Bayessche Netze
- Graphische Modelle im Maschinellen Lernen
- Inferenz in Graphischen Modellen
  - ◆ Exakte Inferenz: Message-Passing
  - ◆ Approximative Inferenz: Sampling
- Ungerichtete Graphische Modelle: Markov Netze



# Exakte Inferenz: Naiv

- Graphisches Modell: Repräsentation von  $p(X_1, \dots, X_N)$

- Naive Inferenz:

$$\text{Notation : } \{ X_1, \dots, X_N \} = \left\{ \underbrace{X_a}_{\text{Anfrage-Variable}}, \underbrace{X_{i_1}, \dots, X_{i_m}}_{\text{Evidenz-Variablen}}, \underbrace{X_{j_1}, \dots, X_{j_k}}_{\text{restliche Variablen}} \right\}$$

$$\text{Berechne für jeden Wert } x_a: \quad p(x_a | x_{i_1}, \dots, x_{i_m}) = \frac{p(x_a, x_{i_1}, \dots, x_{i_m})}{p(x_{i_1}, \dots, x_{i_m})}$$

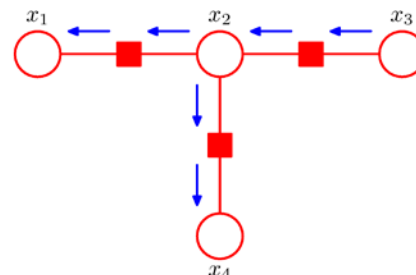
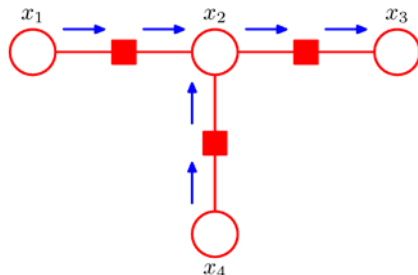
Z Normalisierungsfaktor,  
leicht explizit zu berechnen  
bei univariaten Verteilungen

$$\begin{aligned} &= \frac{1}{Z} p(x_a, x_{i_1}, \dots, x_{i_m}) \\ &= \frac{1}{Z} \sum_{x_{j_1}} \sum_{x_{j_2}} \dots \sum_{x_{j_k}} p(x_1, \dots, x_N) \end{aligned}$$

**Zentrales Problem:** Aussummieren aller restlichen Variablen (exponentiell, wenn naiv gelöst)

# Effizientere Inferenzmethoden?

- Effizientere Methode als naive Inferenz?
  - ◆ Für allgemeine Graphen (vollständig verbunden) nicht möglich!
  - ◆ Aber wenn es Struktur im Modell gibt (Unabhängigkeiten), können wir diese unter Umständen ausnutzen
- Idee: Lokale Berechnungen, die entlang der Graphstruktur propagiert werden
  - ◆ Knoten schicken sich gegenseitig „Nachrichten“, die Ergebnisse von Teilberechnungen enthalten
  - ◆ „Message Passing“, „Belief Propagation“
  - ◆ Laufzeit der Verfahren hängt von Graphstruktur ab (exponentiell im worst-case)



# Graphische Modelle: Inferenz auf linearer Kette

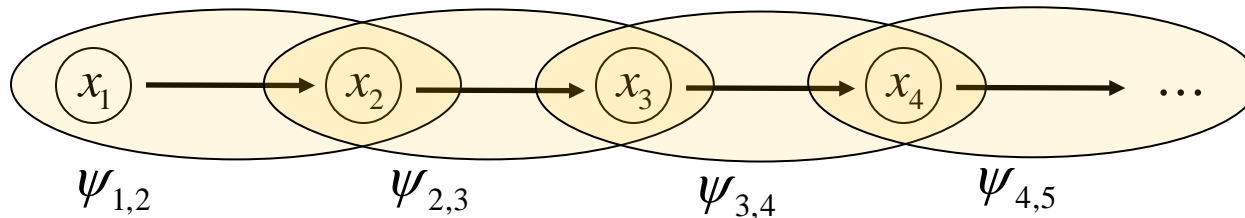
- Wir betrachten zunächst Spezialfall mit besonders einfacher Struktur: lineare Kette von Zufallsvariablen



$$p(x_1, \dots, x_N) = p(x_1) p(x_2 | x_1) p(x_3 | x_2) \cdot \dots \cdot p(x_N | x_{N-1})$$

- Darstellung der gemeinsamen Verteilung als Produkt von *Potenzialen* über Teilmengen von ZV

$$p(x_1, \dots, x_N) = \underbrace{p(x_1) p(x_2 | x_1)}_{\psi_{1,2}(x_1, x_2)} \underbrace{p(x_3 | x_2)}_{\psi_{2,3}(x_2, x_3)} \cdot \dots \cdot \underbrace{p(x_N | x_{N-1})}_{\psi_{N-1,N}(x_N, x_{N-1})}$$



# Inferenz: lineare Kette von ZV

- Zunächst betrachten wir Inferenzaufgabe ohne Evidenz

$$p(x_a) = \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_N} p(x_1, \dots, x_N)$$

Anfrage-Variable

Restliche Variablen (aussummieren)

- Naive Berechnung exponentiell (Mehrfachsumme)
- Idee: Struktur (lineare Kette) ausnutzen, um Berechnung effizienter durchzuführen

# Inferenz: Message-Passing

- Nutze Faktorisierung der gemeinsamen Verteilung in Potenziale (Unabhängigkeiten)

$$\begin{aligned}
 p(x_a) &= \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_N} p(x_1, \dots, x_N) \\
 &= \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_N} \psi_{1,2}(x_1, x_2) \cdots \psi_{N-2,N-1}(x_{N-2}, x_{N-1}) \psi_{N-1,N}(x_{N-1}, x_N) \\
 &= \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_{N-1}} \psi_{1,2}(x_1, x_2) \cdots \psi_{N-2,N-1}(x_{N-2}, x_{N-1}) \underbrace{\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)}_{\mu_\beta(x_{N-1})}
 \end{aligned}$$

- Lokale Teilberechnung: „Nachricht“  $\mu_\beta(x_{N-1})$

- ◆ Berechne für alle Werte von  $x_{N-1}$ :  $\mu_\beta(x_{N-1}) = \sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)$
- ◆ In der Nachricht ist der Knoten  $X_N$  aussummiert
- ◆ Nachricht ist Funktion in Abhängigkeit vom Zustand  $x_{N-1}$ , (z.B. kodiert als Vektor)

# Inferenz: Message-Passing

- Nutze Fakt  
Potenziale

$$\begin{aligned}
 p(x_a) &= \sum_{x_1} \dots \\
 &= \sum_{x_1} \dots \\
 &= \sum_{x_1} \dots
 \end{aligned}$$

Kodierung z.B. als Vektor:  $\mu_\beta(x_{N-1}) =$

$$\begin{pmatrix}
 \sum_{x_N} \psi_{N-1,N}(x_{N-1} = 1, x_N) \\
 \sum_{x_N} \psi_{N-1,N}(x_{N-1} = 2, x_N) \\
 \dots \\
 \sum_{x_N} \psi_{N-1,N}(x_{N-1} = K, x_N)
 \end{pmatrix}$$

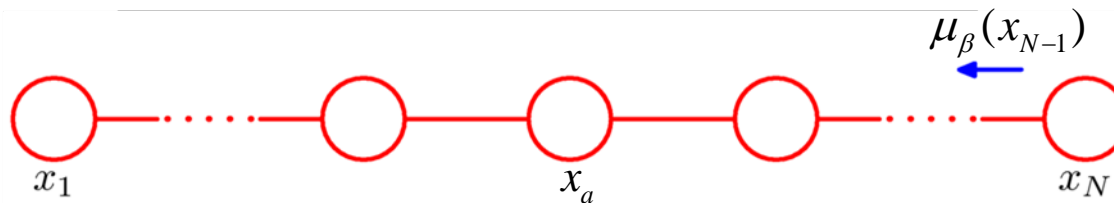
$\underbrace{x_N}_{\mu_\beta(x_{N-1})}$

- Lokale Teilberechnung: „Nachricht“  $\mu_\beta(x_{N-1})$

- ◆ Berechne für alle Werte von  $x_{N-1}$ :  $\mu_\beta(x_{N-1}) = \sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)$
- ◆ In der Nachricht ist der Knoten  $X_N$  aussummiert
- ◆ Nachricht ist Funktion in Abhängigkeit vom Zustand  $x_{N-1}$ , (z.B. kodiert als Vektor)

# Inferenz: Message-Passing

- **Anschauung:** Wir summieren den Knoten  $x_N$  aus, und schicken das Ergebnis weiter an den Knoten  $x_{N-1}$



# Inferenz: Message-Passing

- Wir wenden dieselbe Idee auf die weiteren auszusummierenden Variablen an

$$\begin{aligned}
 p(x_a) &= \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_{N-1}} \psi_{1,2}(x_1, x_2) \cdots \psi_{N-3, N-2}(x_{N-3}, x_{N-2}) \psi_{N-2, N-1}(x_{N-2}, x_{N-1}) \mu_\beta(x_{N-1}) \\
 &= \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_{N-2}} \psi_{1,2}(x_1, x_2) \cdots \psi_{N-3, N-2}(x_{N-3}, x_{N-2}) \underbrace{\sum_{x_{N-1}} \psi_{N-2, N-1}(x_{N-2}, x_{N-1}) \mu_\beta(x_{N-1})}_{\mu_\beta(x_{N-2})} \\
 &= \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_{N-2}} \psi_{1,2}(x_1, x_2) \cdots \psi_{N-3, N-2}(x_{N-3}, x_{N-2}) \mu_\beta(x_{N-2})
 \end{aligned}$$

- Rekursive Teilberechnung: „Nachricht“  $\mu_\beta(x_{N-2})$

- ◆ Berechne  $\mu_\beta(x_{N-2}) = \sum_{x_{N-1}} \psi_{N-2, N-1}(x_{N-2}, x_{N-1}) \mu_\beta(x_{N-1})$
- ◆ In der Nachricht sind die Knoten  $X_N, X_{N-1}$  aussummiert
- ◆ Nachricht ist Funktion in Abhängigkeit vom Zustand  $x_{N-2}$



# Inferenz: Message-Passing

- Weiter nach demselben Prinzip, bis die Nachrichten den Anfrageknoten erreichen

$$\begin{aligned} p(x_a) &= \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_{N-2}} \psi_{1,2}(x_1, x_2) \cdots \psi_{N-3, N-2}(x_{N-3}, x_{N-2}) \mu_\beta(x_{N-2}) \\ &= \dots \\ &= \sum_{x_1} \cdots \sum_{x_{a-1}} \psi_{1,2}(x_1, x_2) \cdots \psi_{a-1, a}(x_{a-1}, x_a) \mu_\beta(x_a) \end{aligned}$$

- Rekursive Teilberechnungen  $\mu_\beta(x_k)$  (für  $k = N-3, \dots, a$ )

$$\mu_\beta(x_{N-3}) = \sum_{x_{N-2}} \psi_{N-3, N-2}(x_{N-3}, x_{N-2}) \mu_\beta(x_{N-2}) \quad X_N, X_{N-1}, X_{N-2} \text{ aussummiert}$$

...

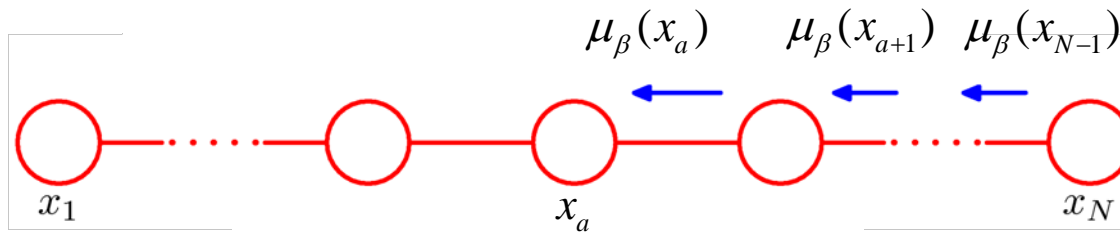
$$\mu_\beta(x_a) = \sum_{x_{a+1}} \psi_{a, a+1}(x_a, x_{a+1}) \mu_\beta(x_{a+1}) \quad X_N, \dots, X_{a+1} \text{ aussummiert}$$

# Inferenz: Message-Passing

- Weiter nach demselben Prinzip, bis die Nachrichten den Anfrageknoten erreichen

$$\begin{aligned} p(x_a) &= \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_{N-2}} \psi_{1,2}(x_1, x_2) \cdots \psi_{N-3, N-2}(x_{N-3}, x_{N-2}) \mu_\beta(x_{N-2}) \\ &= \dots \\ &= \sum_{x_1} \cdots \sum_{x_{a-1}} \psi_{1,2}(x_1, x_2) \cdots \psi_{a-1, a}(x_{a-1}, x_a) \mu_\beta(x_a) \end{aligned}$$

- Nachrichten von rechts ausgetauscht bis zum Anfrageknoten  $x_a$ :  
aller Variablen rechts des Anfrageknotens aussummiert



- Verbleibende Variablen aussummieren: Nachrichten von links

# Inferenz: Message-Passing

- Gleiches Prinzip anwenden auf die Variablen links vom Anfrageknoten

$$\begin{aligned} p(x_a) &= \sum_{x_1} \sum_{x_2} \cdots \sum_{x_{a-1}} \psi_{1,2}(x_1, x_2) \psi_{2,3}(x_2, x_3) \cdots \psi_{a-1,a}(x_{a-1}, x_a) \mu_\beta(x_a) \\ &= \mu_\beta(x_a) \sum_{x_{a-1}} \cdots \sum_{x_2} \sum_{x_1} \psi_{a-1,a}(x_{a-1}, x_a) \cdots \psi_{2,3}(x_2, x_3) \psi_{1,2}(x_1, x_2) \\ &= \mu_\beta(x_a) \sum_{x_{a-1}} \cdots \sum_{x_2} \psi_{a-1,a}(x_{a-1}, x_a) \cdots \psi_{2,3}(x_2, x_3) \underbrace{\sum_{x_1} \psi_{1,2}(x_1, x_2)}_{\mu_\alpha(x_2)} \end{aligned}$$

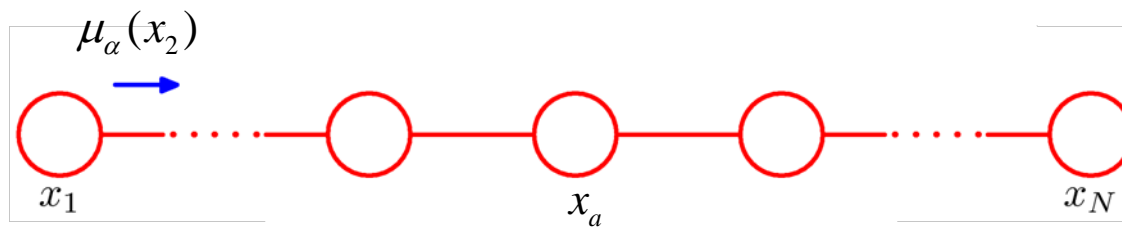
- Lokale Teilberechnung: „Nachricht“  $\mu_\alpha(x_2)$

- ◆ Berechne  $\mu_\alpha(x_2) = \sum_{x_1} \psi_{1,2}(x_1, x_2)$

- ◆ In der Nachricht ist der Knoten  $X_1$  aussummiert

# Inferenz: Message-Passing

- Anschauung: Wir summieren den Knoten  $X_1$  aus, und schicken das Ergebnis weiter an den Knoten  $X_2$



# Inferenz: Message-Passing

- Auf diese Art alle verbleibenden Variablen aussummieren

$$\begin{aligned} p(x_a) &= \mu_\beta(x_a) \sum_{x_{a-1}} \cdots \sum_{x_3} \sum_{x_2} \psi_{a-1,a}(x_{a-1}, x_a) \cdots \psi_{3,4}(x_3, x_4) \psi_{2,3}(x_2, x_3) \mu_\alpha(x_2) \\ &= \mu_\beta(x_a) \sum_{x_{a-1}} \cdots \sum_{x_3} \psi_{a-1,a}(x_{a-1}, x_a) \cdots \psi_{3,4}(x_3, x_4) \underbrace{\sum_{x_2} \psi_{2,3}(x_2, x_3) \mu_\alpha(x_2)}_{\mu_\alpha(x_3)} \\ &= \dots \\ &= \mu_\beta(x_a) \mu_\alpha(x_a) \end{aligned}$$

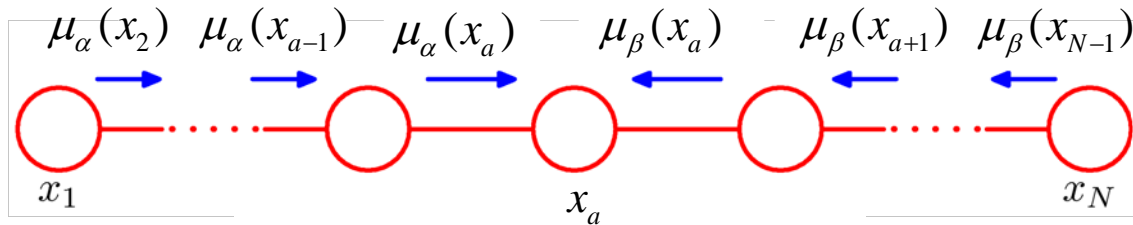
- Rekursive Teilberechnungen:

$$\begin{array}{ll} \text{Berechne } \mu_\alpha(x_3) = \sum_{x_2} \psi_{2,3}(x_2, x_3) \mu_\alpha(x_2) & X_1, X_2 \text{ aussummiert} \\ \dots & \end{array}$$

$$\begin{array}{ll} \text{Berechne } \mu_\alpha(x_a) = \sum_{x_{a-1}} \psi_{a-1,a}(x_{a-1}, x_a) \mu_\alpha(x_{a-1}) & X_1, \dots, X_{a-1} \text{ aussummiert} \end{array}$$

# Inferenz: Message-Passing

- Nachrichten-Austausch Schema



$$p(x_a) = \mu_\beta(x_a) \mu_\alpha(x_a)$$



**Endergebnis: Wahrscheinlichkeit ist Produkt der Nachrichten**

# Inferenz: Message-Passing Algorithmus

- Algorithmus: Message-Passing auf linearer Kette

- ◆ Eingabe:

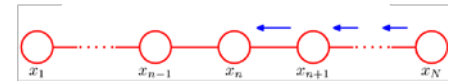
$$p(x_1, \dots, x_N) = \psi_{1,2}(x_1, x_2), \dots, \psi_{N-1,N}(x_{N-1}, x_N)$$

Gesucht:  $p(x_a) = ?$

- ◆ Berechne Nachrichten (rekursiv):

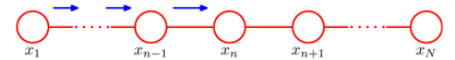
$$\mu_\beta(x_N) = \mathbf{1}$$

$$\text{Für } k = N-1, \dots, a: \quad \mu_\beta(x_k) = \sum_{x_{k+1}} \psi_{k,k+1}(x_k, x_{k+1}) \mu_\beta(x_{k+1})$$



$$\mu_\alpha(x_1) = \mathbf{1}$$

$$\text{Für } k = 2, \dots, a: \quad \mu_\alpha(x_k) = \sum_{x_{k-1}} \psi_{k-1,k}(x_{k-1}, x_k) \mu_\alpha(x_{k-1})$$



- ◆ Ausgabe:

$$p(x_a) = \mu_\alpha(x_a) \mu_\beta(x_a) \quad (\text{Verteilung})$$