Towards a formulation of fuzzy contrastive explanations

Isabelle Bloch and Marie-Jeanne Lesot

Sorbonne Université, CNRS, LIP6

FuzzIEEE, Padova, 2022

Context

- **XAI:** generating explanations for decisions
  - vast domain!
  - diversity of tasks, needs, frameworks
  - diversity of approaches, categorisations, discussion axes

    (e.g. Guidotti et al. 19, Arrieta et al. 20, Rudin et al. 21, Molnar 22, ...)

Context

- XAI: generating explanations for decisions
- **Logical formal framework: structural causal models**

(Halpern and Pearl, 05, 15)

Context

- XAI: generating explanations for decisions
- **Logical formal framework: structural causal models**

(Halpern and Pearl, 05, 15)

- Contrastive nature of explanations (Miller, 19, 20)
  - "why make decision P rather than Q?"
  - ex: "why does this flat cost 300k€ rather than 200k€?"

Context

- XAI: generating explanations for decisions
- **Logical formal framework: structural causal models**

(Halpern and Pearl, 05, 15)

- Contrastive nature of explanations (Miller, 19, 20)
  - "why make decision P rather than Q?"
  - ex: "why does this flat cost 300k€ rather than 200k€?"

$\Longrightarrow$ **Principles for an extension** to the case of **imperfect data and knowledge**

- formal framework of **fuzzy logic** (Zadeh, 65)
- model inputs: imprecisely described data instances
- model description: imprecise knowledge about functional relations
- output level: explanation expression

Structural causal models

(Halpern and Pearl, 05, 15)

- **Signature** $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$
  - exogenous and endogenous variables and their definition domains
  - ex: $\mathcal{U} = \{\text{area, year, length, width}\}$, $\mathcal{V} = \{\text{surface, price}\}$
    $\mathcal{R}(\text{length}) = \mathbb{R}^+$, $\mathcal{R}(\text{year}) = [\![1800, 2022]\!]$
- **Functional relations** $\mathcal{F} = \{F_X, X \in \mathcal{V}\}$
  - ex : $F_{\text{surface}} = \text{length} \times \text{width}$, $F_{\text{price}} = f(\text{surface, year, area})$
- **Context**: values for each exogenous variable $u \in \mathcal{U}$
  - ex : $\mathbf{u}* = \langle \text{Padova}, 1963, 7, 10 \rangle$
- Causal model: $M = (\mathcal{S}, \mathcal{F})$
- Situation: $(M, \mathbf{u})$

Structural causal models

(Halpern and Pearl, 05, 15)

- **Signature** $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$
  - exogenous and endogenous variables and their definition domains
  - ex: $\mathcal{U} = \{\text{area, year, length, width}\}$, $\mathcal{V} = \{\text{surface, price}\}$
    $\mathcal{R}(\text{length}) = \mathbb{R}^+$, $\mathcal{R}(\text{year}) = [\![1800, 2022]\!]$
- **Functional relations** $\mathcal{F} = \{F_X, X \in \mathcal{V}\}$
  - ex : $F_{\text{surface}} = \text{length} \times \text{width}$, $F_{\text{price}} = f(\text{surface, year, area})$
- **Context**: values for each exogenous variable $u \in \mathcal{U}$
  - ex : $\mathbf{u}* = \langle \text{Padova}, 1963, 7, 10 \rangle$
- Assertion $(M, \mathbf{u}) \models [\mathbf{Y} = \mathbf{y}](X = x)$ holds if
  - for model $M = (\mathcal{S}, \mathcal{F})$ and exogenous values $\mathbf{u}$
  - if endogenous variables $\mathbf{Y}$ are forced to $\mathbf{y}$
  - then X takes value x
  - functional notation $\tilde{F}_X(\mathbf{u}, \mathbf{Y} = \mathbf{y}) = x$

Structural causal models

(Halpern and Pearl, 05, 15)

- **Signature** $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$
  - exogenous and endogenous variables and their definition domains
  - ex: $\mathcal{U} = \{\text{area, year, length, width}\}$, $\mathcal{V} = \{\text{surface, price}\}$
    $\mathcal{R}(\text{length}) = \mathbb{R}^+$, $\mathcal{R}(\text{year}) = [\![1800, 2022]\!]$
- **Functional relations** $\mathcal{F} = \{F_X, X \in \mathcal{V}\}$
  - ex : $F_{\text{surface}} = \text{length} \times \text{width}$, $F_{\text{price}} = f(\text{surface, year, area})$
- **Context**: values for each exogenous variable $u \in \mathcal{U}$
  - ex : $\mathbf{u}* = \langle \text{Padova}, 1963, 7, 10 \rangle$
- Example: $(M, \mathbf{u}*) \models [\text{length} = 5.3](\text{surface} = 53)$

Explanations

(Miller, 19, 20)

- $\mathbf{X} = \mathbf{x}$ is a **sufficient cause** of φ in situation $(M, \mathbf{u})$ if
  - $(M, \mathbf{u}) \models \mathbf{X} = \mathbf{x}$ and $(M, \mathbf{u}) \models \varphi$
  - there exist values $\mathbf{x}'$ such that if $\mathbf{X} = \mathbf{x}'$, keeping the other variables unchanged, then φ is not satisfied anymore

Explanations

(Miller, 19, 20)

- $\mathbf{X} = \mathbf{x}$ is a **sufficient cause** of φ in situation $(M, \mathbf{u})$ if
  - $(M, \mathbf{u}) \models \mathbf{X} = \mathbf{x}$ and $(M, \mathbf{u}) \models \varphi$
  - there exist values $\mathbf{x}'$ such that if $\mathbf{X} = \mathbf{x}'$, keeping the other variables unchanged, then φ is not satisfied anymore
- **Actual cause**: minimal sufficient cause
  - partial cause: subset of an actual cause

Explanations

(Miller, 19, 20)

- $\mathbf{X} = \mathbf{x}$ is a **sufficient cause** of φ in situation $(M, \mathbf{u})$ if
  - $(M, \mathbf{u}) \models \mathbf{X} = \mathbf{x}$ and $(M, \mathbf{u}) \models \varphi$
  - there exist values $\mathbf{x}'$ such that if $\mathbf{X} = \mathbf{x}'$, keeping the other variables unchanged, then φ is not satisfied anymore
- **Actual cause**: minimal sufficient cause
  - partial cause: subset of an actual cause
- Example: $(M, \langle \text{Padova}, 1963, 7, 10 \rangle)$, $\varphi = (s = 70)$
  - sufficient cause: $\{(w = 7, l = 10)\}$ or $\{(w = 7)\}$ or $\{(l = 10)\}$
  - actual causes: $\{(w = 7)\}$ or $\{(l = 10)\}$

Explanations

(Miller, 19, 20)

- $\mathbf{X} = \mathbf{x}$ is a **sufficient cause** of φ in situation $(M, \mathbf{u})$ if
  - $(M, \mathbf{u}) \models \mathbf{X} = \mathbf{x}$ and $(M, \mathbf{u}) \models \varphi$
  - there exist values $\mathbf{x}'$ such that if $\mathbf{X} = \mathbf{x}'$, keeping the other variables unchanged, then φ is not satisfied anymore
- **Actual cause**: minimal sufficient cause
  - partial cause: subset of an actual cause
- **Contrastive cause of φ as opposed to ψ in $(M, \mathbf{u})$**:
  pair $(\mathbf{X} = \mathbf{x}, \mathbf{X} = \mathbf{y})$ such that
  - $\mathbf{X} = \mathbf{x}$ is a partial cause of φ in $(M, \mathbf{u})$
  - there exists $\mathbf{W} \subset \mathcal{V}$ and values $\mathbf{w}$ such that $\mathbf{X} = \mathbf{y}$ is a partial cause of ψ in $(M_{\mathbf{W}=\mathbf{w}}, \mathbf{u})$
  - the values $\mathbf{x}$ and $\mathbf{y}$ are incompatible
  - $\mathbf{X}$ is maximal

Explanations

(Miller, 19, 20)

- **Contrastive cause of φ as opposed to ψ in $(M, \mathbf{u})$**:
  pair $(\mathbf{X} = \mathbf{x}, \mathbf{X} = \mathbf{y})$ such that
  - $\mathbf{X} = \mathbf{x}$ is a partial cause of φ in $(M, \mathbf{u})$
  - there exists $\mathbf{W} \subset \mathcal{V}$ and values $\mathbf{w}$ such that $\mathbf{X} = \mathbf{y}$ is a partial cause of ψ in $(M_{\mathbf{W}=\mathbf{w}}, \mathbf{u})$
  - the values $\mathbf{x}$ and $\mathbf{y}$ are incompatible
  - $\mathbf{X}$ is maximal
- Example: $(M, \langle \text{Padova}, 1963, 7, 10 \rangle)$, $\varphi = (s = 70)$
  - sufficient cause: $\{(w = 7, l = 10)\}$ or ...
  - actual causes: $\{(w = 7)\}$ or $\{(l = 10)\}$
  - contrastive cause wrt $\psi = (s = 35)$: $\{(l = 10, l = 5)\}$

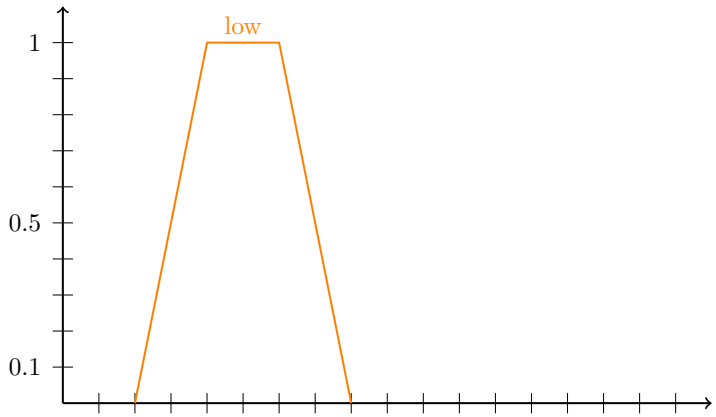
Extension to imperfect data and knwowledge

- Three levels: **fuzzy expert knowledge** and **fuzzy causes**
- Model inputs: imprecisely described data instances
  - $\mathbf{u}* = \langle$Padova, *in the 60s*, *around 7*m, *around 10*$m\rangle$
- Model description: imprecise knowledge about functional relations
  - if surface is *large*, price is *around 300k€*
- Explanation level: imprecise expression
  - why is this flat *expensive* rather than *affordable*?

Fuzzy logic in a nutshell

(Zadeh, 65)

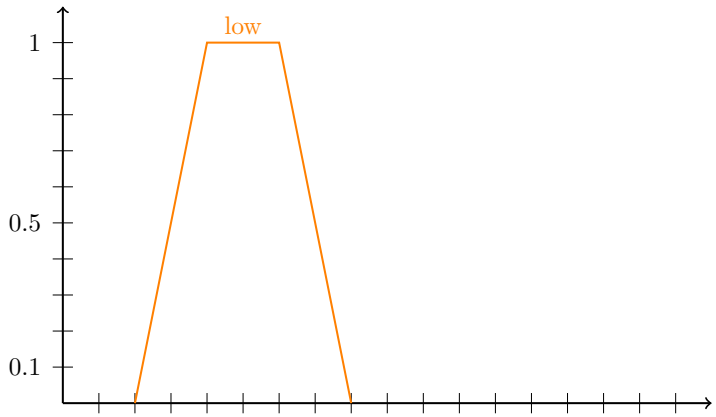
- Fuzzy subset: membership function $\mu_A : \mathcal{U} \longrightarrow [0, 1]$

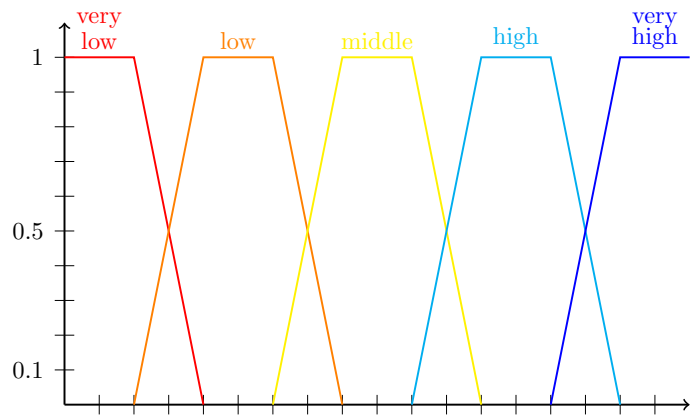


Fuzzy logic in a nutshell

(Zadeh, 65)

- Fuzzy subset: membership function $\mu_A : \mathcal{U} \longrightarrow [0, 1]$

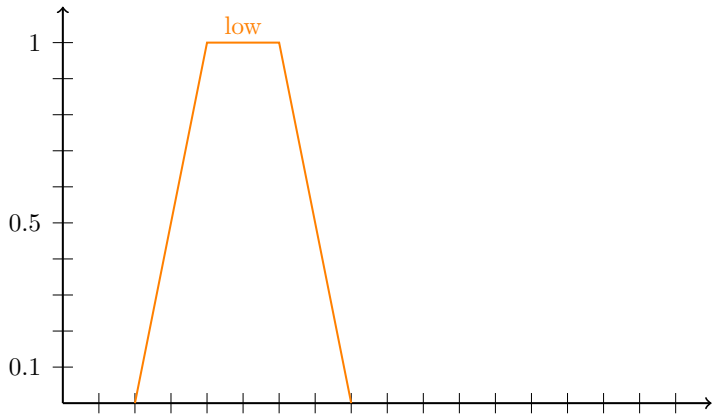


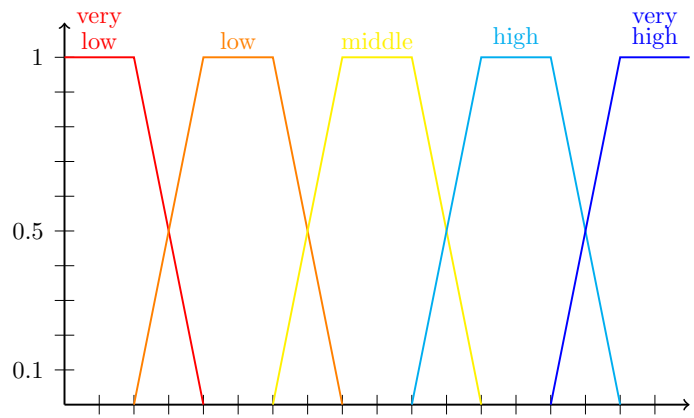


Fuzzy logic in a nutshell

(Zadeh, 65)

- Fuzzy subset: membership function $\mu_A : \mathcal{U} \longrightarrow [0, 1]$



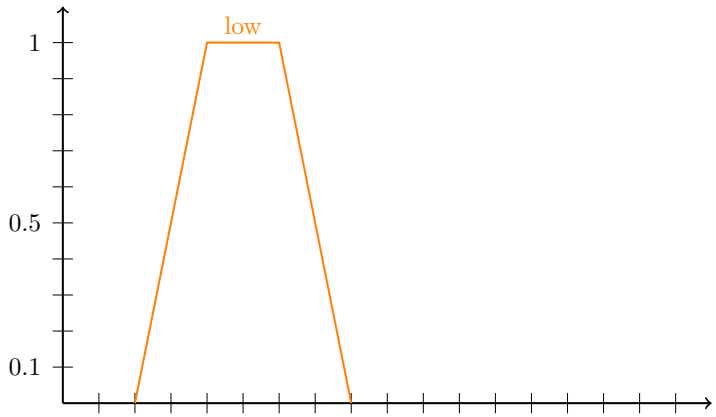


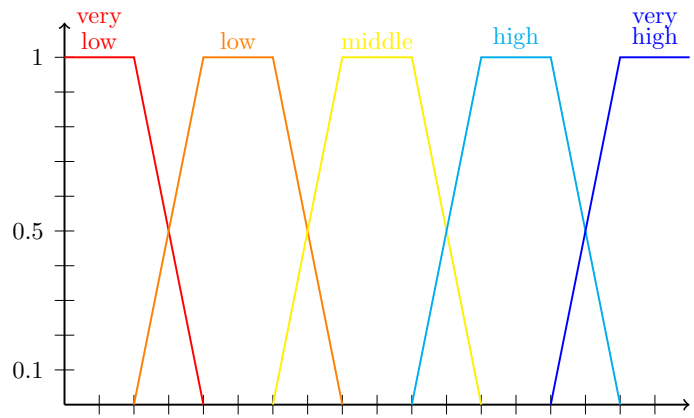
  - set manipulation: intersection, union, ...

Fuzzy logic in a nutshell

(Zadeh, 65)

- Fuzzy subset: membership function $\mu_A : \mathcal{U} \longrightarrow [0, 1]$





  - set manipulation: intersection, union, ...

- Fuzzy logic:
  - predicate semantics: fuzzy subset of the domain
  - formula manipulation: conjunction, disjunction, implication, ...
  - inference: Generalised Modus Ponens

Fuzzy model requirements

- Imprecise variable values
  - fuzzy signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$ and fuzzy context $\mathbf{u}$
  - fuzzy function relations $\mathcal{F}$
  - propagation: extension principle and generalised modus ponens
- Propagation to the logical assertions $(M, \mathbf{u}) \models \varphi$
  - computation of satisfaction degrees

Fuzzy signature

- For any $X \in \mathcal{U} \cup \mathcal{V}$, $\mathcal{R}(X) = (Dom(X), V(X))$ with
  - $Dom(X)$: domain, universe of fuzzy set definitions
  - $V(X) \subseteq FS(Dom(X))$

- Specific cases: examples for $Dom(X) = \mathbb{R}^+$
  - crisp case: $V(X) = \mathbb{R}^+$
  - predefined fuzzy sets: linguistic modalities
    - $V(X) = \{big, middleSized, small\}$ or $V(X) = \{about\ x, x \in \mathbb{R}^+\}$
    - each $v \in V(X)$ associated to $\mu_v : Dom(X) \rightarrow [0, 1]$
- Advantages
  - adaptation to the user considered context
  - legible expression of the explanations

Functional relations

- Case of **crisp functions** as in the classical case
  - **applied to fuzzy inputs**
  - example: surface = width $\times$ length
    width = *about* 10m and length = *about* 7m

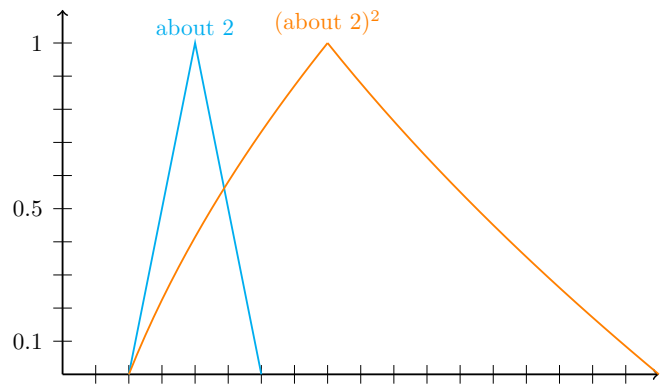
Functional relations

- Case of **crisp functions** as in the classical case
  - applied to fuzzy inputs
  - example: surface = width $\times$ length
    width = *about* 10m and length = *about* 7m

$\Longrightarrow$ **application of the extension principle**

- e.g. binary function $F_Z : Dom(X) \times Dom(Y) \rightarrow Dom(Z)$
- extended to fuzzy inputs described by μ_A and μ_B, $A, B \in V(X)$

$$\mu_Z(z) = \sup\{ \quad \min(\mu_A(x), \mu_B(y)) \mid (x, y) \in Dom(X) \times Dom(Y), z = F_Z(x, y)\}$$



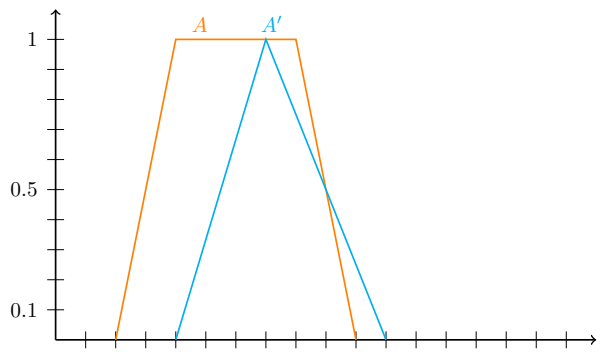
Functional relations

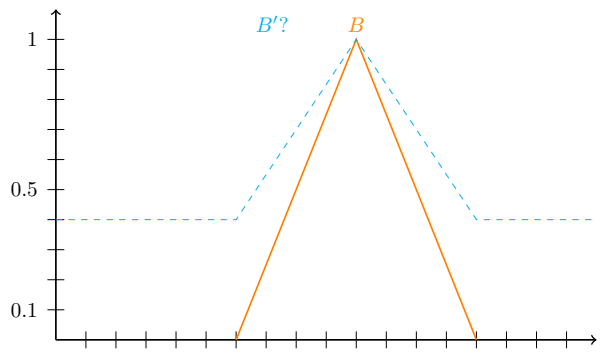
- Case of crisp functions: application of the extension principle
- Case of **general fuzzy functions**
  - if surface is *middleSize*, then price is *high*
  - surface is *around* 70m^2

$\Longrightarrow$ **application of the Generalised Modus Ponens** GMP

- example for a unary function $F_Y : Dom(X) \rightarrow Dom(Y)$ such that $F_Y(A) = B$
- for a fuzzy input A'

$$\mu_{B'}(y) = \sup_{x \in Dom(X)} \top(\mu_{A'}(x), I(\mu_A(x), \mu_B(y)))$$





Fuzzy formulation of causal formulas

- Syntax: $(M, \mathbf{u}) \models \varphi,\ \alpha$
  - $(M, \mathbf{u})$: fuzzy situation
  - φ : logical combination of formulas $\mathbf{X} = \mathbf{x}$ with $x \in V(X)$

- Definition of degree $\alpha \in [0, 1]$
  - by a fuzzy inference algorithm, e.g. the tableau algorithm
  - functional framework $\Longrightarrow$ **using similarity measures**
  - case of atomic formulas:

$$(M, \mathbf{u}) \models X = x,\ \alpha \quad \Longleftrightarrow \quad \alpha = sim(x, F_X(\mathbf{u}))$$

Fuzzy formulation of causal formulas

- Atomic formulas

$$(M, \mathbf{u}) \models X = x,\ \alpha \quad \Longleftrightarrow \quad \alpha = sim(x, F_X(\mathbf{u}))$$

- Composed formulas
  - using fuzzy operators associated to the logical connectives

- Causal formulas

$$(M, \mathbf{u}) \models [\mathbf{Y} = \mathbf{y}]\varphi,\ \alpha \quad \Longleftrightarrow \quad (M_{\mathbf{Y}=\mathbf{y}}, \mathbf{u}) \models \varphi,\ \alpha$$

Fuzzy cause

- $\mathbf{X} = \mathbf{x}$ cause of φ in $(M, \mathbf{u})$ with degree $\alpha = \top(\alpha_1, \alpha_2, \alpha_3)$
  - α_1: satisfaction degree of $\mathbf{X} = \mathbf{x}$ and φ in situation $(M, \mathbf{u})$: $(M, \mathbf{u}) \models (\mathbf{X} = \mathbf{x}) \wedge \varphi,\ \alpha_1$
  - α_2: degree with which $\neg\varphi$ would be satisfied for any $\mathbf{x}' \neq \mathbf{x}$
    - $(M, \mathbf{u}) \models (\mathbf{X} = \mathbf{x}') \wedge \neg\varphi,\ \alpha'$
    - $\alpha_2 = \sup_{\mathbf{x}' \in V(\mathbf{X})} \alpha'$
  - α_3: minimality condition

Minimality discussion

- Crisp case: inclusion of variable sets
  - $\mathbf{X} = \mathbf{x} = \wedge_i (X_i = x_i)$

$$\mathbf{X} = \mathbf{x} \subset \mathbf{Y} = \mathbf{y} \qquad \text{iff} \qquad \{X_i\} \subset \{Y_i\}$$

- Fuzzy case: inclusion of values $x_i \subseteq y_i$
  - $\mu_{x_i}(z) \leq \mu_{y_i}(z)$
  - $(w = \text{about 10m}) \subseteq (w = \text{big})$
- Tradeoff with satisfiability: dominance notion
  - $E_1 = \mathbf{X}_1 = \mathbf{x}_1, \alpha_1$
  - $E_2 = \mathbf{X}_2 = \mathbf{x}_2, \alpha_2$
  - what if $E_1 \subset E_2$ but $\alpha_1 < \alpha_2$?

Fuzzy contrastive cause

- **Fuzzy contrastive cause of φ as opposed to ψ in $(M, \mathbf{u})$**

  pair $(\mathbf{X} = \mathbf{x}, \mathbf{X} = \mathbf{y}), \alpha$ where $\alpha = agg(\beta_0, \alpha, \beta, \alpha_{int})$ with
  - $(M, \mathbf{u}) \models \neg\psi, \beta_0$
  - $\mathbf{X} = \mathbf{x}$ cause of φ with degree α
  - $\mathbf{X} = \mathbf{y}$ cause of ψ in $(M', \mathbf{u})$ with degree β
  - minimal intersection degree α_{int} of $(\mathbf{X} = \mathbf{x}, \mathbf{X} = \mathbf{y})$

Conclusion

- **Fuzzy semantics for contrastive causes**
  - integrating imprecise observations or knowledge
  - using classical fuzzy logic tools
- Directions for future works
  - application to toy data and real data
  - study of the properties according to the aggregation operators and similarity measures
  - case of bifactual causes: changing the considered situation
  - case of uncertain information: possibilistic semantics