Analogies and Theories in Belief Formation

Itzhak Gilboa – Tel Aviv University and HEC, Paris
ISIPTA 2015

Joint works of subsets of

A. Billot, G. Gayer, I. Gilboa, O. Lieberman, A. Postlewaite, D. Samet, L. Samuelson, D. Schmeidler

Background

- Classics:
  - Ramsey (1926), de Finetti (1931,7)
  - von-Neumann-Morgenstern (1944)
  - Savage (1954)
  - Anscombe-Aumann (1963)

- Problems:
  - Descriptive
  - Normative

Background – cont.

- Alternative theories
  - Schmeidler (1989) Choquet EU
  - G-Sch (1989) Maxmin EU
  - Klibanoff, Marinacci, Mukerji (2005) (Nau, Seo…) "Smooth Model"
  - Maccheroni, Marinacci, Rustichini (2006) "Variational Preferences"

- Still the "black box" paradigm

Background – cont.

- Case-Based Decision Theory
  - (w/ Schmeidler, Theory of Case Based Decisions, CUP 2001)
- Probabilities from cases
  - (w/ Schmeidler and others, Case-Based Prediction, World Scientific 2012)
- Analogies and Theories
  - (w/ Samuelson, Schmeidler and others, Analogies and Theories, OUP, 2015)

Statistics and Psychology

- This project touches on both
- And we found ourselves axiomatizing known formulae
- Surprisingly, known in both domains
  - Which goes beyond this project
  - Sometimes, even the mistakes

Probabilities from Cases: Similarity-weighted frequencies

The data: $\left(x_i^{\,1},\ldots,x_i^{\,m},y_i\right)_{i\le n}$

where $\left(x_i^{\,1},\ldots,x_i^{\,m}\right)\in\Re^m$ and $y_i\in\{0,1\}$

We are asked about the probability that $y_p=1$ for a new data point $\left(x_p^1,\ldots,x_p^m\right)$

Similarity-weighted frequencies – Formula (Kernel)

Choose a similarity function $s : \Re^m \times \Re^m \to \Re_{++}$

Given observations $\left(x_i^{\ 1}, \ldots, x_i^{\ m}, y_i\right)_{i \le n}$

and a new data point $\left(x_p^1, \ldots, x_p^m\right)$

estimate $P(y_p = 1)$ by $y_p^s = \dfrac{\sum_{i \le n} s(x_i, x_p) y_i}{\sum_{i \le n} s(x_i, x_p)}$

Similarity-weighted frequencies – Interpretation

- Special cases of $$y_p^s = \frac{\sum_{i \le n} s(x_i, x_p) y_i}{\sum_{i \le n} s(x_i, x_p)}$$
  - If s is constant: an estimate of the expectation

  (in fact, "repeated experiment" is always a matter of subjective judgment of equal similarity)
  - If $s(x_i, x_p) = 1_{\{x_i = x_p\}}$: an estimate of the conditional expectation
- Useful when precise updating leaves us with a sparse database
- Akin to interpolation
- But not to extrapolation!

Axiomatization – Setup

$M = \Re^{m+1}$ observations (case types)

$$\left(\left(x^1,\ldots,x^m\right),y\right)\approx\left(x^1,\ldots,x^m,y\right)$$

A database is a multi-set of observations

$$I : M \to Z_+$$

We will refer to a database as a sequence or a multi-set interchangeably.

Axiomatization I: Observables

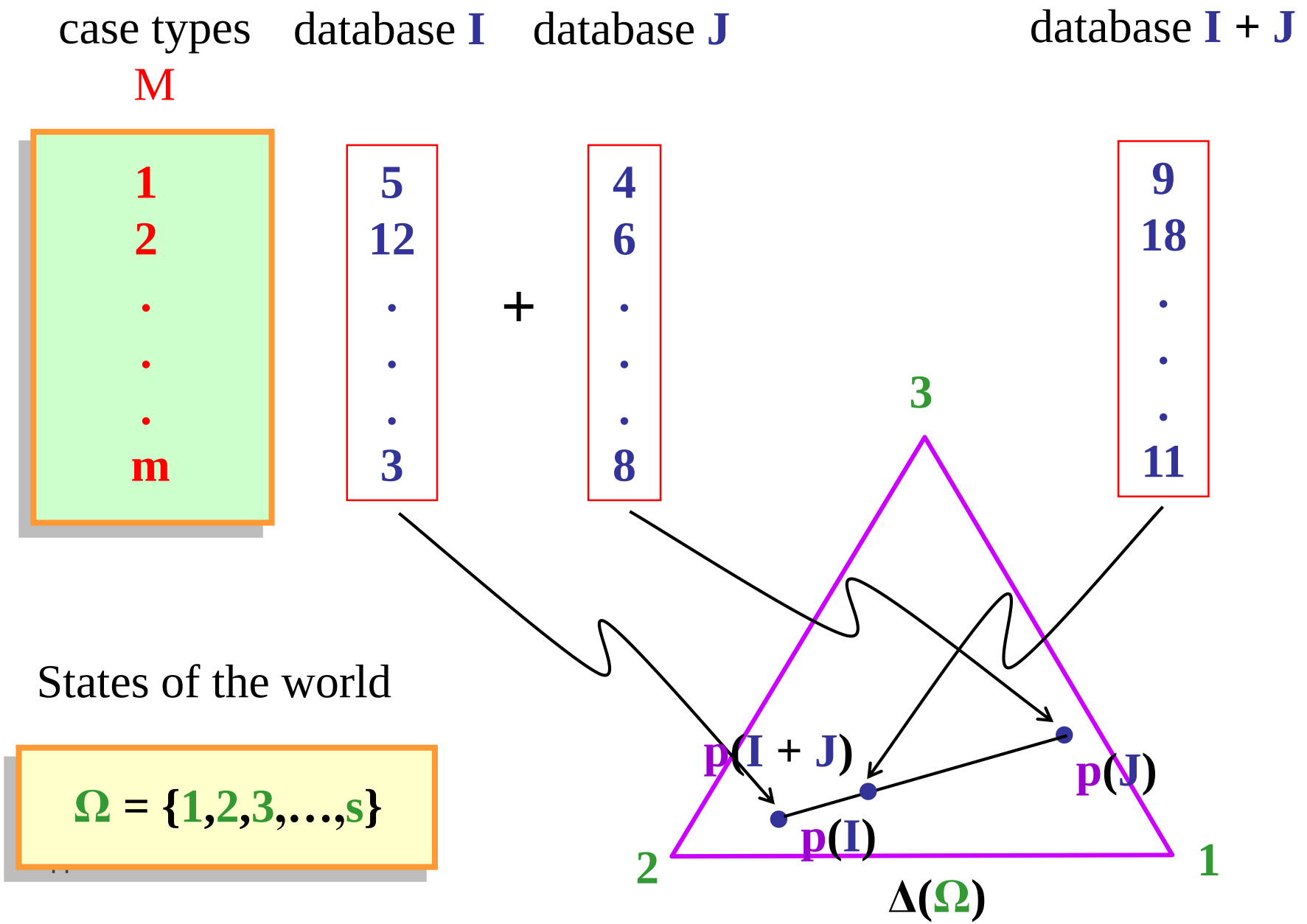
- A state space $\Omega = \{1,..., s\}$
- Fix a new data point $\left(x_p^1,..., x_p^m\right) \in \Re^m$
- Databases

$$I : M \to Z_+$$

- A probability assignment function

$$p : I \mapsto \Delta(\Omega), \quad I \neq 0$$

The combination axiom



The combination axiom

- Formally

$$p(I+J)=\lambda p(I)+(1-\lambda)\,p(J)$$

for some

$$0<\lambda<1$$

Theorem I

- The combination axiom holds, and not all $\{p(I)\}_I$ are collinear

if and only if

- For each $c \in M$ there are $p^c \in \Delta(\Omega)$, not all collinear, and $s_c > 0$ such that

$$p(I) = \frac{\sum_{c \in M} I(c) s_c p^c}{\sum_{c \in M} I(c) s_c}$$

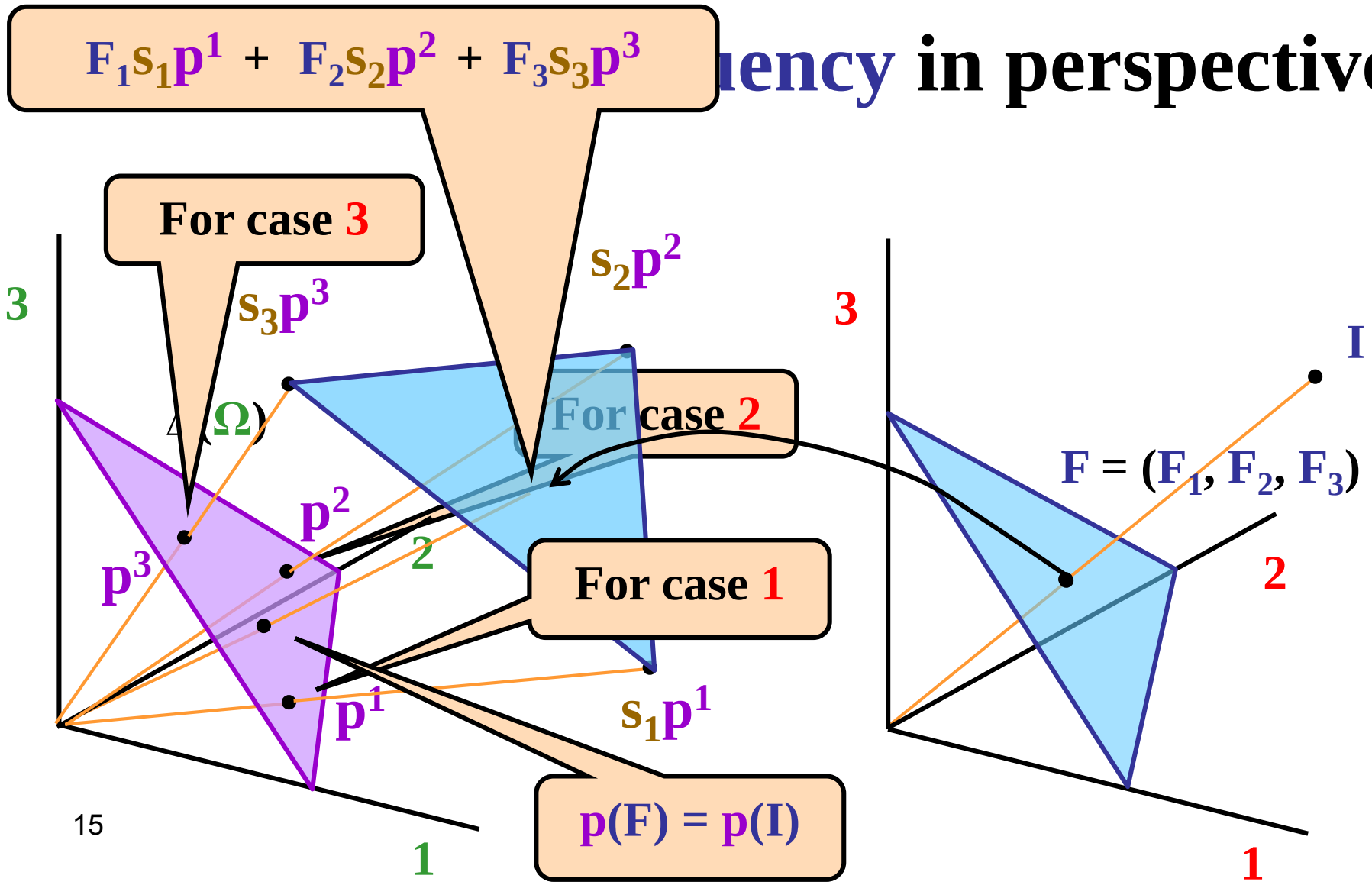
  - In "Probabilities as Similarity-Weighted Frequencies" w/ Billot, Samet, Schmeidler

The perspective

Probability
of states

Frequency
of cases

uency in perspective



Theorem II

Some axioms hold iff there exists a function $s : \Re^m \times \Re^m \to \Re_{++}$ such that $\geq_I$ ranks values by their proximity to

$$y_p^s = \frac{\sum_{i \leq n} s(x_i, x_p) y_i}{\sum_{i \leq n} s(x_i, x_p)}$$

where $x_i = \left(x_i^1, ..., x_i^m\right)$ and $I \approx \left(\left(x_i^1, ..., x_i^m, y_i\right)\right)_{i=1}^n$

The function s is unique up to multiplication by $\lambda > 0$

- In "Empirical Similarity" w/Lieberman and Schmeidler

Theorem III

Some additional axioms hold iff there exists a norm

$$n : \Re^m \to \Re_+$$

such that

$$s(x,z) = e^{-n(x-z)}$$

- Satisfies "multiplicative transitivity":

$$s(x,z) \geq s(x,y)s(y,z)$$

- In "Exponential Similarity" w/ Billot and Schmeidler

The Similarity – whence?

- In “Empirical Similarity” w/Lieberman and Schmeidler we propose an empirical approach:
- Estimate the similarity function from the data
- A parametrized approach: Consider a certain functional form
- Choose a criterion to measure goodness of fit
- Find the best parameters

A functional form

- Consider a weighted Euclidean distance

$$d_w(x_i, x_t) = \sqrt{\sum_{j=1}^{m} w_j (x_{ij} - x_{tj})^2}$$

and

$$s_w(x_i, x_t) = e^{-d_w(x_i, x_t)}$$

Selection criteria

- Find weights that would minimize

$$\sum_i \left(y_i - \hat{y}_i\right)^2$$

- Or: round off $\hat{y}_i$ to get a prediction $y_i^p \in \{0,1\}$
  - and then minimize

How objective is it?

- Modeling choices that can affect the "probability":
  - Choice of X's and of sample
  - Choice of functional form
  - Choice of goodness of fit criterion

- As usual, objectivity may be an unattainable ideal
- But it doesn't mean we shouldn't try.

Statistical inference

- In "Empirical Similarity" w/Lieberman and Schmeidler we also develop statistical inference tools for our estimation procedure
- Assume that the data were generated by a DGP of the type

$$P(Y_t = 1) = \frac{\sum_{i<t} s(X_i, X_t) Y_i}{\sum_{i<t} s(X_i, X_t)} + \varepsilon_t$$

- Estimate the similarity function from the data
- Perform statistical inference

Statistical inference – cont.

- Estimate the weights w_j by maximum likelihood
- Test hypotheses of the form

$$H_0 : w_j = 0$$

- Predict out-of-sample by the maximum likelihood estimators (via the similarity-weighted average formula)

Failures of the combination axiom

- Integration of induction and deduction
  - Learning the parameter of a coin
  - Linear regression

Limited to case-to-case induction, generalizing empirical frequencies

Failures of the combination axiom – cont.

- Second order induction
  - Learning the similarity function

In particular, doesn't allow the similarity function to get more concentrated for large databases

Combination restricted to periods of "no learning".

Combining Theories and Analogies

- History unfolds; a state of the world:

$$\omega = (x_0, y_0, \dots, x_t, y_t, \dots)$$

- Given

$$h_t = (x_0, y_0, \dots, x_t)$$

rank possible subsets of y_t

- Conjectures are subsets of the state space
- A weight on each (measurable subset of) conjecture(s) yields a Dempster-Shafer Belief Function

Learning in the Model

- Conjectures that have been refuted, i.e., that are disjoint from the set defined by h_t, are discarded
- The reasoner continues by aggregating the weights of the non-discarded ones
- Turns out to be the Dempster-Shafer update

Modes of Reasoning

- Bayesian: All the weight is put on singletons
- Case-Based: All the weight is put on "conjectures" of the type: "In period i we'll observe characteristics x, in period t we'll observe characteristics z, and the outcomes will be identical"
- Rule-Based: Every rule is identified with the states that don't refute it

Dynamics of Reasoning

- Under mild assumptions that mean that
  - The reasoner doesn't know the nature of the process
  - The reasoner is "open-minded"

- The reasoner converges away from Bayesian reasoning

Example

- The year is 1965
- For each of the following 60 years, predict war (1) or not (0)

- 2^{60} states
- 2^{60} Bayesian conjectures
- $\binom{60}{2} \cong 1800$ case-based conjectures

Example – cont.

- Suppose that weight $(1-\varepsilon)$ is split evenly among the 2^{60} Bayesian conjectures
- And ε – among the $\binom{60}{2}$ case-based conjectures
- In 2015, we have

$\frac{(1-\varepsilon)}{2^{50}}$ left on the Bayesian

$50 * \frac{\varepsilon}{1800}$ left on the Case-Based