Representations of uncertainty

Bayesian Statistics and Machine Learning

Dominik Endres

October 25, 2017

Philipps Universität Marburg

Motivation

- Representation of **uncertainty** is a major concern for Machine Learning.
- For successful learning, types of uncertainty need to be traded off against each other:
  - randomness
  - uncertainty about the model parameters (curse of dimensionality), i.e. ignorance.

Outline

possible worlds

Most representations of uncertainty start with a set **W of possible worlds** (ignorance) or **elementary outcomes** (randomness). These are worlds or outcomes which are considered possible.

- Example (randomness): tossing a die. Six elementary outcomes possible, represented by set $W = \{w_1, w_2, \ldots, w_6\}$.
- Example (ignorance): polynomial curve fitting. Consider degrees $M < 10$. The possible worlds are also representable by a set $W = \{m_0, \ldots, m_{10}\}$.

Unless stated otherwise, we assume that W is finite.

possible worlds

Most representations of uncertainty start with a set W of **possible worlds** (ignorance) or **elementary outcomes** (randomness). These are worlds or outcomes which are considered possible.

- Example (randomness): tossing a die. Six elementary outcomes possible, represented by set $W = \{w_1, w_2, \ldots, w_6\}$.
- Example (ignorance): polynomial curve fitting. Consider degrees $M < 10$. The possible worlds are also representable by a set $W = \{m_0, \ldots, m_{10}\}$.

Unless stated otherwise, we assume that W is finite.

possible worlds

Most representations of uncertainty start with a set W of **possible worlds** (ignorance) or **elementary outcomes** (randomness). These are worlds or outcomes which are considered possible.

- Example (randomness): tossing a die. Six elementary outcomes possible, represented by set $W = \{w_1, w_2, \ldots, w_6\}$.
- Example (ignorance): polynomial curve fitting. Consider degrees $M < 10$. The possible worlds are also representable by a set $W = \{m_0, \ldots, m_{10}\}$.

Unless stated otherwise, we assume that W is finite.

Randomness vs. ignorance

Two main aspects of uncertainty:

1. Randomness
   - Familiar to people with engineering/physics background
   - Used to represent things (e.g. events, values of variables) which are:
     - truly unknowable before observation (quantum mechanics)
     - impractical to describe deterministically (statistical mechanics)
     - uninteresting (measurement noise, also statistical mechanics)
2. Ignorance
   - Familiar to people with a background in logic/philosophy/computer science
   - Used to represent things (e.g. propositions, values of variables/parameters) which
     - we are not sure about
     - we believe have a (fixed, deterministic, existing) value
   - Formalizing ignorance allows us to reason in its presence!

Events and propositions

The objects which are considered known/possible/probable are **events** (randomness) or **propositions** (ignorance). They are subsets of W.

- Example: (randomness) 'Die shows even number'$=\{w_2, w_4, w_6\} \subseteq W$.
- Example: (ignorance) 'The correct degree M is between 3 and 5'$=\{m_3, m_4, m_5\} \subseteq W$.

Assume your uncertainty was represented by a set $W' \subseteq W$. Then

- you consider an event U **possible**, if $U \cap W' \neq \emptyset$
- you consider an event U **impossible**, if $U \cap W' = \emptyset$
- you consider an event U **certain**, of $W' \subseteq U$.

Example: after die was tossed, someone tells you that $W' = \{w_1, w_2\}$. Then

- you consider $U = \{w_2, w_4, w_6\}$ possible
- you are certain of $U = \{w_1, w_2, w_3\}$.

Events and propositions

The objects which are considered known/possible/probable are **events** (randomness) or **propositions** (ignorance). They are subsets of W.

- Example: (randomness) 'Die shows even number'$=\{w_2, w_4, w_6\} \subseteq W$.
- Example: (ignorance) 'The correct degree M is between 3 and 5'$=\{m_3, m_4, m_5\} \subseteq W$.

Assume your uncertainty was represented by a set $W' \subseteq W$. Then

- you consider an event U **possible**, if $U \cap W' \neq \emptyset$
- you consider an event U **impossible**, if $U \cap W' = \emptyset$
- you consider an event U **certain**, of $W' \subseteq U$.

Example: after die was tossed, someone tells you that $W' = \{w_1, w_2\}$. Then

- you consider $U = \{w_2, w_4, w_6\}$ possible
- you are certain of $U = \{w_1, w_2, w_3\}$.

Events and propositions

The objects which are considered known/possible/probable are **events** (randomness) or **propositions** (ignorance). They are subsets of W.

- Example: (randomness) 'Die shows even number'$=\{w_2, w_4, w_6\} \subseteq W$.
- Example: (ignorance) 'The correct degree M is between 3 and 5'$=\{m_3, m_4, m_5\} \subseteq W$.

Assume your uncertainty was represented by a set $W' \subseteq W$. Then

- you consider an event U **possible**, if $U \cap W' \neq \emptyset$
- you consider an event U **impossible**, if $U \cap W' = \emptyset$
- you consider an event U **certain**, of $W' \subseteq U$.

Example: after die was tossed, someone tells you that $W' = \{w_1, w_2\}$. Then

- you consider $U = \{w_2, w_4, w_6\}$ possible
- you are certain of $U = \{w_1, w_2, w_3\}$.

Certainty, possibility and impossibility

Example: after die was tossed, someone tells you that $W' = \{w_1, w_2\}$. Then

- you consider $U = \{w_2, w_4, w_6\}$ possible
- you are certain of $U = \{w_1, w_2, w_3\}$.

Let the complement of U, $\bar{U} := \{w : w \in W \wedge w \notin U\}$. Then

- U is certain iff $\bar{U}$ is impossible.

This is a very **coarse-grained** representation of uncertainty:

- three-valued: certain, possible, impossible.
- already more fine-grained than boolean logic.

Certainty, possibility and impossibility

Example: after die was tossed, someone tells you that $W' = \{w_1, w_2\}$. Then

- you consider $U = \{w_2, w_4, w_6\}$ possible
- you are certain of $U = \{w_1, w_2, w_3\}$.

Let the complement of U, $\bar{U} := \{w : w \in W \wedge w \notin U\}$. Then

- U is certain iff $\bar{U}$ is impossible.

This is a very **coarse-grained** representation of uncertainty:

- three-valued: certain, possible, impossible.
- already more fine-grained than boolean logic.

Updating of uncertainty, conditioning

Contains a simple notion of **conditioning**, the updating of uncertainty upon receipt of information.

- Before die was tossed: $W = \{w_1, w_2, \ldots, w_6\}$.
- $U = \{w_2, w_4, w_6\}$ is possible.
- $V = \{w_1, w_2, w_3\}$ is possible.

- After die was tossed, we learn: $W' = \{w_4, w_6\}$.
- $U = \{w_2, w_4, w_6\}$ is certain.
- $V = \{w_1, w_2, w_3\}$ is impossible.

Updating of uncertainty, conditioning

Contains a simple notion of **conditioning**, the updating of uncertainty upon receipt of information.

- Before die was tossed: $W = \{w_1, w_2, \ldots, w_6\}$.
- $U = \{w_2, w_4, w_6\}$ is possible.
- $V = \{w_1, w_2, w_3\}$ is possible.

- After die was tossed, we learn: $W' = \{w_4, w_6\}$.
- $U = \{w_2, w_4, w_6\}$ is certain.
- $V = \{w_1, w_2, w_3\}$ is impossible.

Conditioning on elementary events $w \in W$

A very important form of conditioning: information received is in the form

$$W' = \{w\}.$$

You learn that one of the possible worlds is the 'real' one (i.e. the only possible one). A.k.a. conditioning on 'the data'.

- Whole Machine Learning textbooks have been written about it.
- Bayes' rule is an instance of this type of conditioning.
- For any $U \neq \emptyset$: either U is certain ($w \in U$) or impossible ($w \notin U$).

$\Rightarrow$ more fine-grained representation of uncertainty is desirable.

Conditioning on elementary events $w \in W$

A very important form of conditioning: information received is in the form

$$W' = \{w\}.$$

You learn that one of the possible worlds is the 'real' one (i.e. the only possible one). A.k.a. conditioning on 'the data'.

- Whole Machine Learning textbooks have been written about it.
- Bayes' rule is an instance of this type of conditioning.
- For any $U \neq \emptyset$: either U is certain ($w \in U$) or impossible ($w \notin U$).

$\Rightarrow$ more fine-grained representation of uncertainty is desirable.

Probability theory in Neuroscience

REVIEW

nature neuroscience

Probabilistic brains: knowns and unknowns

Alexandre Pouget[1–3], Jeffrey M Beck[1], Wei Ji Ma[4,5] & Peter E Latham[3]

There is strong behavioral and physiological evidence that the brain both represents probability distributions and performs probabilistic inference. Computational neuroscientists have started to shed light on how these probabilistic representations and computations might be implemented in neural circuits. One particularly appealing aspect of these theories is their generality: they can be used to model a wide range of tasks, from sensory processing to high-level cognition. To date, however, these theories have only been applied to very simple tasks. Here we discuss the challenges that will emerge as researchers start focusing their efforts on real-life computations, with a focus on probabilistic learning, structural learning and approximate inference.

erved.

frontiers in HUMAN NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 02 December 2010
doi: 10.3389/fnhum.2010.00215

Attention, uncertainty, and free-energy

*Harriet Feldman and Karl J. Friston**

The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London, UK

Edited by:
Sven Bestmann, University College London, UK

Reviewed by:
William Milberg, Harvard Medical School, USA
Tamer Demiralp, Istanbul University, Turkey

We suggested recently that attention can be understood as inferring the level of uncertainty or precision during hierarchical perception. In this paper, we try to substantiate this claim using neuronal simulations of directed spatial attention and biased competition. These simulations assume that neuronal activity encodes a probabilistic representation of the world that optimizes free-energy in a Bayesian fashion. Because free-energy bounds surprise or the (negative) log-evidence for internal models of the world, this optimization can be regarded as evidence accumulation or (generalized) predictive coding. Crucially, both predictions about the state of

LETTER — Communicated by Richard Zemel

Bayesian Spiking Neurons I: Inference

Sophie Deneve
sophie.deneve@ens.fr
Group for Neural Theory, Département d'Etudes Cognitives, Ecole Normale Supérieure, Collège de France, 75005 Paris, France

We show that the dynamics of spiking neurons can be interpreted as a form of Bayesian inference in time. Neurons that optimally integrate evidence about events in the external world exhibit properties similar to

LETTER — Communicated by Jochen Ditterich

Bayesian Spiking Neurons II: Learning

Sophie Deneve
sophie.deneve@ens.fr
Group for Neural Theory, Département d'Etudes Cognitives, College de France 75005 Paris, France
Ecole Normale Supérieure, College de France 75005 Paris, France

In the companion letter in this issue ("Bayesian Spiking Neurons I: Inference"), we showed that the dynamics of spiking neurons can be interpreted as a form of Bayesian integration, accumulating evidence over time about events in the external world or the body. We proceed to develop a theory of Bayesian learning in spiking neural networks, where

Probability theory in Neuroscience

frontiers in NEUROSCIENCE
ORIGINAL RESEARCH ARTICLE
published: 05 June 2012
doi: 10.3389/fnins.2012.00075

Making decisions with unknown sensory reliability

Sophie Deneve[1,2]*

[1] *Département d'Etudes Cognitives, Group for Neural Theory, Ecole Normale Supérieure, Paris, France*
[2] *Groupe de Neuroscience Théorique, Collège de France, Paris, France*

Edited by:
Gabriel José Corrêa Mograbi, Federal University of Mato Grosso, Brazil

Reviewed by:
Floris P. De Lange, Radboud University Nijmegen, Netherlands
Philippe N. Tobler, University of Zurich, Switzerland

***Correspondence:**
Sophie Deneve, Département d'Etudes Cognitives, Group for Neural Theory, Ecole Normale Supérieure, 29, rue d'Ulm, 75005 Paris, France.
e-mail: sophie.deneve@ens.fr

To make fast and accurate behavioral choices, we need to integrate noisy sensory input, take prior knowledge into account, and adjust our decision criteria. It was shown previously that in two-alternative-forced-choice tasks, optimal decision making can be formalized in the framework of a sequential probability ratio test and is then equivalent to a diffusion model. However, this analogy hides a "chicken and egg" problem: to know how quickly we should integrate the sensory input and set the optimal decision threshold, the reliability of the sensory observations must be known in advance. Most of the time, we cannot know this reliability without first observing the decision outcome. We consider here a Bayesian decision model that simultaneously infers the probability of two different choices and at the same time estimates the reliability of the sensory information on which this choice is based. We show that this can be achieved within a single trial, based on the noisy responses of sensory spiking neurons. The resulting model is a non-linear diffusion to bound where the weight of the sensory inputs and the decision threshold are both dynamically chang-

REVIEW

nature neuroscience

Probabilistic brains: knowns and unknowns

Alexandre Pouget[1–3], Jeffrey M Beck[1], Wei Ji Ma[4,5] & Peter E Latham[3]

There is strong behavioral and physiological evidence that the brain both represents probability distributions and performs probabilistic inference. Computational neuroscientists have started to shed light on how these probabilistic representations and computations might be implemented in neural circuits. One particularly appealing aspect of these theories is their generality: they can be used to model a wide range of tasks, from sensory processing to high-level cognition. To date, however, these theories have only been applied to very simple tasks. Here we discuss the challenges that will emerge as researchers start focusing their efforts on real-life computations, with a focus on probabilistic learning, structural learning and approximate inference.

Probability – randomness perspective

The most prominent representation of uncertainty is **probability**.

Assigns a real number to subsets of possible worlds.

Customary but not necessary: assign real number to every possible world.

Example: suppose a fair die is rolled repeatedly. It would show each possible face 1/6th of the time.

- Assign 1/6 to every $w_i \in W$. Write this as $P(\{w_i\}) = P(w_i) = \frac{1}{6}$.
    - An instance of the **principle of indifference**.
- It seems natural to assign $P(\{w_1, w_2\}) = \frac{1}{6} + \frac{1}{6}$.
    - More generally: if $U, V \in W$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.
- Since something must happen, assign $P(W) = 1$.
- An instance of the **relative frequency** interpretation of probability.

Probability – randomness perspective

The most prominent representation of uncertainty is **probability**. Assigns a real number to subsets of possible worlds.

Customary but not necessary: assign real number to every possible world.

Example: suppose a fair die is rolled repeatedly. It would show each possible face 1/6th of the time.

- Assign $1/6$ to every $w_i \in W$. Write this as $P(\{w_i\}) = P(w_i) = \frac{1}{6}$.
  - An instance of the **principle of indifference**.
- It seems natural to assign $P(\{w_1, w_2\}) = \frac{1}{6} + \frac{1}{6}$.
  - More generally: if $U, V \in W$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.
- Since something must happen, assign $P(W) = 1$.
- An instance of the **relative frequency** interpretation of probability.

Probability – randomness perspective

The most prominent representation of uncertainty is **probability**.
Assigns a real number to subsets of possible worlds.
Customary but not necessary: assign real number to every possible world.
Example: suppose a fair die is rolled repeatedly. It would show each possible face 1/6th of the time.

- Assign 1/6 to every $w_i \in W$. Write this as $P(\{w_i\}) = P(w_i) = \frac{1}{6}$.
  - An instance of the **principle of indifference**.
- It seems natural to assign $P(\{w_1, w_2\}) = \frac{1}{6} + \frac{1}{6}$.
  - More generally: if $U, V \in W$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.
- Since something must happen, assign $P(W) = 1$.
- An instance of the **relative frequency** interpretation of probability.

Probability – randomness perspective

The most prominent representation of uncertainty is **probability**.
Assigns a real number to subsets of possible worlds.
Customary but not necessary: assign real number to every possible world.
Example: suppose a fair die is rolled repeatedly. It would show each possible face 1/6th of the time.

- Assign 1/6 to every $w_i \in W$. Write this as $P(\{w_i\}) = P(w_i) = \frac{1}{6}$.
  - An instance of the **principle of indifference**.
- It seems natural to assign $P(\{w_1, w_2\}) = \frac{1}{6} + \frac{1}{6}$.
  - More generally: if $U, V \in W$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.
- Since something must happen, assign $P(W) = 1$.
- An instance of the **relative frequency** interpretation of probability.

Why not assign a probability to every possible world?

Example: an Urn contains N(red)=30 red balls. It also contains blue and yellow balls such that N(blue)+N(yellow)=70. Now a ball is drawn at random from the Urn.

- The set of possible worlds is $W = \{r, b, y\}$.
- It seems natural to assign $P(r) = 0.3$.
- But what about $P(y)$ and $P(b)$?
- Insufficient information to assign a probability to $\{y\}$ or $\{b\}$
- Assign probability $P(\{b, y\}) = 0.7$.

Why not assign a probability to every possible world?

Example: an Urn contains N(red)=30 red balls. It also contains blue and yellow balls such that N(blue)+N(yellow)=70. Now a ball is drawn at random from the Urn.

- The set of possible worlds is $W = \{r, b, y\}$.
- It seems natural to assign $P(r) = 0.3$.
- But what about $P(y)$ and $P(b)$?
- Insufficient information to assign a probability to $\{y\}$ or $\{b\}$
- Assign probability $P(\{b, y\}) = 0.7$.

Why not assign a probability to every possible world?

Example: an Urn contains N(red)=30 red balls. It also contains blue and yellow balls such that N(blue)+N(yellow)=70. Now a ball is drawn at random from the Urn.

- The set of possible worlds is $W = \{r, b, y\}$.
- It seems natural to assign $P(r) = 0.3$.
- But what about $P(y)$ and $P(b)$?
- Insufficient information to assign a probability to $\{y\}$ or $\{b\}$
- Assign probability $P(\{b, y\}) = 0.7$.

σ-algebra over W

So it (sometimes) makes sense to assign probabilities only to some subsets of the possible worlds.

Question: which subsets are typically used?

Definition: A σ-algebra over W is a set $\mathcal{F}$ of subsets of W that

- contains W, and
- is closed under union, i.e. if U and V are in $\mathcal{F}$, then so is $U \cup V$, and
- is closed under complementation, i.e. if U is in $\mathcal{F}$, then so is $\bar{U} = \{w : w \in W \wedge w \notin U\}$.

Note that a σ-algebra is also closed under intersection.

σ-algebra over W

So it (sometimes) makes sense to assign probabilities only to some subsets of the possible worlds.

Question: which subsets are typically used?

Definition: A σ-algebra over W is a set $\mathcal{F}$ of subsets of W that

- contains W, and
- is closed under union, i.e. if U and V are in $\mathcal{F}$, then so is $U \cup V$, and
- is closed under complementation, i.e. if U is in $\mathcal{F}$, then so is $\bar{U} = \{w : w \in W \wedge w \notin U\}$.

Note that a σ-algebra is also closed under intersection.

Example 1: σ-algebra

Example: an Urn contains N(red)=30 red balls. It also contains blue and yellow balls such that N(blue)+N(yellow)=70. Now a ball is drawn at random from the Urn.

Possible worlds $W = \{r, b, y\}$

σ-Algebra $\mathcal{F} = \{\emptyset, W, \{r\}, \{b, y\}\}$

Probability assignments

- $P(W) = 1$
- $P(\emptyset) = 0$
- $P(r) = 0.3$
- $P(\{b, y\}) = 0.7$

Example 1: σ-algebra

Example: an Urn contains N(red)=30 red balls. It also contains blue and yellow balls such that N(blue)+N(yellow)=70. Now a ball is drawn at random from the Urn.

Possible worlds $W = \{r, b, y\}$

σ-**Algebra** $\mathcal{F} = \{\emptyset, W, \{r\}, \{b, y\}\}$

Probability assignments

- $P(W) = 1$
- $P(\emptyset) = 0$
- $P(r) = 0.3$
- $P(\{b, y\}) = 0.7$

Example 1: σ-algebra

Example: an Urn contains N(red)=30 red balls. It also contains blue and yellow balls such that N(blue)+N(yellow)=70. Now a ball is drawn at random from the Urn.

Possible worlds $W = \{r, b, y\}$

σ-**Algebra** $\mathcal{F} = \{\emptyset, W, \{r\}, \{b, y\}\}$

Probability assignments

- $P(W) = 1$
- $P(\emptyset) = 0$
- $P(r) = 0.3$
- $P(\{b, y\}) = 0.7$

Example 2: powerset algebra

Example: an Urn contains N(red)=30 red balls, N(blue)=20 blue balls and N(yellow)=50 yellow balls. Now a ball is drawn at random from the Urn.

Possible worlds $W = \{r, b, y\}$

σ-Algebra: since we can assign probabilities to all possible worlds (singleton subsets of W), we can assign probabilities to all subsets of W. Thus, all subsets of W are in $\mathcal{F}$:

$\mathcal{F} = \{\emptyset, W, \{r\}, \{b\}, \{y\}, \{r, b\}, \{r, y\}, \{b, y\}\} = 2^W$

2^W is the *powerset* of W.

Probability assignments

- $P(W) = 1$, $P(\emptyset) = 0$
- $P(r) = 0.3$, $P(y) = 0.5$, $P(b) = 0.2$
- $P(\{r, y\}) = 0.8$, $P(\{r, b\}) = 0.5$, $P(\{b, y\}) = 0.7$.

Example 2: powerset algebra

Example: an Urn contains N(red)=30 red balls, N(blue)=20 blue balls and N(yellow)=50 yellow balls. Now a ball is drawn at random from the Urn.

Possible worlds $W = \{r, b, y\}$

σ-**Algebra**: since we can assign probabilities to all possible worlds (singleton subsets of W), we can assign probabilities to all subsets of W. Thus, all subsets of W are in $\mathcal{F}$:

$\mathcal{F} = \{\emptyset, W, \{r\}, \{b\}, \{y\}, \{r, b\}, \{r, y\}, \{b, y\}\} = 2^W$

2^W is the *powerset of W*.

Probability assignments

- $P(W) = 1$, $P(\emptyset) = 0$
- $P(r) = 0.3$, $P(y) = 0.5$, $P(b) = 0.2$
- $P(\{r, y\}) = 0.8$, $P(\{r, b\}) = 0.5$, $P(\{b, y\}) = 0.7$.

Example 2: powerset algebra

Example: an Urn contains N(red)=30 red balls, N(blue)=20 blue balls and N(yellow)=50 yellow balls. Now a ball is drawn at random from the Urn.

Possible worlds $W = \{r, b, y\}$

σ-**Algebra**: since we can assign probabilities to all possible worlds (singleton subsets of W), we can assign probabilities to all subsets of W. Thus, all subsets of W are in $\mathcal{F}$:

$\mathcal{F} = \{\emptyset, W, \{r\}, \{b\}, \{y\}, \{r, b\}, \{r, y\}, \{b, y\}\} = 2^W$

2^W is the *powerset* of W.

Probability assignments

- $P(W) = 1$, $P(\emptyset) = 0$
- $P(r) = 0.3$, $P(y) = 0.5$, $P(b) = 0.2$
- $P(\{r, y\}) = 0.8$, $P(\{r, b\}) = 0.5$, $P(\{b, y\}) = 0.7$.

Definition of probability

Definition: A *probability space* is a tuple $(W, \mathcal{F}, P)$, where $\mathcal{F}$ is a σ-algebra over W and $P : \mathcal{F} \rightarrow [0, 1]$, with the properties:

P1 $P(W) = 1$

P2 If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.

Notes

- P is a *probability measure on* $\mathcal{F}$
- $U \in \mathcal{F}$ are the *measurable sets.*
- If all $w \in W$ are $\{w\} \in \mathcal{F}$, then all subsets of W are measurable.
- $P(\emptyset) = 0$, since $1 = P(W) = P(W \cup \emptyset) = P(W) + P(\emptyset)$
- No notion of conditioning so far!

Assigning probabilities: the principle of indifference

Question: how to choose the values of $P(U)$?

A possible answer is given by the **principle of indifference**: assume all elementary outcomes are equally probable.

- Roll a die with 6 faces: each outcome has probability $\frac{1}{6}$.
- Toss a coin: $W = \{h, t\}$. $P(h) = P(t) = \frac{1}{2}$.

But what if the coin is tossed twice? What is the probability of observing heads twice ?

option 1 $W = \{\{2h\}, \{2t\}, \{h, t\}\}$. Probability $P(2h) = \frac{1}{3}$.

option 2 $W = \{(h, h), (h, t), (t, h), (t, t)\}$. Probability $P((h, h)) = \frac{1}{4}$.

2 seems to be in better agreement with experimental results.

When constructing W, keep as much information about outcomes as possible?

How much is enough ? What if the experiment can't really be repeated ?

Assigning probabilities: the principle of indifference

Question: how to choose the values of $P(U)$?
A possible answer is given by the **principle of indifference**: assume all elementary outcomes are equally probable.

- Roll a die with 6 faces: each outcome has probability $\frac{1}{6}$.
- Toss a coin: $W = \{h, t\}$. $P(h) = P(t) = \frac{1}{2}$.

But what if the coin is tossed twice? What is the probability of observing heads twice ?

option 1 $W = \{\{2h\}, \{2t\}, \{h, t\}\}$. Probability $P(2h) = \frac{1}{3}$.

option 2 $W = \{(h, h), (h, t), (t, h), (t, t)\}$. Probability $P((h, h)) = \frac{1}{4}$.

2 seems to be in better agreement with experimental results.
When constructing W, keep as much information about outcomes as possible?
How much is enough ? What if the experiment can't really be repeated ?

Assigning probabilities: the principle of indifference

Question: how to choose the values of $P(U)$?
A possible answer is given by the **principle of indifference**: assume all elementary outcomes are equally probable.

- Roll a die with 6 faces: each outcome has probability $\frac{1}{6}$.
- Toss a coin: $W = \{h, t\}$. $P(h) = P(t) = \frac{1}{2}$.

But what if the coin is tossed twice? What is the probability of observing heads twice ?

option 1 $W = \{\{2h\}, \{2t\}, \{h, t\}\}$. Probability $P(2h) = \frac{1}{3}$.

option 2 $W = \{(h, h), (h, t), (t, h), (t, t)\}$. Probability $P((h, h)) = \frac{1}{4}$.

2 seems to be in better agreement with experimental results.
When constructing W, keep as much information about outcomes as possible?
How much is enough ? What if the experiment can't really be repeated ?

Justifying probability for describing randomness

Question: why should we use probability with the properties

P1 $P(W) = 1$

P2 If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.

instead of

- Dempster-Shafer belief functions
- Ranking functions
- Relative likelihood
- Plausibility measures
- ....?

This question is relevant, since a **HUGE** amount of effort is expended on e.g. developing machine learning techniques based on probability.

Justifying probability for describing randomness

Question: why should we use probability with the properties

P1 $P(W) = 1$

P2 If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.

instead of

- Dempster-Shafer belief functions
- Ranking functions
- Relative likelihood
- Plausibility measures
- ....?

This question is relevant, since a **HUGE** amount of effort is expended on e.g. developing machine learning techniques based on probability.

Justifying probability for describing randomness

Question: why should we use probability with the properties

P1 $P(W) = 1$

P2 If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.

This question is relevant, since a **HUGE** amount of effort is expended on e.g. developing machine learning techniques based on probability.

Many justifications possible, we consider one made by F. Ramsey based on *rational betting behavior*: when deciding between two bets, a rational agent's preference can be stated as a thresholding function. Ramsey showed that this threshold behaves like a probability.

Complementary bets

Ramsey (and Halpern, 2003) considered bets of the form:

Bet (U, α): If $U \subseteq W$ happens (i.e. the actual world $w \in U$), then I *win* $X(1 - \alpha)$, otherwise I *lose* $X\alpha$. X is some large amount of money.

If I had to chose between different bets e.g. $(U, \alpha)$and (V, β), which one would I prefer? Write

$$(U, \alpha) \succeq (V, \beta)$$

if I like (U, α)at least as much (but possible more) than (V, β). I am trying to bet rationally. Ramsey argues that means that I satisfy 4 *rationality properties*:

Complementary bets

Ramsey (and Halpern, 2003) considered bets of the form:

Bet (U, α): If $U \subseteq W$ happens (i.e. the actual world $w \in U$), then I *win* $X(1-\alpha)$, otherwise I *lose* $X\alpha$. X is some large amount of money.

If I had to chose between different bets e.g. $(U, \alpha)$and (V, β), which one would I prefer? Write

$$(U, \alpha) \succeq (V, \beta)$$

if I like (U, α)at least as much (but possible more) than (V, β).

I am trying to bet rationally. Ramsey argues that means that I satisfy 4 *rationality properties*:

Complementary bets

Ramsey (and Halpern, 2003) considered bets of the form:

Bet (U, α): If $U \subseteq W$ happens (i.e. the actual world $w \in U$), then I *win* $X(1-\alpha)$, otherwise I *lose* $X\alpha$. X is some large amount of money.

If I had to chose between different bets e.g. $(U, \alpha)$and (V, β), which one would I prefer? Write

$$(U, \alpha) \succeq (V, \beta)$$

if I like (U, α)at least as much (but possible more) than (V, β). I am trying to bet rationally. Ramsey argues that means that I satisfy 4 *rationality properties*:

Rationality property 1: partial ordering of bets

RAT1: If $(U, \alpha)$is guaranteed to yield at least as much money as (V, β), then $(U, \alpha) \succeq (V, \beta)$.

Furthermore, if $(U, \alpha)$is guaranteed to yield more money than (V, β), then $(U, \alpha) \succ (V, \beta) \Leftrightarrow (U, \alpha) \succeq (V, \beta), (V, \beta) \not\succeq (U, \alpha)$.

Considering sets of bets B_1, B_2:

RAT1.1: If B_1 is guaranteed to yield at least as much money as B_2, then $B_1 \succeq B_2$.

If B_1 is guaranteed to yield more money than B_2, then $B_1 \succ B_2 \Leftrightarrow B_1 \succeq B_2, B_2 \not\succeq B_1$.

In other words, if the sum of payoffs in B_1 is greater than the sum of payoffs in B_2, then I strictly prefer B_1 to B_2.

Rationality property 1: partial ordering of bets

RAT1: If $(U, \alpha)$is guaranteed to yield at least as much money as (V, β), then $(U, \alpha) \succeq (V, \beta)$.

Furthermore, if $(U, \alpha)$is guaranteed to yield more money than (V, β), then $(U, \alpha) \succ (V, \beta) \Leftrightarrow (U, \alpha) \succeq (V, \beta), (V, \beta) \not\succeq (U, \alpha)$.

Considering sets of bets B_1, B_2:

RAT1.1: If B_1 is guaranteed to yield at least as much money as B_2, then $B_1 \succeq B_2$.

If B_1 is guaranteed to yield more money than B_2, then $B_1 \succ B_2 \Leftrightarrow B_1 \succeq B_2, B_2 \not\succeq B_1$.

In other words, if the sum of payoffs in B_1 is greater than the sum of payoffs in B_2, then I strictly prefer B_1 to B_2.

Rationality property 2: transitivity of partial order

RAT2: If $(U, \alpha) \succeq (V, \beta)$, and $(V, \beta) \succeq (Q, \gamma)$, then $(U, \alpha) \succeq (Q, \gamma)$.

Considering sets of bets B_1, B_2, B_3:
RAT2.1: If $B_1 \succeq B_2$, and $B_2 \succeq B_3$, then $B_1 \succeq B_3$.

Rationality property 2: transitivity of partial order

RAT2: If $(U, \alpha) \succeq (V, \beta)$, and $(V, \beta) \succeq (Q, \gamma)$, then $(U, \alpha) \succeq (Q, \gamma)$.

Considering sets of bets B_1, B_2, B_3:
RAT2.1: If $B_1 \succeq B_2$, and $B_2 \succeq B_3$, then $B_1 \succeq B_3$.

Rationality property 3: comparability of complementary bets

Bet (U, α): If U happens, then I *win* $X(1-\alpha)$, else I *lose* $X\alpha$.
Bet $(\overline{U}, 1-\alpha)$: If $\overline{U}$ happens, then I *win* $X\alpha$, else I *lose* $X(1-\alpha)$.

$(U, \alpha)$and $(\overline{U}, 1-\alpha)$are *complementary* bets .

RAT3: Either $(U, \alpha) \succeq (\overline{U}, 1-\alpha)$, or $(\overline{U}, 1-\alpha) \succeq (U, \alpha)$.

In other words, complementary bets are always comparable.

Point-wise determination of preferences

RAT4: If $(U_i, \alpha_i) \succeq (V_i, \beta_i)$ for all $i = 1, \ldots, k$, then $\{(U_i, \alpha_i)\} \succeq \{(V_i, \beta_i)\}$

Justification of probability

Theorem: If I satisfy RAT1-RAT4, the for each $U \subseteq W$, a number α_U exists such that $(U, \alpha) \succeq (\overline{U}, 1-\alpha)$ for $\alpha < \alpha_U$ and $(\overline{U}, 1-\alpha) \succeq (U, \alpha)$ for $\alpha > \alpha_U$. Furthermore, the function defined by $P(U) = \alpha_U$ is a probability measure.

Notes

- $P(U) = \alpha_U$ fulfills P1 $(\alpha_W = 1)$ and P2 (disjoint additivity).
- The smaller α, the bigger the potential payoff from (U, α). Thus, choose (U, α) for small α, and $(\overline{U}, 1-\alpha)$ for large α.
- The break-even point, i.e. when I have no preference for either (U, α) and $(\overline{U}, 1-\alpha)$, is reached for α_U, which represents my uncertainty about U.

Conditioning

Suppose your uncertainty is represented by $P : \mathcal{F} \to [0,1]$. Now you learn that the event $U \in \mathcal{F}$ has happended (alternatively: that a possible world $w \in U$ is the real one).

Question: how should P be updated to reflect this information?

Let $P|U$ be the updated probability measure.

It seems reasonable to require that

C1 $P|U(\bar{U}) = 0$, since we learned that $w \in U$.

C2 If $V_1, V_2 \subseteq U$: $\frac{P(V1)}{P(V2)} = \frac{P|U(V_1)}{P|U(V_2)}$, if all we have learned is that U is certain.

Conditioning

Let $P|U$ be the update probability measure. It seems reasonable to require that

C1 $P|U(\bar{U}) = 0$, since we learned that $w \in U$.

C2 If $V_1, V_2 \subseteq U$: $\frac{P(V1)}{P(V2)} = \frac{P|U(V_1)}{P|U(V_2)}$, if all we have learned is that U is certain.

Proposition: If $P(U) > 0$ and $P|U$ is a probability measure on $\mathcal{F}$ which fulfills C1 and C2, then

$$P|U(V) = \frac{P(V \cap U)}{P(U)}.$$

It is customary to write $P(V|U)$ for $P|U(V)$.

Justification of conditioning

Conditioning can be justified by a betting argument, very much like probability. Consider the

Bet $(V|U, \alpha)$: if U happens, then if V also happens, then I win $X(1-\alpha)$, while if $\overline{V}$ also happens, then I lose $X\alpha$. If U does not happen, then the bet is called off.

It is then possible to proof the

Theorem: If I satisfy RAT1-RAT4, then for all $U, V \subseteq W$ such that $\alpha_U > 0$, a number $\alpha_{V|U}$ exists such that $(V|U, \alpha) \succeq (\overline{V}|U, 1-\alpha)$ for $\alpha < \alpha_{V|U}$ and $(\overline{V}|U, 1-\alpha) \succeq (V|U, \alpha)$ for $\alpha > \alpha_{V|U}$. Furthermore, $\alpha_{V|U} = \frac{\alpha_{V \cap U}}{\alpha_U}$.

In other words, if I want to behave rationally, I must use probabilistic conditioning.

Important consequence: Bayes' Rule

Bayes' Rule

$$P(U|V) = P(V|U)\frac{P(U)}{P(V)}$$

Proof: assume $P(U), P(V) > 0$. Since $P(V|U) = \frac{P(V \cap U)}{P(U)}$, we find $P(V \cap U) = P(V|U)P(U)$.

Thus $P(U|V) = \frac{P(U \cap V)}{P(V)} = P(V|U)\frac{P(U)}{P(V)}$ $\square$

Take-home messages, part 1

- Representations of uncertainty are usually built upon the concept of the *possible world* or *elementary outcome*
- *Events* or *propositions* are subsets of the possible worlds.
- It is possible to construct a simple representation of uncertainty via set operations alone. Very coarsely grained: certainty,possibility,impossibility.
- *Probability* is a much more finely grained representation of uncertainty: event is assigned a real number $[0, 1]$.
- Not all possible events need to be assigned a probability, only those that are members of a given σ-algebra.

Take-home messages, part 1

- Representations of uncertainty are usually built upon the concept of the *possible world* or *elementary outcome*
- *Events* or *propositions* are subsets of the possible worlds.
- It is possible to construct a simple representation of uncertainty via set operations alone. Very coarsely grained: certainty,possibility,impossibility.
- *Probability* is a much more finely grained representation of uncertainty: event is assigned a real number $[0, 1]$.
- Not all possible events need to be assigned a probability, only those that are members of a given σ-algebra.

Take-home messages, part 2

- Probability is the only rational way of representing uncertainty, given that you accept RAT1-4.
- *Conditioning* is the process of updating your probabilities if new information arrives.
- If the information is in the form 'the real world is in U' or alternatively 'U just happended', then conditioning is done via $P(V|U) = \frac{P(U \cap V)}{P(U)}$.
- Accepting RAT1-4 for conditional bets forces you to accept this form of conditioning.
- Conditioning can be inverted via *Bayes' Rule*: $P(U|V) = P(V|U)\frac{P(U)}{P(V)}$.

Take-home messages, part 2

- Probability is the only rational way of representing uncertainty, given that you accept RAT1-4.
- *Conditioning* is the process of updating your probabilities if new information arrives.
- If the information is in the form 'the real world is in U' or alternatively 'U just happended', then conditioning is done via $P(V|U) = \frac{P(U \cap V)}{P(U)}$.
- Accepting RAT1-4 for conditional bets forces you to accept this form of conditioning.
- Conditioning can be inverted via *Bayes' Rule*: $P(U|V) = P(V|U)\frac{P(U)}{P(V)}$.

Take-home messages, part 2

- Probability is the only rational way of representing uncertainty, given that you accept RAT1-4.
- *Conditioning* is the process of updating your probabilities if new information arrives.
- If the information is in the form 'the real world is in U' or alternatively 'U just happended', then conditioning is done via $P(V|U) = \frac{P(U \cap V)}{P(U)}$.
- Accepting RAT1-4 for conditional bets forces you to accept this form of conditioning.
- Conditioning can be inverted via *Bayes' Rule*: $P(U|V) = P(V|U)\frac{P(U)}{P(V)}$.