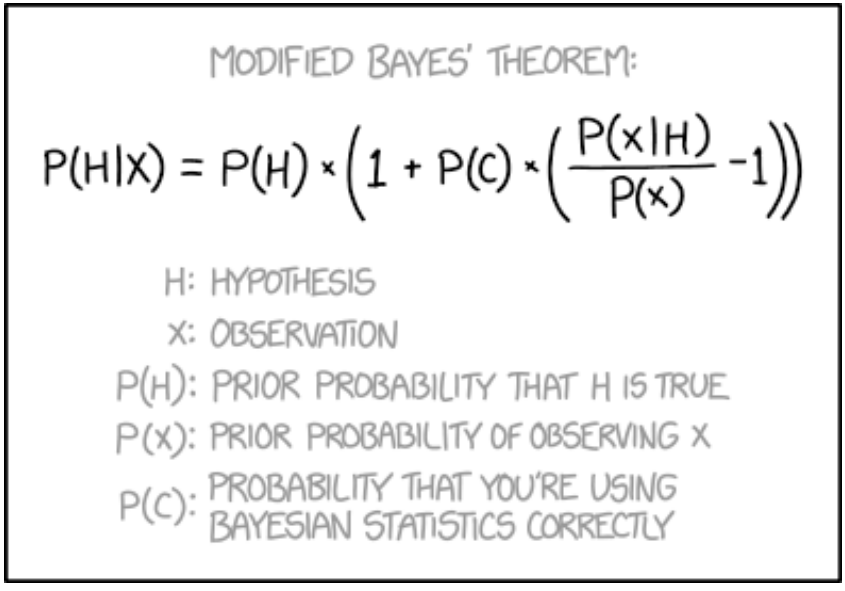


https://xkcd.com/2059/

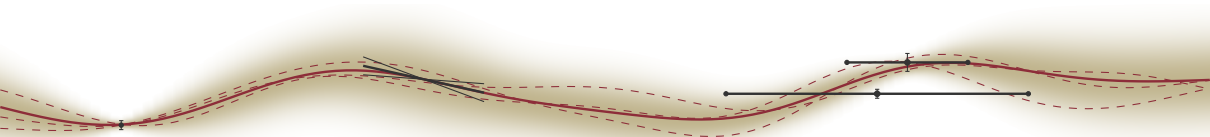
Probabilistic Inference and Learning
Lecture 01
Probabilistic Reasoning

Philipp Hennig

17 October 2018

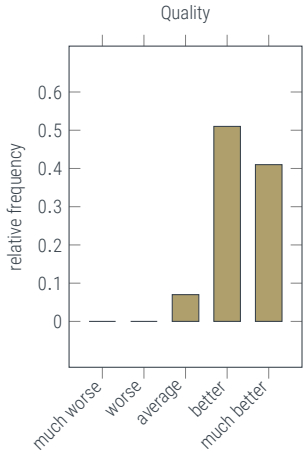
Eberhard Karls
Universität
Tübingen

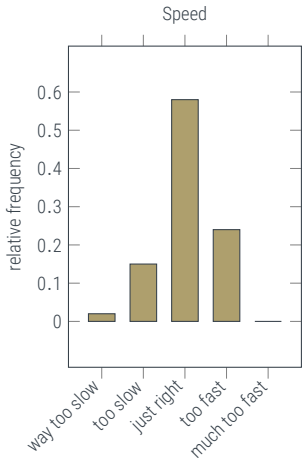
Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning

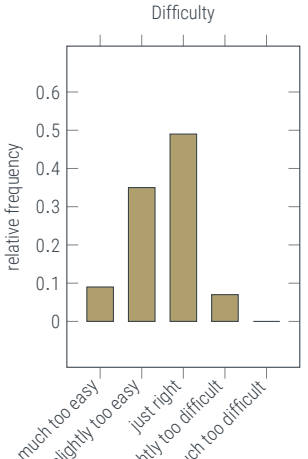


Last Lecture: Debrief

Feedback dashboard







Last Lecture: Debrief

Detailed Feedback

Things you did not like:

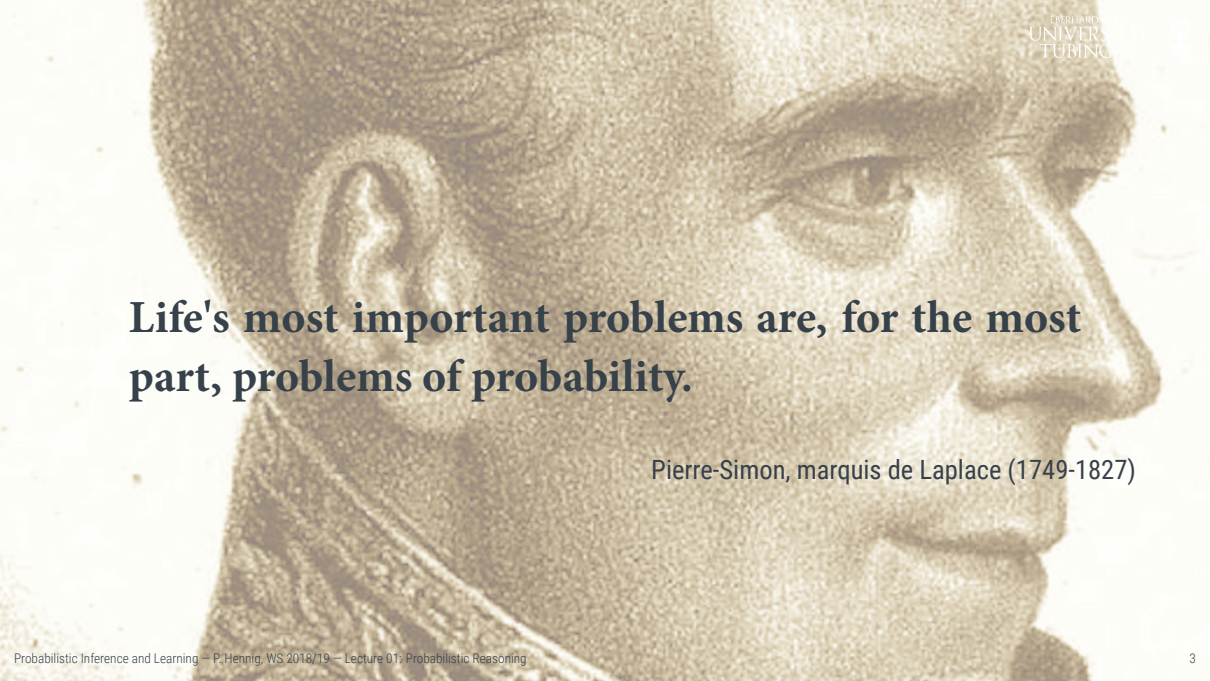
- **"Too easy"**
- "Deriving simple formulas"
- "The rush at the end"
- "Not enough Philosophy"
- "The break"

Things you did not understand:

- **"The example with the bird" (!!!)**
- "The answer to card & color. $^2/_3$?"
- "Kolmogorov's approach'

Things you enjoyed:

- **"The card thing at the beginning" (!!!)**
- **"Derivations on the Board"**
- "Comparing Cox and Kolmogorov", "The connection to Logic and Philosophy"
- "Sketching goals/possibilities"
- "Edgar Allan Poe"
- "Enthusiasm and Motivation", "Interactivity"



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: $p(A, B) = p(A \mid B)p(B) = p(B \mid A)p(A)$

  sum rule: $p(A) + p(\neg A) = 1$
- Corollary: **Bayes' Theorem**

$$\underbrace{p(X \mid D)}_{\text{posterior for } X \text{ given } D} = \frac{\overbrace{p(X)}^{\text{prior for } X} \cdot \overbrace{p(D \mid X)}^{\text{likelihood for } X}}{\underbrace{p(D)}_{\text{evidence for the model}}} = \frac{p(X) \cdot p(D \mid X)}{\sum_{x \in \mathcal{X}} p(D \mid x)p(x)}$$

- This extends deductive reasoning to plausible reasoning

Today:

- Building an intuition for probability
- The computational complexity of probabilistic inference

Lemma (from Bayes' theorem)

$A \Rightarrow B$: $p(B \mid A) = 1$ *implies*	*if A is true, then B is true*
✦ $p(B \mid A) = 1$ *"modus ponens"*	*A is true implies B is true*
✦ $p(B \mid \neg A) \leq p(B)$	*A is false implies B becomes less plausible*
✦ $p(A \mid B) \geq p(A)$	*B is true implies A becomes more plausible*
✦ $p(\neg A \mid \neg B) = 1$ *"modus tollens"*	*B is false implies A is false*

Lemma (from Bayes' theorem)

$p(B \mid A) \geq B$ *implies*	*if A is true, then B becomes more plausible*
✦ $p(B \mid A) \geq p(B)$	*A is true implies B becomes more plausible*
✦ $p(B \mid \neg A) \leq p(B)$	*A is false implies B becomes less plausible*
✦ $p(A \mid B) \geq p(A)$	*B is true implies A becomes more plausible*
✦ $p(\neg A \mid \neg B) \geq p(\neg A)$	*B is false implies A becomes less plausible*

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

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

- 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?

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:

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

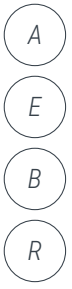
Stating that $p(A, B) = p(A) \cdot p(B)$ means **all** of the following

$$p(A = 1, B = 1) = p(A = 1) \cdot p(B = 1) \qquad p(A = 1, B = 1) = p(A = 1) \cdot p(B = 1)$$
$$p(A = 0, B = 1) = p(A = 0) \cdot p(B = 1) \qquad p(A = 0, B = 0) = p(A = 0) \cdot p(B = 0)$$

Parameter Counting

a simple example

[adapted from Pearl, 1988 / MacKay, 2003 §21]



$A =$ the alarm was triggered

$E =$ there was an earthquake

$B =$ there was a break-in

$R =$ 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:

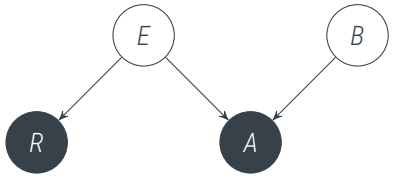
$$p(A, E, B, R) = p(A \mid E, B) \cdot p(R \mid E) \cdot p(E) \cdot p(B)$$

A Graphical Representation

Our first Bayesian network.

[adapted from Pearl, 1988 / MacKay, 2003 §21]

$$p(A, E, B, R) = p(A \mid E, B) \cdot p(R \mid E) \cdot p(E) \cdot p(B)$$



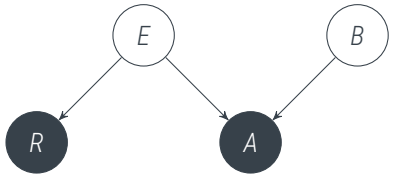
$A =$ the alarm was triggered
$E =$ there was an earthquake
$B =$ there was a break-in
$R =$ an announcement is made on the radio

A Graphical Representation

Our first Bayesian network.

[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

Conditional probability tables:

$p(E = 1) = 10^{-3}$

$p(R = 1 \mid E = 1) = 1.0$	$p(R = 1 \mid E = 0) = 0.0$

$p(B = 1) = 10^{-3}$	$p(A = 1 \mid E = 1, B = 0) = 0.01099$	$p(A = 1 \mid E = 0, B = 0) = 10^{-3}$
	$p(A = 1 \mid E = 1, B = 1) = 0.9901099$	$p(A = 1 \mid E = 0, B = 1) = 0.99001$

- What is the probability that there was a break-in, given that the alarm went off?

$$p(B=1|A=1) = \frac{p(B=1,A=1)}{p(A=1)} = \frac{\sum_{R,E} p(A=1,R,E,B=1)}{\sum_{R,E,B} p(A=1,R,E,B)} = \ldots = 0.495$$

- What is the probability for a break-in, given alarm *and* radio announcement?

$$p(B=1|A=1,R=1) = \frac{p(B=1,A=1,R=1)}{p(A=1,R=1)} = \frac{\sum_{E} p(A=1,R=1,E,B=1)}{\sum_{E,B} p(A=1,R=1,E,B)}$$
$$= \ldots = 0.08$$

The radio announcement is **explaining away** the break-in as the explanation for the alarm.

- Note: $B \perp\!\!\!\perp R$ but $B \not\perp\!\!\!\perp R \mid A$.

What is Probabilistic Reasoning?

One recipe for all your inference needs!

Always write down the probability of everything.

David JC MacKay (1967–2016)

- 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.

Directed Graphical Models

aka. Bayesian networks, Bayes nets, belief networks, …

[Judea Pearl, *Probabilistic Reasoning in Intelligent Systems*, 1988]

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

$$p(X_1, X_2, \ldots, X_D) = \prod_{i=1}^{D} p\left(X_i \mid \text{pa}(X_i)\right)$$

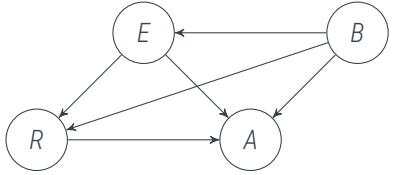
where $\text{pa}(X_i)$ are the parental variables of X_i, that is, $X_i \notin \text{pa}(X_j)\ \forall\ X_j \in \text{pa}(X_i)$. A DGM can be represented by a **Directed Acyclic Graph (DAG)** with the propositional variables as nodes, and arrows from parents to children.

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.

$$p(A, E, B, R) = p(A \mid E, B, R) \cdot p(R \mid E, B) \cdot p(E \mid B) \cdot p(B)$$



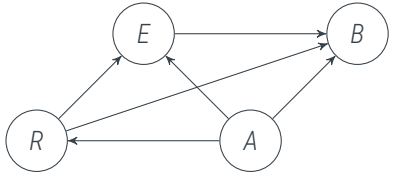
$A =$ the alarm was triggered
$E =$ there was an earthquake
$B =$ there was a break-in
$R =$ an announcement is made on the radio

Every Probability Distribution is a DAG

It's just not always a helpful concept

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

$$p(A, E, B, R) = p(B \mid A, E, R) \cdot p(E \mid A, R) \cdot p(R \mid A) \cdot p(A)$$



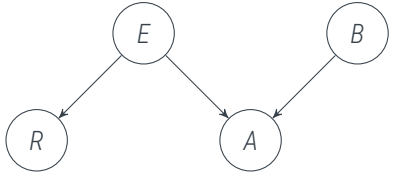
$A =$ the alarm was triggered
$E =$ there was an earthquake
$B =$ there was a break-in
$R =$ an announcement is made on the radio

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

Note that $p(E|B) = p(E)$ implies $p(E, B) = p(E)\ p(B)$, because $p(E, B) = p(E \mid B)p(B)$.

Note that $p(E|B) = p(E)$ implies $p(E, B) = p(E)\ p(B)$, because $p(E, B) = p(E \mid B)p(B)$.

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.*

Note that $p(E|B) = p(E)$ implies $p(E, B) = p(E)\ p(B)$, because $p(E, B) = p(E \mid B)p(B)$.

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.

Example: Two coins.

$A =$ coin 1 shows heads

$B =$ coin 2 shows heads

Then $A \perp\!\!\!\perp B$.

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:* $A \perp\!\!\!\perp B \mid C$

Definition (conditional independence)

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

$$p(A, B|C) = p(A|C)\, p(B|C)$$

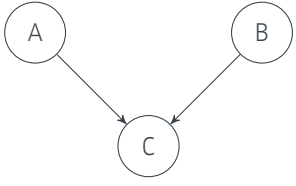
In that case we have $p(A|B, C) = p(A|C)$, i.e. in light of information C, B provides no (further) information about A. Notation: $A \perp\!\!\!\perp B \mid C$

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



$A \perp\!\!\!\perp B$

Definition (conditional independence)

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

$$p(A, B|C) = p(A|C)\, p(B|C)$$

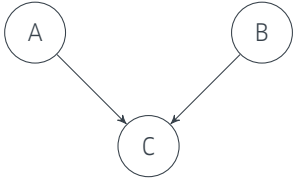
In that case we have $p(A|B, C) = p(A|C)$, i.e. in light of information C, B provides no (further) information about A. Notation: $A \perp\!\!\!\perp B \mid C$

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,*

$$p(A, B|C) = p(A|C)\, p(B|C)$$

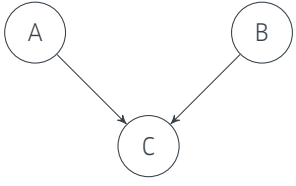
In that case we have $p(A|B, C) = p(A|C)$, i.e. in light of information C, B provides no (further) information about A. Notation: $A \perp\!\!\!\perp B \mid C$

Example: Two coins and a bell.

$A =$ 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 $B \perp\!\!\!\perp C$,

Definition (conditional independence)

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

$$p(A, B|C) = p(A|C)\, p(B|C)$$

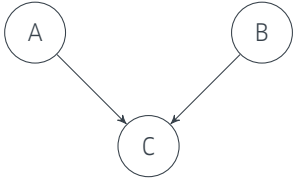
In that case we have $p(A|B, C) = p(A|C)$, i.e. in light of information C, B provides no (further) information about A. Notation: $A \perp\!\!\!\perp B \mid C$

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 $A \not\perp\!\!\!\perp B \mid C$

Definition (conditional independence)

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

$$p(A, B|C) = p(A|C)\, p(B|C)$$

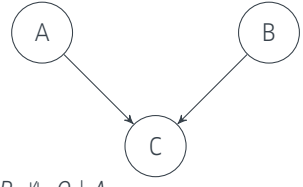
In that case we have $p(A|B, C) = p(A|C)$, i.e. in light of information C, B provides no (further) information about A. Notation: $A \perp\!\!\!\perp B \mid C$

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

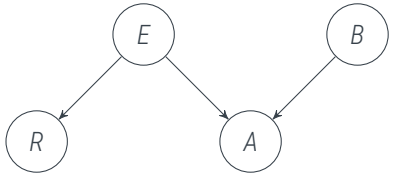


$A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp C$ and $B \perp\!\!\!\perp C$, but $A \not\perp\!\!\!\perp B \mid C$ and $A \not\perp\!\!\!\perp C \mid B$ and $B \not\perp\!\!\!\perp C \mid A$.

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
$E =$ it rained last night
$B =$ there was a break-in
$R =$ an announcement is made on the radio

Conditional probability tables:

$p(E = 1) = 10^{-3}$

$p(R = 1 \mid E = 1) = 1.0$	$p(R = 1 \mid E = 0) = 0.0$

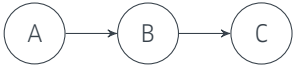
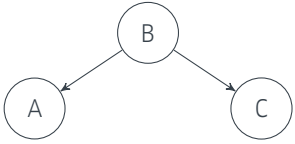
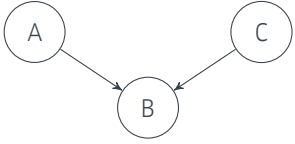
$p(B = 1) = 10^{-3}$	$p(A = 1 \mid E = 1, B = 0) = 0.01099$	$p(A = 1 \mid E = 0, B = 0) = 10^{-3}$
	$p(A = 1 \mid E = 1, B = 1) = 0.9901099$	$p(A = 1 \mid E = 0, B = 1) = 0.99001$

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:

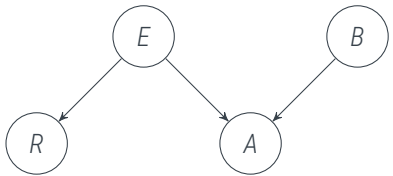
	graph	factorization	implications
(i)		$p(A, B, C) = p(C \mid B) \cdot p(B \mid A) \cdot p(A)$	$A \perp\!\!\!\perp C \mid B$ but not, i.e., $A \not\perp\!\!\!\perp C$
(ii)		$p(A, B, C) = p(A \mid B) \cdot p(C \mid B) \cdot p(B)$	$A \perp\!\!\!\perp C \mid B$ but not, i.e., $A \not\perp\!\!\!\perp C$
(iii)		$p(A, B, C) = p(B \mid A, C) \cdot p(B) \cdot p(A)$	$A \perp\!\!\!\perp C$ but not, i.e., $A \not\perp\!\!\!\perp C \mid B$

Deducing Conditional Independencies

back to our example

	graph	factorization	implications
(i)		$p(A, B, C) = p(C \mid B) \cdot p(B \mid A) \cdot p(A)$	$A \perp\!\!\!\perp C \mid B$ but not, i.e., $A \not\perp\!\!\!\perp C$
(ii)		$p(A, B, C) = p(A \mid B) \cdot p(C \mid B) \cdot p(B)$	$A \perp\!\!\!\perp C \mid B$ but not, i.e., $A \not\perp\!\!\!\perp C$
(iii)		$p(A, B, C) = p(B \mid A, C) \cdot p(B) \cdot p(A)$	$A \perp\!\!\!\perp C$ but not, i.e., $A \not\perp\!\!\!\perp C \mid B$

Which independencies can we infer only from the graph?

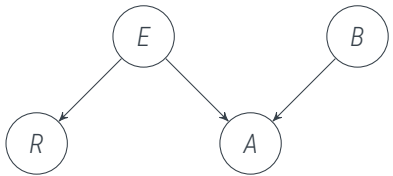


Deducing Conditional Independencies

back to our example

	graph	factorization	implications
(i)		$p(A,B,C) = p(C \mid B) \cdot p(B \mid A) \cdot p(A)$	$A \perp\!\!\!\perp C \mid B$ but not, i.e., $A \not\perp\!\!\!\perp C$
(ii)		$p(A,B,C) = p(A \mid B) \cdot p(C \mid B) \cdot p(B)$	$A \perp\!\!\!\perp C \mid B$ but not, i.e., $A \not\perp\!\!\!\perp C$
(iii)		$p(A,B,C) = p(B \mid A,C) \cdot p(B) \cdot p(A)$	$A \perp\!\!\!\perp C$ but not, i.e., $A \not\perp\!\!\!\perp C \mid B$

Which independencies can we infer only from the graph?



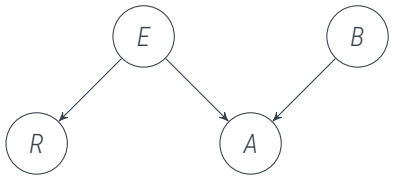
- $R \perp\!\!\!\perp A \mid E$ and $E \perp\!\!\!\perp B$

Deducing Conditional Independencies

back to our example

	graph	factorization	implications
(i)		$p(A,B,C) = p(C \mid B) \cdot p(B \mid A) \cdot p(A)$	$A \perp\!\!\!\perp C \mid B$ but not, i.e., $A \not\perp\!\!\!\perp C$
(ii)		$p(A,B,C) = p(A \mid B) \cdot p(C \mid B) \cdot p(B)$	$A \perp\!\!\!\perp C \mid B$ but not, i.e., $A \not\perp\!\!\!\perp C$
(iii)		$p(A,B,C) = p(B \mid A,C) \cdot p(B) \cdot p(A)$	$A \perp\!\!\!\perp C$ but not, i.e., $A \not\perp\!\!\!\perp C \mid B$

Which independencies can we infer only from the graph?



- $R \perp\!\!\!\perp A \mid E$ and $E \perp\!\!\!\perp B$
- but also $(R \perp\!\!\!\perp B \mid E)$, $(R \perp\!\!\!\perp B)$, $(R \perp\!\!\!\perp B \mid E, A)$, with more work

The Graph for Two Coins and a Bell

DAGs are not a perfect tool

$p(A = 1) = 0.5$

$p(B = 1) = 0.5$

$p(C = 1 \mid A = 1, B = 1) = 1$	$p(C = 1 \mid A = 1, B = 0) = 0$
$p(C = 1 \mid A = 0, B = 1) = 0$	$p(C = 1 \mid A = 0, B = 0) = 1$

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

The Graph for Two Coins and a Bell

DAGs are not a perfect tool

$p(A = 1) = 0.5$

$p(B = 1) = 0.5$

$p(C = 1 \mid A = 1, B = 1) = 1$	$p(C = 1 \mid A = 1, B = 0) = 0$
$p(C = 1 \mid A = 0, B = 1) = 0$	$p(C = 1 \mid A = 0, B = 0) = 1$

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

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

$p(A = 1) = 0.5$

$p(B = 1) = 0.5$

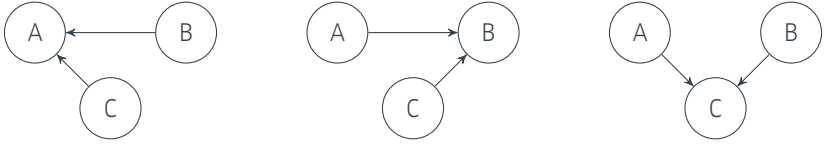
$p(C = 1 \mid A = 1, B = 1) = 1$	$p(C = 1 \mid A = 1, B = 0) = 0$
$p(C = 1 \mid A = 0, B = 1) = 0$	$p(C = 1 \mid A = 0, B = 0) = 1$

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

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. $p(A, B, C) = p(B|C, A) \cdot p(C|A) \cdot p(A) = p(B|C, A) \cdot p(C) \cdot p(A)$

Each corresponds to a graph. Note that each can only express some of the independencies:



Measuring relevance between variables (3)

Definition (conditional independence, extended)

Two ***sets of variables*** $\mathcal{A}$ *and* $\mathcal{B}$ *are* ***conditionally independent*** *given a set of variables* $\mathcal{C}$*, if and only if their conditional distribution factorizes,*

$$p(\mathcal{A}, \mathcal{B}|\mathcal{C}) = p(\mathcal{A}|\mathcal{C})\; p(\mathcal{B}|\mathcal{C})$$

where for $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$*, we define* $p(\mathcal{A}) := p(A_1, A_2, \ldots, A_n)$*. We write* $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{C}$*.*

Measuring relevance between variables (3)

Definition (conditional independence, extended)

Two ***sets of variables*** $\mathcal{A}$ *and* $\mathcal{B}$ *are* ***conditionally independent*** *given a set of variables* $\mathcal{C}$*, if and only if their conditional distribution factorizes,*

$$p(\mathcal{A}, \mathcal{B}|\mathcal{C}) = p(\mathcal{A}|\mathcal{C})\; p(\mathcal{B}|\mathcal{C})$$

where for $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$*, we define* $p(\mathcal{A}) := p(A_1, A_2, \ldots, A_n)$*. We write* $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{C}$*.*

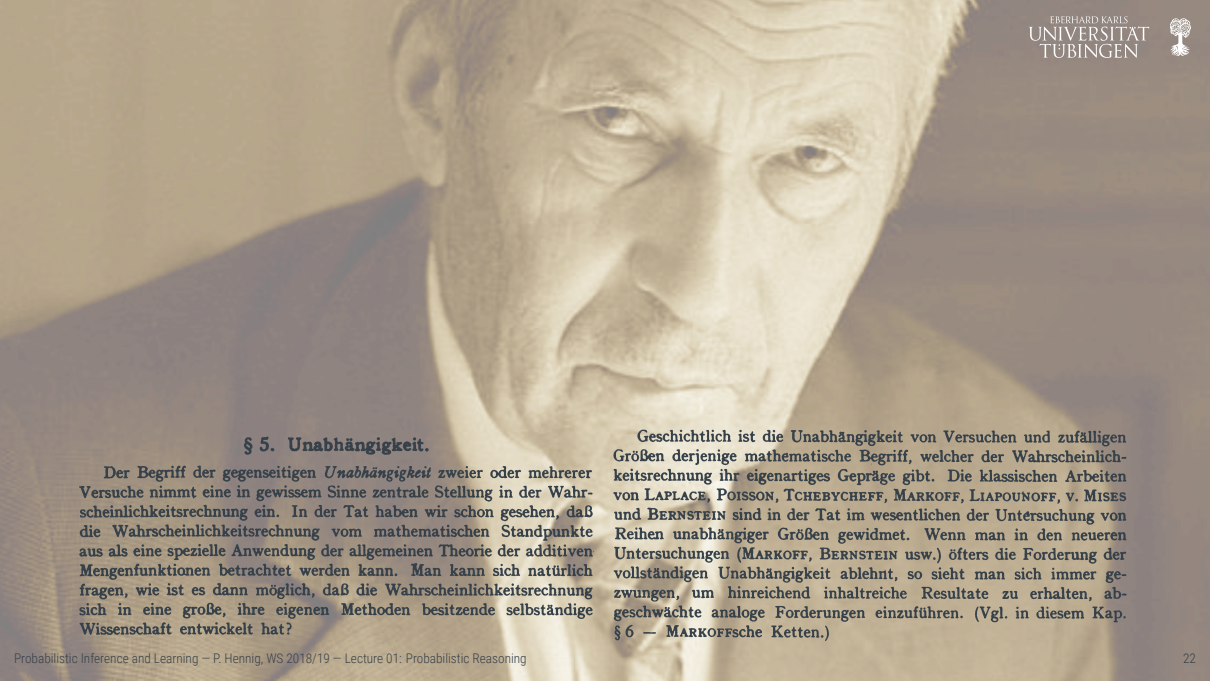
Note: The two previous definitions are special cases of the latter:

$$\begin{aligned} A \perp\!\!\!\perp B \quad &\text{iff} \quad \{A\} \perp\!\!\!\perp \{B\} \\ A \perp\!\!\!\perp B \mid C \quad &\text{iff} \quad \{A\} \perp\!\!\!\perp \{B\} \mid \{C\} \end{aligned}$$

Graphical Models and Conditional Independence

- Multivariate distributions can have **exponentially** many degrees of freedom.
- **(conditional) independence** helps reduce this complexity to make things tractable
- **(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.

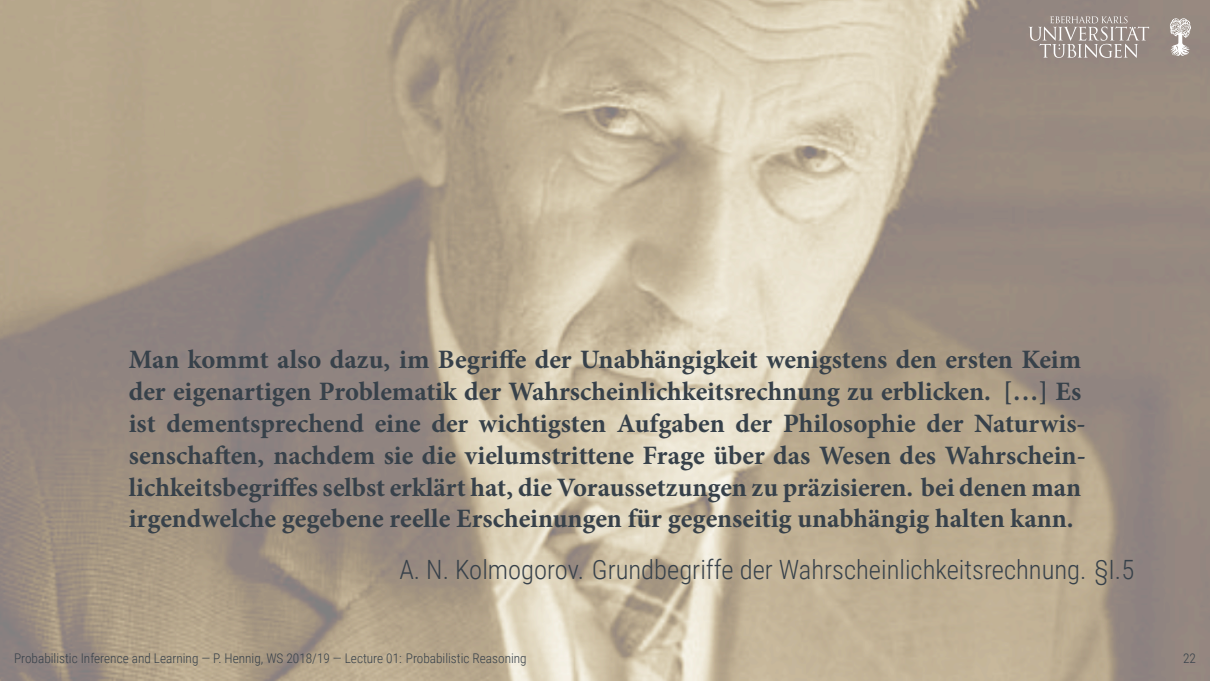
Conditional independence is a tool (and may be required)
to keep inference tractable in multi-variate problems.



§ 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?

Geschichtlich ist die Unabhängigkeit von Versuchen und zufälligen Größen derjenige mathematische Begriff, welcher der Wahrscheinlichkeitsrechnung ihr eigenartiges Gepräge gibt. Die klassischen Arbeiten von LAPLACE, POISSON, TCHEBYCHEFF, MARKOFF, LIAPOUNOFF, v. MISES und BERNSTEIN sind in der Tat im wesentlichen der Untersuchung von Reihen unabhängiger Größen gewidmet. Wenn man in den neueren Untersuchungen (MARKOFF, BERNSTEIN usw.) öfters die Forderung der vollständigen Unabhängigkeit ablehnt, so sieht man sich immer gezwungen, um hinreichend inhaltreiche Resultate zu erhalten, abgeschwächte analoge Forderungen einzuführen. (Vgl. in diesem Kap. § 6 — MARKOFFsche Ketten.)



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. §I.5

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Definition (alphabet)

The **alphabet** $\mathcal{A} = \mathcal{V} \cup \{\neg, \wedge, \vee, \Rightarrow,), (\}$ consists of various symbols:

- a finite set $\mathcal{V}$ of **symbols** $X, Y, Z, \ldots$; aka. non-logical symbols, propositional variables
- **junctors** $\neg, \wedge, \vee, \rightarrowtriangle$; aka. logical symbols, connectives, names of truth functions
- **parentheses** $), ($

In the theory of formal grammars, $\mathcal{F}$ is a context-free language, i.e. the above definition is a context-free grammar (type II of Chomsky's hierarchy).

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Definition (alphabet)

The **alphabet** $\mathcal{A} = \mathcal{V} \cup \{\neg, \wedge, \vee, \Rightarrow,), (\}$ consists of various symbols:

- a finite set $\mathcal{V}$ of **symbols** $X, Y, Z, \ldots$; aka. non-logical symbols, propositional variables
- **junctors** $\neg, \wedge, \vee, \rightarrowtriangle$; aka. logical symbols, connectives, names of truth functions
- **parentheses** $), ($

Definition (formula)

A **formula** A is a finite-length string of symbols build along the following inductive definition:

- all symbols in $\mathcal{V}$ are formulas
- if A and B are formulas, then $\neg A, A \wedge B, A \vee B, A \Rightarrow B$ and (A) are also formulas

$\mathcal{F}$ is the set of all formulas. $\mathcal{F}$ is a subset of all strings: $\mathcal{F} \subset \mathcal{A}^*$.

In the theory of formal grammars, $\mathcal{F}$ is a context-free language, i.e. the above definition is a context-free grammar (type II of Chomsky's hierarchy).

Definition (Boolean assignment)

A Boolean assignment ω assigns every propositional variable in $\mathcal{V}$ a truth value, i.e.

$$\omega : \mathcal{V} \to \{0, 1\}$$

where 0 represents **false** and 1 represents **true**.

Definition (entailment)

A Boolean assignment ω induces a truth function $q : \mathcal{F} \to \{0, 1\}$ that maps all formulas onto truth values as follows:

- $q(X) := \omega(X)$ for propositional variables $X \in \mathcal{V}$
- $q(\neg A) := 1 - q(A)$ and $q(A \wedge B) := q(A)q(B)$ and $q((A)) := q(A)$
- $q(A \wedge B) := q(\neg(\neg A \wedge \neg B))$ and $q(A \Rightarrow B) := q(\neg A \vee B)$

If $q(A) = 1$, we say that ω **entails** A and write $\omega \models A$.

Definition (sample space)

The set Ω of all Boolean assignments ω is called **sample space**.
(Note that, for n propositional variables, it has 2^n elements.)

Definition (sample space)

The set Ω of all Boolean assignments ω is called **sample space**.
(Note that, for n propositional variables, it has 2^n elements.)

Definition (probability mass function)

The **probability mass function** $f : \Omega \to [0, 1]$ assigns each Boolean assignment a probability,

such that $0 \leq f(\omega) \leq 1$ for all $\omega \in \Omega$ and $\sum_{\omega \in \Omega} f(\omega) = 1$.

Definition (sample space)

The set Ω of all Boolean assignments ω is called **sample space**.
(Note that, for n propositional variables, it has 2^n elements.)

Definition (probability mass function)

The **probability mass function** $f : \Omega \to [0, 1]$ assigns each Boolean assignment a probability,

$$\text{such that} \quad 0 \leq f(\omega) \leq 1 \text{ for all } \omega \in \Omega \quad \text{and} \quad \sum_{\omega \in \Omega} f(\omega) = 1.$$

Definition (event)

An **event** $E \subset \Omega$ is a subset of the sample space. Each formula A naturally induces an event E_A :

$$E_A := \{\omega \in \Omega \text{ such that } \omega \vDash A\} \subset \Omega.$$

Different formulas can induce the same event. Note that $\Omega = E_{A \wedge \neg A}$.

Definition (probability)

The **probability** $p(E)$ of an event E is the probability mass of E. That is,

$$p(E) := \sum_{\omega \in E} f(\omega).$$

The probability $p(A)$ of a formula A is defined as the probability $p(E_A)$ of the induced event E_A:

$$p(A) := p(E_A) = \sum_{\omega \in E_A} f(\omega) = \sum_{\omega \models A} f(\omega).$$

Definition (probability)

The **probability** $p(E)$ of an event E is the probability mass of E. That is,

$$p(E) := \sum_{\omega \in E} f(\omega).$$

The probability $p(A)$ of a formula A is defined as the probability $p(E_A)$ of the induced event E_A:

$$p(A) := p(E_A) = \sum_{\omega \in E_A} f(\omega) = \sum_{\omega \models A} f(\omega).$$

Definition (joint probability)

The **joint probability** $p(A, B)$ of several formulas A and B is the probability of their conjunction:

$$p(A, B) := p(A \wedge B).$$

Analogously for more than two. The joint probability of several events is the probability of their intersection (remember events are subsets of Ω):

$$p(E_1, E_2) = p(E_1 \cap E_2).$$

Theorem (Kolmogorov's axioms, and more)

1. $0 \leq p(A) \leq 1$
2. $p(\Omega) = 1$
3. *if* $E_A \cap E_B = \varnothing$, *then* $p(A \vee B) = p(A) + p(B)$
4. $P(A \vee B) = p(A) + p(B) - p(A, B)$
5. $p(A) = p(A, B) + p(A, \neg B)$
6. $p(A) + p(\neg A) = 1$
7. $p(A \vee B) = p(\neg(\neg A \wedge \neg B))$ *and* $p(A \Rightarrow B) = p(\neg A \vee B)$

The first three are Kolmogorov's axioms, which imply 4.,5., and 6.

1., 2., and 7. hold by definition. Proof of 3.: (using $E_A \cap E_B = \varnothing$)

$$p(A \vee B) = \sum_{\omega \in E_{A \vee B}} f(\omega) = \sum_{\omega \in E_A} f(\omega) + \sum_{\omega \in E_B} f(\omega) = p(A) + p(B) \qquad \square$$

Boole was a Bayesian

G. Boole. *An Investigation of the Laws of Thought*, 1854, §XVI, p. 249

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.

George Boole
(1815–1864)

[Boole, 1854; Stefan Harmeling, 2013]

Definition (conditional probability)

For some formula or event B with non-zero probability (i.e. $p(B) > 0$), the **conditional probability** $p(A \mid B)$ is defined as

$$p(A \mid B) := \frac{p(A, B)}{p(B)}.$$

[Boole, 1854; Stefan Harmeling, 2013]

Definition (conditional probability)

For some formula or event B with non-zero probability (i.e. $p(B) > 0$), the **conditional probability** $p(A \mid B)$ is defined as

$$p(A \mid B) := \frac{p(A, B)}{p(B)}.$$

Note: defining $p(A \mid B) = 1$ for $p(B) = 0$ (which looks like "ex falso quod libet") is a bad idea, since it would imply

$$p(A \mid B) + p(\neg A \mid B) = 2 \neq 1.$$

From Propositional Logic to Probabilities

[Boole, 1854; Stefan Harmeling, 2013]

Definition (conditional probability)

For some formula or event B with non-zero probability (i.e. $p(B) > 0$), the **conditional probability** $p(A \mid B)$ is defined as

$$p(A \mid B) := \frac{p(A, B)}{p(B)}.$$

Note: defining $p(A \mid B) = 1$ for $p(B) = 0$ (which looks like "ex falso quod libet") is a bad idea, since it would imply

$$p(A \mid B) + p(\neg A \mid B) = 2 \neq 1.$$

Theorem (Kolmogorov's axioms, Bayes' Theorem and more)

For fixed B, the conditional probability $p(A \mid B)$ fulfills Kolmogorov's axioms (and thus also Bayes' Theorem).

Lemma (implication vs. conditional probability)

Assume $p(A) \geq 0$ (otherwise $p(B \mid A)$ not defined).

1. $p(B \mid A) = \dfrac{p(A) - (1 - p(A \Rightarrow B))}{p(A)}$
2. $p(A) = 1$ *implies* $p(B \mid A) = p(A \Rightarrow B)$
3. $p(A \Rightarrow B) \geq p(B \mid A) \geq p(B, A)$
4. $p(A \Rightarrow B) \geq 1 - p(A)$
5. $p(A \Rightarrow B) = 1$ *if, and only if,* $p(B \mid A) = 1$

Proof:

$$p(B \mid A) \overset{\text{defin.}}{=} \frac{p(B,A)}{p(A)} \overset{\text{sum rule}}{\underset{5.\text{above}}{=}} \frac{p(A) - p(A, \neg B)}{p(A)} \overset{7.\ \text{above}}{=} \frac{p(A) - (1 - p(\neg A \vee B))}{p(A)}$$

$$\overset{7.\ \text{above}}{=} \frac{p(A) - (1 - p(A \Rightarrow B))}{p(A)}$$

□

From Propositional Logic to Probabilities:

- Propositional Logic can be **extended** by assigning probability mass to Boolean assignments.
- Doing so provides a motivation for Kolmogorov's axioms, and additional rules for propositional formulae.
- In particular, $p(A \Rightarrow B) = p(\neg A \vee B)$