ZORN'S LEMMA AND SOME APPLICATIONS

KEITH CONRAD

1. Introduction

Zorn's lemma is a result in set theory that appears in proofs of some non-constructive existence theorems throughout mathematics. We will state Zorn's lemma below and use it in later sections to prove some results in linear algebra, ring theory, and group theory. In an appendix, we will give an application to metric spaces. The statement of Zorn's lemma is not intuitive, and some of the terminology in it may be unfamiliar, but after reading through the explanation of Zorn's lemma and then the proofs that use it you should be more comfortable with how it can be applied.

Theorem 1.1 (Zorn's lemma)**.** *Let S be a partially ordered set. If every totally ordered subset of S has an upper bound, then S contains a maximal element.*

To understand Theorem 1.1, we need to know four terms: partially ordered set, totally ordered subset, upper bound, and maximal element.

A *partial ordering* on a (nonempty) set S is a binary relation on S, denoted $\leq$, which satisfies the following properties:

- for all $s \in S$, $s \leq s$,
- if $s \leq s'$ and $s' \leq s$ then $s = s'$,
- if $s \leq s'$ and $s' \leq s''$ then $s \leq s''$.

When we fix a partial ordering $\leq$ on S, we refer to S (or, more precisely, to the pair $(S, \leq)$) as a partially ordered set.

It is important to notice that we do not assume all pairs of elements in S are comparable under $\leq$: for some s and s' we may have neither $s \leq s'$ nor $s' \leq s$. If all pairs of elements can be compared (that is, for all s and s' in S either $s \leq s'$ or $s' \leq s$) then we say S is *totally ordered* with respect to $\leq$.

Example 1.2. The usual ordering relation $\leq$ on $\mathbf{R}$ or on $\mathbf{Z}^+$ is a partial ordering of these sets. In fact it is a total ordering on either set. This ordering on $\mathbf{Z}^+$ is the basis for proofs by induction.

Example 1.3. On $\mathbf{Z}^+$, declare $a \leq b$ if $a|b$. This partial ordering on $\mathbf{Z}^+$ is different from the one in Example 1.2 and is called ordering by *divisibility*. It is one of the central relations in number theory. (Proofs about $\mathbf{Z}^+$ in number theory sometimes work not by induction, but by starting on primes, then extending to prime powers, and then extending to all positive integers using prime factorization. Such proofs view $\mathbf{Z}^+$ through the divisibility relation rather than through the usual ordering relation.) Unlike the ordering on $\mathbf{Z}^+$ in Example 1.2, $\mathbf{Z}^+$ is not totally ordered by divisibility: most pairs of integers are not comparable under the divisibility relation. For instance, 3 doesn't divide 5 and 5 doesn't divide 3. The subset $\{1, 2, 4, 8, 16, \dots\}$ of powers of 2 is totally ordered under divisibility.

Example 1.4. Let S be the set of all subgroups of a given group G. For $H, K \in S$ (that is, H and K are subgroups of G), declare $H \leq K$ if H is a subset of K. This is a partial ordering, called ordering by *inclusion*. It is not a total ordering: for most subgroups H and K neither $H \subset K$ nor $K \subset H$.

One can similarly partially order the subspaces of a vector space or the ideals (or subrings or all subsets) of a commutative ring by inclusion.

Example 1.5. On $\mathbf{Z}^+$, declare $a \leq b$ if $b|a$. Here one positive integer is "larger" than another if it is a factor. This is called ordering by *reverse divisibility*.

Example 1.6. On the set of subgroups of a group G, declare subgroups H and K to satisfy $H \leq K$ if $K \subset H$. This is a partial ordering on the subgroups of G, called ordering by *reverse inclusion*.

In case you think ordering by reverse inclusion seems weird, let's take a look again at Example 1.3. There positive integers are ordered by divisibility, and nothing seems "backwards." But let's associate to each $a \in \mathbf{Z}^+$ the subgroup $a\mathbf{Z}$ of $\mathbf{Z}$. Every nonzero subgroup of $\mathbf{Z}$ has the form $a\mathbf{Z}$ for a unique positive integer a, $a\mathbf{Z} = b\mathbf{Z}$ if and only if $a = b$ (both a and b are positive), and $a|b$ if and only if $b\mathbf{Z} \subset a\mathbf{Z}$. For instance, $4|12$ and $12\mathbf{Z} \subset 4\mathbf{Z}$. Therefore the ordering by divisibility on $\mathbf{Z}^+$ is essentially the same as ordering by reverse inclusion on nonzero subgroups of $\mathbf{Z}$. Partial ordering by reverse inclusion is used in the construction of completions of groups and rings.

Example 1.7. Let A and B be sets. Let S be the set of functions defined on some subset of A with values in B. The subset can vary with the function. That is, S is the set of pairs (X, f) where $X \subset A$ and $f\colon X \to B$. Two elements (X, f) and (Y, g) in S are equal when $X = Y$ and $f(x) = g(x)$ for all $x \in X$.

We can partially order S by declaring $(X, f) \leq (Y, g)$ when $X \subset Y$ and $g|_X = f$. This means g is an extension of f to a larger subset of A. Let's check the second property of a partial ordering: if $(X, f) \leq (Y, g)$ and $(Y, g) \leq (X, f)$ then $X \subset Y$ and $Y \subset X$, so $X = Y$. Then the condition $g|_X = f$ means $g = f$ as functions on their common domain, so $(X, f) = (Y, g)$.

Example 1.8. If S is a partially ordered set for the relation $\leq$ and $T \subset S$, then the relation $\leq$ provides a partial ordering on T. Thus T is a new partially ordered set under $\leq$. For instance, the partial ordering by inclusion on the subgroups of a group restricts to a partial ordering on the cyclic subgroups of a group.

In these examples, only Example 1.2 is totally ordered. This is typical: most naturally occurring partial orderings are not total orderings. However (and this is important) a partially ordered set can have many subsets that are totally ordered. As a dumb example, every one-element subset of a partially ordered set is totally ordered. A more interesting illustration was at the end of Example 1.3 with the powers of 2 inside $\mathbf{Z}^+$ under divisibility. As another example, if we partially order the subspaces of a vector space V by inclusion then any tower of subspaces

$$W_1 \subset W_2 \subset W_3 \subset \cdots$$

where each subspace is a proper subset of the next one is a totally ordered subset of V.

Here is a result about totally ordered subsets that will be useful at a few points later.

Lemma 1.9. *Let S be a partially ordered set. If $\{s_1, \ldots, s_n\}$ is a finite totally ordered subset of S then there is an s_i such that $s_j \leq s_i$ for all $j = 1, \ldots, n$.*

Proof. The s_i's are all comparable to each other; that's what being totally ordered means. Since we're dealing with a finite set of pairwise comparable elements, there will be one that is greater than or equal to them all in the partial ordering on S. The reader can formalize this with a proof by induction on n, or think about the bubble sort algorithm □

An *upper bound* on a subset T of a partially ordered set S is an $s \in S$ such that $t \leq s$ for all $t \in T$. When we say T has an upper bound in S, we do *not* assume the upper bound is in T itself; it is just in S.

Example 1.10. In $\mathbf{R}$ with its natural ordering, the subset $\mathbf{Z}$ has no upper bound while the subset of negative real numbers has the upper bound 0 (or any positive real). No upper bound on the negative real numbers is a negative real number.

Example 1.11. In the proper subgroups of $\mathbf{Z}$ ordered by inclusion, an upper bound on $\{4\mathbf{Z}, 6\mathbf{Z}, 8\mathbf{Z}\}$ is $2\mathbf{Z}$ since $4\mathbf{Z}$, $6\mathbf{Z}$, and $8\mathbf{Z}$ all consist entirely of even numbers. (Note $4\mathbf{Z} \subset 2\mathbf{Z}$, *not* $2\mathbf{Z} \subset 4\mathbf{Z}$.)

A *maximal* element m of a partially ordered set S is an element that is not below any element to which it is comparable: for all $s \in S$ to which m is comparable, $s \leq m$. Equivalently, m is maximal when the only $s \in S$ satisfying $m \leq s$ is $s = m$. This does *not* mean $s \leq m$ for all s in S since we don't insist that maximal elements are actually comparable to every element of S. A partially ordered set could have many maximal elements.

Example 1.12. If we partially order $\mathbf{Z}^+$ by reverse divisibility (so $a \leq b$ means $b|a$), the number 1 is a maximal element. In fact 1 is the only maximal element. This is not a good example because 1 is comparable to everything in this relation, which is not a typical feature of maximal elements.

Example 1.13. Consider the positive integers *greater than* 1 with the reverse divisibility ordering: $a \leq b$ when $b|a$. The maximal elements here are the positive integers with no positive factor greater than 1 except themselves. These are the prime numbers, so the primes are the maximal elements for the reverse divisibility relation on $\{2, 3, 4, 5, 6, \dots\}$.

Equivalently, if we partially order the *proper* subgroups of $\mathbf{Z}$ by inclusion then the maximal elements are $p\mathbf{Z}$ for prime numbers p.

We now return to the statement of Zorn's lemma:

> If every totally ordered subset of a partially ordered set S has an upper bound, then S contains a maximal element.

All the terms being used here have now been defined.[1] Of course this doesn't mean the statement should be any clearer!

Zorn's lemma is not intuitive, but it turns out to be logically equivalent to more readily appreciated statements from set theory like the Axiom of Choice (which says the Cartesian product of any family of nonempty sets is nonempty). In the set theory appendix to [13], Zorn's lemma is derived from the Axiom of Choice. A proof of the equivalence between Zorn's lemma and the Axiom of Choice is given in the appendix to [16]. The reason for calling Zorn's lemma a lemma rather than an axiom is purely historical. Zorn's lemma is

[1]The hypotheses refer to *all* totally ordered subsets, and a totally ordered subset might be uncountable. Therefore it is a bad idea to write about "totally ordered sequences," since the label "sequence" is often understood to refer to a countably indexed set. Just use the label "totally ordered subset."

also equivalent to the Well-Ordering Principle (which says every nonempty set has a well-ordering: that means a total ordering in which every nonempty subset has a least element), but do *not* confuse the totally ordered subsets in the hypotheses of Zorn's lemma with well-orderings on the whole set. They are different concepts, and you should never invoke the Well-Ordering Principle in the middle of an application of Zorn's lemma unless you really want to make bad mistakes.

Zorn's lemma provides no mechanism to find a maximal element whose existence it asserts. It also says nothing about how many maximal elements there are. Usually, as in Example 1.13, there are many maximal elements.

In a partially ordered set S we can speak about minimal elements just as much as maximal elements: $m \in S$ is called minimal if $m \leq s$ for all $s \in S$ to which m is comparable. Zorn's lemma can be stated in terms of minimal elements: if any totally ordered subset of a partially ordered set S has a lower bound in S then S has a minimal element. There really is no need to use this formulation, in practice, since by reversing the meaning of the partial ordering (that is, using the reverse ordering) lower bounds become upper bounds and minimal elements become maximal elements.

The applications we will make of Zorn's lemma are to algebra, but it shows up in many other areas. For instance, the most important result in functional analysis is the Hahn-Banach theorem, whose proof uses Zorn's lemma. Another result from functional analysis, the Krein-Milman theorem, is proved using Zorn's lemma. (The Krein-Milman theorem is an example where Zorn's lemma is used to prove the existence of something that is more naturally a minimal element than a maximal element.) In topology, the most important theorem about compact spaces is Tychonoff's theorem, and it is proved using Zorn's lemma.

When dealing with objects that have a built-in finiteness condition (such as finite-dimensional vector spaces or finite products of spaces $X_1 \times \cdots \times X_n$), Zorn's lemma can be avoided by using ordinary induction in a suitable way (*e.g.*, inducting on the dimension of a vector space). The essential uses of Zorn's lemma are for truly infinite objects, where one has to make infinitely many choices at once in a rather extreme way.

2. Applications to ideals

The ideals in a commutative ring can be partially ordered by inclusion. The whole ring, which is the unit ideal (1), is obviously maximal for this ordering. But this is boring and useless. Proper ideals that are maximal for inclusion among the proper ideals are called the maximal ideals in the ring. (That is, a maximal ideal is understood to mean a maximal proper ideal.) Let's prove they always exist.

Theorem 2.1. *Every nonzero commutative ring contains a maximal ideal.*

Proof. Let S be the set of proper ideals in a commutative ring $R \neq 0$. Since the zero ideal (0) is a proper ideal, $S \neq \emptyset$. We partially order S by inclusion.

Let $\{I_\alpha\}_{\alpha \in A}$ be a totally ordered set of proper ideals in R. To write down an upper bound for these ideals in S, it is natural to try their union $I = \bigcup_{\alpha \in A} I_\alpha$. As a set, I certainly contains all the I_α's, but is I an ideal? We may be hesitant about this, since a union of ideals is *not* usually an ideal: try $2\mathbf{Z} \cup 3\mathbf{Z}$. But we are dealing with a union of a totally ordered set of ideals, and the total ordering of the ideals will be handy!

If x and y are in I then $x \in I_\alpha$ and $y \in I_\beta$ for two of the ideals I_α and I_β. Since this set of ideals is totally ordered, $I_\alpha \subset I_\beta$ or $I_\beta \subset I_\alpha$. Without loss of generality, $I_\alpha \subset I_\beta$.

Therefore x and y are in I_β, so $x \pm y \in I_\beta \subset I$. Hence I is an additive subgroup of R. The reader can check $rx \in I$ for $r \in R$ and $x \in I$, so I is an ideal in R.

Because I contains every I_α, I is an upper bound on the totally ordered subset $\{I_\alpha\}_{\alpha \in A}$ *provided* it is actually in S: is I a proper ideal? Well, if I is not a proper ideal then $1 \in I$. Since I is the union of the I_α's, we must have $1 \in I_\alpha$ for some α, but then I_α is not a proper ideal. That is a contradiction, so $1 \notin I$. Thus $I \in S$ and we have shown every totally ordered subset of S has an upper bound in S.

By Zorn's lemma S contains a maximal element. This maximal element is a proper ideal of R that is maximal for inclusion among all proper ideals (not properly contained in any other proper ideal of R). That means it is a maximal ideal of R. □

Corollary 2.2. *Every proper ideal in a nonzero commutative ring is contained in a maximal ideal.*

Proof. Let R be the ring and I be a proper ideal in R. The quotient ring R/I is nonzero, so it contains a maximal ideal by Theorem 2.1. The inverse image of this ideal under the natural reduction map $R \to R/I$ is a maximal ideal of R that contains I. □

It is crucial in the proof of Theorem 2.1 to have the multiplicative identity 1 available, which lies in no proper ideal. For instance, the analogue of Theorem 2.1 for groups can fail: the additive group $\mathbf{Q}$ contains no maximal proper subgroups. Indeed, if H is a proper subgroup of $\mathbf{Q}$ then $[\mathbf{Q} : H] = \infty$ (if the index were finite, say n, then $H \supset n\mathbf{Q} = \mathbf{Q}$, so $H = \mathbf{Q}$) and if we pick $r \in \mathbf{Q} - H$ then the subgroup $H + \mathbf{Z}r$ properly contains H with finite index (why?), so $H + \mathbf{Z}r \neq \mathbf{Q}$. (See the exercise below for an instance where groups do contain maximal proper subgroups.) The importance of Theorem 2.1 in the foundations of commutative algebra is one reason that rings should always have a multiplicative identity, at least if you are interested in areas of math that depend on commutative algebra (*e.g.*, number theory and algebraic geometry).

Exercise. If G is a nontrivial *finitely generated* group, use Zorn's lemma and Lemma 1.9 to prove G contains a maximal proper subgroup, and more generally if H is a proper subgroup of G then there is a maximal proper subgroup M of G such that $H \subset M \subset G$. (The step analogous to showing I is proper in the proof of Theorem 2.1 will need the finite generatedness hypothesis on G.)

The proof of Theorem 2.1 exhibits a standard *disconnect* between upper bounds on totally ordered subsets and maximal elements in the whole set. Consider proper ideals of $\mathbf{Z}$ under inclusion. The maximal ideals are $p\mathbf{Z}$ for prime numbers p. The ideals $\{6\mathbf{Z}, 12\mathbf{Z}, 24\mathbf{Z}\}$ are totally ordered under inclusion, and in the proof of Theorem 2.1 the upper bound created on this subset is the union $6\mathbf{Z} \cup 12\mathbf{Z} \cup 24\mathbf{Z} = 6\mathbf{Z}$. (A finite totally ordered subset of a partially ordered set always has one of its members as an upper bound on the subset, by Lemma 1.9.) This upper bound is *not* a maximal ideal in $\mathbf{Z}$. So the task of checking Zorn's lemma can be applied is a completely separate matter from applying Zorn's lemma: an upper bound on a totally ordered subset does not have to be a maximal element of the whole set. Remember that!

The following very important theorem concerns nilpotent elements. In a commutative ring, an element r is *nilpotent* if $r^n = 0$ for some $n \geq 1$.

Theorem 2.3. *The intersection of all prime ideals in a commutative ring is the set of nilpotent elements in the ring.*

This is striking: it tells us that if we know an element in a commutative ring lies in every prime ideal, some power of it must be 0. When $R = \mathbf{Z}$ this result is obvious, since the prime ideals are (0) and (p) for prime numbers p, and the intersection is obviously $\{0\}$, which is the only nilpotent integer. As a somewhat more interesting (and finite) example, let $R = \mathbf{Z}/(12)$. Its prime ideals are $(2)/(12)$ and $(3)/(12)$ (not $(0)/(12)$!), whose intersection is $(6)/(12) = \{0, 6 \bmod 12\}$, which is also (by inspection) the nilpotent elements of $\mathbf{Z}/(12)$.

Proof. Let R be the ring. Pick a nilpotent element r and a prime ideal P. We have $r^n = 0$ for some $n \geq 1$, so $r^n \in P$. Since P is prime we must have $r \in P$. Thus every nilpoent element is in the intersection of all prime ideals.

Now we want to show the intersection of all prime ideals consists only of nilpotent elements: if $r \in P$ for all prime ideals P then $r^n = 0$ for some $n \geq 1$. That is kind of amazing. How could it be shown? The right thing to do is *not* to attempt to prove $r^n = 0$ for some n directly, but rather to prove the contrapositive statement: any non-nilpotent element does not lie in some prime ideal (so anything that is in all prime ideals has to be nilpotent).

Pick a non-nilpotent element r. Since r is not nilpotent, $r^n \neq 0$ for every $n \geq 1$. Consider the set S of all ideals I in R that don't contain any positive power of r:

$$I \in S \Longleftrightarrow \{r^n : n \geq 1\} \cap I = \emptyset.$$

The zero ideal (0) doesn't meet $\{r^n : n \geq 1\}$, because r is not nilpotent, so S is nonempty. We partially order S by inclusion. After checking the conditions for Zorn's lemma can be applied to S, we will show that a maximal element of S is a prime ideal.[2] Since none of the ideals in S contain r, we will have found a prime ideal of R not containing r.

Let $\{I_\alpha\}_{\alpha \in A}$ be a totally ordered set of ideals in S. Their union I is an ideal (same proof as that in Theorem 2.1). Since no I_α contains a positive power of r, their union I does not contain any positive power of r either. Therefore $I \in S$. As $I_\alpha \subset I$ for all α, I is an upper bound in S for the set of I_α's. Thus every totally ordered subset of S contains an upper bound in S.

By Zorn's lemma there is a maximal element of S. This is an ideal P that does not contain any positive power of r and is maximal for this property (with respect to inclusion). We denoted it as P because we're going to show P is a prime ideal. The ideal P is obviously proper. Suppose x and y are in R and $xy \in P$. To prove $x \in P$ or $y \in P$, assume otherwise. Then the ideals $(x) + P$ and $(y) + P$ are both strictly larger than P, so they can't lie in S. That means we have $r^m \in (x) + P$ and $r^n \in (y) + P$ for some positive integers m and n. Write

$$r^m = ax + p_1, \quad r^n = by + p_2$$

where p_1 and p_2 are in P and a and b are in R. Now multiply:

$$r^{m+n} = abxy + axp_2 + byp_1 + p_1p_2.$$

Since P is an ideal and $xy \in P$, the right side is in P. But then $r^{m+n} \in P$, which contradicts P being disjoint from the positive powers of r (because $P \in S$)[3]. Hence $x \in P$ or $y \in P$, so P is prime. By construction P contains no positive power of r, so in particular $r \notin P$. □

[2]Notice we are *not* defining S to be a set of prime ideals, but only a set of ideals with a disjointness property. That any maximal element in S is a prime ideal is going to be proved using maximality.

[3]If we had defined S to be the ideals in R that don't include r, rather than no positive power of r, then at this point we'd be stuck: we'd have $r = ax + p_1$ and $r = bx + p_2$, so then multiplying gives $r^2 \in P$, but that would not be a contradiction since we didn't require ideals of S to avoid containing r^2. Seeing this difficulty is a motivation for the definition of S in the proof.

Remark 2.4. Although P in the proof is maximal with respect to inclusion among all the ideals disjoint from $\{r, r^2, r^3, \dots\}$, there is no reason to expect the ideal P is actually a maximal ideal in R: the name "maximal ideal" means an ideal that is maximal with respect to inclusion among *all* proper ideals, while the proof we just gave used Zorn's lemma on a set of ideals that might not include all proper ideals. Quite generally, if S is a partially ordered set and S' is a subset of S, a maximal element of S' need not be a maximal element of S. Make sure you understand that!

In the proof of Theorem 2.3, we showed that if $r \in R$ is not nilpotent then an ideal that is maximal for the property of not containing $\{r^n : n \geq 1\}$ is prime. The important algebraic property of $\{r^n : n \geq 1\}$ is that it is closed under multiplication and does not include 0. As a check on your understanding, use Zorn's lemma to show that for any subset of a ring R that is closed under multiplication and does not include 0, there is an ideal in R that is maximal for the property of being disjoint from the subset and it is a prime ideal.

Corollary 2.5. *For any ideal I in a commutative ring R,*

$$\bigcap_{P \supset I} P = \{x \in R : x^n \in I \text{ for some } n \geq 1\},$$

where the intersection runs over the prime ideals of R that contain I.

Proof. If $x^n \in I$ for some $n \geq 1$, then for any prime ideal $P \supset I$ we have $x^n \in P$, so $x \in P$. Conversely, suppose x is in every prime ideal of R containing I. The natural map $R \to R/I$ identifies the prime ideals in R that contain I with the prime ideals of R/I, so $\overline{x}$ is in every prime ideal of R/I. Therefore by Theorem 2.3, $\overline{x}$ is nilpotent in R/I. This means $\overline{x}^n = \overline{0}$ for some $n \geq 1$, so $x^n \in I$. □

We used Zorn's lemma to prove Theorem 2.3 since the point of this handout is to hit you over the head with enough applications of Zorn's lemma that the basic principle behind its use becomes transparent, but it turns out that Theorem 2.3 can also be proved using Corollary 2.2 about proper ideals lying in a maximal ideal. The trick is to use a maximal ideal not in R itself but in $R[X]$, as follows.

Proof. We will only address one direction: a non-nilpotent element r in a ring R lies outside some prime ideal. (The other direction is easy; see the first paragraph in the first proof of Theorem 2.3.) We will create such a prime ideal as the kernel of a homomorphism out of R and r won't be in the kernel.

In the polynomial ring $R[X]$, $rX - 1$ does not have a multiplicative inverse. Indeed, if we did have $(rX - 1)(c_nX^n + \cdots + c_1X + c_0) = 1$ then equating coefficients of like powers of X on both sides shows

$$-c_0 = 1, \quad c_0r - c_1 = 0, \quad c_1r - c_2 = 0, \ \dots, \quad c_{n-1}r - c_n = 0, \quad rc_n = 0.$$

Therefore $c_0 = -1$, $c_i = c_{i-1}r$ for $1 \leq i \leq n$, and $rc_n = 0$. So $c_i = -r^i$ for $1 \leq i \leq n$. Therefore $0 = rc_n = -r^{n+1}$, but r is not nilpotent. So $rX - 1$ is not a unit in $R[X]$. (If r *were* nilpotent, say $r^n = 0$, then $rX - 1$ is a unit in $R[X]$ with inverse $-(1 + rX + r^2X^2 + \cdots + r^{n-1}X^{n-1})$.)

The ideal $(rX - 1)$ in $R[X]$ is proper, since $rX - 1$ is not a unit, so this ideal lies inside a maximal ideal M of $R[X]$ by Corollary 2.2. Now consider the composite homomorphism $R \to R[X] \to R[X]/M$, where the first map is inclusion and the second is reduction. Since the target is a field, the kernel back in R is a prime ideal (because the quotient of R by the

kernel embeds into a field and thus must be an integral domain, as subrings of fields are integral domains). Call the kernel P. Since $rX \equiv 1 \bmod M$, r is not in the kernel. Thus P is a prime ideal in R not containing r. We're done. □

We now leave maximal ideals and nilpotent elements, turning our attention to an interesting theorem of I. S. Cohen about finitely generated ideals.

Theorem 2.6 (I.S. Cohen)**.** *If every prime ideal in a commutative ring is finitely generated then every ideal in the ring is finitely generated.*

Proof. We will prove the contrapositive: if there is some ideal in the ring that is not finitely generated then there is a prime ideal in the ring that is not finitely generated. We will find this prime ideal as an ideal maximal with respect to inclusion for the property of not being finitely generated. (Here again we should stress, as in Remark 2.4, that such prime ideals need not be actual maximal ideals in the ring. They are only created as being maximal among non-finitely generated ideals.)

Let S be the collection of all non-finitely generated ideals, so $S \neq \emptyset$ by *assumption*. We partially order S by inclusion. For any totally ordered subset of ideals $\{I_\alpha\}_{\alpha \in A}$ in S, their union I is an ideal containing each I_α. (That I is an ideal follows by the same argument as in the proof of Theorem 2.1.) To know I is an upper bound on the I_α's in S we have to show I is not finitely generated. Well, if I were finitely generated, say $I = (r_1, \ldots, r_k)$, then each of the generators is in some I_α and by total ordering on the ideals, these hypothetical finitely many generators of I are all in a common I_α (Lemma 1.9). But then I lies inside that I_α, so I equals that I_α, which shows that I_α is finitely generated. This is a contradiction, so I is not finitely generated.

Now we can apply Zorn's lemma: there exists a non-finitely generated ideal that is maximal with respect to inclusion among the non-finitely generated ideals. Call such an ideal P. We are going to prove P is a prime ideal. It is certainly a proper ideal. Suppose $xy \in P$ with $x \notin P$ and $y \notin P$. Then $(x) + P$ is an ideal properly containing P, so $(x) + P \notin S$. Therefore this ideal is finitely generated:

$$(x) + P = (r_1, \ldots, r_k).$$

Write $r_i = c_i x + p_i$ for $i = 1, 2, \ldots, k$, where $c_i \in R$ and $p_i \in P$. (We have no right to expect the p_i's generate P. They only occur in the expressions for the r_i's.) Then every r_i is in the ideal $(x, p_1, \ldots, p_k)$, and conversely each p_i is in the ideal $P \subset (x) + P = (r_1, \ldots, r_k)$, so

$$(x) + P = (x, p_1, \ldots, p_k).$$

If we are given any $p \in P$, then since $P \subset (x) + P$ we can write

$$p = cx + a_1 p_1 + \cdots + a_k p_k \tag{2.1}$$

with c and the a_i's all in R. Then $cx = p - \sum a_i p_i \in P$, so c lies in the ideal $J = \{r \in R : rx \in P\}$. Obviously $P \subset J$. Since $xy \in P$ and $y \notin P$, J contains y and therefore J strictly contains P. Thus by maximality of P among all the non-finitely generated ideals in R, J is finitely generated. By (2.1), $p \in xJ + \sum_{i=1}^k Rp_i$, so $P \subset xJ + \sum_{i=1}^k Rp_i$. The reverse inclusion is easy (by the definition of J), so

$$P = xJ + \sum_{i=1}^{k} Rp_i.$$

Since J is finitely generated, this shows P is finitely generated, a contradiction. Hence $x \in P$ or $y \in P$, so P is a prime ideal. □

Exercise. Use Zorn's lemma to prove an analogue of Cohen's theorem for principal ideals: if every prime ideal in a commutative ring is principal than all ideals are principal. (It is false that if every prime ideal has at most 2 generators then all ideals have at most 2 generators, *e.g.*, in $\mathbf{C}[X, Y]$ the prime ideals have 1 or 2 generators but the ideal (X^2, XY, Y^2) can't be generated by 2 elements.)

We now generalize Cohen's theorem to modules in place of rings. When M is an R-module and $\mathfrak{a}$ is an ideal of R, let $\mathfrak{a}M$ denote the submodule of M that is spanned by all finite products am for $a \in \mathfrak{a}$ and $m \in M$. That is, $\mathfrak{a}M$ is the set of all finite sums $\sum_{i=1}^n a_i m_i$ with $n \geq 1$, $a_i \in \mathfrak{a}$ and $m_i \in M$. Check this is a submodule of M. (We need to use finite sums since the set of products am with $a \in \mathfrak{a}$ and $m \in M$ is usually not a submodule since it's not additively closed.)

Theorem 2.7. *Let M be a finitely generated R-module. If every submodule of the form $\mathfrak{p}M$ with prime $\mathfrak{p}$ is finitely generated then every submodule of M is finitely generated.*

When $M = R$, this is Cohen's theorem. However, the proof of Theorem 2.7 has some additional aspects at the end that don't occur in Cohen's theorem, so it seems best to prove Theorem 2.7 separately from Theorem 2.6.

Proof. We will prove the contrapositive: if M has a submodule that is not finitely generated then it has a submodule of the form $\mathfrak{p}M$ with prime $\mathfrak{p}$ that is not finitely generated.

Let S be the set of submodules of M that are not finitely generated, so $S \neq \emptyset$ by *assumption*. Note $M \notin S$. Partially order S by inclusion. By the same kind of argument as in the proof of Theorem 2.6, every totally ordered subset of S has an upper bound in S. (Check the details!) Therefore we can apply Zorn's lemma: S contains a maximal element. Call one of them N. That is, N is a submodule of M that is not finitely generated and (this is the key point) any submodule of M that properly contains N is finitely generated. Note $N \neq M$.

Will we show $N = \mathfrak{p}M$ for some prime ideal $\mathfrak{p}$ of R? No! There is so little control over the maximal elements coming from Zorn's lemma that we can't expect this. Instead we will show, following [14], that

(1) $\mathfrak{p} := \mathrm{Ann}_R(M/N) = \{r \in R : rM \subset N\}$ is a prime ideal of R,
(2) $\mathfrak{p}M$ is not finitely generated.

To show $\mathfrak{p}$ is prime, first we note $\mathfrak{p} \neq R$ since $N \subsetneq M$. If $\mathfrak{p}$ is not prime then there are x and y in R with $xy \in \mathfrak{p}$ but x and y are not in $\mathfrak{p}$. From the definition of $\mathfrak{p}$, these conditions on x and y mean

$$xyM \subset N, \quad xM \not\subset N, \quad yM \not\subset N.$$

Thus $xM + N$ properly contains N, so $xM + N$ is finitely generated. Let a finite spanning set of $xM + N$ be $xm_i + n_i$ $(i = 1, \ldots, k)$. (warning: $n_1, \ldots, n_k$ do not span N, as N is not finitely generated.) Then

$$xM + N = \sum_{i=1}^k Rxm_i + \sum_{i=1}^k Rn_i.$$

(Just check a spanning set of the module on each side is in the other side.) For any $n \in N \subset xM + N$,

$$\begin{aligned} n &= r_1xm_1 + \cdots + r_kxm_k + r_1'n_1 + \cdots + r_k'n_k \\ &= x(r_1m_1 + \cdots + r_km_k) + r_1'n_1 + \cdots + r_k'n_k \end{aligned} \tag{2.2}$$

where $r_i, r_i' \in R$. Thus $x(r_1m_1 + \cdots + r_km_k) \in N$, so $r_1m_1 + \cdots + r_km_k$ lies in

$$L := \{m \in M : xM \subset N\},$$

which is a submodule of M. Note

$$N \subset L, \quad yM \subset L, \quad yM \not\subset N.$$

Therefore N is a proper subset of L, so L is finitely generated. By (2.2),

$$n \in xL + \sum_{i=1}^{k} Rn_i,$$

so

$$N \subset xL + \sum_{i=1}^{k} Rn_i.$$

The reverse inclusion is straightforward (use the definition of L), so

$$N = xL + \sum_{i=1}^{k} Rn_i.$$

The right side is finitely generated, which is a contradiction since N is not finitely generated. Thus $\mathfrak{p}$ is a prime ideal in R.

(At this point, if we had been taking $M = R$ as in Cohen's theorem, then N would be an ideal and $\mathrm{Ann}_R(M/N) = \mathrm{Ann}_R(R/N)$ is equal to N, so $N = \mathfrak{p}$ is prime and we would have finished proving Cohen's theorem.)

It remains to show $\mathfrak{p}M$ is not finitely generated. What we will do is show $N = \mathfrak{p}M + Q$ for some finitely generated R-module Q. Then, since N is not finitely generated, $\mathfrak{p}M$ can't be finitely generated either.

Since, by hypothesis, M is finitely generated, write $M = Re_1 + \cdots + Re_\ell$. (We are not assuming the e_i's are linearly independent in any sense, just a spanning set.) Then M/N is spanned over R by the reductions $\overline{e}_1, \ldots, \overline{e}_k$, so

$$\mathfrak{p} = \mathrm{Ann}_R(M/N) = \bigcap_{i=1}^{\ell} \mathrm{Ann}_R(R\overline{e}_i).$$

The ideal $\mathfrak{p}$ is inside each $\mathrm{Ann}_R(R\overline{e}_i)$, but in fact it must equal one of these: if not then each $\mathrm{Ann}_R(R\overline{e}_i)$ contains an element r_i outside $\mathfrak{p}$, but then the product of those r_i's (over all i) is an element outside of $\mathfrak{p}$ (because $\mathfrak{p}$ is prime) while at the same time the product of the r_i's kills each $\overline{e}_i$, so this product is in each annihilator and hence is in $\mathfrak{p}$. This is absurd, so $\mathfrak{p}$ is the annihilator of some $R\overline{e}_i$. Without loss of generality, $\mathfrak{p} = \mathrm{Ann}_R(R\overline{e}_1) = \{r \in R : re_1 \subset N\}$.

Since $\mathfrak{p} \neq R$, also $\overline{e}_1 \neq 0$ in M/N, and thus $e_1 \notin N$, so $Re_1 \not\subset N$. Therefore $Re_1 + N$ properly contains N so it is finitely generated, say by $r_je_1 + n_j$ $(j = 1, \ldots, d)$. For $n \in N \subset$

$Re_1 + N$, write

$$n = \sum_{j=1}^{d} a_j(r_j e_1 + n_j) = \left(\sum_{j=1}^{d} a_j r_j\right) e_1 + \sum_{j=1}^{d} a_j n_j,$$

with $a_j \in R$. The coefficient of e_1 scales e_1 into N by this equation, so the coefficient of e_1 is in $\mathfrak{p}$. Thus

$$N \subset \mathfrak{p}e_1 + \sum_{j=1}^{d} Rn_j.$$

The reverse inclusion is easy, so

$$N = \mathfrak{p}e_1 + \sum_{j=1}^{d} Rn_j \subset \mathfrak{p}M + \sum_{j=1}^{d} Rn_j \subset N + N = N.$$

Thus $N = \mathfrak{p}M + \sum_{j=1}^{d} Rn_j$, so N not being finitely generated forces $\mathfrak{p}M$ not to be finitely generated, which is what we wanted to show. □

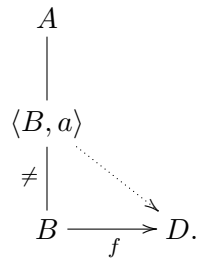
3. Application to divisible groups

An abelian group D is called *divisible* if the function $x \mapsto nx$ is surjective for every $n \geq 1$. (We write the group operation additively.) That is, every element of D is an "nth power" for any $n \geq 1$. For example, $\mathbf{R}/\mathbf{Z}$ and $\mathbf{Q}/\mathbf{Z}$ are divisible groups. Working multiplicatively, the group $\mathbf{C}^\times$ and its subgroups S^1 and μ_∞ (the complex roots of unity) are all divisible groups. The function $x \mapsto e^{2\pi i x}$ sets up isomorphisms $\mathbf{R}/\mathbf{Z} \cong S^1$ and $\mathbf{Q}/\mathbf{Z} \cong \mu_\infty$.

Using Zorn's lemma, we will show homomorphisms into a divisible group have an extension property. In this section the label "homomorphism" always means "group homomorphism." Our argument will be a prototype for the way we use Zorn's lemma in the next section to extend homomorphisms between fields, and the same method is how Zorn's lemma is used to prove the Hahn-Banach theorem in functional analysis.

Theorem 3.1. *Let D be a divisible group. If A is any abelian group and $B \subset A$ is a subgroup, any homomorphism $f\colon B \to D$ can be extended to a homomorphism $\widetilde{f}\colon A \to D$.*

Proof. Pick any $a \in A$ with $a \notin B$. Then the subgroup $\langle B, a\rangle = B + \mathbf{Z}a$ spanned by B and a contains B. As a warm-up we will show how to extend f to this larger subgroup of A. (See the diagram below.) Then we will bring in Zorn's lemma.



Consider how $\langle a\rangle$ can meet B. The set $\{k \in \mathbf{Z} : ka \in B\}$ is a subgroup of $\mathbf{Z}$, so it is 0 or $n\mathbf{Z}$ for some $n \geq 1$. If it is 0 then each element of $\langle B, a\rangle$ is $b + kx$ for unique $b \in B$ and $k \in \mathbf{Z}$. Define $f'\colon \langle B, a\rangle \to D$ by $f'(b + kx) = f(b)$. This is well-defined (why?) and it is a homomorphism with $f'|_B = f$. If instead $\{k \in \mathbf{Z} : ka \in B\} = n\mathbf{Z}$ for some $n \geq 1$ then

some positive multiple of a lies in B, and na is that multiple with n minimal. The function f makes sense at na, but not at a. If we can extend f to a homomorphism $f'\colon \langle B, a\rangle \to D$ then $f'(a)$ would have to satisfy the relation $nf'(a) = f'(na) = f(na)$. Here $f(na)$ is already defined. To pick a choice for $f'(a)$ we need to find $d \in D$ such that $nd = f(na)$. This can be done because D is divisible. Having chosen such a d, define $f'(a) = d$ and more generally

$$f'(b + ka) = f(b) + kd.$$

Is this well-defined? If $b + ka = b' + k'a$ then $(k - k')a = b' - b \in B$, so $k - k' \in n\mathbf{Z}$ by the definition of n. Write $k = k' + n\ell$ for some $\ell \in \mathbf{Z}$, so

$$\begin{aligned} f(b) + kd &= f(b) + (k' + n\ell)d \\ &= f(b) + k'd + \ell(nd) \\ &= f(b) + k'd + \ell f(na) \\ &= f(b + \ell na) + k'd. \end{aligned}$$

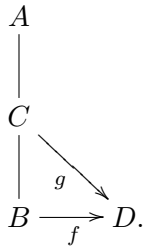
Since $b + \ell na = b + (k - k')a = b + ka - k'a = b' + k'a - k'a = b'$ we have

$$f(b) + kd = f(b') + k'd.$$

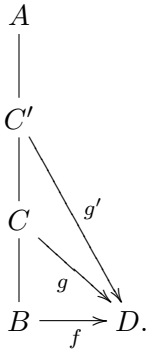
Thus $f'\colon \langle B, a\rangle \to D$ is well-defined. It is left to the reader to show it is a homomorphism.

Now we show Zorn's lemma can be applied. Example 1.7 is the paradigm.

Let S be the set of pairs (C, g) where C is a subgroup between B and A and $g\colon C \to D$ is a homomorphism that extends f (that is, $g|_B = f$). The picture is as follows.



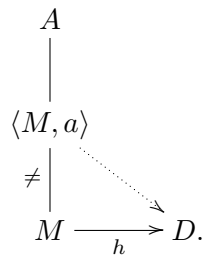
The set S is nonempty since $(B, f) \in S$. Partially order S by declaring $(C, g) \leq (C', g')$ if $C \subset C'$ and $g'|_C = g$. That is, g' extends g to the larger intermediate subgroup C'.



If $\{(C_i, g_i)\}_{i\in I}$ is a totally ordered subset of S, then it has an upper bound in S: use $C = \bigcup_{i\in I} C_i$ as the subgroup (it really is a subgroup, using an argument by now quite familiar from past setups for Zorn's lemma) and let $g\colon C \to D$ by $g(x) = g_i(x)$ if $x \in C_i$. Is this well-defined? Well, supposing x is in C_i and C_j, we need to know $g_i(x) = g_j(x)$.

Either $(C_i, g_i) \leq (C_j, g_j)$ or $(C_j, g_j) \leq (C_i, g_i)$ since the (C_i, g_i)'s are totally ordered in S. If $(C_i, g_i) \leq (C_j, g_j)$ then $C_i \subset C_j$ and $g_j|_{C_i} = g_i$, so $g_j(x) = g_i(x)$. That $g_i(x) = g_j(x)$ if $(C_j, g_j) \leq (C_i, g_i)$ is proved the same way. Because any two elements in C lie in a common C_i by Lemma 1.9, g is a homomorphism because each g_i is a homomorphism. Since $g_i|_B = f$ for all i, $g|_B = f$. Thus $(C, g) \in S$ and (C, g) is an upper bound on $\{(C_i, g_i)\}_{i \in I}$.

We have verified the hypotheses of Zorn's lemma, so S has a maximal element (M, h). That is, M is a group between B and A, $h\colon M \to D$ is a homomorphism and there is no extension of h to a homomorphism out of a larger subgroup of A than M. We will show $M = A$ by *contradiction*. If $M \neq A$ then there is some $a \in A$ with $a \notin M$. By the argument used at the start of this proof, with B and f there replaced by M and h here, h can be extended to a homomorphism $\langle M, a\rangle \to D$.



This contradicts the maximality of (M, h), so $M = A$. □

This application of Zorn's lemma generalizes from abelian groups (**Z**-modules) to modules over any commutative ring R, and is called Baer's characterization of injective R-modules. (An R-module is injective when it is a direct summand of any module it can be embedded in.) See [7, p. 396] or [16, p. 483]. Divisible abelian groups are injective **Z**-modules.

Let's use Zorn's lemma to do something crazy: show there is a "maximal" group. On the set of all groups, define the partial ordering by inclusion. This is a partial ordering. If $\{G_\alpha\}$ is a totally ordered set of groups, let G be the union of the G_α's. Any two elements in G are in a common G_α, so it is easy to define the group law in G (use the group law in G_α), check it is well-defined (independent of the G_α containing the two elements), and G contains all G_α's. So now it seems, by Zorn's lemma, that we should have a maximal group: a group that is not contained in any larger group. This is absurd, since for any group G we can create $G \times \mathbf{Z}$ and literally (if you wish) replace the elements of $G \times \{0\}$ with the elements of G to make G a geniune subset of $G \times \mathbf{Z}$. Then G is properly contained in another group and we have contradicted maximality. What is the error?

The problem here is right at the start when we defined our partially ordered set as the "set of all groups." This is the kind of set-theoretic looseness that leads to paradoxes (set of all sets, *etc.*). In fact, the group-theoretic aspect of the construction was irrelevant for the contradiction: if we just worked with sets and containment, the same argument goes through to show there is a set contained in no other sets, which is false and contains the same error as above: there is no "set of all sets" that one can partially order to make Zorn's lemma apply. Looking back now, perhaps with some suspicion, at the proofs where we created an upper bound on a totally ordered subset of a partially ordered set relative to some kind of inclusion relation, the objects that we formed the union of were always subsets of some larger fixed set (ideals in a ring, subgroups of a group). In that context the union really makes sense. It doesn't make sense to take arbitrary unions of arbitrary sets that *a*

priori don't live in some common set. As an extreme case, we can't take the union of "all sets" to find one set containing all others.

The following exercise applies Zorn's lemma as in Theorem 3.1 to another setting.

Exercise. Let K and F be fields. There need not be a homomorphism $K \to F$ (*e.g.*, $K = \mathbf{Q}$ and $F = \mathbf{F}_2$). But assume some subring of K admits a ring homomorphism to F (*e.g.*, if K has characteristic 0 then $\mathbf{Z} \subset K$ and there is certainly a ring homomorphism $\mathbf{Z} \to F$, while if K and F both have characteristic $p > 0$ then $\mathbf{Z}/(p)$ is a subfield of K and F so there is an inclusion homomorphism $\mathbf{Z}/(p) \hookrightarrow F$).

a) Let S be the set of pairs (A, f) where A is a subring of K and $f\colon A \to F$ is a ring homomorphism, so $S \neq \emptyset$ by hypothesis. Partially order S by $(A, f) \leq (B, g)$ if $A \subset B$ and $g|_A = f$. Show with Zorn's lemma that S contains a maximal pair, which amounts to a subring $A \subset K$ that admits a ring homomorphism $f\colon A \to F$ that can't be extended to a ring homomorphism out of any larger subring of K.

b) When $K = \mathbf{Q}$ and $F = \mathbf{Z}/(2)$, show the ring of fractions $\{m/n : m \in \mathbf{Z}, n \in \mathbf{Z} - \{0\}, n \text{ is odd}\}$ admits a homomorphism to $\mathbf{Z}/(2)$ and is the ring part of a maximal pair (A, f) in S. (Note $\mathbf{Q}$ itself is not part of a maximal pair since there is no ring homomorphism from $\mathbf{Q}$ to $\mathbf{Z}/(2)$.) What is the kernel of this homomorphism?

c) If (A, f) is a maximal pair in S, show every element of A not in $\ker f$ is a unit in A, and therefore $\ker f$ is the only maximal ideal in A. (A ring with only one maximal ideal is called a *local ring*. Trying to find a maximal subring of one field that admits a homomorphism to another field gives rise, using Zorn's lemma, to local rings.)

4. Applications to bases

We want to use Zorn's lemma to prove an arbitrary nonzero vector space has a basis. Let's first make sure we know what the label "basis" means when we are dealing with vector spaces that may turn out to be infinite-dimensional. For a nonzero vector space V over a field F, a *basis* of V is a subset $\mathcal{B}$ of V that is linearly independent (*i.e.*, no finite subset of $\mathcal{B}$ has a nontrivial F-linear relation) and spans V (*i.e.*, every element of V is an F-linear combination of finitely many elements of $\mathcal{B}$).

Notice the finiteness assumptions built into linear independence and spanning sets, even if the basis itself is infinite: linear independence involves a finite linear combination equal to 0, and spanning sets involve finitely many vectors at a time. In analysis one deals with infinite linear combinations once a suitable topology has been introduced on the vector space. Such a topological basis is not covered by our use of the label "basis" here. Our more algebraically-oriented concept involving only finite linear combinations is sometimes called a Hamel basis. For us, bases are used only through finite linear combinations.

Theorem 4.1. *Every nonzero vector space contains a basis.*

Proof. The idea is that a basis can be constructed as a maximal linearly independent set, and this maximal set will be found with Zorn's lemma.

Let V be a nonzero vector space and let S be the set of linearly independent sets in V. For instance, a single nonzero $v \in V$ is a linearly independent set, so $\{v\} \in S$. Thus $S \neq \emptyset$.

For two linearly independent sets L and L' in V, declare $L \leq L'$ if $L \subset L'$. This is the partial ordering on S by inclusion. It is easy to see that any subset of a linearly independent set is also a linearly independent set, so if $L \in S$ then any subset of L is also in S.

Assume $\{L_\alpha\}_{\alpha \in A}$ is a totally ordered subset of S. That is, every L_α is a linearly independent set in V and for any L_α and L_β in our subset we have $L_\alpha \subset L_\beta$ or $L_\beta \subset L_\alpha$. An

upper bound for the L_α's in S is the union $L = \bigcup_{\alpha \in A} L_\alpha$. Well, we need to check L is really a linearly independent set (so $L \in S$); once that is settled then L is an upper bound in S since $L_\alpha \subset L$ for all $\alpha \in A$.

Pick any finite set of vectors $v_1, \ldots, v_n$ in L. We must show they are linearly independent. Each v_k is in some L_α, say $v_1 \in L_{\alpha_1}, \ldots, v_n \in L_{\alpha_n}$. Since the L_α's are totally ordered, one of the sets $L_{\alpha_1}, \ldots, L_{\alpha_n}$ contains the others (Lemma 1.9). That means $v_1, \ldots, v_n$ are all in a common L_α, so they are linearly independent.

Zorn's lemma now tells us that S contains a maximal element: there is a linearly independent set $\mathcal{B}$ in V that is not contained in any larger linearly independent set in V. We will show $\mathcal{B}$ spans V, so it is a basis.

Let W be the span of $\mathcal{B}$. That means W is the set of all finite F-linear combinations $\sum_{i=1}^k c_i v_i$ with $k \geq 1$, $c_i \in F$, and $v_i \in \mathcal{B}$. If $\mathcal{B}$ does not span V then $W \neq V$, so we can pick $v \in V$ with $v \notin W$. Then $\mathcal{B}$ is a proper subset of $\mathcal{B} \cup \{v\}$. We will show $\mathcal{B} \cup \{v\}$ is linearly independent, which contradicts the maximality of $\mathcal{B}$ and thus proves $W = V$.

To prove $\mathcal{B} \cup \{v\}$ is linearly independent, assume otherwise: there is an expression

$$\sum_{i=1}^k c_i v_i = 0$$

where the coefficients are not all 0 and the v_i's are taken from $\mathcal{B} \cup \{v\}$. Since the elements of $\mathcal{B}$ are linearly independent, one of the v_i's with a nonzero coefficient must be v. We can re-index and suppose $v_k = v$, so $c_k \neq 0$. We must have $k \geq 2$, since otherwise $c_1 v = 0$, which is impossible since $v \neq 0$ and the coefficient of v is nonzero. Then

$$c_k v = -\sum_{i=1}^{k-1} c_i v_i.$$

Multiplying both sides by $1/c_k$,

$$v = \sum_{i=1}^{k-1} \left(-\frac{c_i}{c_k}\right) v_i,$$

which shows $v \in W$. But $v \notin W$, so $\mathcal{B} \cup \{v\}$ is a linearly independent set. □

Corollary 4.2. *Every linearly independent subset of a nonzero vector space V can be extended to a basis of V. In particular, every subspace W of V is a direct summand: $V = W \oplus U$ for some subspace U of V.*

Proof. Let $\mathcal{L}$ be a linearly independent subset of V. A basis of V containing $\mathcal{L}$ will be found as a maximal linearly independent subset containing $\mathcal{L}$.

Take S to be the set of linearly independent sets in V that contain $\mathcal{L}$. For instance, $\mathcal{L} \in S$, so $S \neq \emptyset$. The same argument as in the proof of Theorem 4.1 shows every totally ordered subset of S has an upper bound. (If the L_α's are linearly independent sets in V that each contain $\mathcal{L}$ then their union L also contains $\mathcal{L}$, and $L \in S$ because the L_α's are totally ordered, by a kind of argument we've made before.)

By Zorn's lemma there is a maximal element of S. This is a linearly independent set in V that contains $\mathcal{L}$ and is maximal with respect to inclusion among all linearly independent sets in S containing $\mathcal{L}$. The proof that a maximal element of S is a basis of V follows just as in the proof of Theorem 4.1.

To prove any subspace $W \subset V$ is a direct summand, let $\mathcal{L}$ be a basis of W. There is a basis $\mathcal{B}$ of V containing $\mathcal{L}$. Let U be the span of the complement $\mathcal{B} - \mathcal{L}$. It is left to the reader to show $V = W + U$ and $W \cap U = \{0\}$, so $V = W \oplus U$. □

Corollary 4.3. *Every spanning set of a nonzero vector space V contains a basis of V.*

Proof. Let $\mathcal{S}$ be a spanning set of V. Consider the set of linearly independent subsets of $\mathcal{S}$. This is a nonempty set, as $\{v\}$ is linearly independent for any $v \in \mathcal{S}$. Partially order the set of linearly independent subsets of $\mathcal{S}$ by inclusion. If $\{L_i\}$ is a totally ordered subset then $\bigcup_i L_i$ is a linearly independent subset of $\mathcal{S}$ and an upper bound on the L_i's. So by Zorn's lemma there is a maximal element $\mathcal{B}$: a linearly independent subset of $\mathcal{S}$ that is maximal with respect to inclusion. We will show $\mathcal{B}$ is a spanning set for V so it is a basis. Because $\mathcal{S}$ spans V, it is enough to show every element of $\mathcal{S}$ is in the span of $\mathcal{B}$ to know V is spanned by $\mathcal{B}$. If some $v \in \mathcal{S}$ is not in the span of $\mathcal{B}$ then $\mathcal{B} \cup \{v\}$ is a linearly independent set and it is a subset of $\mathcal{S}$ that strictly contains $\mathcal{B}$. This contradicts the maximality of $\mathcal{B}$ in $\mathcal{S}$. □

Remark 4.4. To find a basis of V inside a spanning set $\mathcal{S}$, a natural first idea might be to find a minimal spanning set of V inside of $\mathcal{S}$ rather than a maximal linearly independent subset. The minimality of a spanning set would force its linear independence and thus give us a basis. It is obvious how to use Zorn's lemma here: consider the set of all spanning sets of V inside $\mathcal{S}$, and partially order it by reverse inclusion. If $\{S_i\}$ is a totally ordered subset then the intersection $\bigcap_i S_i$ should be an upper bound on all the S_i's (we're using reverse inclusion, so an upper bound would be a spanning set contained in every S_i), and then Zorn's lemma gives us maximal elements, which will be minimal spanning sets. But there's a problem: how do you prove $\bigcap_i S_i$ is a spanning set? You can't; it's not generally true. For example, let $V = \mathbf{Q}$ as a $\mathbf{Q}$-vector space, enumerate the rationals as $\{r_1, r_2, r_3, \dots\}$ and let S_i equal $\mathbf{Q}$ with the first i rationals removed. Each S_i is a spanning set of $\mathbf{Q}$ as a $\mathbf{Q}$-vector space, and the S_i's are totally ordered by reverse inclusion, but their intersection is *empty*. This is an instructive case where it seems clear how Zorn's lemma should work, but it doesn't work!

Corollary 4.5. *Every nonzero module over a division ring has a basis, any linearly independent subset of the module can be extended to a basis, and every spanning set of the module contains a basis.*

Proof. We never used commutativity of the coefficient field in the proofs of Theorem 4.1 or its previous corollaries, except at the end of the proof of Theorem 4.1 where we wrote c_i/c_k. If we write this more carefully as $c_k^{-1}c_i$ then the proof goes through when the coefficient ring is a division ring: multiply through on the left by c_k^{-1} to solve for v as a linear combination of $v_1, \dots, v_{k-1}$. The proofs of Corollaries 4.2 and 4.3 also go through with a division ring as the coefficient ring. □

Over commutative rings that are not fields, modules need not be free. And even in a free module, a linearly independent subset need not extend to a basis and a submodule need not be a direct summand. For instance, the one-element linearly independent set $\{(4, 6)\}$ in $\mathbf{Z}^2$ can't be extended to a basis of $\mathbf{Z}^2$ and the submodule $(2\mathbf{Z})^2$ is not a direct summand of $\mathbf{Z}^2$. The reason our proofs for vector spaces over fields don't carry over to $\mathbf{Z}$-modules is that nonzero integers generally don't have inverses in $\mathbf{Z}$.

Here's an amusing corollary of the existence of bases in an arbitrary (especially infinite-dimensional) vector space.

Corollary 4.6. *If G is a group with more than two elements then G has a nontrivial automorphism.*

Proof. If G is nonabelian then some element of G is not in the center, so conjugation by that element is a nontrivial automorphism of G. If G is abelian then inversion (sending each element to its inverse) is an automorphism, and it is nontrivial unless every element is its own inverse. If G is abelian and every element is its own inverse then every element is killed by 2 ($x = -x \Rightarrow 2x = 0$), so G is a vector space over $\mathbf{Z}/(2)$. Let $\{e_i\}_{i\in I}$ be a basis of G over $\mathbf{Z}/(2)$. If there is more than one basis element, then exchanging two basis elements while fixing the rest extends to an automorphism of G. The only case remaining is a $\mathbf{Z}/(2)$-vector space with a basis of size at most 1. Such groups are trivial or cyclic of order 2, and their only automorphism is the identity. □

5. Equivalences with Zorn's lemma

We have used Zorn's lemma to prove existence theorems in algebra, but some of these results can be reversed: Zorn's lemma is equivalent to the existence of bases in arbitrary vector spaces [5, 12] and also to the existence of a maximal ideal in any nonzero commutative ring [1, 10]; actually, the existence of a maximal ideal in any unique factorization domain is enough to derive Zorn's lemma. Zorn's lemma is also equivalent to Theorem 3.1 [4].

In topology, Tychonoff's theorem in its general form (allowing compact non-Hausdorff spaces) is equivalent to Zorn's lemma [11], but Tychonoff's theorem for products of compact Hausdorff spaces [18] is strictly weaker than Zorn's lemma since it is consistent with Zermelo-Fraenkel set theory plus a strong negation of Zorn's lemma [8]. Tychonoff's theorem for Hausdorff spaces is equivalent to the existence of maximal ideals in all nonzero Boolean rings, which (by definition) are the rings satisfying $x^2 = x$ for all x. In functional analysis, the Hahn-Banach theorem is weaker than Zorn's lemma [15] while the Krein-Milman theorem is equivalent to it [3].

Surveys on equivalents to, and consequences of, Zorn's lemma are [9, 17].

Appendix A. Application to metric spaces

In a real vector space V, the line between vectors v and w is defined to be the set $\{tv+(1-t)w : 0 \leq t \leq 1\}$. A subset of V is called convex if it contains the line between any two points in the set. This notion of convexity, while very important in analysis, depends heavily on the real vector space structure: we used real scalars between 0 and 1 and also vector addition. There is a notion of convexity in arbitrary metric spaces, whose definition is based on the idea that the "line" between two points should contain only points in which the triangle inequality is an equality.

In a metric space (M, ρ), a subset S will be called *convex* if for any $x \neq y$ in S there is a $z \neq x, y$ in S such that $\rho(x, y) = \rho(x, z) + \rho(z, y)$. We circumvented the lack of a real vector space structure by not defining the line between x and y, but rather the points that ought to lie on any such "lines." (This definition of convex does not quite match the notion in Euclidean space: an *open* star-shaped region of $\mathbf{R}^n$ is not convex in the usual sense but is convex in the abstract sense above. However, for closed subsets of $\mathbf{R}^n$ with the metric induced from $\mathbf{R}^n$, the above notion of convex does match the usual meaning of the term.)

It may appear that our definition is quite weak: we only assume there is one such z. But by repeating the construction with x, y replaced by x, z, we can get more such points,

although we don't have much control over the actual distances we can achieve for points "between" x and y. Zorn's lemma will offer that control when we are in a complete space.

Theorem A.1. *Let (M, ρ) be a complete convex metric space. For distinct points x and y in M and $t \in [0, \rho(x, y)]$, there is a $z \in M$ such that $\rho(x, z) = t$ (and $\rho(x, y) = \rho(x, z) + \rho(z, y)$).*

Proof. We will use Zorn's lemma twice. Also, we will need to use the formulation of completeness in terms of nets, not sequences: any Cauchy net in a complete metric space converges.

First we define some notation. For $a, b \in M$, let

$$[a, b] := \{c \in M : \rho(a, b) = \rho(a, c) + \rho(c, b)\}.$$

For instance, this set contains a and b, and by hypothesis it contains a point besides a and b when $a \neq b$. Intuitively, $[a, b]$ is the set of points lying on geodesics from a to b. It is helpful when reading the following discussion to draw many pictures of line segments with points marked on them. Given t between 0 and $\rho(x, y)$, we will find a $z \in [x, y]$ with $\rho(x, z) = t$.

Some simple properties of these "intervals" are:

(1) $[a, b] = [b, a]$.
(2) If $c \in [a, b]$ and $b \in [a, c]$, then $\rho(b, c) = -\rho(b, c)$, so $b = c$.

Less simple properties are

(3) If $b \in [a, c]$ then $[a, b] \subset [a, c]$ and $[b, c] \subset [a, c]$.
(4) If $b \in [a, d]$ and $c \in [b, d]$ then $[a, c] \subset [a, d]$, $[b, d] \subset [a, d]$, and $[b, c] = [a, c] \cap [b, d] \subset [a, d]$. (We will only need that $[b, c]$ lies in the intersection, not equality.)

Proof of (3): Without loss of generality, we show $[a, b] \subset [a, c]$. For p in $[a, b]$,

$$\begin{aligned} \rho(a, c) &\leq \rho(a, p) + \rho(p, c) \\ &\leq \rho(a, p) + \rho(p, b) + \rho(b, c) \\ &= \rho(a, b) + \rho(b, c) \\ &= \rho(a, c). \end{aligned}$$

Therefore $p \in [a, c]$.

Proof of (4): By (3), $[b, d] \subset [a, d]$ and $[b, c] \subset [b, d]$. Therefore $c \in [a, d]$, so $[a, c] \subset [a, d]$, so

$$\begin{aligned} \rho(a, d) &= \rho(a, c) + \rho(c, d) \\ &\leq \rho(a, b) + \rho(b, c) + \rho(c, d) \\ &= \rho(a, b) + \rho(b, d) \\ &= \rho(a, d). \end{aligned}$$

Therefore the inequality is an equality, so $b \in [a, c]$, so $[b, c] \subset [a, c]$. Thus $[b, c] \subset [a, c] \cap [b, d]$.

For the reverse inclusion, let $p \in [a, c] \cap [b, d]$. Then

$$\begin{aligned} \rho(a, d) &= \rho(a, b) + \rho(b, d) \\ &= \rho(a, b) + \rho(b, c) + \rho(c, d) \\ &\leq \rho(a, b) + \rho(b, p) + \rho(p, c) + \rho(c, d) \\ &= \rho(a, b) + \rho(b, d) - \rho(p, d) + \rho(a, c) - \rho(a, p) + \rho(c, d) \\ &= 2\rho(a, d) - \rho(p, d) - \rho(a, p). \end{aligned}$$

Rearranging terms, $\rho(a,p)+\rho(p,d)\le\rho(a,d)$, so there is equality throughout, so $\rho(b,c)=\rho(b,p)+\rho(p,c)$. Thus $p\in[b,c]$.

Now we are ready to investigate "geodesics" on M coming out of x. Define a partial ordering on M that might be called "closer to x on geodesics" by

$$z_1\le z_2 \text{ if and only if } z_1\in[x,z_2].$$

In particular, $z_1\le z_2$ implies $\rho(x,z_1)\le\rho(x,z_2)$.

Let's check this is a partial ordering.

If $z_1\le z_2$ and $z_2\le z_1$, then $z_1\in[x,z_2]$ and $z_2\in[x,z_1]$, so $z_1=z_2$ by (2).

If $z_1\le z_2$ and $z_2\le z_3$ then $z_1\in[x,z_2]$ and $z_2\in[x,z_3]$. By (1), $z_2\in[z_3,x]$ and $z_1\in[z_2,x]$. Therefore by (4),

$$z_1\in[z_1,z_2]\subset[x,z_3],$$

hence $z_1\le z_3$.

Define

$$A=\{z\in[x,y]:\rho(x,z)\le t\}.$$

This set is nonempty, since it contains x. We will apply Zorn's Lemma to A with its induced partial ordering and show a maximal element of A has distance t from x.

Let $\{z_i\}_{i\in I}$ be a totally ordered subset of A. We want an upper bound. Let

$$s=\sup_{i\in I}\rho(x,z_i)\le t.$$

for any $\varepsilon>0$, there is some i_0 such that

$$s-\varepsilon\le\rho(x,z_{i_0})\le s,$$

so

$$s-\varepsilon\le\rho(x,z_i)\le s$$

for all $i\ge i_0$. For $i_0\le i\le j$, $s-\varepsilon\le\rho(x,z_i)\le s$ and

$$s-\varepsilon\le\rho(x,z_j)=\rho(x,z_i)+\rho(z_i,z_j)\le s$$

so $\rho(z_i,z_j)\le\varepsilon$. Thus $\{z_i\}$ is a Cauchy net, so has a limit ℓ by completeness of M. We show this limit is an upper bound in A.

Taking limits,

$$\rho(x,y)=\rho(x,z_i)+\rho(z_i,y)\Rightarrow\rho(x,y)=\rho(x,\ell)+\rho(\ell,y)$$

$$\rho(x,z_i)\le t\Rightarrow\rho(x,\ell)\le t.$$

Thus $\ell\in A$.

For $i\le j$,

$$\rho(x,z_j)=\rho(x,z_i)+\rho(z_i,z_j).$$

Taking limits over j,

$$\rho(x,\ell)=\rho(x,z_i)+\rho(z_i,\ell),$$

so $z_i\in[x,\ell]$, so $z_i\le\ell$ for all i.

We have justified an application of Zorn's Lemma to A. Let m be a maximal element. That is, $m\in A$, and if $z\in A$ with $m\in[x,z]$ then $z=m$.

Let $B=\{z\in[y,m]:\rho(y,z)\le\rho(x,y)-t\}$. Since $y\in B$, B is nonempty. Our goal is to show $m\in B$, which is *not* obvious. Note that the definition of B depends on the existence of a maximal element of A.

In B, introduce a partial ordering by $z_1\le z_2$ when $z_1\in[y,z_2]$.

As above, every totally ordered subset of B has an upper bound in B, so by Zorn's lemma B contains a maximal element, m'. Since $m \in [x,y]$ and $m' \in [m,y]$, we get by (4) that

$$[m,m'] \subset [x,m'] \cap [m,y] \subset [x,y].$$

For $z \in [m,m']$,

$$\begin{aligned} z \in [x,y] &\Rightarrow \rho(x,y) = \rho(x,z) + \rho(y,z) \\ &\Rightarrow \rho(x,z) \leq t \text{ or } \rho(y,z) \leq \rho(x,y) - t \\ &\Rightarrow z \in A \text{ or } z \in B. \end{aligned}$$

Also,

$$\begin{aligned} \rho(x,y) &= \rho(x,z) + \rho(z,y) \\ &\leq \rho(x,m) + \rho(m,z) + \rho(z,y) \\ &= \rho(x,m) + \rho(m,y) \text{ since } z \in [m,y] \\ &= \rho(x,y) \text{ since } m \in [x,y]. \end{aligned}$$

Therefore $\rho(x,z) = \rho(x,m) + \rho(m,z)$, so $m \in [x,z]$.

We now have

$$\begin{aligned} \rho(x,y) &= \rho(x,z) + \rho(z,y) \text{ since } z \in [x,y] \\ &\leq \rho(x,z) + \rho(z,m') + \rho(m',y) \\ &= \rho(x,m') + \rho(m',y) \text{ since } z \in [x,m'] \\ &= \rho(x,y) \text{ since } m' \in [x,y]. \end{aligned}$$

Therefore $\rho(y,z) = \rho(z,m') + \rho(m',y)$, so $m' \in [y,z]$.

Thus if $z \in A$ then $m \in [x,z] \Rightarrow z = m$. If $z \in B$, then $m' \in [y,z] \Rightarrow z = m'$. Therefore $[m,m'] = \{m,m'\}$, so by *convexity* of M, $m = m'$, hence $m \in A \cap B$. Therefore $\rho(x,m) \leq t$ and $\rho(y,m) \leq \rho(x,y) - t$, so

$$\rho(x,y) = \rho(x,m) + \rho(m,y) \leq \rho(x,y),$$

so $\rho(x,m) = t$. □

The z we constructed in the proof need not be unique. Consider M to be the sphere in $\mathbf{R}^3$ with its surface metric, x and y to be the north and south poles and take z to be any of the points on some chosen line of latitude. It is natural to expect that if $\rho(x,y)$ is small enough, the z in Theorem A.1 is unique, and this would let us construct *paths* in M. For a proof (without Zorn's lemma!) that any complete convex metric space is in fact path connected, see [6, Theorem 14.1, p. 41].

References

[1] B. Banaschewski, *A new proof that "Krull implies Zorn"*, Math. Logic Quart. **40** (1994), 478–480.

[2] J. L. Bell and F. Jellett, *On the relationship between the Boolean prime ideal theorem and two principles in functional analysis*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys. **19** (1971), 191–194.

[3] J. Bell and D. H. Fremlin, *A geometric form of the axiom of choice*, Fund. Math. **77** (1972), 167–170.

[4] A. Blass, *Injectivity, projectivity, and the axiom of choice*, Trans. Amer. Math. Soc. **255** (1979), 31–59.

[5] A. Blass, *Existence of bases implies the axiom of choice*, pp. 31–33 in: "Axiomatic set theory (Boulder, Colo., 1983)", Amer. Math. Soc., Providence, 1984.

[6] L. Blumenthal, "Theory and Applications of Distance Geometry," Clarendon Press, Oxford, 1953.

[7] D. Dummit and R. Foote, "Abstract Algebra," 3rd ed., Wiley, New York, 2004.

[8] J. D. Halpern and A. Lévy, *The Boolean prime ideal theorem does not imply the axiom of choice*, pp. 83–134 in: "Axiomatic Set Theory (Proc. Sympos. Pure Math., Vol. XIII, Part I), Amer. Math. Soc., Providence, 1971.
[9] H. Herrlich, "Axiom of Choice," Springer-Verlag, Berlin, 2006.
[10] W. Hodges, *Krull implies Zorn*, J. London Math. Soc. **19** (1979), 285–287.
[11] J. Kelley, *The Tychonoff product theorem implies the axiom of choice*, Fund. Math. **37** (1950), 75–76.
[12] K. Keremedis, *Bases for vector spaces over the two-element field and the axiom of choice*, Proc. Amer. Math. Soc. **124** (1996), 2527–2531.
[13] S. Lang, "Algebra," 3rd revised ed., Springer, New York, 2002.
[14] A. R. Naghipour, *A Simple Proof of Cohen's Theorem*, Amer. Math. Monthly **112** (2005), 825–826.
[15] D. Pincus, *The strength of the Hahn-Banach theorem*, pp. 203–248 in "Victoria Symposium on Nonstandard Analysis (Univ. Victoria, Victoria, B.C., 1972)," Springer–Verlag, Berlin, 1974.
[16] J. Rotman, "Advanced Modern Algebra," Prentice-Hall, Upper Saddle River, NJ, 2002.
[17] H. Rubin and J. Rubin, "Equivalents of the axiom of choice. II," North-Holland, Amsterdam, 1985.
[18] H. Rubin and D. Scott, *Some topological theorems equivalent to the Boolean prime ideal theorem*, Bull. Amer. Math. Soc. **60** (1954), 389.