The Reflection Theorem: A Study in Meta-Theoretic Reasoning

Lawrence C. Paulson

University of Cambridge, Computer Laboratory, JJ Thomson Avenue, Cambridge CB3 0FD, UK, lcp@cl.cam.ac.uk

Abstract. The reflection theorem has been proved using Isabelle/ZF. This theorem cannot be expressed in ZF, and its proof requires reasoning at the meta-level. There is a particularly elegant proof that reduces the meta-level reasoning to a single induction over formulas. Each case of the induction has been proved with Isabelle/ZF, whose built-in tools can prove specific instances of the reflection theorem upon demand.

1 Introduction

A vast amount of mathematics has been verified using proof tools. The Mizar Mathematical Library[1] is probably the largest single repository, but others exist, built using a variety of theorem-provers. An optimist might conclude that any theorem can be verified given enough effort. A sufficiently large and talented team could enter the whole of Wiles's proof of Fermat's Last Theorem [17] and its mathematical prerequisites into a theorem prover, which would duly assert the formula $\forall nxyz\,(n > 2 \to x^n + y^n \neq z^n)$.

The flaw in this point of view is that mathematicians sometimes reason in ways that are hard to formalize. Typical is Gödel's proof [5] of the relative consistency of the axiom of choice (AC). Gödel begins with a complicated set-theoretic construction. At a crucial stage, he introduces operations on syntax. He defines *absoluteness* in terms of the *relativization* $\phi^{\mathbf{M}}$ of a first-order formula ϕ with respect to a class $\mathbf{M}$. He proceeds to apply absoluteness to his entire construction. His proof is of course correct, but it mixes reasoning about sets with reasoning about the language of sets.

Relativization [8, p. 112] replaces each subformula $\exists x\,\phi$ by $\exists x\,(x \in \mathbf{M} \wedge \phi)$ and dually $\forall x\,\phi$ by $\forall x\,(x \in \mathbf{M} \to \phi)$, bounding all quantifiers by $\mathbf{M}$. In Zermelo-Fraenkel (ZF) set theory, a class is simply a formula and $x \in \mathbf{M}$ denotes $\mathbf{M}(x)$. So relativization combines the two formulas, ϕ and $\mathbf{M}$, to yield a third, $\phi^{\mathbf{M}}$. This suggests that we recursively define the set F of first-order formulas within ZF. Relativization for elements of F is trivial to formalize, but it is useless—we cannot relate the "formulas" in F to real formulas. More precisely, no formula χ expresses the truth of elements of F. If for each formula ϕ we write $\ulcorner\phi\urcorner$ for the corresponding element of F, then some formula ψ is handled incorrectly:

[1] Available via http://mizar.org

$\psi \leftrightarrow \neg\chi(\ulcorner\psi\urcorner)$ can be proved in ZF. This fact is Tarski's theorem on the non-definability of truth [8, p. 41].

Gödel introduced meta-level reasoning in order to make his consistency proof effective. He could have worked entirely with sets and demonstrated how to transform a model of set theory into a model of set theory that also satisfied AC. In the latter approach, if we found a contradiction from the axioms of set theory and AC, then we would know that there existed a contradiction in set theory, but we would have no idea how to find the contradiction. Gödel's methods let us transform the contradiction involving AC into a contradiction involving the axioms of set theory alone.

One way to handle meta-level reasoning is to throw away our set theory provers and work formalistically. We could work in a weak logic, such as PRA, which has been proposed for the QED project for mechanizing mathematics [13]. In this logic, we would define the set of formulas (the set F), an internalized inference system, and the ZF axioms. Instead of proving the theorem ϕ in a ZF prover, we would prove the theorem ZF $\vdash \ulcorner\phi\urcorner$ in PRA. Then we could easily express syntactic operations on formulas.

However, the formalist approach is not easy. The formal language of set theory has no function symbols and its only relation symbols are $=$ and $\in$. An assertion such as $\langle x, y\rangle \in A \cup B$ has to be expressed in purely relational form, say $\exists p\, C\, [\mathsf{isPair}(x, y, p) \wedge \mathsf{isUnion}(A, B, C) \wedge p \in C]$. Expressions such as $\{x \in A \mid \phi(x)\}$ and $\bigcup_{x\in A} B(x)$ require a treatment of variable binding. Theorems would be hard even to state, and their proofs would require reasoning about syntax when we would rather reason about sets. My earlier work with Grąbczewski [12] using Isabelle/ZF [10,11] demonstrates that large amounts of set theory can be formalized without taking such an extreme measure. It is worth trying to see what can be accomplished using a set theory prover, recognizing that we can never formalize arguments performed at the meta-level. As it happens, our result can be proved as a collection of separate theorems that Isabelle can use automatically to prove any desired instance of the reflection theorem.

Overview. The paper introduces the reflection theorem (§2) and the proof eventually formalized (§3). Excerpts from the two Isabelle/ZF theories are presented, concerning normal functions (§4) and the reflection theorem (§5). An interactive Isabelle session demonstrates the reflection theorem being applied (§6), and the paper concludes (§7).

2 The Reflection Theorem

The *reflection theorem* is a simple result that illustrates the issues mentioned above. Let **ON** denote the class of ordinals. Suppose that $\{M_\alpha\}_{\alpha\in\mathbf{ON}}$ is a family of sets that is *increasing* (which means $\alpha < \beta$ implies $M_\alpha \subseteq M_\beta$) and *continuous* (which means $M_\alpha = \bigcup_{\xi\in\alpha} M_\xi$ when α is a limit ordinal). Define the class $\mathbf{M}$ by $\mathbf{M} = \bigcup_{\alpha\in\mathbf{ON}} M_\alpha$, Then the reflection theorem states that if $\phi(x_1, \ldots, x_n)$ is a formula in n variables and α is an ordinal, then for some $\beta > \alpha$ and all $x_1, \ldots,$

$x_n \in M_\beta$ we have (intuitively)

$$\mathbf{M} \models \phi(x_1, \ldots, x_n) \;\leftrightarrow\; M_\beta \models \phi(x_1, \ldots, x_n).$$

I say intuitively, because $\mathbf{M}$ could be $\mathbf{V}$, the universal class; as remarked above, truth in ZF is not definable by a formula. A precise statement of the conclusion requires relativization:

$$\phi^{\mathbf{M}}(x_1, \ldots, x_n) \;\leftrightarrow\; \phi^{M_\beta}(x_1, \ldots, x_n).$$

The reflection theorem reduces truth in the class $\mathbf{M}$ to truth in the set M_β, where β can be made arbitrarily large. It is valuable because classes do not exist in ZF; they are merely notation. The theorem can be applied by letting $\mathbf{M}$ be $\mathbf{V}$ and letting M_α be V_α, the cumulative hierarchy defined by $V_0 = 0$, $V_{\alpha+1} = \mathcal{P}(V_\alpha)$ and $V_\alpha = \bigcup_{\xi\in\alpha} V_\xi$ when α is limit. The reflection theorem is also applied to $\mathbf{L}$, the constructible universe [8, p. 169]; it is an essential part of modern treatments of Gödel's consistency proof that are based on ZF set theory.

Proving the reflection theorem is not difficult, if only we can formalize it. Bancerek [1] proved it in Mizar. Mizar's native Tarski-Grothendieck properly extends ZF: classes really do exist, and we can define $\mathbf{M} \models \ulcorner\phi(x_1, \ldots, x_n)\urcorner$ when $\mathbf{M}$ is a class. This solves the problem concerning the definability of truth. It is ironic that the formalization problems can be solved by working either in the weaker logic PRA or in a stronger logic.

The approach taken below is more in the spirit of set theory: a *theorem* follows from the axioms, while a *meta-theorem* is a mechanical procedure for yielding theorems. Most authors do not formalize the meta-theory. Results such as the following are not meta-theorems, but merely *theorem schemes*:

$$a \in \{x \in A \mid \phi(x)\} \;\leftrightarrow\; a \in A \wedge \phi(a)$$
$$a \in \bigcup_{x\in A} B(x) \;\leftrightarrow\; \exists x\, (x \in A \wedge a \in B(x))$$

Their proofs depend not at all on the structure of the formula ϕ or the expression B. Thanks to Isabelle's higher-order syntax, each is a single Isabelle/ZF theorem, with a trivial proof. The reflection theorem is different: it is proved by reasoning about a formula's structure.

3 Proof Overview

The first task in formalizing the reflection theorem is to find a proof with the least amount of meta-level reasoning. Kunen's proof [8, p. 136] needs a lemma, also proved at the meta-level, about a subformula-closed list of formulas. The proof idea is related to Skolemization and involves finding all existentially quantified subformulas. Drake's proof [4, p. 99] requires the formula to be presented in prenex form and involves a simultaneous construction for the whole quantifier string. In both proofs, the meta-level component is substantial.

Mostowski's proof [9, p. 23], fortunately, is a simple structural induction. Reflection for atomic formulas is trivial. Reflection for $\neg\phi(x)$ and $\phi(x)\wedge\phi'(x)$ follows trivially from induction hypotheses for $\phi(x)$ and $\phi'(x)$. Reflection for $\exists y\,\phi(x,y)$ follows from an induction hypothesis for $\phi(x,y)$. The main complication is that the case for $\exists y\,\phi(x,y)$ adds a variable to the induction hypothesis; we do not want the theorem statement to depend upon the number of free variables in the formula. By assuming that the class $\mathbf{M}$ is closed (in a suitable way) under ordered pairing, it suffices to derive reflection for $\exists y\,\phi(\langle x,y\rangle)$ from reflection for $\phi(\langle x,y\rangle)$, which trivially follows from reflection for $\phi(z)$. The proofs are non-trivial, but they take place entirely within ZF. The only meta-level reasoning is the structural induction itself: noting that it suffices to prove the cases for atomic formulas, $\neg$, $\wedge$ and $\exists$. The simple structure of these lemmas makes it easy to apply reflection to individual formulas and yields an expression for the class of ordinals that reflect the formula. At the end of this paper, we shall see Isabelle doing this automatically.

Mostowski's proof owes its simplicity to the classic technique of strengthening the induction hypothesis. The required conclusion has the form $\forall\alpha\,\exists\beta>\alpha\ldots$; in other words, the possible values of β form an unbounded class. In Mostowski's proof, this class is closed as well as unbounded. A class $\mathbf{X}$ of ordinals is *closed* provided for every nonempty set Y, if $Y\subseteq\mathbf{X}$ then $\bigcup Y\in\mathbf{X}$. (The union $\bigcup Y$ is the supremum, or limit, of the set Y.) It turns out that if $\mathbf{X}$ and $\mathbf{X}'$ are closed and unbounded, then so is $\mathbf{X}\cap\mathbf{X}'$. This fact is crucial; in particular, it gives an immediate proof for the conjunctive case of the reflection theorem: if $\mathbf{X}$ is the class of ordinals for $\phi(x)$ and $\mathbf{X}'$ is the class of ordinals for $\phi'(x)$ then $\mathbf{X}\cap\mathbf{X}'$ is a closed, unbounded class of ordinals for $\phi(x)\wedge\phi'(x)$.

The function $F:\mathbf{ON}\to\mathbf{ON}$ is *normal* provided it is increasing and continuous:

$$\begin{aligned} F(\alpha) &< F(\beta) \quad \text{if } \alpha<\beta \\ F(\alpha) &= \bigcup_{\xi<\alpha} F(\xi) \quad \text{if } \alpha \text{ is a limit ordinal} \end{aligned}$$

Every normal function enjoys a key property: the class of fixedpoints $\{\alpha \mid F(\alpha)=\alpha\}$ is closed and unbounded. This fact has surprising consequences. Consider the enumeration of the cardinals, $\{\aleph_\alpha\}_{\alpha\in\mathbf{ON}}$. Given that even $\aleph_0$ is infinite, one might expect $\alpha<\aleph_\alpha$ to be a trivial theorem, but in fact $\aleph$ is a normal function and the solutions of $\aleph_\alpha=\alpha$ form a closed and unbounded class.

Normal functions are used in the critical case of the reflection theorem, when we have an existential quantifier. Here is a sketch of the argument. At a key stage in the proof, we seek an ordinal β such that for all $x\in M_\beta$ we have

$$\exists y\in\mathbf{M}\;\phi(x,y)\to\exists y\in M_\beta\;\phi(x,y). \tag{1}$$

Let α be an ordinal. If $x\in M_\alpha$ and $y\in\mathbf{M}$ then (since $\mathbf{M}=\bigcup_{\alpha\in\mathbf{ON}} M_\alpha$) we can choose the least $\xi(x)$ such that $y'\in M_\xi$ and $\phi(x,y')$. This ordinal is a function of x, and we can apply the replacement axiom over the set M_α to find

the least upper bound of the set $\{\xi(x)\}_{x\in M_\alpha}$. This map from α to $\bigcup_{x\in M_\alpha}\xi(x)$ can be used to define a normal function, F. Let β be a fixedpoint of F. Then, by construction, if $x \in M_\beta$ and $y \in \mathbf{M}$ then there exists $y' \in M_{F(\beta)}$ such that $\phi(x, y')$. Since $F(\beta) = \beta$ we conclude $y' \in M_\beta$, which establishes (1).

Two points remain before we can proceed to the Isabelle/ZF proofs. First, recall that we can restrict attention to unary formulas, deriving reflection for $\exists y\, \phi(\langle x, y\rangle)$ rather than for $\exists y\, \phi(x, y)$. This requires assuming that the class $\mathbf{M}$ is closed under ordered pairing *in a suitable way*. The natural way is to assume that M_α is closed under ordered pairing whenever α is a limit ordinal. The class of limit ordinals is closed and unbounded, so we can intersect this class with the class found by the proof sketched above and the resulting class will still be closed and unbounded.

For the second point, recall that the conclusion of the reflection theorem is

$$\phi^{\mathbf{M}}(z) \leftrightarrow \phi^{M_\beta}(z).$$

Working with real formulas makes it impossible to formalize the relativizations $\phi^{\mathbf{M}}$ and ϕ^{M_β}. It turns out that we can abstract $\phi^{\mathbf{M}}(z)$ to $\phi(z)$ and $\phi^{M_\beta}(z)$ to $\psi(\beta, z)$ in the crucial case of the existential quantifier, proving

$$\exists y \in \mathbf{M}\ \phi(\langle x, y\rangle) \leftrightarrow \exists y \in M_\beta\ \psi(\beta, \langle x, y\rangle).$$

For the induction hypothesis, we merely need a closed unbounded class of ordinals α such that $\phi(x) \leftrightarrow \psi(\alpha, x)$ for $x \in M_\alpha$. The proof does not require $\psi(\alpha, x)$ to behave like $\phi^{M_\alpha}(x)$. The resulting theorems inductively generate (at the meta-level!) pairs of formulas of the form $\phi^{\mathbf{M}}$ and ϕ^{M_β}.

4 Normal Functions in Isabelle/ZF

Two Isabelle/ZF theories define the concepts covered in this paper: one for normal functions and closed and unbounded classes, the other for the reflection theorem itself. The files (available from the author) prove about 90 lemmas and theorems using about 210 proof commands. The next two sections present highlights, omitting most proofs and many technical lemmas. The formal material presented below was generated automatically from the Isabelle theories. It is similar to what the user sees on the screen when using Proof General.[2] These proofs were not written in ML, as in traditional Isabelle, but as tactic scripts in the Isar language [16].

Iteration of the function `F`, written `iterates(F,x,n)`, corresponds to $F^n(x)$. The following concept is the limit of all such iterations, corresponding to $F^\omega(x)$.

constdefs

```
iterates_omega :: "[i⇒i,i] ⇒ i"
  "iterates_omega(F,x) ≡ ⋃n∈nat. iterates(F,x,n)"
```

The ordinal ω is written nat in Isabelle/ZF because it is the set of natural numbers.

[2] http://www.proofgeneral.org/

4.1 Closed and Unbounded Classes of Ordinals

Classes have no special status in Isabelle/ZF. Although Isabelle's overloading mechanism [15] makes it possible to extend operations such as $\in$, $\cup$, $\cap$ and $\subseteq$ to classes, the theories adopt the traditional approach. A class $\mathbf{M}$ is really a formula ϕ. Membership in a class, $a \in \mathbf{M}$, means $\phi(a)$. Intersection of two classes, $\mathbf{M} \cap \mathbf{N}$, denotes the conjunction of the predicates, $\lambda x.\, \phi(x) \wedge \psi(x)$. A family of classes, $\{\mathbf{M}_z\}_{z\in\mathbf{N}}$, denotes a 2-argument predicate; for example, $a \in \bigcup_{z\in\mathbf{N}} \mathbf{M}_z$ stands for $\exists z\, \psi(z) \wedge \phi(z, a)$. These examples illustrate the extent to which we can reason about classes in ZF.

The theory defines closed and unbounded (c.u.) classes of ordinals. A class has type `i⇒o`, which is the type of functions from sets to truth values.

```
constdefs
  Closed :: "(i⇒o) ⇒ o"
    "Closed(P) ≡ ∀I. I≠0⟶ (∀i∈I. Ord(i) ∧ P(i)) ⟶ P(⋃(I))"

  Unbounded :: "(i⇒o) ⇒ o"
    "Unbounded(P) ≡ ∀i. Ord(i) ⟶ (∃j. i<j ∧ P(j))"

  Closed_Unbounded :: "(i⇒o) ⇒ o"
    "Closed_Unbounded(P) ≡ Closed(P) ∧ Unbounded(P)"
```

The predicate `Ord` recognizes the class of ordinal numbers, which is traditionally written **ON**, while `Limit` recognizes the limit ordinals. The predicate `Card` recognizes the class of cardinals, which is traditionally written **CARD**. All three classes are easily proved to be closed and unbounded.

```
theorem Closed_Unbounded_Ord   [simp]: "Closed_Unbounded(Ord)"
theorem Closed_Unbounded_Limit [simp]: "Closed_Unbounded(Limit)"
theorem Closed_Unbounded_Card  [simp]: "Closed_Unbounded(Card)"
```

4.2 The Intersection of a Family of Closed Unbounded Classes

A key lemma for the reflection theorem is that the intersection of a family of closed unbounded (c.u.) classes is c.u. (The family must be indexed by a set, not a class, for $\bigcap_{\alpha\in\mathbf{ON}} \{\beta \mid \beta > \alpha\}$ is empty.) The constructions below come from Kunen [8, p. 78].

A *locale* [7] lets us fix the class `P` and the index set `A`. It states assumptions that hold for the whole development, namely that `P` is closed and unbounded and that `A` is nonempty. It also contains definitions of functions `next_greater` and `sup_greater`, which are local to the proof.

```
locale cub_family =
  fixes P and A
  fixes next_greater — the next ordinal satisfying class A
  fixes sup_greater  — sup of those ordinals over all A
```

assumes *closed:* "a∈A ⟹ Closed(P(a))"
and *unbounded:* "a∈A ⟹ Unbounded(P(a))"
and *A_non0:* "A≠0"
defines "next_greater(a,x) ≡ μy. x<y ∧ P(a,y)"
and "sup_greater(x) ≡ ⋃a∈A. next_greater(a,x)"

Our result is the culmination of a series of lemmas proved in the scope of this locale. We begin by proving that the intersection is closed.

lemma (**in** *cub_family*) *Closed_INT:* "Closed(λx. ∀i∈A. P(i,x))"

The proof (omitted) is a one-liner. The difficulty is showing that the intersection is unbounded. For a∈A, by the unboundedness of P(a), it contains an ordinal *next_greater*(a,x) greater than x. By reasoning about the μ-operator, which denotes the least ordinal satisfying a formula, these claims are easily verified:

"⟦Ord(x); a∈A⟧ ⟹ P(a, next_greater(a,x))"
"⟦Ord(x); a∈A⟧ ⟹ x < next_greater(a,x)"

I have omitted the **lemma** commands, for brevity.

Now *sup_greater*(x) is the supremum of *next_greater*(a,x) for a∈A. We can iterate this step to reach $\mathit{sup\_greater}^{\omega}(x)$. The point is that $\mathit{sup\_greater}^{\omega}(x)$ belongs to all of the classes, and thus to the intersection. First, a number of trivial facts have to be verified, such as these:

"Ord(x) ⟹ x < iterates_omega(sup_greater,x)"
"a∈A ⟹ next_greater(a,x) ≤ sup_greater(x)"

This is a key stage in the argument. Fixing a∈A, we find that $\mathit{sup\_greater}^{\omega}(x)$ can be written as the supremum of values of the form *next_greater*(a,−), that is, as the supremum of members of P(a).

"⟦Ord(x); a∈A⟧
⟹ iterates_omega(sup_greater,x) =
(⋃n∈nat. next_greater(a, iterates(sup_greater,x,n)))"

Since this class is closed, it must contain $\mathit{sup\_greater}^{\omega}(x)$.

"⟦Ord(x); a∈A⟧ ⟹ P(a, iterates_omega(sup_greater,x))"

The desired result follows immediately. Note that the intersection of a family of classes is expressed as a universally-quantified formula:

theorem *Closed_Unbounded_INT:*
"(⋀a. a∈A ⟹ Closed_Unbounded(P(a)))
⟹ Closed_Unbounded(λx. ∀a∈A. P(a, x))"

Since $2 = \{0, 1\}$ in set theory, the intersection of two classes can be reduced to the general case by using *2* for the index set:

"P(x) ∧ Q(x) ⟷ (∀i∈2. (i=0 ⟶ P(x)) ∧ (i=1 ⟶ Q(x)))"

Thus we obtain the corollary for binary intersections, which is the version used in the reflection theorem:

theorem *Closed_Unbounded_Int:*
"⟦Closed_Unbounded(P); Closed_Unbounded(Q)⟧
⟹ Closed_Unbounded(λx. P(x) ∧ Q(x))"

4.3 Fixedpoints of Normal Functions

Our proof of the reflection theorem uses the lemma that the class of fixedpoints of a normal function is closed and unbounded. The Isabelle/ZF proof follows Drake [4, pp. 113–114]. It begins by defining normal functions as those that are monotonic and continuous over the ordinals:

constdefs

```
mono_Ord :: "(i⇒i) ⇒ o"
  "mono_Ord(F) ≡ ∀i j. i<j ⟶ F(i) < F(j)"

cont_Ord :: "(i⇒i) ⇒ o"
  "cont_Ord(F) ≡ ∀l. Limit(l) ⟶ F(l) = (⋃i<l. F(i))"

Normal :: "(i⇒i) ⇒ o"
  "Normal(F) ≡ mono_Ord(F) ∧ cont_Ord(F)"
```

Among the consequences of these definitions is an equation expressing continuity of normal functions over unions. It follows (with a little effort) from their continuity over limit ordinals.

```
"⟦X≠0; ∀x∈X. Ord(x); Normal(F)⟧
 ⟹ F(Union(X)) = (⋃y∈X. F(y))"
```

From this lemma, it is easy to prove that the class of fixedpoints is closed:

```
"Closed(λi. F(i) = i)"
```

As with the intersection theorem, the work goes into showing that the class is unbounded, by reasoning about suprema. If F is a normal function, then consider $F^\omega(\alpha) = \bigcup_{\alpha\in\omega} F^n(\alpha)$. It is easy to show that $F^\omega(\alpha)$ is a fixedpoint of F.

```
"⟦Normal(F); Ord(a)⟧
 ⟹ F(iterates_omega(F,a)) = iterates_omega(F,a)"
```

Since $\alpha \leq F^\omega(\alpha)$, there are arbitrarily large fixedpoints, which yields the desired result.

theorem *Normal_imp_fp_Closed_Unbounded:*

```
"Normal(F) ⟹ Closed_Unbounded(λi. F(i) = i)"
```

4.4 Function *normalize*

The key construction of the reflection theorem maps an ordinal α to another ordinal, $F(\alpha)$, but F might not be monotonic, so it is not necessarily normal. The usual proof complicates the construction in order to force F to be monotonic. However, we can define a separate operator for this purpose.

Function *normalize* maps a continuous function $F : \mathbf{ON} \to \mathbf{ON}$ to a normal function F' that bounds it above. Continuity of F is needed to show that $F(\alpha) \leq F'(\alpha)$. For a counterexample, consider the successor function $S : \mathbf{ON} \to \mathbf{ON}$, which is not continuous. If S' is normal, then let α be one of its fixedpoints. Then $S'(\alpha) = \alpha < S(\alpha)$.

constdefs
```
  normalize :: "[i⇒i, i] ⇒ i"
    "normalize(F,a) ≡ ...
```

The definition is omitted because it is too technical. It defines `normalize(F,a)` to be the function $F'(\alpha)$ satisfying the transfinite recursion

$$\begin{aligned} F'(0) &= F(0) \\ F'(\alpha+1) &= \max\{F'(\alpha)+1,\, F(\alpha+1)\} \\ F'(\alpha) &= \bigcup_{\xi<\alpha} F'(\xi) \qquad\qquad \text{if } \alpha \text{ is a limit ordinal} \end{aligned}$$

Monotonicity follows directly, since by the definition $F'(\alpha+1) > F'(\alpha)$. The essential properties of `normalize` are easily shown:

theorem *Normal_normalize:*
```
    "(⋀x. Ord(x) ⟹ Ord(F(x))) ⟹ Normal(normalize(F))"
```

theorem *le_normalize:*
```
    "⟦Ord(i); cont_Ord(F); ⋀x. Ord(x) ⟹ Ord(F(x))⟧
     ⟹ F(i) ≤ normalize(F,i)"
```

5 The Reflection Theorem in Isabelle/ZF

Recall that the reflection theorem concerns a class $\mathbf{M} = \bigcup_{\alpha\in\mathbf{ON}} M_\alpha$, where the $\{M_\alpha\}_{\alpha\in\mathbf{ON}}$ are an increasing and continuous family of sets indexed by the ordinals. The constant `mono_le_subset` expresses the notion of *increasing*:

constdefs
```
  mono_le_subset :: "(i⇒i) ⇒ o"
    "mono_le_subset(M) ≡ ∀i j. i≤j ⟶ M(i) ⊆ M(j)"
```

A locale (Fig. 1) fixes `Mset`, which stands for the family $\{M_\alpha\}_{\alpha\in\mathbf{ON}}$. It states the assumptions that `Mset` is increasing, continuous and (at limit stages) closed under ordered pairing. Its definition of the class `M` uses an existential quantifier to express $\bigcup_{\alpha\in\mathbf{ON}} M_\alpha$. It defines reflection as the ternary relation `Reflects(Cl,P,Q)` joining a closed, unbounded class `Cl` with a predicate `P` (supposed to be relativized to $\mathbf{M}$) and a predicate `Q` (supposed to be relativized to M_α).

The locale also defines some items that are needed only to prove the existential case. The ordinal-valued functions mentioned in §3 are formalized as `F0` and `FF`. The class `ClEx(P)` consists of all limit ordinals that are fixedpoints of the normal function `normalize(FF(P))`. This class will be closed and unbounded because it is the intersection of two other c.u. classes, and the restriction to limit ordinals lets us use the assumption that `Mset` is closed under pairing at limit stages.

Now we find ourselves reasoning at the meta-level. Formulas have not been defined within set theory; rather they are part of the language of set theory. Therefore, the induction on the structure of formulas cannot be formalized. We simply state and prove the separate cases of this induction.

```
locale reflection =
  fixes Mset and M and Reflects
  assumes Mset_mono_le : "mono_le_subset(Mset)"
     and Mset_cont    : "cont_Ord(Mset)"
     and Pair_in_Mset : "⟦x ∈ Mset(a); y ∈ Mset(a); Limit(a)⟧
                         ⟹ <x,y> ∈ Mset(a)"
  defines "M(x) ≡ ∃a. Ord(a) ∧ x ∈ Mset(a)"
    and "Reflects(Cl,P,Q) ≡
              Closed_Unbounded(Cl) ∧
              (∀a. Cl(a) ⟶ (∀x∈Mset(a). P(x) ⟷ Q(a,x)))"
  fixes F0 — ordinal for a specific value y
  fixes FF — sup over the whole level, y ∈ Mset(a)
  fixes ClEx — Reflecting ordinals for the formula ∃z. P
  defines "F0(P,y) ≡ μb. (∃z. M(z) ∧ P(<y,z>)) ⟶
                         (∃z∈Mset(b). P(<y,z>))"
    and "FF(P)   ≡ λa. ⋃y∈Mset(a). F0(P,y)"
    and "ClEx(P) ≡ λa. Limit(a) ∧ normalize(FF(P),a) = a"
```

Fig. 1. The locale *reflection*

5.1 Proving Easy Cases of the Reflection Theorem

The base case is when the two formulas are identical, which in practice means that they contain no quantifiers. All ordinals belong to the reflecting class. The proof, **by** *(simp ...)*, is shown to emphasize that the proof is immediate by definition.

```
theorem (in reflection) Triv_reflection [intro]:
     "Reflects(Ord, λx. P(x), λa x. P(x))"
by (simp add: Reflects_def)
```

The reflecting class for a negation equals that for its operand. This proof is also trivial.

```
theorem (in reflection) Not_reflection [intro]:
     "Reflects(Cl,P,Q) ⟹ Reflects(Cl, λx. ~P(x), λa x. ~Q(a,x))"
by (simp add: Reflects_def)
```

The reflecting class for a conjunction is the intersection of those for the two conjuncts. This proof uses *Closed_Unbounded_Int*, our lemma that the intersection of two c.u. classes is c.u. Not shown are the theorems for $\vee$, $\rightarrow$ and $\leftrightarrow$, whose proofs are equally trivial.

```
theorem (in reflection) And_reflection [intro]:
     "⟦Reflects(Cl,P,Q); Reflects(C',P',Q')⟧
      ⟹ Reflects(λa. Cl(a) ∧ C'(a), λx. P(x) ∧ P'(x),
                                     λa x. Q(a,x) ∧ Q'(a,x))"
by (simp add: Reflects_def Closed_Unbounded_Int, blast)
```

The attribute *[intro]*, which appears in each of the theorems above, labels them as *introduction rules* for Isabelle's classical reasoner. This will enable Isabelle to perform reflection automatically.

5.2 Reflection for Existential Quantifiers

This is the most important part of the development. A key lemma is that the function *FO* works as it should: if *y ∈ Mset(a)* then *FO(P,y)* is a large enough ordinal for *∃z∈Mset(FO(P,y)). P(<y,z>)* to hold. The proof is four lines long, using simple reasoning about the μ-operator.

```
"⟦y∈Mset(a); Ord(a); M(z); P(<y,z>)⟧
 ⟹ ∃z∈Mset(FO(P,y)). P(<y,z>)"
```

Similarly, the function *FF* works as it should: if *a* is an ordinal then *FF(P,a)* is large enough for the desired conclusion to hold.

```
"⟦M(z); y∈Mset(a); P(<y,z>); Ord(a)⟧
 ⟹ ∃z∈Mset(FF(P,a)). P(<y,z>)"
```

Similarly again, the normal function derived from *FF* returns an ordinal large enough for the conclusion to hold.

```
"⟦M(z); y∈Mset(a); P(<y,z>); Ord(a)⟧
 ⟹ ∃z∈Mset(normalize(FF(P),a)). P(<y,z>)"
```

To complete the proof, a further locale declares the induction hypothesis. More precisely, it declares half of it: namely that *Cl* consists of ordinals that correctly relate *P* and *Q*. At this point, there is no need to assume that *Cl* is closed and unbounded.

```
locale ex_reflection = reflection +
  fixes P  — the original formula
  fixes Q  — the reflected formula
  fixes Cl — the class of reflecting ordinals
  assumes Cl_reflects:
          "⟦Cl(a); Ord(a)⟧ ⟹ ∀x∈Mset(a). P(x) ⟷ Q(a,x)"
```

Now we can reap the benefits of the previous work, such as the lemmas about *FF*. Translated into mathematical language, the next result states that if $z \in \mathbf{M}$ and $y \in M_\alpha$, where α is an ordinal belonging to the class we have constructed, and $P(\langle y, z\rangle)$ holds, then $Q_\alpha(\langle y, z\rangle)$ holds for some $z \in M_\alpha$.

```
"⟦M(z); y∈Mset(a); P(<y,z>); Cl(a); ClEx(P,a)⟧
 ⟹ ∃z∈Mset(a). Q(a,<y,z>)"
```

This lemma is the opposite and easy direction, for if $z \in M_\alpha$ then obviously $z \in \mathbf{M}$.

```
"⟦z∈Mset(a); y∈Mset(a); Q(a,<y,z>); Cl(a); ClEx(P,a)⟧
 ⟹ ∃z. M(z) ∧ P(<y,z>)"
```

Combining these results, we find that *ClEx* indeed expresses closed and unbounded classes of ordinals for reflection:

```
"Closed_Unbounded(ClEx(P))"

"⟦y∈Mset(a); Cl(a); ClEx(P,a)⟧
 ⟹ (∃z. M(z) ∧ P(<y,z>)) ⟷ (∃z∈Mset(a). Q(a,<y,z>))"
```

It only remains to package up the existential case using the *Reflects* symbol:

```
"Reflects(Cl,P0,Q0)
 ⟹ Reflects(λa. Cl(a) ∧ ClEx(P0,a),
             λx. ∃z. M(z) ∧ P0(<x,z>),
             λa x. ∃z∈Mset(a). Q0(a,<x,z>))"
```

The previous version applies only to formulas that involve ordered pairs. To correct that problem, we can use the projection functions *fst* and *snd*, which return the first and second components of a pair:

```
theorem (in reflection) Ex_reflection [intro]:
     "Reflects(Cl, λx. P(fst(x),snd(x)),
                   λa x. Q(a,fst(x),snd(x)))
      ⟹ Reflects(λa. Cl(a) ∧ ClEx(λx. P(fst(x),snd(x)), a),
                   λx. ∃z. M(z) ∧ P(x,z),
                   λa x. ∃z∈Mset(a). Q(a,x,z))"
```

The dual rule for the universal quantifier is trivial, since $\forall x\, \phi(x)$ is $\neg\exists x\, \neg\phi(x)$.

6 Invoking the Reflection Theorem

We have no formal statement of the reflection theorem in Isabelle/ZF. However, we have a mechanical procedure for applying it in specific cases, which is one interpretation of a meta-theorem. That procedure is simply Isabelle's classical reasoner, *fast*. No modifications are necessary. Declaring each case of the reflection theorem with the *[intro]* attribute flags them as introduction rules, suitable for backward chaining. The many λ-bound variables in these rules pose no problems for *fast*: it searches for proofs using Isabelle's inbuilt inference mechanisms, which employ higher-order unification [6].

Here the reflection theorem is applied to $\phi(x) \equiv \exists y \forall z\, (z \subseteq x \to z \in y)$. I have explicitly written the relavitized formulas, namely $\phi^{\mathbf{M}}$ and ϕ^{M_α}, though this can be automated using ML if necessary. We have no idea what the reflecting class will be, but we can write it as the variable *?Cl* and let Isabelle work it out.

```
lemma (in reflection)
     "Reflects(?Cl,
               λx. ∃y. M(y) ∧ (∀z. M(z) ⟶ z ⊆ x ⟶ z ∈ y),
               λa x. ∃y∈Mset(a). ∀z∈Mset(a). z ⊆ x ⟶ z ∈ y)"
by fast
```

Here, reflection is applied to a more complicated formula. Despite the three quantifiers, the call to *fast* takes only 90 milliseconds.

```
lemma (in reflection)
  "Reflects(?Cl,
      λx. ∃y. M(y) ∧ (∀z. M(z) ⟶
                          (∀w. M(w) ⟶ w∈z ⟶ w∈x) ⟶ z∈y),
      λa x. ∃y∈Mset(a). ∀z∈Mset(a).
                          (∀w∈Mset(a). w∈z ⟶ w∈x) ⟶ z∈y)"
by fast
```

Conducting a single-step proof shows how easy these theorems are to prove and also illustrates how the reflecting class is determined incrementally. For this, let us return to the first example:

```
lemma (in reflection)
     "Reflects(?Cl,
               λx. ∃y. M(y) ∧ (∀z. M(z) ⟶ z ⊆ x ⟶ z ∈ y),
               λa x. ∃y∈Mset(a). ∀z∈Mset(a). z ⊆ x ⟶ z ∈ y)"
```

The outermost connective is $\exists$, so we apply the corresponding instance of the reflection theorem. Observe how the variable *?Cl* in the main goal is replaced by an expression involving an invocation of *ClEx* and a new variable, *?Cl1*. This variable must be replaced by some class *?Cl1* that reflects the remaining subformulas:

```
apply (rule Ex_reflection)
Reflects
 (λa. ?Cl1(a) ∧
      ClEx(λx. ∀z. M(z) ⟶ z ⊆ fst(x) ⟶ z ∈ snd(x), a),
  λx. ∃y. M(y) ∧ (∀z. M(z) ⟶ z ⊆ x ⟶ z ∈ y),
  λa x. ∃y∈Mset(a). ∀z∈Mset(a). z ⊆ x ⟶ z ∈ y)
 1. Reflects
     (?Cl1, λx. ∀z. M(z) ⟶ z ⊆ fst(x) ⟶ z ∈ snd(x),
      λa x. ∀z∈Mset(a). z ⊆ fst(x) ⟶ z ∈ snd(x))
```

Now the outermost connective is $\forall$, so we apply *All_reflection*. The variable *?Cl1* is in its turn replaced by another invocation of *ClEx* and another new variable, *?Cl2*:

```
apply (rule All_reflection)
Reflects
 (λa. (?Cl2(a) ∧
       ClEx(λx. ¬ (snd(x) ⊆ fst(fst(x)) ⟶
                   snd(x) ∈ snd(fst(x))),
            a)) ∧
      ClEx(λx. ∀z. M(z) ⟶ z ⊆ fst(x) ⟶ z ∈ snd(x), a),
  λx. ∃y. M(y) ∧ (∀z. M(z) ⟶ z ⊆ x ⟶ z ∈ y),
  λa x. ∃y∈Mset(a). ∀z∈Mset(a). z ⊆ x ⟶ z ∈ y)
 1. Reflects
     (?Cl2,
      λx. snd(x) ⊆ fst(fst(x)) ⟶ snd(x) ∈ snd(fst(x)),
      λa x. snd(x) ⊆ fst(fst(x)) ⟶ snd(x) ∈ snd(fst(x)))
```

The two formulas are now identical, so *Triv_reflection* completes the proof. It replaces *?C12* by *Ord*, the class of all ordinals.

```
apply (rule Triv_reflection)
Reflects
 (λa. (Ord(a) ∧
       ClEx(λx. ¬ (snd(x) ⊆ fst(fst(x)) ⟶
                   snd(x) ∈ snd(fst(x))),
            a)) ∧
      ClEx(λx. ∀z. M(z) ⟶ z ⊆ fst(x) ⟶ z ∈ snd(x), a),
  λx. ∃y. M(y) ∧ (∀z. M(z) ⟶ z ⊆ x ⟶ z ∈ y),
  λa x. ∃y∈Mset(a). ∀z∈Mset(a). z ⊆ x ⟶ z ∈ y)
No subgoals!
```

We should not use defined predicates such as *Ord* in the formula being reflected. The resulting theorems, although valid, would not be instances of the reflection theorem: *Ord* is itself defined in terms of quantifiers, which need to be relativized. I have defined relativized versions of many set-theoretic concepts, such as order-isomorphism, and proved their equivalence to the originals. These relativized concepts form a vocabulary for specific invocations of the reflection theorem.

7 Conclusions

Gödel worked in von Neumann-Bernays-Gödel (NBG) set theory. Modern versions of his proof are typically expressed in ZF set theory. Either way, the base set theory is proved to be relatively consistent with AC. Bancerek [1] proved the reflection theorem years ago, in Mizar, also following Mostowski [9]. However, Bancerek's proof does not address the issue of meta-level reasoning. It instead uses Tarski-Grothendieck set theory to reason about ZF, a weaker system. It does not suggest how to prove the consistency of the axiom of choice with respect to Tarski-Grothendieck set theory, which is a natural question for users of that theory.

My aim is not simply to mechanize the reflection theorem but to capture the spirit of Gödel's consistency proof. I have not given a general way of eliminating meta-reasoning, but I have shown how to treat one specific case. A number of researchers [2,14] have done mechanical proofs using NBG set theory. Gödel's original proof [5] does not require the reflection theorem, but perhaps other parts of his proof could be mechanized in NBG.

Acknowledgements. Alexander S. Kechris and Ken Kunen gave valuable advice by electronic mail. I learned that there was a suitable proof of the reflection theorem in Kechris's unpublished 1976 lecture notes. Markus Wenzel added his Isar language and proof presentation tools to Isabelle/ZF.

References

1. Grzegorz Bancerek. The reflection theorem. *Journal of Formalized Mathematics*, 2, 1990. http://megrez.mizar.org/mirror/JFM/Vol2/zf_refle.html.
2. Johan G. F. Belinfante. Computer proofs in Gödel's class theory with equational definitions for composite and cross. *Journal of Automated Reasoning*, 22(3):311–339, March 1999.
3. Yves Bertot, Gilles Dowek, André Hirschowitz, Christine Paulin, and Laurent Théry, editors. *Theorem Proving in Higher Order Logics: TPHOLs '99*, LNCS 1690. Springer, 1999.
4. Frank R. Drake. *Set Theory: An Introduction to Large Cardinals*. North-Holland, 1974.
5. Kurt Gödel. The consistency of the axiom of choice and of the generalized continuum hypothesis with the axioms of set theory. In S. Feferman et al., editors, *Kurt Gödel: Collected Works*, volume II. Oxford University Press, 1990. First published in 1940.
6. G. P. Huet. A unification algorithm for typed λ-calculus. *Theoretical Computer Science*, 1:27–57, 1975.
7. Florian Kammüller, Markus Wenzel, and Lawrence C. Paulson. Locales: A sectioning concept for Isabelle. In Bertot et al. [3], pages 149–165.
8. Kenneth Kunen. *Set Theory: An Introduction to Independence Proofs*. North-Holland, 1980.
9. Andrzej Mostowski. *Constructible Sets with Applications*. North-Holland, 1969.
10. Lawrence C. Paulson. Set theory for verification: I. From foundations to functions. *Journal of Automated Reasoning*, 11(3):353–389, 1993.
11. Lawrence C. Paulson. Set theory for verification: II. Induction and recursion. *Journal of Automated Reasoning*, 15(2):167–215, 1995.
12. Lawrence C. Paulson and Krzysztof Grąbczewski. Mechanizing set theory: Cardinal arithmetic and the axiom of choice. *Journal of Automated Reasoning*, 17(3):291–323, December 1996.
13. The QED manifesto. http://www-unix.mcs.anl.gov/qed/, 1995.
14. Art Quaife. Automated deduction in von Neumann-Bernays-Gödel set theory. *Journal of Automated Reasoning*, 8(1):91–147, 1992.
15. Markus Wenzel. Type classes and overloading in higher-order logic. In Elsa L. Gunter and Amy Felty, editors, *Theorem Proving in Higher Order Logics: TPHOLs '97*, LNCS 1275, pages 307–322. Springer, 1997.
16. Markus Wenzel. Isar: A generic interpretative approach to readable formal proof documents. In Bertot et al. [3], pages 167–183.
17. Andrew J. Wiles. Modular elliptic curves and Fermat's Last Theorem. *Annals of Mathematics*, 141(3):443–551, 1995.