Bayessche Graphen

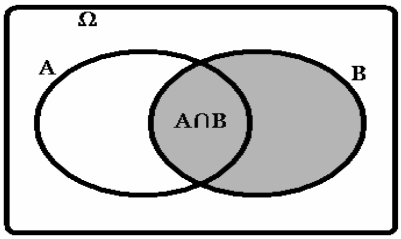
Jan Simon, RWTH Aachen, im Sommer 2008

Zum Inhalt:
Nach einer Wiederholung der relevanten Grundlagen werden Bayessche Graphen als graphentheoretische Darstellung der Bedingtheitsstruktur einer gemeinsamen Wahrscheinlichkeitsverteilung von Zufallsvariablen eingeführt. Dazu wird ein hauptsächlich exemplarischer Zugang gewählt. Einige Ausblicke zeigen auch die Art der formalen Herangehensweise in diesem Themenbereich auf. Grundlagen wie die Konstruktion Bayesscher Graphen, die d-Separation und Inferenz werden eingeführt. Darüber hinaus folgt ein kurzer Abriss des Themas Kausalität im Kontext Bayesscher Graphen.

I Bedingtheit und Unabhängigkeit

Mit dem Konzept der **bedingten Wahrscheinlichkeiten** bietet die Stochastik ein wichtiges Mittel, um zum Ausdruck zu bringen, wie sich Information auf unsere Einschätzung einer dem Zufall unterworfenen Situation auswirkt. Ist Ω eine Grundgesamtheit, so bezeichnet man die Wahrscheinlichkeit eines Ereignisses $A \subseteq \Omega$ unter der Voraussetzung, dass ein zweites Ereignis $B \subseteq \Omega$ mit $P(B)>0$ bereits eingetreten ist, mit der „Wahrscheinlichkeit von A gegeben B“ und schreibt dafür $P(A|B)$. Man berechnet diese Wahrscheinlichkeit gemäß der Formel

$$P(A|B)=P(A\cap B)/P(B). \qquad (1)$$



Die Vorstellung dahinter ist die, dass B durch die Information, dass es eingetreten ist, selber als nun sicheres Ereignis zur neuen Grundgesamtheit wird. Dadurch tritt A also genau dann ein, wenn $A\cap B$ eintritt. Um $P(B|B)=1$ zu erhalten, dividiert man noch durch $P(B)$.

Die Überlegung ermöglicht es uns, formal zu definieren, was es bedeutet, dass zwei Ereignisse A und B keinen Einfluss aufeinander haben, dass sie **stochastisch unabhängig** sind. Offenbar sollte gelten

$$P(A|B)=P(A). \qquad (2)$$

Man definiert hingegen die stochastische Unabhängigkeit häufig über folgende Multiplikationsregel:

Definition: Zwei Ereignisse A, B heißen *stochastisch unabhängig*, falls gilt: $P(A\cap B)=P(A)*P(B)$. (3)

Aus dieser Definition folgt mit (1) die Bedingung (2). Die Definition (3) ist aber gegenüber (2) allgemeingültiger, da sie auch für unmögliche Ereignisse B mit $P(B)=0$ anwendbar ist. Darüber hinaus zeigt sich in ihr deutlicher, dass die Unabhängigkeitseigenschaft symmetrisch ist. Es sei noch darauf hingewiesen, dass stochastische Unabhängigkeit nicht nur von den Ereignissen selber, sondern auch von der verwendeten Verteilung P abhängt.

Häufig nimmt man auch umgekehrt stochastische Unabhängigkeit zweier Ereignisse an, um mit der Formel aus der Definition bequem die Wahrscheinlichkeit des Schnitt-Ereignisses zu berechnen.

Dieses Vorgehen lässt sich verallgemeinern auf Schnitte mit beliebigen Anzahlen von Ereignissen $A_1, A_2, \ldots, A_n$, deren Wahrscheinlichkeit größer als Null ist. Es gilt die

Kettenregel: $P(A_1\cap A_2\cap\ldots\cap A_{n-1}\cap A_n)=P(A_1)\,P(A_2|A_1)\,P(A_3|A_1\cap A_2)\ldots P(A_n|A_1\cap A_2\cap\ldots\cap A_{n-1})$. (4)

Die Regel folgt formal durch Verwenden von (1) und (3). Als Interpretation kann man sich denken, dass die Ereignisse $A_1, \ldots, A_n$ in einer *zeitlichen Reihenfolge* angeordnet sind. Zuerst tritt A_1 ein, dann A_2, wobei A_1 aber bereits eingetreten ist, dann A_3, wobei A_1 und A_2 bereits eingetreten sind u. s. w.. Was schließlich resultiert, ist das Ereignis, dass alle $A_1, \ldots, A_n$, also ihr Schnitt, eingetreten ist. Wir betrachten dazu folgendes Beispiel:

Aus einem Vorrat von sieben weißen und fünf schwarzen Kugeln erhalten drei Personen je zwei Kugeln. Gesucht ist die Wahrscheinlichkeit, dass *jede Person genau eine schwarze und genau eine weiße Kugel* erhalten hat. Wir wählen die Ereignisse

A_1=„Die erste Person erhält genau eine schwarze und genau eine weiße Kugel“

A_2=„Die zweite Person erhält genau eine schwarze und genau eine weiße Kugel“

A_3=„Die dritte Person erhält genau eine schwarze und genau eine weiße Kugel“.

Dann interessiert also der Schnitt $A_1\cap A_2\cap A_3$. Es gilt $P(A_1)=1/7*1/5$ und ferner $P(A_2|A_1)=1/6*1/4$, da ja bereits eine schwarze und eine weiße Kugel an die erste Person ausgeteilt wurde. Schließlich ist $P(A_3|A_1\cap A_2)=1/5*1/3$, da inzwischen zwei schwarze und zwei weiße Kugeln im Vorrat fehlen. Mit der Kettenregel (4) finden wir

$P(A_1\cap A_2\cap A_3)=P(A_1)\,P(A_2|A_1)\,P(A_3|A_1\cap A_2)=1/7*1/5*1/6*1/4*1/5*1/3=1/12600=0{,}000079\ldots$.

Bedingte Wahrscheinlichkeiten lassen sich auch hinsichtlich *Ursache und Wirkung* interpretieren: Die Ursache B ändert die Wahrscheinlichkeit, dass die Wirkung A eintritt. Dann beschreibt P(A|B) im Gegensatz zu P(A) die Wirkung von B auf A, die fördernd oder hemmend ausfallen kann. Wiederum lässt sich formal aus (1) eine wichtige Regel schließen, die

Inversionsformel nach Bayes: $P(A|B)=P(B|A)\,P(A)/P(B)$. (5)

Ihr besonderer Reiz liegt darin, dass sie die bedingten Wahrscheinlichkeiten P(A|B) und P(B|A) verknüpft. So lässt sich nicht nur von der Ursache B auf die Wirkung A schließen (kausales Schließen: P(A|B)), sondern auch *von der Wirkung auf die Ursache* (diagnostisches Schließen: P(B|A)). Häufig fällt es uns Menschen leichter, die Wahrscheinlichkeit P(A|B) in kausaler Richtung anzugeben, obwohl die umgekehrte Richtung interessiert. Die Inversionsformel ermöglicht es dann (falls auch P(A) und P(B) bekannt sind) in diagnostischer Richtung auf P(B|A) zu schließen.

B→A

Beispielsweise sei bekannt, dass eine Alarmanlage bei einem Einbruch recht zuverlässig alarmiert: P(„Alarm“|„Einbruch“)=0,95.

Außerdem sei P(„Einbruch“)=0,0001 bekannt und P(„Alarm“)=0,0004. Die Tatsache, dass P(„Alarm“)>P(„Einbruch“) ist, deutet bereits darauf hin, dass die Alarmanlage bisweilen auch Fehlalarm gibt. Von Interesse ist nun die diagnostische Wahrscheinlichkeit, dass bei Alarm tatsächlich ein Einbruch vorliegt. Die Inversionsformel liefert dann P(„Einbruch“|„Alarm“)=0,9*0,0001/0,0004=0,225.

Die in (3) eingeführte stochastische Unabhängigkeit von Ereignissen lässt sich auf Zufallsvariablen ausweiten:

Definition: Zwei Zufallsvariablen X, Y heißen *stochastisch unabhängig*, falls für alle Werte x, die X annehmen kann, und für alle Werte y, die Y annehmen kann, gilt: $P(X=x\cap Y=y)=P(X=x)*(Y=y)$. (6)

Wir werden die Verwendung von Großbuchstaben für Zufallsvariablen und der entsprechenden Kleinbuchstaben für die möglichen Werte dieser Variablen im gesamten Text beibehalten.

Für die Bayesschen Graphen benötigen wir noch die **bedingte Unabhängigkeit** von Zufallsvariablen:

Definition: Zwei Zufallsvariablen X, Y heißen *bedingt unabhängig gegeben eine Zufallsvariable Z*, falls für alle Werte x, y, z, die X, Y und Z annehmen können (deren Wahrscheinlichkeit positiv ist), gilt: $P(X=x|Y=y\cap Z=z)=P(X=x|Z=z)$. (7)

Unter der Voraussetzung, dass der Wert z, den Z annimmt, bekannt ist, ändert also die Kenntnis des Wertes y, den Y annimmt, nichts mehr an der Wahrscheinlichkeit für X=x. Ist Z festgelegt, so werden dadurch X und Y stochastisch unabhängig. Wir werden bald Beispiele hierzu kennen lernen.

Zuvor überzeugen wir uns aber noch davon, dass auch die bedingte Unabhängigkeit symmetrisch in X und Y ist, was die Definition nicht unmittelbar vermuten lässt. Mit (1) folgern wir

$P(X=x|Y=y\cap Z=z)=P(X=x|Z=z)$
⇨ $P(X=x\cap Y=y\cap Z=z)/P(Y=y\cap Z=z)=P(X=x\cap Z=z)/P(Z=z)$
⇨ $P(X=x\cap Y=y\cap Z=z)/P(X=x\cap Z=z)=P(Y=y\cap Z=z)/P(Z=z)$
⇨ $P(Y=y|X=x\cap Z=z)=P(Y=y|\cap Z=z)$, wovon wir uns überzeugen wollten.

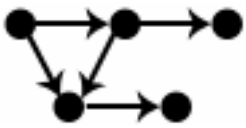
Es ist also gerechtfertigt, in der Definition X und x mit Y und y zu vertauschen.

II Bayessche Graphen

Wir führen **gerichtete Graphen** ein:

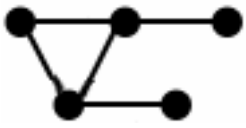
Definition: Ist V eine Menge von Knoten $v_1, v_2, \ldots, v_n$ und E eine Menge von Paaren (v_i, v_j) dieser Knoten, so nennt man das Paar G=(V, E) einen *gerichteten Graphen*. (8)

Ein gerichteter Graph G besteht also aus Knoten (die Menge V) und Pfeilen (die Menge E), die von Knoten zu Knoten zeigen können. Zeigt ein Pfeil von v_i zu v_j, so drückt sich das darin aus, dass das Paar (v_i, v_j) in der Menge E vorkommt.



Ferner versteht man unter einem *gerichteten*, *azyklischen Graphen* einen solchen, in dem es keinen geschlossenen Weg gibt, *der ausschließlich den Pfeilrichtungen folgt*.

Der *ungerichtete Graph* eines gerichteten Graphen (sein *Skelett*) geht aus dem gerichteten durch Ignorieren der Pfeilrichtung hervor. (Formal ersetzt man geordnete Paare durch zweielementige Mengen.)



Darüber hinaus definiert man *Eltern-* und *Kindknoten* eines Knoten als seine unmittelbaren Vorgänger und Nachfolger gemäß den Pfeilrichtungen.

Angenommen, wir möchten eine komplizierte, vom Zufall beeinflusste Situation mit den Zufallsvariablen $X_1, X_2, ..., X_n$ modellieren. Dann interessiert offenbar die Wahrscheinlichkeit, dass ein gewisser Zustand eintritt, etwa

$$P(X_1=x_1 \cap X_2=x_2 \cap ... \cap X_n=x_n) =: P(x_1, x_2, ..., x_n). \quad (9)$$

Die zweite Schreibweise wird aufgrund ihrer Kompaktheit gerne verwendet, wenn klar ist, um welche Zufallsvariablen es sich in welcher Reihenfolge handelt. Der Kleinbuchstabe x_i gibt dabei wieder einen der Werte an, welche die Zufallsvariable X_i annehmen kann. Nach der Kettenregel (4) gilt

$$P(x_1, x_2, ..., x_n) = P(x_1)\, P(x_2|x_1)\, P(x_3|x_1, x_2) \ldots P(x_n|x_1, ..., x_{n-1}). \quad (10)$$

Wir erinnern uns an die bedingte Unabhängigkeit (7) und vermuten, dass im Allgemeinen eine geringere Anzahl von Zufallsvariablen pro Faktor genügt, auf die bedingt wird. Zu X_i gibt es offenbar eine *minimale Teilmenge* von $\{X_1, X_2, ..., X_{i-1}\}$, die wir mit Pa_i bezeichnen, sodass $P(x_i|x_1, ..., x_{i-1})$ in (10) durch $P(x_i|pa_i)$ ersetzt werden kann, *ohne dass Information verloren ginge*. Man nennt Pa_i auch die *Markov-Eltern* von X_i. Gegebenenfalls ist $Pa_i=\{X_1, X_2, ..., x_{i-1}\}$. Damit erhalten wir (mit dem Kunstgriff $Pa_1:=\{\}$)

$$P(x_1, x_2, ..., x_n) = P(x_1|pa_1)\, P(x_2|pa_2)\, P(x_3|pa_3) \ldots P(x_n|pa_n). \quad (11)$$

Das ist eine Verbesserung gegenüber (10), da es leichter ist, bedingte Wahrscheinlichkeiten auszurechnen, wenn auf weniger Variablen bedingt wird. Wir kommen im Abschnitt zur Inferenz darauf zurück.

Es bietet sich an, die Bedingtheitsstruktur der Zufallsvariablen in einem Graphen darzustellen. Damit kommen wir zur zentralen Definition eines **Bayesschen Graphen**, der stochastische und graphentheoretische Konzepte vereint.

Definition: Es sei $G=(V, E)$ ein gerichteter, azyklischer Graph, dessen Knoten Zufallsvariablen sind und in dem die Markov-Eltern Pa_i einer jeden Knotenvariable X_i gerade deren Eltern im Graph sind, sodass diese also eine minimale Knotenmenge bilden, von denen X_i abhängig ist. In diesem Fall heißt G ein *Bayesscher Graph*. (12)

In der Praxis wird jeder Knoten eine Tabelle beinhalten, in der die elternbedingten Wahrscheinlichkeiten für die einzelnen Werte vermerkt sind, die der Knoten annehmen kann. Die Anzahl erforderlicher Tabelleneinträge verhält sich exponentiell in der Anzahl Eltern. Da Bayessche Graphen diese Anzahl minimieren, sind sie (unter anderem) hinsichtlich des Speicherbedarfs sehr effektiv.
Kann beispielsweise jede Zufallsvariable nur zwei Werte annehmen (etwa „ja“ und „nein“), und wird eine Zufallsvariable X auf fünf andere bedingt, so sind 2^5=32 Tabellenzeilen erforderlich, um alle möglichen Zustände der bedingenden Variablen zu erfassen. Entdeckt man, dass zwei dieser Variablen unabhängig von X sind gegeben der übrigen, so viertelt sich die Anzahl der Tabelleneinträge bereits von 32 auf 2^(5-2)=8.

Wir vollziehen jetzt anhand eines Beispiels nach, **wie Bayessche Graphen konstruiert werden** können. Es sei eine Menge $\{X_1, X_2, ..., X_7\}$ von Zufallsvariablen gegeben, die der Einfachheit halber nur die Werte „ja“ und „nein“ annehmen. Unser Beispiel kommt aus der Medizin und die Variablen sind gewisse Diagnosen und Symptome, die bei einem Patienten entweder vorliegen („ja“) oder nicht („nein“).

X_1=Schrumpfniere
X_2=Arteriosklerose (Verhärtung von Arterien)
X_3=Hypertonie (Bluthochdruck)
X_4=arterielle Embolie (fortgeschwemmtes Gerinnsel in den Arterien)
X_5=Schlaganfall
X_6=Taubheit einer Körperseite
X_7=einseitige Lähmung

Um den Bayesschen Graphen zu konstruieren, wählen wir die erste Variable X_1, die Schrumpfniere, als ersten Knoten. Anschließend fügen wir die übrigen Variablen X_i sukzessive ein, wobei wir jedes Mal nach den Markov-Eltern, also einer *minimalen* Menge $Pa_i \subseteq \{X_1, ..., X_{i-1}\}$ an bereits im Graphen befindlichen Variablen suchen, sodass

$$P(x_i|x_1, ..., x_{i-1}) = P(x_i|pa_i) \quad (13)$$

ist. Es sollen also nur die *direkten*, nicht die mittelbaren Zusammenhänge dargestellt werden. Von den ausgewählten Variablen Pa_i, werden dann Pfeile zur neuen Variablen X_i gezogen, sodass die Pa_i im Graphen die *Eltern* von X_i werden. Dadurch ist X_i eingefügt und so fährt man fort bis zur letzten Variablen.

X_2 ist die Arteriosklerose. Bisher befindet sich nur die Schrumpfniere im Graph. Da die Wahrscheinlichkeit für Arteriosklerose im Wesentlichen unabhängig von der Schrumpfniere ist, fügen wir X_2 in unseren Graphen ein, *ohne* einen Pfeil von X_1 her einzuzeichnen.

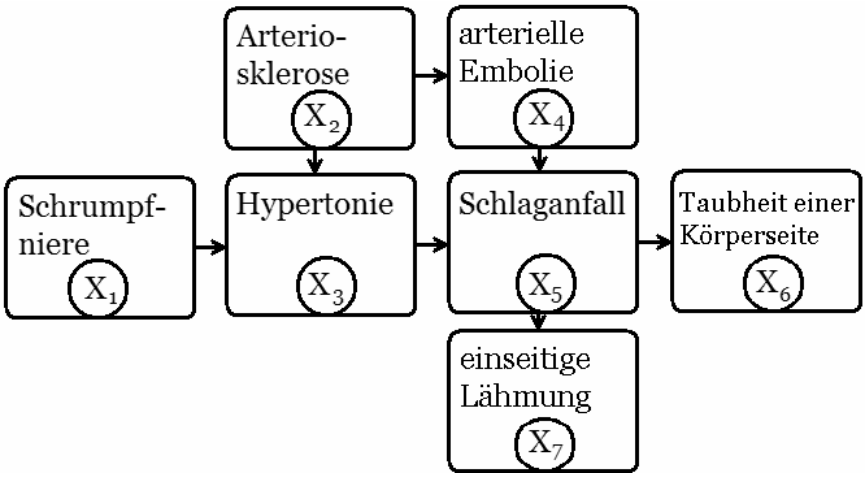
Jetzt ist die Hypertonie X_3 an der Reihe. Die Schrumpfniere steigert als klassische Ursache von Hypertonie deren Wahrscheinlichkeit deutlich. Ebenso sorgt die Arteriosklerose durch die verhärteten Gefäße für erhöhten Blutdruck. Daher zeichnen wir sowohl von X_1 als auch von X_2 aus Pfeile zu X_3.

Die arterielle Embolie, X_4, ist unabhängig von der Schrumpfniere. Sehr wohl entsteht sie aber bei Arteriosklerose gehäuft. Zwar bewirkt auch die Hypertonie manchmal ein Auftreten der arteriellen Embolie, aber meist *in Zusammenhang mit Arteriosklerose*. Ist diese bereits gegeben, so ändert das Wissen über eine Hypertonie nur wenig an der Wahrscheinlichkeit der arteriellen Embolie. Da unsere Menge Pa_4 der Eltern von X_4 minimal sein soll, zeichnen wir also *nur einen Pfeil* von X_2 nach X_4.

Also nächstes kommt der Schlaganfall X_5. Er wird von allen bisher in den Graphen eingefügten Variablen begünstigt, auf direktem Weg allerdings nur durch die Hypertonie X_3 (erhöhter Blutdruck bringt Gefäßwände zum Platzen) und durch die arterielle Embolie X_4 (der Embolus verstopft die Blutgefäße). Gemäß der Minimalbedingung zeigen also nur von X_3 und X_4 aus Pfeile nach X_5.

X_6, die Taubheit einer Körperseite, ist ein häufiges Symptom des Schlaganfalls. Die übrigen Variablen machen diese Taubheit nur dadurch wahrscheinlicher, dass sie den Schlaganfall begünstigen. Wir zeichnen nur einen Pfeil von X_5 nach X_6.

Ebenso verfahren wir mit X_7, der einseitigen Lähmung. Nur mittels des Schlaganfalls nehmen die übrigen Variablen des Graphen auf sie Einfluss.



Damit steht das Gerüst unseres Bayesschen Graphen. Im nächsten Schritt müssen wir für jede Variable die elternbedingten Wahrscheinlichkeitsverteilungen in Tabellenform spezifizieren. Die Knoten X_1 und X_2, die keine Eltern haben, werden nicht auf andere Variablen bedingt. In unserem Beispiel seien folgende Verteilungen gegeben:

X_1 Schrumpfniere

gegeben:	ja	nein
---	0,01	0,99

X_2 Arteriosklerose

gegeben:	ja	nein
---	0,15	0,85

X_3 Hypertonie

gegeben:	ja	nein
X_1 nein, X_2 nein	0,1	0,9
X_1 ja, X_2 nein	0,8	0,2
X_1 nein, X_2 ja	0,5	0,5
X_1 ja, X_2 ja	0,9	0,1

X_4 arterielle Embolie

gegeben:	ja	nein
X_2 nein	0,01	0,99
X_2 ja	0,05	0,95

X_5 Schlaganfall

gegeben:	ja	nein
X_3 nein, X_4 nein	0,001	0,999
X_3 ja, X_4 nein	0,02	0,98
X_3 nein, X_4 ja	0,5	0,5
X_3 ja, X_4 ja	0,6	0,4

X_6 Taubheit einer Körperseite

gegeben:	ja	nein
X_5 nein	0,001	0,999
X_5 ja	0,7	0,3

X_7 einseitige Lähmung

gegeben:	ja	nein
X_5 nein	0,01	0,99
X_5 ja	0,7	0,3

Interessiert uns beispielsweise die Wahrscheinlichkeit eines Schlaganfalls, wenn wir wissen, dass der Patient zwar eine arterielle Embolie (X_4=„ja“) hat, aber keine Hypertonie (X_3=„nein“), so entnehmen wir der Tabelle von X_5 in der Zeile „X_3 nein, X_4 ja“ die Wahrscheinlichkeit 0,5. Im Abschnitt zur Inferenz werden wir kompliziertere Fragen an unseren Bayesschen Graphen stellen.

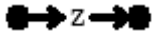
Wir erkennen wieder die Effektivität der Darstellung. Eine herkömmliche gemeinsame Verteilung der Variablen X_1, ..., X_7 mit ihren jeweils zwei Zuständen würde eine vollständige Tabellierung der Wahrscheinlichkeiten aller 2^7=128 Situationen erfordern. Hier kommen wir mit bereits 32 Zahlen aus.

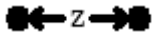
In der Praxis wird der Graph, der die direkten Bedingtheiten darstellt, zumeist von Experten erstellt, während die Wahrscheinlichkeitstabellen häufig durch statistische Erhebungen zustande kommen. Dabei treten bereits bei der Konstruktion des Graphen Schwierigkeiten auf: Es ist nicht nur erforderlich, (direkte) Abhängigkeiten zu erkennen und durch Pfeile zu verdeutlichen. Sondern wann immer zwei Variablen *nicht* verbunden werden, müssen sie auch unabhängig voneinander sein, oder dürfen höchstens mittelbar über andere Variablen des Bayesschen Graphen in Zusammenhang stehen, die durch Pfeile verbunden sind. Diese impliziten Informationen sauber abzubilden, stellt letztendlich das größere Problem dar.

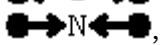
III d-Separation

Bayessche Graphen stellen eine Adaption an die Struktur der Bedingtheiten (der stochastischen Zusammenhänge) von Zufallsvariablen dar. Sie ermöglichen es, das Relevante auf einen Blick zu *sehen* und das Irrelevante zu ignorieren. Allein in der Topologie des Graphen offenbaren sich bedingte (Un-) Abhängigkeiten der Wahrscheinlichkeitsverteilung. Fundamental ist in diesem Zusammenhang das d-Separationskriterium, welchem wir uns nun nähern werden. Zunächst definieren wir, was es bedeutet, dass ein Weg in einem Bayesschen Graphen **blockiert** ist. Dabei seien **X**, **Y** und **Z** stets paarweise disjunkte Mengen von Variablen des Graphen. Unter einem *Weg* verstehen wir eine endliche Folge von durch Pfeile (beliebiger Richtung) verbundenen Knoten, die sich nicht wiederholen dürfen.

Definition: Ein Weg in einem Bayesschen Graphen heißt *von* **Z** *blockiert*, falls in ihm (mindestens) eine der folgenden Konfigurationen vorkommt ($Z \in \mathbf{Z}$, die schwarzen Punkte sind beliebige Knoten des Weges):

1. seriell: 

2. divergent: 

3. konvergent: 

, wobei weder N, noch ein Nachfolger von N Variable aus **Z** sein darf.

Ein Weg, der nicht blockiert ist (in dem also keine der drei Konfigurationen auftritt), heißt *frei*. (14)

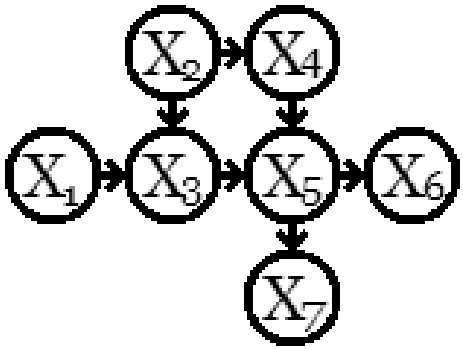
Definition: **X** und **Y** heißen *von* **Z** *blockiert*, falls jeder beliebige Weg von einem Knoten $X \in \mathbf{X}$ zu einem Knoten $Y \in \mathbf{Y}$ durch **Z** im Sinne von (14) blockiert ist. (15)

Im einfachen Fall von **X**={X} und **Y**={Y} bedeutet also Blockade von **X** und **Y** durch **Z**, dass in *jedem* Weg von X nach Y *mindestens eine* der drei obigen Konfigurationen auftaucht. Sind andererseits **X** und **Y** nicht durch **Z** blockiert, so meint dies, dass es *mindestens einen* Weg von X nach Y gibt, in dem *keine* der drei obigen Konfigurationen vorkommt.

Die Idee dahinter (und der Grund für die Sprechweise zu „blockieren“) ist die, dass über die Wege *Informationsflüsse* stattfinden, die unterbrochen (also blockiert) oder frei sein können.

Betrachten wir dazu zunächst die **serielle Konfiguration** X_1 -> X_3 -> X_5 aus unserem Beispiel von vorhin:

Die Information, dass der Patient an Schrumpfniere (X_1=„ja“) leidet, steigert (über eine erhöhte Wahrscheinlichkeit der Hypertonie, X_3) die Wahrscheinlichkeit eines Schlaganfalls X_5. Ist jedoch bereits bekannt, dass der Patient an Hypertonie leidet, so ändert sich die Wahrscheinlichkeit eines Schlaganfalls nicht mehr durch das Wissen, dass er eine Schrumpfniere hat, weil diese nur über die bereits instanzierte Variable Hypertonie auf die Wahrscheinlichkeit des Schlaganfalls Einfluss nimmt. Wird also X_3=„ja“ (oder alternativ auch X_3=„nein“) bekannt, so blockiert dies den seriellen Informationsfluss von X_1 zu X_5.



Wie sieht es mit der **divergenten Konfiguration** aus? Wir wählen beispielhaft X_7 <- X_5 -> X_6. In der Regel sind X_7, die einseitige Lähmungserscheinung, und X_6, die Taubheit einer Körperseite, voneinander abhängig als gemeinsame Symptome derselben Ursache: Wird bekannt, dass der Patient einseitig Taubheit verspürt, so steigt automatisch die Wahrscheinlichkeit eines Schlaganfalls als typische Ursache für diese Taubheit. Ist die Wahrscheinlichkeit eines Schlaganfalls aber erhöht, so steigt auch die Wahrscheinlichkeit des zweiten Symptoms, der Lähmung. Ist im Gegensatz dazu aber bereits im Voraus bekannt, dass der Patient einen Schlaganfall erlitten hat, so ändert das Wissen, dass er an Taubheit einer Körperhälfte leidet, nichts an der Wahrscheinlichkeit, dass er eine einseitige Lähmung hat. Diese ist konstant gleich hoch, da der Schlaganfall ja mit Sicherheit vorliegt. Analog lässt sich argumentieren, wenn bekannt wird, dass die Variablen „nein“ anzeigen. Daher blockiert die Instanzierung von X_5 den Informationsfluss.

Die konvergente Konfiguration verhält sich dem gegenüber umgekehrt: Wir betrachten X_3 -> X_5 <- X_4. In der Regel sind X_3, die Hypertonie (über das Platzen von Gefäßen), und X_4, die arterielle Embolie (über die Verstopfung von Gefäßen), unabhängige Ursachen einer gemeinsamen Wirkung, nämlich des Schlaganfalls X_5. X_3 und X_4 sind also zunächst einmal blockiert. Das ändert sich nun aber durch Bekanntwerden des Schlaganfalls: Dann sorgt eine Verringerung der Wahrscheinlichkeit der arteriellen Embolie für eine Steigerung der Wahrscheinlichkeit der Hypertonie. Schließlich muss die Wirkung ja irgendeinen Grund haben. Sinkt die Wahrscheinlichkeit des einen, so steigt automatisch die des anderen. Bei Konvergenz von Knoten sorgt also die Instanzierung einer Wirkung für eine (gegenläufige) Kopplung der möglichen Ursachen. Dabei spielt es keine Rolle, ob die direkte gemeinsame Wirkung (in unserem Beispiel der Schlaganfall X_5) oder eine Wirkung dieser Wirkung, also ein Nachfolger des Konvergenzknotens (wie beispielsweise die einseitige Lähmung X_7) instanziert wird.

Wir betrachten nun zur Verdeutlichung folgende Beispiele an unserem Bayesschen Graphen:

$\{X_3\}$ und $\{X_4\}$ sind von $\{X_2\}$ blockiert, weil einerseits der Weg X_3 <- X_2 -> X_4 durch Instanzieren von X_2 blockiert wird (divergent), andererseits X_3 -> X_5 <- X_4 ohnehin blockiert ist als Konvergenzkonfiguration, in der weder X_5 noch einer der Nachfolger X_6 oder X_7 instanziert ist (d. h. zu $\{X_2\}$ gehört). Weitere Wege zwischen $\{X_3\}$ und $\{X_4\}$ gibt es nicht.

{X1} und $\{X_6, X_7\}$ sind durch $\{X_3\}$ nicht blockiert. Zwar sind (seriell) die Wege X_1 -> X_3 -> X_5 -> X_6 und auch X_1 -> X_3 -> X_5 -> X_7 durch $\{X_3\}$ blockiert, aber es gibt (und das genügt, damit die Knotenmengen insgesamt nicht blockiert sind) den freien Weg X_1 -> X_3 <- X_2 -> X_4 -> X_5 -> X_6. (Dabei ist die Konvergenz um X_3 frei, weil X_3 instanziert ist.)

$\{X_1\}$ und $\{X_6\}$ sind durch $\{X_5\}$ blockiert: Es genügt bereits, dass in jedem Weg zwischen $\{X_1\}$ und $\{X_6\}$, nämlich einerseits X_1 -> X_3 <- X_2 -> X_4 -> X_5 -> X_6 und andererseits X_1 -> X_3 -> X_5 -> X_6 die seriellen Blockaden X_4 -> X_5 -> X_6 bzw. X_3 -> X_5 -> X_6 vorkommen.

Wie wir durch die Interpretation als Informationsfluss plausibel gemacht haben, gibt die *Topologie des Graphen* Auskunft über bedingte Ab- und Unabhängigkeiten der gemeinsamen Verteilung der Variablen, deren Bedingtheitsstruktur von dem Bayesschen Graphen dargestellt wird. Präzise fasst diesen Zusammenhang von Graphen- und Wahrscheinlichkeitstheorie das

d-Separationskriterium: Es sei ein Bayesscher Graph G gegeben, und X, Y, Z seien verschiedene Knoten von G. Dann gilt:

1. Sind X und Y (als Knoten) von Z blockiert gemäß Definition (15), so sind X und Y (als Variablen) bedingt unabhängig gegeben Z gemäß Definition (7) bezüglich jeder gemeinsamen Verteilung der Knotenvariablen in G, deren Bedingtheitsstruktur von G dargestellt wird.
2. Sind X und Y von Z (als Knoten) frei, so sind X und Y (als Variablen) bedingt abhängig gegeben Z bezüglich *vieler* gemeinsamer Verteilungen der Knotenvariablen in G, deren Bedingtheitsstruktur von G dargestellt wird. (16)

Damit ist die Sprechweise des „Blockierens“ endgültig legitimiert.

Es sei angemerkt, dass sich das Kriterium auch auf disjunkte Mengen **X**, **Y**, **Z** von Variablen verallgemeinern lässt.

Die Sprechweise „bezüglich *vieler* gemeinsamer Verteilungen“ spricht die Existenz einer solchen Verteilung aus, deutet aber darüber hinaus darauf hin, dass dies dann quasi für alle gilt. Das liegt daran, dass Unabhängigkeit, die nicht durch den Bayesschen Graphen dargestellt wird, die Ausnahme ist. Nur im Falle einer sehr speziellen (und dadurch höchst unwahrscheinlichen) Justierung der Wahrscheinlichkeitstabellen gelingt es, die Multiplikationsregel (3) exakt zu erfüllen.

Im nächsten Abschnitt werden wir sehen, wie sich die d-Separation nutzen lässt, um geschickt Informationen aus einem Bayesschen Graphen abzulesen.

IV Inferenz

Wir untersuchen als nächstes die im letzten Abschnitt eingeführten Informationsflüsse quantitativ. Typischerweise interessiert die Wahrscheinlichkeitsverteilung einer oder mehrerer der Variablen, wenn gewisse Beobachtungen, *Evidenzen*, vorliegen, d. h. wenn die Werte einiger Variablen bekannt (instanziert) werden (mit Wahrscheinlichkeit 1). Zum Beispiel könnte die Frage lauten, wie wahrscheinlich es ist, dass ein Patient einen Schlaganfall hat, wenn er Taubheit auf einer Körperseite verspürt, und ferner bekannt ist, dass er keine Schrumpfniere hat. Es ist dann $P(X_5$=„ja“$|X_1$=„nein“, X_6=„ja“) zu bestimmen, wobei X_1=„nein“ und X_6=„ja“ gemeinsam die Evidenz bilden.

Solche Fragen lassen sich, wie wir gleich am Beispiel nachvollziehen werden, immer mit Hilfe von (3), (4) und (11) beantworten, da vollständige Information über die gemeinsame Verteilung $P(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ aller Variablen vorliegt. (Schließlich wurden in den Tabellen der Variablen nur solche Bedingtheiten weg gelassen, die durch andere Bedingtheiten zu bedingten Unabhängigkeiten werden.) Wir kodieren im Folgenden der Einfachheit halber die möglichen Werte, welche die Zufallsvariablen annehmen können, mit 0 für „nein“ und mit 1 für „ja“. Es gilt beispielsweise nach (3)

$$P(X_5 = 1 \mid X_1 = 0, X_6 = 1) = \frac{P(X_1 = 0, X_5 = 1, X_6 = 1)}{P(X_1 = 0, X_6 = 1)} \tag{17}$$

In einen allgemeineren Kontext gestellt lautet das verbleibende Problem nun, dass zwar eine gemeinsame Verteilung $P(\mathbf{X}=\mathbf{x}, \mathbf{Y}=\mathbf{y})$ bekannt ist, tatsächlich aber nur die Wahrscheinlichkeit $P(\mathbf{X}=\mathbf{x})$ für die Werte **x** einer Auswahl **X** von Variablen gesucht ist, in der die übrigen Werte (hier mit **y** bezeichnet) der übrigen Variablen **Y** nicht auftauchen. Dieses Problem wird durch so genanntes *Marginalisieren* gelöst. Man summiert dabei einfach über die möglichen Werte **y** der nicht relevanten Variablen **Y**:

$$P(\mathbf{X}=\mathbf{x})=\sum_{\mathbf{y}} P(\mathbf{X}=\mathbf{x},\mathbf{Y}=\mathbf{y})\,. \tag{18}$$

Dadurch berechnet man – ausgedrückt durch die gemeinsame Verteilung – die Wahrscheinlichkeit, dass **X**=**x** ist, wie gefordert, und die **Y** *irgendwelche Werte* annehmen, was genau die relevante Wahrscheinlichkeit P(**X**=**x**) ergibt. Mit dieser Methode erhalten wir aus (17)

$$P(X_5=1|X_1=0,X_6=1)=\frac{\sum_{x_2,x_3,x_4,x_7\in\{0,1\}} P(X_1=0,X_2=x_2,X_3=x_3,X_4=x_4,X_5=1,X_6=1,X_7=x_7)}{\sum_{x_2,x_3,x_4,x_5,x_7\in\{0,1\}} P(X_1=0,X_2=x_2,X_3=x_3,X_4=x_4,X_5=x_5,X_6=1,X_7=x_7)}\,. \tag{19}$$

Darauf wenden wir nacheinander die Kettenregel (4) und die bedingten Unabhängigkeiten, die dem Bayesschen Graphen zu entnehmen sind, an. Wir sind so bereits in (11) vorgegangen.

$$P(X_5=1|X_1=0,X_6=1)= \tag{20}$$

$$\frac{\sum_{x_2,x_3,x_4,x_7\in\{0,1\}} P(X_1=0)P(X_2=x_2)P(X_3=x_3|X_1=0,X_2=x_2)P(X_4=x_4|X_2=x_2)P(X_5=1|X_3=x_3,X_4=x_4)P(X_6=1|X_5=1)P(X_7=x_7|X_5=1)}{\sum_{x_2,x_3,x_4,x_5,x_7\in\{0,1\}} P(X_1=0)P(X_2=x_2)P(X_3=x_3|X_1=0,X_2=x_2)P(X_4=x_4|X_2=x_2)P(X_5=x_5|X_3=x_3,X_4=x_4)P(X_6=1|X_5=x_5)P(X_7=x_7|X_5=x_5)}$$

Hier wird also (was dank des d-Separationskriteriums (16) erlaubt ist) nur jeweils auf die Eltern bedingt, die aus dem Graphen abzulesen sind. Da X_1 und X_2 keine Eltern haben, wird bei ihnen nicht bedingt. Wir halten fest, dass wir die Wahrscheinlichkeiten in (20) nun bereits in eine Form gebracht haben, wie sie in den Tabellen auf Seite 4 gegeben sind. Doch bevor wir die Zahlenwerte einsetzen, ziehen wir noch diverse Faktoren aus den Summen heraus. Dabei werden möglichst viele Faktoren so weit wie möglich nach links „verschoben“, wobei kein Laufindex seine zugehörige Summe „verlassen“ darf.

$$P(X_5=1|X_1=0,X_6=1)= \tag{21}$$

$$\frac{P(X_1=0)P(X_6=1|X_5=1)\sum_{x_2\in\{0,1\}}P(X_2=x_2)\sum_{x_3\in\{0,1\}}P(X_3=x_3|X_1=0,X_2=x_2)\sum_{x_4\in\{0,1\}}P(X_4=x_4|X_2=x_2)P(X_5=1|X_3=x_3,X_4=x_4)\sum_{x_7\in\{0,1\}}P(X_7=x_7|X_5=1)}{P(X_1=0)\sum_{x_2\in\{0,1\}}P(X_2=x_2)\sum_{x_3\in\{0,1\}}P(X_3=x_3|X_1=0,X_2=x_2)\sum_{x_4\in\{0,1\}}P(X_4|X_2)\sum_{x_5\in\{0,1\}}P(X_5|X_3=x_3,X_4=x_4)P(X_6=1|X_5=x_5)\sum_{x_7\in\{0,1\}}P(X_7=x_7|X_5=x_5)}$$

Eine längere Rechnung liefert nun $P(X_5=1|X_1=0, X_6=1)=0{,}8978\ldots$.

Wie bereits angesprochen, lässt sich auf diesem Weg (Marginalisierung und Ausnutzen der bedingten Unabhängigkeiten) jede Evidenz in einem beliebigen Bayesschen Graphen handhaben. Je nach Struktur des Graphen existieren jedoch alternativ diverse Algorithmen, die Inferenz effektiver realisieren.

Wir schränken unsere genauere Betrachtung ein auf **message–passing in Ketten.** Das heißt, unser Graph ist einfach eine Kette von aufeinander folgenden Knoten; je einer (außer der erste) ist einziges Kind eines anderen. Wir sind an der Wahrscheinlichkeitsverteilung eines Knotens X interessiert.
Für das nun Folgende unterscheiden wir zwei Arten der Evidenz.

1. *Kausale Evidenz* E^+ liegt vor, wenn ein Knoten instanziert wird, der Vorgänger von X ist.
2. *Diagnostische Evidenz* E^- liegt vor, wenn ein Knoten instanziert wird, der Nachfolger von X ist.

Eine Evidenz E setzt sich im Allgemeinen zusammen aus E^+ und E^-. Für jeden Knoten X definieren wir

1. seine *kausale Unterstützung* $\pi(X):=P(x|e^+)$, welche als Einfluss der Ursachen-Evidenz auf die Wahrscheinlichkeit von X zu interpretieren ist, sowie
2. seine *diagnostische Unterstützung* $\lambda(X):=P(e^-|x)$, welche wir als Einfluss der Wirkungs-Evidenz auf die Wahrscheinlichkeit von X auffassen können.

Dabei sind $\pi(X)$ und $\lambda(X)$ als *Vektoren* zu verstehen. In den einzelnen Komponenten stehen die Werte für $P(x|e^+)$ bzw. $P(e^-|x)$, wobei x dabei die möglichen Werte von X durchläuft (x=0 bei der ersten und x=1 bei der zweiten Komponente). In unserem Beispiel haben wir es also mit zweikomponentigen Vektoren

$$\pi(X)=(\,P(X=0|e^+);\,P(X=1|e^+)\,)\quad\text{und}\quad\lambda(X)=(\,P(e^-|X=0);\,P(e^-|X=1)\,) \tag{22}$$

zu tun. Benutzen wir im Folgenden Kleinbuchstaben, so ist eine (allgemeine) Komponente gemeint.

Mit αV (wobei V ein beliebiger Wahrscheinlichkeitsvektor ist) bezeichnen wir die *Normierung*, welche die Summe der Vektorkomponenten verhältniserhaltend zu 1 normiert. Z. B. ist α(2/5; 4/5)=(1/3; 2/3).

Zwei Tatsachen motivieren die Einführung von π und λ:
Einerseits gilt (mit komponentenweiser Multiplikation)

$$P(X|E^+, E^-)= \alpha\, \pi(X)\, \lambda(X). \tag{23}$$

Zur Begründung: Es ist nach (1), und weil X die Knoten E^+ und E^- seriell blockiert (und daher E^+ und E^- nach dem d-Separationskriterium (16) bedingt unabhängig gegeben X sind) für jede Komponente

$$\frac{P(x, e^+, e^-)}{P(x, e^-)} = P(e^+ \mid x, e^-) = P(e^+ \mid x) = \frac{P(x, e^+)}{P(x)}, \text{d. h. } P(x, e^+, e^-) = \frac{P(x, e^-)\, P(x, e^+)}{P(x)}. \tag{24}$$

Mit dieser Hilfsgleichung beweisen wir nun (23) folgendermaßen:

$$P(x \mid e^+, e^-) = \frac{P(x, e^+, e^-)}{P(e^+, e^-)} = \frac{P(e^+)}{P(e^+, e^-)} \frac{P(x, e^+, e^-)}{P(e^+)} = \frac{1}{P(e^- \mid e^+)} \frac{P(x, e^+)}{P(e^+)} \frac{P(x, e^-)}{P(x)} = \frac{1}{P(e^- \mid e^+)} \pi(x)\, \lambda(x). \tag{25}$$

Dabei haben wir sogar die Normierungskonstante $\alpha = \frac{1}{P(e^- \mid e^+)}$ in (23) bestimmt. (Sie ist deshalb als Konstante zu bezeichnen, weil sie – da unabhängig von x - in jeder Komponente des Vektors gleich ist.) In der Praxis braucht α nicht extra ausgerechnet zu werden. Man bestimmt es einfach durch die Bedingung

$$\alpha \sum_x \pi(x)\, \lambda(x) = 1. \tag{26}$$

Die Bedeutung von (23) liegt darin, dass kausale und diagnostische Unterstützung durch *Multiplikation* die gesamte Wahrscheinlichkeit $P(x|e^+, e^-)$ ergeben. Hierin zeigt sich die *Unabhängigkeit* beider Anteile.

Die zweite Tatsache ist die, dass π(X) berechnet werden kann, wenn nur π(Y) des Elternteils Y bekannt ist, ohne dass dazu irgendwelche λ-Werte benötigt werden, und dass λ(X) berechnet werden kann, wenn nur λ(Z) des Kindes Z bekannt ist, ohne dass dazu irgendwelche π-Werte erforderlich sind. Es gilt nämlich

$$\pi(x) = \sum_y P(x \mid y)\, \pi(y) \quad \text{und} \quad \lambda(x) = \sum_z P(z \mid x)\, \lambda(z). \tag{27}$$

wobei P(x|y) bzw. P(z|x) einfach den bedingten Verteilungen (den Tabellen) der Knoten X bzw. Z zu entnehmen sind.
Wir weisen nun die erste Gleichung von (27) nach. Durch Marginalisierung bezüglich y und anschließende Anwendung der Kettenregel (4) folgt

$$\pi(x) = P(x \mid e^+) = \frac{P(x, e^+)}{P(e^+)} = \frac{\sum_y P(x, y, e^+)}{P(e^+)} = \frac{\sum_y P(e^+)\, P(y \mid e^+)\, P(x \mid y, e^+)}{P(e^+)} = \sum_y P(x \mid y, e^+)\, P(y \mid e^+). \tag{28}$$

Nutzen wir jetzt aus, dass Y die Knoten E^+ und X seriell blockiert, so folgt mit dem d-Separationskriterium (16) die erste Gleichung, dadurch, dass wir statt $P(x|y, e^+)$ einfach P(x|y) schreiben.
Analog kann man sich von der Richtigkeit der zweiten Gleichung überzeugen.

Die Propagationsregeln (27) zeigen, wie sich die beiden Anteile der Evidenz (kausaler und diagnostischer) *voneinander entkoppelt* in einer Kette von Variablen *ausbreiten*, was einer algorithmischen Formulierung zugänglich ist: Es sind bloß zwei Schritte erforderlich, die einander nicht beeinflussen, nämlich der π-Fluss von E^+ zu X in Pfeilrichtung und der λ-Fluss von E^- zu X entgegen der Pfeilrichtung. Sobald π(X) und λ(X) bekannt sind, kann schließlich mittels (23) die gesuchte Wahrscheinlichkeitsverteilung des Knotens $P(X|E^+, E^-)$ berechnet werden.
Als Startwerte wählt man sinnvollerweise für die λ-Vektoren aller Knoten den Einsvektor (1; 1), was unter der Regel (27) für λ invariant ist. Die Komponenten des π-Vektors des Wurzelknotens der Kette sind

einfach seine nicht bedingten Wahrscheinlichkeiten für die möglichen Werte 0 und 1. Dann kann man mit (27) die übrigen π-Vektoren ausrechnen.

Damit haben wir ein erstes so genanntes *message-passing* bereits durchgeführt, und zwar dasjenige für den Fall, dass *keine* Evidenz vorliegt. Mit (23) erhalten wir dann die Wahrscheinlichkeiten für die Werte 0 und 1 jedes beliebigen Knoten des Bayesschen Graphen (der ja bloß eine Kette ist).

Bei Evidenzknoten E=0 bzw. E=1 setzen wir $\lambda(E):=(1; 0)$ bzw. (0; 1) und ebenso $\pi(E):=(1; 0)$ bzw. (0; 1).

Einige einfache, aber typische Fälle des message-passing führen wir jetzt beispielhaft durch. Dabei legen wir eine vereinfachte Version des Bayesschen Graphen aus Abschnitt III zugrunde, der nur aus den Knoten X_1, X_3, X_5, und X_6 besteht, sodass wir es mit einer Kette zu tun haben. Außerdem verwenden wir vereinfachte Wahrscheinlichkeitstabellen.

X_1 Schrumpfniere

gegeben:	ja (1)	nein (0)
---	0,01	0,99

X_3 Hypertonie

gegeben:	ja (1)	nein (0)
X_1 nein	0,1	0,9
X_1 ja	0,8	0,2

X_5 Schlaganfall

gegeben:	ja (1)	nein (0)
X_3 nein	0,001	0,999
X_3 ja	0,02	0,98

X_6 Taubheit einer Körperseite

gegeben:	ja (1)	nein (0)
X_5 nein	0,001	0,999
X_5 ja	0,7	0,3

Erstes Beispiel: $P(X_5=1|X_3=1)$

Das ist die einfachste Situation. Der Tabelle von X_5 entnehmen wir direkt die Wahrscheinlichkeit 0,02. Mit message-passing kommen wir natürlich zu demselben Ergebnis, wobei wir X_1 ignorieren dürfen, da X_3 diesen Knoten seriell von X_5 separiert: Mit der Evidenz $\pi(X_3)=(0, 1)$ folgt nach (27) für X_5 wie zu erwarten $\pi(X_5)=(P(X_5=0|X_3=0)*0 + P(X_5=0|X_3=1)*1; P(X_5=1|X_3=0)*0 + P(X_5=1|X_3=1)*1)=(0,98; 0,02)$. Der λ-Vektor von X_5 ist (1; 1), da keine diagnostische Evidenz vorliegt, sodass wir ihn in (23) ignorieren können. (Zwar ist $\lambda(X_3)=(0; 1)$, aber der λ-Fluss erfolgt entgegen der Pfeilrichtung, sodass X_5 unbeeinflusst bleibt.)

Zweites Beispiel: $P(X_5=1|X_1=1)$

Aus $\pi(X_1)=(0; 1)$ folgt $\pi(X_3)=(P(X_3=0|X_1=0)*0 + P(X_3=0|X_1=1)*1; P(X_3=1|X_1=0)*0 + P(X_3=1|X_1=1)*1)$, also $\pi(X_3)=(0,2, 0,8)$. Als nächstes errechnen wir hieraus analog

$\pi(X_5)=(P(X_5=0|X_3=0)*0,2 + P(X_5=0|X_3=1)*0,8; P(X_5=1|X_3=0)*0,2 + P(X_5=1|X_3=1)*0,8)$.

Es folgt nach Einsetzen der Zahlen $\pi(X_5)=(0,9838; 0,0163)$. Wegen $\lambda(X_5)=(1, 1)$ ist $\pi(X_5)$ bereits nach (23) der endgültige Wahrscheinlichkeitsvektor von X_5 und wir lesen ab: $P(X_5=1|X_1=1)=0,0163$.

Drittes Beispiel: $P(X_5=1|X_6=1)$

Hier kommt zum ersten Mal die diagnostische Unterstützung ins Spiel. Es gilt $\lambda(X_6)=(0; 1)$, was sich auf $\lambda(X_5)=(P(X_6=0|X_5=0)*0 + P(X_6=1|X_5=0)*1; P(X_6=0|X_5=1)*0 + P(X_6=1|X_5=1)*1)=(0,001; 0,7)$ auswirkt. Nun errechnen wir noch aus $\pi(X_1)=(0,99; 0,01)$ den Vektor $\pi(X_5)=(0,996967; 0,003033)$ auf analoge Weise, wie schon im zweiten Beispiel. Es folgt dann gemäß (23) für den Wahrscheinlichkeitsvektor von X_5 unter Beachtung dessen, dass die Vektoren komponentenweise multipliziert werden,

$$\alpha\, \pi(X)\, \lambda(X) = \alpha(0,996967; 0,003033)\ (0,001; 0,7) = \alpha(0,000996967; 0,0021231) \approx (0,31953; 0,68047).$$

Wir lesen $P(X_5=1|X_6=1)=0,68047$ ab, was wir übrigens alternativ auch mittels der Inversionsformel (5) hätten berechnen können, wobei dann allerdings auch $P(X_5=1)$ und $P(X_6=1)$ zu bestimmen gewesen wäre.

Viertes Beispiel: $P(X_5=1|X_1=0, X_6=1)$

Jetzt kombinieren wir kausale und diagnostische Evidenzen. Erst in diesem Beispiel profitieren wir wirklich von der Entkopplung beider Anteile. Im dritten Beispiel haben wir bereits $\lambda(X_5)=(0,001; 0,7)$ bestimmt. Außerdem berechnen wir mit $\pi(X_1)=(1; 0)$ analog zu vorhin als $\pi(X_5)=(0,9971; 0,0029)$. Schließlich liefert (23) dann für X_5 den Wahrscheinlichkeitsvektor

$$\alpha\, \pi(X)\, \lambda(X) = \alpha(0,9971; 0,0029)\ (0,001; 0,7) = \alpha(0,0009971; 0,00203) \approx (0,32939; 0,67061),$$

woraus wir $P(X_5=1|X_1=0, X_6=1)=0,67061$ ablesen. Zu wissen, dass der Patient keine Schrumpfniere hat, verringert also die Wahrscheinlichkeit eines Schlaganfalls bei vorliegender Taubheit einer Körperseite geringfügig. Die Tatsache, dass ein anderes Ergebnis als bei der Marginalisierung auf Seite 7 herauskommt, rührt daher, dass wir den Graphen zur Vereinfachung abgewandelt haben.

Prinzipiell funktioniert das message-passing in allgemeineren Graphenstrukturen, wie Bäumen, ähnlich. Es ist dann allerdings erforderlich, sich Gedanken darüber zu machen, wie sich der π-Fluss an Knoten mit mehreren Kindern bzw. der λ-Fluss an Knoten mit mehreren Eltern aufteilt.

V Kausalität

Bisher haben wir nicht darüber gesprochen, was geschieht, wenn wir bei der Konstruktion eines Bayesschen Graphen die Reihenfolge ändern, in der die Zufallsvariablen in den Graphen eingefügt werden. Die Richtung der Pfeile, welche die Bedingtheiten der Variablen anzeigen, hängt (nach Konstruktion des Graphen) nur von dieser Reihenfolge ab. Wird eine Variable zum Beispiel erst spät in den Graphen eingefügt, so zeigen tendenziell viele Pfeile auf sie, wird sie früh eingefügt, so zeigen wenige Pfeile auf sie, eventuell gehen mehr Pfeile von ihr aus. Wir haben auch gesehen, dass der Begriff der (bedingten) stochastischen Unabhängigkeit symmetrisch ist. Es zeigt sich, dass die Pfeile in Bayesschen Graphen eigentlich überflüssig sind. Das d-Separationskriterium (16) ist ebenfalls unabhängig von den Pfeilrichtungen (d. h., es ist unabhängig von der Reihenfolge, in welcher der Graph aus den Variablen aufgebaut wurde). Hinzu kommt, dass es manchmal gar nicht möglich ist, Variablen in kausalen Zusammenhang zu stellen und eine als Ursache der anderen zu bezeichnen. Bei Arteriosklerose und Hypertonie ist dies zum Beispiel nicht ohne Weiteres möglich. Gleichwohl koinzidieren sie. Nun mag man einwenden, dass ein Bayesscher Graph, der genauer und detaillierter ist, als unser Beispiel, diese Koinzidenz durch kausal zusammenhängende Variablen ersetzen kann. Allerdings ist dies oft aufgrund von unangemessen hohem Aufwand oder Unwissen über die genauen Wirkmechanismen nicht realisierbar. Dennoch liegt es in unserer menschlichen Natur, den einzelnen Variablen Eigenschaften, wie Ursache und Wirkung, oder auch eine zeitliche Reihenfolge zuzuordnen, und das mit Pfeilen zum Ausdruck zu bringen. Die Interpretation des d-Separationskriteriums mithilfe von Informationsflüssen ist sehr eingängig und geht verloren, wenn wir auf die Pfeile verzichten oder sie entgegen der *kausal richtigen* Richtung ziehen. Es ist jedoch nicht zu verschweigen, dass die Stochastik Kausalität „eigentlich“ nicht kennt. Schließlich ist auch die Inversionsformel (5) symmetrisch und erlaubt es nicht, zu unterscheiden. Trotzdem gibt es im Wesentlichen drei Gründe, wieso Bayessche Graphen, mit Pfeilen, die – wann immer möglich - die kausalen Wirkzusammenhänge darstellen, von Vorteil sind:

1. Die Glaubwürdigkeit des Modells ist größer. Zwar „funktioniert“ die Stochastik ohne Kausalität, doch die Akzeptanz von Technologien ist größer, wenn sie den Menschen vermittelbar ist. Kausale Bayessche Graphen liefern eine *Erklärung* für die Vorschläge, die sie uns machen. Gerade in Bereichen von Sicherheit und medizinischer Diagnostik, wo Bayessche Graphen eingesetzt werden, ist das ein wesentlicher Gesichtspunkt.
2. Es fällt Experten, die ihr Wissen in einem Bayesschen Graphen ausdrücken sollen, deutlich leichter, diesen nach kausalen Beziehungen zu ordnen, da Menschen nun einmal kausal denkende Wesen sind. Somit ist die Konstruktion der Graphen erleichtert. Wir benötigen einfach keine umfangreichen statistischen Daten, um zu *wissen*, dass eine Rot-Grün-Sehschwäche (im Wesentlichen) unabhängig vom Vornamen ist. Darüber hinaus würde der Graph deutlich unübersichtlicher, wenn er nicht nach kausalen Prinzipien konstruiert wäre. Dadurch würde er fehleranfälliger, benötigte mehr Speicherkapazität und die Inferenz wäre problematischer. (Beispielsweise würde aus der Kette ein Baum, über dessen message-passing wir hier kaum etwas gesagt haben.)
3. Schließlich ist es leichter, einen kausalen Bayesschen Graphen zu *ändern*, wenn die Rahmenbedingungen wechseln. Das ist nicht zu verwechseln damit, dass sich gemessene Daten ändern, die von den Variablen des Graphen repräsentiert werden. Verändertes Wissen über eine Situation ändert vielleicht die probabilistischen, nicht aber die kausalen Zusammenhänge. Gemeint ist vielmehr, eine grundlegende Änderung der Gesamtsituation durch äußere Eingriffe. Beispielsweise ist es denkbar, dass in unserem Beispiel die Lähmung durch Betäubungsmittel *von außen* herbeigeführt wird. In diesem Fall ist es klar, dass dadurch keine direkte Beziehung zur Schrumpfniere entsteht. Orientiert man sich an der Kausalität, so ist es offensichtlich, dass gewisse Änderungen sich nur lokal auf den Graphen auswirken. Gegebenenfalls ist nur eine einzige Variable hinzuzufügen, oder ein Pfeil zu entfernen. Eine Situation zu *verstehen* heißt, sie von außen zu betrachten und auf Veränderungen reagieren zu können. Daher sind kausale Bayessche Graphen auch geeignet für den Einsatz in Systemen, die *intelligent* agieren sollen.

Es verbleibt die Frage, was Kausalität denn eigentlich ist. Wie vorhin dargelegt, existiert sie vom wahrscheinlichkeitstheoretischen Standpunkt aus nicht. Kausalität scheint eine *menschliche Interpretation der Welt* zu sein, die einfach ungemein praktisch ist. In der Mathematik, die philosophischen Fragen aus dem Weg geht, könnte man Kausalität einfach über ihre vorteilhaften Wirkungen charakterisieren. (Das ist eine typische Herangehensweise: Es wird nicht erklärt, was die Objekte sind, sondern es wird gesagt, wie mit ihnen umzugehen ist.)

In diesem Sinn lässt sich ein kausaler Bayesscher Graph einfach als ein solcher definieren, der minimal ist (vgl. Punkt 2 obiger Aufzählung), und bei dem sich Änderungen nur lokal auswirken (vgl. Punkt 3).

Anstatt zu monieren, dass wir uns gerade im Kreis gedreht zu haben scheinen, merken wir an, dass hier ein Weg angelegt ist, mit dem Kausalität (im Sinne der Definition) als *Minimal- und Invarianzprinzip* nachweisbar ist.

Das Thema diskutieren wir jetzt jedoch nicht ausführlicher. Stattdessen sei auf die Literatur verwiesen.

VI Literatur

Zu empfehlen ist insbesondere

„Probabilistic Reasoning in Intelligent Systems“ von Judea Pearl, veröffentlicht 1988 durch Morgan Kaufmann,

sowie in Hinblick auf den fünften Abschnitt zur Kausalität

„Causality“ von Judea Pearl, veröffentlicht 2000 von der Cambridge University Press.