MA5219 LOGIC AND FOUNDATION OF MATHEMATICS I

# Lecture 26: Consistency statements

Tin Lok Wong

15 November, 2018

We conclude the course by pointing out a connection between proofs of finite consistency statements and complexity theory.

From the Incompleteness Theorems, one learns that consistency statements are in general hard to prove. In the case of the Second Incompleteness Theorem, 'hard to prove' simply means 'unprovable', while in the case of the Finite Incompleteness Theorem, 'hard to prove' means 'not polynomial-length provable'. One naturally asks where the limit of this incompleteness phenomenon is. For instance, can one improve 'polynomial' to 'exponential' in the Finite Incompleteness Theorem? As mentioned briefly in the previous lecture, the answer is in general no. Let us state this more properly here. Recall the definition of $\mathrm{PA}^-$ from Lemma 11.1.

**Theorem 26.1** (Pudlák)**.** Let $T$ be a finite $\mathcal{L}_{\mathrm{A}}(\exp)$ theory extending $\mathrm{PA}^-$. Then one can construct a $\Delta_0(\exp)$ formula $\Box^{\leqslant}(w, y)$ with some $k \in \mathbb{N}$ such that

(1) $\{(n, m) \in \mathbb{N}^2 : \mathbb{N} \models \Box^{\leqslant}(n, m)\} = \{(n, \ulcorner\theta\urcorner) \in \mathbb{N}^2 : \theta \text{ is an } \mathcal{L}_{\mathrm{A}}(\exp) \text{ formula and } T \vdash^n \theta\}$;

(2) $\Box^{\leqslant}(w, y)$ satisfies the derivability conditions (M), (N), (IN), and ($\Box$D) listed in the statement of Theorem 25.7 for all $n \in \mathbb{N}$ and all $\mathcal{L}_{\mathrm{A}}(\exp)$ sentences $\sigma, \tau$; and

(3) $T \vdash^{2^{k|n|}+k} \neg\Box^{\leqslant}(\underline{n}, \underline{\bot})$ for all $n \in \mathbb{N}$. □

Although Theorem 26.1 says every finite theory $T \supseteq \mathrm{PA}^-$ has exponential-length proofs of its own finite consistencies, it does not exclude the existence of such a theory $T$ which does not have exponential-length proofs of finite consistencies *of a stronger theory* $T^*$. In view of the general impression that consistency statements are hard to prove, one may actually expect every consistent such theory $T$ to *not* have exponential-length proofs of finite consistencies of some sufficiently strong theory $T^* \supseteq T$. Given a finite consistent theory $T$, the most prominent example of a strictly stronger (but not necessarily consistent) theory is $T^* = T + \mathrm{Con}(T)$ by the Second Incompleteness Theorem, where $\mathrm{Con}(T)$ is some sentence expressing the consistency of $T$ reasonably. This leads to the following conjecture. To avoid introducing new notation, we formulate this conjecture entirely in terms of the finite provability predicate $\Box^{\leqslant}(w, y)$ for $T$. In particular, we write $\mathrm{Con}(T) = \forall w\, \neg\Box^{\leqslant}(w, \underline{\bot})$, and $T + \mathrm{Con}(T) \nvdash \bot$ is rewritten as $T \nvdash \neg\mathrm{Con}(T)$ via (RAA′) and ($\bot$).

**Conjecture 26.2** (Pudlák)**.** Fix a finite consistent $\mathcal{L}_{\mathrm{A}}(\exp)$ theory $T \supseteq \mathrm{PA}^-$. Let $\Box^{\leqslant}(w, y)$ be an $\mathcal{L}_{\mathrm{A}}(\exp)$ formula satisfying both (1) and (2) in the statement of Theorem 26.1. Then for every $k \in \mathbb{N}$, there is $n \in \mathbb{N}$ such that

$$T \nvdash^{2^{k|n|}+k} \neg\Box^{\leqslant}(\underline{n}, \underline{\neg\forall w\, \neg\Box^{\leqslant}(w, \underline{\bot})}).$$

As Pudlák himself observed, his conjecture implies NP ≠ co-NP, and thus also P ≠ NP. Since it is generally believed that radically new techniques have to be developed to prove these inequations in complexity theory, one does not expect to see a proof of Conjecture 26.2 in the near future. Therefore, realistically, to make progress, one considers either weaker conjectures which are easier to prove, or stronger conjectures which are easier to refute. The rest of this lecture is devoted to an example of the latter. More precisely, we claim that if $\mathrm{Con}(T) = \forall w\, \neg\Box^{\leqslant}(w, \underline{\bot})$ is replaced by an arbitrary $T$-unprovable sentence, then the conjecture becomes false in general.

In addition to the derivability conditions listed in Theorem 25.7, we will employ three properties of the usual finite provability predicate. The first one is Theorem 26.1. The second and the third ones are the following duals of (N) and (IN).

**Lemma 26.3.** Provided $T \vdash \mathrm{I}\Delta_0(\exp)$, one can add to the list of conditions in Theorem 26.1 the following clause: there are $p(X), q(X) \in \mathbb{N}[X]$ such that for all $n \in \mathbb{N}$ and all $\mathcal{L}_{\mathrm{A}}(\exp)$ formulas $\theta$,

- if $T \mathrel{|\!\!\overset{n}{\vdash}} \theta$, then $T \mathrel{|\!\!\overset{p(2^n)}{\vdash}} \neg\Box^{\leqslant}(\underline{n}, \underline{\theta})$; and
- $T \mathrel{|\!\!\overset{q(n+\mathrm{len}(\theta))}{\vdash}} \neg\Box^{\leqslant}(\underline{n}, \underline{\theta}) \to \Box^{\leqslant}(\underline{p(2^n)}, \underline{\neg\Box^{\leqslant}(\underline{n}, \underline{\theta})})$.

*Proof sketch.* Very carefully study the actual shape of the $\Delta_0(\exp)$ formula $\Box^{\leqslant}(w, y)$ and the proof of formalized $\Sigma_1$ completeness, i.e., Theorem 12.6. The $2^n$ comes from the exponential number of proofs of length less than $n$. The details of the verification are cumbersome and so are omitted here. □

To establish the claim, we will need to produce a sentence $\sigma$ such that $T + \sigma$ is strictly stronger than the consistent theory $T$ we are given. Recall from Theorem 26.1 that $T$ has exponential-length proofs of its own finite consistencies. If $T + \sigma$ is not much stronger than $T$, then intuitively the finite consistencies of $T + \sigma$ are not much harder to prove than those of $T$, and thus they should have exponential-length proofs in $T$ too. This, in a sense, explains why the witness comparison method is employed in the construction of our $\sigma$ below. The following theorem can be seen as a formalization of First Incompleteness Theorem with explicit bounds on proof lengths. We modify the self-referential sentence slightly so as to keep the lengths of the required proofs exponential. Note that a polynomial in $n$ is an exponential in $|n|$.

**Notation.** Write $T \mathrel{|\!\!\overset{n}{*}} S(n)$ and $T \mathrel{|\!\!\overset{2^n}{*}} S(n)$ respectively for

$$\exists t(X){\in}\mathbb{N}[X] \quad \forall n{\in}\mathbb{N} \quad T \mathrel{|\!\!\overset{t(n)}{\vdash}} S(n) \qquad \text{and} \qquad \exists t(X){\in}\mathbb{N}[X] \quad \forall n{\in}\mathbb{N} \quad T \mathrel{|\!\!\overset{t(2^n)}{\vdash}} S(n).$$

**Theorem 26.4** (Hrubeš)**.** Fix an $\mathcal{L}_{\mathrm{A}}(\exp)$ theory $T \vdash \mathrm{I}\Delta_0(\exp)$. Let $\Box^{\leqslant}(w, y)$ be an $\mathcal{L}_{\mathrm{A}}(\exp)$ formula and $p(X), q(X) \in \mathbb{N}[X]$ such that for all $n \in \mathbb{N}$ and all $\mathcal{L}_{\mathrm{A}}(\exp)$ sentences $\sigma, \tau$,

(M) $T \vdash \forall w, y\ \big(\Box^{\leqslant}(w, y) \to \forall w' \geqslant w\ \Box^{\leqslant}(w', y)\big)$;

(N) If $T \mathrel{|\!\!\overset{n}{\vdash}} \sigma$, then $T \mathrel{|\!\!\overset{p(n)}{\vdash}} \Box^{\leqslant}(\underline{n}, \underline{\sigma})$;

(IN) $T \mathrel{|\!\!\overset{q(|n|+\mathrm{len}(\sigma))}{\vdash}} \Box^{\leqslant}(\underline{n}, \underline{\sigma}) \to \Box^{\leqslant}(\underline{p(n)}, \underline{\Box^{\leqslant}(\underline{n}, \underline{\sigma})})$;

(□D) $T \mathrel{|\!\!\overset{q(|n|+\mathrm{len}(\sigma)+\mathrm{len}(\tau))}{\vdash}} \Box^{\leqslant}(\underline{n}, \underline{\sigma \to \tau}) \to \big(\Box^{\leqslant}(\underline{n}, \underline{\sigma}) \to \Box^{\leqslant}(\underline{p(n)}, \underline{\tau})\big)$;

(C) $T \mathrel{|\!\!\overset{p(n)}{\vdash}} \neg\Box^{\leqslant}(\underline{n}, \underline{\bot})$;

(¬N) If $T \mathrel{|\!\!\overset{n}{\vdash}} \sigma$, then $T \mathrel{|\!\!\overset{p(2^n)}{\vdash}} \neg\Box^{\leqslant}(\underline{n}, \underline{\sigma})$; and

(¬IN) $T \mathrel{|\!\!\overset{q(n+\mathrm{len}(\sigma))}{\vdash}} \neg\Box^{\leqslant}(\underline{n}, \underline{\sigma}) \to \Box^{\leqslant}(\underline{p(2^n)}, \underline{\neg\Box^{\leqslant}(\underline{n}, \underline{\sigma})})$.

Suppose $\sigma$ is an $\mathcal{L}_{\mathrm{A}}(\exp)$ sentence which satisfies

$$T \vdash \sigma \leftrightarrow \neg\exists w\ \big(\Box^{\leqslant}(w, \underline{\sigma}) \wedge \neg\Box^{\leqslant}(2^w, \underline{\neg\sigma})\big).$$

Then there are non-constant $r(X), s(X) \in \mathbb{N}[X]$ such that for all $n \in \mathbb{N}$, the following hold.

(1) If $T \mathrel{|\!\!\overset{r(2^{2^n})}{\vdash}} \bot$, then $T \mathrel{|\!\!\overset{n}{\vdash}} \sigma$.

(2) If $T \mathrel{|\!\!\overset{r(n)}{\vdash}} \bot$, then $T \mathrel{|\!\!\overset{n}{\vdash}} \neg\sigma$.

(3) $T \mathrel{|\!\!\overset{s(2^n)}{\vdash}} \neg\Box^{\leqslant}(\underline{r(2^{2^n})}, \underline{\bot}) \to \neg\Box^{\leqslant}(\underline{n}, \underline{\sigma})$.

(4) $T \vdash^{s(n)}_{*} \neg\Box^{\leqslant}(\underline{r(n)}, \underline{\bot}) \to \neg\Box^{\leqslant}(\underline{n}, \underline{\neg\sigma})$.

*Proof.* Let us introduce the function $x \mapsto |x|$ to $\mathcal{L}_{\mathrm{A}}(\exp)$ in a way similar to how we introduced the function len in Lecture 10. The usual properties of the function $x \mapsto |x|$ can routinely be proved within $\mathrm{I}\Delta_0(\exp)$; cf. Lecture 11. It is a straightforward exercise involving numerous applications of (M), (N) and ($\Box$D) to show that, by choosing a larger $p(X)$ if necessary, one can make the following hold.

(i) $T \vdash^{|n|}_{*} \Box^{\leqslant}(\underline{p(n)}, \underline{\theta(\underline{n})}) \wedge \Box^{\leqslant}(\underline{p^2(n)}, \underline{\neg\theta(\underline{n})}) \to \Box^{\leqslant}(\underline{p^3(n)}, \underline{\bot})$ for all $\mathcal{L}_{\mathrm{A}}(\exp)$ formulas $\theta(x)$.

(ii) $T \vdash^{|n|}_{*} \Box^{\leqslant}(\underline{n}, \underline{\neg\sigma}) \to \Box^{\leqslant}\big(\underline{p^2(n)}, \underline{\exists w\ \big(\Box^{\leqslant}(w, \underline{\sigma}) \wedge \neg\Box^{\leqslant}(2^w, \underline{\neg\sigma})\big)}\big)$.

(iii) $T \vdash^{|n|}_{*} \Box^{\leqslant}(\underline{p(n)}, \underline{\exists w\ \big(\Box^{\leqslant}(w, \underline{\sigma}) \wedge \neg\Box^{\leqslant}(2^w, \underline{\neg\sigma})\big)}) \wedge \Box^{\leqslant}(\underline{p^2(n)}, \underline{\Box^{\leqslant}(\underline{n}, \underline{\neg\sigma})})$
$\to \Box^{\leqslant}(\underline{p^3(n)}, \underline{\Box^{\leqslant}(\underline{|n|}, \underline{\sigma})})$.

The proofs of (3) and (4) are essentially formalizations of those of (1) and (2). So we concentrate on (3) and (4).

(3) On the one hand,

$$\begin{aligned}
& T \vdash^{n}_{*} \Box^{\leqslant}(\underline{n}, \underline{\sigma}) \wedge \Box^{\leqslant}(\underline{2^n}, \underline{\neg\sigma}) \to \Box^{\leqslant}(\underline{p^3(2^n)}, \underline{\bot}) && \text{by (i);} \\
\therefore\ & T \vdash^{2^n}_{*} \Box^{\leqslant}(\underline{n}, \underline{\sigma}) \to \neg\Box^{\leqslant}(\underline{2^n}, \underline{\neg\sigma}) && \text{as } T \vdash^{p^4(2^n)} \neg\Box^{\leqslant}(\underline{p^3(2^n)}, \underline{\bot}) \text{ by (C);} \\
\therefore\ & T \vdash^{2^n}_{*} \Box^{\leqslant}(\underline{n}, \underline{\sigma}) \to \Box^{\leqslant}\big(\underline{p(2^{2^n})}, \underline{\neg\Box^{\leqslant}(\underline{2^n}, \underline{\neg\sigma})}\big) && \text{by } (\neg\mathrm{IN}).
\end{aligned}$$

On the other hand, condition (IN) implies

$$T \vdash^{|n|}_{*} \Box^{\leqslant}(\underline{n}, \underline{\sigma}) \to \Box^{\leqslant}(\underline{p(n)}, \underline{\Box^{\leqslant}(\underline{n}, \underline{\sigma})}).$$

Combining the two, we deduce that

$$\begin{aligned}
& T \vdash^{2^n}_{*} \Box^{\leqslant}(\underline{n}, \underline{\sigma}) \to \Box^{\leqslant}\big(\underline{r(2^{2^n})}, \underline{\Box^{\leqslant}(\underline{n}, \underline{\sigma}) \wedge \neg\Box^{\leqslant}(\underline{2^n}, \underline{\neg\sigma})}\big) && \text{for some } r(X) \in \mathbb{N}[X]; \\
\therefore\ & T \vdash^{2^n}_{*} \Box^{\leqslant}(\underline{n}, \underline{\sigma}) \to \Box^{\leqslant}(\underline{r(2^{2^n})}, \underline{\neg\sigma}) && \text{for some } r(X) \in \mathbb{N}[X], \\
& && \text{by the choice of } \sigma, \text{ plus (N) and } (\Box\mathrm{D}); \\
\therefore\ & T \vdash^{2^n}_{*} \Box^{\leqslant}(\underline{n}, \underline{\sigma}) \to \Box^{\leqslant}(\underline{r(2^{2^n})}, \underline{\bot}) && \text{for some } r(X) \in \mathbb{N}[X], \text{ by (i).}
\end{aligned}$$

(4) By (IN), we know $T \vdash^{|n|}_{*} \Box^{\leqslant}(\underline{n}, \underline{\neg\sigma}) \to \Box^{\leqslant}\big(\underline{p(n)}, \underline{\Box^{\leqslant}(\underline{n}, \underline{\neg\sigma})}\big)$. Combining this with (ii) using (iii), we deduce that

$$T \vdash^{|n|}_{*} \Box^{\leqslant}(\underline{n}, \underline{\neg\sigma}) \to \Box^{\leqslant}\big(\underline{p^4(n)}, \underline{\Box^{\leqslant}(\underline{|n|}, \underline{\sigma})}\big).$$

Consequently, since $T \vdash^{|n|}_{*} \neg\Box^{\leqslant}(\underline{|n|}, \underline{\sigma}) \to \Box^{\leqslant}\big(\underline{p(n)}, \underline{\neg\Box^{\leqslant}(\underline{|n|}, \underline{\sigma})}\big)$ by $(\neg\mathrm{IN})$,

$$\begin{aligned}
& T \vdash^{|n|}_{*} \Box^{\leqslant}(\underline{n}, \underline{\neg\sigma}) \wedge \neg\Box^{\leqslant}(\underline{|n|}, \underline{\sigma}) \to \Box^{\leqslant}\big(\underline{p^6(n)}, \underline{\bot}\big) && \text{by (i);} \\
\therefore\ & T \vdash^{n}_{*} \Box^{\leqslant}(\underline{n}, \underline{\neg\sigma}) \to \Box^{\leqslant}(\underline{|n|}, \underline{\sigma}) && \text{as } T \vdash^{p^7(n)} \neg\Box^{\leqslant}(\underline{p^6(n)}, \underline{\bot}) \text{ by (C);} \\
\therefore\ & T \vdash^{n}_{*} \Box^{\leqslant}(\underline{n}, \underline{\neg\sigma}) \to \Box^{\leqslant}(\underline{p^3(n)}, \underline{\bot}) && \text{by (i).}
\end{aligned}$$

□

From Theorem 26.4, one can easily derive what we claimed on page 126.

**Corollary 26.5** (Hrubeš)**.** Let $T$ be any consistent $\mathcal{L}_{\mathrm{A}}(\exp)$ theory extending $\mathrm{I}\Delta_0(\exp)$. Given an $\mathcal{L}_{\mathrm{A}}(\exp)$ formula $\Box^{\leqslant}(w, y)$ and $p(X), q(X) \in \mathbb{N}[X]$ satisfying (M)–(¬IN) in the statement of Theorem 26.4 for all $n \in \mathbb{N}$ and all $\mathcal{L}_{\mathrm{A}}(\exp)$ sentences $\sigma, \tau$, one can construct an $\mathcal{L}_{\mathrm{A}}(\exp)$ sentence $\sigma$ such that

- $T \not\vdash \sigma$ and $T \not\vdash \neg\sigma$; and
- $\exists k{\in}\mathbb{N}\ \ \forall n{\in}\mathbb{N}\ \ T \overset{2^{k|n|}+k}{\vdash} \neg\Box^{\leqslant}(\underline{n}, \underline{\neg\sigma})$.

*Proof.* Apply the Diagonal Lemma to find an $\mathcal{L}_{\mathrm{A}}(\exp)$ sentence $\sigma$ such that

$$T \vdash \sigma \leftrightarrow \neg\exists w\ \big(\Box^{\leqslant}(w, \underline{\sigma}) \wedge \neg\Box^{\leqslant}(2^w, \underline{\neg\sigma})\big).$$

Let $r(X), s(X) \in \mathbb{N}[X]$ given by an application of Theorem 26.4. As $T$ is consistent, we see from Theorem 26.4(1,2) that $T \not\vdash \sigma$ and $T \not\vdash \neg\sigma$. Moreover, since (C) implies $T \overset{p(r(n))}{\vdash} \neg\Box^{\leqslant}(\underline{r(n)}, \underline{\bot})$ for all $n \in \mathbb{N}$, we know $T \overset{n}{\vdash_*} \neg\Box^{\leqslant}(\underline{n}, \underline{\neg\sigma})$ by Theorem 26.4(4). $\square$

Although Corollary 26.5 produces a sentence $\sigma$ such that $T$ has exponential-length proofs of the finite consistencies of $T+\sigma$, it does not guarantee the same for $T+\neg\sigma$. In fact, Theorem 26.4 only gives a triply-exponential length-bound for $T$-proofs of the finite consistencies of $T+\neg\sigma$. Using a different variation of the witness comparison method, Jeřábek managed to produce another such $\sigma$ such that $T$ also has exponential-length proofs of the finite consistencies of $T+\neg\sigma$, thus restoring the symmetry between $T+\sigma$ and $T+\neg\sigma$ in the usual proof of the First Incompleteness Theorem.