Set theory and logic throughout mathematics

Chris Lambie-Hanson

March 3, 2024

Contents

Chapter 1

Lecture 1: Ordinals and the hydra

1.1 Well-orders

Let us begin by briefly reviewing the definition of *partial order*, *linear order*, and *well-order.*

Definition 1.1.1. Suppose that X is a set and $\leq$ is a binary relation on X. Then $\leq$ is a *partial order* on X (or $(X, \leq)$ is a *partial order*) if it is

1. *Reflexive*: $x \leq x$ for all $x \in X$;
2. *Transitive*: for all $x, y, z \in X$, if $x \leq y$ and $y \leq z$, then $x \leq z$; and
3. *Anti-symmetric*: for all $x, y \in X$, if $x \leq y$ and $y \leq x$, then $x = y$.

A partial order $\leq$ on a set X is a *linear order* if, in addition, it is *total*, i.e., for all $x, y \in X$, we have $x \leq y$ or $y \leq x$.

If $\leq$ is a partial order on a set X and $x, y \in X$, then we will write $x < y$ to mean that $x \leq y$ and $x \neq y$. The relation $<$ is then referred to as the *strict* part of $\leq$.

Definition 1.1.2. Suppose that X is a set and R is a binary relation on X. Then R is *well-founded* if every nonempty subset of X has an R-minimal element. In other words, for every nonempty $Y \subseteq X$, there is $y \in Y$ such that, for all $x \in Y$ with $x \neq y$, $\neg(xRy)$. A well-founded linear order is called a *well order.*

Exercise 1.1.3. Suppose that $\leq$ is a linear order on a set X. Prove that the following are equivalent.

1. $\leq$ is a well order.
2. There are no infinite, strictly decreasing sequences with respect to $\leq$. In other words, there does not exist a sequence $\langle x_0, x_1, x_2, \ldots \rangle$ of elements of X such that, for all n, we have $x_{n+1} < x_n$.

There is a natural way to assert that two partial orders are "essentially" the same, i.e., *isomorphic.*

Definition 1.1.4. Suppose that $\leq_0$ is a partial order on X_0 and $\leq_1$ is a partial order on X_1. Then we say that $\leq_0$ and $\leq_1$ are *isomorphic* if there is a bijection $F : X_0 \to X_1$ such that, for all $x, y \in X_0$, we have

$$x \leq_0 y \iff F(x) \leq_1 F(y).$$

Example 1.1.5. The following are some examples and non-examples of isomorphic partial orders.

1. The open interval $(0, 1)$ and the open interval $(0, 2)$, both with the usual ordering of real numbers, are isomorphic via the bijection $x \mapsto 2x$.

2. The open interval $(0, 1)$ and the closed interval $[0, 1]$ are not isomorphic. One way to see this is to note that $[0, 1]$ has a maximal element and $(0, 1)$ does not, so any order-preserving map from $(0, 1)$ to $[0, 1]$ could not include 1 in its range.

3. Let Y be any nonempty set. Let $X_0 = \mathscr{P}(Y)$ be the *power set* of Y, i.e., the collection of all subsets of Y. Let $\leq_0$ be the partial order on X_0 defined by letting $u \leq v$ if and only if $u \leq v$.

   Let X_1 be the collection of all functions $f : Y \to \{0, 1\}$, and let $\leq_1$ be the partial order on X_1 defined by letting $f \leq g$ if and only if $f(y) \leq g(y)$ for all $y \in Y$.

   Then $\leq_0$ and $\leq_1$ are isomorphic via the bijection $F : X_0 \to X_1$ that sends each $u \in X_0$ to the *characteristic function* of u, i.e., the function $f_u : Y \to \{0, 1\}$ that takes value 1 on all elements in u and value 0 on all elements of Y that are not in u.

1.2 Ordinal numbers

Roughly speaking, an ordinal number can be thought of as a description of the order type of a well-order. In other words, to each well-order, we assign an ordinal, and two well-orders are isomorphic if and only if they are assigned the same ordinal.

Example 1.2.1. For each natural number n, all well-orders of size n are isomorphic; their order type is itself referred to as "n".

However, there are many non-isomorphic countably infinite well-orders. The ordinal describing the order type of the natural numbers,

$$0 < 1 < 2 < 3 < \ldots$$

is denoted "ω". But we can form a new order type by adding a new element (call it ∞) that is larger than all of the natural numbers:

$$0 < 1 < 2 < 3 < \ldots < \infty.$$

The ordinal describing this order type is denoted "$\omega + 1$". Or we can form yet another order type by placing two copies of the natural numbers one after the other:

$$0 < 1 < 2 < 3 < \ldots < 0' < 1' < 2' < 3' < \ldots.$$

The ordinal describing this order type is denoted "$\omega + \omega$".

There is a natural way to order the ordinal numbers themselves. To make this precise, we need the following definition.

Definition 1.2.2. Suppose that $\leq$ is a well-order of a set X. Then an *initial segment* of $(X, \leq)$ is a subset $Y \subseteq X$ such that, for all $y \in Y$ and all $x \in X$, if $x \leq y$, then $x \in Y$. In other words, if $y \in Y$, then Y also contains all elements of X that are smaller than y in the ordering $\leq$.

Exercise 1.2.3. Suppose that $\leq$ is a well-order of a set X and Y is an initial segment of $(X, \leq)$. Then either

- $Y = X$; or
- there is $x \in X$ such that $Y = \{y \in X \mid y < x\}$.

Exercise 1.2.4. Suppose that $(X_0, \leq_0)$ and $(X_1, \leq_1)$ are two well-orders. Then either

1. $(X_0, \leq_0)$ is isomorphic to an initial segment of $(X_1, \leq_1)$; or
2. $(X_1, \leq_1)$ is isomorphic to an initial segment of $(X_0, \leq_0)$.

If both 1. and 2. hold, then $(X_0, \leq_0)$ and $(X_1, \leq_1)$ are isomorphic.

With Exercise 1.2.4 in mind, we can make the following definition.

Definition 1.2.5. Suppose that α and β are ordinals. Then we say that $\alpha \leq_{\text{ord}} \beta$ if, whenever $(X_\alpha, \leq_\alpha)$ is a well-order of type α and $(X_\beta, \leq_\beta)$ is a well-order of type β, then $(X_\alpha, \leq_\alpha)$ is isomorphic to an initial segment of $(X_\beta, \leq_\beta)$.

Exercise 1.2.6. The class of ordinals is well-ordered by $\leq_{\text{ord}}$.

One can perform arithmetic on ordinal numbers. We will make this more precise later, but let us first give an informal description. Let α and β be ordinal numbers, and let $(X_\alpha, \leq_\alpha)$ and $(X_\beta, \leq_\beta)$ be well-orders of type α and β, respectively.

We first describe ordinal addition. The ordinal $\alpha+\beta$ is the ordinal describing the well-ordering formed by placing a copy of $(X_\beta, \leq_\beta)$ after a copy of $(X_\alpha, \leq_\alpha)$ (i.e., every element of X_β is declared to be larger than every element of X_α).

Note that ordinal addition is not commutative: it may not be the case that $\alpha + \beta = \beta + \alpha$. To see this, consider $2 + \omega$ and $\omega + 2$. Represent the ordinal 2 by the order $0' < 1'$, and represent ω by the usual natural numbers. Then $2 + \omega$ is the order type of the order

$$0' < 1' < 0 < 1 < 2 < 3 < \dots$$

This is isomorphic to the usual ordering of the natural numbers, via the map

$$\begin{aligned} 0' &\mapsto 0 \\ 1' &\mapsto 1 \\ n &\mapsto n + 2 \text{ for every natural number } n. \end{aligned}$$

Thus, $2 + \omega = \omega$. On the other hand, $\omega + 2$ is the order type of the order

$$0 < 1 < 2 < \ldots < 0' < 1'$$

This is clearly *not* isomoprhic to the natural numbers; for example, it has a maximal element, whereas that natural numbers do not. Thus, $\omega + 2 \neq \omega$, and in fact $\omega <_{\mathrm{ord}} \omega + 2$.

We next describe ordinal multiplication. The ordinal $\alpha \cdot \beta$ is the ordinal describing the well-ordering formed by starting with a copy of $(X_\beta, \leq_\beta)$ and replacing every element of X_β with a copy of $(X_\alpha, \leq_\alpha)$.

Again, ordinal multiplication is not commutative. For example, $2 \cdot \omega$ is the order type of the following order:

$$0 < 0' < 1 < 1' < 2 < 2' < 3 < 3' < \ldots$$

formed by replacing every natural number n with a copy $n < n'$ of the two-element order. It is not too hard to show that this order is isomorphic to the natural numbers, so $2 \cdot \omega = \omega$. On the other hand, $\omega \cdot 2$ is the order type of the following order:

$$0 < 1 < 2 < 3 < \ldots < 0' < 1' < 2' < 3' < \ldots$$

formed by replacing each element of the two-element order $* < *'$ by a copy of the natural numbers. This is not isomorphic to the set of natural numbers; for instance, it contains elements that are larger than infinitely many other elements, whereas the natural numbers do not. Thus, $\omega \cdot 2 \neq \omega$, and in fact $\omega <_{\mathrm{ord}} \omega \cdot 2$.

We finally describe ordinal exponentiation. If $\alpha = 0$, then $\alpha^\beta = 0$. Otherwise, first let 0_α denote the *minimal* element of $(X_\alpha, \leq_\alpha)$. This must exist, since $\leq_\alpha$ is a well-order. We say that a function $f : X_\beta \to X_\alpha$ is *finitely supported* if the set $\{y \in X_\beta \mid f(y) \neq 0_\alpha\}$ is finite. The ordinal α^β is now defined as follows. Let Z be the set of all finitely-supported functions from X_β to X_α. Now describe an ordering $\preceq$ on Z as follows. Given $f, g \in Z$, set $f \preceq g$ if and only if either

- $f = g$; or
- $f \neq g$ and, letting $y \in X_\beta$ be the $\leq_\beta$-maximal element such that $f(y) \neq g(y)$, we have $f(y) \leq_\alpha g(y)$.

Then let α^β be the ordinal describing the order type of $(Z, \preceq)$.

Exercise 1.2.7. Prove that the order $(Z, \preceq)$ described in the preceding paragraph is indeed a well-order.

A concrete representation of the ordinals.

In practice we often work with a particular concrete realization of the ordinals, and we think of an ordinal α as the set of all ordinals that are strictly less than α (with respect to the ordering $\leq_{\mathrm{ord}}$ introduced above. At first glance, this may appear like a circular definition, but it is not, due to the fact that $\leq_{\mathrm{ord}}$ is itself a well-ordering. In particular, there is a least ordinal, 0. Since there are no ordinals strictly less than 0, we represent 0 as the empty set, $\emptyset$. The next smallest ordinal is 1. It only has one ordinal less than it, namely, 0, so 1 is represented as $\{0\} = \{\emptyset\}$. The first few ordinals are thus represented as follows:

$$\begin{aligned}
0 &= \emptyset \\
1 &= \{0\} = \{\emptyset\} \\
2 &= \{0,1\} = \{\emptyset,\{\emptyset\}\} \\
3 &= \{0,1,2\} = \{\emptyset,\{\emptyset\},\{\emptyset,\{\emptyset\}\}\} \\
4 &= \{0,1,2,3\} = \{\emptyset,\{\emptyset\},\{\emptyset,\{\emptyset\}\},\{\emptyset,\{\emptyset\},\{\emptyset,\{\emptyset\}\}\}\} \\
&\dots \\
\omega &= \{0,1,2,3,4,\dots\} \\
\omega+1 &= \omega \cup \{\omega\} = \{0,1,2,3,4,\dots\} \cup \{\omega\} \\
\omega+2 &= \omega \cup \{\omega,\omega+1\} \\
&\dots \\
\omega+\omega &= \omega\cdot 2 = \{0,1,2,3,4,\dots\} \cup \{\omega,\omega+1,\omega+2,\omega+3,\omega+4,\dots\}
\end{aligned}$$

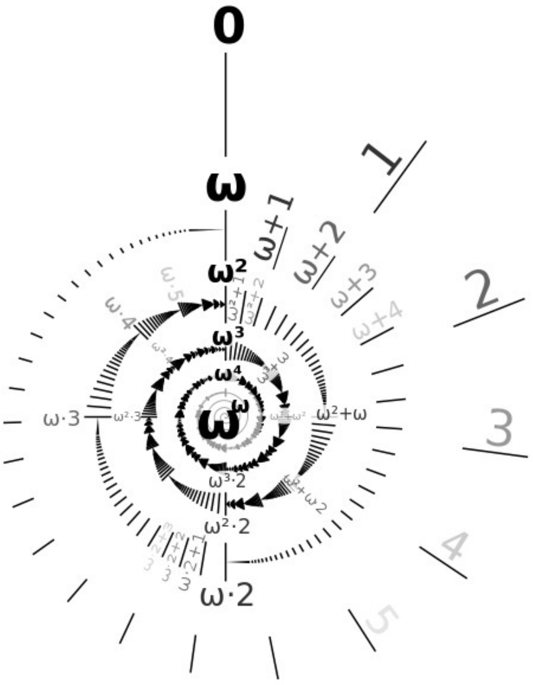


Figure 1.1: A stylized image of the ordinals up to ω^ω

With this concrete representation of the ordinals, we can easily be more precise about ordinal arithmetic. We first introduce the following notions.

Definition 1.2.8. Let X be a nonempty set of ordinals. Then the *supremum* of X, denoted $\sup(X)$, is the least ordinal that is greater than or equal to every element of X.

Exercise 1.2.9. Working with our concrete representation of the ordinals, prove that, for every nonempty set of ordinals X, the supremum of X is equal to the union of all of the elements of X, i.e.,

$$\sup(X) = \bigcup X.$$

Definition 1.2.10. Suppose that β is an ordinal.

1. We say that β is a *successor* ordinal if $\beta = \alpha + 1$ for some ordinal α.
2. If β is not a successor ordinal, we say that β is a *limit* ordinal.

We can now rigorously define ordinal arithmetic by recursion. We first deal with addition. For all ordinals α, we let:

- $\alpha + 0 = \alpha$;
- $\alpha + 1 = \alpha \cup \{\alpha\}$;
- for all ordinals β, we have $\alpha + (\beta + 1) = (\alpha + \beta) + 1$;
- if γ is a nonzero limit ordinal, then $\alpha + \gamma = \sup\{\alpha + \beta \mid \beta < \gamma\}$.

Next, multiplication:

- $\alpha \cdot 0 = 0$;
- $\alpha \cdot 1 = \alpha$;
- for all ordinals β, we have $\alpha \cdot (\beta + 1) = (\alpha \cdot \beta) + \alpha$;
- if γ is a nonzero limit ordinal, then $\alpha \cdot \gamma = \sup\{\alpha \cdot \beta \mid \beta < \gamma\}$.

Finally, exponentiation:

- $\alpha^0 = 1$;
- $\alpha^1 = \alpha$;
- for all ordinals β, we have $\alpha^{\beta+1} = (\alpha^\beta) \cdot \alpha$;
- if γ is a nonzero limit ordinal, then $\alpha^\gamma = \sup\{\alpha^\beta \mid \beta < \gamma\}$.

1.3 The hydra

We end this first lecture with a surprising demonstration of the utility of infinite ordinals: the hydra game. You may be familiar with the Hydra from Greek mythology. It is a fearsome water monster with many heads with the property that, whenever you chop off one of its heads, two heads will grow back in its place. Eventually, the hydra was slain by Heracles, with the assistance of his nephew Iolaus.

We will be examining a game played using a mathematical version of the Hydra introduced in the paper "Accessible independence results for Peano Arithmetic" by Laurie Kirby and Jeff Paris. For us, a *hydra* is a finite tree with a root. In other words, a hydra consists of finitely many nodes and edges. There is a root node at the bottom, which has finitely many edges coming out of it, each leading to another node. In turn, each of these nodes has finitely many edges coming out of it, each leading to a further node, and so on. For example, this is a hydra:

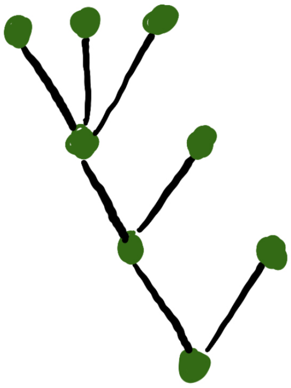


Figure 1.2: A hydra

We will always draw hydras with the root at the bottom. A *terminal node* of a hydra is a non-root node that is connected to only one other node. A *head* of a hydra consists of a terminal node and the single edge that leads to it. For example, the hydra pictured above has five heads:

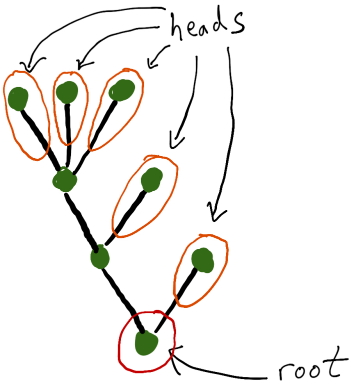


Figure 1.3: A hydra with its root and heads labeled

Given a head, the single node that it is attached to is called its *parent*. If its parent is not the root, then the node that is one step closer to the root from its parent is called its *grandparent*:

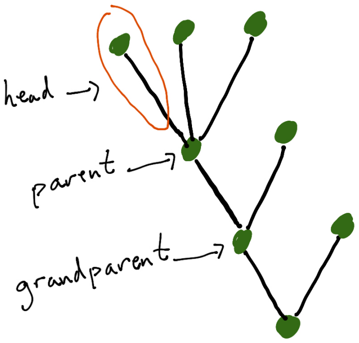


Figure 1.4: A hydra with a labeled head, its parent, and its grandparent

In the hydra game, we start with a hydra and, on each move (starting with Move 1), we chop off one of its heads. Our goal is to reduce the hydra to only a root node in a finite number of moves. However, like its mythological counterpart, the hydra regenerates, according to the following rules:

- If, on Move n, we chop off a head directly connected to the root, then the hydra does not create any new heads.

- If, on Move n, we chop off a head *not* directly connected to the root, then first delete the node and edge that make up that head. Then, move down one edge towards the root, to the edge connecting the parent and the grandparent of the head that was removed. The hydra makes n new copies of the subtree consisting of this edge and everything above it and attaches each of these new copies to the grandparent of the head that was removed.

This is best illustrated with a picture. Suppose that we are about to make Move 2 of a game, and we are confronted with the hydra pictured above. One option is to chop off the head in the bottom right, directly connected to the root:

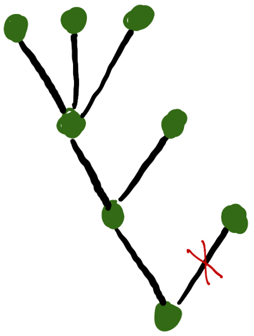


Figure 1.5: Chopping off a head directly connected to the root

Since this head is directly connected to the root, the hydra does not generate any new heads, so on our next move we see the following hydra:

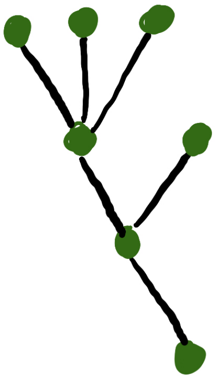


Figure 1.6: The result of the move in Figure 1.5

However, we could have done something different on Move 2 and instead chopped off the head on the upper left:

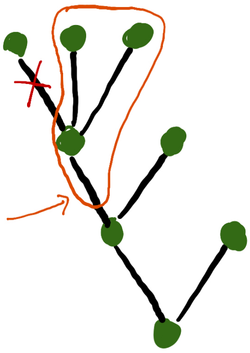


Figure 1.7: Chopping off a head not directly connected to the root

Now, to generate the hydra for the next move, we first remove the head. Then we consider the subtree consisting of the edge between the head's parent and grandparent and everything above it (circled in orange in Figure 1.6). Since we are on Move 2, we make 2 new copies of it and attach them to the grandparent of the removed head, resulting in the following hydra:

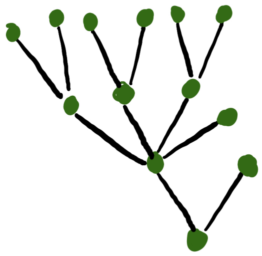


Figure 1.8: The result of making the move in Figure 1.7 on Move 2

Let us see now a complete play of the game, starting from a very simple hydra:

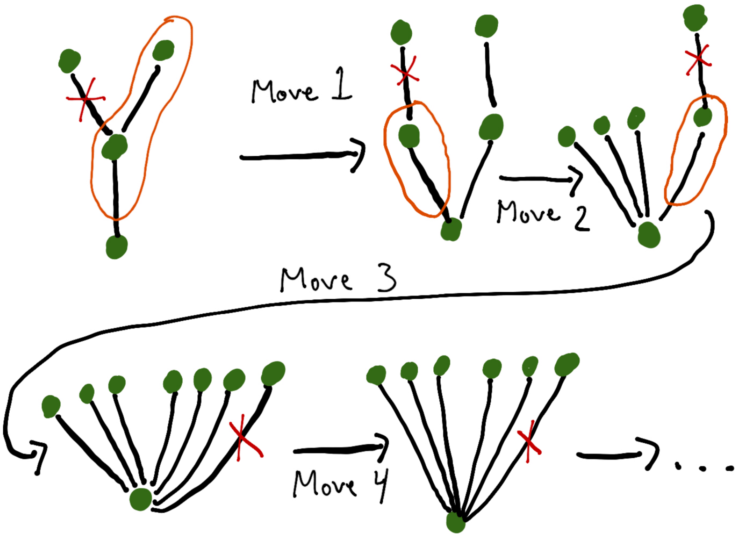


Figure 1.9: A play of the hydra game

We begin with a simple hydra with two heads. In Move 1, we chop off the left head. The head is not connected to the root, so we make one new copy of the circled region and connect it to the head's grandparent (which in this case is the root). The hydra we are left with still has two heads. In Move 2, we chop off the left head. Again, this is not connected to the root, so we make two new copies of the circled region and connect them them to the head's grandparent. In Move 3, we chop off the right head (the only one left that is not directly connected to the root). We make three new copies of the circled region and connect them to the head's grandparent. We are then left with a hydra that has seven heads, but each of them is directly connected to the root. We can thus chop them off one at a time, winning the game after seven more moves.

We thus won this round of the hydra game, but maybe that is only because we started with a very simple hydra. Consider the following diagram, taken from the paper by Kirby and Paris in which the hydra game was introduced, depicting the first three moves in a hydra game starting from a more complicated hydra:

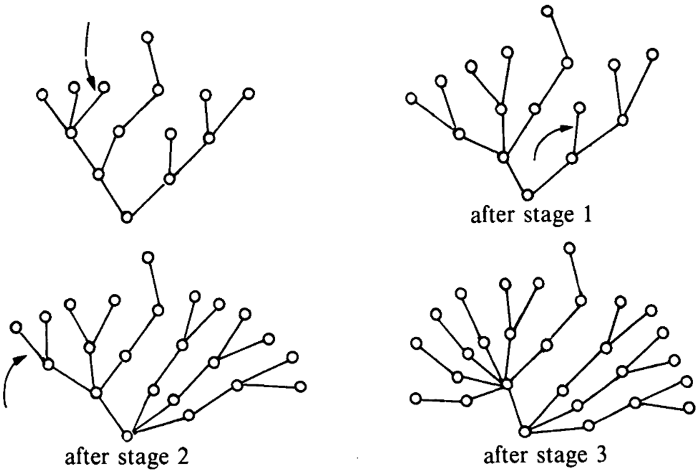


Figure 1.10: The first three moves of a more complicated hydra game

Here, the hydra we end up with after three moves looks, to the untrained eye, to be significantly larger and more complicated than the one we started with, and it seems conceivable that we will never win this hydra game. However, we will prove the following somewhat surprising theorem, showing that not only can we win *every* hydra game, but in fact we *cannot lose.*

Theorem 1.3.1. *In every hydra game, no matter how we play, we will always win after a finite number of moves.*

Proof. We will denote runs of the hydra game by $\langle H_0, H_1, H_2, \ldots \rangle$, where H_0 is the initial hydra that starts the game, H_1 is the hydra resulting after Move 1, H_2 is the hydra resulting after Move 2, and, in general, H_k is the hydra resulting after Move k. If we ever reach an n such that H_n is the hydra consisting of just a root node, then we have won the game, so the complete run of the game is then $\langle H_0, H_1, H_2, \ldots, H_n \rangle$. We must show that every possible run of the hydra game is finite.

To do this, given an arbitrary hydra H, we will assign it an ordinal number, $\#(H)$ in the following way. Starting with the terminal nodes and working our way down to the root, we will assign an ordinal number to each node of the hydra. Each terminal node gets labeled with a 0. Now suppose that u is a non-terminal node of H and we have labeled all of the nodes that are directly above u (i.e., above u and connected to it by an edge). Suppose that there are m such nodes, and they are labeled with ordinal numbers $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_m$ (arranged in non-increasing order). Then label u with the ordinal

$$\omega^{\alpha_1} + \omega^{\alpha_2} + \cdots + \omega^{\alpha_m}.$$

Finally, let $\#(H)$ equal the ordinal number that is assigned to the root of H by this process. Here are a couple of simple examples to illustrate this.

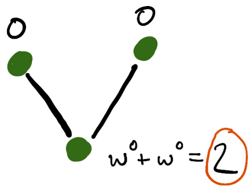


Figure 1.11: A hydra H with $\#(H) = 2$

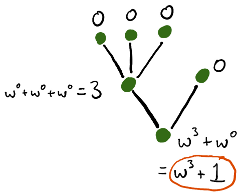


Figure 1.12: A hydra H with $\#(H) = \omega^3 + 1$

If we calculate $\#(H)$ for the first hydra presented above, we find that it is equal to $\omega^{\omega^3+1} + 1$:

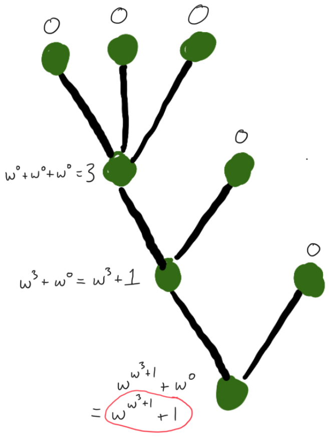


Figure 1.13: A hydra H with $\#(H) = \omega^{\omega^3+1} + 1$.

Now let's see what happens to the ordinal number assigned to this hydra if we make a play of the hydra game and chop off one of its heads. Let's suppose that we are at Move 2 and chop off the left head of this hydra, as depicted in Figure 1.7 above. The resulting hydra H' is depicted in Figure 1.8 above, and we can calculate $\#(H')$ to be $\omega^{\omega^2 \cdot 3+1} + 1$:

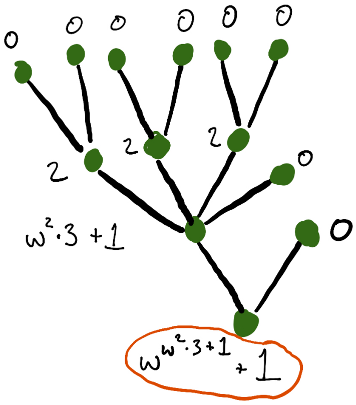


Figure 1.14: A hydra H' with $\#(H') = \omega^{\omega^2 \cdot 3+1} + 1$

Notice that $\omega^2 \cdot 3 < \omega^3$, so $\omega^{\omega^2 \cdot 3+1} + 1 < \omega^{\omega^3+1} + 1$, i.e., $\#(H') < \#(H)$. Thus, by making a move in the hydra game and chopping off a head of H, we created a new hydra H' such that, even though H' has more heads than H, its ordinal value is strictly *smaller*. This is not a coincidence.

Exercise 1.3.2. Calculate the ordinal numbers assigned to the hydras appearing in the run of the hydra game depicted in Figure 1.9 above.

You should have found in the above exercise that the ordinal values assigned to the hydras were strictly decreasing throughout the run of the game. We can in fact prove that this is always the case. The following is the key step of this proof; for now, we leave it as an exercise.

Exercise 1.3.3. Suppose that $\langle H_0, H_1, H_2, H_3, \ldots \rangle$ is a run of the hydra game. Prove that $\#(H_0) > \#(H_1) > \#(H_2) > \#(H_3) > \ldots$. In other words, performing a move in the hydra game always strictly decreases the value of the ordinal assigned to the hydra. (**Hint.** First prove the following basic fact about ordinal arithmetic: for every ordinal α and every natural number n, we have $\omega^\alpha \cdot n < \omega^{\alpha+1}$.)

With the previous exercise, though, we can finish the proof of the theorem! For every run of the hydra game $\langle H_0, H_1, H_2, H_3, \ldots \rangle$, we obtain a strictly decreasing sequence of ordinals $\#(H_0) > \#(H_1) > \#(H_2) > \#(H_3) > \ldots$. Since the ordinals are themselves well-ordered, there can be no infinite strictly decreasing sequences of ordinals. Therefore, *every* run of the hydra game must be finite. In other words, no matter how you play the hydra game, you will always win after some finite number of moves. □

1.4 Peano arithmetic

Our use of infinite ordinals to prove that every hydra game must end after finitely many moves may seem counterintuitive and perhaps unnecessary. The theorem about hydra games is, after all, a statement that is entirely about finite objects. Why should we need to reason about infinite ordinals in order to prove it? However, a remarkable theorem of Kirby and Paris shows that something like this *is* indeed necessary to prove the theorem.

You may have seen before the axioms of Peano Arithmetic (PA), introduced by Giuseppe Peano in the 19th century. These axioms are meant to capture our intuition about the behavior of the natural numbers, and hence about finite discrete objects more broadly.

The language of Peano arithmetic consists of

- the equality sign $=$;
- a constant symbol 0;
- a unary function symbol S.

The intended interpretation of the function S is that it returns the *successor* of its input, i.e., $S(n) = n + 1$. Peano arithmetic has five axioms that are meant to describe the arithmetical properties of the *natural numbers*. These axioms can be stated as follows:

1. 0 is a natural number.
2. For every natural number n, $S(n)$ is also a natural number.
3. For all natural numbers m and n, if $S(m) = S(n)$, then $m = n$.
4. For every natural number n, $0 \neq S(n)$, i.e., 0 is not the successor of any natural number.
5. (Induction) If K is a set such that
   - 0 is in K; and
   - for every natural number n, if n is in K, then $S(n)$ is also in K,

   then K contains every natural number.

Peano Arithmetic captures much of our intuition about the natural numbers, and many theorems about the natural numbers or finite discrete objects can be proven using only PA. For example, much of number theory, as well as the finite Ramsey theorem, can be established in PA. However, Kirby and Paris proved that PA is *not* strong enough to prove that every hydra game must end in a finite number of moves. In fact, they proved the following (stated in a slightly imprecise way):

Theorem 1.4.1 (Kirby–Paris)**.** *If a set of axioms can prove that every hydra game must end in a finite number of moves, then it can also prove the consistency of* PA.

By Gödel's Second Incompleteness Theorem, PA cannot prove its own consistency. Therefore, the Kirby–Paris theorem implies that we cannot prove our theorem about hydra games using PA alone; we must use *something* that goes beyond it.

Chapter 2

Lecture 2: Transfinite induction and recursion

Two of the principal reasons for the centrality of well-orderings in set theory and its applications to other fields of mathematics are the techniques of transfinite induction and transfinite recursion. Let us briefly recall these techniques, in both a formal formulation and a more informal one that better reflects how we actually think about them in practice.

2.1 Transfinite induction

To motivate the statement of transfinite induction, recall classical induction on the natural numbers:

Principle of induction: Suppose that P is a property that can hold of natural numbers and suppse that we know the following:

> For all $n \in \mathbb{N}$, if $P(m)$ holds for all $m < n$, then $P(n)$ holds.

Then $P(n)$ holds for all $n \in \mathbb{N}$.

A similar principle holds for arbitrary well-ordered sets, not just for $\mathbb{N}$.

Theorem 2.1.1 (Transfinite induction)**.** *Suppose that $(X, \preceq)$ is a well-order, (or X is the class of all ordinals, and $\preceq$ is the usual ordering of ordinals) and suppose that P is a property that can hold of elements of X. Suppose moreover that we know the following:*

> *For all $y \in X$, if $P(x)$ holds for all $x \prec y$, then $P(y)$ holds.*

Then $P(y)$ holds for all $y \in X$.

As a simple illustration, let us prove that ordinal addition is associative.

Theorem 2.1.2. *For all ordinals α, β, γ, we have $(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$.*

Proof. The proof is by induction on γ. Thus, fix ordinals α and β. We will prove the following:

> For every ordinal γ, if $(\alpha + \beta) + \varepsilon = \alpha + (\beta + \varepsilon)$ for all $\varepsilon < \gamma$, then $(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$.

Theorem 2.1.1 will them imply that $(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$ for all ordinals γ.

To this end, fix an ordinal γ, and suppose that $(\alpha+\beta)+\varepsilon = \alpha+(\beta+\varepsilon)$ for all $\varepsilon < \gamma$. We must show that $(\alpha+\beta)+\gamma = \alpha+(\beta+\gamma)$. The proof splits into three cases, based on whether $\gamma = 0$, γ is a successor ordinal, or γ is a nonzero limit ordinal.

Case 1: $\gamma = 0$. Recall that, for any ordinal δ, we have $\delta + 0 = \delta$. Thus, we have

$$(\alpha + \beta) + 0 = \alpha + \beta = \alpha + (\beta + 0),$$

as desired.

Case 2: γ is a successor ordinal. Let ε be such that $\gamma = \varepsilon + 1$. Recall that, by definition of ordinal addition, we know that, for all ordinals δ, we have $\delta + (\varepsilon + 1) = (\delta + \varepsilon) + 1$. Then

$$\begin{aligned}(\alpha + \beta) + \gamma &= (\alpha + \beta) + (\varepsilon + 1) \\ &= ((\alpha + \beta) + \varepsilon) + 1 \\ &= (\alpha + (\beta + \varepsilon)) + 1 \\ &= \alpha + ((\beta + \varepsilon) + 1) \\ &= \alpha + (\beta + (\varepsilon + 1)) \\ &= \alpha + (\beta + \gamma),\end{aligned}$$

where the equality between lines 2 and 3 follow from the inductive hypothesis and all other equalities follow from the definition of ordinal addition.

Case 3: γ is a nonzero limit ordinal. In this case recall that, by the definition of ordinal addition, for every ordinal δ,

$$\delta + \gamma = \sup\{\delta + \varepsilon \mid \varepsilon < \gamma\}.$$

Now we have

$$\begin{aligned}(\alpha + \beta) + \gamma &= \sup\{(\alpha + \beta) + \varepsilon \mid \varepsilon < \gamma\} \\ &= \sup\{\alpha + (\beta + \varepsilon) \mid \varepsilon < \gamma\} \\ &= \alpha + \sup\{\beta + \varepsilon \mid \varepsilon < \gamma\} \\ &= \alpha + (\beta + \gamma),\end{aligned}$$

where the equality between lines 1 and 2 follows from the inductive hypothesis and all other equalities follow from the definition of ordinal arithmetic.

This completes all three cases and thus the proof of the theorem. □

Another example of a proof by transfinite induction involves *strictly increasing functions.*

Definition 2.1.3. Suppose that $(A, \leq_A)$ and $(B, \leq_B)$ are two well-orders. Then a function $f : A \to B$ is said to be *strictly increasing* if, for all $x, y \in A$, we have

$$(x <_A y) \Longrightarrow (f(x) <_B f(y)).$$

Exercise 2.1.4. Suppose that $(A, \leq_A)$ is a well-order and $f : A \to A$ is strictly increasing. Prove that $x \leq f(x)$ for all $x \in A$.

2.2 Transfinite recursion

You are probably familiar with the notion of a *sequence* indexed by the natural numbers. For example, the sequence $\langle 1/2^n \mid n < \omega\rangle$ is the sequence $\langle 1, 1/2, 1/4, 1/8, \ldots\rangle$. But we can equally well have sequences indexed by other ordinal numbers.

Definition 2.2.1. Let α be an ordinal. An *α-sequence* is a sequence of the form $\langle x_\eta \mid \eta < \alpha\rangle$, i.e., a sequence that is indexed by the set of ordinals less than η.

Roughly speaking, a construction of a sequence by *recursion* is a construction done by specifying one element at a time, with the choice of a particular element possibly depending on the initial segment of the sequence that has been constructed so far. For example, the ω-sequence $\langle 1/2^n \mid n < \omega\rangle$ can be given a recursive definition as follows:

- $x_0 = 1$;
- for all $n < \omega$, $x_{n+1} = x_n/2$.

In general, it is not hard to see that, if one is given a rule for selecting x_0 and, given x_n, a rule for selecting x_{n+1}, then there is exactly one sequence $\langle x_n \mid n < \omega\rangle$ that satisfies these rules.

Another well-known recursively defined sequence is the *Fibonacci* sequence, $\langle 0, 1, 1, 2, 3, 5, 8, 13, \ldots\rangle$, defined recursively as follows:

- $x_0 = 0$;
- $x_1 = 1$;
- for all $n < \omega$, $x_{n+2} = x_n + x_{n+1}$.

Just as with induction, recursion can be extended to arbitrary well-orders.

Theorem 2.2.2 (Transfinite recursion)**.** *Suppose that α is an ordinal and Z is a nonempty set. Let $\mathcal{S}$ be the set of all sequences of elements of Z of length less than α, and suppose that $F : \mathcal{S} \to Z$ is a function. Then there is a unique sequence $\langle x_\eta \mid \eta < \alpha\rangle$ such that, for all $\xi < \alpha$, we have*

$$x_\xi = F(\langle x_\eta \mid \eta < \xi\rangle).$$

Informally speaking, Theorem 2.2.2 is saying that one can construct sequences by (arbitrarily long) transfinite recursion. In applications of the theorem, one is typically seeking to construct an α-sequence for some ordinal α. The function F in the theorem is describing a rule that tells you how to pick the *next* element of the sequence given what has come so far. The theorem then says that there is exactly one sequence that satisfies all of these rules.

We have in fact already seen recursive constructions; the rigorous definitions of ordinal arithmetic given in Chapter 1 were definitions by transfinite recursion.

In practice, when applying transfinite recursion, we typically want to produce an α-sequence $\langle x_\eta \mid \eta < \alpha\rangle$ of elements of a set Z such that the sequence satisfies certain desired properties. The construction of such a sequence will typically consist of the following two steps:

1. For $\xi < \alpha$, describe a rule for choosing x_ξ based on the sequence $\langle x_\eta \mid \eta < \xi \rangle$ constructed so far. Sometimes this rule will break into cases depending on whether ξ is 0, a successor ordinal, or a nonzero limit ordinal, though sometimes these distinctions will not matter.

2. Show that a sequence constructed according to this rule will have the desired properties.

2.3 Well-ordering principle

Transfinite induction and transfinite recursion gain additional power when paired with the well-ordering principle, which can be stated as follows.

Theorem 2.3.1 (Well-ordering principle)**.** *For every set X, there is a binary relation $\preceq$ on X such that $(X, \preceq)$ is a well-order.*

We have stated the well-ordering principle as a theorem, and it is indeed a theorem of ZFC. Over the axioms of ZF (which are just the axioms of ZFC without the axiom of choice), it turns out that the well-ordering principle is *equivalent* to the axiom of choice, so one could just as well think of it as an alternative formulation of the axiom of choice.

Since the ordinal numbers are themselves well-ordered, we know that, for every *cardinal* number κ, there is a *minimal* ordinal of cardinality κ. In practice, we identify the cardinal κ with this ordinal, which we will also refer to as κ. A key property of this ordinal κ is the following:

> Every proper initial segment of κ has cardinality strictly less than κ.

This is useful enough in practice that it is worth it to state a more refined version of the well-ordering principle.

Theorem 2.3.2 (Well-ordering principle, version 2)**.** *Suppose that X is a set and $\kappa = |X|$. Then there is a sequence $\vec{x} = \langle x_\alpha \mid \alpha < \kappa \rangle$ such that*

- *$X = \{x_\alpha \mid \alpha < \kappa\}$; and*
- *$\vec{x}$ is injective, i.e., for all $\alpha < \beta < \kappa$, we have $x_\alpha \neq x_\beta$.*

2.4 Applications of transfinite recursion to Euclidean space

In this section, we apply the tools of transfinite recursion to construct interesting objects in Euclidean space, focusing in particular on $\mathbb{R}^2$ and $\mathbb{R}^3$. This will involve transfinite recursions of length $|\mathbb{R}|$. We denote the cardinality of $\mathbb{R}$ by $\mathfrak{c}$; sometimes this cardinal is simply referred to as "the continuum". As you may know, the precise value of $\mathfrak{c}$ is not determined by the axioms of ZFC. It could be $\aleph_1$, $\aleph_2$, or more generally any cardinal κ such that the cofinality of κ is uncountable. Note that

$$\mathfrak{c} = |\mathbb{R}| = |\mathbb{R}^2| = |\mathbb{R}^3| = \ldots = |\mathbb{R}^\omega|.$$

Our first example, due to Mazurkiewicz, establishes the existence of a so-called *two-point set.*

Theorem 2.4.1. *There is a subset A of $\mathbb{R}^2$ such that every straight line in $\mathbb{R}^2$ intersects A in exactly two points.*

Proof. Let $\mathcal{L}$ be the set of all lines in $\mathbb{R}^2$. We first claim that $|\mathcal{L}| = \mathfrak{c}$. Here's one way to see that. First, there are *at least* $\mathfrak{c}$-many lines, since, for example, for every real number r, the equation $y = r$ describes a unique horizontal line in $\mathbb{R}^2$. Thus, $|\mathcal{L}| \geq \mathfrak{c}$. On the other hand, to see that $|\mathcal{L}| \leq \mathfrak{c}$, note that, to specify a line in $\mathbb{R}^2$, it suffices to specify two distinct points on the line. There are only $\mathfrak{c} \times \mathfrak{c} = \mathfrak{c}$-many ways of choosing two points in $\mathbb{R}^2$. Thus, $|\mathcal{L}| \leq \mathfrak{c}$.

Using the well-ordering principle, we can fix an injective sequence $\langle \ell_\alpha \mid \alpha < \mathfrak{c}\rangle$ of lines in $\mathbb{R}^2$ such that every element of $\mathcal{L}$ is equal to ℓ_α for some $\alpha < \mathfrak{c}$. We will recursively construct a sequence $\langle A_\alpha \mid \alpha < \mathfrak{c}\rangle$ satisfying the recursion requirements that, for every $\beta < \mathfrak{c}$:

1. A_β is a subset of $\mathbb{R}^2$ of size at most 2;
2. $\bigcup_{\alpha\leq\beta} A_\alpha$ does not contain any three points that lie on the same line;
3. $\bigcup_{\alpha\leq\beta} A_\alpha$ contains exactly two points on the line ℓ_β.

If we can succeed in this construction, then the set $A = \bigcup_{\alpha<\mathfrak{c}} A_\alpha$ will be as required by the theorem. Let us now describe the recursive construction.

Fix an ordinal $\beta < \mathfrak{c}$ and suppose that we have constructed $\langle A_\alpha \mid \alpha < \beta\rangle$ that satisfies the recursion requirements so far. The following describes how to choose a set A_β to continue the construction.

Let $B = \bigcup_{\alpha<\beta} A_\alpha$, i.e., B is the set of points that we have chosen so far. Let $\mathcal{G}$ be the set of all lines passing through two points of B. Note that

$$|\mathcal{G}| \leq |B| \times |B| \leq |\beta| \times |\beta| < \mathfrak{c}.$$

When choosing A_β, we must be careful not to add any new points that are on lines in $\mathcal{G}$ to satisfy requirement (2) above.

Consider the line ℓ_β. We must make sure that $\bigcup_{\alpha\leq\beta} A_\beta$ contains exactly two points on ℓ_β to satisfy requirement (3) above. If B already contains two points from ℓ_β, then we can simply let $A_\beta = \emptyset$ and move on to the next step. If B contains either 0 or 1 points from ℓ_β, then notice that $\ell_\beta \notin \mathcal{G}$, and therefore every line in $\mathcal{G}$ intersects ℓ_β in at most one point. Since $|\mathcal{G}| < \mathfrak{c}$ and the number of points on ℓ_β is exactly $\mathfrak{c}$, we know that $|\ell_\beta \setminus \bigcup \mathcal{G}| = \rfloor$, i.e., there are $\mathfrak{c}$-many points on ℓ_β that are not on any of the lines in $\mathcal{G}$.

If B contains 0 points from ℓ_β, then let A_β consist of precisely 2 points from $\ell_\beta \setminus \bigcup \mathcal{G}$, and if B contains 1 point from ℓ_β, then let A_β consist of precisely 1 point from $\ell_\beta \setminus \bigcup \mathcal{G}$. This completes the description of stage β of the construction; one can check that we have maintained the recursion requirements (1) – (3). Thus, this completes the construction and the proof of the theorem. □

The next results concern "circles" in Euclidean space. These are probably what you intuitively expect them to be. By "circle" we mean the boundary of the circle, not its interior. We consider only nontrivial circles, i.e., circles whose radius is strictly positive. In $\mathbb{R}^2$, one can specify a circle by fixing real numbers x_0 and y_0 and a radius $r > 0$; the circle is then the set of all points $(x, y) \in \mathbb{R}^2$ such that $(x - x_0)^2 + (y - y_0)^2 = r^2$. One can do something similar, but more complicated, in $\mathbb{R}^3$: one can specify a circle by first specifying a plane in $\mathbb{R}^3$ and then specifying a circle in that plane in the same way as we did in $\mathbb{R}^2$. However, for what we want to discuss here, these formal descriptions are not necessary and may even get in the way of intuition. The two basic facts we will need about circles are the following:

Fact 2.4.2. *Suppose that a, b, and c are any three points in $\mathbb{R}^2$ or $\mathbb{R}^3$ that are not all on the same line. Then there is a unique circle that contains all three points.*

Fact 2.4.3. *If C_0 and C_1 are two distinct circles in $\mathbb{R}^2$ or $\mathbb{R}^3$, then C_0 and C_1 intersect in at most two points.*

First, we have an exercise giving a variant on Theorem 2.4.1.

Exercise 2.4.4. There is a subset A of $\mathbb{R}^2$ such that every circle in $\mathbb{R}^2$ intersects A in exactly three points.

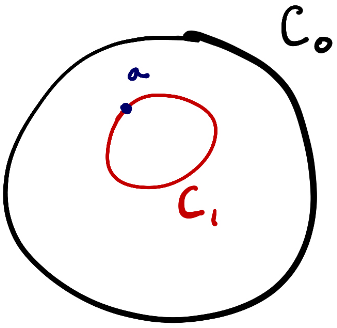
The next example concerns covering Euclidean space by pairwise disjoint circles. Let us say precisely what we mean by this. If $\mathcal{C}$ is a set of circles (in either $\mathbb{R}^2$ or $\mathbb{R}^3$), then we say that $\mathcal{C}$ is *pairwise disjoint* if, for all distinct $C_0, C_1 \in \mathcal{C}$, we have $C_0 \cap C_1 = \emptyset$. In other words, $\mathcal{C}$ is pairwise disjoint if no two distinct element of $\mathcal{C}$ intersect each other.

We say that a pairwise disjoint set $\mathcal{C}$ of circles *covers* $\mathbb{R}^2$ (or $\mathbb{R}^3$) if every element of $\mathbb{R}^2$ (or $\mathbb{R}^3$) is in an element of $\mathcal{C}$. In other words, $\mathcal{C}$ covers $\mathbb{R}^2$ (or $\mathbb{R}^3$) if $\bigcup \mathcal{C} = \mathbb{R}^2$ (or $\bigcup \mathcal{C} = \mathbb{R}^3$). Note that, since $\mathcal{C}$ is pairwise disjoint, if $\mathcal{C}$ covers $\mathbb{R}^2$ or $\mathbb{R}^3$, then every point is in *exactly* one element of $\mathcal{C}$.

We are interested in the question of whether Euclidean spaces can be covered by pairwise disjoint sets of circles and, if they can, what further requirements we can place on these circles. We first show that this is impossible for $\mathbb{R}^2$.

Theorem 2.4.5. $\mathbb{R}^2$ *cannot be covered by a pairwise disjoint set of circles.*

Proof. The key observation about $\mathbb{R}^2$ is the following: every circle in $\mathbb{R}^2$ divides the rest of the plane into a region *inside* the circle and a region outside the circle. If C_0 is a circle and a is a point *inside* of C_0, then any circle C_1 containing a that is disjoint from C_0 must itself lie entirely inside of C_1 (see picture below).



Now suppose for the sake of contradiction that $\mathcal{C}$ is a pairwise disjoint set of circles that covers $\mathbb{R}^2$. Simultaneously recursively define sequences $\langle C_n \mid n < \omega \rangle$ and $\langle a_n \mid n < \omega \rangle$ as follows:

- C_0 is an arbitrary element of $\mathcal{C}$;
- for each $n < \omega$, a_n is the *center* of the circle C_n;
- for each $n < \omega$, C_{n+1} is the unique element of $\mathcal{C}$ that passes through a_n.

In other words, C_0 is an arbitrary element of $\mathcal{C}$, C_1 is the unique element of $\mathcal{C}$ that passes through the center of C_0, C_2 is the unique element of $\mathcal{C}$ that passes through the center of C_1, and so on. By the observation above, for each $n < \omega$, the circle C_{n+1} lies entirely inside of C_n. Moreover, since C_{n+1} passes through the *center* of C_n, the radius of C_{n+1} must be less than half the radius of C_n. For all $n < \omega$, let r_n denote the radius of C_n. We have shown that

$$r_{n+1} < \mathfrak{r}_\mathfrak{n}2$$

for all n, and hence the radii $\langle r_n \mid n < \omega \rangle$ converge to 0. Therefore, the centers $\langle a_n \mid n < \omega \rangle$ of the circles $\langle C_n \mid n < \omega \rangle$ converge to a single limit point; call this limit point b.

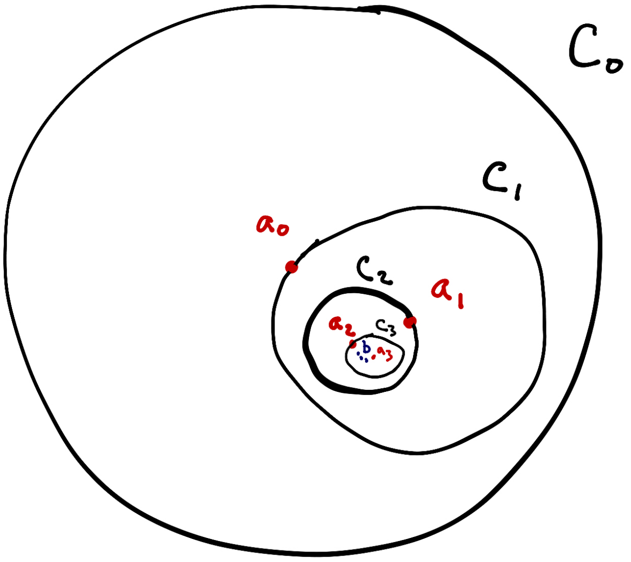


Figure 2.1: The centers $\langle a_n \mid n < \omega \rangle$ of the circles converging to a limit point b.

Notice that b must lie on the *inside* of C_n for every $n < \omega$.

We are assuming that $\mathcal{C}$ covers $\mathbb{R}^2$. We can therefore find a circle $C^* \in \mathcal{C}$ such that $b \in C^*$. Since b is *inside* C_n for all $n < \omega$, we know that C^* cannot be equal to C_n for any $n < \omega$. Let r^* be the radius of C^*. Since the radii $\langle r_n \mid n < \omega \rangle$ converge to 0, we can fix an $n < \omega$ such that $r_n < r^*$. But now we know the following two things:

1. C^* contains a point on the inside of C_n, namely b.

2. The radius of C^*, r^*, is larger than the radius of C_n, r_n. Therefore, C^* cannot be entirely contained in the inside of C_n.

It follows from the two items above that C^* contains points both inside and outside C_n and therefore it must intersect C_n:

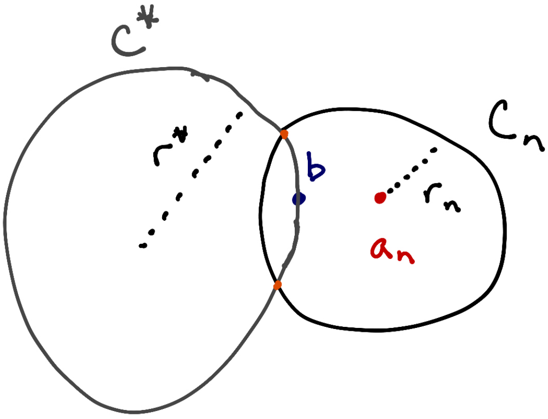


Figure 2.2: C^* contains points both inside and outside C_n, so they must intersect.

However, C^* and C_n are distinct elements of $\mathcal{C}$, which was supposed to be a pairwise disjoint family. This is a contradiction, thus proving the theorem. □

However, perhaps surprising, we now show that $\mathbb{R}^3$ *can* be covered by a pairwise disjoint set of circles. The key difference between $\mathbb{R}^3$ and $\mathbb{R}^2$ with respect to this problem is that, in $\mathbb{R}^3$, unlike in $\mathbb{R}^2$, a circle no longer divides the rest of the space into "inside" and "outside", and this gives us much more freedom to construct interesting sets of circles. For instance, we can have two disjoint circles that are *linked*, like successive rings in a chain.

We will prove, in fact, not only that $\mathbb{R}^3$ can be covered by a pairwise disjoint set of circles, but that all of the circles in this set can be required to have any specified radius (we will construct such a set containing only circles of radius 1).

Theorem 2.4.6. *There is a pairwise disjoint family $\mathcal{C}$ of circles in $\mathbb{R}^3$ such that*

1. *every circle in $\mathcal{C}$ has radius* 1*; and*

2. *$\mathcal{C}$ covers $\mathbb{R}^3$.*

Proof. Let $\langle a_\alpha \mid \alpha < \mathfrak{c} \rangle$ be an injective sequence of points in $\mathbb{R}^3$ such that every point in $\mathbb{R}^3$ is equal to a_α for some $\alpha < \mathfrak{c}$.

We will recursive construct a sequence $\langle C_\alpha \mid \alpha < \mathfrak{c} \rangle$ satisfying the recursion requirements that, for every $\beta < \mathfrak{c}$:

1. C_β is either the empty set or a circle in $\mathbb{R}^3$ of radius 1;

2. for all $\alpha < \beta$, we have $C_\alpha \cap C_\beta = \emptyset$;

3. $a_\beta \in \bigcup_{\alpha \leq \beta} C_\beta$.

If we can succeed in this construction, then the set

$$\mathcal{C} = \{C_\alpha \mid \alpha < \mathfrak{c} \text{ and } C_\alpha \text{ is a circle in } \mathbb{R}^3\}$$

is as required the theorem. Let us now describe the recursive construction.

Fix an ordinal $\beta < \mathfrak{c}$ and suppose that we have constructed $\langle C_\alpha \mid \alpha < \beta \rangle$ that satisfies the recursion requirements so far. The following describes how to choose C_β to continue the construction.

Consider the point a_β. If there is $\alpha < \beta$ such that $a_\beta \in C_\alpha$, then we can simply let $C_\beta = \emptyset$ and move on to the next step. Otherwise, to satisfy requirements (1) and (3), we need to choose C_β to be a circle in $\mathbb{R}^3$ of radius 1 that passes through x_β. To satisfy requirement (2), we must make sure that C_β is disjoint from C_α for all $\alpha < \beta$.

Let $\mathcal{B} = \{C_\alpha \mid \alpha < \beta \text{ and } C_\alpha \neq \emptyset\}$. In other words, $\mathcal{B}$ is the set of all circles chosen so far. Note that, for each $C_\alpha \in \mathcal{B}$, there is a unique plane P_α containing C_α. Moreover, there are precisely $\mathfrak{c}$-many planes that pass through the point a_β. Therefore, since $|\mathcal{B}| \leq |\beta| < \mathfrak{c}$, we can find a plane P^* passing through a_β such that, for every $C_\alpha \in \mathcal{B}$, P^* does not contain C_α.

Now note that, if P is a plane and C is a circle that is not contained in P, then C intersects P in at most two points. Let $Q = \bigcup_{\alpha<\beta} P^* \cap C_\alpha$, i.e., Q is the set of points in P^* that are in C_α for some $\alpha < \beta$. By the previous paragraph, $P^* \cap C_\alpha$ has size at most two for every $\alpha < \beta$. Therefore, we have $|Q| \leq |\beta| \times 2 < \mathfrak{c}$.

Let $\mathcal{D}$ be the set of circles C such that

- C passes through a_β;
- C is contained in P^*;
- C has radius 1.

We would like to choose an element of $\mathcal{D}$ to be our circle C_β. The following is left as an exercise:

Exercise 2.4.7. $|\mathcal{D}| = \mathfrak{c}$.

We need to require that C_β is disjoint from C_α for all $\alpha < \beta$. Since we are going to choose C_β to be contained in P^*, this amounts to choosing C_β so that it is disjoint from Q. Let us call a circle $C \in \mathcal{D}$ *bad* if $C \cap Q \neq \emptyset$. Let $\mathcal{D}^-$ be the set of bad elements of $\mathcal{D}$. For each $C \in \mathcal{D}^-$, choose $q_C \in C \cap Q$.

Our goal is to show that $\mathcal{D} \setminus \mathcal{D}^- \neq \emptyset$. The following is also left as an exercise:

Exercise 2.4.8. If u and v are two distinct points in a plane P, then there are at most two circles that are contained in P, pass through both u and v, and have radius 1.

By the exercise, for each $q \in Q$, there are at most *two* circles $C \in \mathcal{D}^-$ such that $q_C = q$. Since $|Q| < \mathfrak{c}$, it follows that

$$|\mathcal{D}^-| \leq |Q| \times 2 < \mathfrak{c}.$$

Since $|\mathcal{D}| = \mathfrak{c}$, we know that $\mathcal{D}$ contains circles that are not bad, so we can choose C_β to be any element of $\mathcal{D}$ that is not bad. This completes stage β of the construction. By our construction, we know that that:

- C_β is a circle of radius 1 passing through a_β;
- $C_\beta \cap C_\alpha = \emptyset$ for all $\alpha < \beta$,

so we have maintained the recursion requirements. This completes the construction of $\mathcal{C}$ and hence the proof of the theorem. □

Chapter 3

Lecture 3: The axiom of choice and its consequences

The axiom of choice is an incredibly important, useful, and sometimes controversial axiom in set theory. It forms the “C” in “ZFC”, which are the standard axioms of set theory. (The “Z” and “F” stand for “Zermelo” and “Fraenkel”, respectively. “ZF” denotes the axioms of ZFC without the axiom of choice.) In this lecture, we review the axiom of choice and its equivalent formulations and present a couple basic applications thereof.

3.1 The axiom of choice

Roughly speaking, the axiom of choice asserts that, given any collection of nonempty sets, one can form a new collection by choosing one element from each set. More formally, it can be formulated as follows.

Definition 3.1.1 (Axiom of choice)**.** The *axiom of choice* is the following assertion: Whenever I is a set and $\langle X_i \mid i \in I\rangle$ is such that each X_i is a nonempty set, there is a sequence $\langle y_i \mid i \in I\rangle$ such that, for all $i \in I$, we have $y_i \in X_i$.

The axiom of choice can also be phrased in terms of functions:

> Whenever I is a set and F is a function with domain I such that $F(i)$ is a nonempty set for all $i \in I$, there is a function g with domain I such that $g(i) \in F(i)$ for all $i \in I$.

As we shall see, the axiom of choice is incredibly powerful and is essential to proving a number of important theorems. It can also lead to some counterintuitive consequences, which has led to it being somewhat controversial. Let us note now, though, that, due to a theorem of Gödel, there is no cost in consistency strength in assuming the axiom of choice.

Theorem 3.1.2 (Gödel)**.** *If the axioms of* ZF *are consistent, then so are the axioms of* ZFC*.*

(We refer the reader to any standard set theory textbook for a precise statement of the axioms of ZFC.)

There are a number of statements that are equivalent to the axiom of choice over ZF; we mention here two especially important ones. In fact, we have already seen one of them: the well-ordering principle, which states that every set can be well-ordered:

Definition 3.1.3 (Well-ordering principle)**.** The *well-ordering principle* is the following assertion: Whenever X is a set, there is a binary relation $\preceq$ on X such that $(X, \preceq)$ is a well-order.

The second important statement that is equivalent to the axiom of choice is known as *Zorn's lemma.* To state it properly, we first need a preliminary definition.

Definition 3.1.4. Suppose that $(X, \leq)$ is a partial order (recall Definition 1.1.1).

1. A subset $K \subseteq X$ is called a *chain* if K is linearly ordered by $\leq$, i.e., for all $x, y \in K$, either $x \leq y$ or $y \leq x$.
2. If $K \subseteq X$ and $z \in X$, then we say that z is an *upper bound* for K if $x \leq z$ for all $x \in K$.

We can now state Zorn's lemma.

Definition 3.1.5 (Zorn's lemma)**.** Zorn's lemma is the following assertion: Suppose that $(X, \leq)$ is a nonempty partial order such that every chain $K \subseteq X$ has an upper bound. Then X contains a *maximal* element, i.e., there is $y \in X$ such that, for all $z \in X$, we have $y \not< z$.

The following theorem establishes the equivalence of these three important statements. Recall that, if T is a set of axioms and φ and ψ are two statements in the language of T, then we say that φ and ψ are *equivalent* over T if one can prove ψ from $\mathsf{T} \cup \{\varphi\}$ *and* one can prove φ from $\mathsf{T} \cup \{\psi\}$.

We leave its proof as an exercise.

Theorem 3.1.6. *Over the axioms of* ZF*, the following are equivalent:*

1. *the axiom of choice;*
2. *the well-ordering principle;*
3. *Zorn's lemma.*

We say some applications of the well-ordering principle in the previous lecture. We now present two further applications, one using the axiom of choice directly, and the other using Zorn's lemma. Both are related to the general mathematical problem of measuring the *size* of mathematical objects.

3.2 A non-measurable set of real numbers

This section concerns the general task of measuring the size of sets of real numbers. One could of course measure each set of real numbers by its cardinality, but this does not really match with our geometric intuition about $\mathbb{R}$. For example, we already have some intuition about what the measures of certain very simple sets of real numbers should be. For instance, it is natural to measure *intervals* of real numbers by their length:

Definition 3.2.1. If $a \leq b$ are real numbers, then

- the *closed interval* $[a, b]$ is the set $\{x \in \mathbb{R} \mid a \leq x \leq b\}$; and
- the *open interval* (a, b) is the set $\{x \in \mathbb{R} \mid a < x < b\}$.

If I equals either $[a, b]$ or (a, b), then we say that I is an *interval* with endpoints a and b. The *length* of $[a, b]$ (or (a, b)) is denoted $\ell([a, b])$ and is equal to $b - a$.

One can also define intervals with endpoints at $\pm\infty$. For example, if $b \in \mathbb{R}$, then $(-\infty, b] = \{x \in \mathbb{R} \mid x \leq b\}$, or $(b, \infty) = \{x \in \mathbb{R} \mid b < x\}$. The length of such intervals is defined to be ∞.

One can straightforwardly extend this method of measuring sets of real numbers to other simple sets, such as disjoint unions of intervals. For example, it makes sense to say that the measure of the set $[0, 1] \cup (2, 2.5)$ should be $1 + 0.5 = 1.5$. Or, suppose that $\langle I_n \mid n \in \mathbb{N} \rangle$ is a sequence of pairwise disjoint intervals such that, for all $n \in \mathbb{N}$, we have $\ell(I_n) = 1/2^n$. Then it would make sense to say that the measure of the union $\bigcup_{n \in \mathbb{N}} I_n$ should be

$$\sum_{n=0}^{\infty} 1/2^n = 1 + 1/2 + 1/4 + 1/8 + \ldots = 2.$$

It is natural now to ask whether this method of measurement can be extended to measure the size of *all* subsets of $\mathbb{R}$. Of course, in order for this question to make sense, we must ask that this notion of measure satisfies certain nice properties that we would expect to hold of functions that measure the size of sets of real numbers. We extract these properties in the following definition.

Definition 3.2.2. Suppose that $m : \mathscr{P}(\mathbb{R}) \to [0, \infty]$ (i.e., m assigns to every subset $X \subseteq \mathbb{R}$ a measure $m(X)$ that is either a non-negative real number or ∞). We say that m is a *nice measure* if it satisfies the following properties:

1. If $a \leq b$ are real numbers and I is an interval with endpoints a and b, then $m(I) = b - a$.
2. (Monotonicity) If $X \subseteq Y$ are subsets of $\mathbb{R}$, then $m(X) \leq m(Y)$.
3. (Translation invariance) If $X \subseteq \mathbb{R}$ and $a \in \mathbb{R}$, then $m(X) = m(a + X)$, where $a + X$ denotes the set $\{a + x \mid x \in X\}$. In other words, the measure of a set X should not change if we simply shift it horizontally on the number line.

4. (Countable additivity) If $\mathcal{F}$ is a finite or countably infinite collection of subsets of $\mathbb{R}$, then
$$m\left(\bigcup \mathcal{F}\right) = \sum_{X \in \mathcal{F}} m(X).$$
In other words, if $\langle X_n \mid n \in \mathbb{N}\rangle$ is a sequence of pairwise disjoint subsets of $\mathbb{R}$ and $X = \bigcup_{n \in \mathbb{N}} X_n$, then
$$m(X) = \sum_{n \in \mathbb{N}} m(X_n) = m(X_0) + m(X_1) + m(X_2) + \dots.$$

Perhaps surprisingly, we will now show that, assuming the axioms of ZFC, there are no nice measures. In other words, it is *impossible* to extend the notion of length to measure *all* subsets of $\mathbb{R}$ in a way that comports with our intuitions about how measures of size should behave.

Theorem 3.2.3 (Vitali)**.** *There are no nice measures.*

Proof. Suppose for the sake of contradiction that $m : \mathscr{P}(\mathbb{R}) \to [0, \infty]$ is a nice measure. Recall that $\mathbb{Q}$ denotes the set of *rational* numbers. $\mathbb{Q}$ is a countably infinite set, and it is also dense in $\mathbb{R}$, meaning that every nonempty open interval (a, b) contains an element of $\mathbb{Q}$. Given a real number $a \in \mathbb{R}$, let $a + \mathbb{Q}$ denote the set $\{a + q \mid q \in \mathbb{Q}\}$.

Let $\mathcal{F}$ denote the set $\{a + \mathbb{Q} \mid a \in \mathbb{R}\}$. Notice that there will be distinct real numbers $a \neq b$ such that $a + \mathbb{Q} = b + \mathbb{Q}$; in fact, this will happen if and only if the difference $b - a$ is rational. Since $\mathcal{F}$ is a *set*, if $a \neq b$ and $a + \mathbb{Q} = b + \mathbb{Q}$, then $\mathcal{F}$ does not somehow contain *separate copies* $a + \mathbb{Q}$ and $b + \mathbb{Q}$. Rather, it contains one set that equals both $a + \mathbb{Q}$ and $b + \mathbb{Q}$. Moreover, $\mathcal{F}$ consists of *pairwise disjoint* sets. The verifications of these facts form the following exercise.

Exercise 3.2.4. Suppose that $a, b \in \mathbb{R}$. Prove the following.

1. $(a + \mathbb{Q}) = (b + \mathbb{Q})$ if and only if $b - a \in \mathbb{Q}$.
2. If $(a + \mathbb{Q}) \neq (b + \mathbb{Q})$, then $(a + \mathbb{Q}) \cap (b + \mathbb{Q}) = \emptyset$.

Lemma 3.2.5. *For every $a \in \mathbb{R}$, we have $(a + \mathbb{Q}) \cap [0, 1] \neq \emptyset$.*

Proof. Fix $a \in \mathbb{R}$. Since $\mathbb{Q}$ is dense in $\mathbb{R}$, we can fix a rational number q in the interval $(-a, -a + 1)$. Then $a + q \in a + \mathbb{Q}$, and we have the following:

- $a + q \geq a + (-a) = 0$;
- $a + q \leq a + (-a + 1) = 1$.

Therefore, $a + q \in (a + \mathbb{Q}) \cap [0, 1]$. □

We can therefore apply the axiom of choice to the family of pairwise disjoint nonempty sets
$$\{(a + \mathbb{Q}) \cap [0, 1] \mid a \in \mathbb{R}\}$$
to find a set $X \subseteq [0, 1]$ that contains exactly one point from each element of $\mathcal{F}$. In other words, for each $a \in \mathbb{R}$, there is exactly one element in the intersection $X \cap (a + \mathbb{Q})$.

We will eventually reach a contradiction using the fact that our measure m must assign a value to the set X. We first need a couple of lemmas.

Lemma 3.2.6. *If p and q are two distinct rational numbers, then $(p+X)\cap(q+X)=\emptyset$.*

Proof. Suppose for the sake of contradiction that p and q are distinct rational numbers and $(p+X)\cap(q+X)\neq\emptyset$. Fix a number $a\in(p+X)\cap(q+X)$. Then there are $x_p, x_q\in X$ such that $a=p+x_p=q+x_q$. Since $p\neq q$, we must have $x_p\neq x_q$. But rearranging the equation in the previous sentence yields

$$x_p - x_q = q - p.$$

Since $q-p$ is rational, Exercise 3.2.4 implies that $x_p+\mathbb{Q}=x_p+\mathbb{Q}$. In other words, x_p and x_q are both elements of $x_p+\mathbb{Q}$. This contradicts the fact that $x_p, x_q\in X$ and X contains only one element of $x_p+\mathbb{Q}$. □

Let $C=\mathbb{Q}\cap[-1,1]$ be the set of all rational numbers between -1 and 1, and let

$$U=\bigcup_{q\in C}(q+X).$$

Lemma 3.2.7. $[0,1]\subseteq U\subseteq[-1,2]$.

Proof. We first show that $[0,1]\subseteq U$. Fix $a\in[0,1]$, and find $b\in X\cap(a+\mathbb{Q})$. Then $q=a-b$ is a rational number. Moroever, since a and b are both in the interval $[0,1]$, we must have $q\in[-1,1]$. But then $a=q+b\in q+X$, and $q+X\subseteq U$, so $a\in U$.

We next show that $U\subseteq[-1,2]$. Fix $a\in U$. Then there is a rational number $q\in[-1,1]$ and a $b\in X$ such that $a=q+b$. Since $b\in[0,1]$, it follows that $a\in[0-1,1+1]=[-1,2]$. □

We are now ready to reach our contradiction. By Lemma 3.2.7 and properties (1) and (2) of Definition 3.2.2, we know that

$$1\leq m(U)\leq 3.$$

Moreover, by Lemma 3.2.6, we know that the family $\{q+X \mid q\in C\}$ is pairwise disjoint. Also, by definition of U, we have $U=\bigcup_{q\in C}(q+X)$. Thus, by property (4) of Definition 3.2.2, we have

$$m(U)=\sum_{q\in C}m(q+X).$$

By property (3) of Definition 3.2.2, we know that $m(q+X)=m(X)$ for all $q\in C$, so

$$m(U)=\sum_{q\in C}m(X).$$

If $m(X)=0$, then this yields

$$m(U)=\sum_{q\in C}0=0+0+0+\ldots=0,$$

contradicting the fact that $m(U)\geq 1$. On the other hand, if $m(X)>0$, then we have

$$m(U)=\sum_{q\in C}m(X)=m(X)+m(X)+m(X)+\ldots=\infty,$$

contradicting the fact that $m(U) \leq 3$. In either case, we reach a contradiction, therefore completing the proof of the theorem. □

We end this section by noting that some use of the axiom of choice really is necessary in the proof of Theorem 3.2.3, due to the following theorem of Solovay. In the statement of the theorem, an *inaccessible cardinal* is a (relatively small) example of a *large cardinal* (formally, an inaccessible cardinal is an uncountable, regular, strong limit cardinal). It is not necessary to understand precisely what it is; we only emphasize that the theory

$$\text{“}\mathsf{ZFC} + \text{ there exists an inaccessible cardinal”}$$

is considered to be a relatively mild extension of ZFC.

Theorem 3.2.8 (Solovay)**.** *Suppose that the theory*

$$\text{“}\mathsf{ZFC} + \textit{ there exists an inaccessible cardinal”}$$

is consistent. Then so is the theory “ZF + *there exists a nice measure*”.

3.3 Nonprincipal ultrafilters

Recall that, given a set X, the *power set* of X, denoted $\mathscr{P}(X)$, is defined to be the set of all subsets of X, i.e.,

$$\mathscr{P}(X) = \{Z \mid Z \subseteq X\}.$$

Recall also that, if X is a set and $Y \subseteq X$, then $X \setminus Y$ is called the *complement of* Y *in* X, and

$$X \setminus Y = \{x \in X \mid x \notin Y\}.$$

In particular, $X \setminus Y$ satisfies:

- $Y \cup (X \setminus Y) = X$;
- $Y \cap (X \setminus Y) = \emptyset$.

Given a nonempty set X, a *filter* over X can be thought of as a way of specifying what it means to be a “large” subset of X. Formally, it is defined as follows.

Definition 3.3.1. Suppose that X is a nonempty set. A *filter* over X is a set $\mathcal{F} \subseteq \mathscr{P}(X)$ with the following properties:

1. $X \in \mathcal{F}$ and $\emptyset \notin \mathcal{F}$;
2. for all $Y, Z \in \mathscr{P}(X)$, if $Y \subseteq Z$ and $Y \in \mathcal{F}$, then $Z \in \mathcal{F}$;
3. for all $Y, Z \in \mathscr{P}(X)$, if $Y \in \mathcal{F}$ and $Z \in \mathcal{F}$, then $Y \cap Z \in \mathcal{F}$.

Requirements (1)–(3) in Definition 3.3.1 should make intuitive sense if you think of elements of a filter $\mathcal{F}$ over X as being “large” subsets of X. Namely:

1. Requirement (1) says that the entire set X is large and the empty set is not large.

2. Requirement (2) says that, if Y is large and $Z \supseteq Y$, then Z should also be large.

3. Requirement (3) says that if Y and Z are both large, then their intersection $Y \cap Z$ is large.

Exercise 3.3.2. Suppose that X is an infinite set, and let

$$\mathcal{F} = \{Y \subseteq X \mid X \setminus Y \text{ is finite}\}.$$

In other words, $\mathcal{F}$ consists of all subsets of X that contain all but finitely many elements of X. Prove that $\mathcal{F}$ is a filter over X. This filter is called the *cofinite filter over* X, or sometimes the *Fréchet filter over* X.

Note that, if $\mathcal{F}$ is a filter over a set X and $Y \subseteq X$, then it cannot be the case that both Y and $X \setminus Y$ are in $\mathcal{F}$: if it were the case, then requirement (3) of Definition 3.3.1 would imply that $Y \cap (X \setminus Y) \in \mathcal{F}$, i.e., $\emptyset \in \mathcal{F}$, contradicting requirement (1) of Definition 3.3.1. Thus, $\mathcal{F}$ can contain at most one of Y and $X \setminus Y$. If $\mathcal{F}$ contains precisely one of these sets for *every* $Y \subseteq X$, then we call it an *ultrafilter*.

Definition 3.3.3. Suppose that X is a nonempty set. A set $\mathcal{U} \subseteq \mathscr{P}(X)$ is called an *ultrafilter* over X if

- $\mathcal{U}$ is a filter over X;
- for all $Y \in \mathscr{P}(X)$, either $Y \in \mathcal{U}$ or $X \setminus Y \in \mathcal{U}$.

It is easy to describe certain ultrafilteres:

Exercise 3.3.4. Suppose that X is a nonempty set and $x \in X$. Let

$$\mathcal{U} = \{Y \in \mathscr{P}(X) \mid x \in Y\}.$$

Prove that $\mathcal{U}$ is an ultrafilter over X.

Ultrafilters as in Exercise 3.3.4 are called *principal ultrafilters*. More formally:

Definition 3.3.5. Suppose that X is a nonempty set. Then a set $\mathcal{U} \subseteq \mathscr{P}(X)$ is called a *principal ultrafilter over* X if there is $x \in X$ such that

$$\mathcal{U} = \{Y \in \mathscr{P}(X) \mid x \in Y\}.$$

If $\mathcal{U}$ is an ultrafilter over X and $\mathcal{U}$ is *not* a principal ultrafilter over X, then we call it a *nonprincipal ultrafilter over* X.

We have seen that principal ultrafilters exist, so the question naturally arises whether *nonprincipal* ultrafilters exist. It turns out that, if X is a *finite* set, then every ultrafilter over X is finite:

Exercise 3.3.6. Suppose that X is a finite nonempty set and $\mathcal{U}$ is an ultrafilter over X. Prove that $\mathcal{U}$ is a principal ultrafilter over X.

In fact, we have the following:

Exercise 3.3.7. Suppose that X is a set and $\mathcal{U}$ is an ultrafilter over $\mathcal{U}$. Then the following are equivalent:

1. $\mathcal{U}$ is a principal ultrafilter;
2. there is a finite set $Y \subseteq X$ such that $Y \in \mathcal{U}$.

However, if X is infinite, then, assuming the axiom of choice holds, we can prove that nonprincipal ultrafilters over X exist. The following theorem, which we prove using the help of Zorn's lemma, is the key.

Theorem 3.3.8. *Suppose that X is a nonempty set and $\mathcal{F}$ is a filter over X. Then there is an ultrafilter $\mathcal{U}$ over X such that $\mathcal{F} \subseteq \mathcal{U}$.*

Proof. Let P be the set of all filters $\mathcal{G}$ over X such that $\mathcal{F} \subseteq \mathcal{G}$. Note that P is nonempty, since we certainly have $\mathcal{F} \in P$. Consider the binary relation $\subseteq$ on $\mathcal{F}$.

Lemma 3.3.9. *$(P, \subseteq)$ is a partial order.*

Proof. We need to verify all three requirements in Definition 1.1.1. They all follow almost immediately from the definition of the subset relation:

- (Reflexive): For all $\mathcal{G} \in P$, we certainly have $\mathcal{G} \subseteq \mathcal{G}$.
- (Transitive): For all $\mathcal{G}_0, \mathcal{G}_1, \mathcal{G}_2 \in P$, if $\mathcal{G}_0 \subseteq \mathcal{G}_1$ and $\mathcal{G}_1 \subseteq \mathcal{G}_2$, then clearly $\mathcal{G}_0 \subseteq \mathcal{G}_2$.
- (Anti-symmetric): For all $\mathcal{G}, \mathcal{H} \in P$, if $\mathcal{G} \subseteq \mathcal{H}$ and $\mathcal{H} \subseteq \mathcal{G}$, then $\mathcal{G} = \mathcal{H}$.

□

We want to apply Zorn's lemma to the partial order $(P, \subseteq)$. We first need to verify that every $\subseteq$-chain in P has an upper bound. This will follow from the next lemma.

Lemma 3.3.10. *Suppose that $K \subseteq P$ is nonempty and linearly ordered by $\subseteq$ (i.e., for all $\mathcal{G}, \mathcal{H} \in K$, either $\mathcal{G} \subseteq \mathcal{H}$ or $\mathcal{H} \subseteq G$. Then*

$$\bigcup K = \{Y \in \mathscr{P}(X) \mid \exists \mathcal{G} \in K \; [X \in \mathcal{G}]\}$$

is an upper bound for K in $(P, \subseteq)$.

Proof. Let $\mathcal{H} = \bigcup K$. It is clear from the definition that $\mathcal{G} \subseteq \mathcal{H}$ for every $\mathcal{G} \in K$, and therefore also $\mathcal{F} \subseteq \mathcal{H}$. We therefore only need to show that $\mathcal{H}$ is a filter over X. We verify requirements (1)–(3) of Definition 3.3.1.

1. For all $\mathcal{G} \in K$, we have $X \in \mathcal{G}$. Therefore, $X \in \mathcal{H}$. Similarly, for all $\mathcal{G} \in K$, we have $\emptyset \notin \mathcal{G}$. Therefore, $\emptyset \notin \mathcal{H}$.
2. Suppose that $Y, Z \in \mathscr{P}(X)$, $Y \subseteq Z$, and $Y \in \mathcal{H}$. Then there is $\mathcal{G} \in K$ such that $Y \in \mathcal{G}$. Since $\mathcal{G}$ is a filter, it follows that $Z \in \mathcal{G}$. By definition of $\mathcal{H}$, we then have $Z \in \mathcal{H}$.

3. Suppose that $Y, Z \in \mathscr{P}(X)$, $Y \in \mathcal{H}$, and $Z \in \mathcal{H}$. Then there are $\mathcal{G}, \mathcal{G}' \in K$ such that $Y \in \mathcal{G}$ and $Z \in \mathcal{G}'$. Since K is linearly ordered, either $\mathcal{G} \subseteq \mathcal{G}'$ or $\mathcal{G}' \subseteq \mathcal{G}$. Without loss of generality, assume that $\mathcal{G} \subseteq \mathcal{G}'$ (the other case is symmetric). Then, since $Y \in \mathcal{G}$, we have $Y \in \mathcal{G}'$. By assumption, $Z \in \mathcal{G}'$, so, since $\mathcal{G}'$ is a filter, we have $Y \cap Z \in \mathcal{G}'$.

Therefore, $\mathcal{H}$ is indeed a filter over X, completing the proof of the lemma. □

We have now shown that $(P, \subseteq)$ satisfies the hypotheses of Zorn's lemma. Apply Zorn's lemma to find a *maximal* filter $\mathcal{G} \in P$, i.e., an element $\mathcal{G} \in P$ such that there does not exist $\mathcal{H} \in P$ such that $\mathcal{G} \subsetneq \mathcal{H}$.

We claim that $\mathcal{G}$ is an ultrafilter. To show this, suppose for the sake of contradiction that it is *not* an ultrafilter. Then there is a set $Y \in \mathscr{P}(X)$ such that $Y \notin \mathcal{G}$ and $X \setminus Y \notin \mathcal{G}$. Note that $Y \neq \emptyset$, since $X \setminus \emptyset = X$ certainly is in $\mathcal{G}$. We claim that we can find an ultrafilter $\mathcal{H}$ in P such that $\mathcal{G} \cup \{Y\} \subseteq \mathcal{H}$, which will contradict the maximality of $\mathcal{G}$.

To see this, let

$$\mathcal{H} = \mathcal{G} \cup \{Z \cap Y \mid Z \in \mathcal{G}\}.$$

Clearly, $\mathcal{G} \subseteq \mathcal{H}$. Moreover, $Y \in \mathcal{H}$, since $X \in \mathcal{G}$ and $X \cap Y = Y$. Therefore, we will be done if we show that $\mathcal{H}$ is a filter over X. We again verify requirements (1)–(3) of Definition 3.3.1.

1. Since $X \in \mathcal{G}$, we have $X \in \mathcal{H}$. To see that $\emptyset \notin \mathcal{H}$, suppose for the sake of contradiction that $\emptyset \in H$. We know that $\emptyset \notin \mathcal{G}$, so there must be $Z \in \mathcal{G}$ such that $Z \cap Y = \emptyset$. But this means that $Z \subseteq (X \setminus Y)$ (Exercise: prove this!). Therefore, since $\mathcal{G}$ is a filter, we must have $(X \setminus Y) \in \mathcal{G}$. But we chose Y so that $(X \setminus Y) \notin \mathcal{G}$, which yields a contradiction. Thus, $\emptyset \notin \mathcal{G}$.

2. Suppose that $W, W' \in \mathscr{P}(X)$, $W \subseteq W'$, and $W \in \mathcal{H}$. If $\mathcal{W} \in \mathcal{G}$, then, since $\mathcal{G}$ is a filter, it follows that $W' \in \mathcal{G}$ and hence $W' \in \mathcal{H}$. Otherwise, there is $Z \in \mathcal{G}$ such that $W = Z \cap Y$. Let $Z' = W' \cup (X \setminus Y)$.

   Claim 3.3.11. *$Z \subseteq Z'$ and $Z' \cap Y = W'$*

   Proof. To show that $Z \subseteq Z'$, fix $z \in Z$. If $z \in Y$, then $z \in W \subseteq W'$, so $z \in Z'$. If $z \notin Y$, then $z \in (X \setminus Y)$, so $z \in Z'$.

   To show that $Z' \cap Y = W'$, first fix $z \in Z' \cap Y$. Then $z \in Z'$ and $z \notin (X \setminus Y)$. By the definition of Z', we must have $z \in W'$. The other direction is immediate: by the definition of Z', we have $W' \subseteq Z'$. □

   Since $Z \in \mathcal{G}$ and $\mathcal{G}$ is a filter, it follows that $Z' \in \mathcal{G}$. But then, by the definition of $\mathcal{H}$, we have $Z' \cap Y = W' \in \mathcal{H}$.

3. Suppose that $W, W' \in \mathscr{P}(X)$, $W \in \mathcal{H}$, and $W' \in \mathcal{H}$. We want to show that $W \cap W' \in \mathcal{H}$. There are three cases to consider.

   - If both W and W' are in $\mathcal{G}$, then, since $\mathcal{G}$ is a filter, we have $W \cap W' \in \mathcal{G} \subseteq \mathcal{H}$.

- If exactly one of W and W' is in $\mathcal{G}$, suppose without loss of generality that $W \in \mathcal{G}$ and $W' \notin \mathcal{G}$ (the other case is symmetric). Then there is $Z' \in \mathcal{G}$ such that $W' = Z' \cap Y$. Then

  $$W \cap W' = W \cap Z' \cap Y = (W \cap Z') \cap Y.$$

  Since $\mathcal{G}$ is a filter, we have $W \cap Z' \in \mathcal{G}$. Thus, by the definition of $\mathcal{H}$, we have $(W \cap Z') \cap Y \in \mathcal{H}$, i.e., $W \cap W' \in \mathcal{H}$.

- If neither of W nor W' is in $\mathcal{G}$, then there are $Z, Z' \in \mathcal{G}$ such that $W = Z \cap Y$ and $W' = Z' \cap Y$. Then

  $$W \cap W' = (Z \cap Y) \cap (Z' \cap Y) = (Z \cap Z') \cap Y.$$

  Since $\mathcal{G}$ is a filter, we have $(Z \cap Z') \in \mathcal{G}$. Then, as in the previous case, we have $(Z \cap Z') \cap Y \in \mathcal{H}$, i.e., $W \cap W' \in \mathcal{H}$.

This completes the verification that $\mathcal{H}$ is a filter, and hence $\mathcal{H}$ witnesses that $\mathcal{G}$ is not a maximal element of P, since $\mathcal{G} \subsetneq \mathcal{H}$. This is a contradiction; therefore, $\mathcal{G}$ is indeed an ultrafilter. □

Corollary 3.3.12. *For every infinite set X, there is a nonprincipal ultrafilter over X.*

Proof. Let $\mathcal{F} = \{Y \in \mathscr{P}X \mid X \setminus Y \text{ is finite}\}$ be the cofinite filter over X (see Exercise 3.3.2). By Theorem 3.3.8, we can find an ultrafilter $\mathcal{U}$ over X such that $\mathcal{F} \subseteq \mathcal{U}$.

We claim that $\mathcal{U}$ is a nonprincipal ultrafilter. To see this, suppose for the sake of contradiction that $\mathcal{U}$ is a principal ultrafilter. Then there is $x \in X$ such that

$$\mathcal{U} = \{Y \in \mathscr{P}(X) \mid x \in Y\}.$$

In particular, we have $\{x\} \in \mathcal{U}$. But $X \setminus \{x\} \in \mathcal{F}$, since $X \setminus (X \setminus \{x\}) = \{x\}$ is finite. Thus, $X \setminus \{x\} \in \mathcal{U}$ and $\{x\} \in \mathcal{U}$, so

$$(X \setminus \{x\}) \cap \{x\} = \emptyset \in \mathcal{U},$$

contradicting the fact that $\mathcal{U}$ is a filter over X. Thus, $\mathcal{U}$ is a nonprincipal ultrafilter. □

Just as with the nonexistence of a nice measure, the existence of nonprincipal ultrafilters really does require some use of the axiom of choice.

Theorem 3.3.13 (Feferman)**.** *If* ZFC *is consistent, then so is*

$$\text{“}\mathsf{ZF} + \textit{ there is no nonprincipal ultrafilter over } \mathbb{N}\text{”}.$$