The statistics of belief functions

4th BFAS Summer School on Belief Functions and their Applications

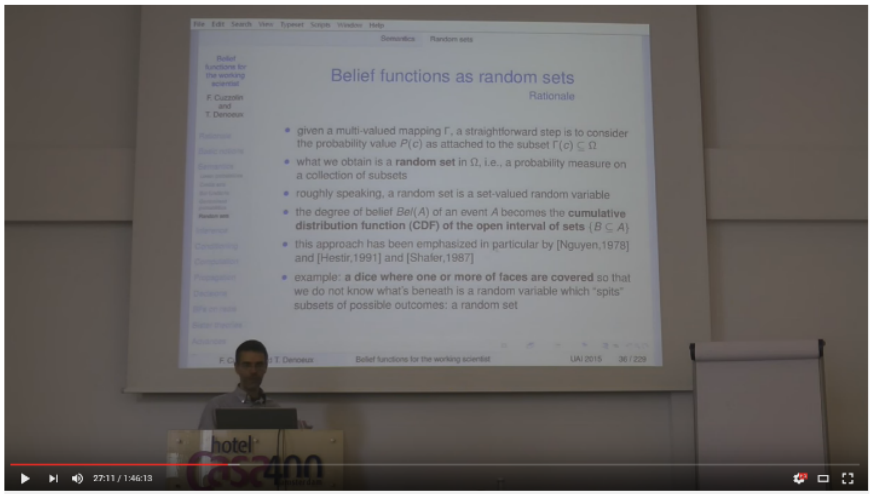
Professor Fabio Cuzzolin

Department of Computing and Communication Technologies
Oxford Brookes University, Oxford, UK

Xi'an, China, July 6 2017

Tutorial web site

IJCAI 2016 Tutorial: Belief functions for the Working scientist



UAI 2015 Amsterdam Tutorial: Belief Functions for the Working Scientist

Previous UAI 2015 tutorial on YouTube

CURRENT VERSION OF IJCAI 2016 PDF SLIDES (Work in progress)

All fields of Artificial Intelligence and applied science are subject to various degrees of uncertainty, caused by missing or scarce data: random sets and their subjective incarnations called **belief functions** naturally arise when one tries to formalize lack of data in a coherent way. This tutorial introduces to the wider AI audience the basic principles of the theory of belief functions, describes practical tools based on them and compares their performance against more classical approaches to uncertainty on cutting edge real-world problems.

The theory of belief functions, sometimes referred to as evidence theory or Dempster-Shafer theory, was first introduced by Arthur P. Dempster in the context of statistical inference, and was later developed by Glenn Shafer as a general framework for modelling epistemic uncertainty. Belief theory and the closely related random set theory form a natural framework for modelling situations in which data are missing or scarce: think of extremely rare events such as volcanic eruptions or power plant meltdowns, problems subject to huge uncertainties due to the number and complexity of the factors involved (e.g. climate change), but also the all-important issue with generalisation from small training sets in machine learning.

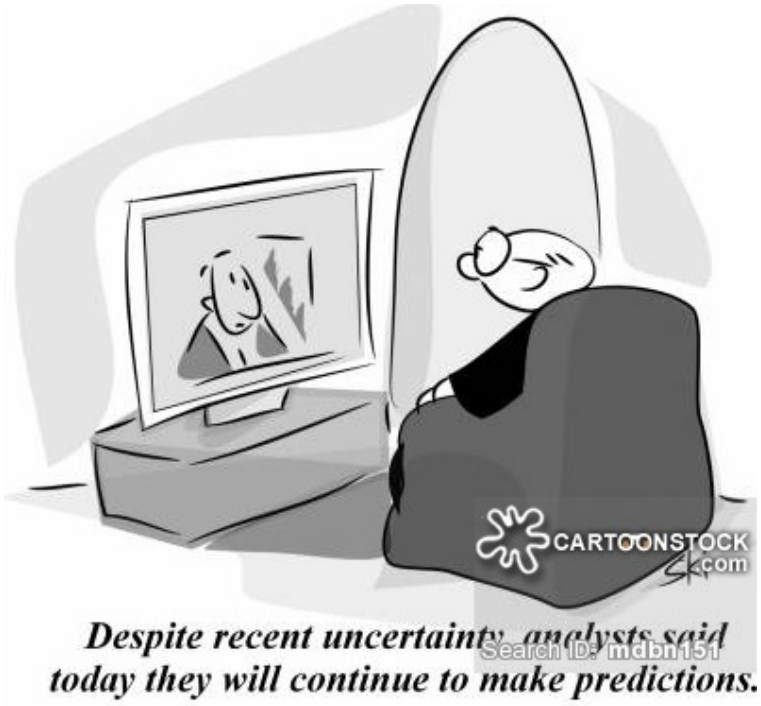
This tutorial is designed to introduce the principles and rationale of random sets and belief function theory to the wider AI audience, survey the key elements of the methodology and the most recent developments, make AI practitioners aware of the set of tools that have been developed for reasoning in the belief function framework on real-world problems.

- http://cms.brookes.ac.uk/staff/FabioCuzzolin/ijcai2016.html

Outline

1. Uncertainty and probability
   - Nature of uncertainty
   - Mathematical probability
   - Interpretations of probability
   - Bayesian interpretation
   - Bayesians vs frequentists
2. Beyond probability
   - Missing data
   - Propositional data
   - Scarce data
   - Pure or no data
   - Rare data
   - Uncertain data
3. Statistics of belief functions
   - Dempster's multi-valued setting
   - Monotone capacities
   - Inner measures
   - Conditional multi-valued mappings
   - Statistics of experts' opinions
   - Probabilistic multivalued RVs
   - Gaussian belief functions
   - Measure-theoretical analysis
   - Frequentist interpretations
   - Dempster's a-equations
   - Fiducial inference and weak beliefs
   - Random set theory
4. New horizons
   - Upper and lower likelihood
   - Generalising logistic regression
   - Frequentist inference with RS
   - Generalised laws
   - The total belief theorem
   - Random set random variables
5. Summarising

What is uncertainty?



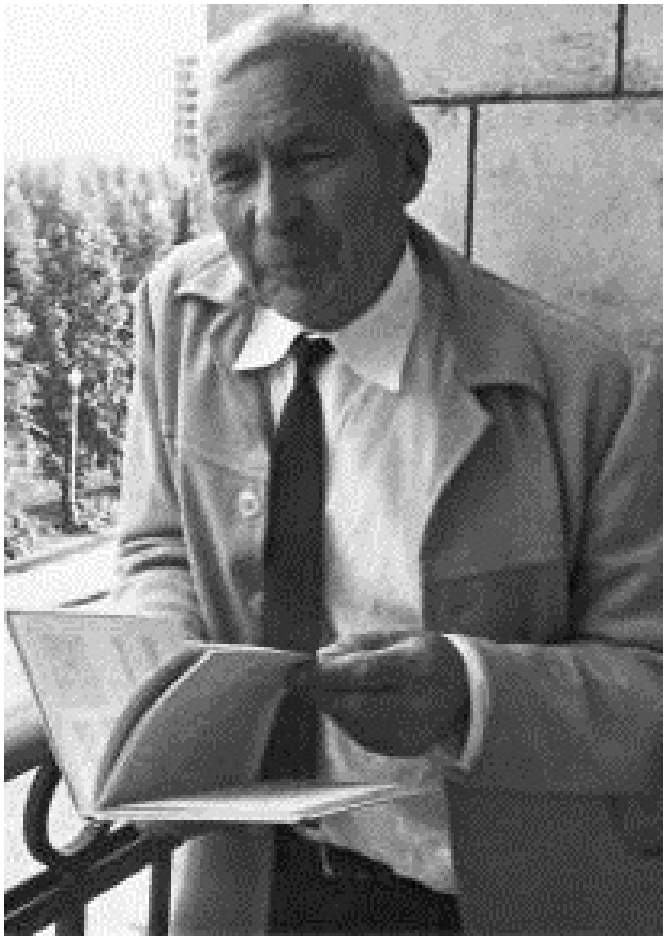
- **uncertainty** $\rightarrow$ lack of information or imperfect information
- a state of limited knowledge, where it is impossible to exactly describe the existing state or future outcomes

'Knightian' Uncertainty



- the difference between **predictable** and **unpredictable** variation is one of the fundamental issues in the philosophy of probability
- 'second order' uncertainty: being uncertain about our very model of uncertainty
- has a consequence on human behaviour: people are averse to unpredictable variations (cfr. *Ellsberg's paradox* in Decision making)
- Chicago economist Frank Knight: "**You cannot be certain about uncertainty**"

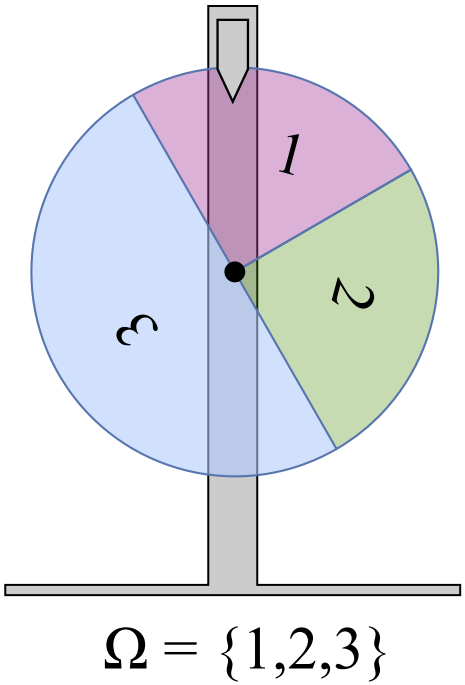
Probability measures

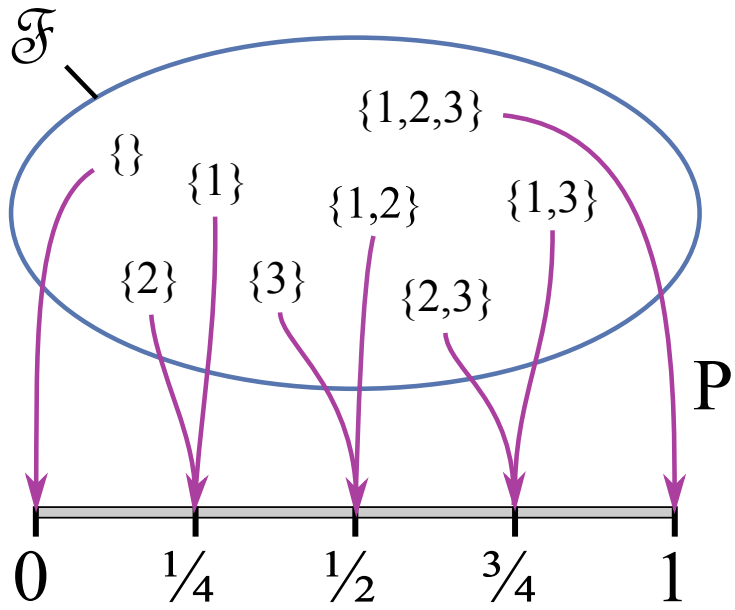


- mainstream mathematical theory of (first order) uncertainty: **mathematical (measure-theoretical) probability**
- mainly due to Russian mathematician Andrey Kolmogorov
- probability is an application of **measure theory**, the theory of assigning numbers to sets
- **additive** probability measure $\rightarrow$ mathematical representation of the notion of chance
- assigns a probability value to every subset of a collection of possible outcomes (of a random experiment, of a decision problem, etc)
- collection of outcomes $\Omega \rightarrow$ sample space, universe
- subset A of the universe $\rightarrow$ **event**

Example: the spinning wheel

Probability measures





- typical example: spinning wheel with 3 possible outcomes
- universe $\Omega = \{1, 2, 3\}$
- eight possible events (right), including the empty set
- probability of $\emptyset$ is 0, probability of Ω is 1
- **additivity** holds: $P(\{1, 2\}) = P(\{1\}) + P(\{2\})$

Probability measures

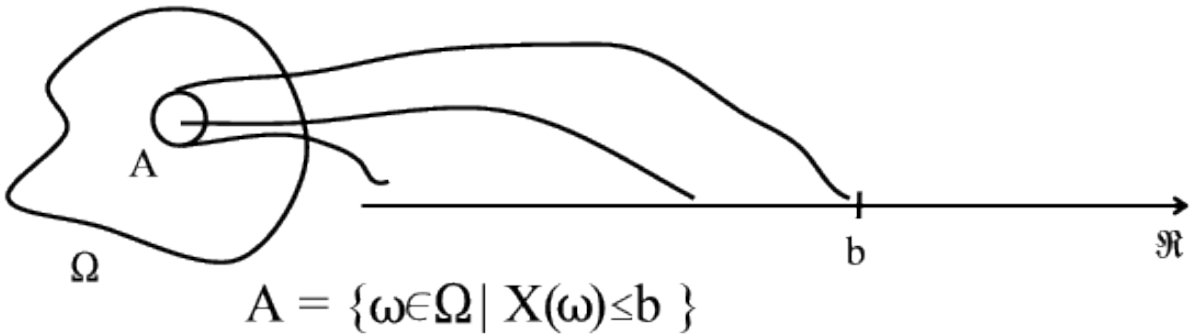
Formal definition

- **probability measure** μ: a real-valued function on a probability space that satisfies countable additivity
- **probability space**: it is a triplet $(\Omega, \mathcal{F}, P)$ formed by a universe Ω, a σ-algebra $\mathcal{F}$ of its subsets, and a probability measure on $\mathcal{F}$
  - not all subsets of Ω belong necessarily to $\mathcal{F}$
- axioms of probability measures:
  - $\mu(\emptyset) = 0$, $\mu(\Omega) = 1$
  - $0 \leq \mu(A) \leq 1$ for all events $A \subseteq \mathcal{F}$
  - additivity: for all countable collection of pairwise disjoint events A_i:

$$\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i)$$

Random variable

- a variable whose value is subject to random variations, i.e. due to 'chance': what chance is is subject to philosophical debate!
- it can take one of a set of possible values, with probability
- mathematically, it is a function X from a sample space Ω (which forms a probability space) to (usually) the real line

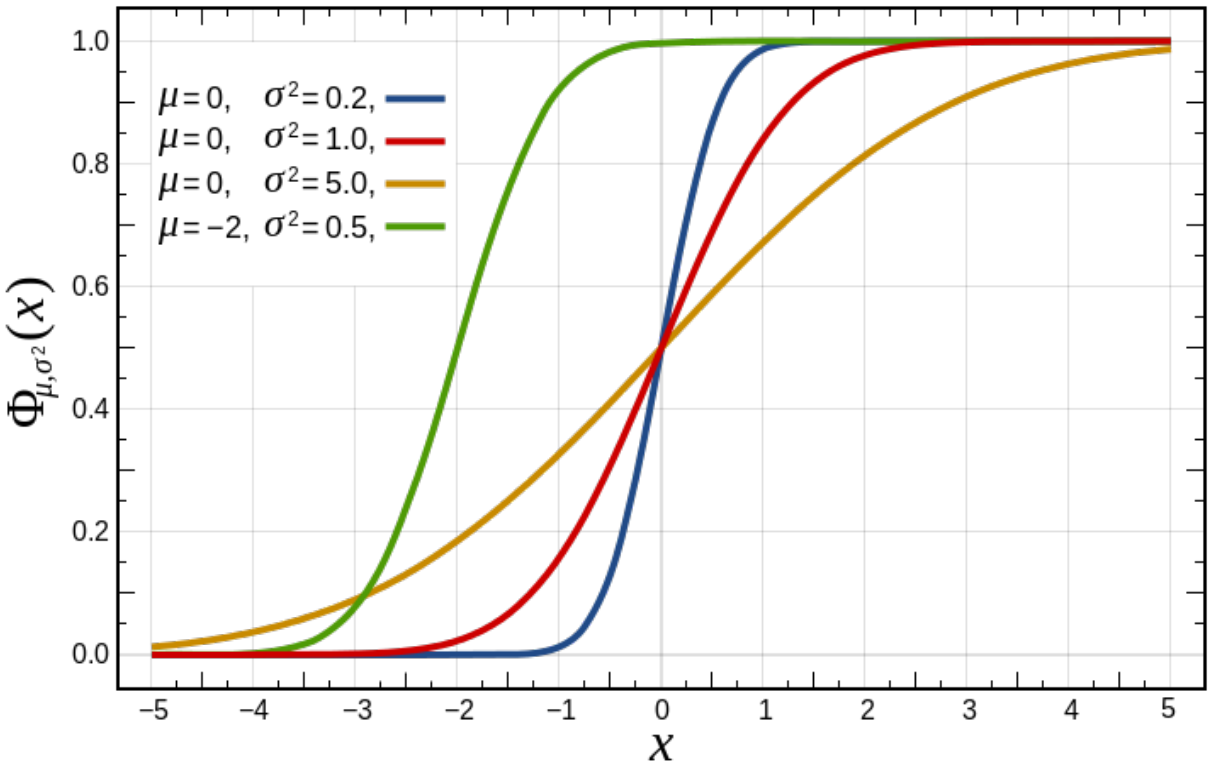


- subject to a condition of **measurability**: each range of values of the real line must have an anti-image in Ω which has a probability value
- this way, we can forget about the initial probability space and record the probabilities of various values of X

Cumulative Distribution Function (CDF)

of a random variable

- $F(x) = P(X \leq x)$

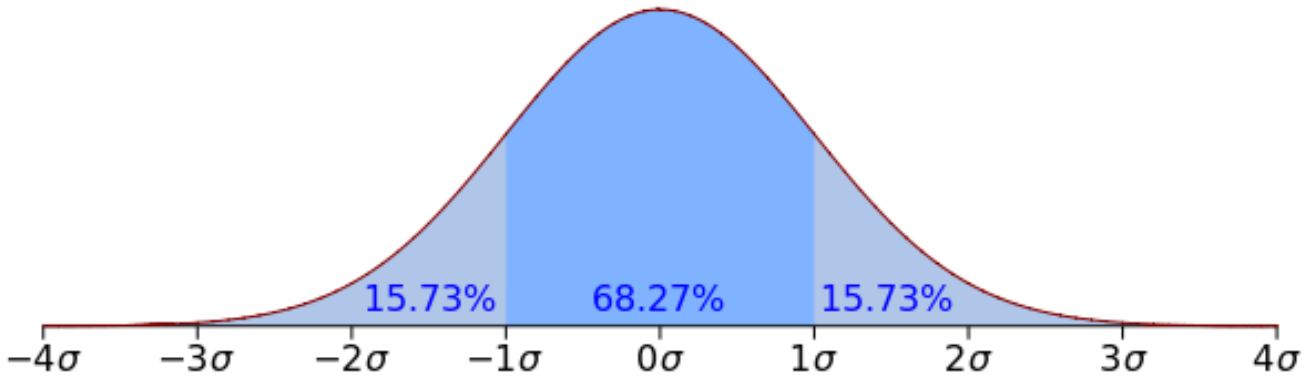


- example: CDF of a Gaussian random variable

Probability Density Function (PDF)

of a continuous random variable

- a random variable is called **continuous** when it can assume values in a non-countable set (e.g. the real line)
- it is described by a **probability density function (PDF)**, which describes the likelihood of the variable taking any continuous value
- the probability of any range of values (e.g., an interval) is the integral of the PDF over the range: $P([a,b]) = \int_a^b f(x)dx$



- example: PDF of a Gaussian random variable

Radon-Nikodym derivative

Measure-theoretic probability theory

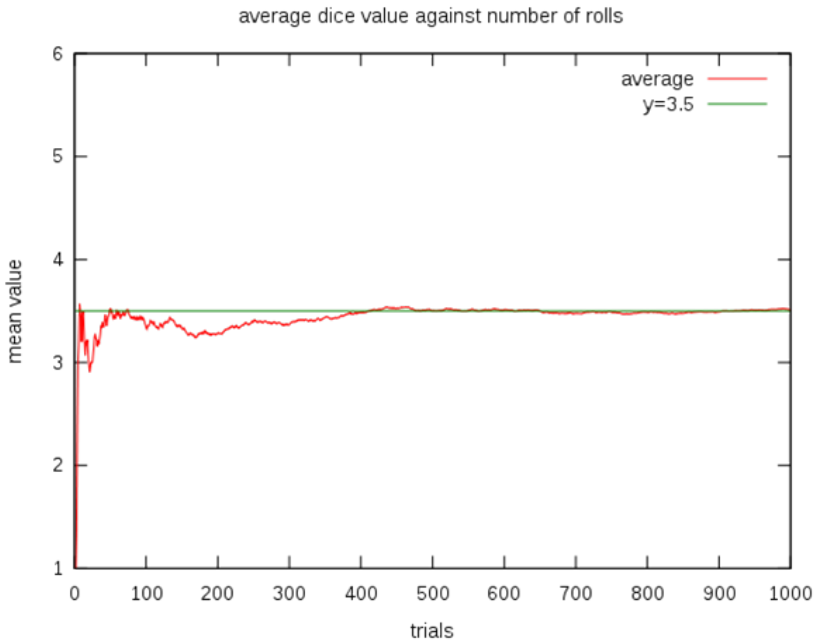
- a continuous random variable with values in a measurable space $(\mathcal{X}, \mathcal{A})$ (usually $\mathbb{R}^n$ with the Borel sets as measurable family) has as probability distribution the measure $X_* P$ on $(\mathcal{X}, \mathcal{A})$
- formally, the PDF of X is the **Radon-Nikodym derivative**, denoted:

$$f = \frac{dX_* P}{d\mu}$$

  where μ is a reference measure on $(\mathcal{X}, \mathcal{A})$
- that is, f is any measurable function such that: $P(X \in A) = \int_A f d\mu$
  - analogous to a derivative in calculus
- modern mathematical probability is really just **an application of measure theory**

Law of large numbers

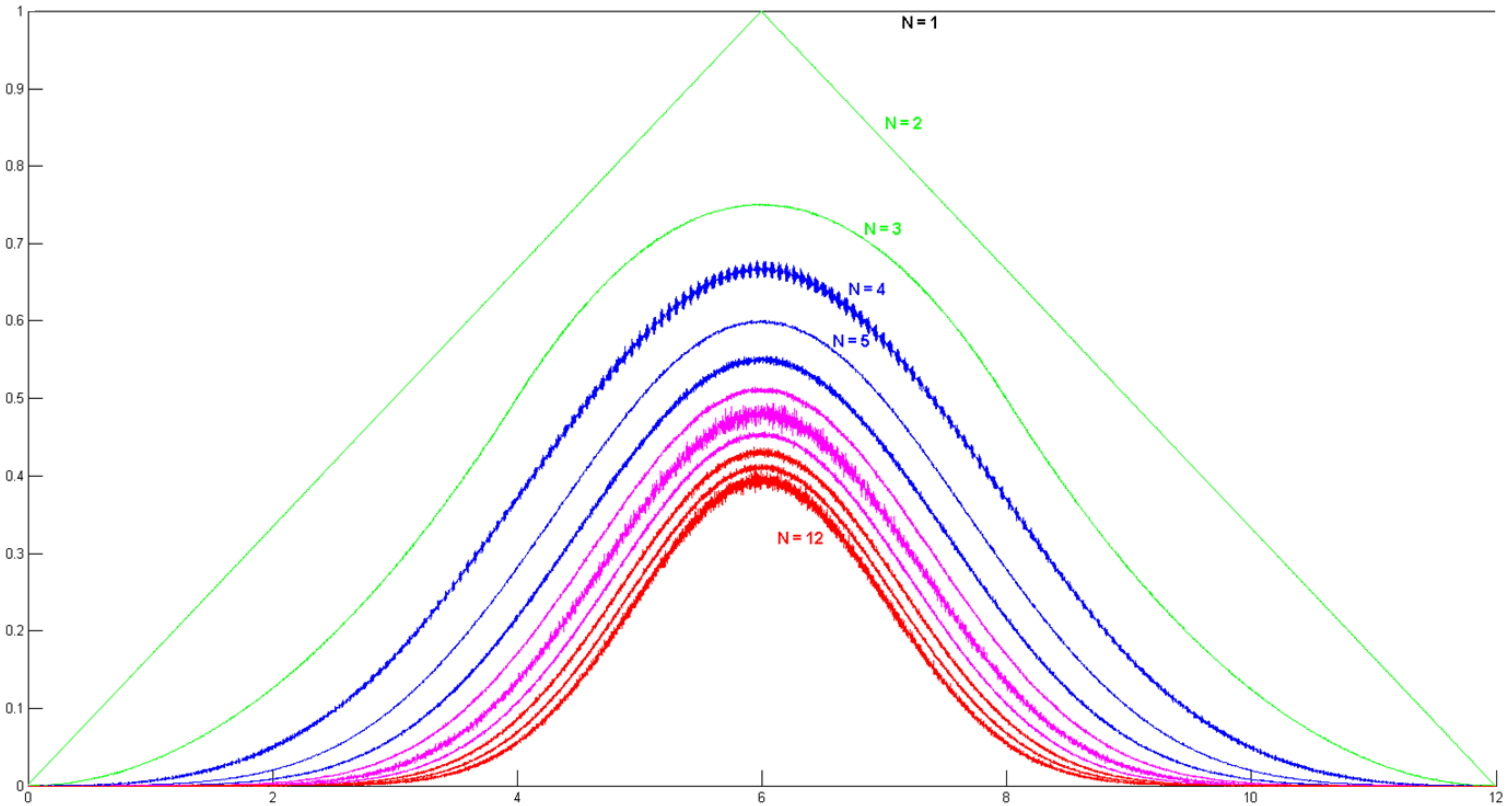
- describes what happens when you repeat the same **random experiment** (extracting a sample from a random variable) an increasing number of times n
- the average of the results (**sample mean**) $\overline{X}_n = \dfrac{X_1 + \ldots + X_n}{n}$ should be close to the expected value (actual mean)
- (**estimator** $\rightarrow$ a statistic (that is, a function of the data) used to infer the value of an unknown parameter in a statistical model)
- **consistent**: converges in probability to the quantity being estimated for $n \rightarrow \infty$



- probabilities become predictable as we run the same trial more and more times!
- comes in two forms:
- **strong law** of large numbers: $P(\lim_{n\rightarrow\infty} \overline{X}_n = \mu) = 1$
- **weak law** of large numbers: $\lim_{n\rightarrow\infty} P(|\overline{X}_n - \mu| > \epsilon) = 0$

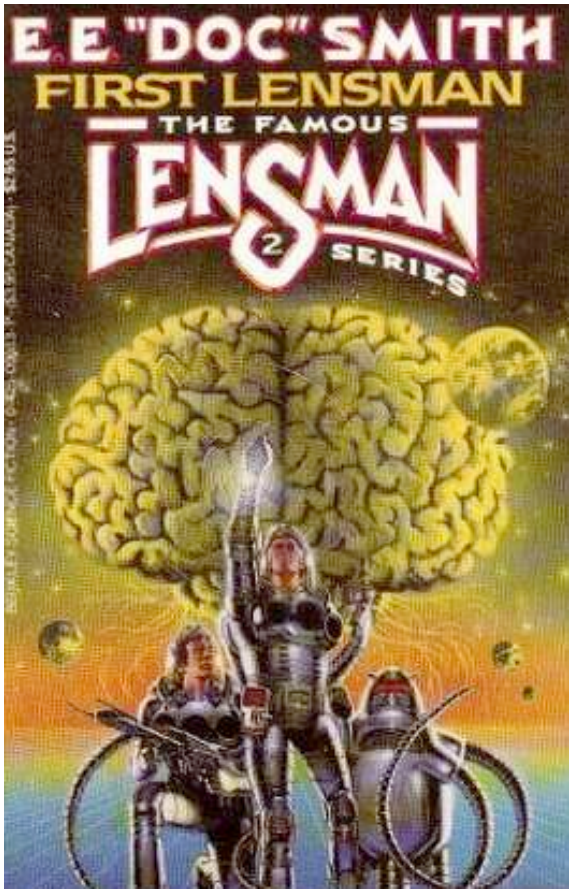
Central limit theorem

- let $X_1, ..., X_n$ independent and identically distributed random variables with the same mean μ and variance σ^2
- we can build the sample average as $\overline{X}_n = \dfrac{X_1 + ... + X_n}{n}$
- the random variable $\sqrt{n}(\overline{X}_n - \mu)$ **tends to a Gaussian** with mean 0 and var σ^2



Does probability really exist?

That sinking feeling



- what is probability, really?
- is it just the way we call our ignorance?
- can it be that, with sufficient information, any phenomenon is predictable in a deterministic way?
- Einstein: God does not roll dice
- Doc Smith's Lensmen series: the Arisians have such mental powers that they compete on foreseen future events to the tiniest detail
- the principles of quantum mechanics seems to suggest that probability is not just a figment of our mathematical imagination

Interpretations of probability

Savage's take



- even assuming that probability is inherent to the physical world, people cannot agree on what it is
- "*It is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel. Doubtless, much of the disagreement is merely terminological and would disappear under sufficiently sharp analysis.*"
- L.J. Savage, 1954

Interpretations of probability

Frequentist, subjective and behavioural

- as a result, probability has multiple competing interpretations
- an objective description of frequencies of events (meaning 'things that happen') at a certain persistent rate, or 'relative frequency' $\rightarrow$ **frequentist** interpretation [Fisher, Pearson]
- a degree of belief on events (interpreted as statements/propositions on the state of the world), regardless of any random process $\rightarrow$ **Bayesian or evidential** probability [de Finetti, Savage]
  - neither frequentist nor Bayesian probability are in constrast with the classical mathematical definition of probability - others are (as we will see)
- the propensity of an agent to act (or gamble, or decide) in case the event happens $\rightarrow$ **behavioural** probability [Walley, Vovk]

Frequentist probability

- the (**aleatory**) **probability** of an event is its **relative frequency in time**
- when tossing a fair coin, frequentists say that the probability of getting a heads is 1/2, because repeated series of large numbers of trials demonstrate that the empirical frequency converges to 1/2 as the number of trials goes to infinity:

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n} \neq \frac{N_A}{N}$$

  where n is the number of trials, n_A those resulting in A
- impossible to actually perform an infinity of repetitions of a random experiment
- what are the consequences for inference? the frequentist interpretation offers guidance in the design of practical 'random' experiments [Fisher, Pearman, Neyman]
- main tool: **statistical hypothesis testing**, **maximum likelihood estimation**

Statistical hypothesis testing

Procedure

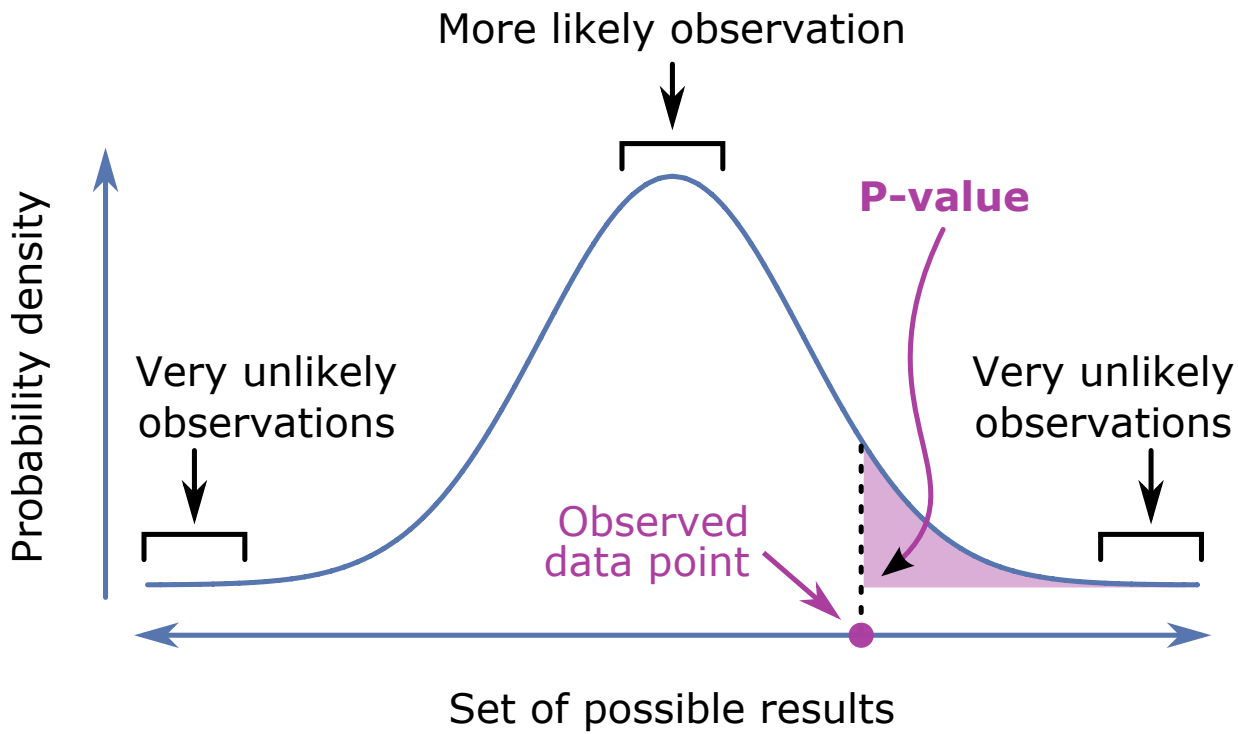
1. state the research hypothesis
2. state the relevant null and alternative hypotheses
3. state the statistical assumptions being made about the sample, e.g. assumptions about the statistical independence or about the form of the distributions of the observations
4. state the relevant **test statistic** T (a quantity derived from the sample)
5. derive the distribution of the test statistic under the null hypothesis from the assumptions
6. set a significance level (α), i.e. a probability threshold below which the null hypothesis will be rejected
7. compute from the observations the observed value t_{obs} of the test statistic T
8. calculate the **p-value**, the probability (under the null hypothesis) of sampling a test statistic at least as extreme as the observed value
9. Reject the null hypothesis, in favor of the alternative hypothesis, if and only if the p-value is less than the significance level threshold

Statistical hypothesis testing

Interpretation

- given the observed data
- assuming a (parameterised) probability distribution generating them, of which we do not know the parameter value
- we test hypotheses on the value of the parameter
- output: yes or no (binary decision)
- interpretation: if the p-value is less than the required significance level the null hypothesis is rejected
- if not, the test has no result - the evidence is insufficient to support a conclusion
- a reductio ad absurdum argument adapted to statistics: a claim is shown to be valid by demonstrating the improbability of the consequence of the opposite
- statistical hypothesis testing is a form of **confirmatory** data analysis, as opposed to **exploratory** data analysis which does not rely on pre-specified hypotheses

P-values



- **the p-value is not the probability that the null hypothesis is true or the probability that the alternative hypothesis is false**: frequentist statistics does not and cannot attach probabilities to hypotheses

Maximum Likelihood Estimation (MLE)

- the term 'likelihood' was popularized in mathematical statistics by Ronald Fisher in 1922: 'On the mathematical foundations of theoretical statistics'
- given a *parametric model* (i.e., a family probability distributions $\{f(.|\theta), \theta \in \Theta\}$ of the data x conditional on a parameter $\theta \in \Theta$) ..
- .. the **likelihood function** is

$$\mathcal{L}(\theta\,;\, x_1, \ldots, x_n) = f(x_1, x_2, \ldots, x_n \mid \theta)$$

- **likelihood principle**: all of the evidence in a sample relevant to model parameters is contained in the likelihood function
  - this is hotly debated, still [Mayo,Gandenberger]
- Fisher argued against 'inverse' (Bayesian) probability as a basis for statistical inferences, and instead proposed inferences based on likelihood functions
- **maximum likelihood** estimation:

$$\{\hat{\theta}_{\text{mle}}\} \subseteq \left\{ \arg\max_{\theta \in \Theta} \mathcal{L}(\theta\,;\, x_1, \ldots, x_n) \right\}$$

Properties of Maximum Likelihood

- maximum-likelihood estimators have no optimum properties for finite samples
- however, they do have good *limiting* properties:
  - **consistency**: the sequence of MLEs converges in probability, for a sufficiently large number of observations, to the (actual) value being estimated
  - **asymptotic normality**: as the sample size increases, the distribution of the MLE tends to the Gaussian distribution with mean on the true parameter (under a number of conditions)
  - **efficiency**: it achieves the *Cramer-Rao lower bound* when the sample size tends to infinity, i.e. no consistent estimator has lower asymptotic mean squared error than the MLE

Bayesian probability



- (**epistemic**) **probability** = degrees of belief of an individual assessing the state of the world
- in the **Bayesian** view, a probability is assigned to a hypothesis, whereas under frequentist inference, a hypothesis is typically tested without being assigned a probability
- special case of **evidential** probability: some prior probability is updated to a posterior probability in the light of new evidence (data)
- once again it makes use of mathematical probabilities
- needs to specify a **prior** probability distribution taking into account the available (prior) information
- sequentially uses **Bayes' rule** to compute a **posterior** distribution when more data becomes available

Bayesian inference

- prior distribution of the parameter(s) before any data is observed: $p(\theta)$
- likelihood: distribution of the observed data $X = \{x_1, ..., x_N\}$ conditional on its parameters: $p(X|\theta)$
- marginal likelihood: $p(X) = \int_\theta p(X|\theta)p(\theta)\,\mathrm{d}\theta$
- **posterior distribution**: distribution of the parameter(s) after taking into account the observed data, as determined by **Bayes' rule**

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \propto p(X|\theta)p(\theta)$$

- *posterior predictive distribution* is the distribution of a new data point, marginalized over the posterior:

$$p(\tilde{x} \mid X) = \int_\theta p(\tilde{x}|\theta)p(\theta|X)\,\mathrm{d}\theta$$

Maximum-A-Posteriori (MAP)

- in full Bayesian inference a distribution over possible data values is obtained, which takes into account the uncertainty in the value of the parameter
- in opposition, **maximum a posteriori** estimates θ as the mode of the posterior distribution:

$$\hat{\theta}_{\text{MAP}}(x) = \arg\max_{\theta} p(\theta|X) = \arg\max_{\theta} p(X|\theta)p(\theta)$$

- MAP and MLE estimates coincide when the prior $p(\theta)$ is uniform
- not very representative of Bayesian methods, as the latter are characterized by the use of distributions to draw inferences
- also, unlike ML estimators, the MAP estimate is not invariant under reparameterization

Dutch book arguments



- de Finetti: a **Dutch book** is made when a clever gambler places a set of bets that guarantee a profit, no matter what the outcome of the bets.
- if a bookmaker follows the rules of the Bayesian calculus, a Dutch book cannot be made: subjective beliefs must follow the laws of probability if they are coherent (avoid sure loss)
- if this 'proof' was prooftight we would not be here!
- indeed, the argument starts by assuming that the price at which we 'sell' or 'buy' an event A are the same: hence, they assume that degrees of belief $Bel(A)$ and disbelief $Bel(A^c)$ add to 1
- also, evidence casts doubt that humans will have coherent beliefs or behave rationally
- a justification of Bayesian reasoning by axiomatisation has been tried, but with no great success

Lindley's paradox

Bayesian vs frequentist inference

- **Lindley's paradox** is a counterintuitive situation in statistics
- Bayesian and frequentist approaches to a hypothesis testing problem give different results for certain choices of the prior distribution
- not really a paradox – the two approaches answer fundamentally different questions
- Lindley's paradox[1] occurs when
  - the result x is 'significant' by a frequentist test of H_0, indicating sufficient evidence to reject H_0 say, at the 5% level, and
  - the posterior probability of H_0 given x is high, indicating strong evidence that H_0 is in better agreement with x than H_1
- this can happen when H_0 is very specific, H_1 less so, and the prior distribution does not strongly favor one or the other

[1] onlinelibrary.wiley.com/doi/10.1002/0470011815.b2a15076/pdf

Bayesian vs frequentist inference

- it is not like they are different ways of solving the same problem
- **they are really designed to solve different problems!**
- the result of a Bayesian approach can be a probability distribution on the parameters given the results of the experiment
- the result of a frequentist approach is either:
  - a 'true or false' (binary) conclusion from a significance test, or
  - a conclusion in the form that a given sample-derived confidence interval covers the true value
- either of these conclusions has a given probability of being correct

Outline

1 Uncertainty and probability
- Nature of uncertainty
- Mathematical probability
- Interpretations of probability
- Bayesian interpretation
- Bayesians vs frequentists

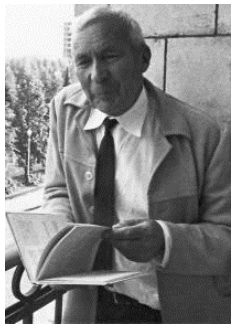
2 Beyond probability
- Missing data
- Propositional data
- Scarce data
- Pure or no data
- Rare data
- Uncertain data

3 Statistics of belief functions
- Dempster's multi-valued setting
- Monotone capacities
- Inner measures
- Conditional multi-valued mappings
- Statistics of experts' opinions
- Probabilistic multivalued RVs
- Gaussian belief functions
- Measure-theoretical analysis
- Frequentist interpretations
- Dempster's a-equations
- Fiducial inference and weak beliefs
- Random set theory

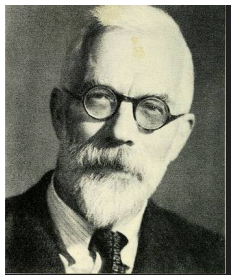
4 New horizons
- Upper and lower likelihood
- Generalising logistic regression
- Frequentist inference with RS
- Generalised laws
- The total belief theorem
- Random set random variables

5 Summarising

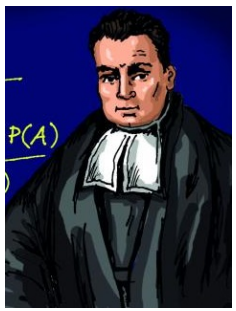
Something is wrong?



- **measure-theoretical mathematical probability** is not general enough:
  - cannot (properly) model missing data
  - cannot (properly) model propositional data
  - cannot really model unusual data (second order uncertainty)



- the **frequentist approach** to probability:
  - cannot really model pure data (without 'design')
  - in a way, cannot even model properly continuous data
  - models scarce data only asymptotically



- **Bayesian reasoning** has several limitations:
  - cannot model no data (ignorance)
  - cannot model 'uncertain' data
  - cannot model pure data (without prior)
  - again, cannot properly model scarce data (only asymptotically)

Fisher has not got it all right

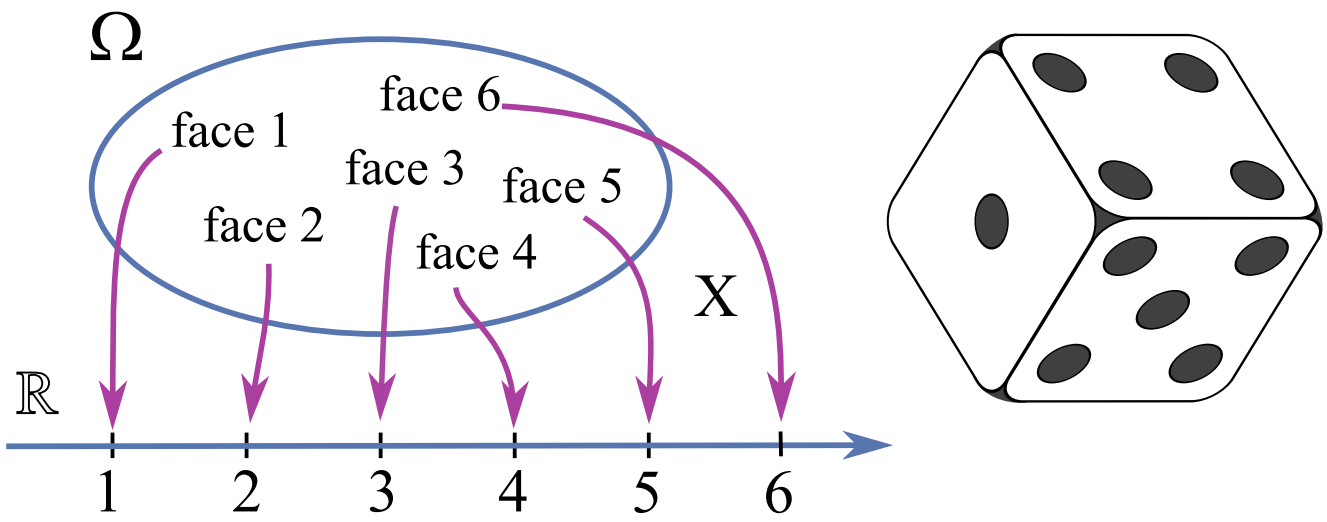
- the setting of hypothesis testing is (arguably) arguable
  - the **scope is quite narrow**: rejecting or not rejecting a hypothesis (although it can provide confidence intervals)
  - the **criterion is arbitrary**: who decides what an 'extreme' realisation is (choice of α)? what is the deal with 0.05 and 0.01?
  - the whole 'tail' idea comes from the fact that, under measure theory, the conditional probability (p-value) of a point outcome x is zero – seems trying to **patch an underlying problem** with the way probability is mathematically defined
- **cannot cope with pure data**, without assumptions on the process (experiment) which generated them (we will come back to this later)
- deals with scarce data only **asymptotically** (see 'scarce data')

The problem(s) with Bayes

- pretty **bad at representing ignorance**
  - Fisher uninformative priors are just not adequate
  - different results on different parameter spaces
- Bayes' rule assumes the new evidence **comes in the form of certainty**: "A is true"
  - in the real world, often this is not the case ('uncertain' or 'vague' evidence)
- **beware the prior!** $\rightarrow$ model selection in Bayesian statistics
  - results from a confusion between the original subjective interpretation, and the objectivist view of a rigorous objective procedure
  - why should we 'pick' a prior? either there is prior knowledge (beliefs) or there is not
  - all will be fine, in the end! asymptotically, the choice of the prior does not matter (really!)

The cloaked die

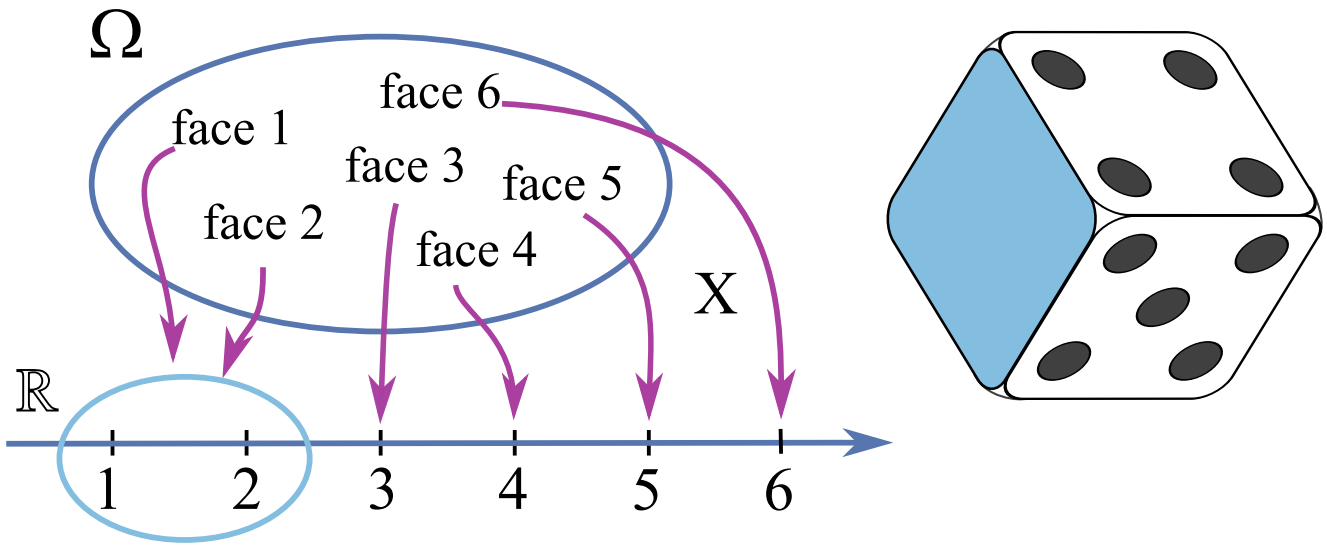
The die as random variable



- a die is a simple example of (discrete) random variable
- there is a probability space $\Omega = \{face1, face2, ..., face6\}$ which maps to a real number: $1, 2, ..., 6$ (no need for measurability here)

The cloaked die

Observations which are sets



- now, imagine that face1 and face2 are cloaked, and we roll the die
- the same probability space $\Omega = \{face1, face2, ..., face6\}$ is still there (nothing has changed in the way the die works)
- however, now the mapping is different: both face1 and face2 are mapped to the set of possible values $\{1, 2\}$ (since we cannot observe the outcome)

Dealing with missing data

with belief functions

- the mapping in the cloaked die example is what is called mathematically a **random set** [Matheron,Kendall,Nguyen], i.e. a set-valued random variable
- **whenever data are missing observations are inherently set-valued**
- how do belief functions cope with missing data?

Belief functions and missing data

Belief functions are (finite) random sets, hence they can naturally describe uncertainty induced by set-valued and missing observations.

Reliable witnesses

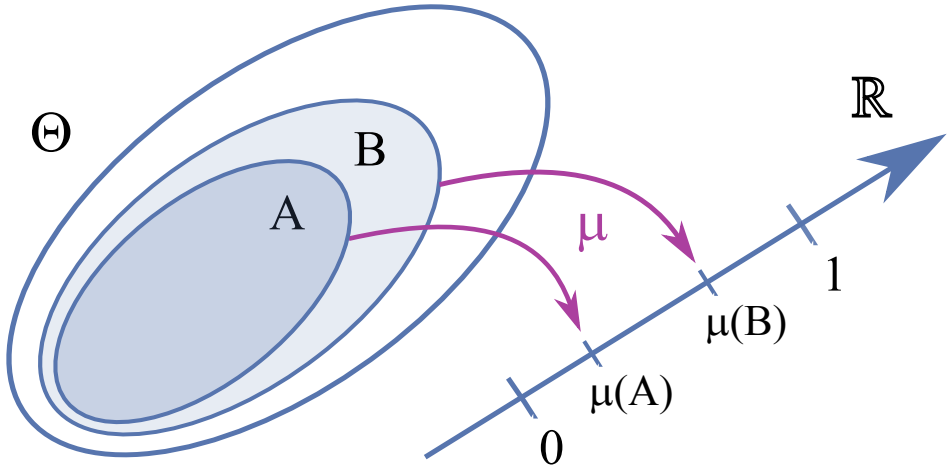
Evidence supporting propositions

- sometimes evidence points directly towards **propositions**, rather than outcomes
- suppose there is a murder, and Peter, John and Mary are under trial for it

$$\Theta = \{\text{Peter}, \text{John}, \text{Mary}\}$$

- there is a witness: he testifies that the person he saw was a man
- this amounts to supporting the proposition $A = \{\text{Peter}, \text{John}\} \subset \Theta$
- should we take this testimony at face value?
- in fact, the witness was tested and the machine reported an 80% chance he was drunk when he reported the crime
- we should partly support the (vacuous) hypothesis that any one among Peter, John and Mary could be the murderer: it is natural to assign 80% chance to proposition A, and 20% chance to proposition Θ

Dealing with propositional evidence

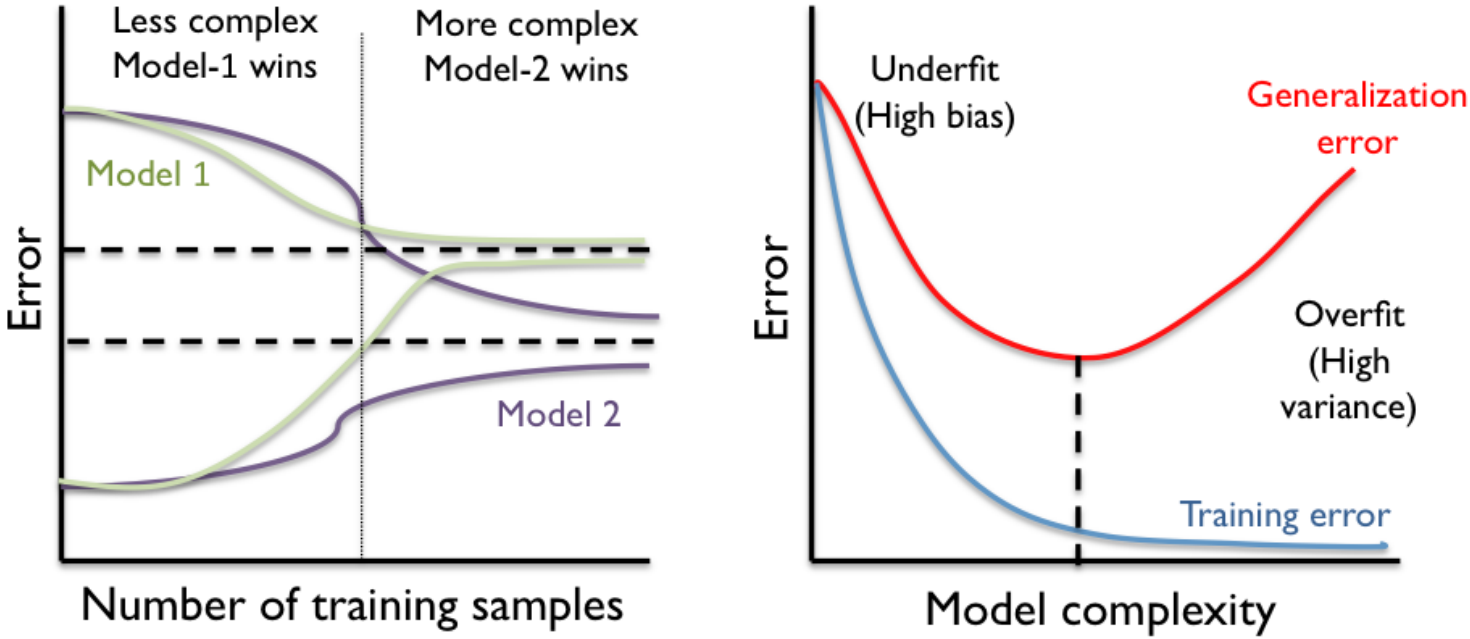


- even when evidence (data) supports propositions, Kolmogorov's **probability forces us to specify support for individual outcomes**
- this is unreasonable - an artificial constraint due to a mathematical model (additivity) that is not general enough
- under a minimal requirement of *monotonitic* measures can potentially suitable to describe probabilities of events: these objects are called **capacities**

Belief functions and propositional data

Belief functions (and random sets in general) are capacities, in which the numbers assigned to subsets are given by a probability distribution (∞-*monotone* capacities).

Machines that can learn



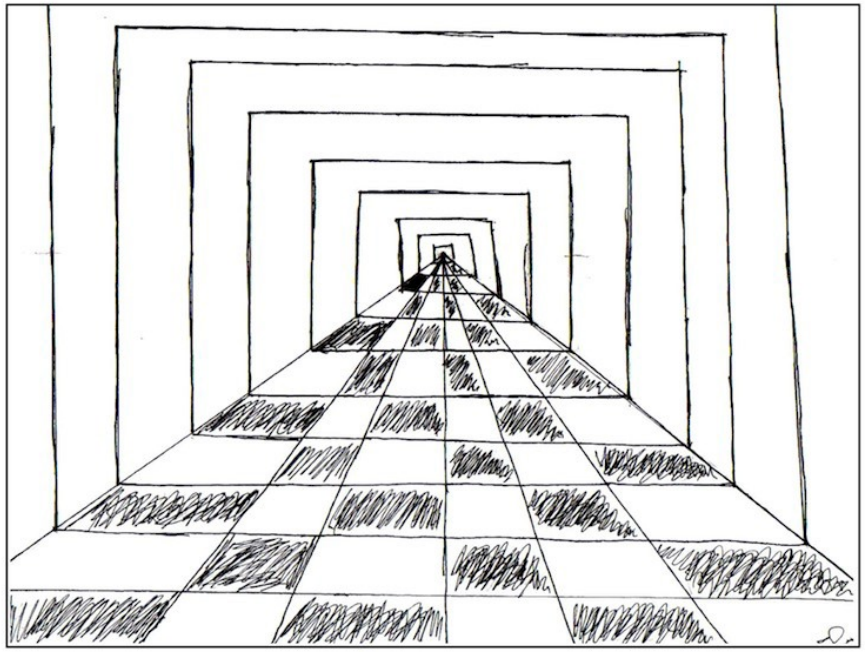
thebayesianobserver.wordpress.com

- we design algorithms that can learn → machine learning
- BUT, we train them on a ridicously small amount of data
- how do we make sure they have learned the right lesson? we compute *generalisation bounds* using statistical learning theory [Vapnik]
- these bounds, however, **assume training and test data distribution to be the same** (!)

Dealing with scarcity

- how do mainstream statistical approaches deal with scarce data?
- a (somewhat naive) objection: probability distributions assume an infinite amount of evidence, so in reality **finite evidence can only provide a constraint on the 'true' probability values**
- unfortunately, those who believe probabilities to be limits of relative frequencies (the frequentists) never really 'estimate' a probability from the data – the only assume ('design') probability distributions for their p-values
  - Fisher: fine, I can never compute probabilities, but I can use the data to test my hypotheses on them
- in opposition, those who do estimate probability distributions from the data (the Bayesians) do not think of probabilities as infinite accumulations of evidence (but as degrees of belief)
  - Bayes: I only need to be able to model a likelihood function of the data
- (well, actually, frequentists **do estimate** probabilities from scarce data when they do stochastic regression)

Asymptotic happiness



- at any rate, both frequentists and Bayesians seem to be happy with **solving their problems 'asymptotically'**
  - limit properties of ML estimates
  - Bernstein-von Mises theorem: asymptotically, the posterior does not depend on the prior
- what about the here and now? how do we explicitly model the uncertainty induced by insufficient data?

Dealing with scarcity

with belief functions

- how do belief functions cope with scarce data?

Belief functions and scarce data

Belief functions cope with scarce data by lower and upper bounds to the probability of each event, whose width reflects the amount of data at disposal.

- note that these bounds are not necessarily induced by 'imprecise' observations
- as we will see later, even a limited amount of precise observations induces Knightian uncertainty which can be expressed by (a pair of) belief functions (cfr. generalised logistic regression)

Modelling pure data

Bayesian approach

- Bayesian reasoning requires modelling the data *and* a prior (actually, you need to pick the proper hypothesis space too!)
  - prior is just a name for beliefs built over a long period of time, from the evidence you have observed – so long a time has passed that all track record of observations is lost, and all is left is a probability distribution
- **why should we 'pick' a prior?** either there is prior knowledge or there is not
- nevertheless we are compelling to picking one, because the mathematical formalism requires it
  - this is the result of a confusion between the original subjective interpretation (where prior beliefs always exist), and the objectivist view of a rigorous objective procedure (where in most cases we do not have any prior knowledge)
- Bayesians then go in 'damage limitation' mode, and try to pick the least damaging prior (see 'ignorance' later)
- **all will be fine, in the end!** (Bernstein-von Mises theorem) Asymptotically, the choice of the prior does not matter (really!)

Choosing the prior

Bayesian inference

- the prior distribution is typically hard to determine – 'solution' $\rightarrow$ pick an 'uninformative' probability
  - **Jeffrey's prior** $\rightarrow$ Gramian of the Fisher information matrix
  - can be improper (unnormalised), and
  - **it violates the strong version of the likelihood principle**: inferences depend not just on the data likelihood but also on the universe of all possible outcomes
- uniform priors can lead to **different results on different spaces**, given the same likelihood functions [Shafer]
  - nice example by Wilson: if I suppose a thousand addresses are equally probable, that make my actual address look implausible
- the asymptotic distribution of the posterior mode depends on the Fisher information and not on the prior (**Bernstein-von Mises theorem**)
- A. W. F. Edwards: "*It is sometimes said, in defence of the Bayesian concept, that the choice of prior distribution is unimportant in practice, because it hardly influences the posterior distribution at all when there are moderate amounts of data. The less said about this 'defence' the better.*"

Modelling pure data

Frequentist approach

- the frequentist approach is **inherently unable to describe pure data**, without making additional assumptions on the data-generating process
- in Nature one cannot 'design' an experiment: **data come your way**, whether you want it or not – you cannot set the 'stopping rules'
  - again, recalls the old image of a scientist 'analysing' (from Greek '*ana*'+'*lysis*', breaking up) a specific aspect of the world in their lab
- the **same data can lead to opposite conclusions**
  - different experiments can lead to the same data, whereas the parametric model employed (family of probability distributions) is linked to a specific experiment
  - apparently, however, frequentists are just fine with this

Dealing with pure data or ignorance

with belief functions

- 'uninformative' priors can be dangerous [Gelman]
- wrong priors can kill a Bayesian model
- priors in general cannot handle multiple hypothesis spaces in a coherent way (families of frames, in Shafer's terminology)

Belief functions and priors

Reasoning with belief functions does not require any prior.

Belief functions and ignorance

Belief functions naturally represent ignorance via the 'vacuous' belief function, assigning mass 1 to the whole hypothesis space.

Extinct statisticians (er.. dinosaurs)

The statistics of rare events



- dinosaurs probably were worrying about the overpopulation risks signalled by their statistical models ..
- .. until a giant asteroid hit them!

What's a rare event?

Very unusual data

- examples of rare events, also called 'tail risks', are: volcanic eruptions, meteor impacts, tsunamis ..
  - in the most extreme cases, these events might have never occurred (e.g. your vote will be decisive in the next presidential election, [Gelman and King, 1998])
- **what is a 'rare' event?**
- an event is rare when it covers a region of the hypothesis space which is seldom sampled -> the **quality** of the data is the issue
- **Knightian uncertainty is presumed to not exist**, with typically bad consequences
- people's aversion to Knightian uncertainty, instead, is well documented (e.g. Ellsberg's paradox in decision making)
  - `http://www.econ.ucla.edu/workingpapers/wp362.pdf`

Dealing with rare events

- probability distributions for the system's behaviour are built in 'normal' times (e.g. while the nuclear plant is working just fine), then used to extrapolate results at the 'tail' of the distribution
- popular statistical procedures (e.g. logistic regression) can sharply underestimate the probability of rare events [King, 2001] (yes, statisticians DO estimate probabilities from data after all!)
- **we should explictly model second-order (Knightian) uncertainties**

Belief functions and rare events

Belief functions can provide cautious predictions of rare events by generalising logistic regression to model second-order uncertainty.

- we will introduce generalising logistic regression at the end

Bayes' rule and certainty

- Bayes' rule is used by Bayesians to reason (in time) when new evidence becomes available
- used by frequentist to condition on the (certain) measurements and generate their p-values
- indeed, it assumes that new evidence **always comes in the form of certain statements**: "the event A is true"
- this is reasonable or even true in many situations: in science and engineering measurements flow in, and this is a form of 'certain' evidence
- applying Bayes' rule to condition on series of measurements to construct likelihood functions (or p-values, if you are a frequentist) then appears very reasonable
- in many real world problems, though, evidence/data is **uncertain**

Uncertain data

- **concepts themselves can be not well defined**, e.g. 'dark' or 'somewhat round' object (qualitative or 'fuzzy' data)
- unreliable sensors can generate faulty (outlier) measurements: **can we still treat these data as 'certain'**? or is more natural to attach to them a degree of reliability, based on the past track record of the 'sensor'?
- but then, can we still apply Bayes' rule, which assumes certainty?
- people ('experts', e.g. doctors) tend to express themselves in terms of likelihoods directly (e.g. 'I think diagnosis A is most likely, otherwise either A or B')
  - if the doctors were frequentists, and were provided with the same data, they would probably apply logistic regression and come up with the same prediction on P(disease|symptoms): unfortunately doctors are not statisticians
- multiple **sensors can provide as output a PDF** on the same space
  - e.g., two Kalman filters based one on color, the other on motion (optical flow)
- can we update a probability using another probability?

Jeffrey's rule of conditioning

A step forward from certainty and Bayes' rule

- an initial probability P **stands corrected by a second probability P', defined only on a number of events**
- suppose P is defined on a σ-algebra $\mathbb{A}$
- there is a new prob measure P' on a sub-algebra $\mathbb{B}$ of $\mathbb{A}$, and the updated probability P'' has to:
  1. meet the prob values specified by P' for events in $\mathbb{B}$
  2. be such that $\forall\, B \in \mathbb{B}$, $X, Y \subset B$, $X, Y \in \mathbb{A}$

$$\frac{P''(X)}{P''(Y)} = \begin{cases} \frac{P(X)}{P(Y)} & \text{if } P(Y) > 0 \\ 0 & \text{if } P(Y) = 0 \end{cases}$$

- there is a unique solution: $P''(A) = \sum_{B \in \mathbb{B}} P(A|B) P'(B)$
- also called the **law of total probability**
- generalises conditioning (obtained when $P'(B) = 1$ for some B)

Belief functions and uncertain evidence

Conditioning versus combination

- what if P' is on the *same* σ-algebra $\mathbb{A}$? Jeffrey's rule cannot be applied!
  - as we saw, this happens when multiple sensors provide predictive PDFs
- belief function deal with uncertain evidence by moving away from the concept of *conditioning* (via Bayes' rule) ..
- .. to that of **combining** pieces of evidence supporting multiple (intersecting) propositions to various degrees

Belief functions and uncertain data

Belief reasoning works by combining existing belief functions with new ones, which are able to encode uncertain evidence.

- belief functions can represent fuzzy concepts as consonant (nested) BFs
- they can represent unreliable measurements as 'discounted' probabilities (by assigning mass to the entire hypothesis set)

Outline

1 Uncertainty and probability
- Nature of uncertainty
- Mathematical probability
- Interpretations of probability
- Bayesian interpretation
- Bayesians vs frequentists

2 Beyond probability
- Missing data
- Propositional data
- Scarce data
- Pure or no data
- Rare data
- Uncertain data

3 Statistics of belief functions
- Dempster's multi-valued setting
- Monotone capacities
- Inner measures
- Conditional multi-valued mappings
- Statistics of experts' opinions
- Probabilistic multivalued RVs
- Gaussian belief functions
- Measure-theoretical analysis
- Frequentist interpretations
- Dempster's a-equations
- Fiducial inference and weak beliefs
- Random set theory

4 New horizons
- Upper and lower likelihood
- Generalising logistic regression
- Frequentist inference with RS
- Generalised laws
- The total belief theorem
- Random set random variables

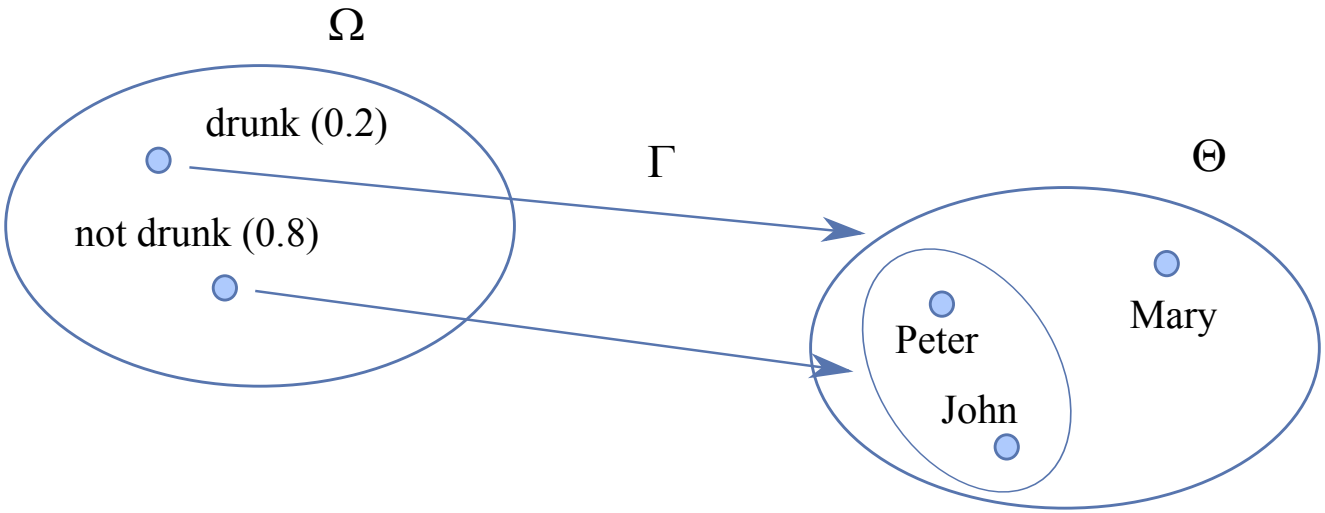
5 Summarising

Belief functions and statistics

- belief functions interpretations
  - Dempster's original multi-valued setting, compatibility relations
  - as generalised probabilities (infinitely monotone capacities)
  - as inner measures
- the statistics of belief functions
  - variations on the multi-valued mapping idea
    - ★ conditional multi-valued mappings [Yen]
    - ★ statistics of expert opinions [Hummel and Landy]
    - ★ probabilistic multi-valued random variables [Lin]
  - Gaussian belief functions [Dempster and Liu]
  - measure-theoretical analysis [Kramosil]
  - frequentist interpretations
    - ★ Walley & Fine
    - ★ confidence structures [Balch]
    - ★ Dempster's (p, q, r) interpretation
  - fiducial inference
    - ★ Almond and Dempster's pivots
    - ★ weak beliefs [Martin, Zhang & Liu]
  - random set theory [Nguyen, Molchanov]

Dempster's multi-valued mappings

- **rationale**: there exists evidence E in the form of probabilities on Ω, which supports degrees of belief on a certain hypothesis space Θ (e.g., the set of outcomes of a trial)



- *elements* of Ω are mapped to *subsets* of Ω: once again this is a **random set**
  - in the example Γ maps $\{not\ drunk\} \in \Omega$ to $\{Peter, John\} \subset \Theta$
- the probability distribution P induces a **mass assignment** $m : 2^\Theta \to [0, 1]$ via the multi-valued (one-to-many) mapping $\Gamma : \Omega \to 2^\Theta$
  - the corresponding mass function is: $m(\{Peter, John\}) = 0.8$, $m(\Omega) = 0.2$

Belief and plausibility functions

.. and compatibility relations

- the total **degree of support (belief)** in A as the probability that the evidence implies A:

$$Bel(A) = P\left(\left\{\omega \in \Omega \middle| \Gamma(\omega) \subseteq A\right\}\right) = \sum_{B \subseteq A} m(B)$$

- the **plausibility** of A as the probability that the evidence does not contradict A:

$$Pl(A) = P\left(\left\{\omega \in \Omega \middle| \Gamma(\omega) \cap A \neq \emptyset\right\}\right) = 1 - Bel(\overline{A})$$

- a multi-valued mapping Γ is equivalent to a **compatibility relation** ..

$$\mathcal{C} = \left\{(\omega, \theta) | \theta \in \Gamma(\omega)\right\} \subset \Omega \times \Theta$$

.. which describes the subset of answers θ in Θ compatible with each $\omega \in \Omega$

- many statistical approaches to BFs are based on generalising the notion of compatibility relation

Belief functions as non-additive probabilities

Completely monotone capacities

Probability measure (additive capacity)

A function $P : \mathcal{F} \to [0, 1]$ over a σ-field $\mathcal{F} \subset 2^\Theta$ such that

- $P(\emptyset) = 0$, $P(\Theta) = 1$;
- if $A \cap B = \emptyset,\ A, B \in \mathcal{F}$ then $P(A \cup B) = P(A) + P(B)$ (*additivity*).

- if we relax the third constraint to allow the function to meet additivity **only as a lower bound** we obtain Shafer's definition of belief function:

Belief function (completely monotone capacity)

A function $Bel : 2^\Theta \to [0, 1]$ from the power set 2^Θ to $[0, 1]$ such that:

- $Bel(\emptyset) = 0$, $Bel(\Theta) = 1$;
- for every n and for every collection $A_1, ..., A_n \in 2^\Theta$ we have that:

$$Bel\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} Bel\left(\bigcap_{i \in I} A_i\right)$$

Belief functions as inner measures

- Shafer has shown that the two definitions of belief functions are equivalent
- BUT, BFs can also be assimilated to **inner measures** [Ruspini & Fagin]
- given a probability measure P defined over a σ-field of subsets $\mathcal{F}$ of a finite set $\mathcal{X}$, the *inner probability* of P is the function P_* defined by:

$$P_*(A) = \max\left\{P(B)\middle| B \subset A,\ B \in \mathcal{F}\right\}, \qquad A \subseteq \mathcal{X}$$

- extends a probability measure on *some* events to the *whole* of $\mathcal{X}$
- now, set $\mathcal{X} = \mathcal{C}$, a compatibility relation between Ω and Θ, and

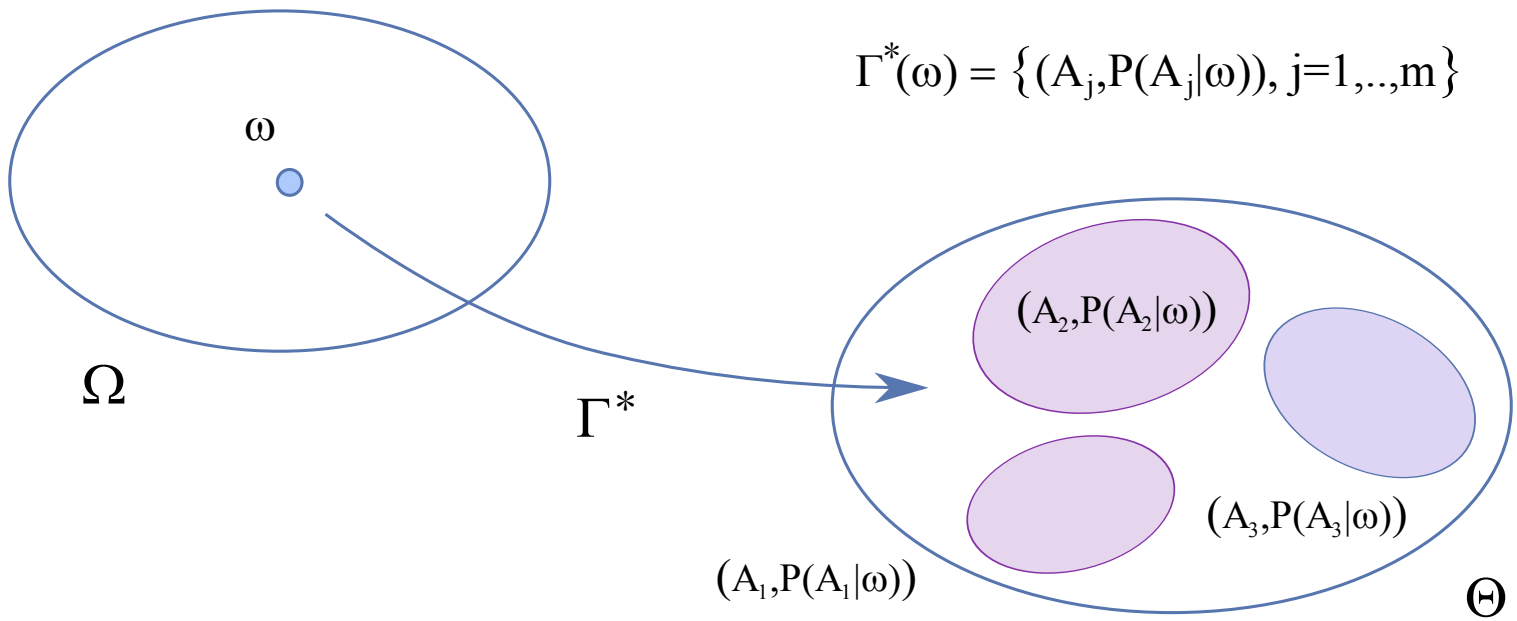
$$\mathcal{F} = \left\{\mathcal{C} \cap (E \times \Theta), \forall E \subset \Omega\right\}$$

  each element $\mathcal{C} \cap (E \times \Theta)$ of $\mathcal{F}$ is the set of compatibility pairs (ω, θ) which involve an element of $E \subset \Omega$: $\omega \in E$
- the inner probability induced by $Q(\mathcal{C} \cap (E \times \Theta)) \doteq P(E)$ is the belief function induced by $\Gamma : \Omega \to 2^\Theta$ with source probability P

Conditional multi-valued mappings

Yen, 1986

- extends the original multivalued mapping of Dempster's formulation
- probabilistic setting which uses conditional probabilities to **express uncertainty in the mapping itself**



- the deterministic multi-valued setting by Dempster is associated with binary conditional probabilities on the mapping: $P(A|\omega) = 1$, $P(A^c|\omega) = 0$

Conditional multi-valued mappings

- **probabilistic multi-valued mapping** from a space Ω to a space Θ is a function

$$\Gamma^* : \Omega \to 2^{2^\Theta \times [0,1]}$$

in which the image of an element ω of Ω is a collection of subset-probability pairs

$$\Gamma^*(\omega) = \left\{ (A_{i_1}, P(A_{i_1}|\omega)), \cdots, (A_{i_m}, P(A_{i_m}|\omega)) \right\}$$

- subject to some sensible constraints: $A_{i_j} \neq \emptyset$ $j = 1, ..., m$; $A_{i_j} \cap A_{i_k} = \emptyset$ whenever $j \neq k$; $P(A_{i_j}|\omega) > 0$ $j = 1, ..., m$; $\sum_j P(A_{i_j}|\omega) = 1$
- each A_{i_j} is called a **granule**
- each focal element of standard belief theory is broken down into a collection of **disjoint** granules, each with an attached (conditional) probability.
- the mass of each focal element $A \subset \Theta$ in this extended multi-valued framework is

$$m(A) = \sum_{\omega \in \Omega} P(A|\omega) P(\omega),$$

obtained by multiplying the conditional probability expressing the mapping by the prior probability of the elements of Ω

Statistics of experts' opinions

Hummel and Landy, 1988

- evidence combination relates to **statistics of experts in a space of 'expert opinions'** who combine information in a Bayesian fashion
- Dempster's rule reduces to the Bayesian updating of Boolean assertions
- consider a set of experts $\mathcal{E}$, each expert $\omega \in \mathcal{E}$ has weight $\mu(\omega)$, and maintains a set of possible outcomes ('labels') for a question with universe Λ
- as in Dempster, there is a multivalued mapping $\Gamma : \mathcal{E} \to 2^\Lambda$ BUT
- **each expert has a probabilistic opinion** p_ω on Λ such that $\forall \omega$

$$p_\omega(\lambda) \geq 0 \ \forall \lambda \in \Lambda, \quad p_\omega(\lambda) > 0 \text{ iff } \lambda \in \Gamma(\omega),$$

  and either $\sum_{\lambda \in \Lambda} p_\omega(\lambda) = 1$ or $p_\omega(\lambda) = 0 \ \forall \lambda$
- the last constraint describes the case in which expert ω has no opinion
- $p_\omega(\lambda)$ similar to Yen's conditionals $P(A|\omega)$

Statistics of experts' opinions

Boolean opinions of experts

- this setting generalises Dempster-Shafer theory: the indicator functions

$$x_\omega(\lambda) = \begin{cases} 1 & p_\omega(\lambda) > 0 \\ 0 & p_\omega(\lambda) = 0 \end{cases}$$

  are called the **Boolean opinions of the experts**, and correspond to $\Gamma(\omega)$
- if we regard the space of experts $\mathcal{E}$ as a sample space, then each $x_\omega(\lambda)$ can be regarded as a sample of a random (Boolean) variable $x(\lambda)$
- the $p_\omega(\lambda)$s can be seen as samples of a random variable $p(\lambda)$
- the state of the system is thus defined by statistics on the set of random variables $\{x(\lambda)\}_{\lambda \in \Lambda}$
- statistics are computed using the weights $\mu(\omega)$ of the individual experts, via the counting measure $\mu(F) = \sum_{\omega \in F} \mu(\omega)$

Statistics of experts' opinions

Belief measures

- given an event $A \subset \Lambda$, we can define

$$\tilde{m}(A) = \frac{\mu(\omega \in \mathcal{E} | x_\omega = x_A)}{\mu(\mathcal{E})} = \frac{\mu(\omega \in \mathcal{E} | \Gamma(\{\omega\}) = A)}{\mu(\mathcal{E})}$$

  where x_A is the indicator function of A
- we can also say that $\tilde{m}(A) = Prob_{\mathcal{E}}(x_\omega = x_A)$, once we express

$$Prob_{\mathcal{E}}(\text{event}) = \frac{\mu(\{\omega \in \mathcal{E} | \text{event is true for } \omega\})}{\mu(\mathcal{E})}$$

  the probability of an event as the **ratio of experts for which the event is true**
- under this interpretation, the **belief on a set** A is the joint probability

$$Bel(A) = \sum_{B \subseteq A} m(B) = Prob_{\mathcal{E}'}(x(\lambda) = 0 \text{ for } \lambda \notin A),$$

  where $\mathcal{E}'$ is the set of experts expressing an opinion

Probabilistic multivalued random variables

Lin, 1997

- generalised the concept of random variable
- a **probabilistic multivalued random variable** (PMRV) from Ω to Θ is a function $\mu : \Omega \times \Theta \to [0, 1]$ such that, for all $\omega \in \Omega$:

$$\sum_{\theta \in \Theta} \mu(\omega, \theta) = 1$$

- we can define the inverse mapping $\mu^{-1}(\theta) = \left\{ \omega \in \Omega \middle| \mu(\omega, \theta) \neq 0 \right\}$
- if (Ω, p) is a probability space, a PMRV μ induces a probability p_Θ on Θ:

$$p_\Theta(\theta) = \sum_{\omega \in \mu^{-1}(\theta)} p(\omega)\mu(\omega, \theta). \tag{1}$$

  a PMRV maps each point in a sample space $\omega \in \Omega$ to a probability distribution on the target one, via: $\mu_\omega(\theta) = \mu(\omega, \theta)$
- (compare Hummel's probabilistic opinions of experts!)

Probabilistic multivalued random variables

- the PMRV induces a basic probability assignment on Ω as follows:

$$m(A) = \sum_{\mu^{-1}(\theta)=A} p_\Theta(\theta) = \sum_{\mu^{-1}(\theta)=A} \sum_{\omega \in A} p(\omega)\mu(\omega, \theta)$$

- Yen's conditional multivalued mappings very close to this!
- whereas in Shafer's definition the probability distribution on Ω is arbitrary and unrelated to the compatibility relation..
- .. here the probability distribution on Θ is induced by the PMRV, which also defines the multivalued mapping μ^{-1}
- a difference emerges when deriving Dempster's combination rule
- consider two PMRVs μ_1, μ_2 from Ω to Θ_1, Θ_2
- since $\mu_1^{-1}(\theta_1) \cap \mu_2^{-1}(\theta_2) = \emptyset$ if and only if $\mu_1^{-1}(\theta_1) = \emptyset$ and $\mu_2^{-1}(\theta_2) = \emptyset$, it follows that the independence assumption behind Dempster's rule implies:

$$\sum_{A \cap B = \emptyset} m_1(A) m_2(B) = 0,$$

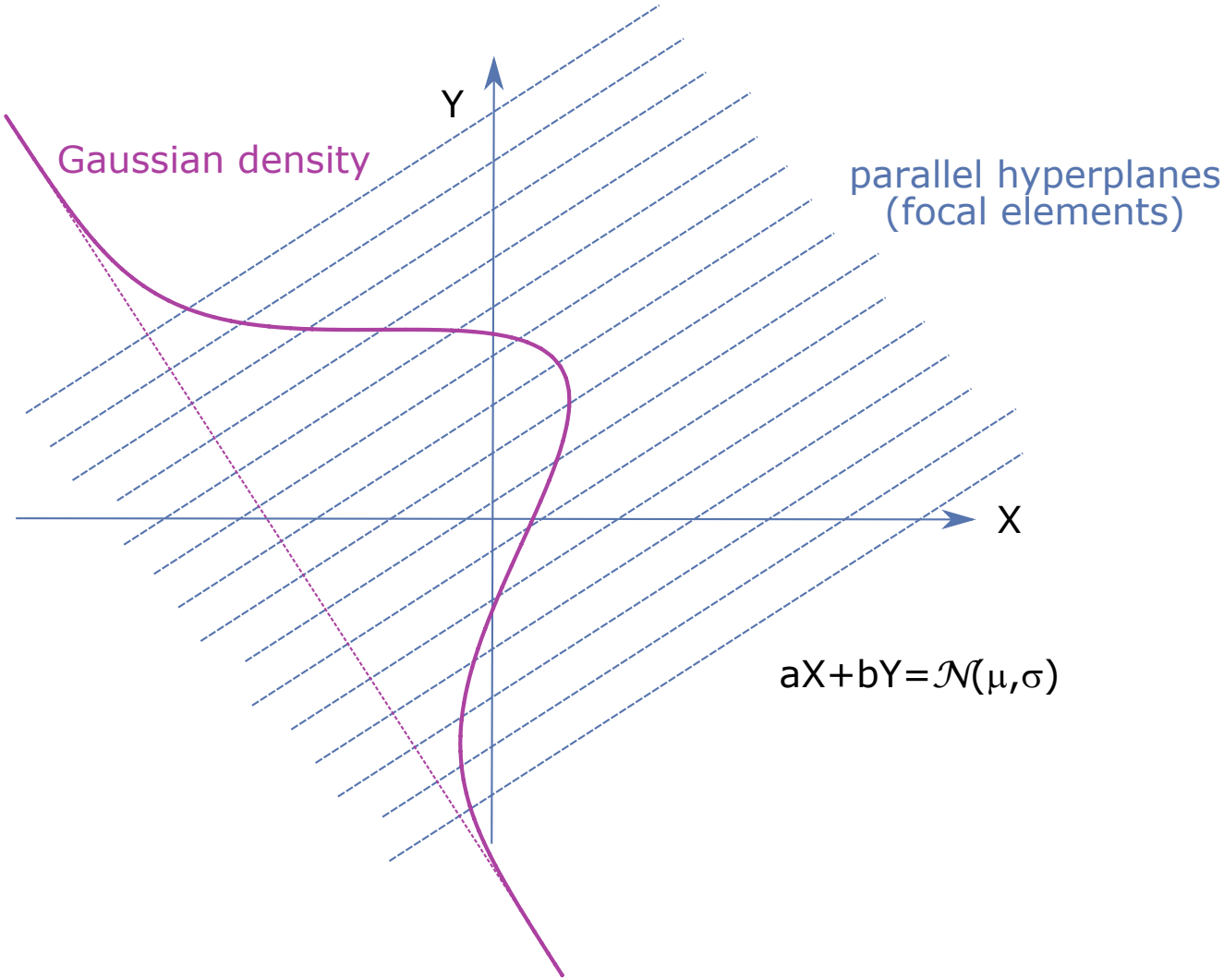
and no normalization is necessary, because conflict does not arise

Gaussian belief functions

Liu, 1996

- notion proposed by A. Dempster and formalized by L. Liu in 1996
- a Gaussian belief function is a Gaussian distribution over the members of the parallel partition of an hyperplane
- encode each proposition (event) as a linear equation, so that all parallel sub-hyperplanes of a given hyperplane are possible focal elements and a Gaussian belief function is a Gaussian distribution over these sub-hyperplanes.
- As focal elements (hyperplanes) cannot intersect, the framework is less general than standard belief theory
- e.g. we can write $3X + 5Y \sim \mathcal{N}(2, 10)$, to mean that the frame of discernment is $\mathbb{R}^2$, the focal elements are the hyperplanes parallel to $3X + 5Y = 0$
- the mass value for each hyperplane $s(w) = \{(x, y)|3x + 5y = w\}$ is the density value $f(w) \sim \mathcal{N}(2, 10)$.

Gaussian belief functions



- a nice matrix calculus can be derived for linear belief functions [Liu 2016]

Measure-theoretical analysis

Kramosil, 1998-2001

- Ivan Kramosil exploited measure theory to expand the theory of belief functions beyond its classical scope
- main contributions include:
  - Boolean and non-standard valued belief functions
  - the extension of belief functions to countable sets
  - a strong law of large numbers for random sets
  - 'signed' belief functions
  - Monte-Carlo estimation of belief functions
  - **belief functions induced by partial compatibility relations**

BFs induced by partial compatibility relations

Kramosil, 1999

- assume a standard compatibility relation $\mathcal{C} \subset \Omega \times \Theta$
- the *total generalised compatibility relation* induced on $\mathcal{P}(\Omega) \times \mathcal{P}(\Theta)$ by $\mathcal{C}$ is the relation $\mathcal{C}^* \subset \mathcal{P}(\Omega) \times \mathcal{P}(\Theta)$ such that:

$$(X, Y) \in \mathcal{C}^* \text{ iff } \exists \omega \in X, \theta \in Y \text{ such that } (\omega, \theta) \in \mathcal{C}$$

- a *partial* generalised compatibility relation is then a relation defined on a proper subset D of $\mathcal{P}(\Omega) \times \mathcal{P}(\Theta)$, such that it can be extended to a proper total relation
- main result: any belief function defined by a fragment of the total generalised compatibility relation induced by an original compatibility relation C, is a lower approximation of the original belief function associated with C [Kramosil, 1998]

Frequentist interpretation

Walley and Fine, 1982

- attempted to formulate a **frequentist theory for lower probabilities**
- problem: estimating a lower probability $\underline{P}$ from a series $\epsilon_1, ..., \epsilon_n$
- they consider the **estimator for** $\underline{P}$

$$\underline{r}_n(A) = \min \Big\{ r_j(A) : k(n) \leq j \leq n \Big\}, \quad k(n) \to \infty$$

  where $r_j(A)$ is the relative frequency of event A after ϵ_j
- let $\underline{P}^\infty$ describe the IID repetitions $\epsilon_1, ...$
- this estimation process succeeds, in the sense that

$$\lim_{n \to \infty} \frac{\underline{P}^\infty(G^c_{n,\delta})}{\underline{P}^\infty(G_{n,\delta})} = 0 \quad \forall \delta > 0$$

  where $G_{n,\delta}$ is the event $|\underline{r}_n(A) - \underline{P}(A)| < \delta$
- parallels Bernoulli's law of large numbers: **the confidence that** $\underline{r}_n(A)$ **is close to** $\underline{P}(A)$ **grows with sample size**

Confidence structures

Balch, 2012

- recent attempt to reconcile belief function theory with frequentist statistics
- **observation-conditional random set** $\Gamma(\omega, x) \subset \Theta \rightarrow$ depends on the observed value x of the observable random variable X
- induces a belief value as

$$Bel_{\theta|x}(A) = P_\omega(\{\omega \in \Omega : \Gamma(\omega, x) \subseteq A\}) \quad \forall x \in \mathcal{X}, A \subset \Theta$$

- a **confidence set** is a set-estimator for a parameter

$$C : \mathcal{X} \rightarrow \mathcal{P}(\Theta), \quad C(x) \subset \Theta$$

(classical **confidence intervals** are the most well-known type of confidence set)
`https://en.wikipedia.org/wiki/Confidence_interval`

- the **Neyman-Pearson confidence** associated with a confidence set is the *frequentist probability of drawing x s.t. $C(x)$ covers the true parameter value*

Confidence structures

- a **confidence structure** is then an observation-conditional random set whose source probabilities P_ω are consistent the Neyman-Pearson confidence (left):

$$P_{x|\theta}\left(\left\{x \in \mathcal{X} : \theta \in \bigcup_{\omega \in A} \Gamma(\omega, x)\right\}\right) \geq P_\omega(A), \quad \forall A \subset \Omega, \theta \in \Theta$$

- alternatively, it is a means for constructing confidence sets of the form

$$C(A; x) = \bigcup_{\omega \in A} \Gamma(\omega, x), \quad A \subset \Omega,$$

  s.t. the coverage probability of $C(A; x)$ is greater than or equal to the source probability, $P_\omega(A)$

- every $B \subset \Theta$ can be put in the form $B = C(A; x)$ for some $A \subset \Omega$, $x \in \mathcal{X} \Rightarrow$ all sets of parameter values are confidence intervals
- $Bel(B = C(A; , x))$ returns then a lower bound on the coverage probability of $B \subset \Theta$ for a given observation x
- various (inference) methods for constructing confidence structures from confidence distributions, pivots and p-values are proposed

Dempster's (p, q, r) interpretation

Dempster, 2011

- a semantics of belief functions is provided whereby every formal assertion A is associated with a triple (p, q, r) where p is the probability 'for' the assertion, q is the probability 'against' the assertion, and r is the probability of 'don't know'

$$p = Bel(A), \quad q = Bel(A^c), \quad r = 1 - p - q$$

- arguments are presented for the necessity of $r = 1 - Bel(A) - Bel(A^c)$
- the methodology is applied to inference and prediction from Poisson counts
- the relation of DS theory to statistical significance testing is elaborated, introducing along the way the concept of *dull* null hypothesis
- this is a hypothesis which assigns mass 1 to an interval of values $L_1 \leq L \leq L_2$
- Poisson's a-probabilities and values for p, q and r can then be readily derived

Dempster's approach to statistical inference

"a-equations"

- consider a statistical model

$$\Big\{ f(x|\theta), x \in \mathbb{X}, \theta \in \Theta \Big\},$$

  where $\mathbb{X}$ is the sample space and Θ the parameter space
- having observed x, how to **quantify the uncertainty about the parameter** θ, without specifying a prior probability distribution?
- suppose that the model is represented by an "a-equation":

$$X = a(\Theta, U)$$

  where U is an **(unobserved) auxiliary variable** with known probability distribution $\mu : U \to [0, 1]$ (**pivotal measure**) independent of θ

Dempster's approach to statistical inference

From *a*-equations to belief functions

- the equation $X = a(\Theta, U)$ defines a multi-valued mapping $\Gamma : U \to 2^{\mathbb{X} \times \Theta}$:

$$\Gamma : u \mapsto \Gamma(u) = \left\{ (x, \theta) \in \mathbb{X} \times \Theta \middle| x = a(\theta, u) \right\} \subset \mathbb{X} \times \Theta$$

- under the usual measurability conditions, the probability space $(U, \mathcal{B}(U), \mu)$ and the multi-valued mapping Γ induce a belief function $Bel_{\Theta \times \mathbb{X}}$ on $\mathbb{X} \times \Theta$
- conditioning $Bel_{\Theta \times \mathbb{X}}$ on θ yields

$$Bel_{\mathbb{X}}(A|\theta) = Bel(\mu\{u : a(\theta, u) \in A\}), \quad A \subseteq \mathbb{X}$$

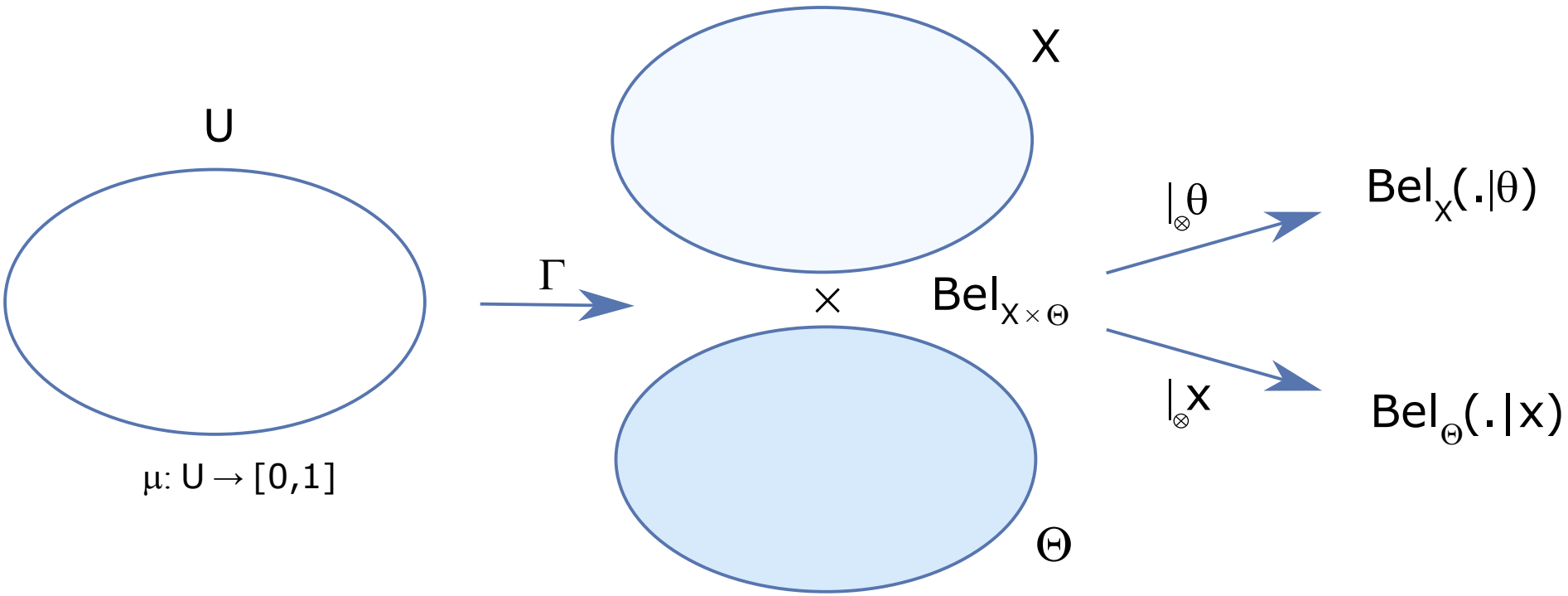
- conditioning it on $X = x$ gives

$$Bel_{\Theta}(B|x) = \frac{\mu\{u : M_x(u) \subseteq B\}}{\mu\{u : M_x(u) \neq \emptyset\}}, \quad B \subseteq \Theta$$

where $M_x(u) = \{\theta : x = a(\theta, u)\}$

Dempster's approach to statistical inference

Graphical representation



Fiducial inference

Fisher, 1930

- inference based on a-equations is an example of **fiducial inference**
- Θ is the parameter of interest, X is a sufficient statistic[1] rather than the observed data, and U is the auxiliary variable (**pivotal quantity**)
- crucial assumption underlying the fiducial argument: each one of (X, Θ, U) is uniquely determined by the a-equation given the other two
- the pivotal quantity U is assumed to have an a priori distribution μ, indep of Θ
- prior to the experiment, X has a sampling distribution that depends on Θ; after the experiment, however, X is no longer a random variable (as it is measured)
- to produce a posterior distribution for Θ, the variability in X prior to the experiment must somehow be transferred, after the experiment, to Θ
- if we "**continue to believe**" that U is distributed according to μ after X is observed we can derive a distribution for Θ, called the **fiducial distribution**

[1] A statistic $t = T(S)$ is a function of the observed sample S. The statistic is sufficient if $P(s|t, \theta) = P(s|t)$.

Fiducial inference

Classical example

- problem: estimating the unknown mean of a $N(0,1)$ population based on a single observation X
- a-equation: $X = \Theta + \Psi^{-1}(U)$, where $\Psi(.)$ is the CDF of the $N(0,1)$ distribution
- assume the pivotal quantity has prior distribution $\mu = Unif(0,1)$
- for a fixed θ, the events $\{\Theta \leq \theta\}$ and $\{U \geq \Psi(X - \theta)\}$ are the same!
- hence, their probabilities need to be the same
- if we 'continue to believe' then, the **fiducial probability** of $\{\Theta \leq \theta\}$ is $\Psi(\theta - X)$
- in other words, the **fiducial distribution of** Θ **given** X is

$$\Theta \sim N(X, 1)$$

- we got a posterior for Θ, without requiring a prior for it

Inferential models

Martin and Liu, 2013

- given Bel_x, Dempster's posterior on Θ given x ..
- .. we say that another BF Bel^* on Θ specifies an **inferential model** (IM) there if

$$Bel^*(A) \leq Bel_x(A) \quad \forall A \subseteq \Theta$$

- **weak belief** (WB) is method for specifying a suitable belief function within an inferential model
- can be done, e.g., by modifying the multi-valued mapping Γ or the pivot distribution μ

Weak beliefs

- if we assume an a-equation, and we observe x generated by it, knowing θ is equivalent to knowing u^*
- inference on θ is equivalent to predicting the value of the unobserved u^*
  - (this is shared by both fiducial and Dempster's inference approaches)
- the 'continue to believe' assumption of both fiducial and DS says that u^* can be predicted by taking draws from the pivotal measure μ
- **WB weakens this assumption by replacing the draw $u \sim \mu$ with a set $S(u)$ containing u, which is equivalent to replacing μ with a belief function**
- given a pivotal measure μ, WB constructs a belief function on U by choosing a set-valued mapping $\mathcal{S} : U \to 2^U$ that satisfies $u \in \mathcal{S}(u)$
- it is a multivalued mapping from the domain of the auxiliary variable onto itself
- the pivot distribution induces belief function onto its own domain

Inference with weak beliefs

- $\mathcal{S}$ is called **predictive random set**, sort of 'smearing' the given pivot distribution μ for u
- so now we have three domains $\mathbb{X}$, U and Θ and two multi-valued mappings, $\Gamma : U \to 2^{\mathbb{X}\times\Theta}$ and $\mathcal{S} : U \to 2^U$
- the two belief functions on U and $\mathbb{X} \times \Theta$ are extended to $U \times \mathbb{X} \times \Theta$
- by combining them there, and marginalising over U, we get a BF whose random set is

$$\Gamma_{\mathcal{S}}(u) = \bigcup_{u' \in \mathcal{S}(u)} \Gamma(u')$$

  which is dominated by the original Dempster's posterior: it is a **valid inferential model**
- various methods for building the mapping $\mathcal{S}$ are discussed, including 'maximal belief'

Belief functions as random sets

Nguyen 1978, Hestir 1991

- given a multi-valued mapping Γ, a straightforward step is to consider the probability value $P(c)$ as attached to the subset $\Gamma(c) \subseteq \Omega$
- a **random set** in Ω, i.e., a probability measure on a collection of subsets
- the degree of belief $Bel(A)$ of an event A becomes the **cumulative distribution function (CDF) of the open interval of sets** $\{B \subseteq A\}$
- **lower inverse** and **upper inverse** of Γ:

$$\Gamma_*(A) \doteq \{c \in \mathcal{C} : \Gamma(c) \subset A, \Gamma(c) \neq \emptyset\} \quad \Gamma^*(A) \doteq \{c \in \mathcal{C} : \Gamma(c) \cap A \neq \emptyset\}$$

- given two σ-fields $\mathcal{A}, \mathcal{B}$ on $\mathcal{C}, \Omega$ respectively, Γ is said **strongly measurable** iff $\forall B \in \mathcal{B}$, $\Gamma^*(B) \in \mathcal{A}$
- the **lower probability measure** on $\mathcal{B}$ is defined as $P_*(B) \doteq P(\Gamma_*(B))$ for all $B \in \mathcal{B}$ - this is nothing but a belief function!
- Nguyen proved that, if Γ is strongly measurable, the CDF $\hat{P}$ of the random set coincides with the lower probability measure:

$$\hat{P}[I(B)] = P_*(B) \quad \forall B \in \mathcal{B}, \qquad I(B) \doteq \{C \in \mathcal{B}, C \subseteq B\}$$

Belief functions as random sets

Molchanov's work

- recently, strong renewed interest in a theory of random sets, thanks to Molchanov [2006] and others
  - developed a theory of calculus with capacities and random sets
  - Radon-Nikodym theorems for capacities and random sets (see Horizons) and derivatives of capacities
  - (conditional) expectations of random sets
  - limit theorems: strong law of large numbers, central limit theorem, Gaussian RSs
  - examined set-valued random processes
- **powerful mathematical framework!** way forward for the theory in my view
- no mentioning of conditioning and combination yet
- connections with mathematical statistics to develop
- special case of **random element** [Frechet], random variable with structured output

Random closed sets

Definition

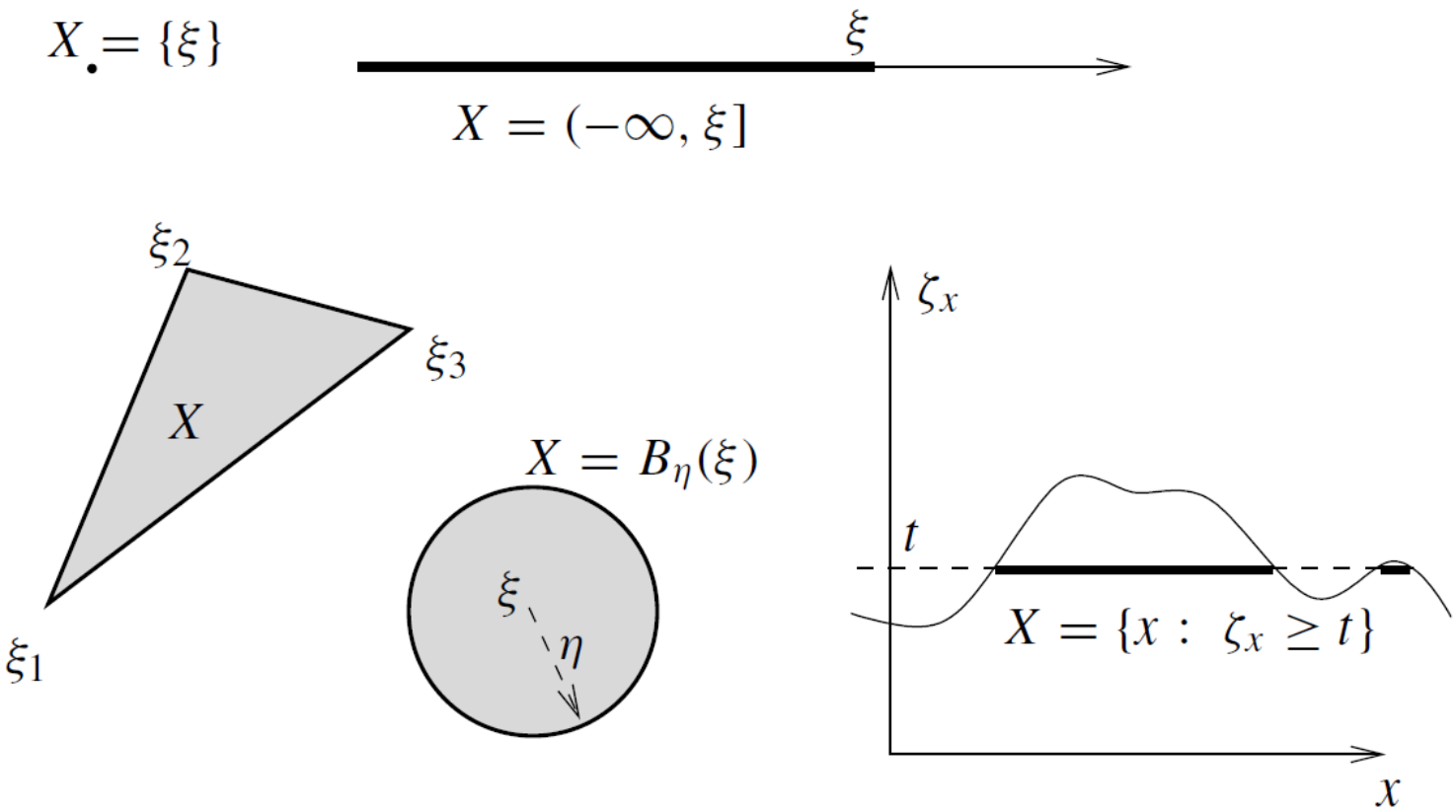
- the family of all sets is too large, we typically restrict ourselves to the case of random elements in the space of closed subsets of a certain topological space $\mathbb{E}$
- the family of closed subsets of $\mathbb{E}$ is denoted by $\mathcal{C}$
- $\mathcal{K}$ denotes the family of all compact subsets of $\mathbb{E}$
- let $(\Omega, \mathcal{F}, P)$ be a probability space
- a map $X : \Omega \to \mathcal{C}$ is called a **random closed set** if, for every compact set K in $\mathbb{E}$:

$$\{\omega : X(\omega) \cap K \neq \emptyset\} \in \mathcal{F}$$

- this is equivalent to strong measurability, whenever the σ-field on Θ is replaced by the family $\mathcal{K}$ of compact subsets of Θ
- the consequence is that the upper probability of K exists $\forall K \in \mathcal{K}$

Random closed sets

Some examples



- if ξ is a random variable, then $X = (-\infty, \xi]$ is a random closed set
- if ξ_1, ξ_2 and ξ_3 are three random vectors in $\mathbb{R}^d$, then the triangle with vertices ξ_1, ξ_2 and ξ_3 is a random closed set

Capacity functionals

Random closed sets

- a functional $T_X : \mathcal{K} \to [0, 1]$ given by

$$T_X(K) = P(\{X \cap K \neq \emptyset\}), K \in \mathcal{K}$$

  is said to be the **capacity functional** of X
- in particular, if $X = \{\xi\}$ is a classical random variable, then $T_X(K) = P(\{\xi \in K\})$ is the probability distribution of the random variable ξ
- the name 'capacity functional' follows from the fact that T_X is a functional on $\mathcal{K}$ which takes values in $[0, 1]$, equals 0 on the empty set, is monotone and upper semicontinuous (i.e., T_X **is a capacity**, and also completely alternating on $\mathcal{K}$)
- $T_X(K)$ is the plausibility measure induced by the multivalued mapping X, restricted to compact subsets
- the links betwen random closed sets and belief/plausibility functions, upper and lower probabilities, and contaminated models in statistics are very briefly hinted at in [Molchanov 2005], Chapter 1, Section 9

Outline

1 Uncertainty and probability
- Nature of uncertainty
- Mathematical probability
- Interpretations of probability
- Bayesian interpretation
- Bayesians vs frequentists

2 Beyond probability
- Missing data
- Propositional data
- Scarce data
- Pure or no data
- Rare data
- Uncertain data

3 Statistics of belief functions
- Dempster's multi-valued setting
- Monotone capacities
- Inner measures
- Conditional multi-valued mappings
- Statistics of experts' opinions
- Probabilistic multivalued RVs
- Gaussian belief functions
- Measure-theoretical analysis
- Frequentist interpretations
- Dempster's a-equations
- Fiducial inference and weak beliefs
- Random set theory

4 New horizons
- Upper and lower likelihood
- Generalising logistic regression
- Frequentist inference with RS
- Generalised laws
- The total belief theorem
- Random set random variables

5 Summarising

New statistical horizons

- random sets (RSs) are a most natural representation of uncertainty
- they are also a straightforward generalisation of mathematical statistics
- how should the theory develop? some modest proposals:
  - learning a belief function from the data: **lower and upper likelihoods**
  - a theory of **generalised logistic regression** for dealing with rare event estimation
  - **parameterised families of random sets** for **frequentist hypothesis testing** ..
  - generalised laws:
    - ★ the central limit theorem for random sets
    - ★ laws of large numbers
    - ★ the **total belief theorem**
  - **random set random variables**: generalising the Radon-Nikodym derivative

Belief likelihood function

Generalising the traditional likelihood function

- the traditional likelihood function is a conditional probability of the data given a parameter $\theta \in \Theta$, i.e. a family of PDF over $\mathbb{X}$ parameterised by θ
- a statistical inference approach which maps traditional likelihood to a belief function have been proposed [Shafer, Wasserman,Denoeux]
- why can we not define directly a '**belief likelihood function**' of a sample $x \in \mathbb{X}$?
- it is natural to define a belief (set-) likelihood function as family of belief functions on $\mathbb{X}$, $Bel_{\mathbb{X}}(.|\theta)$ parameterised by $\theta \in \Theta$
  - this is the input of Smets' Generalised Bayesian Theorem, a collection of 'conditional' belief functions
- note that a belief likelihood takes values on *sets* of outcomes $A \subset \mathbb{X}$ – individual outcomes are just a special case
- seems a **natural setting for computing likelihoods of set-valued observations** $\rightarrow$ coherent with the random set philosophy

Belief likelihood function

of series of repeated trials

- let $Bel_{\mathbb{X}_i}(A|\theta)$, for $i = 1, 2, ..., n$ be a parameterised family of belief functions on $\mathbb{X}_i$, the space of quantities that can be observed at time i, depending on $\theta \in \Theta$
- a series of repeated trials then assumes values in $\mathbb{X}_1 \times \cdots \times \mathbb{X}_n$, whose elements are tuples of the form $\mathbf{x} = (x_1, ..., x_n)$ ('sharp' samples)

Belief likelihood function (BLF) of repeated trials

$$Bel_{\mathbb{X}_1 \times \cdots \times \mathbb{X}_n}(A|\theta) \doteq Bel_{\mathbb{X}_1}^{\uparrow \times_i \mathbb{X}_i} \odot \cdots \odot Bel_{\mathbb{X}_n}^{\uparrow \times_i \mathbb{X}_i}(A|\theta)$$

where $Bel_{\mathbb{X}_j}^{\uparrow \times_i \mathbb{X}_i}$ is the vacuous extension of $Bel_{\mathbb{X}_j}$ to $\mathbb{X}_1 \times \cdots \times \mathbb{X}_n$, and $\odot$ is a combination rule

Lower and upper likelihoods

$$\underline{L}(\mathbf{x}) \doteq Bel_{\mathbb{X}_1 \times \cdots \times \mathbb{X}_n}(\{(x_1, ..., x_n)\}|\theta), \quad \overline{L}(\mathbf{x}) \doteq Pl_{\mathbb{X}_1 \times \cdots \times \mathbb{X}_n}(\{(x_1, ..., x_n)\}|\theta)$$

of the 'sharp' sample $\mathbf{x} = \{(x_1, ..., x_n)\}$

- can we reduce this to the belief values of the individual trials?

Belief likelihood function

Factorisation results

- **yes**, whenever $\cap$ or $\oplus$ are used for computing the BLF of Cartesian products:

BLF factorisation for Cartesian products

When using both $\cap$ or $\oplus$ as a combination rule, and $A = A_1 \times \cdots \times A_n$, $A_i \subseteq \mathbb{X}_i$ for all i:

$$Bel_{\mathbb{X}_1 \times \cdots \times \mathbb{X}_n}(A_1 \times \cdots \times A_n | \theta) = \prod_{i=1}^{n} Bel_{\mathbb{X}_i}(A_i | \theta)$$

- in particular, for 'sharp' tuples $A = \{(x_1, ..., x_n)\}$:

Factorisation for sharp tuples

$$\underline{L}(\mathbf{x} = \{x_1, ..., x_n\}) = \prod_{i=1}^{n} Bel_{\mathbb{X}_i}(x_i) \quad \overline{L}(\mathbf{x} = \{x_1, ..., x_n\}) = \prod_{i=1}^{n} Pl_{\mathbb{X}_i}(x_i)$$

- (Smets' *conditional conjunctive independence*, but just for individual samples $\mathbf{x}$)

Lower and upper likelihoods

Bernoulli trials

- case of Bernoulli: $\mathbb{X}_i = \mathbb{X} = \{1, 0\}$ and the BFs $Bel_{\mathbb{X}_i}$ are equidistributed
- under conditional independence and equidistribution, the traditional likelihood for a series of Bernoulli trials reads as $p^k(1-p)^{n-k}$, where k is the number of successes and n the number of trials
- **let us compute the belief likelihood function for Bernoulli trials!**
- we seek the belief function on $\mathbb{X} = \{1, 0\}$, parameterised by $p = m(\{1\})$, $q = m(\{0\})$ (with $p + q \leq 1$ this time) which best describes the observed sample
- the lower and upper likelihoods of the sample $\mathbf{x} = \{x_1, ..., x_n\}$ are:

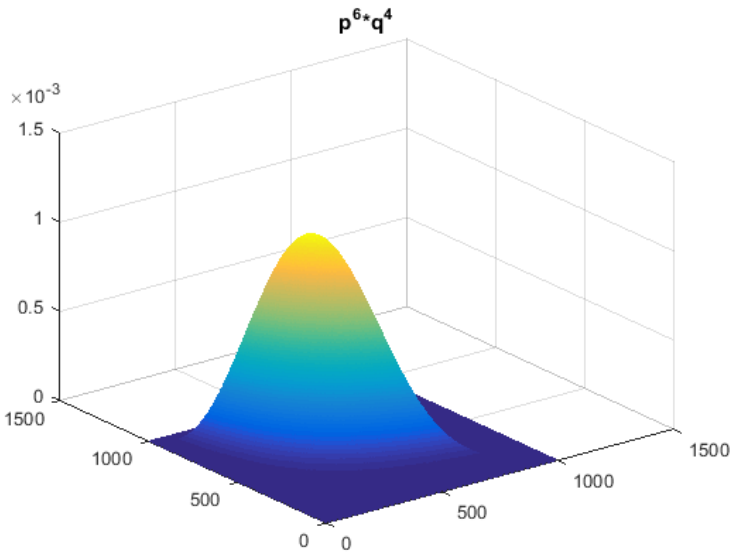
$$\underline{L}(\mathbf{x} = \{x_1, ..., x_n\}) = Bel_{\mathbb{X}}(\{x_1\}) \cdot \dots \cdot Bel_{\mathbb{X}}(\{x_n\}) = p^k q^{n-k}$$

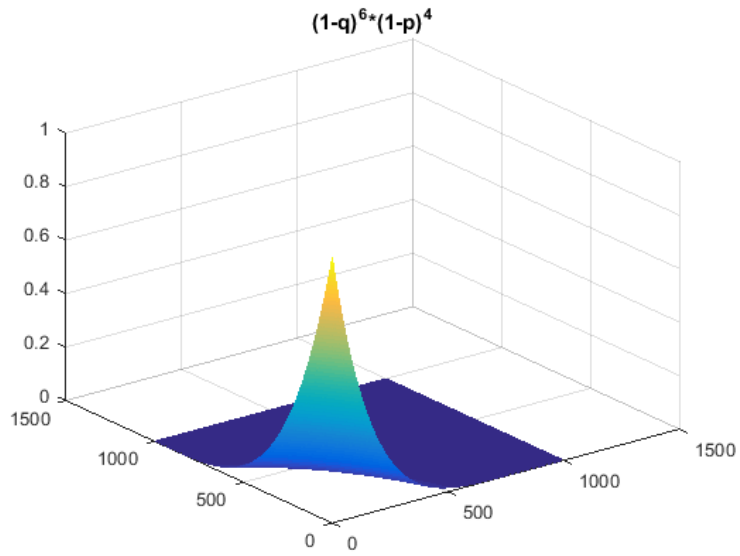
$$\overline{L}(\mathbf{x} = \{x_1, ..., x_n\}) = Pl_{\mathbb{X}}(\{x_1\}) \cdot \dots \cdot Pl_{\mathbb{X}}(\{x_n\}) = (1-q)^k (1-p)^{n-k}$$

- after normalisation, **these are PDFs over the space $\mathcal{B}$ of all belief functions definable on $\mathbb{X}$!**

Lower and upper likelihoods

Bernoulli trials





- both subsume to the traditional likelihood $p^k(1-p)^{n-k}$ for $p+q=1$
- the maximum of the lower likelihood is the traditional ML estimate
  - makes sense: the lower likelihood is highest for the most committed BFs (probabilities)
- upper likelihood (right) has maximum in $p=q=0$ (the vacuous BF on $\{1,0\}$)
- the interval of BFs joining $\max \overline{L}$ with $\max \underline{L}$ is the set of belief functions such that $\frac{p}{q} = \frac{k}{n-k}$, those which **preserve the ratio between the empirical counts**
- results lead us to **think in terms of intervals of belief functions** [cfr. Denouex IBFs]

Logistic regression

- Bernoulli trials are central in statistics: generalising their likelihood allow us to represent uncertainty in a number of regression problems
- in logistic regression, a logistic dependency between data and conditional probability of success is assumed:

$$p_i = P(Y_i = 1|x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}, \quad 1 - p_i = P(Y_i = 0|x_i) = \frac{e^{-(\beta_0 + \beta_1 x_i)}}{1 + e^{-(\beta_0 + \beta_1 x_i)}}$$

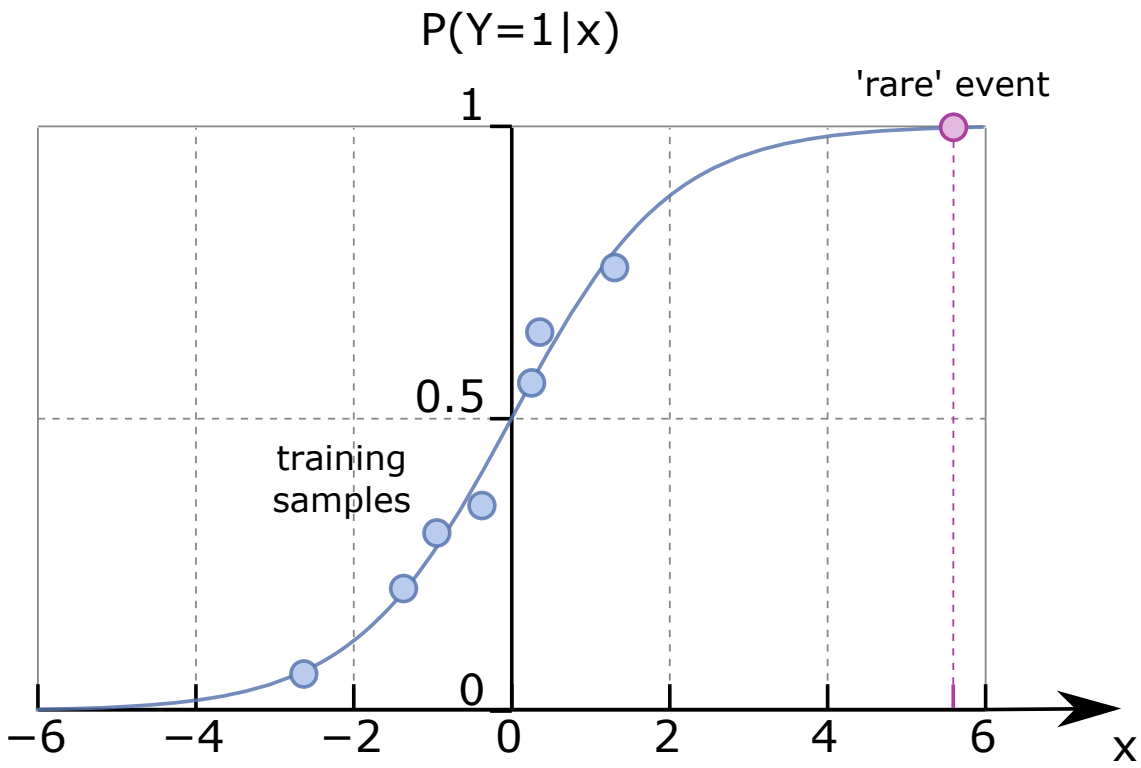
- the parameters β_0, β_1 are estimated by maximum likelihood of the sample, where

$$L(\beta_0, \beta_1 | Y) = \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1 - Y_i}$$

  $Y_i \in \{0, 1\}$ and p_i is a function of β_0, β_1 – yielding a single conditional PDF
- as in the Bernoulli series experiment, we can replace the conditional probability $(p_i, 1 - p_i)$ on $\mathbb{X} = \{1, 0\}$ with a belief function there!

Logistic regression

for rare event estimation



- events are called **rare** when they occur in a region of the sample space which is seldom sampled (e.g. we have no training samples there)
- logistic regression tends to underestimate the probability of success there [King]

Generalised logistic regression

- upper and lower likelihoods can then be computed as

$$\underline{L}(\beta|Y) = \prod_{i=1}^{n} p_i^{Y_i} q_i^{1-Y_i}, \quad \overline{L}(\beta|Y) = \prod_{i=1}^{n} (1-q_i)^{Y_i} (1-p_i)^{1-Y_i}$$

  where $p_i = Bel_i(\{1\})$, $q_i = Bel_i(\{0\})$
- this time the Bel_i are not equally distributed
- we need a logistic-like dependency for both p_i and q_i, this time
- first simple proposal: add a parameter β_2 such that

$$q_i = m(Y_i = 0|x_i) = \beta_2 \frac{e^{-(\beta_0+\beta_1 x_i)}}{1 + e^{-(\beta_0+\beta_1 x_i)}}$$

- we can then find lower and upper optimal estimates for the parameters β

$$\arg\max_{\beta} \underline{L} \mapsto \underline{\beta}_0, \underline{\beta}_1, \underline{\beta}_2 \qquad \arg\max_{\beta} \overline{L} \mapsto \overline{\beta}_0, \overline{\beta}_1, \overline{\beta}_2$$

- the solution is an upper and a lower family of conditional belief functions given x

$$Bel_{\mathbb{X}}(.|\underline{\beta}, x) \quad Bel_{\mathbb{X}}(.|\overline{\beta}, x)$$

- applied to rare event estimation, produces cautious estimates which incorporate our uncertainty about the model itself

Hypothesis testing with random sets

- in hypothesis testing, designing an experiment amounts to choosing a family of probability distributions generating the data
- if parameterised families of random sets can be contructed, they can be plugged in to the frequentist inference machinery (step 2 below)

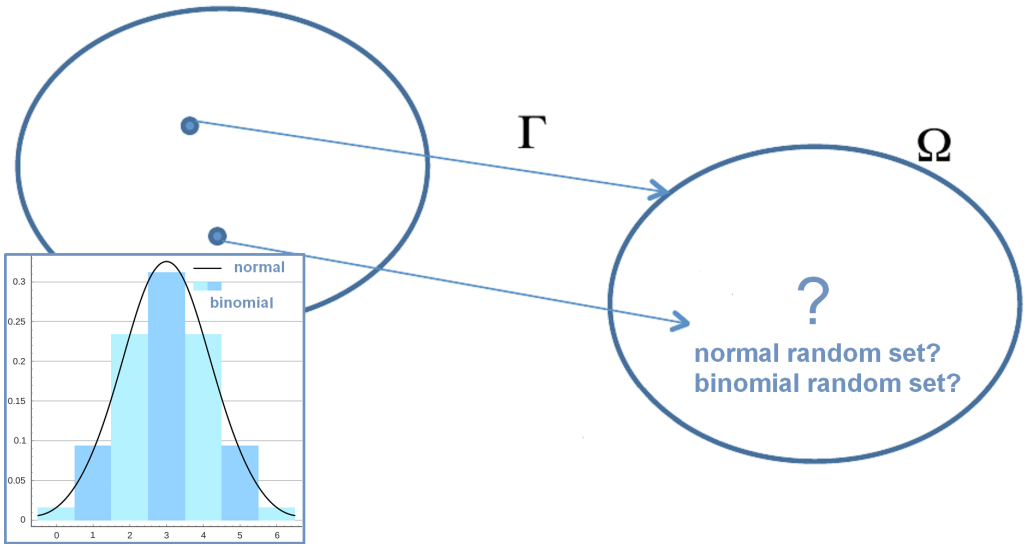
1. state relevant null H_0 and alternative hypotheses
2. state (e.g.) assumptions about the form of the ~~distributions~~ **random set (mass assignment)** describing the observations
3. state the relevant test statistic T (a quantity derived from the sample) $\rightarrow$ **this time the sample contains set-valued observations!**
4. derive the ~~distribution~~ **mass assignment** of the test statistic under the null hypothesis (from the assumptions)
5. set a significance level (α)
6. compute from the observations the observed value t_{obs} of the test statistic $T \rightarrow$ **this will also be set-valued**
7. calculate the ~~p-value, the probability~~ **conditional belief value** under H_0 of sampling a test statistic at least as extreme as the observed value
8. Reject the null hypothesis, in favor of the alternative hypothesis, if and only if ~~the p-value~~ such **conditional belief value** is less than the significance level

Parameterised families of random sets

- should the multivalued mapping Γ be modelled, or derived from the problem?
- e.g.: in the cloaked die example, it is the occlusion which generates the mapping
- it may make sense to model a parameterised family of multivalued mappings

$$\Gamma(.|\theta) : \Omega \to 2^{\Theta}$$

given a (fixed) probability on Ω, this yields a parameterised family of random sets



- the other option is to fix the multi-valued mapping ..
- .. and have the source probability vary with a certain parameterised distribution
- this will also induce a family of random sets

- the 'weak belief' approach is an attempt to do this, really

Central limit theorems for random sets

- the Gaussian distribution is central in probability theory and its applications ('normal' distribution)
  - is the PDF with maximum entropy, among those with given mean and σ
  - central limit theorem shows that all sums of iid random variables is Gaussian
  - whenever test statistics or estimators are functions of sums of random variables, they will have asymptotical normal distributions
- the old proposal by Dempster and Liu merely transfers normal distributions on the real line by Cartesian product with $\mathbb{R}^m$
- other avenue of research: **a central limit theorem for random sets**
- central limit theorem and law(s) of large numbers have been generalised to imprecise probabilities: Introduction to Imprecise Probabilities
- Larry G. Epstein & Kyoungwon Seo (Boston University) [2011]: A Central Limit Theorem for Belief Functions
- Xiaomin Shi (Shandong University) [2015]: Central limit theorems for belief measures

Central limit theorems for random sets

Shi 2015

Shi's central limit theorem

Let Y_i, $i = 1, ...$ equally distributed random variables on $(\Theta, \mathcal{B}(\Theta))$, bounded by a constant M, and denote by $X_i(\theta_1, \theta_2, ..) \doteq Y_i(\theta_i)$, $i \geq 1$. Then we have, for all $\alpha \in \mathbb{R}$:

$$\lim_{n\to\infty} Bel^\infty\left(\frac{\sum_{i=1}^n X_i - n\underline{\mu}}{\sqrt{n}\underline{\sigma}} \geq \alpha\right) = 1 - \mathcal{N}(\alpha) = \lim_{n\to\infty} Bel^\infty\left(\frac{\sum_{i=1}^n X_i - n\overline{\mu}}{\sqrt{n}\overline{\sigma}} < \alpha\right)$$

where for any $A \in \mathcal{B}(\Theta^\infty)$ Bel^∞ is the unique belief measure on $(\Theta^\infty, \mathcal{B}(\Theta^\infty)$

$$Bel^\infty(A) = P^\infty_{Bel}(\{K = K_1 \times K_2 \times ... \in (\mathcal{K}(\Theta))^\infty : K \subset A\}),$$

induced by the i.i.d. product probability measure P^∞_{Bel}, $\mathcal{K}(\Theta)$ is the collection of all compact subspaces of Θ, $\underline{\mu} = E_{P^\infty_{Bel}}[\underline{Z}_i]$, $\overline{\mu} = E_{P^\infty_{Bel}}[\overline{Z}_i]$ and

$$\underline{Z}_i(K_1 \times K_2 \times ...) = \inf_{\theta_i \in K_i} X_i(\theta_1, \theta_2, ...), \quad \overline{Z}_i(K_1 \times K_2 \times ...) = \sup_{\theta_i \in K_i} X_i(\theta_1, \theta_2, ...),$$

for $K_i \in \mathcal{K}(\Theta)$, $i \geq 1$. The expressions for $\underline{\sigma}$, $\overline{\sigma}$ are given in [Shi 2015], Th 3.1.

Generalised laws of large numbers

- the laws of large numbers have also been generalised to belief functions, random sets and monotone capacities [Kramosil 1998, Tern 2014-15, Kerkvliet 2017]

Tern's law of large numbers

Let ν be a completely monotone set function on a measurable space $(\Theta, \mathcal{B})$. Let X be an integrable[a] random variable, and $\{X_n, n\}$ pairwise independent and identically distributed to X. If ν is a capacity, then for every $\epsilon > 0$:

$$\nu\left(E_\nu[X] - \epsilon < \overline{X}_n < E_{\overline{\nu}}[X] + \epsilon\right) \to 1.$$

[a] I.e., its Choquet's lower $E_\nu[X]$ and upper $E_{\overline{\nu}}[X]$ expectations are finite.

- Choquet's integral of a random variable X against a capacity ν:

$$\int X d\nu \doteq \int_0^\infty \nu(\{X \geq t\})dt - \int_{-\infty}^0 [1 - \nu(\{X \geq t\})]dt$$

exists due to the monotonicity of ν

The total belief theorem

Generalising total probability to belief functions

- the generalisation of total probability exists for Walley's imprecise probabilities: it is called **marginal extension**
- however, natural and marginal extensions are not closed operators in the space of belief functions: when applied to a random set the result is not a random set

Theorem

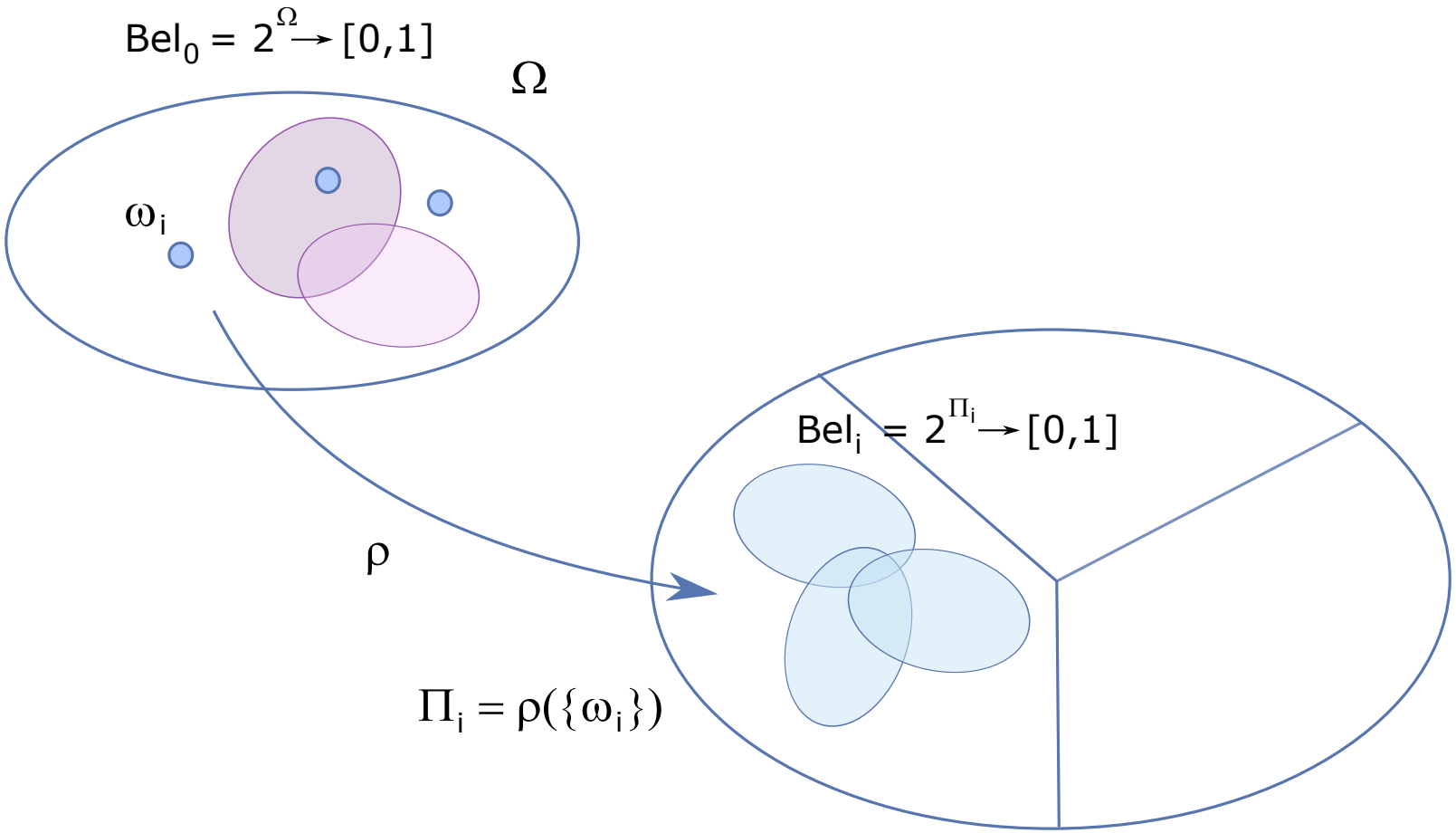
Suppose Θ *and* Ω *are two frames of discernment, and* $\rho : 2^\Omega \to 2^\Theta$ *the unique refining between them. Let* Bel_0 *be a belief function defined over* $\Omega = \{\omega_1, ..., \omega_{|\Omega|}\}$. *Suppose there exists a collection of belief functions* $Bel_i : 2^{\Pi_i} \to [0, 1]$, *where* $\Pi = \{\Pi_1, ..., \Pi_{|\Omega|}\}$, $\Pi_i = \rho(\{\omega_i\})$, *is the partition of* Θ *induced by its coarsening* Ω.
Then, there exists a belief function $Bel : 2^\Theta \to [0, 1]$ *such that:*

1. Bel_0 *is the restriction of* Bel *to* Ω
2. $Bel \oplus Bel_{\Pi_i} = Bel_i \ \forall i = 1, ..., |\Omega|$, *where* Bel_{Π_i} *is the logical belief function with mass*

$$m_{\Pi_i}(A) = 1 \ A = \Pi_i, \ 0 \ otherwise$$

The total belief theorem

Visual representation



- pictorial representation of the total belief theorem

The total belief theorem

Existence of a solution

- assume $\Theta' \supseteq \Theta$, and m a mass function over Θ
- m can be identified with a mass function $\overrightarrow{m}_{\Theta'}$ over the larger frame Θ': for any $E' \subseteq \Theta'$, $\overrightarrow{m}_{\Theta'}(E') = m(E)$ if $E' = E \cup (\Theta' \setminus \Theta)$ and $\overrightarrow{m}_{\Theta'}(E') = 0$ otherwise
- such $\overrightarrow{m}_{\Theta'}$ is called the **conditional embedding** of m into Θ'
- let $\overrightarrow{Bel_i}$ be the conditional embedding of Bel_i into Θ for all $Bel_i : 2^{\Pi_i} \to [0, 1]$, and

$$\overrightarrow{Bel} = \overrightarrow{Bel_1} \oplus \cdots \oplus \overrightarrow{Bel_{|\Omega|}}$$

Total belief theorem: existence [Zhou & Cuzzolin, UAI 2017]

The belief function

$$Bel \doteq Bel_0^{\uparrow\Theta} \oplus \overrightarrow{Bel}$$

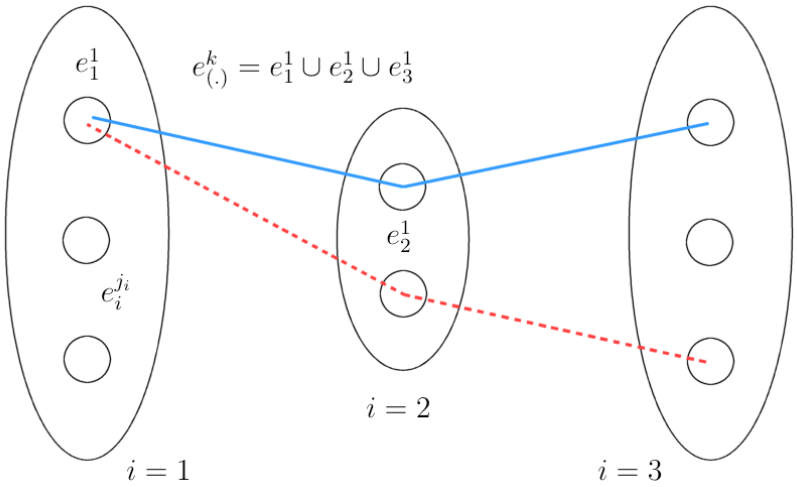
is a valid total belief function

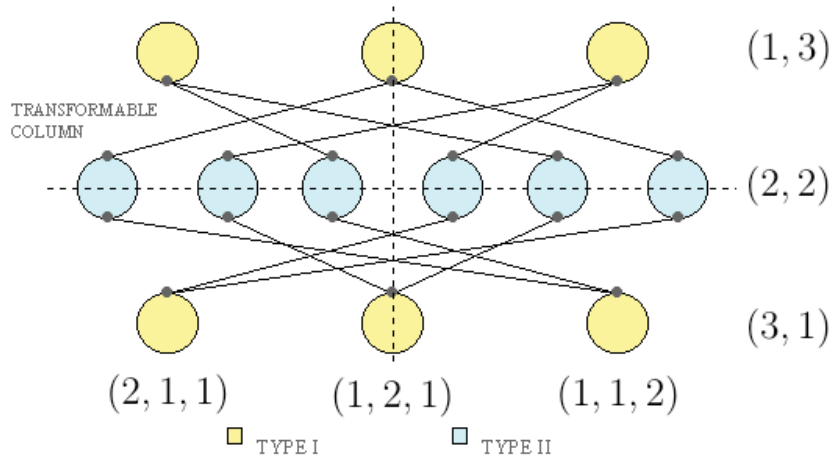
- recalls a result by Spies of generalised Jeffrey's rule

Structure of the solutions

to the total belief theorem

- unlike the classical law of total probability, there are (likely) multiple solutions
- pictorial representation of the structure of the FEs of a total BF *Bel* lying in the image of a focal element of Bel_0 of cardinality 3





- potential solutions correspond to square linear systems, and form a graph whose nodes are linked by linear transformations of columns: $e \mapsto e' = -e + \sum_{i \in \mathcal{C}} e_i - \sum_{j \in \mathcal{S}} e_j$
- at each transformation, the most negative component decreases
- general solution based on simplex-like optimisation?

Random set random variables?

- can we actually build a random variable using as a basis a random set on Ω instead of a probability measure there?
- as usual we need a mapping from Θ to a measurable space (e.g. the real line):
- for a continuous random variable X we can compute its probability density function (PDF) as its **Radon-Nikodym derivative** p such that

$$P[X \in A] = \int_A p d\mu$$

- can we compute a **(generalised) PDF for a random set random variable**?
- the Radon-Nikodym derivative for set functions was studied first by Harding et al [1997]
- Yann Rebille [2009]: A Radon-Nikodym derivative for almost subadditive set functions
- Graf: problem of defining the RND for capacities (rather than probability measures)
- assume capacities μ, ν are monotone, subadditive and continuous from below

Absolute continuity

A capacity ν is *absolutely continuous* with respect to another capacity μ if, for every $A \in \mathcal{F}$, $\nu(A) = 0$ whenever $\phi(A) = 0$.

- same definition as for standard measures
- for standard measures it is equivalent to the integral relation $\mu = \int \nu$

Generalising the Radon-Nikodym derivative

Strong decomposition

The pair (μ, ν) has the strong decomposition property if $\forall \alpha \geq 0$ there exists a measurable set $A_\alpha \in \mathcal{F}$ such that

$$\alpha(\nu(A) - \nu(B)) \leq \mu(A) - \mu(B) \quad \text{if} B \subset A \subset A_\alpha,$$
$$\alpha(\nu(A) - \nu(A \cap A_\alpha)) \geq \mu(A) - \mu(A \cap A_\alpha) \quad \forall A.$$

Radon-Nikodym theorem for capacities

For every two capacities μ and ν, ν is an indefinite integral of μ iff (μ, ν) has the strong decomposition property *and* ν is absolutely continuous with respect to μ.

- **open problems**:
- interpreting the conditions of the theorem (which holds for general capacities) for completely alternating capacities (distributions of random closed sets)
- Nguyen: a constructive approach to RN derivatives for capacities of random sets, similar to the one in constructive measure theory based on derivatives of set functions [Shilov]

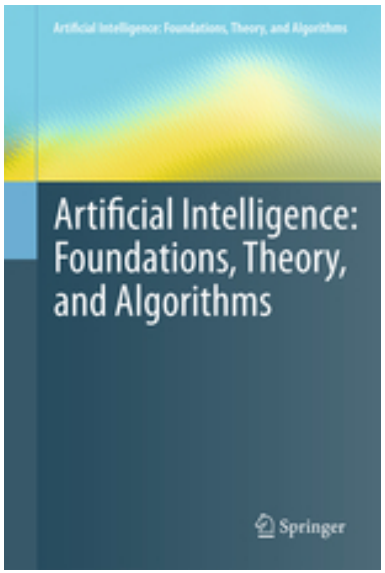
Outline

1. Uncertainty and probability
   - Nature of uncertainty
   - Mathematical probability
   - Interpretations of probability
   - Bayesian interpretation
   - Bayesians vs frequentists
2. Beyond probability
   - Missing data
   - Propositional data
   - Scarce data
   - Pure or no data
   - Rare data
   - Uncertain data
3. Statistics of belief functions
   - Dempster's multi-valued setting
   - Monotone capacities
   - Inner measures
   - Conditional multi-valued mappings
   - Statistics of experts' opinions
   - Probabilistic multivalued RVs
   - Gaussian belief functions
   - Measure-theoretical analysis
   - Frequentist interpretations
   - Dempster's a-equations
   - Fiducial inference and weak beliefs
   - Random set theory
4. New horizons
   - Upper and lower likelihood
   - Generalising logistic regression
   - Frequentist inference with RS
   - Generalised laws
   - The total belief theorem
   - Random set random variables
5. Summarising

What did we learn

- the rationale for a theory of belief functions in terms of data
- how belief theory is placed with respect to measure-theoretic probability and the frequentist and Bayesian approaches to probability
- the various equivalent mathematical definitions of belief functions
- the statistical frameworks and analysis for belief function theory developed along the years
  - evolutions of the multi-valued mapping idea
  - frequentists interpretations
  - fiducial interpretations
  - random set theory
- some open avenues to investigate for a fully fledged theory of statistical inference with belief functions
  - generalised likelihood, logistic regression
  - limit theorem, total probability for random sets
  - random set random variables and processes
  - frequentist inference with random sets

For Further Reading I



F. Cuzzolin

The geometry of uncertainty - The geometry of imprecise probabilities

Springer-Verlag (upcoming)