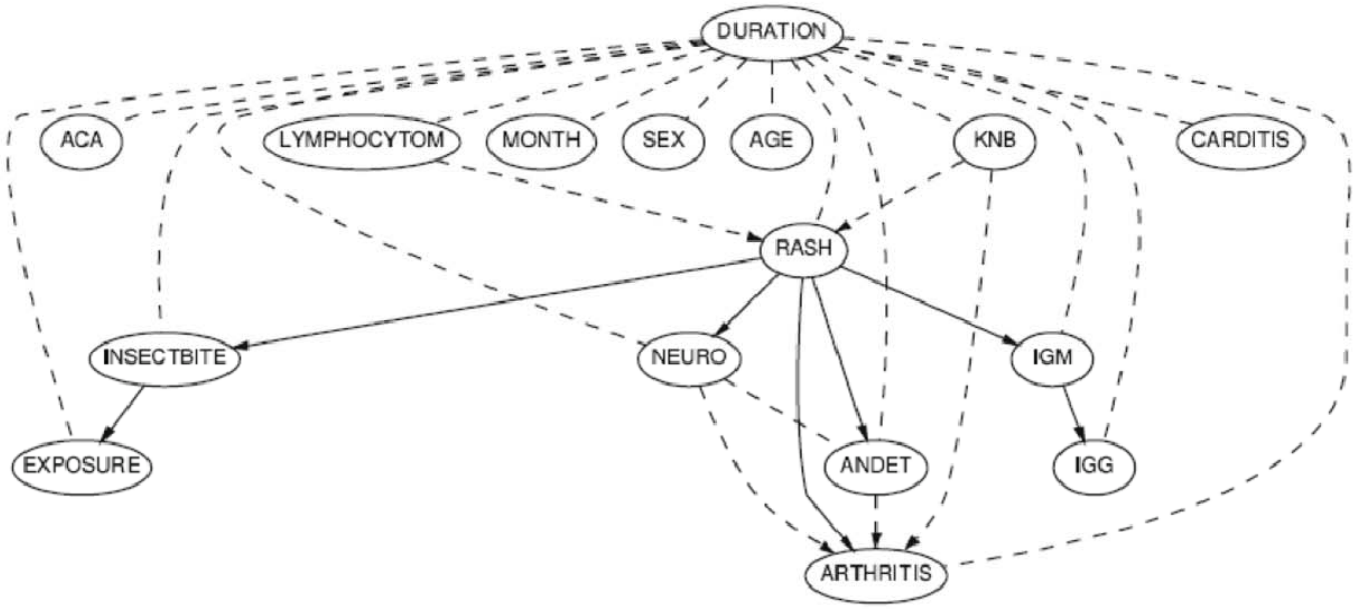
Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Graphische Modelle

Christoph Sawade/Niels Landwehr/Tobias Scheffer

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?



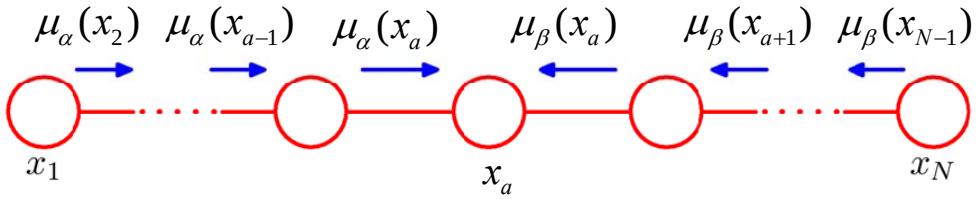
Überblick

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

Exakte Inferenz: Message-Passing

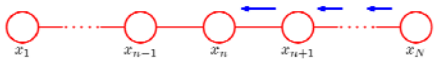
- Message Passing Algorithmus auf linearer Kette

$$p(\mathbf{x}) = \prod_{i=1}^{N} \psi_{i,i+1}(x_i, x_{i+1})$$



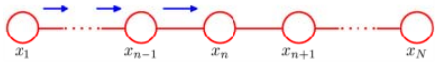
$$\mu_\beta(x_N) = \mathbf{1}$$

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



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

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



$p(x_a) = \mu_\beta(x_a)\mu_\alpha(x_a)$ ← Randverteilung über Anfragevariable x_a: Produkt der Nachrichten

Message-Passing mit Evidenz

- Bisher Randverteilung $p(x_a)$ ohne Evidenz bestimmt
- Was ist wenn wir Evidenz haben?

$$\text{Notation}: \{x_1, ..., x_N\} = \{ \underbrace{x_a}_{\substack{\text{Anfrage-}\\ \text{Variable}}} , \underbrace{x_{i_1}, ..., x_{i_m}}_{\text{Evidenz-Variablen}} , \underbrace{x_{j_1}, ..., x_{j_k}}_{\text{restliche Variablen}} \}$$

- Bedingte Verteilung

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

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

Z einfach zu berechnen

(Normalisierer univariate Verteilung)

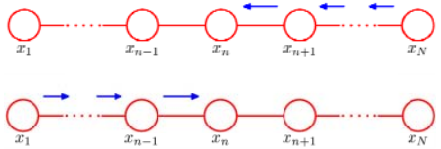
$$= \frac{1}{Z} \sum_{x_{j_1}} ... \sum_{x_{j_k}} p(\underbrace{x_{j_1}, ..., x_{j_k}}_{\text{aussummieren}}, \underbrace{x_a, x_{i_1}, ..., x_{i_m}}_{\text{fest}})$$

Message-Passing mit Evidenz

- Ziel: $p(x_a, x_{i_1}, ..., x_{i_m}) = ?$
- Leichte Modifikation des Message-Passing Algorithmus
  - Wir berechnen wie bisher Nachrichten

$$\mu_\beta(x_{N-1}), ..., \mu_\beta(x_a)$$

$$\mu_\alpha(x_2), ..., \mu_\alpha(x_a)$$



  - Falls x_{k+1} unbeobachtet ist, summieren wir diesen Knoten aus

$$k+1 \notin \{i_1, ..., i_m\} \quad \Rightarrow \quad \mu_\beta(x_k) = \sum_{x_{k+1}} \psi_{k,k+1}(x_k, x_{k+1}) \mu_\beta(x_{k+1})$$

  - Falls x_{k+1} beobachtet ist, verwenden wir nur den entsprechenden Summanden

x_{k+1} beobachteter Wert (Evidenz)

$$k+1 \in \{i_1, ..., i_m\} \quad \Rightarrow \quad \mu_\beta(x_k) = \psi_{k,k+1}(x_k, x_{k+1}) \mu_\beta(x_{k+1})$$

Message-Passing mit Evidenz

- Ebenso für $\mu_\alpha(x_k)$

$$\mu_\alpha(x_k) = \begin{cases} \sum_{x_{k-1}} \psi_{k-1,k}(x_{k-1}, x_k)\mu_\alpha(x_{k-1}) : k-1 \notin \{i_1, ..., i_m\} & \text{(Knoten nicht beobachtet)} \\ \psi_{k-1,k}(x_{k-1}, x_k)\mu_\alpha(x_{k-1}) : k-1 \in \{i_1, ..., i_m\} & \text{(Knoten beobachtet)} \end{cases}$$

- Jetzt gilt

$$p(x_a, x_{i_1}, ..., x_{i_m}) = \mu_\alpha(x_a)\mu_\beta(x_a)$$

- Laufzeit für Inferenz mit Evidenz immer noch $O(NK^2)$

Beispiel: Markov Modelle

- Beispiel für Inferenz auf linearer Kette: Markov Modelle

- Markov Modelle: einfache Modelle für dynamische probabilistische Prozesse
  - Prozess, der verschiede Zustände annehmen kann
  - Zufallsvariable X_t repräsentiert den Zustand zur Zeit t
  - Diskrete Zeitpunkte $t=1,...,T$

- Beispiel: Wetter
  - Zufallsvariable X_t = Wetter am Tag t
  - Zwei mögliche Zustände, Regen und Sonne

Beispiel: Markov Modelle

- Modellierung:
  - Prozess wird in einem zufällig gewählten Zustand gestartet:

    Verteilung über Startzustände $p(x_1)$

  - In jedem Schritt geht der Prozess in einen neuen Zustand über, abhängig nur vom gegenwärtigen Zustand (vereinfachende Annahme!)

    Verteilung über nächsten Zustand $p(x_{t+1} \mid x_t)$

- Unabhängigkeitsannahme:

  $\forall t: \; p(x_{t+1} \mid x_1, ..., x_t) = p(x_{t+1} \mid x_t)$ "Markov" Eigenschaft

- Übergangswahrscheinlichkeiten hängen nicht von t ab:

  $\forall t: \; p(x_{t+1} \mid x_t) = p(x_t \mid x_{t-1})$ "Homogener" Prozess

Beispiel: Markov Modelle

- Gemeinsame Wahrscheinlichkeit über alle Zustände

$$p(x_1,...,x_T) = \prod_{t=1}^{T} p(x_t \mid x_1,...,x_{t-1})$$

$$= p(x_1) \prod_{t=2}^{T} p(x_t \mid x_{t-1})$$

„Zukunft ist unabhängig von der Vergangenheit gegeben Gegenwart“

- Darstellung als graphisches Modell:

Beispiel: Markov Modelle

- Gemeinsame Wahrscheinlichkeit über alle Zustände

$$p(x_1,...,x_T) = \prod_{t=1}^{T} p(x_t \mid x_1,...,x_{t-1})$$

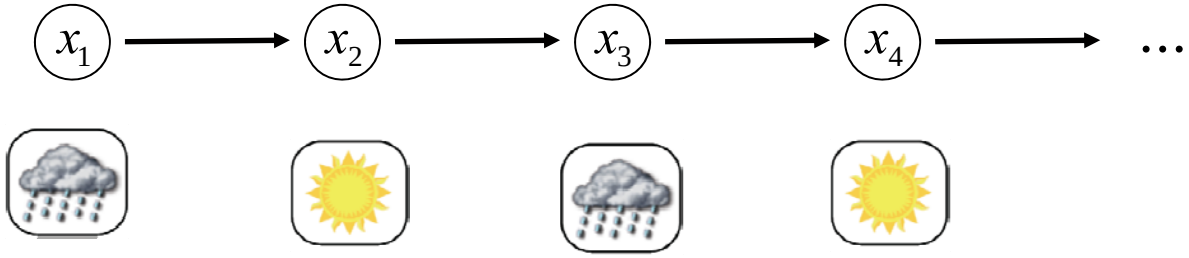
$$= p(x_1) \prod_{t=2}^{T} p(x_t \mid x_{t-1})$$

„Zukunft ist unabhängig von der Vergangenheit gegeben Gegenwart“

- Darstellung als graphisches Modell:

Beispiel: Markov Modelle

- Beispiel Markov Modell:
  - Zustand x_t = Wetter am Tag t
  - Zwei mögliche Zustände, Regen und Sonne



  - Verteilungen

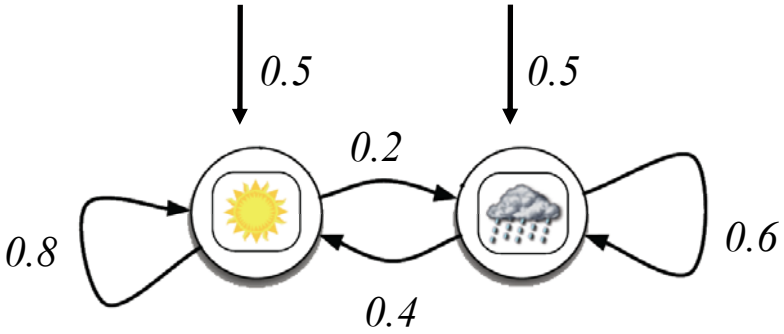
$$p(x_1 = s) = 0.5$$
$$p(x_1 = r) = 0.5$$

$$p(x_t = s \mid x_{t-1} = s) = 0.8$$
$$p(x_t = r \mid x_{t-1} = s) = 0.2$$
$$p(x_t = s \mid x_{t-1} = r) = 0.4$$
$$p(x_t = r \mid x_{t-1} = r) = 0.6$$

Wahrscheinlichkeit, dass morgen die Sonne scheint, gegeben dass es heute regnet.

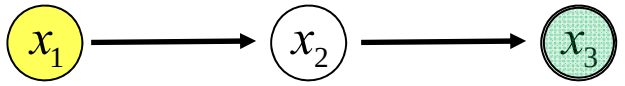
Beispiel: Markov Modelle

- Markov Modelle entsprechen probabilistischen endlichen Automaten



- Beispiel Inferenzproblem:
  - Wie ist das Wetter übermorgen, gegeben dass es heute regnet?

$$p(x_3 \mid x_1 = r) = ?$$



Beispiel: Markov Modelle

- Berechnung mit Message-Passing Algorithmus

- Zu berechnende Nachrichten: $\mu_\alpha(x_1), \mu_\alpha(x_2), \mu_\alpha(x_3);\ \ \mu_\beta(x_3)$

$$\mu_\alpha(x_k) = \begin{cases} \sum_{x_{k-1}} \psi_{k-1,k}(x_{k-1}, x_k)\mu_\alpha(x_{k-1}) : k-1 \notin \{i_1, \ldots, i_m\} & \text{(Knoten nicht beobachtet)} \\ \psi_{k-1,k}(x_{k-1}, x_k)\mu_\alpha(x_{k-1}) : k-1 \in \{i_1, \ldots, i_m\} & \text{(Knoten beobachtet)} \end{cases}$$

Initialisierung: $\mu_\alpha(x_1) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

x_1 beob. Knoten: $\mu_\alpha(x_2) = p(x_1)p(x_2 \mid x_1)\mu_\alpha(x_1) = \begin{pmatrix} 0.5 \cdot 0.4 \\ 0.5 \cdot 0.6 \end{pmatrix} = \begin{pmatrix} 0.2 \\ 0.3 \end{pmatrix}$ ($\mu_\alpha(x_2 = s)$, $\mu_\alpha(x_2 = r)$)

x_2 unbeob. Knoten: $\mu_\alpha(x_3) = \sum_{x_2 \in \{s,r\}} p(x_3 \mid x_2)\mu_\alpha(x_2) = \begin{pmatrix} 0.8 \cdot 0.2 + 0.4 \cdot 0.3 \\ 0.2 \cdot 0.2 + 0.6 \cdot 0.3 \end{pmatrix} = \begin{pmatrix} 0.28 \\ 0.22 \end{pmatrix}$ ($\mu_\alpha(x_3 = s)$, $\mu_\alpha(x_3 = r)$)

Beispiel: Markov Modelle

- Berechnung mit Message-Passing Algorithmus

- Nachrichten: $\mu_\alpha(x_1), \mu_\alpha(x_2), \mu_\alpha(x_3); \quad \mu_\beta(x_3)$

Initialisierung: $\mu_\beta(x_3) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$$p(x_3 \mid x_1 = r) = \frac{1}{Z}\mu_\alpha(x_3)\mu_\beta(x_3) = \frac{1}{Z}\begin{pmatrix} 0.28 \\ 0.22 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{Z}\begin{pmatrix} 0.28 \\ 0.22 \end{pmatrix}$$

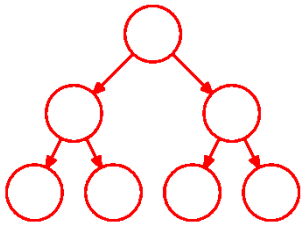
$$Z = 0.28 + 0.22 = 0.5$$

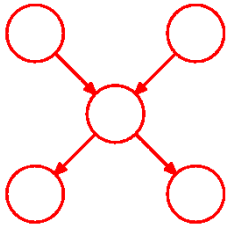
$$p(x_3 = s \mid x_1 = r) = 0.56$$
$$p(x_3 = r \mid x_1 = r) = 0.44$$

Inferenz in Allgemeinen Graphen

- Bisher nur Spezialfall: Inferenz auf linearer Kette
- Die Grundidee des Message-Passing funktioniert auch auf allgemeineren Graphen
- Erweiterung: Exakte Inferenz auf *Polytrees*
  - Polytree: Gerichteter Graph, in dem es zwischen zwei Knoten immer genau einen ungerichteten Pfad gibt
  - Etwas allgemeiner als gerichteter Baum

Gerichteter Baum



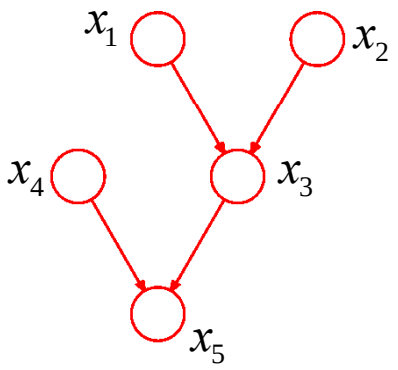


Polytree

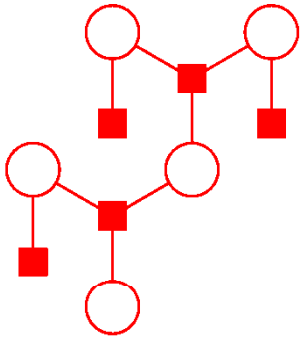
Inferenz in Allgemeinen Graphen

- Grundidee Message-Passing auf Polytrees:
  - Umwandlung in *Faktor-Graph* (ungerichteter Baum)

Ursprünglicher Graph



Faktor-Graph



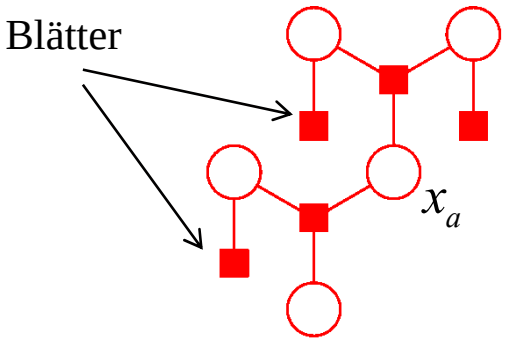
Gemeinsame Verteilung

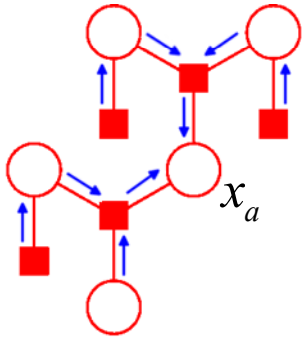
$$p(x_1, x_2, x_3, x_4, x_5) =$$
$$p(x_1)\,p(x_2)\,p(x_3 \mid x_1, x_2)\,p(x_4)\,\underbrace{p(x_5 \mid x_3, x_4)}_{Faktor}$$

- Faktor-Knoten
  - Für jeden Faktor in der gemeinsamen Verteilung gibt es einen Faktor-Knoten
  - Ungerichtete Kanten von den Faktor-Knoten zu den im Faktor auftauchenden Variablen

Inferenz in Allgemeinen Graphen (Skizze)

- Falls der ursprüngliche Graph ein Polytree war, ist der Faktor-Graph ein ungerichteter Baum (dh zykelfrei).

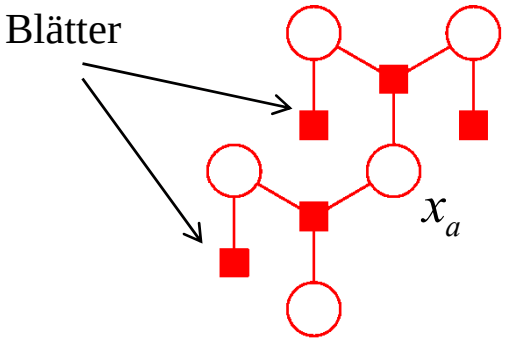


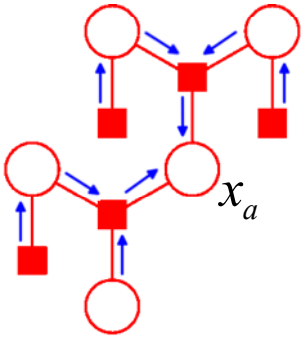


Spezialfall
lineare Kette

- Betrachten Anfragevariable x_a als Wurzel des Baumes
- Nachrichten von den Blättern zur Wurzel schicken (immer eindeutiger Pfad, weil Baum)
- Es gibt zwei Typen von Nachrichten: Faktor-Nachrichten und Variablen-Nachrichten

Inferenz in Allgemeinen Graphen (Skizze)

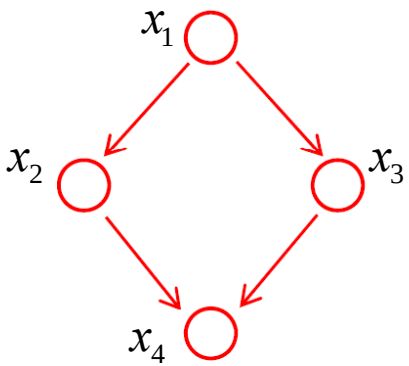


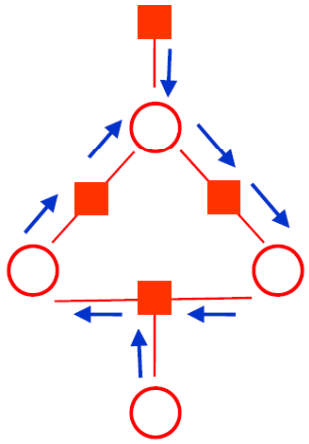


- Nachrichten werden „verschmolzen“, dabei müssen wir über mehrere Variablen summieren
- Grundidee dieselbe wie bei Inferenz auf der linearen Kette: geschicktes Aussummieren
- Laufzeit abhängig von Graphstruktur, exponentiell im worst-case
- Details im Bishop-Textbuch („Sum-Product“ Algorithmus)

Inferenz in Allgemeinen Graphen

- Inferenz in Graphen, die keine Polytrees sind?
- Approximativer Ansatz: Iteratives Message-Passing Schema, wegen Zyklen im Graph nicht exakt





„Loopy Belief Propagation“

$$p(\mathbf{x}) = p(x_1)\,p(x_2 \mid x_1)\,p(x_3 \mid x_1)\,p(x_4 \mid x_2, x_3)$$

- Alternative für exakte Inferenz in allgemeinen Graphen:
  - Graph in einen äquivalenten azyklischen Graphen umwandeln
  - „Junction Tree“ Algorithmus, (i.A. exponentielle Laufzeit)

Überblick

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

Approximative Inferenz

- Exakte Inferenz NP-hart: In der Praxis spielen *approximative* Inferenzverfahren wichtige Rolle

- Wir betrachten Sampling-basierte Verfahren
  - Relativ einfach zu verstehen/implementieren
  - Anytime-Algorithmen (je länger die Laufzeit, desto genauer)

Inferenz: Sampling-basiert

- Grundidee Sampling:
  - Wir interessieren uns für eine Verteilung $p(\mathbf{z})$, $\mathbf{z}$ ist eine Menge von Zufallsvariablen (z.B. bedingte Verteilung über Anfragevariablen in graphischem Modell)
  - Es ist schwierig, $p(\mathbf{z})$ direkt auszurechnen
  - Stattdessen ziehen wir „Samples“ (Stichproben)

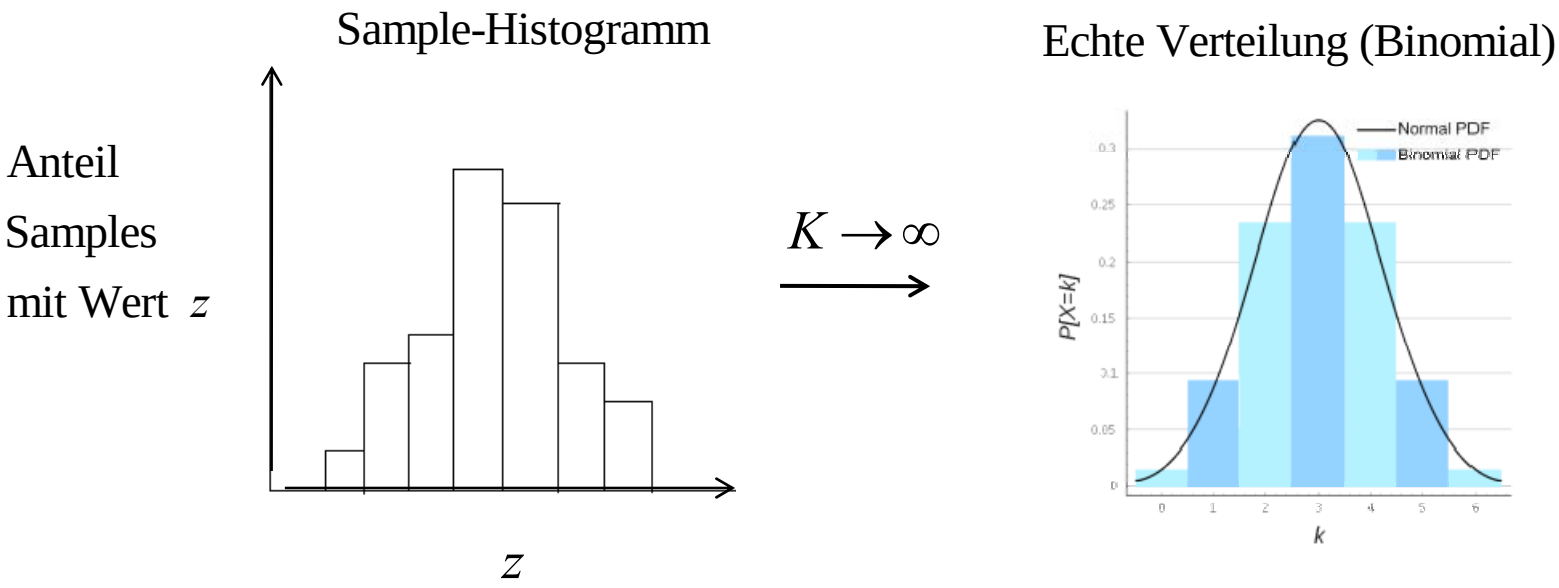
$$\mathbf{z}^{(k)} \sim p(\mathbf{z}) \qquad \text{i.i.d.,} \quad k = 1, \ldots, K,$$

    jedes Sample $\mathbf{z}^{(k)}$ ist eine vollständige Belegung der Zufallsvariablen in $\mathbf{z}$

- Die Samples $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(K)}$ approximieren die Verteilung $p(\mathbf{z})$

Inferenz: Sampling-basiert

- Beispiel:
  - Eindimensionale Verteilung, $\mathbf{z} = \{z\}$
  - Diskrete Variable mit Zuständen {0,…,6}: Anzahl „Kopf“ bei 6 Münzwürfen



Sampling-Inferenz für Graphische Modelle

- Gegeben graphisches Modell, repräsentiert Verteilung durch

$$p(x_1,...,x_N) = \prod_{i=1}^{N} p(x_i \mid pa(x_i))$$

- Etwas allgemeinere Inferenz-Problemstellung: Menge von Anfragevariablen

$$p(\mathbf{x}_I \mid \mathbf{x}_D) \approx ?$$

$\mathbf{x}_I \subseteq \mathbf{x} = \{x_1,...,x_N\}$ Menge von Anfragevariablen

$\mathbf{x}_D \subseteq \mathbf{x} = \{x_1,...,x_N\}$ Menge von Evidenzvariablen

- Wir betrachten zunächst den Fall ohne Evidenz:

$$p(\mathbf{x}_I) \approx ? \qquad \mathbf{x}_I = \{x_{i_1},...,x_{i_m}\} \subseteq \{x_1,...,x_N\}$$

Sampling-Inferenz für Graphische Modelle

- Ziel: Samples aus der Randverteilung $p(\mathbf{x}_I) = p(x_{i_1}, ..., x_{i_m})$

$$\mathbf{x}_I^{(k)} \sim p(\mathbf{x}_I) \qquad k = 1, ..., K$$

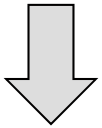
- Es genügt, Samples aus der Gesamtverteilung $p(\mathbf{x}) = p(x_1, ..., x_N)$ zu ziehen:

$$\mathbf{x}^{(k)} = (x_1^{(k)}, ..., x_N^{(k)}) \sim p(x_1, ..., x_N) \qquad k = 1, ..., K$$

- Samples aus der Randverteilung $p(x_{i_1}, ..., x_{i_m})$ erhalten wir einfach durch Projektion der Samples auf die Menge $\{x_{i_1}, ..., x_{i_m}\}$

.

$$\mathbf{x}^{(k)} = (x_1^{(k)}, ..., x_N^{(k)}) \sim p(x_1, ..., x_N) \qquad k = 1, ..., K$$



Projektion

$$\mathbf{x}_I^{(k)} = (x_{i_1}^{(k)}, ..., x_{i_m}^{(k)}) \sim p(x_{i_1}, ..., x_{i_m}) \qquad k = 1, ..., K$$

Inferenz: Ancestral Sampling

- Wie ziehen wir Samples $\mathbf{x}^{(k)} \sim p(\mathbf{x})$?

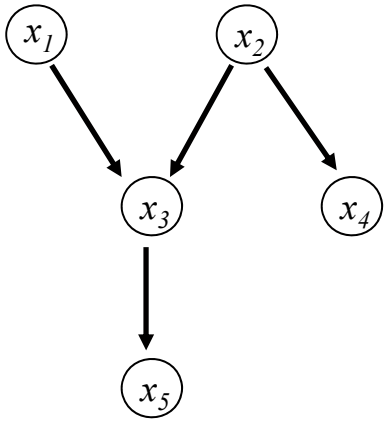
- Einfach bei gerichteten graphischen Modellen: „Ancestral Sampling“
  - Nutze Faktorisierung der gemeinsamen Verteilung

$$\mathbf{x}^{(k)} \sim p(\mathbf{x}) = p(x_1, ..., x_N)$$
$$= \prod_{n=1}^{N} p(x_n \mid pa(x_n))$$

  - Annahme: $pa(x_n) \subseteq \{x_1, ..., x_{n-1}\}$ (sonst umbenennen)

  - „Ziehen entlang der Kanten“

„Ziehen entlang der Kanten“



Inferenz: Ancestral Sampling

- Wir ziehen ein Sample $(x_1^{(k)},...,x_N^{(k)})$, indem wir nacheinander die einzelnen $x_i^{(k)}$ ziehen

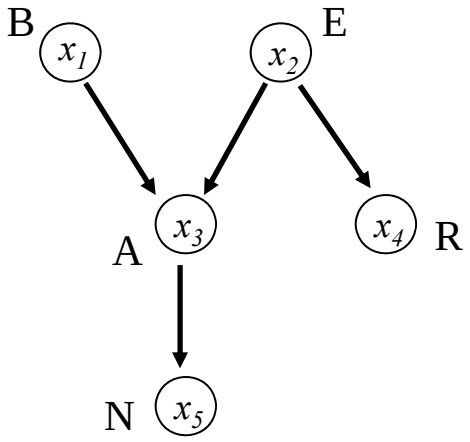
$$x_1^{(k)} \sim p(x_1)$$
$$x_2^{(k)} \sim p(x_2 \mid pa(x_2))$$
$$...$$
$$x_N^{(k)} \sim p(x_N \mid pa(x_N))$$

Schon gezogene Werte

$$\mathbf{x}^{(k)} \sim p(\mathbf{x}) = p(x_1,.,x_N) = \prod_{n=1}^{N} p(x_n \mid pa(x_n))$$

- Beispiel

$$x_1^{(k)} \sim p(x_1) \rightarrow x_1 = 1$$
$$x_2^{(k)} \sim p(x_2) \rightarrow x_2 = 0$$
$$x_3^{(k)} \sim p(x_3 \mid x_1 = 1, x_2 = 0) \rightarrow x_3 = 1$$
$$x_4^{(k)} \sim p(x_4 \mid x_2 = 0) \rightarrow x_4 = 0$$
$$x_5^{(k)} \sim p(x_5 \mid x_3 = 1) \rightarrow x_5 = 1$$



Inferenz: Ancestral Sampling

- Wir ziehen ein Sample $(x_1,...,x_N)^{(k)}$, indem wir nacheinander die einzelnen x_i ziehen

$$x_1^{(k)} \sim p(x_1)$$
$$x_2^{(k)} \sim p(x_2 \mid pa(x_2))$$
$$...$$
$$x_N^{(k)} \sim p(x_N \mid p$$

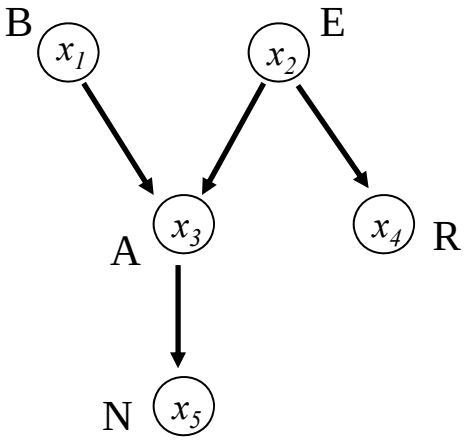
Schon gezogene Werte

P(B=1)
0.1

$$\mathbf{x}^{(k)} \sim p(\mathbf{x}) = p(x_1,..,x_N) = \prod_{n=1}^{N} p(x_n \mid pa(x_n))$$

- Beispiel

$$x_1^{(k)} \sim p(x_1) \rightarrow x_1 = 1$$
$$x_2^{(k)} \sim p(x_2) \rightarrow x_2 = 0$$
$$x_3^{(k)} \sim p(x_3 \mid x_1 = 1, x_2 = 0) \rightarrow x_3 = 1$$
$$x_4^{(k)} \sim p(x_4 \mid x_2 = 0) \rightarrow x_4 = 0$$
$$x_5^{(k)} \sim p(x_5 \mid x_3 = 1) \rightarrow x_5 = 1$$



Inferenz: Ancestral Sampling

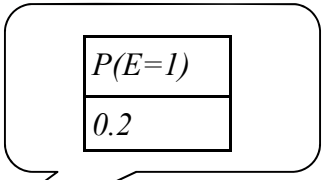
- Wir ziehen ein Sample $(x_1,...,x_N)^{(k)}$, indem wir nacheinander die einzelnen x_i ziehen

$$x_1^{(k)} \sim p(x_1)$$
$$x_2^{(k)} \sim p(x_2 \mid pa(x_2))$$
$$...$$
$$x_N^{(k)} \sim p(x_N \mid pa(x_N))$$

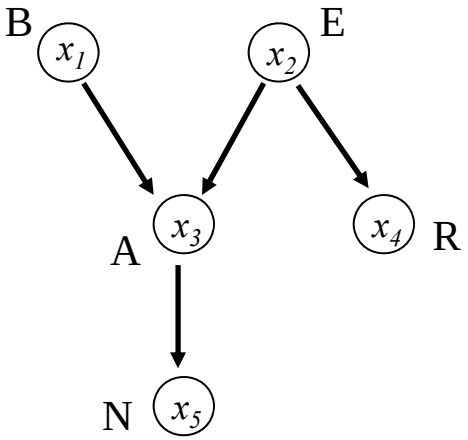
Schon gezogene Werte

$$\mathbf{x}^{(k)} \sim p(\mathbf{x}) = p(x_1,..,x_N) = \prod_{n=1}^{N} p(x_n \mid pa(x_n))$$

- Beispiel



$$x_1^{(k)} \sim p(x_1) \rightarrow x_1 = 1$$
$$x_2^{(k)} \sim p(x_2) \rightarrow x_2 = 0$$
$$x_3^{(k)} \sim p(x_3 \mid x_1 = 1, x_2 = 0) \rightarrow x_3 = 1$$
$$x_4^{(k)} \sim p(x_4 \mid x_2 = 0) \rightarrow x_4 = 0$$
$$x_5^{(k)} \sim p(x_5 \mid x_3 = 1) \rightarrow x_5 = 1$$



Inferenz: Ancestral Sampling

- Wir ziehen ein Sample $(x_1,...,x_N)^{(k)}$, indem wir nacheinander die einzelnen x_i ziehen

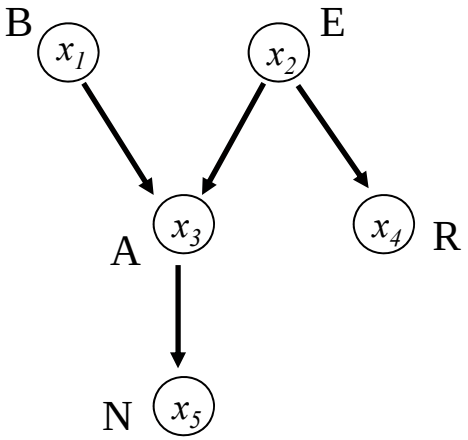
$$x_1^{(k)} \sim p(x_1)$$
$$x_2^{(k)} \sim p(x_2 \mid pa(x_2))$$
$$\ldots$$
$$x_N^{(k)} \sim p(x_N \mid pa$$

B	E	P(A=1|B,E)
0	0	0.01
0	1	0.5
1	0	0.9
1	1	0.95

$$\mathbf{x}^{(k)} \sim p(\mathbf{x}) = p(x_1,..,x_N) = \prod_{n=1}^{N} p(x_n \mid pa(x_n))$$

- Beispiel

$$x_1^{(k)} \sim p(x_1)$$
$$x_2^{(k)} \sim p(x_2) \rightarrow x_2 = 0$$
$$x_3^{(k)} \sim p(x_3 \mid x_1 = 1, x_2 = 0) \rightarrow x_3 = 1$$
$$x_4^{(k)} \sim p(x_4 \mid x_2 = 0) \rightarrow x_4 = 0$$
$$x_5^{(k)} \sim p(x_5 \mid x_3 = 1) \rightarrow x_5 = 1$$



Inferenz: Ancestral Sampling

- Wir ziehen ein Sample $(x_1,...,x_N)^{(k)}$, indem wir nacheinander die einzelnen x_i ziehen

$$x_1^{(k)} \sim p(x_1)$$
$$x_2^{(k)} \sim p(x_2 \mid pa(x_2))$$
$$...$$
$$x_N^{(k)} \sim p(x_N \mid pa$$

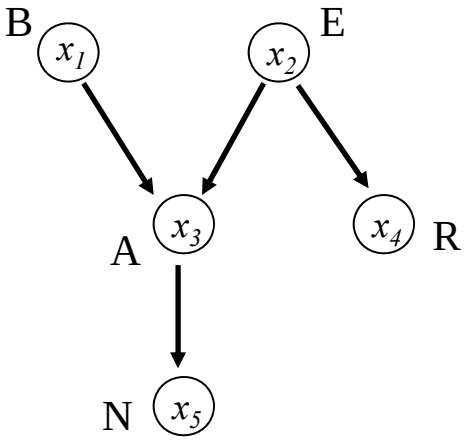
Schon gezogene Werte

E	P(R=1|E)
0	0.01
1	0.5

$$\mathbf{x}^{(k)} \sim p(\mathbf{x}) = p(x_1,.,x_N)$$
$$= \prod_{n=1}^{N} p(x_n \mid pa(x_n))$$

- Beispiel

$$x_1^{(k)} \sim p(x_1)$$
$$x_2^{(k)} \sim p(x_2)$$
$$x_3^{(k)} \sim p(x_3 \mid x_1 = 1, x_2 = 0) \quad \rightarrow x_3 = 1$$
$$x_4^{(k)} \sim p(x_4 \mid x_2 = 0) \quad \rightarrow x_4 = 0$$
$$x_5^{(k)} \sim p(x_5 \mid x_3 = 1) \quad \rightarrow x_5 = 1$$



Inferenz: Ancestral Sampling

- Wir ziehen ein Sample $(x_1,...,x_N)^{(k)}$, indem wir nacheinander die einzelnen x_i ziehen

$$x_1^{(k)} \sim p(x_1)$$
$$x_2^{(k)} \sim p(x_2 \mid pa(x_2))$$
$$...$$
$$x_N^{(k)} \sim p(x_N \mid pa(x_N))$$

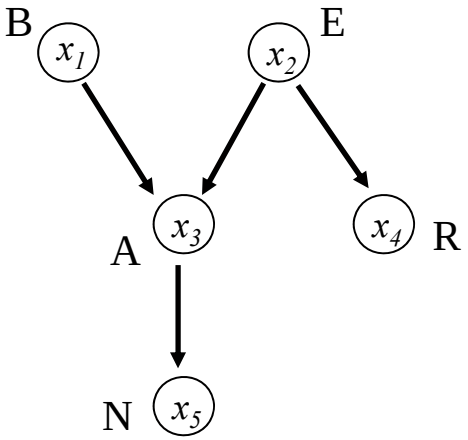
Schon gezogene Werte

$$\mathbf{x}^{(k)} \sim p(\mathbf{x}) = p(x_1,..,x_N) = \prod_{n=1}^{N} p(x_n \mid pa(x_n))$$

- Beispiel

$$x_1^{(k)} \sim p(x_1)$$
$$x_2^{(k)} \sim p(x_2)$$
$$x_3^{(k)} \sim p(x_3 \mid x_1 = \ldots$$
$$x_4^{(k)} \sim p(x_4 \mid x_3 = \ldots) \rightarrow x_4 = 0$$
$$x_5^{(k)} \sim p(x_5 \mid x_3 = 1) \rightarrow x_5 = 1$$

A	P(N=1|A)
0	0.1
1	0.7



Inferenz: Ancestral Sampling

- Wir ziehen ein Sample $(x_1,...,x_N)^{(k)}$, indem wir nacheinander die einzelnen x_i ziehen

$$x_1^{(k)} \sim p(x_1)$$
$$x_2^{(k)} \sim p(x_2 \mid pa(x_2))$$
$$...$$
$$x_N^{(k)} \sim p(x_N \mid pa(x_N))$$

Schon gezogene Werte

$$\mathbf{x}^{(k)} \sim p(\mathbf{x}) = p(x_1,.,x_N) = \prod_{n=1}^{N} p(x_n \mid pa(x_n))$$

- Beispiel

$$x_1^{(k)} \sim p(x_1) \rightarrow x_1 = 1$$
$$x_2^{(k)} \sim p(x_2) \rightarrow x_2 = 0$$
$$x_3^{(k)} \sim p(x_3 \mid x_1 = 1, x_2 = 0) \rightarrow x_3 = 1$$
$$x_4^{(k)} \sim p(x_4 \mid x_2 = 0) \rightarrow x_4 = 0$$
$$x_5^{(k)} \sim p(x_5 \mid x_3 = 1) \rightarrow x_5 = 1$$

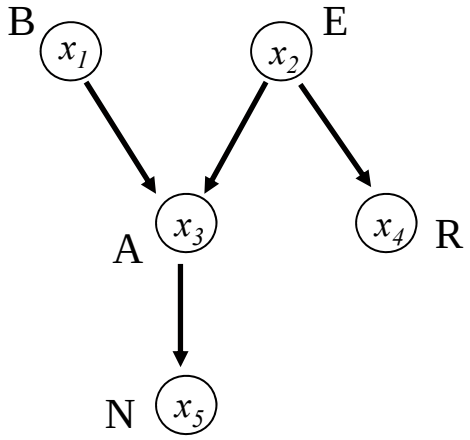
$$\Rightarrow \quad \mathbf{x}^{(k)} = (1,0,1,0,1)$$

Inferenz: Ancestral Sampling

- Beispiel für Schätzung der Randverteilungen aus Samples:

$$\mathbf{x}^{(1)} = (1,0,1,0,1)$$
$$\mathbf{x}^{(2)} = (0,0,0,0,0)$$
$$\mathbf{x}^{(3)} = (0,1,0,1,0)$$
$$\mathbf{x}^{(4)} = (0,1,1,0,1)$$
$$\mathbf{x}^{(5)} = (0,0,0,0,0)$$

$$p(x_3 = 1) \approx 0.4$$
$$p(x_4 = 1) \approx 0.2$$
$$p(x_5 = 1) \approx 0.4$$



- Analyse Ancentral Sampling
  - + Zieht direkt aus der korrekten Verteilung
  - + Effizient
  - - Funktioniert nur ohne Evidenz

Inferenz: Logic Sampling

- Wie erhalten wir Samples unter Evidenz?

Beobachtete Variablen

$$\mathbf{x}_I^{(k)} \sim p(\mathbf{x}_I \mid \mathbf{x}_D) = p(x_{i_1}, ..., x_{i_m} \mid x_{j_1}, ..., x_{j_l})$$

- Logic Sampling: Ancestral Sampling + Zurückweisung von Samples, die nicht mit der Beobachtung konsistent sind
  - Ancestral Sampling: vollständige Samples

$$\mathbf{x}^{(k)} = (x_1^{(k)}, ..., x_N^{(k)}) \sim p(\mathbf{x})$$

  - Fallunterscheidung:

$(x_{j_1}^{(k)}, ..., x_{j_l}^{(k)}) = (x_{j_1}, ..., x_{j_l})$ [Sample konsistent mit Beobachtung]: akzeptiere $\mathbf{x}^{(k)}$

$(x_{j_1}^{(k)}, ..., x_{j_l}^{(k)}) \neq (x_{j_1}, ..., x_{j_l})$ [Sample ìnkonsistent mit Beobachtung]: weise $\mathbf{x}^{(k)}$ zurück

Inferenz: Logic Sampling

- Die im Logic Sampling akzeptierten Samples $\mathbf{x}^{(k)}$ repräsentieren die bedingte Verteilung gegeben Evidenz:

$$\mathbf{x}^{(k)} \sim p(\mathbf{x} \mid \mathbf{x}_D)$$

- Marginale Samples

$$\mathbf{x}_I^{(k)} \sim p(\mathbf{x}_I \mid \mathbf{x}_D)$$

  wieder durch Projektion auf Anfragevariablen

- Problem: Oft werden fast alle Samples verworfen
  - Wahrscheinlichkeit, Sample zu generieren, das mit $\mathbf{x}_D$ konsistent ist, sinkt meist exponentiell schnell mit Größe von D
  - Entsprechend exponentielle Laufzeit, um ausreichende Menge von Samples zu erhalten
  - In der Praxis selten anwendbar

Inferenz: MCMC

- Alternative Strategie zum Erzeugen von Samples: Markov Chain Monte Carlo („MCMC“)
- Idee:
  - Schwierig, direkt Samples aus $p(\mathbf{z})$ zu ziehen
  - Alternativstrategie: Konstruiere Folge von Samples

$$\mathbf{z}^{(0)} \rightarrow \mathbf{z}^{(1)} \rightarrow \mathbf{z}^{(2)} \rightarrow \mathbf{z}^{(3)} \rightarrow \mathbf{z}^{(4)} \rightarrow \mathbf{z}^{(5)} \rightarrow \ldots$$

$\mathbf{z}^{(0)}$ zufällig initialisiert $\qquad$ $\mathbf{z}^{(t+1)} = update(\mathbf{z}^{t})$

durch mehrfache probabilistische Update-Schritte

  - Wenn Updates geeignet gewählt, gilt asymptotisch

$$\mathbf{z}^{(T)} \sim p(\mathbf{z}) \qquad \text{ungefähr, für sehr grosse } T$$

ZV: T-te Variablenbelegung

Markov-Ketten

- Folge der Zustände

$$\mathbf{z}^{(0)} \rightarrow \mathbf{z}^{(1)} \rightarrow \mathbf{z}^{(2)} \rightarrow \mathbf{z}^{(3)} \rightarrow \mathbf{z}^{(4)} \rightarrow \mathbf{z}^{(5)} \rightarrow \ldots$$

bildet Markov-Kette:

$\mathbf{z}^{(t)}$ heisst "Zustand" der Kette zum Zeitpunkt t

$$p(\mathbf{z}^{(0)},\ldots,\mathbf{z}^{(T)}) = p(\mathbf{z}^{(0)})\prod_{t=1}^{N} p(\mathbf{z}^{(t)} \mid \mathbf{z}^{(t-1)})$$

- Dynamik beschrieben durch Transitionswahrscheinlichkeiten

$$\mathrm{T}(\mathbf{z}^{(t)},\mathbf{z}^{(t+1)}) = p(\mathbf{z}^{(t+1)} \mid \mathbf{z}^{(t)})$$

Wahrscheinlichkeit Übergang $\mathbf{z}^{(t)} \rightarrow \mathbf{z}^{(t+1)}$

Neuer Zustand

Aktueller Zustand

Markov-Ketten

- Verteilung über Folgezustand berechnen aus Verteilung über aktuellem Zustand

Folgezustand ↘ Aktueller Zustand ↙

$$p(\mathbf{z}^{(t+1)}) = \sum_{\mathbf{z}_t} p(\mathbf{z}^{(t+1)} \mid \mathbf{z}^{(t)}) p(\mathbf{z}^{(t)})$$

$$= \sum_{\mathbf{z}_t} T(\mathbf{z}^{(t)}, \mathbf{z}^{(t+1)}) p(\mathbf{z}^{(t)})$$

- Stationäre Verteilung
  - Eine Verteilung $p_*(\mathbf{z})$ über die Zustände der Kette heisst „stationär“, falls sie bei einem Schritt der Markov-Kette nicht verändert wird:

$$p_*(\mathbf{z}^{(t+1)}) = \sum_{\mathbf{z}^{(t)}} T(\mathbf{z}^{(t)}, \mathbf{z}^{(t+1)}) p_*(\mathbf{z}^{(t)})$$

Markov-Ketten: Stationäre Verteilung

- Falls die Kette eine stationäre Verteilung erreicht, dh. $p(\mathbf{z}^{(T)}) = p_*(\mathbf{z}^{(T)})$ für ein geeignetes T, so bleibt diese Verteilung erhalten, dh. $p(\mathbf{z}^{(T+N)}) = p_*(\mathbf{z}^{(T+N)})$ für alle N.
  - Kette ist „konvergiert“ zur Verteilung p_*

- Unter bestimmten Bedingungen („Ergodische Ketten“) konvergiert eine Markov-Kette für $t \to \infty$ gegen eine eindeutige stationäre Verteilung, diese heisst dann „Gleichgewichtsverteilung“

Inferenz: MCMC

- Gegeben Bayessches Netz über ZV $\mathbf{x} = \{x_1, \ldots, x_N\}$, definiert Verteilung $p(\mathbf{x})$

- „Markov Chain Monte Carlo“ Methoden
  - Konstruiere aus dem Bayesschen Netz eine Folge von Samples durch iterative probabilistische Updates

$$\mathbf{x}^{(0)} \rightarrow \mathbf{x}^{(1)} \rightarrow \mathbf{x}^{(2)} \rightarrow \mathbf{x}^{(3)} \rightarrow \mathbf{x}^{(4)} \rightarrow \mathbf{x}^{(5)} \rightarrow \ldots$$

$\mathbf{x}^{(t)}$ jeweils Belegung aller Knoten im Netz

$\mathbf{x}^{(0)}$ zufällig initialisiert

$$\mathbf{x}^{(t+1)} = update(\mathbf{x}^{t})$$

  - Ziel: Updates so wählen, dass sich ergodische Markov-Kette mit Gleichgewichtsverteilung $p(\mathbf{x})$ ergibt
  - Einfachste Methode: lokales Ziehen einer Variable, gegeben Zustand der anderen Variablen („Gibbs-Sampling“)

Inferenz: Gibbs Sampling

- Gibbs Sampling: Eine Version von MCMC
- Übergangswahrscheinlichkeiten bestimmt durch wiederholtes lokales Ziehen einer ZV, gegeben den Zustand aller anderen ZV
  - Gegeben alter Zustand $\mathbf{x} = (x_1, ..., x_N)$
  - Ziehen des neuen Zustands $\mathbf{x}' = (x_1', ..., x_N')$:

$$x_1' \sim p(x_1 \mid \overbrace{x_2, ..., x_N}^{\text{Beobachtete (alte) Werte}})$$

$$x_2' \sim p(x_2 \mid x_1', x_3, ..., x_N)$$

$$x_3' \sim p(x_3 \mid x_1', x_2', x_4, ..., x_N)$$

$$...$$

$$x_N' \sim p(x_N \mid x_1', x_2', ..., x_{N-1}')$$

Inferenz: Gibbs Sampling

- Einzelner Gibbs-Schritt einfach, bedingte Verteilung über eine Variable gegeben Evidenz auf **allen** anderen Variablen direkt auszurechnen:

$$\text{Berechnung von } p(x_1 \mid \overbrace{x_2, \ldots, x_N}^{\text{Beobachtete (alte) Werte}}):$$

$$\text{Für } x_1 = 1 \text{ berechne } p_1 = \mathrm{p}(x_1, x_2, \ldots, x_N)$$

$$\text{Für } x_1 = 0 \text{ berechne } p = \mathrm{p}(x_1, x_2, \ldots, x_N) + p_1$$

$$p(x_1 = 1 \mid x_2, \ldots, x_N) = p_1 / p$$

- Satz: Falls $p(x_n \mid x_1, x_2, \ldots, x_{n-1}, x_{n+1}, \ldots, x_{N-1}) \neq 0$ für alle n und alle möglichen Zustände x_i, so ist die resultierende Markov-Kette ergodisch mit Gleichgewichtsverteilung $p(\mathbf{x})$.

Gibbs-Sampling mit Evidenz

- Bisher haben wir Inferenz ohne Evidenz betrachtet
- Wie erhalten wir Samples aus der bedingten Verteilung?

Ziel: $\mathbf{x}^{(T)} \sim p(\mathbf{x} \mid \mathbf{x}_D)$ ungefähr, für sehr grosse T

- Leichte Modifikation der Gibbs-Sampling Methode:
  - Gibbs-Sampling zieht immer eine Variable x_i neu, gegeben Zustand der anderen Variablen
  - Mit Evidenz: Nur die unbeobachteten Variablen werden jeweils neu gezogen, die beobachteten Variablen werden fest auf den beobachteten Wert gesetzt

Inferenz: Gibbs Sampling

- Zusammenfassung Gibbs Sampling Algorithmus:
  - $\mathbf{x}^{(0)} = \text{zufällige Initialisierung aller ZV, konsistent mit Evidenz } \mathbf{x}_D$
  - Für $t = 1,...,T$: $\mathbf{x}^{(t)} = \text{Gibbs-update}(\mathbf{x}^{(t-1)})$
  - Die Samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, ...$ sind asymptotisch verteilt nach $p(\mathbf{x} \mid \mathbf{x}_D)$
- Gibbs Sampling in vielen praktischen Anwendungen brauchbar
  - Einzelne Update-Schritte effizient
  - Garantierte Konvergenz (für $t \to \infty$)
  - Erlaubt, Samples aus $p(\mathbf{x} \mid \mathbf{x}_D)$ zu ziehen, ohne dass Laufzeit explodiert wenn Evidenzmenge groß (im Gegensatz zu Logic Sampling)

Inferenz: Gibbs Sampling

- Gibbs-Sampling: Konvergenz
  - Konvergenz der Markov-Kette $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \ldots$ ist nur garantiert für $t \to \infty$.
  - In der Praxis: „Burn-In“ Iterationen, bevor Samples verwendet werden (verwerfe Samples $\mathbf{x}^{(t)}$ für $t \leq T_{Burn-in}$)
  - Es gibt auch Konvergenztests, um Anzahl der Burn-In Iterationen zu bestimmen

- Gibbs-Sampling: Abhängigkeit
  - Einzelne aufeinander folgende Samples $\mathbf{x}^{(t)}$, $\mathbf{x}^{(t+1)}$ abhängig
  - Lösung: verwende Samples $\mathbf{x}^{(t)}, \mathbf{x}^{(t+L)}, \mathbf{x}^{(t+2L)}, \mathbf{x}^{(t+3L)}, \ldots$
  - Samples dann weitestgehend unabhängig

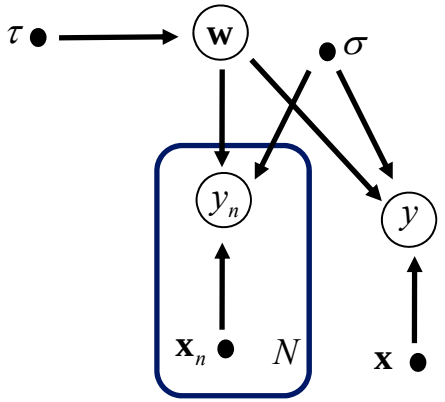
Inferenz: Zusammenfassung

- Exakte Inferenz
  - Message-Passing Algorithmen
  - Exakte Inferenz auf Polytrees (mit Junction-Tree Erweiterung auf allgemeinen Graphen)
  - Laufzeit abhängig von Graphstruktur, exponentiell im worst-case

- Approximative Inferenz
  - Sampling-Methoden: Approximation durch Menge von „Samples“, exakte Ergebnisse für $t \rightarrow \infty$.
    - Ancestral Sampling: einfach, schnell, keine Evidenz
    - Logic Sampling: mit Evidenz, aber selten praktikabel
    - MCMC/Gibbs-Sampling: Effizientes approximatives Ziehen von Samples unter Evidenz

Lernen Graphischer Modelle aus Daten?

- Graphische Modelle im maschinellen Lernen: Problemstellung ist meist Inferenzproblem
  - MAP Parameterschätzung
  - Bayessche Vorhersage



- Auch möglich, die Verteilungen eines graphischen Modells aus Daten zu schätzen

$$p(x_1,...,x_N) = \prod_{i=1}^{N} p(x_i \mid pa(x_i))$$

$p(x_i \mid pa(x_i))$ parametrisierte Verteilung (z.B. diskrete Tabelle)

Parameter können aus Daten geschätzt werden

Maximum-Likelihood Parameterlernen

- Parametrisierung mit Parametervektor θ

$$p(x_1,...,x_N \mid \theta) = \prod_{i=1}^{N} p(x_i \mid pa(x_i),\theta)$$

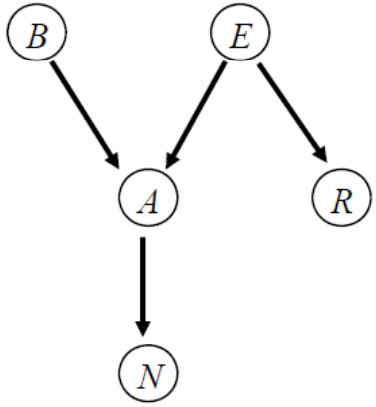
- Beobachtete Daten: Belegungen der Zufallsvariablen $x_1,\ldots,x_N$

Daten:

i.i.d. Beispiele

$L = \{\mathbf{x}_1,....,\mathbf{x}_m\}$

	B	E	A	N	R
$\mathbf{x}_1$	0	0	0	0	0
$\mathbf{x}_2$	0	1	1	1	1
$\mathbf{x}_3$	0	1	0	0	0
$\mathbf{x}_4$	1	0	1	1	0
$\mathbf{x}_5$	0	0	0	1	0
$\mathbf{x}_6$	1	1	1	0	1



Maximum-Likelihood Parameterlernen

- Maximum-Likelihood-Ansatz

$$
\begin{aligned}
\theta_* &= \arg\max_\theta \, p(L \mid \theta) \\
&= \arg\max_\theta \, p(\mathbf{x}_1, ..., \mathbf{x}_m \mid \theta) \\
&= \arg\max_\theta \prod_{j=1}^{m} p(\mathbf{x}_j \mid \theta) && \text{i.i.d.} \\
&= \arg\max_\theta \prod_{j=1}^{m} \prod_{i=1}^{N} p(x_{ji} \mid pa(x_{ji}), \theta) && \text{Faktorisierung GM}
\end{aligned}
$$

Maximum-Likelihood Parameterlernen

- Lösung (diskrete Variablen):

$$p(\underbrace{x_{ji}}_{\substack{\text{Zustand} \\ x_{ji}}} \mid \underbrace{pa(x_{ji})}_{\substack{\text{Zustand} \\ pa(x_{ji})}}, \theta_*) = \frac{C(x_{ji}, pa(x_{ji}))}{C(pa(x_{ji}))}$$

$C(x_{ji}, pa(x_{ji}))$ = Wie oft wurde der gemeinsame Zustand $x_{ji}, pa(x_{ji})$ beobachtet?

$C(pa(x_{ji}))$ = Wie oft wurde der gemeinsame Zustand $pa(x_{ji})$ beobachtet?

Daten:

i.i.d. Beispiele

$L = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$

	B	E	A	N	R
$\mathbf{x}_1$	0	0	0	0	0
$\mathbf{x}_2$	0	1	1	1	1
$\mathbf{x}_3$	0	1	0	0	0
$\mathbf{x}_4$	1	0	1	1	0
$\mathbf{x}_5$	0	0	0	1	0
$\mathbf{x}_6$	1	1	1	0	1

$$p(N = 1 \mid A = 0, \theta_*) = \frac{1}{3}$$

$$p(N = 1 \mid A = 1, \theta_*) = \frac{2}{3}$$

Maximum-Likelihood Parameterlernen

- Einfache Lösung nur möglich bei *vollständig beobachteten* Daten
- Modelle enthalten oft Variablen, die nicht beobachtbar sind
- Maximum-Likelihood-Schätzung dann mit dem EM Algorithmus („Expectation-Maximization“)
- Später mehr!