THE AXIOM OF CHOICE, ZORN'S LEMMA, AND THE WELL ORDERING PRINCIPLE

The *Axiom of Choice* is a foundational statement of set theory:

> *Given any collection $\{S_i : i \in I\}$ of nonempty sets, there exists a choice function*
> $$f : I \longrightarrow \bigcup_i S_i, \qquad f(i) \in S_i \textit{ for all } i \in I.$$

In the 1930's, Kurt Gödel proved that the Axiom of Choice is consistent (in the Zermelo-Frankel first-order axiomatization) with the other axioms of set theory. In the 1960's, Paul Cohen proved that the Axiom of Choice is independent of the other axioms.

Closely following Van der Waerden, this writeup explains how the Axiom of Choice implies two other statements, *Zorn's Lemma* and the *Well Ordering Principle.* In fact, all three statements are equivalent, as is a fourth statement called the *Hausdorff Maximality Principle.*

1. Partial Order

Let $\mathcal{S}$ be any set, and let $\mathcal{P}(\mathcal{S})$ be its power set, the set whose elements are the subsets of $\mathcal{S}$,

$$\mathcal{P}(\mathcal{S}) = \{S : S \text{ is a subset of } \mathcal{S}\}.$$

The *proper subset* relation "$\subsetneq$" in $\mathcal{P}(\mathcal{S})$ satisfies two conditions:

- (*Partial Trichotomy*) For any $S, T \in \mathcal{P}(\mathcal{S})$, at most one of the conditions
$$S \subsetneq T, \qquad S = T, \qquad T \subsetneq S$$
holds, and possibly none of them holds.
- (*Transitivity*) For any $S, T, U \in \mathcal{P}(\mathcal{S})$, if $S \subsetneq T$ and $T \subsetneq U$ then also $S \subsetneq U$.

The two conditions are the motivating example of a *partial order.*

Definition 1.1. *Let T be a set. Let $\prec$ be a relation that may stand between some pairs of elements of T. Then $\prec$ is a* **partial order** *on T if*

- *For any $s, t \in T$, at most one of the conditions*
$$s \prec t, \qquad s = t, \qquad t \prec s$$
holds.
- *For any $r, s, t \in T$, if $r \prec s$ and $s \prec t$ then also $r \prec t$.*

If for any $s, t \in T$, also at least *one of the three conditions in the first bullet holds then the partial order is a* **total order***.*

A relation $\prec$ on some pairs from a set T need have nothing to do with any sort of numerical order.

Visibly, any subset of a partially ordered set is again partially ordered, and similarly for a totally ordered set.

If $\prec$ is a partial order on T then for any $s, t \in T$, the definitions of the relations

$$\succ, \qquad \preceq, \qquad \succeq$$

in terms of the basic relation $\prec$ are clear, and so we feel free to use these additional relations without further comment.

Definition 1.2. *Let $(T, \prec)$ be a partially ordered set. A subset S of T is an* **initial segment** *if*

$$\textit{for each } s \in S, \textit{ also } r \in S \textit{ for all } r \prec s.$$

The **section** *of any element $t \in T$ is*

$$T_t = \{s \in T : s \prec t\}.$$

Thus any section is an initial segment, while T is an initial segment that is not a section.

2. Well Ordering

Definition 2.1. *Let $(T, \prec)$ be a partially ordered set. Then T is* **well ordered** *by $\prec$ if*

every nonempty set of T has a least element.

That is, T is well ordered by $\prec$ if every nonempty subset S of T contains an element $s_o \in S$ such that $s_o \preceq s$ for all $s \in S$. By partial trichotomy there is at most one such element.

Complementing the general observations that any section of a partially ordered set $(T, \prec)$ is an initial segment and that T is an initial segment but not a section, we note that if T is well ordered by $\prec$ then also any initial segment other than T is a section. Indeed, given an initial segment $S \subsetneq T$, the complement of S in T has a least element t, and one quickly sees that $S = T_t$.

3. Closed Sets and the Fundamental Lemma

Definition 3.1. *Let $(T, \prec)$ be a partially ordered set, and let S be a subset of T. Any element $t \in T$ such that*

$$s \preceq t \quad \textit{for all } s \in S$$

is called an **upper bound** *of S. If t_o is an upper bound of S such that*

$$t_o \preceq t \quad \textit{for all upper bounds } t \textit{ of } S$$

then t_o is the **least upper bound** *of S.*

Partial trichotomy shows that a least upper bound, if it exists, is unique, justifying the *the* least upper bound in the definition.

Definition 3.2. *Let $(T, \prec)$ be a partially ordered set. Any totally ordered subset of T is called a* **chain** *in T. If the ordering of T has the property that every chain in T has a least upper bound in T then T is* **closed**.

Lemma 3.3 (Fundamental Lemma)**.** *Let $(T, \prec)$ be a closed partially ordered set. Suppose that a mapping*

$$f : T \longrightarrow T$$

satisfies the condition

$$f(t) \succeq t \quad \textit{for all } t \in T.$$

Then there exists some $t_o \in T$ such that $f(t_o) = t_o$.

Proof. Since T is closed, there is a least upper bound function on chains in T,

$$g : \{\text{chains in } T\} \longrightarrow T, \quad g(C) = \text{least upper bound of } C.$$

Given a chain C, each of its sections C_t is a chain in turn and hence has a least upper bound $g(C_t)$. A chain C that is well ordered and satisfies

$$fg(C_t) = t \quad \text{for all } t \in C$$

is called an *fg-chain.* Every initial segment of an fg-chain is again an fg-chain.

Let C and D be fg-chains. We will show that one of them is an initial segment of the other. Suppose that D is not an initial segment of C.

The set of initial segments I of D is well ordered by containment; indeed, each $I \neq D$ is a section $I = D_{d(I)}$ as noted at the end of section 2 above, so given a set $\{I_i\}$ of such segments, the set $\{d(I_i)\}$ has a least element $d(I_j)$ and thus $\{I_i\}$ has least element I_j. Since D is not an initial segment of C, it is sensible to define

$$A = \text{the first initial segment of } D \text{ that is not an initial segment of } C.$$

If A has no last element then each element a of A lies in the proper initial segment $A_a \cup \{a\}$ of A. The proper initial segments of A are initial segments of C by definition of A, and thus A is an initial segment of C, contradicting its own definition. So we may assume that A has a last element t.

The initial segment A_t is both an initial segment of C and an initial segment of D. If $A_t = C$ then C is an initial segment of D, and we are done. (Recall that we are trying to prove that if the fg-chain D is not an initial segment of the fg-chain C then C must be an initial segment of D.) Thus it remains to show that the condition $A_t \neq C$ is impossible. Let $A_t \neq C$ and let u be the first element of $C - A_t$. Then

$$D_t = A_t = C_u$$

and hence, since C and D are fg-chains,

$$t = fg(D_t) = fg(C_u) = u.$$

Thus

$$A = A_t \cup \{t\} = C_u \cup \{u\} \quad \text{is an initial segment of } C,$$

contrary to the definition of A, and showing that the supposition $A_t \neq C$ is impossible. This completes the argument that given any two fg-chains in T, one is always an initial segment of the other.

Recall that T is closed, that $f : T \longrightarrow T$ is such that $f(t) \succeq t$ for all $t \in T$, and that we are trying to find some $t_o \in T$ such that $f(t_o) = t_o$. To do so, let S be the union of all fg-chains in T. Then

(1) the argument just given shows that S is linearly ordered and thus a chain,
(2) S is well ordered,
(3) $t = fg(S_t)$ for all $t \in S$, and thus S is an fg-chain,
(4) No proper superset of S is an fg-chain.

Now take

$$t_o = fg(S) \succeq g(S),$$

an upper bound of S. If $t_o \notin S$ then $S \cup \{t_o\}$ is an fg-chain, contradicting (4). Thus $t_o \in S$, and hence $t_o \preceq g(S)$, so by the previous display $t_o = g(S)$, and finally,

$$f(t_o) = fg(S) = t_o,$$

as desired. □

4. Zorn's Lemma

Definition 4.1. *Let $(T, \prec)$ be a partially ordered set. A* **maximal** *element of T is an element m of T satisfying the condition*

$$m \not\prec t \quad \textit{for all } t \in T.$$

That is:

> To say that an element is maximal is not necessarily to say *it is bigger than all others*, but rather *no other is bigger.*

We now assume the Axiom of Choice.

Theorem 4.2 (Zorn's Lemma)**.** *Let $(T, \prec)$ be a closed partially ordered set. Then T contains at least one maximal element.*

Proof. Suppose that T has no maximal element. Then for each $t \in T$ there is a nonempty subset of T,

$$S_t = \{u \in T : t \prec u\}.$$

The Axiom of Choice says that consequently there exists a function $f : T \longrightarrow T$ such that $f(t) \in S_t$ for all $t \in T$. That is,

$$f(t) \succ t \quad \text{for all } t \in T.$$

This contradicts the Fundamental Lemma. □

5. The Well Ordering Principle

Again we assume the Axiom of Choice.

Theorem 5.1 (Well Ordering Principle)**.** *Let T be a set. Then T can be well ordered.*

Proof. By the Axiom of Choice, there exists a function φ that assigns to each proper subset S of T an element of $T - S$. Define a *φ-chain* to be a well ordered subset $(S, \prec)$ of T such that

$$\varphi(S_t) = t \quad \text{for all } t \in S.$$

(Here $S_t = \{s \in S : s \prec t\}$ is a section of S as before.)

The argument of the Fundamental Lemma applies with φ-chains in place of fg-chains. The union S of all φ-chains is well ordered, is a φ-chain, and no proper superset of S is again a φ-chain.

If $S \subsetneq T$ then we could adjoin $\varphi(S) \in T - S$ to S as a terminal element, obtaining a proper superset of S that is again a φ-chain, contradiction. Thus $S = T$, showing that T is well ordered. □

6. Transfinite Induction

Theorem 6.1 (Transfinite Induction Principle)**.** *Let $(T, \prec)$ be well ordered. Consider a proposition form $\mathcal{P}$ over T. If*

$$\textit{for every } t \in T, \quad \mathcal{P}(s) \textit{ for all } s \in T_t \implies \mathcal{P}(t)$$

then

$$\mathcal{P}(t) \textit{ for all } t \in T.$$

Proof. Otherwise there is a least $t \in T$ such that $\mathcal{P}(t)$ is false. But then $P(s)$ for all $s \in T_t$, and so $P(t)$ is true after all, contradiction. □

7. An Application to Field Constructions

Given a field k and a nonconstant polynomial $f \in k[X]$, constructing a smallest-possible field extension of k in which f has a root is easy, and so is constructing a smallest-possible field extension of k in which f splits completely into linear factors. But constructing a smallest-possible field extension $\overline{k}$ of k in which *every* nonconstant polynomial in $k[X]$ has a root, and in fact even every nonconstant polynomial in $\overline{k}[X]$ splits into linear factors, requires Zorn's Lemma. We carry out the three field constructions in succession.

7.1. **Root Fields.** Let k be any field, and let $f(X) \in k[X]$ be irreducible and have positive degree. We want to construct a superfield K of k in which f has a root. To do so, consider the quotient ring

$$R = k[X]/\langle f\rangle,$$

where $\langle f\rangle$ is the principal ideal $f(X)k[X]$ of $k[X]$. That is, R is the usual ring of polynomials over k subject to the additional rule $f(X) = 0$. Specifically, the ring-elements are cosets and the operations are

$$\begin{aligned}(g + \langle f\rangle) + (h + \langle f\rangle) &= (g + h) + \langle f\rangle,\\ (g + \langle f\rangle)(h + \langle f\rangle) &= gh + \langle f\rangle.\end{aligned}$$

The ring forms a vector space over k whose dimension is $\deg(f)$.

The fact that f is irreducible makes the ideal $\langle f\rangle$ maximal, and consequently R is a field, not only a ring. Indeed, consider an ideal I of $k[X]$ that contains $\langle f\rangle$ and some $g \notin \langle f\rangle$. Thus $f \nmid g$, and so $(f, g) = 1$ (because f is irreducible). So there exist $F, G \in k[X]$ such that

$$Ff + Gg = 1 \quad \text{in } k[X].$$

Since the ideal I contains f and g, it contains 1, making it all of R.

Now use the field R to create a set K of symbols that is a superset of k and is in bijective correspondence with R. That is, there is a bijection

$$\sigma : R \xrightarrow{\sim} K, \qquad \sigma(a + \langle f\rangle) = a \text{ for all } a \in k.$$

Endow K with addition and multiplication operations that turn the set bijection into a field isomorphism. The operations on K thus extend the operations on k. Name a particular element of K,

$$r = \sigma(X + \langle f\rangle).$$

Then

$$\begin{aligned}f(r) &= f(\sigma(X + \langle f\rangle)) && \text{by definition of } r\\ &= \sigma(f(X + \langle f\rangle)) && \text{since algebra passes through } \sigma\\ &= \sigma(f(X) + \langle f\rangle) && \text{by the nature of algebra in } R\\ &= \sigma(\langle f\rangle) && \text{by the nature of algebra in } R\\ &= 0 && \text{by construction of } \sigma.\end{aligned}$$

Thus K is a superfield of k containing an element r such that $f(r) = 0$.

For example, since the polynomial $f(X) = X^3 - 2$ is irreducible over $\mathbb{Q}$, the corresponding quotient ring

$$R = \mathbb{Q}[X]/\langle X^3 - 2\rangle = \{a + bX + cX^2 + \langle X^3 - 2\rangle : a, b, c \in \mathbb{Q}\}$$

is a field. And from R we construct a field (denoted $\mathbb{Q}(r)$ or $\mathbb{Q}[r]$) such that $r^3 = 2$. Yes, we know that there exist cube roots of 2 in the superfield $\mathbb{C}$ of $\mathbb{Q}$, but the construction given here is purely algebraic and makes no assumptions about the nature of the starting field k to which we want to adjoin a root of a polynomial.

7.2. **Splitting Fields.** Again let k be a field and consider a nonunit polynomial $f(X) \in k[X]$. We can construct an extension field

$$k_1 = k(r_1),$$

where r_1 satisfies some irreducible factor of f. Let

$$f_2(X) = f(X)/(X - r_1) \in k_1[X].$$

We can construct an extension field

$$k_2 = k_1(r_2) = k(r_1, r_2),$$

where r_2 satisfies some irreducible factor of f_2. Continue in this fashion until reaching a field where the original polynomial f factors down to linear terms. The resulting field is the **splitting field of f over** k, denoted

$$\mathrm{spl}_k(f).$$

Continuing the example of the previous section, compute that

$$\frac{X^3 - 2}{X - r} = X^2 + rX + r^2 \quad \text{in } \mathbb{Q}(r)[X].$$

Let $s = rt$ where $t^3 = 1$ but $t \neq 1$. Then, working in $\mathbb{Q}(r, t)$ we have

$$s^2 + rs + r^2 = r^2t^2 + r^2t + r^2 = r^2(t^2 + t + 1) = r^2 \cdot 0 = 0,$$

Thus $s = rt$ satisfies the polynomial $X^2 + rX + r^2$, and now compute that

$$\frac{X^2 + rX + r^2}{X - rt} = X - rt^2 \quad \text{in } \mathbb{Q}(r, t)[X].$$

That is,

$$X^3 - 2 = (X - r)(X - rt)(X - rt^2) \in \mathbb{Q}(r, t)[X],$$

showing that

$$\mathrm{spl}_{\mathbb{Q}}(X^3 - 2) = \mathbb{Q}(r, t).$$

7.3. **Algebraic Closure.** Again let k be a field. Associate to each nonconstant monic irreducible polynomial $f \in k[X]$ an indeterminate X_f, and let S denote the set of such indeterminates,

$$S = \{X_f : f \in k[X] \text{ nonconstant monic irreducible}\}.$$

Consider the ring of polynomials in the elements of S,

$$A = k[S],$$

and consider the ideal of A generated by all the $f(X_f)$,

$$I_o = \langle \{f(X_f) : X_f \in S\} \rangle.$$

We show that I_o is a proper ideal of A by showing that no relation exists of the form

$$\sum_f g_f(S) f(X_f) = 1,$$

where the sum is finite and each g_f is a polynomial in finitely many elements of S. Any such relation occurs in in a subring of A consisting of the polynomials in finitely many variables,

$$A_o = k[\{X_f \text{ appearing in the relation}\}].$$

Using the earlier results from this section, we may construct a superfield K of k where for each variable X_f occurring in the relation, the polynomial f has a root r_f. There is a homomorphism

$$A_o \longrightarrow K, \quad X_f \longmapsto r_f \text{ for each } X_f \text{ appearing in the relation.}$$

Let R denote the set of roots r_f from a moment ago. Pass the relation through the homomorphism to get

$$\sum_f g_f(R) f(r_f) = 1.$$

But this is impossible since each $f(r_f) = 0$. This completes the argument that I_o is a proper ideal of A.

Now consider the set of superideals of I_o that are again proper ideals of A

$$\mathcal{I}(A) = \{I : I_o \subseteq I \subsetneq A,\ I \text{ is an ideal of } A\}.$$

This set is partially ordered by proper containment. Given any chain C in $\mathcal{I}(A)$, consider the union of all elements of all ideals of the chain,

$$J = \bigcup_{I \in C} I \subset A.$$

Each element x of J lies in some ideal I_x of C. Thus, any two elements x, y of J lie in a common ideal I of C since one of I_x and I_y contains or equals the other. Thus $x + y \in I$ as well since I is an ideal, and consequently $x + y \in J$. That is, J is closed under addition. Similar arguments show that J is in fact an ideal of A. Also, $1 \notin J$ since $1 \notin I$ for each ideal I of the chain C. In sum, $J \in \mathcal{I}(A)$, and we have shown that

$$\mathcal{I}(A) \textit{ is closed.}$$

Consequently, by Zorn's Lemma,

$$\mathcal{I}(A) \textit{ has a maximal element } M,$$

which is to say,

$$A \textit{ contains a maximal superideal } M \textit{ of } I_o.$$

Because the ideal is maximal, the corresponding quotient is a field,

$$\overline{k} = A/M.$$

And in the quotient each nonconstant monic irreducible polynomial $f \in k[X]$ has a root, $X_f + M$.

We want to establish two more properties of $\overline{k}$:

(1) It is *not too big*, in the sense that every element of $\overline{k}$ satisfies a polynomial over k.

(2) It is *big enough*, in the sense that every nonconstant polynomial $g \in \overline{k}[X]$ has a root in $\overline{k}[X]$, and therefore every such g has all of its roots in $\overline{k}[X]$.

Establishing these properties is facilitated by a bit of vocabulary.

Definition 7.1. *Let k be a field. An element α of a superfield K of k is* **algebraic over** k *if α is a root of some nonzero polynomial in $k[X]$. A superfield K of k is algebraic over k if each of its elements is algebraic over k. The field k is* **algebraically closed** *if no proper superfield is algebraic over k.*

If α is algebraic over k then the polynomials in $k[X]$ satisfied by α form a nonzero ideal. Because $k[X]$ is a PID, the ideal has a unique monic generator (*monic* means that the coefficient of the highest power of X is 1). This generator is called the **minimal polynomial** of α over k. The minimal polynomial is irreducible, since otherwise one of its proper factors is again satisfied by α, contradicting its property of dividing all such polynomials.

Proposition 7.2. *Every element of $\overline{k}$ is algebraic over k.*

Proof. Every element α of $\overline{k}$ is the coset of a polynomial in finitely many X_f, and so it lies in the subfield of $\overline{k}$ generated by finitely many r_i. The subfield forms a finite-dimensional vector space over k, and so the powers of α,

$$\{1, \alpha, \alpha^2, \cdots\}$$

must be linearly dependent over k. That is, α satisfies a nonzero polynomial over k. □

Proposition 7.3. *Any nonconstant polynomial in $\overline{k}[X]$ has all of its roots in $\overline{k}$ as well. That is, $\overline{k}$ is algebraically closed.*

Proof. It suffices to consider any irreducible polynomial in $\overline{k}[X]$, and it suffices to show that one root of the polynomial lies in $\overline{k}$. Thus, consider a monic irreducible polynomial in $\overline{k}[X]$,

$$x^n + c_1 x^{n-1} + \cdots + c_n, \quad c_i \in \overline{k},$$

and let α be one of its roots. Each ring $k[c_i]$ is a finite-dimensional vector space over k, and hence so is the ring

$$R_o = k[c_1, \ldots, c_n].$$

Let

$$R = R_o[\alpha] = k[c_1, \ldots, c_n, \alpha].$$

If $\{v_i : 1 \leq i \leq m\}$ is a basis of R_o over k then

$$\{v_i \alpha^j : 1 \leq i \leq m, \, 0 \leq j < n\}$$

is a basis of R as a vector space over k. Thus R has a basis $\{r_1, \ldots, r_\ell\}$ as a vector space over k. Multiply α by each basis element to get

$$\begin{aligned}
\alpha r_1 &= a_{11} r_1 + a_{12} r_2 + \cdots + a_{1\ell} r_\ell, \\
\alpha r_2 &= a_{21} r_1 + a_{22} r_2 + \cdots + a_{1\ell} r_\ell, \\
&\vdots \\
\alpha r_\ell &= a_{\ell 1} r_1 + a_{\ell 2} r_2 + \cdots + a_{\ell\ell} r_\ell,
\end{aligned}$$

or, letting $\vec{r}$ denote the column vector with entries $r_1, \ldots, r_\ell$,

$$\alpha \vec{r} = A \vec{r}, \quad A \in k^{\ell \times \ell}.$$

Thus α is an eigenvalue of A, a root of the characteristic polynomial of A, a polynomial with coefficients in the original ground field k. But the field $\overline{k}$ already contains all roots of all nonconstant polynomials in $k[X]$. That is, $\alpha \in \overline{k}$ as desired. □

In sum, $\overline{k}$ is algebraic over k and algebraically closed. One can further show that every algebraic extension of k embeds (injects homomorphically) in $\overline{k}$ and that $\overline{k}$ is unique up to isomorphism, but we omit these arguments.