



Decision Making under Weakly Structured Information

Day 3 at the **SIPTA Summer School 2026**

Held on 27.7.2026 – 31.07.2026 at LMU Munich

Christoph Jansen

School of Computing & Communications, Lancaster University Leipzig

Julian Rodemann

RI Lab, CISPA Helmholtz Center; Department of Statistics, LMU Munich

Some Words About Us



- **Christoph** Jansen (c.jansen@lancaster.ac.uk)
- BSc, MSc, PhD, Habilitation at LMU
- More often than not supervised by Thomas Augustin
- Now Assistant Professor in Data Science at Lancaster University Leipzig
- Happy to be here

Some Words About Us



- **Julian** Rodemann (mail@julian-rodemann.de)
- MSc and PhD at LMU supervised by Thomas Augustin
- Now PostDoc at the Rational Intelligence Lab at CISPA led by Krikamol Muandet (also here)
- Equally happy to be here

Some Words About Thomas



- Thomas Augustin
- Passed away January this year
- Professor of Statistics at LMU
- Taught us most of the things we know about Decision Theory
- Would have loved to see this event finally taking place in Munich

Some Words About Other Friends Supervised by Thomas



Georg Schollmeyer

You will see him speaking on Thursday



Hannah Blocher

Not here, but many of her ideas are in the slides.

In Memoriam: Thomas Augustin (1965–2026)

- Professor of Statistics at LMU Munich, and the driving force behind **this** SIPTA Summer School taking place in Munich.
- Passed away unexpectedly on 10 January 2026.
- Taught us decision theory and shaped a whole generation of researchers in imprecise probability.
- He was in charge of organising the 2026 SIPTA School



Thomas's Contributions to IP and Statistics / Special Issue

- Obituary [Coolen et al., 2026] just appeared in IJAR:
https://authors.elsevier.com/c/1nWB0_KD6ZnKMW



The screenshot shows the article page for "Thomas Augustin's Contributions to Imprecise Probability and Statistics" in the International Journal of Approximate Reasoning. The article is by F. Coolen, F. Cozman, T. Denoeux, C. Jansen, H. Küchenhoff, R. Martin, E. Miranda, J. Plass, J. Rodemann, G. Schollmeyer, and R. Seising. It is a research article with full text access. The page includes a "View PDF" button, a "Cite" button, an "Add to Mendeley" button, and a "Share" button. The article ID is 10.1016/j.ijar.2026.109781. The page also features a "Highlights" section and an "Outline" section.

- **Call for Papers:** Special Issue *In Memoriam of Thomas Augustin*
- Topics of interest include
 - ...
 - *Anything related to IP*
- <https://www.sciencedirect.com/journal/international-journal-of-approximate-reasoning/about/call-for-papers>

Thomas Augustin: Short Bio

- Born 1965 in Burghausen; started in mechanical engineering at TU Munich before finding his home in statistics at LMU (Diploma 1992).
- PhD 1997 (*summa cum laude*) and Habilitation 2002; Professor at LMU from 2003.
- Doctoral student of Kurt Weichselberger, starting point of a lifelong engagement with interval and imprecise probability.
- A pillar of SIPTA: present at the very first ISIPTA (Ghent, 1999) and every edition since; co-editor of the standard reference *Introduction to Imprecise Probabilities* (2014).
- Dean of Studies at LMU for nearly 20 years at the Department of Statistics

Thomas and Decision Theory

“Statistik ist die Stangenware, aber Entscheidungstheorie ist der Maßanzug.”

— Franz Fersch, quoted by Thomas with full agreement

Statistics is off-the-peg clothing; decision theory is a tailor-made suit.

For Thomas, decision theory was the principled way to frame a statistical problem: formally, concisely, and no more complicated than necessary. Much of what follows today carries his fingerprints:

- Resolving **Ellsberg's paradox** via interval probability [Weichselberger and Augustin, 1998].
- One of the first LP-based algorithms for **Γ -maximin** and for checking **E-admissibility** [Utkin and Augustin, 2005]
- The **maximality** algorithm built on that foundation [Jansen et al., 2017], and much of the work on weakly structured information you will see later.

- A gentle, kind, and deeply fair human, who cared about others before himself.
- Endlessly curious across fields — he fondly called his own group “ein philosophisch angehauchter Gemischtwarenladen” (a philosophically flavoured hodgepodge).
- Active in core decision theory until the very end; one of his last projects [Jansen et al., 2026] was submitted just weeks before he died.

With his passing we not only lost an academic giant, but also a mentor and friend.

These slides are dedicated to his memory.

Roadmap

A Short Motivation

General Overview

Classical Decision Theory

Decision Theory for Statistics and Machine Learning

Decision Making Under Imprecise Probabilities

Decision Making Under Weakly Structured Information

Application: DT under WSI for Statistical Multicriteria Benchmarking

A Short Motivation

Learning Decision Theory: An Interdisciplinary Investment

Decision-theoretic thinking is heavily relied on in many fields. Some examples:

Learning Decision Theory: An Interdisciplinary Investment

Decision-theoretic thinking is heavily relied on in many fields. Some examples:

Theoretical Economics:

Modeling rational agents who choose actions to maximize expected utility.

Learning Decision Theory: An Interdisciplinary Investment

Decision-theoretic thinking is heavily relied on in many fields. Some examples:

Theoretical Economics:

Modeling rational agents who choose actions to maximize expected utility.

Philosophy:

Analyzing rational choice, preference, and the normative foundations of decision making. Compare yesterday's lectures of Arthur and Catrin 😊.

Learning Decision Theory: An Interdisciplinary Investment

Decision-theoretic thinking is heavily relied on in many fields. Some examples:

Theoretical Economics:

Modeling rational agents who choose actions to maximize expected utility.

Philosophy:

Analyzing rational choice, preference, and the normative foundations of decision making. Compare yesterday's lectures of Arthur and Catrin 😊.

Statistics:

Developing principled methods for inference and estimation based on minimizing expected loss. Compare Julian's part later today.

Learning Decision Theory: An Interdisciplinary Investment

Decision-theoretic thinking is heavily relied on in many fields. Some examples:

Theoretical Economics:

Modeling rational agents who choose actions to maximize expected utility.

Philosophy:

Analyzing rational choice, preference, and the normative foundations of decision making. Compare yesterday's lectures of Arthur and Catrin 😊.

Statistics:

Developing principled methods for inference and estimation based on minimizing expected loss. Compare Julian's part later today.

Decision Theory: A truly interdisciplinary investment!
But wait: **What about AI?**



Decision Theory for AI and ML

Already in 1988, Eric Horvitz (CSO of Microsoft) and co-authors state:

“Despite their different perspectives, artificial intelligence (AI) and decision science have common roots and strive for similar goals.”

([Horvitz et al., 1988, p. 1])

Decision Theory for AI and ML

Already in 1988, Eric Horvitz (CSO of Microsoft) and co-authors state:

“Despite their different perspectives, artificial intelligence (AI) and decision science have common roots and strive for similar goals.”

([Horvitz et al., 1988, p. 1])

But while a decision-theoretic perspective is now standard in mathematical statistics, in the context of AI and ML it is still said in 2021:

“A decision-theoretic grounding promises to support the systematic development of ML methods in a principled (rather than ad hoc) manner”

([Hüllermeier, 2021, p. 5])

Decision Theory for AI and ML

Already in 1988, Eric Horvitz (CSO of Microsoft) and co-authors state:

“Despite their different perspectives, artificial intelligence (AI) and decision science have common roots and strive for similar goals.”

([Horvitz et al., 1988, p. 1])

But while a decision-theoretic perspective is now standard in mathematical statistics, in the context of AI and ML it is still said in 2021:

“A decision-theoretic grounding promises to support the systematic development of ML methods in a principled (rather than ad hoc) manner”

([Hüllermeier, 2021, p. 5])

Bottom line:

- AI and ML promise to benefit greatly from decision theory.
- This promise is still being fulfilled far too little.

General Overview

Decision Theory in One Slide

Informal description of the model:

- An **agent** has to choose among different **acts** X from a set $\mathcal{G}$.
- The **consequence** that choosing X yields depends on which **state of nature** s from a set S is the true one.

Decision Theory in One Slide

Informal description of the model:

- An agent has to choose among different acts X from a set $\mathcal{G}$.
- The consequence that choosing X yields depends on which state of nature s from a set S is the true one.

Formal description of the model:

- Let C denote some non-empty set of consequences.
- Each act X corresponds to a mapping $X : S \rightarrow C$.
- The set $\mathcal{G}$ is a subset of $C^S = \{X : S \rightarrow C\}$.

Decision Theory in One Slide

Informal description of the model:

- An agent has to choose among different acts X from a set $\mathcal{G}$.
- The consequence that choosing X yields depends on which state of nature s from a set S is the true one.

Formal description of the model:

- Let C denote some non-empty set of consequences.
- Each act X corresponds to a mapping $X : S \rightarrow C$.
- The set $\mathcal{G}$ is a subset of $C^S = \{X : S \rightarrow C\}$.

Goal: Determining a **choice function**

$$ch : 2^{\mathcal{G}} \rightarrow 2^{\mathcal{G}} \text{ with } ch(\mathcal{D}) \subseteq \mathcal{D} \text{ for all } \mathcal{D} \in 2^{\mathcal{G}}$$

that best possibly utilizes the available information.

Thanks for your attention!



Running Example

You want to get out, but you are not sure how the weather will be.

Which outfit to choose?

Acts:

-  $\approx$ coat and umbrella ($= X_1$)
-  $\approx$ only T-shirt ($= X_2$)
-  $\approx$ only coat ($= X_3$)

States:

-  $\approx$ sunny ($= s_1$)
-  $\approx$ cloudy ($= s_2$)
-  $\approx$ rainy ($= s_3$)
-  $\approx$ lightening ($= s_4$)

Running Example

Running Example

Running Example

Running Example

Running Example

Running Example

Running Example

Running Example

Running Example

Running Example

Running Example

Running Example

Running Example

Running Example

Task. Spend five minutes reflecting on:

- How would a **general** choice function look like in this setting?
- How would you construct a **specific** choice function in this setting?
- What **assumptions** does your choice function rely on?

Main Information Sources for Constructing Cochoice Functions

Choice functions are usually constructed relying on two building blocks:

$\mathcal{I}_u$: Information on the mechanism generating the states of the world.

$\mathcal{I}_p$: Information about the agent's preferences on the consequences.

Main Information Sources for Constructing Cohoice Functions

Choice functions are usually constructed relying on two building blocks:

$\mathcal{I}_u$: Information on the mechanism generating the states of the world.

$\mathcal{I}_p$: Information about the agent's preferences on the consequences.

Examples for $\mathcal{I}_u$:

- $\text{prob}(\text{☀}) = 0.2$
- $\text{prob}(\text{☀}) \leq 2 \cdot \text{prob}(\text{☁})$
- $\text{prob}(\text{☀}) \in [0, 0.3]$
- $\text{prob}(\text{☀} \mid \text{☂}) = 1$

Main Information Sources for Constructing Cohoice Functions

Choice functions are usually constructed relying on two building blocks:

$\mathcal{I}_u$: Information on the **mechanism generating the states of the world**.

$\mathcal{I}_p$: Information about the agent's **preferences on the consequences**.

Examples for $\mathcal{I}_u$:

- $\text{prob}(\text{☀}) = 0.2$
- $\text{prob}(\text{☀}) \leq 2 \cdot \text{prob}(\text{☁☁})$
- $\text{prob}(\text{☀}) \in [0, 0.3]$
- $\text{prob}(\text{☀} \mid \text{☂}) = 1$

Examples for $\mathcal{I}_p$:

- $(\text{☂}, \text{☀}) \succ (\text{👔}, \text{☀})$
- $(\text{☂}, \text{☁}) \sim (\text{☂}, \text{☁})$
- $(\text{☂}, \text{☀}) \bowtie (\text{👔}, \text{☁})$
- $(\text{👔}, \text{☀}) \gg (\text{☂}, \text{⚡})$

Running Example Continued

Running Example Continued

Possible Assumptions (I):

$\mathcal{I}_U$: States are picked by an omniscient and almighty antagonist.

$\mathcal{I}_P$:  $\succ$  $\succ$  $\succ$  $\succ$  $\succ$ 

Running Example Continued

Possible Assumptions (I):

$\mathcal{I}_U$: States are picked by an omniscient and almighty antagonist.

$\mathcal{I}_P$:  $\succ$  $\succ$  $\succ$  $\succ$  $\succ$ 

An **optimal** choice function given this information is given by:

$$\begin{aligned}
 ch_{\text{ma}}(\{\text{umbrella}\}) &= \{\text{umbrella}\} & ch_{\text{ma}}(\{\text{red t-shirt}\}) &= \{\text{red t-shirt}\} & ch_{\text{ma}}(\{\text{blue jacket}\}) &= \{\text{blue jacket}\} \\
 ch_{\text{ma}}(\{\text{umbrella}, \text{red t-shirt}\}) &= \{\text{red t-shirt}\} & ch_{\text{ma}}(\{\text{red t-shirt}, \text{blue jacket}\}) &= \{\text{blue jacket}\} & ch_{\text{ma}}(\{\text{umbrella}, \text{blue jacket}\}) &= \{\text{blue jacket}\} \\
 ch_{\text{ma}}(\{\text{umbrella}, \text{red t-shirt}, \text{blue jacket}\}) &= \{\text{blue jacket}\}
 \end{aligned}$$

Running Example Continued

Possible Assumptions (II):

$\mathcal{I}_U$: States are picked by an omniscient and almighty supporter.

$\mathcal{I}_P$:  $\succ$  $\succ$  $\succ$  $\succ$  $\succ$ 

Running Example Continued

Possible Assumptions (II):

$\mathcal{I}_U$: States are picked by an omniscient and almighty supporter.

$\mathcal{I}_P$:  $\succ$  $\succ$  $\succ$  $\succ$  $\succ$ 

An **optimal** choice function given this information is given by:

$$ch_{mi}(\{\text{umbrella}\}) = \{\text{umbrella}\} \quad ch_{mi}(\{\text{red t-shirt}\}) = \{\text{red t-shirt}\} \quad ch_{mi}(\{\text{blue jacket}\}) = \{\text{blue jacket}\}$$

$$ch_{mi}(\{\text{umbrella}, \text{red t-shirt}\}) = \{\text{red t-shirt}\} \quad ch_{mi}(\{\text{red t-shirt}, \text{blue jacket}\}) = \{\text{red t-shirt}\}$$

$$ch_{mi}(\{\text{umbrella}, \text{blue jacket}\}) = \{\text{umbrella}, \text{blue jacket}\} \quad ch_{mi}(\{\text{umbrella}, \text{red t-shirt}, \text{blue jacket}\}) = \{\text{red t-shirt}\}$$

Running Example Continued

Possible Assumptions (III):

$\mathcal{I}_U$: $s \sim \pi$, but we only know $\pi(\text{Sun}) \geq \pi(\text{Cloud}) \geq \pi(\text{Rain cloud}) \geq \pi(\text{Lightning bolt})$ (*)

$\mathcal{I}_P$:  $\succ$  $\succ$  $\succ$  $\succ$  $\succ$  (**)

Running Example Continued

Possible Assumptions (III):

$$\mathcal{I}_U: s \sim \pi, \text{ but we only know } \pi(\text{Sun}) \geq \pi(\text{Cloud}) \geq \pi(\text{Rain}) \geq \pi(\text{Lightning}) \quad (*)$$

$$\mathcal{I}_P: \text{Star-struck} \succ \text{Happy} \succ \text{Neutral} \succ \text{Sad} \succ \text{Angry} \succ \text{Skull} \quad (**)$$

A **possible** choice function ch_{sd} can be constructed based on the following idea:
Exclude an act X whenever there is an act X' such that

$$\mathbb{E}_\pi(u \circ X' - u \circ X) > 0$$

for all $u : C \rightarrow [0, 1]$ that respect $(**)$ and all π that respect $(*)$.

Running Example Continued

Possible Assumptions (III):

$$\mathcal{I}_U: s \sim \pi, \text{ but we only know } \pi(\text{Sun}) \geq \pi(\text{Cloud}) \geq \pi(\text{Rain}) \geq \pi(\text{Lightning}) \quad (*)$$

$$\mathcal{I}_P: \text{Star-struck} \succ \text{Happy} \succ \text{Neutral} \succ \text{Sad} \succ \text{Angry} \succ \text{Skull} \quad (**)$$

A **possible** choice function ch_{sd} can be constructed based on the following idea:
Exclude an act X whenever there is an act X' such that

$$\mathbb{E}_\pi(u \circ X' - u \circ X) > 0$$

for all $u : C \rightarrow [0, 1]$ that respect $(**)$ and all π that respect $(*)$.

Task. Determine ch_{sd} entirely!

Running Example Continued

Possible Assumptions (III):

$$\mathcal{I}_U: s \sim \pi, \text{ but we only know } \pi(\text{Sun}) \geq \pi(\text{Cloud}) \geq \pi(\text{Rain}) \geq \pi(\text{Lightning}) \quad (*)$$

$$\mathcal{I}_P: \text{Happy with star} \succ \text{Happy} \succ \text{Neutral} \succ \text{Sad} \succ \text{Angry} \succ \text{Skull} \quad (**)$$

A **possible** choice function ch_{sd} can be constructed based on the following idea:
Exclude an act X whenever there is an act X' such that

$$\mathbb{E}_{\pi}(u \circ X' - u \circ X) > 0$$

for all $u : C \rightarrow [0, 1]$ that respect $(**)$ and all π that respect $(*)$.

Solution. We have $ch_{sd}(M) = M$ for all $M \subseteq \{\text{Umbrella}, \text{Red t-shirt}, \text{Blue jacket}\}$

Interpreting Choice Functions

Depending on the quality of the underlying information, the choice sets $ch(\mathcal{D})$ can be given two different interpretations:

Interpreting Choice Functions

Depending on the quality of the underlying information, the **choice sets** $ch(\mathcal{D})$ can be given two different **interpretations**:

Strong interpretation:

- $ch(\mathcal{D})$ is the set of **equally optimal** acts from $\mathcal{D}$.
- The agent is **indifferent** between these acts.

Interpreting Choice Functions

Depending on the quality of the underlying information, the **choice sets** $ch(\mathcal{D})$ can be given two different **interpretations**:

Strong interpretation:

- $ch(\mathcal{D})$ is the set of equally optimal acts from $\mathcal{D}$.
- The agent is indifferent between these acts.

Weak interpretation:

- $ch(\mathcal{D})$ is the set of all **non-neglectable** acts from $\mathcal{D}$.
- These acts are **incomparable** for the agent.

Interpreting Choice Functions

Depending on the quality of the underlying information, the **choice sets** $ch(\mathcal{D})$ can be given two different **interpretations**:

Strong interpretation:

- $ch(\mathcal{D})$ is the set of equally optimal acts from $\mathcal{D}$.
- The agent is indifferent between these acts.

Weak interpretation:

- $ch(\mathcal{D})$ is the set of all non-neglectable acts from $\mathcal{D}$.
- These acts are incomparable for the agent.

Obvious comment: If only **weakly structured information** is available, we often have to work with weakly interpretable choice functions.

Running Example Continued

Task. Which of the three choice functions allow for the strong interpretation? Which only admit the weak interpretation? Why?

Running Example Continued

Task. Which of the three choice functions allow for the strong interpretation? Which only admit the weak interpretation? Why?

Solution. We have the following:

Strong: ch_{ma} and ch_{mi}

Explanation: Both function produce choice sets contain elements between the decision maker is truly indifferent. For instance, $ch_{mi}(\{\text{☂}, \text{🧥}\}) = \{\text{☂}, \text{🧥}\}$ means the decision maker is indefferent between umbrella and coat, since under their best case they produce the same outcome.

Weak: ch_{sd}

Explanation: The acts remaining in the choice sets are incomparable to the decision maker. For instance, $ch_{sd}(\{\text{☂}, \text{🧥}\}) = \{\text{☂}, \text{🧥}\}$ means that there exist (u, π) and (u', π') both compatible with $(\mathcal{I}_P, \mathcal{I}_U)$ satisfying:

$$\mathbb{E}_{\pi}(u \circ \text{☂} - u \circ \text{🧥}) > 0 \quad \wedge \quad \mathbb{E}_{\pi'}(u' \circ \text{🧥} - u' \circ \text{☂}) > 0$$

Standard Ways of Constructing Choice Functions

Criterion-Based Choice Functions:

Given the information base $(\mathcal{I}_U, \mathcal{I}_P)$, define a **criterion** $\Phi_{(\mathcal{I}_U, \mathcal{I}_P)} : \mathcal{G} \rightarrow \mathbb{R}$.

Use this to define, for all $M \subseteq \mathcal{G}$, the choice function

$$ch_{\Phi_{(\mathcal{I}_U, \mathcal{I}_P)}}(M) := \operatorname{argmax}_{X \in M} \Phi_{(\mathcal{I}_U, \mathcal{I}_P)}(X)$$

Standard Ways of Constructing Choice Functions

Criterion-Based Choice Functions:

Given the information base $(\mathcal{I}_U, \mathcal{I}_P)$, define a **criterion** $\Phi_{(\mathcal{I}_U, \mathcal{I}_P)} : \mathcal{G} \rightarrow \mathbb{R}$.

Use this to define, for all $M \subseteq \mathcal{G}$, the choice function

$$ch_{\Phi_{(\mathcal{I}_U, \mathcal{I}_P)}}(M) := \operatorname{argmax}_{X \in M} \Phi_{(\mathcal{I}_U, \mathcal{I}_P)}(X)$$

Binary-Comparison Induced Choice Functions:

Given the information base $(\mathcal{I}_U, \mathcal{I}_P)$, define a **preorder** $\succsim_{(\mathcal{I}_U, \mathcal{I}_P)}$ on $\mathcal{G}$.

Use this to define, for all $M \subseteq \mathcal{G}$, the choice function

$$ch_{\succsim_{(\mathcal{I}_U, \mathcal{I}_P)}}(M) := \left\{ X : \nexists Y \in M \text{ such that } Y \succ_{(\mathcal{I}_U, \mathcal{I}_P)} X \right\}$$

Standard Ways of Constructing Choice Functions

Criterion-Based Choice Functions:

Given the information base $(\mathcal{I}_U, \mathcal{I}_P)$, define a **criterion** $\Phi_{(\mathcal{I}_U, \mathcal{I}_P)} : \mathcal{G} \rightarrow \mathbb{R}$.

Use this to define, for all $M \subseteq \mathcal{G}$, the choice function

$$ch_{\Phi_{(\mathcal{I}_U, \mathcal{I}_P)}}(M) := \operatorname{argmax}_{X \in M} \Phi_{(\mathcal{I}_U, \mathcal{I}_P)}(X)$$

Binary-Comparison Induced Choice Functions:

Given the information base $(\mathcal{I}_U, \mathcal{I}_P)$, define a **preorder** $\succsim_{(\mathcal{I}_U, \mathcal{I}_P)}$ on $\mathcal{G}$.

Use this to define, for all $M \subseteq \mathcal{G}$, the choice function

$$ch_{\succsim_{(\mathcal{I}_U, \mathcal{I}_P)}}(M) := \left\{ X : \nexists Y \in M \text{ such that } Y \succ_{(\mathcal{I}_U, \mathcal{I}_P)} X \right\}$$

Other Choice Functions:

Utilize the information base $(\mathcal{I}_U, \mathcal{I}_P)$ to define a choice function that does not fall under the two previous categories. (We'll see examples later!)

Runnig Example Continued

Task. Answer the following two questions:

- In what sense are criterion-based choice functions a strict sub-category of binary-comparison induced ones?
- Which is the tightest category the three choice functions from before belong to, respectively?

Runnig Example Continued

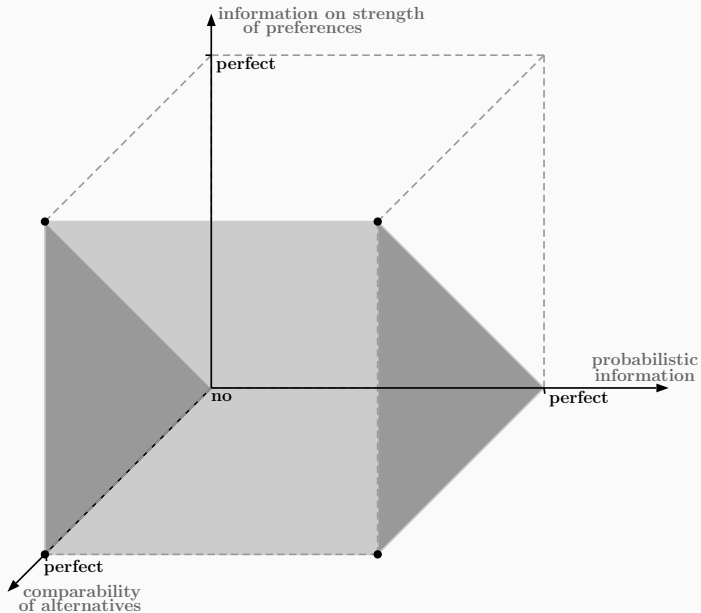
Task. Answer the following two questions:

- In what sense are criterion-based choice functions a strict sub-category of binary-comparison induced ones?
- Which is the tightest category the three choice functions from before belong to, respectively?

Solution.

- We can use the criterion Φ to define a weak order on the acts by setting $X \succsim_{\Phi} Y \iff \Phi(X) \geq \Phi(Y)$. Then, we have $ch_{\Phi} = ch_{\succsim_{\Phi}}$. However, not any b.-c. induced CF is criterion-based (e.g., $\exists$ incomparabilities).
- ch_{ma} and ch_{mi} are criterion-based.
- ch_{sd} is induced by pairwise comparisons, however, not by any single criterion function (in general, note that in this concrete example it would be possible though).

Visual Outline of Today



Classical Decision Theory

Why 'Classical'?

Classical Decision Theory

$\mathcal{I}_P$ Information on preferences is characterized by a cardinal utility function $u : C \rightarrow \mathbb{R}$ that is unique up to positive affine transformations.

$\mathcal{I}_U$ Information on the generation of states is characterized by a unique probability measure on (S, Σ_S) .

Why 'Classical'?

Classical Decision Theory

$\mathcal{I}_P$ Information on preferences is characterized by a cardinal utility function $u : C \rightarrow \mathbb{R}$ that is unique up to positive affine transformations.

$\mathcal{I}_U$ Information on the generation of states is characterized by a unique probability measure on (S, Σ_S) .

How restrictive are these assumptions? (Part I)

A cardinal utility function in the above sense can e.g. be obtained e.g. by von Neumann and Morgenstern's theory [von Neumann et al., 1944]:

- The decision maker ranks all simple lotteries on C by preference.
- If the DM's preferences satisfy three behavioral axioms, they can be represented (unique up to plts) by a cardinal utility function $u : C \rightarrow \mathbb{R}$.

Why 'Classical'?

Classical Decision Theory

$\mathcal{I}_P$ Information on preferences is characterized by a cardinal utility function $u : C \rightarrow \mathbb{R}$ that is unique up to positive affine transformations.

$\mathcal{I}_U$ Information on the generation of states is characterized by a unique probability measure on (S, Σ_S) .

How restrictive are these assumptions? (Part II)

A unique probability measure representing the uncertainty on the states can e.g. be obtained by De Finetti's theory [De Finetti, 1937]:

- The decision maker ranks all utility-valued gambles on S by preference.
- If the DM's preferences satisfy five axioms, they can be represented as the expectation order wrt a unique probability measure on S .

Why 'Classical' ?

La prévision : ses lois logiques, ses sources subjectives

par

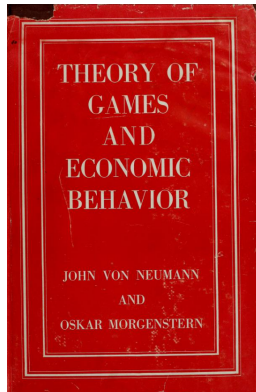
Bruno de FINETTI.

AVANT-PROPOS.

Dans les conférences que j'ai eu l'honneur de faire à l'Institut Henri Poincaré les 2, 3, 8, 9 et 10 mai 1935, et dont les pages qui suivent reproduisent le texte, je me suis proposé de donner une vue d'ensemble sur deux sujets dont je me suis particulièrement occupé, et d'éclaircir la liaison étroite qui les unit. Il s'agit, d'une part, de la définition de la probabilité (que je considère comme une entité purement subjective) et de la signification de ses lois, et, d'autre part, des notions et de la théorie des événements et nombres aléatoires « équivalents » ; le lien entre les deux questions réside en ce que cette dernière théorie constitue la solution, pour le cas le plus typique, du problème du raisonnement inductif selon la conception subjective de la probabilité (et éclaircit ainsi, en général, la façon dont le problème de l'induction est posé). D'ailleurs, même s'il n'en était pas ainsi, autrement dit même si l'on n'acceptait pas le point de vue subjectif que nous avons adopté dans la première partie, la seconde n'en conserverait pas moins sa validité et constituerait un chapitre intéressant du calcul des probabilités.

L'exposé est divisé en six chapitres, dont les deux premiers ont pour objet la première question, les deux suivants la seconde, et les deux derniers l'examen des conclusions qu'on peut en tirer. La plupart des questions traitées ont été développées, tantôt en détail, tantôt succinctement, mais toujours d'une façon fragmentaire ⁽¹⁾, dans mes

⁽¹⁾ Un exposé plus complet de mon point de vue, sous forme d'essai purement critique et philosophique, sans formules, constitue l'ouvrage [22].



vNM: obtain utilities with probabilities externally given

De Finetti: obtain probabilities with utilities externally given

vNM + De Finetti: Classical Decision Theory

Why 'Classical'?

Classical Decision Theory

$\mathcal{I}_P$ Information on preferences is characterized by a cardinal utility function $u : C \rightarrow \mathbb{R}$ that is unique up to positive affine transformations.

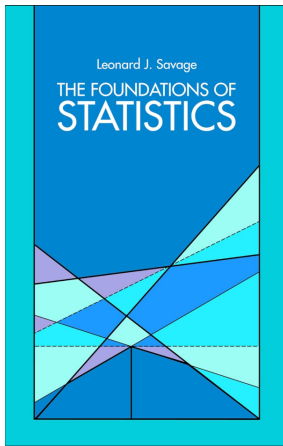
$\mathcal{I}_U$ Information on the generation of states is characterized by a unique probability measure on (S, Σ_S) .

How restrictive are these assumptions? (Part III)

A simultaneous pair of probability (unique) π and utility (unique up to plts) u can e.g. be obtained by Savage's theory [Savage, 1954]:

- The decision maker directly ranks all acts $X : S \rightarrow C$ by preference.
- If the DM's preferences satisfy seven axioms (Savage's P1 to P7), they can be represented as the expectation order based on π and u .

Why 'Classical'?



Savage: derive utility and probability from preferences over acts!

Sometimes referred to as the *crowning glory of decision theory*.

Finite Cardinal Decision Problem

A special case is the **finite decision problems with cardinal utilities**.

Finite Cardinal Decision Problem

A special case is the **finite decision problems with cardinal utilities**.

Assume we have:

- Acts $\mathcal{G} = \{X_1, \dots, X_n\}$
- States $S = \{s_1, \dots, s_m\}$
- A cardinal utility function $u : C \rightarrow \mathbb{R}$

We write u_{ij} for $u \circ X_i(s_j)$.

Finite Cardinal Decision Problem

A special case is the **finite decision problems with cardinal utilities**.

Assume we have:

- Acts $\mathcal{G} = \{X_1, \dots, X_n\}$
- States $S = \{s_1, \dots, s_m\}$
- A cardinal utility function $u : C \rightarrow \mathbb{R}$

We write u_{ij} for $u \circ X_i(s_j)$.

A convenient form for analyzing these problems are **utility tables**:

	s_1	s_2		s_{m-1}	s_m
X_1	u_{11}	u_{12}		u_{1m-1}	u_{1m}
$\vdots$	...	...		...	...
X_n	u_{n1}	u_{n2}		u_{nm-1}	u_{nm}

Criterion-Based Choice Functions in Classical DT

Given a cardinal utility u and a precise probability measure π , the following criteria are commonly for constructing choice functions:

Criterion-Based Choice Functions in Classical DT

Given a cardinal utility u and a precise probability measure π , the following criteria are commonly for constructing choice functions:

Bayes-Criterion (a.k.a. Maximizing Expected Utility):

$$\Phi_{\text{Bayes}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \mathbb{E}_{\pi}(u \circ X)$$

Criterion-Based Choice Functions in Classical DT

Given a cardinal utility u and a precise probability measure π , the following criteria are commonly for constructing choice functions:

Bayes-Criterion (a.k.a. Maximizing Expected Utility):

$$\Phi_{\text{Bayes}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \mathbb{E}_{\pi}(u \circ X)$$

Hurwicz-Criterion: Fix **optimism parameter** $\alpha \in [0, 1]$.

$$\Phi_{\text{Hurwicz}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \alpha \cdot \inf_{s \in S} u \circ X(s) + (1 - \alpha) \cdot \sup_{s \in S} u \circ X(s)$$

For $\alpha = 0 \rightarrow \Phi_{\text{maxmax}}$; for $\alpha = 1 \rightarrow \Phi_{\text{maxmin}}$.

Criterion-Based Choice Functions in Classical DT

Given a cardinal utility u and a precise probability measure π , the following criteria are commonly for constructing choice functions:

Bayes-Criterion (a.k.a. Maximizing Expected Utility):

$$\Phi_{\text{Bayes}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \mathbb{E}_{\pi}(u \circ X)$$

Hurwicz-Criterion: Fix **optimism parameter** $\alpha \in [0, 1]$.

$$\Phi_{\text{Hurwicz}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \alpha \cdot \inf_{s \in S} u \circ X(s) + (1 - \alpha) \cdot \sup_{s \in S} u \circ X(s)$$

For $\alpha = 0 \rightarrow \Phi_{\text{maxmax}}$; for $\alpha = 1 \rightarrow \Phi_{\text{maxmin}}$.

Hodges&Lehmann-Criterion: Fix **trust parameter** $\mu \in [0, 1]$.

$$\Phi_{\text{H\&L}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \mu \cdot \inf_{s \in S} u \circ X(s) + (1 - \mu) \cdot \mathbb{E}_{\pi}(u \circ X)$$

For $\mu = 0 \rightarrow \Phi_{\text{Bayes}}$; for $\mu = 1 \rightarrow \Phi_{\text{maxmin}}$.

DIY: Something to Not Fall Asleep

Task. Work on the following questions:

- Explain what you think is the rationale underlying the three criteria introduced on the previous slide: When are they suitable and why?
- Let $\alpha = 0.1$ and $\mu = 0.2$, and let π be induced by $(0.2, 0.3, 0.4, 0.1)$. Suppose we have been given the following utility table:

	s_1	s_2	s_3	s_4
X_1	2	4	6	1
X_2	0	9	4	4
X_3	1	1	7	7
X_4	4	3	6	5

Determine the choice functions induced by the three criteria introduced on the previous slides.

Solution

Answers. We have the following:

First part is discussed in the lecture.

For the second part, we extend the utility table by several relevant columns:

	s_1	s_2	s_3	s_4	Φ_{Bayes}	Φ_{maxmax}	Φ_{maxmin}	$\Phi_{\text{H\&L}}$	Φ_{Hurwicz}
X_1	2	4	6	1	4.1	6	1	3.48	5.5
X_2	0	9	4	4	4.7	9	0	3.76	8.1
X_3	1	1	7	7	4	7	1	3.4	6.4
X_4	4	3	6	5	4.6	6	3	4.88	5.7

The unique optimal acts for the choice functions induced by the respective criterion are given as the rows with the **bold** entries.

Robustness of Bayes-Acts

Measuring Stability of Bayes-Optimal Acts

Consider the following setup:

- Let $(\mathcal{G}, S, u)$ be a finite cardinal decision problem.
- Let Δ_S denote the probability simplex over the state space (S, Σ_S) .
- Let $\pi_0 \in \Delta_S$, and let $\mathcal{G}_{\pi_0}$ the set of Bayes-optimal acts from $\mathcal{G}$ wrt π .
- For $\varepsilon \in [0, 1]$ fixed, let

$$B(\pi_0, \varepsilon) := \left\{ \pi \in \Delta_S : |\pi_0(s_j) - \pi(s_j)| \leq \varepsilon \text{ for all } j = 1, \dots, m \right\}$$

(This subsection is strongly based on [\[Jansen and Schollmeyer, 2026\]](#).)

How much do you trust your prior choice π_0 ?



Measuring Stability of Bayes-Optimal Acts

Consider the following setup:

- Let $(\mathcal{G}, S, u)$ be a finite cardinal decision problem.
- Let Δ_S denote the probability simplex over the state space (S, Σ_S) .
- Let $\pi_0 \in \Delta_S$, and let $\mathcal{G}_{\pi_0}$ the set of Bayes-optimal acts from $\mathcal{G}$ wrt π .
- For $\varepsilon \in [0, 1]$ fixed, let

$$B(\pi_0, \varepsilon) := \{\pi \in \Delta_S : |\pi(s_j) - \pi_0(s_j)| \leq \varepsilon \text{ for all } j = 1, \dots, m\}$$

Fix an act $X \in \mathcal{G}$, and define the following two quantities:

- i) The **robustness radius** of X with respect to π_0 is defined by:

$$rob(X, \pi_0) := \sup \{\varepsilon : X \in \mathcal{G}_\pi \text{ for all } \pi \in B(\pi_0, \varepsilon)\}$$

- ii) The **contamination need** of X with respect to π_0 is defined by

$$con(X, \pi_0) := \inf \{\varepsilon : X \in \mathcal{G}_\pi \text{ for some } \pi \in B(\pi_0, \varepsilon)\}$$

Measuring Stability of Bayes-Optimal Acts

Intuition behind the formulas:

- $rob(X, \pi_0)$ measures how dependent the **optimality** of X is on the correct specification of π_0 .
- $con(X, \pi_0)$ measures how dependent the **sub-optimality** of X is on the correct specification of π_0 .

Wow, that's awesome! But
how can we compute those
measures??



Computing The Contamination Need

In the situation before, the contamination need of X with respect to π_0 can be computed via the following linear program:

$$\text{con}(X, \pi_0) = \min_{\pi, \varepsilon} \left\{ \varepsilon : \begin{array}{l} \forall j : \pi_j \geq 0 \\ \sum_{j=1}^m \pi_j = 1 \\ \forall j : \pi_j \leq \pi_{0j} + \varepsilon \\ \forall j : \pi_{0j} - \varepsilon \leq \pi_j \\ \forall Y \neq X : \sum_{j=1}^m \pi_j (u \circ X(s_j) - u \circ Y(s_j)) \geq 0 \end{array} \right\}$$

Computing The Robustness Radius

In the situation before, fix $\varepsilon \in [0, 1]$, $X \in \mathcal{G}_{\pi_0}$ and $Y \neq X$.

Consider the LP:

$$R_{X,Y}(\varepsilon) = \min_{\pi} \left\{ \sum_{j=1}^m \pi_j (u \circ X(s_j) - u \circ Y(s_j)) : \begin{array}{l} \forall j : \pi_j \geq 0 \\ \sum_{j=1}^m \pi_j = 1 \\ \forall j : \pi_j \leq \pi_{0j} + \varepsilon \\ \forall j : \pi_{0j} - \varepsilon \leq \pi_j \end{array} \right\}$$

Then, setting $R(\varepsilon) := \min_{Y \neq X} R_{X,Y}(\varepsilon)$, the robustness radius of X with respect to π_0 can be computed via:

$$\text{rob}(X, \pi_0) = \sup \{ \varepsilon \geq 0 : R(\varepsilon) \geq 0 \}$$

Computing The Robustness Radius, continued

```
1: Input: decision problem  $(\mathcal{G}, S, u)$ , prior  $\pi_0$ , act  $X \in \mathcal{G}_{\pi_0}$ , tolerance  $\delta > 0$ 
2: Initialize:  $\varepsilon_L \leftarrow 0$ ,  $\varepsilon_U \leftarrow 1$ 
3: Define:  $R(\varepsilon)$  as in Proposition above
4: while  $\varepsilon_U - \varepsilon_L > \delta$  do
5:    $\varepsilon \leftarrow (\varepsilon_L + \varepsilon_U)/2$ 
6:   Compute  $R(\varepsilon)$  by solving  $R_{X,Y}(\varepsilon)$  for all  $Y \neq X$ 
7:   if  $R(\varepsilon) \geq 0$  then
8:      $\varepsilon_L \leftarrow \varepsilon$ 
9:   else
10:     $\varepsilon_U \leftarrow \varepsilon$ 
11:   end if
12: end while
13: return  $\varepsilon_L = 0$ 
```

Going from here: A Stability Criterion

Question: Can we design a meaningful decision criterion that balances both stability measures (*rob* and *con*) from before?

Going from here: A Stability Criterion

Question: Can we design a meaningful decision criterion that balances both stability measures (*rob* and *con*) from before?

Idea: In the situation before, we fix

- a **selection cost function** $c : \mathcal{G} \rightarrow \mathbb{R}_{\geq 0}$, and
- a **weight parameter** $\lambda \geq 0$.

Define the **cost-adjusted stability score** by

$$S_{\lambda, \pi_0}(X) := \begin{cases} \frac{\text{rob}(X, \pi_0)}{\max_{Y \in \mathcal{G}} \text{rob}(Y, \pi_0)} - \lambda \frac{c(X)}{\max_{Y \in \mathcal{G}} c(Y)}, & \text{if } X \in \mathcal{G}_{\pi_0}, \\ -\frac{\text{con}(X, \pi_0)}{\max_{Y \in \mathcal{G}} \text{con}(Y, \pi_0)} - \lambda \frac{c(X)}{\max_{Y \in \mathcal{G}} c(Y)}, & \text{if } X \notin \mathcal{G}_{\pi_0} \end{cases}$$

A **cost-adjusted stability optimal act** is any

$$X^* \in \arg \max_{X \in \mathcal{G}} S_{\lambda, \pi_0}(X).$$

A Toy Case Study: Portfolio Selection (Self-Study)

Situation:

We want to illustrate the cost-adjusted stability criterion using **financial return data**, interpreting the problem as decision making under uncertainty.

Rather than relying on historical observations directly, we use them to approximate the payouts under a finite set of economically meaningful scenarios.

The uncertainty is then between these scenarios.

A Toy Case Study: Portfolio Selection (Self-Study)

Constructing the Action Space:

We consider $|\mathcal{G}| = 6$ portfolio strategies constructed from exchange-traded funds (ETF's) representing major asset classes.

The underlying dataset consists of monthly log-returns over the period 2015–2024. ETF price data are obtained from Yahoo Finance, and portfolio construction follows standard asset-pricing methodology.

The six portfolios are defined as fixed convex combinations of the underlying ETF assets, representing heterogeneous investment strategies:

	Equity	Bonds	Commodities	Real Estate	Intl. Equity
X_1 (Equity Core)	0.70	0.10	0.05	0.10	0.05
X_2 (Balanced Equity)	0.50	0.20	0.05	0.15	0.10
X_3 (Multi-Asset)	0.30	0.30	0.10	0.20	0.10
X_4 (Bond Dominant)	0.15	0.70	0.05	0.05	0.05
X_5 (Real-Asset Tilt)	0.25	0.15	0.25	0.30	0.05
X_6 (Equal-Weight)	0.20	0.20	0.20	0.20	0.20

A Toy Case Study: Portfolio Selection (Self-Study)

Constructing the State Space:

To obtain a tractable representation of economic uncertainty, we define the state space as a set of four economic regimes:

$$S = \{s_1, s_2, s_3, s_4\} = \{\text{Expansion, Recession, Stagnation, Recovery}\}.$$

The returns under these regimes are approximated from the data using k -means clustering in the joint space of broad market performance and risk.

This induces a partition $\{T_1, \dots, T_4\}$ of the sample, where each T_j corresponds to the set of time periods assigned to cluster j . Each observation is assigned to the regime whose centroid is closest in Euclidean distance.

A Toy Case Study: Portfolio Selection (Self-Study)

Constructing the Utility Function:

The utility of $X \in \mathcal{M}$ in state s_j is defined as the conditional mean return:

$$u(a, s_j) := \frac{1}{|T_j|} \sum_{t \in T_j} r_t(a),$$

where T_j denotes the set of time periods classified into regime s_j , and $r_t(a)$ is the return of portfolio a at time t .

	Expansion	Recovery	Stagnation	Recession
X_1	0.021	0.059	-0.011	-0.049
X_2	0.018	0.051	-0.011	-0.037
X_3	0.015	0.051	-0.011	-0.029
X_4	0.006	0.024	-0.007	-0.024
X_5	0.014	0.042	-0.008	-0.045
X_6	0.015	0.041	-0.009	-0.024

A Toy Case Study: Portfolio Selection (Self-Study)

Constructing the Utility Function:

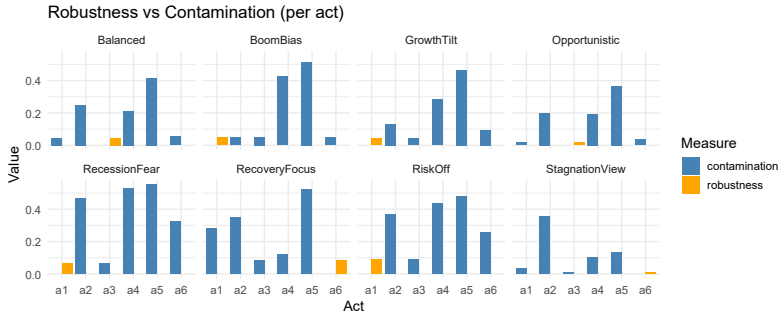
We consider a collection of eight prior distributions on the state space S . Each prior $\pi_0 \in \Delta_S$ represents a stylized belief about the likelihood of regimes and is constructed to reflect distinct economic narratives.

Concretely, we consider:

- four **regime-focused** priors with largest mass on a single state: a boom-biased prior, a recession-focused prior, a stagnation-oriented prior, and a recovery-focused prior;
- four **balanced and mixed** specifications: a uniform prior, a growth-tilted prior, a defensive prior, and an opportunistic prior.

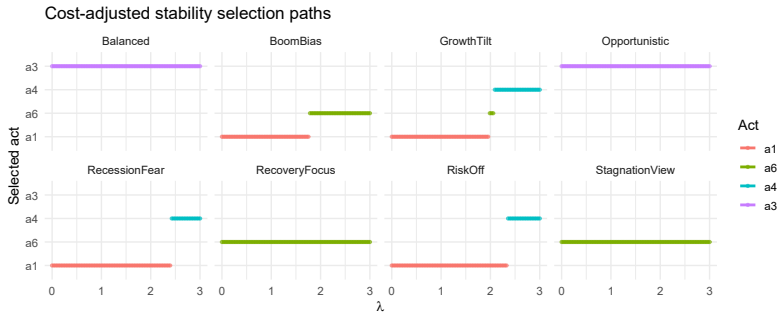
A Toy Case Study: Portfolio Selection (Self-Study)

Results:



A Toy Case Study: Portfolio Selection (Self-Study)

Results:



Statistical Decision Theory



Figure 1: Wittelsbacher Bridge in Munich

Decision Theory as a Bridge

Imprecise Probability



Erik



Arthur



Catrin

Decision Theory



Christoph



Julian

Statistics



Georg



Ryan

Machine Learning



Eyke



Yusuf



Alan

MON
IP

TUE
IP

WED
Decision Theory

THU
Statistics

FRI
ML

Statistics (from German: *Statistik*, orig. "description of a **state**, a country"^[1])

Figure 2: <https://en.wikipedia.org/wiki/Statistics>

- Etymology of Statistics (roughly): Latin: status (condition/state) → German: Statistik → English: statistics.
- In the eighteenth century, *Statistik* referred to the systematic study of the condition and affairs of a state (Thomas emphasized this root of Statistics in *Staatskunde*)
- The "science of the state": collecting information **so that the state could act** (with consequences)

Statistical decision theory

- ▶ Consider a decision model $\mathcal{G}$ with a cardinal utility $u : \mathcal{C} \rightarrow \mathbb{R}$
- ▶ Suppose we have data $\mathbf{Z} : \Omega \rightarrow \mathcal{Z}$, for which we know $\mathbf{Z} \sim \xi_s$ given $s \in S$ is the true state, i.e., S parametrizes our model.
- ▶ Instead of directly choosing acts X from $\mathcal{G}$, we now choose decision functions $d : \mathcal{Z} \rightarrow \mathcal{G}$ from a suitable $\mathbb{D} \subset \mathcal{G}^{\mathcal{Z}}$.

Statistical decision theory

- ▶ Consider a decision model $\mathcal{G}$ with a cardinal utility $u : C \rightarrow \mathbb{R}$
- ▶ Suppose we have data $\mathbf{Z} : \Omega \rightarrow \mathcal{Z}$, for which we know $\mathbf{Z} \sim \xi_s$ given $s \in S$ is the true state, i.e., S parametrizes our model.
- ▶ Instead of directly choosing acts X from $\mathcal{G}$, we now choose decision functions $d : \mathcal{Z} \rightarrow \mathcal{G}$ from a suitable $\mathbb{D} \subset \mathcal{G}^{\mathcal{Z}}$ ("statistical decision functions" [Wald, 1949])
- ▶ **Pre-data methods:** Evaluate the choice of $d \in \mathbb{D}$ under $s \in S$ by
$$U(d, s) := \mathbb{E}_{\xi_s}(g_{(d,s)}) \quad \text{with} \quad g_{(d,s)}(z) := u(d(z)(s))$$
and utilize the resulting values $[U(d, s)]_{d,s}$ for constructing a choice function on $2^{\mathbb{D}}$ (e.g. Maximin estimation).

Statistical decision theory

- ▶ Consider a decision model $\mathcal{G}$ with a cardinal utility $u : \mathcal{C} \rightarrow \mathbb{R}$
- ▶ Suppose we have data $\mathbf{Z} : \Omega \rightarrow \mathcal{Z}$, for which we know $\mathbf{Z} \sim \xi_s$ given $s \in S$ is the true state, i.e., S parametrizes our model.
- ▶ Instead of directly choosing acts X from $\mathcal{G}$, we now choose decision functions $d : \mathcal{Z} \rightarrow \mathcal{G}$ from a suitable $\mathbb{D} \subset \mathcal{G}^{\mathcal{Z}}$.
- ▶ **Pre-data methods:** Evaluate the choice of $d \in \mathbb{D}$ under $s \in S$ by
$$U(d, s) := \mathbb{E}_{\xi_s}(g_{(d,s)}) \quad \text{with} \quad g_{(d,s)}(z) := u(d(z)(s))$$
and utilize the resulting values $[U(d, s)]_{d,s}$ for constructing a choice function on $2^{\mathbb{D}}$ (e.g. Maximin estimation).
- ▶ **Post-data methods:** Determine (utilizing u and $(\xi_s)_{s \in S}$) an optimal decision $X_z \in \mathcal{G}$ for every $z \in \mathcal{Z}$ separately and choose $d^* \in \mathbb{D}$ given by $d^*(z) := X_z$ for all $z \in \mathcal{Z}$ (e.g. Maximum Likelihood estimation).

Maximum Likelihood Estimation as a post-data method

- ▶ Random variable $\mathbf{Z} : \Omega \rightarrow \mathcal{Z}$ with $\mathbf{Z} \sim \xi_s$ given $s \in S$ is true.
- ▶ $(\xi_s)_{s \in S}$ have densities $(f_s)_{s \in S}$ w.r.t. some $(\sigma$ -finite) measure λ .
- ▶ For an observation $z \in \mathcal{Z}$ of $\mathbf{Z}$, pick

$$s_{ML} \in \operatorname{argmax}\{f_s(z) : s \in S\}$$

and choose

$$X_z \in \operatorname{argmax}\{u \circ X(s_{ML}) : X \in \mathcal{G}\}$$

- ▶ Define the Maximum-Likelihood decision function d_{ML} by setting

$$d_{ML} : \mathcal{Z} \rightarrow \mathcal{G} \quad , \quad z \mapsto X_z$$

Maximin-estimation as a pre-data method

- ▶ Random variable $\mathbf{Z} : \Omega \rightarrow \mathcal{Z}$ with $\mathbf{Z} \sim \xi_s$ given $s \in S$ is true.
- ▶ $(\xi_s)_{s \in S}$ have densities $(f_s)_{s \in S}$ w.r.t. some (σ -finite) measure λ .
- ▶ Compute for $d \in \mathbb{D}$ and $s \in S$ the values

$$U(d, s) := \mathbb{E}_{\xi_s}(g_{(d,s)}) \quad \text{with} \quad g_{(d,s)}(z) := u(d(z)(s))$$

- ▶ Then every decision function $d_{MM} : \mathcal{Z} \rightarrow \mathcal{G}$ satisfying

$$d_{MM} \in \operatorname{argmax} \left\{ \inf_{s \in S} U(d, s) : d \in \mathbb{D} \right\}$$

is a Maximin decision function.

Hypothesis testing as a pre-data method

► Let $S = S_0 \dot{\cup} S_1$ (null and alternative) and consider the two acts $\mathcal{G} = \{X_0, X_1\}$, where $X_1 \hat{=}$ “reject H_0 ” and $X_0 \hat{=}$ “retain H_0 ”, $C = \{\text{correct, error I, error II}\}$ and $u(\text{correct}) = 0$, $u(\text{error I}) = -1$, $u(\text{error II}) = -c$; hence $u \circ X_1 = -\mathbb{1}_{S_0}$ and $u \circ X_0 = -c \cdot \mathbb{1}_{S_1}$ with $c > 0$.

Hypothesis testing as a pre-data method

► Let $S = S_0 \dot{\cup} S_1$ (null and alternative) and consider the two acts $\mathcal{G} = \{X_0, X_1\}$, where $X_1 \hat{=}$ “reject H_0 ” and $X_0 \hat{=}$ “retain H_0 ”, $C = \{\text{correct, error I, error II}\}$ and $u(\text{correct}) = 0$, $u(\text{error I}) = -1$, $u(\text{error II}) = -c$; hence $u \circ X_1 = -\mathbb{1}_{S_0}$ and $u \circ X_0 = -c \cdot \mathbb{1}_{S_1}$ with $c > 0$.

► A decision function $d : \mathcal{Z} \rightarrow \mathcal{G}$ is a test: it is determined by its critical region $K_d := d^{-1}(X_1)$, with **power function** $\beta_d(s) := \xi_s(K_d)$.

Hypothesis testing as a pre-data method

► Let $S = S_0 \dot{\cup} S_1$ (null and alternative) and consider the two acts $\mathcal{G} = \{X_0, X_1\}$, where $X_1 \hat{=}$ “reject H_0 ” and $X_0 \hat{=}$ “retain H_0 ”, $C = \{\text{correct, error I, error II}\}$ and $u(\text{correct}) = 0$, $u(\text{error I}) = -1$, $u(\text{error II}) = -c$; hence $u \circ X_1 = -\mathbb{1}_{S_0}$ and $u \circ X_0 = -c \cdot \mathbb{1}_{S_1}$ with $c > 0$.

► A decision function $d : \mathcal{Z} \rightarrow \mathcal{G}$ is a test: it is determined by its critical region $K_d := d^{-1}(X_1)$, with **power function** $\beta_d(s) := \xi_s(K_d)$.

► The pre-data evaluation $U(d, s) = \mathbb{E}_{\xi_s}(g_{(d,s)})$ with $g_{(d,s)}(z) = u(d(z)(s))$ is

$$U(d, s) = \begin{cases} -\beta_d(s) & s \in S_0 \text{ (error I)} \\ -c(1 - \beta_d(s)) & s \in S_1 \text{ (error II)} \end{cases}$$

Hypothesis testing as a pre-data method

► Let $S = S_0 \dot{\cup} S_1$ (null and alternative) and consider the two acts $\mathcal{G} = \{X_0, X_1\}$, where $X_1 \hat{=}$ “reject H_0 ” and $X_0 \hat{=}$ “retain H_0 ”, $C = \{\text{correct, error I, error II}\}$ and $u(\text{correct}) = 0$, $u(\text{error I}) = -1$, $u(\text{error II}) = -c$; hence $u \circ X_1 = -\mathbb{1}_{S_0}$ and $u \circ X_0 = -c \cdot \mathbb{1}_{S_1}$ with $c > 0$.

► A decision function $d : \mathcal{Z} \rightarrow \mathcal{G}$ is a test: it is determined by its critical region $K_d := d^{-1}(X_1)$, with **power function** $\beta_d(s) := \xi_s(K_d)$.

► The pre-data evaluation $U(d, s) = \mathbb{E}_{\xi_s}(g_{(d,s)})$ with $g_{(d,s)}(z) = u(d(z)(s))$ is

$$U(d, s) = \begin{cases} -\beta_d(s) & s \in S_0 \text{ (error I)} \\ -c(1 - \beta_d(s)) & s \in S_1 \text{ (error II)} \end{cases}$$

Neyman-Pearson/Karlin-Rubin is a choice function on $2^{\mathbb{D}_\alpha}$: restrict to $\mathbb{D}_\alpha = \{d : \sup_{s \in S_0} \beta_d(s) \leq \alpha\}$, then maximize β_d on S_1 to get the UMP test

Question: Why pre-data?

Hypothesis testing as a pre-data method

► Let $S = S_0 \dot{\cup} S_1$ (null and alternative) and consider the two acts $\mathcal{G} = \{X_0, X_1\}$, where $X_1 \hat{=}$ “reject H_0 ” and $X_0 \hat{=}$ “retain H_0 ”, $C = \{\text{correct, error I, error II}\}$ and $u(\text{correct}) = 0$, $u(\text{error I}) = -1$, $u(\text{error II}) = -c$; hence $u \circ X_1 = -\mathbb{1}_{S_0}$ and $u \circ X_0 = -c \cdot \mathbb{1}_{S_1}$ with $c > 0$.

► A decision function $d : \mathcal{Z} \rightarrow \mathcal{G}$ is a test: it is determined by its critical region $K_d := d^{-1}(X_1)$, with **power function** $\beta_d(s) := \xi_s(K_d)$.

► The pre-data evaluation $U(d, s) = \mathbb{E}_{\xi_s}(g_{(d,s)})$ with $g_{(d,s)}(z) = u(d(z)(s))$ is

$$U(d, s) = \begin{cases} -\beta_d(s) & s \in S_0 \text{ (error I)} \\ -c(1 - \beta_d(s)) & s \in S_1 \text{ (error II)} \end{cases}$$

Neyman-Pearson/Karlin-Rubin is a choice function on $2^{\mathbb{D}_\alpha}$: restrict to $\mathbb{D}_\alpha = \{d : \sup_{s \in S_0} \beta_d(s) \leq \alpha\}$, then maximize β_d on S_1 to get the UMP test

Question: Why pre-data? $U(d, \cdot)$ averages over data that was *not* observed.

DIY: Fisher's Lady Tasting Tea [Fisher, 1935]



Sir Ronald A. Fisher

DIY: Fisher's Lady Tasting Tea (1935), Munich Version



Hop Tea

DIY: Fisher's Lady Tasting Hop Tea (2026)



VS



DIY: Fisher's Lady Tasting Beer (2026)

Eight glasses, four from tap, four from bottles. The Lady must tell which four are from the tap (i.e., she has to pick exactly four). Let $\mathbf{Z} \in \mathcal{Z} = \{0, \dots, 4\}$ count the correctly named glasses. Under H_0 (guessing) by hypergeometric:

k	0	1	2	3	4
$\xi_{s_0}(\{k\})$	1/70	16/70	36/70	16/70	1/70

As alternatives take $S_1 = \{s_\theta : \theta \in (0, 1]\}$: with probability θ she really can tell (then $\mathbf{Z} = 4$), otherwise she guesses.

Exercise (inspired by Thomas's Decision Theory Lecture):

- (a) Determine ξ_{s_θ} and write the experiment as a problem $(\mathcal{G}, S, u)$ as above.
- (b) How large is $\mathbb{D}$ if no randomization is allowed? Compute $[U(d, s)]$ for the monotone tests $K_t = \{\mathbf{Z} \geq t\}$, $t = 0, \dots, 5$.
- (c) Fisher('s exact test) rejects iff $\mathbf{Z} = 4$: compute its size and show it is UMP at level $\alpha = 0.05$ among the K_t . **Bonus:** What happens if all 2^5 tests are allowed?
- (d) **Another Bonus:** For fixed θ : for which $(\pi(s_0), c)$ is Fisher's test a **Bayes** decision function?

Solution (a): the alternative and the decision problem

1. **Determine ξ_{s_θ} .** The hypergeometric ξ_{s_0} is fixed by the design (8 glasses, 4 from the tap, she picks 4, hence $\binom{8}{4} = 70$); θ is the **mixing weight** between “she really tells” and “she guesses”:

$$\xi_{s_\theta}(\{k\}) = \theta \cdot \mathbb{1}_{\{k=4\}} + (1 - \theta) \cdot \xi_{s_0}(\{k\})$$

k	0	1	2	3	4
$\xi_{s_\theta}(\{k\})$	$\frac{1-\theta}{70}$	$\frac{16(1-\theta)}{70}$	$\frac{36(1-\theta)}{70}$	$\frac{16(1-\theta)}{70}$	$\theta + \frac{1-\theta}{70}$

In other words, on $\{0, 1, 2, 3\}$ the shape of the distribution under the null is only **rescaled** by $(1 - \theta)$ and all removed mass now is on $k = 4$.

2. **Part: Write this testing problem as a decision problem $(\mathcal{G}, S, u)$.**

States: $S = \{s_\theta : \theta \in [0, 1]\}$, with $S_0 = \{s_0\}$ simple and $S_1 = \{s_\theta : \theta > 0\}$ **composite**.

Acts: $\mathcal{G} = \{X_0, X_1\}$; consequences C and utility u as on the testing slide.

Decisions: tests $d : \mathcal{Z} \rightarrow \mathcal{G}$, i.e. critical regions $K_d = d^{-1}(X_1) \subseteq \mathcal{Z} = \{0, \dots, 4\}$, where $\mathbf{Z} \sim \xi_{s_\theta}$.

Solution (b): size of $\mathbb{D}$, and the null tail G_0

How large is $\mathbb{D}$? A non-randomized test is any decision function $\mathcal{Z} \rightarrow \{X_0, X_1\}$, i.e. any subset $K \subseteq \mathcal{Z}$ of a set with 5 elements: $|\mathbb{D}| = 2^5 = 32$. Only the six nested ones are **monotone**:

$$\underbrace{\{0, 1, 2, 3, 4\}}_{K_0} \supset \underbrace{\{1, 2, 3, 4\}}_{K_1} \supset \underbrace{\{2, 3, 4\}}_{K_2} \supset \underbrace{\{3, 4\}}_{K_3} \supset \underbrace{\{4\}}_{K_4} \supset \underbrace{\emptyset}_{K_5}$$

Null tail. Set $G_0(t) := \xi_{s_0}(\{\mathbf{Z} \geq t\})$, i.e. cumulate $\frac{1}{70}(1, 16, 36, 16, 1)$ from the top:

t	0	1	2	3	4	5
$G_0(t)$	1	$\frac{69}{70}$	$\frac{53}{70}$	$\frac{17}{70}$	$\frac{1}{70}$	0
$1 - G_0(t)$	0	$\frac{1}{70}$	$\frac{17}{70}$	$\frac{53}{70}$	$\frac{69}{70}$	1

Power. $\beta_{K_t}(s_0) = G_0(t)$; and as the “tells”-branch always lands on $\mathbf{Z} = 4$,

$$\beta_{K_t}(s_\theta) = \theta \cdot \mathbb{1}_{\{t \leq 4\}} + (1 - \theta) G_0(t)$$

Solution (b): the value table $[U(d, s)]$

Recall $U(d, s_0) = -\beta_d(s_0)$ and $U(d, s_\theta) = -c(1 - \beta_d(s_\theta))$. For $t \leq 4$, $1 - \beta_{K_t}(s_\theta) = (1 - \theta)(1 - G_0(t))$. In other words, she must be **guessing** and the guess must fall in the retain region.

t	K_t	size $\beta(s_0)$	$U(K_t, s_0)$	$\beta(s_\theta)$	$U(K_t, s_\theta)$
0	$\{0, 1, 2, 3, 4\}$	1	-1	1	0
1	$\{1, 2, 3, 4\}$	$69/70$	$-69/70$	$\theta + (1 - \theta)\frac{69}{70}$	$-c(1 - \theta)\frac{1}{70}$
2	$\{2, 3, 4\}$	$53/70$	$-53/70$	$\theta + (1 - \theta)\frac{53}{70}$	$-c(1 - \theta)\frac{17}{70}$
3	$\{3, 4\}$	$17/70$	$-17/70$	$\theta + (1 - \theta)\frac{17}{70}$	$-c(1 - \theta)\frac{53}{70}$
4	$\{4\}$	$1/70$	$-1/70$	$\theta + (1 - \theta)\frac{1}{70}$	$-c(1 - \theta)\frac{69}{70}$
5	$\emptyset$	0	0	0	$-c$

Solution (c): size of Fisher's test, and UMP among the K_t

Since $S_0 = \{s_0\}$, we have $\text{size}(K_4) = \sup_{s \in S_0} \beta_{K_4}(s) = \xi_{s_0}(\{4\}) = \frac{1}{70} \approx 0.0143 \leq 0.05$.

Obviously, $G_0(t)$ decrease in t , so only K_4, K_5 are of level 0.05. Their powers satisfy, for every $\theta > 0$,

$$\beta_{K_5}(s_\theta) = 0 < \theta + (1 - \theta)\frac{1}{70} = \beta_{K_4}(s_\theta),$$

so K_4 has higher power than its only admissible competitor **uniformly in θ** $\Rightarrow$ K_4 is UMP at level 0.05 among the monotone tests.

Solution (c), Bonus: all 2^5 tests are allowed

Then Fisher's test is not most powerful any more. Level $0.05 = \frac{3.5}{70}$ forces K to avoid $1, 2, 3$ (null mass $\frac{16}{70}$ each), so the only candidates are $\emptyset, \{0\}, \{4\}, \{0, 4\}$, and $\text{size}(\{0, 4\}) = \frac{2}{70} \approx 0.0286 \leq 0.05$. Since

$$\beta_{\{0,4\}}(s_\theta) - \beta_{K_4}(s_\theta) = \xi_{s_\theta}(\{0\}) = \frac{1-\theta}{70} > 0 \quad (\theta < 1),$$

$K^* = \{0, 4\}$ is the unique non-randomized MP level-0.05 test.

Interpretation: $\mathbf{Z}$ is discrete, so the α -budget cannot be spent exactly. A common way to fix this is randomization.

Solution (d), when is Fisher's test a Bayes rule?

Fix θ and a prior π on $\{s_0, s_\theta\}$, $p_0 := \pi(s_0)$. Maximizing $\Phi_{\text{Bayes}}(d) = \mathbb{E}_\pi(U(d, \cdot))$ can be done **pointwise in z** : reject at z iff not rejecting costs more than rejecting,

$$(1 - p_0) c \xi_{s_\theta}(\{z\}) > p_0 \xi_{s_0}(\{z\}) \iff \text{LR}(z) > \kappa, \quad \kappa := \frac{p_0}{(1 - p_0) c}.$$

With $\text{LR} = 1 - \theta$ on $\{0, 1, 2, 3\}$ and $\text{LR}(4) = 1 + 69\theta$, Fisher's test (reject *only* at $z = 4$) is Bayes iff

$$\underbrace{1 - \theta \leq \kappa}_{\text{retain on } \{0, 1, 2, 3\}} \quad \text{and} \quad \underbrace{\kappa \leq 1 + 69\theta}_{\text{reject at } 4},$$

that is

$$\begin{aligned} 1 - \theta &\leq \frac{p_0}{(1 - p_0) c} \leq 1 + 69\theta \\ \iff \quad \frac{p_0}{(1 + 69\theta)(1 - p_0)} &\leq c \leq \frac{p_0}{(1 - \theta)(1 - p_0)} \end{aligned}$$

Outlook: Imprecise statistical decision theory

There are two different points where imprecision can kick in:

- 1.) **Imprecise distribution assumption:** Instead of having $\mathbf{Z} \sim \xi_s$ if s is the true state, we now assume $\mathbf{Z} \sim \mathcal{M}_s$, where $(\mathcal{M}_s)_{s \in S}$ is a family of credal sets. This makes necessary to modify the definition of the gain function to, e.g.,

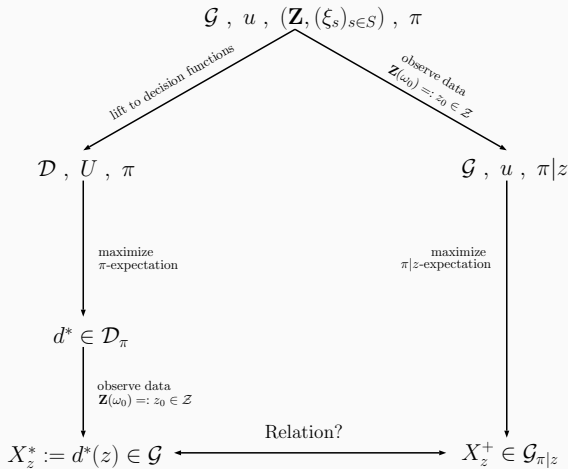
$$\tilde{U}(d, s) := \inf_{\xi \in \mathcal{M}_s} \mathbb{E}_{\xi}(g_{(d,s)}) \quad \text{with} \quad g_{(d,s)}(z) := u(d(z)(s))$$

- 2.) **Robust Bayesian analysis:** Instead of assuming a precise prior π on the state space, we here assume a set of priors Π . In this case, expectation maximization has to be adapted to ranking decision function according to the criterion function

$$B_{\Pi}(d) := \inf_{\pi \in \Pi} \mathbb{E}_{\pi}(U(d, \cdot))$$

Imprecise statistical decision theory, cont.

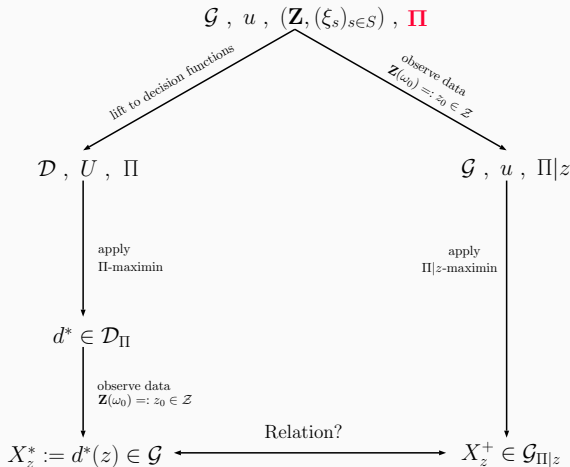
Two ways for including precise prior knowledge π :



Theorem: Define $d^+ : \mathcal{Z} \rightarrow \mathcal{G}$, $z \mapsto X_z^+$. Then $d^+ \in \mathcal{D}_\pi$.

Imprecise statistical decision theory, cont.²

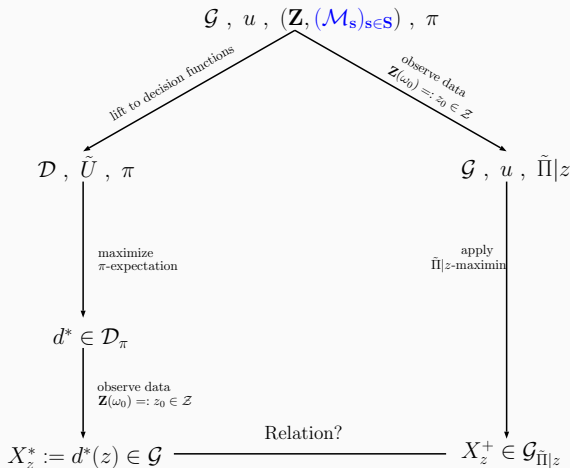
Under imprecise prior knowledge:



In general: $d^+ \notin \mathcal{D}_\Pi$.

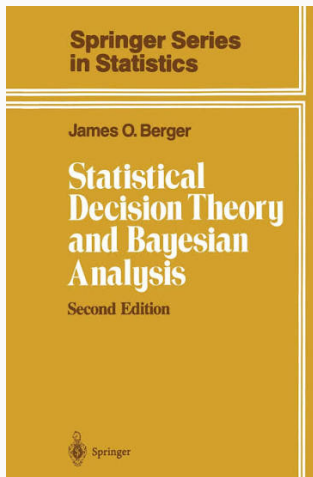
Imprecise statistical decision theory, cont.³

Imprecise distributional assumption:



In general: $d^+ \notin \mathcal{D}_\pi$.

Further Reading: Statistical Decision Theory



This book covers the basic material of statistical decision theory in an accessible manner. The prerequisites are relatively modest.

- **Statistical level:** moderately advanced statistics
- **Mathematical level:** some calculus

Machine Learning as Decision Problem(s)

(Supervised) Machine Learning

► **Instances** $x \in \mathcal{X}$, **outcomes** $y \in \mathcal{Y}$; data are generated i.i.d. by an **unknown** probability measure $P \in \mathbb{P}(\mathcal{X} \times \mathcal{Y})$.

► **Hypothesis space** $\mathcal{H}$ of predictors $h : \mathcal{X} \rightarrow \mathbb{P}(\mathcal{Y})$, evaluated by a **loss** $L : \mathcal{Y} \times \mathbb{P}(\mathcal{Y}) \rightarrow \mathbb{R}$.

► The learner would like to solve

$$h^* \in \operatorname{argmin}\{R(h) : h \in \mathcal{H}\} \quad , \quad R(h) := \mathbb{E}_{(x,y) \sim P} L(y, h(x))$$

► Since P is unknown, one minimizes instead, given $D = \{(x_i, y_i)\}_{i=1}^N$,

$$R_{\text{emp}}(h) := \frac{1}{N} \sum_{i=1}^N L(y_i, h(x_i)) \quad (\text{Empirical Risk Minimization})$$

► A **learning algorithm** is a map $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^N \rightarrow \mathcal{H}$.

Claim: Supervised Machine Learning $\cong$ Pre-data + Post-data decisions

Two Statistical Decision Problems

Both are instances of $(\mathcal{G}, S, u)$ with data $\mathbf{Z} \sim \xi_s$ and decision functions $d : \mathcal{Z} \rightarrow \mathcal{G}$. Just like above, the state $s \in S$ indexes the candidate measures $P_s \in \mathbb{P}(\mathcal{X} \times \mathcal{Y})$.

	Problem I: prediction	Problem II: learning
observation $\mathbf{Z}$	query $x \in \mathcal{X}$	sample $D \in (\mathcal{X} \times \mathcal{Y})^N$
sampling model ξ_s	$\mathcal{X}$ -marginal of P_s	$P_s^{\otimes N}$
acts $\mathcal{G}$	predictions $p \in \mathbb{P}(\mathcal{Y})$	hypotheses $h \in \mathcal{H}$
decision functions $\mathbb{D}$	hypotheses $h : \mathcal{X} \rightarrow \mathbb{P}(\mathcal{Y})$	learners $\mathcal{A} : \mathcal{Z} \rightarrow \mathcal{H}$
post-data solution	(pointwise) Bayes act	empirical risk minimization
pre-data value $U(d, s)$	$-R_{P_s}(h)$ (the risk)	$-\mathbb{E}_D[R_{P_s}(\mathcal{A}(D))]$ (generalization)

Problem I, Post-Data: Pointwise Bayes Acts

The query x has been observed; assume for the moment that the state s is known.

- **Residual states:** $\mathcal{Y}$ (which outcome materializes)
- **Consequences:** $C := \mathcal{Y} \times \mathbb{P}(\mathcal{Y})$ (what was true, what was predicted)
- **Acts:** $\mathcal{G}_x := \{X_p : p \in \mathbb{P}(\mathcal{Y})\}$ with $X_p(y) := (y, p)$
- **Utility:** $u(y, p) := -L(y, p)$
- $\mathcal{I}_U$: the true conditional $\pi_x := P_s(\cdot \mid x)$

Problem I, Post-Data: Pointwise Bayes Acts

The query x has been observed; assume for the moment that the state s is known.

- **Residual states:** $\mathcal{Y}$ (which outcome materializes)
- **Consequences:** $C := \mathcal{Y} \times \mathbb{P}(\mathcal{Y})$ (what was true, what was predicted)
- **Acts:** $\mathcal{G}_x := \{X_p : p \in \mathbb{P}(\mathcal{Y})\}$ with $X_p(y) := (y, p)$
- **Utility:** $u(y, p) := -L(y, p)$
- $\mathcal{I}_U$: the true conditional $\pi_x := P_s(\cdot \mid x)$

The Bayes criterion then comes down to

$$\Phi_{\text{Bayes}}(X_p) = \mathbb{E}_{\pi_x}(u \circ X_p) = -\mathbb{E}_{y \sim \pi_x} L(y, p) =: -R(p \mid x)$$

Problem I, Post-Data: Pointwise Bayes Acts

The query x has been observed; assume for the moment that the state s is known.

- **Residual states:** $\mathcal{Y}$ (which outcome materializes)
- **Consequences:** $C := \mathcal{Y} \times \mathbb{P}(\mathcal{Y})$ (what was true, what was predicted)
- **Acts:** $\mathcal{G}_x := \{X_p : p \in \mathbb{P}(\mathcal{Y})\}$ with $X_p(y) := (y, p)$
- **Utility:** $u(y, p) := -L(y, p)$
- $\mathcal{I}_U$: the true conditional $\pi_x := P_s(\cdot \mid x)$

The Bayes criterion then comes down to

$$\Phi_{\text{Bayes}}(X_p) = \mathbb{E}_{\pi_x}(u \circ X_p) = -\mathbb{E}_{y \sim \pi_x} L(y, p) =: -R(p \mid x)$$

- ▶ L is a **proper scoring rule** $\iff X_{\pi_x} \in \text{ch}_{\text{Bayes}}(\mathcal{G}_x)$ for every π_x ("telling the truth is a Bayes act") [Gneiting and Raftery, 2007]
- ▶ $-\Phi_{\text{Bayes}}$ of a Bayes act is the **generalized entropy** of π_x [DeGroot, 1962, Grünwald and Dawid, 2004]. This is **aleatoric** uncertainty only (because we assume s to be known)

DIY: Pointwise Bayes Acts (Binary $\mathcal{Y}$)

Fix x and a state s , so $Y \in \{0, 1\}$ with $q := \pi_x(1) = P(Y = 1 \mid x)$ known. A prediction is $p \in [0, 1]$ and the corresponding act is X_p ; recall $\Phi_{\text{Bayes}}(X_p) = -R(p \mid x)$ with $R(p \mid x) = \mathbb{E}_{y \sim q} L(y, p)$.

Exercise: For each loss, find the Bayes act $p^* = \arg \min_p R(p \mid x)$ as well as $\min_p R(p \mid x)$.

(a) **Brier Loss:** $L(y, p) = (y - p)^2$.

(b) **Absolute Loss:** $L(y, p) = |y - p|$.

Which loss makes *truthful reporting* $p^* = q$ optimal (i.e. is **proper**)? What does the other one reward instead?

DIY: Pointwise Bayes Acts (Binary $\mathcal{Y}$)

Fix x and a state s , so $Y \in \{0, 1\}$ with $q := \pi_x(1) = P(Y = 1 \mid x)$ known. A prediction is $p \in [0, 1]$ and the corresponding act is X_p ; recall $\Phi_{\text{Bayes}}(X_p) = -R(p \mid x)$ with $R(p \mid x) = \mathbb{E}_{y \sim q} L(y, p)$.

Exercise: For each loss, find the Bayes act $p^* = \arg \min_p R(p \mid x)$ as well as $\min_p R(p \mid x)$.

(a) **Brier Loss:** $L(y, p) = (y - p)^2$.

(b) **Absolute Loss:** $L(y, p) = |y - p|$.

Which loss makes *truthful reporting* $p^* = q$ optimal (i.e. is **proper**)? What does the other one reward instead?

Solution.

(a) $R(p) = q(1 - p)^2 + (1 - q)p^2$, so $R'(p) = 2(p - q) \Rightarrow p^* = q$ (**proper**). Value $R(q) = q(1 - q)$ (called the **Gini** generalized entropy).

(b) $R(p) = q(1 - p) + (1 - q)p = q + (1 - 2q)p$ is linear in p , so $p^* = \mathbb{1}_{\{q > \frac{1}{2}\}}$ (if $q \neq \frac{1}{2}$), i.e., a **hard label**, not q (**improper**): it rewards guessing the majority class (overconfident, not calibrated)

Problem II, Pre-Data: Generalization

- ▶ A learning algorithm is a decision function of Problem II,

$$d_A : \mathcal{Z} \rightarrow \mathcal{G}_{\mathcal{H}} \quad , \quad D \mapsto X_{A(D)} \quad , \quad \mathbb{D} \subseteq \mathcal{G}_{\mathcal{H}}^{\mathcal{Z}}$$

so choosing a learner = choosing an element of $\mathbb{D}$.

Problem II, Pre-Data: Generalization

- ▶ A **learning algorithm** is a **decision function** of Problem II,

$$d_{\mathcal{A}} : \mathcal{Z} \rightarrow \mathcal{G}_{\mathcal{H}} \quad , \quad D \mapsto X_{\mathcal{A}(D)} \quad , \quad \mathbb{D} \subseteq \mathcal{G}_{\mathcal{H}}^{\mathcal{Z}}$$

so choosing a learner = choosing an element of $\mathbb{D}$.

- ▶ Its pre-data evaluation is

$$U(d_{\mathcal{A}}, s) = \mathbb{E}_{\xi_s}(g_{(d_{\mathcal{A}}, s)}) = -\mathbb{E}_{D \sim P_s^{\otimes N}} \left[R_{P_s}(\mathcal{A}(D)) \right]$$

i.e. exactly the negative **expected generalization error (population risk)** of $\mathcal{A}$.

Problem II, Pre-Data: Generalization

- ▶ A **learning algorithm** is a **decision function** of Problem II,

$$d_{\mathcal{A}} : \mathcal{Z} \rightarrow \mathcal{G}_{\mathcal{H}} \quad , \quad D \mapsto X_{\mathcal{A}(D)} \quad , \quad \mathbb{D} \subseteq \mathcal{G}_{\mathcal{H}}^{\mathcal{Z}}$$

so choosing a learner = choosing an element of $\mathbb{D}$.

- ▶ Its pre-data evaluation is

$$U(d_{\mathcal{A}}, s) = \mathbb{E}_{\xi_s}(g_{(d_{\mathcal{A}}, s)}) = -\mathbb{E}_{D \sim P_s^{\otimes N}} \left[R_{P_s}(\mathcal{A}(D)) \right]$$

i.e. exactly the negative **expected generalization error (population risk)** of $\mathcal{A}$.

- ▶ **Question:** How would we embed ERM as a post-data method?

A Dictionary For Friday

Statistical decision theory	Machine learning
states $s \in S$	candidate data-generating processes P_s
consequences C	(outcome, prediction) pairs resp. loss lotteries
cardinal utility $u : C \rightarrow \mathbb{R}$	negative loss $-L$
acts of Problem I	predictions $p \in \mathbb{P}(\mathcal{Y})$
decision functions of Problem I	hypotheses $h \in \mathcal{H}$
pre-data value $U(d_h, s)$	risk $-R_{P_s}(h)$
acts of Problem II	hypotheses, $\mathcal{G}_{\mathcal{H}} \hat{=} \mathcal{H}$
decision functions of Problem II	learning algorithms $\mathcal{A}$
sample space $\mathcal{Z}$, model ξ_s	$(\mathcal{X} \times \mathcal{Y})^N$, $P_s^{\otimes N}$ (i.i.d. assumption)
post-data method d_{ML}	empirical risk minimization
pre-data value $U(d_{\mathcal{A}}, s)$	negative expected generalization risk
$\mathcal{I}_U$ on S	second-order belief Q / credal set $\mathcal{K}$
choice function ch	set-valued / credal predictor

Changing the Criterion Changes the Learning Paradigm

Post-data: criteria on $\mathcal{G}_{\mathcal{H}}$, with the information about S extracted from D .

Information state / criterion	Resulting learner
plug-in $\delta_{s_{ML}}$	empirical risk minimization
posterior $\pi(\cdot D)$, Φ_{Bayes}	Bayesian ML: minimize $R_{\bar{P}}(h)$, $\bar{P} = \int P_s d\pi(s D)$
posterior + plug-in (MAP)	penalized / regularized ERM
credal set $\mathcal{M}(D)$, Γ -maximin	robust Bayesian ML
credal set $\mathcal{M}(D)$, E -admissibility	set-valued output: a credal set of hypotheses

Changing the Criterion Changes the Learning Paradigm

Post-data: criteria on $\mathcal{G}_{\mathcal{H}}$, with the information about S extracted from D .

Information state / criterion	Resulting learner
plug-in $\delta_{s_{ML}}$	empirical risk minimization
posterior $\pi(\cdot D)$, Φ_{Bayes}	Bayesian ML: minimize $R_{\bar{P}}(h)$, $\bar{P} = \int P_s d\pi(s D)$
posterior + plug-in (MAP)	penalized / regularized ERM
credal set $\mathcal{M}(D)$, Γ -maximin	robust Bayesian ML
credal set $\mathcal{M}(D)$, E -admissibility	set-valued output: a credal set of hypotheses

Pre-data: criteria on $\mathbb{D}$ using the values $[U(d, s)]_{d,s}$.

Criterion	Resulting learner
Φ_{maxmin} over S	minimax / distributionally robust learning
Φ_{Bayes} w.r.t. a prior π	Bayes-optimal learner
minimax regret	minimax-regret learning

Changing the Criterion Changes the Learning Paradigm

Post-data: criteria on $\mathcal{G}_{\mathcal{H}}$, with the information about S extracted from D .

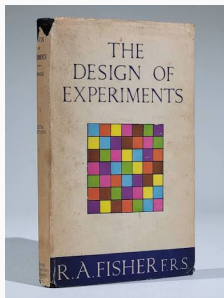
Information state / criterion	Resulting learner
plug-in $\delta_{s_{ML}}$	empirical risk minimization
posterior $\pi(\cdot D)$, Φ_{Bayes}	Bayesian ML: minimize $R_{\bar{P}}(h)$, $\bar{P} = \int P_s d\pi(s D)$
posterior + plug-in (MAP)	penalized / regularized ERM
credal set $\mathcal{M}(D)$, Γ -maximin	robust Bayesian ML
credal set $\mathcal{M}(D)$, E -admissibility	set-valued output: a credal set of hypotheses

Pre-data: criteria on $\mathbb{D}$ using the values $[U(d, s)]_{d,s}$.

Criterion	Resulting learner
Φ_{maxmin} over S	minimax / distributionally robust learning
Φ_{Bayes} w.r.t. a prior π	Bayes-optimal learner
minimax regret	minimax-regret learning

Outlook to Friday: A deep ensemble or a credal wrapper is nothing but a non-singleton **choice set** $ch(\mathcal{G}_{\mathcal{H}})$

Yet Another Decision Problem: Querying Data



R. A. Fisher:
The Design of
Experiments

The whole art and practice of scientific experimentation is comprised in the skillful interrogation of Nature. Observation has provided the scientist with a picture of Nature in some aspect, which has all the imperfections of a voluntary statement. He wishes to check his interpretation of this statement by asking specific questions. . .

— Joan Fisher Box, *R. A. Fisher: The Life of a Scientist* (1978);
cited after [Pearl and Mackenzie, 2018, p. 144]

Yet Another Decision Problem: Querying Data

- **States:** S as above, still indexing $s \mapsto P_s$
- **Acts:** simply (a subset of the) instance space $\mathcal{G} := \mathcal{Z}$ or $\mathcal{G} := \mathcal{X}$
- **Consequences:** pairs (s, z) (or (s, x))
- **Utility:**
 - Given prior π on S , Information Gain (IG): $H(\pi(s)) - H(\pi(s \mid z))$ with $H(\cdot)$ the entropy.
 - Given $D = \{(x_i, y_i)\}_{i=1}^N$, likelihood $\ell_s(D \cup z)$

Yet Another Decision Problem: Querying Data

- **States:** S as above, still indexing $s \mapsto P_s$
- **Acts:** simply (a subset of the) instance space $\mathcal{G} := \mathcal{Z}$ or $\mathcal{G} := \mathcal{X}$
- **Consequences:** pairs (s, z) (or (s, x))
- **Utility:**
 - Given prior π on S , Information Gain (IG): $H(\pi(s)) - H(\pi(s | z))$ with $H(\cdot)$ the entropy.
 - Given $D = \{(x_i, y_i)\}_{i=1}^N$, likelihood $\ell_s(D \cup z)$

The second decision problem has the (possibly infinite) **utility table**

	s_1		s_m	...
z_1	$\ell_{s_1}(D \cup z_1)$		$\ell_{s_m}(D \cup z_1)$	...
$\vdots$	...		...	...
z_n	$\ell_{s_1}(D \cup z_n)$		$\ell_{s_m}(D \cup z_n)$	...

We will use this as a running example to illustrate some IP-based decision rules following Rodemann, Jansen, Schollmeyer, Augustin (2023): *In all likelihoods: Robust selection of pseudo-labeled data. ISIPTA*. [Rodemann et al., 2023]

Selection of Pseudo-Labeled Data

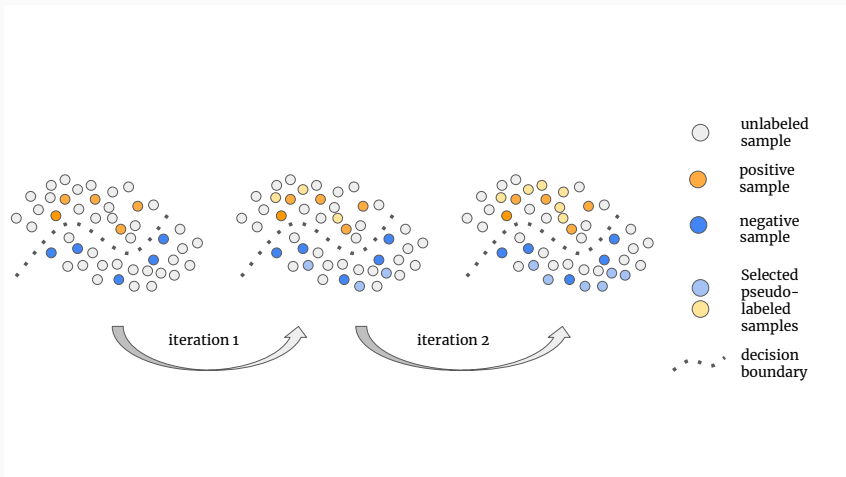
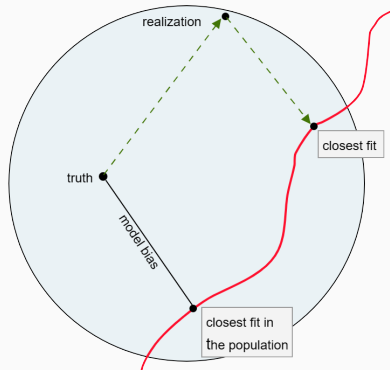
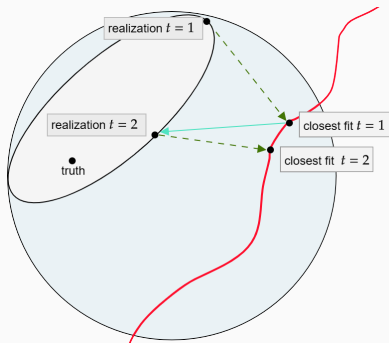


Figure 3: Sketch of Self-Training for Binary Classification

Reciprocal Learning, Informal



(a) Classical statistical learning

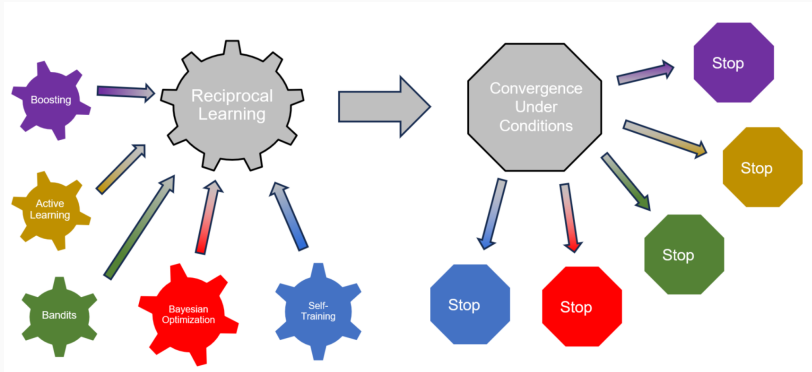


(b) Reciprocal learning

(a) Classical statistical learning fits a model from the model space (restricted by the red curve) to a realized sample from the sample space (blue-grey), see [Hastie et al., 2009, Fig. 7.2].

(b) In reciprocal learning, the sample changes *in response* to the model fit.

Short Commercial Break: Reciprocal Learning



Full Paper (arXiv)



arxiv.org/abs/2408.06257

Short Paper (NeurIPS)



[paper](#) · [slides](#) · [poster](#) · [OpenReview](#)

Julian Rodemann, Christoph Jansen, Georg Schollmeyer (2024).
Reciprocal Learning, *Advances in Neural Information Processing Systems 37 (NeurIPS)*.

Task-Aware Calibration: Provably Optimal Decoding in LLMs

Tim Tomov^{*1,2,3} Dominik Fuchsgruber^{*1,2} Rajeev Verma⁴ Stephan Günnemann^{1,2,3}
{tim.tomov,d.fuchsgruber,s.guennemann}@tum.de & r.verma@uva.nl

¹School of Computation, Information & Technology, Technical University of Munich
²Munich Data Science Institute ³Munich Center for Machine Learning ⁴University of Amsterdam

Abstract

LLM decoding often relies on the model's predictive distribution to generate an output. Consequently, misalignment with respect to the true generating distribution leads to suboptimal decisions in practice. While a natural solution is to calibrate the model's output distribution, for LLMs, this is ill-posed at the combinatorially vast level of free-form language. We address this by building on the insight that in many tasks, these free-form outputs can be interpreted in a semantically meaningful latent structure, for example, discrete class labels, integers, or sets. We introduce *task calibration* as a paradigm to calibrate the model's predictive distribution in the task-induced latent space. We apply a decision-theoretic result to show that Minimum Bayes Risk (MBR) decoding on the *task-calibrated* latent distribution is the optimal decoding strategy on latent model beliefs. Empirically, it consistently improves generation quality across different tasks and baselines. We also introduce Task Calibration Error (TCE), an application-aware calibration metric that quantifies the excess loss due to miscalibration. Our work demonstrates that task calibration enables more reliable model decisions across various tasks and applications.

Decision Making Under Imprecise Probabilities

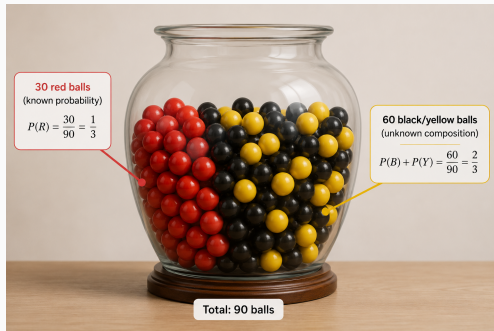
Motivation: Ellsberg's Paradox (1961)

In his paper [Ellsberg, 1961], Daniel Ellsberg performed a thought experiment among statisticians and economists, i.e., people familiar with probability theory.

Motivation: Ellsberg's Paradox (1961)

In his paper [Ellsberg, 1961], Daniel Ellsberg performed a thought experiment among statisticians and economists, i.e., people familiar with probability theory.

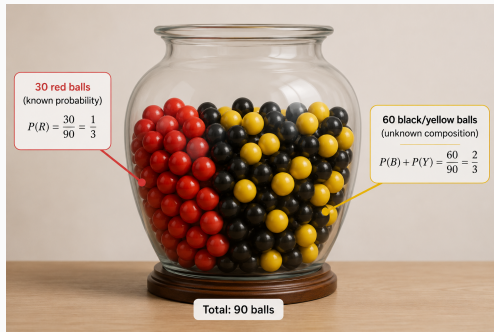
Concretely, he asked people to imagine the following setup:



Motivation: Ellsberg's Paradox (1961)

In his paper [Ellsberg, 1961], Daniel Ellsberg performed a thought experiment among statisticians and economists, i.e., people familiar with probability theory.

Concretely, he asked people to imagine the following setup:



The key distinction:

Risk: $P(R)$ is known

Ambiguity: Only $P(\{B, Y\})$ is known

The Four Gambles

Ellsberg presented the following two pairs of gambles to the participants:

Gamble A: Receive \$100 if the drawn ball is red.

$$P(R) = \frac{1}{3}$$

vs.

Gamble B: Receive \$100 if the drawn ball is black.

$$P(B) \in [0, \frac{2}{3}]$$

The Four Gambles

Ellsberg presented the following two pairs of gambles to the participants:

Gamble A: Receive \$100 if the drawn ball is red.

$$P(R) = \frac{1}{3}$$

vs.

Gamble B: Receive \$100 if the drawn ball is black.

$$P(B) \in [0, \frac{2}{3}]$$

Gamble C: Receive \$100 if the drawn ball is red or yellow.

$$P(\{R, Y\}) \in [\frac{1}{3}, 1]$$

vs.

Gamble D: Receive \$100 if the drawn ball is black or yellow.

$$P(\{B, Y\}) = \frac{2}{3}$$

The Reactions

[...] Responses do vary. There are those who do not violate the axioms, or say they won't, even in these situations (e.g., G. Debreu, R. Schlaiffer, P. Samuelson); such subjects tend to apply the axioms rather than their intuition, and when in doubt, to apply some form of the Principle of Insufficient Reason. Some violate the axioms cheerfully, even with gusto (J. Marschak, N. Dalkey); others sadly but persistently, having looked into their hearts, found conflicts with the axioms and decided, in Samuelson's phrase, to satisfy their preferences and let the axioms satisfy themselves. Still others (H. Raiffa) tend, intuitively, to violate the axioms but feel guilty about it and go back into further analysis.

*The important finding is that, after rethinking all their "offending" decisions in the light of the axioms, a number of people who are not only sophisticated but reasonable decide that they wish to persist in their choices. This includes people who previously felt a "first-order commitment" to the axioms, many of them surprised and some dismayed to find that they wished, in these situations, to violate the Sure-thing Principle. Since this group included **L. J. Savage**, when last tested by me (I have been reluctant to try him again), it seems to deserve respectful consideration.[...]*

Observed Preferences: Ambiguity Aversion

The largest group of participants provided the following combinations of answers:



Observed Preferences: Ambiguity Aversion

The largest group of participants provided the following combinations of answers:



Interpretation:

- People prefer the gamble with a **known probability** over one involving an unknown probability.
- The unknown black/yellow composition creates **ambiguity**.
- Risk is acceptable; ambiguity is avoided.

Observed Preferences: Ambiguity Aversion

The largest group of participants provided the following combinations of answers:



Interpretation:

- People prefer the gamble with a **known probability** over one involving an unknown probability.
- The unknown black/yellow composition creates **ambiguity**.
- Risk is acceptable; ambiguity is avoided.

Task. Show that there is no pair of prior probability π and cardinal utility u such that both majority choices are simultaneously reflected by the expected utility order based on (π, u) .

Conflict with Expected Utility Theory

Solution. Assume such (π, u) exists. Set $p = \pi(B)$, then $\pi(Y) = \frac{2}{3} - p$.

Expected utility maximization then implies:

$$E_{\pi}(u \circ A) = \frac{1}{3}u(100) \quad \text{and} \quad E_{\pi}(u \circ B) = pu(100)$$

Therefore: $A \succ B \iff \frac{1}{3} > p$

For the second pair:

$$E_{\pi}(u \circ C) = (1 - p)u(100) \quad \text{and} \quad E_{\pi}(u \circ D) = \frac{2}{3}u(100)$$

Hence: $D \succ C \iff \frac{2}{3} > 1 - p \iff p > \frac{1}{3}$

This produces a contradiction:

$$A \succ B \Rightarrow p < \frac{1}{3} \quad D \succ C \Rightarrow p > \frac{1}{3}$$

No single probability π can explain both choices.

Sets of Probabilities to the Rescue

In [Weichselberger and Augustin, 1998], the following way of resolving the paradox is discussed:

- (1) Model uncertainty using the set of all compatible probability measures:

$$\mathcal{M} := \left\{ \pi : \pi(R) = \frac{1}{3} \wedge \pi(\{B, Y\}) = \frac{2}{3} \right\}$$

- (2) Rank options based on the worst-case expected utility

$$\underline{\mathbb{E}}_{\mathcal{M}}(u \circ X) := \inf_{\pi \in \mathcal{M}} \mathbb{E}_{\pi}(u \circ X)$$

Sets of Probabilities to the Rescue

In [Weichselberger and Augustin, 1998], the following way of resolving the paradox is discussed:

- (1) Model uncertainty using the set of all compatible probability measures:

$$\mathcal{M} := \left\{ \pi : \pi(R) = \frac{1}{3} \wedge \pi(\{B, Y\}) = \frac{2}{3} \right\}$$

- (2) Rank options based on the worst-case expected utility

$$\underline{\mathbb{E}}_{\mathcal{M}}(u \circ X) := \inf_{\pi \in \mathcal{M}} \mathbb{E}_{\pi}(u \circ X)$$

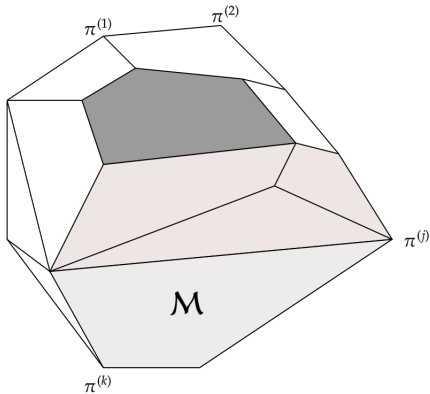
Task. Verify that the proposed criterion based on sets of probabilities indeed manages to recover both expressed majority preferences at the same time!

Decision Theory with Imprecise Probabilities

$\mathcal{I}_P$ Information on preferences is characterized by a cardinal utility function $u : C \rightarrow \mathbb{R}$ that is unique up to positive affine transformations.

$\mathcal{I}_U$ Information on the generation of states is characterized by **credal set** of probability measure on (S, Σ_S) .

Imprecise Probabilities for Decision Making



Imprecise Probabilities for Decision Making

Decision Theory with Imprecise Probabilities

$\mathcal{I}_P$ Information on preferences is characterized by a cardinal utility function $u : C \rightarrow \mathbb{R}$ that is unique up to positive affine transformations.

$\mathcal{I}_U$ Information on the generation of states is characterized by **credal set** of probability measure on (S, Σ_S) .

Modelling $\mathcal{I}_U$ via Convex Credal sets

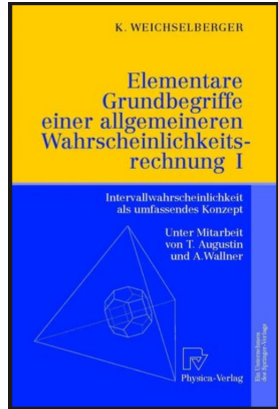
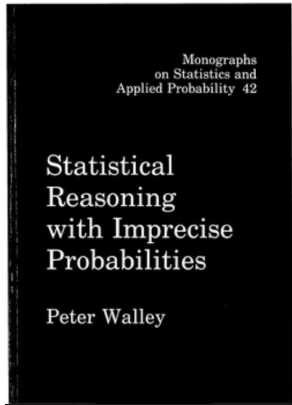
The uncertainty among the elements of S is described by a polyhedral **credal set** of probability measures of the form

$$\mathcal{M} = \left\{ \pi \in \Delta_S : \underline{b}_\ell \leq \mathbb{E}_\pi(f_\ell) \leq \bar{b}_\ell \text{ for } \ell = 1, \dots, r \right\}$$

where Δ_S is the set of all probability measures on (S, Σ_S) and

- $f_1, \dots, f_r : S \rightarrow \mathbb{R}$ are real-valued mappings and
- $\underline{b}_\ell \leq \bar{b}_\ell$, $\ell = 1, \dots, r$, are lower and upper expectation bounds.

Imprecise Probabilities for Decision Making



Walley: behaviorally motivated axioms, subjective theory for DM

→ generalizing De Finetti, Savage, ...

Weichselberger: interpretation free axioms and theory for DM

→ generalizing Kolmogorov

Criterion-Based Choice Functions in DT under IP

Given a cardinal utility u and a credal set $\mathcal{M}$, the following criteria are commonly used for constructing choice functions:

Criterion-Based Choice Functions in DT under IP

Given a cardinal utility u and a credal set $\mathcal{M}$, the following criteria are commonly used for constructing choice functions:

Γ -Maximin criterion:

$$\Phi_{\Gamma\text{-Maximin}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \underline{\mathbb{E}}_{\mathcal{M}}(u \circ X) := \inf_{\pi \in \mathcal{M}} \mathbb{E}_{\pi}(u \circ X)$$

Criterion-Based Choice Functions in DT under IP

Given a cardinal utility u and a credal set $\mathcal{M}$, the following criteria are commonly used for constructing choice functions:

Γ -Maximin criterion:

$$\Phi_{\Gamma\text{-Maximin}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \underline{\mathbb{E}}_{\mathcal{M}}(u \circ X) := \inf_{\pi \in \mathcal{M}} \mathbb{E}_{\pi}(u \circ X)$$

Γ -Maximax criterion:

$$\Phi_{\Gamma\text{-Maximax}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \overline{\mathbb{E}}_{\mathcal{M}}(u \circ X) := \sup_{\pi \in \mathcal{M}} \mathbb{E}_{\pi}(u \circ X)$$

Criterion-Based Choice Functions in DT under IP

Given a cardinal utility u and a credal set $\mathcal{M}$, the following criteria are commonly used for constructing choice functions:

Γ -Maximin criterion:

$$\Phi_{\Gamma\text{-Maximin}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \underline{\mathbb{E}}_{\mathcal{M}}(u \circ X) := \inf_{\pi \in \mathcal{M}} \mathbb{E}_{\pi}(u \circ X)$$

Γ -Maximax criterion:

$$\Phi_{\Gamma\text{-Maximax}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \overline{\mathbb{E}}_{\mathcal{M}}(u \circ X) := \sup_{\pi \in \mathcal{M}} \mathbb{E}_{\pi}(u \circ X)$$

Γ -Maximix criterion: Fix **ambiguity aversion parameter** $\eta \in [0, 1]$.

$$\Phi_{\Gamma\text{-Maximix}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \eta \cdot \underline{\mathbb{E}}_{\mathcal{M}}(u \circ X) + (1 - \eta) \cdot \overline{\mathbb{E}}_{\mathcal{M}}(u \circ X)$$

For $\eta = 0 \rightarrow \Phi_{\Gamma\text{-Maximin}}$; for $\eta = 1 \rightarrow \Phi_{\Gamma\text{-Maximax}}$.

Running Example (Pseudo-Label Selection): Γ -Maximin

Recall the querying-data problem with $\mathcal{G} := \mathcal{X}$ *unlabeled* data and $\hat{y}(\cdot)$ the current predictor:

- *States*: $s \in S$, indexing $s \mapsto P_s$
- *Acts*: "add z with its pseudo-label $\hat{y}(z)$ "
- *Utility*: $(u \circ x)(s) = \ell_s(D \cup (x, \hat{y}(x)))$
- *Beliefs*: credal set $\mathcal{M}$ on (S, Σ_S)

Γ -Maximin for PLS

$$\Phi_{\Gamma\text{-Maximin}} : \mathcal{G} \rightarrow \mathbb{R}, \quad X_z \mapsto \inf_{\pi \in \mathcal{M}} \mathbb{E}_{\pi} \ell_s(D \cup (x, \hat{y}(x)))$$

Add the $x^* \in \arg \max_{x \in \mathcal{G}} \Phi_{\Gamma\text{-Maximin}}(x)$.

- For precise π , $\mathbb{E}_{\pi}(u \circ x)$ is the posterior predictive of $\mathcal{D} \cup (x, \hat{y}(x)) \Rightarrow$ Bayes-optimal PLS.
- Γ -Maximin takes its lower envelope: add the instance that stays plausible under the $\pi \in \mathcal{M}$ contradicting the (possibly overfitted) likelihood the most.

DIY: Something to Not Fall Asleep

Task. Work on the following questions:

- Explain what you think is the rationale underlying the three criteria introduced on the previous slide: When are they suitable and why?
- Let $\eta = 0.2$, and let $\mathcal{M}$ be given by

$$\mathcal{M} = \{ \pi : \pi(s_1) \geq \pi(s_2) \geq \pi(s_3) \geq \pi(s_4) \}$$

Suppose we have been given the following utility table:

	s_1	s_2	s_3	s_4
X_1	2	4	6	1
X_2	0	9	4	4
X_3	1	1	7	7
X_4	4	3	6	5

Determine the choice functions induced by the three criteria introduced on the previous slides.

Solution

Answers. We have the following:

First part is discussed in the lecture.

For the second part, we extend the utility table by several relevant columns:

	s_1	s_2	s_3	s_4	$\Phi_{\Gamma-\text{Maximin}}$	$\Phi_{\Gamma-\text{Maximin}}$	$\Phi_{\Gamma-\text{Maximix}}$
X_1	2	4	6	1	2	4	3.6
X_2	0	9	4	4	0	4.5	3.6
X_3	1	1	7	7	1	4	3.4
X_4	4	3	6	5	3.5	4.5	4.3

The optimal acts for the choice functions induced by the respective criterion are given as the rows with the **bold** entries.

Non-Criterion-Based Choice Functions

Given a cardinal utility u and a credal set $\mathcal{M}$, there are also ways of constructing choice functions that do not rely on real-valued criteria:

Non-Criterion-Based Choice Functions

Given a cardinal utility u and a credal set $\mathcal{M}$, there are also ways of constructing choice functions that do not rely on real-valued criteria:

Walley's Maximality (a.k.a. Maximality):

([Walley, 1991])

Call X **maximal** in $\mathcal{D} \subseteq \mathcal{G}$ wrt $\mathcal{M}$, if

$$\forall Y \in \mathcal{D} \quad \exists \pi_Y \in \mathcal{M} : \mathbb{E}_{\pi_Y}(u \circ X) \geq \mathbb{E}_{\pi_Y}(u \circ Y)$$

Denote the choice functions that selects all maximal acts by $ch_{\max}$.

Non-Criterion-Based Choice Functions

Given a cardinal utility u and a credal set $\mathcal{M}$, there are also ways of constructing choice functions that do not rely on real-valued criteria:

Walley's Maximality (a.k.a. Maximality):

([Walley, 1991])

Call X **maximal** in $\mathcal{D} \subseteq \mathcal{G}$ wrt $\mathcal{M}$, if

$$\forall Y \in \mathcal{D} \quad \exists \pi_Y \in \mathcal{M} : \mathbb{E}_{\pi_Y}(u \circ X) \geq \mathbb{E}_{\pi_Y}(u \circ Y)$$

Denote the choice functions that selects all maximal acts by $ch_{\max}$.

Levi's E-Admissibility (a.k.a. E-Admissibility):

([Levi, 1974])

Call X **E-admissible** in $\mathcal{D} \subseteq \mathcal{G}$ wrt $\mathcal{M}$, if

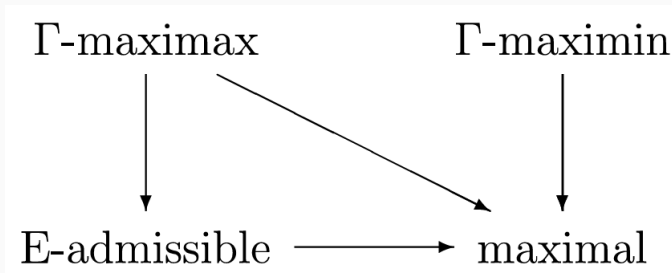
$$\exists \pi \quad \forall Y \in \mathcal{D} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$$

Denote the choice functions that selects all E-admissible acts by ch_{ead} .

Connections Between IP Choice Functions

A very nice overview on IP Decision Making is given in [\[Troffaes, 2007\]](#).

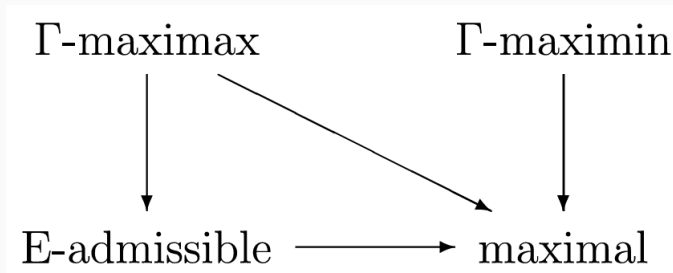
The following diagram is taken from there:



Connections Between IP Choice Functions

A very nice overview on IP Decision Making is given in [Troffaes, 2007].

The following diagram is taken from there:



Quick Question.

Why does being E-admissible always imply being maximal?

Running Example (Pseudo-Label Selection): Maximality

- Reading: *drop* a candidate instance only if some competitor has a higher expected likelihood under *every* $\pi \in \mathcal{M}$.
- The witnessing π_y may vary with the competitor $\Rightarrow \text{ch}_{\max}$ is induced by pairwise comparisons ($\text{ch}_{\succsim} \equiv \text{ch}_{\max}$ for $x \succsim x' :\Leftrightarrow \forall \pi \in \mathcal{M} : \mathbb{E}_{\pi}(u \circ x) \geq \mathbb{E}_{\pi}(u \circ x')$).
- Set-valued output $\Rightarrow$ batch-wise self-training: add all maximal instances in one iteration, tie-break by a secondary criterion, or shrink $\mathcal{M}$.

Running Example (Pseudo-Label Selection): E-Admissibility

- Reading: add every instance that would be the *Bayes-optimal* pseudo-label for at least one prior you are willing to entertain.
- Here one π must work against *all* competitors at once $\Rightarrow$ $\text{ch}_{\text{ead}} \subseteq \text{ch}_{\text{max}}$, and ch_{ead} is *not* induced by pairwise comparisons.
- A Γ -Maximin choice need *not* be E-admissible: an instance that hedges across S can be optimal under no single $\pi \in \mathcal{M}$.
- Checking a candidate: one LP; relax to E_ϵ -admissibility for stability.

DIY: Something to Not Fall Asleep

Task. Work on the following questions:

- Prove that $ch_{\max}$ is induced by pairwise comparisons.
- Prove that ch_{ead} is not induced by pairwise comparisons.

Hint: *For the latter, verify that the utility table*

	s_1	s_2
X_1	10	0
X_2	0	10
X_3	4	4

together with the credal set

$$\mathcal{M} = \{\pi : 0.3 \leq \pi(s_1) \leq 0.8\}$$

provides a counter-example.

(Example and results adapted from [Schervish et al., 2003].)

E-Admissibility versus Maximality

Solution.

- Define a preorder $\succsim$ among act by setting:

$$X \succsim Y \quad :\Leftrightarrow \quad \forall \pi \in \mathcal{M} : \mathbb{E}_\pi(u \circ X) \geq \mathbb{E}_\pi(u \circ Y)$$

It is then straightforward to check that $ch_{\succsim} \equiv ch_{\max}$.

- Assume (for contradiction) there is a preorder $\succsim$ that induces ch_{ead} .

It is easy to verify that all three following must hold:

- (I) $ch_{\text{ead}}(\{X_1, X_3\}) = \{X_1, X_3\}$, i.e., $X_1 \not\succ X_3$
- (II) $ch_{\text{ead}}(\{X_2, X_3\}) = \{X_2, X_3\}$, i.e., $X_2 \not\succ X_3$
- (III) $ch_{\text{ead}}(\{X_1, X_2, X_3\}) = \{X_1, X_2\}$, i.e., $(X_1 \succ X_3 \vee X_2 \succ X_3)$

A contradiction, hence ch_{ead} is not induced by pairwise comparisons.

Computation

Computation: Checking for E-Admissibility

In the case of a finite cardinal decision problem $(\mathcal{G}, S, u)$, checking whether a given act is E -admissible can be done by the following LP:

$$\max_{\gamma} \left\{ \begin{array}{l} \sum_{j=1}^m \gamma_j \leq 1 \\ \underline{b}_s \leq \sum_{j=1}^m f_s(s_j) \cdot \gamma_j \leq \bar{b}_s \text{ for all } s \\ \sum_{j=1}^m (u_{ij} - u_{zj}) \cdot \gamma_j \leq 0 \text{ for all } i \end{array} \right\}$$

Then $X_z \in \mathcal{G} = \{X_1, \dots, X_n\}$ is E -admissible iff the optimal outcome is 1.

(Compare [Utkin and Augustin, 2005].)

Computation: Checking for Maximality

In the case of a finite cardinal decision problem $(\mathcal{G}, S, u)$, checking whether a given act is maximal can be done by the following LP:

$$\max_{\gamma} \left\{ \sum_{i=1}^n \left(\sum_{j=1}^m \gamma_{ij} \right) : \begin{array}{l} \sum_{j=1}^m \gamma_{ij} \leq 1 \text{ for all } i \\ \underline{b}_s \leq \sum_{j=1}^m f_s(s_j) \cdot \gamma_{ij} \leq \bar{b}_s \text{ for all } s, i \\ \sum_{j=1}^m (u_{ij} - u_{zj}) \cdot \gamma_{ij} \leq 0 \text{ for all } i \end{array} \right\}$$

Then $X_z \in \mathcal{G} = \{X_1, \dots, X_n\}$ is maximal iff the optimal outcome is n .

(Compare [Jansen et al., 2017].)

A choice function trades off between ch_{ead} and ch_{max} can be obtained as follows:

E_ε -Admissibility

A choice function trades off between ch_{eas} and ch_{max} can be obtained as follows:

E_ε -Admissibility:

Let $\varepsilon \geq 0$. Call X is called E_ε -admissible in $\mathcal{D} \subseteq \mathcal{G}$ wrt $\mathcal{M}$, if there exists a family of probability measures $(\pi_Y)_{Y \in \mathcal{G}}$ from $\mathcal{M}$ such that:

- i) $\mathbb{E}_{\pi_Y}(u \circ X) \geq \mathbb{E}_{\pi_Y}(u \circ Y)$ for all $Y \in \mathcal{D}$ and
- ii) $\| \pi_X - \pi_Y \| \leq \varepsilon$ for all $X, Y \in \mathcal{G}$, where $\| \cdot \|$ denotes a norm on Δ_S .

Denote the choice function that selects all E_ε -admissible acts by ch_ε .

E_ε -Admissibility

A choice function trades off between $ch_{\text{e ad}}$ and ch_{max} can be obtained as follows:

E_ε -Admissibility:

Let $\varepsilon \geq 0$. Call X is called E_ε -admissible in $\mathcal{D} \subseteq \mathcal{G}$ wrt $\mathcal{M}$, if there exists a family of probability measures $(\pi_Y)_{Y \in \mathcal{G}}$ from $\mathcal{M}$ such that:

- i) $\mathbb{E}_{\pi_Y}(u \circ X) \geq \mathbb{E}_{\pi_Y}(u \circ Y)$ for all $Y \in \mathcal{D}$ and
- ii) $\| \pi_X - \pi_Y \| \leq \varepsilon$ for all $X, Y \in \mathcal{G}$, where $\| \cdot \|$ denotes a norm on Δ_S .

Denote the choice function that selects all E_ε -admissible acts by ch_ε .

Intuition: The choice function ch_ε selects those acts that are maximal AND not too far from being also E-admissible.

Computation: Checking for E_ε -Admissibility

In the case of a finite cardinal decision problem $(\mathcal{G}, S, u)$, checking whether a given act is E_ε -admissible (wrt L_1) can be done by the following LP:

$$\max_{\gamma} \left\{ \sum_{i=1}^n \left(\sum_{j=1}^m \gamma_{ij} \right) : \begin{array}{l} \sum_{j=1}^m \gamma_{ij} \leq 1 \text{ for all } i \\ \underline{b}_s \leq \sum_{j=1}^m f_s(s_j) \cdot \gamma_{ij} \leq \bar{b}_s \text{ for all } s, i \\ \sum_{j=1}^m (u_{ij} - u_{zj}) \cdot \gamma_{ij} \leq 0 \text{ for all } i \\ \sum_{j=1}^m |\gamma_{i_1 j} - \gamma_{i_2 j}| \leq \varepsilon \text{ for all } i_1 > i_2 \end{array} \right\}$$

Then $X_z \in \mathcal{G} = \{X_1, \dots, X_n\}$ is E_ε -admissible iff the optimal outcome is n .

(Compare [Jansen et al., 2022].)

Stability Revisited

Stability of E-admissible acts

Informally, we are interested in the question:

“How stable is a given act X in its property of being E -admissible?”

Stability of E-admissible acts

Informally, we are interested in the question:

“How stable is a given act X in its property of being E -admissible?”

To approach this question, we introduce the following concept:

Supporter set

Let $X \in \mathcal{G}$ be E -admissible. We define the *supporter set* of X as

$$\mathcal{M}_X = \left\{ \pi \in \mathcal{M} : X \in \text{ch}_{u,\pi}(\mathcal{G}) \right\}$$

i.e. as set of all compatible probabilities w.r.t. which X maximizes EU.

Stability of E-admissible acts

Informally, we are interested in the question:

“How stable is a given act X in its property of being E -admissible?”

To approach this question, we introduce the following concept:

Supporter set

Let $X \in \mathcal{G}$ be E -admissible. We define the *supporter set* of X as

$$\mathcal{M}_X = \left\{ \pi \in \mathcal{M} : X \in \text{ch}_{u,\pi}(\mathcal{G}) \right\}$$

i.e. as set of all compatible probabilities w.r.t. which X maximizes EU.

Formal version of the question:

“How to define sensible measures for characterizing $\mathcal{M}_X$?”

The maximal extent

One possible measure is the following:

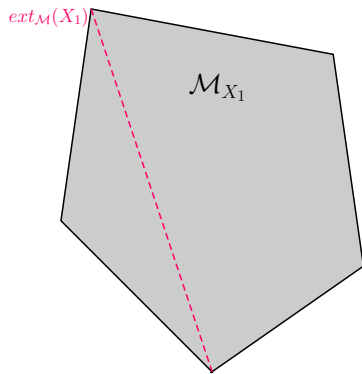
Maximal extent

Let $\| \cdot \|$ denote some norm on $\mathcal{M}$. For an E-admissible act $X \in \mathcal{G}$ we call the expression

$$\text{ext}_{\mathcal{M}}(X) = \sup_{\pi, \pi' \in \mathcal{M}_X} \| \pi - \pi' \|$$

the *maximal extent* of X with respect to $\| \cdot \|$.

Maximal extent of an E -admissible act X_1



The maximal extent

One possible measure is the following:

Maximal extent

Let $\| \cdot \|$ denote some norm on $\mathcal{M}$. For an E-admissible act $X \in \mathcal{G}$ we call the expression

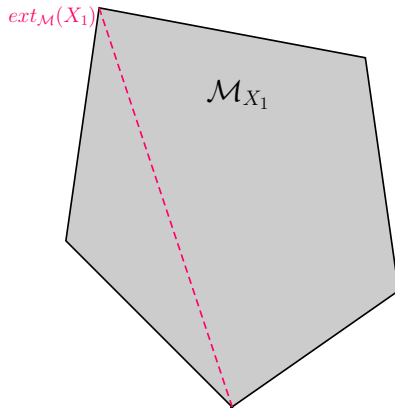
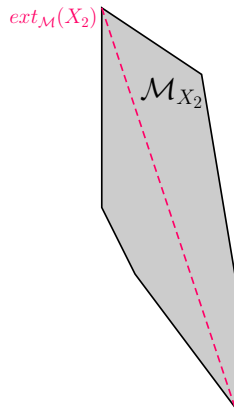
$$\text{ext}_{\mathcal{M}}(X) = \sup_{\pi, \pi' \in \mathcal{M}_X} \| \pi - \pi' \|$$

the *maximal extent* of X with respect to $\| \cdot \|$.

Shortcoming: The measure $\text{ext}_{\mathcal{M}}(X)$ describes the set $\mathcal{M}_X$ solely by means of its **most extreme** direction. This might be misleading in certain situations!

The maximal extent

Maximal extents of two different acts X_1 and X_2



The uniform extent

A measure avoiding this shortcoming is given by:

Uniform extent

For an E-admissible act $X \in \mathcal{G}$, the *uniform extent* is the expression

$$\text{uxt}_{\mathcal{M}}(X) := \max \left\{ \varepsilon : \exists \pi \in \mathcal{M}_X \text{ s.t. } C_\varepsilon(\pi) \subseteq \mathcal{M}_X \right\}$$

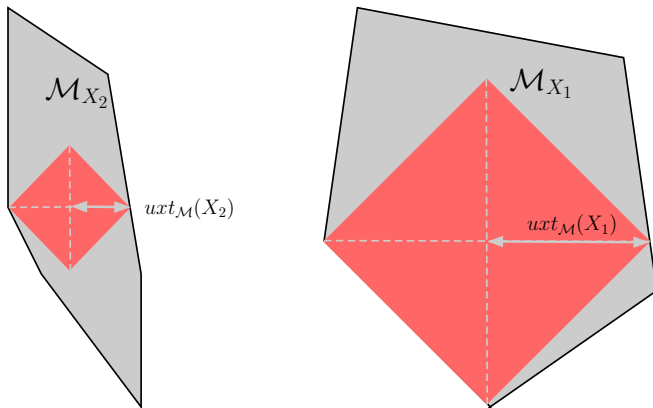
where $C_\varepsilon(\pi) := \text{conv}(\{\pi_\varepsilon^{1+}, \pi_\varepsilon^{1-}, \pi_\varepsilon^{2+}, \pi_\varepsilon^{2-}, \dots, \pi_\varepsilon^{m+}, \pi_\varepsilon^{m-}\})$ with

$$\pi_\varepsilon^{j*+}(\{s_j\}) = \begin{cases} \pi(\{s_j\}) + \varepsilon & \text{if } j = j^* \\ \pi(\{s_j\}) - \frac{\varepsilon}{m-1} & \text{if } j \neq j^* \end{cases}$$

$$\pi_\varepsilon^{j*-}(\{s_j\}) = \begin{cases} \pi(\{s_j\}) - \varepsilon & \text{if } j = j^* \\ \pi(\{s_j\}) + \frac{\varepsilon}{m-1} & \text{if } j \neq j^* \end{cases}$$

The uniform extent

Uniform extents of two different acts X_1 and X_2



Computation: Uniform extent

The uniform extent of an act X can be computed by the following linear optimization problem:

Proposition

Let X_z be E -admissible in $\mathcal{G}$. Consider the linear programming problem

$$\varepsilon \longrightarrow \max_{(\gamma_1, \dots, \gamma_m, \varepsilon)}$$

with constraints $(\gamma_1, \dots, \gamma_m, \varepsilon) \geq 0$ and

(1) $\sum_{j=1}^m \gamma_j = 1$

(2) $\gamma_j \geq \varepsilon$ for all $j = 1, \dots, m$

Computation: Uniform extent

The uniform extent of an act X can be computed by the following linear optimization problem:

Proposition

Let X_z be E -admissible in $\mathcal{G}$. Consider the linear programming problem

$$\varepsilon \longrightarrow \max_{(\gamma_1, \dots, \gamma_m, \varepsilon)}$$

with constraints $(\gamma_1, \dots, \gamma_m, \varepsilon) \geq 0$, (1), (2) and

$$(3) \quad \underline{b}_\ell \leq \sum_{j=1}^m f_\ell(s_j) \cdot \gamma_j + \varepsilon \cdot d(j^*, \ell) \leq \bar{b}_\ell \quad \text{for all } \ell = 1, \dots, r, j^* = 1, \dots, m$$

$$(4) \quad \underline{b}_\ell \leq \sum_{j=1}^m f_\ell(s_j) \cdot \gamma_j - \varepsilon \cdot d(j^*, \ell) \leq \bar{b}_\ell \quad \text{for all } \ell = 1, \dots, r, j^* = 1, \dots, m$$

where

$$d(j^*, \ell) = f_\ell(s_{j^*}) - \frac{1}{(m-1)} \sum_{j \neq j^*} f_\ell(s_j)$$

Computation: Uniform extent

The uniform extent of an act X can be computed by the following linear optimization problem:

Proposition

Let X_z be E -admissible in $\mathcal{G}$. Consider the linear programming problem

$$\varepsilon \longrightarrow \max_{(\gamma_1, \dots, \gamma_m, \varepsilon)}$$

with constraints $(\gamma_1, \dots, \gamma_m, \varepsilon) \geq 0$, (1), (2), (3), (4) and

$$(5) \quad \sum_{j=1}^m (u_{ij} - u_{zj}) \cdot \gamma_j + \varepsilon \cdot c(j^*, i) \leq 0 \quad \text{for } i = 1, \dots, k, j^* = 1, \dots, m$$

$$(6) \quad \sum_{j=1}^m (u_{ij} - u_{zj}) \cdot \gamma_j - \varepsilon \cdot c(j^*, i) \leq 0 \quad \text{for } i = 1, \dots, k, j^* = 1, \dots, m$$

where

$$c(j^*, i) = (u_{ij^*} - u_{zj^*}) - \frac{1}{(m-1)} \sum_{j \neq j^*} (u_{ij} - u_{zj})$$

Computation: Uniform extent

The uniform extent of an act X can be computed by the following linear optimization problem:

Proposition

Let X_z be E -admissible in $\mathcal{G}$. Consider the linear programming problem

$$\varepsilon \longrightarrow \max_{(\gamma_1, \dots, \gamma_m, \varepsilon)}$$

with constraints $(\gamma_1, \dots, \gamma_m, \varepsilon) \geq 0$, (1), (2), (3), (4), (5) and (6).

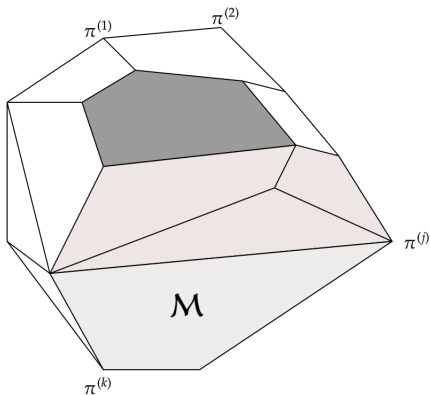
Then the uniform extent $\text{uxt}_{\mathcal{M}}(X_z)$ of X_z is given by the optimal value.

Decision Making Under Weakly Structured Information

Decision Theory under Weakly Structured Information

$\mathcal{I}_U$ Information on the generation of states is characterized by credal set of probability measure on (S, Σ_S) .

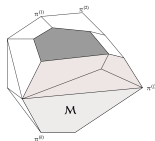
Weakly Structured Information for Decision Making



Weakly Structured Information for Decision Making

Decision Theory under Weakly Structured Information

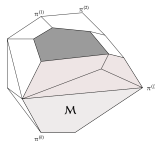
$\mathcal{I}_U$ Information on the generation of states is characterized by **credal set** of probability measure on (S, Σ_S) .



Weakly Structured Information for Decision Making

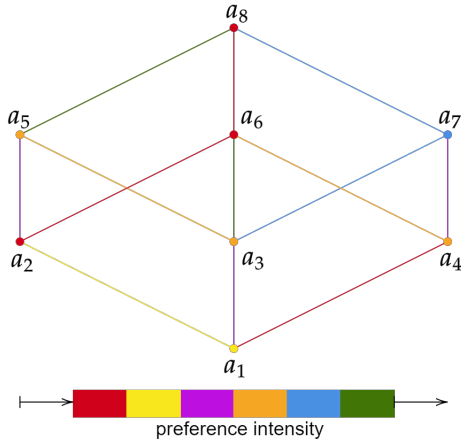
Decision Theory under Weakly Structured Information

$\mathcal{I}_U$ Information on the generation of states is characterized by **credal set** of probability measure on (S, Σ_S) .



$\mathcal{I}_P$ Information on preferences is characterized by a **set $\mathcal{U}$ of cardinal utility functions** $u : C \rightarrow \mathbb{R}$ on the consequences.

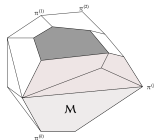
Weakly Structured Information for Decision Making



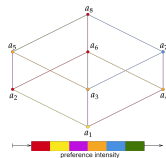
Weakly Structured Information for Decision Making

Decision Theory under Weakly Structured Information

$\mathcal{I}_U$ Information on the generation of states is characterized by **credal set** of probability measure on (S, Σ_S) .



$\mathcal{I}_P$ Information on preferences is characterized by a **set $\mathcal{U}$ of cardinal utility functions** $u : C \rightarrow \mathbb{R}$ on the consequences.



Notation: Binary relation R has **strict part** P_R and **indifference part** I_R .

Preference system & Consistency

Let A denote a set of consequences. Let further

- $R_1 \subseteq C \times C$ be a binary relation on C
- $R_2 \subseteq R_1 \times R_1$ be a binary relation on R_1

The triplet $\mathcal{A} = [C, R_1, R_2]$ is called a **preference system** on C .

We call $\mathcal{A}$ **consistent** if there is $u : C \rightarrow [0, 1]$ with for all $a, b, c, d \in C$:

$$(a, b) \in R_1 \Rightarrow u(a) \geq u(b) \quad (\text{with } = \text{ iff } \in I_{R_1}).$$

$$((a, b), (c, d)) \in R_2 \Rightarrow u(a) - u(b) \geq u(c) - u(d) \quad (\text{with } = \text{ iff } \in I_{R_2}).$$

The set of all representations u of $\mathcal{A}$ is denoted by $\mathcal{U}$.

Interpretation of the components of $\mathcal{A}$:

- $(a, b) \in R_1$: *"a is at least as desirable as b"*
- $((a, b), (c, d)) \in R_2$: *"exchanging b by a is at least as desirable as d by c"*

Modelling $\mathcal{U}$: Preference Systems

Notation: Binary relation R has **strict part** P_R and **indifference part** I_R .

Preference system & Consistency

Let A denote a set of consequences. Let further

$R_1 \subseteq C \times C$ be a binary relation on C

$R_2 \subseteq R_1 \times R_1$ be a binary relation on R_1

The triplet $\mathcal{A} = [C, R_1, R_2]$ is called a **preference system** on C .

We call $\mathcal{A}$ **consistent** if there is $u : C \rightarrow [0, 1]$ with for all $a, b, c, d \in C$:

- $(a, b) \in R_1 \Rightarrow u(a) \geq u(b)$ (with $=$ iff $\in I_{R_1}$).
- $((a, b), (c, d)) \in R_2 \Rightarrow u(a) - u(b) \geq u(c) - u(d)$ (with $=$ iff $\in I_{R_2}$).

The set of all representations u of $\mathcal{A}$ is denoted by $\mathcal{U}$.

(Compare [Jansen et al., 2018].)

Normalization & Regularization

Let $\mathcal{A} = [C, R_1, R_2]$ be consistent and assume there exist $a_*, a^* \in A$ such that $(a^*, a) \in R_1$ and $(a, a_*) \in R_1$ for all $a \in A$. Then

$$\mathcal{N}_{\mathcal{A}} := \left\{ u \in \mathcal{U}_{\mathcal{A}} : u(a_*) = 0 \wedge u(a^*) = 1 \right\}$$

is called the **normalized representation set** of $\mathcal{A}$.

Further, for $\delta \in [0, 1)$, we denote by $\mathcal{N}_{\mathcal{A}}^{\delta}$ the set of all $u \in \mathcal{N}_{\mathcal{A}}$ satisfying

$$u(a) - u(b) \geq \delta \quad \wedge \quad u(c) - u(d) - u(e) + u(f) \geq \delta$$

for all $(a, b) \in P_{R_1}$ and for all $((c, d), (e, f)) \in P_{R_2}$.

We call $\mathcal{A}$ **δ -consistent** if $\mathcal{N}_{\mathcal{A}}^{\delta} \neq \emptyset$.

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

The set of potential benefits $\mathcal{X} = \{b_1, \dots, b_5\}$ is given by

b_1	b_2	b_3	b_4	b_5
overtime premium	child care	advanced training	promotion prospects	flexible hours

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

The set of potential benefits $\mathcal{X} = \{b_1, \dots, b_5\}$ is given by

b_1	b_2	b_3	b_4	b_5
overtime premium	child care	advanced training	promotion prospects	flexible hours

Then $C \subseteq \mathbb{R}^+ \times 2^{\mathcal{X}}$ and $\mathcal{A} = [C, R_1, R_2]$ is given by

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

The set of potential benefits $\mathcal{X} = \{b_1, \dots, b_5\}$ is given by

b_1	b_2	b_3	b_4	b_5
overtime premium	child care	advanced training	promotion prospects	flexible hours

Then $C \subseteq \mathbb{R}^+ \times 2^{\mathcal{X}}$ and $\mathcal{A} = [C, R_1, R_2]$ is given by

$$R_1 = \left\{ ((y_1, B_1), (y_2, B_2)) : y_1 \geq y_2 \wedge B_1 \supseteq B_2 \right\}$$

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

The set of potential benefits $\mathcal{X} = \{b_1, \dots, b_5\}$ is given by

b_1	b_2	b_3	b_4	b_5
overtime premium	child care	advanced training	promotion prospects	flexible hours

Then $C \subseteq \mathbb{R}^+ \times 2^{\mathcal{X}}$ and $\mathcal{A} = [C, R_1, R_2]$ is given by

$$R_1 = \left\{ ((y_1, B_1), (y_2, B_2)) : y_1 \geq y_2 \wedge B_1 \supseteq B_2 \right\}$$

$$R_2 = \left\{ \left(((y_1, B_1), (y_2, B_2)), ((y_3, B_3), (y_4, B_4)) \right) : \begin{array}{l} y_1 - y_2 \geq y_3 - y_4 \wedge \\ B_1 \supseteq B_3 \supseteq B_4 \supseteq B_2 \end{array} \right\}$$

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

The set of potential benefits $\mathcal{X} = \{b_1, \dots, b_5\}$ is given by

b_1	b_2	b_3	b_4	b_5
overtime premium	child care	advanced training	promotion prospects	flexible hours

Then $C \subseteq \mathbb{R}^+ \times 2^{\mathcal{X}}$ and $\mathcal{A} = [C, R_1, R_2]$ is given by

$$R_1 = \left\{ ((y_1, B_1), (y_2, B_2)) : y_1 \geq y_2 \wedge B_1 \supseteq B_2 \right\}$$

$$R_2 = \left\{ \left(((y_1, B_1), (y_2, B_2)), ((y_3, B_3), (y_4, B_4)) \right) : \begin{array}{l} y_1 - y_2 \geq y_3 - y_4 \wedge \\ B_1 \supseteq B_3 \supseteq B_4 \supseteq B_2 \end{array} \right\}$$

" $(y_1, B_1) \leftarrow (y_2, B_2)$ is more desirable than $(y_3, B_3) \leftarrow (y_4, B_4)$ if it yields a higher increase of salary and a superset increase of benefits."

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

The set of potential benefits $\mathcal{X} = \{b_1, \dots, b_5\}$ is given by

b_1	b_2	b_3	b_4	b_5
overtime premium	child care	advanced training	promotion prospects	flexible hours

Then $C \subseteq \mathbb{R}^+ \times 2^{\mathcal{X}}$ and $\mathcal{A} = [C, R_1, R_2]$ is given by

$$R_1 = \left\{ ((y_1, B_1), (y_2, B_2)) : y_1 \geq y_2 \wedge B_1 \supseteq B_2 \right\}$$

$$R_2 = \left\{ \left(((y_1, B_1), (y_2, B_2)), ((y_3, B_3), (y_4, B_4)) \right) : \begin{array}{l} y_1 - y_2 \geq y_3 - y_4 \wedge \\ B_1 \supseteq B_3 \supseteq B_4 \supseteq B_2 \end{array} \right\}$$

*" $(y_1, B_1) \leftarrow (y_2, B_2)$ is more desirable than $(y_3, B_3) \leftarrow (y_4, B_4)$ if it yields a **higher increase of salary** and a **superset increase of benefits**."*

Preference System: Toy Example

Suppose, an agent is looking for a new job.

Jobs are compared w.r.t. salary after tax and additional benefits.

The set of potential benefits $\mathcal{X} = \{b_1, \dots, b_5\}$ is given by

b_1	b_2	b_3	b_4	b_5
overtime premium	child care	advanced training	promotion prospects	flexible hours

Then $C \subseteq \mathbb{R}^+ \times 2^{\mathcal{X}}$ and $\mathcal{A} = [C, R_1, R_2]$ is given by

$$R_1 = \left\{ ((y_1, B_1), (y_2, B_2)) : y_1 \geq y_2 \wedge B_1 \supseteq B_2 \right\}$$

$$R_2 = \left\{ \left(((y_1, B_1), (y_2, B_2)), ((y_3, B_3), (y_4, B_4)) \right) : \begin{array}{l} y_1 - y_2 \geq y_3 - y_4 \wedge \\ B_1 \supseteq B_3 \supseteq B_4 \supseteq B_2 \end{array} \right\}$$

*" $(y_1, B_1) \leftarrow (y_2, B_2)$ is more desirable than $(y_3, B_3) \leftarrow (y_4, B_4)$ if it yields a higher increase of salary and a **superset increase of benefits**."*

Checking consistency

For a finite preference system $\mathcal{A} = [C, R_1, R_2]$, we can check for consistency by solving the following LP:

Checking consistency

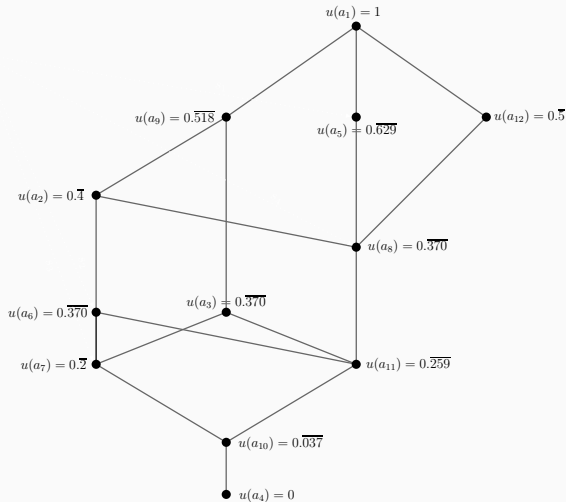
Let $C = \{c_1, \dots, c_k\}$ and consider the LP:

$$\varepsilon^* = \min_{\vec{u}, \varepsilon} \left\{ \begin{array}{l} 0 \leq (u_1, \dots, u_k, \varepsilon) \leq 1 \\ u_p = u_q \text{ for all } (c_p, c_q) \in I_{R_1} \\ u_q + \varepsilon \leq u_p \text{ for all } (c_p, c_q) \in P_{R_1} \\ u_p - u_q = u_r - u_s \text{ for all } ((c_p, c_q), (c_r, c_s)) \in I_{R_2} \\ u_r - u_s + \varepsilon \leq u_p - u_q \text{ for all } ((c_p, c_q), (c_r, c_s)) \in P_{R_2} \end{array} \right\}$$

Then $\mathcal{A}$ is consistent iff the optimal outcome is strictly positive.

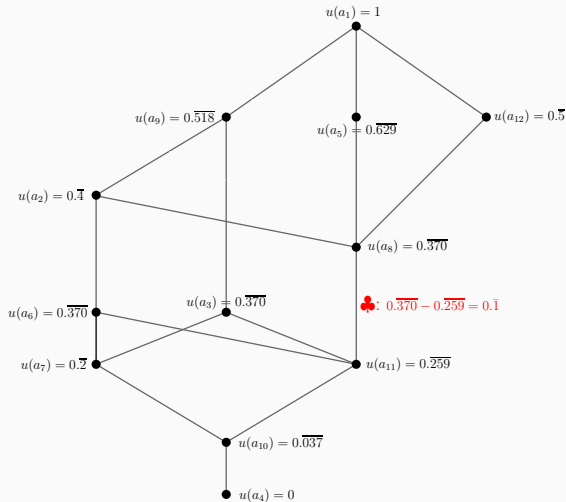
Example, continued

In our example the optimal outcome is > 0 . Thus, the preference system is consistent and **one possible** representation is given by



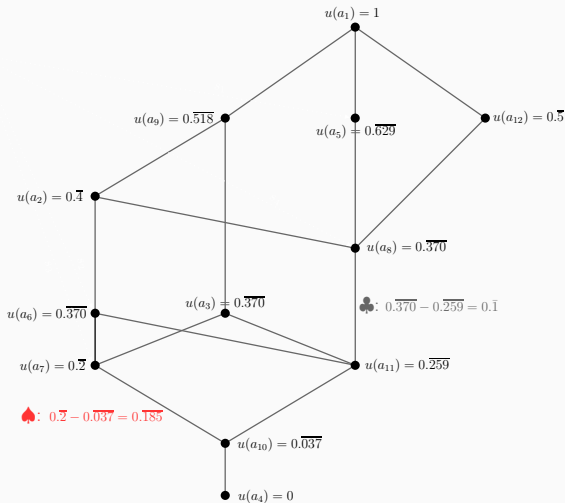
Example, continued

In our example the optimal outcome is > 0 . Thus, the preference system is consistent and **one possible** representation is given by



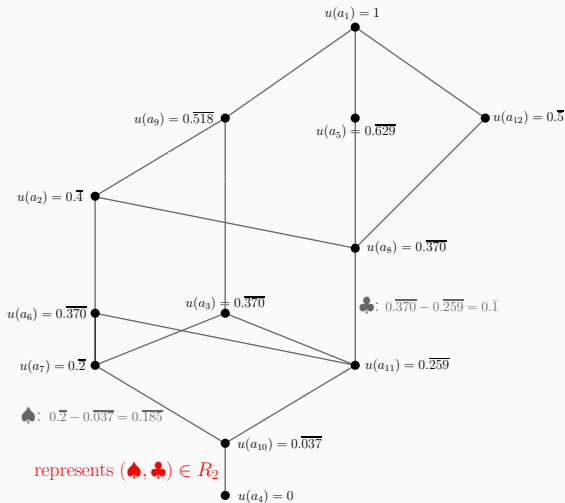
Example, continued

In our example the optimal outcome is > 0 . Thus, the preference system is consistent and **one possible** representation is given by



Example, continued

In our example the optimal outcome is > 0 . Thus, the preference system is consistent and **one possible** representation is given by



Criterion-Based Choice Functions in DT under WSI

Given a consistent preference system $\mathcal{A}$ and a credal set $\mathcal{M}$, the following criteria are commonly for constructing choice functions:

Criterion-Based Choice Functions in DT under WSI

Given a consistent preference system $\mathcal{A}$ and a credal set $\mathcal{M}$, the following criteria are commonly for constructing choice functions:

δ -Maximin criterion: Choose $\delta \in (0, 1)$ and set

$$\Phi_{\delta\text{-Maximin}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \underline{\mathbb{E}}_{\delta}(X) := \inf_{(u, \pi) \in \mathcal{N}_{\mathcal{A}}^{\delta} \times \mathcal{M}} \mathbb{E}_{\pi}(u \circ X)$$

Criterion-Based Choice Functions in DT under WSI

Given a consistent preference system $\mathcal{A}$ and a credal set $\mathcal{M}$, the following criteria are commonly for constructing choice functions:

δ -Maximin criterion: Choose $\delta \in (0, 1)$ and set

$$\Phi_{\delta\text{-Maximin}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \underline{E}_{\delta}(X) := \inf_{(u, \pi) \in \mathcal{N}_{\mathcal{A}}^{\delta} \times \mathcal{M}} \mathbb{E}_{\pi}(u \circ X)$$

δ -Maximax criterion: Choose $\delta \in (0, 1)$ and set

$$\Phi_{\delta\text{-Maximax}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \overline{E}_{\delta}(X) := \sup_{(u, \pi) \in \mathcal{N}_{\mathcal{A}}^{\delta} \times \mathcal{M}} \mathbb{E}_{\pi}(u \circ X)$$

Criterion-Based Choice Functions in DT under WSI

Given a consistent preference system $\mathcal{A}$ and a credal set $\mathcal{M}$, the following criteria are commonly for constructing choice functions:

δ -Maximin criterion: Choose $\delta \in (0, 1)$ and set

$$\Phi_{\delta\text{-Maximin}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \underline{E}_{\delta}(X) := \inf_{(u, \pi) \in \mathcal{N}_{\mathcal{A}}^{\delta} \times \mathcal{M}} \mathbb{E}_{\pi}(u \circ X)$$

δ -Maximax criterion: Choose $\delta \in (0, 1)$ and set

$$\Phi_{\delta\text{-Maximax}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \overline{E}_{\delta}(X) := \sup_{(u, \pi) \in \mathcal{N}_{\mathcal{A}}^{\delta} \times \mathcal{M}} \mathbb{E}_{\pi}(u \circ X)$$

Γ -Maximix criterion: Fix $\delta \in (0, 1)$ and $\eta \in [0, 1]$.

$$\Phi_{\delta\text{-Maximix}} : \mathcal{G} \rightarrow \mathbb{R} \quad , \quad X \mapsto \eta \cdot \underline{E}_{\delta}(u \circ X) + (1 - \eta) \cdot \overline{E}_{\delta}(u \circ X)$$

For $\eta = 0 \rightarrow \Phi_{\delta\text{-Maximin}}$; for $\eta = 1 \rightarrow \Phi_{\delta\text{-Maximax}}$.

Lower and Upper δ -Expectation

Let $C = \{c_1, \dots, c_k\}$ such that $(c_1, b), (b, c_n) \in R_1$ for all $b \in C$.

Let $\mathcal{E}(\mathcal{M}) := \{\pi^{(1)}, \dots, \pi^{(T)}\}$ denote the extreme points of $\mathcal{M}$, and let $X \in \mathcal{G}$.

For $\varepsilon^* \geq \delta > 0$, consider $LP_1^\delta, \dots, LP_T^\delta$ given by:

$$\min_{\vec{u}} \left\{ \sum_{i=1}^k u_i \pi^{(t)}(X^{-1}(c_i)) : \begin{array}{l} u_1 = 1, u_k = 0 \\ 0 \leq (u_1, \dots, u_n, \delta) \leq 1 \\ u_p = u_q, (c_p, c_q) \in I_{R_1} \\ u_q + \delta \leq u_p, (c_p, c_q) \in P_{R_1} \\ u_p - u_q = u_r - u_s, ((c_p, c_q), (c_r, c_s)) \in I_{R_2} \\ u_r - u_s + \delta \leq u_p - u_q, ((c_p, c_q), (c_r, c_s)) \in P_{R_2} \end{array} \right\}$$

Let $\underline{v}(t, \delta)$ the optimal outcomes of problem LP_t^δ .

Then, we have $\underline{\mathbb{E}}_\delta(X) = \min_t \underline{v}(t, \delta)$. (For $\overline{\mathbb{E}}_\delta(X)$ replace all min by max.)

Real world toy example, cont.

The agent specifies a set $S = \{s_1, \dots, s_4\}$ economic scenarios which might affect the offers in different ways.

This leads to the following decision problem $\mathcal{G} = \{J_1, J_2, J_3\}$:

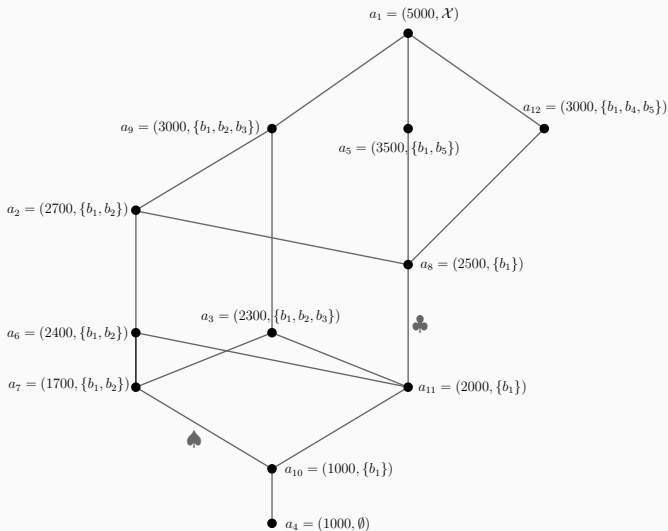
	s_1	s_2	s_3	s_4
J_1	$\underbrace{(5000, \mathcal{X})}_{=:a_1}$	$\underbrace{(2700, \{b_1, b_2\})}_{=:a_2}$	$\underbrace{(2300, \{b_1, b_2, b_3\})}_{=:a_3}$	$\underbrace{(1000, \emptyset)}_{=:a_4}$
J_2	$\underbrace{(3500, \{b_1, b_5\})}_{=:a_5}$	$\underbrace{(2400, \{b_1, b_2\})}_{=:a_6}$	$\underbrace{(1700, \{b_1, b_2\})}_{=:a_7}$	$\underbrace{(2500, \{b_1\})}_{=:a_8}$
J_3	$\underbrace{(3000, \{b_1, b_2, b_3\})}_{=:a_9}$	$\underbrace{(1000, \{b_1\})}_{=:a_{10}}$	$\underbrace{(2000, \{b_1\})}_{=:a_{11}}$	$\underbrace{(3000, \{b_1, b_4, b_5\})}_{=:a_{12}}$

The agent can rank the different scenarios by their probability to occur, i.e. specify the “ordinal” credal set

$$\mathcal{M} = \left\{ \pi \in \mathcal{P} : \pi(\{s_1\}) \geq \pi(\{s_2\}) \geq \pi(\{s_3\}) \geq \pi(\{s_4\}) \right\}$$

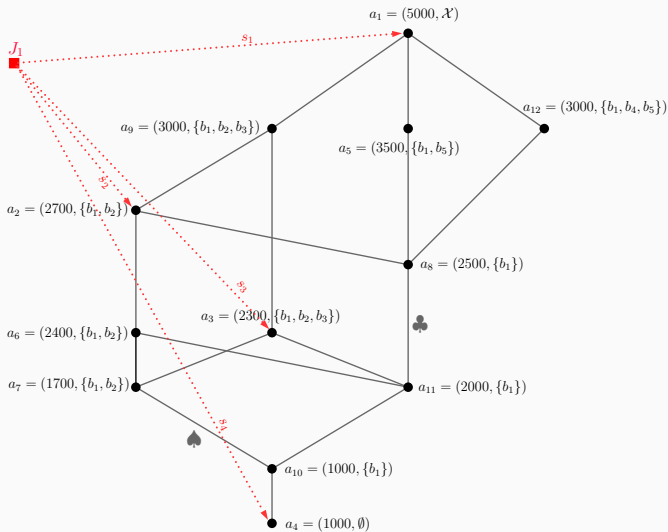
Hasse diagram of R_1

We can visualize the problem by means of a **Hasse diagram**:



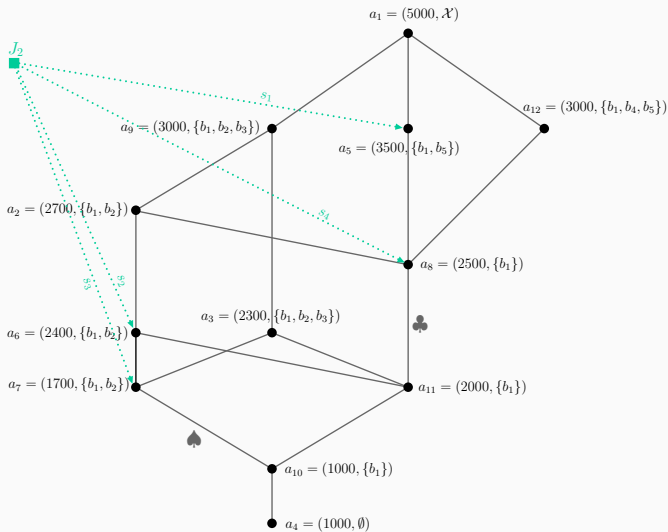
Hasse diagram of R_1

The first job offer looks as follows:



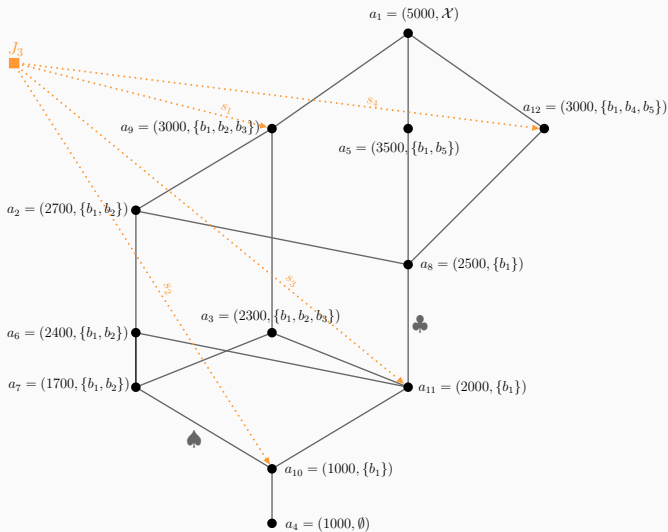
Hasse diagram of R_1

The second job offer looks as follows:



Hasse diagram of R_1

The third job offer looks as follows:

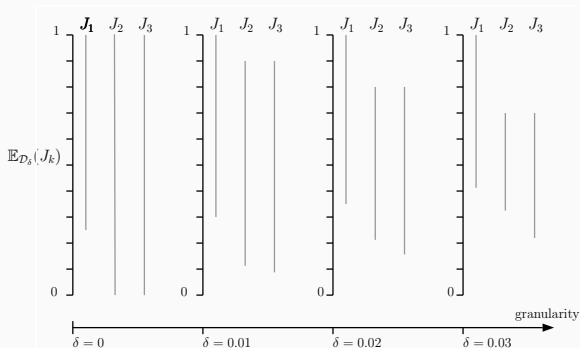


Real world toy example, continued

In our example the set of extreme points is given by

$$\mathcal{E}(\mathcal{M}) = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix} \right\}$$

Solving the linear optimization problem from before gives:



Choice Functions induced by Pairwise-Comparisons under WSI

Given a consistent preference system $\mathcal{A}$ and a credal set $\mathcal{M}$, the following binary relations can be used for constructing choice functions:

Choice Functions induced by Pairwise-Comparisons:

Define the following six binary relations:

$$(X, Y) \in R_{\exists\exists} \quad :\Leftrightarrow \quad \exists u \in \mathcal{U}_{\mathcal{A}} \exists \pi \in \mathcal{M} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$$

$$(X, Y) \in R_{\exists\forall}^1 \quad :\Leftrightarrow \quad \exists u \in \mathcal{U}_{\mathcal{A}} \forall \pi \in \mathcal{M} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$$

$$(X, Y) \in R_{\exists\forall}^2 \quad :\Leftrightarrow \quad \exists \pi \in \mathcal{M} \forall u \in \mathcal{U}_{\mathcal{A}} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$$

$$(X, Y) \in R_{\forall\exists}^1 \quad :\Leftrightarrow \quad \forall u \in \mathcal{U}_{\mathcal{A}} \exists \pi \in \mathcal{M} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$$

$$(X, Y) \in R_{\forall\exists}^2 \quad :\Leftrightarrow \quad \forall \pi \in \mathcal{M} \exists u \in \mathcal{U}_{\mathcal{A}} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$$

$$(X, Y) \in R_{\forall\forall} \quad :\Leftrightarrow \quad \forall \pi \in \mathcal{M} \forall u \in \mathcal{U}_{\mathcal{A}} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$$

For $R \in \{R_{\exists\exists}, R_{\exists\forall}^1, R_{\exists\forall}^2, R_{\forall\exists}^1, R_{\forall\exists}^2, R_{\forall\forall}\}$, we define a choice function for all $\mathcal{D} \subseteq \mathcal{G}$ by setting:

$$ch_R(\mathcal{D}) := \{X \in \mathcal{D} : \nexists Y \in \mathcal{G} \text{ s.t. } (Y, X) \in P_R\}$$

Characterizing $R_{\exists\exists}$ and $R_{\forall\forall}$

Let $C = \{c_1, \dots, c_k\}$, $\mathcal{E}(\mathcal{M}) := \{\pi^{(1)}, \dots, \pi^{(T)}\}$, and let $X, Y \in \mathcal{G}$.

Consider the LPs $LP_1, \dots, LP_T$ given by:

$$\min_{\vec{u}}, \max_{\vec{u}} \left\{ D_t(X, Y) : \begin{array}{l} 0 \leq (u_1, \dots, u_n) \leq 1 \\ u_p = u_q, (c_p, c_q) \in I_{R_1} \\ u_q \leq u_p, (c_p, c_q) \in P_{R_1} \\ u_p - u_q = u_r - u_s, ((c_p, c_q), (c_r, c_s)) \in I_{R_2} \\ u_r - u_s \leq u_p - u_q, ((c_p, c_q), (c_r, c_s)) \in P_{R_2} \end{array} \right\}$$

where $D_t(X, Y) := \sum_{i=1}^k u_i \cdot [\pi^{(t)}(X^{-1}(\{c_i\})) - \pi^{(t)}(Y^{-1}(\{c_i\}))]$.

For $t = 1, \dots, T$, denote by $\underline{v}(t)$ and $\bar{v}(t)$ the min and max of problem LO_t :

- i) $(X, Y) \in R_{\forall\forall}$ if and only if $\min_t \underline{v}(t) \geq 0$
- ii) $(X, Y) \in R_{\exists\exists}$ if $\max_t \bar{v}(t) > 0$

Real world toy example, continued

Let's solve these LPs in our job seeker example to find out how the different job offers J_1 , J_2 and J_3 relate to each other regarding $R_{\exists\exists}$ and $R_{\forall\forall}$. The results are summarized in the following table:

	(J_1, J_2)	(J_1, J_3)	(J_2, J_1)	(J_2, J_3)	(J_3, J_1)	(J_3, J_2)
$R_{\exists\exists}$	$\in$	$\in$	$\in$	$\in$	$\in$	$\in$
$R_{\forall\forall}$	$\notin$	$\notin$	$\notin$	$\notin$	$\notin$	$\notin$

Choice Functions induced by Global Comparisons

Given a consistent preference system $\mathcal{A}$ and a credal set $\mathcal{M}$, the following non-binary strategies can be used to construct a choice function under WSI:

Choice Functions induced by Global-Comparisons:

We define the following four admissibility concepts. An act $X \in \mathcal{G}$ is called:

- i) $\mathcal{A}|\mathcal{M}$ -**admissible** iff
$$\exists u \in \mathcal{U}_{\mathcal{A}} \exists \pi \in \mathcal{M} \forall Y \in \mathcal{G} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$$
- ii) $\mathcal{A}$ -**admissible** iff
$$\exists u \in \mathcal{U}_{\mathcal{A}} \forall \pi \in \mathcal{M} \forall Y \in \mathcal{G} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$$
- iii) $\mathcal{M}$ -**admissible** iff
$$\exists \pi \in \mathcal{M} \forall u \in \mathcal{U}_{\mathcal{A}} \forall Y \in \mathcal{G} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$$
- iv) $\mathcal{A}|\mathcal{M}$ -**dominant** iff
$$\forall u \in \mathcal{U}_{\mathcal{A}} \forall \pi \in \mathcal{M} \forall Y \in \mathcal{G} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$$

Denote by $ch_{\mathcal{A}|\mathcal{M}}$, $ch_{\mathcal{A}}$, $ch_{\mathcal{M}}$ and $ch_{\mathcal{A}|\mathcal{M}}^d$ the choice functions induced by these concepts, respectively.

Characterizing $R_{\exists\exists}$ and $R_{\forall\forall}$

Let $C = \{c_1, \dots, c_k\}$, $\mathcal{E}(\mathcal{M}) := \{\pi^{(1)}, \dots, \pi^{(T)}\}$, and let $X \in \mathcal{D} \subseteq \mathcal{G}$.

Consider the LP:

$$\varepsilon^* = \min_{\vec{u}, \varepsilon} \left\{ \varepsilon : \begin{array}{l} 0 \leq (u_1, \dots, u_k, \varepsilon) \leq 1 \\ u_p = u_q \text{ for all } (c_p, c_q) \in I_{R_1} \\ u_q + \varepsilon \leq u_p \text{ for all } (c_p, c_q) \in P_{R_1} \\ u_p - u_q = u_r - u_s \text{ for all } ((c_p, c_q), (c_r, c_s)) \in I_{R_2} \\ u_r - u_s + \varepsilon \leq u_p - u_q \text{ for all } ((c_p, c_q), (c_r, c_s)) \in P_{R_2} \\ D_t(X, Y) \geq 0 \text{ for all } Y \in \mathcal{D} \text{ and } t = 1, \dots, T \end{array} \right\}$$

Then X is $\mathcal{A}$ -admissible in $\mathcal{D}$ if and only if the optimal outcome of this optimization problem is strictly greater than 0.

Real world toy example, continued

We want to check which of the jobs J_1, J_2, J_3 are $\mathcal{A}$ -admissible. The LP for job J_1 for instance has ε as objective and the constraints:

$$(C_1) \quad u_1 \geq u_5$$

$$u_1 \geq u_9$$

$$(C_2) \quad \frac{1}{2}(u_1 + u_2) \geq \frac{1}{2}(u_5 + u_6)$$

$$\frac{1}{2}(u_1 + u_2) \geq \frac{1}{2}(u_9 + u_{10})$$

$$(C_3) \quad \frac{1}{3}(u_1 + u_2 + u_3) \geq \frac{1}{3}(u_5 + u_6 + u_7)$$

$$\frac{1}{3}(u_1 + u_2 + u_3) \geq \frac{1}{3}(u_9 + u_{10} + u_{11})$$

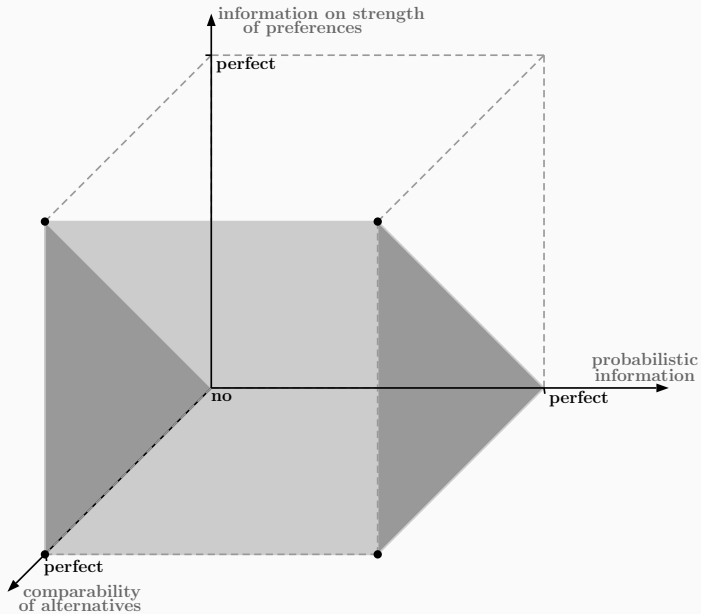
$$(C_4) \quad \frac{1}{4}(u_1 + u_2 + u_3 + u_4) \geq \frac{1}{4}(u_5 + u_6 + u_7 + u_8)$$

$$\frac{1}{4}(u_1 + u_2 + u_3 + u_4) \geq \frac{1}{4}(u_9 + u_{10} + u_{11} + u_{12})$$

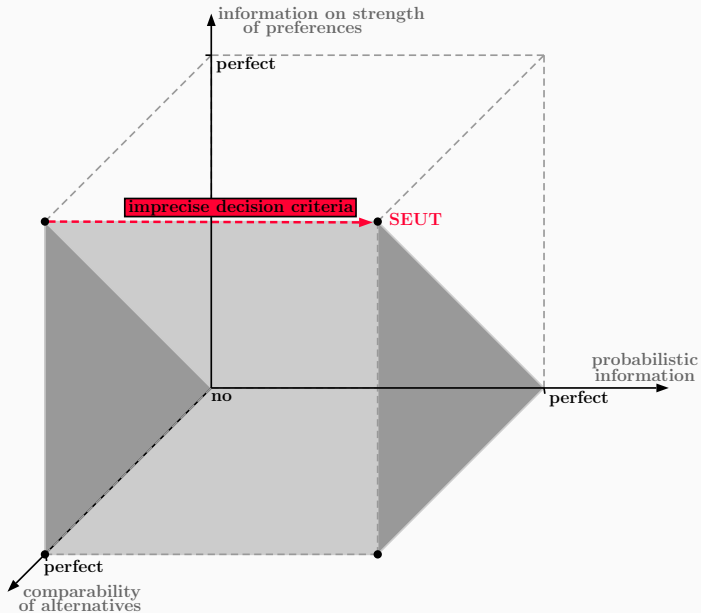
Solving the LP gives $0.\overline{037}$. Thus, job J_1 $\mathcal{A}$ -admissible.

In contrast, the LPs for jobs J_2 and J_3 , both give 0, indicating neither of them is $\mathcal{A}$ -admissible.

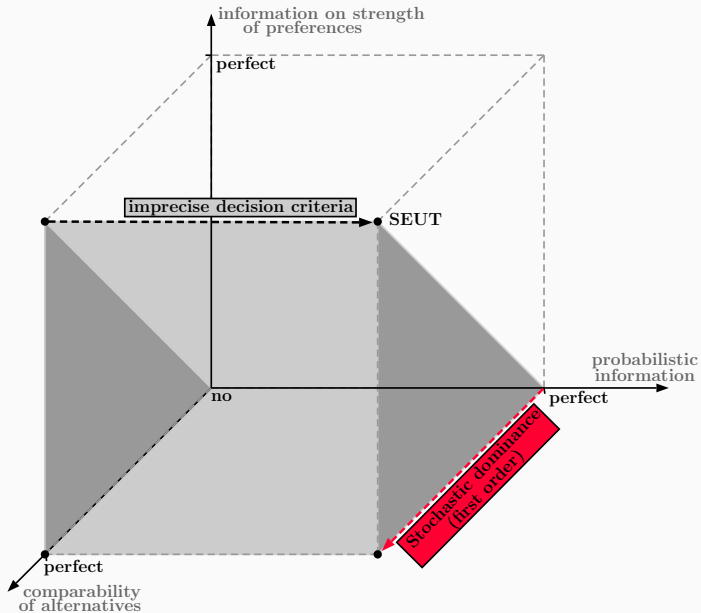
Coming Back



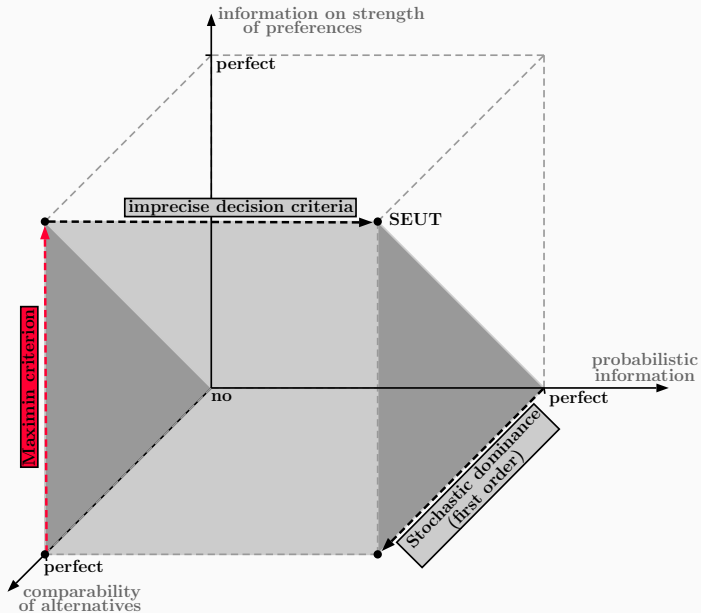
Coming Back



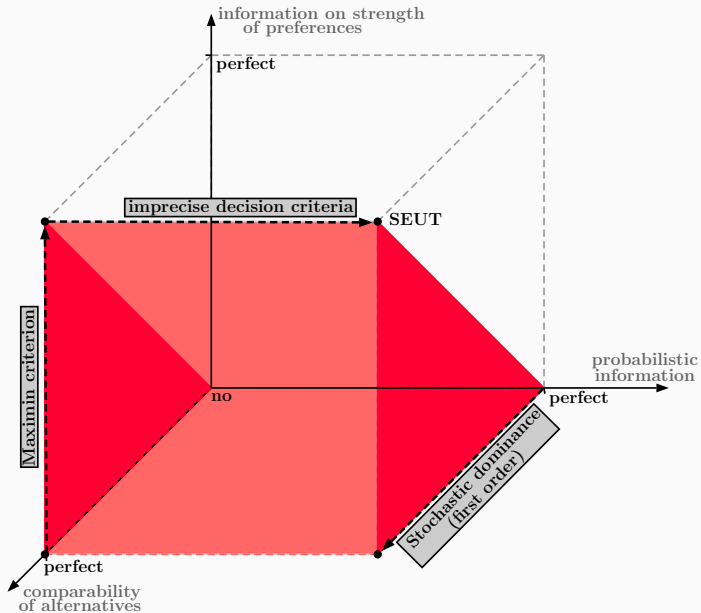
Coming Back



Coming Back



Coming Back



Weakly Structured Information in Statistics and Machine Learning

Observation I:

Statistical methods are usually tailored for data situations that can be clearly assigned to a **standard scale of measurement**.

Motivation

Observation I:

Statistical methods are usually tailored for data situations that can be clearly assigned to a **standard scale of measurement**.

Observation II:

Non-standard data can often not clearly be assigned to a standard scale.

Motivation

Observation I:

Statistical methods are usually tailored for data situations that can be clearly assigned to a **standard scale of measurement**.

Observation II:

Non-standard data can often not clearly be assigned to a standard scale.

Taken together:

Statistical methods are often not well-suited for non-standard data!

Motivation

Observation I:

Statistical methods are usually tailored for data situations that can be clearly assigned to a **standard scale of measurement**.

Observation II:

Non-standard data can often not clearly be assigned to a standard scale.

Taken together:

Statistical methods are often not well-suited for non-standard data!

Idea:

Use the notion of a preference system to model data with scales of measurement which not correspond to one of these extreme poles.

Further Reading

What follows is mainly based on these three papers:

- C. Jansen, G. Schollmeyer, J. Rodemann, H. Blocher, T. Augustin: [Statistical Multicriteria Benchmarking via the GSD-Front](#). *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.
- C. Jansen, G. Schollmeyer, H. Blocher, J. Rodemann, T. Augustin: [Robust statistical comparison of random variables with locally varying scale of measurement](#). *Uncertainty in Artificial Intelligence (UAI 2023)*.
- C. Jansen, M. Nalenz, G. Schollmeyer, T. Augustin: [Statistical Comparisons of Classifiers by Generalized Stochastic Dominance](#). *Journal of Machine Learning Research (JMLR 2022)*.

Standard Scales of Measurement

Consider some random variable $X : \Omega \rightarrow A$ mapping to some set A .

Standard Scales of Measurement

Consider some random variable $X : \Omega \rightarrow A$ mapping to some set A .

Ordinal Data:

If A is structured only by a **preorder** $\succsim$, we call X of **ordinal** scale.

Then: Any analysis of the variable X has to be **invariant** under $\succsim$ -monotone scale transformations $u : A \rightarrow \mathbb{R}$, as these encode the same information.

Standard Scales of Measurement

Consider some random variable $X : \Omega \rightarrow A$ mapping to some set A .

Ordinal Data:

If A is structured only by a preorder $\succsim$, we call X of ordinal scale.

Then: Any analysis of the variable X has to be invariant under $\succsim$ -monotone scale transformations $u : A \rightarrow \mathbb{R}$, as these encode the same information.

Cardinal Data:

If there exists a **natural scale** $u^* : A \rightarrow \mathbb{R}$, we call X of **cardinal** scale.

Then: Any analysis of the variable X can be based on u^* **alone**.

Preference Systems in Statistics

Question: What if X does not belong to either extreme pole?

In other words: What if the structuredness of A varies along its subsets?

Preference Systems in Statistics

Question: What if X does not belong to either extreme pole?

In other words: What if the structuredness of A varies along its subsets?

A **preference system** $\mathcal{A} = [A, R_1, R_2]$ helps to formalize:

- R_1 formalizes the available **ordinal information**.
- R_2 describes the available **cardinal information**.
- A is **locally almost cardinal** on subsets where R_1 and R_2 are dense.
- A is **locally at most ordinal** on subsets where R_2 is sparse.

Preference Systems in Statistics

Question: What if X does not belong to either extreme pole?

In other words: What if the structuredness of A varies along its subsets?

A **preference system** $\mathcal{A} = [A, R_1, R_2]$ helps to formalize:

- R_1 formalizes the available ordinal information.
- R_2 describes the available cardinal information.
- A is locally almost cardinal on subsets where R_1 and R_2 are dense.
- A is locally at most ordinal on subsets where R_2 is sparse.

In general, we then have

$$\{u^*\} \subsetneq \mathcal{U}_{\mathcal{A}} \subsetneq \mathcal{U}_{R_1}$$

where $\mathcal{U}_{R_1}$ is the set of all scales compatible with the ordinal information only.

Comparing PS-Valued Variables

Idea: Weaken the expectation order to a **preorder** by demanding expectation dominance for all scales u compatible with the preference system $\mathcal{A}$.

Comparing PS-Valued Variables

Idea: Weaken the expectation order to a **preorder** by demanding expectation dominance for all scales u compatible with the preference system $\mathcal{A}$.

This idea leads to the following precise version of $R_{\forall\forall}$ from before. (Concretely, by precise, we mean we have $\mathcal{M} = \{\pi\}$ in the definition of $R_{\forall\forall}$.)

Generalized Stochastic Dominance (GSD)

Define the following preorder among ps-valued variables $X, Y : \Omega \rightarrow A$:

$$X \succsim_{\text{GSD}} Y \quad :\Leftrightarrow \quad \forall u \in \mathcal{U}_{\mathcal{A}} : \mathbb{E}_{\pi}(u \circ X) \geq \mathbb{E}_{\pi}(u \circ Y)$$

The relation $\succsim_{\text{GSD}}$ is called **generalized stochastic dominance (GSD)**.

Important Special Case: Mixed-Scaled Dimensions

Situation: Consider an r -dimensional space $A \subseteq \mathbb{R}^r$ and assume that

- the first $0 \leq z \leq r$ dimensions are of **cardinal scale** and
- the remaining dimensions are of purely **ordinal scale**.

Important Special Case: Mixed-Scaled Dimensions

Situation: Consider an r -dimensional space $A \subseteq \mathbb{R}^r$ and assume that

- the first $0 \leq z \leq r$ dimensions are of **cardinal scale** and
- the remaining dimensions are of purely **ordinal scale**.

Question: How can we utilize the cardinal dimensions without making unjustified assumptions about the ordinal ones?

Important Special Case: Mixed-Scaled Dimensions

Situation: Consider an r -dimensional space $A \subseteq \mathbb{R}^r$ and assume that

- the first $0 \leq z \leq r$ dimensions are of **cardinal scale** and
- the remaining dimensions are of purely **ordinal scale**.

Question: How can we utilize the cardinal dimensions without making unjustified assumptions about the ordinal ones?

Idea: Utilize the cardinal information only on those parts of A where there is no possible conflict with the ordinal information.

Important Special Case: Mixed-Scaled Dimensions

Situation: Consider an r -dimensional space $A \subseteq \mathbb{R}^r$ and assume that

- the first $0 \leq z \leq r$ dimensions are of **cardinal scale** and
- the remaining dimensions are of purely **ordinal scale**.

Question: How can we utilize the cardinal dimensions without making unjustified assumptions about the ordinal ones?

Idea: Utilize the cardinal information only on those parts of A where there is no possible conflict with the ordinal information.

Formalization: Consider A to be a subsystem of $\mathcal{P} = [\mathbb{R}^r, R_1^*, R_2^*]$, where

$$R_1^* = \left\{ (x, y) : x_j \geq y_j \ \forall j \leq r \right\}$$

$$R_2^* = \left\{ ((x, y), (x', y')) : \begin{array}{l} x_j - y_j \geq x'_j - y'_j \ \forall j \leq z \\ x_j \geq x'_j \geq y'_j \geq y_j \ \forall j > z \end{array} \right\}$$

Running Example (Pseudo-Label Selection): GSD

Suppose K competing specifications yield K likelihoods, so the utility of data point x in state s is a vector

$$c(s, z) = \left(\ell_s^{(1)}(\mathcal{D} \cup (z, \hat{y}(z))), \dots, \ell_s^{(K)}(\mathcal{D} \cup (z, \hat{y}(z))) \right) \in C \subseteq [0, 1]^K$$

Every single dimension is cardinal $\Rightarrow$ preference system $\mathcal{A} = [C, R_1, R_2]$ with representations $\mathcal{U}_{\mathcal{A}}$ (mixed-scaled case).

GSD for PLS (under a credal set)

$$x \succsim_{\text{GSD}} x' \quad :\Leftrightarrow \quad \forall u \in \mathcal{U}_{\mathcal{A}} \quad \forall \pi \in \mathcal{M} : \mathbb{E}_{\pi}(u \circ x) \geq \mathbb{E}_{\pi}(u \circ x')$$

$$\text{ch}_{\text{GSD}}(\mathcal{G}) = \{x \in \mathcal{G} : \nexists x' \in \mathcal{G} \text{ with } (x', x) \in P_{\succsim_{\text{GSD}}}\}$$

- This is the relation $R_{\forall\forall}$: add every instance not dominated for *all* compatible scales *and all* priors
- Too many survivors? Pass to $\mathcal{N}_{\mathcal{A}}^{\delta}$ and/or shrink $\mathcal{M}$.

Addressing Sampling Uncertainty

Practical Problem: Usually, we do not know π but only *i.i.d.* samples

$$\mathbf{X} = (X_1, \dots, X_n) \quad \text{and} \quad \mathbf{Y} = (Y_1, \dots, Y_m)$$

of X and Y are available.

Addressing Sampling Uncertainty

Practical Problem: Usually, we do not know π but only *i.i.d.* samples

$$\mathbf{X} = (X_1, \dots, X_n) \quad \text{and} \quad \mathbf{Y} = (Y_1, \dots, Y_m)$$

of X and Y are available.

Approach: Perform a statistical test for GSD.

Addressing Sampling Uncertainty

Practical Problem: Usually, we do not know π but only *i.i.d.* samples

$$\mathbf{X} = (X_1, \dots, X_n) \quad \text{and} \quad \mathbf{Y} = (Y_1, \dots, Y_m)$$

of X and Y are available.

Approach: Perform a statistical test for GSD.

Hypotheses:

$$H_0 : Y \preceq_{\text{GSD}} X \quad \text{vs.} \quad H_1 : \neg Y \preceq_{\text{GSD}} X$$

Addressing Sampling Uncertainty

Practical Problem: Usually, we do not know π but only *i.i.d.* samples

$$\mathbf{X} = (X_1, \dots, X_n) \quad \text{and} \quad \mathbf{Y} = (Y_1, \dots, Y_m)$$

of X and Y are available.

Approach: Perform a statistical test for GSD.

Hypotheses:

$$H_0 : Y \preceq_{\text{GSD}} X \quad \text{vs.} \quad H_1 : \neg Y \preceq_{\text{GSD}} X$$

Good News: As the worst case under the null hypothesis H_0 is

$$\pi_X = \pi_Y$$

performing a **permutation test** gives a valid level α test.

The Choice of the Test Statistic

Observation: It holds $X \succsim_{\text{GSD}} Y$ if and only if

$$D(X, Y) := \inf_{u \in \mathcal{U}_{\mathcal{A}}} (\mathbb{E}_{\pi}(u \circ X) - \mathbb{E}_{\pi}(u \circ Y)) \geq 0.$$

The Choice of the Test Statistic

Observation: It holds $X \succsim_{\text{GSD}} Y$ if and only if

$$D(X, Y) := \inf_{u \in \mathcal{U}_{\mathcal{A}}} (\mathbb{E}_{\pi}(u \circ X) - \mathbb{E}_{\pi}(u \circ Y)) \geq 0.$$

Consequence: A natural test statistic is the empirical version of $D(X, Y)$, i.e.,

$$d_{\mathbf{X}, \mathbf{Y}} : \Omega \rightarrow \mathbb{R}$$
$$\omega \mapsto \inf_{u \in \mathcal{N}_{\mathcal{A}_{\omega}}} \sum_{z \in (\mathbf{XY})_{\omega}} u(z) \cdot (\hat{\pi}_{\mathbf{X}}^{\omega}(\{z\}) - \hat{\pi}_{\mathbf{Y}}^{\omega}(\{z\}))$$

where $\hat{\pi}_{\mathbf{X}}^{\omega}$, $\hat{\pi}_{\mathbf{Y}}^{\omega}$, $(\mathbf{XY})_{\omega}$, and $\mathcal{A}_{\omega}$ are the corresponding empirical analoga.

The Choice of the Test Statistic

Observation: It holds $X \succsim_{\text{GSD}} Y$ if and only if

$$D(X, Y) := \inf_{u \in \mathcal{U}_{\mathcal{A}}} (\mathbb{E}_{\pi}(u \circ X) - \mathbb{E}_{\pi}(u \circ Y)) \geq 0.$$

Consequence: A natural test statistic is the empirical version of $D(X, Y)$, i.e.,

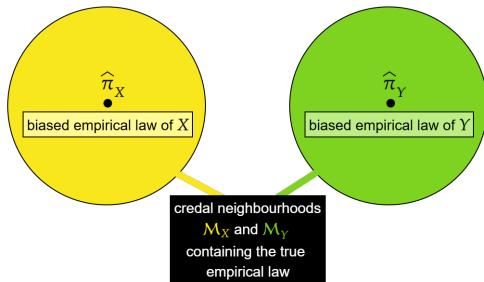
$$d_{\mathbf{X}, \mathbf{Y}} : \Omega \rightarrow \mathbb{R}$$
$$\omega \mapsto \inf_{u \in \mathcal{N}_{\mathcal{A}_{\omega}}} \sum_{z \in (\mathbf{XY})_{\omega}} u(z) \cdot (\hat{\pi}_{\mathbf{X}}^{\omega}(\{z\}) - \hat{\pi}_{\mathbf{Y}}^{\omega}(\{z\}))$$

where $\hat{\pi}_{\mathbf{X}}^{\omega}$, $\hat{\pi}_{\mathbf{Y}}^{\omega}$, $(\mathbf{XY})_{\omega}$, and $\mathcal{A}_{\omega}$ are the corresponding empirical analoga.

Good news: The test statistic $d_{\mathbf{X}, \mathbf{Y}}$ can be computed by solving one single **linear programming problem**. (Essentially the one for $R_{\forall\forall}$ from earlier...)

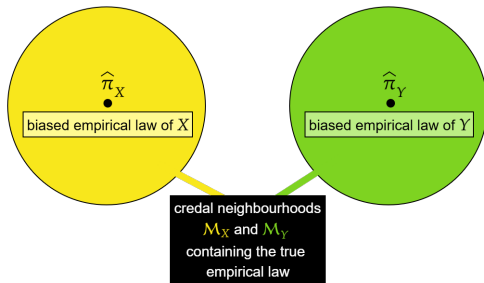
Credal Sets For Robustification

Idea: Use **credal sets** to robustify the permutation test to small deviations from the *i.i.d.* assumption.



Credal Sets For Robustification

Idea: Use **credal sets** to robustify the permutation test to small deviations from the *i.i.d.* assumption.



Adapted Resampling Scheme: Replace

$$\begin{aligned} \cdot d_{X,Y}(\omega_0) & \quad \text{by} \quad \inf_{(\pi_1, \pi_2) \in \mathcal{M}_X^{\omega_0} \times \mathcal{M}_Y^{\omega_0}} \tilde{d}_{X,Y}(\omega_0) \\ \cdot d_I(\omega_0) & \quad \text{by} \quad \sup_{(\pi_1, \pi_2) \in \mathcal{M}_X^{\omega_0} \times \mathcal{M}_Y^{\omega_0}} \tilde{d}_I(\omega_0) \end{aligned}$$

Results in: Valid (yet conservative) statistical test!

A special class of credal sets with a very intuitive interpretation are

γ -contamination models

Here, for $\omega \in \Omega$, $\gamma \in [0, 1]$, and $Z \in \{X, Y\}$ fixed, we set

$$\mathcal{M}_Z^\omega = \left\{ \pi : \pi \geq (1 - \gamma) \cdot \hat{\pi}_Z^\omega \right\}.$$

A special class of credal sets with a very intuitive interpretation are

γ -contamination models

Here, for $\omega \in \Omega$, $\gamma \in [0, 1]$, and $Z \in \{X, Y\}$ fixed, we set

$$\mathcal{M}_Z^\omega = \left\{ \pi : \pi \geq (1 - \gamma) \cdot \hat{\pi}_Z^\omega \right\}.$$

Enables quantifying robustness:

The observed p -values of the robustified test can then be computed as a function of the contamination size γ :

$$f(\gamma) := 1 - \frac{1}{N} \cdot \sum_{I \in \mathcal{I}_N} \mathbb{1}_{\{d_{\mathbf{X}, \mathbf{Y}}(\omega_0) - d_I > \frac{2\gamma}{(1-\gamma)}\}}$$

Why use multiple criteria in benchmark studies?

Reason 1: Performance is a latent construct

The application at hand suggests a very clear evaluation concept, which is too complex to be expressed in terms of a single metric.

Example: *Robustness* as stability under perturbations of both X and Y .

Why use multiple criteria in benchmark studies?

Reason 1: Performance is a latent construct

The application at hand suggests a very clear evaluation concept, which is too complex to be expressed in terms of a single metric.

Example: *Robustness* as stability under perturbations of both X and Y .

Reason 2: Quality is a multidimensional concept

It may be desirable to trade-off various competing quality dimensions.

Example: Trade-off between *accuracy* and *computation time*.

Why use multiple criteria in benchmark studies?

Reason 1: **Performance is a latent construct**

The application at hand suggests a very clear evaluation concept, which is too complex to be expressed in terms of a single metric.

Example: *Robustness* as stability under perturbations of both X and Y .

Reason 2: **Quality is a multidimensional concept**

It may be desirable to trade-off various competing quality dimensions.

Example: Trade-off between *accuracy* and *computation time*.

Take-away:

*Using **multiple criteria** should be **standard** rather than the exception.*

Five Challenges in (Multicriteria) Benchmarking

Setup: Let

- $\mathcal{D}$ denote the universe of **data sets**,
- $\mathcal{C}$ denote the finite set of all relevant **classifiers**,
- $(\phi_i : \mathcal{C} \times \mathcal{D} \rightarrow [0, 1])_{i \in \{1, \dots, r\}}$ denote a family of **quality criteria**,
- $\Phi := (\phi_1, \dots, \phi_r) : \mathcal{D} \times \mathcal{C} \rightarrow [0, 1]^r$ be the **multidimensional criterion**.

Five Challenges in (Multicriteria) Benchmarking

Setup: Let

- $\mathcal{D}$ denote the set of all relevant data sets,
- $\mathcal{C}$ denote the finite set of all relevant classifiers,
- $(\phi_i : \mathcal{C} \times \mathcal{D} \rightarrow [0, 1])_{i \in \{1, \dots, r\}}$ denote a family of quality criteria,
- $\Phi := (\phi_1, \dots, \phi_r) : \mathcal{D} \times \mathcal{C} \rightarrow [0, 1]^r$ be the multidimensional criterion.

Assumptions:

- For $0 \leq z \leq n$, the criteria $\phi_1, \dots, \phi_z$ are of **cardinal scale**.
- The remaining criteria are of purely **ordinal scale**.

Five Challenges in (Multicriteria) Benchmarking

data sets		D_1	$\dots$	D_s
classifier				
C_1		$\begin{pmatrix} \phi_1(C_1, D_1) \\ \vdots \\ \phi_r(C_1, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_r(C_1, D_s) \end{pmatrix}$
$\vdots$		$\vdots$	$\vdots$	$\vdots$
C_q		$\begin{pmatrix} \phi_1(C_q, D_1) \\ \vdots \\ \phi_r(C_q, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_q, D_s) \\ \vdots \\ \phi_r(C_q, D_s) \end{pmatrix}$

Five Challenges in (Multicriteria) Benchmarking

classifier \ data sets	data sets	
	D_1	D_2
C_1	$\begin{pmatrix} 0.8 \\ \vdots \\ 0.7 \end{pmatrix}$	$\begin{pmatrix} \phi_1(C_1, D_1) \\ \vdots \\ \phi_1(C_1, D_2) \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.8 \end{pmatrix}$	$\begin{pmatrix} \phi_1(C_q, D_1) \\ \vdots \\ \phi_1(C_q, D_2) \end{pmatrix}$

Challenge 1: Intra-dataset incomparability

On a **fixed** data set D it may hold

$$\phi_1(C_1, D) > \phi_1(C_2, D) \wedge \phi_2(C_1, D) < \phi_2(C_2, D).$$

Five Challenges in (Multicriteria) Benchmarking

data sets classifier	D_1	D_2
C_1	$\begin{pmatrix} 0.8 \\ \vdots \\ 0.8 \end{pmatrix}$	$\begin{pmatrix} \phi(C_1, D_1) \\ \vdots \\ \phi(C_1, D_2) \end{pmatrix}$
$\vdots$	$\vdots$	
C_q	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$	

Challenge 2: Conflicting datasets

Even if, for all $i \in \{1, \dots, n\}$, we have $\phi_i(C_1, D_1) > \phi_i(C_2, D_1)$,

Five Challenges in (Multicriteria) Benchmarking

data sets		D_1	$\dots$	D_s
classifier				
C_1		$\begin{pmatrix} 0.8 \\ \vdots \\ 0.8 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.6 \\ \vdots \\ \phi_r(C_1, D_s) \end{pmatrix}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	C_q	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.9 \\ \vdots \\ \phi_r(C_q, D_s) \end{pmatrix}$

Challenge 2: Conflicting datasets

Even if, for all $i \in \{1, \dots, n\}$, we have $\phi_i(C_1, D_1) > \phi_i(C_2, D_1)$, there may exists some $i_0 \in \{1, \dots, n\}$ such that

$$\phi_{i_0}(C_1, D_2) < \phi_{i_0}(C_2, D_2).$$

Five Challenges in (Multicriteria) Benchmarking

classifier \ data sets	data sets		
	D_1	$\dots$	D_s
C_1	$\begin{pmatrix} \text{medium} \\ \vdots \\ 0.8 \end{pmatrix}$	$\dots$	$\begin{pmatrix} \text{bad} \\ \vdots \\ 0.7 \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} \text{good} \\ \vdots \\ 0.93 \end{pmatrix}$	$\dots$	$\begin{pmatrix} \text{excellent} \\ \vdots \\ 0.64 \end{pmatrix}$

Challenge 3: Mixed-scaled quality metrics

Even if some of the quality metrics are only of **ordinal scale**, we still want to capture the entire information encoded in the metrics with **cardinal scale**.

Five Challenges in (Multicriteria) Benchmarking

data sets		D_1	$\dots$	D_s
classifier				
C_1		$\begin{pmatrix} 0.8 \\ \vdots \\ 0.8 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.8 \\ \vdots \\ 0.8 \end{pmatrix}$
$\vdots$		$\vdots$	$\vdots$	$\vdots$
C_q		$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.7 \\ \vdots \\ 0.7 \end{pmatrix}$

Challenge 4: Lack of inferential guarantees

Even if a decision can be made for a sample $(D_1, \dots, D_s)$ of data sets,

Five Challenges in (Multicriteria) Benchmarking

data sets		D_1^*	$\dots$	D_s^*
classifier				
C_1		$\begin{pmatrix} 0.7 \\ \vdots \\ 0.9 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.75 \\ \vdots \\ 0.4 \end{pmatrix}$
$\vdots$		$\vdots$	$\vdots$	$\vdots$
C_q		$\begin{pmatrix} 0.85 \\ \vdots \\ 0.67 \end{pmatrix}$	$\dots$	$\begin{pmatrix} 0.33 \\ \vdots \\ 0.98 \end{pmatrix}$

Challenge 4: Lack of inferential guarantees

Even if a decision can be made for a sample $(D_1, \dots, D_s)$ of data sets, no clear decision might be possible for a different sample $(D_1^*, \dots, D_s^*)$.

Five Challenges in (Multicriteria) Benchmarking

classifier \ data sets			
	D_1	i.i.d.!!	D_s
C_1	$\begin{pmatrix} \phi_1(C_1, D_1) \\ \vdots \\ \phi_r(C_1, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_r(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} \phi_1(C_q, D_1) \\ \vdots \\ \phi_r(C_q, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_q, D_s) \\ \vdots \\ \phi_r(C_q, D_s) \end{pmatrix}$

Challenge 5: Non-robustness under deviations from i.i.d.

Even if our classifier ranking comes with inferential guarantees under i.i.d. sampling of data sets,

Five Challenges in (Multicriteria) Benchmarking

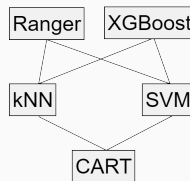
data sets classifier	D_1^*	contamination!!	D_s^*
C_1	$\begin{pmatrix} \phi_1(C_1, D_1) \\ \vdots \\ \phi_r(C_1, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_1, D_s) \\ \vdots \\ \phi_r(C_1, D_s) \end{pmatrix}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_q	$\begin{pmatrix} \phi_1(C_q, D_1) \\ \vdots \\ \phi_r(C_q, D_1) \end{pmatrix}$	$\dots$	$\begin{pmatrix} \phi_1(C_q, D_s) \\ \vdots \\ \phi_r(C_q, D_s) \end{pmatrix}$

Challenge 5: Non-robustness under deviations from i.i.d.

Even if our classifier ranking comes with inferential guarantees under i.i.d. sampling of data sets, these are invalid under **contaminated sampling**.

Roadmap

Start with the GSD relation $\succsim$ among classifiers
(borrowing ideas from decision theory ideas)



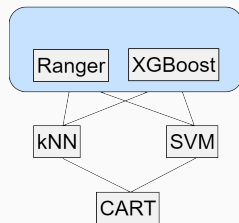
Roadmap

Start with the GSD relation $\succsim$ among classifiers
(borrowing ideas from decision theory ideas)



$$\text{gsd}(C) = \{C : \nexists C' \succ C\}$$

(more informative than Pareto)



Roadmap

Start with the GSD relation $\succsim$ among classifiers
(borrowing ideas from decision theory ideas)

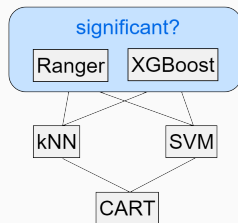


$$\text{gsd}(C) = \left\{ C : \nexists C' \succ C \right\}$$

(more informative than Pareto)



$H_0 : C \notin \text{gsd}(C) \text{ vs. } H_1 : C \in \text{gsd}(C)$
(providing inferential guarantees under i.i.d)



Roadmap

Start with the GSD relation $\succsim$ among classifiers
(borrowing ideas from decision theory ideas)



$$\text{gsd}(C) = \left\{ C : \nexists C' \succ C \right\}$$

(more informative than Pareto)

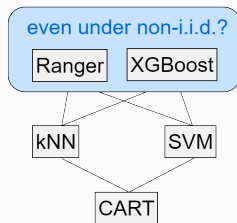


$H_0 : C \notin \text{gsd}(C) \text{ vs. } H_1 : C \in \text{gsd}(C)$

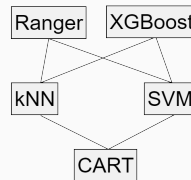
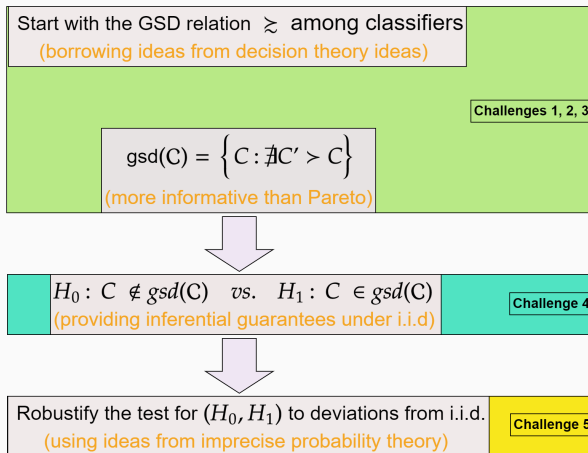
(providing inferential guarantees under i.i.d.)



Robustify the test for (H_0, H_1) to deviations from i.i.d.
(using ideas from imprecise probability theory)



Roadmap



Lifting GSD to Classifier Comparison

Main Idea: Embed the range $\Phi(\mathcal{C} \times \mathcal{D})$ into $\mathcal{P} = [\mathbb{R}^r, R_1^*, R_2^*]$.

Lifting GSD to Classifier Comparison

Main Idea: Embed the range $\Phi(\mathcal{C} \times \mathcal{D})$ into $\mathcal{P} = [\mathbb{R}^r, R_1^*, R_2^*]$.

For this:

- Interpret data sets in $\mathcal{D}$ as realizations of a variable $T : \Omega \rightarrow \mathcal{D}$.
- Associate each $C \in \mathcal{C}$ with the variable $\Phi_C := \Phi(C, T(\cdot))$ on Ω .

Lifting GSD to Classifier Comparison

Main Idea: Embed the range $\Phi(\mathcal{C} \times \mathcal{D})$ into $\mathcal{P} = [\mathbb{R}^r, R_1^*, R_2^*]$.

For this:

- Interpret data sets in $\mathcal{D}$ as realizations of a variable $T : \Omega \rightarrow \mathcal{D}$.
- Associate each $C \in \mathcal{C}$ with the variable $\Phi_C := \Phi(C, T(\cdot))$ on Ω .

Dominance and GSD-Front

Define the following preorder among classifiers $C, C' \in \mathcal{C}$:

$$C \succsim C' \quad :\Leftrightarrow \quad \Phi_C \succsim_{\text{GSD}} \Phi_{C'}$$

The relation $\succsim$ is called the **dominance relation** among classifiers.

The **GSD-front** is the set

$$\text{gsd}(\mathcal{C}) := \{C \in \mathcal{C} : \nexists C' \in \mathcal{C} \text{ s.t. } C' \succ C\},$$

where $\succ$ denotes the strict part of $\succsim$.

Theoretical and Empirical GSD-Front

We associate the following two sets to the dominance relation:

The GSD-Front

Let $T_1, \dots, T_s$ be *i.i.d.* copies of T and let $\rho \in [0, 1]$.

i) The **GSD-front** is the set

$$\text{gsd}(\mathcal{C}) := \{C \in \mathcal{C} : \nexists C' \in \mathcal{C} \text{ s.t. } C' \succ C\},$$

where $\succ$ denotes the strict part of $\succsim$.

ii) The ρ -**empirical GSD-front** is the (random) subset of $\mathcal{C}$ defined by

$$\text{egsd}_s^\rho(\mathcal{C}) = \left\{ C : \nexists C' \in \mathcal{C} \text{ s.t. } \begin{array}{l} d_{(\Phi_{C'}, \Phi_C)} \geq -\rho \\ d_{(\Phi_C, \Phi_{C'})} < 0 \end{array} \right\}.$$

Consistent Estimability of the GSD-Front

Estimating the GSD-Front

Denote by $\mathcal{I}_\Phi$ the set of all sets $\{a : u(a) \geq c\}$, where $c \in [0, 1]$ and $u \in \mathcal{U}_{\mathcal{P}_\Phi}$.

Assume the following:

- The VC-dimension of $\mathcal{I}_\Phi$ is finite.
- The relation $\succsim$ is antisymmetric.
- $\rho : \mathbb{N} \rightarrow [0, 1]$ converges to 0 with at most $\Theta(1/\sqrt[4]{s})$

Then $(\text{egsd}_s^{\rho(s)}(\mathcal{C}))_{s \in \mathbb{N}}$ is consistent, i.e.,

$$\pi \left(\left\{ \omega \in \Omega : \lim_{s \rightarrow \infty} \text{egsd}_s^{\rho(s)}(\mathcal{C}) = \text{gsd}(\mathcal{C}) \right\} \right) = 1,$$

where set convergence is defined via the trivial metric.

Consistent Tests for the GSD-Front

Intuitive Goal: **Inferential** comparisons of the **multivariate** and **mixed-scaled** performance of a new classifier C with a set $\mathcal{C}$ of state-of-the-art classifiers.

Consistent Tests for the GSD-Front

Intuitive Goal: **Inferential** comparisons of the **multivariate** and **mixed-scaled** performance of a new classifier C with a set $\mathcal{C}$ of state-of-the-art classifiers.

Stated formally: Develop a statistical test for the pair

$$H_0 : C \notin \text{gsd}(\mathcal{C}) \quad \text{vs.} \quad H_1 : C \in \text{gsd}(\mathcal{C})$$

Consistent Tests for the GSD-Front

Intuitive Goal: **Inferential** comparisons of the **multivariate** and **mixed-scaled** performance of a new classifier C with a set $\mathcal{C}$ of state-of-the-art classifiers.

Stated formally: Develop a statistical test for the pair

$$H_0 : C \notin \text{gsd}(\mathcal{C}) \quad \text{vs.} \quad H_1 : C \in \text{gsd}(\mathcal{C})$$

Or equivalently: Note that H_0 can be rewritten as:

$$H_0 : \exists C' \in \mathcal{C} \setminus \{C\} : C' \succsim C.$$

Thus, H_0 is false iff the hypothesis $H_0^{C'} : C' \succsim C$ is false **for every** $C' \in \mathcal{C} \setminus \{C\}$.

Consistent Tests for the GSD-Front

Intuitive Goal: **Inferential** comparisons of the **multivariate** and **mixed-scaled** performance of a new classifier C with a set $\mathcal{C}$ of state-of-the-art classifiers.

Stated formally: Develop a statistical test for the pair

$$H_0 : C \notin \text{gsd}(\mathcal{C}) \quad \text{vs.} \quad H_1 : C \in \text{gsd}(\mathcal{C})$$

Or equivalently: Note that H_0 can be rewritten as:

$$H_0 : \exists C' \in \mathcal{C} \setminus \{C\} : C' \succsim C.$$

Thus, H_0 is false iff the hypothesis $H_0^{C'} : C' \succsim C$ is false **for every** $C' \in \mathcal{C} \setminus \{C\}$.

Good news:

- $(H_0^{C'}, \neg H_0^{C'})$ can be tested by using the permutation **test from before**.
- $(H_0, \neg H_0)$ can be tested by running these tests **multiple times**.
- Robustification techniques from before can be **easily adapted**.

Benchmarking Experiments: Setup

We use 80 binary classification datasets from **OpenML** ([Van Rijn et al., 2013]).

Benchmarking Experiments: Setup

We use 80 binary classification datasets from **OpenML** ([Van Rijn et al., 2013]).

We compare the performance of **Support Vector Machine (SVM)** with

- **Random Forest (RF)**,
- **Decision Tree (CART)**,
- **Logistic Regression (LR)**,
- **Elastic net (GLMNet)**,
- **Extreme Gradient Boosting (xGBoost)**, and
- **k-Nearest Neighbors (kNN)**.

Benchmarking Experiments: Setup

We use 80 binary classification datasets from **OpenML** ([Van Rijn et al., 2013]).

We compare the performance of **Support Vector Machine (SVM)** with

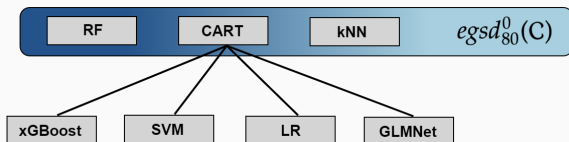
- **Random Forest (RF)**,
- **Decision Tree (CART)**,
- **Logistic Regression (LR)**,
- **Elastic net (GLMNet)**,
- **Extreme Gradient Boosting (xGBoost)**, and
- **k-Nearest Neighbors (kNN)**.

Comparison is based on the multivariate metric Φ composed of

- **Accuracy** (cardinal)
- **Test Computation Time** (treated as ordinal)
- **Training Computation Time** (treated as ordinal)

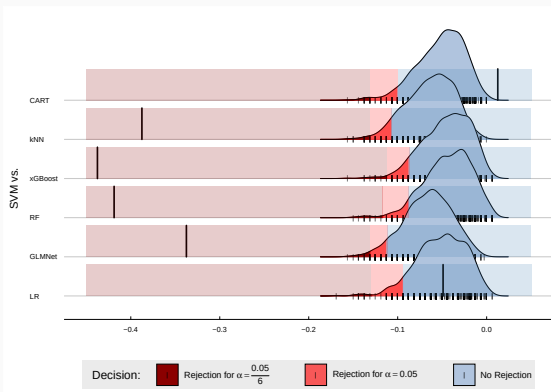
Benchmarking Experiments: Empirical GSD-Front

The Hasse graph of the empirical GSD relation:

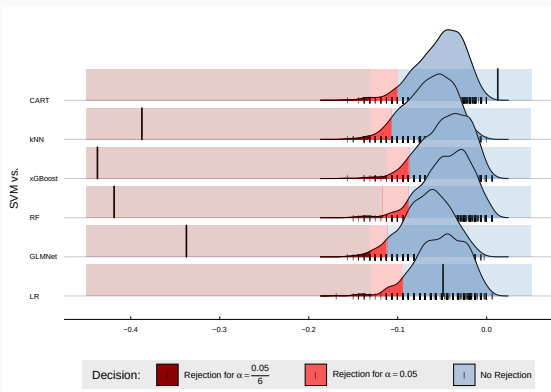


The blue shaded region symbolizes the 0-empirical GSD-front.

Benchmarking Experiments: Tests for GSD-Front



Benchmarking Experiments: Tests for GSD-Front

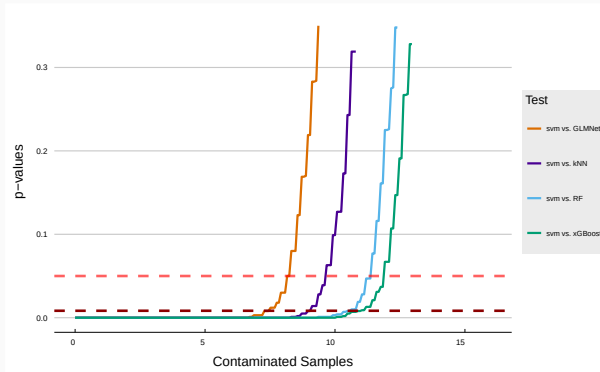


Result: On a 5% level, **SVM** significantly lies in the GSD-front of

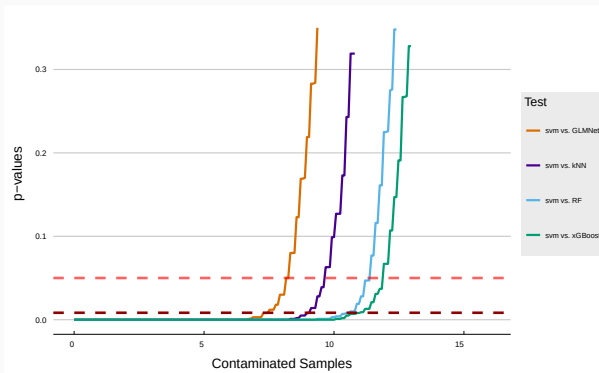
{SVM , kNN , xGBoost , RF , GLMNet}

if quality is measured by trading-off **time** and **accuracy**.

Benchmarking Experiments: Robustness



Benchmarking Experiments: Robustness

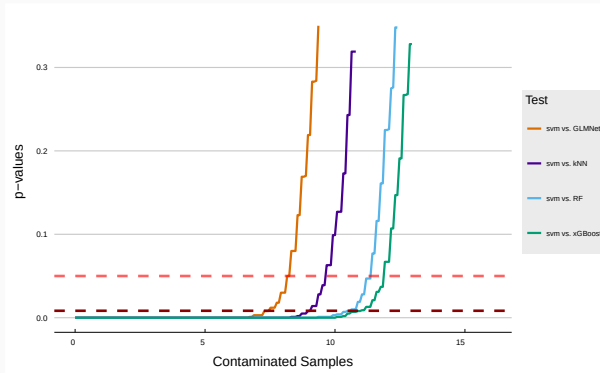


Result: Our test decision remains valid if up to

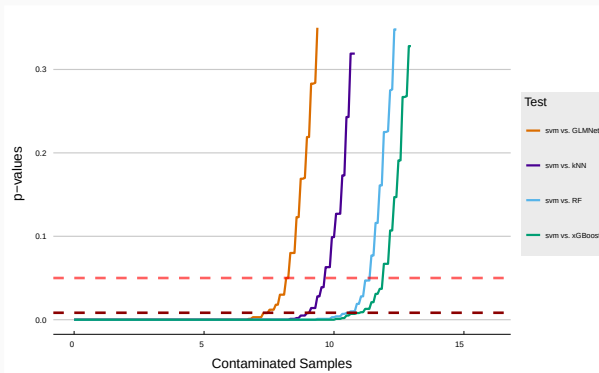
$$7 = \min\{7, 8, 11\}$$

data sets from our benchmark suite come from arbitrary distributions.

Benchmarking Experiments: Robustness



Benchmarking Experiments: Robustness



Result: Our test decision remains valid if up to

$$7 = \min\{7, 8, 11\}$$

data sets from our benchmark suite come from arbitrary distributions.



Some highlights of the paper:

- Introduces a decision model that avoids state-of-the-world assumptions.
- Represents decision problems empirically via act–consequence protocols.
- Develops (credal-set) methods for inference on empirical choice functions.

Thanks for your attention!





Coolen, F., Cozman, F., Denoeux, T., Jansen, C., Küchenhoff, H., Martin, R., Miranda, E., Plass, J., Rodemann, J., Schollmeyer, G., et al. (2026).

Thomas augustin's contributions to imprecise probability and statistics.

International Journal of Approximate Reasoning, page 109781.



De Finetti, B. (1937).

La prévision: ses lois logiques, ses sources subjectives.

Annales de l'Institut Henri Poincaré, 7:1–68.



DeGroot, M. H. (1962).

Uncertainty, information, and sequential experiments.

The Annals of Mathematical Statistics, 33(2):404–419.



Ellsberg, D. (1961).

Risk, ambiguity, and the savage axioms.

The Quarterly Journal of Economics, 75:643–669.



Fisher, R. A. (1935).

The Design of Experiments.

Oliver and Boyd, Edinburgh.



Gilboa, I. and Schmeidler, D. (1989).

Maxmin expected utility with non-unique prior.

J Math Econ, 18:141–153.



Gneiting, T. and Raftery, A. E. (2007).

Strictly proper scoring rules, prediction, and estimation.

Journal of the American Statistical Association, 102(477):359–378.



Grünwald, P. D. and Dawid, A. P. (2004).

Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory.

The Annals of Statistics, 32(4):1367–1433.



Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009).

The elements of statistical learning: data mining, inference, and prediction, volume 2.

Springer.



Horvitz, E., Breese, J., and Henrion, M. (1988).

Decision theory in expert systems and artificial intelligence.

International Journal of Approximate Reasoning, 2:247–302.



Hüllermeier, E. (2021).

Prescriptive machine learning for automated decision making: Challenges and opportunities.

arXiv.



Jansen, C., Augustin, T., and Schollmeyer, G. (2017).

Decision theory meets linear optimization beyond computation.

In Antonucci, A., Cholvy, L., and Papini, O., editors, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty. ECSQARU 2017. Lecture Notes in Computer Science, vol 10369.*, pages 329–339. Springer, Cham.



Jansen, C. and Schollmeyer, G. (2026).

Robust bayes acts under prior perturbations: Contamination, stability, and selection paths.



Jansen, C., Schollmeyer, G., and Augustin, T. (2018).

Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences.

International Journal of Approximate Reasoning, 98:112–131.



Jansen, C., Schollmeyer, G., and Augustin, T. (2022).

Quantifying Degrees of E-admissibility in Decision Making with Imprecise Probabilities, pages 319–346.

Springer International Publishing, Cham.



Jansen, C., Schollmeyer, G., Augustin, T., and Rodemann, J. (2026).

Empirical decision theory.

Information Sciences, page 123906.



Kofler, E. and Menges, G. (1976).

Entscheiden bei unvollständiger Information.

Springer, Berlin.



Levi, I. (1974).

On indeterminate probabilities.

The Journal of Philosophy, 71:391–418.



Pearl, J. and Mackenzie, D. (2018).

The book of why: the new science of cause and effect.

Basic books.



Rodemann, J., Jansen, C., Schollmeyer, G., and Augustin, T. (2023).

In all likelihoods: Robust selection of pseudo-labeled data.

In *International Symposium on Imprecise Probability: Theories and Applications*, pages 412–425. PMLR.



Savage, L. (1954).

The Foundations of Statistics.

Wiley.



Schervish, M. J., Seidenfeld, T., Kadane, J. B., and Levi, I. (2003).

Extensions of expected utility theory and some limitations of pairwise comparisons.

In *Proceedings of ISIPTA 2003*.



Troffaes, M. (2007).

Decision making under uncertainty using imprecise probabilities.

International Journal of Approximate Reasoning, 45:17–29.



Utkin, L. and Augustin, T. (2005).

Powerful algorithms for decision making under partial prior information and general ambiguity attitudes.

In *Proceedings of the Fourth International Symposium on Imprecise Probability: Theories and Applications*, pages 349–358. SIPTA.



Van Rijn, J. N., Bischl, B., Torgo, L., Gao, B., Umaashankar, V., Fischer, S., Winter, P., Wiswedel, B., Berthold, M. R., and Vanschoren, J. (2013).

OpenML: A collaborative science platform.

In *ECML PKDD 2013*, pages 645–649.



von Neumann, J., Morgenstern, O., Kuhn, H., and Rubinstein, A. (1944).

Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition).

Princeton University Press.



Wald, A. (1949).

Statistical decision functions.

The Annals of Mathematical Statistics, 20(2):165–205.



Walley, P. (1991).

Statistical Reasoning with Imprecise Probabilities.

Chapman and Hall, London.



Weichselberger, K. and Augustin, T. (1998).

Analysing Ellsberg's paradox by means of interval-probability.

In *Econometrics in Theory and Practice: Festschrift for Hans Schneeweiß*, pages 291–304. Springer.