Safe Probability

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematical Institute – Leiden University

Safe Probability (Informal)

A probability distribution $P^{\circ}(X, Y)$ is called **safe** for prediction of random variable Y given X relative to a class of decision problems $\mathcal{D}$ if:

> for every problem in $\mathcal{D}$, acting as if P° were correct will have the same consequences as it would if P° were indeed **correct** - even though P° may in fact be **wrong**

Safe Probability (Informal)

A probability distribution $P°(X,Y)$ is called **safe** for prediction of random variable Y given X relative to a class of decision problems $\mathcal{D}$ if:

for every problem in $\mathcal{D}$, acting as if $P°$ were correct (taking the action that maximizes expected utility under $P°$) were will have the same consequences as it would if $P°$ were indeed **correct** - even though $P°$ may in fact be **wrong**

Safe Probability (Informal)

A probability distribution $P^{\circ}(X, Y)$ is called **safe** for prediction of random variable Y given X relative to a class of decision problems $\mathcal{D}$ if:

for every problem in $\mathcal{D}$, acting as if P° were correct will have the same consequences as it would if P° were indeed **correct** - even though P° may in fact be **wrong**

We call P° a **pragmatic** distribution

Why develop Safe Probability?

- explains why people/algorithms often do something seemingly wrong and get away with it (or not)

- clarifies much of the discussion between subjective and objective Bayesians, entropy maximizers, imprecise probabilists, fiducialists, and frequentists
  - and points towards a unified view?

Why Safe Probability at PROGIC?

- For logicians: may point towards new (?) pragmatic concept of truth (conditions)…
- For probabilists: points towards interesting generalization of measure theory

Safe Probability (Strong Version)

- A probability distribution $P^{\circ}(X, Y)$ is called **safe** for prediction of random variable Y given X relative to a class of decision problems $\mathcal{D}$ if:

  for every problem in $\mathcal{D}$, acting as if P° were correct will have exactly the same consequences as it would if P° were **correct** - even though P°may in fact be **wrong**

Menu

1. **Example 1: Marginals**
   - Extends to calibrated P°
2. Example 2: Monty Hall
   - P° not a marginal/not calibrated, still safe
3. Example 3: Objective Bayes, Jeffreys' Prior
4. A Unification of Bayesian, Imprecise Probability and Frequentist Ideas?

Safety with Marginal Probabilities

When Ignorance is Bliss, G. & Halpern ’04

- Suppose you’re a doctor and a patient comes in with a light fever and a swollen cheek.
- You know from the literature that, in the general population, 90% of people with these symptoms have the mumps
- You also observe the patient’s gender. You know that the probability may very well depend significantly on the patient’s gender but you have no idea how exactly the relation goes…

Safety with Marginals

$$\mathcal{X} = \mathcal{Y} = \{0, 1\}, \mathcal{P}^* = \{P^*(X, Y) : P^*(Y = 1) = 0.9\}$$

- **Given:** marginal probability of Y. Y may depend on X, but we have no idea how
- **Task:** predict Y given X.
- Suppose we observe $X = 0$. Now conditional probability could be anything...

$$\mathcal{P}^*(Y = 1 \mid X = 0) := \{P(Y = 1 \mid X = 0) \ : P \in \mathcal{P}^*\} = [0, 1]$$

- Similarly if we observe $X = 1$:

$$\mathcal{P}^*(Y = 1 \mid X = 1) := \{P(Y = 1 \mid X = 1) \ : P \in \mathcal{P}^*\} = [0, 1]$$

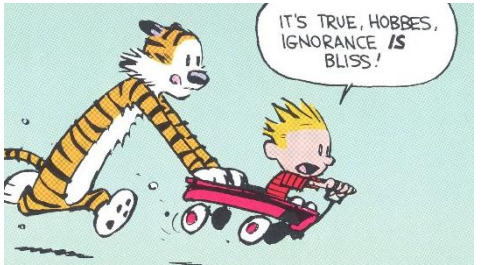
- How to predict? Reporting the full set of conditional probabilities given X does not give anything useful...

Safety with Marginals

- **Given:** marginal probability of Y. Y may depend on X, but we have no idea how
- **Task:** predict Y given X.
- Reporting the full set of conditional probabilities given X does not give anything useful...
- ...so what to do?

One option is to simply **ignore** X and predict Y with its marginal, irrespective of X , i.e. use

$$P^{\circ}(Y = 1 \mid X = 1) = P^{\circ}(Y = 1 \mid X = 0) = 0.9$$



First Safety Result

- Let now $\mathcal{X}, \mathcal{Y}$ be arbitrary, and suppose we know the marginal $P(Y)$. Let $\mathcal{P}^* = \{P^*(X,Y) : P^*(Y) = P(Y)\}$
- Use (=base decisions on) pragmatic P° with $P^\circ(Y|X = x) = P^*(Y)$, i.e. **act as if** X, Y independent
- For every loss function $L: \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, all $P^* \in \mathcal{P}^*$:

$$\mathbf{E}_{(X,Y)\sim P^\circ}\left[L\left(Y, \delta_{P^\circ|X}\right)\right] = \mathbf{E}_{(X,Y)\sim P^*}\left[L\left(Y, \delta_{P^\circ|X}\right)\right]$$

First Safety Result

- Let now $\mathcal{X}, \mathcal{Y}$ be arbitrary, and suppose we know the marginal $P(Y)$. Let $\mathcal{P}^* = \{P^*(X,Y) : P^*(Y) = P(Y)\}$
- Use (=base decisions on) pragmatic P° with $P^\circ(Y|X = x) = P^*(Y)$, i.e. **act as if** X, Y independent
- For every loss function $L: \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, all $P^* \in \mathcal{P}^*$:

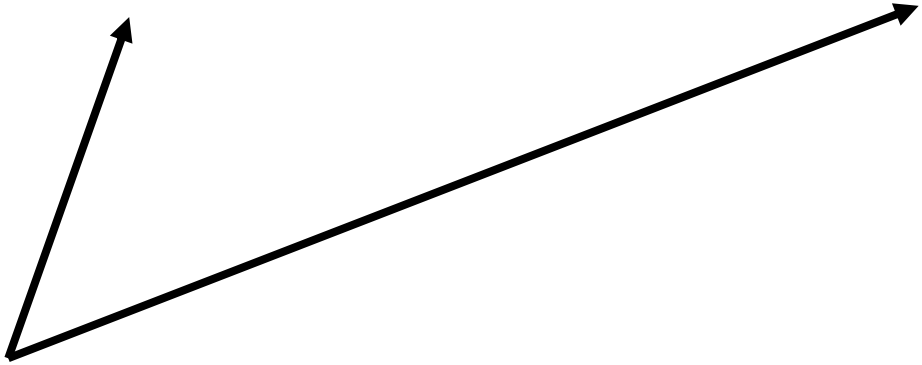
$$\mathbf{E}_{(X,Y)\sim P^\circ}\left[L\left(Y, \delta_{P^\circ|X}\right)\right] = \mathbf{E}_{(X,Y)\sim P^*}\left[L\left(Y, \delta_{P^\circ|X}\right)\right]$$

- We say that predicting Y using $P^\circ|X$ is **safe** for every fixed loss function L that depends on Y only.

Safe Decision Problems

For every loss function $L: \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, all $P^* \in \mathcal{P}^*$:

$$\mathbf{E}_{(X,Y)\sim P^\circ}\left[L\left(Y, \delta_{P^\circ|X}\right)\right] = \mathbf{E}_{(X,Y)\sim P^*}\left[L\left(Y, \delta_{P^\circ|X}\right)\right]$$



Bayes act $\delta_{P^\circ} = \arg\min_{a\in\mathcal{A}} \mathbf{E}_{Y\sim P^\circ}[L(Y,a)]$

Safe Decision Problems

For every loss function $L: \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, all $P^* \in \mathcal{P}^*$:

$$\mathbf{E}_{(X,Y)\sim P^\circ}\left[L\left(Y, \delta_{P^\circ|X}\right)\right] = \mathbf{E}_{(X,Y)\sim P^*}\left[L\left(Y, \delta_{P^\circ|X}\right)\right]$$

Decision-Maker's pragmatic distribution

"true" distribution

Bayes act $\delta_{P^\circ} = \arg\min_{a\in\mathcal{A}} \mathbf{E}_{Y\sim P^\circ}[L(Y,a)]$

Example: 0/1-Loss

$$\mathbf{E}_{(X,Y)\sim P^\circ}\left[L(Y,\delta_{P^\circ|X})\right] = \mathbf{E}_{(X,Y)\sim P^*}\left[L(Y,\delta_{P^\circ|X})\right]$$

Example: $0/1$-loss:

$\mathcal{A} = \mathcal{Y}$, $L(y,a) = 0$ if $y = a$, 1 otherwise

With $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and $\mathcal{P}^*$ as before, for $x \in \{0,1\}$:

$$\delta_{P^\circ|X=x} = \arg\min_{a\in\mathcal{A}} \mathbf{E}_{Y\sim P^\circ|X=x}\left[L(Y,a)\right] = 1$$

and then above just says $P^\circ(Y \neq 1) = P^*(Y \neq 1)$

…but example extends to arbitrary $\mathcal{A}, \mathcal{Y}, L$

Data-Oriented Re-Interpretation

- Suppose we get an i.i.d. sample $(X_1, Y_1), (X_2, Y_2), \ldots$
- If we think that it comes from P°, we would use $\delta_{P^\circ|X_i}$ at time i and be very certain that in long run average loss $\frac{1}{n}\sum_{i=1}^{n} L(Y_i, \delta_{P^\circ|X_i})$ would converge to $\mathbf{E}_{(X,Y)\sim P^\circ}[L(Y, \delta_{P^\circ|X})]$
  - In our binary example we would be very certain to be correct about 90% of the time

Data-Oriented Re-Interpretation

- Suppose we get an i.i.d. sample $(X_1, Y_1), (X_2, Y_2), \ldots$
- If we think that it comes from P°, we would use $\delta_{P^\circ|X_i}$ at time i and be very certain that in long run average loss $\frac{1}{n}\sum_{i=1}^{n} L(Y_i, \delta_{P^\circ|X_i})$ would converge to $\mathbf{E}_{(X,Y)\sim P^\circ}[L(Y, \delta_{P^\circ|X})]$
  - In our binary example we would be very certain to be correct about 90% of the time
- Even though $P^*$is wrong, the above conclusions based on P° are still correct if data is from any $P^* \in \mathcal{P}^*$

If we act according to P°, the world behaves as if P° were correct, even though it is not

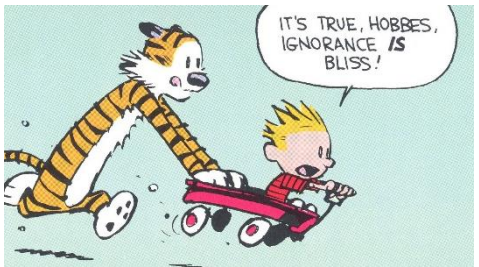
Remainder of this talk: much less trivial examples in which $P^\circ|X$ does depend on X yet is still not 'correct'

An Unsafe Decision Problem

- Suppose the relevant loss function may depend on X
- Then using P° that ignores X is unsafe, for we may very well have

$$\mathbf{E}_{(X,Y)\sim P^*}[L_{g(X)}(Y, \delta_{P^\circ|X})] \gg \mathbf{E}_{(X,Y)\sim \tilde{P}}[L_{g(X)}(Y, \delta_{P^\circ|X})]$$

- In our example, a misclassification of the disease may e.g. have worse consequences if X is female
- $L_0(y, y) = L_1(y, y) = 0$;
- $L_0(y, a) = 1, L_1(y, a) = 100$ if $y \neq a$
- e.g.…$g(x) = x$



Safe Probability

A probability distribution $P(X,Y)$ is called **safe** relative to a class of decision problems $\mathcal{D}$ if for every problem in $\mathcal{D}$, acting as if P were correct will have the same consequences as it would if P were **correct** - even though P may in fact be **wrong**

- i.e. for all $L \in \mathcal{D}$, all $P^* \in \mathcal{P}^*$ (one version of safety)

$$\mathbf{E}_{(X,Y)\sim P^*}\left[L\left(Y,\delta_{P^\circ|X}\right)\right] = \mathbf{E}_{(X,Y)\sim P^\circ}\left[L\left(Y,\delta_{P^\circ|X}\right)\right]$$

Safe Probability

A probability distribution $P(X,Y)$ is called **safe** relative to a class of decision problems $\mathcal{D}$ if for every problem in $\mathcal{D}$, acting as if P were correct will have the same consequences as it would if P were **correct** - even though P may in fact be **wrong**

- i.e. for all $L \in \mathcal{D}$, all $P^* \in \mathcal{P}^*$ (one version of safety)

$$\mathbf{E}_{(X,Y)\sim P^*}\left[L\left(Y,\delta_{P^\circ|X}\right)\right] = \mathbf{E}_{(X,Y)\sim P^\circ}\left[L\left(Y,\delta_{P^\circ|X}\right)\right]$$

- but we may also require instead (one-sided version):

$$\mathbf{E}_{(X,Y)\sim P^*}\left[L\left(Y,\delta_{P^\circ|X}\right)\right] \leq \mathbf{E}_{(X,Y)\sim P^\circ}\left[L\left(Y,\delta_{P^\circ|X}\right)\right]$$

Safe Probability

A probability distribution $P(X,Y)$ is called **safe** relative to a class of decision problems $\mathcal{D}$ if for every problem in $\mathcal{D}$, acting as if P were correct will have the same consequences as it would if P were **correct** - even though P may in fact be **wrong**

- i.e. for all $L \in \mathcal{D}$, all $P^* \in \mathcal{P}^*$ (one version of safety)

$$\mathbf{E}_{(X,Y)\sim P^*}\left[L\left(Y,\delta_{P^\circ|X}\right)\right] = \mathbf{E}_{(X,Y)\sim P^\circ}\left[L\left(Y,\delta_{P^\circ|X}\right)\right]$$

- but we may also require instead (one-sided version):

$$\mathbf{E}_{(X,Y)\sim P^*}\left[L\left(Y,\delta_{P^\circ|X}\right)\right] \leq \mathbf{E}_{(X,Y)\sim P^\circ}\left[L\left(Y,\delta_{P^\circ|X}\right)\right]$$

or (weaker version)

$$\mathbf{E}_{(X,Y)\sim P^*}\left[L\left(Y,\delta_{P^\circ|X}\right)\right] = \mathbf{E}_{X\sim P^*}\mathbf{E}_{Y\sim P^\circ|X}\left[L\left(Y,\delta_{P^\circ|X}\right)\right]$$

Safe Probability

- G. ‘18 describes various notions of safety, varying in wideness of $\mathcal{D}$ (from all decision problems to a single specific RV) and the meaning of consequences (use of expectation, inequality…)
- Strongest notion: *validity*
  - P° safe for every prediction problem that can be defined on your sample space
  - Frequentist would say “P° is true”, Bayesian would say “P° fully and correctly describes my beliefs”
- Weakest notion: *unbiasedness relative to fixed U*
  - $\mathbf{E}_{P^\circ}[U] = \mathbf{E}_{P^*}[U]$, otherwise nothing can be said

Safe Probability

Strongest safety notion: validity

Weakest notion: unbiasedness relative to fixed RV

…inbetween are several other kinds of safety, such as:

- calibration
  - weather forecasting!
  - previous example (ignoring X) is very special case
- fiducial/confidence safety
  - use (Fisher) fiducial and objective Bayes posteriors safely for some, but not all prediction tasks...
- …and more to be developed!

The Weather Forecaster

- A weather forecaster predicts daily precipitation probabilities $P^{\circ}(U = \text{rain} \mid V)$, based on measurements of air pressure and temperature taken all over the world
- V is **giant** vector. WF will probably not be able to give accurate predictions given the air pressure in Honolulu, although her predictions do depend thereon.
- We don't mind this, but we do want her to be **calibrated:**

  given that the **says** "the probability is p", it should be approximately p

Calibration is a form of Safety

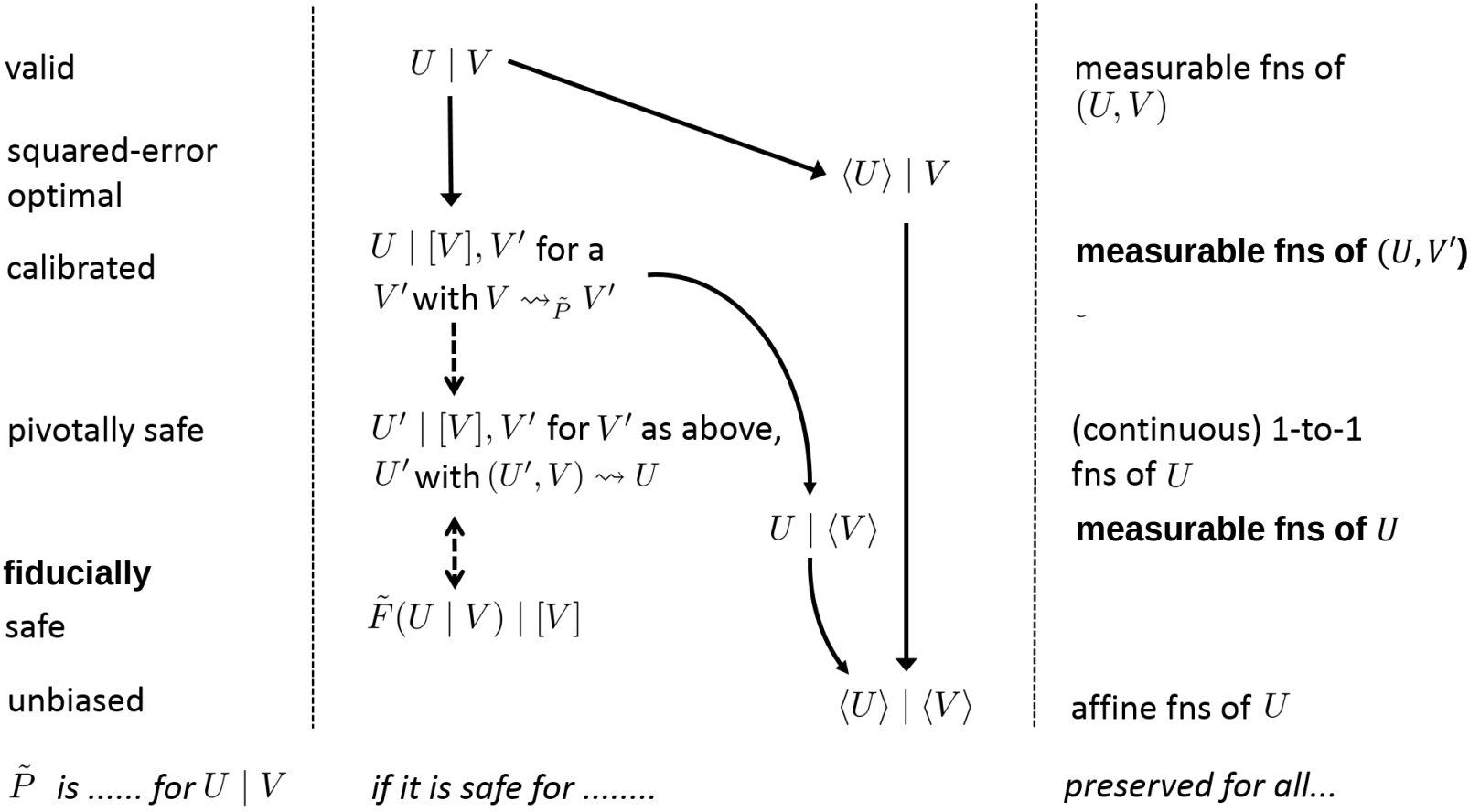
- We say that $P^\circ(U|V)$ is **calibrated**

  .....if for all $P^* \in \mathcal{P}^*, \vec{p} \in \text{range}(P^\circ(U \mid V))$

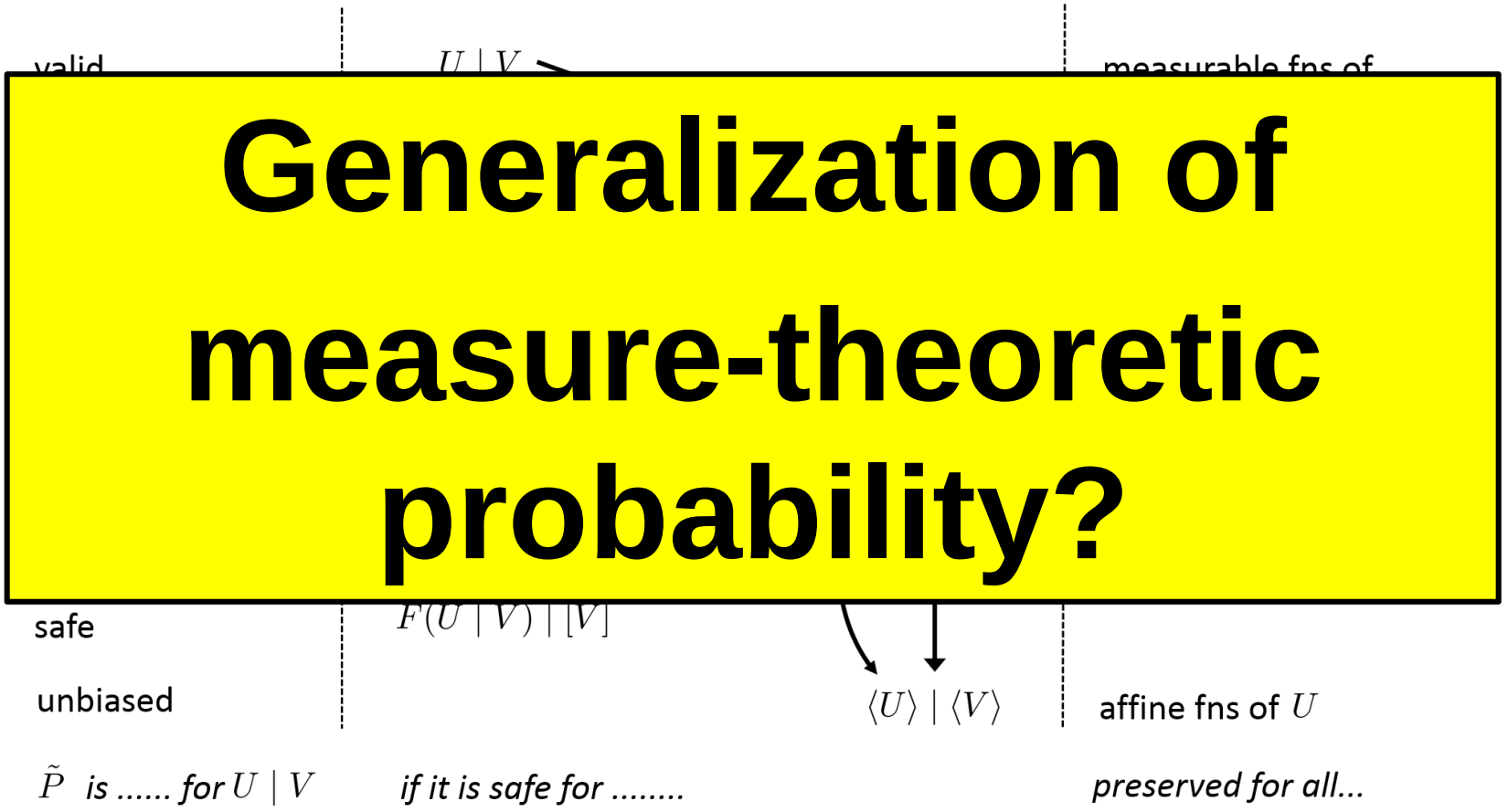
  $$P^*(U \mid P^\circ(U \mid V) = \vec{p}) = \vec{p}$$

- Calibration implies loss-safety as defined earlier for all loss functions that can be written as a function of U . Validity implies loss-safety for all loss functions that can be written as a function of (U, V).

Main Result of G. ’18: The Hierarchy



Main Result of G. ’18: The Hierarchy

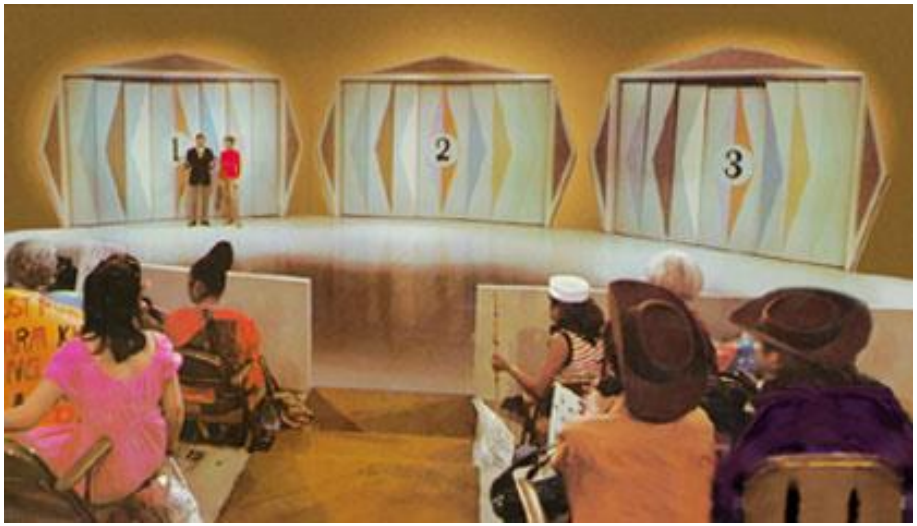


Generalization of measure-theoretic probability?

Menu

1. Example 1: Marginals
   - Extends to calibrated P°
2. **Example 2: Monty Hall**
   - P° not a marginal/not calibrated, still safe
3. Example 3: Objective Bayes, Jeffreys' Prior
4. A Unification of Bayesian, Imprecise Probability and Frequentist Ideas?

Monty Hall



- There are three doors in the TV studio. Behind one door is a car, behind both other doors a goat. You choose one of the doors. Monty Hall opens one of the other two doors, and shows that there is a goat behind it. You are now allowed to switch to the other door that is still closed. Is it smart to switch?

The Monty Hall Wikipedia Wars

(Gill 11, Mlodinow 08)

- Both sides **agree**:
  1. It is better to switch!
  2. To model problem correctly, you must take Monty's Protocol into account – what does Monty do when he has a choice?

The Monty Hall **Wikipedia Wars**

(Gill 11, Mlodinow 08)

- Both sides **agree**:
  1. It is better to switch!
  2. To model problem correctly, you must take Monty's Protocol into account – what does Monty do when he has a choice?
- "war" is about how to **prove** that switching is better:
  - "strictly Bayesian": via conditioning, with **additional assumption** that Monty chooses by tossing a fair coin
  - show that switching is a dominating strategy ("credal sets"/"imprecise probability")

Strict Bayesians vs Imprecise Probabilists

- **Strictly Bayesian**: Your uncertainty can be modeled by a single distribution
  - Ramsey (‘31), De Finetti (‘37), Savage (‘54), Cox* (‘61), …
  - Expected Utility in economics

- **Imprecise**: Your uncertainty can be modeled by a set of distributions
  - Keynes (1921), Seidenfeld (‘83), Walley (‘91), others…
  - Knightian Uncertainty
  - ‘Multiple Priors’ in economics

The Model on which both sides agree

- Suppose Contestant invariably chooses door a.
- Let RV Y denote location of car: $Y \in \{a, b, c\}$
- Let RV X denote Monty's action:

$$X \in \{\ \mathrm{open}(b),\ \mathrm{open}(c)\ \}$$

$X = \mathrm{open}(c)$ means Monty opens door c.
$X = \mathrm{open}(b)$ means Monty opens door b.

The Model on which they agree

- Suppose Contestant invariably chooses door a.
- Let Y denote location of car.
- Let X denote Monty's action.

$$P(Y = a) = P(Y = b) = P(Y = c) = \frac{1}{3}$$

$$P(X = \text{ open}(b) \mid Y = b) = P(X = \text{ open}(c) \mid Y = c) = 0$$

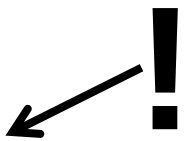
The Sets-of-Probabilities Side

- Suppose Contestant invariably chooses door a.
- Let Y denote location of car.
- Let X denote Monty's action.

$$P(Y = a) = P(Y = b) = P(Y = c) = \frac{1}{3}$$

$$P(X = \text{ open}(b) \mid Y = b) = P(X = \text{ open}(c) \mid Y = c) = 0$$

$$P(X = \text{ open}(b) \mid Y = \mathbf{a}) =$$
$$1 - P(X = \text{ open}(c) \mid Y = \mathbf{a}) \in [0, 1]$$



The Strict Bayesian Side

- Suppose Contestant invariably chooses door a.
- Let Y denote location of car.
- Let X denote Monty's action.

$$P(Y = a) = P(Y = b) = P(Y = c) = \frac{1}{3}$$

$$P(X = \text{open}(b) \mid Y = b) = P(X = \text{open}(c) \mid Y = c) = 0$$

$$P(X = \text{open}(b) \mid Y = \mathbf{a}) = P(X = \text{open}(c) \mid Y = \mathbf{a}) = \frac{1}{2}$$

Assuming an Unbiased Monty

- Let's be a strict Bayesian and **pretend that choices in protocol were made by fair coin tosses:**

$$P(X = \text{open}(b) \mid Y = \mathbf{a}) = P(X = \text{open}(c) \mid Y = \mathbf{a}) = \frac{1}{2}$$

...implying the familiar result

$$P(Y = a \mid X = \text{open}(b)) = \frac{1}{3}$$

$$P(Y = c \mid X = \text{open}(b)) = \frac{2}{3}$$

Assuming an Unbiased Monty...

- ..., i.e. use

$$P^{\circ}(Y = b \mid X = \text{open}(c)) := 2/3$$

This is

1. safe and
2. minimax optimal

....under all **symmetric** decision problems

Hence there is a useful middle ground between strict Bayes and imprecise probability

Safety

Unbiased Monty is

1. **"safe"** *under* all symmetric loss functions: for all $P^* \in \mathcal{P}^*$:

$$\mathbf{E}_{P^\circ}[\mathsf{Loss}(Y, \delta_{P^\circ}(X))] = \mathbf{E}_{P^*}[\mathsf{Loss}(Y, \delta_{P^\circ}(X))]$$

where

$\mathcal{A} =$ set of actions

$\mathsf{Loss} : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$

$\delta_{P^\circ}(x) := \arg\min_{q \in \mathcal{A}} E_{P^\circ|X=x}[\mathsf{Loss}(Y, q)] =$ Bayes act rel. to P°

Safety

Unbiased Monty is

1. **"safe"** *under* all symmetric loss functions: for all $P^* \in \mathcal{P}^*$:

$$\mathbf{E}_{P^\circ}[\mathsf{Loss}(Y, \delta_{P^\circ}(X))] = \mathbf{E}_{P^*}[\mathsf{Loss}(Y, \delta_{P^\circ}(X))]$$

Example:

$\mathcal{A} = \{a, b, c\}$

$\mathsf{Loss} : \mathcal{Y} \times \mathcal{A} \to \{0, 1\}$

$\mathsf{Loss}(Y, \widehat{y}) = \mathbf{1}_{Y \neq \widehat{y}}$

$\delta_{P^\circ}(\ \mathsf{open}(c)) = b\ ;\ \delta_{P^\circ}(\ \mathsf{open}(b)) = c.$

=1/3

Safety

credal set

Unbiased Monty is

1. **"safe"** *under* all symmetric loss functions: for all $P^* \in \mathcal{P}^*$:

$$\mathbf{E}_{P^\circ}[\text{Loss}(Y, \delta_{P^\circ}(X))] = \mathbf{E}_{P^*}[\text{Loss}(Y, \delta_{P^\circ}(X))]$$

Bayes act based on P°

'true' distribution

Decision-Maker's pragmatic distribution

Safety

Unbiased Monty is

1. **"safe"** *under* all symmetric loss functions: for all $P^* \in \mathcal{P}^*$:

$$\mathbf{E}_{P^\circ}[\mathrm{Loss}(Y, \delta_{P^\circ}(X))] = \mathbf{E}_{P^*}[\mathrm{Loss}(Y, \delta_{P^\circ}(X))]$$

Second Example: logarithmic scoring rule **=H(1/3)**

$\mathcal{A} =$ set of prob. mass fn. on $\{a, b, c\}$

$\mathrm{Loss} : \mathcal{Y} \times \mathcal{A} \to [0, \infty]$

$\mathrm{Loss}(Y, q) = -\log q(Y)$

$\delta_{P^\circ}(\text{ open}(c)) = \left(\frac{1}{3}, \frac{2}{3}, 0\right)$; $\delta_{P^\circ}(\text{ open}(b)) = \left(\frac{1}{3}, 0, \frac{2}{3}\right)$

Safety & Optimality

Unbiased Monty is

1. **"safe"** *under* all symmetric loss functions: for all $P^* \in \mathcal{P}^*$:

$$\mathbf{E}_{P^\circ}[\mathsf{Loss}(Y, \delta_{P^\circ}(X))] = \mathbf{E}_{P^*}[\mathsf{Loss}(Y, \delta_{P^\circ}(X))]$$

2. **minimax optimal:**

$$\max_{P^* \in \mathcal{P}^*} \mathbf{E}_{P^*}[\mathsf{Loss}(Y, \delta_{P^\circ}(X))] = \min_{\text{all } \delta} \max_{P^* \in \mathcal{P}^*} \mathbf{E}_{P^*}[\mathsf{Loss}(Y, \delta(X))]$$

Safety & Optimality

Unbiased Monty is

1. **"safe"** *under* all symmetric loss functions: for all $P^* \in \mathcal{P}^*$:

$$\mathbf{E}_{P^\circ}[\text{Loss}(Y, \delta_{P^\circ}(X))] = \mathbf{E}_{P^*}[\text{Loss}(Y, \delta_{P^\circ}(X))]$$

2. **minimax optimal:**

$$\max_{P^* \in \mathcal{P}^*} \mathbf{E}_{P^*}[\text{Loss}(Y, \delta_{P^\circ}(X))] = \min_{\text{all } \delta} \max_{P^* \in \mathcal{P}^*} \mathbf{E}_{P^*}[\text{Loss}(Y, \delta(X))]$$

3. "admissible"

1/3

"pretty adequate"

What about nonsymmetric losses?

- ‘asymmetric’ means e.g. that if the car is behind door B, it is a Ferrari; if it is behind door C, it is a Fiat Panda
- Now pretending that Monty chooses by tossing a fair coin is neither safe nor minimax optimal!
- In our first example, we had safety under all loss functions that were chosen independently of observation
- In this example, we additionally need symmetry : this safety is strictly weaker (applies to strictly smaller set of loss functions)

Maximum Entropy Monty?

- Natural question: can safe probabilities be understood as ‘maximum entropy’ probabilities?
- Answer: in some settings, inferring P° by MaxEnt is safe for an interestingly large class of decision problems, in other settings it simply has no good safety properties

Aside: a big frustration of mine…my first paper on safety-like notions was Maximum Entropy and the Glasses You are Looking Through G., *Proceedings UAI 2000*

This paper introduces ‘safe’, ‘risky’ and ‘silly’ uses of MaxEnt. By now, 21 years later, safety has come of age, but I’m still frustrated that that original paper was largely ignored…

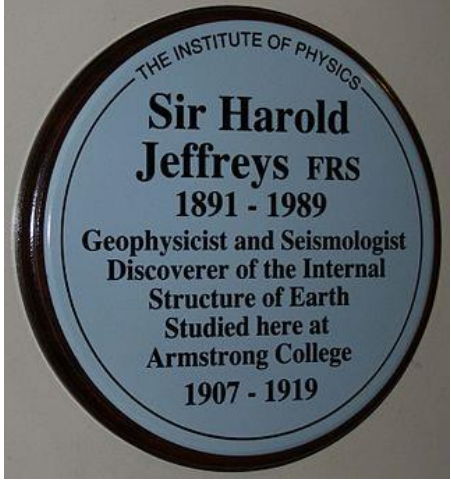
Menu

1. Example 1: Marginals
   - Extends to calibrated P°
2. Example 2: Monty Hall
   - P° not a marginal/not calibrated, still safe
3. Example 3: Objective Bayes, Jeffreys' Prior
4. A Unification of Bayesian, Imprecise Probability and Frequentist Ideas?

Objective Bayes

- Model $\mathcal{P}^* = \{P_\theta : \theta \in \Theta\}$ for data $Y^n = (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$
- Bayesian statistics proceeds by
  1. postulating prior distribution $w(\theta)$ on Θ …
  2. …thus constructing joint distribution P on $\Theta \times \mathcal{Y}^n$ with $P(\theta \in \Theta') = \int_{\Theta'} w(\theta) d\theta$ and $P(\, Y^n \mid \Theta = \theta \,) := P_\theta(Y^n)$

Objective Bayes



- Model $\mathcal{P}^* = \{P_\theta : \theta \in \Theta\}$ for data $Y^n = (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$
- Bayesian statistics proceeds by
  1. postulating prior distribution $w(\theta)$ on Θ …
  2. …thus constructing joint distribution P on $\Theta \times \mathcal{Y}^n$ with $P(\theta \in \Theta') = \int_{\Theta'} w(\theta) d\theta$ and $P(\, Y^n \mid \Theta = \theta \,) := P_\theta(Y^n)$
- **Objective Bayesians** claim*: if we have no clear prior knowledge about θ, we can use a special "default" prior w that represents "ignorance"
  - For 1-dimensional Θ, often claimed to be **Jeffreys' prior**
  - Sir Jeffreys ('46,'61), Bernardo ('79), Berger, …

Example: Bernoulli meets Jeffreys

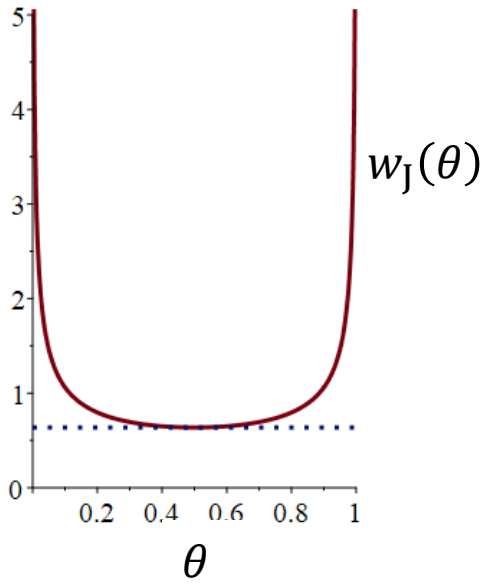
Model $\mathcal{P}^* = \{P_\theta : \theta \in \Theta\}$ for data $Y^n = (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$

Bernoulli model: $Y_i \in \{0,1\}$, $P_\theta(Y^n) = \theta^{n_1}(1-\theta)^{n_0}$

Jeffreys' prior for Bernoulli is $w_J(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}\pi}$

Bayes marginal distribution becomes

$P_J(Y^n) = \int_{0..1} w_J(\theta)\theta^{n_1}(1-\theta)^{n_0}\, d\theta$



Example: Bernoulli meets Jeffreys

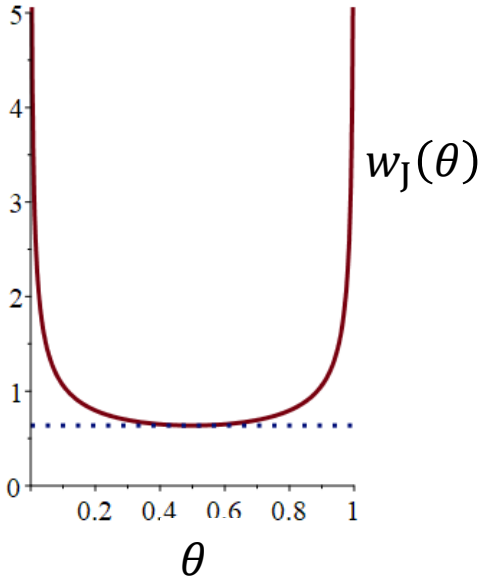
Model $\mathcal{P}^* = \{P_\theta : \theta \in \Theta\}$ for data $Y^n = (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$

Bernoulli model: $Y_i \in \{0,1\}$, $P_\theta(Y^n) = \theta^{n_1}(1-\theta)^{n_0}$

Jeffreys' prior for Bernoulli is $w_J(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}\pi}$

Bayes marginal distribution becomes

$P_J(Y^n) = \int_{0..1} w_J(\theta)\theta^{n_1}(1-\theta)^{n_0}\, d\theta$



Hard-core Objective Bayesian might say: **for any two random variables U, V that can be written as fn of Y^n, reasonable to predict U given V with $P_J(U|V)$.**

But this would give for example:

$$P_J\left(\text{frequency of } 1s \leq \frac{1}{10}\right) \approx 10 P_J(0.45 \leq \text{frequency of 1s} \leq 0.55)$$

Un-Safe Objective Bayes

- We have for large n,

$$P_{\mathrm{J}}\left(\text{frequency of } 1s \leq \frac{1}{10}\right) \approx 10\mathrm{P}_{\mathrm{J}}(0.45 \leq \text{frequency of } 1s \leq\ 0.55)$$

Thus, if you really want to follow the recommendation to predict with Jeffreys' prior, you would be willing to play the following game: 10000 outcomes will be generated ; then:

- If frequency of 1s is between 0.45-0.55, you pay 90$
- If between 0 and 0.05 you get 10$
- Otherwise nothing happens

Who in this zoom would actually want to play this game!?

Safe Objective Bayes

- So, is Jeffreys' prior all bad? No of course not.
- For example, we have the following safety-like property (important in model selection):

For all $\theta^* \in \Theta$, (i.e. all $P^* \in \mathcal{P}^*$),

$$\mathbf{E}_{\theta \sim W_J} \mathbf{E}_{Y^n \sim P_\theta} \left[\log \frac{P_\theta(Y^n)}{P_{W_J}(Y^n)} \right] = \mathbf{E}_{Y^n \sim P_{\theta^*}} \left[\log \frac{P_{\theta^*}(Y^n)}{P_{W_J}(Y^n)} \right] + o(1)$$

- Safe Probability can take the sting out of objective Bayes by 'blocking' some possible inferences but not others
- ...much more to say on safety and objective Bayes

Menu

1. Example 1: Marginals
   - Extends to calibrated P°
2. Example 2: Monty Hall
   - P° not a marginal/not calibrated, still safe
3. Example 3: Objective Bayes, Jeffreys' Prior
4. A Unification of Bayesian, Imprecise Probability and Frequentist Ideas?

Bayes, Imprecise, Frequentist at the same time?

- In this work we view a probability distribution primarily as a tool summarizing how you would act in various situations

Bayes, Imprecise, Frequentist at the same time?

- In this work we view a probability distribution primarily as a tool summarizing how you would act in various situations (≈ choose between various options)
  - very (subjective) Bayesian in spirit! (Savage)
  - but the non-Bayesian thing is the explicit restriction to a limited set of situations

Bayes, Imprecise, Frequentist at the same time?

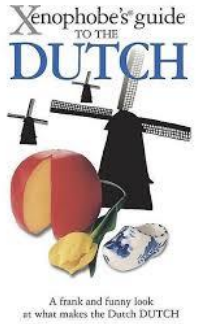
- In this work we view a probability distribution primarily as a tool summarizing how you would act in various situations (≈ choose between various options)
  - very (subjective) Bayesian in spirit!
  - but the non-Bayesian thing is the explicit restriction to a limited set of situations
- The very definitions of safety ensure that under the hood there are always sets of probabilities
  - imprecise probability!
- …and the interpretation of $\mathcal{P}^*$as possible 'truths' is frequentist….

But: aren't we contradicting Ramsey, De Finetti, Savage, …?

Ramsey, De Finetti, Savage, Cox, Fishburn and many others all gave strong reasons to act like a Bayesian.

e.g. when you advocate a non-Bayesian method, followers of De Finetti might say "but that is silly! I can make **Dutch boo**k against you!"

- Meaning a set of gambles all of which you'd accept but in the end would guarantee you a sure loss

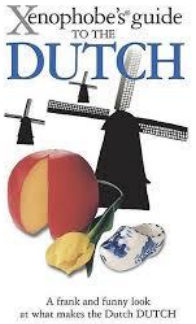


But: aren't we contradicting Ramsey, De Finetti, Savage, …?

Ramsey, De Finetti, Savage, Cox, Fishburn and many others all gave strong reasons to act like a Bayesian.

e.g. when you advocate a non-Bayesian method, followers of De Finetti might say "but that is silly! I can make **Dutch book** against you!"

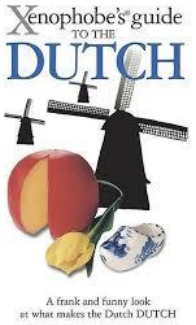
- Meaning a set of gambles all of which you'd accept but in the end would guarantee you a sure loss



More generally they stipulate a set of innocuous looking axioms that every reasonable decision maker should obey.

They then show that adhering to the axioms implies you are a Bayesian. So it may seem safe probability contradicts these axioms! **Does it!?!?**

But: aren't we contradicting Ramsey, De Finetti, Savage, …?



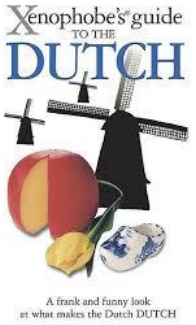
"being a Bayesian" ≈ there must a single (not a set!) distribution $P°$ on $\mathcal{X} \times \mathcal{Y}$ such that under **every** loss function $L: \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, you estimate the loss incurred with action a given $X = x$ to be

$$\mathbf{E}_{Y \sim P°|X=x}\left[L(x, Y, a)\right]$$

…and hence always prefer to play the Bayes act relative to $P°|X = x$ and L, minimizing expected loss

With safe probability, for a given $P°$ we are non-Bayesian in that we would block such an inference for some loss L (e.g. Monty Hall: nonsymmetric L)

But: aren't we contradicting Ramsey, De Finetti, Savage, …?



"being a Bayesian" ≈ there must a single (not a set!) distribution $P°$ on $\mathcal{X} \times \mathcal{Y}$ such that under **every** loss function $L: \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, you estimate the loss incurred with action a given $X = x$ to be

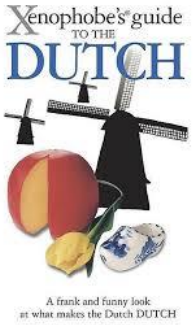
$$\mathbf{E}_{Y \sim P° | X = x} [L(x, Y, a)]$$

…and hence always prefer to play the Bayes act relative to $P°|X = x$ and L, minimizing expected loss

With safe probability, for a given $P°$ we are non-Bayesian in that we would block such an inference for some loss L (e.g. Monty Hall: nonsymmetric L)

Alternatively, we can also set up safe probability such that we choose $P°$ as a function of L !

But: aren't we contradicting Ramsey, De Finetti, Savage, …?



- We can set up safe probability such that we choose $P°$ as a function of L !
- Then $P°$ can be thought of as being conditional on L, and the same $P° \mid X = x,\ \bar{L} = L$ is used for making all decisions. You are then acting like a proper Bayesian!
- …even though you have may introduced dependencies (the choice of loss influencing the probabilities of outcomes Y) that you do not really think are there – but that does not contradict any of the typical axioms!

This argument is not water-tight yet (esp. not for Savage's axioms)

Collaborations welcome!

Why develop Safe Probability?

- explains why people (and algorithms) sometimes get away with doing something that's wrong (and sometimes don't)
- Clarifies much of the discussion between strict Bayesians, imprecise probabilists and frequentists (and points towards a unified view?)

Why Safe Probability at PROGIC?

- explains why people (and algorithms) sometimes get away with doing something that's wrong (and sometimes don't)
- Clarifies much of the discussion between strict Bayesians, imprecise probabilists and frequentists (and points towards a unified view?)

- For logicians: may point towards new (?) pragmatic concept of **truth** (conditions)…
- For probabilists: does point towards interesting generalization of **measure theory**

Read/Do more?

- G., **Safe Probability,** *Journal of Statistical Planning and Inference,* 2018 (full theory, too complicated...)
- G. & J. Halpern, Making Decisions using Sets of Probabilties, *Journal of AI Research,* 2011 ('pre-work')
- G., Maximum Entropy and the Glasses you are Looking Through, *Proceedings UAI 2000*
- Appendix of G. and T. van Ommen: Inconsistency of Bayesian Inference under Misspecification, *Bayesian Analysis,* 2017
- ...more examples (e.g. optional stopping) on Friday!