Probabilistic Team Semantics

Jonni Virtema

Hasselt University, Belgium
jonni.virtema@gmail.com

Joint work with Arnaud Durand (Université Paris Diderot), Miika Hannula (University of Helsinki), Juha Kontinen (University of Helsinki), and Arne Meier (Leibniz Universität Hannover)

August 3, 2018

Teams as collections of measurements

- Multiteams (multisets of assignments) vs.

	x	y	z
s_1	a	a	b
s_2	a	a	b
s_3	b	c	c
s_4	a	b	c

	x	y	z	#
s_1	a	a	b	2
s_2	b	c	c	1
s_3	a	b	c	1

	x	y	z	prob.
s_1	a	a	b	$\frac{1}{2}$
s_2	b	c	c	$\frac{1}{4}$
s_3	a	b	c	$\frac{1}{4}$

Teams as collections of measurements

- Multiteams (multisets of assignments) vs. probabilistic teams (distributions over assignments)

	x	y	z
s_1	a	a	b
s_2	a	a	b
s_3	b	c	c
s_4	a	b	c

	x	y	z	#
s_1	a	a	b	2
s_2	b	c	c	1
s_3	a	b	c	1

	x	y	z	prob.
s_1	a	a	b	$\frac{1}{2}$
s_2	b	c	c	$\frac{1}{4}$
s_3	a	b	c	$\frac{1}{4}$

Distributions of data

Consider:

- A collection of data from some repetitive science experiment.
- Data obtained from a poll.
- Any collection of data, that involves meaningful duplicates of data.

One natural way to represent the data is to use multisets (sets with duplicates).

Claim:

Often the multiplicities themselves are not important; the distribution of data is.

Distributions of data

Consider:

- A collection of data from some repetitive science experiment.
- Data obtained from a poll.
- Any collection of data, that involves meaningful duplicates of data.

One natural way to represent the data is to use multisets (sets with duplicates).

Claim:
Often the multiplicities themselves are not important; the distribution of data is.

Distributions

Definition

A distribution is a mapping $f : A \to \mathbb{Q}_{[0,1]}$ from a set A of values to the closed interval $[0, 1]$ of rational numbers such that the probabilities sum to 1, i.e.,

$$\sum_{a \in A} f(a) = 1.$$

- A multiteam is a pair (X, m), where X is a set of assignments and $m : X \to \mathbb{N}^{>0}$ is a multiplicity function (a database with duplicates).
- A probabilistic team is a pair (X, p), where X is a set of assignments and $p : X \to \mathbb{Q}_{[0,1]}$ is a distribution (distribution of data).

Distributions

Definition

A distribution is a mapping $f : A \to \mathbb{Q}_{[0,1]}$ from a set A of values to the closed interval $[0, 1]$ of rational numbers such that the probabilities sum to 1, i.e.,

$$\sum_{a \in A} f(a) = 1.$$

- A multiteam is a pair (X, m), where X is a set of assignments and $m : X \to \mathbb{N}^{>0}$ is a multiplicity function (a database with duplicates).
- A probabilistic team is a pair (X, p), where X is a set of assignments and $p : X \to \mathbb{Q}_{[0,1]}$ is a distribution (distribution of data).

Probabilistic teams

- Modelling of data that is inherently a probability distribution.
- Abstraction of data with duplicates.
- There is close connection between multiteams and probabilistic teams.

We introduce a logic that describe properties of probabilistic teams.

We consider the expansion of first-order logic with the marginal identity atoms

$$(x_1, \ldots, x_n) \approx (y_1, \ldots, y_n)$$

and with the probabilistic conditional independence atoms

$$\overline{y} \perp\!\!\!\perp_{\overline{x}} \overline{z}.$$

Probabilistic teams

- Modelling of data that is inherently a probability distribution.
- Abstraction of data with duplicates.
- There is close connection between multiteams and probabilistic teams.

We introduce a logic that describe properties of probabilistic teams.

We consider the expansion of first-order logic with the marginal identity atoms

$$(x_1, \ldots, x_n) \approx (y_1, \ldots, y_n)$$

and with the probabilistic conditional independence atoms

$$\overline{y} \perp\!\!\!\perp_{\overline{x}} \overline{z}.$$

Probabilistic atoms

We define that

$$\mathfrak{A} \models_{\mathbb{X}} \vec{x} \approx \vec{y} \text{ iff the distribution of values for } \vec{x} \text{ and } \vec{y} \text{ in } \mathbb{X} \text{ coincide.}$$

Probabilistic atoms

We define that

$$\mathfrak{A} \models_{\mathbb{X}} \vec{x} \approx \vec{y} \text{ iff the distribution of values for } \vec{x} \text{ and } \vec{y} \text{ in } \mathbb{X} \text{ coincide.}$$

We define that

$$\mathfrak{A} \models_{\mathbb{X}} \overline{y} \perp\!\!\!\perp_{\overline{x}} \overline{z} \text{ iff for every fixed value for } \vec{x},$$

the value distribution of $\vec{y}$ remains unchanged if any value for $\vec{z}$ is given.

Probabilistic atoms

Let $\mathbb{X} = (X, p)$ be a probablistic team and $\vec{x}, \vec{a}$ be tuples of variables and values of length k. We define

$$|\mathbb{X}|_{\vec{x}=\vec{a}} := \sum_{\substack{s \in X \\ s(\vec{x})=\vec{a}}} p(s).$$

We define that

$$\mathfrak{A} \models_{\mathbb{X}} \vec{x} \approx \vec{y} \text{ iff } \; |\mathbb{X}|_{\vec{x}=\vec{a}} = |\mathbb{X}|_{\vec{y}=\vec{a}}, \text{ for each } \vec{a} \in A^k.$$

We define that

$$\mathfrak{A} \models_{\mathbb{X}} \overline{y} \perp\!\!\!\perp_{\overline{x}} \overline{z} \text{ iff for all assignments } s \text{ for } \vec{x}, \vec{y}, \vec{z}$$

$$|\mathbb{X}|_{\vec{x}\vec{y}=s(\vec{x}\vec{y})} \times |\mathbb{X}|_{\vec{x}\vec{z}=s(\vec{x}\vec{z})} = |\mathbb{X}|_{\vec{x}\vec{y}\vec{z}=s(\vec{x}\vec{y}\vec{z})} \times |\mathbb{X}|_{\vec{x}=s(\vec{x})}.$$

Semantics of complex formulae

Definition

Let $\mathfrak{A}$ be a structure over a finite domain A, and $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$ a probabilistic team of $\mathfrak{A}$. The satisfaction relation $\models_{\mathbb{X}}$ for first-order logic is defined as follows:

$$\mathfrak{A} \models_{\mathbb{X}} x = y \Leftrightarrow \text{for all } s \in X : \text{if } \mathbb{X}(s) > 0, \text{ then } s(x) = s(y)$$

$$\mathfrak{A} \models_{\mathbb{X}} x \neq y \Leftrightarrow \text{for all } s \in X : \text{if } \mathbb{X}(s) > 0, \text{ then } s(x) \neq s(y)$$

$$\mathfrak{A} \models_{\mathbb{X}} R(\overline{x}) \Leftrightarrow \text{for all } s \in X : \text{if } \mathbb{X}(s) > 0, \text{ then } s(\overline{x}) \in R^{\mathfrak{A}}$$

$$\mathfrak{A} \models_{\mathbb{X}} \neg R(\overline{x}) \Leftrightarrow \text{for all } s \in X : \text{if } \mathbb{X}(s) > 0, \text{ then } s(\overline{x}) \notin R^{\mathfrak{A}}$$

$$\mathfrak{A} \models_{\mathbb{X}} (\psi \wedge \theta) \Leftrightarrow \mathfrak{A} \models_{\mathbb{X}} \psi \text{ and } \mathfrak{A} \models_{\mathbb{X}} \theta$$

Semantics of complex formulae

Definition

Let $\mathfrak{A}$ be a structure over a finite domain A, and $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$ a probabilistic team of $\mathfrak{A}$. The satisfaction relation $\models_{\mathbb{X}}$ for first-order logic is defined as follows:

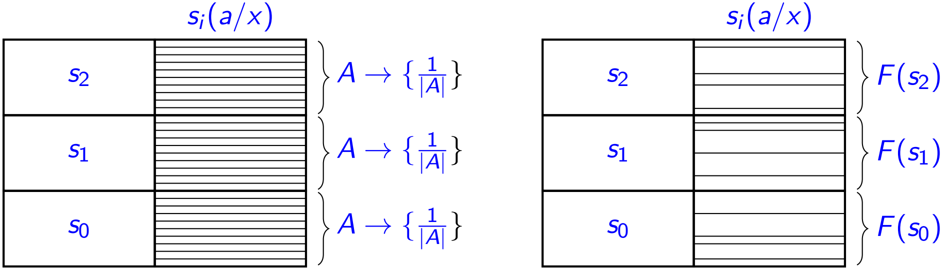
$$\mathfrak{A} \models_{\mathbb{X}} (\psi \vee \theta) \Leftrightarrow \mathfrak{A} \models_{\mathbb{Y}} \psi \text{ and } \mathfrak{A} \models_{\mathbb{Z}} \theta \text{ for some } \mathbb{Y}, \mathbb{Z} \; s.t. \; \mathbb{Y} \sqcup \mathbb{Z} = \mathbb{X}$$

$$\mathfrak{A} \models_{\mathbb{X}} \forall x \psi \Leftrightarrow \mathfrak{A} \models_{\mathbb{X}[A/x]} \psi$$

$$\mathfrak{A} \models_{\mathbb{X}} \exists x \psi \Leftrightarrow \mathfrak{A} \models_{\mathbb{X}[F/x]} \psi \text{ holds for some } F\colon X \to p_A.$$

Above p_A denote the set those distributions that have domain A.

Intuition of the quantifiers



- Universal quantification (i.e., the set $\mathbb{X}[A/x]$) is depicted on left.
- Existential quantification (i.e., the set $\mathbb{X}[F/x]$) is depicted on right.
- Height of a box corresponds to the probability of an assignment.

Intuition behind the disjunction

Question: How do we split distributions?
Answer: We rescale.

Let $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$ and $\mathbb{Y}\colon Y \to \mathbb{Q}_{[0,1]}$ be probabilistic teams and $k \in \mathbb{Q}_{[0,1]}$ be a rational number.

We denote by $\mathbb{X} \sqcup_k \mathbb{Y}$ the k-scaled union of $\mathbb{X}$ and $\mathbb{Y}$, that is, the probabilistic team $\mathbb{X} \sqcup_k \mathbb{Y}\colon X \cup Y \to \mathbb{Q}_{[0,1]}$ defined s.t. for each $s \in X \cup Y$,

$$(\mathbb{X} \sqcup_k \mathbb{Y})(s) := \begin{cases} k \cdot \mathbb{X}(s) + (1-k) \cdot \mathbb{Y}(s) & \text{if } s \in X \text{ and } s \in Y, \\ k \cdot \mathbb{X}(s) & \text{if } s \in X \text{ and } s \notin Y, \\ (1-k) \cdot \mathbb{Y}(s) & \text{if } s \in Y \text{ and } s \notin X. \end{cases}$$

We then write that $Z = \mathbb{X} \sqcup \mathbb{Y}$ if $Z = \mathbb{X} \sqcup_k \mathbb{Y}$, for some k.

Intuition behind the disjunction

Question: How do we split distributions?
Answer: We rescale.

Let $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$ and $\mathbb{Y}\colon Y \to \mathbb{Q}_{[0,1]}$ be probabilistic teams and $k \in \mathbb{Q}_{[0,1]}$ be a rational number.

We denote by $\mathbb{X} \sqcup_k \mathbb{Y}$ the k-scaled union of $\mathbb{X}$ and $\mathbb{Y}$, that is, the probabilistic team $\mathbb{X} \sqcup_k \mathbb{Y}\colon X \cup Y \to \mathbb{Q}_{[0,1]}$ defined s.t. for each $s \in X \cup Y$,

$$(\mathbb{X} \sqcup_k \mathbb{Y})(s) := \begin{cases} k \cdot \mathbb{X}(s) + (1-k) \cdot \mathbb{Y}(s) & \text{if } s \in X \text{ and } s \in Y, \\ k \cdot \mathbb{X}(s) & \text{if } s \in X \text{ and } s \notin Y, \\ (1-k) \cdot \mathbb{Y}(s) & \text{if } s \in Y \text{ and } s \notin X. \end{cases}$$

We then write that $Z = \mathbb{X} \sqcup \mathbb{Y}$ if $Z = \mathbb{X} \sqcup_k \mathbb{Y}$, for some k.

Intuition behind the disjunction

Question: How do we split distributions?
Answer: We rescale.

Let $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$ and $\mathbb{Y}\colon Y \to \mathbb{Q}_{[0,1]}$ be probabilistic teams and $k \in \mathbb{Q}_{[0,1]}$ be a rational number.

We denote by $\mathbb{X} \sqcup_k \mathbb{Y}$ the k-scaled union of $\mathbb{X}$ and $\mathbb{Y}$, that is, the probabilistic team $\mathbb{X} \sqcup_k \mathbb{Y}\colon X \cup Y \to \mathbb{Q}_{[0,1]}$ defined s.t. for each $s \in X \cup Y$,

$$(\mathbb{X} \sqcup_k \mathbb{Y})(s) := \begin{cases} k \cdot \mathbb{X}(s) + (1-k) \cdot \mathbb{Y}(s) & \text{if } s \in X \text{ and } s \in Y, \\ k \cdot \mathbb{X}(s) & \text{if } s \in X \text{ and } s \notin Y, \\ (1-k) \cdot \mathbb{Y}(s) & \text{if } s \in Y \text{ and } s \notin X. \end{cases}$$

We then write that $Z = \mathbb{X} \sqcup \mathbb{Y}$ if $Z = \mathbb{X} \sqcup_k \mathbb{Y}$, for some k.

Example

Consider a database table that lists results of experiments as a multiteam or as the related probabilistic team using the counting measure.

- Records: Outcomes of measurements obtained simultaneously in two locations.
- Attributes: Test1 and Test2 ranging over types of measurements, and Outcome1 and Outcome2 ranging over outcomes of the measurements.

The probabilistic independence atom $\mathrm{Test1} \perp\!\!\!\perp \mathrm{Test2}$ expresses that the types of measurements are independently picked in the two locations.

The marginal identity atom $(\mathrm{Test1}, \mathrm{Outcome1}) \approx (\mathrm{Test2}, \mathrm{Outcome2})$ expresses that the distributions of tests and results are the same in both test sites.

The formula $\mathrm{Test1} = \mathrm{Test2} \vee (\mathrm{Test1} \neq \mathrm{Test2} \wedge \mathrm{Outcome1} \perp\!\!\!\perp \mathrm{Outcome2})$ expresses that there is no correlation between outcomes of different measurements in the two test sites.

More examples

- The formula $\forall \vec{y}\, \vec{x} \approx \vec{y}$ states that the probabilities for $\overline{x}$ are uniformly distributed over all value sequences of length $|\overline{x}|$.
- The probability of $P(x)$ is at least twice the probability of $Q(x)$.
- Can we characterise the expressive power of $\mathrm{FO}(\approx, \perp\!\!\!\perp)$ in the probabilistic setting?

More examples

- The formula $\forall \vec{y}\, \vec{x} \approx \vec{y}$ states that the probabilities for $\overline{x}$ are uniformly distributed over all value sequences of length $|\overline{x}|$.
- The probability of $P(x)$ is at least twice the probability of $Q(x)$.
- Can we characterise the expressive power of $\mathrm{FO}(\approx, \perp\!\!\!\perp)$ in the probabilistic setting?

Benchmark logic

- In team semantics context fragments of second-order logic are captured.
- $\mathrm{FO}(\perp)$ (team semantics) is as expressive as existential second-order logic.
- We define a two-sorted variant of ESO in which we allow the quantification of rational distributions.
- This logic characterises the expressive power of $\mathrm{FO}(\approx, \perp\!\!\!\perp)$.

Benchmark logic

- In team semantics context fragments of second-order logic are captured.
- $\mathrm{FO}(\perp)$ (team semantics) is as expressive as existential second-order logic.
- We define a two-sorted variant of ESO in which we allow the quantification of rational distributions.
- This logic characterises the expressive power of $\mathrm{FO}(\approx, \perp\!\!\!\perp)$.

Probabilistic structures

Definition

Let τ and σ be a relational and a functional vocabulary. A probabilistic $\tau \cup \sigma$-structure is a tuple

$$\mathfrak{A} = (A, \mathbb{Q}_{[0,1]}, (R_i^{\mathfrak{A}})_{R_i \in \tau}, (f_i^{\mathfrak{A}})_{f_i \in \sigma}),$$

where

- A (i.e. the domain of $\mathfrak{A}$) is a finite nonempty set,
- $\mathbb{Q}_{[0,1]}$ is the set of rational numbers in the closed interval $[0, 1]$,
- each $R_i^{\mathfrak{A}}$ is a relation on A (i.e., a subset of $A^{\mathrm{ar}(R_i)}$),
- each $f_i^{\mathfrak{A}}$ is a probability distribution from $A^{\mathrm{ar}(f_i)}$ to $\mathbb{Q}_{[0,1]}$ (i.e., a function such that $\sum_{\vec{a} \in A^{\mathrm{ar}(f_i)}} f_i(\vec{a}) = 1$).

Second-order logic for probabilistic structures

- As first-order terms we have first-order variables.
- The set of numerical σ-terms i is defined via the grammar

$$i ::= f(\vec{x}) \mid i \times i \mid \mathrm{SUM}_{\vec{x}}\, i(\vec{x}, \vec{y}),$$

  where $\vec{x}, \vec{y}$ are tuples of first-order variables, $f \in \sigma$ and σ is a set of functions.
- The value of a numerical term i in a structure $\mathfrak{A}$ under an assignment s is denoted by $[i]_s^{\mathfrak{A}}$ and defined as follows:

$$[f(\overline{x})]_s^{\mathfrak{A}} := f^{\mathfrak{A}}(s(\overline{x})), \qquad [i \times j]_s^{\mathfrak{A}} := [i]_s^{\mathfrak{A}} \cdot [j]_s^{\mathfrak{A}},$$

$$[\mathrm{SUM}_{\vec{x}}\, i(\vec{x}, \vec{y})]_s^{\mathfrak{A}} := \sum_{\vec{a} \in A^{|\vec{x}|}} [i(\vec{a}, \vec{y})]_s^{\mathfrak{A}},$$

  where $\cdot$ and $\sum$ are the multiplication and sum of rational numbers.

Second-order logic for probabilistic structures

- As first-order terms we have first-order variables.
- The set of numerical σ-terms i is defined via the grammar

$$i ::= f(\vec{x}) \mid i \times i \mid \mathrm{SUM}_{\vec{x}}\, i(\vec{x},\vec{y}),$$

  where $\vec{x},\vec{y}$ are tuples of first-order variables, $f \in \sigma$ and σ is a set of functions.
- The value of a numerical term i in a structure $\mathfrak{A}$ under an assignment s is denoted by $[i]_s^{\mathfrak{A}}$ and defined as follows:

$$[f(\overline{x})]_s^{\mathfrak{A}} := f^{\mathfrak{A}}(s(\overline{x})), \qquad [i \times j]_s^{\mathfrak{A}} := [i]_s^{\mathfrak{A}} \cdot [j]_s^{\mathfrak{A}},$$

$$[\mathrm{SUM}_{\vec{x}}\, i(\vec{x},\vec{y})]_s^{\mathfrak{A}} := \sum_{\vec{a} \in A^{|\vec{x}|}} [i(\vec{a},\vec{y})]_s^{\mathfrak{A}},$$

  where $\cdot$ and $\sum$ are the multiplication and sum of rational numbers.

Second-order logic for probabilistic structures

Definition

The formulae of ESOf$_{\mathbb{Q}}$ is defined via the following grammar:

$$\phi ::= x = y \mid x \neq y \mid i = j \mid i \neq j \mid R(\vec{x}) \mid \neg R(\vec{x}) \mid \phi \wedge \phi \mid \phi \vee \phi \mid \exists x \phi \mid \forall x \phi \mid \exists f \phi,$$

where i is a numerical term, R is a relation symbol, f is a function variable, $\vec{x}$ is a tuple of first-order variables.

Semantics of ESOf$_{\mathbb{Q}}$ is defined via probabilistic structures and assignments analogous to FO. In addition to the clauses of first-order logic, we have:

$$\mathfrak{A} \models_s i = j \Leftrightarrow [i]^{\mathfrak{A}}_s = [j]^{\mathfrak{A}}_s, \qquad \mathfrak{A} \models_s i \neq j \Leftrightarrow [i]^{\mathfrak{A}}_s \neq [j]^{\mathfrak{A}}_s,$$

$\mathfrak{A} \models_s \exists f \phi \Leftrightarrow \mathfrak{A}[h/f] \models_s \phi$ for some probability distribution $h \colon A^{\mathrm{ar}(f)} \to \mathbb{Q}_{[0,1]}$,

where $\mathfrak{A}[h/f]$ denotes the expansion of $\mathfrak{A}$ that interprets f to h.

Second-order logic for probabilistic structures

Definition

The formulae of $\mathrm{ESOf}_{\mathbb{Q}}$ is defined via the following grammar:

$$\phi ::= x = y \mid x \neq y \mid i = j \mid i \neq j \mid R(\vec{x}) \mid \neg R(\vec{x}) \mid \phi \wedge \phi \mid \phi \vee \phi \mid \exists x \phi \mid \forall x \phi \mid \exists f \phi,$$

where i is a numerical term, R is a relation symbol, f is a function variable, $\vec{x}$ is a tuple of first-order variables.

Semantics of $\mathrm{ESOf}_{\mathbb{Q}}$ is defined via probabilistic structures and assignments analogous to FO. In addition to the clauses of first-order logic, we have:

$$\mathfrak{A} \models_s i = j \Leftrightarrow [i]^{\mathfrak{A}}_s = [j]^{\mathfrak{A}}_s, \qquad \mathfrak{A} \models_s i \neq j \Leftrightarrow [i]^{\mathfrak{A}}_s \neq [j]^{\mathfrak{A}}_s,$$

$$\mathfrak{A} \models_s \exists f \phi \Leftrightarrow \mathfrak{A}[h/f] \models_s \phi \text{ for some probability distribution } h \colon A^{\mathrm{ar}(f)} \to \mathbb{Q}_{[0,1]},$$

where $\mathfrak{A}[h/f]$ denotes the expansion of $\mathfrak{A}$ that interprets f to h.

Translating from $\mathrm{FO}(\perp\!\!\!\perp, \approx)$ to $\mathrm{ESOf}_{\mathbb{Q}}$

For a probabilistic team $\mathbb{X}\colon X \to \mathbb{Q}_{[0,1]}$, we let $f_{\mathbb{X}} : A^n \to \mathbb{Q}_{[0,1]}$ be the probability distribution such that $f_{\mathbb{X}}(s(\overline{x})) = \mathbb{X}(s)$ for all $s \in X$.

Theorem

For every $\phi(\overline{x}) \in \mathrm{FO}(\perp\!\!\!\perp, \approx)$ there is a formula $\phi^*(f) \in \mathrm{ESOf}_{\mathbb{Q}}$ with one free function variable f s.t. for all structures $\mathfrak{A}$ and nonempty probabilistic teams $\mathbb{X}$

$$\mathfrak{A} \models_{\mathbb{X}} \phi(\overline{x}) \iff (\mathfrak{A}, f_{\mathbb{X}}) \models \phi^*(f)$$

and vice versa.

The proof utilises the observation that independence atoms and marginal identity atoms can be used to express multiplication and SUM in $\mathbb{Q}_{[0,1]}$.

Translating from $\mathrm{FO}(\perp\!\!\!\perp, \approx)$ to $\mathrm{ESOf}_{\mathbb{Q}}$

For a probabilistic team $\mathbb{X} \colon X \to \mathbb{Q}_{[0,1]}$, we let $f_{\mathbb{X}} : A^n \to \mathbb{Q}_{[0,1]}$ be the probability distribution such that $f_{\mathbb{X}}(s(\overline{x})) = \mathbb{X}(s)$ for all $s \in X$.

Theorem

For every $\phi(\overline{x}) \in \mathrm{FO}(\perp\!\!\!\perp, \approx)$ there is a formula $\phi^*(f) \in \mathrm{ESOf}_{\mathbb{Q}}$ with one free function variable f s.t. for all structures $\mathfrak{A}$ and nonempty probabilistic teams $\mathbb{X}$

$$\mathfrak{A} \models_{\mathbb{X}} \phi(\overline{x}) \iff (\mathfrak{A}, f_{\mathbb{X}}) \models \phi^*(f)$$

and vice versa.

The proof utilises the observation that independence atoms and marginal identity atoms can be used to express multiplication and SUM in $\mathbb{Q}_{[0,1]}$.