数理解析研究所講究録 1268 巻 2002 年 126-137

Gödel's incompleteness theorem and forcing

Yasuhito Kawano
kawano@theory.brl.ntt.co.jp
NTT Communication Science Labs., NTT Corporation *

Abstract

A new approach to the P versus NP problem based on Gödel's incompleteness theorem and forcing is proposed. The assertion that the paradox of the unexpected hanging can be regarded as a kind of partial extension of the liar paradox is applied. Pronouncements are formalized in the language of extended Buss' bounded arithmetic. The pronouncements are shown to be both consistent and inconsistent when P=NP and an assumption hold. As a consequence, it is proved that P≠NP is reducible from another separation problem of an ideal from a Boolean algebra.

Keywords
P versus NP, Computational complexity, Forcing,
Gödel's incompleteness theorem, Bounded arithmetic, Paradox

1 Introduction

The structure of weak arithmetic systems described by S_2^i, which is called *bounded arithmetic*, is closely related to the structure of the polynomial time hierarchy. P≠NP is said to be deducible from the separation of bounded arithmetic theories. One well-known method for separating arithmetic theories combines Gödel's incompleteness theorem and the truth definition (cf. §7.6 in [3], page 140 in [7], and §10.5 in [8]). However, the separation of bounded arithmetic theories has not yet been proved using such a method. This is because the truth definition of Σ_i^b-formulas cannot be represented by a bounded formula in the language of S_2 (cf. [9]).

We generalize the above separation method by applying *the paradox of the unexpected hanging* [4], which is as follows. "One Saturday, the prisoner was told by the judge, 'You will be hanged at noon on one day next week. You will be hanged on a day you cannot predict.' The prisoner's lawyer proved that the hanging could not be executed by making the following argument: 'If the day of the hanging is Saturday, the prisoner will be able to predict it Friday afternoon because Saturday is the last day of the next week. This contradicts the judge's second pronouncement. Saturday is thus excluded. The execution can only take place on a day before Saturday. In the same way, each final day can be eliminated one by one.' "

It is now asserted that this problem is not a paradox because the day of the hanging is not predictable by the prisoner even if he (or she) makes full use of all the information contained in the pronouncements (cf. [2]). The author agrees with this assertion. However, this problem can still be applied to the separation of logical systems, given the following observation: *The paradox of the unexpected hanging can be regarded as a kind of partial extension of the liar paradox.* The separation method based on Gödel's incompleteness theorem and the truth definition can be naturally generalized by applying this observation. If we apply this generalized method to bounded arithmetic theories, the role of the truth definition is replaced by forcing. The $\mathcal{M}$-generic maximal filter constructed by forcing on bounded arithmetic naturally corresponds to the number meaning the day of the hanging.

If we apply this method to the P versus NP problem, it is proved that P≠NP is reducible from another separation problem of an ideal $\mathcal{I}$ from a Boolean algebra $\mathcal{B}$. (The precise definitions of $\mathcal{B}$ and $\mathcal{I}$ will be given in §6.2.) The proof sketch is described as follows. Suppose P=NP and $\mathcal{B} \neq \mathcal{I}$. There is an

*This work was done while the author belonged to NTT East Corporation.

extended Buss' bounded arithmetic in which P=NP is made true by adding an axiom representing P=NP. Furthermore, we formalize the two pronouncements in the language of this theory. First, we construct a model of the theory, on the assumption of $\mathcal{B} \neq \mathcal{I}$, such that the first pronouncement is true in it. Second, we show that the second pronouncement is proved in the theory. (This means that the pronouncements are consistent.) Finally, we prove that 0=1 is true in the model by developing the lawyer's argument. (This means that the pronouncements are inconsistent.) This is trivially a contradiction. Hence, we can conclude that $\mathcal{B} \neq \mathcal{I}$ implies P≠NP.

2 Paradox of the unexpected hanging

The paradox of the unexpected hanging is as follows [4]:

"The prisoner was sentenced by the judge on Saturday, 'The hanging will take place at noon on one of the seven days of next week. But you will not know which day it is until you are so informed on the morning of the hanging.'

The judge was known to be a man who always kept his word. The prisoner, accompanied by his lawyer, went back to his cell. After careful consideration, the lawyer proved that the judge's sentence could not possibly be carried out. 'They obviously cannot hang you next Saturday because Saturday is the last day of the week. On Friday afternoon you would still be alive and you would know with absolute certainty that the hanging would be on Saturday. You would know this before you were told so on Saturday morning. That would violate the judge's decree.'

'Saturday, then, is positively ruled out. This leaves Friday as the last day they can hang you. But they cannot hang you on Friday because by Thursday afternoon only two days would remain: Friday and Saturday. Since Saturday is not a possible day, the hanging would have to be on Friday. Your knowledge of that fact would violate the judge's decree again. So Friday is out. This leaves Thursday as the last possible day. But Thursday is out because if you're alive Wednesday afternoon, you'll know that Thursday is to be the day.'

'In exactly the same way you can rule out Wednesday, Tuesday and Monday. That leaves only tomorrow. But they cannot hang you tomorrow because you know it today.'

In brief, the judge's decree seems to be self-refuting. There is nothing logically contradictory in the two pronouncements that make up his decree; nevertheless, it can not be carried out in practice."

When Garner published the details of this paradox in an article titled "Mathematical Games" in Scientific American [4], there was a great public response. We will first rearrange the pronouncements from the viewpoint of application and make some assertions about the paradox. These interpretations do not relate to the work by Gardner.

The pronouncements are rearranged as follows:

- **First pronouncement**

  The hanging will take place on one of the seven days of next week.

- **Second pronouncement**

  The day of the hanging cannot be predicted by the prisoner.

The paradox of the unexpected hanging consists of two contradictory assertions:

- **Judge's assertion**

  The pronouncements are consistent.

- **Lawyer's assertion**

  The pronouncements are inconsistent.

Most people do not perceive an incompatibility between the pronouncements. Since the paradox of the unexpected hanging is said to be a problem of *time*, this imperception reflects our unconscious recognition of time. Results of an inference made now should be taken as a historical fact when making an inference in the future. When an inference made now is based on the results of an inference made in the future, the premise of the inference made now is changed to match the results of the inference made now. Thus, the inference including a time element may contain inconsistency if the inference is not restricted. This restriction can be represented by choosing the prisoner's ability to infer as a weak system. The lawyer uses a tacit understanding of the prisoner's ability to infer in his proof.

Conversely, this paradox can be used to show that the prisoner's ability to infer is weak. To show this, we first prove that the judge's assertion is right and then that the lawyer's assertion is right based on the assumption that the prisoner's ability to infer is sufficiently strong. More simply, if both the judge's and lawyer's assertions are proved based on the assumption that the prisoner's ability to infer is sufficiently strong, we can conclude that the prisoner's ability to infer is weak. With this method, we can intuitively prove that an inference including the notion of time is properly weaker than a standard one.

Furthermore, we can describe this paradox as follows:

Assertion. The structure of the inference-deducing inconsistency used by the lawyer in the paradox of the unexpected hanging is essentially the same as the structure of the inference deducing inconsistency in the liar paradox. The lawyer's inference consists of seven iterations of the inference in the liar paradox. The paradox of the unexpected hanging can be regarded as a kind of partial extension of the liar paradox in this sense.

Explanation. The inferences used in the liar paradox and the paradox of the unexpected hanging are compared. We first describe the one found in the liar paradox.

$$\begin{aligned} \phi &\iff \text{“}\phi \text{ is false”} \\ &\iff \text{“}\phi \text{ is true” is negative.} \end{aligned}$$

The upper arrow is the definition of ϕ. The lower arrow is the rewriting according to the definition of "false." The inference-deducing inconsistency from the positive assertion of ϕ is as follows.

Begin

1. ϕ (assumption)
2. “ϕ is true.” (immediately implied by 1)
3. If “ϕ is true,” then ϕ is negative. (by the definition of ϕ)
4. ϕ is negative. (from 2 and 3)

This contradicts 1.

End

To describe the paradox of the unexpected hanging, we first introduce the symbols used in the inference. The prisoner's ability to infer is denoted by T, a logical system. The day of the hanging is labeled $0, 1, \cdots, 6$. The first day is labeled 0 and called the 0-th day. $\phi(x)$ is the formula used to determine the day of the hanging. On the afternoon of the $(x-1)$-th day, the prisoner knows that the hanging has not been executed before the x-th day. Hence, the prisoner's ability to infer on the afternoon of the $(x-1)$-th day, denoted by T_x, can be represented by $T + \{(\forall y < x)\neg\phi(y)\}$.

The first pronouncement can be represented as follows:

$$(\exists x < 7)\phi(x).$$

The second pronouncement can be represented as follows:

For all x such that x is less than 7,

$$\begin{aligned} \phi(x) &\Longrightarrow \text{“}\phi(x) \text{ is not predictable by } T_x.\text{”} \\ &\Longleftrightarrow \text{“}\phi(x) \text{ is predictable by } T''_x \text{ is negative.} \end{aligned}$$

The inference used by the lawyer is naturally described as follows:

Begin

$x := 6$.

1. $(\exists y \leq x)\phi(y)$ (the first pronouncement)
2. “$\phi(x)$ is predictable by T_x.” (immediately implied by 1)
3. If “$\phi(x)$ is predictable by T_x,” then $\phi(x)$ is negative. (by the second pronouncement)
4. $\phi(x)$ is negative. (from 2 and 3)

If $x = 0$, then 4 contradicts 1, else $(\exists y \leq x - 1)\phi(y)$ is obtained.
$x := x - 1$ and go to 1.

End

Comparing these inferences, the assertion is easily grasped. “ϕ is true” in the liar paradox corresponds to “$\phi(x)$ is predictable by T_x” in the paradox of the unexpected hanging. □

A special framework, like temporal logic, is not needed to formalize this paradox. However, the notion of predictability must be expressed. Since it is natural to express this notion by provability, a strong theory in which meta-notions can be represented should be used to formalize this paradox. Peano arithmetic is one such theory. However, the pronouncements will be inconsistent if expressed naturally in Peano arithmetic because Peano arithmetic is too strong for formalizing them. bounded arithmetic is better because it is neither too weak nor too strong.

3 Relation to the classical separation method

Our separation method using the paradox of the unexpected hanging looks quite new. However, it is related to the well-known classical method (cf. §7.6 in [3]). We will now explain the relation between our proof and the one for $I\Sigma_1 \neq I\Sigma_2$, because our plan for proving the main theorem corresponds to the structure of the proof of $I\Sigma_1 \neq I\Sigma_2$.

The proof consists of two statements.

1. $I\Sigma_2 \vdash Con(I\Sigma_1)$
2. $I\Sigma_1 \nvdash Con(I\Sigma_1)$

Intuitively, $I\Sigma_1$ and $I\Sigma_2$ are separated by $Con(I\Sigma_1)$. In other words, both consistency and inconsistency of $I\Sigma_1 + \{\neg Con(I\Sigma_1)\}$ are proved based on the assumption $I\Sigma_1 = I\Sigma_2$. Inconsistency is proved by showing that there is no model of $I\Sigma_1 + \{\neg Con(I\Sigma_1)\}$, and the truth definition of Σ_1-formulas plays an important role in the proof. Consistency is proved by Gödel's incompleteness theorem; this proof is based on the liar paradox.

In our main proof, we will use T, which will be defined in the next section, instead of $I\Sigma_1$ $(= I\Sigma_2)$. Both consistency and inconsistency of $T + \{$the two formalized pronouncements$\}$ will be proved. Here, T is roughly defined as the theory $S_2 + \{\text{P=NP}\}$. Inconsistency is proved by the lawyer's argument, based on an extension of the argument of the liar paradox. Consistency is proved by showing that there is a model of $T + \{$the two formalized pronouncements$\}$, that can be proved by forcing. Our method thus twistingly corresponds to the proof of $I\Sigma_1 \neq I\Sigma_2$.

4 Relation to P versus NP problem

We have explained why the paradox of the unexpected hanging is related to separation of logical systems. We will now intuitively explain why it is related to the P versus NP problem.

As a concrete example of NP-complete problems, we consider the clique problem (cf. page 47 in [5]). Given a graph (V, E) and a positive integer $J \leq |V|$, let $V_0, V_1, \cdots, V_n$ be an enumeration of subsets of V (vertex). Then, the clique problem consists of $n+1$ subproblems. For each $i \leq n$, it is easily checked whether V_i is a clique and $J \leq |V_i|$, so each subproblem is in P. In other words, the clique problem is a set of many subproblems such that each of them is in P. It is believed that there is no polynomial time computable function calculating a true one from $n+1$ subproblems (i.e. P$\neq$NP). This means that there is no algorithm much better than checking them one by one. If P=NP, there is a polynomial time computable function such that it calculates a true one from $n+1$ subproblems.

Notions in the paradox of the unexpected hanging can be expressed by using notions of computational theory. For example, the prisoner's ability to infer is expressed by a Turing machine, predictability is expressed by non-deterministic polynomial time computability, and so on. "Is the day of the hanging predictable by the prisoner's ability to infer?" can then be considered as a problem of computational complexity. Here, the length of the input is a log of the maximum time to execute, because pronouncements can be coded by words whose length is bounded by a polynomial length of a log of the maximum time to execute. On the other hand, the proof by the lawyer can be expressed by an instantaneous description (ID). However, the length of the ID is exponentially longer than the input length, because he eliminates candidates of the hanging one by one. Since we now regard predictability as polynomial time computability, that hanging will occur on the first day is perhaps unpredictable by the prisoner. (Because the prisoner needs an exponential length ID to prove that the hanging is the first day.) However, if P=NP, there may be a polynomial time computable function such that it calculates the day of the hanging, like the case of the clique problem.

5 Expressing notions in the paradox of the unexpected hanging

We show definitions how we express the notions used in the paradox of the unexpected hanging in bounded arithmetic.

5.1 Prisoner's ability to infer

The prisoner's ability to infer is expressed by a computational system such as a Turing machine. In this paper, the prisoner's ability to infer is expressed by a theory T of bounded arithmetic defined by extending theory S_2. Let $\tilde{S}_2$ be a theory such that the language of $\tilde{S}_2$ has all symbols in the language of S_2 plus all symbols introduced in §2.4–§2.5 of [3] plus symbols for representing P=NP later, and the axioms of $\tilde{S}_2$ has all axioms of S_2 plus all axioms defining symbols introduced in §2.4–§2.5 of [3]. The language of T is completely the same as that of $\tilde{S}_2$. T has all the axioms in $\tilde{S}_2$. The only difference between T and $\tilde{S}_2$ is that T contains axioms representing P=NP. This approach was also used by Takeuti [11].

Definition 1 (The definition of $(\exists x \leq \mathbf{t}(a))\mathbf{A}(x,a)$) *$(\exists x \leq \mathbf{t}(a))\mathbf{A}(x,a)$ is an NP-complete formula w.r.t. $\tilde{S}_2$ such that $\mathbf{t}$ is a term and $\mathbf{A}(a_1, a_2)$ is a sharply bounded formula in the language of $\tilde{S}_2$. That is, for any Σ^b_1-formula $A(a)$ in the language of $\tilde{S}_2$, there is a polynomial time computable function $f_A(a)$ such that*

$$\tilde{S}_2 \vdash A(a) \leftrightarrow (\exists x \leq \mathbf{t}(f_A(a)))\mathbf{A}(x, f_A(a)).$$

Remark. We selected one NP-complete formula $(\exists x \leq \mathbf{t}(a))\mathbf{A}(x,a)$ and used it consistently in the following argument. T is constructed by adding a new axiom for it, so T depends on the choice of this NP-complete formula. The main theorem will hold independently of its selection. The reader who wants to avoid confusion due to this ambiguity can choose a concrete formula; $(\exists x \leq \mathbf{t}(a))\mathbf{A}(x,a)$, for example, is defined as a formalization of the class of clique problems. However, we do not know of any concrete

form of $\mathbf{f}(a)$ that can be used in our next definition. □

Definition 2 (Definition of f) *Since P=NP, there is a polynomial time computable function* $\mathbf{f}$ *that satisfies*

$$(\exists x \leq \mathbf{t}(a))\mathbf{A}(x,a) \leftrightarrow \mathbf{f}(a) \leq \mathbf{t}(a) \wedge \mathbf{A}(\mathbf{f}(a),a).$$

We selected such a function, $\mathbf{f}(a)$*, and use it consistently in this paper.*

Now we define T.

Definition 3 (Definition of T) *The language of* T *is the same as the language of* $\tilde{S}_2$*.* T *has all the axioms of* $\tilde{S}_2$*. Additionally,* T *has the following new axioms:*

$$x \leq \mathbf{t}(a), \mathbf{A}(x,a) \rightarrow \mathbf{f}(a) \leq \mathbf{t}(a), \text{ and}$$
$$x \leq \mathbf{t}(a), \mathbf{A}(x,a) \rightarrow \mathbf{A}(\mathbf{f}(a),a).$$

Furthermore, T *has defining axioms to calculate the value of* $\mathbf{f}(x)$ *for each value of* x*: these axioms are introduced using limited iteration (§1.1 in [3]).*

5.2 Judge

The judge is expressed by the 'universe' in which the prisoner thinks about the day of the hanging. It is naturally expressed by a model of T. However, it is not necessarily a standard model. In this paper, a non-standard arithmetic model $M[G]$ constructed by forcing will be selected as the judge.

5.3 Maximum time to execute

The maximum time to execute is naturally expressed by a number in the model. In this paper, it is a non-standard number, denoted by n. Each day of the hanging is labeled by numbers $0, 1, 2, \cdots, n$. The first day is labeled 0, and the last day is n.

5.4 The day of the hanging

The day of the hanging is determined by the number x ($\leq n$) such that $\phi(x)$ is true in the model. It is available that $\phi(x)$ is true for more than two values of x. (Though this means that the hanging will be executed more than two times.) In this paper, $\phi(a)$ is defined as a Π_1-formula meaning "The prisoner cannot predict a hanging on the $(a-1)$-th day" as follows.

Let f_0 be the function defined by

$$f_0(a) \stackrel{\text{def}}{=} \ulcorner((\exists x_0 < a_1)^\urcorner ** (Sub(a, \ulcorner a_1 \urcorner, \ulcorner x_0 \urcorner) * \ulcorner)^\urcorner);$$

i.e.

$$\begin{aligned} f_0 \ : \ & \mathbb{N} \longrightarrow \mathbb{N} \\ & \ulcorner \psi(a_1)^\urcorner \mapsto \ulcorner((\exists x_0 < a_1)\psi(x_0))^\urcorner. \end{aligned}$$

We define $\psi(a_1, a_2)$ as

$$\psi(a_1,a_2) \stackrel{\text{def}}{=} (\forall x_1)\neg Prf_T(x_1, \overrightarrow{FSub}(f_0(a_2), \ulcorner \vec{a} \urcorner, \vec{a})),$$

where $\vec{a} = (a_1, a_2)$. We then set $\xi \stackrel{\text{def}}{=} \ulcorner \psi(a_1,a_2)^\urcorner$, and define

$$\phi(a_1) \stackrel{\text{def}}{=} \psi(a_1, \xi).$$

Consequently, $\phi(a_1)$ is a Π_1-formula in the language of $\tilde{S}_2$.

Then, $\tilde{S}_2^1$ proves

$$\begin{aligned}
\phi(0) &\leftrightarrow (\forall w)\neg Prf_T(w, \ulcorner(\exists x < 0)\phi(x)\urcorner) \\
\phi(1) &\leftrightarrow (\forall w)\neg Prf_T(w, \ulcorner(\exists x < 1)\phi(x)\urcorner) \\
\phi(2) &\leftrightarrow (\forall w)\neg Prf_T(w, \ulcorner(\exists x < 2)\phi(x)\urcorner) \\
&\cdots\cdots .
\end{aligned}$$

Trivially, $\phi(0)$ is equivalent to the consistency statement of T. It can be said that the formula $\phi(a)$ is an extension of the Gödel sentence. Intuitively, $\phi(a)$ is a formalization of "The prisoner cannot predict a hanging on the $(a-1)$-th day."

5.5 Predictability

Defining predictability is very important when we express notions in the paradox of the unexpected hanging in a formal system. Predictability will be expressed by bounded provability in this paper.

By the well-known method (cf. [3]), the notion "w is the Gödel number of a proof in T of a sequent whose Gödel number is v" is formalized in the language of bounded arithmetic. This formula will be denoted by $Prf_T(w, v)$.

As we express predictability by bounded provability, we have to determine a term that bounds lengths of proofs: in other words, the 'horizon' of the prisoner's thought. For any term $t(a)$, there is a number κ_2 such that $t(a)$ is bounded by $\underbrace{(a+2)\sharp(a+2)\sharp\cdots\sharp(a+2)}_{\kappa_2}$. $\underbrace{(a+2)\sharp(a+2)\sharp\cdots\sharp(a+2)}_{\kappa_2}$ is abbreviated as $\sharp^{\kappa_2}(a+2)$. We determine κ_2 as a very large number, as follows.

Definition 4 (Definitions of κ_2) *κ_2 is a number such that*

$$T \vdash A(\vec{a}) \rightarrow Thm_T(\sharp^{\kappa_2}(a_4+2), \overrightarrow{FSub}(\ulcorner A(\vec{a})\urcorner, \ulcorner\vec{a}\urcorner, \vec{a})),$$

where $\vec{a} = (a_1, a_2, a_4)$ and $A(a_1, a_2, a_4)$ is any Π_2^b-formula,

$$(a_1 < \forall x \leq a_2)\, Thm_T(a_4, \overrightarrow{FSub}(\eta, \ulcorner(x, a_2)\urcorner, (x, a_2))).$$

(η is a Gödel number of a formula.)

The existence of κ_2 is guaranteed from the theorem by Buss [3], since we assume P=NP. In contrast, the existence of κ_2 is not guaranteed if P$\neq$NP. Intuitively, $\sharp^{\kappa_2}(a+2)$ is an exponent of $|a|$ since P$\neq$NP. Let $\tau(a)$ be a term $\sharp^{\kappa_2^2}(a+2)$. We define τ-provability as predictability.

5.6 Preliminary description of pronouncements

We define $\phi(a_1)$ to mean that 'The hanging takes place on the a_1-th day.' For example, $\phi(0)$ means that 'The hanging takes place on the first day.' Let a_2 be a free variable denoting the maximum time to execution. The first and second pronouncements depend on ϕ and a_2, so they are denoted by $J_1(\phi, a_2)$ and $J_2(\phi, a_2)$. More precise ones are given in Lemmas 2 and 3.

Additionally, $\varphi(a_1)$ is defined as $(\exists x_0 < a_1)\phi(x_0)$, i.e.

$$\varphi(a_1) \stackrel{\text{def}}{=} (\exists x_0 < a_1)\phi(x_0).$$

The first pronouncement is then naturally denoted by $\varphi(a_2)$. However, we make a further claim about the recognition by the prisoner. This claim is represented by "There is a short proof of the first pronouncement." This is represented by a Σ_1^b-formula $\overline{\varphi}(a)$ defined by

$$\overline{\varphi}(a_2) \stackrel{\text{def}}{=} Thm_T(\sigma(a_2), FSub(\ulcorner\varphi(a_2)\urcorner, \ulcorner a_2\urcorner, a_2)),$$

where $\sigma(a_2)$ is $(a_2+2)\sharp(a_2+2)$.

The first pronouncement is then defined as

$$J_1(\phi, a_2) \quad : \quad \varphi(a_2) \wedge \overline{\varphi}(a_2).$$

Next, we present a preliminary formalization of the second pronouncement. The prisoner's ability to infer before he hears the pronouncements is defined as T. The prisoner on the afternoon of the (a_1-1)-th day knows that the hanging did not occur before the a_1-th day. Therefore, it can be represented by $T+\{(\forall x < a_1)\neg\phi(x)\}$.

'The hanging on the a_1-th day cannot be predicted by the prisoner' is intuitively represented by

$$\text{There is no short proof of } \phi(a_1) \text{ in } T+\{(\forall x < a_1)\neg\phi(x)\}.$$

This is almost the same as

$$\text{There is no short proof of } \varphi(a_1+1) \text{ in } T$$

by deduction theorem. Let $\tau(a_2)$ be $\sharp^{\kappa_2^2}(a_2+2)$, where κ_2 is the number defined in Definition 4. This term is the length bound on the proofs available to the prisoner. 'The hanging on the a_1-th day cannot be predicted by the prisoner' can then be represented by

$$\neg \mathit{Thm}_T(\tau(a_2), \mathit{FSub}(\ulcorner\varphi(a_1+1)\urcorner, \ulcorner a_1\urcorner, a_1)).$$

The lawyer interprets the second pronouncement to mean that "If the prisoner can predict the day of execution, the hanging cannot take place on that day." Pronouncement $J_2(\phi, a_2)$ is thus denoted by

$$J_2(\phi, a_2) \ : \ (\mathit{Thm}_T(\tau(a_2), \mathit{FSub}(\ulcorner\varphi(a_1+1)\urcorner, \ulcorner a_1\urcorner, a_1)) \rightarrow \neg\phi(a_1)) \text{ for all } a_1 < a_2.$$

6 Main theorem

Main Theorem. $\mathcal{B} \neq \mathcal{I}$ implies P≠NP.

The precise definitions of $\mathcal{B}$ and $\mathcal{I}$ will be given in §6.2. The outline of the proof of our main theorem is described as follows. More detailed explanations will be given in §6.1–§6.4

Outline of proof. Suppose P=NP and $\mathcal{B} \neq \mathcal{I}$. Our purpose is to show a contradiction.

(6.1 $T+\{\overline{\varphi}(n+1)\}+\mathit{exp}$ is consistent.) We show that there is a countable non-standard model, named N, of $T+\{\overline{\varphi}(n+1)\}+\mathit{exp}$. N will be used as the ground model of forcing. The existence of N is proved by Corollary 8.14 in [14]. A non-standard number $n \in N$, which means the maximum time to execution minus one, will also be selected in this subsection.

(6.2 $T+\{\varphi(n+1)\}+\{\overline{\varphi}(n+1)\}$ is consistent.) A non-standard model, $M[G]$, of T will be constructed by forcing, which is the same method as [12]. $M[G] \models \overline{\varphi}(n+1)$ since the forcing extension on bounded arithmetic preserves the basic properties of the ground model. On the other hand, $M[G] \models \varphi(n+1)$ is made true since a non-standard number α such that $M[G] \models \phi(\alpha)$ is added by forcing. Here, the condition $\mathcal{B} \neq \mathcal{I}$ is necessary for constructing the model $M[G]$.

Then, we obtain $M[G] \models T+J_1(\phi, n+1)$.

(6.3 $T+J_1(\phi, n+1)+J_2(\phi, n+1)$ is consistent.) The second pronouncement $J_2(\phi, n+1)$ is proved in T. This means $M[G] \models T+J_1(\phi, n+1)+J_2(\phi, n+1)$.

(6.4 $T + J_1(\phi, n+1) + J_2(\phi, n+1)$ is inconsistent.) We show that $T + J_1(\phi, n+1) + J_2(\phi, n+1)$ implies inconsistency.

Since $M[G]$ is a model of T, we have $M[G] \models 0 = 1$. This is a contradiction. (End of outline)

6.1 $T + \{\overline{\varphi}(n+1)\} + exp$ is consistent.

Lemma 1 *There is a countable non-standard model N of $T + exp$ and a non-standard number $n \in N$ such that*

$$N \models \overline{\varphi}(n+1),$$

where $n + 1 = 2^{n_0}$ for some $n_0 \in N$.

We select a countable non-standard model N and a non-standard number $n \in N$, and use them consistently in this paper.

6.2 $T + \{\varphi(n+1)\} + \{\overline{\varphi}(n+1)\}$ is consistent.

Definition 5 (Definition of M) *M is defined as the initial segment of N:*

$$M \stackrel{\text{def}}{=} \{x \in N | \text{ there exists some } n\sharp \cdots \sharp n \text{ such that } x \leq n\sharp \cdots \sharp n\}.$$

Then, M is a model of T. Obviously,

$$M \models \overline{\varphi}(n+1).$$

M is determined according to the non-standard element $n \in N$.

n_0 is defined as $|n|$. Let $\mathcal{B}$ be the same as the Boolean algebra defined in [12]. A new number, α ($< n+1$), which is probably not included in M, is added to model M by forcing. Intuitively, this number corresponds to the day of the hanging. Only the $a_2 = n+1 (= 2^{n_0})$ case will be considered in the following argument.

ϕ_ν is defined by

$$\phi_\nu(a_1) \stackrel{\text{def}}{=} (\forall w \leq \sharp^\nu(n+1)) \neg Prf_T(w, FSub(\ulcorner \varphi(a_1) \urcorner, \ulcorner a_1 \urcorner, a_1)).$$

The second-order bounded formula corresponding to $\phi_\nu(x)$ is denoted by the same symbol, $\phi_\nu(X)$, where X is a subset of $\{i | i < n_0\}$. Since any polynomial time computable function is represented by a circuit, $\phi_\nu(X)$ can be translated into a Boolean circuit.

Let $\mathcal{C}_\nu$ be this Boolean circuit. In it, each input terminal is labeled as a variable or a constant (0 or 1). Without loss of generality, we can assume that $\mathcal{C}_\nu$ has n_0 input variable terminals and one output terminal, because $|X|$ is always bounded by n_0. Input variables will be denoted by $x_0, x_1, \cdots, x_{n_0 \dot{-} 1}$. Nodes except for input terminals are called gates. They are labeled $\wedge$, $\vee$, or $\neg$. The gates of $\mathcal{C}_\nu$ can be coded by a computable function because $\phi_\nu(x)$ consists of polynomial time computable functions.

If we let X be a subset of $\{i | i < n_0\}$, we can regard X as an input by setting $x_i = 1$ if $i \in X$ and $x_i = 0$ if $i \notin X$ for any $i < n_0$. The value of the output when X is input into $\mathcal{C}_\nu$ is denoted by $\mathcal{C}_\nu(X)$.

X_G is defined as a function from $\{i | i < n_0\}$ to $\mathcal{B}$

$$X_G : i \mapsto x_i.$$

We define $i_G(X)$ as

$$i_G(X) = \{x < y | X(x) \in G\}.$$

The value of the output when X_G is input into $\mathcal{C}_\nu$ is denoted by $\mathcal{C}_\nu(X_G)$; α $(< n+1)$ is defined as the number corresponding to X_G, i.e.

$$(\text{The } j\text{-th bit of } \alpha) = 1 \quad \text{iff} \quad j \in i_G(X_G).$$

We transform $\mathcal{C}_\nu$ into a conjunctive normal form, ψ_ν, in N. $\mathcal{C}_\nu$ cannot be transformed into ψ_ν in M because many clauses appear in ψ_ν. Let

$$\begin{aligned} \psi_\nu &= \bigwedge_{i \in 2^{\mathcal{K}_\nu}} \psi_{\nu,i} \\ \psi_{\nu,i} &= \bigvee_{j \in C_{\nu,i}} x_j \vee \bigvee_{j \in \bar{C}_{\nu,i}} \bar{x}_j, \end{aligned} \tag{1}$$

where each $\psi_{\nu,i}$ $(i \in 2^{\mathcal{K}_\nu})$ is called a clause. $\{x_0, \cdots, x_{n_0-1}\}$ are atomic Boolean variables, and x_j and $\bar{x}_j$ correspond to $Bit(j,x) = 1$ and $Bit(j,x) = 0$, respectively. $2^{\mathcal{K}_\nu}$ is the set of indexes of clauses. $C_{\nu,i}$ and $\bar{C}_{\nu,i}$ are subsets of $\{j | j < n_0\}$. We assume $C_{\nu,i} \cap \bar{C}_{\nu,i} = \emptyset$, because otherwise $\bigvee_{j \in C_{\nu,i}} x_j \vee \bigvee_{j \in \bar{C}_{\nu,i}} \bar{x}_j = 1$, meaning it can be eliminated from the decomposition (1) of ψ_ν. However, $C_{\nu,i} \cup \bar{C}_{\nu,i} = \{j | j < n_0\}$ does not generally hold.

We introduce the ideal $\mathcal{I}$ that will be used for the forcing.

Definition 6 (Definitions of $\mathcal{I}$) *$\mathcal{I}$ is defined as the M_0-complete ideal generated from*

$$\{\neg\psi_{\nu,i} | \nu \in \mathbb{N}, i \in 2^{\mathcal{K}_\nu}\} \tag{2}$$

in $\mathcal{B}$, where $\psi_{\nu,i}$ is the clause defined in (1).

More precisely, $b \in \mathcal{I}$ iff there is a function $\gamma : c \to \mathcal{B}$ such that

1. *$c \in M_0$,*
2. *there is a finite subset $\{\neg\psi_{\nu_1^k, i_1^k}, \cdots, \neg\psi_{\nu_{j_k}^k, i_{j_k}^k}\}$ of (2) such that*

$$\neg\psi_{\nu_1^k, i_1^k} \vee \cdots \vee \neg\psi_{\nu_{j_k}^k, i_{j_k}^k} \geq \gamma(k) \text{ in } \mathcal{B}$$

   for all $k < c$, and

3. *$b \leq \bigvee_{k<c} \gamma(k)$.*

Obviously, $\mathcal{I} \subseteq \mathcal{B}$. $\mathcal{I}$ cannot be defined in M.

It is proved that for any $\nu \in \mathbb{N}$ and for any $i \in 2^{\mathcal{K}_\nu}$, $\{b \in \mathcal{B} | b \leq \psi_{\nu,i}\}$ is definable and dense over $\mathcal{I}$. There does not exist an $\mathcal{M}$-generic maximal filter over $\mathcal{I}$ without the condition $\mathcal{B} \neq \mathcal{I}$, because it should be contained in $\mathcal{B} \setminus \mathcal{I}$.

Definition 7 *G is an $\mathcal{M}$-generic maximal filter over $\mathcal{I}$. $M[G]$ is defined in the same way as in [12].*

The existence of G is guaranteed by the assumption $\mathcal{B} \neq \mathcal{I}$. Then, $M[G]$ is a model of T and $M[G]$ is a bounded extension of M by [12] since P=NP is assumed.

Lemma 2 $M[G] \models \varphi(n+1) \wedge \overline{\varphi}(n+1)$

6.3 $T + J_1(\phi, n+1) + J_2(\phi, n+1)$ is consistent.

The formalization of the second pronouncement is proved by theory T. It is strictly described as follows.

Lemma 3 *There is a finite number, κ_0, such that for each value of $\kappa_2 \geq \kappa_0$, $\tilde{S}_2^1$ proves*

$$a_1 \leq a_2, Thm_T(\tau(a_2), FSub(\ulcorner\varphi(a_1+1)\urcorner, \ulcorner a_1 \urcorner, a_1)) \to \neg\phi(a_1),$$

where $\tau(a_2) = \sharp^{\kappa_2^2}(a_2+2)$. Furthermore, there is a polynomial function of κ_2 such that for each value of κ_2 there is a proof for the above sequent, whose length is bounded by a polynomial function of κ_2.

Then, we can conclude $M[G] \models T + J_1(\phi, n+1) + J_2(\phi, n+1)$.

Gödel's incompleteness theorem and forcing

Yasuhito Kawano
kawano@theory.brl.ntt.co.jp
NTT Communication Science Labs., NTT Corporation *

Abstract

A new approach to the P versus NP problem based on Gödel's incompleteness theorem and forcing is proposed. The assertion that the paradox of the unexpected hanging can be regarded as a kind of partial extension of the liar paradox is applied. Pronouncements are formalized in the language of extended Buss' bounded arithmetic. The pronouncements are shown to be both consistent and inconsistent when P=NP and an assumption hold. As a consequence, it is proved that P≠NP is reducible from another separation problem of an ideal from a Boolean algebra.

Keywords
P versus NP, Computational complexity, Forcing,
Gödel's incompleteness theorem, Bounded arithmetic, Paradox

1 Introduction

The structure of weak arithmetic systems described by S_2^i, which is called *bounded arithmetic*, is closely related to the structure of the polynomial time hierarchy. P≠NP is said to be deducible from the separation of bounded arithmetic theories. One well-known method for separating arithmetic theories combines Gödel's incompleteness theorem and the truth definition (cf. §7.6 in [3], page 140 in [7], and §10.5 in [8]). However, the separation of bounded arithmetic theories has not yet been proved using such a method. This is because the truth definition of Σ_i^b-formulas cannot be represented by a bounded formula in the language of S_2 (cf. [9]).

We generalize the above separation method by applying *the paradox of the unexpected hanging* [4], which is as follows. "One Saturday, the prisoner was told by the judge, 'You will be hanged at noon on one day next week. You will be hanged on a day you cannot predict.' The prisoner's lawyer proved that the hanging could not be executed by making the following argument: 'If the day of the hanging is Saturday, the prisoner will be able to predict it Friday afternoon because Saturday is the last day of the next week. This contradicts the judge's second pronouncement. Saturday is thus excluded. The execution can only take place on a day before Saturday. In the same way, each final day can be eliminated one by one.' "

It is now asserted that this problem is not a paradox because the day of the hanging is not predictable by the prisoner even if he (or she) makes full use of all the information contained in the pronouncements (cf. [2]). The author agrees with this assertion. However, this problem can still be applied to the separation of logical systems, given the following observation: *The paradox of the unexpected hanging can be regarded as a kind of partial extension of the liar paradox.* The separation method based on Gödel's incompleteness theorem and the truth definition can be naturally generalized by applying this observation. If we apply this generalized method to bounded arithmetic theories, the role of the truth definition is replaced by forcing. The $\mathcal{M}$-generic maximal filter constructed by forcing on bounded arithmetic naturally corresponds to the number meaning the day of the hanging.

If we apply this method to the P versus NP problem, it is proved that P≠NP is reducible from another separation problem of an ideal $\mathcal{I}$ from a Boolean algebra $\mathcal{B}$. (The precise definitions of $\mathcal{B}$ and $\mathcal{I}$ will be given in §6.2.) The proof sketch is described as follows. Suppose P=NP and $\mathcal{B} \neq \mathcal{I}$. There is an

*This work was done while the author belonged to NTT East Corporation.

6.4 $T + J_1(\phi, n+1) + J_2(\phi, n+1)$ is inconsistent.

Inconsistency of the pronouncements cannot be deduced by coding the lawyer's logic because an extremely long proof would be needed. The longer the maximum time to execution, the exponentially larger the proof. There is not a term capable of bounding such a series of proofs. The assumption NP=co-NP makes it possible to deduce inconsistency using a short proof.

Lemma 4 *T proves*

$$a_1 \leq a_2, \varphi(a_2), \overline{\varphi}(a_2) \rightarrow \varphi(a_1).$$

Proof of Main Theorem. Suppose P=NP and $\mathcal{B} \neq \mathcal{I}$. Lemmas 1 – 4 are thus true. If we set $a_1 = 0$ in the assertion of Lemma 4,

$$T \vdash \varphi(a_2), \overline{\varphi}(a_2) \rightarrow (\exists x < 0)\phi(x), \tag{3}$$

since $\varphi(0) = (\exists x < 0)\phi(x)$ by the definition of φ. Let $M[G]$ be the model defined in Definition 7. Then, $M[G]$ is a model of T, so (3) implies

$$M[G] \models \varphi(n+1) \wedge \overline{\varphi}(n+1) \rightarrow (\exists x < 0)\phi(x)$$

if we set $a_2 = n + 1$. We have already proved $M[G] \models \varphi(n+1) \wedge \overline{\varphi}(n+1)$ in Lemma 2. Hence, $M[G] \models (\exists x < 0)\phi(x)$. This implies $M[G] \models (\exists x)(x < 0)$. However, $M[G] \models (\forall x)(x \geq 0)$ by the third axiom of BASIC [3]. Hence, $M[G] \models 0 = 1$. This is a contradiction. □

References

[1] J. L. Balcázar, J. Díaz, J. Gabarró, Structural Complexity I, second edition, (Springer, Berlin, 1995).

[2] B. H. Bunch, Mathematical Fallacies and Paradoxes, (Van Nostrand Reinhold Company, New York, 1982).

[3] S. Buss, Bounded Arithmetic, (Bibliopolice, Napoli, 1986).

[4] M. Gardner, Mathematical Games, A new paradox, and variations on it, about a man condemned to be hanged, Scientific American **208** (1963) pp.144–154.

[5] M. R. Garey, D. S. Johnson, Computers and Intractablity, (W. H. Freeman and Company, New York,1979).

[6] P. Hájek, P. Pudlák, Metamathematics of First-Order Arithmetic, (Springer, Berlin, 1993).

[7] R. Kaye, Models of Peano Arithmetic, Oxford Logic Guides **15**, (Oxford university press, New York, 1991).

[8] J. Krajíček, Bounded arithmetic, propositional logic, and complexity theory, (Cambridge university press, New York, 1995).

[9] G. Takeuti, Bounded Arithmetic and Truth Definition, Annals of Pure and Applied Logic **39** (1988) pp.75–104.

[10] G. Takeuti, RSUV isomorphisms, in: P. Clote and J. Krajíček, ed., Arithmetic, Proof Theory and Computational Complexity, Oxford Logic Guides **23** (Oxford university press, New York, 1993) pp.364–386.

[11] G. Takeuti, Incompleteness theorem and S_2^i versus S_2^{i+1}, Lecture Notes in Logic **12** (1998).

G. Takeuti, M. Yasumoto, Forcing on Bounded Arithmetic, Gödel '96, Lecture N (1996) pp.120–138.

G. Takeuti, M. Yasumoto, Forcing on Bounded Arithmetic II, Journal of Symbolic (1998) pp.860–868.

A. J. Wilkie, J. B. Paris, On the scheme of induction for bounded arithmetic form Pure and Applied Logic **35** (1987) pp.261–302.