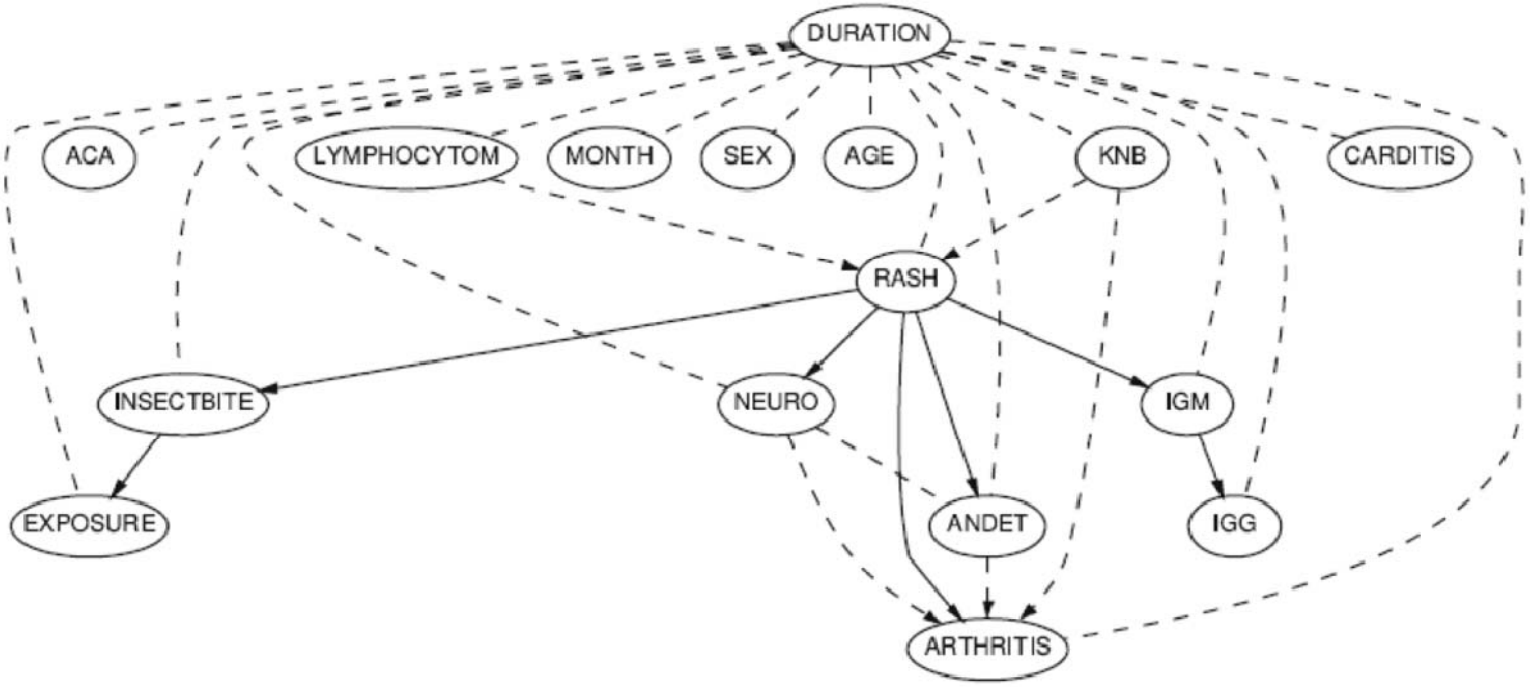
Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Graphische Modelle

Christoph Sawade/Niels Landwehr/Tobias Scheffer

Graphische Modelle

- Modellierung einer Domäne mit verschiedenen Zufallsgrössen
- Gemeinsame Verteilung, insb. Abhängigkeiten



Überblick

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

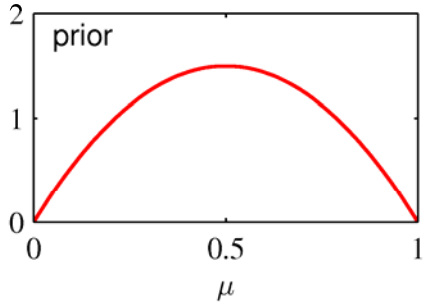
Erinnerung: Parameterschätzung Münzwurf

- Parameterschätzung Münzwurf:
  - Einzelner Münzwurf Bernouilli-verteilt mit Parameter μ
  - Wir haben N unabhängige Münzwürfe gesehen, als Ausprägung $L=\{x_1,\ldots, x_N\}$ der ZV $X_1,\ldots, X_N$
  - Schätzung $\hat{\mu}$ für Münzwurfparameter μ

- Prior: Beta-Verteilung über μ

$$p(\mu \mid \alpha_k, \alpha_z) = Beta(\mu \mid \alpha_k, \alpha_z)$$

α_k, α_z Hyperparameter



- Likelihood

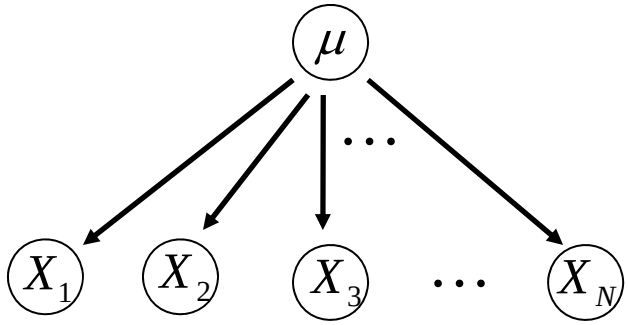
$$p(X_1,\ldots, X_N \mid \mu) = \prod_{i=1}^{N} \underbrace{p(X_i \mid \mu)}_{\text{Bernoulli}}$$

Erinnerung: Parameterschätzung Münzwurf

- Zufallsvariablen in Münzwurfszenario sind $X_1, ..., X_N, \mu$
- Gemeinsame Verteilung von Daten und Parameter (gegeben „Hyperparameter“ α_k, α_z): Prior x Likelihood

$$p_{\alpha_k,\alpha_z}(X_1, ..., X_N, \mu) = p_{\alpha_k,\alpha_z}(\mu) \prod_{i=1}^{N} \underbrace{p(X_i \mid \mu)}_{\text{Bernoulli}}$$

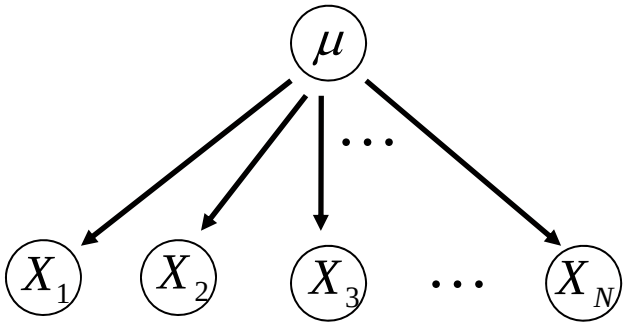
- Darstellung als graphisches Modell:



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

Plate-Modelle

- Unabhängige Münzwürfe: Darstellung als graphisches Modell

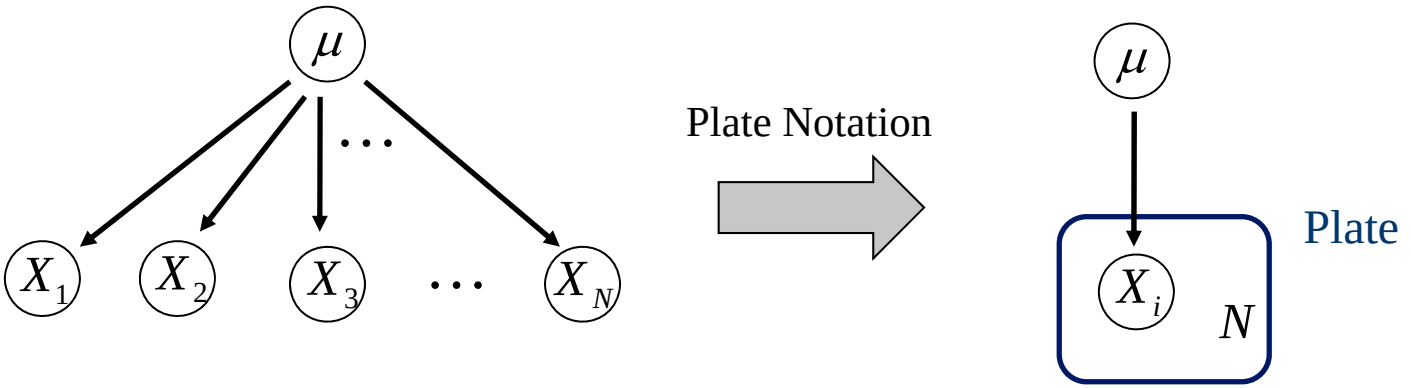


- Knoten $X_1,\ldots, X_N$ sind von gleicher Form
  - Gleicher Wertebereich
  - Gleiche bedingte Verteilungen $p(X_i \mid \mu) = p(X_j \mid \mu).$

- Kurznotation in der Form einer „Schablone“: Plate Notation

Plate-Modelle

- Plate Notation



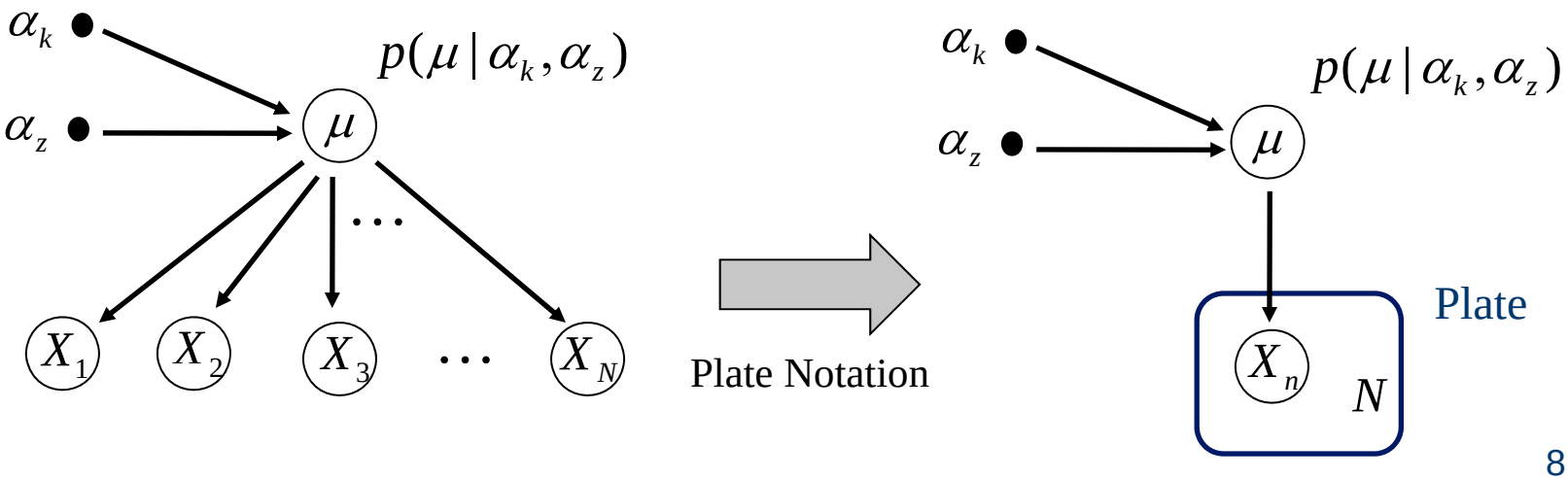
- Ein „Plate“ ist eine abkürzende Notation für N Variablen der gleichen Form
  - Bezeichnet mit Anzahl der Variablen, N
  - Variablen haben Index (z.B. X_i).
- Plate-Modelle werden im Maschinellen Lernen oft verwendet

Plate-Modelle: Hyperparameter

- „Hyperparameter“ α_k, α_z sind keine Zufallsvariablen
  - Wir modellieren nur die gemeinsame Verteilung über $X_1, ..., X_N, \mu$ gegeben Hyperparameter

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

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



Erinnerung: Bayessche Lineare Regression

- Regressionslernen

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

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

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

- Matrixnotation

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

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

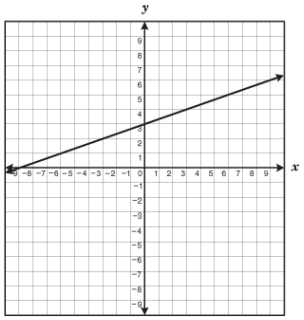
Erinnerung: Bayessche Lineare Regression

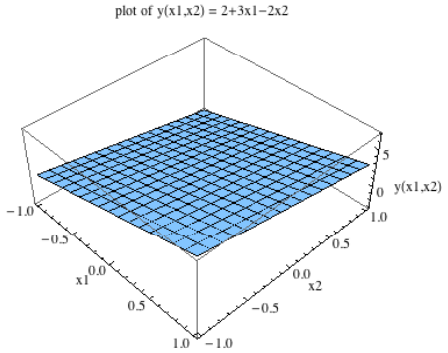
- Lineare Regression

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

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

w „Parametervektor“, „Gewichtsvektor“





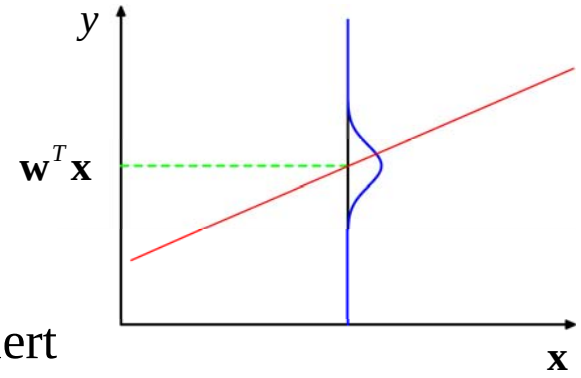
Erinnerung: Bayessche Lineare Regression

- Diskriminatives Setting: $\mathbf{x}_i$ fest, Verteilung über Label y_i: Lineares Modell plus Gaußsches Rauschen

$$p(y \mid \mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x} + N(y \mid 0, \sigma^2)$$
$$= N(y \mid \mathbf{w}^T \mathbf{x}, \sigma^2)$$

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

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



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

$$\underbrace{p(\mathbf{w} \mid L)}_{\text{posterior}} \propto \underbrace{p(\mathbf{y} \mid X, \mathbf{w})}_{\text{likelihood}} \underbrace{p(\mathbf{w})}_{\text{prior}}$$

Erinnerung: Bayessche Lineare Regression

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

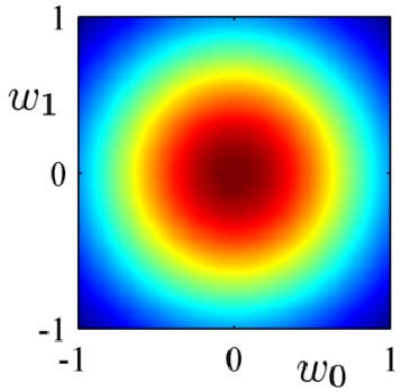
$$p(\mathbf{y} \mid X, \mathbf{w}, \sigma^2) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \mathbf{w}, \sigma^2) \qquad i.i.d.$$

$$= \prod_{i=1}^{N} N(y_i \mid \mathbf{w}^T \mathbf{x}_i, \sigma^2)$$

- Normalverteilter Prior über Modelle

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

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



Isotrope multivariate Normalverteilung, Mittelwert **0,** Varianz τ^2

Bayessche Lineare Regression als Graphisches Modell

- Was sind Zufallsvariablen?
  - Datenpunkte $y_1,\ldots, y_N$, Parameter $\mathbf{w}$
  - Nicht: $\mathbf{x}_1,\ldots, \mathbf{x}_N$, Hyperparameter σ^2, τ^2

- Gemeinsame Verteilung über Daten und Parameter

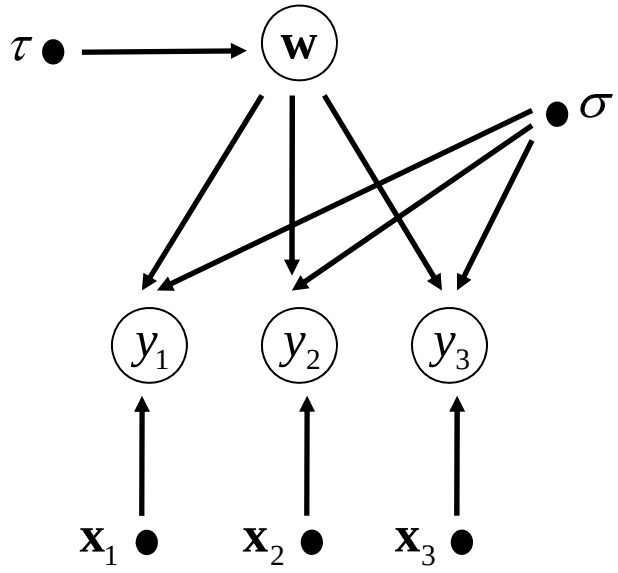
$$
\begin{aligned}
p(y_1,...,y_N,\mathbf{w}\,|\,\mathbf{x}_1,...,\mathbf{x}_N,\sigma^2,\tau^2) &= p(\mathbf{w}\,|\,\tau^2)\,p(y_1,...,y_N\,|\,\mathbf{w},\mathbf{x}_1,...,\mathbf{x}_n,\sigma^2) \\
&= p(\mathbf{w}\,|\,\tau^2)\prod_{n=1}^{N} p(y_n\,|\,\mathbf{w},\mathbf{x}_n,\sigma^2) \qquad i.i.d. \\
&= N(\mathbf{w}\,|\,\mathbf{0},\tau^2 I)\prod_{n=1}^{N} N(y_n\,|\,\mathbf{w}^T\mathbf{x}_n,\sigma^2)
\end{aligned}
$$

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

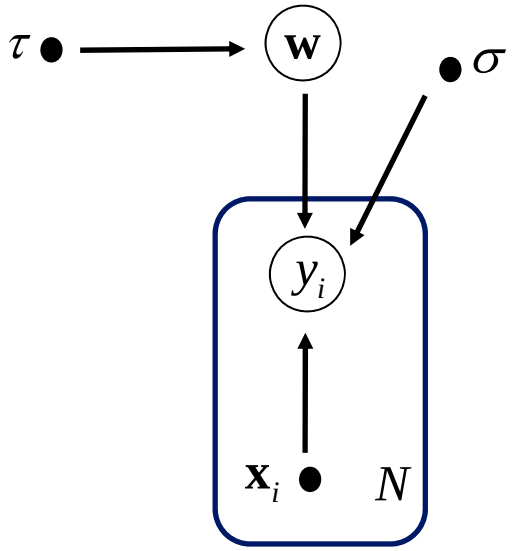
Bayessche Lineare Regression als Graphisches Modell

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

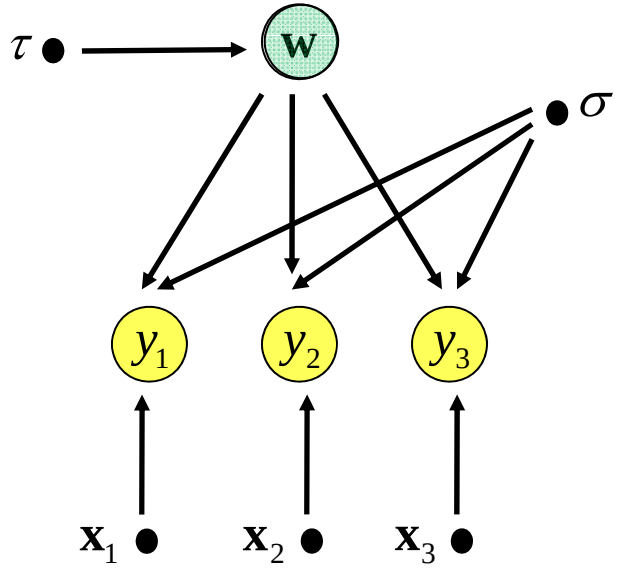
Graphisches Modell, N=3

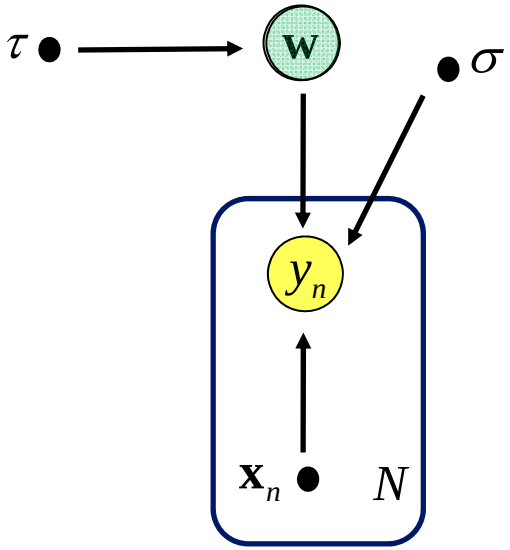


Graphisches Modell, Plate-Notation



MAP Parameterschätzung als Inferenzproblem





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

Bayes-optimale Vorhersage

- Klassifikation mit MAP Modell:

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

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

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

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

- Besser als Klassifikation mit MAP Modell ist Bayessche Vorhersage:

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

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

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)

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

Graphisches Modell, N=3

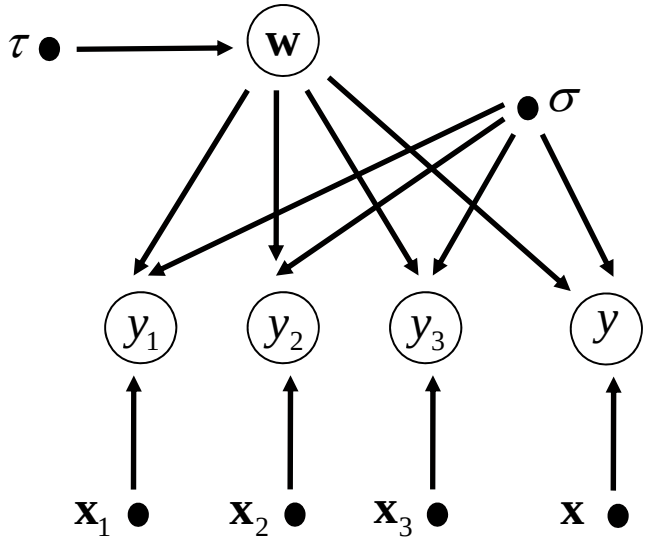
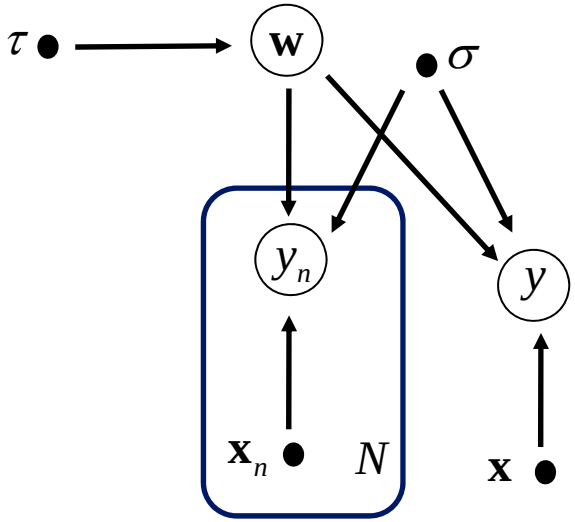
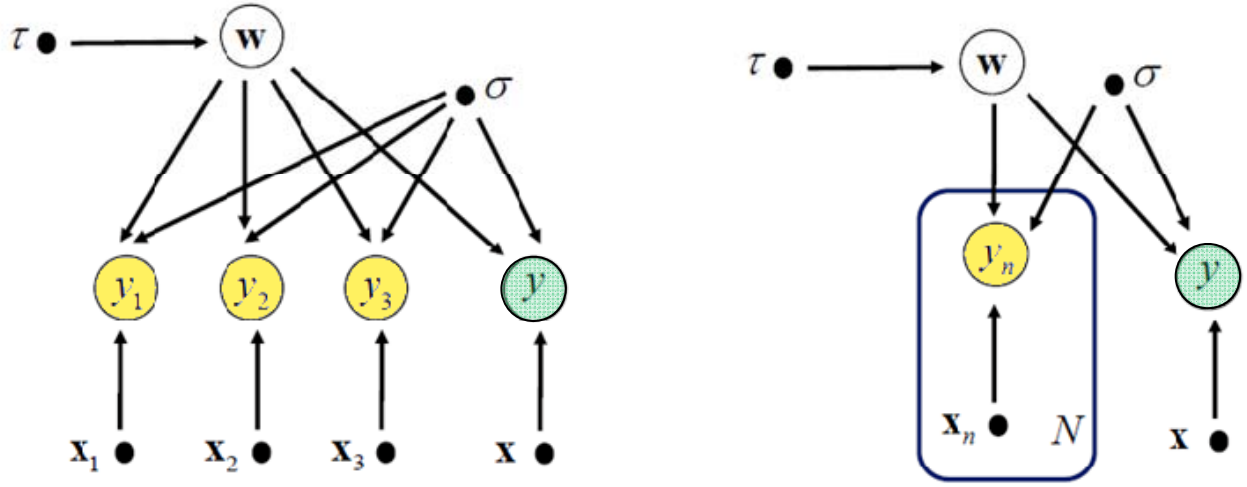


Plate Notation



Bayessche Lineare Regression als Graphisches Modell



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

Graphische Modelle: Zusammenfassung

- Kombination von Wahrscheinlichkeitstheorie und Graphentheorie
- Modellierung der gemeinsamen Verteilung einer Menge von ZV $X_1,\ldots, X_N$
  - Graphstruktur drückt Unabhängigkeitsannahmen aus
  - Kompakt, intuitiv verständlich
  - Bedingte Unabhängigkeiten zwischen Variablenmengen können an der Graphstruktur abgelesen werden (d-separation)
  - Effiziente Inferenzalgorithmen (siehe unten)

Graphische Modelle: Zusammenfassung

- Probabilistische Modelle/Fragestellungen des Maschinellen Lernens können als GM modelliert werden
  - Parameterschätzung Münzwurf
  - Bayessche Lineare Regression
  - MAP Modell, Bayessche Vorhersage: Inferenzprobleme

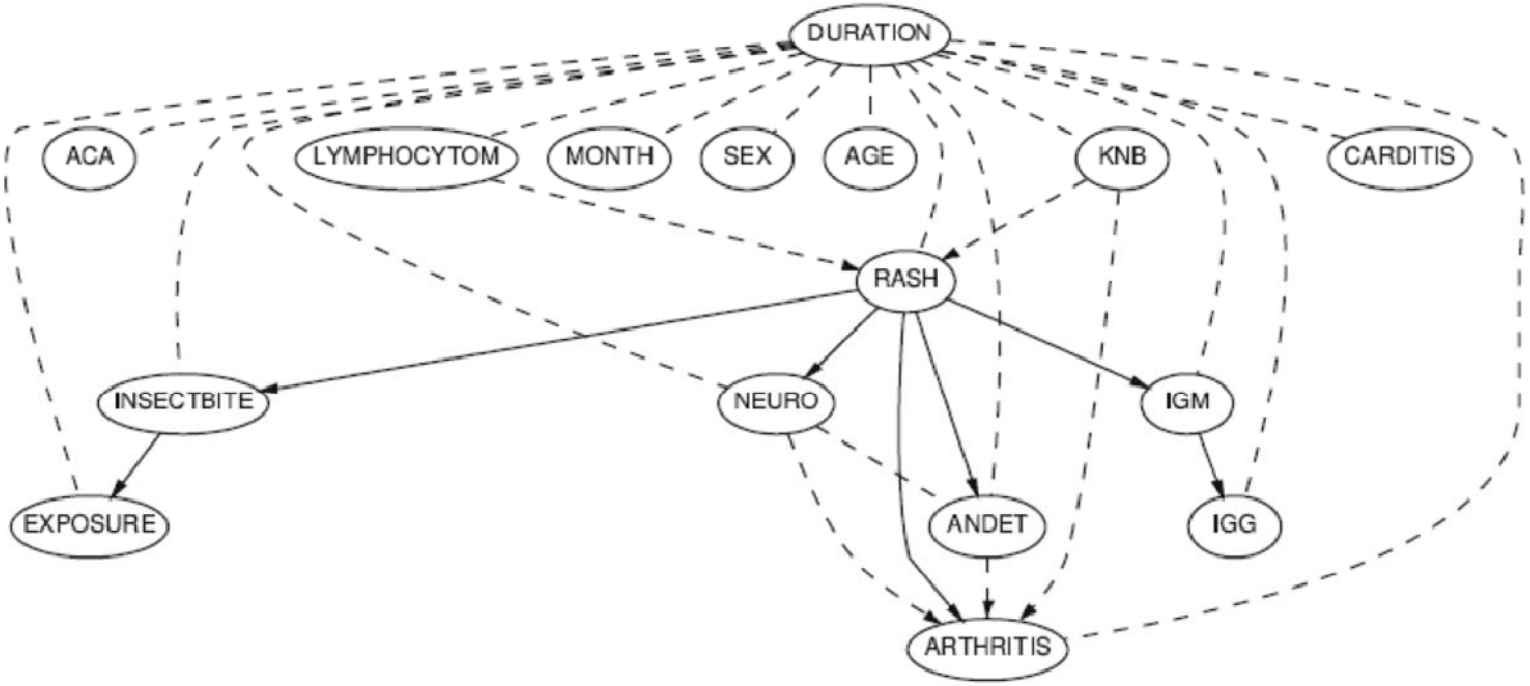
- Entscheidende Fragestellung: Inferenz

Überblick

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

Graphische Modelle: Inferenz

- Wir haben eine Domäne durch gemeinsame Verteilung aller Zufallsgrössen modelliert
- Inferenz: Wahrscheinlichkeit dafür, dass Variablen bestimmte Werte annehmen, gegeben Informationen über andere Variablen?



Problemstellung Inferenz

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

Bedingte Verteilung über ZV X_a

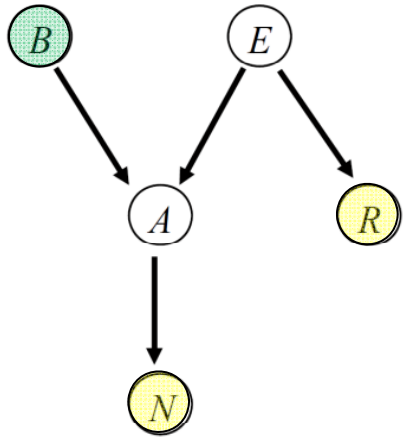
Evidenz: beobachtete Werte für ZV $X_{i_1},\ldots,X_{i_m}$

Berechne $\forall x_a \quad p(x_a \mid x_{i_1},\ldots,x_{i_m})$

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

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 Bayessche Netze: exakte Inferenz NP-hart
  - Es gibt Algorithmen für exakte Inferenz in allgemeinen Bayesschen Netzen, 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

- Bayessches Netz: Repräsentation von $p(X_1, ..., X_N)$

- Naive Inferenz:

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

Berechne für jeden Wert x_a: $p(x_a \mid x_{i_1}, ..., x_{i_m}) = \frac{p(x_a, x_{i_1}, ..., x_{i_m})}{p(x_{i_1}, ..., x_{i_m})}$

Z Normalisierungsfaktor,
leicht explizit zu berechnen
bei univariaten Verteilungen

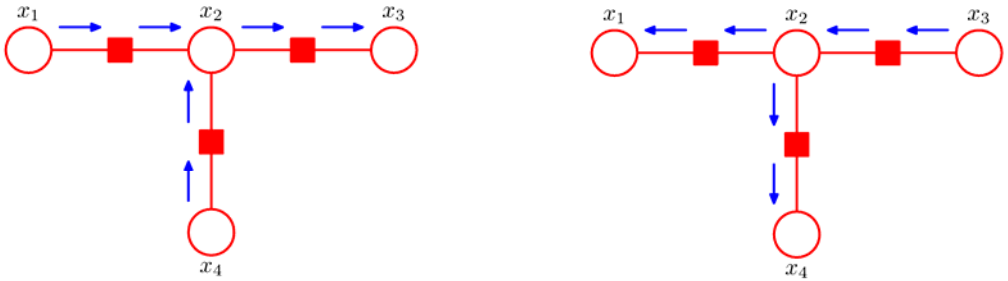
$$= \frac{1}{Z} p(x_a, x_{i_1}, ..., x_{i_m})$$

$$= \frac{1}{Z} \sum_{x_{j_1}} \sum_{x_{j_2}} \cdots \sum_{x_{j_k}} p(x_1, ..., x_N)$$

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

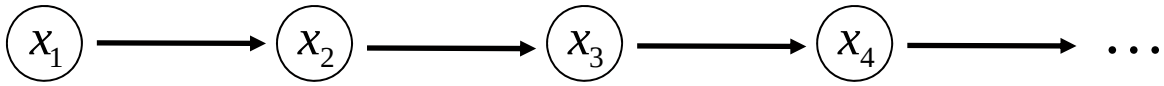
Effizientere Inferenzmethoden?

- 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

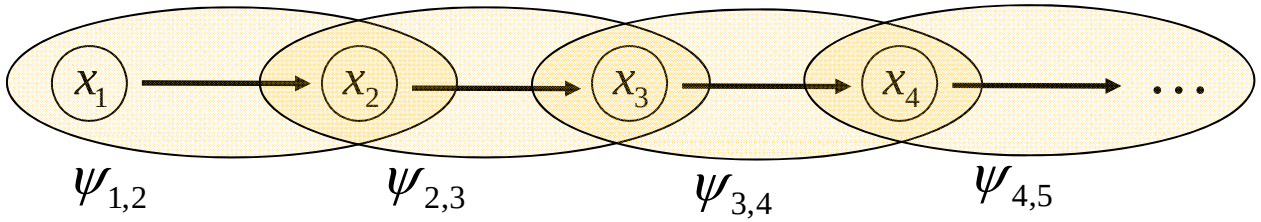
- Wir betrachten zunächst Spezialfall mit besonders einfacher Struktur: lineare Kette von Zufallsvariablen



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

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

$$p(x_1,...,x_N) = \overbrace{p(x_1)p(x_2|x_1)}^{\psi_{1,2}(x_1,x_2)}\overbrace{p(x_3|x_2)}^{\psi_{2,3}(x_2,x_3)}\cdot....\cdot\overbrace{p(x_N|x_{N-1})}^{\psi_{N-1,N}(x_N,x_{N-1})}$$



Inferenz: lineare Kette von ZV

- Zunächst betrachten wir Inferenzaufgabe ohne Evidenz

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

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)

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

$$= \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_N} \psi_{1,2}(x_1, x_2) \cdot ... \cdot \psi_{N-2,N-1}(x_{N-2}, x_{N-1}) \psi_{N-1,N}(x_{N-1}, x_N)$$

$$= \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_{N-1}} \psi_{1,2}(x_1, x_2) \cdot ... \cdot \psi_{N-2,N-1}(x_{N-2}, x_{N-1}) \underbrace{\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)}_{\mu_\beta(x_{N-1})}$$

- Lokale Teilberechnung: „Nachricht“ $\mu_\beta(x_{N-1})$
  - Berechne $\mu_\beta(x_{N-1}) = \sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)$
  - 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

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

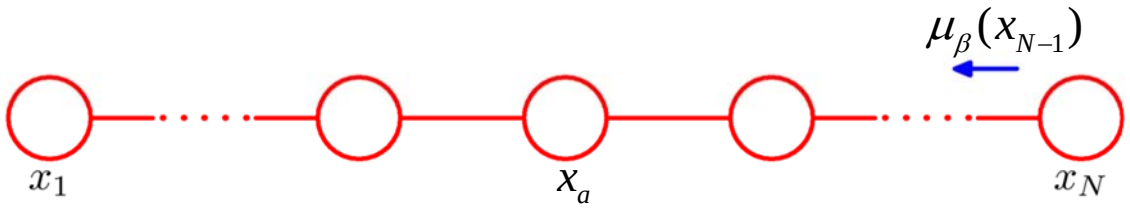
Kodierung z.B. als Vektor: $\mu_\beta(x_{N-1}) = \begin{pmatrix} \sum_{x_N} \psi_{N-1,N}(x_{N-1}=1, x_N) \\ \sum_{x_N} \psi_{N-1,N}(x_{N-1}=2, x_N) \\ \dots \\ \sum_{x_N} \psi_{N-1,N}(x_{N-1}=K, x_N) \end{pmatrix}$

$$\psi_{N-2,N-1}(x_{N-2}, x_{N-1}) \underbrace{\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)}_{\mu_\beta(x_{N-1})}$$

- Lokale Teilber[illegible]nung: „Nachricht“ $\mu_\beta(x_{N-1})$
  - Berechne $\mu_\beta(x_{N-1}) = \sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)$
  - 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

$$p(x_a) = \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_{N-1}} \psi_{1,2}(x_1, x_2) \cdot \ldots \cdot \psi_{N-3,N-2}(x_{N-3}, x_{N-2}) \psi_{N-2,N-1}(x_{N-2}, x_{N-1}) \mu_\beta(x_{N-1})$$

$$= \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_{N-2}} \psi_{1,2}(x_1, x_2) \cdot \ldots \cdot \psi_{N-3,N-2}(x_{N-3}, x_{N-2}) \underbrace{\sum_{x_{N-1}} \psi_{N-2,N-1}(x_{N-2}, x_{N-1}) \mu_\beta(x_{N-1})}_{\mu_\beta(x_{N-2})}$$

$$= \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_{N-2}} \psi_{1,2}(x_1, x_2) \cdot \ldots \cdot \psi_{N-3,N-2}(x_{N-3}, x_{N-2}) \mu_\beta(x_{N-2})$$

- 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

$$p(x_a) = \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_{N-2}} \psi_{1,2}(x_1, x_2) \cdot \ldots \cdot \psi_{N-3,N-2}(x_{N-3}, x_{N-2}) \mu_\beta(x_{N-2})$$
$$= \ldots$$
$$= \sum_{x_1} \cdots \sum_{x_{a-1}} \psi_{1,2}(x_1, x_2) \cdot \ldots \cdot \psi_{a-1,a}(x_{a-1}, x_a) \mu_\beta(x_a)$$

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

$$\mu_\beta(x_{N-3}) = \sum_{x_{N-2}} \psi_{N-3,N-2}(x_{N-3}, x_{N-2}) \mu_\beta(x_{N-2})$$

X_N , X_{N-1} ,X_{N-2} aussummiert

$$\ldots$$

$$\mu_\beta(x_a) = \sum_{x_{a+1}} \psi_{a,a+1}(x_a, x_{a+1}) \mu_\beta(x_{a+1})$$

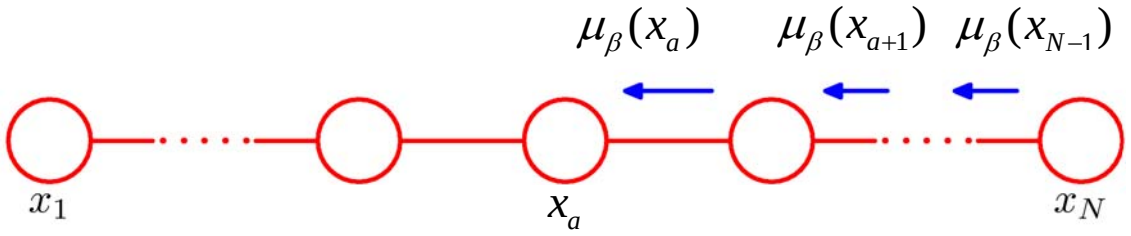
X_N ,…, X_{a+1} aussummiert

Inferenz: Message-Passing

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

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

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



- Verbleibende Variablen aussummieren: Nachrichten von links

Inferenz: Message-Passing

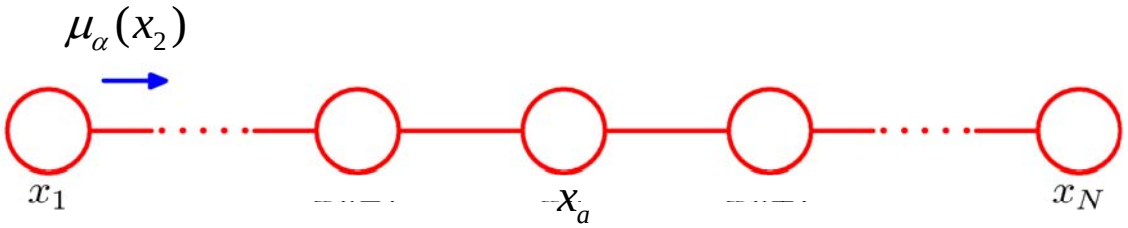
- Gleiches Prinzip anwenden auf die Variablen links vom Anfrageknoten

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

- 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

$$p(x_a) = \mu_\beta(x_a) \sum_{x_{a-1}} \cdots \sum_{x_3} \sum_{x_2} \psi_{a-1,a}(x_{a-1}, x_a) \cdot \ldots \cdot \psi_{3,4}(x_3, x_4) \psi_{2,3}(x_2, x_3) \mu_\alpha(x_2)$$

$$= \mu_\beta(x_a) \sum_{x_{a-1}} \cdots \sum_{x_3} \psi_{a-1,a}(x_{a-1}, x_a) \cdot \ldots \cdot \psi_{3,4}(x_3, x_4) \underbrace{\sum_{x_2} \psi_{2,3}(x_2, x_3) \mu_\alpha(x_2)}_{\mu_\alpha(x_3)}$$

$$= \ldots$$

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

- Rekursive Teilberechnungen:

Berechne $\mu_\alpha(x_3) = \sum_{x_2} \psi_{2,3}(x_2, x_3)\, \mu_\alpha(x_2)$ $\quad X_1, X_2$ aussummiert

…

Berechne $\mu_\alpha(x_a) = \sum_{x_{a-1}} \psi_{a-1,a}(x_{a-1}, x_a) \mu_\alpha(x_{a-1})$ $\quad X_1, \ldots, X_{a-1}$ aussummiert

Inferenz: Message-Passing

- Auf diese Art alle verbleibenden Variablen aussummieren

$$p(x_a) = \mu_\beta(x_a) \sum_{x_{a-1}} \cdots \sum_{x_3} \sum_{x_2} \psi_{a-1,a}(x_{a-1}, x_a) \cdot \ldots \cdot \psi_{3,4}(x_3, x_4) \psi_{2,3}(x_2, x_3) \mu_\alpha(x_2)$$

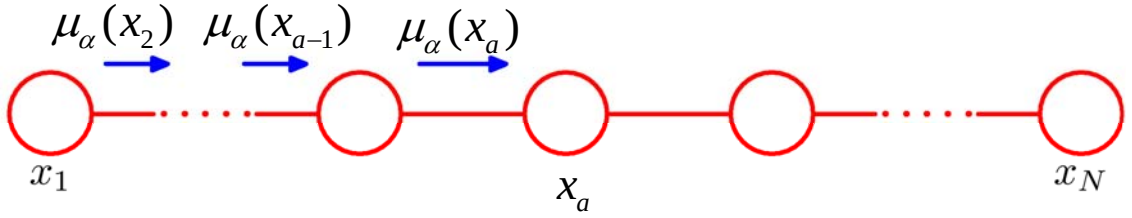
$$= \mu_\beta(x_a) \sum_{x_{a-1}} \cdots \sum_{x_3} \psi_{a-1,a}(x_{a-1}, x_a) \cdot \ldots \cdot \psi_{3,4}(x_3, x_4) \underbrace{\sum_{x_2} \psi_{2,3}(x_2, x_3) \mu_\alpha(x_2)}_{\mu_\alpha(x_3)}$$

$$= \ldots$$

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

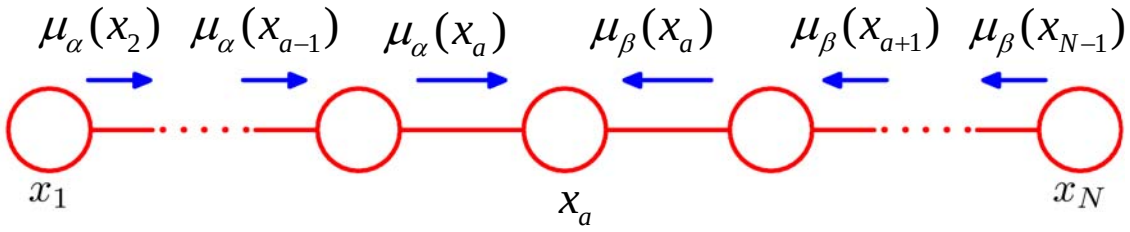
Endergebnis: Wahrscheinlichkeit ist Produkt der Nachrichten

- Veranschaulichung



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

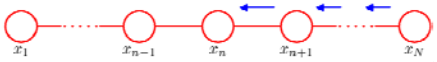
- Algorithmus: Message-Passing auf linearer Kette
  - Eingabe:

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

    Gesucht: $p(x_a) = ?$
  - Berechne Nachrichten (rekursiv):

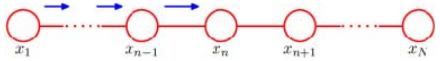
    $\mu_\beta(x_N) = \mathbf{1}$

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

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

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

  - Ausgabe:

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

Inferenz: Message-Passing

- Laufzeit:
  - Berechnung einer Nachricht:

$$\forall\, x_k: \quad \mu_\beta(x_k) = \sum_{x_{k+1}} \psi_{k,k+1}(x_k, x_{k+1})\mu_\beta(x_{k+1})$$

$$\Rightarrow O(K^2) \text{ für Berechnung einer Nachricht (K diskrete Zustände)}$$

  - N Nachrichten insgesamt

$$\Rightarrow O(NK^2) \text{ Gesamtlaufzeit}$$

  - Viel besser als naive Inferenz mit $O(K^N)$

Grund für Effiziente Inferenz: Struktur der gemeinsamen Verteilung

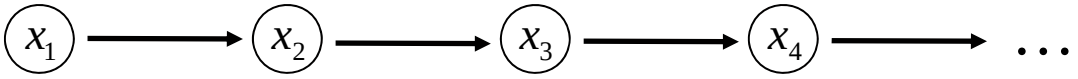
- Partielles, rekursives Aussummieren möglich aufgrund der einfachen Struktur der Verteilung

Faktorisierung der Gesamtverteilung erlaubt „Vorziehen“ der Summen

$$p(x_a) = \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_N} \psi_{1,2}(x_1, x_2) \cdot \ldots \cdot \psi_{N-2,N-1}(x_{N-2}, x_{N-1}) \psi_{N-1,N}(x_{N-1}, x_N)$$

$$= \sum_{x_1} \cdots \sum_{x_{a-1}} \sum_{x_{a+1}} \cdots \sum_{x_{N-1}} \psi_{1,2}(x_1, x_2) \cdot \ldots \cdot \psi_{N-2,N-1}(x_{N-2}, x_{N-1}) \underbrace{\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)}_{\mu_\beta(x_{N-1})}$$

$$p(x_1, \ldots, x_N) = \overbrace{p(x_1)p(x_2|x_1)}^{\psi_{1,2}(x_1,x_2)} \overbrace{p(x_3|x_2)}^{\psi_{2,3}(x_2,x_3)} \cdot \ldots \cdot \overbrace{p(x_N|x_{N-1})}^{\psi_{N-1,N}(x_N,x_{N-1})}$$



- Für weniger einfache Strukturen wird Inferenz aufwendiger

Inferenz: Message-Passing

- Inferenz für alle Marginalverteilungen $p(x_1),\ldots,p(x_N)$
  - Message-Passing Algorithmus N Mal durchführen: $O(N^2K^2)$
  - Geht besser:
    - Beta-Nachrichten für Knoten x_{N-1} bis x_1 berechnen

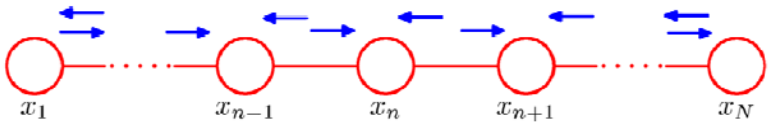
$$\mu_\beta(x_N)=\mathbf{1}$$

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

    - Alpha-Nachrichen für Knoten x_2 bis x_N berechnen

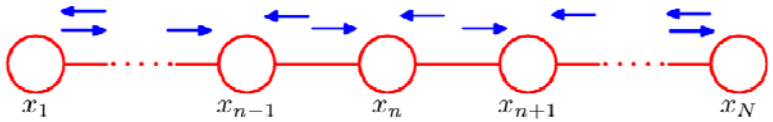
$$\mu_\alpha(x_1)=\mathbf{1}$$

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



Inferenz: Message-Passing

- Inferenz für alle Marginalverteilungen $p(x_1),..., p(x_N)$



- Jetzt gilt

$$a = 1,..., N: \quad p(x_a) = \mu_\alpha(x_a)\mu_\beta(x_a)$$

  - Randverteilungen für alle (einzelnen) Zufallsvariablen gleichzeitig berechnet
  - Gesamtlaufzeit $O(NK^2)$