On Provability Logic

Vítězslav Švejdar[*][†]

April 1, 2000

The original publication is available at NJPL.

Abstract This is an introductory paper about provability logic, a modal propositional logic in which necessity is interpreted as formal provability. We discuss the ideas that led to establishing this logic, we survey its history and the most important results, and we emphasize its applications in metamathematics. Stress is put on the use of Gentzen calculus for provability logic. We sketch our version of a decision procedure for provability logic and mention some connections to computational complexity.

1 Introduction

Some logical concepts have unique and non-problematic meaning. For instance, there are no logical schools forcing their own definition of *algorithm*. Various definitions of algorithm appeared to be equivalent and the notion of algorithm seems to be *absolute*. More or less the same can be said about the notion of *proof*. Of course, there are non-classical logics and a proof acceptable from the classical point of view may be non-acceptable e.g. from intuitionistic point of view. But, in the classical framework or after the logical framework has been fixed, there are no doubts about what a correct proof is. The notion of proof can also be claimed to be absolute.

On the other hand, there is no generally accepted definition e.g. of an *efficient algorithm*. Efficient algorithms can be identified with those working in polynomial time, but other (non-equivalent) definitions may also be useful. The notion of *modality* is even "less absolute". Tens of non-equivalent modal logics can be found e.g. in the book [HC96], and none of them seems to play the role of the most important or the authors' favorite.

The non-existence of "the modal logic" can be explained by the fact that *nested modalities* are rare in natural language. We seldom say that it is necessary

[*] The work on this paper was supported partially by a grant GA CR 401/98/0383 and partially by a grant 162/97 of Charles University

[†] Charles University, Prague, vitezslavdotsvejdaratcunidotcz, http://www1.cuni.cz/~svejdar/. Palachovo nám. 2, 116 38 Praha 1, Czech Republic.

that something is possible and thus there is no agreement whether for instance the modal propositional formula $\Diamond p \to \Box\Diamond p$ should be accepted as a modal tautology.

This paper is devoted to the *provability logic* which is a modal propositional logic based on the idea that something is necessary if it can be *proved*. By provability we mean provability in some fixed sufficiently strong formal axiomatic theory like Peano arithmetic PA. If T is sufficiently strong, then logical syntax including the notion of provability can be formalized within T. This fact means that we can ask which facts about provability in T (or in some other theory S) are provable in T itself.

Thus provability logic is a logic in which necessity is understood as formal provability and which is closely connected to and can have application in metamathematics of the most important mathematical theories. It should be clear from the words "provability" and "provable" in the last sentence of the previous paragraph that considering nested modalities is very natural in provability logic. Provability logic is not intended to help in investigating modalities in natural language. But surprisingly, usual modal methods can be used in its study and it shares some properties (and axioms and rules) with the traditional modal systems like S4.

This paper is meant as an introductory paper; its purpose is to encourage the reader to take an interest in the field. We survey and comment the (from our point of view) most interesting facts and ideas about provability logic and mention their applications in metamathematics. We put stress on the use of Gentzen calculus for provability logic invented in [SV82]. We present no new results and we omit all difficult proofs; but we do show some proofs. Most of the material can be found e.g. in [Smo85], [Smo84] and [Boo93]. In the final part of the paper we discuss our version of a decision procedure for provability logic and mention some connections to computational complexity.

I thank Roy Dyckhoff for bringing my attention to the paper [SV82]. Further I thank the unknown referees for useful remarks and comments on the preliminary version of this paper.

2 Arithmetization of logical syntax

Peano arithmetic PA is an axiomatic theory formulated in *arithmetical language*, for which one can take the set $\{+, \cdot, 0, \mathrm{S}, <\}$ containing two binary function symbols, one constant, one unary function symbol and one binary predicate. The underlying logic is the classical predicate logic with equality. The set of *axioms* of PA is usually specified as that containing several simple axioms (e.g. $\forall x(x \cdot 0 = 0)$) and the induction scheme: each sentence of the form

$$\text{Ind:} \qquad \forall y_1 \,.\,.\, \forall y_n(\varphi(0, \underline{y}) \;\&\; \forall x(\varphi(x, \underline{y}) \to \varphi(\mathrm{S}(x), \underline{y})) \to \forall x \varphi(x, \underline{y})),$$

where $\underline{y}$ is an abbreviation for $y_1, .\,.\,, y_n$, is an axiom.

The structure $\mathbb{N} = \langle \mathrm{N}, +^{\mathbb{N}}, \cdot^{\mathbb{N}}, 0^{\mathbb{N}}, \mathrm{S}^{\mathbb{N}}, <^{\mathbb{N}} \rangle$, where N is the set $\{0, 1, 2, \ldots\}$ of all natural numbers, $+^{\mathbb{N}}$ and $\cdot^{\mathbb{N}}$ addition and multiplication of natural numbers, $<^{\mathbb{N}}$ the strict ordering on natural numbers, $0^{\mathbb{N}}$ the number zero and $\mathrm{S}^{\mathbb{N}}$ the successor function $x \mapsto x + 1$, is called the *standard model (of arithmetic)*.

The axioms of PA allow one to prove general facts about natural numbers, like $\forall x \forall y(x \cdot y = y \cdot x)$, and also facts about particular numbers. For instance the sentence

$$\forall x \forall y(x \cdot y = \mathrm{S}(\mathrm{S}(\mathrm{S}(0))) \;\to\; x = \mathrm{S}(0) \;\vee\; y = \mathrm{S}(0)) \qquad \text{(A)}$$

(which is provable in PA) can be read $\mathsf{the\ number\ three\ is\ a\ prime}$. The term $\mathrm{S}(\mathrm{S}(\mathrm{S}(0)))$ is denoted $\overline{3}$. More generally, the n-th *numeral* is defined as the term $\mathrm{S}(\mathrm{S}\,.\,.\,(0)\,.\,.)$ with n occurrence of the symbol S.

As an exercise we suggest the reader to formulate the fact that $\mathsf{there\ are\ infinitely\ many\ powers\ of\ two}$ in the arithmetical language. The solution is not completely trivial but requires no theoretical knowledge. With some further effort, the resulting sentence can be proved in PA.

We see that some notion (like divisibility or a power of two) can be expressible in the arithmetical language even if no symbol of the language directly corresponds to it. If it is the case a natural question is what Peano arithmetic can prove about that notion.

Terms, formulas, proofs and other syntactical objects are finite sequences of symbols. Since the nature of symbols is irrelevant, these objects can be identified with finite sequences of natural numbers. Any finite sequence of natural numbers can be coded by a single natural number. Thus formulas and other syntactical objects can be identified with natural numbers: they *are* natural numbers, and numerals like $\overline{0 = \mathrm{S}(0)}$ or $\overline{\mathsf{three\ is\ a\ prime}}$ are possible. Here we continue the practice of hiding arithmetical formulas under their informal reading typeset in sans serif font. Thus the sentence (A) above equals to some natural number n (i.e. has numerical code n), and $\overline{\mathsf{three\ is\ a\ prime}}$ is a numeral containing exactly n occurrences of "S". Coding of finite sequences of natural numbers (can be chosen so that it) is definable inside Peano arithmetic. Thus inside Peano arithmetic we can work with syntactical objects and attempt to prove their properties. This is often expressed by saying that logical syntax is arithmetizable.

The result of arithmetization of logical syntax is that not only the notion of a prime or a power of two, but also the notion of PA-proof is expressible in arithmetical language. *Proof predicate* is an arithmetical formula $\mathsf{Prf}(y, x)$ expressing that $\mathsf{the\ number}\ x\ \mathsf{is\ a\ sentence\ and\ the\ number}\ y\ \mathsf{is\ its\ proof\ in\ PA}$ (of course, we do not claim that y is uniquely determined by x; a sentence can have many different proofs). Having the proof predicate, the formula $\mathsf{Pr}(x)$ is defined as $\exists y \mathsf{Prf}(y, x)$ and the sentence Con is defined as $\neg\mathsf{Pr}(\overline{0 = \mathrm{S}(0)})$. The formula $\mathsf{Pr}(x)$ expresses that $\mathsf{the\ number}\ x\ \mathsf{is\ a\ sentence\ provable\ in\ PA}$. Since $\neg(0 = \mathrm{S}(0))$ is easily proved in PA, the sentence $0 = \mathrm{S}(0)$ represents a contradiction, and the sentence Con is the consistency statement, saying that Peano arithmetic is not contradictory. Important properties of the formula $\mathsf{Pr}(x)$ are expressed by the *(Löb) derivability conditions*:

D1: If $\mathrm{PA} \vdash \varphi$, then $\mathrm{PA} \vdash \mathsf{Pr}(\overline{\varphi})$

D2: $\mathrm{PA} \vdash \mathsf{Pr}(\overline{\varphi \to \psi}) \to (\mathsf{Pr}(\overline{\varphi}) \to \mathsf{Pr}(\overline{\psi}))$

D3: $\mathrm{PA} \vdash \mathsf{Pr}(\overline{\varphi}) \to \mathsf{Pr}(\overline{\mathsf{Pr}(\overline{\varphi})})$,

where $\vdash$ denotes provability and φ and ψ are arbitrary arithmetical sentences. An example on the use of D1 is this: the sentence $\mathsf{Pr}(\overline{\textsf{three is a prime}})$ is provable in PA. The condition D2 says that PA knows that provable sentences are closed on the rule modus ponens. Hence D2 is a formalized version of modus ponens. Similarly, D3 is a formalized version of D1. Both D1 and D3 express (on different levels) that if something is provable then it is provable that it is provable. Besides D1–D3 we shall need another property of the provability predicate, namely

Def: $\mathrm{PA} \vdash \varphi$ if and only if $\mathbb{N} \models \mathsf{Pr}(\overline{\varphi})$,

expressing that the formula $\mathsf{Pr}(x)$ *defines* the set of all PA-provable sentences in the standard model. Note that the implication $\Rightarrow$ in the condition Def follows from D1 and from the fact that $\mathbb{N}$ is a model of PA. Also note that, by a similar argument and for any sentence φ, the implication $\mathrm{PA} \vdash \mathsf{Pr}(\overline{\varphi}) \Rightarrow \mathrm{PA} \vdash \varphi$ follows from the condition Def.

A useful tool for proving metamathematical properties of Peano arithmetic is the *self-reference theorem*: for each arithmetical formula $\psi(x,y)$ (containing no free variables except possibly x and y) there is an arithmetical sentence φ such that the equivalence $\varphi \equiv \psi(\overline{\varphi}, \overline{\neg\varphi})$ is provable in PA. In other words, the self-reference theorem says that any equation of the form $\mathrm{PA} \vdash \varphi \equiv \psi(\overline{\varphi}, \overline{\neg\varphi})$, for an unknown sentence φ, has a solution. It is not necessary that *both* variables x and y appear free in ψ. So all equations of the form $\mathrm{PA} \vdash \varphi \equiv \psi(\overline{\varphi})$ or $\mathrm{PA} \vdash \varphi \equiv \psi(\overline{\neg\varphi})$ also have a solution. If $\mathrm{PA} \vdash \varphi \equiv \psi(\overline{\varphi})$ then, inside PA, we know that φ is equivalent to the statement the sentence $\overline{\varphi}$ has the property ψ. So the solution of the equation $\mathrm{PA} \vdash \varphi \equiv \psi(\overline{\varphi})$ can be viewed as a sentence saying I have the property ψ.

Let us list some of the most prominent examples on the use of the self-reference theorem. *Gödel sentence* is a sentence ν satisfying $\mathrm{PA} \vdash \nu \equiv \neg\mathsf{Pr}(\overline{\nu})$, *Rosser sentence* is a sentence ρ satisfying

$$\mathrm{PA} \vdash \rho \equiv \forall y(\mathsf{Prf}(y, \overline{\rho}) \to \exists v {\leq} y\, \mathsf{Prf}(v, \overline{\neg\rho})), \tag{B}$$

and *Henkin sentence* is a sentence κ satisfying $\mathrm{PA} \vdash \kappa \equiv \mathsf{Pr}(\overline{\kappa})$. The sentences κ and ν say I am provable in PA and I am not provable in PA respectively. The sentence ρ says below any proof of myself there is another proof of my negation.

Proposition 1 *If* $\mathrm{PA} \vdash \nu \equiv \neg\mathsf{Pr}(\overline{\nu})$*, then both* ν *and* $\neg\nu$ *are unprovable in* PA.

Proof Assume that $\mathrm{PA} \vdash \nu$. Then, by the condition D1, $\mathrm{PA} \vdash \mathsf{Pr}(\overline{\nu})$. From $\mathrm{PA} \vdash \nu \equiv \neg\mathsf{Pr}(\overline{\nu})$ we have $\mathrm{PA} \vdash \neg\nu$. A contradiction with the consistency of PA.

So $\mathrm{PA} \nvdash \nu$. Hence, by the condition Def, $\mathbb{N} \not\models \mathsf{Pr}(\overline{\nu})$. Assume that $\mathrm{PA} \vdash \neg\nu$. Then, again by the equivalence $\mathrm{PA} \vdash \nu \equiv \neg\mathsf{Pr}(\overline{\nu})$, we have $\mathrm{PA} \vdash \mathsf{Pr}(\overline{\nu})$. Since $\mathbb{N}$ is a model of PA, we have $\mathbb{N} \models \mathsf{Pr}(\overline{\nu})$. This is a contradiction with $\mathbb{N} \not\models \mathsf{Pr}(\overline{\nu})$ proved above. So $\mathrm{PA} \nvdash \neg\nu$. ■

Proposition 1 is basically Gödel first incompleteness theorem, while Proposition 2 below is Gödel second incompleteness theorem.

Proposition 2 *(a) If* $\mathrm{PA} \vdash \nu \equiv \neg\mathsf{Pr}(\overline{\nu})$*, then* $\mathrm{PA} \vdash \mathsf{Con} \equiv \nu$*. So* $\mathrm{PA} \nvdash \mathsf{Con}$*.*
(b) $\mathrm{PA} \vdash \mathsf{Con} \to \neg\mathsf{Pr}(\overline{\mathsf{Con}})$*.*

We are not going to prove Proposition 2. We only make some comments to make the statements plausible. Note that the argument in the first paragraph of proof of Proposition 1, showing that if $\mathrm{PA} \vdash \nu$ then PA is contradictory, is purely syntactical in the sense that it does not mention the structure $\mathbb{N}$ and the condition Def. So it should not be surprising that this argument can be formalized inside PA, and it can be checked that D1–D3 are sufficient to do this. The result is a proof of a sentence if PA is consistent then $\mathrm{PA} \nvdash \nu$, i.e. a proof of the sentence $\mathsf{Con} \to \neg\mathsf{Pr}(\overline{\nu})$. Together with $\mathrm{PA} \vdash \nu \equiv \neg\mathsf{Pr}(\overline{\nu})$ this yields $\mathrm{PA} \vdash \mathsf{Con} \to \nu$. The converse implication $\mathrm{PA} \vdash \nu \to \mathsf{Con}$ can also be proved using D1–D3 and its proof is even simpler. Having $\mathrm{PA} \vdash \mathsf{Con} \to \nu$, the conclusion $\mathrm{PA} \nvdash \mathsf{Con}$ is straightforward using Proposition 1. From $\mathrm{PA} \vdash \mathsf{Con} {\to} \nu$ we have $\mathrm{PA} \vdash \mathsf{Pr}(\overline{\mathsf{Con} \to \nu})$ by D1, and $\mathrm{PA} \vdash \mathsf{Pr}(\overline{\mathsf{Con}}) \to \mathsf{Pr}(\overline{\nu})$ by D2. So $\mathrm{PA} \vdash \neg\mathsf{Pr}(\overline{\nu}) \to \neg\mathsf{Pr}(\overline{\mathsf{Con}})$. This together with $\mathrm{PA} \vdash \mathsf{Con} \to \neg\mathsf{Pr}(\overline{\nu})$ yields the statement of (b). Note that in Proposition 2(b) the formalization went one step deeper: the sentence in (b) can be viewed as a formalization of the statement in (a).

So PA is incomplete, cannot prove its own consistency, but it knows (i.e. can prove) about itself that it can prove its consistency only if it is contradictory.

The provability predicate and the Gödel sentence asserting its own unprovability can be constructed for any recursively axiomatizable consistent theory T extending Peano arithmetic, and the Gödel sentence can be proved to be unprovable in T and equivalent in T to the consistency statement of T. So no such T can prove its own consistency. The Gödel sentence for T (and the consistency statement for T) can be proved independent over T under the additional assumption that T is *sound*, i.e. that all arithmetical sentences provable in T hold in $\mathbb{N}$.

The Rosser sentence ρ can also be proved independent over PA, and its advantage is that both proofs (of $\mathrm{PA} \nvdash \rho$ and $\mathrm{PA} \nvdash \neg\rho$) are purely syntactical. Hence both are formalizable in Peano arithmetic: $\mathrm{PA} \vdash \mathsf{Con} \to \neg\mathsf{Pr}(\overline{\rho})$ and $\mathrm{PA} \vdash \mathsf{Con} \to \neg\mathsf{Pr}(\overline{\neg\rho})$. The Rosser sentence of a theory T (constructed from the proof predicate of T) can be used to show that any consistent recursively axiomatizable extension T of PA, even if not sound, is incomplete.

The self-reference theorem asserts that any self-referential equation has a solution, but does not say that the solution is *unique*. This is why no definite article appears in formulation of Proposition 1: the Proposition says and its proof shows that *any* solution of the Gödel equation is independent over PA. But later, in Proposition 2, it became clear that any Gödel sentence ν is PA-equivalent to the sentence Con. So the Gödel equation in fact has a unique solution (up to PA-provable equivalence). Is this always the case, i.e. do all self-reference equations have a unique solution? Can PA prove the sentence $\mathsf{Con} \rightarrow \neg\mathsf{Pr}(\overline{\neg\nu})$, i.e. does PA know that ν is independent over it? What are the properties ot the Henkin sentence? We shall see that modal logic can throw some light on questions like these.

The first incompleteness theorem was proved in [Göd31]. C. Smoryński writes in [Smo85] about the intended paper with roman "II" that it never materialized, perhaps because the logical community accepted the second incompleteness theorem before Gödel succeeded in typing the paper. The self-reference theorem first appeared explicitly in [Car34]. As introductory texts about arithmetization of logical syntax we recommend [Fef60], [Smo85] or [Boo93]. The derivability conditions were formulated in [Löb55], where M. H. Löb solved the following problem of L. Henkin: is any sentence κ asserting its own provability provable in PA?

3 Arithmetical semantics of modal logic

Formulas of provability logic are the usual modal propositional formulas. They are built up from propositional atoms $p, q, \ldots, p_0, p_1, \ldots$ using the unary modal operator $\Box$ and logical connectives $\rightarrow$, $\neg$, $\&$, $\vee$, $\bot$. The symbol $\Box$ stands for necessity, $\Box A$ is read "A is necessary" or simply "box A". The symbol $\bot$ is a logical constant (i.e. an operator of arity zero) for falsity. Thus $\Box\bot \vee \Box\neg\bot$ or $p \rightarrow \Box(\Box q \;\&\; p)$ are examples of modal formulas. The choice of the set of logical connectives is not absolutely essential; the set $\{\rightarrow, \neg\}$ (just as in classical propositional logic) would also do. But we shall see that it is very convenient to have the symbol $\bot$. Other operators are also allowed: $\Diamond A$ is a shortening for $\neg\Box\neg A$, $\top$ is $\neg\bot$, and $A \equiv B$ stands for $(A \rightarrow B) \;\&\; (B \rightarrow A)$.

An *(arithmetical) evaluation* is a function e from propositional atoms to arithmetical sentences satisfying the following conditions:

- $e(\bot) = (0 = \mathrm{S}(0))$
- $e(\neg A) = \neg e(A)$, $e(A \bowtie B) = e(A) \bowtie e(B)$ for any (binary) logical connective $\bowtie$
- $e(\Box A) = \mathsf{Pr}(\overline{e(A)})$.

We mention two examples of how the arithmetical evaluations work. Under any evaluation e, the value of a formula $\Box A \vee \Box\neg A$ has the form $\mathsf{Pr}(\overline{\psi}) \vee \mathsf{Pr}(\overline{\neg\psi})$, where the sentence ψ is determined by values of propositional atoms occurring in A. The value of the formula $\neg\Box\bot$ is the same under any e because $\neg\Box\bot$ contains no propositional atoms, and we have $e(\neg\Box\bot) = \neg\mathsf{Pr}(\overline{0 = \mathrm{S}(0)}) = \mathsf{Con}$.

A modal formula A is a PA-*tautology* if PA $\vdash e(A)$ for each arithmetical evaluation e. For example, the formula $\Box A \rightarrow \Box\Box A$ is a PA-tautology for any choice of A because any its value has the form $\mathsf{Pr}(\overline{\varphi}) \rightarrow \mathsf{Pr}(\overline{\mathsf{Pr}(\overline{\varphi})})$ which, by D3, is always provable in PA. The formula $p \rightarrow (\Box q \rightarrow p)$ is a PA-tautology because, under any e, its value is an arithmetical sentence which is a (classical) propositional tautology, and all tautologies are provable in all axiomatic theories. This example generalizes to the following statement: any modal formula which (in an obvious sense) is a propositional tautology is also a PA-tautology.

Let us mention some less obvious examples of PA-tautologies and non-tautologies. By Propositions 1 and 2, neither $\neg\Box\bot$ nor $\Box\bot$ are PA-tautologies. By Proposition 2(b), $\neg\Box\bot \rightarrow \neg\Box\neg\Box\bot$ is a PA-tautology. Consider once again the Gödel sentence ν. The fact that $\neg\mathsf{Pr}(\overline{\nu}) \rightarrow \nu$ is provable has the consequence that $\mathsf{Pr}(\overline{\nu}) \rightarrow \nu$ is *not* provable because otherwise ν would also be provable. This argument shows that $\Box p \rightarrow p$ is *not* a PA-tautology: there is an evaluation e, namely that sending p to ν, such that PA $\nvdash e(\Box p \rightarrow p)$. The fact that $\Box p \rightarrow p$ is not a PA-tautology may look surprising, but it is natural: PA cannot claim that all provable statements are true because it knows that a contradictory theory proves any sentence and it cannot claim about itself that it is consistent.

We observe that PA-tautologies are modal formulas that express general facts about provability and about self-referential sentences. For instance the modal formula $\neg\Box\bot \rightarrow (\Box(p \equiv \neg\Box p) \rightarrow \neg\Box p)$, saying that "under the assumption of consistency, any statement asserting its own unprovability is unprovable", is a modal version of Gödel first incompleteness theorem. It is more or less evident from Proposition 1 and will become completely clear in the next Section that this formula is a PA-tautology. Similarly, the formula $\neg\Box\bot \rightarrow \neg\Box\neg\Box\bot$, which we know for sure to be a PA-tautology, is a modal version of the second incompleteness theorem.

At first sight it is not evident whether the set of all PA-tautologies is recursive. But usual methods used in (modal) logic can be used to solve this and other problems. More specifically, an axiomatization of the set of all PA-tautologies can be established and further studied.

4 Calculi, Kripke semantics and conclusions

To obtain an axiomatization of the set of all PA-tautologies, a good idea is to start with the modal versions of D1–D3, i.e. to accept all propositional tautologies and all formulas of the form $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$ and $\Box A \rightarrow \Box\Box A$

as axioms and to accept $A \,/\, \Box A$ and modus ponens $A, A \to B \,/\, B$ as rules of inference. The resulting modal logic is known as K4.

The logic K4 is evidently sound with respect to arithmetical semantics (i.e. proves only PA-tautologies) but it is not complete, i.e. does not prove *all* PA-tautologies. An example of a PA-tautology unprovable in K4 is $\neg\Box\bot \to \neg\Box\neg\Box\bot$. This is understandable because K4 (i.e. the conditions D1–D3 plus modus ponens plus the fact that all propositional tautologies are PA-tautologies) does not capture an important fact, namely the self-reference theorem.

In [Löb55] M. H. Löb proved that any Henkin sentence, i.e. any sentence κ satisfying $\mathrm{PA} \vdash \kappa \equiv \mathsf{Pr}(\overline{\kappa})$, is provable in PA. Actually Löb proved a stronger statement: if $\mathrm{PA} \vdash \mathsf{Pr}(\overline{\kappa}) \to \kappa$, then $\mathrm{PA} \vdash \kappa$. Formalized version of this statement (not mentioned in [Löb55]), saying that $\mathrm{PA} \vdash \mathsf{Pr}(\overline{\mathsf{Pr}(\overline{\kappa}) \to \kappa}) \to \mathsf{Pr}(\overline{\kappa})$, is also true. Löb theorem was originally considered a curiosity not related to anything else. But during the sixties it became clear that its modal version, i.e. the *Löb rule* and the *Löb axiom*

$$\Box A \to A \,/\, A \quad \text{and} \quad \Box(\Box A \to A) \to \Box A$$

could be important. The reader can convince her- or himself by deriving second incompleteness theorem from Löb axiom: a simple substitution of $\bot$ for A is sufficient. Both Löb rule and Löb axiom are sound w.r.t. arithmetical semantics and they are interderivable over K4. So we can officially define the *Hilbert calculus* for *provability logic* GL: it results by adding Löb axiom scheme to the logic K4 defined above. Provability logic is sometimes denoted only G or only L. The letters stand for Gödel and Löb.

More on the history of GL can be found in [BS91]. We only quote from [BS91] that an explicit formulation of Löb axiom first appeared in 1963 in a paper by T. Smiley on a modal treatment of ethics (!). By 1971 several people (Segerberg, Kripke, de Jongh) independently proved Kripke completeness theorem formulated in Propositions 8 and 9 below. And then, in the seventies, de Jongh in Amsterdam and Boolos and Kripke in the USA considered seriously the conjecture that Löb axiom is *the only* missing axiom, i.e. that GL is complete w.r.t. the arithmetical semantics. This was finally proved to be true by Solovay ([Sol76]) in 1975:

Proposition 3 *A modal formula is provable in* GL *if and only if it is a* PA-*tautology.*

Solovay's proof of Proposition 3 uses a delightful but rather complicated *plural* self-reference and it also uses the main facts about Kripke semantics formulated in Propositions 8 and 9 below. So we omit even a sketch of the proof, but below, after Proposition 8, we will show an application of Proposition 3.

In the rest of this Section we deal with sequent calculus for GL and with Kripke semantics of GL. We will try to explain some of the most important results and to provide the reader with an insight to some techniques. But we still omit all longer proofs.

$$\dfrac{\dfrac{\dfrac{\dfrac{\langle\, \Box p, p, \Box\Box p \Rightarrow \Box p \,\rangle}{\langle\, \Box p \Rightarrow \Box\Box p \,\rangle} \qquad \dfrac{\langle\, \Box p, p \to \neg\Box p, p \Rightarrow \;\rangle}{\langle\, \Box\Box p, \Box(p \to \neg\Box p), \Box p \Rightarrow \Box\bot \,\rangle}}{\langle\, \Box(p \to \neg\Box p), \Box p \Rightarrow \Box\bot \,\rangle}}{\langle\, \Box(p \to \neg\Box p), \Box p, \neg\Box\bot \Rightarrow \;\rangle}}{\langle\, \Box(p \to \neg\Box p), \neg\Box\bot \Rightarrow \neg\Box p \,\rangle}$$

Figure 1: An example proof in the calculus $\mathrm{LK}_{\mathrm{GL}}$

For the *Gentzen (sequent) calculus* LK for classical propositional logic we take the calculus from [Tak75] with the only change that a sequent is a pair of finite sets of formulas rather than a pair of finite sequences of formulas. A sequent consisting of sets Γ and Δ is written $\langle\, \Gamma \Rightarrow \Delta \,\rangle$; its intuitive meaning is "if *all* formulas in Γ hold, then *some* formula in Δ also holds". The sets Γ and Δ are *antecedent* and *succedent* of the sequent $\langle\, \Gamma \Rightarrow \Delta \,\rangle$. The calculus has two logical rules for each logical connective. As a sample we give the rule for introducing implication to antecedent:

$$\to\!\mathrm{l}: \qquad \langle\, \Gamma \Rightarrow \Delta, A \,\rangle, \langle\, \Gamma, B \Rightarrow \Delta \,\rangle \,/\, \langle\, \Gamma, A \to B \Rightarrow \Delta \,\rangle,$$

where Δ, A is a shorthand for $\Delta \cup \{A\}$ etc. Besides logical rules the calculus has initial sequents (i.e. rules without premises) of the form $\langle\, \Gamma, A \Rightarrow \Delta, A \,\rangle$ and $\langle\, \Gamma, \bot \Rightarrow \Delta \,\rangle$, and two structural rules: the cut-rule and the weakening rule allowing to add any formula to (any side of) any sequent. There is no ($\bot$r)-rule.

A sequent calculus $\mathrm{LK}_{\mathrm{GL}}$ is obtained by adding to LK a single modal rule

$$\Box\mathrm{r}: \qquad \langle\, \Box\Gamma, \Gamma, \Box A \Rightarrow A \,\rangle \,/\, \langle\, \Box\Gamma \Rightarrow \Box A \,\rangle,$$

where $\Box\Gamma = \{\, \Box A \,;\, A \in \Gamma \,\}$. Note that (i) both the premise and the conclusion of ($\Box$r) contain exactly one formula in succedent, (ii) all formulas in the conclusion start with $\Box$, and (iii) the rule ($\Box$r) satisfies the subformula property. There is no ($\Box$l)-rule: if a formula $\Box A$ is in an antecedent of a provable sequent, it must either have appeared by weakening or come from some initial sequent.

In Fig. 1 we give an example of a proof in $\mathrm{LK}_{\mathrm{GL}}$. It is a proof of a modal version of Gödel first incompleteness theorem. It contains two applications of the ($\Box$r)-rule, a cut on the formula $\Box\Box p$, and ends with two applications of the negation rules. The left leaf is an initial sequent, the right one is not but is easily proved using the ($\to$l)-rule. Some weakenings have been omitted. As an exercise the reader may try to find a (simpler) cut-free proof. From Proposition 9 below it is clear that the cut-elimination theorem holds for $\mathrm{LK}_{\mathrm{GL}}$. The aspects of a direct (syntactical) proof of the cut-elimination theorem are discussed in [SV82].

Let p be a propositional atom. Then $A_p(q)$ or only $A(q)$ denotes the result of substituting q for p in A. We suppose that the atom q does not occur in A and we sometimes write $A(p)$ instead of A. We say that p is *boxed* in A if all occurrences of p in $A(p)$ are in a scope of some $\Box$.

Proposition 4 *(a) Let $A(p)$ be a modal formula and q a propositional atom not occurring in $A(p)$. Then $\langle \Box(p \equiv q) \Rightarrow \Box(A(p) \equiv A(q)) \rangle$ is a sequent provable in LK_{GL}.*
(b) If, moreover, p is boxed in A, then $\langle \Box(p \equiv q) \Rightarrow A(p) \equiv A(q) \rangle$ is a sequent provable in LK_{GL}.

Proof By an induction on the complexity of A, see [SV82] or [Smo85]. ■

Proposition 5 *Let $\langle \Gamma, \Pi \Rightarrow \Delta, \Lambda \rangle$ be a sequent provable in LK_{GL}. Then there is a modal formula D containing only atoms common to sequents $\langle \Gamma \Rightarrow \Delta \rangle$ and $\langle \Pi \Rightarrow \Lambda \rangle$ and such that $\langle \Gamma \Rightarrow \Delta, D \rangle$ and $\langle \Pi, D \Rightarrow \Lambda \rangle$ are sequents provable in LK_{GL}.*

Proof By an induction on the number of steps in a cut-free proof of the sequent $\langle \Gamma, \Pi \Rightarrow \Delta, \Lambda \rangle$ in $\mathrm{LK}_{\mathrm{GL}}$, see [SV82]. ■

Proposition 4 is the substitution theorem and Proposition 5 is the interpolation theorem for GL.

Proposition 6 *Let $A(p)$ be a modal formula not containing q, let p be boxed in $A(p)$. Then the sequent $\langle \Box(p \equiv A(p)), \Box(q \equiv A(q)) \Rightarrow \Box(p \equiv q) \rangle$ is provable in LK_{GL}.*

Proof Having Proposition 4, we can write an almost complete formal proof in $\mathrm{LK}_{\mathrm{GL}}$:

1: $\langle \Box(p \equiv q) \Rightarrow A(p) \equiv A(q) \rangle$; Proposition 4(b)
2: $\langle p \equiv A(p), q \equiv A(q), A(p) \equiv A(q) \Rightarrow p \equiv q \rangle$
3: $\langle p \equiv A(p), q \equiv A(q), \Box(p \equiv q) \Rightarrow p \equiv q \rangle$; Cut on 1 and 2
4: $\langle \Box(p \equiv A(p)), \Box(q \equiv A(q)) \Rightarrow \Box(p \equiv q) \rangle$; ($\Box$r)

The sequent (2) is tautological. Besides ($\Box$r), some weakenings were used to obtain the sequent (4) from (3). ■

We show that on the arithmetical side, the Proposition 6 implies a unique solvability of some self-referential equations. Assume, for example, that λ is an arithmetical sentence and that φ_1 and φ_2 are two solutions of an equation given by the formula $\mathsf{Pr}(x) \to \lambda$. So both sentences $\varphi_1 \equiv (\mathsf{Pr}(\overline{\varphi_1}) \to \lambda)$ and $\varphi_2 \equiv (\mathsf{Pr}(\overline{\varphi_2}) \to \lambda)$ are provable in PA. Also, by D1,

$$\mathrm{PA} \vdash \mathsf{Pr}(\overline{\varphi_1 \equiv (\mathsf{Pr}(\overline{\varphi_1}) \to \lambda)}) \quad \text{and} \quad \mathrm{PA} \vdash \mathsf{Pr}(\overline{\varphi_2 \equiv (\mathsf{Pr}(\overline{\varphi_2}) \to \lambda)}). \qquad \text{(C)}$$

Take the formula $\Box p \to r$ for $A(p)$. Soundness of GL w.r.t. arithmetical semantics and Proposition 6 yield

$$\mathrm{PA} \vdash \mathsf{Pr}(\overline{\varphi_1 \equiv (\mathsf{Pr}(\overline{\varphi_1}) \to \lambda)}) \mathbin{\&} \mathsf{Pr}(\overline{\varphi_2 \equiv (\mathsf{Pr}(\overline{\varphi_2}) \to \lambda)}) \to \mathsf{Pr}(\overline{\varphi_1 \equiv \varphi_2}). \qquad \text{(D)}$$

From (C) and (D) we get $\mathrm{PA} \vdash \mathsf{Pr}(\overline{\varphi_1 \equiv \varphi_2})$, and, since $\mathbb{N}$ is a model of PA, also $\mathbb{N} \models \mathsf{Pr}(\overline{\varphi_1 \equiv \varphi_2})$. Now the condition Def yields $\mathrm{PA} \vdash \varphi_1 \equiv \varphi_2$. So we have shown that, up to equivalence provable in PA, the equation $\mathrm{PA} \vdash \varphi \equiv \mathsf{Pr}(\overline{\varphi}) \to \lambda$ has a unique solution φ. Completely identical argument applies to any other equation $\mathrm{PA} \vdash \varphi \equiv \psi(\overline{\varphi}, \overline{\neg\varphi})$ under the assumption that ψ is an arithmetical version of some modal formula, i.e. that $\psi(x, y)$ is built up using only logical connectives and the formula Pr and that all occurrences of x and y in ψ appear inside some formula Pr. Let us call such equation a *Gödelian equation* and its solution a *Gödelian sentence.* Any Gödelian sentence is uniquely determined by the equation it satisfies. An example of a self-referential equation which is not Gödelian is the Rosser equation (B) above. And indeed, it can be shown that the sentence ρ is *not* uniquely determined by the equation (B), see [GS79].

Proposition 7 *Let $A(p)$ be a modal formula such that p is boxed in A. Then there is a modal formula D containing only atoms occurring in A other than p such that $D \equiv A(D)$ is provable in LK_{GL}.*

Proof This was proved in [SV82] and is discussed also in [Smo85]. We reproduce the proof from [SV82].

Take a new (i.e. not occurring in A) atom q. Then A and atoms p and q are as in Proposition 6, and all sequents (1)–(4) from its proof are provable in $\mathrm{LK}_{\mathrm{GL}}$. We can continue the formal proof using the cut rule on (1) and (4):

5: $\langle \Box(p \equiv A(p)), \Box(q \equiv A(q)) \Rightarrow A(p) \equiv A(q) \rangle$
6: $\langle \Box(p \equiv A(p)), \Box(q \equiv A(q)), A(p) \Rightarrow A(q) \rangle$; 5

Now apply Proposition 5 to the sequent (6): there is a formula D not containing p and q such that both sequents

7: $\langle \Box(p \equiv A(p)), A(p) \Rightarrow D \rangle$
8: $\langle \Box(q \equiv A(q)), D \Rightarrow A(q) \rangle$

are provable in $\mathrm{LK}_{\mathrm{GL}}$. It is evident that a sequent obtained by substituting any formula for a propositional atom in a provable sequent is provable as well. So the following sequents are provable in $\mathrm{LK}_{\mathrm{GL}}$:

9: $\langle \Box(D \equiv A(D)), A(D) \Rightarrow D \rangle$; 7, substitution
10: $\langle \Box(D \equiv A(D)), D \Rightarrow A(D) \rangle$; 8, substitution
11: $\langle \Box(D \equiv A(D)) \Rightarrow D \equiv A(D) \rangle$; 9, 10
12: $\langle \Rightarrow \Box(D \equiv A(D)) \rangle$; ($\Box$r) on 11
13: $\langle \Rightarrow D \equiv A(D) \rangle$; Cut on 11 and 12

■

Recall that the Gödel sentence ν appeared equivalent to the consistency statement Con. Proposition 7 explains that it was not a coincidence: any Gödelian equation has an explicitly definable solution, i.e. a solution expressible in terms of Pr, $\bot$ and other logical connectives. As an example, take the equation $\mathrm{PA} \vdash \varphi \equiv \mathsf{Pr}(\overline{\varphi}) \rightarrow \lambda$. A guess based on the knowledge of a proof in [Löb55], or a look at [Smo85], p. 123, shows that $\varphi := \mathsf{Pr}(\overline{\lambda}) \rightarrow \lambda$ is its only solution.

A *Kripke frame* is a pair $\langle W, R\rangle$ where W is a nonempty set of *nodes* (or *possible worlds*) and R is a binary relation on W. The relation R is called *accessibility relation* of the frame $\langle W, R\rangle$. A relation $\Vdash$ between modal formulas and nodes of a frame $\langle W, R\rangle$ is a *forcing relation* on $\langle W, R\rangle$ if it preserves all logical connectives (i.e. satisfies $a \Vdash A \,\&\, B$ iff $a \Vdash A$ and $a \Vdash B$, etc.) and satisfies the condition

$$a \Vdash \Box A \;\Leftrightarrow\; \forall b(a\; R\; b \;\Rightarrow\; b \Vdash A) \tag{E}$$

for each node a and a modal formula A. A *Kripke model* is a triple $\langle W, R, \Vdash\rangle$ where $\Vdash$ is a forcing relation on a frame $\langle W, R\rangle$. A formula is *valid* in a model $\langle W, R, \Vdash\rangle$ if it is forced in any node $a \in W$. A model $\langle W, R, \Vdash\rangle$ is a *countermodel* for a formula A if A is not valid in it, i.e. if there exists a node $a \in W$ such that $a \nVdash A$. The notions of forcing, validity and countermodel naturally extend from formulas to sequents: a sequent $\langle\, \Gamma \;\Rightarrow\; \Delta \,\rangle$ is forced in a node $a \in W$ of a model $\langle W, R, \Vdash\rangle$ if $a \Vdash \bigwedge \Gamma \rightarrow \bigvee \Delta$, the same sequent is valid in $\langle W, R, \Vdash\rangle$ if it is forced in each node $a \in W$, and $\langle W, R, \Vdash\rangle$ is a countermodel for $\langle\, \Gamma \;\Rightarrow\; \Delta \,\rangle$ if some $a \in W$ forces all formulas in Γ and no formulas in Δ. More on Kripke semantics of modal logic can be found in [HC96] or in any other source on modal logic.

An example of a Kripke model is in Fig. 2. It has five nodes and its accessibility relation has four pairs indicated by arrows. For each node we have indicated which atoms are forced it it. If an atom is not mentioned it it understood that it is not forced. We have $a \nVdash \Box(p \rightarrow q)$ and thus $a \Vdash \neg\Box(p \rightarrow q)$ because there are nodes—namely c is such a node—that are accessible from a, force p, but do not force q. For the propositional constant $\bot$, the condition that all connectives are preserved by $\Vdash$ says that $\bot$ is nowhere forced. Since no nodes are accessible from b (as well as from c, d, and e), we have $b \Vdash \Box\bot$. The formula $\neg\Box\bot \rightarrow p$ is valid in our model since the only node that forces the formula $\neg\Box\bot$ is a, and indeed, a forces p. The model in Fig. 2 is a countermodel e.g. for the formula $\neg\Box\bot \rightarrow \neg p$.

A relation R is *reversely well-founded* on W if each nonempty subset of W has a maximal element. A model $\langle W, R, \Vdash\rangle$ is a *Kripke model for provability logic* if R is transitive and reversely well-founded. Note that any reversely well-founded relation is irreflexive and that a transitive relation on a finite set is reversely well-founded if and only if it is irreflexive. So the model in Fig. 2 is a model for provability logic.

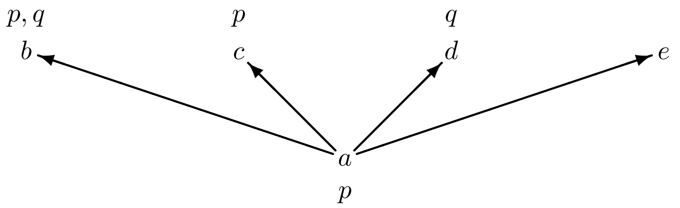


Figure 2: Example Kripke model

Proposition 8 *Each sequent provable in $\mathrm{LK_{GL}}$ is valid in any transitive reversely well-founded Kripke model. So $\mathrm{LK_{GL}}$ is sound w.r.t. Kripke semantics of provability logic.*

Proof We verify soundness of the modal rule ($\Box$r). Soundness of all other rules is straightforward. So let $\langle W, R, \Vdash\rangle$ be a Kripke model with reversely well-founded R and let $\langle\, \Box\Gamma, \Gamma, \Box A \;\Rightarrow\; A \,\rangle$ be a sequent valid in $\langle W, R, \Vdash\rangle$. We show that $\langle\, \Box\Gamma \;\Rightarrow\; \Box A \,\rangle$ is also valid in $\langle W, R, \Vdash\rangle$. Assume not. Then there is a node $a \in W$ such that $a \nVdash \Box A$ and $a \Vdash \Box B$ whenever $B \in \Gamma$. Take $Y = \{\, b\,;\; a\; R\; b \text{ and } b \nVdash A \,\}$. Since $a \nVdash \Box A$, the set Y is nonempty. Let b_0 be a maximal element of Y. Since $a\; R\; b_0$ and $a \Vdash \Box B$, we have $b_0 \Vdash B$ for each $B \in \Gamma$. By transitivity of R, all nodes accessible from b_0 are also accessible from a. So $b_0 \Vdash \Box B$ for each $B \in \Gamma$. If we had $c \nVdash A$ for some c accessible from b_0, then, again by transitivity of R, b_0 would not be maximal in Y. So $b_0 \Vdash \Box A$. We have reached at a contradiction: the assumption that $\langle\, \Box\Gamma, \Gamma, \Box A \;\Rightarrow\; A \,\rangle$ is valid in $\langle W, R, \Vdash\rangle$ is violated in b_0. ■

We show an example on the use of Proposition 8. The formula

$$\Box(p \vee q) \vee \Box(p \vee \neg q) \vee \Box(\neg p \vee q) \vee \Box(\neg p \vee \neg q) \tag{F}$$

has a Kripke countermodel. Indeed, the model in Fig. 2 works. So, by Proposition 3, the formula (F) is not arithmetically valid. So we have an evaluation e such that

$$\mathrm{PA} \nvdash \mathsf{Pr}(\overline{e(p) \vee e(q)}) \vee \,..\, \vee \mathsf{Pr}(\overline{\neg e(p) \vee \neg e(q)}).$$

This together with the condition D1 says that none of the four disjunctions $e(p) \vee e(q), .., \neg e(p) \vee \neg e(q)$ is provable in PA. So we have shown the existence of arithmetical sentences φ and ψ, namely $\varphi := e(p)$ and $\psi := e(q)$, that are mutually independent in the sense that the theories $\mathrm{PA} \cup \{\varphi, \psi\}$, $\mathrm{PA} \cup \{\varphi, \neg\psi\}$, $\mathrm{PA} \cup \{\neg\varphi, \psi\}$ and $\mathrm{PA} \cup \{\neg\varphi, \neg\psi\}$ are all consistent.

The following Proposition is the Kripke completeness theorem for the calculus $\mathrm{LK_{GL}}$. By *length* of a sequent $\langle\, \Sigma \;\Rightarrow\; \Omega \,\rangle$ we mean the total number of all occurrences of logical connectives and propositional atoms in the sequent $\langle\, \Sigma \;\Rightarrow\; \Omega \,\rangle$.

Proposition 9 *If a sequent $\langle \Sigma \Rightarrow \Omega \rangle$ of length n is provable in LK_{GL} then it has a cut-free proof in LK_{GL} of depth at most $\mathcal{O}(n^2)$. Otherwise it has a Kripke countermodel of depth at most n, in which each node has at most n immediate successors.*

Proof We describe a procedure that attempts to construct a proof of a given sequent $\langle \Sigma \Rightarrow \Omega \rangle$. If the attempt fails, the unsuccessful proof is converted to a Kripke countermodel for $\langle \Sigma \Rightarrow \Omega \rangle$.

The first part of our procedure creates a finite tree $\mathcal{T}$ labeled by sequents. Initialization consists of declaring the input sequent $\langle \Sigma \Rightarrow \Omega \rangle$ to be the root of $\mathcal{T}$. Then the first part of the procedure proceeds in steps; in each step it chooses and processes a top sequent in $\mathcal{T}$ that is (so far) not declared to be a leaf. If there are no non-leaf top sequents, the first part of the procedure is finished.

Let $\langle \Gamma \Rightarrow \Delta \rangle$ be a top sequent which is not declared to be a leaf. If $\Gamma \cup \Delta$ contains a formula the outermost symbol of which is a connective, the procedure chooses some such formula A and performs a *propositional step*. This step means to append one or two new sequents above $\langle \Gamma \Rightarrow \Delta \rangle$ using once or twice the corresponding propositional rule in reverse, thus obtaining one or two new top sequents. The cases where A is an implication in succedent or a conjunction in succedent are treated as follows:

$$\dfrac{\langle \Gamma, B \Rightarrow \Lambda, C \rangle}{\langle \Gamma \Rightarrow \Lambda, B \to C \rangle} \qquad\qquad \dfrac{\langle \Gamma \Rightarrow \Lambda, B \rangle \quad \langle \Gamma \Rightarrow \Lambda, C \rangle}{\langle \Gamma \Rightarrow \Lambda, B \,\&\, C \rangle}$$

If A is a conjunction in antecedent or a disjunction in succedent we have two new sequents but only one new top sequent:

$$\dfrac{\dfrac{\langle \Pi, B, C \Rightarrow \Delta \rangle}{\langle \Pi, B \,\&\, C, C \Rightarrow \Delta \rangle}}{\langle \Pi, B \,\&\, C \Rightarrow \Delta \rangle} \qquad\qquad \dfrac{\dfrac{\langle \Gamma \Rightarrow \Lambda, B, C \rangle}{\langle \Gamma \Rightarrow \Lambda, B \vee C, C \rangle}}{\langle \Gamma \Rightarrow \Lambda, B \vee C \rangle}$$

Each of the remaining four cases is similar to some of these. The propositional steps of the procedure are the same as in classical propositional logic.

We show on an example how the procedure works. Assume that the input sequent $\langle \Sigma \Rightarrow \Omega \rangle$ is

$$\langle \Box(\Box p \to q) \vee \Box\neg(p \vee q), \neg\Box q \Rightarrow \Box(\Box\bot \to \neg p), p \rangle.$$

Then our procedure uses the $(\neg\mathrm{l})$- and $(\vee\mathrm{l})$-rule to obtain a four-element tree

$$\dfrac{\dfrac{\langle \Box(\Box p \to q) \Rightarrow \Box q, \Box(\Box\bot \to \neg p), p \rangle^b \qquad \langle C \Rightarrow \Box q, \Box(\Box\bot \to \neg p), p \rangle^c}{\langle \Box(\Box p \to q) \vee \Box\neg(p \vee q) \Rightarrow \Box q, \Box(\Box\bot \to \neg p), p \rangle^a}}{\langle \Box(\Box p \to q) \vee \Box\neg(p \vee q), \neg\Box q \Rightarrow \Box(\Box\bot \to \neg p), p \rangle}$$

with two top sequents b, c containing only propositional atoms and boxed formulas (i.e. formulas starting with the symbol $\Box$). C denotes the formula $\Box\neg(p \vee q)$.

Let's say that a sequent is *critical* if it contains only propositional atoms and boxed formulas. The formula $\bot$ is also treated as an atom. We are now going to specify what the procedure does with critical sequents, i.e. we are going to describe *critical steps*. A critical sequent has the form $\langle \Box\Gamma, \Pi \Rightarrow \Box\Delta, \Lambda \rangle$, where $\Pi \cup \Lambda$ contains only propositional atoms. If $\Pi \cap \Lambda \neq \emptyset$ or $\bot \in \Pi$ or $\Delta = \emptyset$ then the sequent becomes a leaf. Otherwise the procedure appends all sequents of the form $\langle \Box\Gamma, \Gamma, \Box A \Rightarrow A \rangle$, where $A \in \Delta$, as immediate successors of the sequent $\langle \Box\Gamma, \Pi \Rightarrow \Box\Delta, \Lambda \rangle$. Note that if some of its immediate successors is provable, then $\langle \Box\Gamma, \Pi \Rightarrow \Box\Delta, \Lambda \rangle$ is also provable in $\mathrm{LK}_{\mathrm{GL}}$. Also note that propositional atoms are discarded during this step.

As said above, the propositional and critical steps are repeated while there is any top sequent which is not a leaf. On the right hand side of our example, above the sequent c, the procedure gets

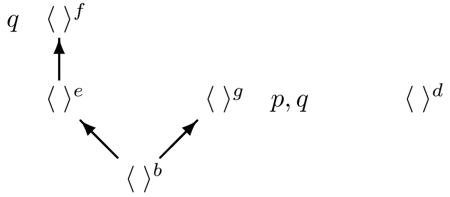
$$\dfrac{\dfrac{\dfrac{\langle C, \Box q \Rightarrow p, q \rangle^d}{\langle C, \Box q \Rightarrow q, p \vee q \rangle}}{\langle C, \neg(p \vee q), \Box q \Rightarrow q \rangle} \quad \dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\langle C, \Box(\Box\bot \to \neg p), \Box\bot, p \Rightarrow p, q \rangle}{\langle C, \Box(\Box\bot \to \neg p), \Box\bot, p \Rightarrow p, p \vee q \rangle}}{\langle C, \Box(\Box\bot \to \neg p), \Box\bot, p \Rightarrow p \vee q \rangle}}{\langle C, \Box(\Box\bot \to \neg p), \Box\bot \Rightarrow \neg p, p \vee q \rangle}}{\langle C, \Box(\Box\bot \to \neg p) \Rightarrow \Box\bot \to \neg p, p \vee q \rangle}}{\langle C, \neg(p \vee q), \Box(\Box\bot \to \neg p) \Rightarrow \Box\bot \to \neg p \rangle}}{\langle \Box\neg(p \vee q) \Rightarrow \Box q, \Box(\Box\bot \to \neg p), p \rangle^c}$$

where C still denotes the formula $\Box\neg(p \vee q)$. We suggest the reader to write down what happens on the left hand side, above the sequent b.

Each critical step adds a boxed formula to the antecedent. Since there are at most n boxed formulas, each branch of the resulting tree contains at most n critical sequents. The distance of two consecutive critical sequents on any branch is bounded by $2n$ because each propositional step removes one logical connective except that two steps are necessary to remove a disjunction from succedent or a conjunction from antecedent, which is visible on the right top of our example. So the resulting tree is finite and the first part of the procedure always terminates.

Our procedure then continues by marking each sequent in the tree as "positive" or "negative". A leaf is positive if it is an initial sequent, i.e. if its antecedent and succedent have some atom in common or if its antecedent contains $\bot$. Otherwise it is negative. In the latter case it contains no boxed formulas in succedent and can easily be verified to have a one-element Kripke countermodel. Any other critical sequent is positive if *some* of its immediate successors is positive, and negative otherwise. A non-critical sequent has one or two immediate successors (since each propositional rule has at most two premises), and is marked positive if and only if *all* its immediate successors are positive. In our example, all sequents on the left branch growing from the sequent c are negative, but all sequents on the right branch are positive, which is sufficient for c to be also positive. The (non-critical) sequent a is negative because if the reader wrote down what happens above b, then he found that b is negative. Doing this, he also must have discovered three more negative critical sequents e, f, and g.

If the root sequent $\langle\, \Sigma \Rightarrow \Omega\, \rangle$ is positive (which is not the case in our example), the procedure discards all negative sequents, then it discards further sequents so that each critical non-leaf sequent has exactly one immediate successor, and finally it adds some weakenings. The result is the desired proof of the original sequent $\langle\, \Sigma \Rightarrow \Omega\, \rangle$. If, on the other hand, the root sequent $\langle\, \Sigma \Rightarrow \Omega\, \rangle$ is negative, the Kripke countermodel $\langle W, R, \Vdash \rangle$ for $\langle\, \Sigma \Rightarrow \Omega\, \rangle$ is constructed as follows. W is the set of all negative critical sequents. R is inherited from the tree $\mathcal{T}$. The forcing relation $\Vdash$ is determined by the condition that each atom p is forced in a critical sequent $\langle\, \Gamma \Rightarrow \Delta\, \rangle$ if and only if it is an element of Γ. In our example the model is



The fact that $\langle W, R, \Vdash \rangle$ is a countermodel for $\langle\, \Sigma \Rightarrow \Omega\, \rangle$ follows from the following Sublemma.

Sublemma *Let $s' \in \mathcal{T}$ and let $s \in W$ be such that s is the first critical sequent on the path (in $\mathcal{T}$ viewed as a directed graph) going from s' to s. If a modal formula A is in the antecedent of s' then $s \Vdash A$. If it is in the succecedent of s' then $s \Vdash \neg A$.*

The first sequent on the path going from s' to s can be s itself, what happens if s' is critical.

The Sublemma is proved by an induction on the complexity of A. We show three typical cases, leaving the rest to the reader. If A is an atom in the succedent of s' then A is not in the antecedent of s, otherwise s would be a positive leaf. So $s \nVdash A$. Now assume that $A = \Box B$ and A is in the succedent of s'. We want to show $s \nVdash \Box B$. Since s is negative and has a boxed formula $\Box B$ in succedent, it has an immediate successor t' having B in succedent. By the construction, t' is negative. Choose a path from t' to some critical t such that all nodes on this path are negative and t is the first critical sequent on this path. We have $t \in W$ and $s\ R\ t$. The induction hypothesis applied to t', t, and B yields $t \nVdash B$. So $s \nVdash \Box B$. Assume finally that $A = \Box B$ and A is in the antecedent of s'. We want to show $s \Vdash \Box B$. Let $t \in W$ be arbitrary such that $s\ R\ t$. Let s_1 be the last critical sequent on the path from s to t which is different from t. Let t' be the immediate successor of s_1 on this path. Since boxed formulas never disappear from antecedent, $\Box B$ is in the antecedent of s_1. Then B is in the antecedent of t' and the induction hypothesis is applicable to t', t, and B. ■

In our example the (negative) node b is the first critical sequent on a path going from the root sequent $s' = \langle\, \Sigma \Rightarrow \Omega\, \rangle$. Hence, by the Sublemma, s' is not forced in b and the whole model is a countermodel for s'. Of course, the node d is immaterial for this purpose. Note that the Sublemma is not applicable to s' and d because d is not the first critical sequent on the path going from s' to d.

Let us remark that a similar decision procedure can be specified also for the intuitionistic propositional logic.

5 What else?

We see that provability logic can be helpful in studying and understanding the properties of self-referential equations and sentences. An immediate consequence of Proposition 9 is that GL has the finite model property and is decidable, which are properties shared by many modal logics. A less common property of GL is that its Kripke semantics is *not compact*, see [Smo85] or [BS91]. In fact GL is the only natural logic known to the author that has a reasonable but non-compact semantics.

In the proof of Proposition 9 we have tried to point out a phenomenon called *alternation* by theoretical computer scientists: in some cases a node is positive iff all successors of it are positive, while in other cases it is positive iff some successor of it is positive. Alternation does not occur e.g. in the decision algorithm of classical propositional logic and is typical for problems in *PSPACE*, which is a class of decision problems solvable by an algorithm the memory requirements of which grow only polynomially with the size of the input. More is true: the decision problem of GL is *PSPACE*-complete, i.e. belongs to the subclass of the most difficult problems in *PSPACE*. In a forthcoming paper [Šve03] we show that the decision problem of GL remains *PSPACE*-complete even if the number of propositional atoms is restricted to one.

A key step in proving *PSPACE*-completeness of the decision problem of GL is the construction of a sequence $A_0, A_1, A_2, \ldots$ of modal formulas such that the length of A_n grows polynomially with n, each A_n has a Kripke countermodel, but the size of a minimal countermodel for A_n (i.e. the number of nodes in that countermodel) grows exponentially with n. The following analogical question, concerning provable formulas, is probably open: is it possible to construct a sequence $A_0, A_1, A_2, \ldots$ of modal formulas provable in GL such that the length of A_n grows polynomially with n, but the size of a minimal proof of A_n grows exponentially with n?

In Section 3 we have shown that the so called *reflection scheme* $\Box A \to A$ is not PA-tautological because there exists an arithmetical evaluation e such that PA $\nvdash e(\Box p \to p)$. This is not the last word about the reflection scheme. We can easily check that for each evaluation e we have $\mathbb{N} \models e(\Box A \to A)$: if $\mathbb{N} \models e(\Box A)$, i.e. $\mathbb{N} \models \mathsf{Pr}(\overline{e(A)})$, then, by the condition Def, PA $\vdash e(A)$. Since $\mathbb{N}$ is a model of PA, we have $\mathbb{N} \models e(A)$. This argument leads us to the notion of $\mathbb{N}$-tautology and to another modal logic GL$^\omega$. A modal formula A is an $\mathbb{N}$-*tautology* if $\mathbb{N} \models e(A)$ for each evaluation e. It is evident that any PA-tautology is an $\mathbb{N}$-tautology and that any instance of the reflection scheme is also an $\mathbb{N}$-tautology. So the logic GL$^\omega$ has an axiom set consisting of all formulas provable in GL and of all formulas of the form $\Box A \to A$, and its only rule of inference is modus ponens. The necessitation rule $A \,/\, \Box A$ is not accepted as an inference rule of GL$^\omega$ because the set of all $\mathbb{N}$-tautologies is not closed under this rule. Any formula is provable in GL$^\omega$ if and only if it is an $\mathbb{N}$-tautology, which means that GL$^\omega$ is complete w.r.t. its arithmetical semantics. The moral from this result, proved by Solovay in [Sol76] and called Solovay second completeness theorem in [Smo85], can be formulated as follows: the reflection scheme captures the only general knowledge about provability in PA that cannot be formalized in PA itself. GL$^\omega$ is decidable and its computational complexity is the same as that of GL.

In Section 2 we have mentioned that the provability predicate can be constructed for any recursively axiomatizable extension T of PA. Further generalization is possible: the provability predicate Pr_T satisfying a straightforward modification of the Löb conditions D1–D3 can be constructed for any recursively axiomatizable extension T of IΔ_0+Ω_1. Here IΔ_0 is Peano arithmetic with the induction scheme restricted to Δ_0-formulas (i.e. to formulas not containing unbounded quantifiers) and Ω_1 is a single axiom asserting the totality of the function $x \mapsto x^{|x|}$, where $|x|$ denotes the length of the binary expansion of x. Such provability predicate can be used to translate the symbol $\Box$. This means that we can ask which general principles about provability in T can be proved in T itself, i.e. we can ask what is the provability logic of the theory T. A perhaps surprising answer to this question is that the provability logic of many reasonable theories is the same. More precisely, GL axiomatizes the provability logic of any sound recursively axiomatizable theory extending IΔ_0+Exp, where Exp is an axiom asserting that the exponential function $x \mapsto 2^x$ is total. The axiom Exp is stronger than Ω_1. Both theories IΔ_0+Ω_1 and IΔ_0+Exp are extensively studied and have connections to computational complexity. GL can easily be verified to be sound w.r.t. the provability logic of IΔ_0+Ω_1, but the Solovay's proof of Proposition 3 needs the axiom Exp, and it is not known whether GL is complete w.r.t. the provability logic of IΔ_0+Ω_1. Despite of considerable effort and promising partial results, see [BV93], the problem what the provability logic of IΔ_0+Ω_1 is remains open.

Formal provability is not the only metamathematical concept that can be studied and explicated using modal logic. There are extensions of GL that are applicable to various kinds of *Rosser sentences*, see [GS79] or [Smo85]. Further interesting extensions of GL fall under the head of *interpretability logic*. The language of interpretability logic has a binary "modality" $\rhd$ in addition to the symbol $\Box$. Arithmetical semantics of interpretability logic is obtained by adding the clause $e(A \rhd B) = \mathsf{Intp}_T(\overline{e(A)}, \overline{e(B)})$ to the definition of arithmetical evaluation. $\mathsf{Intp}_T(x, y)$ is an arithmetical formula saying that there is an interpretation of the theory $T \cup \{y\}$ in the theory $T \cup \{x\}$, where (syntactical) interpretation has its usual meaning defined e.g. in [TMR53]. If T is as usually then the theory $T \cup \{\neg\mathsf{Con}(T)\}$ is interpretable in T. If, in addition, T is consistent then $T \cup \{\mathsf{Con}(T)\}$ is *not* interpretable in T (for both facts see e.g. [Fef60]). So the modal formulas $\neg\bot \rhd \Box\bot$ and $\neg\Box\bot \to \neg(\neg\bot \rhd \neg\Box\bot)$ are examples of tautologies of interpretability logic. As an introduction to this large and rich area we recommend Albert Visser's survey paper [Vis98].

References

[Boo93] G. Boolos. *The Logic of Provability.* Cambridge University Press, 1993.

[BS91] G. Boolos and G. Sambin. Provability: the emergence of a mathematical modality. *Studia Logica*, L(1):1–23, 1991.

[BV93] A. Berarducci and R. Verbrugge. On the provability logic of bounded arithmetic. *Ann. Pure Appl. Logic*, 61(1–2):75–93, 1993.

[Car34] R. Carnap. *Logische Syntax der Sprache.* Springer, 1934.

[Fef60] S. Feferman. Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, 49:35–92, 1960.

[Göd31] K. Gödel. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatsh. Math. Phys.*, 38:173–198, 1931.

[GS79] D. Guaspari and R. M. Solovay. Rosser sentences. *Ann. Math. Logic*, 16:81–99, 1979.

[HC96] G. Hughes and M. Cresswell. *A New Introduction to Modal Logic.* Routledge, London, New-York, 1996.

[Löb55] M. H. Löb. Solution of a problem of Leon Henkin. *J. Symb. Logic*, 20:115–118, 1955.

[Smo84] C. Smoryński. Modal logic and self-reference. In D. M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, number 164–167 in Synthese Library, chapter II.9, pages 441–496. Kluwer, Dordrecht, 1984.

[Smo85] C. Smoryński. *Self-Reference and Modal Logic.* Springer, New York, 1985.

[Sol76] R. M. Solovay. Provability interpretations of modal logic. *Israel J. Math.*, 25:287–304, 1976.

[SV82] G. Sambin and S. Valentini. The modal logic of provability: The sequential approach. *J. Phil. Logic*, 11:311–342, 1982.

[Šve03] V. Švejdar. The decision problem of provability logic with only one atom. *Archive Math. Logic*, 42(8):763–768, 2003.

[Tak75] G. Takeuti. *Proof Theory.* North-Holland, Amsterdam, 1975.

[TMR53] A. Tarski, A. Mostowski, and R. M. Robinson. *Undecidable Theories.* North-Holland, Amsterdam, 1953.

[Vis98] A. Visser. An overview of interpretability logic. In M. Kracht, M. Rijke, H. Wansing, and M. Zakharyaschev, editors, *Advances in Modal Logic 1 (AiML'96)*, volume 87 of *CSLI Lecture Notes*, pages 307–359, Berlin, October 1998. CSLI Publ., 1998.