<table>
<thead>
<tr>
<th>#</th>
<th>date</th>
<th>content</th>
<th>Ex</th>
<th>#</th>
<th>date</th>
<th>content</th>
<th>Ex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.04.</td>
<td>Introduction</td>
<td>1</td>
<td>14</td>
<td>09.06.</td>
<td>Logistic Regression</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>21.04.</td>
<td>Reasoning under Uncertainty</td>
<td>2</td>
<td>15</td>
<td>15.06.</td>
<td>Exponential Families</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>27.04.</td>
<td>Continuous Variables</td>
<td>3</td>
<td>16</td>
<td>16.06.</td>
<td>Graphical Models</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>28.04.</td>
<td>Monte Carlo</td>
<td>4</td>
<td>17</td>
<td>22.06.</td>
<td>Factor Graphs</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>04.05.</td>
<td>Markov Chain Monte Carlo</td>
<td>5</td>
<td>18</td>
<td>23.06.</td>
<td>The Sum-Product Algorithm</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>05.05.</td>
<td>Gaussian Distributions</td>
<td>6</td>
<td>19</td>
<td>29.06.</td>
<td>Example: Topic Models</td>
<td>13</td>
</tr>
<tr>
<td>7</td>
<td>11.05.</td>
<td>Parametric Regression</td>
<td>7</td>
<td>20</td>
<td>30.06.</td>
<td>Mixture Models</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>12.05.</td>
<td>Understanding Deep Learning</td>
<td>8</td>
<td>21</td>
<td>06.07.</td>
<td>EM</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>18.05.</td>
<td>Gaussian Processes</td>
<td>9</td>
<td>22</td>
<td>07.07.</td>
<td>Variational Inference</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>19.05.</td>
<td>An Example for GP Regression</td>
<td>10</td>
<td>23</td>
<td>13.07.</td>
<td>Example: Topic Models</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>25.05.</td>
<td>Understanding Kernels</td>
<td>11</td>
<td>24</td>
<td>14.07.</td>
<td>Example: Inferring Topics</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>08.06.</td>
<td>GP Classification</td>
<td>13</td>
<td>26</td>
<td>21.07.</td>
<td>Revision</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Life's most important problems are, for the most part, problems of probability.

Pierre-Simon, marquis de Laplace (1749-1827)
Catch-up from Last Time:

- Probabilities are the mathematical formalization of uncertainty
- Two basic rules
  - Product rule: \( P(A, B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A) \)
  - Sum rule: \( P(A) = P(A, B) + P(A, \neg B) \)
- Corollary: Bayes’ Theorem
  \[
  \frac{P(X \mid D)}{P(D)} = \frac{P(X) \cdot P(D \mid X)}{\sum_{x \in X} P(D \mid x)P(x)}
  \]
- This extends deductive reasoning to plausible reasoning

Today:

- Building an intuition for probability
- The computational complexity of probabilistic inference
Plausible reasoning extends Boolean Logic

Catch-Up from last time

Lemma (from Bayes’ theorem)

\[ A \implies B : P(B \mid A) = 1 \implies \text{if } A \text{ is true, then } B \text{ is true} \]

- \[ P(B \mid A) = 1 \text{ “modus ponens”} \]
- \[ P(B \mid \neg A) \leq P(B) \]
- \[ P(A \mid B) \geq P(A) \]
- \[ P(\neg A \mid \neg B) = 1 \text{ “modus tollens”} \]

Lemma (from Bayes’ theorem)

\[ P(B \mid A) \geq P(B) \implies \text{if } A \text{ is true, then } B \text{ becomes more plausible} \]

- \[ P(B \mid A) \geq P(B) \]
- \[ P(B \mid \neg A) \leq P(B) \]
- \[ P(A \mid B) \geq P(A) \]
- \[ P(\neg A \mid \neg B) \geq P(\neg A) \]
Boole was a Bayesian

**Principle 1st.** If \( p \) be the probability of the occurrence of any event, \( 1 - p \) will be the probability of its non-occurrence.

2nd. The probability of the concurrence of two independent events is the product of the probabilities of those events.

3rd. The probability of the concurrence of two dependent events is equal to the product of the probability of one of them by the probability that if that event occur, the other will happen also.

4th. The probability that if an event, \( E \), take place, an event, \( F \), will also take place, is equal to the probability of the concurrence of the events \( E \) and \( F \), divided by the probability of the occurrence of \( E \).
Computational Difficulties of Probability Theory

Uncertainty is a global notion

The joint distribution of \( n = 26 \) propositional variables \( A, B, \ldots, Z \) has \( 2^n \) free parameters

\[
\begin{align*}
[1] & \quad P(A, B, \ldots, Z) = \ldots \\
[2] & \quad P(\neg A, B, \ldots, Z) = \ldots \\
[3] & \quad P(A, \neg B, \ldots, Z) = \ldots \\
& \quad \vdots \\
[67\,108\,863] & \quad P(\neg A, \neg B, \ldots, Z) = \ldots \\
[67\,108\,864] & \quad P(\neg A, \neg B, \ldots, \neg Z) = 1 - \sum P(\ldots)
\end{align*}
\]

- requires not just large memory, but computing marginals like \( P(A) \) is also very expensive
- nb: just committing to a single guess is much (exponentially in \( n \)) cheaper
- can we specify the joint distribution with fewer numbers?
Computing with Probabilities

- Probabilistic reasoning extends propositional logic
- instead of tracking a single *true* value, we have to assign probabilities to *combinatorially many* hypotheses

Being uncertain is potentially *much* more expensive in terms of computation and memory than simply committing to a single hypothesis. This is *the* challenge of probabilistic reasoning in practice.
A note on notation

somewhat unfortunate, but very helpful in the remainder

So far, $A$ was a propositional variable that forms formulae:

$$P(A) = \text{probability that formula } A \text{ is true}$$
$$P(\neg A) = 1 - P(A) = \text{probability that formula } \neg A \text{ is true}$$

From now on $A$ is a propositional variable with values in $\{0, 1\}$, i.e. $P(A)$ is a function of two possible input values $A = 1$ and $A = 0$, i.e. with slightly unusual notation:

$$P(A = 1) = \text{probability that formula } A \text{ is true}$$
$$P(A = 0) = 1 - P(A = 1) = \text{probability that formula } A \text{ is false}$$

Stating that $P(A, B) = P(A) \cdot P(B)$ means all of the following

$$P(A = 1, B = 1) = P(A = 1) \cdot P(B = 1) \quad P(A = 1, B = 0) = P(A = 1) \cdot P(B = 0)$$
$$P(A = 0, B = 1) = P(A = 0) \cdot P(B = 1) \quad P(A = 0, B = 0) = P(A = 0) \cdot P(B = 0)$$
Independence
Chiefly a computational concept

Definition (independence)

Two variables $A$ and $B$ are independent, if and only if their joint distributions factorizes into so-called marginal distributions, i.e.

\[ P(A, B) = P(A) P(B) \]

In that case $P(A|B) = P(A)$. Notation: $A \perp \perp B$. Information about $B$ does not give information about $A$ and vice versa.
Independence
Chiefly a computational concept

Definition (independence)

Two variables $A$ and $B$ are independent, if and only if their joint distributions factorizes into so-called marginal distributions, i.e.

$$P(A, B) = P(A) \cdot P(B)$$

In that case $P(A|B) = P(A)$. Notation: $A \perp \perp B$. Information about $B$ does not give information about $A$ and vice versa.

Example: Two coins.

$A = \text{coin 1 shows heads}$
$B = \text{coin 2 shows heads}$

Then $A \perp \perp B$. 
### Definition (conditional independence)

Two variables $A$ and $B$ are **conditionally independent** given variable $C$, if and only if their conditional distribution factorizes,

$$P(A, B|C) = P(A|C) \cdot P(B|C)$$

In that case we have $P(A|B, C) = P(A|C)$, i.e. in light of information $C$, $B$ provides no (further) information about $A$. Notation: $A \perp \perp B \mid C$
Conditional Independence
Chiefly a computational concept

Definition (conditional independence)

Two variables \(A\) and \(B\) are **conditionally independent** given variable \(C\), if and only if their conditional distribution factorizes,

\[
P(A, B|C) = P(A|C) \cdot P(B|C)
\]

In that case we have \(P(A|B, C) = P(A|C)\), i.e. in light of information \(C\), \(B\) provides no (further) information about \(A\). Notation: \(A \perp \perp B \mid C\)

Example: Two coins and a bell.

- \(A =\) coin 1 shows heads
- \(B =\) coin 2 shows heads
- \(C =\) bell rings if both coins show the same result

\(A \perp \perp B\)
Definition (conditional independence)

Two variables $A$ and $B$ are **conditionally independent** given variable $C$, if and only if their conditional distribution factorizes,

$$P(A, B|C) = P(A|C) P(B|C)$$

In that case we have $P(A|B, C) = P(A|C)$, i.e. in light of information $C$, $B$ provides no (further) information about $A$. Notation: $A \perp \perp B \mid C$

Example: Two coins and a bell.

- $A =$ coin 1 shows heads
- $B =$ coin 2 shows heads
- $C =$ bell rings if both coins show the same result

$A \perp \perp B$ and $A \perp \perp C$
Definition (conditional independence)

Two variables $A$ and $B$ are **conditionally independent** given variable $C$, if and only if their conditional distribution factorizes,

$$P(A, B|C) = P(A|C) P(B|C)$$

In that case we have $P(A|B, C) = P(A|C)$, i.e. in light of information $C$, $B$ provides no (further) information about $A$. Notation: $A \perp \perp B \mid C$

Example: Two coins and a bell.

- $A = \text{coin 1 shows heads}$
- $B = \text{coin 2 shows heads}$
- $C = \text{bell rings if both coins show the same result}$

$A \perp \perp B$ and $A \perp \perp C$ and $B \perp \perp C$,
Definition (conditional independence)

Two variables $A$ and $B$ are conditionally independent given variable $C$, if and only if their conditional distribution factorizes,

$$P(A, B|C) = P(A|C) P(B|C)$$

In that case we have $P(A|B, C) = P(A|C)$, i.e. in light of information $C$, $B$ provides no (further) information about $A$. Notation: $A \perp \perp B \mid C$

Example: Two coins and a bell.

$A =$ coin 1 shows heads
$B =$ coin 2 shows heads
$C =$ bell rings if both coins show the same result

$A \perp \perp B$ and $A \perp \perp C$ and $B \perp \perp C$, but $A \not\perp \perp B \mid C$
Conditional Independence

Chiefly a computational concept

Example: Stefan Harmeling

Definition (conditional independence)

Two variables $A$ and $B$ are **conditionally independent** given variable $C$, if and only if their conditional distribution factorizes,

$$P(A, B|C) = P(A|C) P(B|C)$$

In that case we have $P(A|B, C) = P(A|C)$, i.e. *in light of information $C$, $B$ provides no (further) information about $A*.* Notation: $A \perp \perp B \mid C$

Example: Two coins and a bell.

$A =$ coin 1 shows heads
$B =$ coin 2 shows heads
$C =$ bell rings if both coins show the same result

$A \perp \perp B$ and $A \perp \perp C$ and $B \perp \perp C$, but $A \not\perp \perp B \mid C$ and $A \not\perp \perp C \mid B$ and $B \not\perp \perp C \mid A$. 
Computing with Probabilities

- Probabilistic reasoning extends propositional logic
- Instead of tracking a single true value, we have to assign probabilities to combinatorially many hypotheses
- Two variables $A$ and $B$ are conditionally independent given variable $C$, if and only if their conditional distribution factorizes,

$$P(A, B|C) = P(A|C) \cdot P(B|C)$$
Parameter Counting
a simple example [adapted from Pearl, 1988 / MacKay, 2003 §21]

\[ A = \text{the alarm was triggered} \]
\[ E = \text{there was an earthquake} \]
\[ B = \text{there was a break-in} \]
\[ R = \text{an announcement is made on the radio} \]

Joint probability distribution has \( 2^4 - 1 = 15 = 8 + 4 + 2 + 1 \) parameters:

\[
P(A, E, B, R) = P(A \mid R, E, B) \cdot P(R \mid E, B) \cdot P(E \mid B) \cdot P(B).
\]

Removing irrelevant conditions (domain knowledge!) reduces to \( 8 = 4 + 2 + 1 + 1 \) parameters:

\[
P(A, E, B, R) = P(A \mid E, B) \cdot P(R \mid E) \cdot P(E) \cdot P(B).
\]
A Graphical Representation

Our first Bayesian network.

\[ P(A, E, B, R) = P(A \mid E, B) \cdot P(R \mid E) \cdot P(E) \cdot P(B) \]

- **A** = the alarm was triggered
- **E** = there was an earthquake
- **B** = there was a break-in
- **R** = an announcement is made on the radio
Conditional Probability Tables

For the burglar/alarm problem

\[ P(A, E, B, R) = P(A \mid E, B) \cdot P(R \mid E) \cdot P(E) \cdot P(B) \]

\[ P(B = 1) = 10^{-3} \quad P(B = 0) = 1 - 10^{-3} \]
\[ P(E = 1) = 10^{-3} \quad P(E = 0) = 1 - 10^{-3} \]

and

\[ P(R = 1 \mid E = 1) = 1.0 \quad P(R = 1 \mid E = 0) = 0.0 \]

and, using \( f = 10^{-3} \), \( \alpha_b = 0.99 \), \( \alpha_e = 0.01 \)

\[ P(A = 0 \mid B = 0, E = 0) = (1 - f) \]
\[ P(A = 0 \mid B = 1, E = 0) = (1 - f)(1 - \alpha_b) \]
\[ P(A = 0 \mid B = 0, E = 1) = (1 - f)(1 - \alpha_e) \]
\[ P(A = 0 \mid B = 1, E = 1) = (1 - f)(1 - \alpha_b)(1 - \alpha_e) \]
\[ P(A = 1 \mid B = 0, E = 0) = f \]
\[ P(A = 1 \mid B = 1, E = 0) = 1 - (1 - f)(1 - \alpha_b) \]
\[ P(A = 1 \mid B = 0, E = 1) = 1 - (1 - f)(1 - \alpha_e) \]
\[ P(A = 1 \mid B = 1, E = 1) = 1 - (1 - f)(1 - \alpha_b)(1 - \alpha_e) \]
Conditional Probability Tables
For the burglar/alarm problem

\[ P(A, E, B, R) = P(A \mid E, B) \cdot P(R \mid E) \cdot P(E) \cdot P(B) \]

\[ P(B = 1) = 10^{-3} \quad P(B = 0) = 1 - 10^{-3} \]
\[ P(E = 1) = 10^{-3} \quad P(E = 0) = 1 - 10^{-3} \]

and

\[ P(R = 1 \mid E = 1) = 1.0 \quad P(R = 1 \mid E = 0) = 0.0 \]

and, using \( f = 10^{-3} \), \( \alpha_b = 0.99 \), \( \alpha_e = 0.01 \)

\[ P(A = 0 \mid B = 0, E = 0) = 0.999 \]
\[ P(A = 0 \mid B = 1, E = 0) = 0.00999 \]
\[ P(A = 0 \mid B = 0, E = 1) = 0.98901 \]
\[ P(A = 0 \mid B = 1, E = 1) = 0.0098901 \]
\[ P(A = 1 \mid B = 0, E = 0) = 0.001 \]
\[ P(A = 1 \mid B = 1, E = 0) = 0.99001 \]
\[ P(A = 1 \mid B = 0, E = 1) = 0.01099 \]
\[ P(A = 1 \mid B = 1, E = 1) = 0.9901099 \]
Inference

Conditional dependence

What is the probability that there was a break-in and/or an earthquake, given that the alarm went off?

\[
P(B, E \mid A = 1) = \frac{P(A = 1 \mid B, E)P(B)P(E)}{P(A = 1)}
\]

\[
P(A = 1) = P(A = 1 \mid B = 0, E = 0)P(B = 0)P(E = 0)
+ P(A = 1 \mid B = 0, E = 1)P(B = 0)P(E = 1)
+ P(A = 1 \mid B = 1, E = 0)P(B = 1)P(E = 0)
+ P(A = 1 \mid B = 1, E = 1)P(B = 1)P(E = 1)
\]

\[
= 0.000998 + 0.000989 + 0.000010979 + 0.00000099 = 0.002
\]

thus — note conditional dependence!

\[
P(B = 0, E = 0 \mid A = 1) = 0.4993 \quad P(B = 1, E = 0 \mid A = 1) = 0.4947
\]

\[
P(B = 0, E = 1 \mid A = 1) = 0.0055 \quad P(B = 1, E = 1 \mid A = 1) = 0.0005
\]

\[
P(B = 0 \mid A = 1) = P(B = 0, E = 0 \mid A = 1) + P(B = 0, E = 1 \mid A = 1) = 0.505
\]

\[
P(B = 1 \mid A = 1) = P(B = 1, E = 0 \mid A = 1) + P(B = 1, E = 1 \mid A = 1) = 0.495
\]
What is the probability for a break-in, given alarm \textit{and} radio announcement?

\[
P(B = 0 \mid E = 1, A = 1) = \frac{P(B = 0, E = 1 \mid A = 1)}{P(E = 1 \mid A = 1)}
\]

\[
= \frac{P(B = 0, E = 1 \mid A = 1)}{P(B = 0, E = 1 \mid A = 1) + P(B = 1, E = 1 \mid A = 1)} = 0.92
\]

\[
P(B = 1 \mid E = 1, A = 1) = \frac{P(B = 1, E = 1 \mid A = 1)}{P(E = 1 \mid A = 1)}
\]

\[
= \frac{P(B = 1, E = 1 \mid A = 1)}{P(B = 0, E = 1 \mid A = 1) + P(B = 1, E = 1 \mid A = 1)} = 0.08
\]

The radio announcement is \textbf{explaining away} the break-in as the explanation for the alarm.
What is Probabilistic Reasoning?
One recipe for all your inference needs!

**Always write down the probability of everything.**

- identify all relevant variables: $A, R, E, B$
- define **joint probability** $P(A, R, E, B)$ aka. **the generative model**
- **observations** fix certain variables: $A = 1$
- **inference** takes place exclusively by Bayes’ Theorem
  n.b.: this requires integrating out (**marginalizing**) latent variables not being inferred.
Definition (Bayesian Network, preliminary definition — more in later lectures)

A Directed Graphical Model (DGM), aka. Bayesian Network is a probability distribution over variables \( \{X_1, \ldots, X_D\} \) that can be written as

\[
P(X_1, X_2, \ldots, X_D) = \prod_{i=1}^{D} p(X_i \mid \text{pa}(X_i))
\]

where \( \text{pa}(X_i) \) are the parental variables of \( X_i \), that is, \( X_i \notin \text{pa}(X_j) \ \forall \ X_j \in \text{pa}(X_i) \). A DGM can be represented by a Directed Acyclic Graph (DAG) with the propositional variables as nodes, and arrows from parents to children.
Every Probability Distribution is a DAG
It’s just not always a helpful concept

By the Product Rule, every joint can be factorized into a (dense) DAG.

\[
P(A, E, B, R) = P(A \mid E, B, R) \cdot P(R \mid E, B) \cdot P(E \mid B) \cdot P(B)
\]

- \(A\) = the alarm was triggered
- \(E\) = there was an earthquake
- \(B\) = there was a break-in
- \(R\) = an announcement is made on the radio
Every Probability Distribution is a DAG

It’s just not always a helpful concept

The direction of the arrows is **not** a causal statement.

\[ P(A, E, B, R) = P(B \mid A, E, R) \cdot P(E \mid A, R) \cdot P(R \mid A) \cdot P(A) \]

- **A** = the alarm was triggered
- **E** = there was an earthquake
- **B** = there was a break-in
- **R** = an announcement is made on the radio
Every Probability Distribution is a DAG

It’s just not always a helpful concept

But the representation is particularly interesting when it reveals independence.

\[ P(A, E, B, R) = P(A \mid E, B) \cdot P(R \mid E) \cdot P(E) \cdot P(B) \]

- \( A \) = the alarm was triggered
- \( E \) = there was an earthquake
- \( B \) = there was a break-in
- \( R \) = an announcement is made on the radio
Deducing Conditional Independencies

back to our example [adapted from Pearl, 1988 / MacKay, 2003 §21]

\[
P(A, E, B, R) = P(A \mid E, B) \cdot P(R \mid E) \cdot P(E) \cdot P(B)
\]

A = the alarm was triggered
E = there was an earthquake
B = there was a break-in
R = an announcement is made on the radio

Which independencies can we infer only from the graph?
Atomic Independence Structures

DAGs imply conditional independence, but not dependence!

For uni- and bi-variate graphs, conditional independence is trivial.
For tri-variate sub-graphs, there are three possible structures:

<table>
<thead>
<tr>
<th>graph</th>
<th>factorization</th>
<th>implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>$A \perp \perp C \mid B$</td>
<td>$A \perp \perp C \mid B$ but not, i.g., $A \not\perp \not\perp C$</td>
</tr>
<tr>
<td>(ii)</td>
<td>$A \perp \perp C \mid B$</td>
<td>$A \perp \perp C \mid B$ but not, i.g., $A \not\perp \not\perp C$</td>
</tr>
<tr>
<td>(iii)</td>
<td>$A \perp \perp C$</td>
<td>$A \perp \perp C$ but not, i.g., $A \not\perp \not\perp C \mid B$</td>
</tr>
</tbody>
</table>
Deducing Conditional Independencies

back to our example

<table>
<thead>
<tr>
<th>graph</th>
<th>factorization</th>
<th>implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>[P(A, B, C) = P(C</td>
<td>B) \cdot P(B</td>
</tr>
<tr>
<td>(ii)</td>
<td>[P(A, B, C) = P(A</td>
<td>B) \cdot P(C</td>
</tr>
<tr>
<td>(iii)</td>
<td>[P(A, B, C) = P(B</td>
<td>A, C) \cdot P(C) \cdot P(A)]</td>
</tr>
</tbody>
</table>

Which independencies can we infer only from the graph?
### Deducing Conditional Independencies

#### back to our example

<table>
<thead>
<tr>
<th>graph</th>
<th>factorization</th>
<th>implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>[ P(A, B, C) = P(C</td>
<td>B) \cdot P(B</td>
</tr>
<tr>
<td>(ii)</td>
<td>[ P(B) ]</td>
<td>( A \indep C</td>
</tr>
<tr>
<td>(iii)</td>
<td>[ P(A, B, C) = P(B</td>
<td>A, C) \cdot P(C) \cdot P(A) ]</td>
</tr>
</tbody>
</table>

Which independencies can we infer only from the graph?

- \( R \indep A | E \) and \( E \indep B \)
Deducing Conditional Independencies

back to our example

<table>
<thead>
<tr>
<th>graph</th>
<th>factorization</th>
<th>implications</th>
</tr>
</thead>
</table>
| (i) $A \rightarrow B \rightarrow C$ | $P(A, B, C) = P(C | B) \cdot P(B | A) \cdot P(A)$ | $A \perp \perp C | B$  
but not, i.g., $A \n \perp \perp C$ |
| (ii) $A \leftarrow B \rightarrow C$ | $P(A, B, C) = P(A | B) \cdot P(C | B) \cdot P(B)$ | $A \perp \perp C | B$  
but not, i.g., $A \n \perp \perp C$ |
| (iii) $A \rightarrow B \leftarrow C$ | $P(A, B, C) = P(B | A, C) \cdot P(C) \cdot P(A)$ | $A \perp \perp C$  
but not, i.g., $A \n \perp \perp C | B$ |

Which independencies can we infer only from the graph?

- $R \perp \perp A | E$ and $E \perp \perp B$
- but also $(R \perp \perp B | E)$, $(R \perp \perp B)$, $(R \perp \perp B | E, A)$, with more work
The Graph for Two Coins and a Bell

DAGs are not a perfect tool

\[
\begin{align*}
P(A = 1) &= 0.5 & P(C = 1 \mid A = 1, B = 1) &= 1 & P(C = 1 \mid A = 1, B = 0) &= 0 \\
P(B = 1) &= 0.5 & P(C = 1 \mid A = 0, B = 1) &= 0 & P(C = 1 \mid A = 0, B = 0) &= 1
\end{align*}
\]

These CPTs imply \(P(A \mid B) = P(A), P(B \mid C) = P(B)\) and \(P(C \mid A) = P(C)\) and \(P(C \mid B) = P(C)\).
The Graph for Two Coins and a Bell

DAGs are not a perfect tool

\[
\begin{align*}
P(A = 1) &= 0.5 & P(C = 1 | A = 1, B = 1) &= 1 & P(C = 1 | A = 1, B = 0) &= 0 \\
P(B = 1) &= 0.5 & P(C = 1 | A = 0, B = 1) &= 0 & P(C = 1 | A = 0, B = 0) &= 1
\end{align*}
\]

These CPTs imply \( P(A|B) = P(A) \), \( P(B|C) = P(B) \) and \( P(C|A) = P(C) \) and \( P(C | B) = P(C) \).

We thus have three factorizations:

1. \( P(A, B, C) = P(C|A, B) \cdot P(A|B) \cdot P(B) = P(C|A, B) \cdot P(A) \cdot P(B) \)
2. \( P(A, B, C) = P(A|B, C) \cdot P(B|C) \cdot P(C) = P(A|B, C) \cdot P(B) \cdot P(C) \)
3. \( P(A, B, C) = P(B|C, A) \cdot P(C|A) \cdot P(A) = P(B|C, A) \cdot P(C) \cdot P(A) \)
The Graph for Two Coins and a Bell

DAGs are not a perfect tool

\[
P(A = 1) = 0.5 \quad P(C = 1 \mid A = 1, B = 1) = 1 \quad P(C = 1 \mid A = 1, B = 0) = 0
\]
\[
P(B = 1) = 0.5 \quad P(C = 1 \mid A = 0, B = 1) = 0 \quad P(C = 1 \mid A = 0, B = 0) = 1
\]

These CPTs imply \( P(A \mid B) = P(A) \), \( P(B \mid C) = P(B) \), and \( P(C \mid A) = P(C) \) and \( P(C \mid B) = P(C) \).

We thus have three factorizations:

1. \( P(A, B, C) = P(C \mid A, B) \cdot P(A \mid B) \cdot P(B) = P(C \mid A, B) \cdot P(A) \cdot P(B) \)
2. \( P(A, B, C) = P(A \mid B, C) \cdot P(B \mid C) \cdot P(C) = P(A \mid B, C) \cdot P(B) \cdot P(C) \)
3. \( P(A, B, C) = P(B \mid C, A) \cdot P(C \mid A) \cdot P(A) = P(B \mid C, A) \cdot P(C) \cdot P(A) \)

Each corresponds to a graph. Note that each can only express some of the independencies:
Graphical Models and Conditional Independence

- Multivariate distributions can have \textit{exponentially} many degrees of freedom.
- \textbf{(conditional) independence} helps reduce this complexity to make things tractable.
- \textbf{(directed) graphical models} provide a notation from which conditional independence can be read off using simple rules.
- Every probability distribution is a DAG, but not every independence structure of a distribution is captured by a DAG of it.
- We will return to graphs later in the course.

\textbf{Conditional independence is a tool (and may be required) to keep inference tractable in multi-variate problems.}
§ 5. Unabhängigkeit.

Der Begriff der gegenseitigen Unabhängigkeit zweier oder mehrerer Versuche nimmt eine in gewissem Sinne zentrale Stellung in der Wahrscheinlichkeitsrechnung ein. In der Tat haben wir schon gesehen, daß die Wahrscheinlichkeitsrechnung vom mathematischen Standpunkte aus als eine spezielle Anwendung der allgemeinen Theorie der additiven Mengenfunktionen betrachtet werden kann. Man kann sich natürlich fragen, wie ist es dann möglich, daß die Wahrscheinlichkeitsrechnung sich in eine große, ihre eigenen Methoden besitzende selbständige Wissenschaft entwickelt hat?

Man kommt also dazu, im Begriffe der Unabhängigkeit wenigstens den ersten Keim der eigenartigen Problematik der Wahrscheinlichkeitsrechnung zu erblicken. [...] Es ist dementsprechend eine der wichtigsten Aufgaben der Philosophie der Naturwissenschaften, nachdem sie die vielumstrittene Frage über das Wesen des Wahrscheinlichkeitsbegriffes selbst erklärt hat, die Voraussetzungen zu präzisieren. bei denen man irgendwelche gegebene reelle Erscheinungen für gegenseitig unabhängig halten kann.

A. N. Kolmogorov. Grundbegriffe der Wahrscheinlichkeitsrechnung. §1.5