Algorithmic Model Theory
SS 2016

Prof. Dr. Erich Grädel and Dr. Wied Pakusa

Mathematische Grundlagen der Informatik
RWTH Aachen

Contents

1 The classical decision problem

The classical decision problem was generally considered as the main problem of mathematical logic until its unsolvability was proved by Church and Turing in 1936/37.

> Das Entscheidungsproblem ist gelöst, wenn man ein Verfahren kennt, das bei einem vorgelegten logischen Ausdruck durch endlich viele Operationen die Entscheidung über die Allgemeingültigkeit bzw. Erfüllbarkeit erlaubt. (. . .) Das Entscheidungsproblem muss als das Hauptproblem der mathematischen Logik bezeichnet werden. [1]
>
> (D. Hilbert and W. Ackermann, Grundzüge der theoretischen Logik, 1928)

By a *logical expression*, Hilbert and Ackermann meant what we now call a formula of first-order logic (FO). Historically, the classical decision problem was part of Hilbert's formalist programme for the foundations of mathematics. Its importance stems from the fact that first-order logic provides a framework to express almost all aspects of mathematics.

We present three equivalent formulations of the classical decision problem.

Satisfiability: Construct an algorithm that decides for any given formula of FO whether it has a model.

Validity: Construct an algorithm that decides for any given formula of FO whether it is valid, i.e. whether it holds in all models where it is defined.

[1] The Entscheidungsproblem is solved when we know a procedure that allows for any given logical expression to decide by finitely many operations its validity or satisfiability. [. . .] The Entscheidungsproblem must be considered the main problem of mathematical logic.

Provability: Construct an algorithm that decides for any given formula ψ of FO whether $\vdash \psi$, meaning that ψ is provable from the empty set of axioms in some complete formal system such as the sequent calculus.

Since ψ is satisfiable if, and only if, $\neg\psi$ is not valid, satisfiability and validity are equivalent problems with respect to computability. The equivalence with provability is a much more intricate result and in fact a consequence of Gödel's Completeness Theorem.

Theorem 1.1 (Completeness Theorem (Gödel))**.** For any given set of sentences $\Phi \subseteq \mathrm{FO}(\tau)$ and any sentence $\psi \in \mathrm{FO}(\tau)$ it holds that

$$\Phi \models \psi \iff \Phi \vdash \psi \,.$$

In particular $\emptyset \models \psi \Leftrightarrow \emptyset \vdash \psi$.

Corollary 1.2. The set of valid first-order formulae is recursively enumerable.

1.1 Basic notions on decidability

In our formulation of the decision problem it was not precisely specified what an algorithm is. It was not until the 1930s that Church, Kleene, Gödel, and Turing provided precise definitions of an abstract algorithm. Their approaches are today known to be equivalent. We introduce the concept of a Turing machine.

Definition 1.3. A *Turing machine* (TM) M is a tuple $M = (Q, \Sigma, \Gamma, q_0, F, \delta)$, where

- Q is a finite set of (control) states,
- Σ, Γ are finite alphabets, where Σ is the working alphabet with a special blank symbol $\Box \in \Sigma$, and $\Gamma \subseteq \Sigma \setminus \{\Box\}$ is the input alphabet,
- $q_0 \in Q$ is the initial state,
- $F \subseteq Q$ is the set of final states and
- $\delta : (Q \setminus F) \times \Sigma \to Q \times \Sigma \times \{-1, 0, 1\}$ is the transition function.

A *configuration* is a triple $C = (q, p, w) \in Q \times \mathbb{N} \times \Sigma^*$, representing the situation that M is in state q, reads tape cell p and that the inscription of the infinite tape is $w = w_0 \ldots w_k$, followed by infinitely many blank-symbols. The transition function δ induces a partial function on the set of all configurations $C \mapsto \mathrm{Next(C)}$, where for $\delta(q, w_p) = (q', a, m)$, the successor configuration of C is defined as $\mathrm{Next(C)} = (q', p + m, w_0 \ldots w_{p-1} a w_{p+1} \cdots w_k)$. A *computation* of the TM M on an input word $x \in \Gamma^*$ is a sequence

$$C_0, C_1, \ldots$$

where $C_0 = C_0(x) := (q_0, 0, x)$ is the input configuration and $C_{i+1} = \mathrm{Next(C_i)}$ for all i.

M halts on x if the computation of M on x is finite and ends in a final configuration $C_f = (q, p, w)$ with $q \in F$. Further

$$L(M) := \{x \in \Gamma^* : M \text{ halts on } x\}.$$

A Turing machine M computes a partial function $f_M : \Gamma^* \to \Sigma^*$ with domain $L(M)$ such that $f_M(x) = y$ if and only if the computation of M on x ends in (q, p, y) for some $q \in F$, $y \in \Sigma^*$ and $p \in \mathbb{N}$.

Definition 1.4. A *Turing acceptor* is a Turing machine M with $F = F^+ \cup F^-$. We say that M *accepts* x if the computation of M on x ends in a state in F^+ and M *rejects* x if the computation of M on x ends in a state in F^-.

Definition 1.5.

- $L \subseteq \Gamma^*$ is *recursively enumerable (r.e.)* if there exists a TM M with $L(M) = L$.
- $L \subseteq \Gamma^*$ is *co-recursively enumerable (co-r.e.)* if $\overline{L} := \Gamma^* \setminus L$ is r.e..
- A (partial) function $f : \Gamma^* \to \Sigma^*$ is *(Turing) computable* if there is a TM M with $f_M = f$.
- $L \subseteq \Gamma^*$ is *decidable* (or *recursive*), if there is a Turing acceptor M such that for all $x \in \Gamma^*$

  $$x \in L \Rightarrow M \text{ accepts } x$$

$x \notin L \Rightarrow M$ rejects x

or, equivalently, if its characteristic function $\chi_L : \Gamma^* \to \{0,1\}$ is Turing computable.

Theorem 1.6. A language $L \subseteq \Gamma^*$ is decidable if, and only if, L is r.e. and co-r.e.

Definition 1.7. Let $A \subseteq \Gamma^*, B \subseteq \Sigma^*$. We say that A is *(many-to-one) reducible* to B, $A \leq B$, if there is a total computable function $f : \Gamma^* \to \Sigma^*$ such that for all $x \in \Gamma^*$ we have $x \in A \Leftrightarrow f(x) \in B$.

Lemma 1.8.

- $A \leq B$, B decidable $\Rightarrow A$ decidable
- $A \leq B$, B r.e. $\Rightarrow A$ r.e.
- $A \leq B$, A undecidable $\Rightarrow B$ undecidable.

There surely are undecidable languages since there are only countably many Turing machines but uncountably many languages. Unfortunately, among these there are quite relevant classes of languages. For example we cannot decide whether a TM halts on a given input.

Definition 1.9 (Halting Problems)**.** The *general halting problem* is defined as

$$H := \{\rho(M)\#\rho(x) : M \text{ Turing machine, } x \in L(M)\}$$

where $\rho(M)$ and $\rho(x)$ are encodings of the TM M and the input x over a fixed alphabet $\{0,1\}$ such that the computation of M on x can be reconstructed from the encodings $\rho(M)$ and $\rho(x)$ in an effective way. This means that there is a universal TM U which, given $\rho(M)$ and $\rho(x)$, simulates the computation of M on x and halts if, and only if, M halts on x. Thus, $L(U) = H$ from which we conclude that H is r.e..

We introduce two special variants of the halting problem.

- *The self-application problem:* $H_0 := \{\rho(M) : \rho(M) \in L(M)\}$.
- *Halting on the empty word:* $H_\varepsilon := \{\rho(M) : \varepsilon \in L(M)\}$.

Theorem 1.10. H, H_0, and H_ε are undecidable.

Proof.

- H_0 is not co-r.e. and thus undecidable. Otherwise $\overline{H}_0 = L(M_0)$ for some TM M_0. Then
$$\rho(M_0) \in \overline{H}_0 \;\Leftrightarrow\; \rho(M_0) \in L(M_0) \;\Leftrightarrow\; \rho(M_0) \in H_0.$$
- H_0 is a special case of H, hence $H_0 \leq H$, and H is undecidable.
- We can reduce H to H_ε, thus H_ε is undecidable. Q.E.D.

We next establish the much more general result that in fact, no non-trivial semantic property of Turing machines can be decided algorithmically. In particular, for any fixed function, there is no algorithm that decides whether a given program computes precisely that function, i.e. we cannot algorithmically prove the correctness of a program. Note that this does not mean that we cannot prove the correctness of a single given program. Instead the statement is that we cannot do so algorithmically for all programs.

Theorem 1.11 (Rice)**.** Let $\mathcal{R}$ be the set of all computable functions and let $S \subseteq \mathcal{R}$ be a set of computable functions such that $S \neq \emptyset$ and $S \neq \mathcal{R}$. Then $\text{code}(S) := \{\rho(M) : f_M \in S\}$ is undecidable.

Proof. Let $\Uparrow$ be the everywhere undefined function, with domain $\text{Def}(\Uparrow) = \emptyset$. Obviously, $\Uparrow$ is computable. Assume that $\Uparrow \notin S$ (otherwise consider $\mathcal{R} \setminus S$ instead of S. Clearly if $\text{code}(\mathcal{R} \setminus S)$ is undecidable then so is $\text{code}(S)$.)

As $S \neq \emptyset$, there exists a function $f \in S$. Let M_f be a TM that computes f, i.e. $f_{M_f} = f$. We define a reduction $H_\varepsilon \leq \text{code}(S)$ by describing a total computable function $\rho(M) \mapsto \rho(M')$ such that

$$M \text{ halts on } \varepsilon \Leftrightarrow f_{M'} \in S.$$

Specifically, given $\rho(M)$, we construct the encoding of a TM M' which, given an input x, proceeds as follows:

- first simulate M on ε (i.e. apply the universal TM U to $\rho(M)\#\varepsilon$);
- then simulate M_f on x (i.e. apply the universal TM U to $\rho(M_f)\#\rho(x)$).

It is clear that the reduction function is computable. Furthermore, if M halts on ε then $f_{M'}(x) = f(x)$ for all inputs x, i.e. $f_{M'} = f$, so $f_{M'} \in S$. If M does not halt on ε then M' does not halt on x for any x, i.e. $f_{M'} = \Uparrow$, so $f_{M'} \notin S$. Q.E.D.

Definition 1.12 (Recursive inseparability)**.** Let $A, B \subseteq \Gamma^*$ be two disjoint sets. We say that A and B are *recursively inseparable* if there exists no decidable set $C \subseteq \Gamma^*$ such that $A \subseteq C$ and $B \cap C = \emptyset$.

Example. $(A, \overline{A})$ are recursively inseparable if, and only if, A is undecidable.

Lemma 1.13. Let $A, B \subseteq \Gamma^*, A \cap B = \emptyset$ be recursively inseparable. Let $X, Y \subseteq \Sigma^*, X \cap Y = \emptyset$, and let f be a total computable function such that $f(A) \subseteq X$ and $f(B) \subseteq Y$. Then X and Y are recursively inseparable.

Proof. Assume there exists a decidable set $Z \subseteq \Sigma^*$ such that $X \subseteq Z$ and $Y \cap Z = \emptyset$. Consider $C = \{x \in \Gamma^* : f(x) \in Z\}$. C is decidable, $A \subseteq C, B \cap C = \emptyset$, thus C separates A, B. Q.E.D.

Notation: We write $(A, B) \leq (X, Y)$ if such a function f exists.

Example. $(A, \overline{A}) \leq (B, \overline{B}) \Leftrightarrow A \leq B$.

As a preparation for Trakhtenbrot's Theorem, we consider the following refinements of H_ε:

$$\begin{aligned} H_\varepsilon^+ &:= \{\rho(M) : M \text{ accepts } \varepsilon\} \\ H_\varepsilon^- &:= \{\rho(M) : M \text{ rejects } \varepsilon\} \\ H_\varepsilon^\infty &:= \{\rho(M) : \text{ the computation of } M \text{ on } \varepsilon \text{ is infinite} \\ &\qquad\qquad \text{and does not cycle.}\} \end{aligned}$$

H_0^+, H_0^-, H_0^∞ are defined analogously, with respect to self-application.

Theorem 1.14. $H_\varepsilon^+, H_\varepsilon^-$ and H_ε^∞ are pairwise recursively inseparable.

Proof. $(H_\varepsilon^+, H_\varepsilon^\infty)$: We show that every set C with $H_\varepsilon^+ \subseteq C$ and $H_\varepsilon^\infty \cap C = \emptyset$ is undecidable by reducing the halting problem H_ε to C. Define a reduction $\rho(M) \mapsto \rho(M')$ as follows. From a given code $\rho(M)$ construct the code of a TM M' that simulates M and simultaneously counts the number of computation steps since the start. If M halts (accepting or rejecting), M' accepts.

It is clear that the reduction function is computable. If M halts on ε then M' halts on ε as well and accepts, so $\rho(M') \in H_\varepsilon^+ \subseteq C$. If M does not halt on ε then M' does not halt either, and never cycles, so $\rho(M') \in H_\varepsilon^\infty$ and as $H_\varepsilon^\infty \cap C = \emptyset$, we have $\rho(M') \notin C$.

The statement for H_ε^- and H_ε^∞ is proven analogously.

$(H_\varepsilon^-, H_\varepsilon^+)$: Show that $(H_0^-, H_0^+) \leq (H_\varepsilon^-, H_\varepsilon^+)$ and that (H_0^-, H_0^+) are recursively inseparable.

- $(H_0^-, H_0^+) \leq (H_\varepsilon^-, H_\varepsilon^+)$:
  For a given input TM M construct a TM M' that ignores its own input and simulates M on $\rho(M)$. Obviously, M' can be constructed effectively, say by a computable function h. Now $h(M)$ accepts ε iff M accepts $\rho(M)$ and $h(M)$ rejects ε iff M rejects $\rho(M)$.
- (H_0^-, H_0^+) recursively inseparable:
  Assume there exists a decidable C with $H_0^- \subseteq C$ and $H_0^+ \subseteq \overline{C}$. Consider a machine M_0 that decides C. There are two cases:
  (1) M_0 accepts $\rho(M_0)$. Then $\rho(M_0) \in C$ by definition of M_0. Then $\rho(M_0) \notin H_0^+$ by definition of C. On the other hand, if M_0 accepts $\rho(M_0)$ then $\rho(M_0) \in H_0^+$ (by definition of H_0^+), a contradiction.
  (2) M_0 rejects $\rho(M_0)$. Then $\rho(M_0) \notin C$ by definition of M_0. Then $\rho(M_0) \notin H_0^-$ by definition of C. On the other hand, if M_0 rejects $\rho(M_0)$ then $\rho(M_0) \in H_0^-$ (by definition of H_0^-), a contradiction.

Q.E.D.

1.2 Trakhtenbrot's Theorem

In the following, we consider FO, more precisely first-order logic with equality. We restrict ourselves to a countable signature

$$\tau_\infty := \{R_j^i : i, j \in \mathbb{N}\} \cup \{f_j^i : i, j \in \mathbb{N}\}$$

where each R^i_j is a relation symbol of arity i and each f^i_j is a function symbol of arity i. We write formulae in $\mathrm{FO}(\tau_\infty)$ as words over the fixed finite alphabet

$$\Gamma := \{R, f, x, 0, 1, [,]\} \cup \{=, \neg, \wedge, \vee, \rightarrow, \leftrightarrow, \exists, \forall.(,)\},$$

using the following encoding of relation symbols, function symbols, and variables:

relation symbols:	R^i_j	$\longmapsto$	$R[\text{bin } i][\text{bin } j]$
function symbols:	f^i_j	$\longmapsto$	$f[\text{bin } i][\text{bin } j]$
variables:	x_j	$\longmapsto$	$x[\text{bin } j]$.

In this way, every formula $\varphi \in \mathrm{FO}$ can be viewed as a word in Γ^*.

Let $X \subseteq \mathrm{FO}$ be a class of formulae. We analyse the following decision problems:

$$\begin{aligned}
\mathit{Sat}(X) &:= \{\psi \in X : \psi \text{ has a model}\}\\
\mathit{Fin\text{-}Sat}(X) &:= \{\psi \in X : \psi \text{ has a finite model}\}\\
\mathit{Val}(X) &:= \{\psi \in X : \psi \text{ is valid}\}\\
\mathit{Non\text{-}Sat}(X) &:= X \setminus \mathit{Sat}(X)\\
\mathit{Inf\text{-}Axioms}(X) &:= \mathit{Sat}(X) \setminus \mathit{Fin\text{-}Sat}(X)\\
&= \{\psi \in X : \psi \text{ is an infinity axiom, i.e. } \psi \text{ has a}\\
&\qquad\qquad \text{model but no finite model}\}.
\end{aligned}$$

Theorem 1.15. Let $X \subseteq \mathrm{FO}$ be decidable. Then

(1) $\mathit{Val}(X)$ is r.e.
(2) $\mathit{Non\text{-}Sat}(X)$ is r.e.
(3) $\mathit{Sat}(X)$ is co-r.e.
(4) $\mathit{Fin\text{-}Sat}(X)$ is r.e.
(5) $\mathit{Inf\text{-}Axioms}(X)$ is co-r.e.

Proof. (1) φ is valid $\Leftrightarrow\ \vdash \varphi$ (Completeness Theorem). Thus we can systematically enumerate all proofs and halt if a proof for φ is listed.
(2) φ valid $\Leftrightarrow \neg\varphi$ is not satisfiable.

(3) Follows from Item (2).
(4) Systematically generate all finite models and halt if a model of φ is found.
(5) $\mathrm{FO} \setminus \mathit{Inf\text{-}Axioms}(X) = \mathit{Non\text{-}Sat}(X) \cup \mathit{Fin\text{-}Sat}(X)$ is r.e. Q.E.D.

Definition 1.16. A class $X \subseteq \mathrm{FO}$ has the *finite model property* (FMP) if every satisfiable $\varphi \in X$ has a finite model, i.e. if $\mathit{Sat}(X) = \mathit{Fin\text{-}Sat}(X)$.

Theorem 1.17. Suppose that $X \subseteq \mathrm{FO}$ is decidable and that X has the FMP. Then $\mathit{Sat}(X)$ is decidable.

Proof. $\mathit{Sat}(X)$ is co-r.e. and since $\mathit{Sat}(X) = \mathit{Fin\text{-}Sat}(X)$ and $\mathit{Fin\text{-}Sat}(X)$ is r.e. also $\mathit{Sat}(X)$ is r.e. Thus $\mathit{Sat}(X)$ is decidable. Q.E.D.

In this case also $\mathit{Fin\text{-}Sat}(X)$, $\mathit{Non\text{-}Sat}(X)$, $\mathit{Val}(X)$ are decidable and of course $\mathit{Inf\text{-}Axioms}(X) = \emptyset$ is decidable.

Theorem 1.18 (Trakhtenbrot)**.** There is a finite vocabulary $\tau \subseteq \tau_\infty$ such that $\mathit{Fin\text{-}Sat}(\mathrm{FO}(\tau)), \mathit{Non\text{-}Sat}(\mathrm{FO}(\tau))$ and $\mathit{Inf\text{-}Axioms}(\mathrm{FO}(\tau))$ are pairwise recursively inseparable and therefore undecidable.

The proof of Trakhtenbrot's theorem introduces a proof strategy that can be applied in many other undecidability proofs. (Do not focus on the technicalities but on the general idea to construct the reduction formulae.)

Proof. Let M be a deterministic Turing acceptor. We show that there is an effective reduction $\rho(M) \mapsto \psi_M$ such that

(1) M accepts $\varepsilon \implies \psi_M$ has a finite model.
(2) M rejects $\varepsilon \implies \psi_M$ is unsatisfiable.
(3) The computation of M on ε is infinite and non-periodic $\implies \psi_M$ is an infinity axiom.

Then the theorem follows by Lemma 1.13.

Let M be a Turing acceptor with states $Q = \{q_0, \dots, q_r\}$, initial state q_0, alphabet $\Sigma = \{a_0, \dots, a_s\}$ (where $a_0 = \Box$), final states $F = F^+ \cup F^-$ and transition function δ.

ψ_M is defined over the vocabulary $\tau = \{0, f, q, p, w\}$ where 0 is a constant, f, q, p are unary functions and w is a binary function. Define the term $\boldsymbol{k}$ as $f^k 0$.

By constructing a formula we intend to have a model $\mathfrak{A}_M = (A, 0, f, q, p, w)$ describing a run of M on the input ε where

- universe $A = \{0, 1, 2, \ldots, n\}$ or $A = \mathbb{N}$;
- $f(t) = t + 1$ if $t + 1 \in A$ and $f(t) = t$, if t is the last element of A;
- $q(t) = i$ iff M is at time t in state q_i;
- $p(t)$ is the head position of M at time t;
- $w(s, t) = i$ iff symbol a_i is at time t on tape-cell s.

Note that we cannot enforce this model, but if ψ_M is satisfiable this one will be among its models.

$$\psi_M := \text{START} \land \text{COMPUTE} \land \text{END}$$

$$\text{START} := (q0 = 0 \land p0 = 0 \land \forall x\, w(x, 0) = 0).$$

[Enforces input configuration on ε at time 0]

$$\text{COMPUTE} := \text{NOCHANGE} \land \text{CHANGE}$$

$$\text{NOCHANGE} := \forall x \forall y (py \neq x \to w(x, fy) = w(x, y))$$

[content of currently not visited tape cells does not change]

$$\text{CHANGE} := \bigwedge_{\delta:(q_i,a_j)\mapsto(q_k,a_\ell,m)} \forall y(\alpha_{i,j} \to \beta_{k,\ell,m})$$

where

$$\alpha_{ij} := (qy = \boldsymbol{i} \land w(py, y) = \boldsymbol{j})$$

[M is at time y in state q_i and reads the symbol a_j]

$$\beta_{k,\ell,m} := (qfy = \boldsymbol{k} \land w(py, fy) = \boldsymbol{\ell} \land \text{MOVE}_m)$$

and

$$\text{MOVE}_m := \begin{cases} pfy = py & \text{if } m = 0 \\ pfy = fpy & \text{if } m = 1 \\ \exists z(fz = py \land pfy = z) & \text{if } m = -1. \end{cases}$$

$$\text{END} := \bigwedge_{\substack{\delta(q_i,a_j) \text{ undef.} \\ q_i \notin F^+}} \forall y\, \neg\alpha_{ij}$$

[The only way the computation ends is in an accepting state]

Remark 1.19.

- $\rho(M) \mapsto \psi_M$ is an effective construction.
- If M accepts ε, the intended model is finite and is indeed a model $\mathfrak{A}_M \models \psi_M$, thus $\psi_M \in \mathit{Fin\text{-}Sat}(FO(\tau))$.
- If the computation of M on ε is infinite, the intended model is infinite and $\mathfrak{A}_M \models \psi_M$.

It remains to show that if M rejects ε, then ψ_M is unsatisfiable, and if the computation of M on ε is infinite and aperiodic, then ψ_M is an infinity axiom.

Suppose $\mathfrak{B} = (B, 0, f, q, p, w) \models \psi_M$.

Definition 1.20. $\mathfrak{B}$ *enforces* at time t the configuration (q_i, j, w) with $w = a_{i_0} \ldots a_{i_m} \in \Sigma^*$ if

(1) $\mathfrak{B} \models q\boldsymbol{t} = \boldsymbol{i}$,
(2) $\mathfrak{B} \models p\boldsymbol{t} = \boldsymbol{j}$,
(3) for all $k \leq m$, $\mathfrak{B} \models w(\boldsymbol{k}, \boldsymbol{t}) = \boldsymbol{i_k}$ and for all $k > m$, $\mathfrak{B} \models w(\boldsymbol{k}, \boldsymbol{t}) = 0$.

Since $\mathfrak{B} \models \psi_M$, the following holds:

- $\mathfrak{B}$ enforces $C_0 = (q_0, 0, \varepsilon)$ at time 0 (since $\mathfrak{B} \models$ START.)
- If $\mathfrak{B}$ enforces at time t a non-final configuration C_t, then $\mathfrak{B}$ enforces the configuration $C_{t+1} = \mathrm{Next(C_t)}$ at time $t + 1$.
- Especially, the computation of M cannot reach a rejecting configuration. It follows that if M rejects ε, then ψ_M is unsatisfiable.
  Consider an infinite and aperiodic computation of M, and assume $\mathfrak{B} \models \psi_M$ is finite. Since $\mathfrak{B}$ is finite, it enforces a periodic computation in contradiction to the assumption that the computation of M is aperiodic.

$$\begin{tikzcd} C_0 \vdash \ldots \vdash C_r \arrow[rr, bend right, dash] & \vdash \ldots & \vdash C_{t-1} \end{tikzcd}$$

We have shown:

- If M accepts ε, then ψ_M has a finite model.
- If M rejects ε, then ψ_M is unsatisfiable.
- If the computation of M is infinite and aperiodic, then ψ_M is an infinity axiom. Q.E.D.

We now know that the sets of all finitely satisfiable, all unsatisfiable and all only infinitely satisfiable formulae are undecidable for FO(τ) where τ consists of only three unary functions and one binary function. This raises a number of questions.

(1) For which other vocabularies σ do we have similar undecidability results for FO(σ)?
(2) For which σ is satisfiability of FO(σ) decidable?
(3) Is there a complete classification? In this case, we want to find minimal vocabularies σ such that the above problems are undecidable, i.e. vocabularies such that any further restriction yields a class of formulae for which satisfiability is decidable.

We first define what it means that a fragment of FO is as hard for satisfiability as the whole FO.

Definition 1.21. $X \subseteq \mathrm{FO}$ is a *reduction class* if there exists a computable function $f : \mathrm{FO} \to X$ such that $\psi \in \mathit{Sat}(\mathrm{FO}) \Leftrightarrow f(\psi) \in \mathit{Sat}(X)$.

Let $X, Y \subseteq \mathrm{FO}$. A *conservative reduction of* X *to* Y is a computable function $f : X \to Y$ with

- $\psi \in \mathit{Sat}(X) \Leftrightarrow f(\psi) \in \mathit{Sat}(Y)$, and
- $\psi \in \mathit{Fin\text{-}Sat}(X) \Leftrightarrow f(\psi) \in \mathit{Fin\text{-}Sat}(Y)$.

X is a *conservative reduction class* if there exists a conservative reduction of FO to X.

Corollary 1.22. Let X be a conservative reduction class. Then $\mathit{Fin\text{-}Sat}(X)$, $\mathit{Inf\text{-}Axioms}(X)$ and $\mathit{Non\text{-}Sat}(X)$ are pairwise recursively inseparable, and thus $\mathit{Fin\text{-}Sat}(X)$, $\mathit{Sat}(X)$, $\mathit{Val}(X)$, $\mathit{Non\text{-}Sat}(X)$, $\mathit{Inf\text{-}Axioms}(X)$ are undecidable.

Proof. A conservative reduction from FO to X yields a uniform reduction from $\mathit{Fin\text{-}Sat}(\mathrm{FO})$, $\mathit{Inf\text{-}Axioms}(\mathrm{FO})$ and $\mathit{Non\text{-}Sat}(\mathrm{FO})$ to $\mathit{Fin\text{-}Sat}(X)$, $\mathit{Inf\text{-}Axioms}(X)$ and $\mathit{Non\text{-}Sat}(X)$, respectively. Q.E.D.

It is indeed possible to give a complete classification of those vocabularies σ such that FO(σ) is decidable.

Theorem 1.23. If $\sigma \subseteq \{P_0, P_1, \ldots\} \cup \{f\}$ consists of at most one unary function f and an arbitrary number of monadic predicates $P_0, P_1, \ldots$, then $\mathit{Sat}(\mathrm{FO}(\sigma))$ is decidable. In all other cases, $\mathit{Sat}(\mathrm{FO}(\sigma))$, $\mathit{Inf\text{-}Axioms}(\mathrm{FO}(\sigma))$ and $\mathit{Non\text{-}Sat}(\mathrm{FO}(\sigma))$ are pairwise recursively inseparable, and FO(σ) is a conservative reduction class.

A full proof of this classification theorem is rather difficult. In particular, the decidability of the monadic theory of one unary function, which implies the decidability part, is a difficult theorem due to Rabin. On the other side, one has to show that Trakhtenbrot's theorem applies to the vocabularies

$\tau_1 = \{E\}$ where E is a binary relation,
$\tau_2 = \{f, g\}$ where f, g are unary functions,
$\tau_3 = \{F\}$ where F is a binary function,

and hence also to all extensions of τ_1, τ_2, τ_3.

Of course, one may also look at other syntactic restrictions besides restricting the vocabulary. One possibility is to restrict the number of variables. This is only interesting for relational formulae. If we have functions, satisfiability is undecidable even for formulae with only one variable, as we shall see later.

Define FO^k as first-order logic with relational symbols only and a fixed collection of k variables, say $x_1, \ldots, x_k$.

Theorem 1.24.

- FO^2 has the finite model property and is decidable (see Sect. 1.6).
- FO^3 is a conservative reduction class.

A further important possibility is to restrict the structure of quantifier prefixes of formulae in prenex normal form, and to combine this with restrictions on the vocabulary, and the presence or absence of equality. This leads to the notion of a *prefix-vocabulary class* in first-order logic, and indeed, also for these fragments of FO there is a complete classification of those with a solvable satisfiability problem, and those that are conservative reduction classes.

A full description of this classification exceeds the scope of this course by far (see E. Börger, E. Grädel, and Y. Gurevich, The Classical

Decision Problem, 1997). Instead we shall present some of the fundamental methods for establishing such results, and illustrate these with applications to specific fragments of first-order logic.

1.3 Domino problems

Domino problems are a simple and yet general tool for proving undecidability results (and lower bounds in complexity theory) without the need of explicit encodings of Turing machine computations.

The informal idea is the following: a domino problem is given by a finite set of dominoes or tiles, each of them an oriented unit square with coloured edges; the question is whether it is possible to cover the first quadrant in the Cartesian plane by copies of these tiles, without holes and overlaps, such that adjacent dominoes have matching colours on their common edge. The set of tiles is finite, but there are infinitely many copies of each tile available; rotation of the tiles is not allowed. Variants of this problem require a tiling of a different geometric object (a finite square, a rectangle, or a torus) and/or that certain places (e.g. the origin, the bottom row or the diagonal) are tiled by specific tiles.

Here is a more abstract defintion.

Definition 1.25. A *domino system* is a structure $\mathcal{D} = (D, H, V)$ with

- a finite set D (of dominoes),
- horizontal and vertical compatibility relations $H, V \subseteq D \times D$.

The intuitive meaning of H and V is that

- $(d, d') \in H$ if the right colour of d is equal to the left colour of d',
- $(d, d') \in V$ if the top colour of d is equal to the bottom colour of d' (see Figure 1.1).

A *tiling* of $\mathbb{N} \times \mathbb{N}$ by $\mathcal{D}$ is a function $t : \mathbb{N} \times \mathbb{N} \to D$ such that for all $x, y \in \mathbb{N}$

- $(t(x,y), t(x+1,y)) \in H$ and
- $(t(x,y), t(x,y+1)) \in V$.

A periodic tiling of $\mathbb{N} \times \mathbb{N}$ by $\mathcal{D}$ is a tiling t for which there exist two integers $h, v \in \mathbb{N}$ such that $t(x,y) = t(x+h,y) = t(x,y+v)$ for all $x, y \in \mathbb{N}$.

The decision problem DOMINO is described as

DOMINO := $\{\mathcal{D}$: there exists a tiling of $\mathbb{N} \times \mathbb{N}$ by $\mathcal{D}\}$

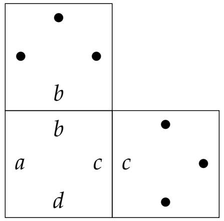


Figure 1.1. Domino adjacency condition

Theorem 1.26 (Berger, Robinson)**.** DOMINO is co-r.e. and undecidable.

In this general form, this is quite a difficult result. A simpler variant is the so-called origin-constrained domino problem, that requires that a specific domino must be placed at the point $(0,0)$. With this requirement, it is straightforward to encode Turing machine computations by domino tilings (successive rows of the tiling correspond to successive configurations in the computation), and thus to reduce halting problems to tiling problems for domino systems. The origin constraint is used to encode the beginning of the computation (and to avoid that the entire space can be tiled by a domino corresponding to the blank symbol) Without an origin constraint, the problem is more difficult to handle; an essential part of the proof is the construction of a set of dominoes that admits only non-periodic tilings.

There are several extensions and variations of this result.

Theorem 1.27. A domino system $\mathcal{D}$ admits a tiling of $\mathbb{Z} \times \mathbb{Z}$ if, and only if, it admits a tiling of $\mathbb{N} \times \mathbb{N}$.

Proof. It is clear that a tiling of $\mathbb{Z} \times \mathbb{Z}$ also gives a tiling of $\mathbb{N} \times \mathbb{N}$. The converse is a nice application of König's Lemma. Suppose that t is a tiling of $\mathbb{N} \times \mathbb{N}$ by $\mathcal{D}$. There exists at least one domino d such that for all n there exist $i, j > n$ with $t(i,j) = d$. Fix such a d. Further, for every $k \in \mathbb{N}$, let S_k be the square $\{-k, \dots, -1, 0, 1, \dots, k\} \times \{-k, \dots, -1, 0, 1, \dots, k\}$.

We define a finitely branching tree whose nodes are the correct tilings t_k of S_k by $\mathcal{D}$ such that $t_k(0,0) = d$. The root is the unique such tiling of S_0 and the children of a tiling t_k are the possible extensions to tilings t_{k+1} of S_{k+1}. This tree contains paths of any finite length. By König's Lemma it also contains an infinite path from the root, which means that $\mathcal{D}$ admits a tiling of $\mathbb{Z} \times \mathbb{Z}$. Q.E.D.

The undecidability result from Theorem 1.26 can be strengthened to a recursive inseparability result.

Theorem 1.28. The set of domino systems admitting a periodic tiling of $\mathbb{N} \times \mathbb{N}$, those that admit no tiling of $\mathbb{N} \times \mathbb{N}$ and those that admit a tiling but not a periodic one are pairwise recursively inseparable.

The proof of Theorem 1.28 reduces the halting problems $H_\varepsilon^+, H_\varepsilon^-, H_\varepsilon^\infty$, to the domino problems. There exists a recursive function that associates with every TM M a domino system $\mathcal{D}$ satisfying

- If $M \in H_\varepsilon^+$ then $\mathcal{D}$ admits a periodic tiling of $\mathbb{N} \times \mathbb{N}$.
- If $M \in H_\varepsilon^-$ then $\mathcal{D}$ admits no tiling of $\mathbb{N} \times \mathbb{N}$.
- If $M \in H_\varepsilon^\infty$ then $\mathcal{D}$ admits a tiling of $\mathbb{N} \times \mathbb{N}$ but no periodic one.

Definition 1.29. A computable function f is a *conservative reduction from domino systems to* X if, for all domino systems $\mathcal{D}$, $f(\mathcal{D}) = \varphi_\mathcal{D}$ is in X and the following holds:

- $\mathcal{D}$ admits a periodic tiling of $\mathbb{N} \times \mathbb{N} \Rightarrow \psi_\mathcal{D}$ has a finite model
- $\mathcal{D}$ admits no tiling of $\mathbb{N} \times \mathbb{N} \Rightarrow \psi_\mathcal{D}$ is unsatisfiable
- $\mathcal{D}$ admits a tiling of $\mathbb{N} \times \mathbb{N}$ but no periodic one $\Rightarrow \psi_\mathcal{D}$ is an infinity axiom.

Proposition 1.30. Let $X \in$ FO. If there exists a conservative reduction from domino systems to X then X is a conservative reduction class.

Proof. Since *Fin-Sat*(FO) and *Non-Sat*(FO) are recursively enumerable and *Inf-Axioms*(FO) is co-recursively enumerable, we can associate with every first-order formula ψ a Turing machine M such that

- $\psi \in$ *Fin-Sat*(FO) $\Rightarrow \rho(M) \in H_\varepsilon^+$,
- $\psi \in$ *Non-Sat*(FO) $\Rightarrow \rho(M) \in H_\varepsilon^-$,

- $\psi \in$ *Inf-Axioms*(FO) $\Rightarrow \rho(M) \in H_\varepsilon^\infty$.

According to the assumption, there is a reduction $\mathcal{D} \mapsto \varphi_\mathcal{D}$ from domino systems to X. Thus, the domino method yields a conservative reduction from FO to X.

Q.E.D.

1.4 Applications of the domino method

We now apply the domino method to obtain several reduction classes.

The Kahr-Moore-Wang class KMW is the class of all first-order sentences of form $\forall x \exists y \forall z \varphi$, where φ is a quantifier-free formula without equality, whose vocabulary contains only binary relation symbols.

Theorem 1.31. The Kahr-Moore-Wang class is a conservative reduction class.

Proof. It suffices to construct a conservative reduction from domino systems to KMW, i.e., a mapping $\mathcal{D} \mapsto \psi_\mathcal{D}$ over a vocabulary consisting of binary relation symbols $(P_d)_{d \in D}$ such that

(1) $\mathcal{D}$ admits a periodic tiling of $\mathbb{N} \times \mathbb{N} \Rightarrow \psi_\mathcal{D}$ has a finite model
(2) $\mathcal{D}$ admits no tiling of $\mathbb{N} \times \mathbb{N} \Rightarrow \psi_\mathcal{D}$ is unsatisfiable
(3) $\mathcal{D}$ admits a tiling of $\mathbb{N} \times \mathbb{N}$ but no periodic one $\Rightarrow \psi_\mathcal{D}$ is an infinity axiom

For a tiling $t : \mathbb{N} \times \mathbb{N} \to D$, an intended model of $\psi_\mathcal{D}$ is $\mathbb{N}$ with the interpretation $P_d = \{(i,j) \in \mathbb{N} \times \mathbb{N} : t(i,j) = d\}$ for all $d \in D$. We define $\psi_\mathcal{D}$ by

$$\psi_\mathcal{D} := \forall x \exists y \forall z \Big(\bigwedge_{d \neq d'} P_d xz \to \neg P_{d'} xz \\ \wedge \bigvee_{(d,d') \in H} (P_d xz \wedge P_{d'} yz) \wedge \bigvee_{(d,d') \in V} (P_d zx \wedge P_{d'} zy) \Big).$$

Obviously $\psi_\mathcal{D}$ is of the desired format, i.e. $\psi_\mathcal{D} \in$ KMW.

(1) Suppose that $\mathcal{D}$ admits a periodic tiling t of $\mathbb{N} \times \mathbb{N}$, such that $t(x,y) = t(x+h,y) = t(x,y+v)$ for all x,y. We construct a finite model

of $\psi_{\mathcal{D}}$ as follows. Let $m := lcm(h,v)$ be the least common multiple of h and v. Then t induces a tiling

$$t' : \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/m\mathbb{Z} \to D$$

with $t'(x,y) = t(x(\mod m), y(\mod m))$.

It follows that $\mathfrak{A} = (\mathbb{Z}/m\mathbb{Z}, (P_d)_{d\in D})$ with $P_d = \{(i,j) : t'(i,j) = d\}$ is a finite model for $\psi_{\mathcal{D}}$ (for x in $\mathbb{Z}/m\mathbb{Z}$ choose $y := x+1 \pmod m$).

(2) By analogous arguments, it follows, that whenever $\mathcal{D}$ admits a tiling of $\mathbb{N}\times\mathbb{N}$, then $\psi_{\mathcal{D}}$ has a model over $\mathbb{N}$.

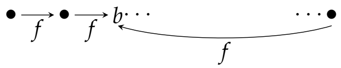
(3) Finally we prove that if $\psi_{\mathcal{D}}$ has a model, then $\mathcal{D}$ admits a tiling of $\mathbb{N}\times\mathbb{N}$, and if that model is finite, we even obtain a periodic tiling.

Consider the Skolem normal form $\varphi_{\mathcal{D}}$ of $\psi_{\mathcal{D}}$:

$$\begin{aligned}\varphi_{\mathcal{D}} := \forall x \forall z(& \bigwedge_{d\neq d'} P_d xz \to \neg P_{d'} xz \\ & \wedge \bigvee_{(d,d')\in H} (P_d xz \wedge P_{d'} fxz) \wedge \bigvee_{(d,d')\in V} (P_d zx \wedge P_{d'} zfx).\end{aligned}$$

If $\psi_{\mathcal{D}}$ is satisfiable, then also $\varphi_{\mathcal{D}}$ has a model $\mathfrak{B} = (B, f, (P_d)_{d\in D})$. Define a tiling $t : \mathbb{N}\times\mathbb{N} \to D$ as follows: choose any $b \in B$, and for all $i,j \in \mathbb{N}$, set $t(i,j) := d$ for the unique $d \in D$ such that $\mathfrak{B} \models P_d(f^i b, f^j b)$. Since $\mathfrak{B} \models \varphi_{\mathcal{D}}$, it follows that t is a correct tiling.

Now suppose that $\mathfrak{B} \models \varphi_{\mathcal{D}}$ is finite.



Choose $b \in B$ such that, for some $n \geq 1$, $f^n b = b$. Then the defined tiling t is periodic. Q.E.D.

Corollary 1.32. FO^3 is a conservative reduction class.

Later we shall prove that FO^2 has the FMP.

Consider now formula classes $X \subseteq$ FO over functional vocabularies. One can prove that FO(τ) is a conservative reduction class if τ contains

- two unary functions or

- one binary function.

This is even true for sentences of the form $\forall x\varphi$ where φ is quantifier-free.

We stablish, again via a conservative reduction from domino problems, a weaker result from which the above mentioned ones can be obtained by interpretation arguments (see exercises).

Theorem 1.33. The class $\mathcal{F}$, consisting of all sentences $\forall x\varphi$ where φ is a quantifier-free formula whose vocabulary consists only of unary function symbols, is a conservative reduction classes.

Proof. We define a conservative reduction $\mathcal{D} = (D,H,V) \mapsto \psi_{\mathcal{D}}$ where $\psi_{\mathcal{D}} \in \mathcal{F}$ has the vocabulary $\{f, g, (h_d)_{d\in D}\}$ where all function symbols are unary. The intended model is $\mathbb{N}\times\mathbb{N}$ with successor functions f and g. The subformula $\forall x(fgx = gfx)$ ensures that the models of $\psi_{\mathcal{D}}$ contain a two-dimensional grid. The fact that a position x is tiled by $d \in D$ is expressed by requiring that $h_d x = x$, i.e. that x is a fixed point of h_d.

$$\begin{aligned}\psi_{\mathcal{D}} := \forall x\big(fgx = gfx \wedge & \bigwedge_{d\neq d'} (h_d x = x \to h_{d'} x \neq x) \\ & \wedge \bigvee_{(d,d')\in H} (h_d x = x \wedge h_{d'} fx = fx) \\ & \wedge \bigvee_{(d,d')\in V} (h_d x = x \wedge h_{d'} gx = gx)\big)\ .\end{aligned}$$

We claim that there exists a tiling $t : \mathbb{N}\times\mathbb{N} \to \mathcal{D}$ if and only if $\psi_{\mathcal{D}}$ is satisfiable.

$'' \Rightarrow ''$ Assume that t is a correct tiling. Construct the (intended) model $\mathfrak{A} = (\mathbb{N}\times\mathbb{N}, f, g, (h_d)_{d\in\mathcal{D}})$ with

– $f(i,j) = (i+1, j)$,

– $g(i,j) = (i, j+1)$,

– $h_d(i,j) \begin{cases} = (i,j) & \text{if } t(i,j) = d \\ \neq (i,j) & \text{otherwise.} \end{cases}$

Clearly $\mathfrak{A} \models \psi_{\mathcal{D}}$.

"$\Leftarrow$" Consider $\mathfrak{B} = (B, f, g, (h_d)_{d\in\mathcal{D}}) \models \psi_{\mathcal{D}}$.
Choose an arbitrary $b \in B$ and define $t : \mathbb{N} \times \mathbb{N} \to D$ by

$$t(i,j) := d \text{ iff } \mathfrak{B} \models h_d f^i g^j b = f^i g^j b.$$

Note that every point in B is a fixed-point of exactly one of the functions h_d, and t is well-defined and a a correct tiling. Further, if $\mathfrak{B}$ is finite, then σ is periodic, and thus the reduction is conservative.

Q.E.D.

Exercise 1.1. Prove that the more restricted class $\mathcal{F}_2 \subseteq \mathcal{F}$ consisting of sentences in $\mathcal{F}$ that contain just two unary function symbols, is also a conservative reduction class.

Hint: Transform sentences $\forall x\varphi$ with unary function symbols $f_1, \ldots, f_m$ into sentences $\forall x\tilde{\varphi} := \forall x\varphi[x/hx, f_i/hg^i]$ where h, g are fresh unary function symbols.

1.5 The finite model property

We study the finite model property (FMP) for fragments of FO as a mean to show that these fragments are decidable, and also to better understand their expressive power and algorithmic complexity.

Recall that a class $X \subseteq$ FO has the *finite model property* if $Sat(X) = Fin\text{-}Sat(X)$. Since for any decidable class X, $Fin\text{-}Sat(X)$ is r.e. and $Sat(X)$ is co-r.e., it follows that $Sat(X)$ is decidable if X has the FMP. In many cases, the proof that a class has the finite model property provides a bound on the model's cardinality, and thus a complexity bound for the satisfiability problem. To prove completeness for complexity classes we make use of a bounded variant of the domino problem.

We shall illustrate the power of this method by a few examples.

Definition 1.34. The *atomic k-type* of $a_1, \ldots, a_k$ in $\mathfrak{A}$ is defined as

$$\begin{aligned}\text{atp}_{\mathfrak{A}}(a_1, \ldots, a_k) := \{\gamma(x_1 \ldots, x_k) : \gamma \text{ atomic formula or negated}\\ \text{atomic formula such that } \mathfrak{A} \models \gamma(a_1, \ldots, a_k)\}.\end{aligned}$$

In the examples that we consider here, the structures contain unary or binary relations only. Hence, to describe a structure it suffices to define its universe and to specify the atomic 1-types and 2-types for all of its elements.

Example 1.35. Let $\mathfrak{A}$ be the structure $(A, E_1, \ldots, E_m)$ where the E_i are binary relations. Then for $a \in A$:

$$\text{atp}_{\mathfrak{A}}(a) = \{E_i xx : \mathfrak{A} \models E_i aa\} \cup \{\neg E_i xx : \mathfrak{A} \models \neg E_i aa\}.$$

The *monadic class* (also called the Löwenheim class) is the class of first-order sentences over a vocabulary the contains only unary predicates.

Theorem 1.36. The monadic class has the FMP.

Proof. Let $\mathfrak{A} = (A, P_1^{\mathfrak{A}}, \ldots, P_n^{\mathfrak{A}}) \models \varphi$ where $\text{qr}(\varphi) = m$. For each sequence of bits $\alpha = \alpha_1 \ldots \alpha_n \in \{0,1\}^n$ we define $P_\alpha^{\mathfrak{A}} = Q_1 \cap Q_2 \cap \ldots \cap Q_n$, where $Q_i = P_i^{\mathfrak{A}}$ if $\alpha_i = 1$ and $Q_i = A \setminus P_i^{\mathfrak{A}}$ if $\alpha_i = 0$. Notice that the sets $P_\alpha^{\mathfrak{A}}$ define a partition of A, and that α completely describes the atomic 1-type of any $a \in P_\alpha^{\mathfrak{A}}$.

We construct $\mathfrak{B}$ by taking $\min(|P_\alpha^{\mathfrak{A}}|, m)$ elements into each $P_\alpha^{\mathfrak{B}}$. Observe that $\mathfrak{B}$ is completly specified in this way, with $P_i^{\mathfrak{B}} = \bigcup_{\alpha|\alpha_i=1} P_\alpha^{\mathfrak{B}}$). We show that $\mathfrak{A} \equiv_m \mathfrak{B}$ using the Ehrenfeucht-Fraïssé Theorem.

The following is a winning strategy for Duplicator in the Ehrenfeucht-Fraïssé game with m moves on $(\mathfrak{A}, \mathfrak{B})$: Answer any element chosen by Spoiler by an element with the same atomic type in the other structure, respecting equalities and inequalities with previously chosen elements. Due to the construction it is certainly possible to do that for m moves, so Duplicator wins the game. Hence $\mathfrak{A} \equiv_m \mathfrak{B}$, and therefore $\mathfrak{B} \models \varphi$. Q.E.D.

From the proof we see that the constructed finite model $\mathfrak{B}$ is in fact a submodel of the arbitrary model $\mathfrak{A}$ that we started with. Thus we have in fact established a stronger result than the finite model property, namely the *finite submodel property* of the monadic class: every infinite model of a sentence in the monadic class has a finite substructure which is also a model of that sentence.

In general it need not be the case that classes with the FMP also have the finite submodel property.

1.6 The two-variable fragment of FO

We denote relational first-order logic over k variables by FO^k, i.e.

$$\mathrm{FO}^k := \{\varphi \in \mathrm{FO} : \varphi \text{ relational, } \varphi \text{ only contains } k \text{ variables}\}.$$

We have shown that the Kahr-Moore-Wang class KMW, and hence also FO^3, are conservative reduction classes. We now prove that FO^2 has the finite model property and is thus decidable. Note that FO^k formulae are not necessarily in prenex normal form. A further motivation for the study of FO^2 is that propositional modal logic can be viewed as a fragment of FO^2 (in fact ML can be proven to be precisely the bisimulation invariant fragment of FO^2).

Before we proceed to prove the finite model property for FO^2, as a first step we establish a normal form for formulae in FO^2.

Lemma 1.37 (Scott). For each sentence $\psi \in \mathrm{FO}^2$ one can construct in polynomial time a sentence $\varphi \in \mathrm{FO}^2$ of the form

$$\varphi := \forall x \forall y \alpha \wedge \bigwedge_{i=1}^{n} \forall x \exists y \beta_i$$

such that $\alpha, \beta_1, \dots, \beta_n$ are quantifier free and such that ψ and φ are satisfiable over the same universe. Moreover, we have $|\varphi| = \mathcal{O}(|\psi| \cdot \log|\psi|)$.

Proof. First of all, we can assume that formulae $\varphi \in \mathrm{FO}^2$ only contain unary and binary relation symbols. This is no restriction since relations of higher arity can be substituted by introducing new binary and unary relation symbols. For example, if R is a relation of arity three, one could add a unary relation R_x and three binary relations $R_{x,x,y}$, $R_{x,y,x}$ and $R_{x,y,y}$ and replace each atom $R(x,x,x)$ (or $R(y,y,y)$) by $R_x(x)$ (or $R_x(y)$) and atoms as $R(x,x,y)$ or $R(x,y,x)$ by $R_{x,x,y}(x,y)$ and $R_{x,y,x}(x,y)$ respectively. By adding appropriate new subformulae one can ensure that the semantics are preserved, i.e. that the newly introduced relations partition a ternary relation in the intended sense. For example we would introduce as a new subformula $\forall x(R_x(x) \leftrightarrow R_{x,x,y}(x,x))$.

With ψ containing at most binary relations, we iterate the following steps until ψ has the desired form. We choose a subformula $Qy\eta$ of ψ ($Q \in \{\forall, \exists\}, \eta$ quantifier free) and add a new unary relation R:

$$\begin{aligned} \psi' &:= \psi[Qy\eta/Rx] \\ \psi &\mapsto \psi' \wedge \forall x(Rx \leftrightarrow Qy\eta). \end{aligned}$$

R captures those x that satisfy $Qy\eta$. The resulting formula φ is not yet of the desired form, but it is equivalent to the following:

(a) if $Q = \exists$, then

$$\varphi \equiv \psi' \wedge \forall x \forall y(\eta \rightarrow Rx) \wedge \forall x \exists y(Rx \rightarrow \eta)$$

(b) else if $Q = \forall$, then

$$\varphi \equiv \psi' \wedge \forall x \forall y(Rx \rightarrow \eta) \wedge \forall x \exists y(\eta \rightarrow Rx)$$

Now use that conjunctions of $\forall\forall$-formulae are equivalent to a $\forall\forall$-formula and obtain $\psi \equiv \forall x \forall y \alpha \wedge \bigwedge_{i=1}^{n} \forall x \exists y \beta_i$. Q.E.D.

Theorem 1.38. FO^2 has the finite model property. In fact, every satisfiable formula $\psi \in \mathrm{FO}^2$ has a model with at most $2^{|\psi|}$ elements.

Proof. The proof strategy is as follows: we start with a model $\mathfrak{A}$ of ψ and proceed by constructing a new model $\mathfrak{B}$ of ψ such that $|\mathfrak{B}| \leq 2^{\mathcal{O}(|\psi|)}$. For the construction the following definitions will be essential.

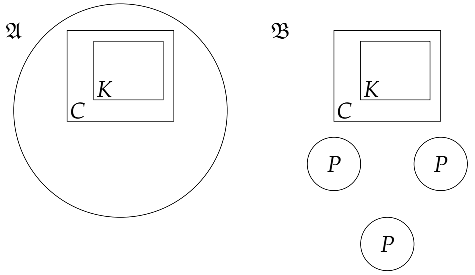
An element $a \in A$ is said to be a *king of* $\mathfrak{A}$ if its atomic 1-type is unique in $\mathfrak{A}$, i.e. if $\mathrm{atp}_{\mathfrak{A}}(b) \neq \mathrm{atp}_{\mathfrak{A}}(a)$ for all $b \neq a$. We let

- $K := \{a \in A : a \text{ is a king of } \mathfrak{A}\}$ be the set of kings of $\mathfrak{A}$, and
- $P := \{\mathrm{atp}_{\mathfrak{A}}(a) : a \in A, a \notin K\}$ be the set of atomic 1-types which are realized at least twice in $\mathfrak{A}$.

Since $\mathfrak{A} \models \forall x \exists y \beta_i$ for $i = 1, \dots, n$, there exist (Skolem) functions $f_1, \dots, f_n : A \rightarrow A$ such that $\mathfrak{A} \models \beta_i(a, f_i a)$ for all $a \in A$. The *court* of $\mathfrak{A}$ is defined as

$$C := K \cup \{f_i k : k \in K, i = 1, \dots, n\}.$$

Let $\mathfrak{C}$ be the substructure of $\mathfrak{A}$ induced by C. We construct a model $\mathfrak{B} \models \psi$ with universe $B = C \cup (P \times \{1, \dots, n\} \times \{0, 1, 2\})$.



To specify $\mathfrak{B}$ we set $\mathfrak{B}|_C = \mathfrak{C}$ and for all other elements we specify the 1- and 2-types (in this way fixing $\mathfrak{B}$ on the remaining part). However,

(1) This must be done consistently:
  - $\mathrm{atp}_{\mathfrak{A}}(b, b')$ and $\mathrm{atp}_{\mathfrak{A}}(b, b'')$ must agree on $\mathrm{atp}_{\mathfrak{A}}(b)$, and
  - $\gamma(x, y) \in \mathrm{atp}_{\mathfrak{B}}(b, b') \Leftrightarrow \gamma(y, x) \in \mathrm{atp}_{\mathfrak{B}}(b', b)$.

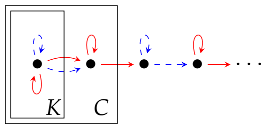
(2) Of course we have to ensure that $\mathfrak{B} \models \psi$.

We illustrate the construction with the following example.

Example 1.39. Consider the formula ψ over the signature $\tau = \{R, B\}$ (red edges and blue edges).

$$\begin{aligned}
\psi \quad = \quad & \exists x (Rxx \wedge Bxx) \\
\wedge \quad & \forall x \forall y ((Rxx \wedge Bxx \wedge Ryy \wedge Byy \to x = y) \\
& \wedge (Rxx \vee Bxx) \\
& \wedge (Rxy \wedge Ryx \to x = y) \\
& \wedge (Bxy \wedge Byx \to x = y) \\
& \wedge (Bxy \wedge x \neq y \to Ryy)) \\
\wedge \quad & \forall x \exists y (x \neq y \wedge (Rxx \to Rxy) \\
\wedge \quad & (Bxx \to Bxy)).
\end{aligned}$$

Let $\mathfrak{A} \models \psi$, then $\mathfrak{A}$ looks like follows:



In this case $P = \{\{Rxx, \neg Bxx\}, \{\neg Rxx, Bxx\}\}$ and the universe of $\mathfrak{B}$ is $B = C \cup (P \times \{1\} \times \{0, 1, 2\})$.

We proceed to construct $\mathfrak{B}$ by specifying the 1-types and 2-types of its elements as follows.

(1) The atomic 1-types of elements (p, i, j) are set to $\mathrm{atp}_{\mathfrak{B}}((p, i, j)) = p$.

(2) The atomic 2-types $\mathrm{atp}_{\mathfrak{B}}(b, b')$ will be set so that $\mathfrak{B} \models \forall x \exists y \beta_i$ for $i = 1, \dots, m$.
Choose for each $p \in P$ an element $h(p) \in A$ with $\mathrm{atp}_{\mathfrak{A}}(h(p)) = p$.
Find for each $b \in \mathfrak{B}$ and each i a suitable element b' such that $\mathfrak{B} \models \beta_i(b, b')$ (by defining $\mathrm{atp}_{\mathfrak{B}}(b, b')$ appropriately).

  (a) If b is a king, set $b' := f_i(b) \in C \subseteq B$. Then $\mathfrak{B} \models \beta_i(b, b')$.

  (b) If $b \in C \setminus K$ (non-royal member of the court), distinguish:
    - If $f_i(b) \in K$, then set $b' := f_i(b) \in K \subseteq B$.
    - Otherwise it holds that $\mathrm{atp}_{\mathfrak{A}}(f_i(b)) = p \in P$.
      In this case, set $b' := (p, i, 0)$. Now set $\mathrm{atp}_{\mathfrak{B}}(b, b') := \mathrm{atp}_{\mathfrak{A}}(b, f_i(b))$. Thus $\mathfrak{B} \models \beta_i(b, b')$ since $\mathfrak{A} \models \beta_i(b, f_i(b))$.

  (c) If $b = (p, j, \ell)$ for some $p \in P, j \in \{1, \dots, n\}, \ell \in \{0, 1, 2\}$, let $a := h(p)$ and consider $f_i(a)$.
  If $f_i(a) \in K$, set $b' = f_i(a)$ and $\mathrm{atp}_{\mathfrak{B}}(b, b') := \mathrm{atp}_{\mathfrak{A}}(a, b')$.
  If $f_i(a) \notin K$, then $\mathrm{atp}_{\mathfrak{A}}(f_i(a)) = p' \in P$.
  Set $b' := (p', i, (\ell + 1) \pmod 3)$.
  Then set $\mathrm{atp}_{\mathfrak{B}}(b, b') := \mathrm{atp}_{\mathfrak{A}}(a, f_i(a))$, and thus $\mathfrak{B} \models \beta_i(b, b')$.

To complete the construction of $\mathfrak{B}$, let $b_1, b_2 \in B$ be such that $\mathrm{atp}_{\mathfrak{B}}(b_1, b_2)$ is not yet specified. Choose $a_1, a_2 \in A$ so that

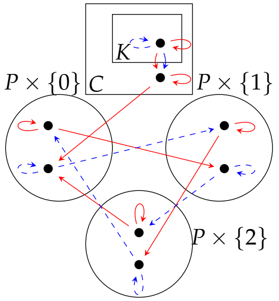
$$\begin{aligned}
\mathrm{atp}_{\mathfrak{A}}(a_1) &= \mathrm{atp}_{\mathfrak{B}}(b_1) \text{ and} \\
\mathrm{atp}_{\mathfrak{A}}(a_2) &= \mathrm{atp}_{\mathfrak{B}}(b_2)
\end{aligned}$$

and set

$$\mathrm{atp}_{\mathfrak{B}}(b_1,b_2) := \mathrm{atp}_{\mathfrak{A}}(a_1,a_2).$$

Since $\mathfrak{A} \models \alpha(a_1,a_2)$, also $\mathfrak{B} \models \alpha(b_1,b_2)$.

For the previously considered example, $\mathfrak{B}$ looks as follows:



Overall, we obtain $\mathfrak{B} \models \forall x \forall y \alpha \wedge \bigwedge_{i=1}^{n} \forall x \exists y \beta_i = \psi$, and the size of B is restricted by

$$|B| = \underbrace{|C|}_{\leq |K|(n+1)} + 3n|P| = \mathcal{O}(n \cdot \#\,(\text{atomic 1-types}))\,.$$

For k relation symbols, there are 2^k atomic 1-types, hence $|B| = 2^{\mathcal{O}(|\psi|)}$.

Q.E.D.

This result implies that $Sat(\mathrm{FO}^2)$ is in NEXPTIME (indeed it is NEXPTIME-complete), since we can simply guess a finite structure $\mathfrak{A}$ of exponential size (in the length of ψ) and verify that $\mathfrak{A} \models \psi$.

Corollary 1.40. $Sat(\mathrm{FO}^2) \in \mathrm{NEXPTIME} = (\bigcup_k \mathrm{NTIME}(2^{n^k}))$.

This is a typical complexity level for decidable fragments of FO. In fact, $Sat(\mathrm{FO}^2)$ is even complete for NEXPTIME. For showing this, we reduce a bounded version of the domino problem to $Sat(\mathrm{FO}^2)$.

Definition 1.41. Let $\mathcal{D} = (D,H,V)$ be a domino system and let $Z(t)$ denote $\mathbb{Z}/t\mathbb{Z} \times \mathbb{Z}/t\mathbb{Z}$. For a word $w = w_0,\dots,w_{n-1} \in D^n$ we say that $\mathcal{D}$ tiles $Z(t)$ with initial condition w if there is $\tau : Z(t) \to D$ such that

- if $\tau(x,y) = d$ and $\tau(x+1,y) = d'$ then $(d,d') \in H$ for all $(x,y) \in Z(t)$,
- if $\tau(x,y) = d, \tau(x,y+1) = d'$ then $(d,d') \in V$ for all $(x,y) \in Z(t)$ and
- $\tau(i,0) = w_i$ for all $i = 0,\dots,n-1$.

Let $\mathcal{D}$ be a domino system and $T : \mathbb{N} \to \mathbb{N}$ a mapping. Define

$$\mathrm{DOMINO}(\mathcal{D},T) := \{w \in D^* : \mathcal{D} \text{ tiles } Z(T(|w|)) \text{ with initial condition } w\}\,.$$

One can describe computations of a (in this case non-deterministic) Turing machine by domino tilings in such a way that the input condition of the domino problem relates to the initial configuration of the Turing machine. The restrictions on the size of the tiled rectangle correspond to the time and space restrictions of the Turing machine. To prove that a problem A is NEXPTIME-hard, it then suffices to show that $\mathrm{DOMINO}(\mathcal{D},2^n) \leq_p A$.

Our goal is to show that $\mathrm{DOMINO}(\mathcal{D},2^n)$ reduces to $Sat(X)$ for relatively simple classes $X \subseteq \mathrm{FO}$. Set

$$X = \{\varphi \in \mathrm{FO}^2 : \varphi = \forall x \forall y\, \alpha \wedge \forall x \exists y\, \beta, \text{ s.t. } \alpha,\, \beta \text{ quantifier-free, without } =, \text{ and with only monadic predicates}\}\,.$$

We show that $Sat(X)$ is NEXPTIME-complete and hence also $Sat(\mathrm{FO}^2)$ is NEXPTIME-complete.

Lemma 1.42. For each domino system $\mathcal{D} = (D,H,V)$ there exists a polynomial time reduction $w \in D^n \mapsto \psi_w \in X$ such that $\mathcal{D}$ tiles $Z(2^n)$ with initial condition w if and only if ψ_w is satisfiable.

Proof. The intended model of ψ_w is a description of a tiling $\tau : Z(2^n) \to D$ in the universe $Z(2^n)$.

Let $z = (a,b) \in Z(2^n)$ with $a = \sum_{i=0}^{n-1} a_i 2^i$ and $b = \sum_{i=0}^{n-1} b_i 2^i$. Encode the tuple as $(a_o,\dots,a_{n-1},b_0,\dots,b_{n-1}) \in \{0,1\}^{2n}$.

To encode the tiling, we define ψ_w with the monadic predicates X_i, X_i^*, Y_i, Y_i^*, N_i for $0 \leq i < n$ and $P_d (d \in D)$ with the following intended meaning:

$$\begin{array}{rll} X_i z & \text{iff} & a_i = 1. \\ X_i^* z & \text{iff} & a_j = 1 \text{ for all } j < i. \\ Y_i z & \text{iff} & b_j = 1. \\ Y_i^* z & \text{iff} & b_j = 1 \text{ for all } j < i. \\ N_i z & \text{iff} & z = (i, 0). \\ P_d z & \text{iff} & \tau(z) = d. \end{array}$$

ψ_w will have the form $\psi_w = \forall x \forall y \alpha \wedge \forall x \exists y \beta$, where β accounts for the correct interpretation of $X_i, X_i^*, Y_i, Y_i^*, N_i$ and ensures that every element has a successor, and α accounts for the description of a correct tiling.

Now β is the the following formula:

$$\begin{aligned} \beta = {} & X_0^* x \wedge Y_0^* x \\ \wedge \quad & \bigwedge_{i=1}^{n-1} X_i^* x \leftrightarrow (X_{i-1}^* x \wedge X_{i-1} x) \\ \wedge \quad & \bigwedge_{i=1}^{n-1} Y_i^* x \leftrightarrow (Y_{i-1}^* x \wedge Y_{i-1} x) \\ \wedge \quad & \bigwedge_{i=0}^{n-1} X_i y \leftrightarrow (X_i x \oplus X_i^* x) \\ \wedge \quad & \bigwedge_{i=0}^{n-1} Y_i y \leftrightarrow (Y_i x \oplus (Y_i^* x \wedge X_{n-1} x \wedge X_{n-1}^* x)) \\ \wedge \quad & N_0 x \leftrightarrow (\bigwedge_{i=0}^{n-1} \neg X_i x \wedge \neg Y_i x) \\ \wedge \quad & \bigwedge_{i=0}^{n-1} N_i x \leftrightarrow N_{i+1} y. \end{aligned}$$

We define the following shorthands for use in α:

$$H(x,y) \quad := \quad \bigwedge_{i=0}^{n-1} (Y_i y \leftrightarrow Y_i x) \wedge \bigwedge_{i=0}^{n-1} (X_i y \leftrightarrow (X_i x \oplus X_i^* x))$$

$$V(x,y) \quad := \quad \bigwedge_{i=0}^{n-1} (X_i y \leftrightarrow X_i x) \wedge \bigwedge_{i=0}^{n-1} (Y_i y \leftrightarrow (Y_i x \oplus Y_i^* x)).$$

Now α is defined to be

$$\begin{aligned} \alpha = {} & \bigwedge_{d \neq d'} \neg (P_d x \wedge P_{d'} x) \\ \wedge \quad & (H(x,y) \rightarrow \bigvee_{(d,d') \in H} (P_d x \wedge P_{d'} y)) \\ \wedge \quad & (V(x,y) \rightarrow \bigvee_{(d,d') \in V} (P_d x \wedge P_{d'} y)) \\ \wedge \quad & (\bigwedge_{i=i}^{n-1} (N_i x \rightarrow P_{w_i} x)). \end{aligned}$$

Claim 1.43. ψ_w is satisfiable if and only if $\mathcal{D}$ tiles $Z(2^n)$ with initial condition w.

Proof. We show both directions.

($\Leftarrow$) Consider the intended model, ψ_w holds in it.
($\Rightarrow$) Consider $\mathfrak{C} = (C, X_1, \ldots) \models \psi_w$ and define a mapping

$$\begin{aligned} f: \; & C && \rightarrow Z(2^n) \\ & c && \mapsto (a, b) \equiv (a_0, \ldots, a_{n-1}, b_0, \ldots, b_{n-1}) \end{aligned}$$

$$\begin{array}{rll} \text{with } a_i = 1 & \text{iff} & \mathfrak{C} \models X_i c \quad \text{and} \\ b_i = 1 & \text{iff} & \mathfrak{C} \models Y_i c. \end{array}$$

As $\mathfrak{C} \models \forall x \exists y \beta$, f is surjective. Choose for each $z \in Z(2^n)$ an element $c \in f^{-1}(z)$ and set $\tau(z) = d$ for the unique d that satisfies $\mathfrak{C} \models P_d c$. Then τ is a correct tiling with initial condition w. Q.E.D.

Since the length of ψ_w is $|\psi_w| = O(n \log n)$, the above claim completes the proof of the lemma. Q.E.D.