Probability Logic & Inductive Learning

Radboud University Nijmegen

Rutger Kuyper

30 June 2013

A primer on first-order logic

A **model** is some (possibly infinite) collection of objects, together with a description of some of their basic properties.

Formulas describe properties of the model. They are built using:

- The basic properties;
- **Connectives**: $\wedge$ (and), $\vee$ (or), $\rightarrow$ (implication) and $\neg$ (negation);
- **Quantifiers**: $\forall$ (for all) and $\exists$ (exists).

$\forall x \varphi(x)$: for all objects in the model we are studying, property φ holds.

Can first-order formulas be learned?

Can we choose between φ and $\neg\varphi$, using just a finite number of observations?

Example. Are all swans white? Let

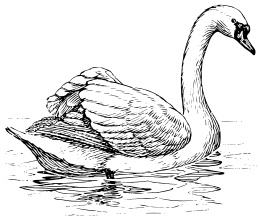
$$\varphi = \text{“All swans are white”}$$
$$\neg\varphi = \text{“Some swan is not white”}$$

Can first-order formulas be learned?

Can we choose between φ and $\neg\varphi$, using just a finite number of observations?

Example. Are all swans white? Let

$$\varphi = \text{“All swans are white”}$$
$$\neg\varphi = \text{“Some swan is not white”}$$

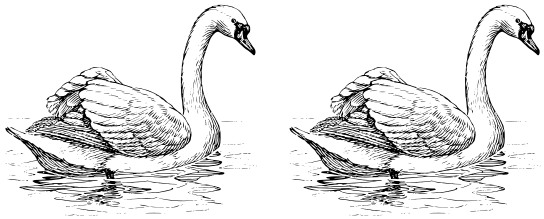


Can first-order formulas be learned?

Can we choose between φ and $\neg\varphi$, using just a finite number of observations?

Example. Are all swans white? Let

$$\varphi = \text{“All swans are white”}$$
$$\neg\varphi = \text{“Some swan is not white”}$$

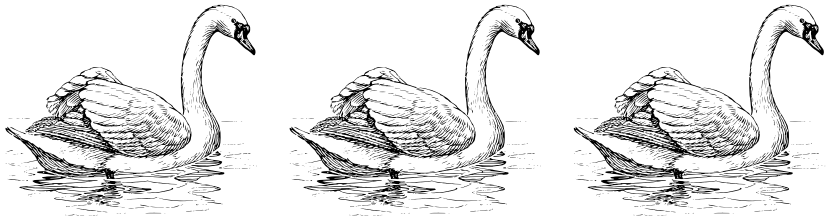


Can first-order formulas be learned?

Can we choose between φ and $\neg\varphi$, using just a finite number of observations?

Example. Are all swans white? Let

$$\varphi = \text{“All swans are white”}$$
$$\neg\varphi = \text{“Some swan is not white”}$$

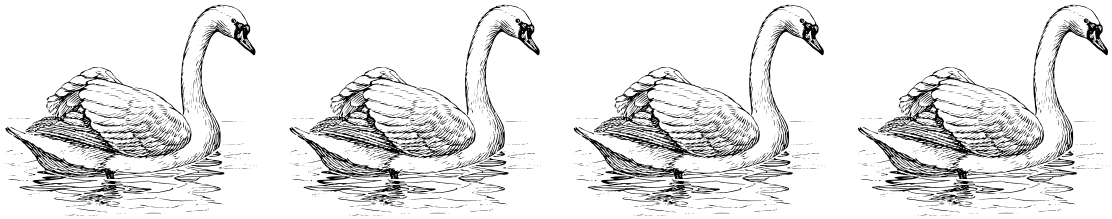


Can first-order formulas be learned?

Can we choose between φ and $\neg\varphi$, using just a finite number of observations?

Example. Are all swans white? Let

$$\varphi = \text{“All swans are white”}$$
$$\neg\varphi = \text{“Some swan is not white”}$$

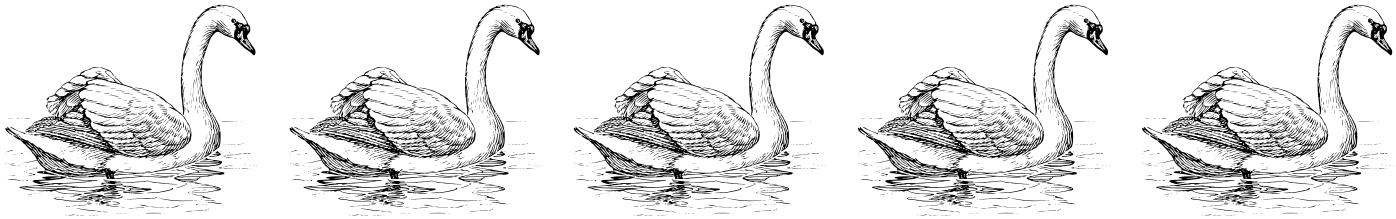


Can first-order formulas be learned?

Can we choose between φ and $\neg\varphi$, using just a finite number of observations?

Example. Are all swans white? Let

$$\varphi = \text{“All swans are white”}$$
$$\neg\varphi = \text{“Some swan is not white”}$$

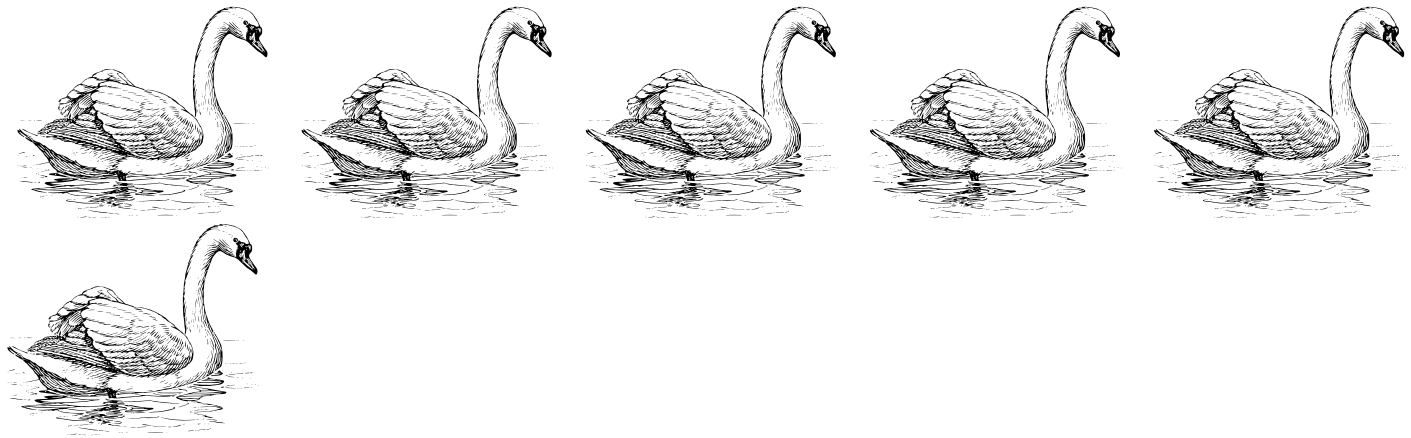


Can first-order formulas be learned?

Can we choose between φ and $\neg\varphi$, using just a finite number of observations?

Example. Are all swans white? Let

$$\varphi = \text{“All swans are white”}$$
$$\neg\varphi = \text{“Some swan is not white”}$$

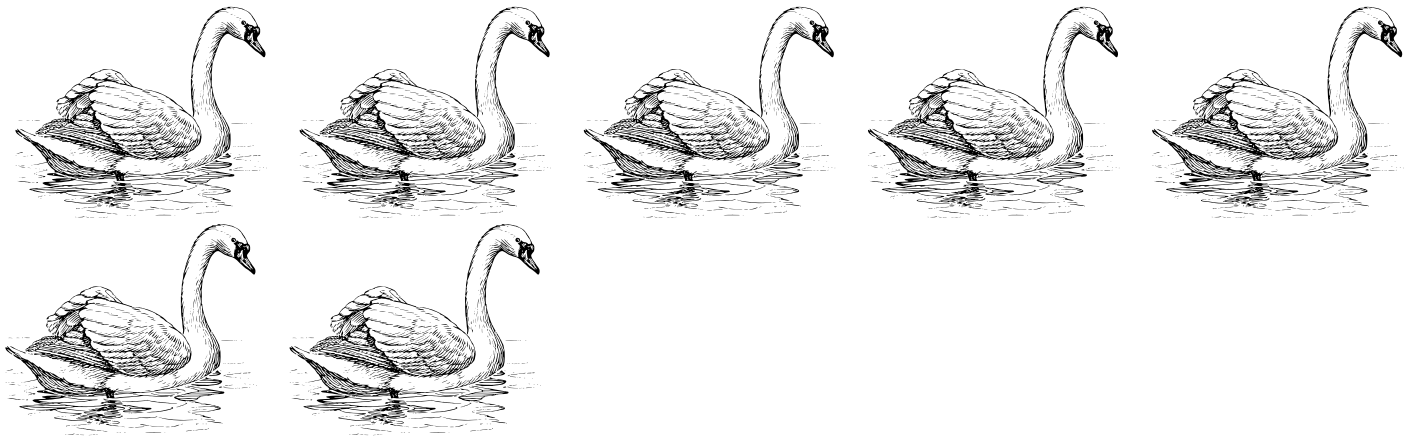


Can first-order formulas be learned?

Can we choose between φ and $\neg\varphi$, using just a finite number of observations?

Example. Are all swans white? Let

$$\varphi = \text{“All swans are white”}$$
$$\neg\varphi = \text{“Some swan is not white”}$$

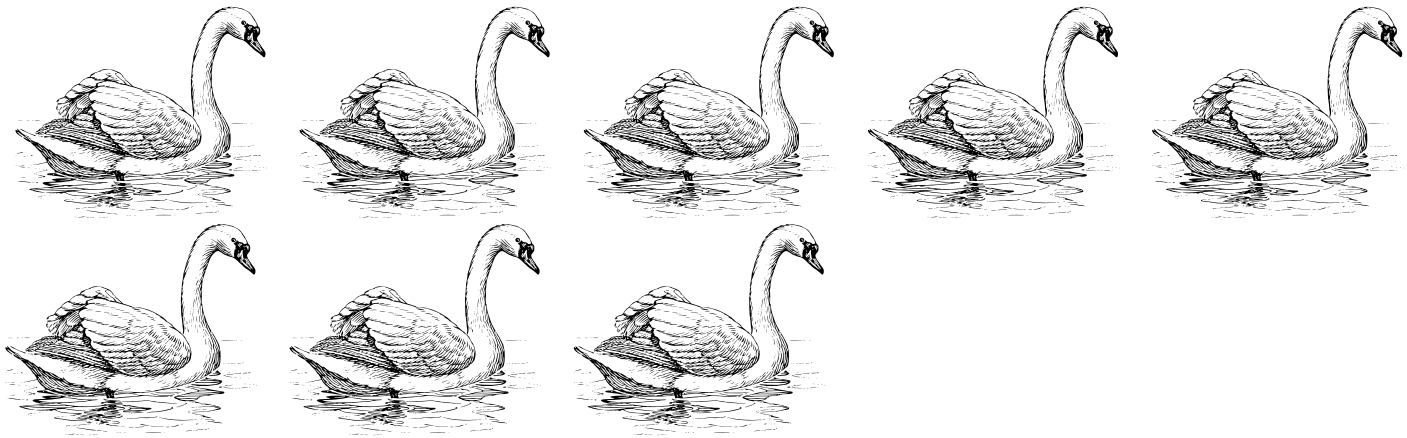


Can first-order formulas be learned?

Can we choose between φ and $\neg\varphi$, using just a finite number of observations?

Example. Are all swans white? Let

$$\varphi = \text{“All swans are white”}$$
$$\neg\varphi = \text{“Some swan is not white”}$$

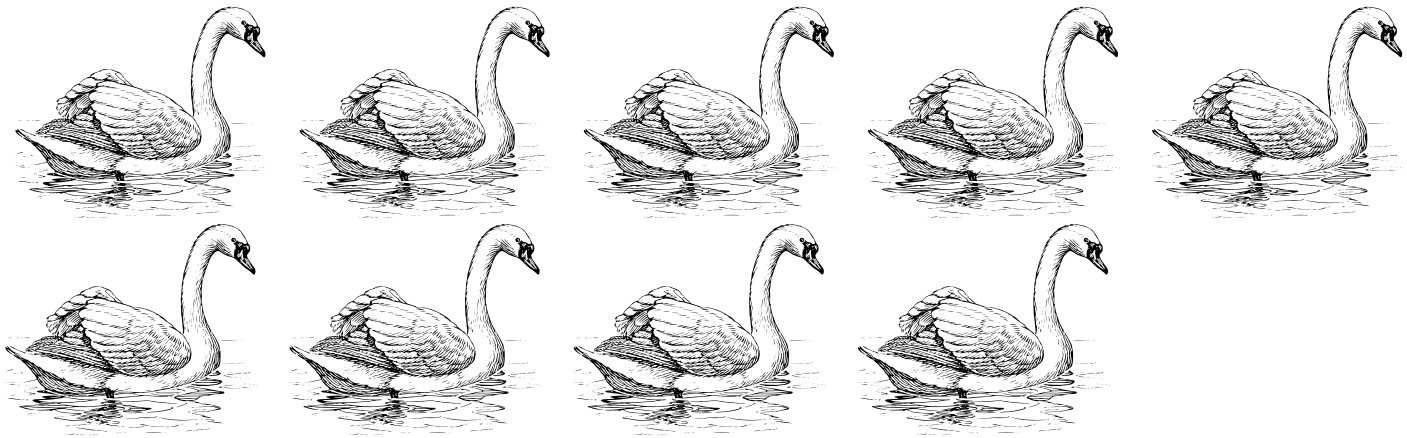


Can first-order formulas be learned?

Can we choose between φ and $\neg\varphi$, using just a finite number of observations?

Example. Are all swans white? Let

$$\varphi = \text{“All swans are white”}$$
$$\neg\varphi = \text{“Some swan is not white”}$$

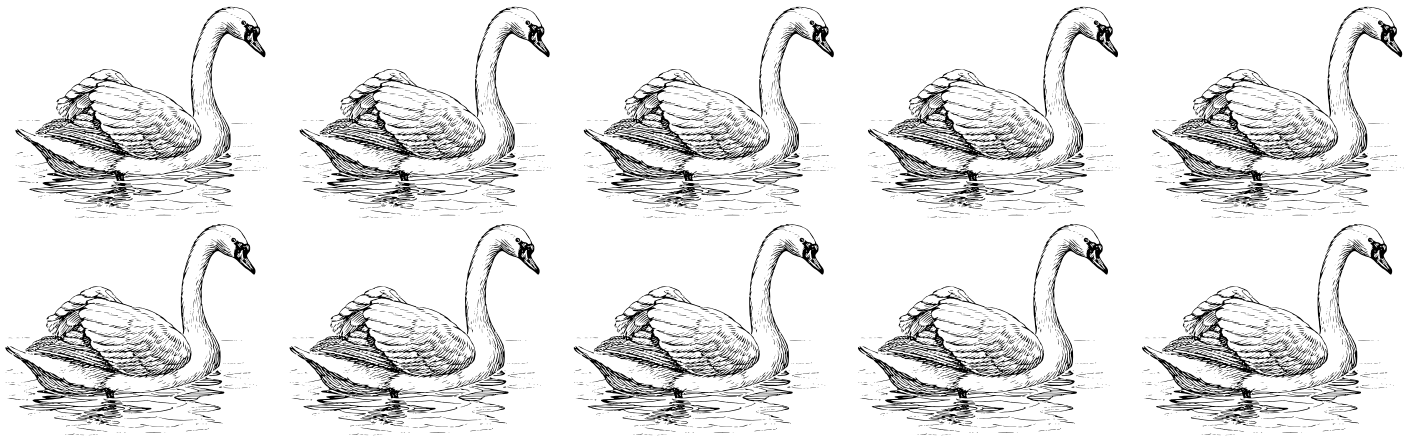


Can first-order formulas be learned?

Can we choose between φ and $\neg\varphi$, using just a finite number of observations?

Example. Are all swans white? Let

$$\varphi = \text{“All swans are white”}$$
$$\neg\varphi = \text{“Some swan is not white”}$$



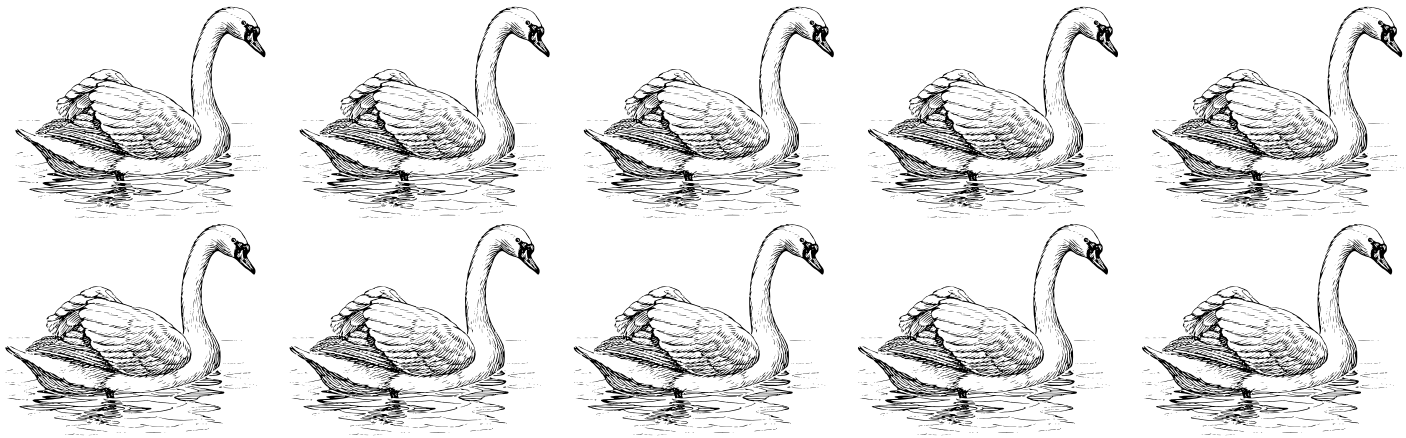
Can first-order formulas be learned?

Can we choose between φ and $\neg\varphi$, using just a finite number of observations?

Example. Are all swans white? Let

$$\varphi = \text{“All swans are white”}$$

$$\neg\varphi = \text{“Some swan is not white”}$$



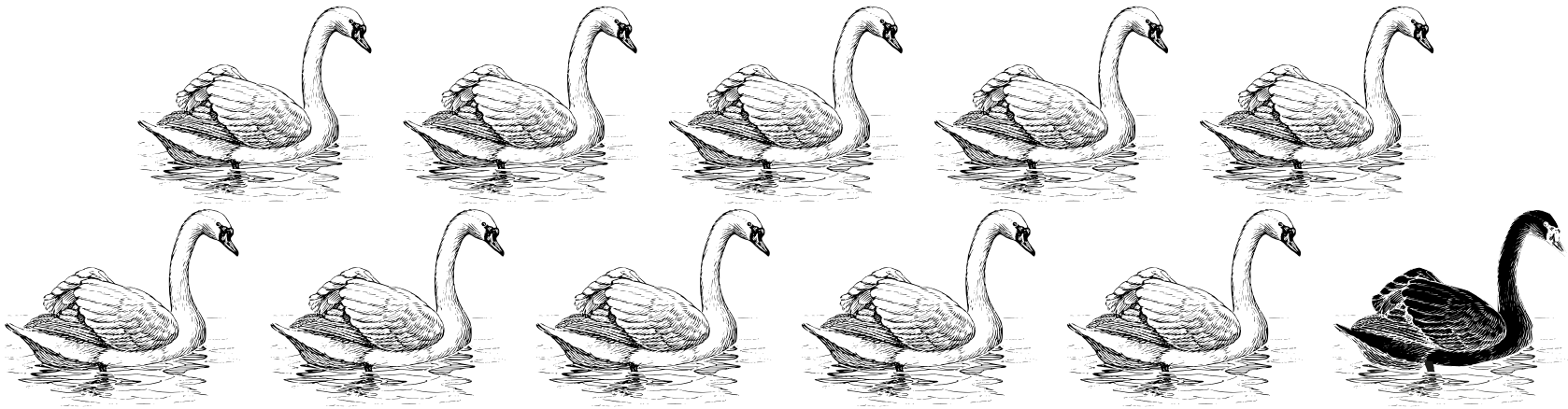
Can first-order formulas be learned?

Can we choose between φ and $\neg\varphi$, using just a finite number of observations?

Example. Are all swans white? Let

$$\varphi = \text{“All swans are white”}$$

$$\neg\varphi = \text{“Some swan is not white”}$$



Can first-order formulas be learned?

More abstractly: given an unknown first-order model $\mathcal{M}$ and a first-order formula φ, can we choose between

$$\mathcal{M} \models \varphi$$

and

$$\mathcal{M} \models \neg\varphi$$

using just a finite sample of atomic truths; i.e. can we choose which of these two options holds if we only know the atomic truths of finitely many elements of $\mathcal{M}$?

Towards a logic for inductive learning

We want to change our interpretation of the quantifiers to make it learnable, while keeping it as classical as possible.

Towards a logic for inductive learning

We want to change our interpretation of the quantifiers to make it learnable, while keeping it as classical as possible.

We have just seen that we cannot say that a universal statement ($\forall$) holds after seeing just a finite number of objects. Is there anything we can say after just a finite number of observations?

Towards a logic for inductive learning

We want to change our interpretation of the quantifiers to make it learnable, while keeping it as classical as possible.

We have just seen that we cannot say that a universal statement ($\forall$) holds after seeing just a finite number of objects. Is there anything we can say after just a finite number of observations?

We can say that **many** swans are white, in the sense that if we randomly pick a swan, then it is white with high probability.

Towards a logic for inductive learning

We want to change our interpretation of the quantifiers to make it learnable, while keeping it as classical as possible.

We have just seen that we cannot say that a universal statement ($\forall$) holds after seeing just a finite number of objects. Is there anything we can say after just a finite number of observations?

We can say that **many** swans are white, in the sense that if we randomly pick a swan, then it is white with high probability.

What about an existential statement ($\exists$)?

Towards a logic for inductive learning

We want to change our interpretation of the quantifiers to make it learnable, while keeping it as classical as possible.

We have just seen that we cannot say that a universal statement ($\forall$) holds after seeing just a finite number of objects. Is there anything we can say after just a finite number of observations?

We can say that **many** swans are white, in the sense that if we randomly pick a swan, then it is white with high probability.

What about an existential statement ($\exists$)?

As soon as we see a single black swan, we know that they exist. Thus, we want the existential quantifier to retain its classical interpretation.

ε-logic

To formalise this, we need to assign probabilities to the objects of the model we are studying. For example, we assign each swan probability $\frac{1}{N}$, where N is the total number of swans. (For the mathematicians: we take any probability measure over the universe of our first-order model.)

ε-logic

Next, we fix an error parameter $\varepsilon \in [0, 1]$. We can now say when the quantifiers ε-hold:

ε-logic

Next, we fix an error parameter $\varepsilon \in [0, 1]$. We can now say when the quantifiers ε-hold:

We interpret the universal quantifier as follows: $\forall x \varphi(x)$ ε-holds if the probability with which $\varphi(x)$ ε-holds is at least $1 - \varepsilon$.

ε-logic

Next, we fix an error parameter $\varepsilon \in [0, 1]$. We can now say when the quantifiers ε-hold:

We interpret the universal quantifier as follows: $\forall x \varphi(x)$ ε-holds if the probability with which $\varphi(x)$ ε-holds is at least $1 - \varepsilon$.

The interpretation of the existential quantifier is classical: $\exists x \varphi(x)$ ε-holds if there exists an a in our model such that $\varphi(a)$ ε-holds.

Paraconsistency

Our logic is **paraconsistent**: a formula φ and its negation $\neg\varphi$ can ε-hold at the same time. Indeed: $\forall x(\mathrm{is\_White}(x))$ will $\frac{1}{10}$-hold, while its negation $\exists x(\neg\mathrm{is\_White}(x))$ will also $\frac{1}{10}$-hold.

Learnability of ε-logic

Theorem. (Terwijn) *ε-logic is indeed learnable for $\varepsilon > 0$, in a way closely related to Valiant's pac-model.*

Related probability logics

Two logics which are related to ours:

1. Keisler's $\mathcal{L}_{\omega P}$ with probability quantifiers $(Px \geq r)$, but not classical $\exists$.
2. Valiant's "Robust Logics".

Assumption

For the rest of this talk, we will assume our first-order language contains **only relations and constants, and no functions or equality**.

Validity, satisfiability and paraconsistency

The ε-validity problem: decide, for a formula φ, if φ ε-holds in every imaginable model. Such formulas exist, for example $\forall x(x = x)$.

The ε-satisfiability problem: decide, for a formula φ, if there exists some model in which φ ε-holds.

Validity, satisfiability and paraconsistency

The ε-validity problem: decide, for a formula φ, if φ ε-holds in every imaginable model. Such formulas exist, for example $\forall x(x = x)$.

The ε-satisfiability problem: decide, for a formula φ, if there exists some model in which φ ε-holds.

(NB for the mathematicians: because of paraconsistency, it is not the case that φ is ε-satisfiable if and only if $\neg\varphi$ is not ε-valid.)

Computational hardness

Validity for classical first-order logic is **undecidable**: there is no computer program that can tell you for any formula φ if it holds in all models.

Computational hardness

Validity for classical first-order logic is **undecidable**: there is no computer program that can tell you for any formula φ if it holds in all models.

However, it is **computably enumerable**: there is a computer program that, if you let it run eternally, keeps outputting formulas, such that:

1. every formula which is valid appears in the list;
2. every formula which appears in the list is valid.

Computational hardness

Validity for classical first-order logic is **undecidable**: there is no computer program that can tell you for any formula φ if it holds in all models.

However, it is **computably enumerable**: there is a computer program that, if you let it run eternally, keeps outputting formulas, such that:

1. every formula which is valid appears in the list;
2. every formula which appears in the list is valid.

Theorem. *For $\varepsilon \in (0,1) \cap \mathbb{Q}$ we have that ε-validity is not computably enumerable (in fact, it is Π^1_1-hard).*

ε-satisfiability

In a certain sense, classical satisfiability is the complement of classical validity: again, it is undecidable, but it is **co-computably enumerable**: there is a computer program that, if you let it run eternally, keeps outputting formulas, such that:

1. every formula which is **not** satisfiable appears in the list;
2. every formula which appears in the list is **not** satisfiable.

ε-satisfiability

In a certain sense, classical satisfiability is the complement of classical validity: again, it is undecidable, but it is **co-computably enumerable**: there is a computer program that, if you let it run eternally, keeps outputting formulas, such that:

1. every formula which is **not** satisfiable appears in the list;
2. every formula which appears in the list is **not** satisfiable.

Theorem. *For $\varepsilon \in (0,1) \cap \mathbb{Q}$ we have that ε-satisfiability is not co-computably enumerable (in fact, it is Σ^1_1-complete).*

0-logic

Theorem. (Terwijn) 0-*validity coincides with classical validity (and hence is computably enumerable).*

Theorem. 0-*satisfiability is decidable.*

A Downward Löwenheim-Skolem theorem

For all ε there exists a sentence φ such that φ is ε-satisfiable in an uncountable model, but not in any countable model.
However, we do have:

Theorem. (Kuyper–Terwijn) *Every ε-model is elementary ε-equivalent to a model of cardinality 2^ω. In other words, every ε-model is elementary ε-equivalent to a model on $[0,1]$.*

Models with Lebesgue measure

Theorem. *Every ε-model is elementary ε-equivalent to a model based on $[0,1]$ with the Lebesgue measure.*

Compactness

Theorem. (Kuyper–Terwijn) *Compactness fails for ε-logic if $\varepsilon \in (0,1) \cap \mathbb{Q}$: there is a set Γ of formulas such that every finite subset has an ε-model, but Γ itself does not have any ε-model.*

Theorem. *Compactness holds for $\varepsilon = 0$.*

A few open questions

- What is the least cardinal κ such that every ε-model is elementary ε-equivalent to an ε-model of cardinality κ? We have seen that $\aleph_1 \leq \kappa \leq 2^{\omega}$.
- Do properties like Craig interpolation, Beth definability and Robinson consistency hold for ε-logic?
- What is the complexity of ε-validity?

Complete definition (1)

1. For every atomic formula φ:

$$(\mathcal{M}, \mathcal{D}) \models_\varepsilon \varphi \text{ if } \mathcal{M} \models \varphi.$$

2. We treat the logical connectives $\wedge$ and $\vee$ classically, e.g.

$$(\mathcal{M}, \mathcal{D}) \models_\varepsilon \varphi \wedge \psi \text{ if } (\mathcal{M}, \mathcal{D}) \models_\varepsilon \varphi \text{ and } (\mathcal{M}, \mathcal{D}) \models_\varepsilon \psi.$$

3. The existential quantifier is treated classically as well:

$$(\mathcal{M}, \mathcal{D}) \models_\varepsilon \exists x \varphi(x)$$

if there exists an $a \in \mathcal{M}$ such that $(\mathcal{M}, \mathcal{D}) \models_\varepsilon \varphi(a)$.

Complete definition (2)

4. The case of negation is split into sub-cases as follows:
   - 4.1 For φ atomic, $(\mathcal{M}, \mathcal{D}) \models_\varepsilon \neg\varphi$ if $(\mathcal{M}, \mathcal{D}) \not\models_\varepsilon \varphi$.
   - 4.2 $\neg$ distributes in the classical way over $\wedge$ and $\vee$, e.g.

$$(\mathcal{M}, \mathcal{D}) \models_\varepsilon \neg(\varphi \wedge \psi) \text{ if } (\mathcal{M}, \mathcal{D}) \models_\varepsilon \neg\varphi \vee \neg\psi.$$

   - 4.3 $(\mathcal{M}, \mathcal{D}) \models_\varepsilon \neg\neg\varphi$ if $(\mathcal{M}, \mathcal{D}) \models_\varepsilon \varphi$.
   - 4.4 $(\mathcal{M}, \mathcal{D}) \models_\varepsilon \neg(\varphi \to \psi)$ if $(\mathcal{M}, \mathcal{D}) \models_\varepsilon \varphi \wedge \neg\psi$.
   - 4.5 $(\mathcal{M}, \mathcal{D}) \models_\varepsilon \neg\exists x\varphi(x)$ if $(\mathcal{M}, \mathcal{D}) \models_\varepsilon \forall x\neg\varphi(x)$.
   - 4.6 $(\mathcal{M}, \mathcal{D}) \models_\varepsilon \neg\forall x\varphi(x)$ if $(\mathcal{M}, \mathcal{D}) \models_\varepsilon \exists x\neg\varphi(x)$.
5. $(\mathcal{M}, \mathcal{D}) \models_\varepsilon \varphi \to \psi$ if $(\mathcal{M}, \mathcal{D}) \models_\varepsilon \neg\varphi \vee \psi$.
6. Finally, we define $(\mathcal{M}, \mathcal{D}) \models_\varepsilon \forall x\varphi(x)$ if

$$\Pr_{\mathcal{D}}\big[a \in \mathcal{M} \mid (\mathcal{M}, \mathcal{D}) \models_\varepsilon \varphi(a)\big] \geq 1 - \varepsilon.$$