Safe Probability

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematisch Instituut – Universiteit Leiden

Prelude: Kelly Gambling

- Suppose we observe sequence $X_1, X_2, \ldots$ of 0s and 1s
- At each point in time i, we can buy a ticket $T_{i,1}$ that pays off \$2 iff $X_i = 1$, and a ticket $T_{i,0}$ that pays off \$2 iff $X_i = 0$. Both tickets cost \$1
- Crucially: we are allowed to divide our capital any way we like and re-invest our capital at each point in time
  - e.g. By putting 50% of your capital at time i on $T_{i,1}$ and 50% on $T_{i,0}$ you make sure that your capital remains the same

Prelude: Kelly Gambling

- At each time i, we can buy a ticket $T_{i,1}$ that pays off \$2 iff $X_i = 1$, and a ticket $T_{i,0}$ that pays off \$2 iff $X_i = 0$. Both tickets cost \$1
- A **gambling strategy** in this game is a function $q : \bigcup_{n \geq 0} \{0,1\}^n \to \Delta_2$ and thus defines a probability distr. on $\{0,1\}^\infty$ via setting

$$\tilde{P}(X_i = 1 \mid X^{i-1}) := q(1|X^{i-1}) \;\; ; \;\; \tilde{P}(X^n) := \prod_{i=1}^n \tilde{P}(X_i|X^{i-1})$$

- If we follow such a strategy and start with \$1, our capital after n rounds will be $\tilde{P}(X^n) \cdot 2^n$

How to design a gambling strategy?

- A **gambling strategy** in this game is formally equivalent to a probability distribution $\tilde{P}$ on infinite sequences. **Which strategy should we adopt?**

How to design a gambling strategy?

- A **gambling strategy** in this game is formally equivalent to a probability distribution $\tilde{P}$ on infinite sequences. **Which strategy should we adopt?**
- Strict Subjective Bayesian: *think* very long about the situation, come up with a subjective distribution P^*, and then play the distribution $\tilde{P}$ maximizing expected gain (we may have $\tilde{P} \neq P^*$)
- Imprecise Probabilist: come up with a *set* of distributions $\mathcal{P}^*$, and then play the distribution $\tilde{P}$ optimal relative to $\mathcal{P}^*$, with optimality defined relative to some additional criterion (which one?)

How to design a gambling strategy?

- Strict Subjective Bayesian: determine subjective P^*, and then play optimal $\tilde{P}$ (we may have $\tilde{P} \neq P^*$)
- Imprecise: determine *set* $\mathcal{P}^*$ and play "optimal" $\tilde{P}$
- Information Theorist: pick any gambling strategy which you think might gain you a lot. E.g. if you think frequency might converge to $p \neq 0.5$, you might play Laplace rule of succession...

$$\tilde{P}(X_{n+1} = 1 \mid X^n) = \tfrac{n_1+1}{n+2} \qquad \tilde{P}(X^n) = \int_0^1 p^{n_1}(1-p)^{n_0} dp$$

How to design a gambling strategy?

- Strict Subjective Bayesian: determine subjective P^*, and then play optimal $\tilde{P}$ (we may have $\tilde{P} \neq P^*$)
- Imprecise: determine *set* $\mathcal{P}^*$ and play "optimal" $\tilde{P}$
- Information Theorist: pick any gambling strategy which you think might gain you a lot. E.g. if you think frequency might converge to $p \neq 0.5$, you might play Laplace rule of succession...

$\tilde{P}(X_{n+1} = 1 \mid X^n) = \frac{n_1+1}{n+2}$ $\quad$ $\tilde{P}(X^n) = \int_0^1 p^{n_1}(1-p)^{n_0} dp$

...if your hypothesis about frequence is correct, you gain **exponential amount of money** even if **at the same time you think data are *not Bernoulli (or not even stationary)***

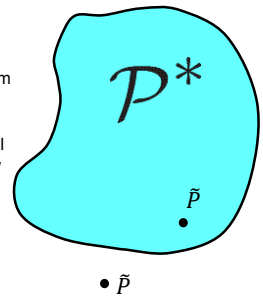
Starting Point

- Adopting a Bayesian predictive distribution like the Laplace Rule of Succession if you think data are not Bernoulli is o.k. (and I think, rational!) for some prediction tasks...
  - Sequential gambling, **Data Compression**

  ...but not for others:
  - 0/1-loss prediction (no fractional bets!) when you are only asked to predict X_i in the situation that $X_{i-1} = 1$
- I want to design a theory which can cope with such 'partially useable' distributions

A Middle Ground between strict Bayes and imprecise probability

- Set of distrs $\mathcal{P}^*$ has unique representative, as in 'objective Bayes', fiducial inference, Maximum Entropy, data compression...
- One absolutely crucial difference: we **restrict** use of $\tilde{P}$ to subset of all possible prediction tasks: we know in advance that $\tilde{P}$ should not be taken to seriously
- **Provides unifying and demistifying view**



Menu

1. The Setting
2. Definition 1, Example 1: Dilation
3. Definition 2, Example 1 cont.
4. Definition 3-4, Example 2: Calibration
5. Example 3: Fiducial Distributions
6. Desert: Monty Hall Problem, Decision Safety

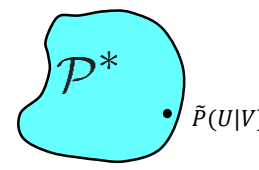
The Setting

- Let $\mathcal{P}^*$ be a set of distributions on a space Ω, representing Decision-Maker (DM)'s uncertainty about a domain
- DM has to make predictions/assertions about some U (or a function thereof), upon observing V. Both U and V are RVs (random variables) on Ω, taking values in $\mathcal{U}$ and $\mathcal{V}$, resp.
- She does so using a pragmatic distribution $\tilde{P}(U|V)$, defined as a conditional distribution of U given V , i.e. a function mapping each $v \in \mathcal{V}$ to a distribution $\tilde{P}(U \mid V = v)$ on $\mathcal{U}$
  - Whenever $\mathcal{U}$ finite, we think of $\tilde{P}(U) = (\tilde{p}(u_1), \ldots, \tilde{p}(u_m))^T$ as a column vector

The Setting

- A Bayesian would have $\mathcal{P}^* = \{P^*\}$ a singleton and could then set $\tilde{P}(U \mid V) = P^*(U \mid V)$
- Note that P^* is a distribution on Ω, inducing a joint $P^*(U, V)$ which in turn induces $P^*(U \mid V)$, while $\tilde{P}(U \mid V)$ is directly defined as a conditional

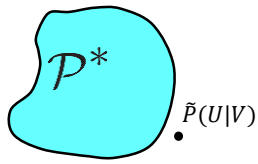
(hence $\tilde{P}$ in picture to be taken with grain of salt)



The Setting

- A Bayesian would have $\mathcal{P}^* = \{P^*\}$ a singleton and could then set $\tilde{P}(U \mid V) = P^*(U \mid V)$
- We have to do something else – sometimes eqv. to conditioning on a special element of $\mathcal{P}^*$, sometimes really different...

$\tilde{P}$ is really a probability update rule!!



First Definition: Weak Safety

- We say that $\tilde{P}(U|V)$ is **safe** for $\langle U \rangle | \langle V \rangle$ if for all $P^* \in \mathcal{P}$:

$$\mathrm{E}_{P^*}[U] = \mathrm{E}_{V \sim P^*} \mathrm{E}_{U \sim \tilde{P}|V}[U]$$

- i.e.

$$\sum_{u \in \mathcal{U}} p^*(u) \cdot u = \sum_{v \in \mathcal{V}} p^*(v) \sum_{u \in \mathcal{U}} \tilde{p}(u|v) \cdot u$$

First Definition

- We say that $\tilde{P}(U|V)$ is **safe** for $\langle U \rangle | \langle V \rangle$ if for all $P^* \in \mathcal{P}$:

$$\mathrm{E}_{P^*}[U] = \mathrm{E}_{V \sim P^*} \mathrm{E}_{U \sim \tilde{P}|V}[U]$$

- i.e.

$$\mathrm{E}_{V \sim P^*} \mathrm{E}_{U \sim P^*|V}[U] = \mathrm{E}_{V \sim P^*} \mathrm{E}_{U \sim \tilde{P}|V}[U]$$

First Definition

- We say that $\tilde{P}(U|V)$ is **safe** for $\langle U \rangle | \langle V \rangle$ if for all $P^* \in \mathcal{P}$:

$$\mathrm{E}_{P^*}[U] = \mathrm{E}_{V \sim P^*} \mathrm{E}_{U \sim \tilde{P}|V}[U]$$

- i.e. we can expect our expectation of U to be 'correct' (in a relative sense)
- we will usually want somewhat stronger versions of 'safety'

First Example: Dilation

$$\mathcal{U} = \mathcal{V} = \{0,1\}, \mathcal{P} = \{P^* : \mathrm{E}_{P^*}[U] = 0.9\}$$

- **Given:** marginal probability of U. U may depend on V, but we have no idea how
- **Task:** predict U given V.
- Suppose we observe $V = 0$. Now conditional probability could be anything...
  $\mathcal{P}(U \mid V = 0) := \{P \mid V = 0 \ : P \in \mathcal{P}\} = [0,1]$
- Similarly if we observe $V = 1$:
  $\mathcal{P}(U \mid V = 1) := \{P \mid V = 1 \ : P \in \mathcal{P}\} = [0,1]$

Dilation

Seidenfeld & Wasserman, '93

Before observing V we had precise probability $P^*(U = 1)$
after we only know $P^*(U = 1 \mid V = v)$ is in large superset

"extra information ⇒ less knowledge
no matter what you observe!"

Ignoring instead of Dilating

- Pointwise conditioning gives dilation
- Instead we may decide to ignore V, i.e. act as if U and V are independent, and predict with the pragmatic distribution

$$\tilde{P}(U = 1 \mid V = 1) = \tilde{P}(U = 1 \mid V = 0) = 0.9$$

- **Proposition:** $\tilde{P}(U|V)$ is **safe** for $\langle U \rangle | \langle V \rangle$

$$\mathrm{E}_{P^*}[U] = \mathrm{E}_{V \sim P^*} \mathrm{E}_{U \sim \tilde{P}|V}[U]$$

- i.e. $P^*(U = 1) = \mathrm{E}_{V \sim P^*}[\tilde{P}(U = 1 \mid V)]$

First Example of 'Safety'

- **REALITY**: U may be dependent on V
- **PRAGMATICS**: we nevertheless decide to predict U with a distribution that assumes U and V are independent

- Our predictions will be **just as accurate as we would expect them to be if our pragmatic distribution $\tilde{P}$ were 'correct'**
- ...as long as we only use $\tilde{P}$ only for certain, not all prediction tasks...

Definition 2, Preparation

- We write $X \rightsquigarrow Y$ if there exists a function ϕ such that $\phi(X) \equiv Y$ ("X **determines** Y ")

- $\tilde{P}(U|V)$ can be used to predict not just U, but also any U' determined by (U,V) , i.e. with $(U,V) \rightsquigarrow U'$:

$$\tilde{P}(U' = u' \mid V = v) := \sum_{u \in \mathcal{U}: \phi(u,v) = u'} \tilde{P}(U = u \mid V = v)$$

Definition 2

- Recall: $\tilde{P}(U|V)$ is **safe** for $\langle U' \rangle | \langle V \rangle$ if $(U,V) \rightsquigarrow U'$ and for all $P^* \in \mathcal{P}$:

$$\mathrm{E}_{P^*}[U'] = \mathrm{E}_{V \sim P^*} \mathrm{E}_{U' \sim \tilde{P}|V}[U']$$

- We say that $\tilde{P}(U|V)$ is **safe** for $U \mid \langle V \rangle$ if **for all** U' with $(U,V) \rightsquigarrow U'$, all $P^* \in \mathcal{P}$:

$$\mathrm{E}_{P^*}[U'] = \mathrm{E}_{V \sim P^*} \mathrm{E}_{U' \sim \tilde{P}|V}[U']$$

Definition 2

- Recall: $\tilde{P}(U|V)$ is **safe** for $\langle U' \rangle | \langle V \rangle$ if $(U,V) \rightsquigarrow U'$ and for all $P^* \in \mathcal{P}$:

$$\mathrm{E}_{P^*}[U'] = \mathrm{E}_{V \sim P^*} \mathrm{E}_{U' \sim \tilde{P}|V}[U']$$

- We say that $\tilde{P}(U|V)$ is **safe** for $U \mid \langle V \rangle$ if **for all** U' with $(U,V) \rightsquigarrow U'$, all $P^* \in \mathcal{P}$:

$$\mathrm{E}_{P^*}[U'] = \mathrm{E}_{V \sim P^*} \mathrm{E}_{U' \sim \tilde{P}|V}[U']$$

$$P^*(U) = \mathrm{E}_{V \sim P^*}[\tilde{P}(U \mid V)]$$

Example 1(b) - dilation again

$$\mathcal{U} = \mathcal{V} = \{0, 1, 2\}, \mathcal{P} = \{P^* : \mathrm{E}_{P^*}[U] = 0.9\}$$

- **Task:** predict U given V.
- Again we decide to ignore V and set e.g. for all $v \in \mathcal{V}$:

$$\tilde{P}(U = 1 \mid V = v) = 0.9, \tilde{P}(U = 2 \mid V = v) = 0$$

- Then $\tilde{P}$ is safe for $\langle U \rangle | \langle V \rangle$ but not for $U \mid \langle V \rangle$

Example 1(c)

$$\mathcal{U} = \mathcal{V} = \{0, 1, 2\}, \mathcal{P} = \{P^* : P^*(U) = (0, 0.3, 0.3)^T\}$$

- **Task:** predict U given V.
- Again we decide to ignore V and set e.g. for all $v \in \mathcal{V}$:

$$\tilde{P}(U = 1 \mid V = v) = 0.9, \tilde{P}(U = 2 \mid V = v) = 0$$

- Then, **again**, $\tilde{P}$ is safe for $\langle U \rangle | \langle V \rangle$ but not for $U | \langle V \rangle$

Example 1(c): use the marginal

$$\mathcal{U} = \mathcal{V} = \{0, 1, 2\}, \mathcal{P} = \{P^* : P^*(U) = (0, 0.3, 0.3)^T\}$$

- **Task:** predict U given V.
- Again we decide to ignore V and set e.g. for all $v \in \mathcal{V}$:

$$\tilde{P}(U \mid V = v) = P^*(U)$$

- Then $\tilde{P}$ is safe for $\langle U \rangle | \langle V \rangle$ and **also** for $U | \langle V \rangle$

Definition 3, Preparation

- Recall: $\tilde{P}(U|V)$ is **safe** for $\langle U' \rangle | \langle V \rangle$ if $(U, V) \rightsquigarrow U'$ and for all $P^* \in \mathcal{P}$:

$$\mathrm{E}_{P^*}[U'] = \mathrm{E}_{V \sim P^*} \mathrm{E}_{U' \sim \tilde{P}|V}[U']$$

- Leave out '$(U, V) \rightsquigarrow U'$' part from now on, for brevity

Definition 3

- Recall: $\tilde{P}(U|V)$ is **safe** for $\langle U' \rangle | \langle V \rangle$ if for all $P^* \in \mathcal{P}$:

$$\mathrm{E}_{P^*}[U'] = \mathrm{E}_{V \sim P^*} \mathrm{E}_{U' \sim \tilde{P}|V}[U']$$

- We say that $\tilde{P}(U|V)$ is **safe** for $\langle U' \rangle | \boldsymbol{V}$ if for all $P^* \in \mathcal{P}$:

$$\mathrm{E}_{P^*}[U' \mid V] \equiv \mathrm{E}_{\tilde{P}}[U' \mid V]$$

Definition 3

- Recall: $\tilde{P}(U|V)$ is **safe** for $\langle U' \rangle | \langle V \rangle$ if for all $P^* \in \mathcal{P}$:

$$\mathrm{E}_{P^*}[U'] = \mathrm{E}_{V \sim P^*} \mathrm{E}_{U' \sim \tilde{P}|V}[U']$$

- We say that $\tilde{P}(U|V)$ is **safe** for $\langle U' \rangle | \boldsymbol{V}$ if for all $P^* \in \mathcal{P}$:

$$\mathrm{E}_{P^*}[U' \mid V] = \mathrm{E}_{\tilde{P}}[U' \mid V]$$

- Our expectation of U' is (relatively) correct

Definition 3, 3b

- Recall: $\tilde{P}(U|V)$ is **safe** for $\langle U' \rangle | \langle V \rangle$ if for all $P^* \in \mathcal{P}$:

$$\mathrm{E}_{P^*}[U'] = \mathrm{E}_{V \sim P^*} \mathrm{E}_{U' \sim \tilde{P}|V}[U']$$

- We say that $\tilde{P}(U|V)$ is **safe** for $\langle U' \rangle | \boldsymbol{V}$ if for all $P^* \in \mathcal{P}$:

$$\mathrm{E}_{P^*}[U' \mid V] = \mathrm{E}_{\tilde{P}}[U' \mid V]$$

- We say that $\tilde{P}(U|V)$ is **safe** for $\boldsymbol{U' | V}$ if for all $P^* \in \mathcal{P}$:

$$P^*[U' \mid V] \equiv \tilde{P}[U' \mid V]$$

i.e. P^* is unique and $\tilde{P}$ is almost surely 'correct'

Definition 3c...

- We can now also combine definitions, e.g. $\tilde{P}(U|V)$ is **safe** for $\langle U' \rangle | \langle V \rangle, W$ if ....(details omitted)...

Ex. 2, Calibration: preparation

- Recall: $\tilde{P}(U|V)$ can be used to predict not just U, but also any U' determined by (U,V) , i.e. with $(U,V) \rightsquigarrow U'$

$$\tilde{P}(U' = u' \mid V = v) := \sum_{(u,v) \in \mathcal{U} \times \mathcal{V} : \phi(u,v) = u'} \tilde{P}(U = u \mid v)$$

Ex. 2, Calibration: preparation

- Recall: $\tilde{P}(U|V)$ can be used to predict not just U, but also any U' determined by (U,V) , i.e. with $(U,V) \rightsquigarrow U'$
- Similarly, $\tilde{P}(U|V)$ can also be used to predict not just given V, but also given any V' with $V \rightsquigarrow V'$ and **extra condition** that for all $v_1, v_2, v_1 \neq v_2$:

$$\phi_{V \rightsquigarrow V'}(v_1) = \phi_{V \rightsquigarrow V'}(v_2) \Rightarrow \tilde{P}(U \mid V = v_1) = \tilde{P}(U \mid V = v_2)$$

- For such $V', \tilde{P}(U|V')$ is well-defined
- **Example:** earlier $\tilde{P}$ that treated U, V as independent:

$$V \in \{0,1\}, V' \equiv \mathbf{0}$$

Ex. 2, preparation

- $\tilde{P}(U|V)$ can be used to predict not just given V, but also given any V' with $V \rightsquigarrow V'$ and **extra condition** that for all $v_1, v_2, v_1 \neq v_2$:

$$\phi_{V \rightsquigarrow V'}(v_1) = \phi_{V \rightsquigarrow V'}(v_2) \Rightarrow \tilde{p}(U \mid V = v_1) = \tilde{P}(U \mid V = v_2)$$

- **compact restatement:** $\tilde{P}(U|V)$ can also be used to predict (i.e. induces a unique definition of $\tilde{P}(U|V')$) based on any V' with $V' \rightsquigarrow \tilde{P}(U \mid V)$

Ex. 2., Calibration

- We say that $\tilde{P}(U|V)$ is **strongly calibrated** for $U' | V$
  .....if it is **safe** for $U' \mid \tilde{P}(U|V)$!
  ...i.e. for all $P^* \in \mathcal{P}, \vec{p} \in \mathrm{Range}(P(U \mid V))$

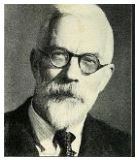
$$P^*(U' \mid \tilde{P}(U \mid V) = \vec{p}) = \vec{p}$$

Ex. 2., Calibration

- We say that $\tilde{P}(U|V)$ is **strongly calibrated** for $U' | V$
  .....if it is **safe** for $U' \mid \tilde{P}(U|V)$!
  ...i.e. for all $P^* \in \mathcal{P}, \vec{p} \in \mathrm{Range}(P(U \mid V))$

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

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



Ex. 3, Fiducial Distributions

- Determining a distribution on parameters $\tilde{p}(\theta \mid X^n)$ without a prior, i.e. **cooking a Bayesian omelet without breaking the Bayesian eggs**
- Introduced by Fisher (1935, *Annals of Eugenics*) Once almost on a par with Bayes and frequentist approaches; but turned out to suffer severe difficulties
  - e.g., Seidenfeld '92
- Yet it is making a small comeback under the name **confidence distributions** (Hjort & Schweder, 2000)

Fiducial Distributions

- Simple Example: normal location family

$\mathcal{M} = \{p_\theta \mid \theta \in \mathbb{R}\}$, $p_\theta(X^n) = \prod_{i=1}^n p_\theta(X_i), p_\theta(X_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_i-\theta)^2}$

- Fisher observed that the density of the ML estimator $\hat{\theta} = n^{-1}\sum_{i=1}^n X_i$ satisfies

$$p_\theta(\hat{\theta}) = \sqrt{\frac{n}{2\pi}} \cdot e^{-\frac{\sqrt{n}}{2}\cdot(\hat{\theta}-\theta)^2}$$

which is **symmetric** in $\theta, \hat{\theta}$ ... so that for each $X^n \in \mathbb{R}^n$ $\tilde{p}(\theta \mid X^n) := p_\theta(\hat{\theta})$ must give a distribution on θ ...

Fiducial Distributions

- Simple Example: normal location family

$\mathcal{M} = \{p_\theta \mid \theta \in \mathbb{R}\}$, $p_\theta(X^n) = \prod_{i=1}^n p_\theta(X_i), p_\theta(X_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_i-\theta)^2}$

- Fisher observed that the density of the ML estimator $\hat{\theta} = n^{-1}\sum_{i=1}^n X_i$ satisfies

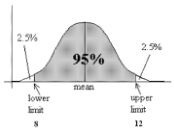
$$p_\theta(\hat{\theta}) = \sqrt{\frac{n}{2\pi}} \cdot e^{-\frac{\sqrt{n}}{2}\cdot(\hat{\theta}-\theta)^2}$$

which is **symmetric** in $\theta, \hat{\theta}$ ... so that for each $X^n \in \mathbb{R}^n$ $\tilde{p}(\theta \mid X^n) := p_\theta(\hat{\theta})$ must give a distribution on θ ...
...Fisher now boldly treated this is a sort-of posterior...

Fiducial Distribution

- Can do similar reversal for other 1-parameter distributions.
  - For scale and location families, the fiducial distr is equal to the Bayes' posterior with the improper Jeffreys' prior
  - For other families, no 100% Bayes interpretation (Lindley, Seidenfeld)
  - For Bayesians this seems flawed: there must be a prior
  - For Frequentists this seems flawed: θ is fixed, not a random variable!!

$$\tilde{p}(\theta \mid X^n) := \sqrt{\frac{n}{2\pi}} \cdot e^{-\frac{\sqrt{n}}{2}\cdot(\hat{\theta}-\theta)^2}$$



Fiducial Distributions and Confidence

- It has long been known that fiducial distributions are "o.k." if used to determine **confidence** intervals...
  suppose $X_1, X_2, \ldots$ i.i.d. $\sim P_{\theta^*}$, for any $\theta^* \in \mathbb{R}$
  Set $\theta^+ = \theta^+(X^n)$ and $\theta^- = \theta^-(X^n)$ such that

$$\tilde{P}(\theta \leq \theta^-(X^n) \mid X^n) = 0.025$$
$$\tilde{P}(\theta \geq \theta^+(X^n) \mid X^n) = 0.025$$

- Then: $P_{\theta^*}(\ \theta^* \notin [\theta^-, \theta^+] \ ) = 0.05$

Fiducial Distributions and Confidence

- It has long been known that fiducial distributions are "o.k." if used to determine **confidence** intervals...
- ...but are not o.k. "in general" (but what exactly does this mean? And what are they o.k. for?)

Using Fiducial Distributions Safely

- Let $\mathcal{P}$ be **any** set of distributions on $\Theta \times \mathbb{R}^n = \mathcal{U} \times \mathcal{V}$ such that for all $\theta^* \in \mathbb{R}, x^n \in \mathbb{R}^n : p(x^n|\theta^*) := p_{\theta^*}(x^n)$
  ($\mathcal{P}$ may e.g. only contain degenerate distr on Θ)
- We define some `pragmatic posterior' $\tilde{P}(\theta \mid X^n)$ and...
- **DEF:** we say that $\tilde{P}(\theta \mid X^n)$ is fiducially safe if it is safe for $\tilde{F}(. \mid X^n) \mid \langle X^n \rangle$, where

$$\tilde{F}(u \mid X^n) = \tilde{P}(\theta \leq u \mid X^n)$$

is the distribution function of $\tilde{P}(\theta \mid X^n)$

Using Fiducial Distributions Safely

- Let $\mathcal{P}$ be **any** set of distributions on $\Theta \times \mathbb{R}^n = \mathcal{U} \times \mathcal{V}$ such that for all $\theta^* \in \mathbb{R}, x^n \in \mathbb{R}^n : p(x^n|\theta^*) := p_{\theta^*}(x^n)$
- **DEF:** we say that $\tilde{P}(\theta \mid X^n)$ if it is safe for $\tilde{F}(. \mid X^n) \mid \langle X^n \rangle$
- **PROP:** if defined in the usual way, 'fiducial' distributions are fiducially safe

Using Fiducial Distributions Safely

- Let $\mathcal{P}$ be **any** set of distributions on $\Theta \times \mathbb{R}^n = \mathcal{U} \times \mathcal{V}$ such that for all $\theta^* \in \mathbb{R}, x^n \in \mathbb{R}^n : p(x^n|\theta^*) := p_{\theta^*}(x^n)$
- **DEF:** we say that $\tilde{P}(\theta \mid X^n)$ if it is safe for $\tilde{F}(. \mid X^n) \mid \langle X^n \rangle$
- **PROP:** if defined in the usual way, 'fiducial' distributions are fiducially safe
- This means they can be safely used to predict any RV U' determined by $\tilde{F}(\theta \mid X^n)$
  - For example, $\mathbf{1}_{\theta \notin [\theta^-, \theta^+]}$ is fiducially safe to predict...

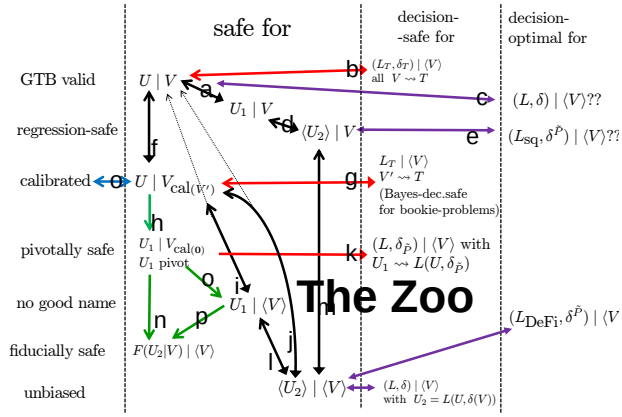
$$P_{\theta^*}(\ \theta^* \notin [\theta^-, \theta^+]\ ) = \mathrm{E}_{X^n \sim P_{\theta^*}}\left[\ \tilde{P}(\theta \notin [\theta^-, \theta^+] \mid X^n)\ \right] = 0.05$$

Using Fiducial Distributions Safely

- Let $\mathcal{P}$ be **any** set of distributions on $\Theta \times \mathbb{R}^n = \mathcal{U} \times \mathcal{V}$ such that for all $\theta^* \in \mathbb{R}, x^n \in \mathbb{R}^n : p(x^n|\theta^*) := p_{\theta^*}(x^n)$
- **DEF:** we say that $\tilde{P}(\theta \mid X^n)$ is fiducially safe for $\theta \mid \langle X^n \rangle$ if it is safe for $\tilde{F}(. \mid X^n) \mid \langle X^n \rangle$
- **PROP:** if defined in the usual way, 'fiducial' distributions are fiducially safe
- This means they can be safely used to predict any RV U' determined by $\tilde{F}(\theta \mid X^n)$
  - For example, $\mathbf{1}_{\theta \notin [\theta^-, \theta^+]}$ is fiducially safe to predict...
  - ....but $\mathbf{1}_{\theta \notin [12.3, 19.8]}$ is not!

Dilation-Fiducial Duality

- **DILATION-REALITY**: U may be **dependent** on V
- **DILATION-PRAGMATICS**: we nevertheless decide to predict U with a distribution that assumes U and V are **independent**
- **FIDUCIAL-REALITY:** U may be **independent** of V, it may even be fixed – but its value is unknown
- **FIDUCIAL-PRAGMATICS:** we nevertheless predict U with a distribution that assumes U and V are **dependent**



Desert: Monty Hall (3-door) Problem

Monty Hall 1970

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
  - credal set (imprecise probability, ambiguity)-based: make no assumptions on Monty and show e.g. that switching is minimax optimal

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
  - credal set (imprecise probability, ambiguity)-based: make no assumptions on Monty and show switching is e.g. dominating strategy

This is really what Gilboa called the 'eternal discussion'

The Model on which they 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:

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

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

The Model on which they agree

- Suppose Contestant invariably chooses door a.
- Let RV Y denote location of car.
- Let RV 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 Point Probabilists' Side

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

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

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

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

The Sets-of-Probabilities Side

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

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

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

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

!

Dilation

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

$$P(Y = a \mid X = \text{open}(b)) \in \left[0, \frac{1}{2}\right]$$

Instance of (partial) **dilation** (Seidenfeld,Wasserman 93):

Before observing X we had precise probability $P(Y = a)$
after we only know $P(Y = a \mid X = x)$ is in large superset

Dilation

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

$$P(Y = a \mid X = \text{open}(b)) \in \left[0, \frac{1}{2}\right]$$

Instance of (partial) **dilation** (Seidenfeld,Wasserman 93):

Before observing X we had precise probability $P(Y = a)$
after we only know $P(Y = a \mid X = x)$ is in large superset

"**extra information ⇒ less knowledge**
no matter what you observe!"

Assuming an Unbiased Monty

- To avoid dilation, tempting to become precise probabilist 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

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

is

1. "safe" and
2. minimax optimal

....under all **symmetric** decision problems and all, **even asymmetric Kelly gambling** problems

Safety for Decision Problems

Unbiased Monty is

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

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

where

$\mathcal{A} =$ set of actions

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

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

Safety

Unbiased Monty is

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

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

Example: =1/3

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

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

$\text{Loss}(Y, \hat{y}) = \mathbf{1}_{Y \neq \hat{y}}$

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

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 $P^* \in \mathcal{P}^*$:

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

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

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

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

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

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

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!
- Except for asymmetric versions of log-loss! Then fair-coin assumption is still both safe and minimax optimal!

$$\text{Loss}(Y, q) = -\log \frac{q(Y)}{a(Y)}$$

Assuming an Unbiased Monty...

- ..., i.e. use

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

is

1. "safe"
2. minimax optimal
3. admissible

hence ***PRETTY ADEQUATE***

....under all **symmetric** decision problems and all, **even asymmetric Kelly gambling** problems

Unbiased Monty

- Straightforward imprecise probability gives dilation
- Straightforward subjective Bayes is problematic for me: *why* would Monty be unbiased??
- Safe Probability Approach: *if* you are willing to make some assumption about loss function, it is *safe* to assume that Monty tosses a fair coin

Dependence on Task

- Safe Probability Approach: *if* you are willing to make some assumption about loss function, it is *safe* to assume that Monty tosses a fair coin
- This means that if you are told that the loss function is asymmetric, you may want to **change** your distribution
  - Similarly, if you're told in dilation problem that the probability that you have to make a prediction depends on V, you don't want to ignore V any more
  - Similarly, if, in 'objective Bayes', you change the sampling plan, you want to change the prior
  - **This is the price we pay for cooking a Bayesian omelet with imprecise eggs**

Conclusion: Towards A Theory of "Safe Probability"

- Compromise between 'strict' Bayes and imprecise probability theory
  - $\mathcal{P}^*$ has unique representative P° as in Minimum Description Length, **objective Bayes**, fiducial inference, 'MaxEnt...
  - One absolutely crucial difference: we **restrict** use of P° to subset of all possible prediction tasks; eqv.

we 'condition' on the task