# Diagonalization & The Fixed Point Lemma

Brendan Cordy

**Question**: How can we write a statement which is true iff Alice is reading it?

*Alice is reading this statement*

Ok, that was easy. What about if we're not allowed to use the word 'this'? We would like to do something like...

*Alice is reading the statement 'Alice is reading the statement'*

But that doesn't work. If Alice is reading it, then she's reading the statement *Alice is reading the statement 'Alice is reading the statement'*, which isn't the same as *Alice is reading the statement*, which our statement above claims she's reading. This is like trying to write a computer program that prints its own source code. Simply writing print(stuff) won't work, because the output won't contain the print command, and more print commands will only dig you deeper.

We can use a very clever trick to close the loop. To make it work, we need to allow variables and substitution in our statements. Something like $r(x) =$ *a rabbit eats* $x$ is a valid statement, and it makes sense to substitute other statements for $x$. If we substitute $c =$ *a delicious carrot*, then we obtain $r(c) =$ *a rabbit eats a delicious carrot*. We can make nonsensical substitutions as well, like $b =$ *the sky is blue*, to obtain $r(b) =$ *a rabbit eats the sky is blue*.

Note that statements in quotes are interpreted literally. For example, the statement *a rabbit eats a delicious carrot* means that a rabbit eats a delicious carrot, while *a rabbit eats 'a delicious carrot'* means that there is a rabbit on this piece of paper eating those words quoted above. Similarly, the statement *my favourite variable is the variable* $x$ has one variable, while the statement *my favourite variable is 'the variable* $x$*'* has no variables.

Now the crux of the trick. The diagonal function takes any quoted statement '$s(x)$' and replaces it with $s($'$s(x)$'$)$. We call this process diagonalization. Consider, for example, the quoted statement '*Alice is reading the statement* $x$'. Its diagonalization is *Alice is reading the statement 'Alice is reading the statement* $x$*'*, which looks very much like something we've tried before. It seems like we're not making much progress, but in fact we're very close to a solution.

The trick is to refer to the diagonal function <u>inside</u> the statement. Consider the following...

Alice is reading the diagonalization of 'Alice likes $x$'

That statement is true exactly when Alice is reading the statement *Alice likes 'Alice likes* $x$*'*. Now we've got the problem cornered. Can you see how to do it? We just need to make the statements inside and outside the quotes match...

Alice is reading the diagonalization of 'Alice is reading the diagonalization of $x$'

That's it! We've closed the loop and created a sentence which is true iff Alice is reading it. Even better, the construction we've just performed is an informal proof of a key part of Gödel's first incompleteness theorem, often known as the diagonal lemma or fixed-point lemma. Let's prove it, and you'll see that it's really the same argument with more formal symbols.

Recall that any formula $\psi$ in a suitable first-order language $\mathcal{L}_A$ for arithmetic can be encoded as a Gödel number, denoted $\ulcorner\psi\urcorner$. Similarly, any $k \in \mathbb{N}$ can be decoded to produce an $\mathcal{L}_A$-expression denoted $\varphi_k$. These two operations are mutually inverse, that is, $\ulcorner\varphi_k\urcorner = k$, and $\varphi_{\ulcorner\psi\urcorner} = \psi$. For the details of how this can be done, see [1].

We can use these encoding and decoding operations to define the diagonal function, $d : \mathbb{N} \to \mathbb{N}$, which can be understood as $d(n) = \ulcorner\varphi_n(\overline{n})\urcorner$, so that $d(\ulcorner\psi\urcorner) = \ulcorner\psi(\overline{\ulcorner\psi\urcorner})\urcorner$. In fact, a little extra care is required to give a correct definition, because $\varphi_n$ may not be a wff with one free variable. We can get around this problem by simply redefining $d(n) = 0$ when this problem arises (for simplicity, let's assume this is the case in the following), or by letting $d(n) = \ulcorner\forall x(x = \overline{n} \to \varphi_n)\urcorner$, which is defined even when $\varphi_n$ is not a wff. Again, see [1].

One more preliminary remark. From this point on, to promote readability, I will deliberately not make a distinction between $k \in \mathbb{N}$ and the term $\overline{k}$ of $\mathcal{L}_A$ which designates it. The informal definition of the diagonal function above, for example, would be written as $d(n) = \ulcorner\varphi_n(n)\urcorner$, with the overline on the $n$ omitted.

**Fixed-Point Lemma**: If $T$ is a first-order theory which represents the diagonal function, then for any formula $\psi(x)$ with one free variable, there is a sentence $\gamma$ such that $T$ proves $\gamma \leftrightarrow \psi(\ulcorner\gamma\urcorner)$.

**Pf** Let $\delta(x, y)$ represent the diagonal function, so for every $k \in \mathbb{N}$ at which $\ulcorner\varphi_k(k)\urcorner$ is well-defined, $T$ proves the sentence $\forall y(\delta(k, y) \leftrightarrow y = \ulcorner\varphi_k(k)\urcorner)$. Now define $\beta(x)$ by the formula given below.

$$\beta(x) = \exists y(\delta(x, y) \wedge \psi(y))$$

This formula says that the diagonalization of $x$ has property $\psi$. If $\psi(y)$ means that $y$ is being read by Alice, then $\beta(x)$ is precisely the statement *there is a $y$ which is the diagonalization of $x$ and $y$ is being read by Alice.* In other words, *Alice is reading the diagonalization of $x$.*

Now we substitute $\ulcorner\beta(x)\urcorner$ for $x$ in the formula $\beta(x)$, just as we did before when we wrote *Alice is reading the diagonalization of 'Alice is reading the diagonalization of x'.*

$$\begin{aligned}
\beta(\ulcorner\beta(x)\urcorner) &= \exists y(\delta(\ulcorner\beta(x)\urcorner, y) \wedge \psi(y)) \\
&\leftrightarrow_1 \exists y(y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(y)) \\
&\leftrightarrow_2 \exists y(y = \ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(y)) \\
&\leftrightarrow_3 \psi\,(\,\ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner\,)
\end{aligned}$$

Therefore, the formula $\gamma = \beta(\ulcorner\beta(x)\urcorner)$ has the desired property. ■

The three equivalences above which link $\forall y(\delta(\ulcorner\beta(x)\urcorner, y) \to \psi(y))$ and $\psi(\ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner)$ above need to be provable in $T$, so to properly complete the proof, we must show these formally.

Note that $\ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner$ and $\ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner$ are abbreviations for the same term of the language $\mathcal{L}_A$, by definition of the encoding and decoding operations, and hence the equivalence $\leftrightarrow_2$ holds because the two sentences involved are literally the same sentence in the object language. The two other equivalences are shown in the appendix, using natural deduction in the style of [3]. If you're not familiar with this style of writing formal deductions, it should be an easy exercise to rewrite them in another style.

Applying this lemma to the formula $\neg Pr(x)$, where $Pr(x)$ is a provability predicate for $T$, yields a Gödel sentence, and applying it to the negation of Rosser's modified provability predicate yields a Rosser sentence. In this sense, the fixed-point lemma is 'the trick' in these incompleteness results, and 'the work' is in the process of arithmetizing formulas and deductions in $T$.

**Motivation & Acknowledgements**

I've spent a long time thinking about classical incompleteness results, and at some point it occurred to me that I ought to be able to prove the fixed-point lemma on the spot, but I couldn't remember how to come up with the formula $\beta(x)$. I recalled reading [2], which contains the puzzle and solution given on the first page, and I knew that the solution to this puzzle was an informal argument for the fixed-point lemma, so I decided to make the connection explicit by working out the corresponding rigorous argument. This article is the result.

## References

[1] B. Cordy. *Tarksi's Undefinability Theorem.* Available at qubd.github.io, 2015.

[2] R. Smullyan. *Satan, Cantor, and Infinity: Mind-Boggling Puzzles.* Knopf, 1992.

[3] D. Van Dalen. *Logic and Structure, 4th Ed.* Springer, 2008.

**Appendix - Formal Derivations of Equivalences $\leftrightarrow_1$ and $\leftrightarrow_3$**

Both directions of $\leftrightarrow_1$ can be shown by using the same argument. With the abbreviations $D(y)$ for $\delta(\ulcorner\beta(x)\urcorner, y)$, $E(y)$ for $y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner$, and $P(y)$ for $\psi(y)$, we need to show both directions of $\exists y(D(y) \wedge P(y)) \leftrightarrow \exists y(E(y) \wedge P(y))$, and we're free to use the assumptions $\forall y(D(y) \to E(y))$ and $\forall y(D(y) \leftarrow E(y))$ (because $T$ represents the diagonal function). The derivations are given without abbreviations below, though the abbreviations were of great help while constructing them.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\forall y(\delta(\ulcorner\beta(x)\urcorner, y) \to y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner)}{\delta(\ulcorner\beta(x)\urcorner, y) \to y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner}\,\forall E \quad \dfrac{[\delta(\ulcorner\beta(x)\urcorner, y) \wedge \psi(y)]^1}{\delta(\ulcorner\beta(x)\urcorner, y)}\,\wedge E}{y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner}\,\to E \quad \dfrac{[\delta(\ulcorner\beta(x)\urcorner, y) \wedge \psi(y)]^1}{\psi(y)}\,\wedge E}{y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(y)}\,\wedge I}{\exists y(y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(y))}\,\exists I}{\exists y(y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(y))}\,\exists E_1}{\exists y(\delta(\ulcorner\beta(x)\urcorner, y) \wedge \psi(y)) \to \exists y(y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(y))}\,\to I_2$$

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\forall y(y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner \to \delta(\ulcorner\beta(x)\urcorner, y))}{y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner \to \delta(\ulcorner\beta(x)\urcorner, y)}\,\forall E \quad \dfrac{[y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(y)]^1}{y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner}\,\wedge E}{\delta(\ulcorner\beta(x)\urcorner, y)}\,\to E \quad \dfrac{[y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(y)]^1}{\psi(y)}\,\wedge E}{\delta(\ulcorner\beta(x)\urcorner, y) \wedge \psi(y)}\,\wedge I}{\exists y(\delta(\ulcorner\beta(x)\urcorner, y) \wedge \psi(y))}\,\exists I}{\exists y(\delta(\ulcorner\beta(x)\urcorner, y) \wedge \psi(y))}\,\exists E_1}{\exists y(y = \ulcorner\varphi_{\ulcorner\beta(x)\urcorner}(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(y)) \to \exists y(\delta(\ulcorner\beta(x)\urcorner, y) \wedge \psi(y))}\,\to I_2$$

The equivalence $\leftrightarrow_3$ is easier to derive. As above, when constructing or interpreting the derivations it helps to introduce abbreviations for the relevant predicates, and the term $\ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner$ as well.

$$\dfrac{\dfrac{\dfrac{\dfrac{[y = \ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(y)]^1}{y = \ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner}\,\wedge E \quad \dfrac{[y = \ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(y)]^1}{\psi(y)}\,\wedge E}{\psi(\ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner)}\,RI_4}{\psi(\ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner)}\,\exists E_1}{\exists y(y = \ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(y)) \to \psi(\ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner)}\,\to I_2$$

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{}{x = x}\,RI_1}{\ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner = \ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner}\,RI_4 \quad [\psi(\ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner)]^1}{\ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner = \ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(\ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner)}\,\wedge I}{\exists y(y = \ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(y))}\,\exists I}{\psi(\ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner) \to \exists y(y = \ulcorner\beta(\ulcorner\beta(x)\urcorner)\urcorner \wedge \psi(y))}\,\to I_1$$